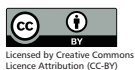


# Functional interface for quick access to high-quality omics data

実験情報を利用した質の高い公共オミックスデータの検索と目次サイトの構築

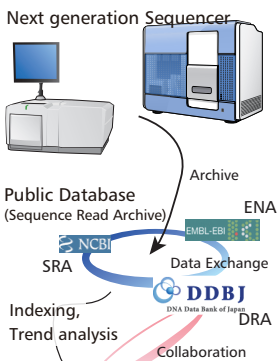


Takeru Nakazato\*, Tazro Ohta, Hidemasa Bono  
仲里 猛留 大田 達郎 坊農 秀雅

Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS)  
情報・システム研究機構 ライフサイエンス統合データベースセンター

\*: nakazato@dbcls.rois.ac.jp  
twitter ID: chalkless

## Backgrounds



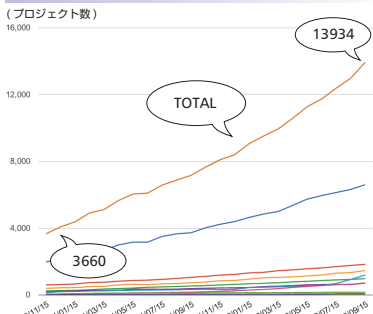
DBCLS DBCLS SRA  
<http://sra.dbcls.jp/>

The next-generation sequencing (NGS) data is archived in public databases, 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.

次世代シーケンサ (NGS) データも公共データベースである Sequence Read Archive (SRA) に登録され、日米欧の3局でデータ交換がなされている。その数は、プロジェクト数で14000 (2012年12月現在) に及んでいる。DBCLSでは、登録データに対して、目次作成、データの傾向分析を行い、NGS データの検索サイトを構築、提供している。

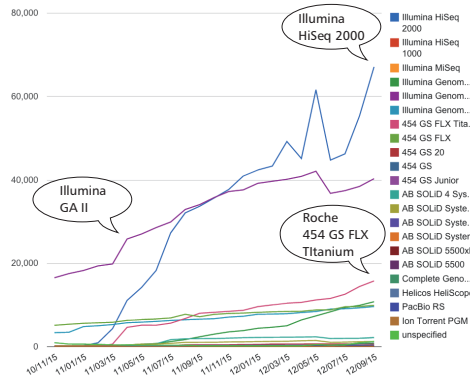
## Results and Discussions

### Statistics



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

(結果配列ファイル数)



Species	Count
<i>Homo sapiens</i>	1261
unidentified	882
<i>Mus musculus</i>	818
<i>Drosophila melanogaster</i>	284
<i>Mustela putorius furo</i>	233
<i>Caenorhabditis elegans</i>	198
metagenome sequence	184
marine metagenome	165
<i>Arabidopsis thaliana</i>	160
unclassified phages	139
<i>Escherichia coli</i> str. K-12 substr. MG1655	136
soil metagenome	127
<i>Saccharomyces cerevisiae</i>	124
<i>Human rhinovirus A</i>	95
<i>Salmo salar</i>	84
<b>Total</b>	<b>16806 (studies)</b>

### Disease List

Disease Type: ANY

Total: 305

Disease	疾病名	# of submission
Best Neoplasms	乳房腫瘍	43
Prostatic Neoplasms	前立腺腫瘍	22
Disease Models, Animal	疾患モデル(動物)	21
Genetic Predisposition to Disease	遺伝的素因(疾病)	20
Disease Progression	病勢悪化	15
Translocation, Genetic	転座	14
Cell Transformation, Neoplastic	腫瘍細胞形質転換	12
Lung Neoplasms	肺腫瘍	11
Staphylococcal Infections	ブドウ球菌感染症	10
Chromosome Aberrations	染色体異常	9

Total: 305

### Publication List

Study Type: (Whole Genome Sequencing) Platform: Species: Search

PMID	Article Title	Journal	Vol	Issue	Page	Date	SRA ID	SRA Title
2187245	Efficient alignment of pyrosequencing reads for re-sequencing applications	BMC bioinformatics	12	1	163	2011	SRA003729	Plasmodium falciparum 3D7
21819913	A novel and well-defined benchmarking method for second generation read mapping	BMC bioinformatics	12	-	210	2011	SRA003786	Validation of management transcripts identified by paired-end sequencing in natural populations of <i>Drosophila melanogaster</i>
21579222	A novel and well-defined benchmarking method for second generation read mapping	BMC bioinformatics	12	-	210	2011	SRA003535	<i>Drosophila</i> Genetic Reference Panel
21579222	Sequence-specific error profiles of Illumina sequencers	Nucleic Acids Res	39	18	6483-6491	2011-May-18	DR000324	Whole genome resequencing of <i>S. subtilis</i> subtile 168 (NA5T)
21573386	A variable region within the genome of <i>Streptococcus pneumoniae</i> contributes to strain-strain variation in virulence	PLoS One	6	5	e18500	2011	SRA028234	Genomic comparisons between invasive and non-invasive serotype 1 isolates of <i>Streptococcus pneumoniae</i>
21453472	Shotgun sequencing of <i>Yersinia enterocolitica</i> strain W23703 (Shiga 2, serotype O:3)	BMC genomics	12	-	168	2011	ERA011964	Shiga 2 sequencing of <i>Yersinia enterocolitica</i> strain W23703 (Shiga 2, serotype O:3) for evidence for oscillation between <i>Yersinia enterocolitica</i> strains
21421196	Draft genome sequence of <i>Calobacter autotrophicus</i> strain RCCT, a thermotolerant from the Great Artesian Basin of Australia	J Bacteriol	193	10	2864-5	2011-May	DR000322	Whole genome shotgun sequencing of <i>Calobacter autotrophicus</i>
21415300	Second-order selection for availability in a large <i>Escherichia coli</i> population	Science	331	6023	1433-6	2011-Mar-18	SRA024331	Second-order selection for availability in a large <i>Escherichia coli</i> population
21342585	Repeat-aware modeling and correction of short-read errors	BMC bioinformatics	12	Suppl 1	S52	2011	SRA001725	Paired-end sequencing of the genome of <i>Salmonella enterica</i> serovar MD165 using the Illumina Genome Analyzer
21337166	Comparative whole genome sequencing reveals phenotypic RNA gene duplication in opportunistic <i>Schistosoma haematobium</i> from the Maldives	Nucleic Acids Res	39	11	4728-42	2011-Jun-1	SRA028885	Reversion of <i>AMV15</i> phenotype suppression of <i>AMV15</i>

