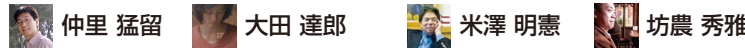


# Functional interface for quick access to disease-relevant NGS data

次世代シーケンサによる疾患に関連するデータを使い倒すために

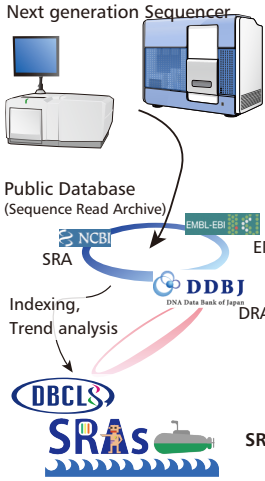
○ Takeru Nakazato\*, Tazro Ohta, Akinori Yonezawa, Hidemasa Bono

\*: nakazato@dbcls.rois.ac.jp



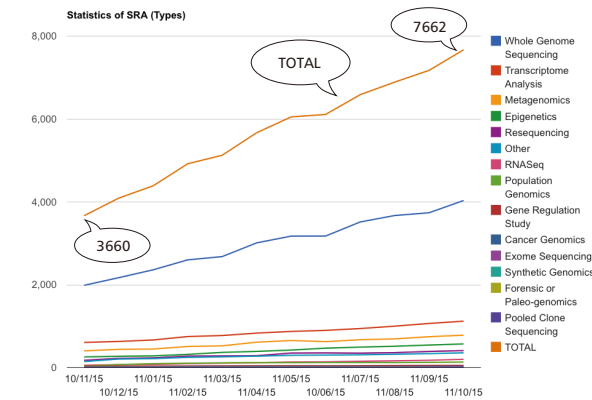
Database Center for Life Science (DBCLS), Res. Org. of Info. and Systems (ROIS)  
情報・システム研究機構 ライフサイエンス統合データベースセンター

## Backgrounds and motivations



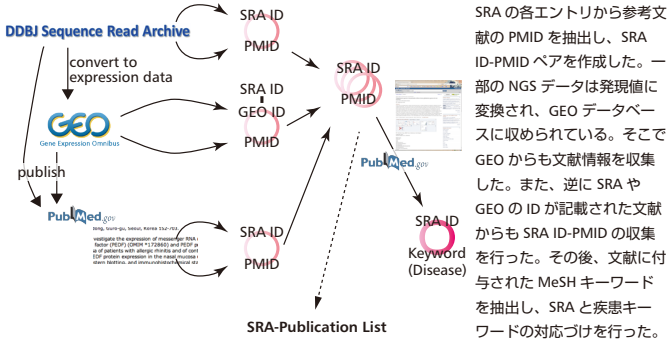
ここ最近、次世代シーケンサ (NGS) による成果が次々に発表されている。マイクロアレイのデータが GEO に登録されると同様に、NGS データも公共データベースである Sequence Read Archive (SRA) に登録され、日米欧の 3 局でデータ交換がなされている。その数は、プロジェクト数で 8300 (2011 年 12 月現在) に及んでいる。DBCLS では、登録データに対して、目次作成、データの傾向分析を行い、NGS データの検索サイトを構築、提供している。今回、NGS データに対する 1 つの切り口として疾患に注目し、データの整理と、その傾向を分析した。また、NGS データを疾患から検索するインターフェースを構築した。

