

AJACS名古屋2

次世代シーケンサー（NGS）と 関連するデータベース・ツール

仲里 猛留

NAKAZATO, Takeru



@chalkless

情報・システム研究機構 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

Database Center for Life Science (DBCLS),
Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)



2018/12/5

@愛知県がんセンター



NBDC の広報サイト

バイオサイエンス ×DB=∞

 Web https://events.biosciencedbc.jp/[ホーム](#) | [シンポジウム](#) | [講習会](#) | [学会・展示会](#) | [記事](#)

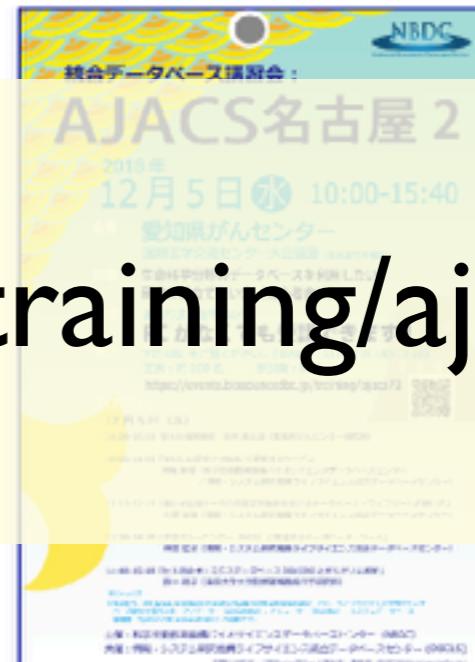
統合データベース講習会：AJACS名古屋2

統合データベース講習会：AJACSは、生命科学系のデータベースやツールの使い方、データベースを統合する活動を紹介する初心者向けの講習会です。

今回の講習会では、生命科学系データベースのカタログ、横断検索、アーカイブ、ヒトデータに関するサービス・ツール等の紹介に加えて、遺伝子発現データベース、次世代シークエンスデータベース、疾患マルチオミクスデータベースについても紹介します。パソコンを使用しない講義形式の講習です。

https://events.biosciencedbc.jp/training/ajacs73

- 対象： 生命科学分野のデータベースを利用したい、研究に役立てたい方（初心者向け）。
- 日時： 2018年12月5日（水）10:00-15:40
(開場および受付は、開始時間の30分前より)
- 会場： 愛知県がんセンター 国際医学交流センター 大会議室
(愛知県名古屋市千種区鹿子殿1番1号)
[\[公共交通機関でのアクセス（市内）\]](#) [\[敷地内案内図（歩行者）\]](#)
- 定員： 約100名
- 参加費： 無料
- PC： 不要です。
- 申込： 申し込み受付は終了しましたが、お席に若干の余裕がございますので、受講希望の方は、統合データベース講習会事務局までお問合せください。



NGSとは？

次世代シーケンサー（NGS）とは

- 次世代シーケンサ
- 次世代シーケンサー
- 新型シーケンサ



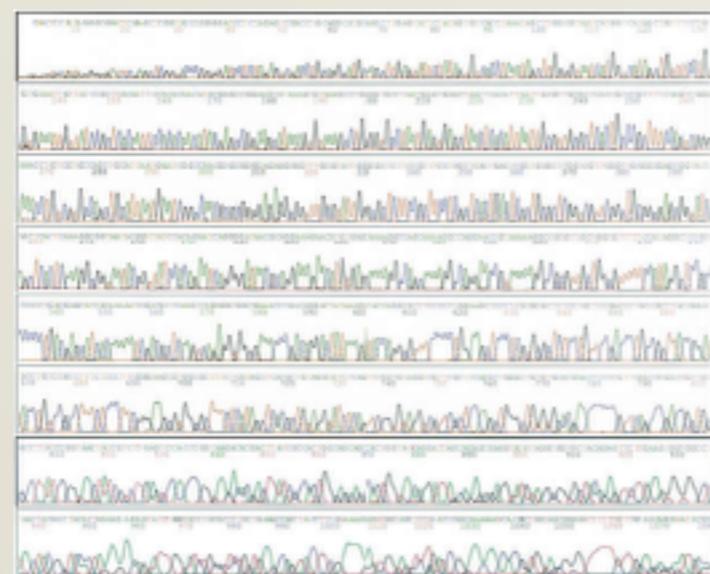
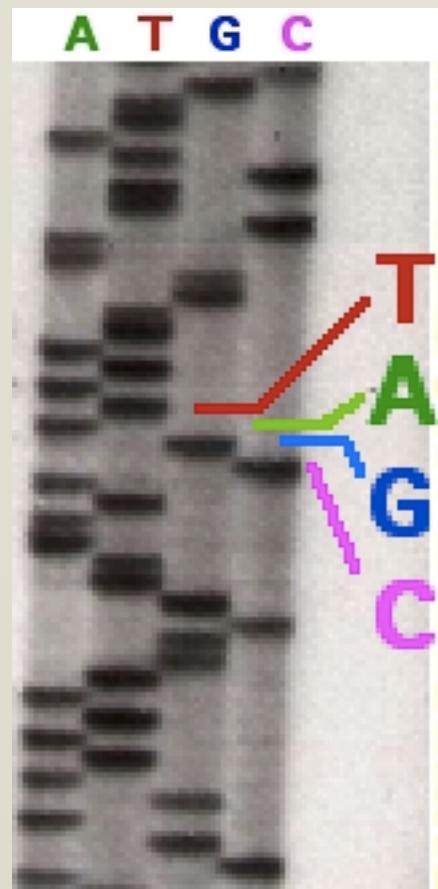
Illumina HiSeq



PacBio

- New-generation Sequencing (NGS)
- Next-generation Sequencing (NGS)
- 他にmassively parallel DNA sequencingとか…
- 最近は、
High-throughput DNA sequencing (technology)
をよく使う印象（略語はNGS）

何が「次世代」か？



電気泳動式

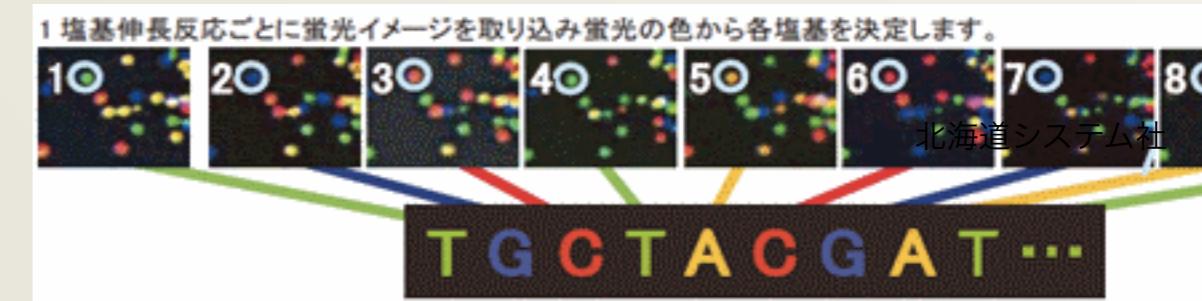
キャピラリ式

NGS

750 (base/lane) \times 48/4 lanes
= 9kbase

500 (base/lane) \times 96 lane
= 48kbase

2 \times 300 (base/seq) \times 25M seq/run
= 15 Gbase



目的別 必要スペック

アプリケーション別に必要なリードスペック

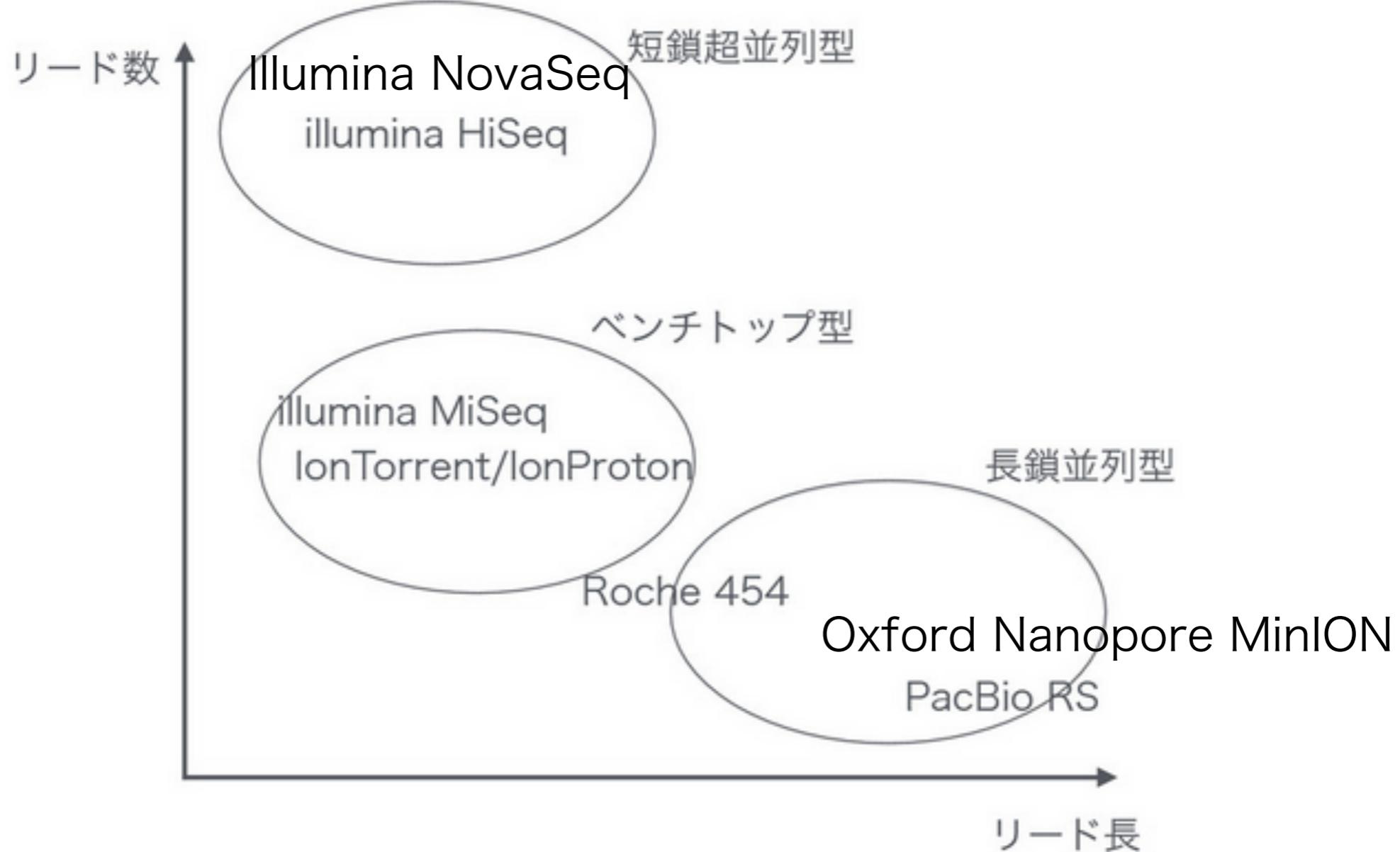
application / 実験種	total bases / 総塩基数	read length / リード長	read number (M) / リード数
ヒトゲノムリシーケンス	90-150Gb	2x100	900-1500
ターゲットリシーケンス	<1Gb	2x100	10
exome sequence	5~7Gb	2x100	70
RNA-Seq	5Gb	2x100	50
TSS-Seq	1Gb	1x50	20
small RNA	0.35Gb	1x35	>10
微生物ゲノム	>150Mb	2x100	>1.5
真核生物ゲノム	>4Gb	2x100	>40
Bisulfite-Seq	90-150Gb	2x100	900-1500
ChIP-Seq	>6Gb	1x100	60

注: 対象のゲノムサイズなどで数字が変わることがあります。また、既に情報が古くなっている可能性もあります

細胞工学別冊 次世代シーケンサー目的別アドバンストメソッド p21より引用

目的別 機器の選択

Sequencers by Read spec (ざっくり)



NGSのデータベース

ライフ系データベース：BLASTとPubMed

PubMed

The screenshot shows a web browser with two tabs open. The left tab is for PubMed (www.ncbi.nlm.nih.gov/pubmed/?term=NGS) and the right tab is for NCBI BLAST (blast.ncbi.nlm.nih.gov/Blast.cgi#). The PubMed page displays search results for 'ngs' gene function, showing 1 article and 1441 more. The NCBI BLAST page shows the results for a nucleotide sequence query (9EN9RPMN01R), including a graphic summary of conserved domains and a distribution plot of blast hits.

PubMed - NCBI

www.ncbi.nlm.nih.gov/pubmed/?term=NGS

NCBI Resources How To Sign In to NCBI

PubMed NGS

RSS Save search Advanced

Show additional filters

Article types Clinical Trial Review More ...

Text availability Abstract available Free full text available Full text available

Publication dates 5 years 10 years Custom range...

