# RNA-seq Coverage Effects on Biological Pathways and GO Tag Clouds

Chien-Ming Chen[1], Tsan-Huang Shih[1], Tun-Wen Pai[1,2*], Zhen-Long Liu[1], Margaret Dah-Tsyr Chang[3]

[1]Dept. of Computer Science and Engineering, [2]Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan

[3] Dept. of Medical Science, National Tsing Hua University, Hsinchu, Taiwan

*twp@mail.ntou.edu.tw

*Abstract* — **RNA-seq data analysis not only detects novel transcripts, promoters, and single nucleotide polymorphisms in a transcriptome scale, but also shows quantitative measurement of gene expression. In order to perform differential expression analysis for unraveling biological functions, we proposed a workflow which integrated annotations from KEGG biological pathways and Gene Ontology associations for manipulating multiple RNA-seq datasets. The developed system started from mapping short reads onto reference genes, and then performed normalization procedures on read coverage to evaluate and compare expression levels within various gene clusters. Different levels of gene expression were indicated by diverse color shades and graphically shown in designed temporal pathways. Besides, representative GO terms associated with differentially expressed gene cluster were also visually displayed by a GO tag cloud representation. Three different public RNA-seq datasets were applied to demonstrate that the proposed workflow could provide effective and efficient analysis on differential gene expression for either cross-strain comparison or an identical sample sequenced at different time points.**

***Keywords - Gene Ontology, temporal pathway, RNA-seq, cross-strain comparison.***

## I. INTRODUCTION

The technologies of high-throughput sequencing (next generation sequencing, NGS) exploited dynamic complementary DNA sequencing in an approach termed high-throughput RNA sequencing (RNA-seq) [1]. The advantages of RNA-seq technology compared with tiling microarray and EST sequencing could be summarized in raising quality, cutting down experimental time and cost-efficient scale [2]. Therefore, it is overwhelmingly adopted in recent years for transcriptome analysis. Up to now, deep sequencing researches were employed widely in complex disease gene expression such as cancer studies, quantitative analysis of transcript expression such as organism diversity and evolution, antisense transcriptome analysis, and discovering of novel isoforms [3-5]. Another major advantage of RNA-seq is the ability of quantitative measurement of each expressed element at transcriptome scale, which facilitates researchers to discover differential gene expression under various circumstances [6]. With rapid growing of bioinformatics algorithms and biotechnologies, correlative RNA-seq researches and applications are getting more concerned in recent years [7]. Typical RNA-seq experiment generates a large number of short reads for transcriptome analysis, and these reads could be mapped/aligned to expressed genes by reference mapping tools [8]. The expression level of each gene could be determined according to the number of times a nucleotide being read within a gene during the sequencing process. However, most analyses still focused on evaluating the existence of a specific gene, or a small set of genes related to a selected function at a time. Therefore, some important associated information might be ignored due to limited analytical scale or non-quantitative measurements of gene analysis. In order to comprehensively analyze differentiated gene expression from various transcriptome datasets, transformed profiles from associated gene expression accounts individually into function-orientated gene cluster becomes an important, macro-based and systematic approach.

Two functional annotation methods for clustered gene groups including biological pathways and GO terms were employed in this study. A biological pathway is one of the most meaningful clustering representations for biological function analysis. It represents a consequent chain of chemical reactions catalyzed by cells, enzymes, or ligands. Each pathway includes a signal transduction starting with a signal to another receptors and ending with changes in cellular behaviors [9]. The expression level of each gene within a regulatory network was usually different in distinct organs and tissues. It was also dynamically changed due to various environmental conditions, different disease stages, or distinct phases within a cell cycle. By integrating coverage information of RNA-seq reads within biological pathways, differential gene expressions among different strains or sequencing time points could reveal dynamic status in biological function within a gene cluster.

