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

「使える」SRA データにすばやくたどりつくために

How to provide quick access to SRA entries of interest

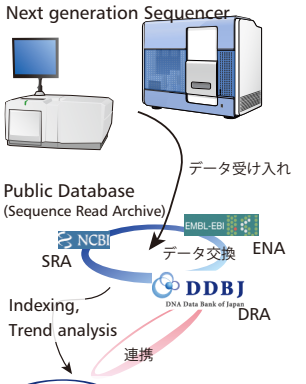
○ 仲里 猛留 *、 大田 達郎、 坊農 秀雅
Takeru Nakazato Tazro Ohta Hidemasa Bono

*: nakazato@dbcls.rois.ac.jp

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
Database Center for Life Science (DBCLS), Res. Org. of Info. and Systems (ROIS)

※ デモしますので発表者 までお気軽に。
http://sra.dbcls.jp/

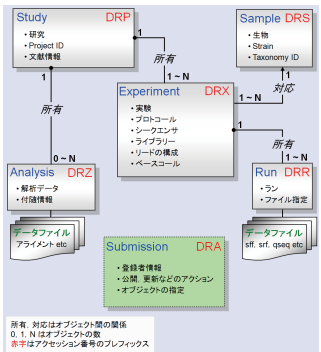
Backgrounds



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

NGS データも、マイクロアレイのデータが GEO に登録されるのと同様に、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、プロジェクト数でおよそ 11300 (2012 年 5 月現在) に及んでいる。DBCLS では、DDBJ と連携して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。

SRA のデータ構造



http://trace.ddbj.nig.ac.jp/dra/documentation.shtml より

Submission	Study	Experiment	Run	Sample	Analysis	
✓				✓		28790
✓			✓			20258
✓	✓	✓	✓	✓		9553
✓		✓			✓	3995
✓		✓	✓			2407
✓		✓				2146
✓	✓					834
✓		✓	✓	✓		514
✓		✓	✓		✓	491
✓						441
✓					✓	181
✓	✓	✓	✓			123
✓	✓	✓		✓		94
✓	✓	✓	✓	✓	✓	73
		⋮				
		Total (submissions)				70003

Results and Discussions

統計値より

(2012-5-15 現在)

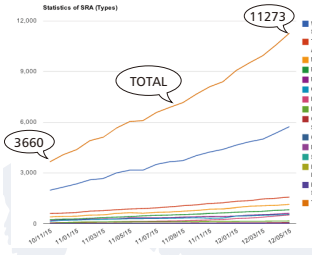
Big Project List

No.	SRA#	Study#	TITLE	STUDY_TYPE	Exps	Runs	UPDATE_DATE
1	SRA026056	SRP000677	The Cancer Genome Atlas Project at NCI/NHGRI	Cancer Genomics	30167	141161	2009-04-10
2	SRA051306	SRP011970	Genetic Dissection of Model Complex Trait Using the Drosophila Synthetic Population Resource	Other	2640	2640	2012-03-27
3	SRA023558	SRP003279	ARRA Autism Sequencing Collaboration	Resequencing	2257	397	2010-08-03
4	ERA007057	ERP000190	Plasmodium_falciparum_natural_genome_variation	Whole Genome Sequencing	2067	2152	2011-11-16
5	SRA026080	SRP003341	Women's Health Initiative Clinical Trial and Observational Study Sequencing Project	Resequencing	1741	1805	2010-08-03
6	ERA015449	ERP000426	ZF_MrSol	Other	1718	1814	2011-11-08
7	SRA029586	SRP004061	Framingham Heart Study Allelic Spectrum Sequencing Discovery	Resequencing	1638	456	2010-11-02
8	SRA050250	SRP011021	Prostate Cancer Genome Sequencing Project	Resequencing	1554	328	2012-02-23
9	SRA026060	SRP002860	Human Microbiome Project 16S rRNA 454 Clinical Production Phase II	Metagenomics	1517	1894	2010-03-17
10	SRA026060	SRP002163	Human Microbiome Project Metagenomes Production Phase	Metagenomics	1179	775	2010-06-09
11	SRA050859	SRP011540	Melanoma Genome Sequencing Project	Resequencing	1081	281	
12	SRA051296	SRP011949	Swedish Schizophrenia Population-Based Case-control Exome Sequencing	Resequencing	1067	1032	2012-03-26
13	SRA023558	SRP003277	Ciliopathies Exome Sequencing Initiative	Resequencing	1062	1066	2010-08-16
14	ERA095466	ERP001277	Pig_X_Y_clone_sequencing	Other	959	959	2012-03-09
15	SRA000288	SRP000033	1000Genomes Project Pilot 3	Resequencing	959	336	2009-10-30
16	ERA011763	ERP000310	Streptococcus pneumoniae global lineages	Whole Genome Sequencing	942	1140	2011-12-07
17	ERA000159	ERP000071	Bacterial variation in human body habitats across space and time	Metagenomics	815	815	2010-08-02
18	ERA046838	ERP000872	Streptococcus_pneumoniae_samples_from_Malawi	Whole Genome Sequencing	702	702	2011-08-25
19	SRA000288	SRP000031	1000Genomes Project Pilot 1	Whole Genome Sequencing	670	2715	2010-03-17
20	ERA011044	ERP000270	Investigating the diversity of Vibrio cholerae isolates.	Whole Genome Sequencing	647	833	2011-11-25

