Functional indexing and curation of next-generation sequencing data

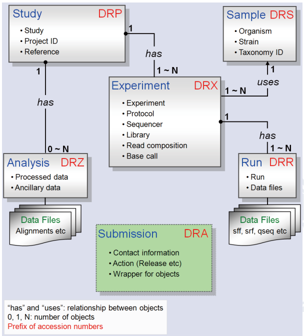
Takeru Nakazato*, Hidemasa Bono, Toshihisa Takagi
Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS)
*: nakazato@dbcls.rois.ac.jp

Backgrounds and motivations

Nucleotide Sequences	gene expression	journal articles	**next-generation sequencing data**
⬇	⬇	⬇	⬇
DDBJ/EMBL/ GenBank	GEO ArrayExpress	PubMed	**Short Read Archive (SRA)**

The next-generation sequencing data is archived in short read archive (SRA) in NCBI, EBI, and DDBJ as well as nucleotide sequences in GenBank and gene expression data in GEO.

The data structure of SRA

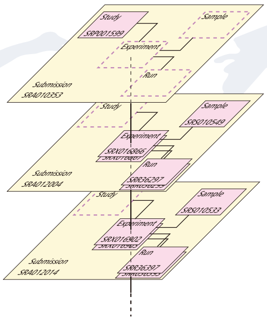


The doposited NGS data contains not only short read sequences but also conditions of experiments including project title, species or cell line names of samples, and sequencing platforms as a meta data. The meta data consists of six files with XML format: submission, study, experiment, run, sample, and 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.

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

Submission	Study	Experiment	Run	Sample	Analysis	
✓			✓			7066
✓	✓	✓	✓	✓		1545
✓		✓				500
✓		✓	✓			228
✓				✓		142
✓						139
✓	✓					106
✓		✓	✓	✓		89
✓	✓	✓		✓		33
✓	✓	✓	✓			18
✓	✓			✓		17
✓					✓	2
✓		✓		✓		1
✓			✓	✓		1
					total	9887



Methods

```
<?xml version="1.0" encoding="UTF-8"?>
<EXPERIMENT_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <EXPERIMENT alias="4NG_TG-P3_044sA-FLX" accession="SRX003641">
    <TITLE/>
    <STUDY_REF refname="1000Genomes Project Pilot 3" accession="SRP000033"/>
    <DESIGN>
      <DESIGN_DESCRIPTION>454 Sequencing of Human Single-Direction Library from Nimblegen
Capture</DESIGN_DESCRIPTION>
      <SAMPLE_DESCRIPTOR refname="NA19179" accession="SRS000790"/>
      <LIBRARY_DESCRIPTOR>
        ...
      </LIBRARY_DESCRIPTOR>
      <SPOT_DESCRIPTOR>
        ...
    </DESIGN>
    <PLATFORM>
      <LS454>
        <INSTRUMENT_MODEL>GS FLX</INSTRUMENT_MODEL>
        <FLOW_SEQUENCE>TACG</FLOW_SEQUENCE>
        <FLOW_COUNT>100</FLOW_COUNT>
      </LS454>
      ...
    </PLATFORM>
    ...
```

SRA008310.experiment.xml

We made connections among each type of corresponding matadata by extracting accession numbers assigned as a reference from XML files. We also obtained informations of experiments such as titles and platforms from each XML files.

Results and Discussions

Statistics

Study Types

Whole Genome Sequencing	1366
Transcriptome Analysis	463
Metagenomics	390
Epigenetics	198
Other	110
Resequencing	70
Gene Regulation Study	19
Population Genomics	17
RNASeq	12
Cancer Genomics	10
Forensic or Paleo-genomics	2
Synthetic Genomics	1
Total	2658

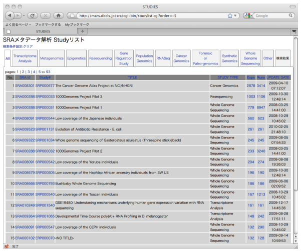
Platforms

Illumina Genome Analyzer II	11727
454 GS FLX	4321
Illumina Genome Analyzer	3058
Solexa 1G Genome Analyzer	1481
454 Titanium	1314
unspecified	923
GS FLX	822
AB SOLiD System 3.0	187
GS 20	164
AB SOLiD System 2.0	158
454 GS 20	98
AB SOLiD System	76
Helicos HeliScope	14
454 GS	9
Total	24352

