

AJACS下総

# 次世代シーケンサー（NGS）を用いた解析と 関連データベース・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)



2017/12/19  
@千葉大学

## MotDB

### AJACS67

#### 統合データベース講習会: AJACST下総

統合データベース講習会は、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、アーカイブの使い方を加えて、遺伝子発現データベース、ゲノムデータベース、次世代シーケンスデータベースなどを用いてお話をさせていただきます。参考に全員がパソコンでエクセルを使っていかがの講習会です。

<http://motdb.dbcls.jp/?AJACS67>

#### 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

2017年12月19日（火）9:00-17:30（開場および受付は、開始時間の30分前より）

#### 会場

千葉大学のなな同窓会館多目的ホール  
(千葉県千葉市中央区亥鼻1丁目8-1 千葉大学亥鼻（医学部）キャンパス)  
【[アクセス/キャンパスマップ](#)】

#### 定員

約50名

#### 費用



このセクションはtweet OK

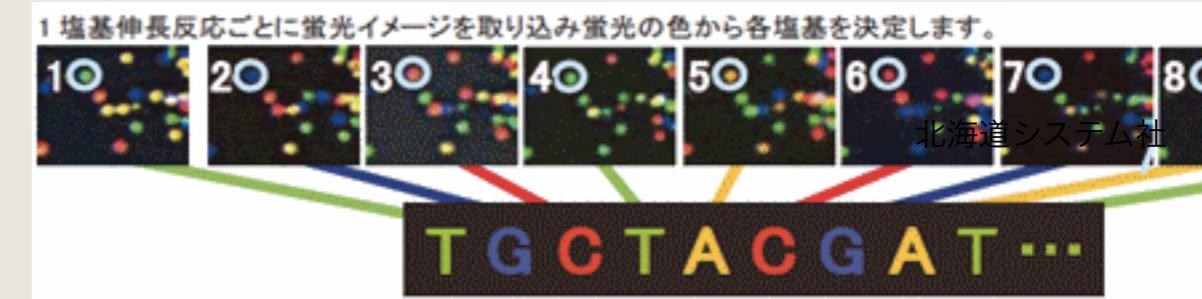
ハッシュタグは #AJACS

# 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
- 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) × 48/4 lanes  
= 9kbase

500 (base/lane) × 96 lane  
= 48kbase

2 × 300 (base/seq) × 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

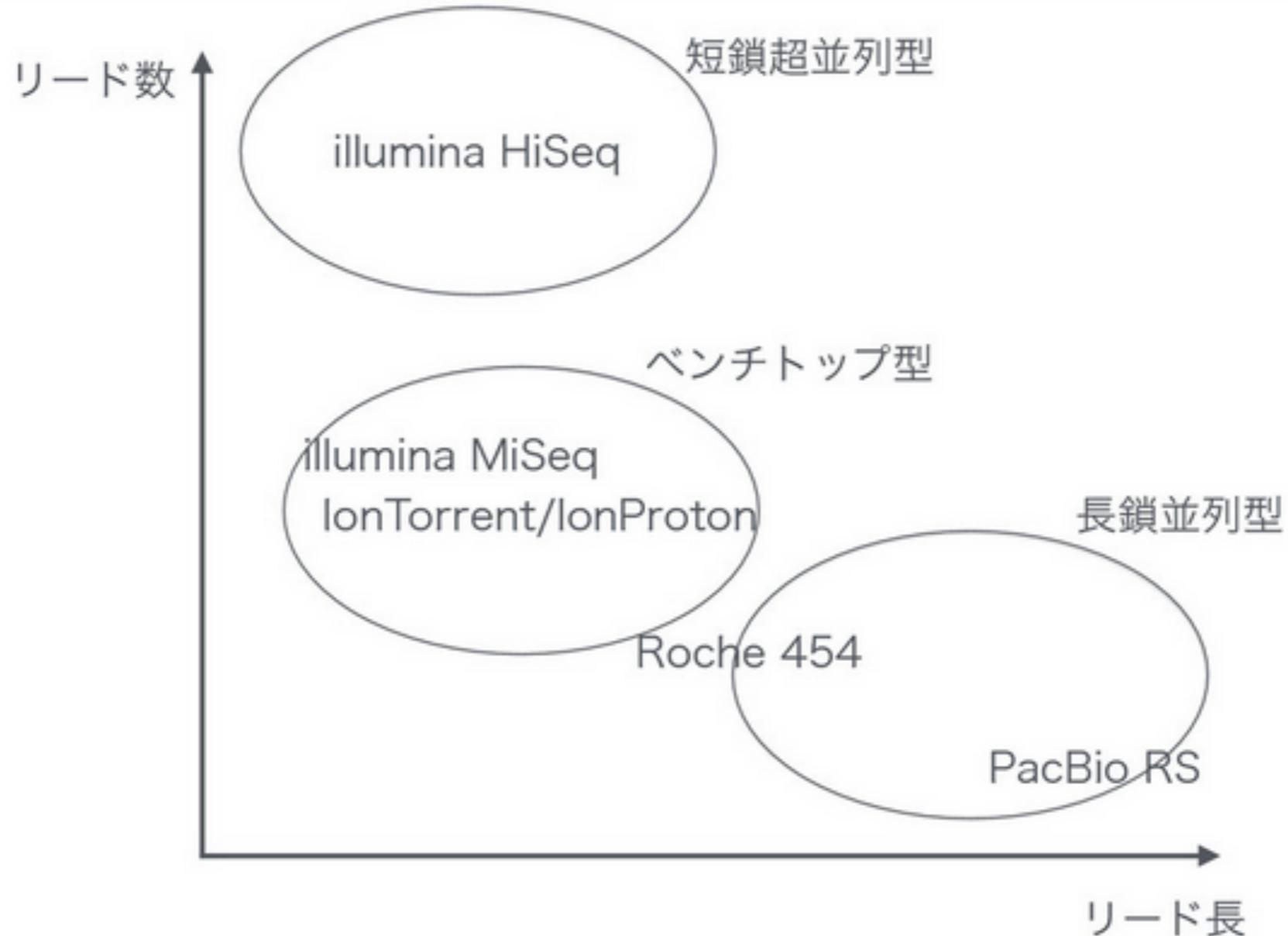
application / 実験種	total bases / 総塩基数	read length / リード長	read number (M) / リード数
ヒトゲノムリシーケンス	90-150Gb	2x100	900-1500
ターゲットリシーケンス	<1Gb	2x100	10
exome sequence	5~7Gb	2x100	70
RNA-Seq	5Gb	2x100	50
TSS-Seq	1Gb	1x50	20
small RNA	0.35Gb	1x35	>10
微生物ゲノム	>150Mb	2x100	>1.5
真核生物ゲノム	>4Gb	2x100	>40
Bisulfite-Seq	90-150Gb	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 a PubMed search results page for the term "NGS". The right tab is a NCBI BLAST search results page for the same term.

**PubMed Tab:**

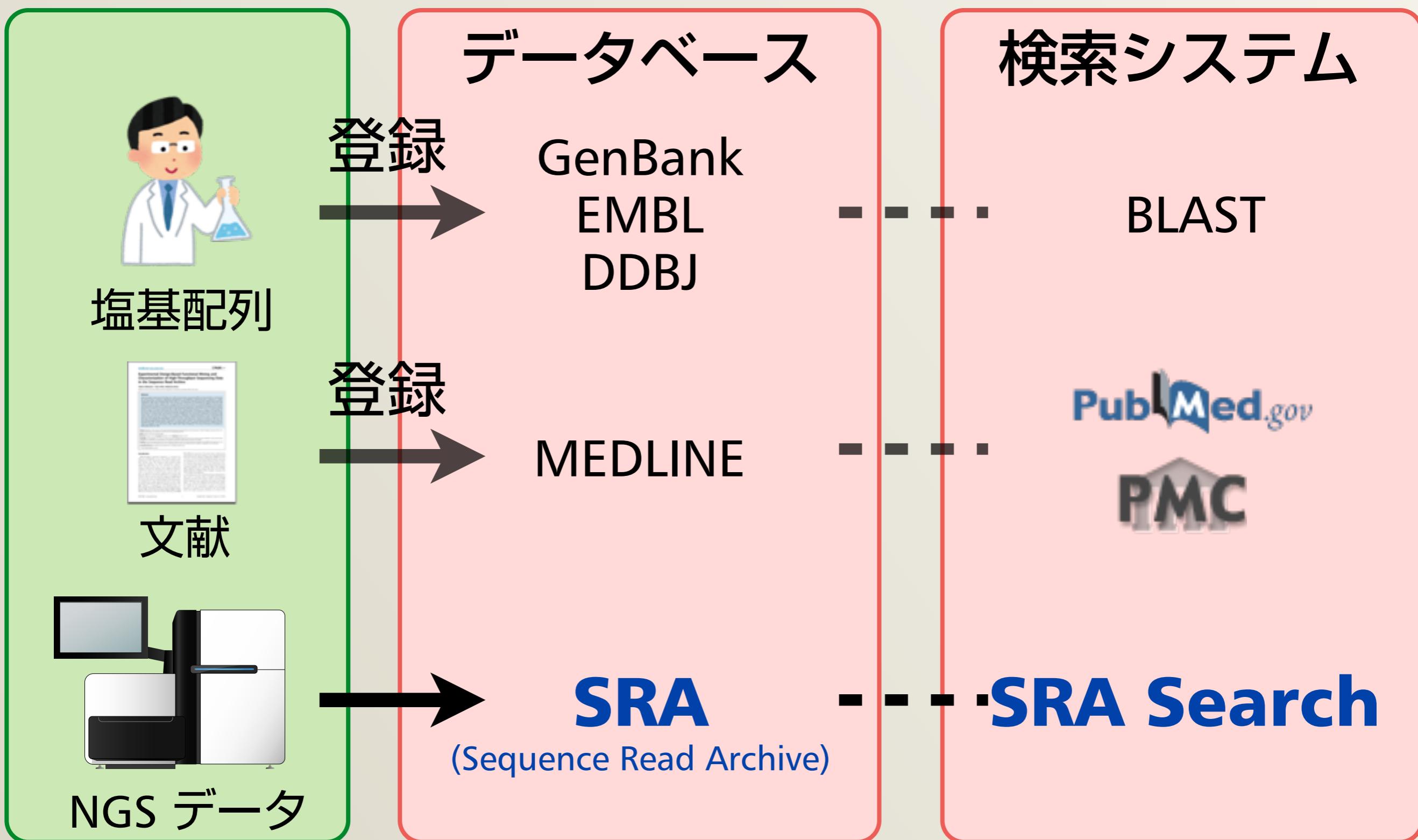
- URL: [www.ncbi.nlm.nih.gov/pubmed/?term=NGS](http://www.ncbi.nlm.nih.gov/pubmed/?term=NGS)
- Search term: NGS
- Display Settings: Summary, 20 per page, Sort by Date
- Results: 1 to 20 of 1441
- Articles listed (titles and authors):
  - An In-Solution Hybridisation Method for rich Clinical Samples for Analysis by NGS
  - Transactivating mutation of the MYOD1 rhabdomyosarcoma.
  - Introduction to Statistical Methods for Microbiome Data Analysis
  - High-Throughput Approaches for Microbial Community Analysis
  - NGSmethDB: an updated genome resource for Next Generation Sequencing

**BLAST Tab:**

- URL: [blast.ncbi.nlm.nih.gov/Blast.cgi#](http://blast.ncbi.nlm.nih.gov/Blast.cgi)
- Basic Local Alignment Search Tool
- NCBI BLAST/blastdb/Formatting Results - 9EN9RPMN01R
- Nucleotide Sequence (3065 letters)
- RID: 9EN9RPMN01R (Expires on 11-29 21:37 pm)
- Query ID: 1clj171457
- 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\_264
  - Superfamilies: Sulfate\_264, Sulfate\_transp, Sulfate\_transp\_superfamily
  - Multi-domains: Sulfate\_264
- Distribution of 112 Blast Hits on the Query Sequence:
  - Color key for alignment scores: <40 (black), 40-50 (blue), 50-60 (light blue), 60-80 (green), 80-200 (yellow), >=200 (red).
  - Query sequence length: 3065.
  - Approximate distribution: 1 (black), 600 (blue), 1200 (light blue), 1800 (green), 2400 (yellow), 3000 (red).