ユーザーのニーズに合わせ、目的別、機器別、生物種別に SRA に登録されたデータを検索・閲覧できるサイトを構築した。また、より信頼度の高いデータにアクセスするため、文献の出ているデータのリスト化も行った。QC 結果を閲覧できる鎖鋸システムとも連携しており、今後、一体的な運用ができるよう改良を行う。

Study Types

Whole Genome Sequencing	5741
Transcriptome Analysis	1574
Metagenomics	1143
Epigenetics	819
Resequencing	613
Other	544
RNASeq	513
Population Genomics	157
Gene Regulation Study	66
Cancer Genomics	38
Exome Sequencing	34
Pooled Clone Sequencing	20
Forensic or Paleo-genomics	6
Synthetic Genomics	5
Total	11273 (studies)



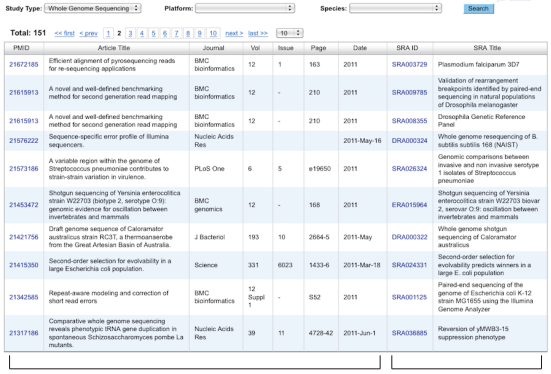
Species of samples (top 15)

Homo sapiens	1035
unidentified	884
Mus musculus	562
Drosophila melanogaster	241
metagenome sequence	180
Caenorhabditis elegans	168
marine metagenome	154
Escherichia coli str. K-12 substr. MG1655	135
Arabidopsis thaliana	117
soil metagenome	110
Saccharomyces cerevisiae	103
Mustela putorius furo	99
Human rhinovirus A	88
Plasmodium falciparum	50
ecological metagenomes	44
Total	13208 (studies)

Platform

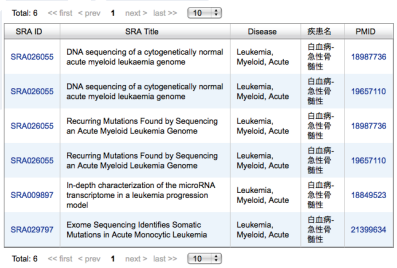
Illumina Genome Analyzer II	3463
454 GS FLX Titanium	2693
Illumina HiSeq 2000	2064
454 GS FLX	1704
Illumina Genome Analyzer	1110
Illumina Genome Analyzer IIx	493
454 GS 20	322
AB SOLiD System 3.0	169
AB SOLiD 4 System	74
unspecified	61
Illumina MiSeq	61
AB SOLiD System 2.0	61
Ion Torrent PGM	53
AB SOLiD System	36
PacBio RS	30
Helicos HeliScope	24
454 GS	15
454 GS Junior	12
none	7
Illumina HiSeq 1000	5
Complete Genomics	4
AB SOLiD 5500xl	1
AB SOLiD 5500	1
Total	12463 (studies)

文献より



Publications using NGS

疾患より

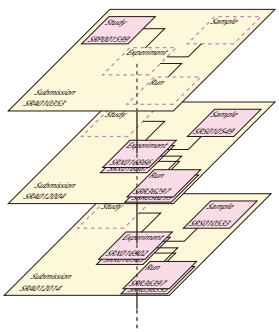


Corresponding NGS data

データのクオリティより

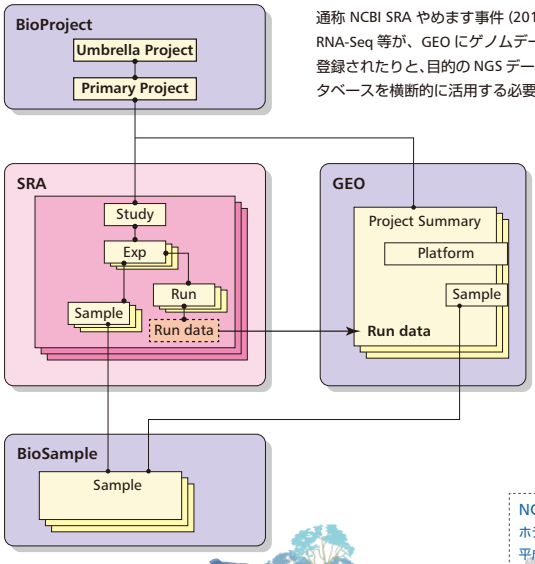
O12：公共 NGS データベース Sequence Read Archive における シーケンスクオリティによるデータ検索（大田 ）参照

SRA に登録された NGS データは、リード（検出された塩基配列）のほかに、実験情報（プロジェクト名、機器名、生物種名）がメタデータとして 6 つの XML ファイルの形で記載され、登録されている。しかしながら、1 つの登録 (Submission) にすべてのメタデータが付与されているわけはない。
今回、これらのメタ情報の関連づけと、情報抽出を行うことで、目的の SRA エントリにすばやくたどりつくためのエントリ間の整理を行った。



Future works

NGS もデータベースの横断的な活用へ



通称 NCBI SRA やめます事件 (2011 年 2 月) の顛末として、RNA-Seq 等が、GEO にゲノムデータが DDBJ/GenBank/EMBL に登録されたりと、目的の NGS データを取得するのに、複数のデータベースを横断的に活用する必要がある。

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。

NGS 現場の会 第 2 回研究会
ホテル阪急エキスポパーク（大阪）
平成 24 年 5 月 23 日〜25 日