

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)

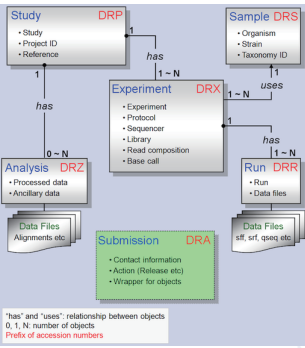
*: nakazato@dbcls.rois.ac.jp



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

Backgrounds and motivations

The data structure of SRA



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

The next-generation sequencing (NGS) data is archived in SRA, ENA, and DRA as public repositories. The deposited 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 reposted as a new submission.

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

Results and Discussions

Statistics (as of Oct. 14, 2011)

Study Types

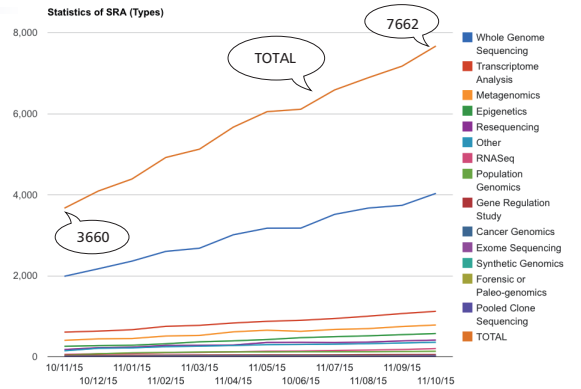
Whole Genome Sequencing	4072
Transcriptome Analysis	1128
Metagenomics	844
Epigenetics	575
Resequencing	410
Other	354
RNASeq	204
Population Genomics	128
Gene Regulation Study	47
Cancer Genomics	27
Exome Sequencing	20
Pooled Clone Sequencing	6
Synthetic Genomics	5
Forensic or Paleo-genomics	5
Total (studies)	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 Ix	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
<i>Homo sapiens</i>	655
<i>Mus musculus</i>	358
metagenome sequence	180
<i>Drosophila melanogaster</i>	179
<i>Caenorhabditis elegans</i>	131
marine metagenome	124
<i>E. coli</i> str. K-12 substr. MG1655	102
<i>Arabidopsis thaliana</i>	81
<i>Saccharomyces cerevisiae</i>	75
...	
Total (studies)	11730



Publication List

PMID	Article Title	Journal	Vol	Issue	Page	Date	SRA ID	SRA Title	
2187185	Efficient alignment of nanopore sequencing reads for re-sequencing applications	BMC Bioinformatics	12	1	163	2011	SRAC02378	Plasmodium falciparum 3D7	
21815913	A novel and well-defined benchmarking method for second generation read mapping	BMC Bioinformatics	12	-	210	2011	SRAC03978	Validation of resequencing breakpoints identified by paired-end sequencing in natural populations of <i>Drosophila melanogaster</i>	
21815913	A novel and well-defined benchmarking method for second generation read mapping	BMC Bioinformatics	12	-	210	2011	SRAC03835	Drosophila Genetic Reference Panel	
21576222	Sequence-specific error profile of Illumina sequencing	Nucleic Acids Res				2011 May 16	DRAC02324	Whole genome resequencing of <i>B. subtilis</i> strains 168 (NA517)	
21873366	A variable region within the genome of <i>Streptococcus pneumoniae</i> contributes to strain-strain variation in virulence	PLoS One	6	5	e19650	2011	SRAC02824	Genomic comparisons between invasive and non-invasive serotype 1 isolates of <i>Streptococcus pneumoniae</i>	
21434472	Shotgun sequencing of <i>Trinectes orientalis</i> strain VZ2703 (Dongye 2, serotype O3) genetic evidence for isolation from invertebrates and mammals	BMC genomics	12	-	166	2011	ERA015964	Shotgun sequencing of <i>Trinectes orientalis</i> strain VZ2703 (Dongye 2, serotype O3)	
21421756	Draft genome sequence of <i>Calamotriton aurantius</i> strain RCT2, a thymopneumoniae from the Great Artesian Basin of Australia	J Biol Chem	193	10	28645	2011 May	DRAC02322	Whole genome shotgun sequencing of <i>Calamotriton aurantius</i>	
21415300	Second-order selection for availability in a large <i>Escherichia coli</i> population	Science	331	6023	14334	2011 Mar 18	SRAC04331	Second-order selection for availability predicts winners in a large <i>E. coli</i> population	
21342885	Repeat aware modeling and correction of short read errors	BMC Bioinformatics	12	Suppl 1	-	352	2011	SRAC01123	Paired-end sequencing of the genome of <i>Escherichia coli</i> K-12 strain MG1655 using the Illumina Genome Analyzer
21317386	Comparative whole genome sequencing reveals phenotypic RNA gene duplication in spontaneous <i>Salmonella enteritidis</i> pent6 mutants	Nucleic Acids Res	39	11	4728-42	2011 Jun 1	SRAC03688	Reversion of <i>rfaH</i> -15 suppression phenotype	

Publications using NGS

Corresponding NGS data

鎖鑑 (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.

