YM500v2: a small RNA sequencing (smRNA-seq) database for human cancer miRNome research

Wei-Chung Cheng1,†, I-Fang Chung2,†, Cheng-Fong Tsai2,3, Tse-Shun Huang4, Chen-Yang Chen2,3, Shao-Chuan Wang2, Ting-Yu Chang3, Hsing-Jen Sun2, Jeffrey Yung-Chuan Chao5,6, Cheng-Chung Cheng7, Cheng-Wen Wu5,8,9 and Hsei-Wei Wang2,3,5,10,11,*

1Research Center for Tumor Medical Science, China Medical University, Taichung 40402, Taiwan, 2Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan, 3VGH-YM Genomic Research Center, National Yang-Ming University, Taipei 11221, Taiwan, 4Institute of Engineering in Medicine, University of California, San Diego, La Jolla, CA 92093, USA, 5Institute of Clinical Medicine, Medical College, National Yang-Ming University, Taipei 11221, Taiwan, 6Department of Radiation Oncology, Taichung Veterans’ General Hospital, Taichung 40705, Taiwan, 7Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, National Defence Medical Center, Taipei 11490, Taiwan, 8Institute of Biomedical Science, Academia Sinica, Taipei 11529, Taiwan, 9Institute of Biochemistry and Molecular Biology, National Yang Ming University, Taipei 11221, Taiwan, 10Institute of Education and Research, Taipei City Hospital, Taipei 10341, Taiwan

ABSTRACT

We previously presented YM500, which is an integrated database for miRNA quantification, isomiR identification, arm switching discovery and novel miRNA prediction from 468 human smRNA-seq datasets. Here in this updated YM500v2 database (http://ngs.ym.edu.tw/ym500/), we focus on the cancer miRNome to make the database more disease-orientated. New miRNA-related algorithms developed after YM500 were included in YM500v2, and, more significantly, more than 8000 cancer-related smRNA-seq datasets (including those of primary tumors, paired normal tissues, PBMC, recurrent tumors, and metastatic tumors) were incorporated into YM500v2. Novel miRNAs (miRNAs not included in the miRBase R21) were not only predicted by three independent algorithms but also cleaned by a new in silico filtration strategy and validated by wetlab data such as Cross-Linked ImmunoPrecipitation sequencing (CLIP-seq) to reduce the false-positive rate. A new function ‘Meta-analysis’ is additionally provided for allowing users to identify real-time differentially expressed miRNAs and arm-switching events according to customer-defined sample groups and dozens of clinical criteria tidying up by proficient clinicians. Cancer miRNAs identified hold the potential for both basic research and biotech applications.

INTRODUCTION

MicroRNAs (miRNAs), a small RNA species of ~22 nt in length, control gene expression and are involved in the regulation of a variety of biological processes including development, differentiation, cell proliferation, metabolism and inflammation, as well as in human diseases (1). It is estimated that more than 60% of human protein-coding genes contain miRNA target sites with their 3' UTR regions (2). miRNAs can function as oncogenes or tumor suppressors, as do miR-21 and let-7, respectively, and improved knowledge of them has been one of the defining developments in cancer research over the past decade. The abnormal expression of miRNAs and dysregulation of factors that regulate miRNAs contribute to the progression of tumor (3). Owing to the stability of miRNAs within clinical samples, miRNAs have been regarded as potential prognostic indicators in cancers (4) and as biomarkers for cancer classification (5–8). Many miRNA signatures have been proposed for patient prognosis and clinical response and are being investigated in clinical trials (9). In addition, miRNAs have emerged as therapeutic targets for cancer treatments. The function of miRNAs can be efficiently and specifically inhibited by artificial miRNA mimics or antagonists, supporting their potentials as novel therapy tactics for diseases (9–

*To whom correspondence should be addressed. Tel: +886 2 28267109; Fax: +886 2 28122880; Email: hwwang@ym.edu.tw
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
MRX34, a synthetic miR-34a mimic loaded in liposome, is the first miRNA-based therapy and has entered the clinics for cancer therapy in 2013 (14).

In the past few years, because of the increasing use of next generation sequencing (NGS), enormous amounts of small RNA sequencing data have been generated from large-scale cancer projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). These data help researchers to investigate the roles of miRNAs in cancer progression. However, translating such massive amounts of data into information that can be easily interpreted and accessed remains a challenge.

