公共塩基配列データベース検索ツールとそれを活用した解析事例

情報・システム研究機構 (ROIS) データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター(DBCLS)
大田 達郎、仲里 猛留、坊農 秀雅

DBCLS Database Center for Life Science

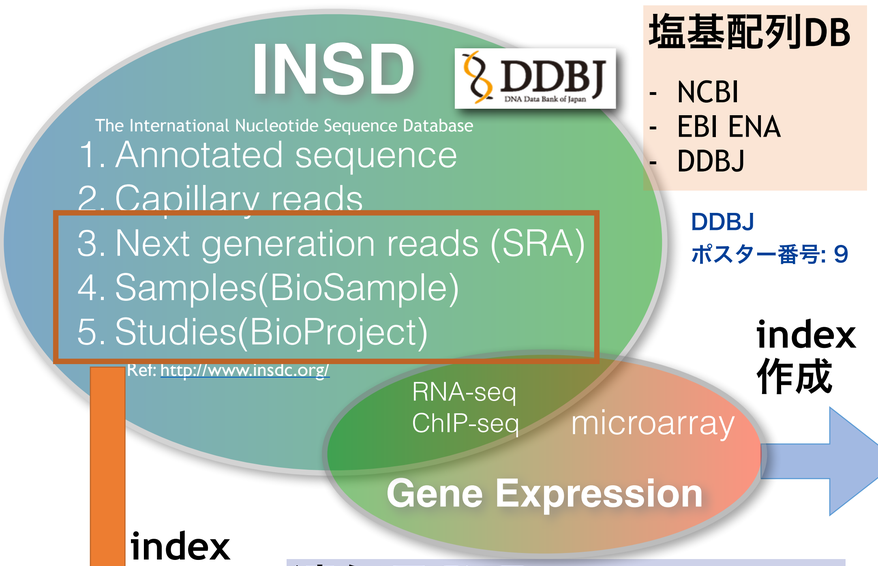
© 2016 DBCLS Togo Picture Gallery / CC-BY 4.0

DBCLSでは、大量に蓄積してきた次世代シークエンサーの塩基配列データを検索できる手段の開発を行ってきた[1,2]。それはDBCLS SRA (http://sra.dbcls.jp/)として実装され、実験条件のデータ（メタデータ）に対してキーワード検索が可能となっており、DDBJの検索エンジンとしても利用されている。また、遺伝子発現DB目次AOE (All of gene expression; https://aoe.dbcls.jp/）では、DBCLS SRAのAPIを利用することで、複数の公共遺伝子発現DBが統合的に検索可能となっている[3]。さらに、これらの検索ツールをフル活用したデータ解析研究にも取り組んでおり、低酸素刺激による遺伝子発現変動のメタ解析を活用解析事例として報告する[4]。

参考文献

[1] Nakazato T, Ohta T, Bono H; doi: 10.1371/journal.pone.0077910
[2] Ohta T, Nakazato T, Bono H ;doi: 10.1093/gigascience/gix029
[3] Bono H; doi: 10.1101/626754
[4] Bono H; doi: 10.1101/267310

公共塩基配列DBと遺伝子発現DB



塩基配列DB

- NCBI
- EBI ENA
- DDBJ

DDBJ
ポスター番号: 9

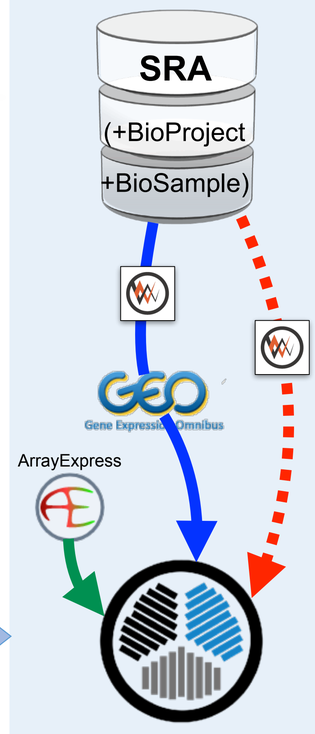
index 作成

index 作成

遺伝子発現DB

- NCBI Gene Expression Omnibus (**GEO**)
  - https://www.ncbi.nlm.nih.gov/geo/
- EBI **ArrayExpress**
  - https://www.ebi.ac.uk/arrayexpress/
- DDBJ Genomic Expression Archive (**GEA**)
  - https://www.ddbj.nig.ac.jp/gea/