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 Page > >>

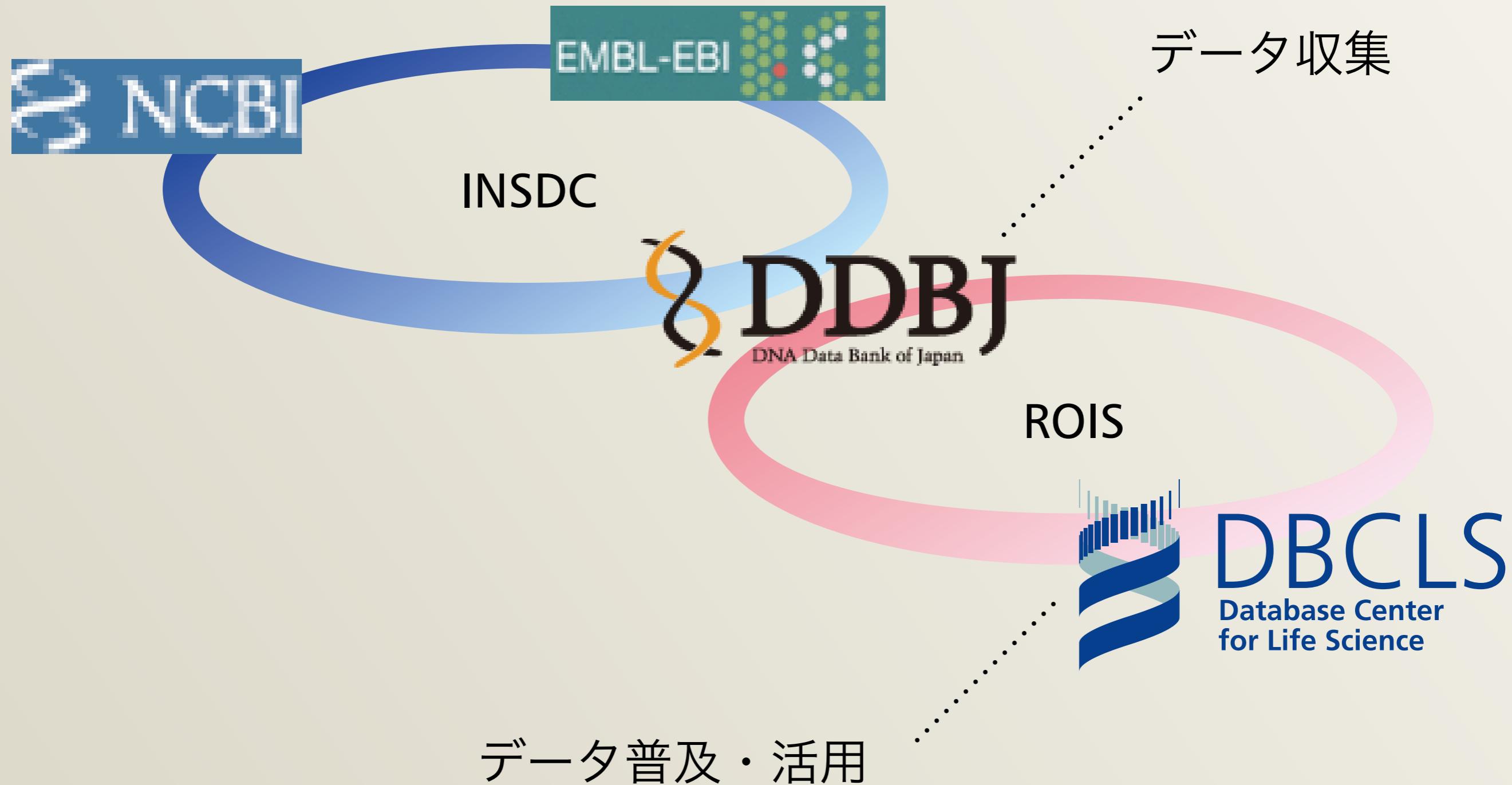
Filtered by

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)

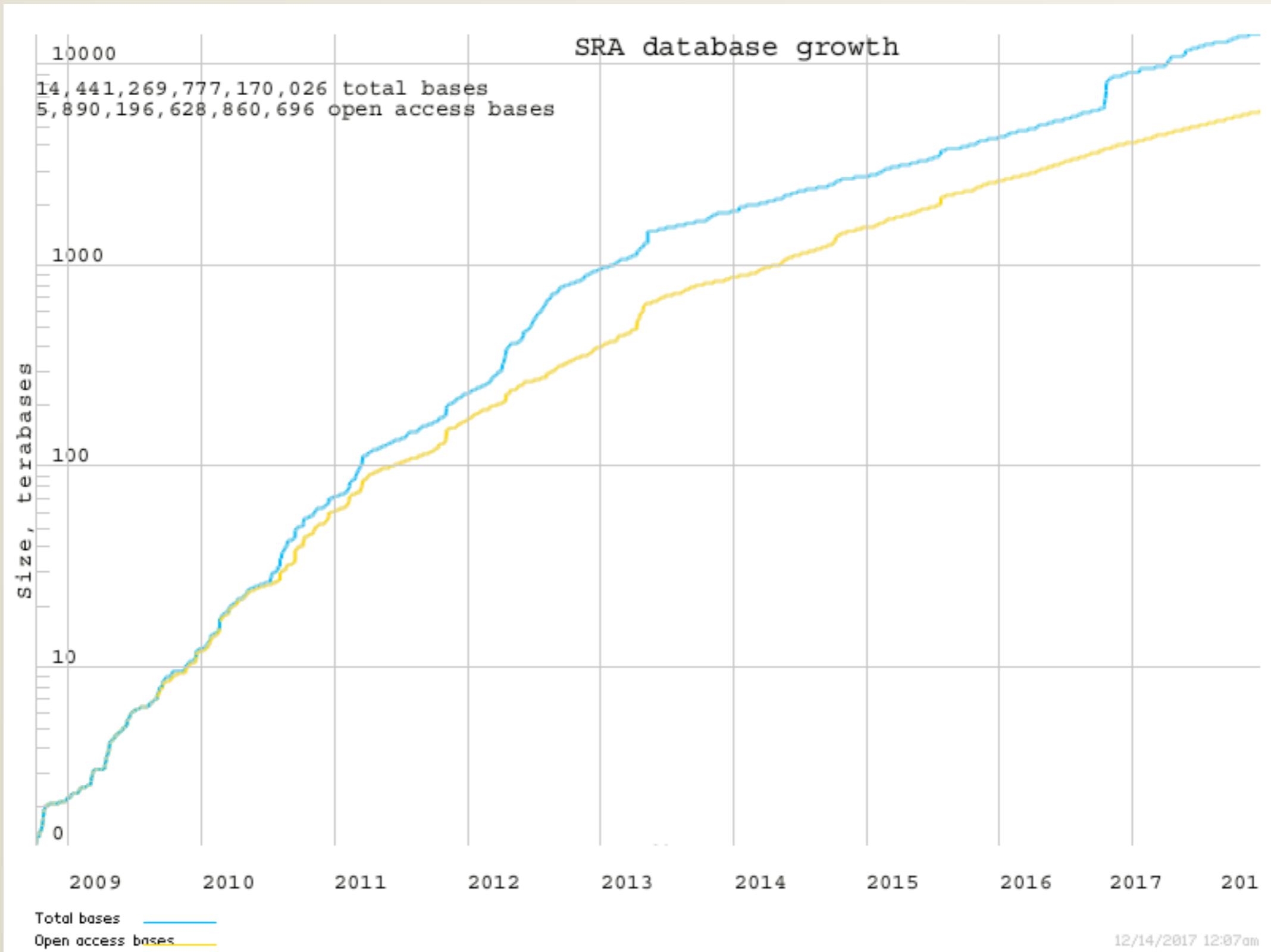
#	META_FILE	STUDY_ID	STUDY_NAME	STUDY_TYPE	ORGANISM	FILE_FORMATTED	CENTER_NAME
1	<a href="#">ERA007211.submission.xml</a> <?xml version="1.0" encoding="UTF-8"?> <SUBMISSION accession="ERA007211" alias="UU-SNP_1"	<a href="#">ERA007211</a>	<a href="#">ERP000177 GEUVADIS_RNASEQ_FIVE_INDIVIDUAL_PILOT</a>	<a href="#">Transcriptome Analysis</a>	<a href="#">Enterobacteria</a> <a href="#">phage phiX174</a> <a href="#">sensu lato</a> <a href="#">Homo sapiens</a>	124.1G	<a href="#">UNIGE</a>
2	<a href="#">ERA007211.run.xml</a> .org/2001/XMLSchema-instance"> <RUN alias="618MTAAXX_3" center_name="UU-SNP" run_center="UU-SNP" accession="ERR011434"	<a href="#">ERR011434</a> <a href="#">ERR011435</a> <a href="#">ERR011436</a> <a href="#">ERR011437</a> <a href="#">ERR011438</a> <a href="#">ERR011439</a>	<a href="#">ERP000177 GEUVADIS_RNASEQ_FIVE_INDIVIDUAL_PILOT</a>	<a href="#">Transcriptome Analysis</a>	<a href="#">Enterobacteria</a> <a href="#">phage phiX174</a> <a href="#">sensu lato</a> <a href="#">Homo sapiens</a>	124.1G	<a href="#">UNIGE</a>
3	<a href="#">SRA114122.submission.xml</a> <?xml version="1.0" encoding="UTF-8"?> <SUBMISSION alias="Lus SNP discovery" submission	<a href="#">SRP033223</a>	<a href="#">Lithum usitatissimum Genome sequencing</a>	<a href="#">Whole Genome Sequencing</a>	<a href="#">Lithium</a> <a href="#">usitatissimum</a>	24G	<a href="#">BioProject</a>
4	<a href="#">SRA024703.study.xml</a> .org/2001/XMLSchema-instance"> <STUDY alias="RNA-seq for SNP" center_name="ouc" accession="SRP003814"> <DESCRIPTOR> <STUDY	<a href="#">SRP003814</a>					
	<a href="#">SRA072630.submission.xml</a> <?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). The URL in the browser is [trace.ddbj.nig.ac.jp/jga/index.html](http://trace.ddbj.nig.ac.jp/jga/index.html). The page features the DDBJ logo and the text "Japanese Genotype-phenotype Archive". It includes navigation links for Home, Studies, and Submission. A search bar with a Google logo and a search button is present. The main content area has sections for Overview, Data Submission/Application, and Data Use.

**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 title bar reads "NBDCヒトデータベース" (NBDC Human Database). The main menu includes "ホーム" (Home), "データの利用" (Data Use), "データの提供" (Data Provision), "ガイドライン" (Guidelines), "NBDCヒトデータ審査委員会" (NBDC Human Data Review Committee), "成果発表" (Results), and "アクセス統計" (Access Statistics). There are two red warning boxes at the top: one stating "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" and another stating "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". A sidebar on the right lists recent news items: "新着情報" (Recent Information) with entries for "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)" and "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)". At the bottom, there's a search bar with "Search NBDC Human Database Beacon for Alternative Alleles" and "API help", and a note that "NBDC Human Database Beacon is a member of CA4GH 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
<a href="#">ERR015534</a>	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			<a href="#">File 1</a>	<a href="#">File 1</a>				
PRJEB2202	SAMEA733146	ERS012465	ERX005906	ERR015535	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			<a href="#">File 1</a>	<a href="#">File 1</a>				
PRJEB2202	SAMEA733149	ERS012460	ERX005901	ERR015536	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			<a href="#">File 1</a>	<a href="#">File 1</a>				
PRJEB2202	SAMEA733145	ERS012464	ERX005905	ERR015537	9606	Homo sapiens	Illumina Genome Analyzer II	SINGLE			<a href="#">File 1</a>	<a href="#">File 1</a>				

# 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	<a href="#">E-MTAB-305 in ArrayExpress</a>
----------	--

**Navigation**

Submission	<a href="#">ER4010380</a>	FTP
Experiment	<a href="#">ERX005900</a>	FASTQ
	<a href="#">ERX005901</a>	SRA
	<a href="#">ERX005902</a>	SRA
	<a href="#">ERX005903</a>	SRA
	<a href="#">ERX005904</a>	SRA
	<a href="#">ERX005905</a>	SRA
	<a href="#">ERX005906</a>	SRA
	<a href="#">ERX005907</a>	SRA
	<a href="#">ERX005908</a>	SRA
	<a href="#">ERX005909</a>	SRA
	<a href="#">ERX005910</a>	SRA
Sample	<a href="#">ERS012159</a>	
	<a href="#">ERS012160</a>	
	<a href="#">ERS012161</a>	
	<a href="#">ERS012462</a>	
	<a href="#">ERS012463</a>	
	<a href="#">ERS012464</a>	
	<a href="#">ERS012465</a>	
	<a href="#">ERS012466</a>	
	<a href="#">ERS012467</a>	
	<a href="#">ERS012468</a>	
	<a href="#">ERS012469</a>	

# DBCLS SRA 公共NGSデータの検索サイト

The screenshot shows the DBCLS SRA web interface. At the top left is the logo "DBCLS SRA". In the center is a circular icon containing a stylized "W" shape, with the text "DISCOVER Interesting & Available SRA Data" to its right. Below this is a large heading "Trends & Search SRA data" and a link "http://sra.dbcls.jp/".

**Species**

Species	Count
Homo sapiens	301929
Mus musculus	88239
human gut metagenome	40061
Oryza sativa	29622
soil metagenome	22162

**Study Type**

Study Type	Count
Whole Genome Sequencing	27402
Other	15292
Transcriptome Analysis	7758
Metagenomics	5105
Population Genomics	723

**Platform**

Platform	Count
Illumina HiSeq 2000	572366
Illumina MiSeq	97400
454 GS FLX Titanium	87105
Illumina Genome Analyzer II	68519
Illumina HiSeq 2500	59793

**Usage**

→ [for more detail](#)

- i Click bars or bubbles and you can see more details of selected data.
- i Input keywords and you can view ratio of feature selection.
- i You can see result of combined search criteria too.

**Search Conditions**

Free Keyword

# DBCLS SRA – リスト表示

Search Result

52.193.26.230/search?species=Homo%20sapiens&type=&instrument=&search\_query=



DBCLS SRA

Home blog

DBCLS SRA Metadata Search

Search query: undefined

948projects

Study ID	Study Title	Study Type	Sequencing Instrument	Scientific Name	Number of Sequencing	PubMed ID
DRP000003	Comprehensive identification and characterization of the nucleosome structure.	Transcriptome Analysis	Illumina Genome Analyzer	Homo sapiens	9	<a href="#">20400770</a>
DRP000004	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Illumina Genome Analyzer	Homo sapiens	3	<a href="#">20400770</a>
DRP000006	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Illumina Genome Analyzer	Homo sapiens	3	<a href="#">20400770</a>
DRP000006	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Illumina Genome Analyzer	Homo sapiens	3	<a href="#">20400770</a>
DRP000007	Comprehensive identification and characterization of the binding sites of	Transcriptome	Illumina	Homo sapiens	2	<a href="#">20400770</a>

# DBCLS SRA – 詳細表示

Comprehensive identification of transcription start sites of putative non-coding RNAs by multifaceted use of massively parallel sequencer.

