SCIENTIFIC REPORTS

OPEN

TACCO, a Database Connecting Transcriptome Alterations, Pathway Alterations and Clinical Outcomes in Cancers

Received: 6 September 2018
Accepted: 19 February 2019
Published online: 07 March 2019

Po-Hao Chou[1], Wei-Chao Liao[1,2,3], Kuo-Wang Tsai[4], Ku-Chung Chen[5], Jau-Song Yu[1,6,7] & Ting-Wen Chen[8]

Because of innumerable cancer sequencing projects, abundant transcriptome expression profiles together with survival data are available from the same patients. Although some expression signatures for prognosis or pathologic staging have been identified from these data, systematically discovering such kind of expression signatures remains a challenge. To address this, we developed TACCO (Transcriptome Alterations in CanCer Omnibus), a database for identifying differentially expressed genes and altered pathways in cancer. TACCO also reveals miRNA cooperative regulations and supports construction of models for prognosis. The resulting signatures have great potential for patient stratification and treatment decision-making in future clinical applications. TACCO is freely available at http://tacco.life.nctu.edu.tw/.

Although considerable cancer sequencing data are already publicly available, systematically discovering meaningful correlations from these data is still challenging for cancer biologists lacking related computer skills. Large cancer sequencing projects, such as The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), have produced large amounts of sequencing data and made these data publicly available[1,2]. Although countless next-generation sequencing analysis tools and pipelines for processing high-throughput genomic and transcriptomic sequencing data have been developed, using these tools or pipelines still requires some basic command-line knowledge and sometimes even certain programming skills. Therefore, an easy-to-use interface that allows investigators to manage, integrate, and visualize cancer sequencing data across multiple cancer types without the need for computer skills would be a valuable tool for utilizing public cancer genomics data and advancing the cancer research field.

A number of databases, including FireBrowse, cBioPortal, OncoLnc, CancerMiner, GEPIA, miRCancerdb and MiRGator, are available for exploring transcriptome changes in cancers[3–8]. Using these databases, researchers can identify differentially expressed genes (DEGs), perform pathway analyses using these DEGs, explore correlations between expression levels of miRNAs and their target genes and analyze associations between the expression of individual genes and overall survival, among other functionalities. A five-miRNA (micro RNA) signature was recently proposed for stratification of patients with pancreatic adenocarcinoma into high-risk and low-risk groups with 5-year overall survival rates of 10.2% and 47.8%, respectively[9]. Similarly, other combined expression signatures have been proposed for lung adenocarcinoma, head and neck squamous cell carcinomas, glioblastomas, and breast cancers[10–13]. These signatures can potentially be used as clinical markers in personalized medicine; however, currently available databases only provide connections between the expression level of a single

[1]Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan. [2]Department of Otolaryngology-Head & Neck Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan. [3]Center for General Education Chang Gung University, Taoyuan, Taiwan. [4]Department of Medical Education and Research, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan. [5]Department of Biochemistry and Molecular Cell Biology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan. [6]Department of Cell and Molecular Biology, Chang Gung University, Taoyuan, Taiwan. [7]Liver Research Center, Chang Gung Memorial Hospital, Linkou, Taiwan. [8]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan. Correspondence and requests for materials should be addressed to T.-W.C. (email: afratw@gmail.com)

gene and survival data[3,6,7]. Therefore, a cancer transcriptome database that incorporates a feature that allows prognosis model construction would be extremely valuable.

In addition to survival signatures, another important, but often neglected, factor is miRNA-mRNA regulatory networks. Dysregulation of miRNA expression is significant in cancer formation and development[14]. miRNAs are 22-nucleotide long non-coding RNAs that target and regulate the expression of hundreds of target mRNAs; moreover, one gene may be targeted by multiple miRNAs. Thus, transcriptome alterations in cancer are a consequence of these multiple-to-multiple regulatory relationships among miRNAs and their target genes[15–17]. However, this type of combinatorial regulation of miRNAs has not been considered or investigated in previous cancer transcriptome databases. These miRNA cooperative modules can be taken into consideration by simply adding an analysis of how many miRNAs co-target the same genes. This additional information about such cooperative miRNAs can be helpful in selecting target genes for subsequent analysis or validation.

To fulfill all the analytical requirements for cancer transcriptomes, we propose the database, Transcriptome Alterations in CanCer Omnibus (TACCO). TACCO aims to provide an interactive interface that enables researchers to specify a group of significant differentially expressed miRNAs (DEmiRNAs) or DEGs, and subsequently perform pathway enrichment analysis and model construction for prognosis. TACCO will be useful for developing models for prognosis and thus should prove beneficial to the entire cancer research community.

Results and Discussion

Browse the expression levels of genes of interest in different cancer types.

An overview of TACCO is shown in Fig. 1. TACCO provides gene and miRNA expression data for 26 and 22 cancer types, respectively. TACCO is the first cancer transcriptome database that includes miRNA-target correlations and provides the signature construction for prognosis and pathological staging. On the browse page, the user can either select or key in a gene symbol or miRNA ID of interest to explore expression fold changes, average expression levels in normal and tumor tissue, and p-values calculated from expression levels in tumor and adjacent normal tissues for different cancer types. TACCO also presents correlations between the expression levels of miRNA and target genes for cancer types for which both miRNA and gene expression data are available. While Pearson's r and Spearman's ρ are suitable for discovering linear correlation and rank correlation, respectably, both correlation analyses have been used in exploring miRNA-mediated regulation of target genes[5,8,18]. Therefore, TACCO calculates both Pearson's r and Spearman's ρ, and offers a distribution plot.

Identify DEGs from a volcano plot.

