EPITRANS: A Database that Integrates Epigenome and Transcriptome Data

Soo Young Cho^{2,3,4,7}, Jin Choul Chai^{1,7}, Soo Jun Park⁵, Hyemyung Seo¹, Chae-Bong Sohn⁶, and Young Seek Lee^{1,*}

Epigenetic modifications affect gene expression and thereby govern a wide range of biological processes such as differentiation, development and tumorigenesis. Recent initiatives to define genome-wide DNA methylation and histone modification profiles by microarray and sequencing methods have led to the construction of databases. These databases are repositories for international epigenetic consortiums or provide mining results from PubMed, but do not integrate the epigenetic information with gene expression changes. In order to overcome this limitation, we constructed EPITRANS, a novel database that visualizes the relationships between gene expression and epigenetic modifications. EPITRANS uses combined analysis of epigenetic modification and gene expression to search for cell function-related epigenetic and transcriptomic alterations (Freely available on the web at <http://epitrans.org>).

INTRODUCTION

Gene expression is regulated by various proteins and complex biological processes. Epigenetic modifications are expression regulators that govern a wide range of biological processes (Bird, 2007). Epigenomics is the study of all epigenetic modifications in DNA or histones in a cell.

Researchers have identified genome wide epigenetic modification profiles using microarray and sequencing methods. For example, the Human Epigenome Project (HEP) aims to identify genome-wide DNA methylation patterns in major tissues. The NIH Roadmap Epigenomics Mapping Consortium was started with the goal to provide a public resource of human epigenomic data in order to catalyze basic biology and disease-oriented research (Eckhardt et al., 2006).

Furthermore, secondary databases are being developed. For example, Human Histone Modification Database (HHMD) provides epigenetic alterations in various cancer types using ChIP-

seq and MeDIP-seq data (Zhang et al., 2010). MethylomeDB gives information on DNA methylation in the human brain (Xin et al., 2012). PubMeth and MeInfoText provide cancer-related DNA methylation by text mining (Fang et al., 2011; Ongenaert et al., 2008). DiseaseMeth and MethyCancer catalogue disease-related alterations in DNA methylation patterns using omics data and text mining methods (He et al., 2008; Lv et al., 2012). MethDB has been constructed to store cytosine methylation data (Grunau et al., 2001). NCBI epigenomics provides genome-wide maps of epigenetic factors from various epigenomic data sets and identifies significant differences for epigenetic profiles in selected samples (Fingerman et al., 2011). The Histone Database and ChromDB are protein sequence repositories displaying associations with chromatin structure (Gendler et al., 2008; Sullivan et al., 2002). However, these databases are simple epigenomics repositories, and do not compare different sets of epigenetic modifications in each promoter region.

To visualize the relationships between gene expression and epigenetic modifications, we integrated 364 microarray and 279 Next Generation Sequencing (NGS) data sets for 28 human tissues into a new database that we named EPITRANS. Only paired samples for epigenome (DNA methylation or histone modification) and transcriptome (microarray and RNA-seq) were selected. The associations between gene expression and epigenetic alterations were predicted by the Pearson correlation and comparison between differentially expressed genes and differentially modified epigenetic regions.

MATERIALS AND METHODS

Resource for database

EPITRANS focuses on gene expression changes associated with epigenetic modifications. We selected paired transcriptome and epigenome data sets generated from the same tissue or cell line. In total, 364 microarray data sets were downloaded from GEO, and 279 NGS data sets were downloaded from

¹Department of Molecular and Life Sciences, Hanyang University, Ansan 425-791, Korea, ²Laboratory of Developmental Biology and Genomics, College of Veterinary Medicine, Research Institute for Veterinary Science, Brain Korea 21 Program for Veterinary Science, ³Interdisciplinary Program for Bioinformatics, Program for Cancer Biology and BIO-MAX Institute, Seoul National University, Seoul 151-742, Korea, ⁴MRC Harwell, Mammalian Genetics Unit, Harwell Science and Innovation Campus, Oxfordshire, United Kingdom, ⁵Bio-Medical IT Convergence Research Department, ETRI, Daejeon 305-700, Korea, ⁶Department of Electronics and Communications Engineering, Kwangju University, Seoul 139-701, Korea, ⁷These authors contributed equally to this work.