Another functional annotation method often applied to describe gene products is Gene Ontology (GO). The GO is a set of structured vocabularies defined by Gene Ontology Consortium [10], which is aimed to provide a universal standard of functional annotation for gene products. All terms in GO are connected with each other by directed acyclic graphs with hierarchy relationship. Each term belongs to one of the three independent ontologies: biological process (BP), molecular function (MF) and cellular component (CC), which respectively represent different aspects of gene in temporal,


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

functional and spatial domains. Today, the GO is frequently used as a *de facto* standard of gene annotation, and various studies have shown that the GO terms can provide conserved function information in a group of genes through over representation analysis [11,12]. Several existing tools based on GO approach provided transmission from expressed data to gene annotation. For example, GOMiner [13] is a tool for analyzing microarray through GO properties to identify specific functions with gene-by-gene approach. DAVID [14] is an integrated functional analyzer to annotate and categorize gene functions from gene/protein identifier lists. They mapped those identifiers to common GO terms or gene-interaction maps through bioinformatics resources. These tools performed well in known gene function and gene network analyses. Especially, GO plays an important role in function annotation and categorization of unidentified and unannotated sequences.

Through previously discovered functional features, the sequenced RNA-seq could be transformed from quantitatively measured coverage rates to gene expression levels as a global view of biological system responses. In this study, we have focused on the evaluation of dynamic expression of specific gene clusters among different conditions. The differential gene expression levels among various RNA-seq datasets regarding a mapped pathway or a GO term would be statistically analyzed and graphically shown by a novel representation for on-line users.

## II. Materials and Methods

### A. System Flowchart

To exploit RNA-seq accounts associated with differential gene expression under different circumstances, more than single RNA-seq experimental results should be input for comparison. The sequenced reads for transcriptome profiling might be obtained from different tissues, different strains, or under various environments. The RNA-seq datasets could also be obtained at different time points such as various embryonic stages or a few hours after drug treatment. An analytical system shown in Figure 1 was designed to reveal quantitative coverage rates of RNA-seq data and systematic changes through novel visualization approaches. The proposed system includes 4 major phases: reference mapping, coverage rate counting and normalization, functional gene pathway mapping, and GO tag cloud visualization. At first, the reads from multiple sequenced RNA-seq datasets were mapped to known reference genes by any reference mapping tool. A reference mapping tool provided details of how each read mapped to the known coding regions in a selected target species. According to the mapped results, the initiative coverage rates could be calculated for each expressed gene. Next, coverage rates were normalized to balance the experimental results under different conditions in order to eliminate bias caused by different output length in total. The derived scores from normalization procedures were used as corresponding expression level for each gene. After successful retrieval of all gene expression scores, active genes would be identified and assigned to biological pathways to dynamically display their expression differences incorporated with biological functions. Theses mapped genes possessing most differential effectiveness among all experiments could be selected and clustered automatically according to KEGG pathway annotations. Besides, the GO term over representation analysis was also applied for comparing different gene groups with differential gene expression, and the results of major variations were displayed by a tag cloud technique. This visualization approach could facilitate users in a way of clear and intuitive recognition of dynamic status of molecular function regarding the differences of function level among various RNA-seq datasets.