Species Humans Other Animals

Clear all Show additional filters

Display Settings: Summary, 20 per page, Sort by relevance

See 1 article about ngs gene function
See also: ngs notochord granular surface in t

Results: 1 to 20 of 1441

1. An In-Solution Hybridisation Method for rich Clinical Samples for Analysis by NG Smith M, Campino S, Gu Y, Clark TG, C Dondorp AM, Kwiatkowski DP, Quail MA Open Genomics J. 2012;5. doi: 10.2174/18756 PMID: 24273626 [PubMed] Related citations

2. Transactivating mutation of the MYOD1 rhabdomyosarcoma. Szuhai K, de Jong D, Leung WY, Fletcher J Petrol. 2013 Nov 25. doi: 10.1002/pech.4307 PMID: 24272621 [PubMed - as supplied by publisher] Related citations

3. Introduction to Statistical Methods for M Zararsiz G, Coşgun E. Methods Mol Biol. 2014;1107:129-55. doi: 10.1007/978-1-4614-8635-5_5 PMID: 24272435 [PubMed - in process] Related citations

4. High-Throughput Approaches for Microbiome Dedeoğlu BG. Methods Mol Biol. 2014;1107:91-103. doi: 10.1007/978-1-4614-8635-5_8 PMID: 24272433 [PubMed - in process] Related citations

NGSmethDB: an updated genome reso

NCBI Blast:Nucleotide Seq X

blast.ncbi.nlm.nih.gov/Blast.cgi#

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST/blastdb/Formatting Results - 9EN9RPMN01R

Edit and Resubmit Save Search Strategies >Formatting options >Download YouTube How to read this page Blast report description

Nucleotide Sequence (3065 letters)

RID 9EN9RPMN01R (Expires on 11-29 21:37 pm)

Query ID Icl|171457

Description None

Molecule type nucleic acid

Query Length 3065

Database Name refseq_protein

Description NCBI Protein Reference Sequences

Program BLASTX 2.2.28+ >Citation

Other reports: >Search Summary >Taxonomy reports

Graphic Summary

Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

RF #1 Specific hits Sulfate_transp Superfamilies Sulfate_transp Superfamily Multi-domains

Distribution of 112 Blast Hits on the Query Sequence

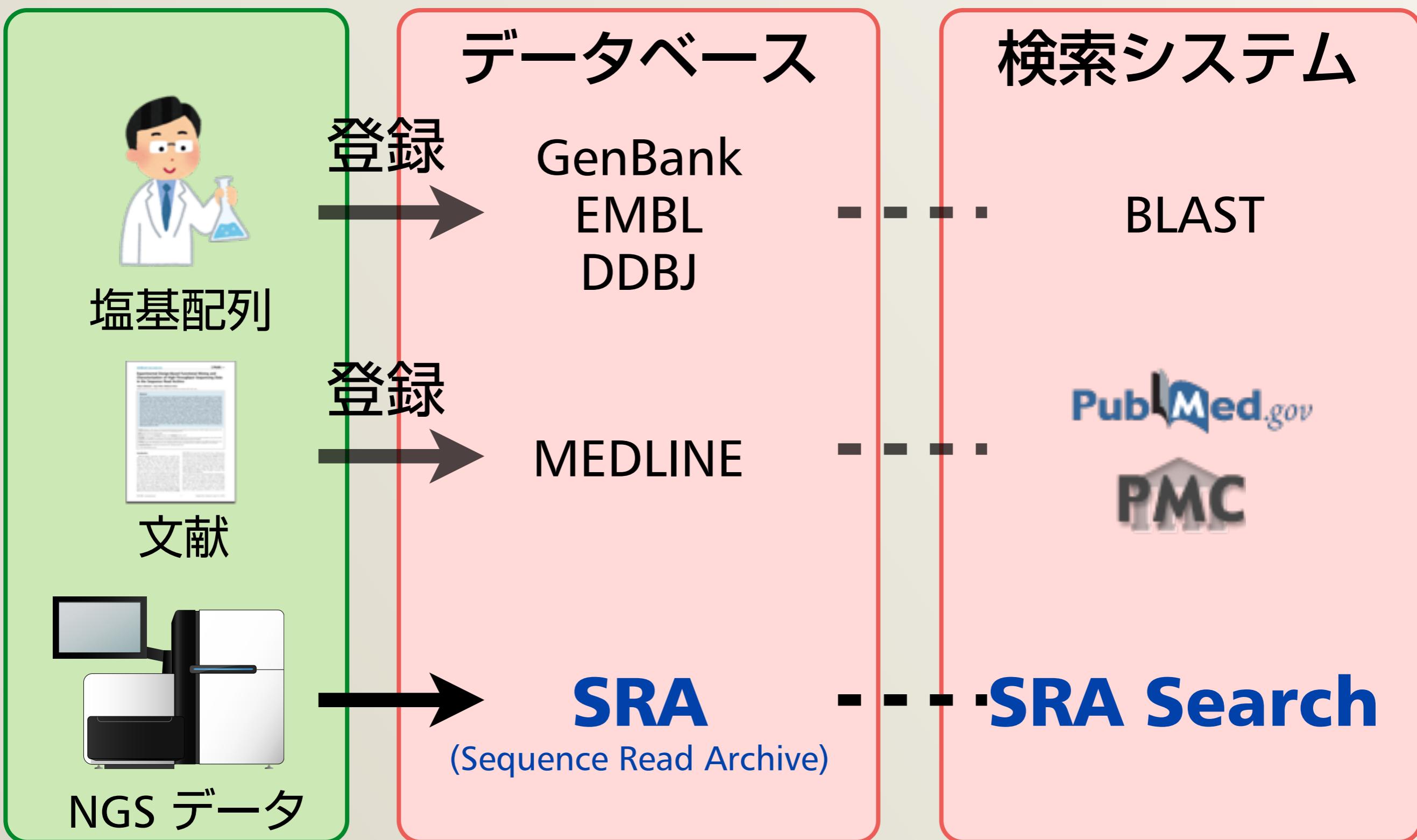
Color key for alignment scores

<40	40-50	50-60	60-80	80-200	>=200
-----	-------	-------	-------	--------	-------

Query 1 600 1200 1800 2400 3000

BLAST

研究データと公共データベース



Result List - DRA Search x

trace.ddbj.nig.ac.jp/DRASearch/query?keyword=SNP&show=20

Send Feedback Search Home DRA Home

DRASearch

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword : SNP

Show 20 records Sort by Search Clear

SRA:

Search Results (1748 records)

<< < 1 / 88 > >>

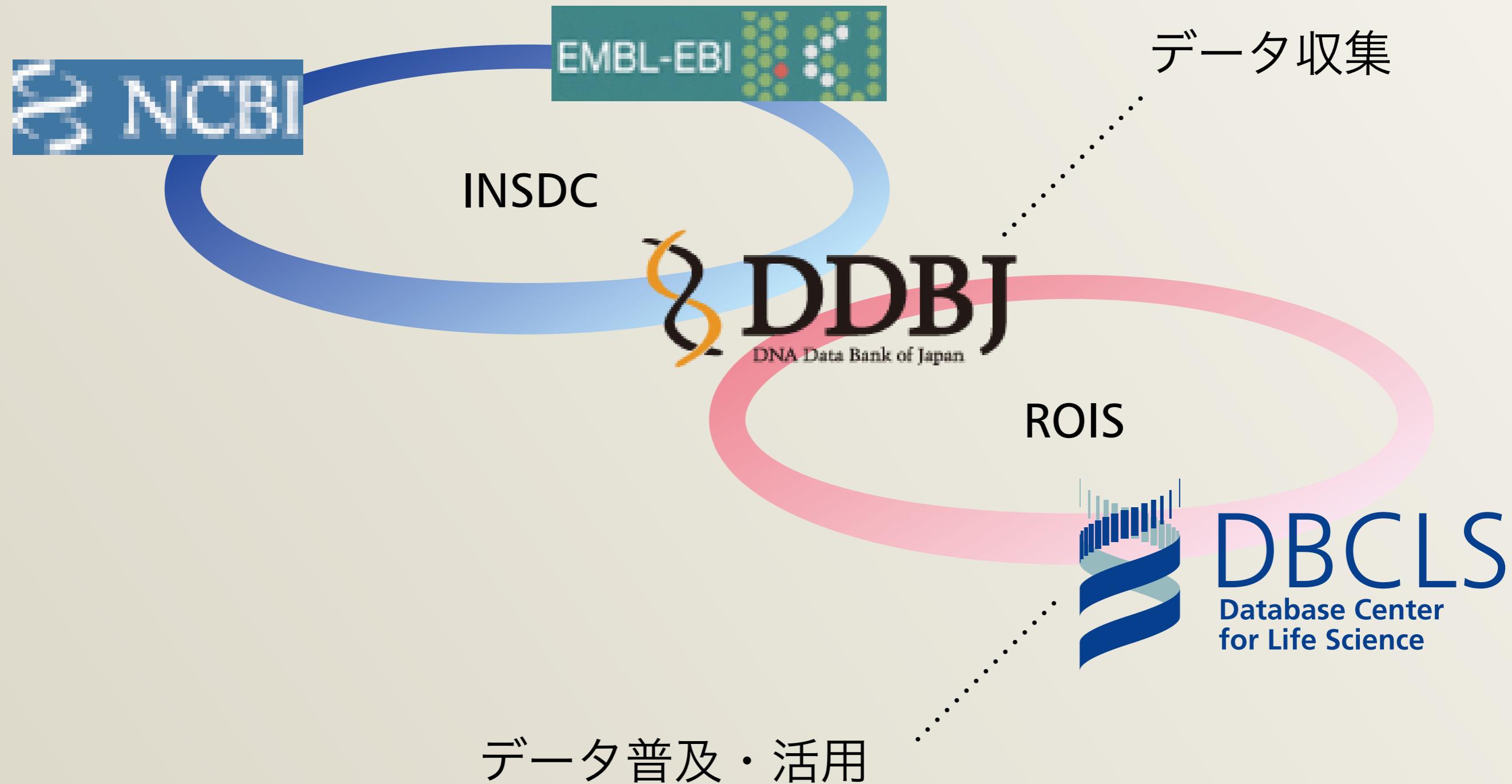
Filtered by

document type:run(1210) study(256) experiment(172) submission(70) sample(39) analysis(1)
organism:Homo sapiens(1208) Ovis aries(74) Anguilla anguilla(66) Mus musculus(29) Oncorhynchus mykiss(9)
Oryza sativa Japonica Group(6)

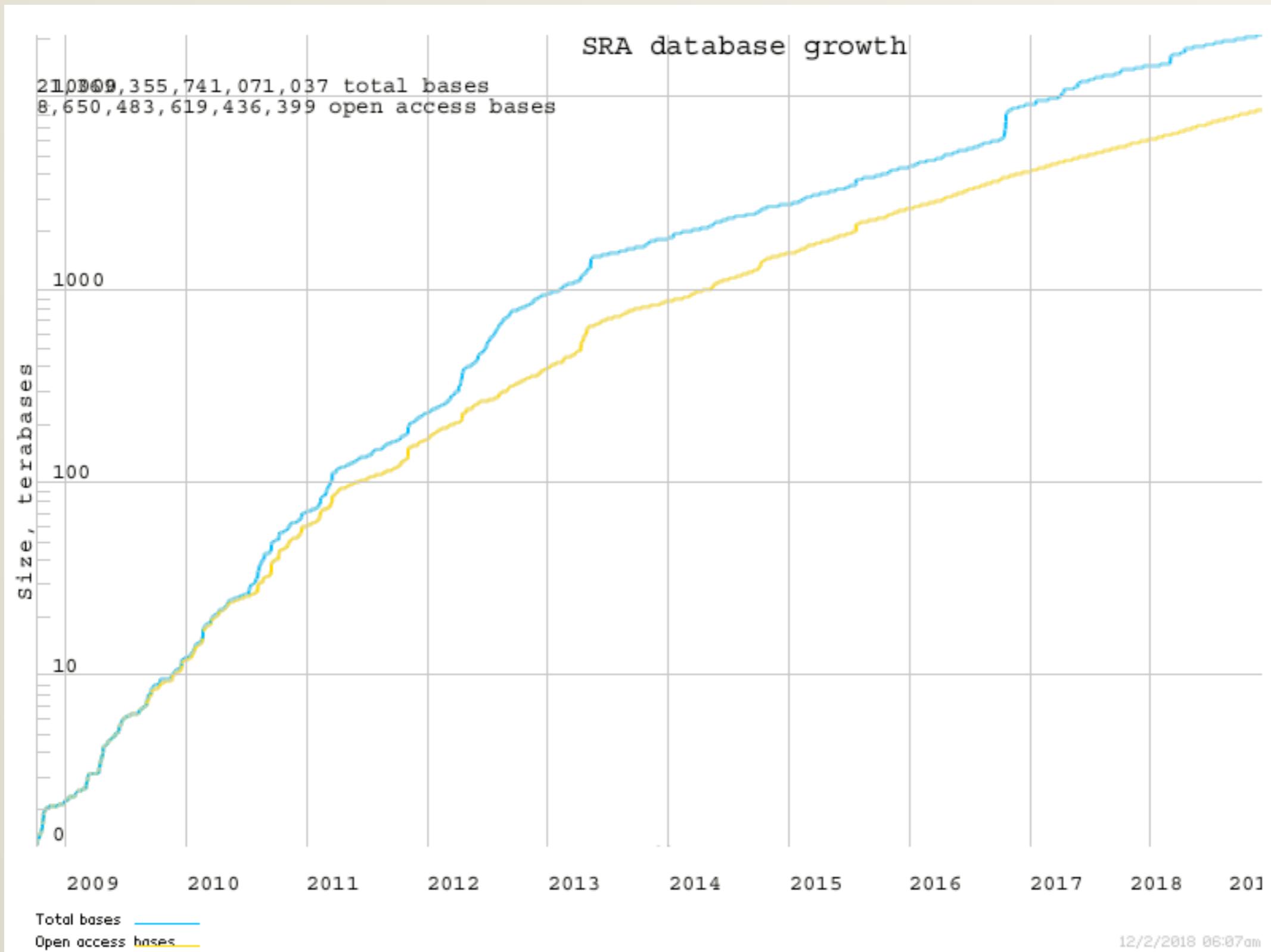
#	META_FILE	STUDY_ID	STUDY_NAME	STUDY_TYPE	ORGANISM	FILE_FORMATTED	CENTER_NAME
1	ERA007211.submission.xml <?xml version="1.0" encoding="UTF-8"?> <SUBMISSION accession="ERA007211" alias="UU-SNP_1"	ERA007211	ERP000177 GEUVADIS_RNASEQ_FIVE_INDIVIDUAL_PILOT	Transcriptome Analysis	Enterobacteria phage phiX174 sensu lato Homo sapiens	124.1G	UNIGE
2	ERA007211.run.xml .org/2001/XMLSchema-instance"> <RUN alias="618MTAAXX_3" center_name="UU-SNP" run_center="UU-SNP" accession="ERR011434"	ERR011434 ERR011435 ERR011436 ERR011437 ERR011438 ERR011439	ERP000177 GEUVADIS_RNASEQ_FIVE_INDIVIDUAL_PILOT	Transcriptome Analysis	Enterobacteria phage phiX174 sensu lato Homo sapiens	124.1G	UNIGE
3	SRA114122.submission.xml <?xml version="1.0" encoding="UTF-8"?> <SUBMISSION alias="Lus SNP discovery" submission	SRP033223	Lithum usitatissimum Genome sequencing	Whole Genome Sequencing	Lithium usitatissimum	24G	BioProject
4	SRA024703.study.xml .org/2001/XMLSchema-instance"> <STUDY alias="RNA-seq for SNP" center_name="ouc" accession="SRP003814"> <DESCRIPTOR> <STUDY	SRP003814					
	SRA072630.submission.xml <?xml version="1.0"						

