

AJACS伊予

# 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 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), Research Organization of Information and Systems (ROIS)



2015/9/25

## データベース

## 検索システム



塩基配列

登録



GenBank  
EMBL  
DDBJ



BLAST



文献

登録



MEDLINE



PubMed.gov

PMC



NGSデータ

登録



**SRA**



**SRA Search**

# SRAを検索してみましょう

<http://trace.ddbj.nig.ac.jp/DRASearch/>

[illegible]

http://sra.dbcls.jp/

DBCLS SRA



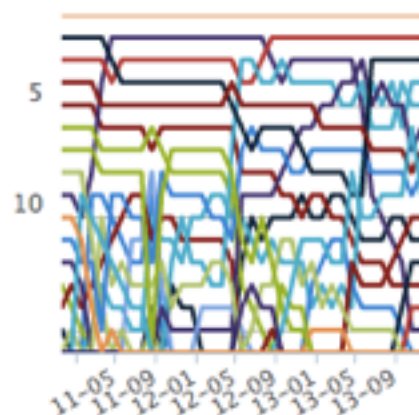
DISCOVER

Interesting & Available SRA Data

## Trends in SRA data

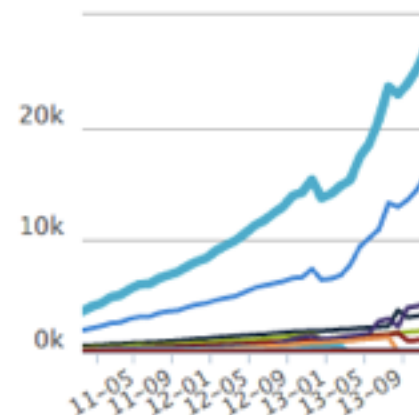
→ for more detail

### Species



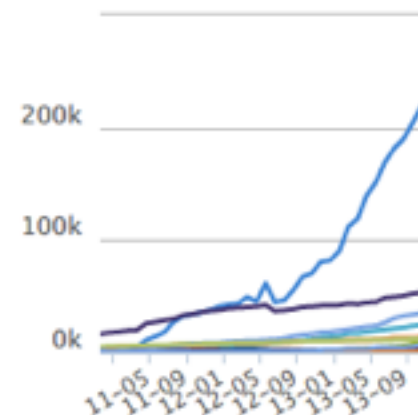
<i>Homo sapiens</i>	138367
<i>Mus musculus</i>	25944
human metagenome	13458
<i>Saccharomyces cerevisiae</i>	9519
<i>Streptococcus pneumoniae</i>	9240
<i>Staphylococcus aureus</i>	8553
<i>Drosophila melanogaster</i>	8143
<i>Danio rerio</i>	7945
human gut metagenome	7036

### Study Type



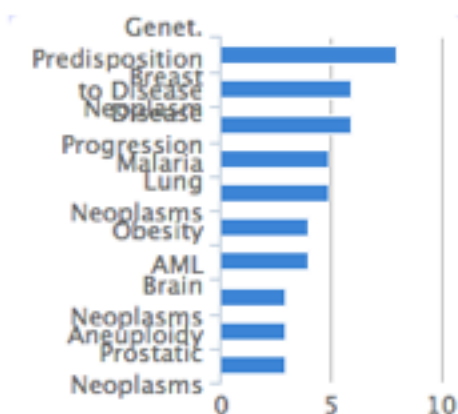
Whole Genome Sequencing	14493
Other	4106
Transcriptome Analysis	3239
Metagenomics	1885
Epigenetics	1088
Population Genomics	173
RNASeq	80
Exome Sequencing	59
Cancer Genomics	44
Pooled Clone Sequencing	28

### Platform



Illumina HiSeq 2000	219371
Illumina Genome Analyzer II	53189
454 GS FLX Titanium	35057
Illumina Genome Analyzer IIx	24676
Illumina Genome Analyzer	13340
454 GS FLX	12290
Illumina MiSeq	9111
unspecified	7285

