All of gene expression (AOE): integrated index for public gene expression databases

Hidemasa Bono
Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Abstract

Gene expression data have been archived as microarray and RNA-seq data in two public databases, Gene Expression Omnibus (GEO) and ArrayExpress (AE). In 2018, DNA DataBank of Japan started similar repository called Genomic Expression Archive (GEA). These databases are useful resource for the functional interpretation of genes.

But, these databases have been separately maintained and these may lack RNA-seq data whose original sequence data is available in the sequence read archive (SRA).

We constructed an index for those gene expression data repositories, called all of gene expression (AOE) to integrate publicly available gene expression data. The web interface of AOE can graphically query data in addition to the application programming interface. By collecting gene expression data by RNA-seq from SRA, AOE also includes data not included in GEO and AE.

AOE can be accessible as a search tool from GEA website and is freely available from http://aoe.dbcls.jp/.

Introduction

After the invention of microarray, it became possible to measure the abundance of all transcripts in the genomic scale, which is now called transcriptome. Since then, gene expression data from those experiments have been archived in the public repository. They are the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) [1] and the EBI ArrayExpress (AE; https://www.ebi.ac.uk/arrayexpress/) [2].

Unlike the International Nucleotide Sequence Database [3], these two databases for gene expression have not exchanged their data each other while AE once had imported data in GEO. In that situation, users are needed to search both databases in order to get data from public gene expression data of interest thoroughly because these databases have been separately maintained. Furthermore, DNA DataBank of Japan (DDBJ) recently started similar repository called the Genomic Expression Archive (GEA; https://www.ddbj.nig.ac.jp/gea/) [4]. The integration of public gene expression databases is really needed.

In addition, these databases may lack transcriptome data by sequencing (RNA-seq) where original sequence data is accessible in the nucleotide sequence repository of high-throughput sequencing platforms, Sequence Read Archive (SRA) [5]. It is because the data deposition to GEO and AE is not mandatory while the original sequencing data is deposited to SRA.

We therefore developed an index of public gene expression databases, called All of gene expression (AOE). The aim of AOE is to integrate gene expression data and make them searchable at a time.We have maintained AOE for 5years, and it turned out to be useful for functional genomics researches.We report here the detailed description and utility of AOE. AOE is freely accessible from http://aoe.dbcls.jp/ .

Results

Status of gene expression databases

Gene expression data in NCBI Gene Expression Omnibus (GEO) had been continuously imported into EBI ArrayExpress (AE), and thus we were able to get all data deposited to GEO from AE theoretically. Therefore, All Of gene Expression (AOE) was originally indexed for AE only.

Meanwhile, AE discontinued importing GEO data in 2017. At that point, we investigated data-series entries in these two databases by matching GEO series IDs (GSExxxxx). It showed that there were over thirty thousand entries missing in ArrayExpress (Figure 1). Furthermore, even GEO did not represent whole transcriptome data in public as over ten thousand entries in AE were missing in GEO. Thus, we decided to include those missing entries in AOE. In other words, we started indexing GEO data and other public transcriptome data including DDBJ Genomic Expression Archive (GEA) for the search of all gene expression data in public.

An index of gene expression data series from metadata

AOE was originally developed to give graphical web interface to search EBI ArrayExpress, which is one of public databases of gene expression data as described above. We call this dataset only from AE 'AOE level1'. Data for this level basically contain only ID for AE, and the entries imported from GEO contain IDs for BioProject and GEO also.

After the import of GEO data to AE was discontinued, AOE start importing GEO data directly utilizing DBCLS SRA API [6]. By subtracting existing GEO data in AE, new entries have been included to AOE. We call the merged dataset which included GEO data 'AOE level2'. Data for this level contain IDs for BioProject and GEO, not that for AE.

There are still missing gene expression data which are not included in AE and GEO, but are registered as transcriptome sequencing data in SRA. The final merged dataset is called 'AOE level3' which represents a real public gene expression dataset. Data for this level contain BioProject ID only.

Figure 2 shows schematic view of data flow to make index data for AOE summarized above.

As AOE was designed to index the public gene expression data, 'experimental series'-wise data have been indexed for the search. Individual hybridization data for microarray and run data for RNA-seq are directly linked to original databases. All codes to parse public databases are freely accessible from DBCLS AOE GitHub page (https://github.com/dbcls/AOE/).

Graphical web interface

Gathering all three levels of data described above, AOE enables visualization and exploration of gene expression data. AOE provides an interactive web interface (http://aoe.dbcls.jp/) to retrieve data of interest (Figure3). Users can easily filter data by organisms and quantification methods.