ちなみに、昔は Short Read Archive

NGSデータの収集と提供



登録されたNGSデータの伸び



JGA (Japanese Genotype-Phenotype Archive)

Controlled-access データのアーカイブ

The screenshot shows the homepage of the Japanese Genotype-phenotype Archive (JGA) at trace.ddbj.nig.ac.jp/jga/index.html. The page features a header with the DDBJ logo, the title "Japanese Genotype-phenotype Archive", and navigation links for Home, Studies, and Submission. The main content area includes sections on Overview, Data Use, and Data Submission, along with a search bar and links to the National Bioscience Database Center (NBDC).

概要

Japanese Genotype-phenotype Archive (JGA) は個人を特定される可能性のある遺伝学的なデータと表現型情報を保存し、提供しています。データが収集された個人との間の同意に基づく協定により、JGA のデータ利用は特定の研究目的に制限されています。JGA は厳格なプロトコールに従い、情報を管理、格納、提供しています。登録処理が終わった全てのデータは暗号化されます。JGA チームには [こちらから連絡](#)することができます。

なお、JGA に登録されるデータおよびデータの利用についての審査は独立行政法人科学技術振興機構 (JST)/バイオサイエンスデータベースセンター (NBDC) が実施しています。JGA は科学技術振興機構 National Bioscience Database Center (NBDC) と共同で運営されています。

データの利用

JGA はデータを格納する際にそのデータに適用される利用制限ポリシーを登録しますが、利用者のデータ利用の可否については JST/NBDC が審査します。利用者は NBDC にデータの利用を申請し、JGA は NBDC からの利用承認連絡を受け、利用者にデータへの安全なアクセスを提供します。

データの登録

JGA は JST/NBDC で承認された匿名化されたデータだけを受け付けています (ヒトを対象とした研究データの登録について)。登録者は JST/NBDC に JGA へのデータ提供を申請し、NBDC からデータ提供の承認連絡を受けた登録者は JGA に連絡します。JGA チーム

登録などはJST NBDCに

The screenshot shows the homepage of the NBDC Human Database Beacon. The URL in the address bar is humandbs.biosciencedbc.jp. The page title is "NBDCヒトデータベース". The navigation menu includes "ホーム", "データの利用", "データの提供", "ガイドライン", "NBDCヒトデータ審査委員会", "成果発表", and "アクセス統計". There are two red warning boxes at the top: one about maintenance on Jan 24, 2017, and another about JGA maintenance from Feb 10 to 20, 2017. A section titled "NBDCヒトデータベースについて" discusses the database's purpose and collaboration with DDBJ. A "新着情報" box lists recent publications. At the bottom, there are links for searching the database and information about its membership in the C44CH Beacon Network.

● ● ● ホーム - NBDCヒトデータベース ×

humandbs.biosciencedbc.jp

NBDC National Bio-Science Database Center NBDC ヒトデータベース

English サイト内検索 検索

ホーム データの利用 データの提供 ガイドライン NBDCヒトデータ審査委員会 成果発表 アクセス統計

⚠ メンテナンスのため、NBDCヒトデータベースのサイトを一時停止します。予定日時：2017/1/24 9:00-18:00

⚠ メンテナンスのため、JGAが一時停止します。停止するサービスの詳細は[こちら](#)。停止期間：2017/2/10 18:00 - 2017/2/20 午前

NBDCヒトデータベースについて

ヒトに関するデータは、次世代シーケンサーをはじめとした解析技術の発達に伴って膨大な量が産生されつつあり、それらを整理・格納して、生命科学の進展のために有効に活用するためのルールや仕組みが必要です。

国立研究開発法人科学技術振興機構(JST)バイオサイエンスデータベースセンター(NBDC)では、個人情報の保護に配慮しつつヒトに関するデータの共有や利用を推進するために、ヒトに関する様々なデータを共有するためのプラットフォーム『NBDCヒトデータベース』を設立するとともに、国立遺伝学研究所 DNA Data Bank of Japan (DDBJ)と協力して、ヒトに関するデータを公開しています。

本Webサイトを通じて、ヒトに関するデータの利用及びヒトに関するデータの提供を行なうことができます。データ共有についての概要是[こちら](#)をご参照下さい。

新着情報

2017/01/12 東京医科歯科大学 医歯学総合研究科（医系）分子腫瘍医学からの制限公開データ（Type I）を公開しました (hum0041)

2016/12/27 大阪大学大学院 医学系研究科 外科学講座 消化器外科学からの制限公開データ（Type I）を公開しました (hum0039)

ニュースへ

Search NBDC Human Database Beacon for Alternative Alleles [API help]

NBDC Human Database Beacon is a member of C44CH Beacon Network.

SRAの検索

NCBIのインターフェース

ncRNA expression, including snoRNAs

保護された通信 https://www.ncbi.nlm.nih.gov/sra/ERX005900[accn]

ERX005900: ncRNA expression, including snoRNAs, across 11 tissues using polyA-neutral amplification
1 ILLUMINA (Illumina Genome Analyzer II) run: 31.8M spots, 1.1G bases, 1Gb downloads

Design: ncRNA expression, including snoRNAs, across 11 tissues using polyA-neutral amplification

Submitted by: TRON

Study: ncRNA expression, including snoRNAs, across 11 tissues using polyA neutral amplification
[PRJEB2202](#) • [ERP000257](#) • All experiments • All runs

Sample: **Protocol:** We purchased total RNA from Ambion (Austin, USA). Each tissue samples was pooled from multiple donors. Librarys were prepared as in Armour CD, Castle JC, Chen R, Babak T, Loerch P, et al. (2009) Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* 6: 647-649.
[SAMEA733156](#) • [ERS012469](#) • All experiments • All runs

Organism: [Homo sapiens](#)

Library:
Name: adipose_RNA
Instrument: Illumina Genome Analyzer II
Strategy: OTHER
Source: GENOMIC
Selection: RANDOM
Layout: SINGLE
Construction protocol: We purchased total RNA from Ambion (Austin, USA). Each tissue samples was pooled from multiple donors. Librarys were prepared as in Armour CD, Castle JC, Chen R, Babak T, Loerch P, et al. (2009) Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* 6: 647-649.

Spot descriptor:
1 forward

Experiment attributes:
Experimental Factor: ORGANISM/PART: adipose

Runs: 1 run, 31.8M spots, 1.1G bases, 1Gb

Run	# of Spots	# of Bases	Size	Published
ERR015534	31,807,065	1.1G	1Gb	2011-03-18

BioProject
BioSample
Taxonomy

Recent activity

Turn 01 Clear

ERP000257 (11) SRA

Uberon, an integrative multi-species anatomy ontology

Tumour resistance in induced pluripotent stem cells derived from naked mole-rats

Tumour resistance in induced pluripotent stem cells derived from naked mole-rats PubMed

bono hidemasa (47) PubMed

See more...

EBIのインターフェース

www.ebi.ac.uk/ena/data/view/ PRJEB2202

ncRNA expression, including snoRNAs, across 11 tissues using polyA-neutral amplification

Navigation Read Files Portal Attributes

Bulk Download Files ⚠ (Please use Firefox to launch the bulk downloader app.)

Download: 1 - 11 of 11 results in TEXT

Select columns

Showing results 1 - 10 of 11 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	NCBI SRA file (ftp)	NCBI SRA file (galaxy)	CRAM Index files (ftp)	CRAM Index files (galaxy)
PRJEB2202	SAMEA733155	ERS012459	ERX005900	ERR015534	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			File 1	File 1				
PRJEB2202	SAMEA733146	ERS012465	ERX005906	ERR015535	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			File 1	File 1				
PRJEB2202	SAMEA733149	ERS012460	ERX005901	ERR015536	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			File 1	File 1				
PRJEB2202	SAMEA733145	ERS012464	ERX005905	ERR015537	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			File 1	File 1				

DDBJのインターフェース

Result List - DRA Search X ERP000257 - DRA Search X

trace.ddbj.nig.ac.jp/DRASearch/study?acc=ERP000257

DRA Search Send Feedback Search Home DRA Home

ERP000257

Study Detail

Title	ncRNA expression, including snoRNAs, across 11 tissues using polyA-neutral amplification
Study Type	Transcriptome Analysis
Abstract	
Description	ncRNA expression, including snoRNAs, across 11 tissues using polyA-neutral amplification
Center Name	TRON

SRA Links

Url Link	E-MTAB-305 in ArrayExpress
----------	--

Navigation

Submission	ERA010380	FTP
Experiment	ERX005900	FASTQ
	ERX005901	SRA
	ERX005902	SRA
	ERX005903	SRA
	ERX005904	SRA
	ERX005905	SRA
	ERX005906	SRA
	ERX005907	SRA
	ERX005908	SRA
	ERX005909	SRA
	ERX005910	SRA
Sample	ERS012159	
	ERS012160	
	ERS012161	
	ERS012462	
	ERS012463	
	ERS012464	
	ERS012465	
	ERS012466	
	ERS012467	
	ERS012468	
	ERS012469	

DBCLS SRA 公共NGSデータの検索サイト

① 保護されていない通信 | sra.dbcls.jp

DBCLS Research Services Contact About
DRA Home DDBJ flat file search

DDBJ Search > sra ✓ SRA BioProject BioSample

Keyword :
Accession :
Advanced search

Show 20 records Sort by ACCESSION

http://sra.dbcls.jp/

Data Last Update 2018-11-15

Organism Name

Organism Name	Count
Homo sapiens	39219
Mus musculus	26402
Panicum virgatum	4769
soil metagenome	4493
Arabidopsis thaliana	4179
Populus trichocarpa	3775
marine metagenome	3193
Zea mays	2900
Rattus norvegicus	2845
Saccharomyces cerevisiae	2562

Project Datatype

Project Datatype	Count
Transcriptome or Gene expression	78048
Genome sequencing and assembly	47205
genome sequencing	28631
transcriptome	28935
raw sequence reads	21095
metagenome	20268
Other	20127
RefSeq Genome	12598
Genome sequencing	12012
Epigenomics	9206