Sathira Nuenkanya N, Yamashita Riu R, Tanimoto Kousuke K, Kanai Akinori A, Preuchi Takeko T, Kanematsu Goutaro G, Nakai Kenta K, Suzuki Yutaka Y, Sugano Sumio S DNA roccoh : an international journal for rapid publication of reports on genes and genomes, 2010/06/16

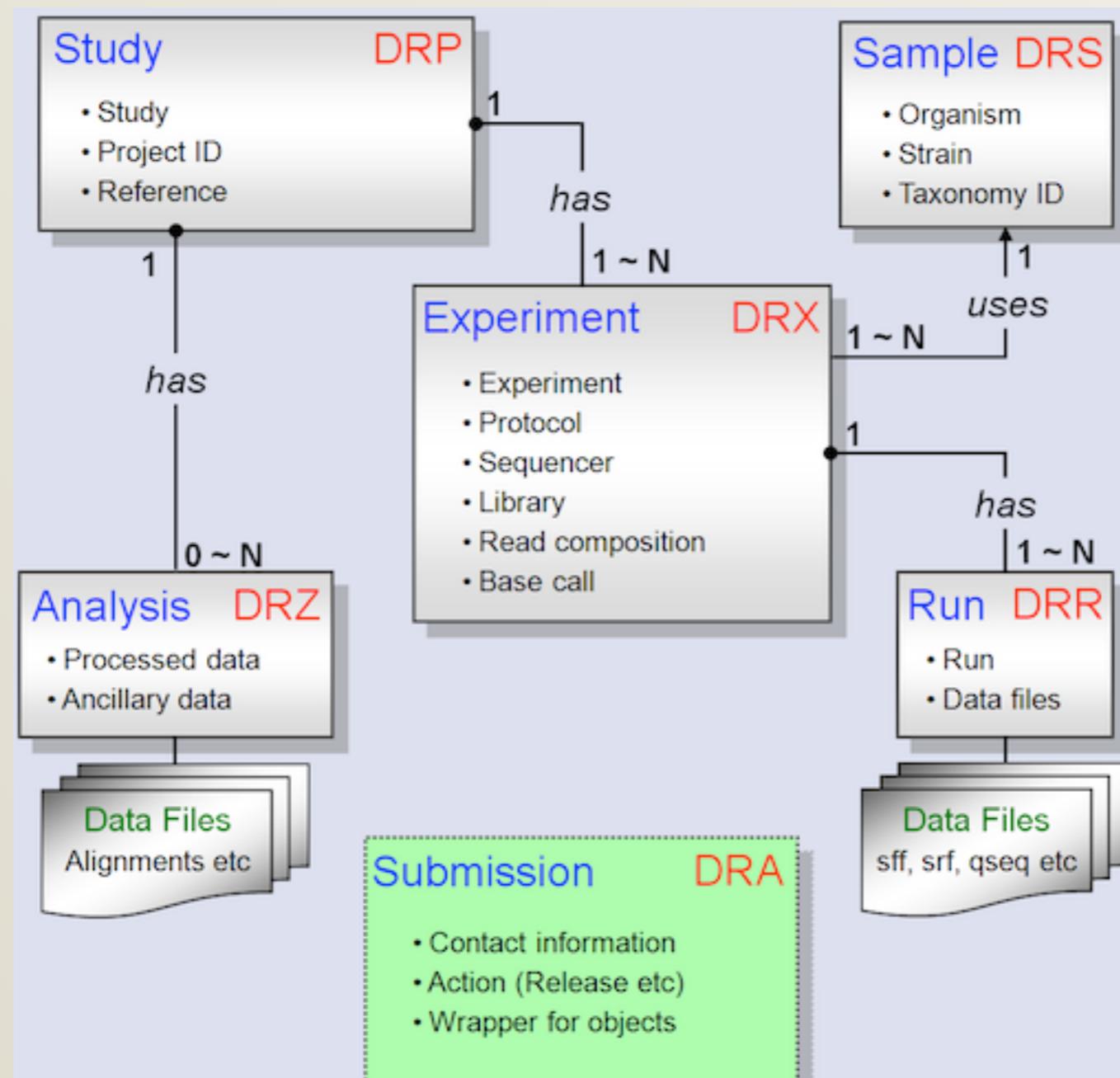
On the basis of integrated transcriptome analysis, we show that not all transcriptional start sites clusters (TSCs) in the intergenic regions (ITSCs) have the same properties; thus, it is possible to discriminate the ITSCs that are likely to have biological relevance from the other noise-level ITSCs. We used a total of 251,933,381 short-read sequence tags generated from various types of transcriptome analyses in order to characterize 6039 ITSCs, which have significant expression levels. We analyzed and found that 23% of these ITSCs were located in the proximal regions of the RefSeq genes. These RefSeq-linked ITSCs showed similar expression patterns with the neighboring RefSeq genes, had widely fluctuating transcription start sites and lacked ordered nucleosomal positioning. Those ITSCs seemed not to form independent transcriptional units, simply representing the by-products of the neighboring RefSeq genes, in spite of their significant expression levels. Similar features were also observed for the TSCs located in the antisense regions of the RefSeq genes. Furthermore, for the remaining ITSCs that were not associated with any RefSeq genes, we demonstrate that integrative interpretation of the transcriptome data provides essential information to specify their biological functions in the hypoxic responses of the cells.

PubMed PMC

## Methods

Cell culture and tissues  
Construction of the TSS Seq libraries and analysis of the TSS tags  
Construction of the Nucleosome Seq library and analysis of the nucleosome tags

# データ構造（概略版）

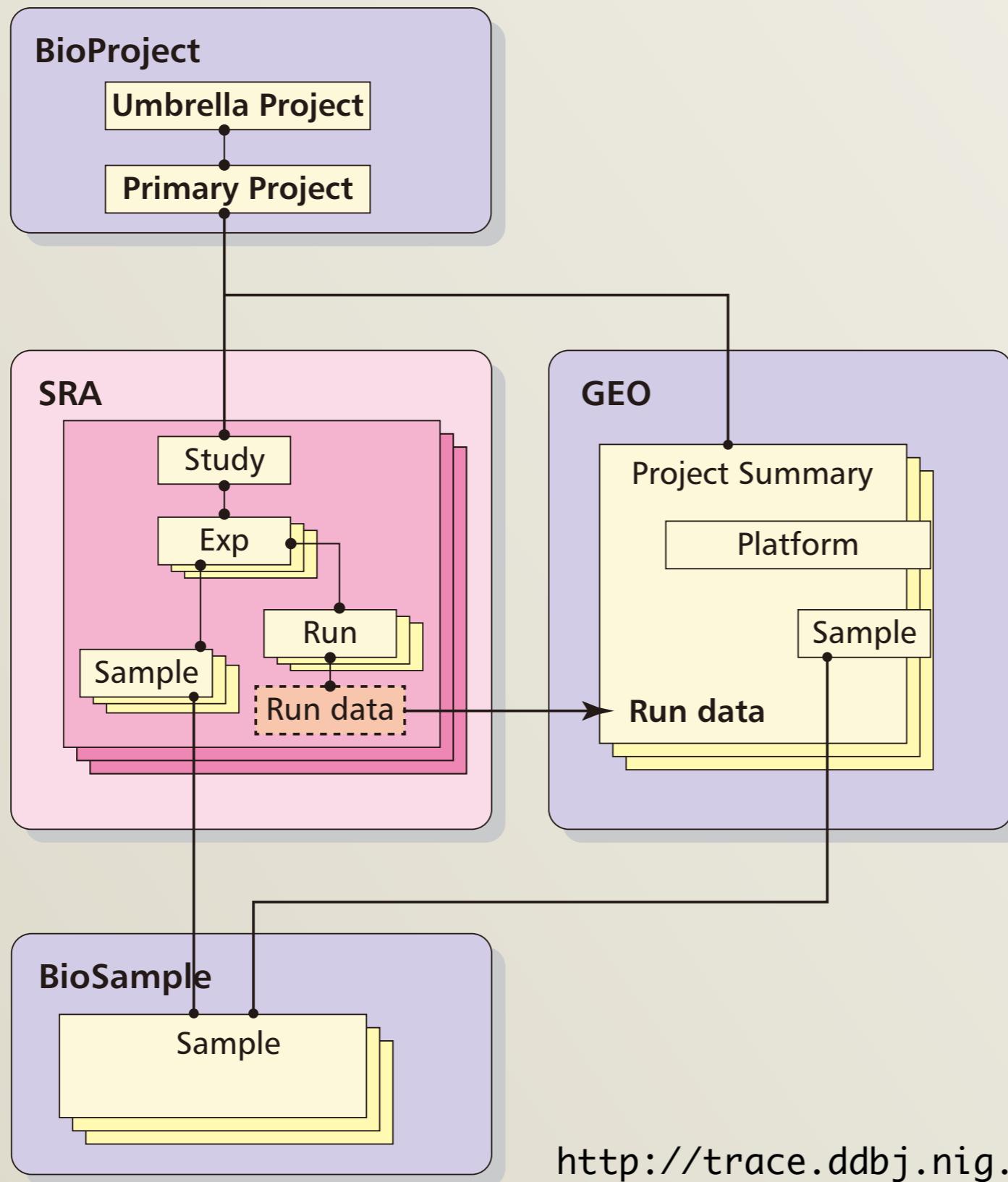


"has" and "uses": relationship between objects

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、…

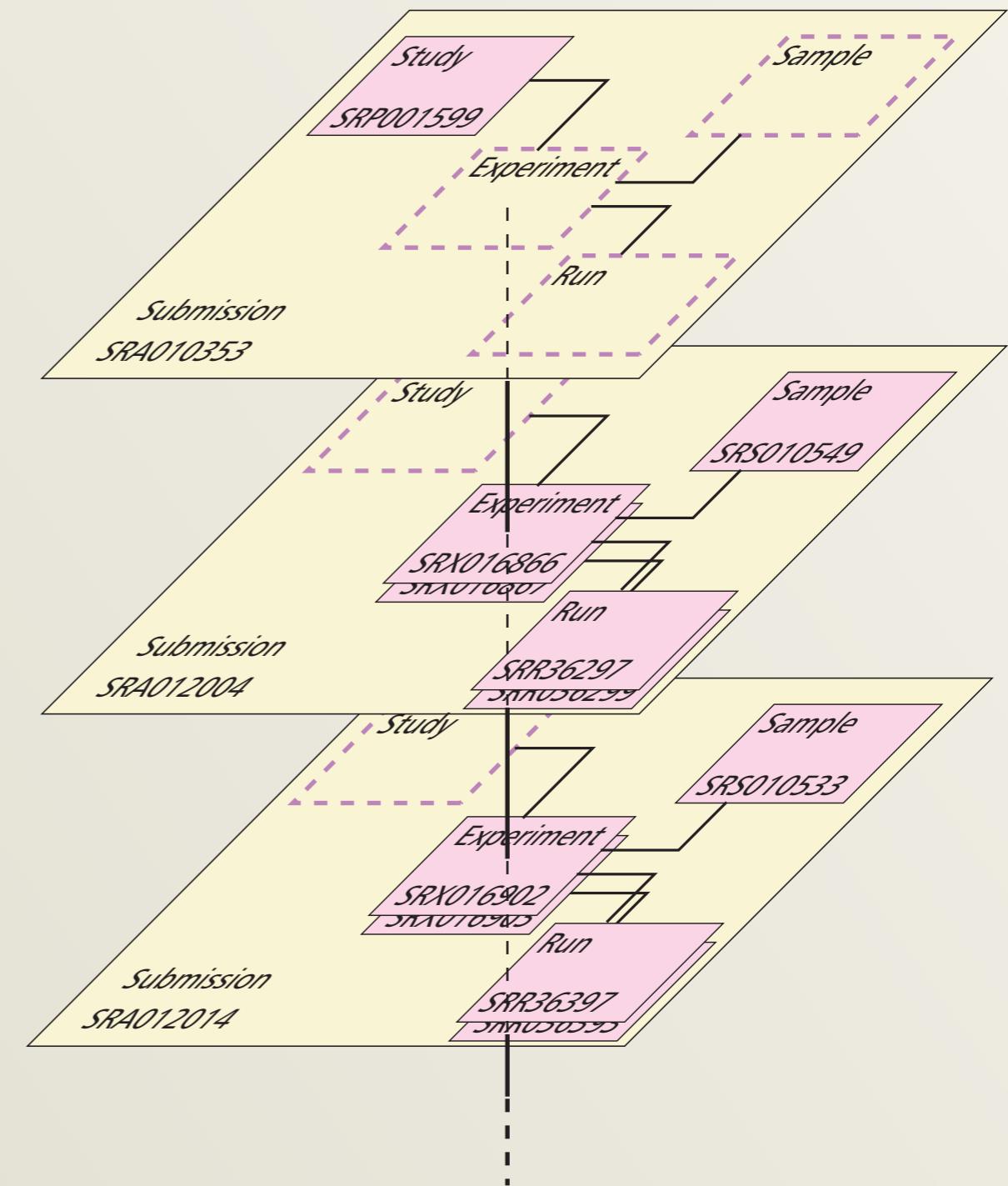
データベースの構造

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



ワインが飲みたい

どれにする?



どうしようかなあ...