TACCO illustrates transcriptome changes in the form of volcano plots together with slider bars for both p-values and fold changes. Hence, users can use customized criteria for identification of DEGs or look specifically for upregulated/downregulated gene lists. The volcano plots and number of DEGs refreshes on the fly upon user modification of the p-value or fold-change filter. After users apply their own criteria to identify a group of significant DEGs, these genes can be used in pathway enrichment analysis or model construction for survival prediction. TACCO also analyzes the number of DEmiRNAs that target the same gene, allowing users to investigate miRNA cooperative regulatory networks. To further investigate miRNA cooperative modules in cancers, we implemented KEGG pathway enrichment analysis in TACCO. We analyzed enriched pathways for genes co-targeted by at least 1, 2, 3, 4, or 5 DEmiRNAs for several cancer types which have related genes included in the KEGG database. We then calculated the percentage of genes among all targeted genes that have been reported in specific-cancer pathways. We found that the ratio of co-targeted genes that are found in the cancer pathways are positively correlated with the number of regulating DEmiRNA (Fig. 2). Although there are other cancer transcriptome databases, TACCO is the only one that provides information on miRNA co-target regulation. In our experience, the cooperative behavior of miRNAs—an important factor to consider in investigating alterations in the cancer transcriptome—may often be ignored.

Specify enriched pathways in KEGG, MSigDB or GO categories.

For each cancer type, TACCO provides GSEA analysis for GO categories, gene sets from MSigDB and KEGG pathways[19–23]. Users can survey the enriched pathways together with the GSEA plot, normalized enrichment score adjusted p-value, and Q-value. Additionally, if DEGs were selected or users are interested in a specific gene list, TACCO also offers pathway enrichment analysis for subgroups of genes. TACCO utilizes a hypergeometric test to examine overrepresented pathways. For example, users can take the 803 up-regulated and 736 down-regulated genes (genes having absolute fold changes larger than 2 and p-values smaller than 0.01) in breast invasive carcinoma for the KEGG pathway and GO term enrichment analysis. KEGG pathway enrichment plots show the enrichment in many cancer-related pathways (Fig. 3a). TACCO also provides hyperlinks for all these pathways to the KEGG database and highlights the DEGs in the pathway (Fig. 3b). As for GO categories, TACCO provides a list of enriched GO terms and depicts a directed acyclic graph. The directed acyclic graph shows the parent-child relationships between GO terms and TACCO highlights the overrepresented GO terms (Fig. 4).

Construct a model for prognosis or pathologic staging.

The significant DEGs identified in cancers are likely related to tumorigenesis; thus, their expression levels are potentially correlated with clinical outcome or cancer stage. Hence TACCO provides model construction for prognosis or pathologic staging. In addition to DEGs/DEmiRNAs, users can upload a specific gene list and select a cancer type for model construction. To identify a signature for prognosis, TACCO first evaluates the power of each gene or miRNA to distinguish patients with a good outcome from those with a bad outcome and then uses Lasso regression, Ridge regression, Classification and Regression Tree (CART), Random forest or Generalized Linear Models (GLM) to construct prediction models. TACCO also produces a Kaplan-Meier survival plot and log-rank p-value for the prediction results. For pathologic staging, TACCO evaluates the power of each gene or miRNA to distinguish patients from different cancer stage or TNM categories and then use aforementioned algorithms for prediction model

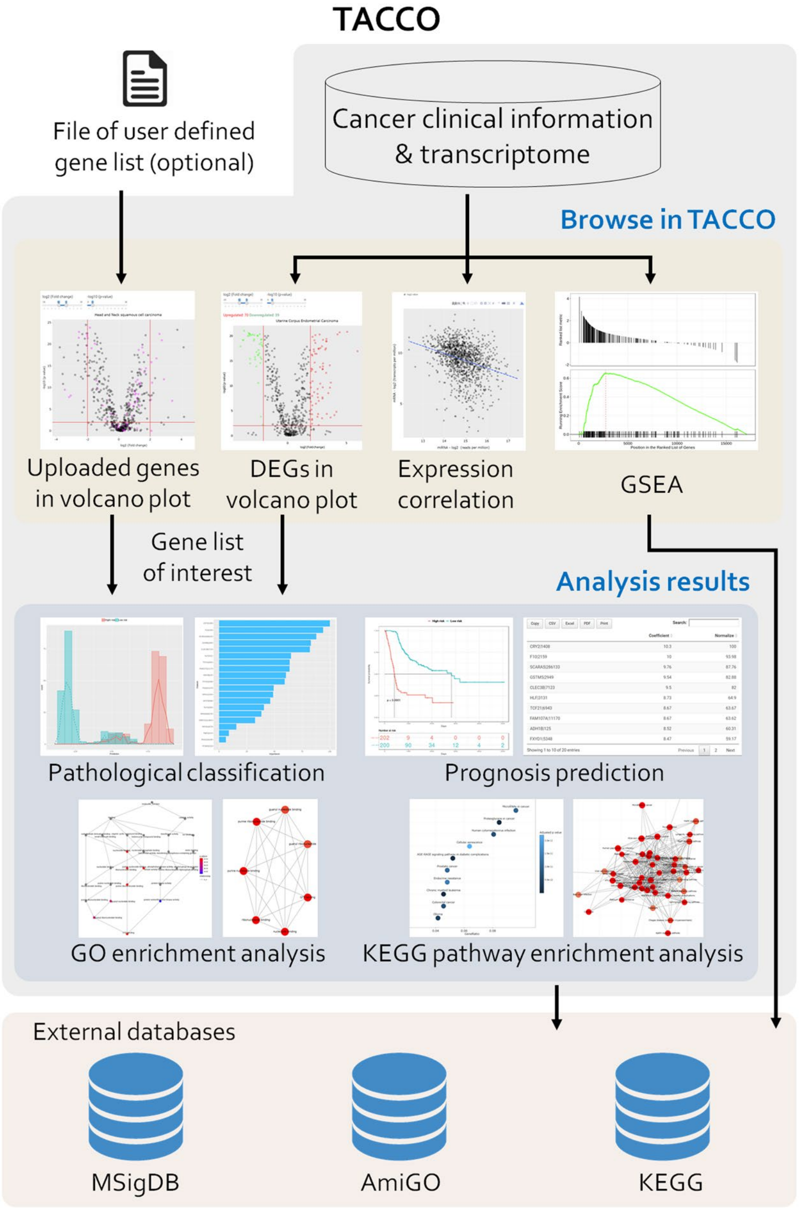

Figure 1. Overview of TACCO. TACCO was constructed using transcriptome data downloaded from several databases and provides GSEA results for gene sets from MSigDB, GO terms and KEGG pathways in 26 cancer types. In addition to GSEA, users can either identify DEGs/DEmiRNAs in TACCO or upload a gene set of interest. After a gene list is defined, users can construct a prediction model for pathological staging or clinical prognosis using the clinical or survival data in TACCO. For prognosis analysis, Kaplan-Meier survival plots with log-rank test p-values are provided for prediction results. Users can also perform a GO enrichment analysis, carry out KEGG pathway enrichment analysis[21] and get the details of specific pathway by visiting external databases.

