All of gene expression (AOE): an 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

Email: bono@dbcls.rois.ac.jp

Abstract

Background

Gene expression data have been archived as microarray and RNA-seq datasets in two public databases, Gene Expression Omnibus (GEO) and ArrayExpress (AE). In 2018, the DNA DataBank of Japan started a similar repository called the Genomic Expression Archive (GEA). These databases are useful resources for the functional interpretation of genes. However, these databases have been separately maintained, and may lack RNA-seq data with original sequence data available in the Sequence Read Archive (SRA).

Findings

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 the SRA, AOE also includes data not included in GEO and AE.

Conclusions

AOE is accessible as a search tool from the GEA website and is freely available at https://aoe.dbcls.jp/.

Keywords

Transcriptome, public database, RNA-seq

Findings

Introduction

After the invention of the microarray, it became possible to measure the abundance of all transcripts at the genomic scale, which is now called the transcriptome. Since then, gene expression data from those experiments have been archived in public repositories. 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 stopped exchanging data with each other since AE stopped importing data from GEO. Therefore, users need to search both databases to get comprehensive public gene expression data of interest, because these databases have been separately maintained. Furthermore, the DNA DataBank of Japan (DDBJ) recently started a similar repository called the Genomic Expression Archive (GEA; https://www.ddbj.nig.ac.jp/gea/) [4]. Integration of these public gene expression databases is required.

In addition, these databases may lack sequencing transcriptome data (RNA-seq) where the original sequence data are accessible in the nucleotide sequence repository of high-throughput sequencing platforms, the Sequence Read Archive (SRA) [5]. This is because data deposition to GEO and AE is not mandatory when the original sequencing data are deposited to the 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 all searchable together. We have maintained AOE for five years, and it has been useful for functional genomics research. Here, we report a detailed description and utility of AOE. AOE is freely accessible from https://aoe.dbcls.jp/.

Status of gene expression databases

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

However, AE discontinued GEO data import 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 publicly represent whole transcriptome data, 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 the DDBJ Genomic Expression Archive (GEA), to allow all public gene expression data to be searched.

An index of gene expression data series from metadata

AOE was originally developed to give a graphical web interface to search EBI ArrayExpress, which is one of the public gene expression databases described above. We call this dataset from AE only 'AOE level 1'. Data at this level basically contain only IDs for AE, and the entries imported from GEO contain IDs for both BioProject and GEO.

After the import of GEO data to AE was discontinued, AOE began importing GEO data by directly utilizing DBCLS SRA application programming interface (API) [6]. By subtracting the GEO data already existing in AE, new entries were included in AOE. We call the merged dataset that includes GEO data 'AOE level 2'. Data at this level contain IDs for BioProject and GEO, not for AE.

There were still gene expression data missing that were not included in AE and GEO, but were registered as transcriptome sequencing data in SRA. The final merged dataset is called 'AOE level 3', and represents a real public gene expression dataset. Data at this level contain BioProject IDs only.

Figure 2 shows a schematic view of the data flow to make the index data for AOE, as summarized above. The constructed data are archived at the Life Science Database Archive at the National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST), and is available with the DOI: 10.18908/lsdba.nbdc00467-000 (https://doi.org/10.18908/lsdba.nbdc00467-000).

As AOE was designed to index 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 the original databases. All codes to parse public databases are freely accessible from the 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 (https://aoe.dbcls.jp/) to retrieve data of interest (Figure 3). Users can easily filter data by organism and quantification method.

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

A tutorial movie "How to use AOE: All Of gene Expression" is available at 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). A precise description for the AOE API is available from the AOE website or directly from the 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 begun as an index for the ArrayExpress (AE) database maintained at EBI, because AE had exported gene expression data from Gene Expression Omnibus (GEO), which is the largest gene expression database maintained at NCBI. That meant that AE contained all gene expression data, 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.

The existence of a great deal of data at AOE level 3 shows that not all sequencing gene expression data are stored in GEO. This indicates that GEO is insufficient as a complete public gene expression database. However, much of the data at AOE level 3 are heterogenous and metadata for those can lack several descriptions, which are curated and cleanly described in GEO and AE.

A similar approach has also been undertaken by 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 the DDBJ Genomic Expression Archive (GEA), and a link to AOE can be found on the official GEA website.