名前	Soleil Hikumo Rouge
タイプ	赤
ワイナリー	旭洋酒
生産地	山梨
ブドウ品種	ピノノワール+ベイリーA
製造年	2012年

→ メタデータによる選択

赤で重くないやつ

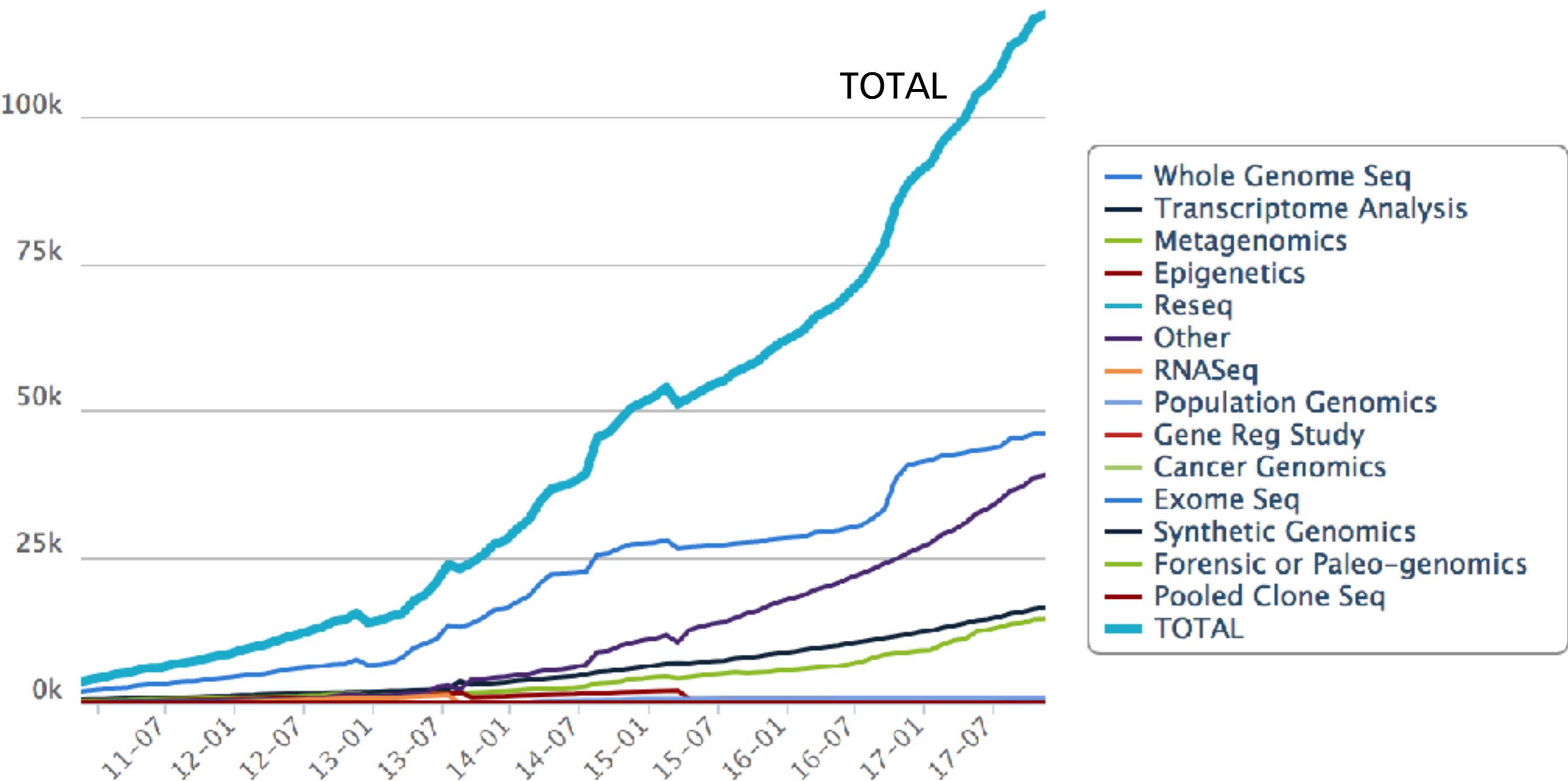
→ 中身のクオリティによる足切り



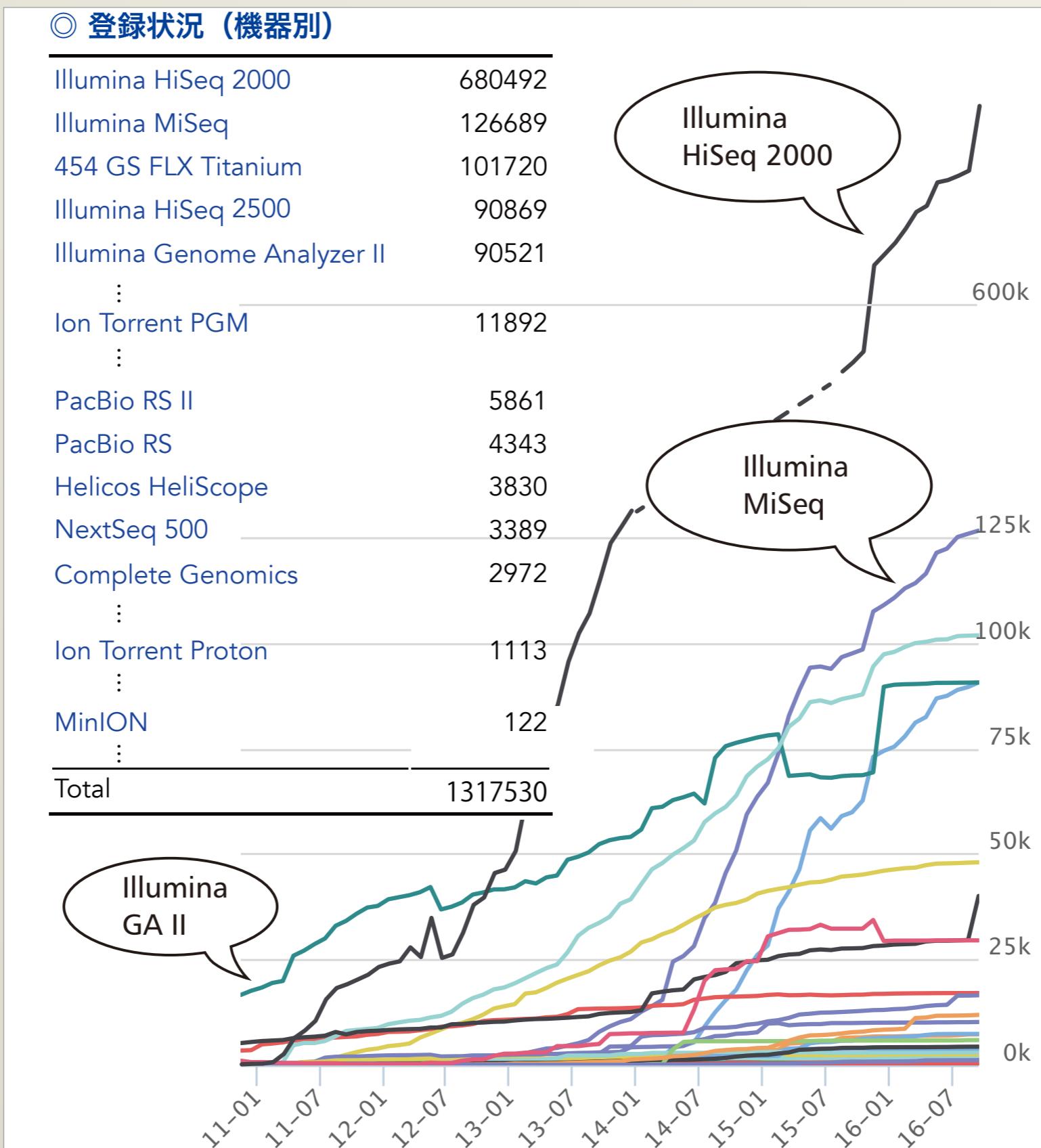
# データ数の推移 – 目的別

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



# 論文からの検索

Species:		Study Type:		Platform:		Rows:		Search	
First	Previous	Next	Last	Page:	1	Go	Rows:		
pmid	article_title	journal	vol	issue	page	date	sra_id_orig	sra_id	sra_title
26578681	Dual RNA-seq of Non-typeable <i>Haemophilus influenzae</i> and Host Cell Transcriptomes Reveals Novel Insights into Host-Pathogen Cross Talk	mBio	6	6	-	2015	SRA216498	SRA216498	Dual RNA-sequencing of non-typeable <i>Haemophilus influenzae</i> and host cell transcriptomes reveals new aspects of host-pathogen interface
26575290	Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma	eLife	4	-	-	2015-Nov-17	SRA220947	SRA220947	Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma
26573221	A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes	BMC genomics	16	1	952	2015	SRA205374	SRA205374	Short and Long RNA sequencing of human mature erythrocytes
26571212		Nature cell biology	-	-	-	2015-Nov-16	SRA180725	SRA180725	
26560027	neuroblastoma mediated by a LMO1 super-enhancer polymorphism	Nature	-	-	-	2015-Nov-11	SRA236526	SRA236526	Genetic predisposition to neuroblastoma mediated by a single nucleotide polymorphism within a LMO1 oncogene super-enhancer element

Publication info

data info  
(SRA)

# NGSデータを使って論文を出したら

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

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

## Availability of data and materials

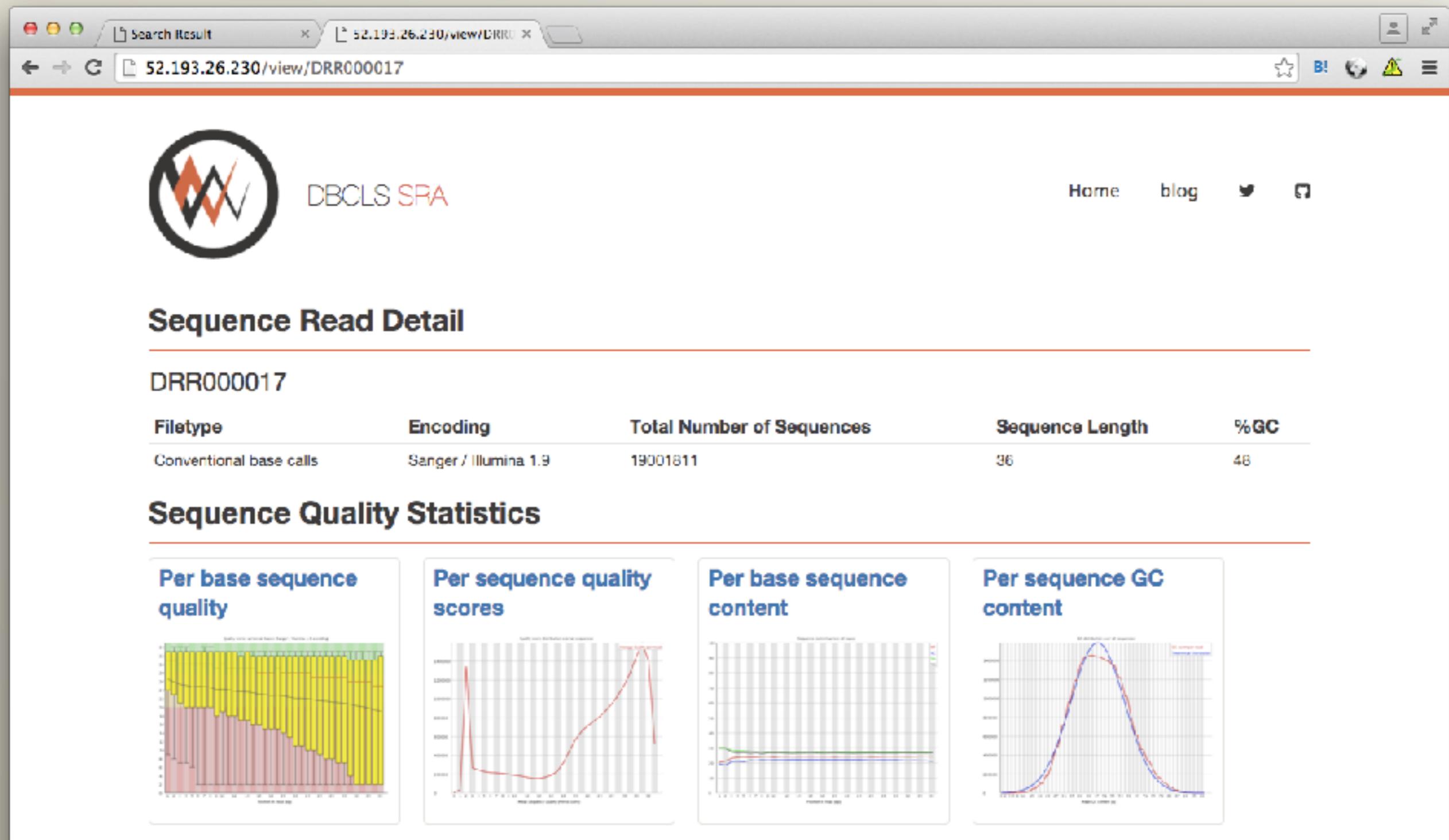
The nucleotide sequences for BmEno1, BmEno2, and BmEnoC were submitted to DDBJ/ENA/GenBank (Accession Nos. LC170036, LC170037, and LC170038, respectively). The RNA-seq reads supporting the conclusions of this article are available in the Sequence Read Archive (SRA) with accession ID DRA005094, <https://www.ncbi.nlm.nih.gov/sra/>.

## 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 T, Nakamura K, Nakanishi T, et al. Recombinant BmEno1 protein inhibits the proliferation of human hepatocellular carcinoma cells. *Int J Biochem Cell Biol* 2003;35:121–127.
11. Wang X, Li Y, Guo Y, et al. High-expression of BmEno1 gene for silk fibroin production. *Appl Biochem Biotechnol* 2006;121:11–18.
12. Xia Q, Li Y, Zhang L, et al. BmEno1 gene expression in mulberry silkworm midgut. *Appl Biochem Biotechnol* 2006;121:19–26.
13. Brewster A, Kondo T, Nakamura K, et al. Enolase from Bombyx mori: its molecular cloning and expression analysis. *Appl Biochem Biotechnol* 2000;85:11–18.