AOE indexデータ作成フロー

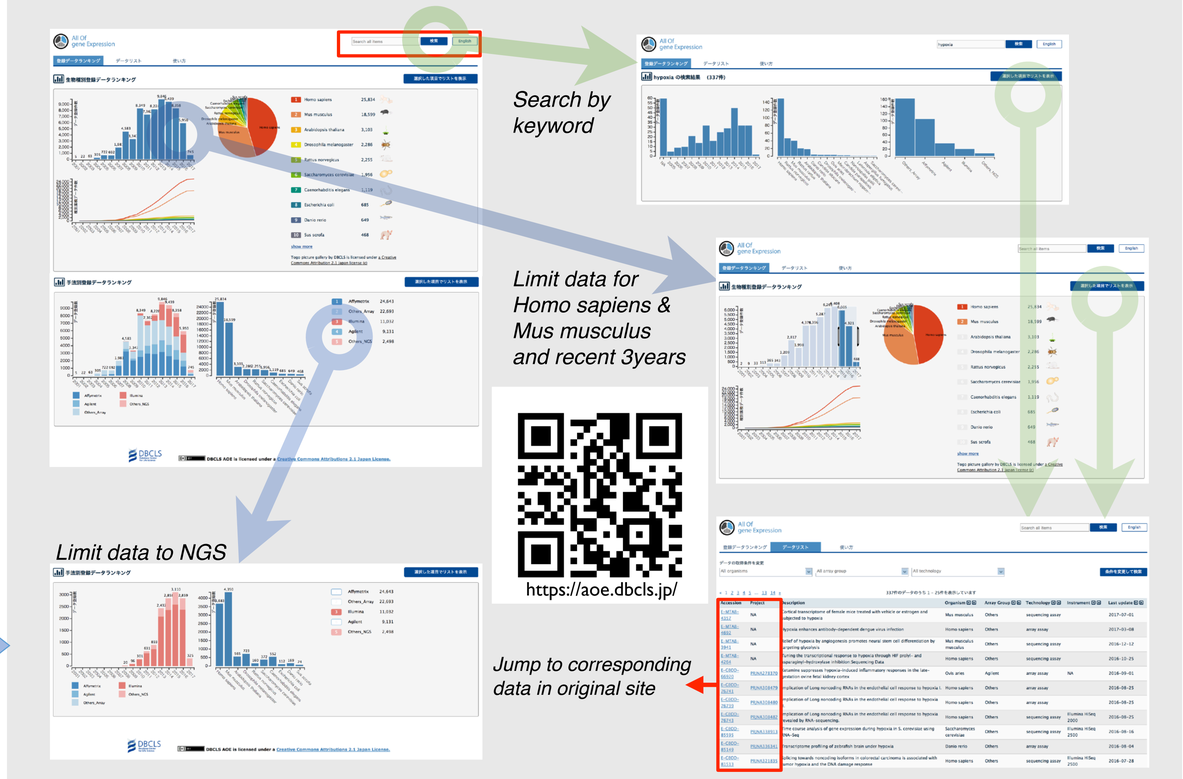


統計値 可視化

		# of series	Cumulative
1	**AE (including GEO) + GEA**	72,770 (GEO-derived 59,374)	72,770
2	**GEO** not in AE	37,849	110,619
3	**SRA (RNA-seq)** not in AE + GEO	13,137	123,756