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

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 the Gene Expression Omnibus (GEO).

For the AE data type, several files are required to make an AOE index. These are Array Design Format (ADF), Investigation Description Format (IDF), and Sample and Data Relationship Format (SDRF) files. These files are routinely acquired from the 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 the 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 an 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 the AE type of data, ADF, IDF and SDRF files are required to make an index for AOE. Data from the DDBJ Genomic Expression Archive (GEA) also consist of the AE type of data, and are available from its FTP site (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/gea/). We made use of the AE type of data to construct an initial AOE index set (called AOE level 1).

GEO data in the Sequence Read Archive (SRA), BioProject and BioSample are used to make an index for AOE. These data have been stored in the DBCLS SRA, and the application programming interface (API) for metadata for those has also been maintained in DBCLS. AOE used this API to retrieve data needed to make the index (AOE level 2).

Finally, we collected RNA-seq data in SRA, making use of the DBCLS SRA API. Most of this fraction of data are in AOE level 2, but many entries can be found in this filter (AOE level 3). A concatenated tab-delimited file of the index constructed is archived in the Life Science Database Archive at National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST) with DOI: 10.18908/lsdba.nbdc00467-000 (https://doi.org/10.18908/lsdba.nbdc00467-000).

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 are accessible from GitHub (https://github.com/dbcls/AOE/).

Visualization of datasets

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

Availability of supporting source code

Project name: All Of gene Expression (AOE)
Project home page: https://aoe.dbcls.jp/en
GitHub page: https://github.com/dbcls/AOE/
Operating system: UNIX (macOS and linux)
Programming language: Python, Perl
License: MIT

Availability of supporting data

The constructed data are archived at the Life Science Database Archive at the National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST), and is available with the DOI: 10.18908/lsdba.nbdc00467-000 (https://doi.org/10.18908/lsdba.nbdc00467-000).
License: CC-BY 4.0

Acknowledgments

The author wishes to thank Dr. Naoya Oishi for the design and development of the web and application programming interface for AOE, Dr. Yuichi Kodama for helpful advice for the integration of the DDBJ Genomic Expression Archive (GEA), and Dr. Takeru Nakazato and Dr. Tazro Ohta for their help using the DBCLS SRA API for updating AOE entries from NCBI databases. The tutorial movie for the 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 supercomputer system at the National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS), Japan. The author thanks Catherine Perfect, MA (Cantab), from Edanz Group (www.edanzediting.com/ac), for editing a draft of this manuscript.

Funding

This work was supported by the 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 the AOE index. Level 1 from ArrayExpress data, level 2 from GEO data in SRA via DBCLS SRA API, and level 3 from RNA-seq data in SRA not in GEO via DBCLS SRA API.