For example, users can search by keyword 'hypoxia'. AOE currently reports that 436 items with three histograms (by year, by organism and by quantification method). After looking at the histograms, user can limit the data for 'Homo sapiens' by dragging the bar in the histogram by organism because there were many items in 'Homo sapiens'. Then, AOE redraws the histograms with the selected data. Additionally, user can limit the data for 'Illumina' by dragging the bar in the histogram by quantification method. The selected data can be retrieved by clicking 'Retrieve' button.

A tutorial movie "How to use AOE: All Of gene Expression" is available from https://doi.org/10.7875/togotv.2018.146. The movie originated from the contents of TogoTV, which provides tutorial videos for useful databases and web tools in the life sciences [7], and is available on the TogoTV original website (https://togotv.dbcls.jp/en/) and YouTube (https://youtube.com/togotv/).

Application programming interface

AOE can also query via Application programming interface (API). Precise description for AOE API is available from AOE website or directly from DBCLS AOE GitHub website at https://github.com/dbcls/AOE/blob/master/API_documentation.md.

Discussion

We have developed and maintained an index of public gene expression databases, called All of gene expression (AOE).

AOE was originally started as an index for ArrayExpress (AE) maintained at EBI because AE had exported gene expression data in Gene Expression Omnibus (GEO) which is the largest gene expression database maintained at NCBI. That means we could find all gene expression data in AE including those deposited to GEO.

Newly submitted data contain BioProject IDs, and this feature makes it possible to integrate multiple levels of indices and resolve complicated relationships among IDs while old AE entries do not have BioProject ID.

Existence of many data in AOE level3 shows that not all gene expression data by sequencing are stored in GEO. This indicates that GEO is not enough for whole public gene expression database. But, many data in AOE level3 are heterogenous and metadata for those can lack several descriptions, which are curated and cleanly described in GEO and AE.

Similar approach is also undertaken in EBI called the Omics Discovery Index (OmicsDI; https://www.omicsdi.org/), which provides a knowledge discovery framework across heterogeneous omics data (genomics, proteomics, transcriptomics and metabolomics) [8]. OmicsDI aims to integrate various types of omics data, and is not focused on gene expression data.

AOE is focused on gene expression data. It is also designed to be a search interface for DDBJ Genomic Expression Archive (GEA), and linked from the official website for GEA.

Because the web interface for AOE is simple and user friendly, AOE can be used for biologists who are not familiar with database search. AOE can also be used for professionals to make reference expression datasets for specific organisms. We have also developed Reference Expression dataset (RefEx) for human and mouse [9]. We are planning to implement RefEx for other organisms making use of these reference expression datasets retrieved by AOE.

Materials and methods

Acquisition of public gene expression data

AOE consists of two major types of data source. One is EBI ArrayExpress (AE), and the other is data in NCBI including Gene Expression Omnibus (GEO).

For AE data type, several files are required to make AOE index, which are Array Design Format (ADF), Investigation Description Format (IDF) and Sample and Data Relationship Format (SDRF) files.These files are routinely acquired from ArrayExpress FTP site (ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/data/). ADF files are located in subdirectories in ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/data/array/ with file extension .adf.txt. IDF and SDRF files are located in subdirectories in ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/data/experiment/ with file extensions .idf.txt and .sdrf.txt respectively.

For data from NCBI, in addition to the file describing ID relationships in the Sequence Read Archive (SRA) named SRA_Accessions.tab from ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Metadata/ , metadata from SRA, BioProject and Biosample are used to make index for AOE as all data from GEO have BioProject IDs and BioSample IDs even if the gene expression quantification for that data was not transcriptome sequencing.

Organizing metadata from different sources

For AE type of data, ADF, IDF and SDRF files are required to make index for AOE. Data from DDBJ Genomic Expression Archive (GEA) also consist of AE type of data and available from its FTP site (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/gea/). We make use of AE type of data to construct an initial set of AOE index (called AOE level1).

GEO data in the Sequence Read Archive (SRA), BioProject and Biosample are used to make index for AOE. These data have been stored in the DBCLS SRA and the application programming interface (API) for metadata for those have also been maintained in DBCLS. AOE uses this API to retrieve data needed to make the index (AOE level2).

Finally we collected RNA-seq data in SRA making use of DBCLS SRA API. Most of this fraction of data are in AOE level2, but not a few entries can be found in this filter (AOE level3).