construction. To illustrate the performance of TACCO, we tested two survival-associated miRNAs signatures provided from previous studies in pancreatic adenocarcinoma and lung adenocarcinoma[9,13]. We uploaded the reported list of miRNAs to TACCO and successfully constructed the prediction models which can distinguish low and high risk groups. As shown in Fig. 5, TACCO generates Kaplan-Meier survival plot for the predicted low and high risk groups from the 480 patients with lung adenocarcinoma as well as the importance of each predictor i.e. miRNA. Identification of a model or a signature that can predict overall or disease-free survival for various cancer types would be helpful, potentially guiding treatment decisions in the clinic.

In addition to GLM, TACCO utilizes Lasso regression and Ridge regression to identify transcriptome signatures for prognosis. Regression is one of the most commonly used machine-learning tasks, and the traditional and most popular of such methods are ordinary least squares regression and stepwise regression. However, both are known to be sensitive to random errors and are weak in terms of feature selection, prompting the development of Ridge regression and Lasso regression methods[24]. Lasso regression is a forward variable selection method that

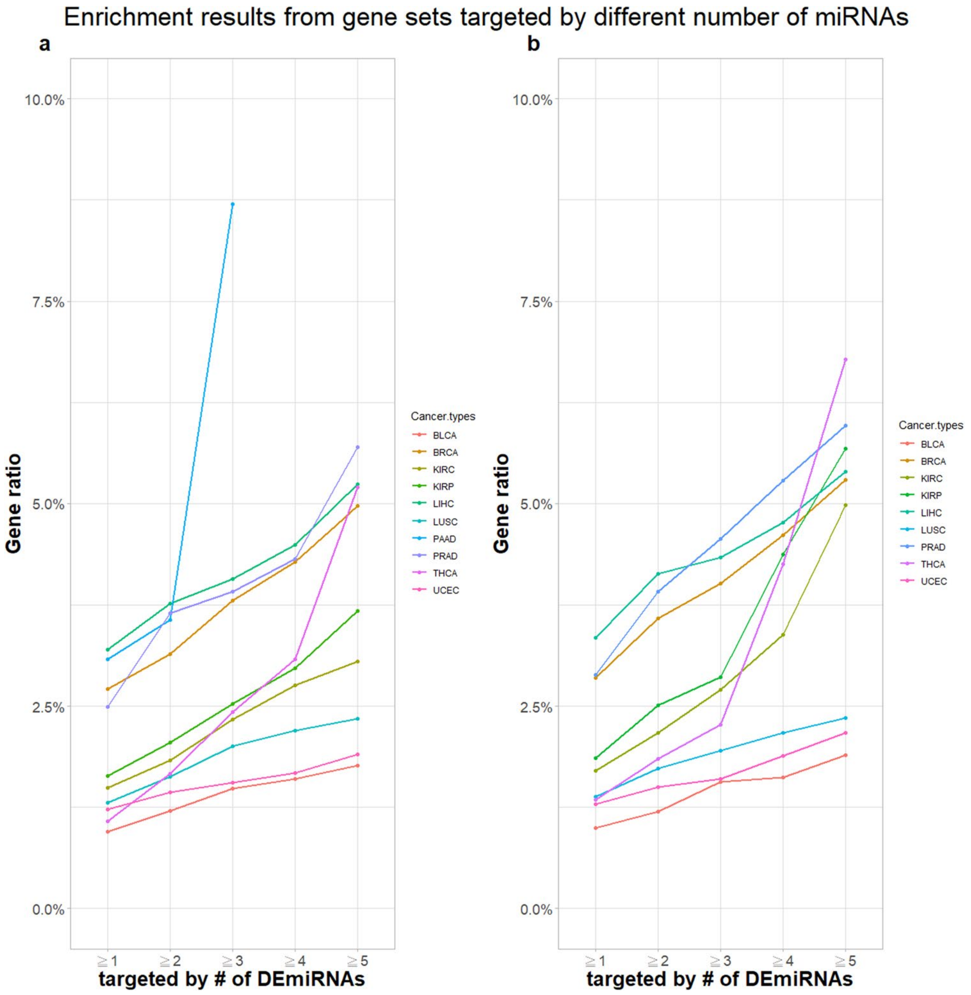

Figure 2. Genes targeted by a larger number of differentially expressed miRNAs (DEmiRNAs) are more frequently involved in cancer pathways. Both significantly upregulated and downregulated miRNAs were considered in (**a**), and only significantly upregulated miRNAs were considered in (**b**). For each cancer type, gene sets composed of genes targeted by different numbers of DEmiRNAs were used in KEGG pathway enrichment analyses. The percentage of genes involved in a given specific cancer type pathway were plotted. For example, in (A) a total of 92 genes are targeted by more than three DEmiRNAs in pancreatic cancer (PAAD; blue line), of which 8 were in KEGG hsa05212 (pancreatic cancer pathway); hence, the gene ratio is 8.7%. There is no additional data for PAAD because there are too few genes targeted by more than 4 or 5 DEmiRNAs for KEGG pathway enrichment analysis. Abbreviations: BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; PAAD, pancreatic adenocarcinoma; LUSC, lung squamous cell carcinoma; PRAD, prostate adenocarcinoma; THCA, thyroid carcinoma; and UCEC, uterine corpus endometrial carcinoma.

can choose one predictor out of a group of correlated variables. Lasso regression can also improve prediction accuracy in models with a limited number of predictors and provide better model interpretability. In addition, Lasso regression has recently been used to generate reliable models for survival prediction using transcriptome or protein expression data[25–27]. Thus, TACCO exploits Lasso regression for selection of transcriptome signatures and construction of prediction models for prognosis. For users who want to include all the uploaded or selected features in the prediction model, TACCO also provides Ridge regression. Meanwhile, decision tree based methods such as Random forest and CART are also included in TACCO because these two algorithms are also useful in signature construction for survival prediction[28,29].