# QC result



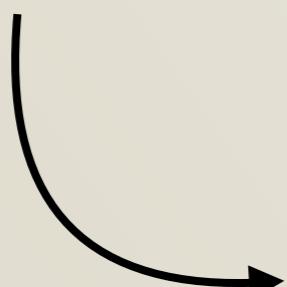
# 疾患からの検索

Disease Type: ANY

Total: 171 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 10

Disease	疾患名	# of submission
Genetic Predisposition to Disease	遗传的素因(疾患)	9
Breast Neoplasms	乳房腫瘍	8
Disease Progression	病勢悪化	Total: 6 << first < prev 1 next > last >> 10
Obesity	肥満	
Malaria	マラリア	
Chromosome Aberrations	染色体異常	
HIV Infections	HIV感染症	
Chromosome Breakage	染色体切断	
Polypliody	多倍数性	
Disease Models, Animal	疾患モデル(動物)	

Total: 171 << first < prev 1 2 3 4 5 6 7



SRA ID	SRA Title	Disease	疾患名	PMID
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">18987736</a>
SRA026055	DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">19657110</a>
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">18987736</a>
SRA026055	Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">19657110</a>
SRA009897	In-depth characterization of the microRNA transcriptome in a leukemia progression model	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">18849523</a>
SRA029797	Exome Sequencing Identifies Somatic Mutations in Acute Monocytic Leukemia	Leukemia, Myeloid, Acute	白血病-急性骨髓性	<a href="#">21399634</a>

Total: 6 << first < prev 1 next > last >> 10

# 生物種からの検索

Project List (from taxonomy) x

sra.ncbi.nlm.nih.gov/cgi-bin/taxon2study.cgi?type=&platform=&taxon\_id=4530&taxon\_tree=on&taxon\_n=Oryza+sativa

Project List from taxonomy (β version)

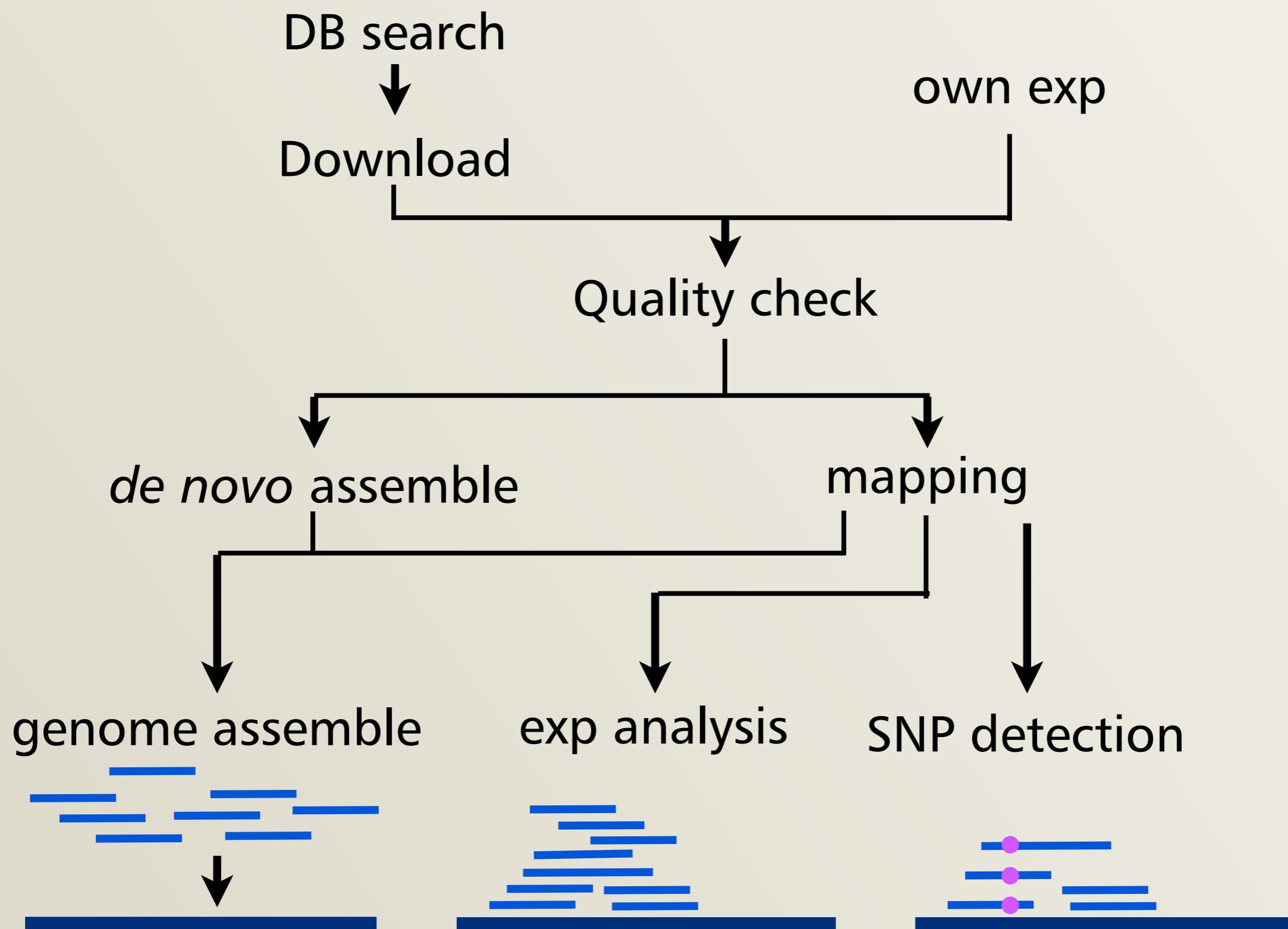
Study Type: Platform: Taxon ID: 4530 incl. child taxonomy (ex. strains)  Species name: Oryza sativa Search → back to DBCLS SRA top

Oryza > Oryza sativa → TAB-delimited format

Total: 580 << first < prev 13 14 15 16 17 18 19 20 21 22 next > last >> 10

SRA ID	Study ID	Study Title	Study Type	Taxon ID	Taxon Name	Exps	Runs
SRA216709	SRP051189	Oryza sativa Transcriptome or Gene expression	Transcriptome Analysis	4530	Oryza sativa	0	0
SRA240900	SRP064996	Oryza sativa japonica Genome sequencing	Whole Genome Sequencing	4530	Oryza sativa	1	0
SRA244535	SRP055515	Small RNA populations in Argonaute complexes purified from Rice stripe Tenuivirus (RSV)-infected rice	Other	4530	Oryza sativa	6	0
DRA000685	DRP000716	Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles	Transcriptome Analysis	39946	Oryza sativa Indica Group	7	57
ERA000212	ERP000096	Novel exon splicing junctions and novel transcriptional active regions in Oryza sativa	Transcriptome Analysis	39946	Oryza sativa Indica Group	1	18
ERA009070	ERP000235	The indica genome sequence by next-generation sequencing	Whole Genome Sequencing	39946	Oryza sativa Indica Group	5	0
SRA010796	SRP001724	GSE19050: Degradome sequencing reveals endogenous small RNA targets in rice ( <i>Oryza sativa</i> L. ssp. <i>indica</i> )	Transcriptome Analysis	39946	Oryza sativa Indica Group	1	0
SRA012190	SRP002084	Single-base resolution DNA methylomes of rice and functional roles of DNA methylation	Epigenetics	39946	Oryza sativa Indica Group	22	0
SRA026538	SRP004490	Plant genome sequencing data	Whole Genome Sequencing	39946	Oryza sativa Indica Group	93	0
SRA036013	SRP006587	Oryza sativa Project	Whole Genome Sequencing	39946	Oryza sativa Indica Group	1	0

# NGSデータ解析の流れ



# 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" (Flowchart of DRA registration) which includes a sub-section "1. 登録アカウントを作成" (Create registration account). On the right side, there is a sidebar titled "In this page" with links to various DRA-related topics like "DRAについて", "メタデータ", etc., and sections for "PDF Download" and "Submit". At the bottom, there is a footer with a link to the submission page.

## 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

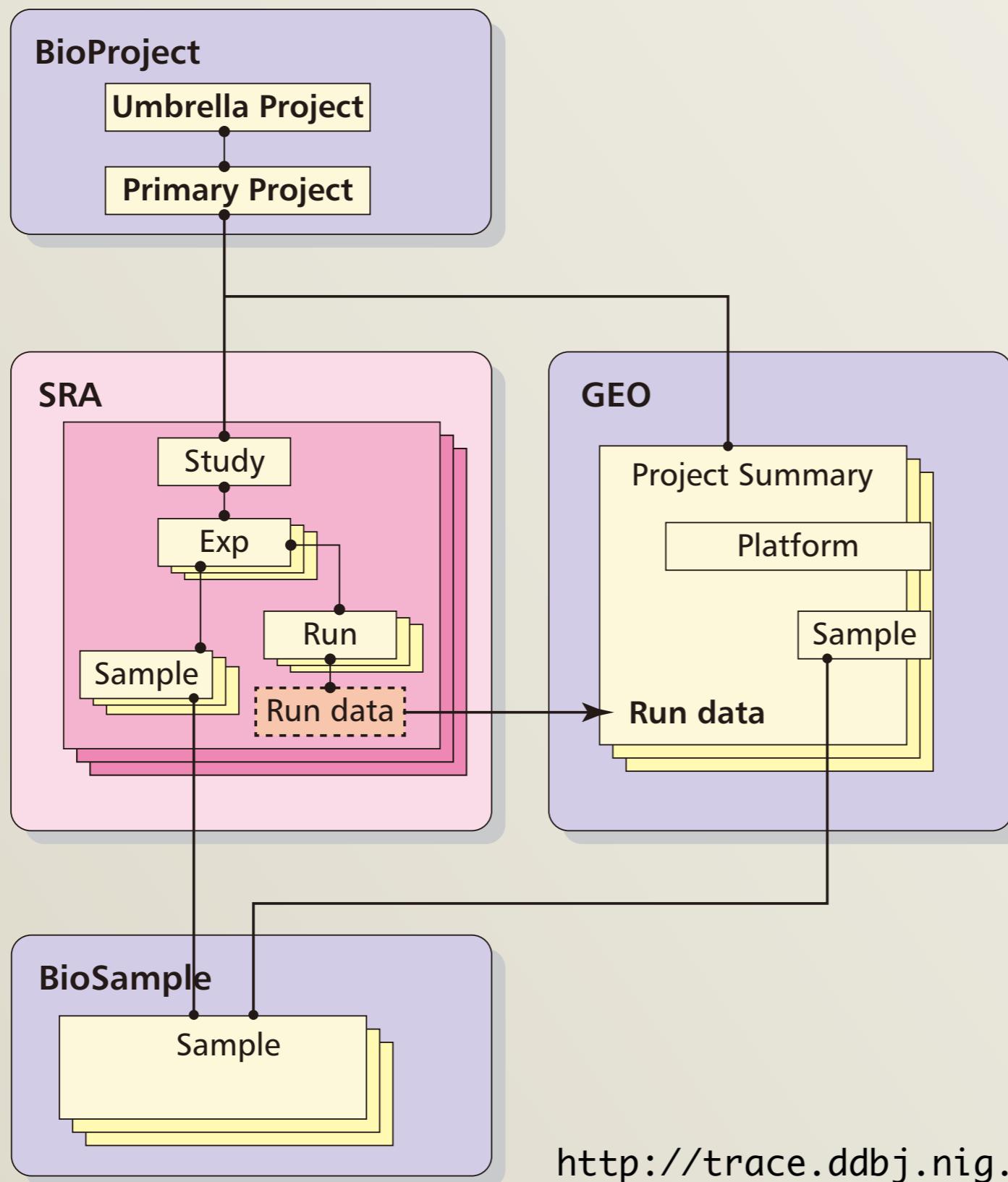
Login & Submit  
Contact

### 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ション番号が発行される

# DRAへの登録 (Sample)

The screenshot shows a Microsoft Excel spreadsheet titled "bioample". 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 is a table with columns labeled A through L. 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 corresponding to these headers. Row 2: Bmori-strainXXXwt-tissue-1, Bombyx mori, 7091, PRJDB####, XXX, Kyushu Univ/NBRP. Row 3: Bmori-strainXXXwt-tissue-2, Bombyx mori, 7091, PRJDB####, XXX, Kyushu Univ/NBRP. Row 4: Bmori-strainXXXwt-tissue-3, Bombyx mori, 7091, PRJDB####, XXX, Kyushu Univ/NBRP.

	A	D	E	F	G	L	
1	*sample_name	*organism	*taxonomy_id	bioproject_id	strain	biomaterial_provider	collection
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: chalkless-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

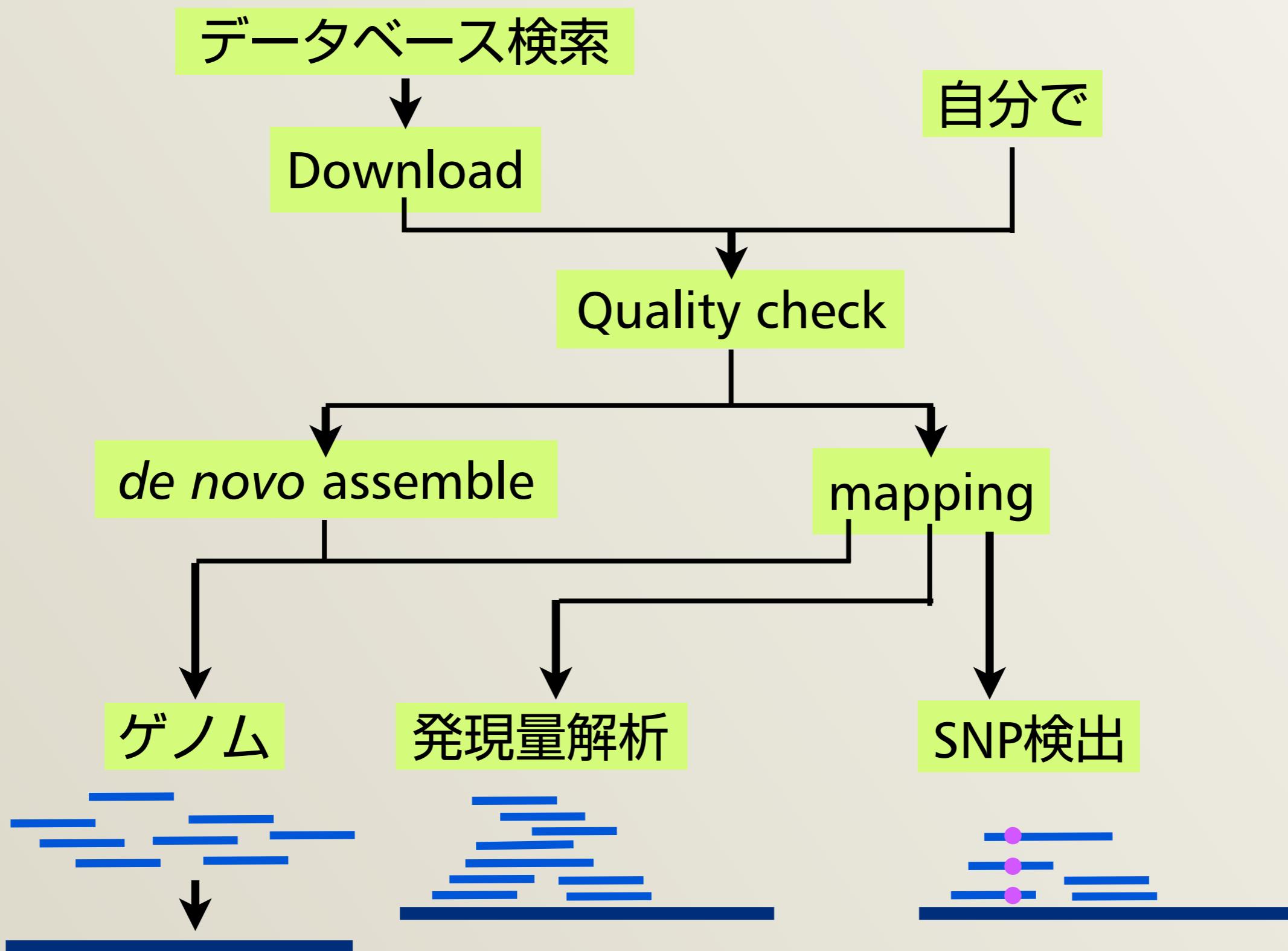
[Add another Experiment\(s\)](#) 1 [Copy Experiment #1](#)