DBCLS SRA – 詳細表示

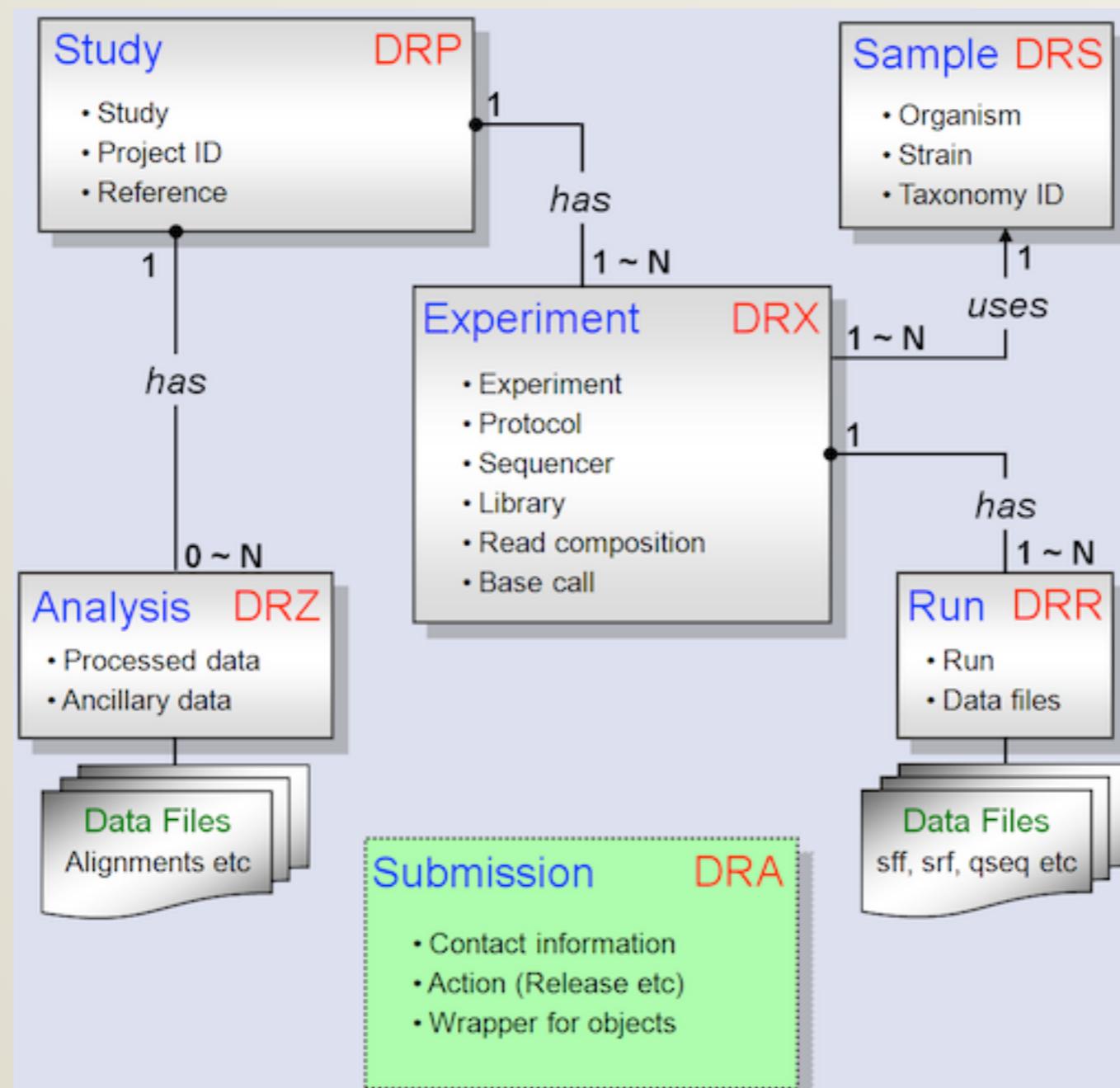
The screenshot shows a web browser window with two tabs, both titled "DDBJ Search". The active tab displays the URL sra.dbcls.jp/details.html?db=sra&accession=DRP001175. The page header includes links to "DRA Home", "DDBJ flat file search", and navigation icons. The main content area is titled "DDBJ Search > SRA". It features search fields for "Keyword" and "Accession", an "Advanced search" button, and search controls for "Show 20 records Sort by ACCESSION". Below these are buttons for "Search" and "Clear". A "Details for DRP001175" section contains a table with study information, including the title "RNA Sequencing to improve annotation and detect expression changes during anhydrobiosis", study type "Other", center name "BioProject", center project name "Ramazzottius varieornatus", and XREF ID "27649274". An orange "JSON" button is located next to this table. A "Related entries" section shows a table for "BioProject: PRJDB2359" with fields such as Title, Description, Organism name, Tax id, Archive, Organization name, and Submitted date.

Study	
DRP001175	
Study Title	RNA Sequencing to improve annotation and detect expression changes during anhydrobiosis
Study Type	Other
Center Name	BioProject
Center Project Name	Ramazzottius varieornatus
XREF ID	27649274

Related entries

BioProject: PRJDB2359	
Title	RNA Sequencing to improve annotation and detect expression changes during anhydrobiosis
Description	RNA Sequencing to improve annotation and detect expression changes during anhydrobiosis
Organism name	Ramazzottius varieornatus
Tax id	947166
Archive	DDBJ
Organization name	Center for Genetic Resource Information, Comparative Genomics Laboratory, National Institute of Genetics, Research Organization of Information and Systems
Submitted	2013-08-16T00:00:00Z

データ構造（概略版）

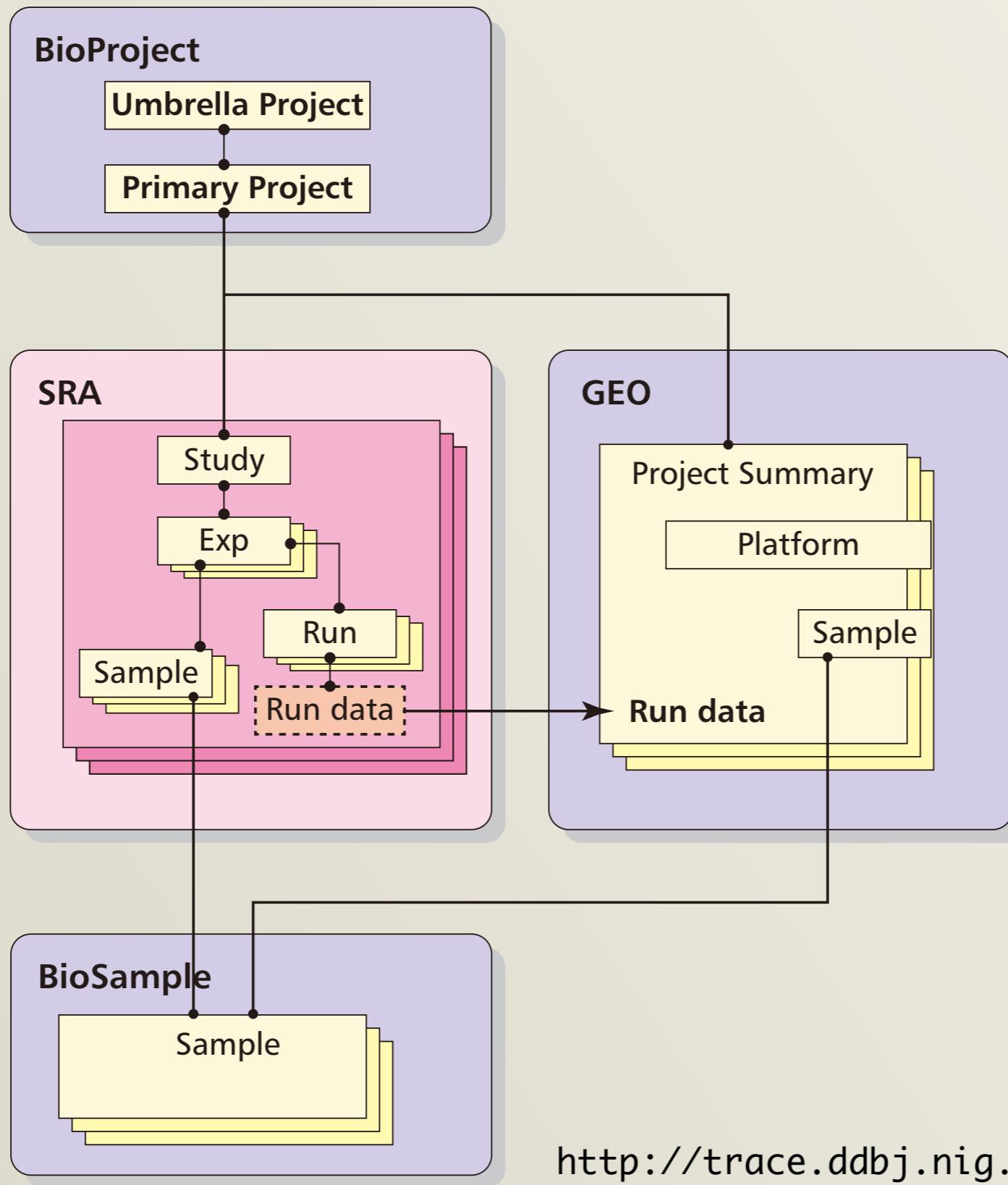


"has" and "uses": relationship between objects

0, 1, N: number of objects

Prefix of accession numbers

データ構造（詳細版）



SRAの検索は意外とツラい

目的が多種多様

ゲノム、発現解析、エピゲ、メタゲ、…

対象生物種も多種多様

ヒト、マウス、メタゲノム、微生物、…

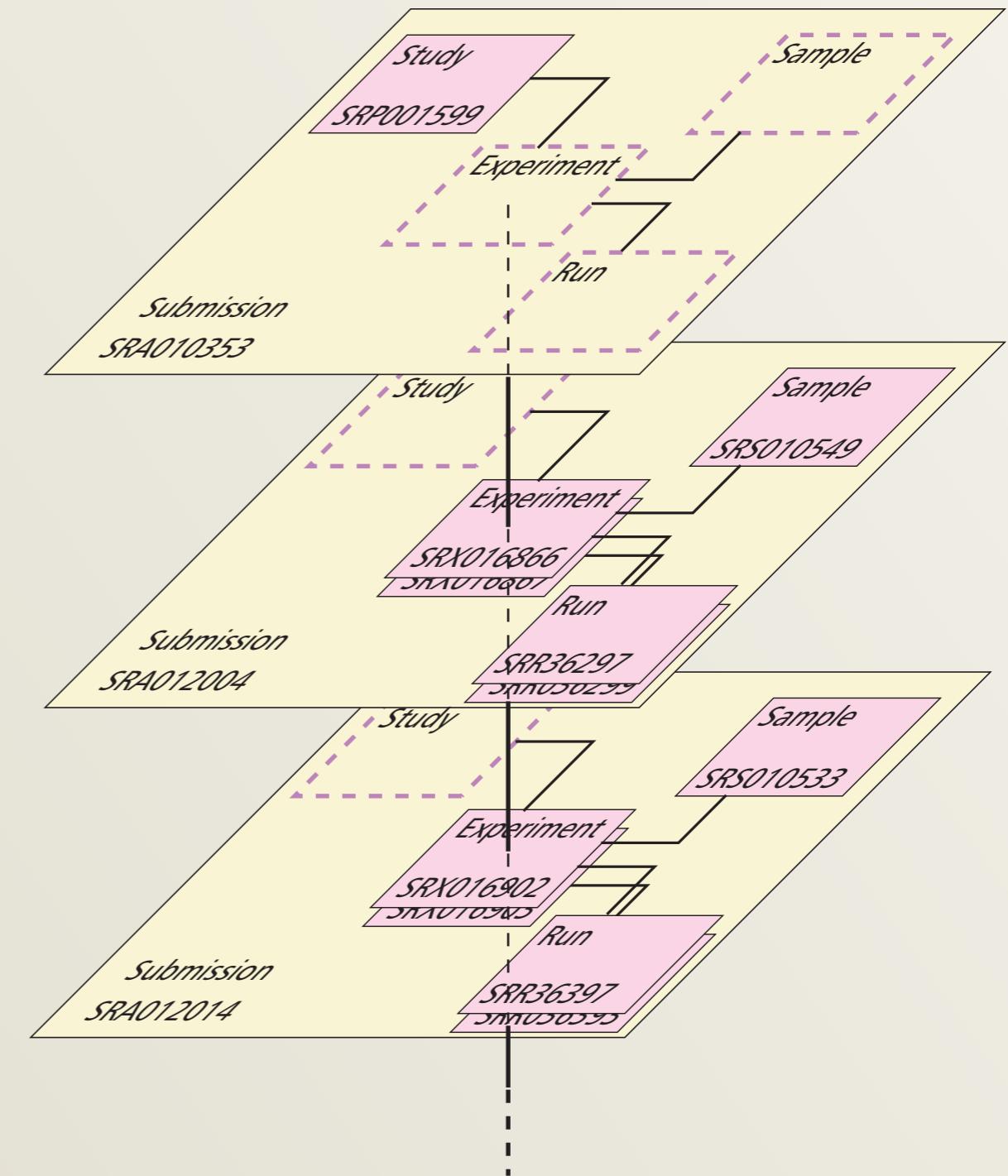
データベースの構造

study：プロジェクト情報

experiment：個々の実験情報

すべてのメタデータが1つの登録に入っているわけではない

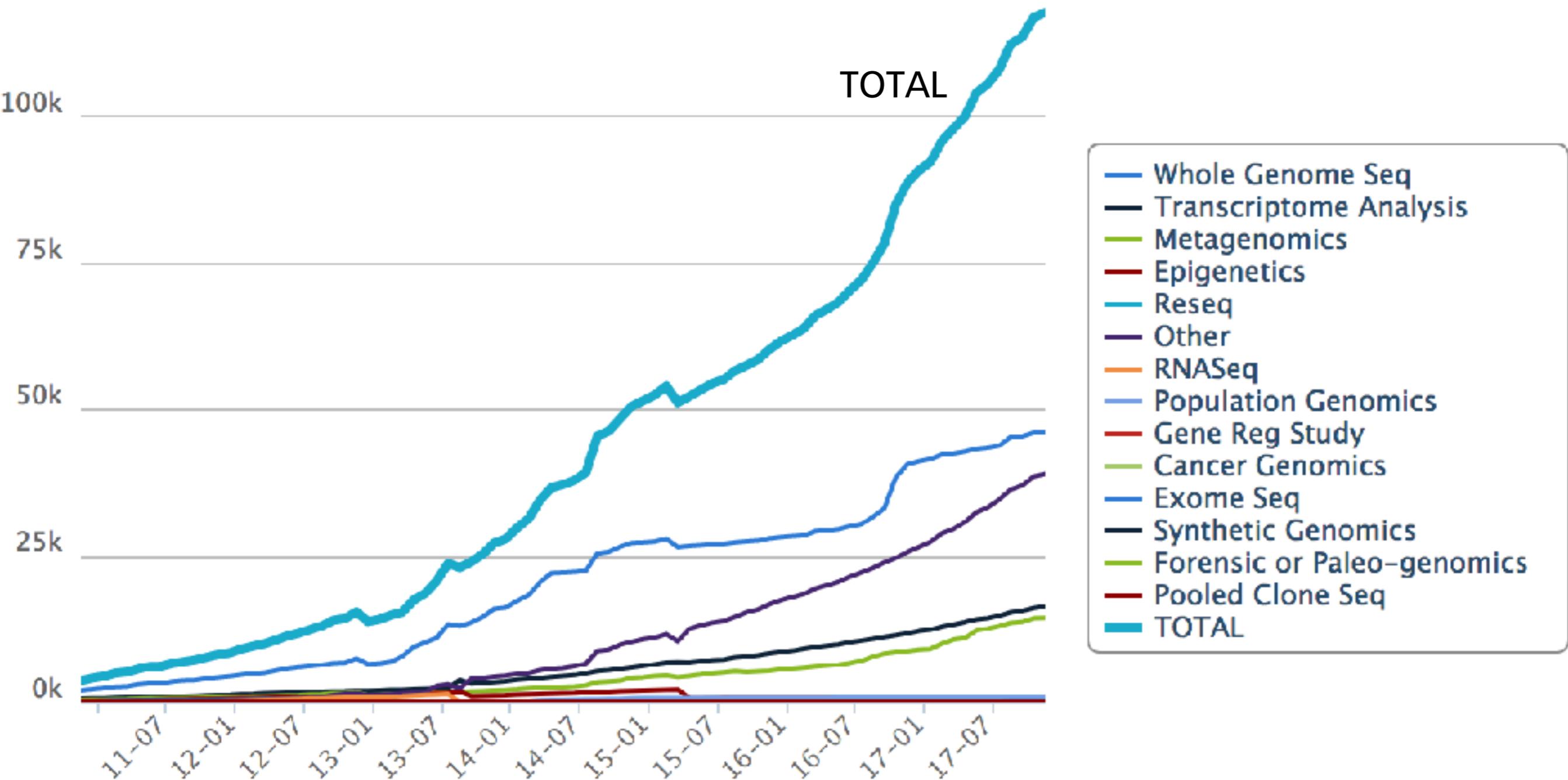
Submission	Study	Experiment	Run	Sample	Analysis	
✓				✓		55452
✓			✓			22025
✓	✓	✓	✓	✓	✓	11228
✓		✓	✓			6066
✓		✓				2915
✓	✓	✓	✓	✓	✓	2608
✓	✓					2430
✓						927
✓	✓	✓	✓			116
✓	✓	✓		✓		107
✓					✓	95
✓	✓	✓	✓	✓	✓	85
✓	✓	✓		✓		58
✓			✓	✓		48
✓	✓		✓	✓		40
⋮						
Total (submissions)						104256