*Correspondence: yslee@hanyang.ac.kr

Received September 5, 2013; accepted September 10, 2013; published online November 8, 2013

Keywords: database, epigenome, transcriptome

Table 1. Statistics of analyses data and curated results

	Microarray platform		NGS platform		Total
	DNA methylation and histone modification	mRNA	DNA methylation and histone modification	mRNA	
# of studies	10	10	2	2	12
# of samples	362	140	251	28	781
# of curated tissues	17	11	14	6	48
# of curated diseases	6	6	7	-	19

Fig. 1. Data flow chart; all omics data were downloaded from public resources. Epigenetic signals in the promoter and CDS regions are identified by read count. mRNA expression is calculated by the RPKM method.

ENCODE and the human reference epigenome mapping project (Chadwick, 2012; Consortium, 2011). All data were hand-curated, resulting in the annotations of 23 human tissue types (Table 1). Detail information for curated sample are provide help menu in EPITRANS.

Data analysis for NGS and microarray

An expression value of transcriptome and epigenome was calculated by the RPKM method (Mortazavi et al., 2008). Expression values (RPKM) in promoter (2.1 kb) and CDS region were calculated for each gene. Expression values were transformed to z-scores and correlation coefficients between gene expression and epigenetic modifications were calculated (Fig. 1).

The microarray data were analyzed by LIMMA package. In microarray data for epigenome and transcriptome, we identified expression signal for genes and specific region and predicted differentially expressed genes and differentially modified epigenetic regions.

Integration between epigenome and transcriptome

To predict epigenetic effects, we used an epigenetic code and calculated correlation coefficients between gene expression

and epigenetic modification. The epigenetic code is hypothesized to be a defining rule of the specific epigenetic modification, and is based on histone and DNA methylation. A gene expression change with epigenetic modification was identified by a p-value < 0.05 and correlation $\geq |0.4|$.

In 138 microarray data sets, we predicted differentially expressed genes (DEG) and differentially modified epigenetic regions (DMER) using the LIMMA package (Smyth, 2004). We identified epigenetic gene regulation using overlaps between DEG and DMER (p value < 0.05 and $\log_2 \geq |2|$). Activators were identified by up DEG and DMER. Repressors were identified by down DEG and DMER.

RESULTS AND DISCUSSION

Using genome-wide experimental data sets in various cell lines and tissues, EPITRANS integrates gene regulation patterns associated with epigenetic modifications. EPITRANS provides the following information.

Defining the epigenetic factors for gene expression

Every promoter region of a given gene accommodates a dis-

Fig. 2. Data contents in EPITRANS; users may search significant epigenetic modifications and gene expression by gene, tissue, and disease name. The gene regulation with epigenetic modification is visualized by correlation and body map.

tinct set of modifications. To compare different sets of epigenetic modifications of all genes, EPITRANS links global (across 23 tissues) gene expression changes with the corresponding tissue (Fig. 2, Data contents in EPITRANS); users may search significant epigenetic modifications and gene expression by gene, tissue, and disease name. The gene regulation with epigenetic modification is visualized by correlation and body map-specific epigenetic modifications and displays the epigenetic signals in the promoter and CDS regions. Thus, by integrating ChIP-seq (epigenetic markers) with RNA-seq (mRNA expression) data of 23 human tissues, we constructed global (across paired samples between epigenome and transcriptome) gene regulation profiles. The gene expression pattern associated with an epigenetic modification was visualized by histogram and dot plot.

In EPITRANS, users can find relationships between epigenetic factors and gene expression and compare the effects of different epigenetic factors in the same gene. Using P value and correlation coefficients, significant global gene regulation with epigenetic factors may be identified.

Gene expression change by epigenetic factors in human tissues

Five hundred and four experiments for 23 tissues containing more than three replicated samples of the same tissue were selected. We identified significantly regulated genes in each tissue. Using that information, we constructed a human body map of gene expression changes according to the type of epigenetic modification. In the body map, the orange-labeled tissue name indicates that the gene in those tissues is regulated by some epigenetic factors. The gray label means that epigenetic gene regulation was not identified or omics data were not collected. The body map gives an overview of gene expression pattern that is regulated by epigenetic change in human tissues.

The gray label means that epigenetic gene regulation was not identified or omics data were not collected.

Additional information for epigenetic alteration

Omics approaches help to identify candidate genes or biological markers in disease or other cellular processes. Also, many papers provide useful information about epigenetic gene regulation. To construct a portal site of epigenomics, we added other epigenetic resources (DiseaseMeth, HHMD, MethyCancer, MethylomeDB and PubMeth). Users may search for epigenetic alterations in specific diseases and associate significant regulated genes with epigenetic modifications.