#	Alias	BioSample Used	Library Name	Library Source	Library Selection	Library Strategy
1	chalkless-0001_Experiment_0001	SAMD00068752		TRANSCRIPTOMIC	cDNA	RNA-Seq
2	chalkless-0001_Experiment_0002	SAMD00068753		TRANSCRIPTOMIC	cDNA	RNA-Seq
3	chalkless-0001_Experiment_0003	SAMD00068754		TRANSCRIPTOMIC	cDNA	RNA-Seq

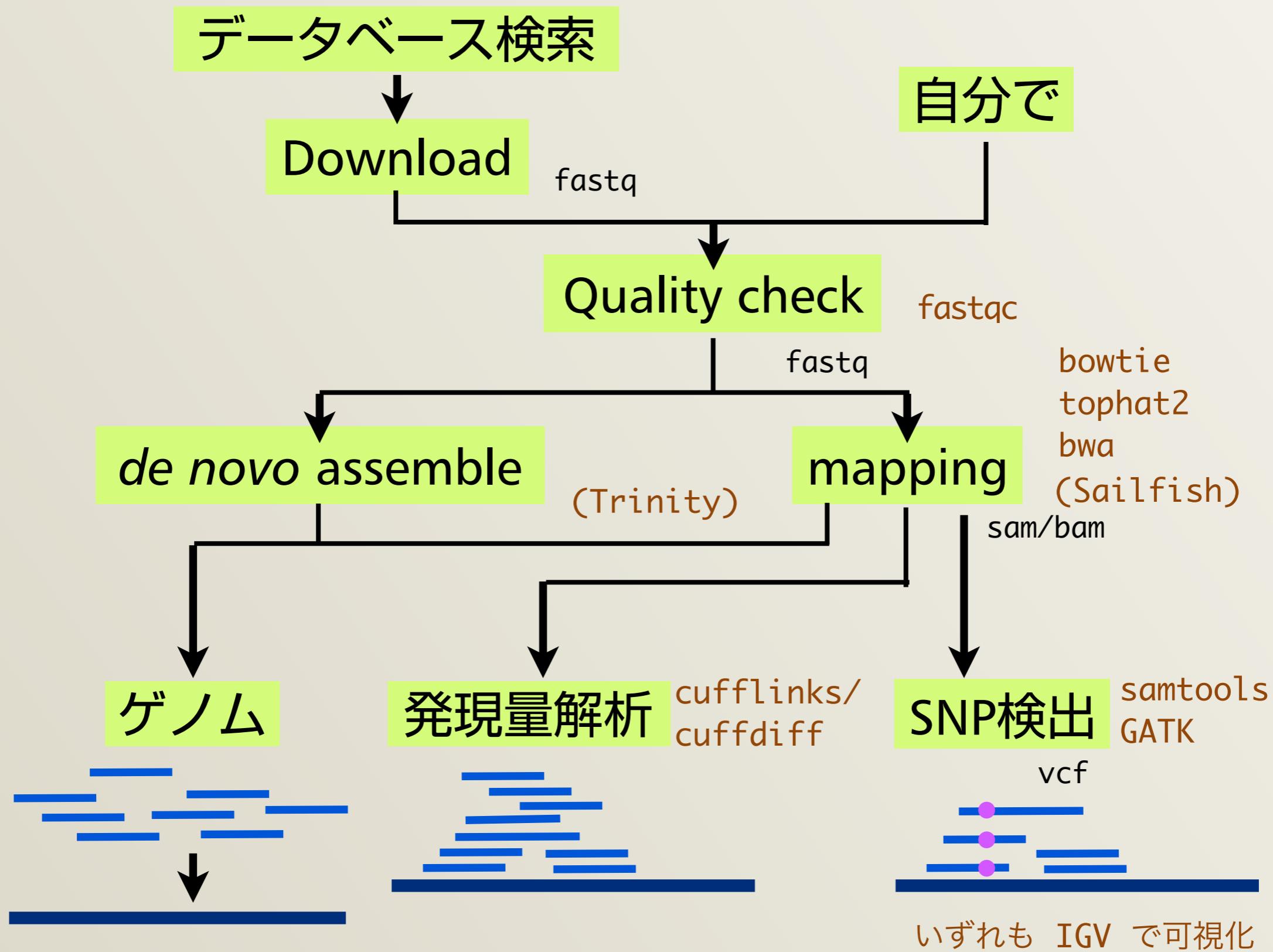
[Save](#) [Next \(RUN\)](#)

# 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）



# 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



この画像を表示

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



次世代シーケンサーDRY解析教本(細胞工学別冊) 単行本 -

2015/10/15

清水厚志 (監修), 坊農秀雅 (監修)

★★★★★ 5件のカスタマーレビュー

▶ その他 (2) の形式およびエディションを表示する

Kindle版

¥ 5,400

今すぐお読みいただけます: 無料アプリ

単行本

¥ 5,832

¥ 4,013 より 11 中古品の出品

¥ 5,832 より 1 新品

1/26 木曜日 にお届けするには、今から**23 時間 8 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)

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)。

NGSデータ解析を丁寧に解説。Kindle版あり

# 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 加者必読)	Bio-Linux8とBioのイ ンストール&実際の PC環境の構築	<ul style="list-style-type: none"><li>Bio-Linux8(第2回および3回)で利用するova(主催・共催機関 ファイル)の導入確認</li><li>共有フォルダ設定完了確認</li><li>基本的なLinuxコマンドの習得状況確認</li><li>R本体およびパッケージのインストール確認</li><li>講習会室の用意予習内容の再確認</li><li>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</li></ul>	担当講師 (敬称略)	講義資料 (PDF: 4MB) 動画(総合TV)
7月20日 (水)	10:30- 16:15	第1回 実習 (農学生命情報科学 特設)	ゲノム解析、塩基配 列解析	<ul style="list-style-type: none"><li>NGS解析手順、ウェブツール(DDG Pipeline)との連携</li><li>k-mer解析( k語や過短塩基に基づく各種解析 )の基礎と応用</li><li>塩基ごとの出現頻度解析(<math>k=1</math>)、2塩基並基 の出現頻度解析(<math>k=2</math>)</li><li>塩基配列解析を行ったための基本スキルの復習</li></ul>	吉田 玲二 (准正人学)	講義資料 (PDF: 7.3MB) 動画(2) 総合TV

## 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

# 解析について詳細な情報を探したい

The screenshot shows a web browser window with the title bar 'Rで)塩基配列解析'. The address bar contains the URL 'www.iu.a.u-tokyo.ac.jp/~kadota/r\_seq.html'. The main content area displays the following text:

**(Rで)塩基配列解析**  
(last modified 2017/01/23, since 2010)

このウェブページのR関連部分は、[インストール](#) | [についての往來手帳](#) (Windows2015.04.04版とMacintosh2015.04.03版)に従って フリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は[基本的な利用法](#)(Windows2015.04.03版と Macintosh2015.04.03版)で自習してください。本ウェブページを体系的にまとめた書籍もあります。(2015/04/03)

**What's new?**

- 下記の統観です。間違いではないような気がしてきましたw。「正規化 | リンブル間 | 2群間 | 梱製あり | median(Anders 2010)」の例題3の size factorsとnormalization factorsの変換式や、実際のDESeqで得られたサイズファクター数値分布的にも妥当。。。ですかね。。。 (2017/01/04) **NEW**
- NGSハンズオン講習会2016の2016年7月22日の講義資料のスライド107で、effective library sizesからのサイズファクター(size factors)への変換式が cf.llbsizes/mcan(cf.llbsizes)となっています。じつはこれって間違いで本当はmcan(cf.llbsizes)/cf.llbsizesだったんじゃないかと。。。ふと思いついています。當時誰からも叱咎されませんでしたが(話の流れ的にはTips的なところだったからかもしれません)、ヘビーユーザの方は特にTCC正規化係数を他のパッケージ(特にsize factorsに変換して使う場合は)と組み合わせる場合は、ちょっと気にかけておいてください。間違いが確定であると思われた方はレポートいただければ幸いですml\_2m。 (2017/01/04) **NEW**
- 「解説 | 一般 | アライメント | ...」周辺の項目名を整理しました。 (2016/12/29) **NEW**
- NGSハンズオン講習会2016の講義映像 ([YouTube](#)と [Youtube](#))が公開されました。 (2016/12/07)
- [日本乳酸菌学会誌のNGS関連連載の第8回分](#)PDFを公開しました。 (2016/11/30)
- バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けの講義資料を作成しました。「参考資料 | 講習会、講義、講演資料」の2016.09.12-16の日付のものです。日本乳酸菌学会誌のNGS連載第1-4回あたりのダイジェスト版のようなものです。NGSハンズオン講習会2016の予習事項が撇々と記載してあるので、受講を断念したヒトも、ここからだとスムーズに頭に入っていくかもしれません。 (2016/08/26)
- 著者情報の項目の情報量が多くなってきたので、「[重複](#)、[学会誌](#)」と「[講習会](#)、[講義](#)、[講演資料](#)」の2つに分割しました。 (2016/08/23)

---

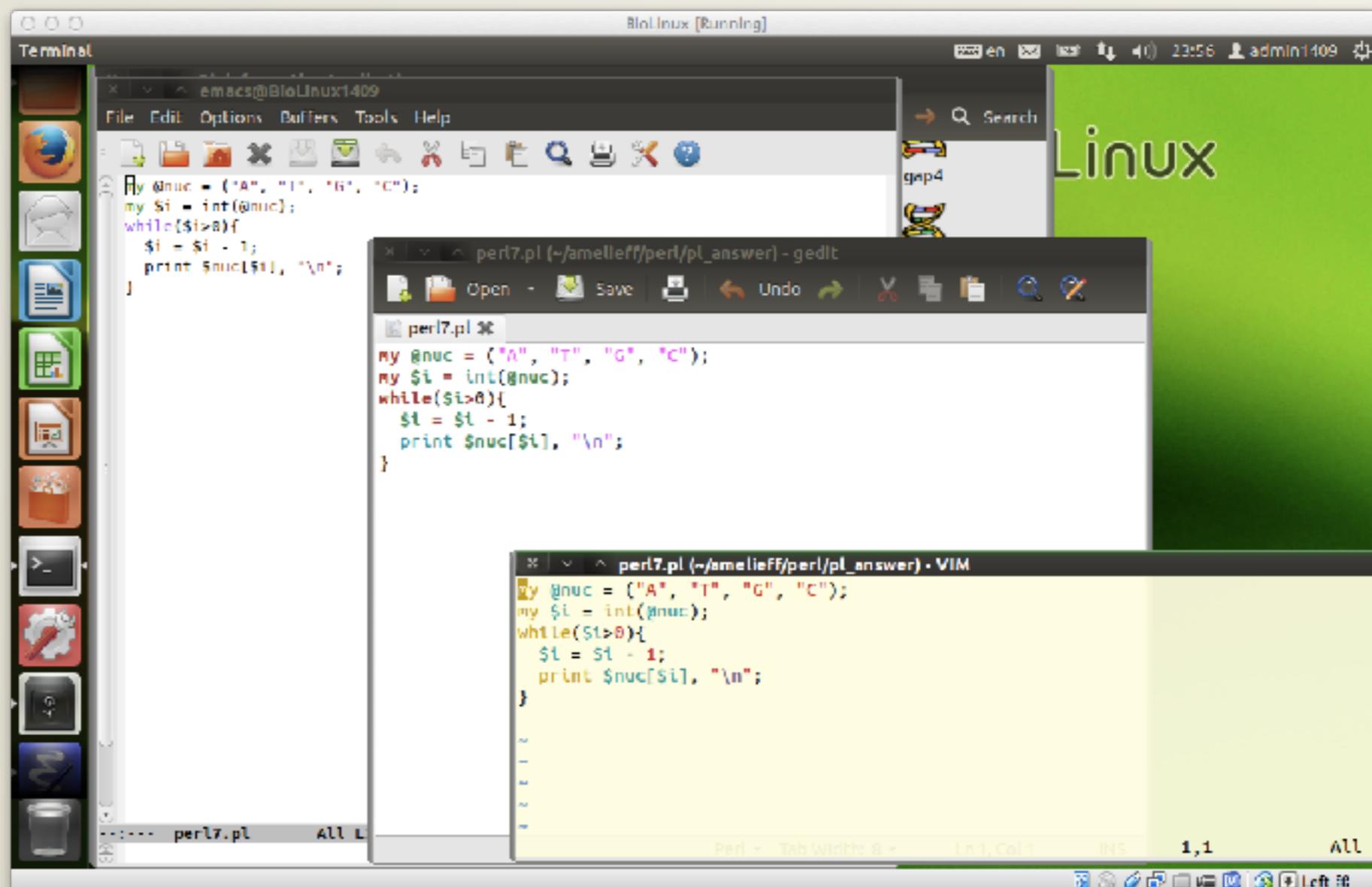
- [門田からメール返信をもらえない場合は](#) (last modified 2016/08/23)
- [はじめに](#) (last modified 2015/03/31)
- [著者資料 | 書籍、学会誌](#) (last modified 2016/11/30)

[トップページへ](#)

門田さん (東大アグリバイオ)

[http://www.iu.a.u-tokyo.ac.jp/~kadota/r\\_seq.html](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html)

# 解析環境・コマンドラインベース



BioLinux（カスタマイズVer.）  
NGSハンズオン講習会で使う解析環境

ツールをひとつおりインストール済  
VirtualBoxの上で仮想環境構築

[http://www.iu.a.u-tokyo.ac.jp/~kadota/  
r\\_seq.html#bioinfo\\_ngs\\_sokushu\\_2016\\_20160719](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#bioinfo_ngs_sokushu_2016_20160719)

# 解析環境・ウェブベース

The screenshot shows the "Selecting Tools for Basic Analysis" section of the DDBJ pipeline. 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". On the left, there's a sidebar with sections for "ACCOUNT" (login ID: nakazato, Logout, Change password) and "ANALYSIS" (Data entry: DRA Start, FTP upload, HTTP upload, DRA Import, Preprocessing Start, step-1, Preprocessing, Mapping / de novo Assembly, step-2, Workflow: Genome (SNP/Short Indel), RNA-seq (tag count), ChIP-seq). Below that is a "JOB STATUS" section showing step1: Preprocessing, step1: Mapping, step1: de novo Assembly, and step2-All status. A "HELP" link is also present.

