TogoEx: the integration of gene expression data

Hidemasa Bono¹ and Takeya Kasukawa²

¹ DBCLS, Japan ² RIKEN, Japan

BioHackathon series:
BioHackathon EUROPE
Paris, France, 2019

Submitted: 24 Jun 2020

License
Authors retain copyright and
release the work under a Creative
Commons Attribution 4.0
International License (CC-BY).

Published by [BioHackrXiv.org](https://biohackxiv.org)

Abstract

Gene expression data have potential to obtain many new biomedical findings by reusing and integrating publicly available datasets since most of the datasets were usually analyzed only focusing on what their data producers have interests to (like genes of interests). However, we have several issues to overcome for the efficient reuse of the public expression data: current available metadata for gene expression data is not sufficient enough to re-use; and human resources are too lacking to tackle the integration of gene expression data. In the ELIXIR Biohackathon-Europe 2019, we discussed what was necessary to achieve a new analysis interface to utilize the FANTOM6 data (quantified transcriptomic responses in human dermal fibroblasts using Capped Analysis of Gene Expression (CAGE) by knock-down of long non-coding RNAs (lncRNAs)). We then considered possible solutions against the issues, and decided implementation plans and scheduled to implement it.

Introduction

Gene expression data have been archived in public repositories, the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) and the EBI ArrayExpress (AE; <https://www.ebi.ac.uk/arrayexpress/>). Unlike the International Nucleotide Sequence Database (<https://www.insdc.org/>), these two databases have not been exchanging data with each other. Furthermore, the DNA DataBank of Japan (DDBJ) in the National Institute of Genetics started a similar repository called the Genomic Expression Archive (GEA; <https://www.ddbj.nig.ac.jp/gea/>) in 2018. Hence there is a need for integration of these public gene expression databases. Thus, we have maintained an index of these public gene expression databases, called All Of gene Expression (AOE; <https://aoe.dbcls.jp/en>) (Bono, 2020) to integrate gene expression data and make them all searchable together.

During the DBCLS/NDBC 2019 BioHackathon in Fukuoka, Japan, we worked on the refinement of Reference Expression Dataset (RefEx; <https://refex.dbcls.jp/?lang=en>) (Ono, Ogasawara, Okubo, & Bono, 2017) for the gene expression analysis in order to make gene expression data reusable. The gene expression data to integrate in the hackathon were mainly project-based, and those included those of FANTOM5 (Kawaji, Kasukawa, Forrest, Carninci, & Hayashizaki, 2017), GTEx v8, fly and rice.

We felt that users wanted to reuse data from FANTOM6 project which was just published with bioRxiv preprint (<https://doi.org/10.1101/700864>; <http://fantom.gsc.riken.jp/6/>). The data consists of the large-scale expression profiles after the knock-down of long non-coding RNAs (lncRNAs) using several technologies including antisense oligonucleotide (ASO) and small interfering RNA (siRNA).

The original purpose of the knock-down experiments in the FANTOM6 project was identifying what genes are affected by the perturbation of a specific lncRNAs, and then predicting the functions of the lncRNAs. However, other researchers can reuse the data for annotating their lists of genes with specific phenotypes (for example, up-/down-regulated in specific

experimental conditions/tissues). They can compare their list of genes with the gene sets affected by the knock-down of a lncRNA of interest and can find potential key regulators related to the phenotype.

There are several hurdles to establish an interface for the analysis described above. The first issue concerns the metadata of the FANTOM6 datasets. The current FANTOM6 metadata is designed only for understanding the dataset by readers of the FANTOM6 preprints and papers. Thus, it is not enough for reusing them. For example, the sample information is not annotated based on well-known ontologies (e.g. cell ontology, cell line ontology, and tissue ontology). Another issue is related to a social aspect of the data management. Human resources for this type of studies are always lacking especially in the field of the biological data engineers and curators, it is not easy to perform the integration of large-scale gene expression data.

We tackled the problems in the reuse of the FANTOM6 data at the ELIXIR BioHackathon-Europe 2019 by pointing out what are existing issues with their possible solutions, and we decided plans and schedules for the implementation of the search interface.

Results

During the ELIXIR BioHackathon-Europe 2019, we discussed the following points about the reuse of gene expression data.

1. Requirements in gene expression data production.
2. Reusing Cap Analysis of Gene Expression (CAGE) data for lncRNA knock-down from FANTOM6 in other research projects.

Requirements in gene expression data production.

We clarified what is necessary in the data production of gene expression data, especially for the large-scale projects including FANTOM, GTEx and Human Cell Atlas. For making large-scale reusable expression data for many research projects, rich “metadata” is one of the most important resources to understand each dataset. The metadata contains the information of the source samples, treatment of the samples, sequencing conditions, and bioinformatics analysis procedures, and these are used to choose suitable datasets for individual research projects, and to properly evaluate the results.