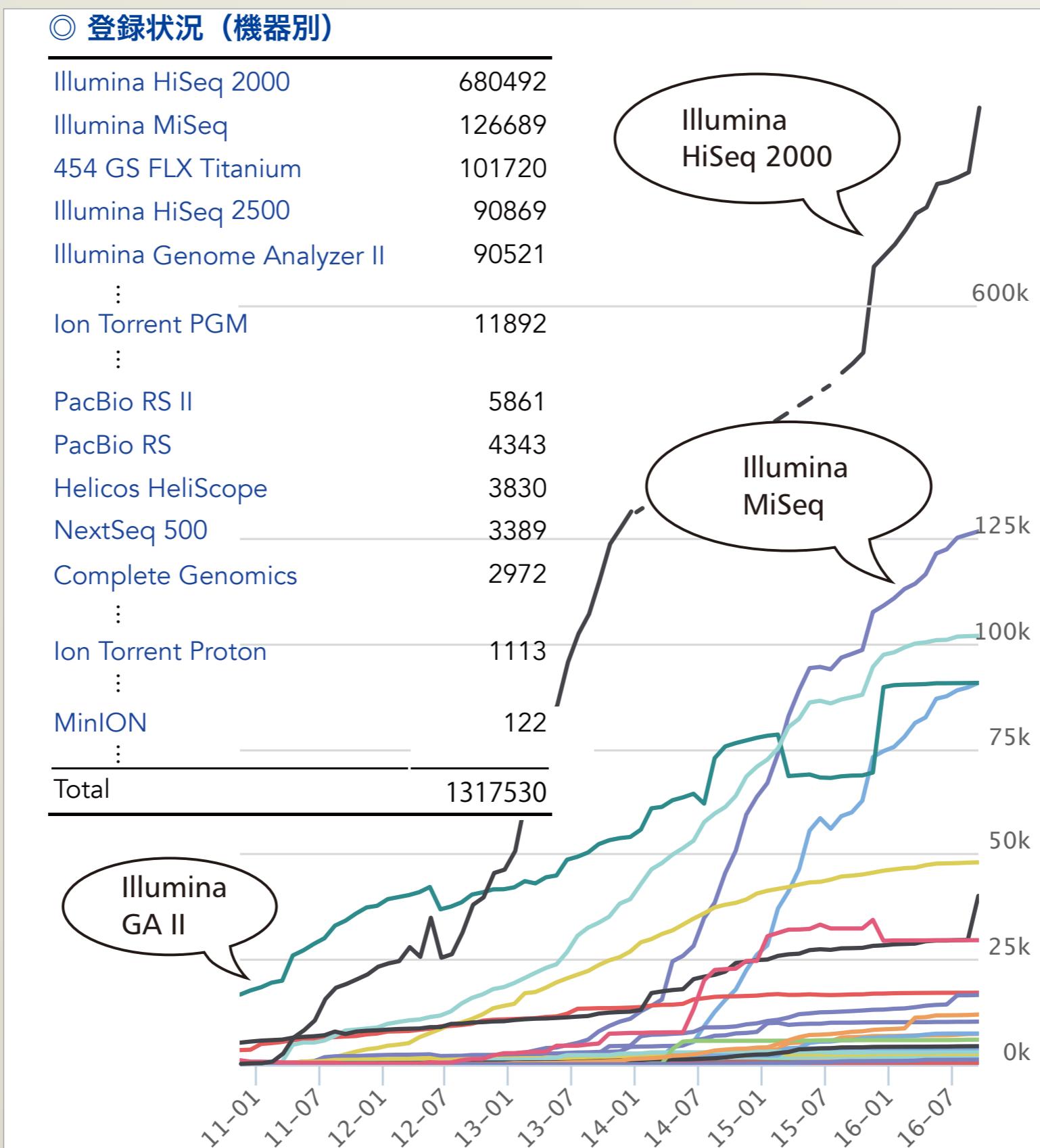
データ数の推移 – 目的別

Zoom 1m 3m 6m YTD 1y All

From Nov 15, 2010 To Nov 15, 2017



データ数の推移 – 機器別



NGSデータを使って論文を出したら

Kikuchi et al. *BMC Genomics* (2017) 18:83

他の研究者のデータでも
明記してリスクトを！

This work was also supported by the NIG Collaborative Research Program (2014-A171 and 2015-A155).

Availability of data and materials

The nucleotide sequences for BmEno1, BmEno2, and BmEnoC were submitted to DDBJ/ENA/GenBank (Accession Nos. LC170036, LC170037, and LC170038, respectively). The RNA-seq reads supporting the conclusions of this article are available in the Sequence Read Archive (SRA) with accession ID DRA005094, <https://www.ncbi.nlm.nih.gov/sra/>.

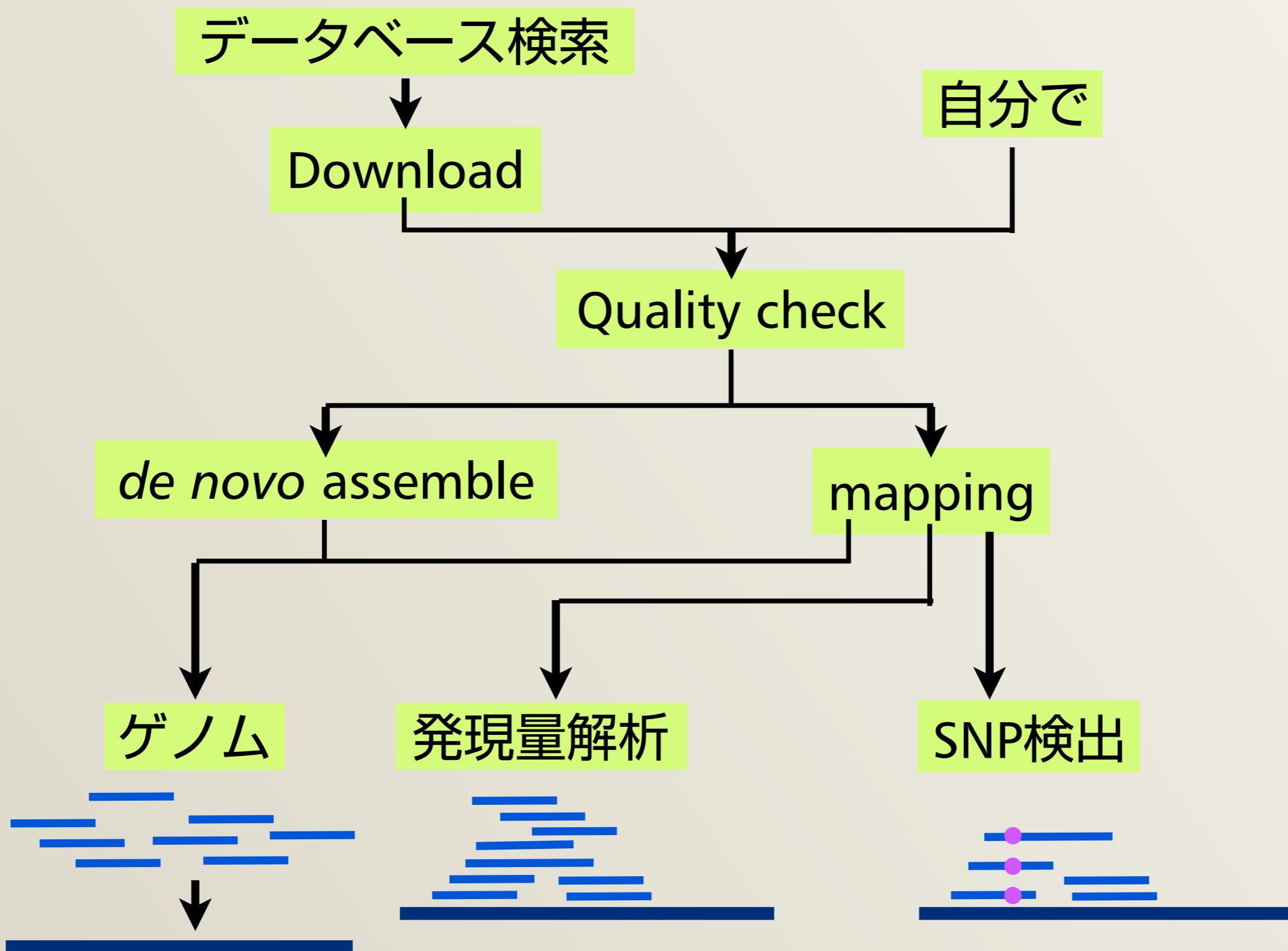
Authors' contributions

Conceived and designed the experiments: AK and HT Performed the experiments: AK, YN, Kal, and AT Contributed reagents/materials/analysis tools: KU, LP, TY, AF and DC Analyzed the data: AK, TN, LP and HT

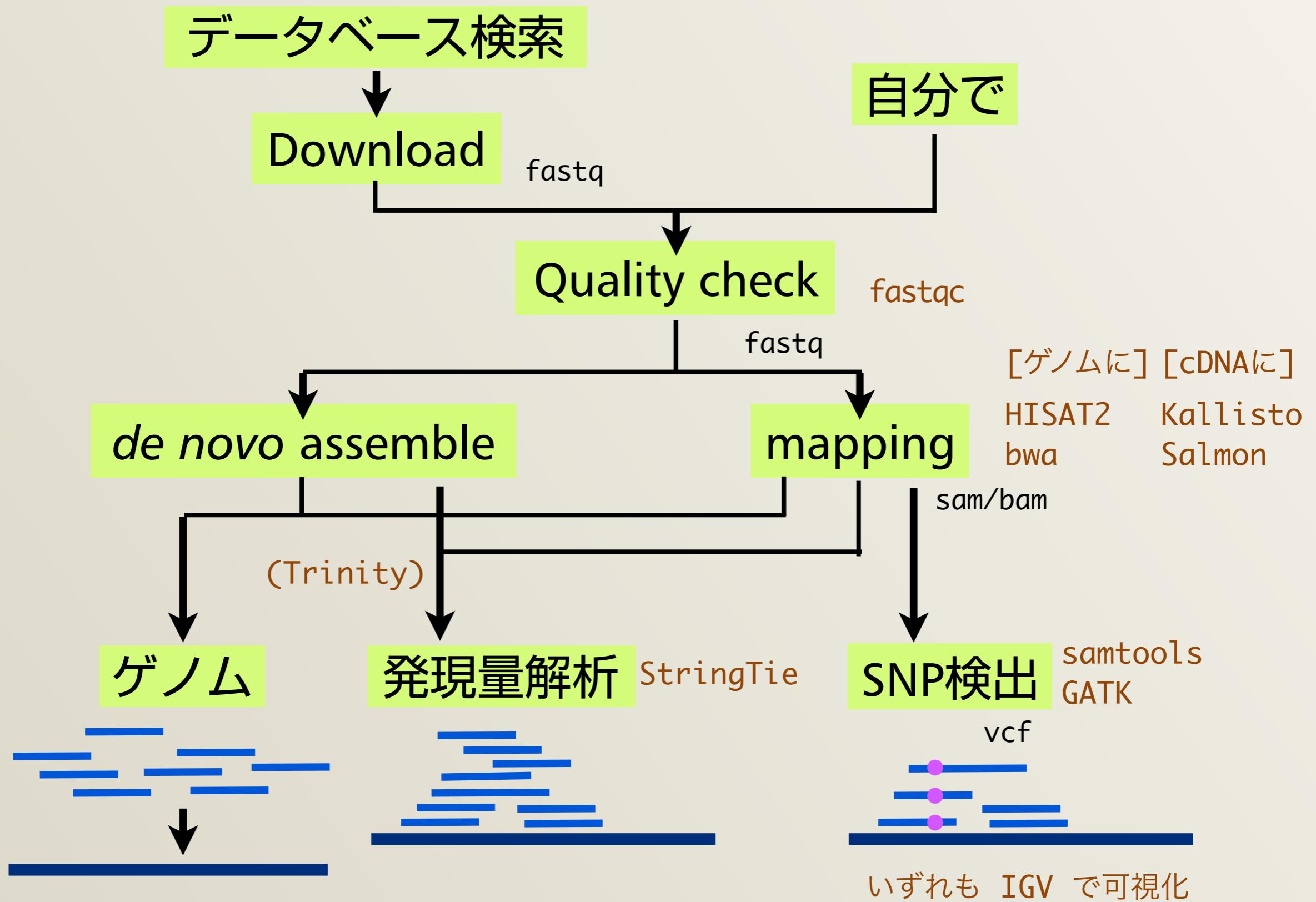
10. Tomita M, Kondo K, Nakamura T, et al. Recombinant BmEno1 enhances the silk fiber production in Bombyx mori. *J Appl Polym Sci*. 2003;88:222–227.
11. Wang J, Li X, Guo X, et al. High-efficiency transgenesis for silk fiber production in silkworm. *Transgenic Res*. 2010;19:103–112.
12. Xia Q, Zhang L, Zhou J, et al. BmEno1 overexpression in mulberry leaves enhances silk fiber production in silkworm. *Plant Mol Biol*. 2012;83:231–242.
13. Brewster A, Enola S, et al. *Enolase 1* expression in *Bombyx mori* during development. *Arch Insect Biochem Physiol*. 2000;43:13–22.