### Disease



Genetic Predisposition to Disease	8
Breast Neoplasms	6
Disease Progression	6
Malaria	5
Lung Neoplasms	5
Obesity	4
Leukemia, Myeloid, Acute	4
Brain Neoplasms	3
Aneuploidy	3
Prostatic Neoplasms	3



# メタゲノム

環境中の塩基配列を網羅的に解読


<http://microbedb.jp/>

Search Result

MicrobeDB.jp — 微生物エン

MDB

microbedb.jp/MDB/search/?jsessionid=A21D2849260D943615FA4A1B0E5F04B5?query=hot+spring



hot spring

Search

[Sign In](#)

Environment

help

ID	Label	Definition	Synonyms
<a href="#">MEO_0000029</a>	hot spring	A spring that is produced by the emergence of geothermally-heated groundwater from the Earth's crust.	spring, hot spring, hot spring, thermal feature, thermal spring
<a href="#">MEO_0000736</a>	hot spring water		
<a href="#">MEO_0000813</a>	serpentine hot spring	A hot spring whose water venting from serpentinites.	serpentine hot springs, serpentine, antigorite, clinochrysotile, lizardite, orthochrysotile, serpentines, serpentinite, serpentinites
<a href="#">MEO_0000730</a>	alkaline hot spring	A hot spring whose water has an alkaline pH.	alkaline hot springs
<a href="#">MEO_0000729</a>	acid hot spring	A hot spring whose water has an acidic pH.	acid hot springs, acidic hot spring, acidic hot springs
<a href="#">MEO_0000121</a>	calcite hot spring		calcite hot springs

Strain

help

Number	Name	Isolated From	Temperature	Application	Related Taxonomy name by		
					NCBI	sequence	DB link
<a href="#">JCM 10698</a>	<a href="#">Caldimonas manganoxidans</a> Takeda et al. 2002	Hot spring water, Matsue, Japan	45	Production of thermostable poly(3-hydroxybutyrate) depolymerase	<a href="#">Caldimonas manganoxidans</a>		<a href="#">Caldimonas manganoxidans</a>
<a href="#">JCM 10600</a>	<a href="#">Acidisphaera rubrifaciens</a> Hiraishi et al. 2000	Hot spring water, Hakone, Kanagawa Pref., Japan	30	Production of bacteriochlorophyll a and carotenoids	<a href="#">Acidisphaera rubrifaciens</a>		

データベース検索

Download

fastq

自分で

Quality check

fastqc

fastq

*de novo* assemble

mapping

bowtie  
tophat2  
bwa  
(Sailfish)