The main content area is titled "Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE". It features two tabs: "Reference Genome Mapping" (selected) and "de novo Assembly".

**Reference Genome Mapping:** This section displays a table comparing tools based on input data, evaluation, analysis, and output format.

Tool	Help	Version	Input data		Evaluation		Analysis		Output format			
			Base space	Color space	Paired-end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed
Bowtie 1.0.11	http://bowtie-bio.sourceforge.net/bowtie1/index.shtml#	0.8.1	<input type="checkbox"/>	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>				
Bowtie 2.0.7	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml#	0.12.7	<input type="checkbox"/>	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>				
TopHat 2.0.11	http://tophat.cbcb.umd.edu/	1.0.11	<input type="checkbox"/>	<input checked="" type="checkbox"/>				<input checked="" type="checkbox"/>				
Bowtie2 2.2.8	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml#	2.2.8	<input type="checkbox"/>	<input checked="" type="checkbox"/>								
TopHat2 2.1.0	http://tophat.cbcb.umd.edu/	2.1.0	<input type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>					

A note in the table states: "For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1."

**de novo Assembly:** This section displays a table comparing tools based on input data, evaluation, analysis, and comment.

Tool	Help	Version	Base space	Color space	Paired-end	M5S(WGS)	Comment
SOAPdenovo 2.04-240	http://soap.genomics.org.cn/	2.04-240	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		
ABYSS 2.0.2	http://abyss.broadinstitute.org/	1.3.2	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		Maximum K-mer value is 64.

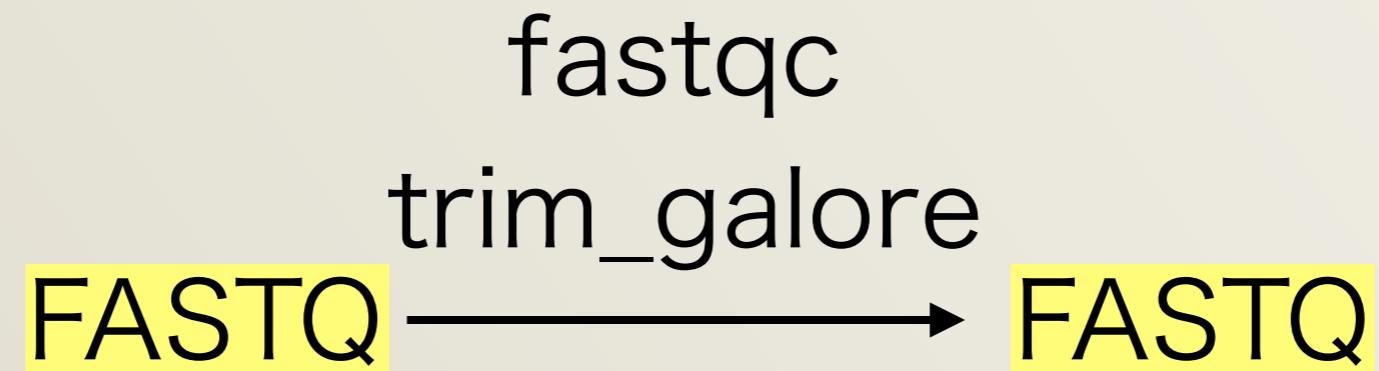
DDBJ Read Annotation Pipeline

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

※ 要利用申請

クオリティチェック・  
トリミング

# クオリティチェック・トリミングの流れ



# 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/
```

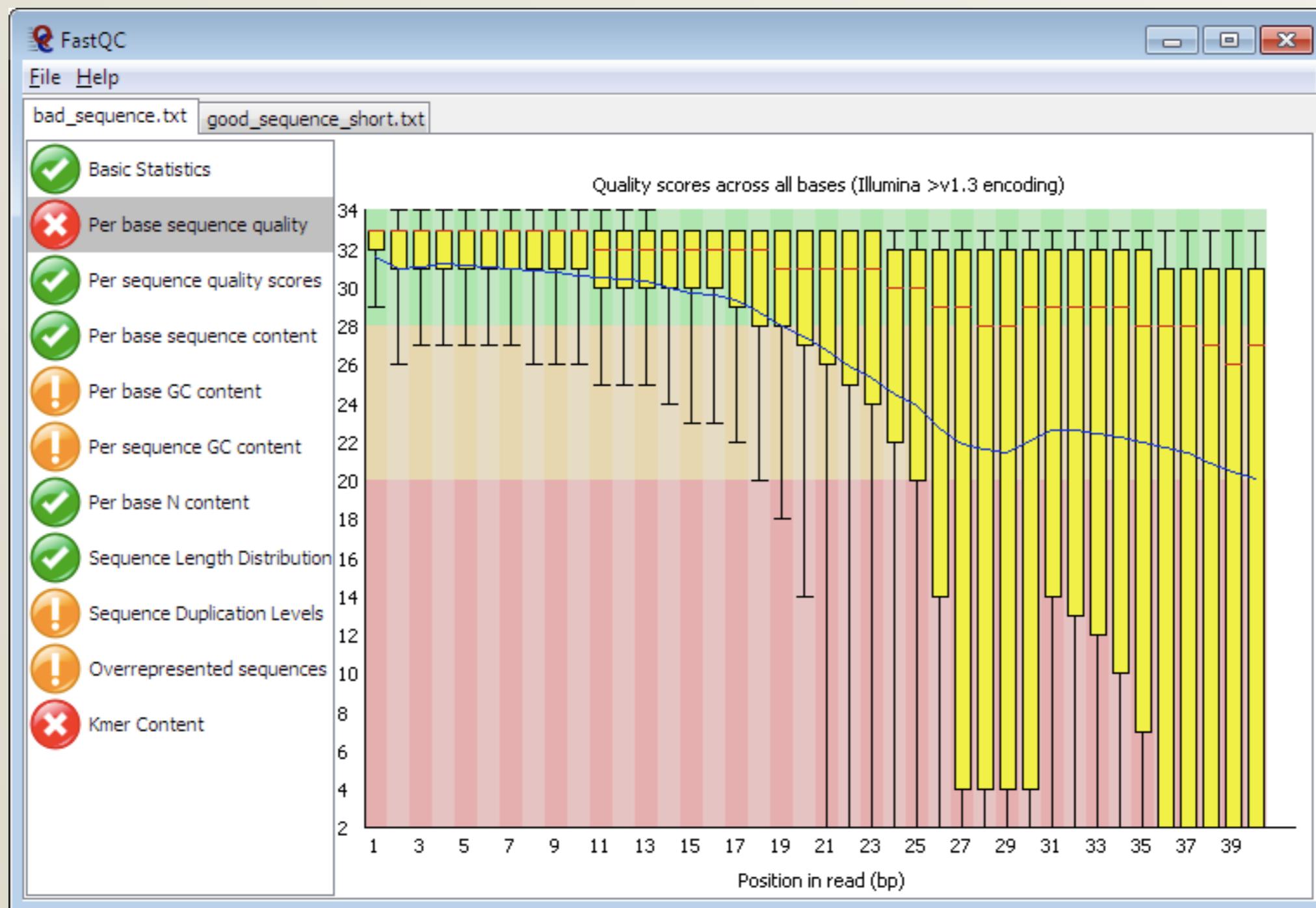
コマンド名 pair-end Illuminaデータ fastqcもかける 出力先

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



# 発現解析 (mapping)

# マッピングの流れ

to phat2  
FASTQ → bam

FASTA



# コマンド例

## tophat2

```
$ tophat -p 2 -G annotation.gtf -o results/
```

コマンド名 プロセス数 アノテーションファイル 出力先

```
Hsapieins.genome.fasta DRR1234567.trimmed.fastq
```

マップ先（ゲノム等） マップするリードファイル

## 形式変換

```
$ samtools view -h DRR1234567.bam -o DRR1234567.sam
```

コマンド名 変換前ファイル

出力先ファイル

# sam/bamフォーマット

(Sequence Alignment/Map Format)

SRR445820.39542705	0	chr17	1	0	4M1T31M *	0	0	AAAGCTTCTCACCTGTCCTGCATAGATAATTGCA	?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	CGCCCATGTCCAGCTTAACCTGCATCCCTAGAAAGTG	IIIIIIIIIIIIHH
SRR445820.32725447	16	chr17	213	31	36M *	0	0	TAGAAAGTGAAGGCACCGCCCCAAAGACACGCCCATGT	CGCGGGGDGGGBGA
SRR445820.43349427	0	chr17	??1	??1	36M *	0	0	△△GGG△△CCC△△AGG△△ACCGCCC△△TTA	TTTTTTTTTTTTTTT

↑ 各リードの名前  
↑ mapされた  
染色体/scaffold  
↑ mapされた場所  
(何塩基目)  
↑ 各リードの配列  
↑ 各リードの  
クオリティ

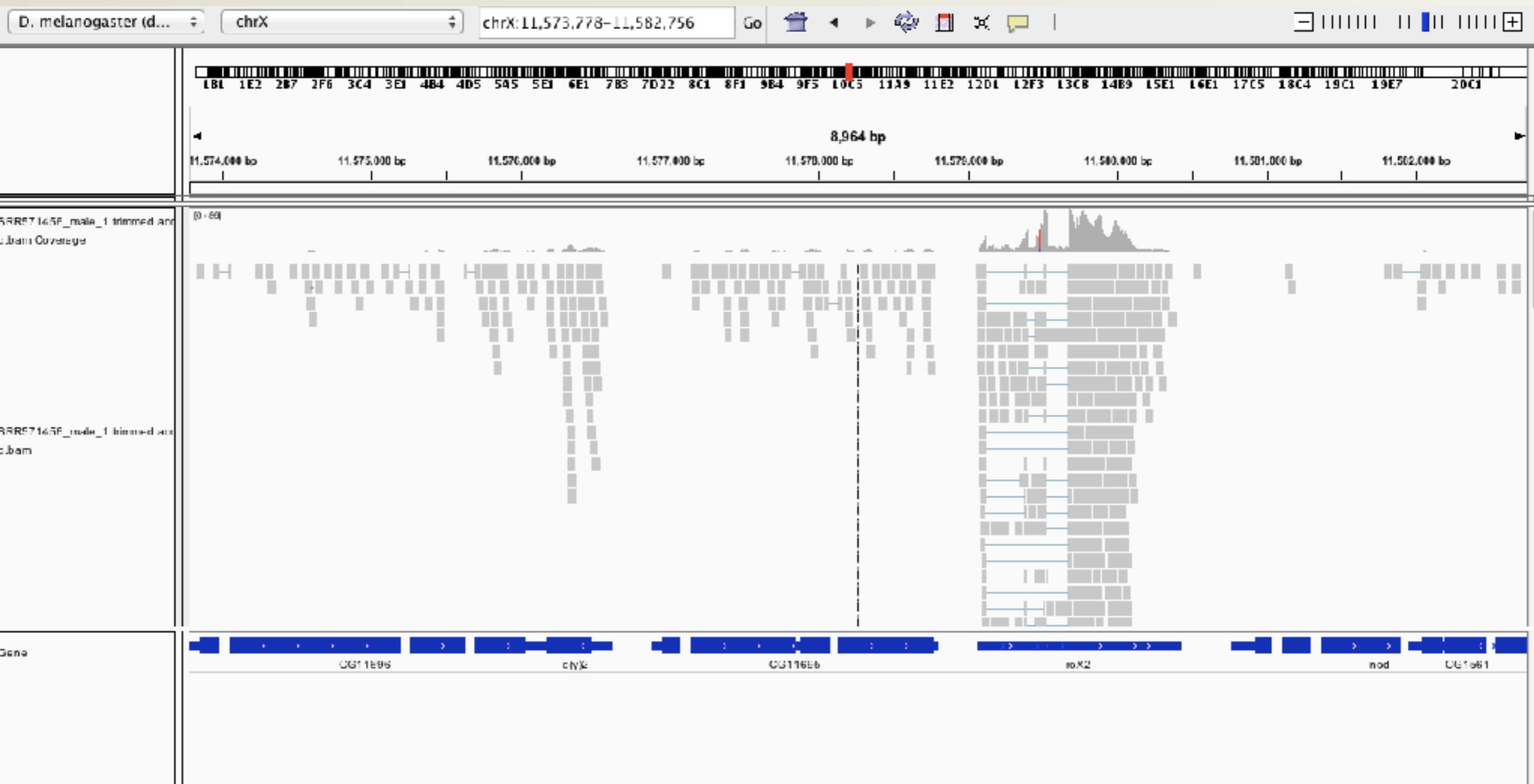
※ その他、マッピングの状況など

※ 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)

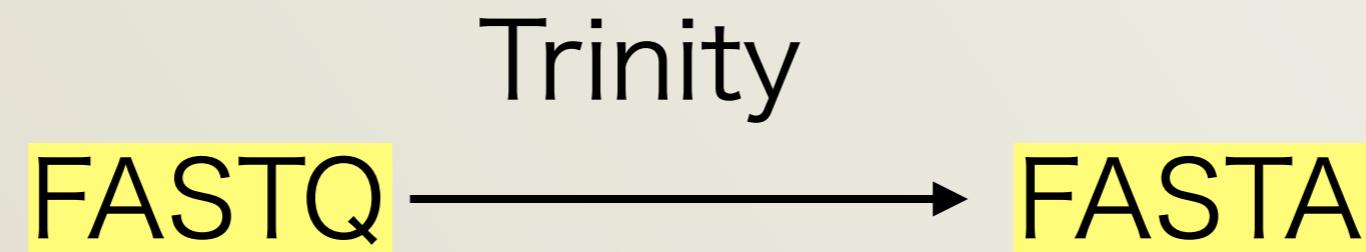
(sam だとデータサイズが非常に大きくなるのでbamにして圧縮

# IGVによる可視化



# 発現解析 (de novo)

# de novo assemble (発現データ) の流れ



# コマンド例

## Trinity

