

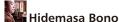
# DBCLS SRA: Functional mining and characterization of public NGS data











Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), JAPAN \*: nakazato@dbcls.rois.ac.jp

# **Backgrounds** Next generation Sequencer **Public Database** (Sequence Read Archive Data Exchange SRA **DDBJ** Indexing, Trend analysis Collaboration DBCLS

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

### **Solutions** FREE!

45502

(as of Aug. 15, 2014)

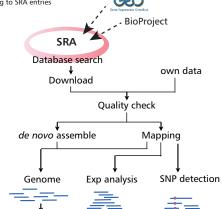
**DBCLS SRA** http://sra.dbcls.jp/

**Statistics** 

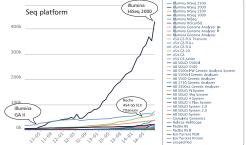
Project Type

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



# Seq platform



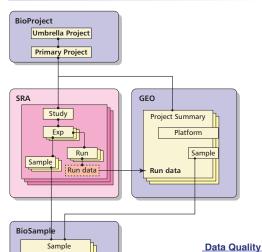
Species (top 15)

### **Publication List**

to get NGS data with sufficient good quality to publish articles

PMID	Article Title	Journal	Vol	hous	Page	Date	SRA_ID_orig	SRA ID	SRA TH
25110290	Proteomics informed by transcriptomics identifies novel secreted proteins in Dermacenter anderson solliva	International journal for parasitratogy			-	2014 Aug-7	SRA183490	SRA163490	Dermicer andersoni Salivary o Transcript
25098249	Molecular Amelysis of the Cold Tolerant Antarctic Necratode, Panagralaimus clavidi	PloS one	9	a	e104526	2014	SRA180488	SRA180488	Panagrolo david Genome sequencia
25089621	Diffect of Incubation on Bacterial Communities of Eggstells in a Temperate Bird, the Eurosian Magpie (Pica pica)	PloS one	9	8	e100959	2014	SRA135159	SRA135159	Bacteria Motagono
25083000	Xanthoreonas assesspedis virulence is promoted by a transcription activator like (TAL) effector mediated induction of a SWEET suger transporter in cassava	Molecular plans-encrobe interactions : MPM	٠			2014 Aug-1	58A177754	SRA177754	TAL effect induced of engression changes Cassena
25083871	MicroRNA Directly Enhances Machaedrial Translation during Neschi Differentiation	o.i	158	3	607-19	2014	SRA163611	SRA163611	GSES750 MicroPhilip Directly Enhance Misochon Translatio during Mi Differenti
25081556	Genomic sequencing identifies nevel Back as fruinglemain inp trings brang and Cryst socies that have high toxicity to Scarabacoides larvae	Applied microtidagy and biotechnology	٠	-	-	2014 Aug 1	SRA163620	SRA183620	Bacillus thuringler strain:190 Geogra- sequenci
25000338	Novel Compound Helerozygous Mutations in MYDZA Associated with Usher Syndrome 1 in a Chinese Family	PloS one	9	7	e100415	2014	SRA189921	SRA169921	Home say Targeted Leous (Li
25078300	THE FECAL VIRONE OF CATS IN AN AMMAL	The Journal of general				2014	SRA179328	SRA178328	CAT feca

## The data structure of SRA



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

SRA ID	SRA Title	Disease	疾患名	PMID	
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	1898773	
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	19657111	
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	1898773	
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	19657111	
SRA009897	In-depth characterization of the microRNA transcriptome in a leukemia progression model	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	1884952	
SRA029797	Exome Sequencing Identifies Somatic Mutations in Acute Monocytic Leukemia	Leukemia, Myeloid, Acute	白血病- 急性骨 髄性	2139963	

Public NGS data

Disease Corresponding publication

NGS data

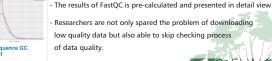
Recently, some NGS run data is not reposited to SRA but GEO and GenBank. Moreover, project and sample information is captured in external other databases: BioProject and BioSample. This complexity of data structure prevents researchers from retrieving the NGS

**Y** Sequence Quality Statistics









Researchers are not only spared the problem of downloading low quality data but also able to skip checking process of data quality.

data of interests.

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

Per base N content







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