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

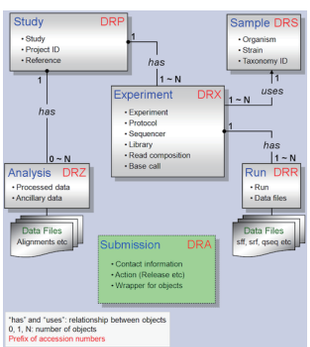
SRAs: The Survey of Read Archives



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

Backgrounds and motivations

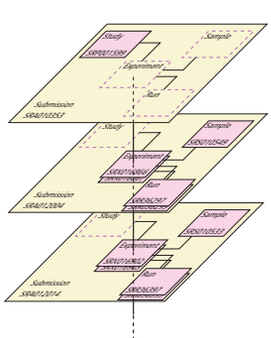
The data structure of SRA



Original content on http://trace.ddbj.nig.ac.jp/dra/documentation_e.shtml

The next-generation sequencing (NGS) data is archived in SRA, ENA, and DRA as public repositories. The doposited NGS data contains not only sequence read sequences but also conditions of experiments including project title, species or cell line names of samples, and sequencing platforms as a meta data. The meta data consists of six files with XML format: submission, study, experiment, run, sample, and analysis.
However, each submission has not all of those meta data because additional experiments or runs to be assigned to a previous project are often performed and reposited as a new submission.

Submission	Study	Experiment	Run	Sample	Analysis	
✓			✓			19856
✓				✓		14904
✓	✓	✓	✓	✓		6474
✓		✓			✓	4173
✓		✓				2180
✓		✓	✓			1543
✓	✓					621
✓						410
✓		✓	✓	✓		315
✓					✓	182
✓	✓	✓		✓		141
✓		✓	✓		✓	97
✓	✓	✓	✓	✓	✓	72
			⋮			
			Total (submissions)			51147



Results and Discussions

Statistics (as of Oct. 14, 2011)

Study Types

Whole Genome Sequencing	4072
Transcriptome Analysis	1128
Metagenomics	844
Epigenetics	575
Resequencing	410
Other	354
RNASeq	204
Population Genomics	128
Gene Regulation Study	47
Cancer Genomics	27
Exome Sequencing	20
Pooled Clone Sequencing	6
Synthetic Genomics	5
Forensic or Paleo-genomics	5
Total (studies)	7825

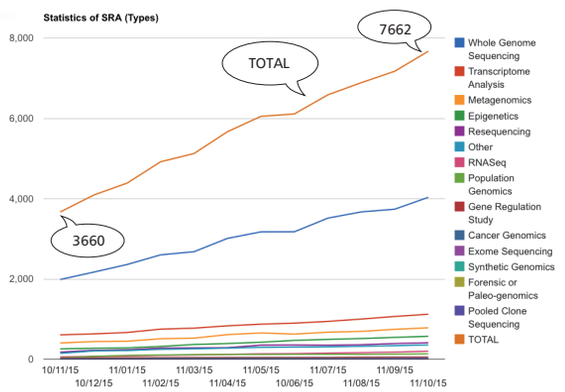
Platforms

Illumina Genome Analyzer II	2407
454 GS FLX Titanium	2128
454 GS FLX	1517
Illumina Genome Analyzer	912
Illumina HiSeq 2000	693
454 GS 20	314
Illumina Genome Analyzer IIx	192
AB SOLiD System 3.0	74
unspecified	54
AB SOLiD System 2.0	50
AB SOLiD 4 System	34
AB SOLiD System	32
Helicos HeliScope	18
454 GS	10
Ion Torrent PGM	8
PacBio RS	3
Complete Genomics	3
Illumina HiSeq 1000	1
454 GS Junior	1
Total (studies)	8451

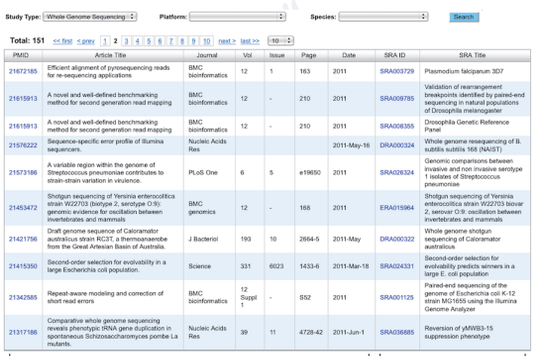
Species of samples (top 10)

unidentified	717
Homo sapiens	655
Mus musculus	358
metagenome sequence	180
Drosophila melanogaster	179
Caenorhabditis elegans	131
marine metagenome	124
E. coli str. K-12 substr. MG1655	102
Arabidopsis thaliana	81
Saccharomyces cerevisiae	75
...	
Total (studies)	11730

SRAs: the survey of read archives
http://sra.dbcls.jp/



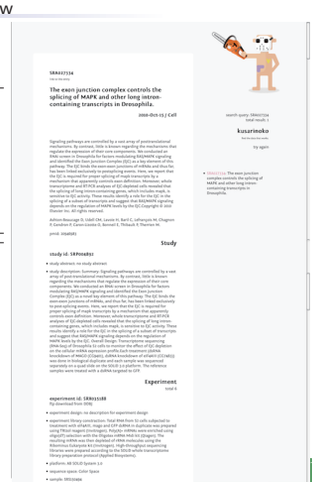
Publication List



Publications using NGS | Corresponding NGS data

鎖鋸 (Kusarinoko): Detail view

In detail, please ask Tazro!



Publications using NGS

Corresponding NGS data

Conclusions

- The next-generation sequencing (NGS) data archived in short read archive (SRA) contains not only short read sequences but also the conditions of experiments as a meta data.
- We categorized NGS data by study types, sequencer platforms, and sample species.
  We developed a web service the Surver of Read Archives (SRAs) as a yellow page, and provides statistics such as the number of projects.
- We constructed a publication list that refers NGS data, and developed a web service called Kusarinoko that shows integrated metadata and information extracted from articles and SRA.

Genome Informatics
(2011 Cold Spring Harbor Laboratory conferences)
Cold Spring Harbor Laboratory, NY, US
Nov. 2-5, 2011