```
$ Trinity --seqType fq
```

コマンド名 ファイル形式指定

```
--left SRR1234567.R1.fastq --right SRR1234567.R2.fastq
```

対象ファイル・その1

対象ファイル・その2

```
--max_memory 24G --CPU 16
```

利用メモリ

プロセス数

# その後の発現解析

# 発現量解析：マッピング後の場合

- FPKMの計算
  - Fragments Per Kilobase of exon per Million mapped fragments
  - mapされたうちのその遺伝子にはりついたフラグメント（リード）の量
  - 「何本」はりついたか数えるとはりつけた遺伝子の長さに依存するので長さで正規化していると思えばよい

```
$ cufflinks          ← コマンド  
-p 8                ← プロセス数（同時に計算する数）  
-g Homo_sapiens/.../genes.gtf    ← 遺伝子(exon) の情報  
tophat/.../ERR266335_P0.bam      ← 対応させるNGSデータ  
-o tophat/.../cufflinks_results  ← 出力先
```

1	Cufflinks	transcript	12190	13639	1000	+	.	gene_id "CUF..."
1	Cufflinks	exon	12190	12227	1000	+	.	gene_id "CUFF..."
1	Cufflinks	exon	12595	12721	1000	+	.	gene_id "CUFF..."

# 発現量解析 : de novoの場合

```
$ ./align_and_estimate_abundance.pl  
--thread_count 12  
--transcripts trinity_out_dir/Trinity.fasta --seqType fq  
--left DRR1234567.R1.fastq --right DRR123456.R2.fastq  
--est_method RSEM --aln_method bowtie2  
--trinity_mode --prep_reference --output_dir rsem_outdir
```

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
TRINITY_DN0_c0_g1	TRINITY_DN0_c0_g1_i1	390.00	158.40	3.00	4.04	5.75
TRINITY_DN10000_c0_g1	TRINITY_DN10000_c0_g1_i1	1199.29	961.07	101.00	22.42	31.88
TRINITY_DN10001_c0_g1	TRINITY_DN10001_c0_g1_i1	497.00	260.68	1.20	0.98	1.39

# 遺伝子機能アノテーション

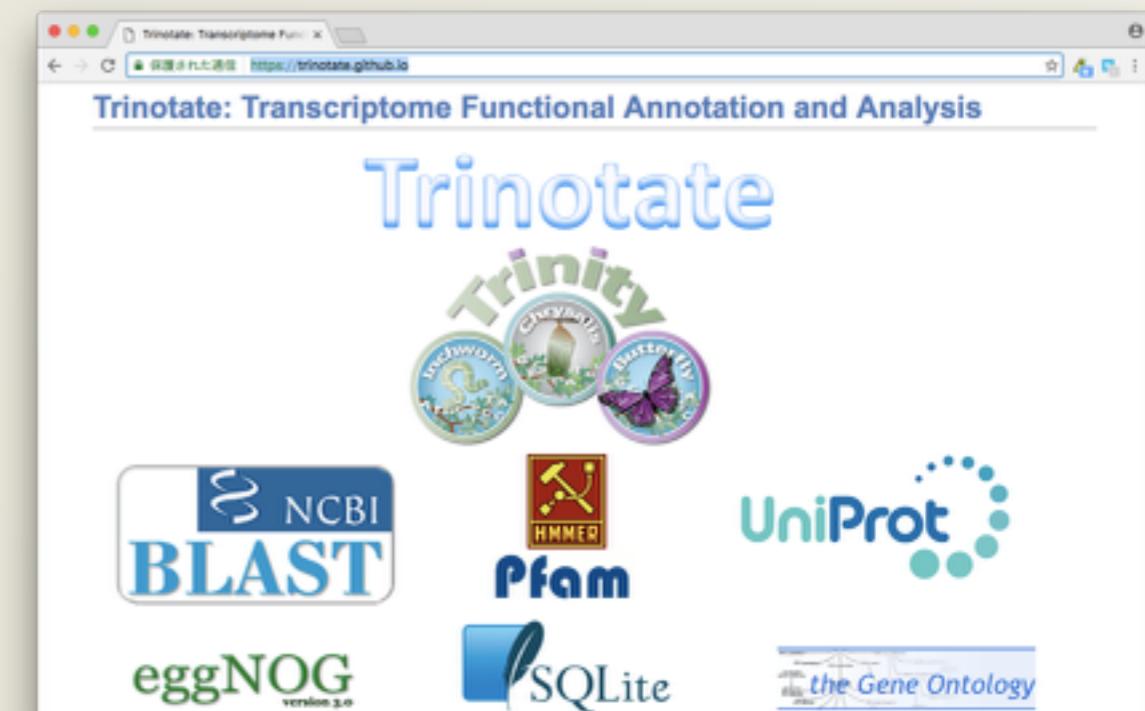
- BLASTで類似性のある遺伝子を検索

```
$ 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

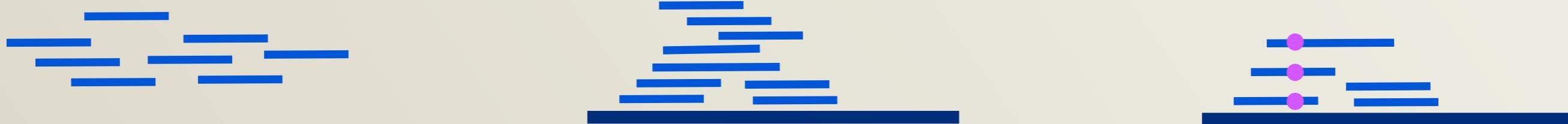
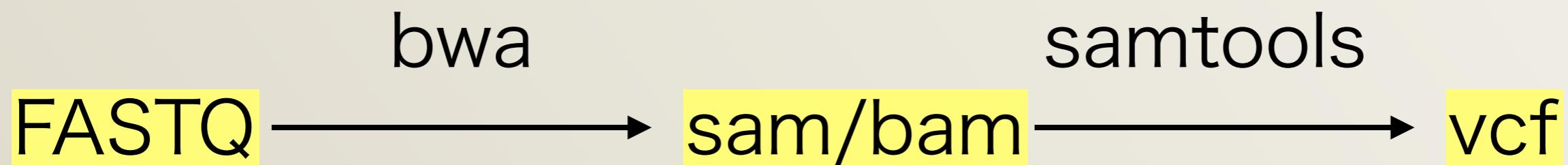
```
$ hmmsearch --cpu 8 -domtblout  
TrinotatePFAM.out Pfam-A.hmm  
transdecoder.pep > pfam.log
```



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

# SNV/Indel解析

# SNV/Indel解析の流れ



# コマンド例

bwa

細かく mapping

```
$ bwa aln -t 2 genome.fasta DRR1234567.fastq > DRR1234567.sai  
$ bwa samse genome.fasta DRR1234567.sai DRR1234567.fastq > DRR1234567.sam
```

samtools

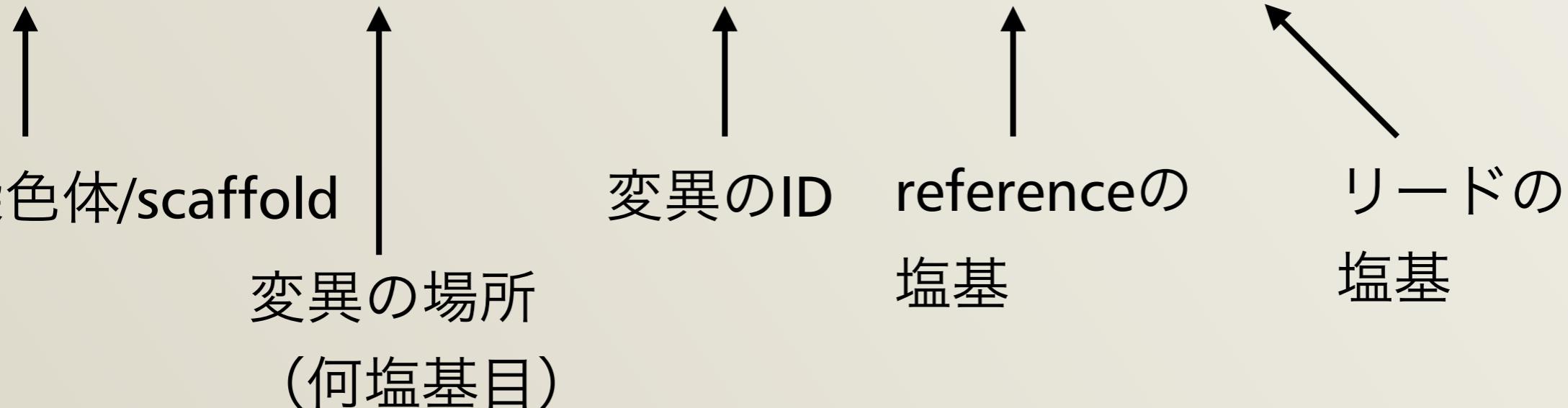
variant call

```
$ samtools mpileup -Bugf in_genome.fasta in_sorted.bam  
| ./bcftools view -bvcg -    ← 結果をbcftoolsに渡す  
> out_raw.vcf                ← 結果をout_raw.vcfに出力
```

# 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

# 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](#)

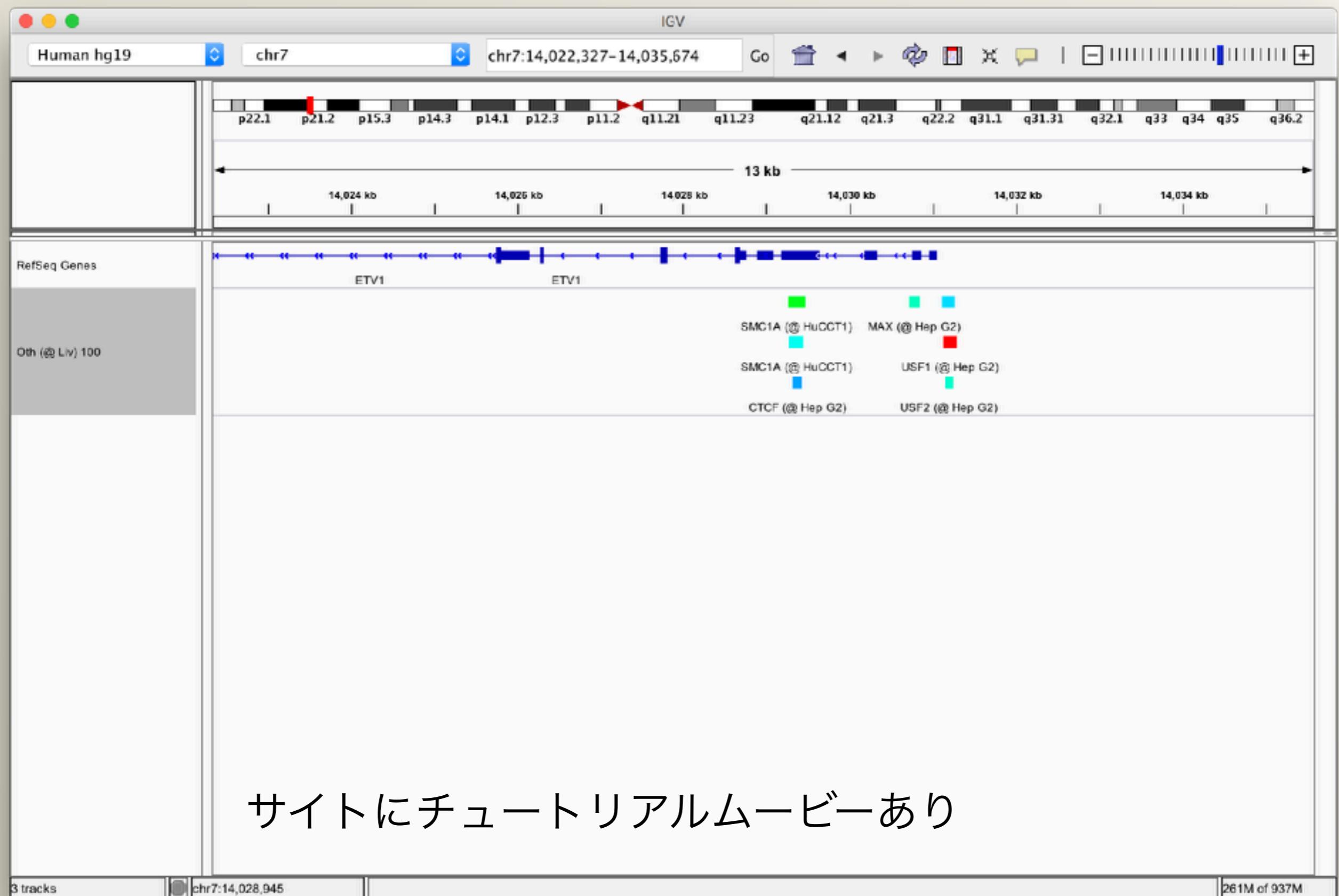
## [in silico ChIP](#)

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