An example–miR-17/92 miRNA cluster. The oncogenic miR-17/92 miRNA cluster (has-miR-17, hsa-miR-18a, hsa-miR-19a, hsa-miR-19b, hsa-miR-20a and hsa-miR-92a) is known to be frequently overexpressed and play a prognostic role in lung cancer[30,31]. We used TACCO to investigate the role of miR-17/92 cluster in lung cancer. We first explored the negative regulation of miR-17/92 cluster on their target genes. For example, significantly negative correlation (p-value $= 1.57*10^{-13}$, Spearman's rank correlation) was found between the expression levels of has-miR-19b and PTEN and their spearman's correlation coefficient is -0.321 in 510 lung adenocarcinoma samples. We found significant weak or moderate negative correlations between many miRNAs

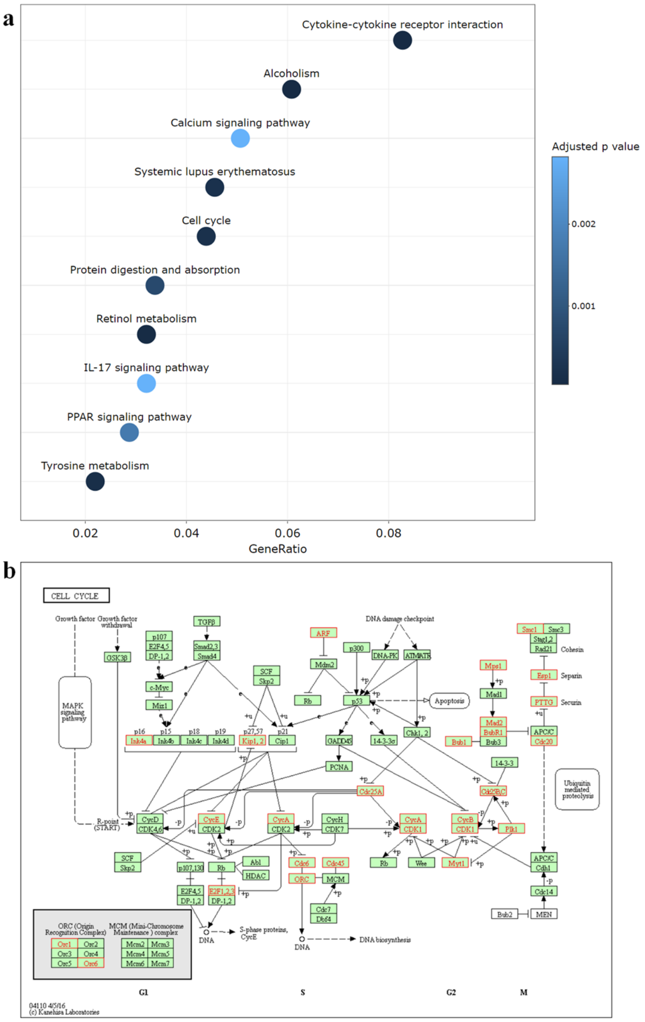

Figure 3. KEGG pathway enrichment results for differentially expressed genes (DEGs) in breast invasive carcinoma. (**a**) Enriched KEGG pathways are listed in a distribution plot in which the x-axis shows the ratio of genes included in DEGs for each pathway and the colors represent the adjusted enriched p-values. (**b**) TACCO provides links to KEGG pathway plots[21]. Here the enriched pathway cell cycle is shown where all the genes included in DEG list are written in red and highlighted in the red boxes.

and their target genes which is expectable from heterogeneous tumor samples and also the complex transcriptome regulatory networks behind. We then uploaded a list of the all the miRNA products of miR-17/92 cluster and explored the expression changes of these miRNAs in lung adenocarcinoma. We found has-miR-20a-3p was 5.09 times up-regulated, this was consistent with previous studies[32,33] (Fig. 6a). We also downloaded the targeted gene list of miR-17/92 cluster and then uploaded the gene list to TACCO to explore the expression changes of these targeted genes (Fig. 6b). We further carried out KEGG pathway enrichment with all the target genes for these miRNAs and found multiple cancer-related pathways enriched, including p53 signaling pathway, cellular senescence, proteoglycans in cancer and PI3K-Akt signaling pathway etc. We further explored the PI3K-Akt signaling pathway in KEGG database and found several important genes including AMPK, Ras, PI3K, Raf-1 and ERK are all targeted by the miR-17/92 miRNA cluster (Fig. 6c). We also tried to construct a model from these miRNAs to distinguish between patients with distant metastasis and without metastasis. Combining these miRNAs and their target genes, TACCO can provide a signature to differentiate patients with (M1) and without (M1) distant metastasis in lung adenocarcinoma. The signature composite of 121 genes include the top important TCP1, ARMT1, PIP4K2A which were already reported to correlated with metastasis in other cancer types[34–36]. Even though the signature misclassified few M1 as M0, most of the patients were correctly grouped (Fig. 6d).