sam/bam

ゲノム

発現量解析

cufflinks/  
cuffdiff

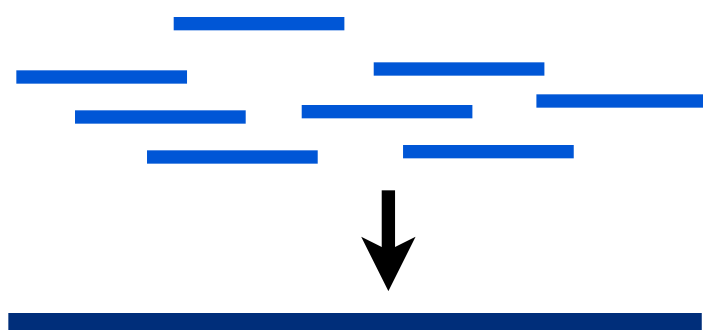
SNP検出

samtools  
GATK

vcf



いずれも IGV で可視化



# 解析コマンド（例）

※ 実際にやるならその時に誰かのマネをすればよい  
あくまでもイメージをつかむだけ

## Quality Check

```
fastqc -o SRR067385.qc -f fastq SRR067385.fastq
```

コマンド名

出力ファイル名

ファイル形式

対象ファイル名

## Mapping

```
bowtie2 -x hg19. -U SRR1294107.fastq > SRR1294107.sam
```

コマンド名

マップ先

マッピング対象

結果ファイル

## 形式変換

```
samtools view -Sb SRR1294107.sam -o SRR1294107.bam
```

## SNP analysis

```
samtools mpileup -B -g in_genome.fasta in_sorted.bam |  
./bcftools view -bvcg - > out_raw.vcf
```

# FASTQ データ

```
@DRR001107.1 GEZQ5F001EEA7F length=77
GCAACATTCAACACATATGTGTTGAATGTTGCACGACGGNGTG...
+DRR001107.1 GEZQ5F001EEA7F length=77
C@BBBECCECDBBBAAAAAA<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	4行目： シーケンスクオリティ
数十億	
	+

メタデータ = 実験情報  
プロジェクト名、生物種、シーケンサー、...



# sam/bam データ

(Sequence Alignment/Map Format)

SRR445820.39542705	0	chr17	1	0	4M1I31M	*	0	0	AAAGCTTCTCACCTGTTCTGTCATAGATAATTGCA	?5>7(+2;'1..'+'<
SRR445820.29211975	16	chr17	88	42	36M	*	0	0	CCACGACCAACTCCCTGGGCCTGGCACCAGGGAGCT	#####BDB8DACC
SRR445820.7156374	16	chr17	138	42	36M	*	0	0	CCAGCGAATACCTGCATCCCTAGAAGTGAAGCCACC	BBB=:;BBEABFBFB
SRR445820.22614977	0	chr17	156	30	36M	*	0	0	CCTAGAAGTGAAGCCACCGCCCAAAGACACGCCCAT	GGGD>DBB3D=?=?=<
SRR445820.19222309	0	chr17	185	42	36M	*	0	0	CGCCCATGTCCAGCTTAACCTGCATCCCTAGAAGTG	IIIIIIIIIIIIHH
SRR445820.32725447	16	chr17	213	31	36M	*	0	0	TAGAAGTGAAGGCACCGCCCAAAGACACGCCCATGT	CGCGGGGDGGGBGAA
SRR445820.43349427	0	chr17	221	31	36M	*	0	0	AAGGCACCGCCCAAAGACACCGCCCATGTCCAGCTTA	IIIIIIIIIIIIII

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

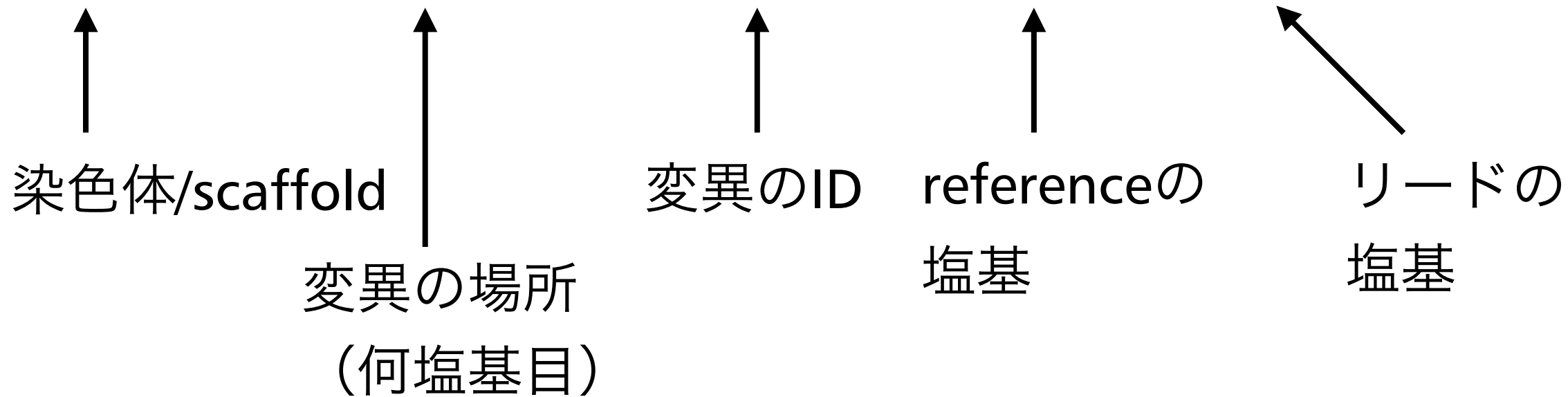
(人間が読めるデータからコンピューター用に変換)

(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)

# 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

# ブラウザでNGSデータ解析

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

The screenshot shows the DDBJ Read Annotation Pipeline website. The browser address bar displays <https://p.ddbj.nig.ac.jp/pipeline/Login.do>. The page features the DDBJ logo and the title "DDBJ Read Annotation Pipeline". There are language selection buttons for "English" and "Japanese". A description states: "DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data."

The "LOGIN" section includes links for "New account" and "Login as 'guest'", a "User ID:" field, a "Password:" field, and a "Login" button. Below the login fields is a "Check current jobs" button with a note: "\* by the guest account."