Previously, we developed YM500 (15), a database that includes integrated pipelines for miRNA quantification, isomiR identification, arm switching discovery and novel miRNA prediction from smRNA-seq. YM500 provides researchers with integrated miRNA-related information with various graphical visualization pages from 468 human and 141 mouse smRNA-seq datasets via a user-friendly web interface. No particular biological question or disease was emphasized in YM500. Here we present YM500v2, an updated version of the database. In addition to including more smRNA-seq results (>8000 TCGA cancer-related datasets), we focus on human cancer research in this version. Other new features also include novel miRNA identification and validation: novel miRNAs, i.e. those not included in miRBase R21, were identified from thousands of samples, and a new strategy was developed for filtering out false-positive discoveries. Especially, dozens of CLIP-seq datasets are provided as experimental evidence for in silico predicted novel miRNAs. Regarding isomiRs, two new statistical charts are provided for a specific isomiR. YM500v2 also contains a new function ‘Meta-analysis’, which allows researchers to identify differentially expressed miRNAs and arm-switching events in two customer-defined groups of samples.

DATA COLLECTION AND PRE-PROCESSING

There are 8105 smRNA-seq datasets from TCGA, including those of primary tumors, paired normal tissues, peripheral blood mononuclear cell (PBMC), recurrent tumors and metastatic tumors, incorporated into YM500v2 (Supplemental Table S1). Raw data were downloaded from CGHub (https://cghub.ucsc.edu/) and were pre-processed by in-house scripts. To avoid the disturbance of reads with poor quality, we filtered out reads that do not meet the following criterion: Phred Score >30 in >90% sequence. About 30–55% of reads were excluded in this step, but millions of reads still remained for each dataset. As shown in Supplemental Figure S1, this filtration step dramatically improves data quality, which is important for novel miRNA and isomiR identification. The other pre-processing steps were the same as those followed in our previous study (15).

Clinical information of each sample was manually curated by proficient clinicians based on clinical data obtained in TCGA. Each sample was annotated with 35 clinical characteristics.

NOVEL miRNA PREDICTION

In this version, 22 CLIP-seq datasets are incorporated as experimental evidence for the putative novel miRNAs, and a new strategy was developed to reduce the false-positive rate. It has been noted that there are numerous advantages in conducting novel miRNA prediction by performing a single analysis of the pooled data (16–18). After pre-processing, we pooled 8105 and 468 smRNA-seq datasets from TCGA and the Gene Expression Omnibus (human datasets in YM500), respectively, comprising approximately 11.1 billion reads from 193 million unique sequences. When analyzing the sequences, we found that there are 76.7% of sequences with only one read in all the samples and 0.41% of sequences with more than 100 reads in ~8600 samples.
Figure 2. The screen shot of the results for ‘Meta-Analysis’.
Figure 3. Two new charts for isomiRs. Panel (A) shows the percentage of samples that expressed the isomiR in various tissues, and panel (B) indicates the mean expression of a specific isomiR.

SUPPLEMENTAL FIGURE S2. This result indicates that most sequences only appear in one sample. We assume that if a sequence does exist in humans, multiple reads from multiple independent experiments will appear to support its existence. Thus, to reduce the false-positive rate, only sequences that met the criterion of being supported by >100 reads in >10 independent datasets were used for novel miRNA prediction according to the pipeline described in our previous study (15). In brief, the pooled dataset was used for novel miRNA prediction by miRDeep2 (19), mireap (20) and miRanalyzer (21). After unifying the prediction results of the three algorithms and filtering out novel miRNAs similar to known transcripts, 3467 putative novel miRNAs were predicted by at least one algorithm and 1408 of those were supported by CLIP-seq data (Supplemental Figure S3). The alignment results of both smRNA-seq and CLIP-seq are provided in the web interface for each putative novel miRNA (Figure 1).

META-ANALYSIS

YM500v2 adds a new function, Meta-analysis, which allows researchers to identify differentially expressed miRs and arm-switching events from two user-defined, specific sets of samples. We utilize an R/Bioconductor package, DESeq (22,23) and the algorithm described in our previous study (15) to identify differentially expressed miRs and arm-switching events, respectively. As shown in Supplemental Figure S4, users can select one or multiple datasets and define what sample type they would like to investigate, such as 'Primary Solid Tumor' or 'Solid Tissue Normal'. In addition, we also provide a list of clinical criteria, such as ICD-O-3 histology, tumor stage, distant metastasis and lymph node status, to help researchers to select a subgroup of well-defined cancer samples according to one or multiple clinical parameters. After selecting two groups of samples, users could overview the detailed clinical information of the two groups of selected samples before submitting this job to the server for real-time calculation. The user would then receive a notification email with a Result ID, and could then see a visualization of the results (Figure 2) in the ‘Result and Download’ section when the job is completed.