データ解析 概略

NGSデータ解析の流れ



NGSデータ解析の流れ（詳細版）



FASTQフォーマット

```
@DRR001107.1 GEZQ5F001EEA7F length=77
GCAACATTCAACACATATGTGTTGAATGTTGCACGACGGNGTG...
+DRR001107.1 GEZQ5F001EEA7F length=77
C@BBECCECDBBBAAAAAA<441111<?@>?=?????44!000...
```

4行1組

1行目： @ + タイトル

2行目： 塩基配列

×

3行目： + (+ タイトル)

数千万
数十億

4行目： シーケンスクオリティ

[参考] FASTAフォーマット

塩基配列を表現するフォーマット

```
>AB084425.1 eel SLC26A6
GACCCAAACTGATAGGTGATGTTCACGTAGTGGC
CATCGCCTGATAGACGGTTTCGCCCTTGACGTT
GGAGTCCACGTTCTTAATAGTGACTCTGAGTAAA ...
```

1行目： > + タイトル

2行目以降： 塩基配列

コマンド例

Quality Check

```
$ fastqc --nogroup -o DRR1234567.fastq
```

コマンド名 おまじない

対象ファイル名

trimming

```
$ trim_galore --paired --illumina --fastqc -o trimmed/
```

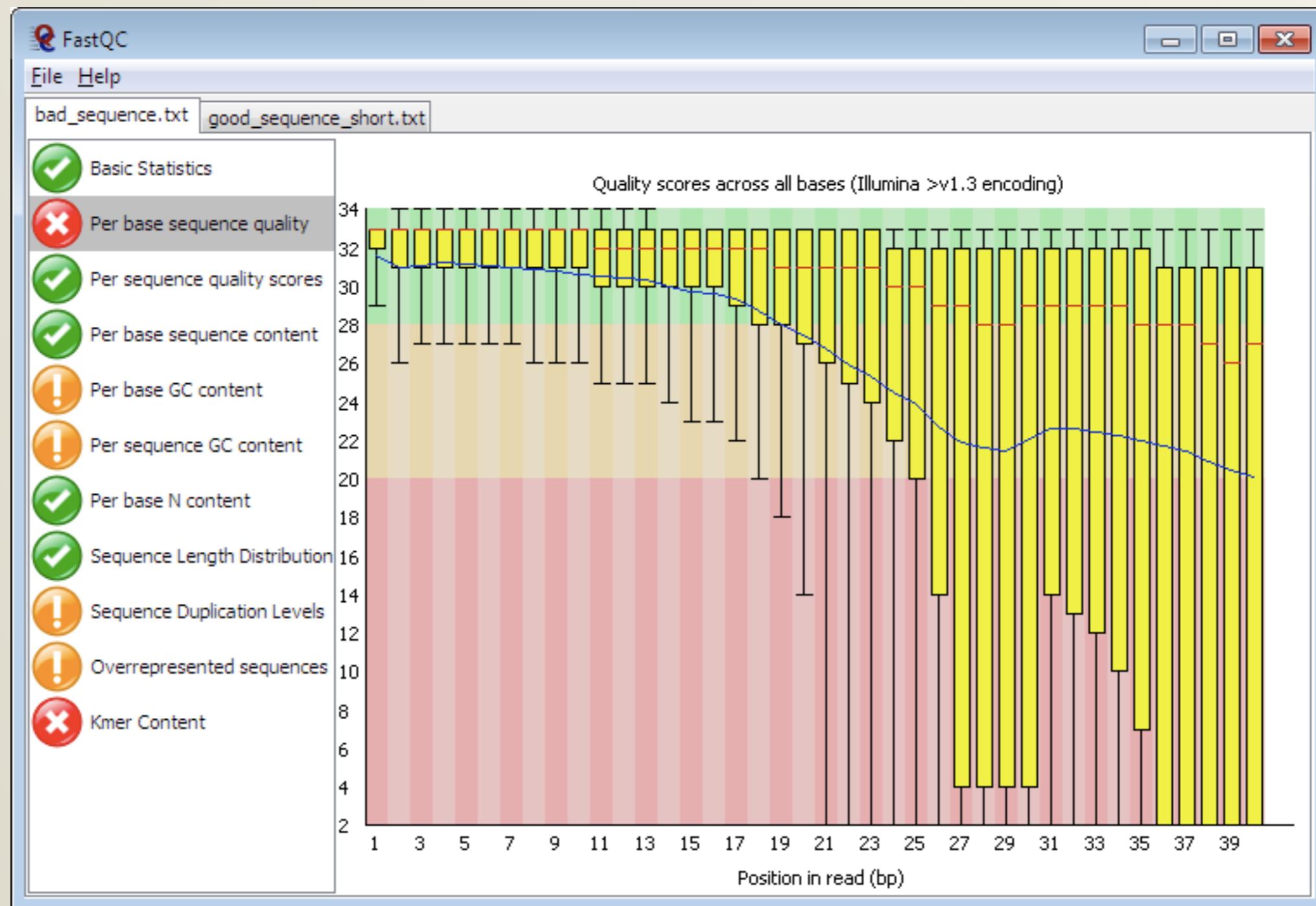
コマンド名 pair-end Illuminaデータ fastqcもかける 出力先

DRR1234567.R1.fastq DRR1234567.R2.fastq

対象ファイル名・その1

対象ファイル名・その2

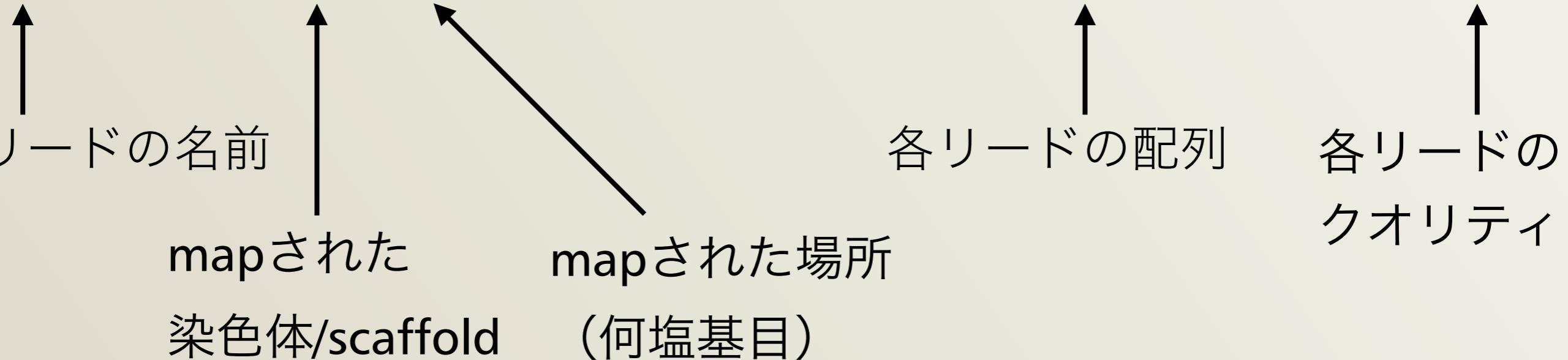
fastqc結果例



sam/bamフォーマット

(Sequence Alignment/Map Format)

SRR445820.39542705	0	chr17	1	0	4M1T31M	*	0	0	AAAGCTTCTCACCTGTTCCCTGCATAGATAATTGCA	?5>7(+2; '1..'+<
SRR445820.29211975	16	chr17	88	42	36M	*	0	0	CCACGACCAACTCCCTGGGCCTGGCACCAGGGAGCT	#### #####BDB8DACC
SRR445820.7156374	16	chr17	138	42	36M	*	0	0	CCAGCGAATACCTGCATCCCTAGAAAGTGAAGCCACC	BBB-:;BBEABFBFB
SRR445820.22614977	0	chr17	156	30	36M	*	0	0	CCTAGAAAGTGAAGCCACCGCCCCAAAGACACGCCCAT	GGGD>DBB3D=?=<
SRR445820.19222309	0	chr17	185	42	36M	*	0	0	CGCCCATG1CCAGCTTAACCTGCATCCCTAGAAAGTG	IIIIIIIIIIIIHH
SRR445820.32725447	16	chr17	213	31	36M	*	0	0	TAGAAAGTGAAGGCACCGCCCCAAAGACACGCCCATGT	CGCGGGGDGGGBGA
SRR445820.43349427	0	chr17	221	31	36M	*	0	0	AAAGCGACCGCCCCAAAGACACGCCCATGTCCAGCTTA	TTTTTTTTTTTTTTTT



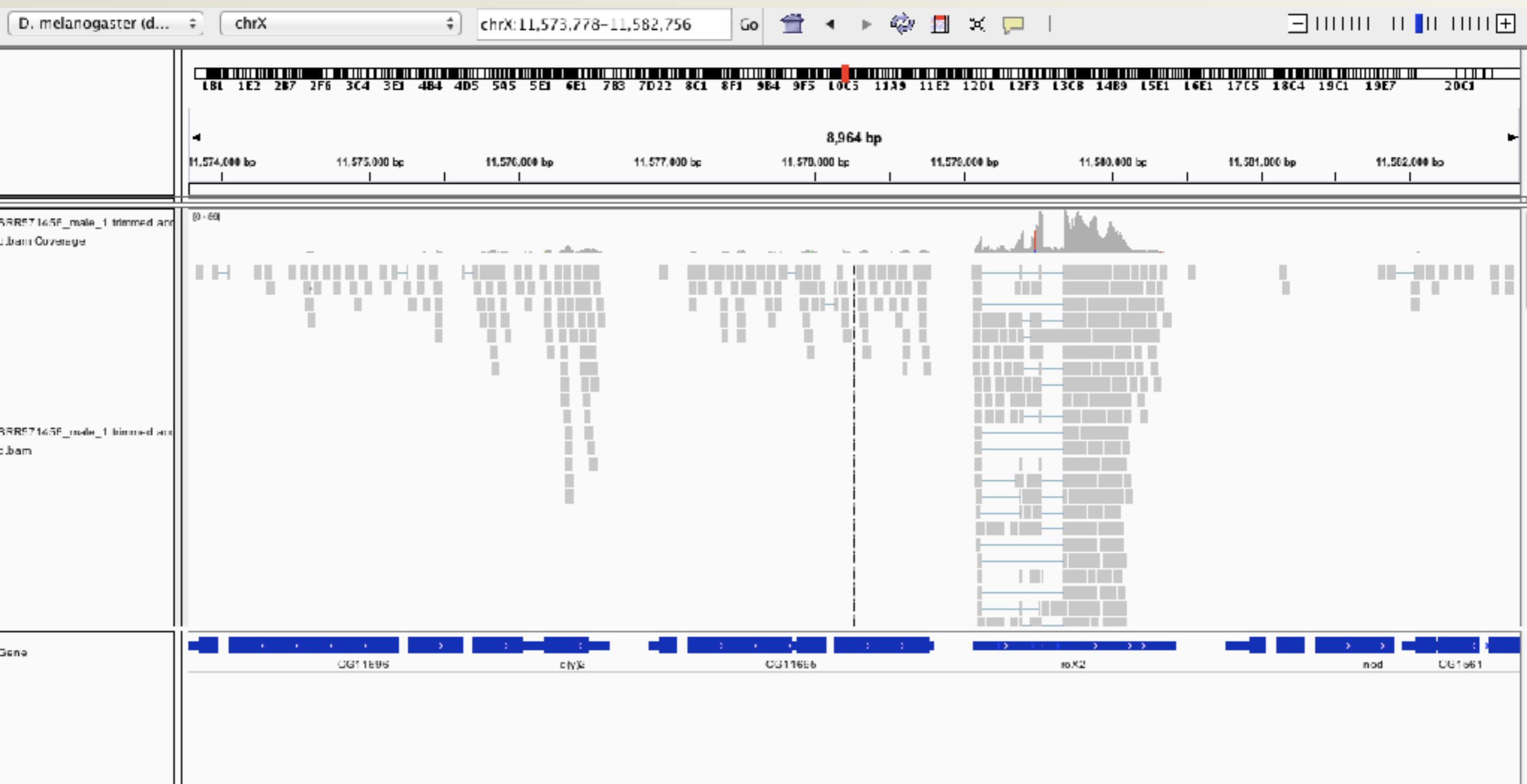
※ その他、マッピングの状況など

※ bam は sam をバイナリにしたもの

(人間が読めるデータからコンピューター用に変換)