Comparison with other existing databases. Although other cancer transcriptome databases that come with interactive graphical user interfaces are available, TACCO is the only one that provides prediction model construction capability and information on cooperative miRNA modules (Table 1). As shown in this study, these cooperative regulatory interactions may be a non-negligible factor in studying transcriptome alterations. Four databases—MiRGator, GEPIA, cBioPortal and TACCO—provide identification of dysregulated genes or

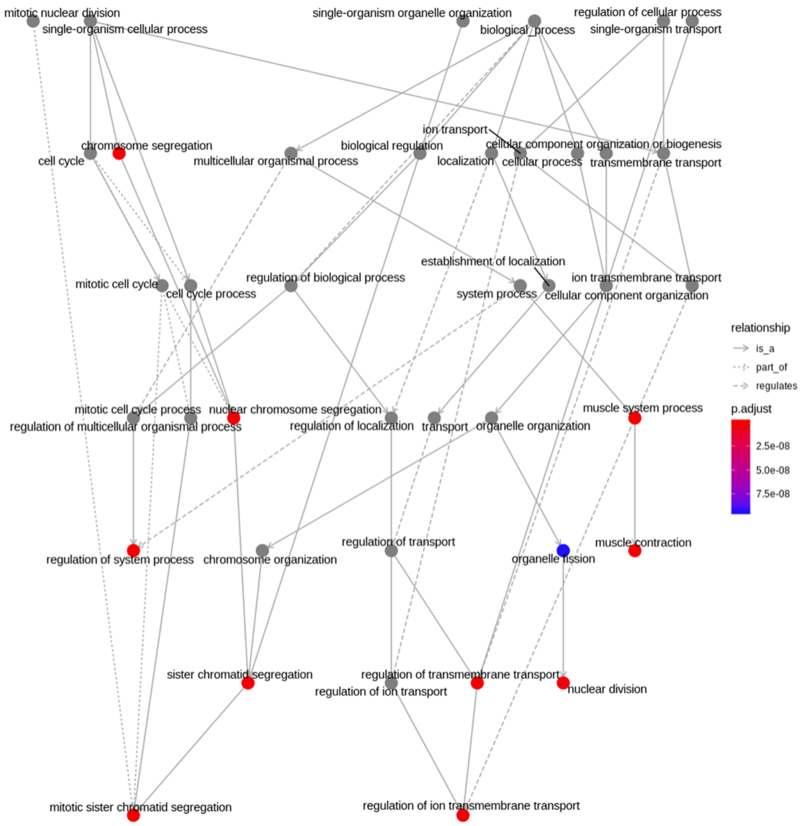
mitotic nuclear division
single-organism cellular process
single-organism organelle organization
biological_process
regulation of cellular process
single-organism transport
ion transport
chromosome segregation
cell cycle
biological regulation
multicellular organismal process
localization
cellular component organization or biogenesis
cellular process
transmembrane transport
establishment of localization
regulation of biological process
mitotic cell cycle
cell cycle process
system process
ion transmembrane transport
cellular component organization
relationship
is_a
part_of
regulates
p.adjust
2.5e-08
5.0e-08
7.5e-08
mitotic cell cycle process
nuclear chromosome segregation
regulation of multicellular organismal process
regulation of localization
transport
organelle organization
muscle system process
regulation of system process
chromosome organization
regulation of transport
organelle fission
muscle contraction
sister chromatid segregation
regulation of transmembrane transport
regulation of ion transport
nuclear division
mitotic sister chromatid segregation
regulation of ion transmembrane transport

Figure 4. GO term enrichment result for differentially expressed genes (DEGs) in breast invasive carcinoma. The directed acyclic graph for overrepresented Biological process GO terms identified for DEGs are depicted. The circles represent GO terms, lines represent the relationship between GO terms and colors represent the adjusted enriched p-values.

miRNAs, but only GEPIA and TACCO provide an interface that allows users to specify customized criteria for DEmiRNA or DEG identification. TACCO additionally provides volcano plots and the number of upregulated/downregulated DEGs, which refresh on the fly. Although FireBrowse, GEPIA and OncoLnc also offer survival analysis for single genes[3,7], none of these databases provide model construction for survival prediction.

Conclusion

We propose a cancer transcriptome database, TACCO, that aims to link transcription alterations and transcriptome regulatory networks with alterations in downstream pathways and clinical outcomes in different cancer types. TACCO provides a user-friendly interface for assessing correlations between the expression of miRNAs and their target genes, identifying DEGs and altered pathways in cancers, and investigating miRNA co-target regulatory networks. Additionally, TACCO constructs models for prognosis from DEG lists or user-defined gene lists. Collectively, the analytical capabilities and model construction features present in TACCO make it feasible for researchers or clinicians to systematically investigate transcriptome regulatory network alterations and clinical outcomes in cancers. Accordingly, we believe that TACCO will shed light on important questions in the field of cancer research.

Materials and Methods

Identification of DEGs.

Expression levels of miRNAs for 22 cancer types and mRNAs for 26 cancer types (Supplementary Table 1) were download from Broad GDAC Firehose version stddata__2016_01_28. All miRNA IDs were converted from MIMAT ID to miRBase nomenclature (hsa-miR-133a-3p, hsa-miR-557 etc.) based on miRBase 22 release[37]. Expression levels of mRNAs were obtained from RNAseqV2 data which were derived from RSEM[38]. For each cancer type, all genes with median expression levels greater than 0.01 transcript per million (TPM) across all samples were considered to be expressed. For each expressed gene, fold-change in expression levels between tumor and normal tissues, corresponding p-value, and Benjamini-Hochberg adjusted p-value were calculated using the EBSeq, Wilcoxon rank-sum test and multiple test correction in R[39,40]. Finally, volcano plots were generated using R package ggplot2. Based on validated interactions between miRNA and mRNA downloaded from miRTarBase 7.0[41], TACCO also lists target genes for DEmiRNAs and calculates the number of DEmiRNAs that target these genes.