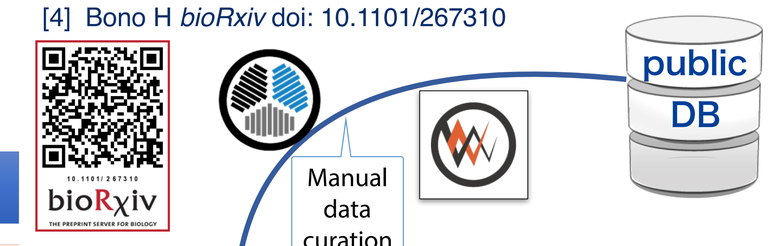
AOE indexは駿府城の三重の堀の構造のように漏れ無くデータを統合

All of gene expression (AOE) https://aoe.dbcls.jp/

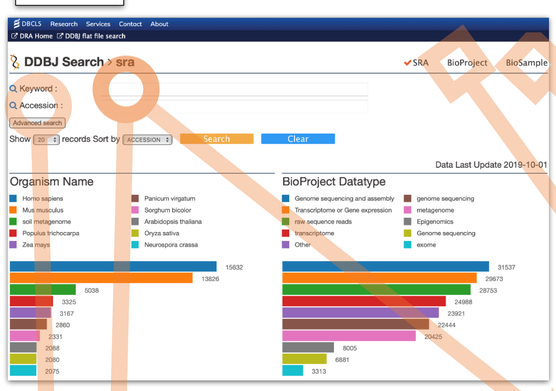


利用実例: 低酸素トランスクリプトームのメタ解析事例

[4] Bono H *bioRxiv* doi: 10.1101/267310



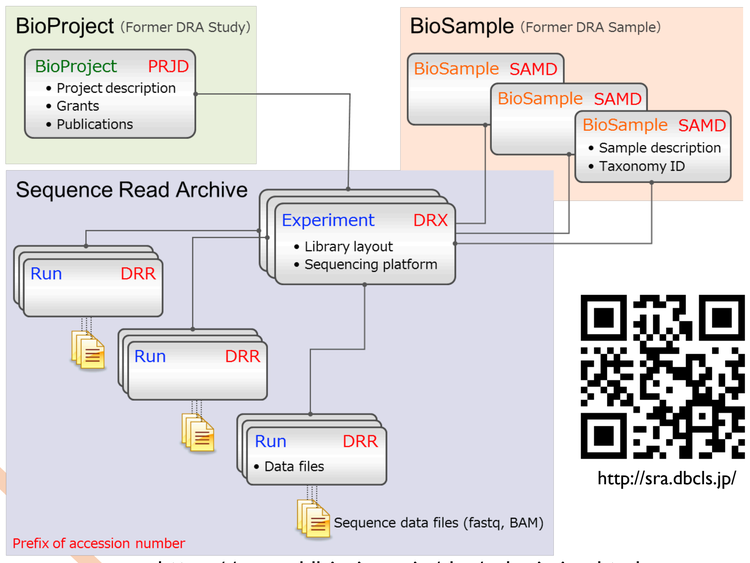
DBCLS SRA (DDBJ Search) http://sra.dbcls.jp/



- キーワード
- DBのID

SRA
- SRA (例: DRA008013)
- SRP (例: SRP045624)
- SRX (例: ERX182686)
- SRR (例: ERR266349)

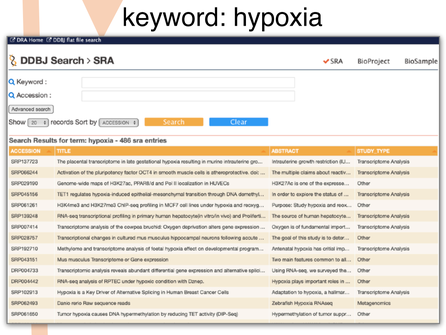
BioProject (例: PRJDB4976)
BioSample (例: SAMD00079763)