Species of samples (top 12)

Human Metagenome	76656
Homo sapiens	2380
Human	1051
Mus musculus	757
Drosophila melanogaster	609
Plasmodium falciparum	591
human metagenome	400
Oryza sativa Indica Group	240
human skin metagenome	178
Metagenomic	160
Caenorhabditis elegans	150
Arabidopsis thaliana	137
...	
Total	93157

Data visualization


STUDIES
http://mars.dbcls.jp/sra/cgi-bin/studylist.cgi?order=-5
SRAメタデータ解析 Studyリスト
検索条件設定:クリア

pages: 1 | 2 | 3 | 4 | 5 >> 93

No.	SRA id	Study#	TITLE	STUDY TYPE	Exps	Runs	UPDATE DATE
1	SRA008301	SRP000677	The Cancer Genome Atlas Project at NCI/NHGRI	Cancer Genomics	2878	3414	2009-04-10 07:12:07
2	SRA000288	SRP000033	1000Genomes Project Pilot 3	Resequencing	1003	1106	2009-10-30 12:48:14
3	SRA000288	SRP000031	1000Genomes Project Pilot 1	Whole Genome Sequencing	779	8947	2008-03-25 14:41:00
4	SRA008091	SRP000544	Low coverage of the Japanese individuals	Whole Genome Sequencing	560	623	2008-10-29 10:45:02
5	SRA009523	SRP001131	Evolution of Antibiotic Resistance - E. coli	Whole Genome Sequencing	261	261	2010-02-25 21:48:10
6	SRA009322	SRP001034	Whole genome sequencing of Gasterosteus aculeatus (Threespine stickleback)	Whole Genome Sequencing	245	245	2009-08-05 07:54:33
7	SRA000288	SRP000032	1000Genomes Project Pilot 2	Whole Genome Sequencing	233	3240	2008-03-25 14:41:00
8	SRA008091	SRP000542	Low coverage of the Yoruba individuals	Whole Genome Sequencing	204	274	2008-08-08 19:36:03
9	SRA008679	SRP000805	Low coverage of the HapMap African ancestry individuals from SW US	Whole Genome Sequencing	196	190	2009-10-30 12:48:14
10	SRA008666	SRP000793	Bushbaby Whole Genome Sequencing	Whole Genome Sequencing	186	186	2009-06-06 02:09:52
11	SRA008091	SRP000540	Low coverage of the Toscan individuals	Whole Genome Sequencing	167	1213	2008-10-29 10:45:02
12	SRA010249	SRP001540	GSE19480: Understaning mechanisms underlying human gene expression variation with RNA sequencing	Transcriptome Analysis	161	161	2009-12-17 14:45:36
13	SRA009364	SRP001065	Developmental Time Course poly(A)+ RNA Profiling in D. melanogaster	Transcriptome Analysis	148	242	2009-08-28 17:51:11
14	SRA008091	SRP000547	Low coverage of the CEPH individuals	Whole Genome Sequencing	132	290	2008-10-29 10:45:02
15	ERA000102	ERP000070	<NO TITLE>	Whole Genome Sequencing	132	128	2009-09-14 10:59:53

完了

jump to the NCBI SRA site
jump to the corresponding experiments list
jump to thecorresponding runs list

We developed an index site of NGS data as yellow pages to make NGS data more searchable and re-usable.

This service shows a project list, and corresponding lists of experiments and runs by clicking the numbers of assigned experiments and runs. Researchers can also restrict the study entries by types of the interests such as transcriptome analysis or whole genome sequencing.

This web service is freely available on http://mars.dbcls.jp/sra/.

The information described in meta data contains errors and spelling variation such as "*Homo sapiens*" and "human" because the information was originally written by researchers who provided corresponding short read sequences. We will curate extracted informations by correcting these misspellings and disambiguate spelling variations.

Conclusions

- The next-generation sequencing (NGS) data is archived in short read archive (SRA) and the archived data contains not only short read sequences but also the conditions of experiments as a meta data.
- Additional experiments and runs are often deposited as a new submission. We therefore made connections among submissions by extracting accession numbers as a reference from XML files.
- We developed an index site of NGS data as yellow pages and the service is available on http://mars.dbcls.jp/sra/.

Genome Informatics
(2010 Joint Wellcome Trust/CSHL conferences)
Hinxton, UK
Sept. 14-19, 2010