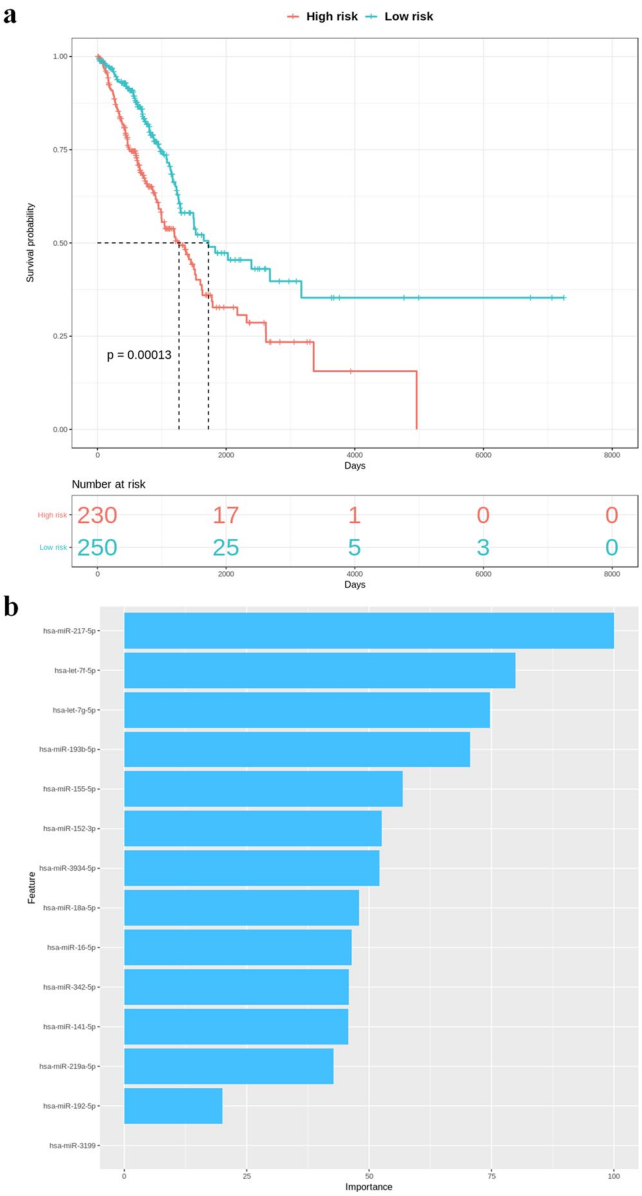

Figure 5. Model for prognosis constructed with TACCO. From a list of miRNAs TACCO constructs a prediction model for lung adenocarcinoma patients and uses this model to stratify samples into high and low risk groups. The patients from high and low risk were tested with log-rank test. TACCO provides tables and plots for the prediction results include (**a**) the Kaplan-Meier survival plot together with the number of patient surviving at each specific time point and (**b**) the importance for each predictor in the prediction model.

Correlation between expression levels of mirna and target genes.

For exploring regulatory relationships between miRNAs and their target genes, TACCO provides a tool that analyzes correlations between the expression levels of miRNAs and their target genes. When browsing TACCO, the user can select a specific gene or miRNA. TACCO calculates both parametric and non-parametric correlation coefficients (i.e., Pearson's r and Spearman's ρ) for expression levels of the interacting mRNA/target gene pair. In cases where a more normal expression distribution is needed, TACCO also offers correlation coefficients for log-transformed expression values. All correlation coefficients are listed in a table from which the user can select a gene or miRNA of interest to explore in detail. When the user selects a gene or miRNA from the table, a distribution plot for expression levels of the selected miRNA/target gene and a regression line with iteratively reweighted least squares is generated on the fly.

Pathway enrichment analysis.

TACCO provides Gene Set Enrichment Analysis (GSEA) for interpreting expression data for different cancer type[22]. TACCO also offers pathway enrichment analysis for either selected DEGs or an uploaded gene list. TACCO exploits the hypergeometric test, clusterProfiler, from the R package to identify enriched KEGG pathways or Gene Ontology (GO) terms[42]. TACCO generates a directed acyclic graph for