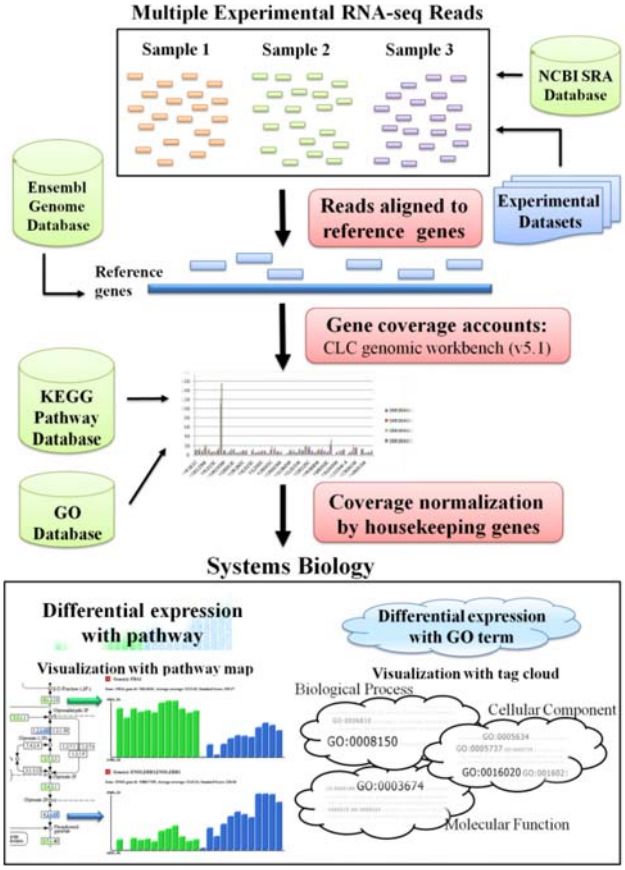


Figure 1. A system flowchart for analyzing multiple sets of RNA-Seq accounts.

### B. Reference Mapping and Normalization

The first step mapped RNA-seq reads to a set of reference gene list in advance. The referred gene sequences should contain UTR and coding regions only, but not intron segments for preventing coverage bias. In this study, we selected Ensembl database as the resource of primary gene datasets, which provided not only detailed coordinate and annotation information, but also corresponding GO information for most of collected genes [15]. There were several reference mapping tools available from either commercial software or open-source projects [16-18]. Most of these tools could generate mapped results with SAM/BAM formats, and provided information of how reads mapped to the references. The average coverage rate or depth for each gene could be counted according to the number of accumulated times at each nucleotide position. Since the coverage rates were obtained from NGS reads directly, coverage between different experiments should be normalized to prevent bias caused by the throughput deviation from each individual NGS run [19]. According to previous reports, one of the normalization methods based on housekeeping genes performed better in the benchmark than simply utilized the total reads from each experiment [20]. If a


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

set of stable housekeeping genes was available for a specified species, the average coverage rate among all housekeeping genes could be used as a referencing factor for normalization processes. All gene coverage rates were then multiplied by the scaling parameter linearly for read coverage normalization..

### C. Biological Pathway

Here we adopted Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database as the fundamental resources for differential gene expression analysis among RNA-seq datasets [21]. In each KEGG pathway, a rectangular component represents a set of genes or enzymes; a cyclic component represents compounds; a linkage regulates genes in corresponding metabolic reactions. According to the distribution of expression levels normalized from RNA-seq coverage, the proposed system colored different levels of variation in each component for visualizing distinct gene expression. There were 10 levels in variation scores for each gene in a pathway map. The level of quantitatively changed value for each functional gene was normalized to $l_i$ according to the following equation, where $S_i$ was the value of standard deviation for the $i^{th}$ node, $S_{max}$ was the maximum standard deviation in a specified KEGG map, and the *ZC* counts were denoted as the number of zero-crossings appeared as the sign change of slope value in each specific gene for various RNA-seq datasets. To obtain *ZC*, the sign of slope value was extracted between two continuous gene expression levels first, and then the total number of sign change was counted to represent a *ZC* value.

$$l_i = S_i \times 10 / S_{max}$$

Hence, all gene boxes within a pathway directly show their dynamic change conditions of gene expression according to sequenced datasets, and a gene node would be filled with red rectangle boxes when its corresponding *ZC* value was greater than 1. Different red layers were selected according to the previously defined level $l_i$. When the *ZC* of coverage rate distribution within a gene node was equal to 1 from multiple datasets, the trend of gene expression would be further examined and see if it was strictly increased or decreased. For the growing up condition, the gene box was depicted by a blue triangle, while growing down situation represented by a green triangle.