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



※ シーケンサ別、生物種別の傾向については、上記 SRAs サイトをご覧ください。発表者にお尋ねください。

## Methods



PubMed	Article Title	Journal	Vol	Issue	Date	SRA ID	SRA Title	
2162186	Efficient alignment of pyrosequencing reads for re-sequencing applications	BMC bioinformatics	12	1	163	2011	SRA003729	Plasmodium falciparum 3D7
21615913	A novel and well-defined benchmarking method for second-generation read mapping	BMC bioinformatics	12	-	210	2011	SRA009785	Validation of rearrangement breakpoints identified by paired-end sequencing in natural populations of <i>Drosophila melanogaster</i>
21618913	A novel and well-defined benchmarking method for second-generation read mapping	BMC bioinformatics	12	-	210	2011	SRA008335	Drosophila Genetic Reference Panel
21596222	Sequence-specific error profile of Illumina sequencers	Nucleic Acids Res	39	11	4278-42	2011-May-16	DR0000324	Whole genome resequencing of <i>B. subtilis</i> strain 106 (D4021)
21573188	A variable region within the genome of <i>Streptococcus pneumoniae</i> contributes to strain-variant virulence	PLoS One	6	5	e19650	2011	SRA028324	Genomic comparisons between invasive and non-invasive serotype 1 isolates of <i>Streptococcus pneumoniae</i>
21453472	Shiga toxin sequencing of <i>Yersinia enterocolitica</i> strain W22703 (biovar 2, serotype O:3) genetic evidence for recombination between invertebrates and mammals	BMC genomics	12	-	168	2011	ERA015964	Shiga toxin sequencing of <i>Yersinia enterocolitica</i> strain W22703 (biovar 2, serotype O:3) genetic evidence for recombination between invertebrates and mammals
21417756	Draft genome sequence of <i>Calceolaria australis</i> strain R121, a thermophilic fern from the Great Australian Basin of Australia	J. Biotechnol	193	10	2644-5	2011-May	DR0000332	Whole genome shotgun sequencing of <i>Calceolaria australis</i>
21415300	Second-order selection for availability in a large E. coli population	Science	331	6023	1433-6	2011-Mar-18	SRA024331	Second-order selection for availability in a large E. coli population
21412655	Repeat-aware modeling and correction of short read errors	BMC bioinformatics	12	Suppl 1	-	2011	SRA001125	Paired-end sequencing of the genome of <i>Escherichia coli</i> K-12 strain MG1625 using the Illumina Genome Analyzer
21317196	Comparative whole genome sequencing reveals phenotypic RNA gene duplication in spontaneous <i>Salmonella enterica</i> serovar Paratyphi A mutants	Nucleic Acids Res	39	11	4278-42	2011-Jun-1	SRA038885	Revision of JMW65-15 suppressive phenotype

Publications using NGS

Corresponding NGS data

## Results and Discussions

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

Statistics (as of Dec 11, 2011)	Big Projects relevant to diseases (top 10)
The Cancer Genome Atlas Project at NCI/NHGRI	30167
ARRA Autism Sequencing Collaboration	5222
Towards a Genomic Understanding of Myeloma	894
Genomic Sequencing of Head and Neck Cancer	883
Whole Exome Sequencing in Early-Onset Myocardial Infarction	621
High-throughput semi-quantitative analysis of insertional mutations in heterogeneous tumors	379
NCI Cancer Genome Characterization Initiative (CGCI)	266
MicroRNA sequence and expression analysis in breast tumors by deep sequencing	245
Frequent mutations of ubiquitin mediated proteolysis pathway in clear cell renal cell carcinoma	219
Frequent mutations of genes in transitional cell carcinoma of the bladder	219

## Data Visualization

### Frequency List (top 10)

Disease	疾患名	# of submission
Genetic Predisposition to Disease	遺伝的素因(疾患)	9
Breast Neoplasms	乳腺腫瘍	8
Disease Progression	病勢悪化	8
Obesity	肥満	7
Malaria	マラリア	6
Chromosome Aberrations	染色体異常	6
HIV Infections	HIV感染症	5
Chromosome Breakage	染色体切断	5
Polyoidy	多倍体性	5
Disease Models, Animal	疾患モデル(動物)	4

## The sample of SRA entries relevant to disease

SRA ID	SRA Title	Disease	疾患名	PMID
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukemia genome	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	18987736
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukemia genome	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	19657110
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	18987736
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	19657110
SRA008987	In-depth characterization of the microRNA transcriptome in a leukemia progression model	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	18840523
SRA029797	Exome Sequencing Identifies Somatic Mutations in Acute Monocytic Leukemia	Leukemia, Myeloid, Acute	白血球-急性骨髄性白血病	21399634

## Tree view of diseases with SRA entries



## Conclusions

- ・ 公共データベース SRA に登録された次世代シーケンサによるデータを疾患の切り口から整理した。
- ・ 整理したデータを頻度別、性質別に見られるようにし Survey of Read Archives (SRAs) より閲覧可能にした。 <http://sra.dbcls.jp/>
- ・ GEO に収録された「発現データ」も含め、実データにさらにアクセスしやすくする整理が課題

第 34 回 日本分子生物学会年会  
パシフィコ横浜  
平成 23 年 12 月 13 日~16 日

Copyright © 2011 Takeru Nakazato (DBCLS)  
このポスターは、クリエイティブ・コモンズ・ライセンス (CC-BY) の下で  
再利用可能です。