(sam だとデータサイズが非常に大きくなるのでbamにして圧縮)

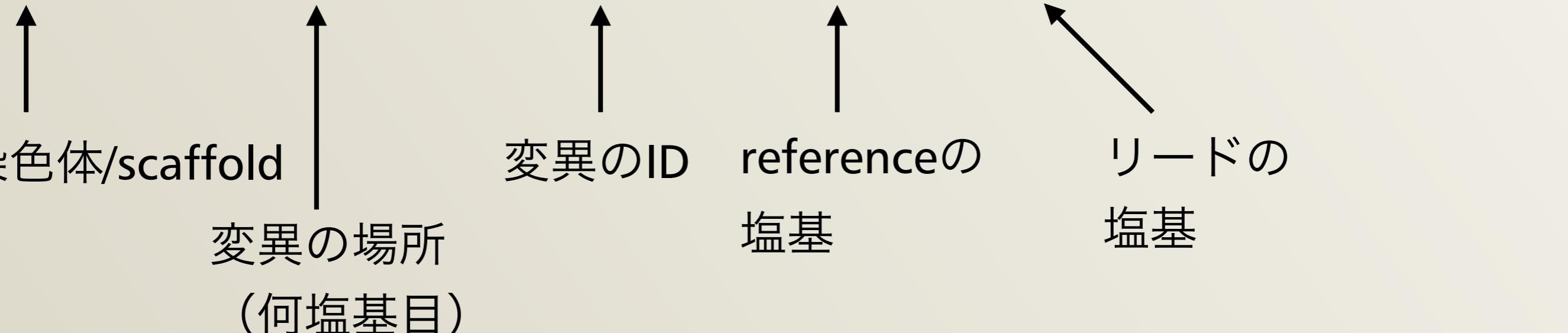
IGVによる可視化



vcfデータ

(variant call format)

3	178738432	rs6790867	T	C	1061.77	PASS	AC=2
3	178739594	rs1542	C	G	1466.77	PASS	AC=2;AF=1.00
3	178740415	rs146675821	G	GA	168.74	PASS	AC=2
3	178740422	rs7641761	T	A	316.77	PASS	AC=2
3	178740425	rs61798175	T	A	313.77	PASS	AC=2



※ この場合、変異のIDとはdbSNPのIDをさしています

遺伝子機能アノテーション

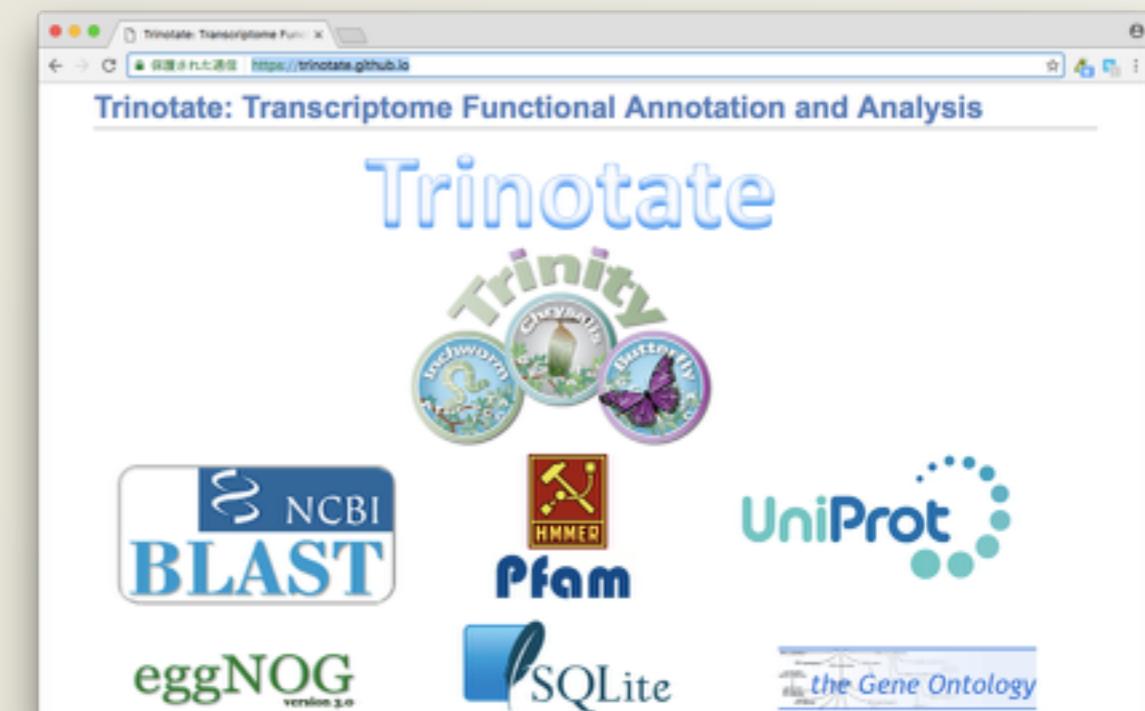
- BLASTで類似性のある遺伝子を検索

```
$ blastx -query Trinity.fasta -db uniprot_sprot.pep  
-num_threads 8 -max_target_seqs 1 -outfmt 6 > blastx.outfmt6
```

TRINITY_DN15083_c2_g1_i1	tr 022669 022669_PANGI	99.160	119	1	0	1	357	85	203	2.33e-79	242
TRINITY_DN15083_c2_g1_i2	tr A0A089WZX0 A0A089WZX0_KALFE	92.884	267	19	0	74	874	1	267	2.20e-175	498
TRINITY_DN15083_c2_g1_i3	tr Q1KLZ3 Q1KLZ3_9ROSI	86.364	66	9	0	95	292	1	66	6.90e-31	117
TRINITY_DN15083_c2_g1_i4	tr I3SIW2 I3SIW2_LOTJA	97.458	118	3	0	1	354	84	201	8.34e-79	238
TRINITY_DN15083_c2_g1_i5	tr Q1KLZ3 Q1KLZ3_9ROSI	95.270	148	7	0	3	446	26	173	2.31e-99	294

- hmmerでドメイン検索

```
$ hmmsearch --cpu 8 -domtblout  
TrinotatePFAM.out Pfam-A.hmm  
transdecoder.pep > pfam.log
```



<https://trinotate.github.io/>

ChIP-Seq

ChIP-Atlas

[ChIP-Atlas](#) [Peak Browser](#) [Target Genes](#) [Colocalization](#) [in silico ChIP](#) [Documentation](#)

[Find an experiment](#) ▾

ChIP-Atlas

ChIP-Atlas is an integrative and comprehensive database for visualizing and making use of public ChIP-seq data. ChIP-Atlas covers almost all public ChIP-seq data submitted to the SRA (Sequence Read Archives) in NCBI, DDBJ, or ENA, and is based on over 46,000 experiments.

[Watch movie introduction](#)

The four main features of ChIP-Atlas are:

[Peak Browser](#)

graphically visualizes protein binding on given genomic loci with genome browser (IGV).

[Watch Movie](#)

[Target Genes](#)

predicts target genes bound by given transcription factors.

[Watch Movie](#)

[Colocalization](#)

predicts partner proteins colocalizing with given transcription factors.

[Watch Movie](#)

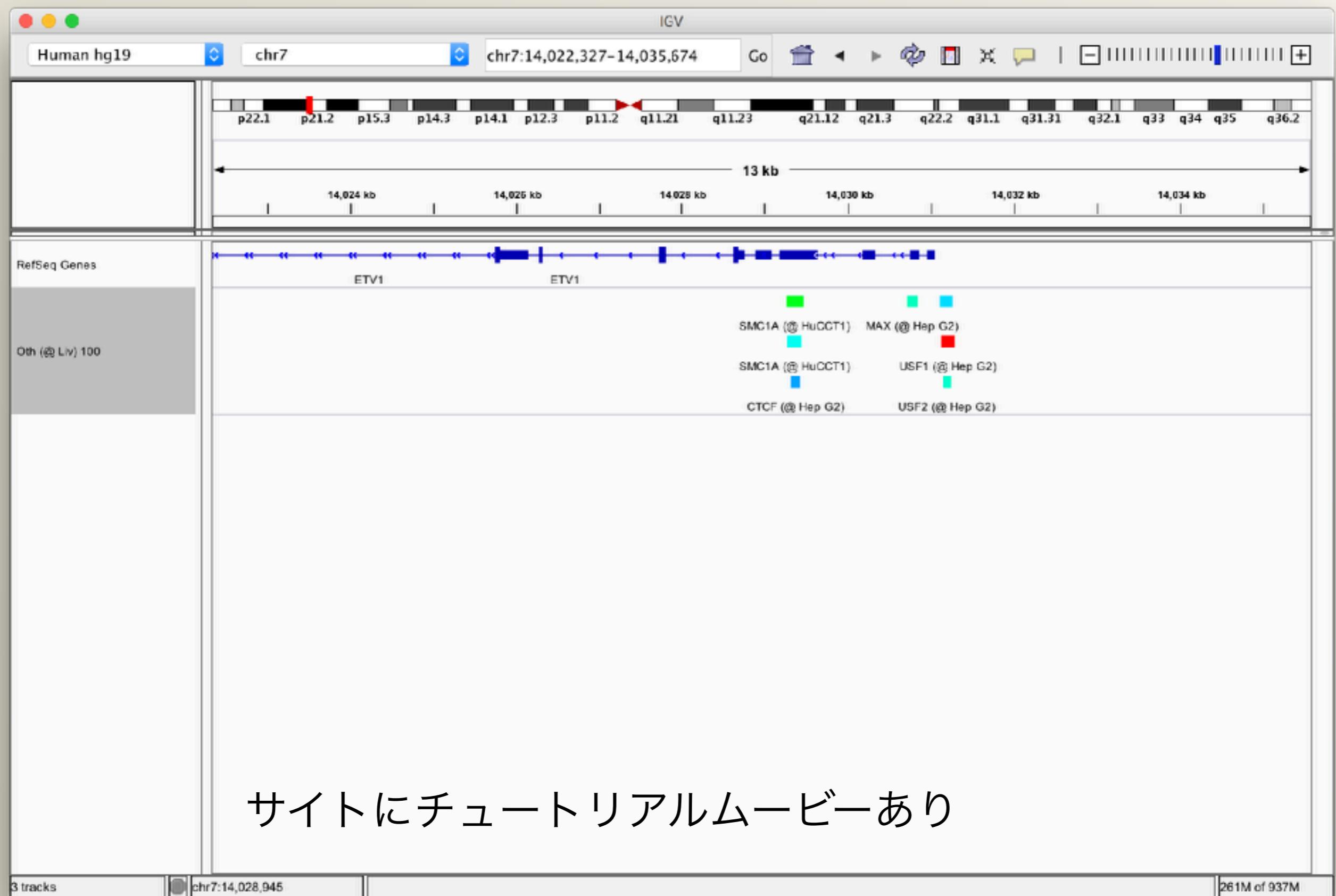
[in silico ChIP](#)

predicts proteins bound to given genomic loci and genes.

[Watch Movie](#)

すでにゲノムにマップして可視化できるようにしたサイトが
<http://chip-atlas.org/>

ChIP-Atlas の情報の可視化



メタゲノム

MicrobeDB.jp

[Sign In](#)



Gene: psbA
Taxonomy: *Streptococcus glycerinaceus*
Mapping: *Escherichia coli* O157:H7 str. Sakai
Environment: hot spring
SRS: rumen
Strain: *Bifidobacterium*
Disease: Cholera
MiGap: GAF

SRAへの登録

DRAへの登録

The screenshot shows a web browser window with the title bar "DRA Handbook". The address bar contains the URL "trace.ddbj.nig.ac.jp/dra/submission.html#DRA_へのデータ登録". The main content area displays the "DRAへのデータ登録" page. On the left, there are three callout boxes: one about recording human subjects data, one about sequencing data, and one about patent-related data. Below these is a section titled "DRA登録の流れ" (Flow of DRA registration) with a sub-section "1. 登録アカウントを作成" (Create registration account). On the right, there is a sidebar titled "In this page" with links to various DRA-related topics, a "PDF Download" section with a link to "PDFをダウンロード", a "Submit" section with links to "Login & Submit" and "Contact", and an "Archives" section with links to "News by year", "FAQs", and "Handbooks".