### D. Gene Ontology

In order to present functions related with differential expression gene set, we employed GO over representation analysis to select the representative GO terms automatically. Since GO is a hierarchy structured term system, and annotation for each child term could be considered the inheritance from its parents. Hence, the terms located at higher level always possess more common annotated chances for each gene [10]. Normalization for each GO term based on the appearance in the whole genome set has been applied in order to omit the bias from a hierarchical structure of term-level. In GO, annotations for each gene marked with the 3-character evidence code, which indicated the type of evidence supporting the annotation. For example, the evidence code IDA represented the annotation supported by Inferred from Direct Assay, and the ISS for Inferred from Sequence or Structural Similarity. Previous study indicated that evidence code could significantly influence the accuracies in a GO-based classifier, and suggested to use the computationally predicted annotations with caution [22]. Therefore, the GO analysis was frequently computed in separation based on different evidence code groups. Five evidence code groups according to GO consortium including *electronic*, *experimental*, *computational analysis*, *author statement,* and *curatorial statement* were applied and categorized in the proposed system.

To efficiently identify important GO terms from dynamic changes among RNA-seq datasets, the system developed a novel tag cloud visualization method for GO variations. To generate a tag cloud from identified differential gene expression of GO terms, the system assigned size weighting coefficients for different GO terms from the mapped gene set. The size of a GO term entry in a tag cloud indicated quantitative changes among multiple RNA-seq experiments. Therefore, a linear accumulation formula was applied for weighting coefficient assignment. This formula simply counted the differences of coverage rate of a specific gene among multiple experimental sets of short reads. According to the definitions, if an identified GO term possessed dramatic changes in gene expression levels, these terms would be defined with a higher weighting coefficient, and the text size of the identified term entry in the tag cloud would be drawn according to their weighting values. Here, the variant weighting scores were initially normalized into 10 different levels according to the average distribution. Larger GO terms shown in the tag cloud graph represented the GO term containing genes with higher gene differential expression.

## III. RESULT

In this study, we employed a few experimental RNA-seq datasets to evaluate our proposed system. The three query datasets were collected from NCBI SRA database including: “SRP002237”, “SRP005380-DatasetN1”, and “SRP005380-DatasetN2” [23]. The first SRP002237 dataset included 24 sets of RNA-seq experiments and these cDNA datasets were sequenced for a study of natural selection on *cis*- and *trans*-regulation in yeast [24] . Of which 12 datasets were obtained from co-culture yeast (run: SRR039256~SRR039267) which originated from two strains of *S. cerevisiae* yeast, another 12 datasets were obtained from hybrid yeast (run: SRR039244~SRR039255) which originated from their hybrid offspring ($F_1$ hybrids). The second dataset of “SRP005380-DatasetN1” contained 4 datasets sequenced at four different time points: 0, 1, 2 and 3 hour in meiosis processes, and the third dataset of “SRP005380-DatasetN2” included only 2 datasets sequenced at two different time points: 0 and 4 hour. The latter two datasets were used to analyze meiotic diploid of *S. cerevisiae* temporally [25]. Calculation of average coverage rates of all yeast genes from these three RNA-seq datasets were shown in Table 1. All selected RNA-seq reads were produced by Illumina high-throughput sequencing technologies.

The first step of analytical pipeline mapped short reads to the *S. cerevisiae* genes. Here, we adopted a reference mapping tool, CLC Genomic workbench (version 5.1), to obtain aligned short reads on reference genes [16]. To successfully utilize the data from SRA website, the “fastq-dump” program from SRA toolkit was executed to obtain FASTQ sequences. Next, the extracted FASTQ reads were imported into the system by removing failed reads. After all reads were imported, a “Map


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Read to References" toolkit from CLC genomic workbench was performed for reference mapping. The resulting data produced by mapping tool contained the information of coverage rate for each aligned gene. It should be noted that the reference mapping tools in this step is not necessarily a commercial software, and it could be substituted by any open-source reference mapping tools such as Bowtie or SOAP [17, 18].