Data parsers to make a tab-delimited text file for visualization are implemented in Perl and UNIX commands. All shell and Perl scripts for those can be accessible from GitHub (https://github.com/dbcls/AOE/).

Visualization of datasets

For visualizing datasets, we employed specially coded Python3 scripts and we also use D3.js a JavaScript library for manipulating documents based on data (https://d3js.org/). It enables data selection by mouse operation. For example, user can select data by release date by dragging the histogram generated in the keyword search.

Acknowledgement

Author wishes to thank Dr. Naoya Oishi for the design and development of web and application programming interface for AOE, Dr. Yuichi Kodama for helpful advice for the integration of DDBJ Genomic Expression Archive (GEA), Dr. Takeru Nakazato and Dr. Tazro Ohta for their help to use DBCLS SRA API for updating AOE entries from NCBI databases. The tutorial movie for AOE web interface was created under the TogoTV project in DBCLS with editorial direction by Dr. Hiromasa Ono. The computing resource was partly provided by the super computer system at National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS), Japan.

Funding

This work was supported by National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST).

References

1. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. **41**, D991–D995 (2013).
2. Kolesnikov, N. et al. ArrayExpress update--simplifying data submissions. *Nucleic Acids Res*. **43**, D1113-1116 (2015).
3. Karsch-Mizrachi, I., Takagi, T., & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res*. 46, D48-D51 (2018) DOI: 10.1093/nar/gkx1097
4. Kodama, Y. et al. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Res*. **47**, D69–D73 (2019).
5. Kodama, Y., Shumway, M. & Leinonen, R. International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
6. Ohta, T., Nakazato, T. & Bono, H. Calculating quality of public high-throughput sequencing data to obtain suitable subset for reanalysis from the Sequence Read Archive *GigaScience*, **6**, gix029 (2017). DOI: 10.1093/gigascience/gix029
7. Kawano, S., Ono, H., Takagi, T. & Bono, H. Tutorial videos of bioinformatics resources: online distribution trial in Japan named TogoTV. *Brief Bioinform.* **13**, 258-268 (2012)
8. Perez-Riverol, Y. et al. Discovering and linking public comics data sets using the Omics Discovery Index. *Nat. Biotech.* **35**, 406-409 (2017)
9. Ono, H., Ogasawara, O., Okubo, K. & Bono, H. RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. *Sci. Data* **4**, 170105 (2017)

Figures

Figure 1

The number of overlapping data-series entries in ArrayExpress (left) and GEO (right).

Figure 2

Three levels of AOE data flow to make AOE index. Level1 from ArrayExpress data. Level2 from GEO data in SRA via DBCLS SRA API. Level3 from RNA-seq data in SRA not in GEO via DBCLS SRA API.