Publications using NGS

Corresponding NGS data

### Data Quality

Please ask Tazro!

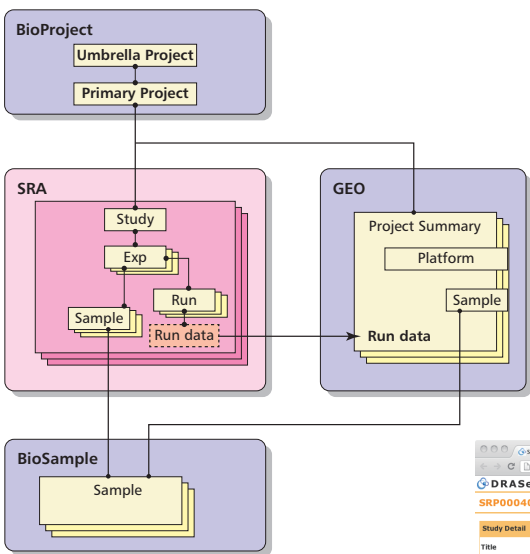


3P-0047: Advanced interface of public NGSseq database for efficient data search

Tazro Ohta, Takeru Nakazato, Hidemasa Bono

公共の次世代シーケンサデータを効率良く検索するためのインターフェース開発  
大田 達郎, 仲里 猛留, 坊農 秀雅

## Future works



3局では、複数のデータベースを俯瞰するため、プロジェクトのデータベースである BioProject と、サンプルのデータベースである BioSample を立ち上げており、今後、これらによる統合的な活用ができるよう、開発を行っている。

### Project List

No.	Accession	Project Title	Platform	Species	Seq. Type	Exp.	Run	Size
1	SRA051306	SRP011870 Genetic Dissection of Model Complex. Trait Using the <i>Drosophila</i> Synthetic Population Resource	Other	2640	2640	2012-03-27		
2	ERA007057	ERPO00190 Plasmodium, falciparum, natural, genome, variation	Whole Genome Sequencing	2447	2532	2012-05-29		
3	ERA015449	ERPO00426 ZF_Mr-SL	Other	2440	2536	2012-04-25		
4	SRA026060	SRP002163 Human Microbiome Project Metagenomes Production Phase	Metagenomics	2306	6017	2010-06-09		
5	SRA023558	SRP003279 ARRA Autism Sequencing Collaboration	Resequencing	2257	1385	2010-08-03		
6	SRA056418	SRP014601 Jackson Heart Study Allelic Spectrum Sequencing Discovery	Other	2030	1948	2012-05-03		
7	ERA071297	ERPO01040 High-resolution, QTL, mapping in an outbred population of mice using low coverage next generation sequencing	Whole Genome Sequencing	1767	1767	2012-04-03		
8	SRA026080	SRP003341 Women's Health Initiative Clinical Trial and Observational Study Sequencing Project	Resequencing	1741	1805	2010-08-03		
9	SRA050290	SRP011021 Prostate Cancer Genome Sequencing Project	Resequencing	1640	3028	2012-02-23		
10	SRA023556	SRP004051 Framingham Heart Study Allelic Spectrum Sequencing Discovery	Resequencing	1630	456	2011-11-02		
11	SRA026066	SRP002865 Human Microbiome Project 16S rRNA 454 Clinical Production Phase II	Metagenomics	1517	1823	2010-03-17		
12	ERA112942	ERPO01384 Defining the core <i>Arabidopsis thaliana</i> root microbiome	Metagenomics	1277	1277	2012-06-04		
13	SRA050859	SRP011540 Malanoma Genome Sequencing Project	Resequencing	1170	2500	2012-02-23		
14	ERA005746	ERPO00129 The Genomics of Speciation in <i>Caryophyllus</i>	Population Genomics	1080	1213	2012-03-20		
15	SRA051296	SRP011948 Swedish Schizophrenia Population-Based Case-control Exome Sequencing	Other	1067	1032	2012-03-26		
16	SRA023556	SRP003277 Cilopathies Exome Sequencing Initiative	Resequencing	1062	1066	2010-08-16		
17	ERA095468	ERPO01277 Pig X, Y, clone sequencing	Other	1038	1033	2012-03-09		
18	ERA068391	ERPO01012 <i>Staphylococcus aureus</i> _BSAC_study	Whole Genome Sequencing	1030	1030	2012-02-03		
19	SRA052357	SRP012894 Genetic Analysis of Hirschsprung Disease	Resequencing	981	344	2012-03-26		
20	ERA011044	ERPO00270 Investigating the diversity of <i>Vibrio cholerae</i> isolates.	Whole Genome Sequencing	981	1167	2012-05-08		

8102 projects

Coming soon... 1912 projects

BioProject

GEO

DRA (DDBJ SRA)

<http://trace.ddbj.nig.ac.jp/DRAsearch/>

第35回 日本分子生物学会年会  
福岡国際会議場 マリンメッセ福岡  
平成24年12月11日~14日