TABLE I. AVERAGE COVERAGE RATE IN DATASETS

| Datasets | Average coverage rates |
|---|---|
| ***SRP002237*** | |
| SRR039244~55 (Hybrid) | 24.83 |
| SRR039256~67 (Co-culture) | 25.18 |
| ***SRP005380 Dataset-1*** | |
| SRR094602_0hr | 10.22 |
| SRR094603_1hr | 8.94 |
| SRR094604_2hr | 12.71 |
| SRR094605_3hr | 13.05 |
| ***SRP005380 Dataset-2*** | |
| SRR094606_0hr | 24.29 |
| SRR094607_4hr | 28.79 |

In the next phase, statistical analysis was performed for average coverage rate of each gene, and the normalization procedures were carried out by featuring a housekeeping gene list. Here we selected 14 genes from the *S. cerevisiae* housekeeping gene TAFs family [26]. Next, according to the expression levels from the selected housekeeping genes, previously defined biological pathways from KEGG dataset [27] and GO term association were automatically evaluated. All orthologous genes within a gene node from an identified pathway were individually annotated with normalized expression level among various RNA-seq datasets. Accordingly, these analyzed gene expression levels of all mapped genes among various datasets were visually displayed through temporal/cross-strains pathway maps and GO tag cloud representation.

For the SRP002237 dataset, there were in total 95 yeast KEGG pathway maps identified and retrieved after gene clustering procedures. The depth of coverage variation in RNA-seq for each gene in an identified pathway map was color coded for transforming gene expression quantities into systematic visual representation. For example, the pheromone signal transfer pathway in the MAPK signaling pathway (Map ID: 04011) was shown in Figure 2. The trend of average coverage rate for each gene among co-culture and hybrid datasets were calculated, and details of individual gene expression were statistically shown after clicking on the color coded gene boxes. In the statistical plot of expression levels, the green bars represented gene expression levels for the first co-cultural experimental RNA-seq, and the blue bars showed the expression quantities for the second hybrid experiments. To easily recognize the trend of differential feature of gene expression, an ascending blue triangle within a gene box represented the depth of coverage being increased from the first experiment to the second one; a descending green triangle within a gene box denoted the gene expression levels being decreased in an opposite trend. Figure 2 showed that the average coverage rate of the Ste2 gene in the co-culture experiment was 2083.16 and decreased to an average of 14.08 for the hybrid experiment. Oppositely, the average coverage of Mcm1 gene in co-culture experiment was 94.31 and increased to an average of 126.94 for hybrid experiment. With the information of coverage rate variations between different generations, the developed system could imply differential gene expression in a specific biological pathway, which could provide useful information for selecting appropriate genes for various applications such as *cis*- and *trans*- changes in regulatory evolution of genes.

Regarding the same datasets, most GO term variations in cellular component category were shown in Figure 3 by a tag cloud visualization approach. The larger size symbol of a GO term represented its corresponding genes possessing higher coverage variation rates among different RNA-seq datasets. Here, relatively high RNA-seq coverage variations of the top 4 GO terms were shown in Figure 4 including nuclear matrix (GO:0016363), eukaryotic translation elongation factor 1 complex (GO:0005853), actin cytoskeleton (GO:0015629), and 3-isopropylmalate dehydratase complex (GO:0009316). Users could click on any GO term on the text cloud to visualize the coverage rates among different RNA-seq datasets. Corresponding RNA-seq variations of these top ranked 4 GO terms were shown in Figure 4. From this example, the gene expression levels for the co-culture yeast genes at the GO terms of "histone deacetylase complex", "Set3 complex", and "cytosolic small ribosomal subunit" were significantly higher than hybrid yeast by observing bar chart distributions in Figure 4 (a), (c), and (d). Reversely, the hybrid yeast gene at GO term of "eukaryotic translation elongation factor 1 complex" was significantly higher than co-culture yeast according to bar charts in Figure 4(b).