For the data production in the large-scale projects, it usually takes longer time spanning a few years, and each expression data for a sample is gradually added to the finished dataset. The process includes sample collection, sample treatment, sequencing, and bioinformatics analysis, which starts in various timings and sometimes requires the starting over of the process due to experimental failures. Its metadata must be collected at the same timing along with each data production processes, and it should not do after the completion of all data production due to several reasons. For example, the source of metadata could only be found as handwritten texts in lab notes, and technicians and research scientists who did experiments might be left from laboratories. However, the simultaneous metadata collection is not straightforward. If experiment protocols are changed during the project, the format of metadata should be adjusted to new protocols. If the nomenclature of terms in the metadata is not fixed, the metadata could be varied by researchers who prepared it. Thus, the metadata collection is a long-term and laborious work.

To be able to produce high-quality metadata, we need good data engineers and data curators. The data engineers implement systems to collect necessary information from wet researchers and technicians. The data curators decide and adjust rules for the data and metadata format. The data curators also make correction of the data and the metadata both manually and computationally if there are mistakes and errors in the data. However, especially in Japan, we

are not easy to hire such skilled people. Possible reasons might be: we have not provided good career paths for them; and we have not developed good training courses for learning required skills for data engineers and curators. This is a social problem, and we need much efforts and long time to solve them.

Reusing Cap Analysis of Gene Expression (CAGE) data for lncRNA knock-down from FANTOM6 in other research projects.

The FANTOM6 consortium performed knock-down (KD) experiments with lncRNA antisense oligonucleotides, followed by the expression profiling using the CAGE protocol (KD-CAGE). The CAGE protocol can identify active transcription start sites in the genome with their expression (i.e. promoter activity) levels. The original purpose of the lncRNA KD-CAGE is for predicting the function of each lncRNA by observing the changes of expression levels after its knocking down. In the pilot phase of the FANTOM6 project, they performed around 300 KD-CAGE assays with the human dermal fibroblast cells. The data of both raw and analysis results is available in the public repository and their web site (<https://fantom.gsc.riken.jp/6/>).

Although the original aim of the lncRNA KD-CAGE is to predict the function of lncRNAs, the expression profiles can potentially be reused for other analysis. For example, if the genes changed by the knock-down of a specific lncRNA are matched with the list of differentially expressed genes in another biological phenotype/stimulus, we can infer the relationship between the phenotype or stimulus and the lncRNA.

To achieve such type of the analyses, we discussed the actual method for those. Through the discussion, we concluded that an enrichment analysis may be suitable for this purpose. A user input is a list of interesting genes. The list is compared with the lists of lncRNA-related genes obtained by the (re-)analysis of the FANTOM6 KD-CAGE expression profiles, and (adjusted) P-values of enrichments are then calculated with Fisher's exact test or a similar statistical method. To achieve the method, we need to do the following tasks as a first step. First, we retrieve the list of lncRNA-related genes from the FANTOM6 KD-CAGE expression profiles. Next, we implement a web-based system for performing the analysis, which is very useful for wet researches who do not have enough computational skills.

In addition to the above functions, the system should also support or consider the following extensions. First, the system should be extensible to the expansion of the FANTOM6 KD dataset. The FANTOM6 project is ongoing, and more data will be released later, using different lncRNAs and cell types. Along with the future releases of the datasets, the system should be able to easily process and load the new additional data. Second, the system should incorporate public KD datasets in addition to the FANTOM6 data. The analysis is not limited to the FANTOM6 KD-CAGE data. Thus, if we can collect more KD data from public repositories, we can make the system to be able to provide more rich results. These are future plans after the prototype system implementation.

We estimate the development of the method and the prototype system can be achieved in 2020. We will keep discussion and evaluation of the method.

Conclusion

After a stimulus discussion in the ELIXIR BioHackathon-Europe 2019, we could clarify the current issues in the transcriptome data production, and we also could successfully make a plan to reuse the CAGE data for lncRNA knock-down in FANTOM6 project. While the Reference Expression dataset in DBCLS (RefEx) currently includes transcriptome data for normal tissues and cell lines, we discussed how to reuse transcriptome data for knock-down. This is only a

starting point to reuse public knock-down data, but these 'reuse-ready' data can be useful for molecular biology in the near future.

Future work

We will continue current effort to integrate gene expression data to increase the usability of these data in life science.

By gathering and then curating gene expression datasets, those data not in NCBI GEO, EBI AE and DDBJ GEA can be included in AOE. We will also work for it.

Jupyter notebooks created

We tried to code with Jupyter Notebook in bash kernel, but it seems that CWL is suitable for our purpose. We are willing to use BioHackrXiv to publish our activity in the ELIXIR BioHackathon-Europe 2019.

Acknowledgement

This work was funded by ELIXIR, the research infrastructure for life-science data, as part of the ELIXIR BioHackathon-Europe 2019.

References

- Bono, H. (2020). All of gene expression (AOE): An integrated index for public gene expression databases. *PLOS ONE*, 15, e0227076. doi:[10.1371/journal.pone.0227076](https://doi.org/10.1371/journal.pone.0227076)
- Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P., & Hayashizaki, Y. (2017). The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci Data*, 4, 170113. doi:[10.1038/sdata.2017.113](https://doi.org/10.1038/sdata.2017.113)
- Ono, H., Ogasawara, O., Okubo, K., & Bono, H. (2017). RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. *Sci Data*, 4, 170105. doi:[10.1038/sdata.2017.105](https://doi.org/10.1038/sdata.2017.105)