We integrated data on human mRNA expression patterns and epigenetic modifications. In EPITRANS, users can find relationships between epigenetic factors and gene expression and compare the effects of different epigenetic factors on a given gene. In a case study, FAIM expression was regulated by hyper-methylation of H3K79 and shows positive correlation with H3K79ME2 in EPITRANS (Krivtsov et al., 2008). Another case study reported that CDKN2A expression was down-regulated by H3K9ME in colon and gastric cancer (Kondo et al., 2003; Meng et al., 2007). EPITRANS shows negative correlation between CDKN2A and H3K9ME and provides comprehensive information on epigenetic modifications. With the development of NGS technologies, the amount of publicly available epigenomic data has been increasing rapidly. We plan to update the data from TCGA and other public resources regularly. EPITRANS will provide useful information for epigenetic target gene prediction and marker selection. We also expect that EPITRANS will help to identify novel classes of epigenetic regulators.

ACKNOWLEDGMENTS

This work was supported by a National Research Foundation of

Korea (NRF) grant funded by the Korean government [Ministry of Science, ICT & Future Planning (MSIP)] (No. 2011-0001395 and 2011-0030775).

REFERENCES

- Bird, A. (2007). Perceptions of epigenetics. *Nature* 447, 396-398.
- Chadwick, L.H. (2012). The NIH roadmap epigenomics program data resource. *Epigenomics* 4, 317-324.
- Consortium, T.E.P. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9, e1001046.
- Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* 38, 1378-1385.
- Fang, Y.C., Lai, P.T., Dai, H.J., and Hsu, W.L. (2011). MelInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics* 12, 471.
- Fingerman, I.M., McDaniel, L., Zhang, X., Ratzat, W., Hassan, T., Jiang, Z., Cohen, R.F., and Schuler, G.D. (2011). NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res.* 39, D908-912.
- Gendler, K., Paulsen, T., and Napoli, C. (2008). ChromDB: the chromatin database. *Nucleic Acids Res.* 36, D298-302.
- Grunau, C., Renault, E., Rosenthal, A., and Roizes, G. (2001). MethDB--a public database for DNA methylation data. *Nucleic Acids Res.* 29, 270-274.
- He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., Yang, L., Sun, Z.S., Yang, H., and Wang, J. (2008). MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.* 36, D836-841.
- Kondo, Y., Shen, L., and Issa, J.P. (2003). Critical role of histone methylation in tumor suppressor gene silencing in colorectal cancer. *Mol. Cell Biol.* 23, 206-215.
- Krivtsov, A.V., Feng, Z., Lemieux, M.E., Faber, J., Vempati, S., Sinha, A.U., Xia, X., Jesneck, J., Bracken, A.P., Silverman, L.B., et al. (2008). H3K79 methylation profiles define murine and human MLL-AF4 leukemias. *Cancer Cell* 14, 355-368.
- Lv, J., Liu, H., Su, J., Wu, X., Li, B., Xiao, X., Wang, F., Wu, Q., and Zhang, Y. (2012). DiseaseMeth: a human disease methylation database. *Nucleic Acids Res.* 40, D1030-1035.
- Meng, C.F., Zhu, X.J., Peng, G., and Dai, D.Q. (2007). Re-expression of methylation-induced tumor suppressor gene silencing is associated with the state of histone modification in gastric cancer cell lines. *World J. Gastroenterol.* 13, 6166-6171.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621-628.
- Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S., and Van Criekinge, W. (2008). PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.* 36, D842-846.
- Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, Article3.
- Sullivan, S., Sink, D.W., Trout, K.L., Makalowska, I., Taylor, P.M., Baxeavanis, A.D., and Landsman, D. (2002). The histone database. *Nucleic Acids Res.* 30, 341-342.
- Xin, Y., Chanrion, B., O'Donnell, A.H., Milekic, M., Costa, R., Ge, Y., and Haghighi, F.G. (2012). MethyloMeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res.* 40, D1245-1249.
- Zhang, Y., Lv, J., Liu, H., Zhu, J., Su, J., Wu, Q., Qi, Y., Wang, F., and Li, X. (2010). HHMD: the human histone modification database. *Nucleic Acids Res.* 38, D149-154.