ISOMIRS

Two statistical charts are added to the ‘IsomiR’ section to help researchers realize the expression profile of an isomiR across distinct tissues. miR-21-5p+CA, an isomiR of miR-21-5p adenylated by PAPD5 reported to lead the degradation of miR-21-5p (24), is used as an example (Figure 3). Figure 3A shows the percentage of samples that expressed this isomiR in various tissues, and Figure 3B indicates the mean expression of a specific isomiR.

DISCUSSION

The decreasing cost of sequencing technology has led to large amounts of small RNA sequencing data from cancer-related studies. To make the most of these valuable yet massive amounts of data, we have updated our YM500 database to make it focus on the human cancer miRNome. There are more than 8000 cancer-related samples incorporated into YM500v2, and a new function ‘Meta-analysis’, which
allows researchers to fully utilize these thousands of cancer samples, is also added. With the updated database, researchers will have the opportunity, for example, to identify a specific set of miRNAs for a distinct question if they could define samples according to a clarified biological or clinical goal. To achieve this, our ‘Meta-analysis’ function allows a user to select two groups of samples using dozens of clinical criteria and then identify differentially expressed miRNAs and arm-switching events for the two groups (Figure 2).

Putative novel miRNAs in YM500v2 were identified from thousands of samples, and more importantly, a new strategy was developed for filtering out false-positive discoveries. Compared with the previous version of YM500, YM500v2 has a lower number of novel miRNAs (from ~11 000 to ~3500), but has more novel miRNAs identified by three prediction algorithms (from 90 to 189). Moreover, 40.61% (1408/3467) of the putative novel miRNAs could be supported by CLIP-seq data, and a web interface displays the alignment results of both smRNA-seq and CLIP-seq.

Accumulating evidence suggests that a miRNA locus can produce a series of isomiRs during the miRNA biogenesis process (24–27). The phenomenon of isomiRs contributes largely to the dynamic miRNome and constantly presents a challenge for miRNA study. A few isomiRs have been proven to be functional regulatory molecules (24,27–31), and more studies about isomiRs should be performed and taken into consideration (32). Several in silico tools have been developed for identifying isomiRs from smRNA-seq data (33–36), and such tools will greatly promote the discovery of isomiRs and their relevant functional roles. YM500v2 provides the existing evidence of isomiRs from enormous smRNA-seq datasets for researchers whenever they discover significant isomiRs in their studies. A representative example of isomiRs is shown in Figure 3.

NGS has become the norm for large-scale cancer research, and cancer-related smRNA-seq data will accumulate rapidly in the future. YM500 will be updated periodically to incorporate new smRNA-seq data. We have developed a pipeline to process new data and to incorporate them into the database semiautomatically. Newly incorporated smRNA-seq data will also be re-annotated for metanalysis. YM500 will continue to be an informative and valuable database for miRNA studies.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT
We are grateful to the National Center for High-performance Computing for computer time and facilities and thank the TCGA research network for the availability of data.

FUNDING
Ministry of Science and Technology (MOST) [101-2320-B-010-059-MY3, 102-2314-B-010-045, 102-2321-B-010-010, 103-2622-B-010-002]; National Health Research Institutes [NHRI-EX102-1025451]; Veterans General Hospitals University System of Taiwan (VGHUST) Joint Research Program; Tsou’s Foundation [VGHUST102-G7-3-2]; UST-UCSD International Center for Excellence in Advanced Bioengineering sponsored by the Taiwan NSC I-RiCE Program [102-2911-I-009-101]; National Yang-Ming University (Ministry of Education, Aim for the Top University Plan). Funding for open access charge: MOST [101-2320-B-010-059-MY3, 102-2314-B-010-045, 102-2321-B-010-010, 103-2622-B-010-002]; National Health Research Institutes [NHRI-EX102-1025451]; VGHUST Joint Research Program; Tsou’s Foundation [VGHUST102-G7-3-2]; USTUCSD International Center for Excellence in Advanced Bioengineering sponsored by the Taiwan NSC I-RiCE Program [102-2911-I-009-101]; National Yang-Ming University (Ministry of Education, Aim for the Top University Plan).

Conflict of interest statement.
None declared.

REFERENCES
Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.