DRAへのデータ登録

ヒトを対象とした研究データの登録について

ヒトを対象とした全ての研究において DDBJ に送付するデータの由来である個人(被験者)の情報・プライバシーは、適用されるべき法律、規定、登録者が所属している機関の方針に従い、登録者の責任において保護されている必要があります。

原則として、被験者を直接特定し得る参照情報は、登録データから取り除いてください。

ヒトを対象とした研究データを登録する場合は「[ヒトを対象とした研究データの登録について](#)」をご覧ください。

次世代シーケンサからのデータを DRA に登録するためにはメタデータとシーケンスデータが必要です。

解析後の配列データは DDBJ へ登録します。DDBJ Mass Submission System (MSS) が次世代シーケンサから生み出されるゲノムや大量データの登録受付先になります。

特許に関連するデータの登録

登録するデータが特許に関連する場合は、「[特許に関連する塩基配列の登録に関する注意、データの優先権](#)」の内容を必ずご確認ください。

DRA登録の流れ

1. 登録アカウントを作成

- [D-way 登録アカウントを作成](#)
- [公開鍵と center name をアカウントに登録し、DRA 登録を可能に](#)

In this page

DRAについて

メタデータ

データファイル

DRAへのデータ登録

DRA登録の流れ

DRAへのデータ登録方法

登録の更新

補足: MD5 値

PDF Download

PDFをダウンロード

Submit

Login & Submit

Contact

Archives

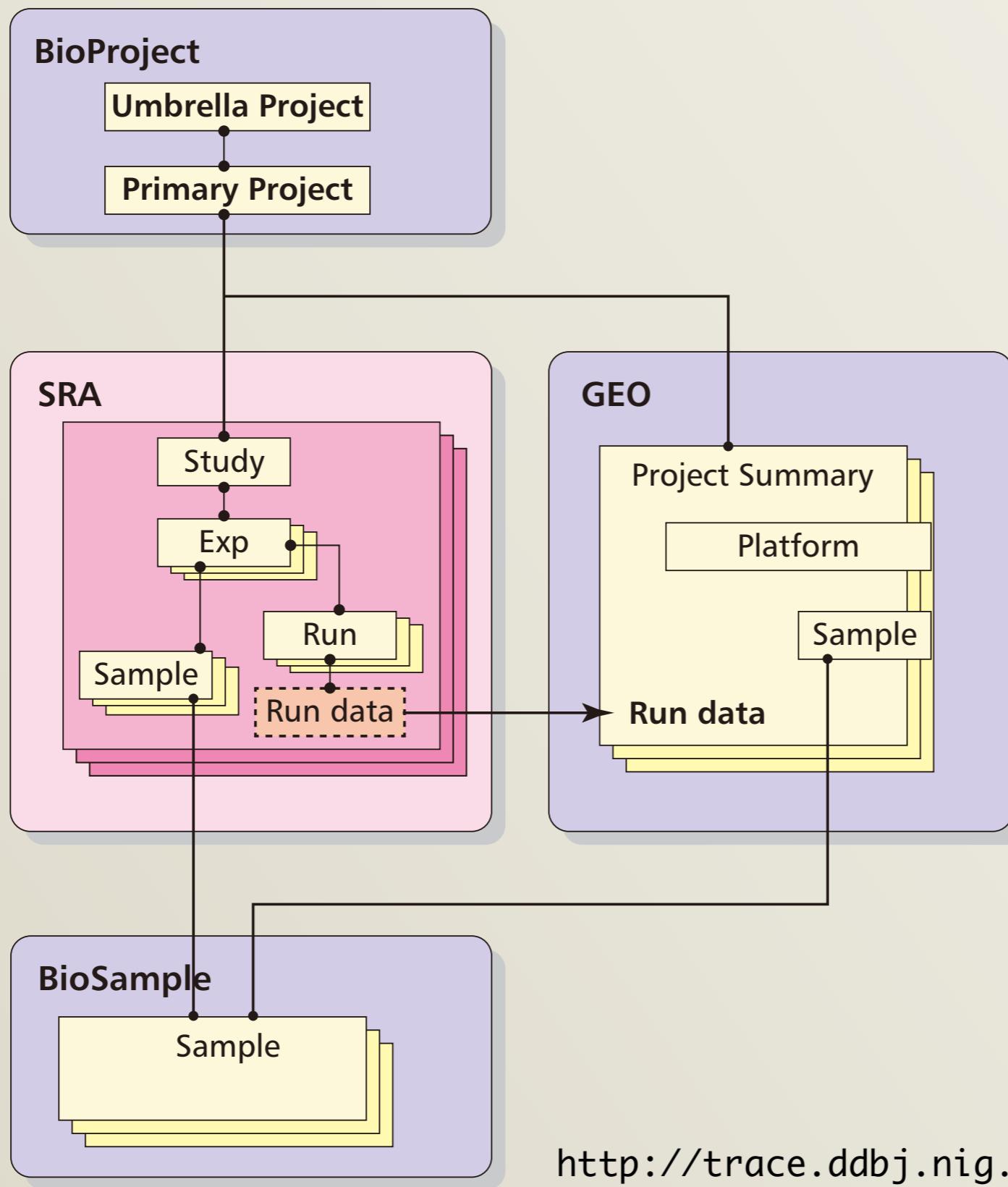
News by year: [Latest](#)

FAQs

Handbooks

<http://trace.ddbj.nig.ac.jp/dra/submission.html>

データ構造（再掲）



DRAへの登録フロー

DRA 登録の流れ

目次:

1. 登録アカウントを作成

- [D-way 登録アカウントを作成](#)
- [公開鍵と center name をアカウントに登録し、DRA 登録を可能に](#)

2. DRA 登録を作成しデータファイルをアップロード

- 新規 DRA 登録を作成 ([アカウントに DRA 登録権限を付与しておきます](#))
- BioProject, BioSample, Experiment と Run を投稿する前にデータファイルを scp でアップロード

3. プロジェクトとサンプル情報を登録

[BioProject \(Study\)](#)

- 研究プロジェクトの内容
- 「なぜ」そのサンプルをシークエンスしたのか

[BioSample \(Sample\)](#)

- 生物学的、物理的にユニークなサンプル
- 「何を」シークエンスしたのか

メタデータをタブ区切りテキストファイルで登録できます

シークエンスデータ

(FASTQ or BAM) に加え
メタデータ（実験情報）を
書いてDDBJに登録

4. Experiment と Run を登録

[DRA Experiment](#)

- 特定のサンプルから構築したライブラリーについての説明
- 「どのように」シークエンスをしたのか
- 複数の Experiment は一つの Sample を参照できるが、逆はできない

[DRA Run](#)

- Experiment と Run を投稿した後、データファイルの検証処理を開始
- Run にリンクしている全てのデータファイルは 1 つの SRA ファイルにマージされます

5. シークエンスデータファイルの検証処理

- シークエンスデータファイルをアーカイブ用 SRA ファイルに変換する処理を開始
- 検証処理を通った登録が査定されアクセスion番号が発行される

DRAへの登録 (Sample)

The screenshot shows a Microsoft Excel spreadsheet titled "bioamp". The ribbon menu is visible at the top, showing tabs for Home, Insert, Print Layout, Number, Data, Review, and View. The Home tab is selected. The formula bar shows the cell reference "A12". The main content of the spreadsheet consists of a table with columns labeled A, D, E, F, G, L, and collection. The first row contains column headers: *sample_name, *organism, *taxonomy_id, bioproject_id, strain, biomaterial_provider, and collection. Rows 2, 3, and 4 contain data for Bombyx mori samples with taxonomy ID 7091, project ID PRJDB####, and strain XXX, all associated with Kyushu Univ/NBRP.

	A	D	E	F	G	L	collection
1	*sample_name	*organism	*taxonomy_id	bioproject_id	strain	biomaterial_provider	
2	Bmori-strainXXXwt-tissue-1	Bombyx mori	7091	PRJDB####	XXX	Kyushu Univ/NBRP	
3	Bmori-strainXXXwt-tissue-2	Bombyx mori	7091	PRJDB####	XXX	Kyushu Univ/NBRP	
4	Bmori-strainXXXwt-tissue-3	Bombyx mori	7091	PRJDB####	XXX	Kyushu Univ/NBRP	

DRAへの登録 (Experiment)

Submission: chalkess-0001

保護された通信 https://trace.ddbj.nig.ac.jp/D-way/contents/dra/metadefine?serial=1&type=experiment

Submit/Update DRA metadata

SUBMISSION BIOPROJECT BIOSAMPLE EXPERIMENT RUN ANALYSIS (optional)

- For more information, please see the [Experiment metadata](#).
詳細は [Experiment メタデータ](#) を参照してください。
- Click the "Save" button and fix aliases to edit metadata in TSV file.
"Save" をクリックして alias を確定してから TSV で内容を編集してください。

[Save](#) [Next \(RUN\)](#)

Edit Experiment by using tab-delimited text (TSV) file

[Download TSV file](#)

[Upload TSV file](#) ファイルを選択 指定されていません

Experiment

[Add another Experiment\(s\)](#) 1 [Copy Experiment #1](#)

#	Alias	=BioSample Used	Library Name	=Library Source	=Library Selection	=Library Strategy
1	chalkess-0001_Experiment_0001	SAMD00058752		TRANSCRIPTOMIC	cDNA	RNA-Seq
2	chalkess-0001_Experiment_0002	SAMD00058753		TRANSCRIPTOMIC	cDNA	RNA-Seq
3	chalkess-0001_Experiment_0003	SAMD00058754		TRANSCRIPTOMIC	cDNA	RNA-Seq

[Save](#) [Next \(RUN\)](#)

参考リソース

参考図書・その1～実験もやる人向け

本・医学・薬学・看護学・歯科学・基礎医学



次世代シーケンス解析スタンダード～NGSのポテンシャルを活かしきる

WET&DRY 単行本 - 2014/8/23

二階堂 愛 (編集)

★★★★☆ 3件のカスタマーレビュー

▶ その他 () の形式およびエディションを表示する

単行本

¥ 5,940

¥ 5,682 より 4 中古品の出品

¥ 5,940 より 1 新品

住所からお届け予定日を確認 詳細

9/1 木曜日 にお届けするには、今から**14 時間 57 分**以内に「お急ぎ便」または「当日お急ぎ便」を選択して注文を確定してください（有料オプション。Amazonプライム会員は無料）

amazon student

Amazon Student会員なら、この商品は+10%Amazonポイント還元(Amazonマーケットプレイスでのご注文は対象外)。無料体験でもれなくポイント1,000円分プレゼントキャンペーン実施中。

実験デザイン・サンプルの用意から解析まで

参考図書・その2～解析を詳しく



NGSデータ解析を丁寧に解説。Kindle版あり

参考図書・その3～初心者向け入門書



Dr. Bonoの生命科学データ解析 単行本 – 2017/9/29

坊農秀雅 (著)



1 件のカスタマーレビュー

▶ その他 () の形式およびエディションを表示する

単行本

¥ 3,240

¥ 4,138 より 7 中古品の出品

¥ 3,240 より 4 新品

クリスマスイブまでにお届け。注文確定時に配送方法をご確認ください。

12/19 火曜日 にお届けするには、今から**6 時間 6 分以内にお急ぎ便**を選択して注文を確定してください (有料オプション。Amazon プライム会員は無料)

詳細な解析をひとつおり知りたい

The screenshot shows a web browser displaying the NBDC (National Bioscience Database Center) website. The URL in the address bar is <https://biosciencedbc.jp/human/human-resources/workshop/h28-2>. The page title is "H28年度 NGSハンズオン講習会カリキュラム". The curriculum table lists two sessions:

実施日	実施時間	大項目	タイトル	内容（予定）	担当講師 (敬称略)	講義資料・ 動画(総合TV)
7月19日 (火)	10:30- 16:15	はじめに(講習会終了後) 加藤必成	Bio-Linux8とBioconda インストール&使用説明 PC環境の構築	・Bio-Linux8(第2回および3回)で利用するova ・主催・共催機関 ・共有フォルダ設定完了確認 ・基本的なLinuxコマンドの習得状況確認 ・R本体およびパッケージのインストール確認 ・講習会概要の学習予習内容の再確認 ・講習会期間中に貸し込まれるノートPCを用いた各種動作確認	講義資料 (PDF: 4MB)	カリキュラム (PDF: 72KB)
7月20日 (水)	10:30- 16:15	第1回 塙山勝行 (農学生命情報科学 特論)	ゲノム解析、塙山記 統計解析	・NGS解析手順、ウェブツール(DDG Bio Pipeline)との連携 ・k-mer解析(k語の過度重複に基づく各種解析)の基礎と応用 ・塙山ごとの出現頻度解析($k=1$)、2塙山塙基の出現頻度解析($k=2$) ・塙基配列解析を行ったための基本スキルの復習	塙山 勝行 (敬称略)	講義資料 (PDF: 7.3MB) 動画 (ZIP: 2.2MB)

NGSハンズオン講習会

JST-NBDC+東大アグリバイオ

<https://biosciencedbc.jp/human/human-resources/workshop/h28-2>

こここのところ毎年やっています
統合TVで録画・公開済

※「NGS 講習会」でググれ

解析環境・ウェブベース

The screenshot shows the "Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE" interface. The top navigation bar includes links for "Select Query Files", "Select Tools" (which is highlighted in orange), "Sel QuerySet", "Sel GenomeSet", "Sel Map Options", and "Confirmation". Below the navigation is a "Running Status" section showing "login ID [nakazeto]" with "Logout" and "Change password" options. The main content area is titled "Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE". It features two sections: "Reference Genome Mapping" and "de novo Assembly".

Reference Genome Mapping

Tool	Help	Version	Input data			Evaluation		Analysis			Output format		
			Base space	Color space	Paired-end	Depth	Coverage	Error rate	SNP	InDel	.gff	.bed	SAM
bowtie 1		0.6.1	✓	✓	✓	✓	✓					✓	
Bowtie 2		0.12.7	✓	✓	✓	✓	✓	✓	✓			✓	
TopHat 1		1.0.11	✓	✓	✓	✓	✓	✓				✓	
Bowtie2		2.2.8	✓	✓	✓	✓	✓	✓	✓	✓		✓	
TopHat2		2.1.0	✓	✓	✓	✓	✓	✓				✓	

For reads longer than about 50 bp, Bowtie2 is generally better, more sensitive, and uses less memory than Bowtie1.

de novo Assembly

Tool	Help	Version	Base space	Color space	Paired-end	M5S(WGS)	Comment
SOAPdenovo-rt		2.04-240	✓		✓		
ABYSS 2		1.3.2	✓		✓		Maximum K-mer value is 64.

DDBJ Read Annotation Pipeline

<http://p.ddbj.nig.ac.jp/>

※ 要利用申請