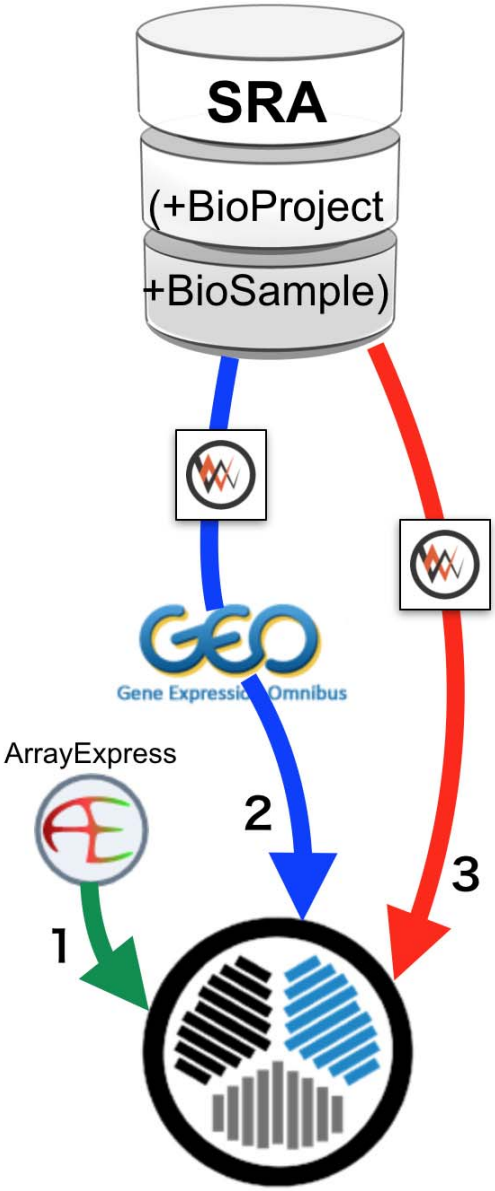
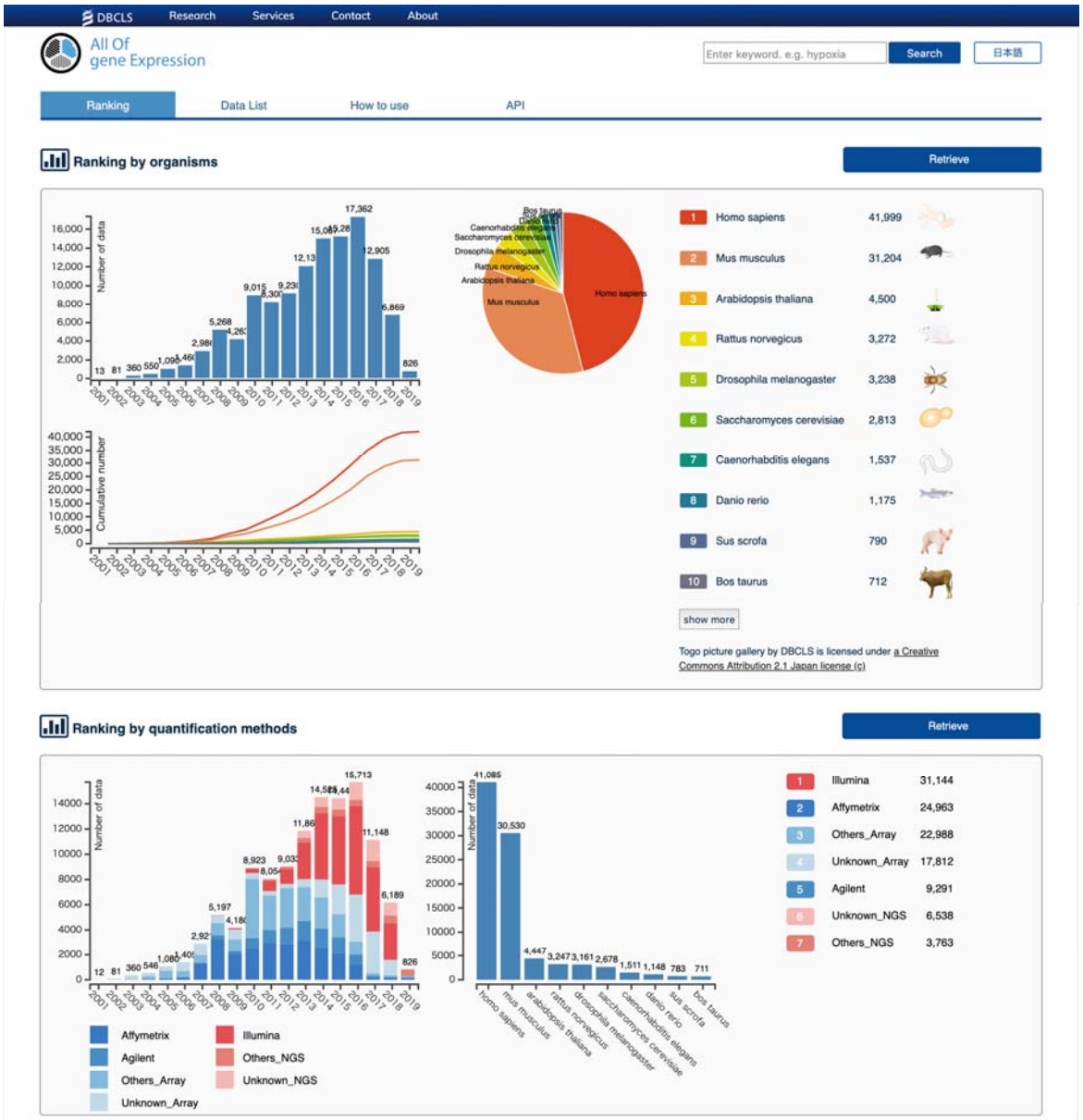


Figure 3

The AOE web interface.