The other two yeast RNA-seq datasets of "SRP005380-DatasetN1" and "SRP005380-DatasetN2" were applied for temporal pathway analysis. In these two testing cases, there were also 95 yeast KEGG pathways identified and retrieved through gene mapping and identification processes. The data visualization method was exactly the same as described in the previous case. The "SRP005380_N1" dataset contained 4 RNA-seq datasets which were sequenced at each hour, and "SRP005380_N1" contained only two datasets which were sequenced at two time points with 4 hour difference. From the comparison results of both datasets, several mapped pathways provided differential gene expression at significant levels. For example, both datasets reflected higher differential gene expression rates in meiotic pathway map (ID:04113). Figure 5(a) represented the meiosis yeast pathway map for SRP005380_N1 and Figure 5(b) for SRP005380_N2. It was observed that gene of Mej1 possessed common status of decreased gene expression within these two datasets and gene of Glc7 for increased conditions simultaneously. In addition to the temporal pathway analysis for these two RNA-seq datasets, associated GO term analysis was also performed. The coverage accounts of different time points for each gene and its associated GO terms were accumulated and compared for temporal GO term variation analysis. For example, the GO term of "GO:0005737" (cytoplasm), "GO:0016020" (membrane), "GO: 0005634" (nucleus) at Cellular Component (CC) level revealed with higher gene expression variations than other GO terms.


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

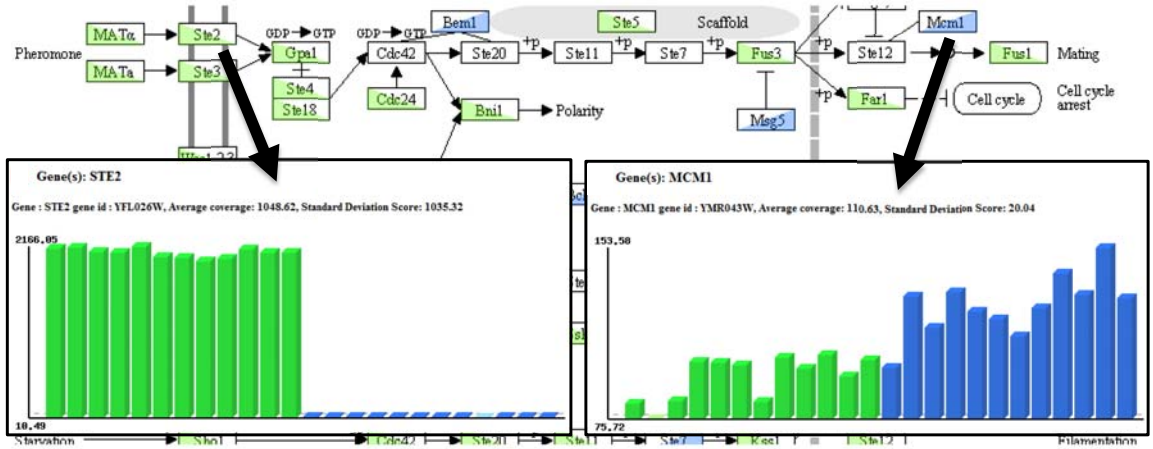


Figure 2. Variations of gene coverage rate in MAPK signaling pathway (Map ID:04011) from SRP002237 RNA-seq datasets. System responded the comparative results between two different experimental datasets.


Figure 3. GO term variations associated with cellular component (CC) in electronic evidence codes for SRP002237. The differences of average coverage rate between two experiments were more than 100 units, and the variations were normalized to show in tag cloud representation.