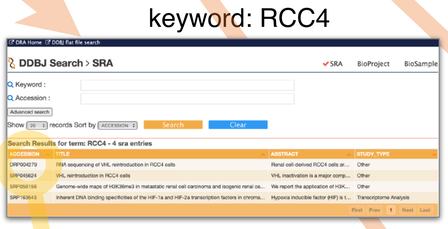
http://sra.dbcls.jp/

https://www.ddbj.nig.ac.jp/dra/submission.html

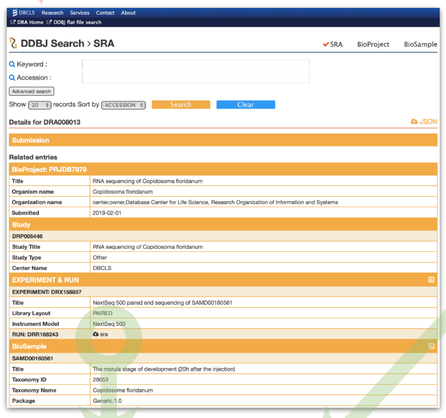
keyword: hypoxia



keyword: RCC4

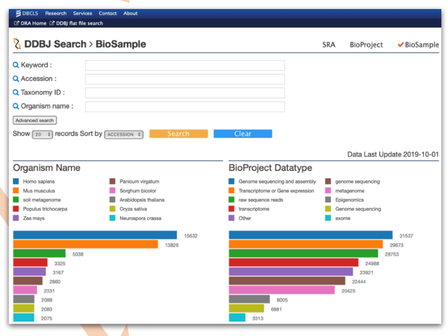


Accession: DRA008013

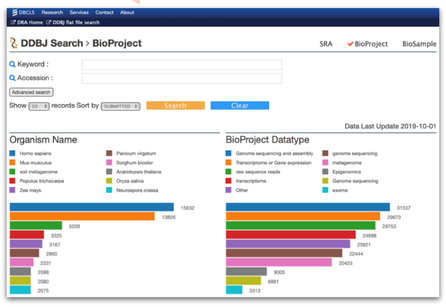


RUNファイルがダウンロード開始
＝ダウンロード先の可能な直リンク

BioSample:



BioProject:



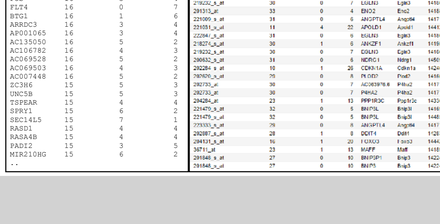
JSON data

up/down/ unchanged Classification

Human RNA-seq

Gene	up	down	unchanged
EGR3	20	2	1
FOSB	19	0	4
SDAD1P1	17	6	0
PRRC1	17	3	3
EGR1	17	0	6
DDIT4	17	6	0
BHLHE40	17	6	0
AC147067	17	2	4
AC132938	17	3	3
ZNF295	16	6	1
ZCWPW2	16	3	4
ZFP36	16	5	2
PGK1P2	16	6	1
OSER1	16	3	4
MAFF	16	6	1
LINC01220	16	6	1
LINC00482	16	3	4
HIST1H2AC	16	1	6
GDF15	16	0	7
FXYD6	16	5	2
FOS	16	3	4
FLT4	16	0	7
BTG1	16	1	6
ARRDC3	16	3	4
AP001065	16	3	4
AC135050	16	5	2
AC106782	16	4	3
AC069528	16	5	2
AC069503	16	4	3
AC027448	16	5	2
ZC3H6	15	5	3
UNC5B	15	5	3
TSPEAR	15	4	4
SPRY1	15	6	2
SEC14L5	15	7	1
RASD1	15	4	4
RASA4B	15	4	4
PADI2	15	3	5
MIR210HG	15	6	2

Curated datasets



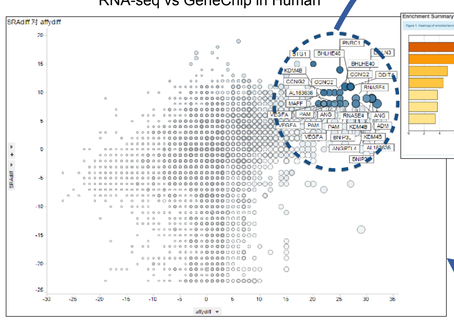
Summary of datasets

	curated hypoxia-normoxia pairs
human GeneChip	**37** pairs from 11 data series
mouse GeneChip	**53** pairs from 8 data series
Human RNA-seq	**23** pairs from 7 data series

Expression quantification

Gene Set Enrichment Analysis

RNA-seq vs GeneChip in Human



Data visualization & evaluation

Human-mouse GeneChip

まとめ

- 高速にSRAと関連DBを検索するウェブツールを開発 (DBCLS SRA)
- 散在する公共発現DBを統合検索するウェブツールを開発 (AOE)
- それらを使った公共DBのメタ解析を低酸素トランスクリプトームで実施

今後の課題

- これらのツールの安定的な稼働と利用普及
- 利用可能なDBはRDFを使って参照、つなげて利用するようにしていく利用例作成

謝辞

- SRA: 国立遺伝学研究所 生命情報・DDBJセンター 児玉悠一
- システム開発: dogrun inc 大石直哉

Licensed under CC-BY 4.0 ©2019 HIDEMASA BONO (DBCLS)