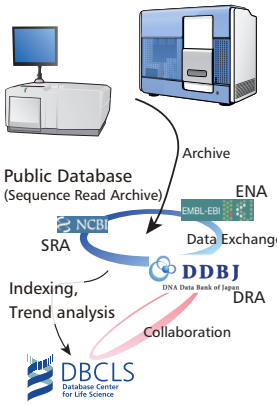
Database Center for Life Science (DBCLS)
Research Organization of Information and Systems (ROIS)

DBCLS SRA: Functional mining and characterization of public NGS data

Takeru Nakazato*, Tazro Ohta, Hidemasa Bono
Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), JAPAN
*: nakazato@dbcls.rois.ac.jp

Backgrounds

Next generation Sequencer



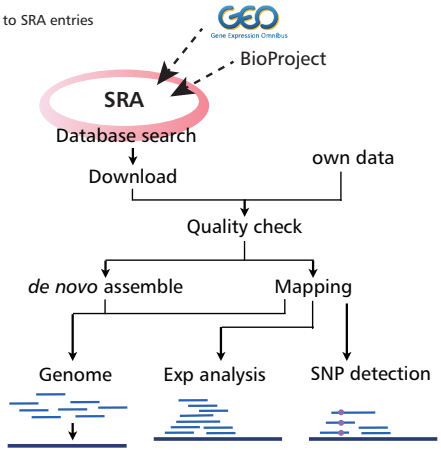
The next-generation sequencing (NGS) data is archived in public database, namely the sequence read archive (SRA), and the data is collaborately maintained by DDBJ, EBI, and NCBI. In Japan, Database Center for Life Science (DBCLS) has developed infrastructure for researchers to access and re-use these data easily by providing index and stats pages and constructing a portal site for life science databases and tools in collaboration with DDBJ.

Solutions

FREE!

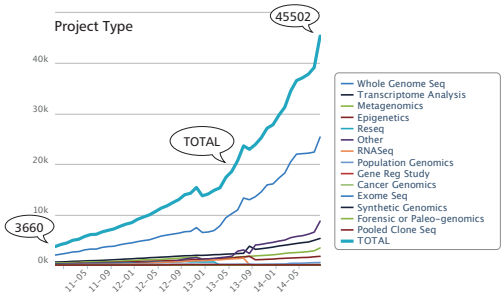
DBCLS SRA
http://sra.dbcls.jp/

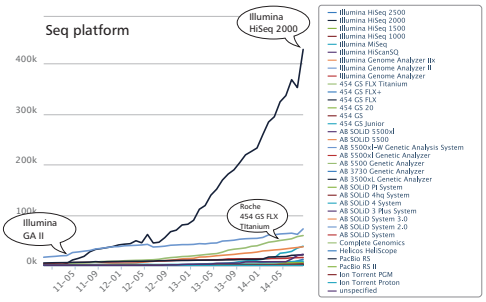
Exp-design based characterization and quality check of public NGS data
Search engine for public NGS data
Also available publication list referring to SRA entries



Statistics

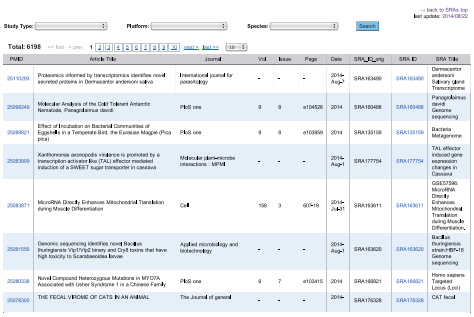
(as of Aug. 15, 2014)





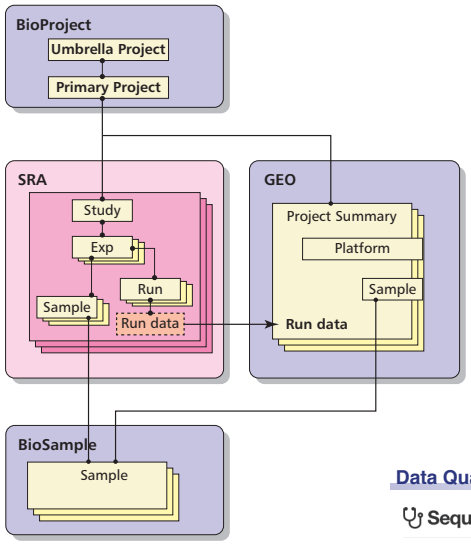
Publication List

to get NGS data with sufficient good quality to publish articles



Publications using NGS | Corresponding NGS data

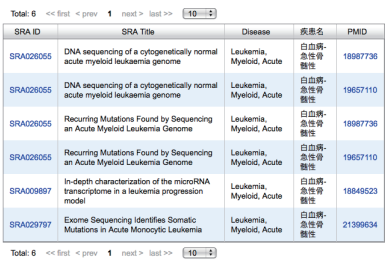
The data structure of SRA



Species (top 15)

Homo sapiens	262584
Mus musculus	50126
human gut metagenome	21731
human metagenome	18150
Plasmodium falciparum	17468
Staphylococcus aureus	15843
Streptococcus pneumoniae	15374
Saccharomyces cerevisiae	12545
soil metagenome	11750
Drosophila melanogaster	10728
Danio rerio	10293
Anopheles gambiae	8231
Mycobacterium tuberculosis	8086
rhizosphere metagenome	7727
Caenorhabditis elegans	7392
Total	759588 (experiments)

NGS data relevant to diseases

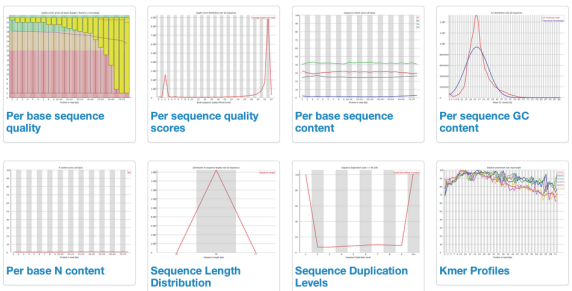


Public NGS data | Disease | Corresponding publication

Recently, some NGS run data is not reposited to SRA but GEO and GenBank. Moreover, project and sample information is captured in external other databases: BioProject and BioSample. This complexity of data structure prevents researchers from retrieving the NGS data of interests.

Data Quality

Sequence Quality Statistics



- The results of FastQC is pre-calculated and presented in detail view.
- Researchers are not only spared the problem of downloading low quality data but also able to skip checking process of data quality.

Reference:
Experimental design-based functional mining and characterization of high-throughput sequencing data in the Sequence Read Archive.
Nakazato T., Ohta T., Bono H.,
PLOS One, 8 (10): e77910 (2013)
PMID: 24167589

ECCB'14
(European Conference on Computational Biology)
Strasbourg Convention Centre, Strasbourg, France
Sep. 7-10, 2014