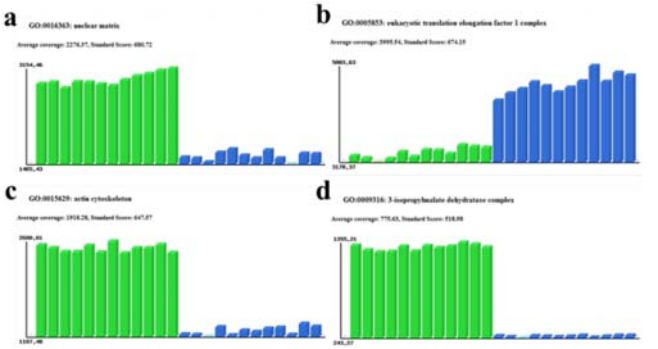


Figure 4. Top 4 variation of GO terms with CC in IEA of SRP002237: (a) GO:0016363, (b) GO:0005853, (c) GO:0015629, and (d) GO:0009316.

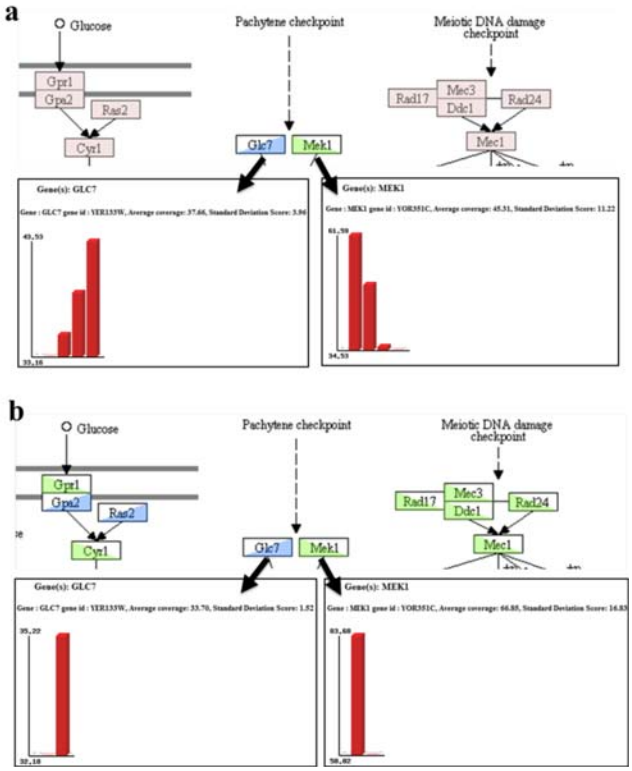


Figure 5. Meiosis yeast pathway maps (MAP:04113) with gene expression indication for (a) SRP005380_N1 and (b) SRP005380_N2. Mek1 gene for decreasing gene expression and GLC7 gene for increasing status in both RNA-seq datasets of Meiosis experiments.

## IV. CONCLUSION

In this study, we have proposed a workflow of differential gene expression for cross-strains or temporally separated RNA-seq datasets. The system mapped short reads to reference genes and measured gene expression level quantitatively through a normalized procedure. Accordingly, the KEGG pathway database was integrated for selecting a group of functional associated genes and to display various levels of differential gene expression regarding biological functional variation. The tag cloud representation for GO annotation with selectable evidence code consideration was also applied for visualizing functional conservation within highly dynamic expression genes. We employed public available RNA-seq reads as testing datasets to demonstrate the workflow could clearly indicate how differential gene expressions were connected to the biological function levels. This workflow can be applied under various experiment conditions invoked with different gene expression, and it is useful for further detailed experiment design.

## ACKNOWLEDGEMENT

This work is supported by the Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University and National Science Council, Taiwan, R.O.C. (NSC 101-2321-B-019-001 and NSC 100-2627-B-019-006 to T.-W. Pai)

## REFERENCES

[1] S. Marguerat and J. Bahler, "RNA-seq: from technology to biology," Cell Mol Life Sci, vol. 67, pp. 569-79, Feb 2010.

[2] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," Nat Rev Genet, vol. 10, pp. 57-63, Jan 2009.

