

Indices for NGS Data and Gene Expression Data Registered in Public Databases



Tazro Ohta, Takeru Nakazato, Hidemasa Bono
bono@dbcls.rois.ac.jp

Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

DBCLS SRA (DDBJ search) is a search engine for public NGS data. It provides a project list categorized by study types, sequencing platforms, and sample species.



<http://sra.dbcls.jp/>



<https://github.com/dbcls/DDBJSearch/>



Search by

- Keywords
- IDs in public DB

SRA

- SRA (example: DRA008013)
- SRP (example: SRP045624)
- SRX (example: ERX182686)
- SRR (example: ERR266349)
- BioProject (example: PRJDB4976)
- BioSample (example: SAMD00079763)

keyword: *RCC4*

keyword: *hypoxia*

Accession: *DRA008013*

Direct link to RUN file in SRA format

BioSample:

BioProject:

JSON data

Paper for DDBJ search

<https://www.ddbj.nig.ac.jp/dra/submission.html>

Abstract

It is important for public data repositories to promote the reuse of archived data. In the growing field of omics sciences, however, the increasing number of submissions of high-throughput sequencing (HTS) data to public repositories prevents users from choosing a suitable data set from among the large number of search results. Repository users need to be able to sort a thousands of results to obtain a suitable subset of high quality data for reanalysis. We calculated the quality of sequencing data archived in a public data repository, the Sequence Read Archive (SRA), by using the quality control software FastQC. We obtained quality values for 1,75,752 experiments, which can be used to evaluate the suitability of data for reuse. We also visualized the data distribution in SRA by integrating the quality information and metadata of experiments and samples. We provide quality information of all of the archived sequencing data, which enable users to obtain sufficient quality sequencing data for reanalysis. The calculated quality data are available to the public in various formats. Our data also provide an example of utilization of the reuse of public data by adding metadata to published research data for a third party.

Keywords: high-throughput sequencing, sequencing quality, public data, database

<https://doi.org/10.1093/gigascience/gix029>

AOE (All Of gene Expression) is an index of public gene expression databases (EBI ArrayExpress + NCBI Gene Expression Omnibus (GEO) + DDBJ Genomic Expression Archive (GEA)).



<https://aoe.dbcls.jp/en>



<https://github.com/dbcls/AOE/>



AOE top page

Limit data by...

Browse by organisms

Search results

Link to original databases

Limit data for RNA-Seq

Limit data for microarray

Limit data for 2017 and 2018

Limit data for Illumina (RNA-Seq)

Retrieve data

Select 'Oryza sativa'

Retrieve species-specific data (Oryza sativa)

Paper for AOE is in bioRxiv!

<https://doi.org/10.1101/626754>

Images @ © 2016 DBCLS TogoTV / CC-BY-4.0

	# of series	Cumulative
1 AE (including GEO) + GEA	72,770 (GEO-derived 59,374)	72,770
2 GEO not in AE	37,849	110,619
3 SRA (RNA-seq) not in AE + GEO	13,137	123,756

