

公共 NGS データから非モデル生物のデータをより簡単に得るための検索

仲里 猛留 (Takeru Nakazato)

nakazato@dbcls.rois.ac.jp

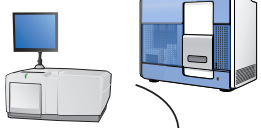
@chalkless



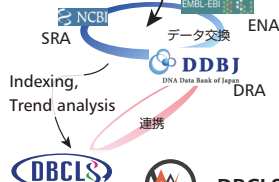
情報・システム研究機構 ライフサイエンス統合データベースセンター (DBCLS)

Backgrounds and motivations

Next generation Sequencer



Public Database (Sequence Read Archive)



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 23000 (2013 年 9 月現在) と前年の倍近くに及んでいる。DBCLS では、DDBJ と連携し、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である DBCLS SRA を構築、提供している。

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

Results and Discussions

生物種別 (Top 15)

<i>Homo sapiens</i>	2019
<i>Mus musculus</i>	1325
unidentified	867
<i>Drosophila melanogaster</i>	507
<i>Caenorhabditis elegans</i>	282
soil metagenome	271
[TaxonID]	248
<i>Arabidopsis thaliana</i>	243
marine metagenome	197
<i>Saccharomyces cerevisiae</i>	191
<i>Escherichia coli</i> str. K-12 substr. MG1655	174
Bacteria	106
human gut metagenome	93
<i>Danio rerio</i>	88
<i>Zea mays</i>	83
Total	25701 (studies)

現場の声

イネなんですけど、*japonica* とか *indica* とかもあって探すの大変なんですよね。

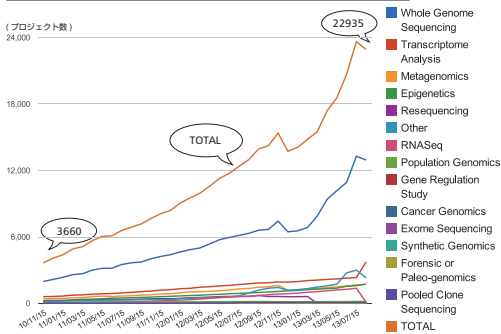
カイコをやっているんですけど、昆虫全体とかで見たんですけど。

作ってみました (ざっくりですけど)

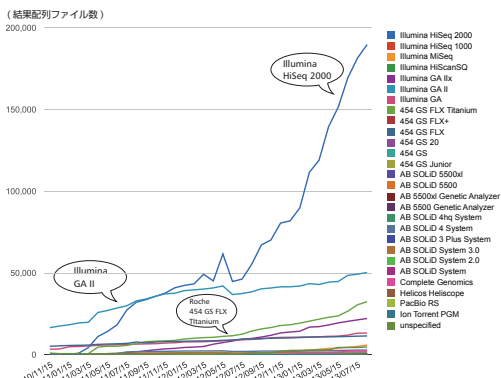
Data Visualization

Statistics (as of Sep 2, 2013)

目的別



シーケンサー別



文献からの検索

PMID	Article Title	Journal	Year	Issue	Page	Date	SRA ID	SRA Title
2187185	Efficient alignment of pyrosequencing reads for re-sequencing applications	BMC Bioinformatics	12	1	183	2011	SRX007279	Phasmodium-falciparum 3D7
21615913	A novel and well-defined benchmarking method for second generation read mapping	BMC Bioinformatics	12	-	210	2011	SRX009785	Isolation of resequencing breakpoints identified by paired-end sequencing in natural populations of <i>Drosophila melanogaster</i>
21615914	A novel and well-defined benchmarking method for second generation read mapping	BMC Bioinformatics	12	-	210	2011	SRX008355	<i>Drosophila</i> Genetic Reference Panel
21578222	Sequence-specific error profile of Illumina sequencers	Nucleic Acids Res	2011-May-15	-	-	2011-May-15	DRX000324	Whole genome resequencing of <i>S. latitans</i> (strain '98 (N4517))
21573386	A variable region within the genome of <i>Streptococcus pneumoniae</i> contributes to strain-variant variation in virulence	PLoS One	6	5	e19850	2011	SRX008324	Genomic comparisons between invasive and non-invasive serotype pneumococci
21453472	Shiga toxin sequencing of <i>Yersinia enterocolitica</i> strain W22703 (type 2, serotype O:3) genetic evidence for co-circulation between invertebrates and mammals	BMC Genomics	12	-	168	2011	ERA15964	Whole genome shotgun sequencing of <i>Caenorhabditis elegans</i> strain W22703 bivar 2, serovar O:3: modification between invertebrates and mammals
21421196	Draft genome sequence of <i>Caenorhabditis aurantiaca</i> strain H232, a <i>Thermomonas</i> from the Great Australian Bight of Australia	J Bacteriol	193	10	2664-5	2011-May	DRX000322	Whole genome resequencing of <i>S. latitans</i> (strain '98 (N4517))
21419300	Second-order selection for virulence in a large <i>Escherichia coli</i> population	Science	331	6023	1433-6	2011-Mar-18	SRX024331	Second-order selection for virulence in a large <i>E. coli</i> population
21342865	Repeat aware modeling and correction of short read errors	BMC Bioinformatics	12	Suppl 1	S82	2011	SRX001125	Paired-end sequencing of the genome of <i>Escherichia coli</i> K-12 strain MG1655 using the Illumina Genome Analyzer
21373786	Comparative whole genome sequencing reveals phenotypic RNA gene duplication in spontaneous <i>Schistosoma japonicum</i> mutants	Nucleic Acids Res	39	11	4728-42	2011-Jun-1	SRX036685	Revision of JMW03-15 suppressor phenotype

Publications using NGS

Corresponding NGS data

Taxonomy ID 入力 下位概念も検索するときはチェック

※ 随時アップデートするので、機能追加+インターフェースの変更もある予定

イネ (とその亜種) の例

4530	<i>Oryza sativa</i>	64
39947	<i>Oryza sativa japonica</i> group	35
39946	<i>Oryza sativa indica</i> group	19
1050722	<i>Oryza sativa Indica</i> Group x <i>Oryza sativa Japonica</i> Group	1
1080340	<i>Oryza sativa Japonica</i> Group x <i>Oryza sativa Indica</i> Group	1

イネのものだけだと 64 件

indica/*japonica* など入れると 120 件

カイコ (とその上位、近縁) の例

7088	Lepidoptera (チョウ目)	
37569	Bombycoidea (カイコガ上科)	
26	...	
153	7091 <i>Bombyx mori</i> (カイコ)	18
	7092 <i>Bombyx mandarina</i> (クワコ)	3
37572	Papilionoidea (アゲハチョウ上科)	
91	...	
	7143 Papilionidae (アゲハチョウ科)	6
...	...	
	40037 Heliconiinae (ドクチョウ亜科)	76

カイコだけだと 18 件

近縁も入れて 26 件 (タバコスズメガが 6 件)

チョウ目全体だと 153 件

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生物種の切り口から整理した。
- 非モデル生物について、上位概念、下位概念を検索ができるようにした。
- 整理したデータは DBCLS SRA より閲覧可能。 <http://sra.dbcls.jp/>

NGS 現場の会 第 3 回研究会
神戸国際会議場
平成 25 年 9 月 4 日~5 日