[3] D. J. Sugarbaker, W. G. Richards, G. J. Gordon, *et al.*, "Transcriptome sequencing of malignant pleural mesothelioma tumors," Proc Natl Acad Sci U S A, vol. 105, pp. 3521-6, Mar 4 2008.

[4] C. A. Maher, C. Kumar-Sinha, X. Cao, *et al.*, "Transcriptome sequencing to detect gene fusions in cancer," Nature, vol. 458, pp. 97-101, Mar 5 2009.

[5] Q. Zhao, O. L. Caballero, S. Levy, *et al.*, "Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line," Proc Natl Acad Sci U S A, vol. 106, pp. 1886-91, Feb 10 2009.

[6] B. T. Wilhelm and J.-R. Landry, "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing," Methods, vol. 48, pp. 249-57, 2009.

[7] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," Nat Methods, vol. 8, pp. 469-77, Jun 2011.

[8] J. Shendure and H. Ji, "Next-generation DNA sequencing," Nat Biotechnol, vol. 26, pp. 1135-45, Oct 2008.

[9] L. T. Macneil and A. J. Walhout, "Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression," Genome Res, Mar 30 2011.

[10] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nat Genet, vol. 25, pp. 25-9, May 2000.

[11] T. Beissbarth and T. P. Speed, "GOstat: find statistically overrepresented Gene Ontologies within a group of genes," Bioinformatics, vol. 20, pp. 1464-5, Jun 12 2004.

[12] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson, "Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration," Bioinformatics, vol. 24, pp. 1650-1, Jul 15 2008.

[13] B. R. Zeeberg, W. Feng, G. Wang, *et al.*, "GoMiner: a resource for biological interpretation of genomic and proteomic data," Genome Biol, vol. 4, p. R28, 2003.

[14] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," Nat Protoc, vol. 4, pp. 44-57, 2009.

[15] P. Flicek, M. R. Amode, D. Barrell, *et al.*, "Ensembl 2012," Nucleic Acids Res, vol. 40, pp. D84-90, Jan 2012.

[16] C. bio, "CLC Genomics Workbench Product Sheet," ed.

[17] R. Li, C. Yu, Y. Li, *et al.*, "SOAP2: an improved ultrafast tool for short read alignment," Bioinformatics, vol. 25, pp. 1966-7, Aug 1 2009.

[18] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," Nat Methods, vol. 9, pp. 357-9, Apr 2012.

[19] S. Anders and W. Huber, "Differential expression analysis for sequence count data," Genome Biol, vol. 11, p. R106, 2010.

[20] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," BMC Bioinformatics, vol. 11, p. 94, 2010.

[21] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," Nucleic Acids Res, vol. 28, pp. 27-30, Jan 1 2000.

[22] M. F. Rogers and A. Ben-Hur, "The use of gene ontology evidence codes in preventing classifier assessment bias," Bioinformatics, vol. 25, pp. 1173-7, May 1 2009.

[23] Y. Kodama, M. Shumway, and R. Leinonen, "The Sequence Read Archive: explosive growth of sequencing data," Nucleic Acids Res, vol. 40, pp. D54-6, Jan 2012.

[24] J. J. Emerson, L. C. Hsieh, H. M. Sung, *et al.*, "Natural selection on cis and trans regulation in yeasts," Genome Res, vol. 20, pp. 826-36, Jun 2010.

[25] J. Pan, M. Sasaki, R. Kniewel, *et al.*, "A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation," Cell, vol. 144, pp. 719-31, Mar 4 2011.

[26] K. L. Huisinga and B. F. Pugh, "A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae," Mol Cell, vol. 13, pp. 573-85, Feb 27 2004.

[27] M. Kotera, M. Hirakawa, T. Tokimatsu, S. Goto, and M. Kanehisa, "The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals," Methods Mol Biol, vol. 802, pp. 19-39, 2012.


978-1-4673-4398-5/12/$31.00 ©2012 IEEE