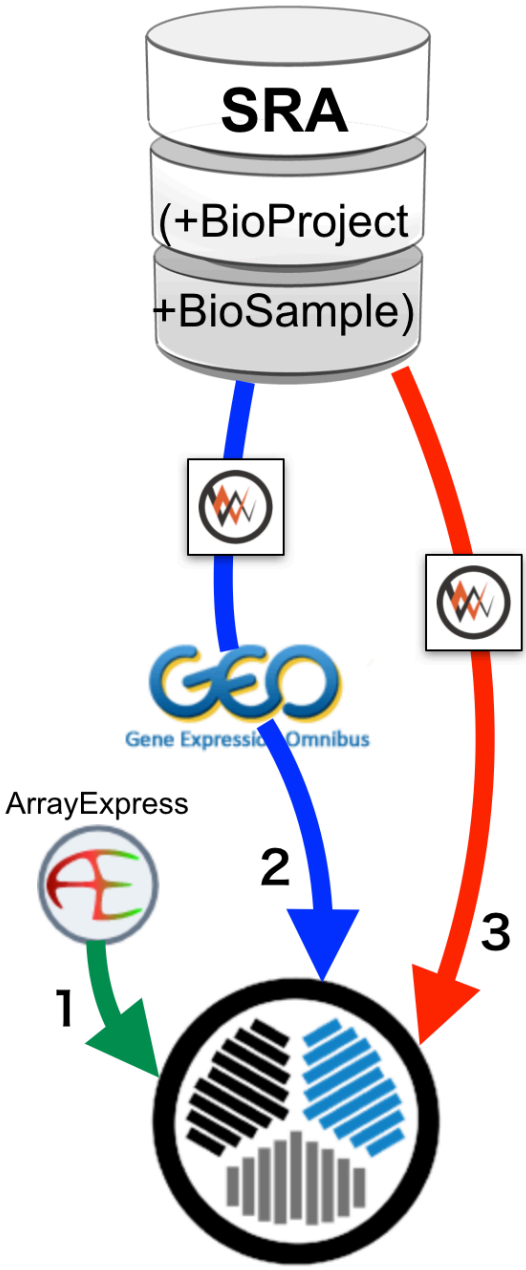
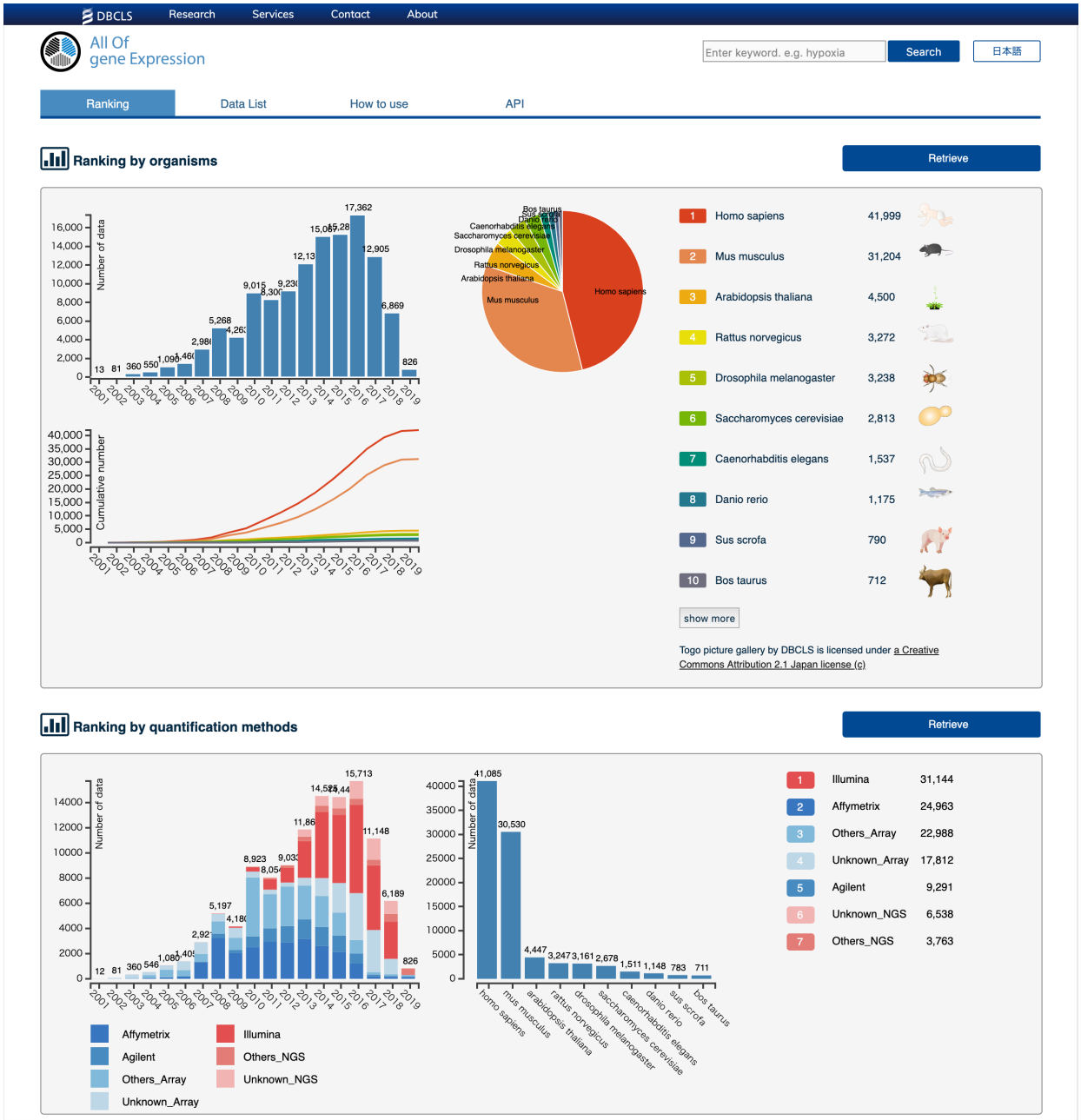


Figure 3

Web interface of AOE.