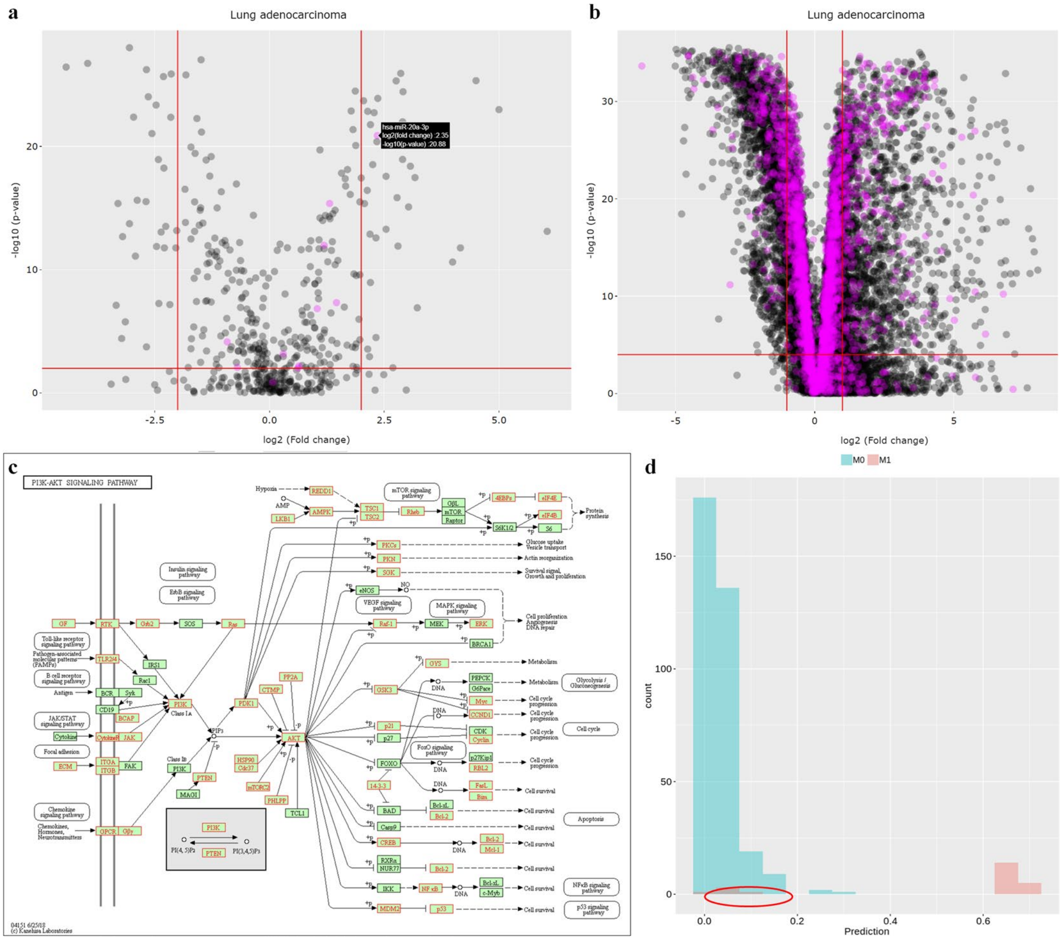

Figure 6. Explore the role of miR-17/92 miRNA cluster in lung adenocarcinoma with TACCO. (**a**) The volcano plot of miRNAs in lung adenocarcinoma. The miR-17/92 miRNA cluster includes 6 miRNAs, i.e. has-miR-17, hsa-miR-18a, hsa-miR-19a, hsa-miR-19b, hsa-miR-20a and hsa-miR-92a. The product of these 6 miRNAs are labeled in red. The most significantly upregulated miRNA is hsa-miR-20a-5p which has log2 fold change value equals to 2.35. (**b**) The volcano plot of genes in lung adenocarcinoma. The genes targeted by miR-17/92 miRNA cluster were labeled in purple. (**c**) PI3K-AKT signaling pathway is found enriched with genes targeted by the miRNAs located in miR-17/92 miRNA clusters. The targeted genes are in red box[21]. (**d**) The M stage (distant metastasis) classification results of the model constructed from miR-17/92 miRNA cluster and their target genes. The x-axis is the prediction values for classification. Patients with prediction value larger than 0.5 and smaller than 0.5 are predicted to be in M1 group and M0 group, respectively. The red and green color represent patients clinically diagnosed as from M1 and M0 groups, respectively. We found the prediction values for all M0 stage patients are smaller than 0.5 and hence correctly classified. As indicated by red circle, a few M1 stage patients have prediction values around 0.1 and wrongly classified as M0 group.

enriched GO terms using R package, enrichplot. TACCO also displays a button to visit the KEGG website, where all the selected genes are highlighted within a red box.

Identification of signatures for prognosis or pathological staging. All clinical and survival information for patients was downloaded with the R package, curatedTCGAData[43]. For each cancer type, all patients are divided into two groups based on their median survival/disease-free survival, in days. Differentially expressed mRNAs and miRNAs capable of distinguishing the two groups (Wilcoxon rank-sum test, $p\text{-value} < 0.05$) are selected as candidate features for signature identification[44]. Algorithms from the caret package[45], including Lasso regression, Ridge regression, Random forest, Classification and Regression Tree (CART) and General Linear Model (GLM) were provided for prediction model construction. Prediction models constructed from 5-fold or 3-fold (for <150 patients) cross-validation is used to categorize patients into better (Low risk) or worse (High risk) surviving groups[45]. Finally, TACCO evaluates the prediction results using a Kaplan-Meier survival plot (KM plot) and log-rank test. Both KM plot and log-rank test results are generated to allow comparisons of survival data from predicted High-risk and Low-risk groups. TACCO also utilizes samples from early/late stage or different TNM stages to construct models for pathological staging with the same classification strategy.

	TACCO	FireBrowse[#]	cBioPortal[#]	OncoLnc	GEPIA	CancerMiner	miRCancerdb	MiRGator
Pathway GSEA	✓	✓						
Identification of DEGs	✓		✓		✓			
Identification of DEmiRNA	✓		✓					✓
Correlation between the expression of miRNAs and target genes	✓	✓				✓	✓	✓
Pathway enrichment analysis	✓	✓	✓					✓
Survival analysis for single mRNAs/miRNAs		✓[*]		✓	✓※			
Analysis of multiple miRNAs that co-target an mRNA	✓							
Survival prediction model construction for multiple miRNAs and mRNAs	✓							

Table 1. Comparison of TACCO with other cancer transcriptome databases. [#]Database for both genome and transcriptome. [*]Only significant mRNA/miRNA are listed in the database. ※GEPIA provides survival analysis for mRNAs, but not miRNA.

Website construction and availability. TACCO was built using Python, JavaScript, R and R Shiny on a Linux operating system and can be updated with new cancer sequencing projects based on current database schema. All tables and figures generated in TACCO can be downloaded by users. TACCO is available at http://tacco.life.nctu.edu.tw/ and can be explored with multiple web browsers, including Chrome, Internet Explorer, Firefox, and Safari.

References

1. Grossman, R. L. *et al*. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**, 1109–1112, https://doi.org/10.1056/NEJMp1607591 (2016).
2. Zhang, J. *et al*. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**, bar026, https://doi.org/10.1093/database/bar026 (2011).
3. Tang, Z. *et al*. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* **45**, W98–W102, https://doi.org/10.1093/nar/gkx247 (2017).
4. Jacobsen, A. *et al*. Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* **20**, 1325–1332, https://doi.org/10.1038/nsmb.2678 (2013).
5. Cho, S. *et al*. MiRGatorv3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic Acids Res* **41**, D252–257, https://doi.org/10.1093/nar/gks1168 (2013).
6. Cerami, E. *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404, https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).
7. Anaya, J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peer J Computer Science*. https://doi.org/10.7717/peerj-cs.67 (2016).
8. Ahmed, M., Nguyen, H., Lai, T. & Kim, D. R. miRCancerdb: a database for correlation analysis between microRNA and gene expression in cancer. *BMC research notes* **11**, 103, https://doi.org/10.1186/s13104-018-3160-9 (2018).
9. Shi, X. H. *et al*. A Five-microRNA Signature for Survival Prognosis in Pancreatic Adenocarcinoma based on TCGA Data. *Scientific reports* **8**, 7638, https://doi.org/10.1038/s41598-018-22493-5 (2018).
10. Wong, N. *et al*. Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer medicine* **5**, 1619–1628, https://doi.org/10.1002/cam4.718 (2016).
11. Volinia, S. & Croce, C. M. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci USA* **110**, 7413–7417, https://doi.org/10.1073/pnas.1304977110 (2013).
12. Kim, Y. W. *et al*. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. *Neuro-oncology* **15**, 829–839, https://doi.org/10.1093/neuonc/not024 (2013).
13. Yerukala Sathipati, S. & Ho, S. Y. Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. *Scientific reports* **7**, 7507, https://doi.org/10.1038/s41598-017-07739-y (2017).
14. Schickel, R., Boyerinas, B., Park, S. M. & Peter, M. E. MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death. *Oncogene* **27**, 5959–5974, https://doi.org/10.1038/onc.2008.274 (2008).
15. Shao, T. *et al*. Survey of miRNA-miRNA cooperative regulation principles across cancer types. *Brief Bioinform*. https://doi.org/10.1093/bib/bby038 (2018).
16. Peter, M. E. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene* **29**, 2161–2164, https://doi.org/10.1038/onc.2010.59 (2010).
17. Chen, W. S. *et al*. Co-modulated behavior and effects of differentially expressed miRNA in colorectal cancer. *BMC Genomics* **14**(Suppl 5), S12, https://doi.org/10.1186/1471-2164-14-S5-S12 (2013).
18. Muniategui, A., Pey, J., Planes, F. J. & Rubio, A. Joint analysis of miRNA and mRNA expression data. *Brief Bioinform* **14**, 263–278, https://doi.org/10.1093/bib/bbs028 (2013).
19. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740, https://doi.org/10.1093/bioinformatics/btr260 (2011).
20. Liberzon, A. *et al*. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425, https://doi.org/10.1016/j.cels.2015.12.004 (2015).
21. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361, https://doi.org/10.1093/nar/gkw1092 (2017).
22. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550, https://doi.org/10.1073/pnas.0506580102 (2005).
23. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
24. Muthukrishnan, R. & Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. *IEEE International Conference on Advances in Computer Applications* (2016).
25. Zhang, H. *et al*. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian. *Cancer. Cell* **166**, 755–765, https://doi.org/10.1016/j.cell.2016.05.069 (2016).

26. Ternes, N., Rotolo, F. & Michiels, S. Robust estimation of the expected survival probabilities from high-dimensional Cox models with biomarker-by-treatment interactions in randomized clinical trials. *BMC Med Res Methodol* **17**, 83, https://doi.org/10.1186/s12874-017-0354-0 (2017).
27. Kaneko, S., Hirakawa, A. & Hamada, C. Enhancing the Lasso Approach for Developing a Survival Prediction Model Based on Gene Expression Data. *Comput Math Methods Med* **2015**, 259474, https://doi.org/10.1155/2015/259474 (2015).
28. Barlin, J. N. *et al*. Classification and regression tree (CART) analysis of endometrial carcinoma: Seeing the forest for the trees. *Gynecol Oncol* **130**, 452–456, https://doi.org/10.1016/j.ygyno.2013.06.009 (2013).
29. Li, J. *et al*. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700–1710, https://doi.org/10.1136/gutjnl-2013-305806 (2014).
30. Mogilyansky, E. & Rigoutsos, I. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ* **20**, 1603–1614, https://doi.org/10.1038/cdd.2013.125 (2013).
31. Liu, F. *et al*. Prognostic role of miR-17-92 family in human cancers: evaluation of multiple prognostic outcomes. *Oncotarget* **8**, 69125–69138, https://doi.org/10.18632/oncotarget.19096 (2017).
32. Thompson, T. A. *et al*. Induction of apoptosis by organotin compounds *in vitro*: neuronal protection with antisense oligonucleotides directed against stannin. *J Pharmacol Exp Ther* **276**, 1201–1216 (1996).
33. Osada, H. & Takahashi, T. let-7 and miR-17-92: small-sized major players in lung cancer development. *Cancer Sci* **102**, 9–17, https://doi.org/10.1111/j.1349-7006.2010.01707.x (2011).
34. Dainty, K. *Investigation into the Role of ARMT1 in Oestrogen Receptor Positive Breast Cancer*, University of Otago (2017).
35. Coghlin, C. *et al*. Characterization and over-expression of chaperonin t-complex proteins in colorectal cancer. *J Pathol* **210**, 351–357, https://doi.org/10.1002/path.2056 (2006).
36. Paula, L. M. *et al*. Analysis of molecular markers as predictive factors of lymph node involvement in breast carcinoma. *Oncol Lett* **13**, 488–496, https://doi.org/10.3892/ol.2016.5438 (2017).
37. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**, D68–73, https://doi.org/10.1093/nar/gkt1181 (2014).
38. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, https://doi.org/10.1186/1471-2105-12-323 (2011).
39. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
40. Leng, N. *et al*. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043, https://doi.org/10.1093/bioinformatics/btt087 (2013).
41. Chou, C. H. *et al*. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **46**, D296–D302, https://doi.org/10.1093/nar/gkx1067 (2018).
42. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287, https://doi.org/10.1089/omi.2011.0118 (2012).
43. Ramos, M., Waldron, L., Schiffer, L., Obenchain, V. & Martin, M. curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects (2018).
44. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. *Journal of Statistical Software* **33**, 1 (2010).
45. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1–26 (2008).

Acknowledgements

This study was supported by grants from the Ministry of Science and Technology, Taiwan (MOST-106-2311-B-182-005 and MOST-107-2311-B-182-001); the Chang Gung Memorial Hospital, Taiwan (CIRPD3B0013); and the "Molecular Medicine Research Center, Chang Gung University" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Author Contributions

T.W.C. designed and supervised the project and wrote the manuscript. P.H.C. designed and construed the database. J.S.Y. and K.C.C. contributed to revision of the manuscript. W.C.L. and K.W.T. provided computer resources and biological interpretations. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40629-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.