A "Pipeline Flow" diagram illustrates the workflow: "USER" uploads "Reads, metadata" via "File Upload" to the "DDBJ Read Archive". From the archive, data flows to "Basic Analysis" (containing "Mapping" and "de novo assembly") and then to "High-level Analysis" (containing "SNP/Indel detection", "RNA-seq", and "Contig annotation"). The final output is "Pipeline".

On the left, there is a "Tweets" section with a tweet from @pipeline\_info dated 3 Sep, stating: "DDBJ Pipeline has recovered from the system trouble. Please re-submit the jobs if you find something wrong in your results."

On the right, there are sections for "Manual & tutorial" and "Data submission for analyzed results and sequenced data".

**Manual & tutorial**

- [Japanese Tutorial \(FAQ\)](#)
- [English manual](#)
- [DBCLS togotv Tutorial video 1 \(JP\) - Reference Genome Mapping](#)
- [DBCLS togotv Tutorial video 2 \(JP\) - De novo Assembly](#)
- [Tutorial : How to upload and register query files to DDBJ Pipeline \(JP\)](#)
- [Tutorial : How to run HGAP for PacBio sequence read on DDBJ Pipeline \(JP\)](#)

**Data submission for analyzed results and sequenced data**

- [DRA](#) : NGS raw sequence reads
- [DDBJ-INSDC](#) : Annotated nucleotide sequences

**Citation**

- Nagasaki, H. et al., "DDBJ Read Annotation Pipeline: A cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data, DNA Res, 20:383-390, 2013.

# 実際の解析は以下を参考に...



## 講習会ページ MotDB

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



## 動画配信ページ 統合TV

<http://togotv.dbcls.jp/>

※ YouTube にもあります



# より詳しい解析は...

## 人材育成カリキュラム (NGS) 速習コース



バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | 速習コース **NEW**

2014年9月1-12日にJST-NBDCと東大農アグリバイオ主催で「バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)速習コース」が開催されました。2014年12月に速習コースの動画が統合TVおよびYoutubeから公開されました。ハッシュタグは#AJACS。

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)関連:

- [NBDCの速習コース案内サイト](#) (速習コース主催機関)
- [HPCIの速習コース受講申込受付サイト](#) (速習コース共催機関)
  - 講義日程のPDF ([20140901-12 bioinformatics intensive course program ver.1.pdf](#))
- カリキュラムを策定した[NBDC運営委員会人材育成分科会](#)
- 「NBDCで実施した調査」の[バイオインフォマティクス人材育成のためのカリキュラム](#)
  - 「バイオインフォマティクス人材育成カリキュラム (次世代シーケンサ)」のPDF ([generation-sequencer.pdf](#))
  - 「カリキュラムで習得できる技能」のPDF ([learning-skills.pdf](#))
  - 「カリキュラム フロー図」のPDF ([flow-diagram.pdf](#))
- 速習コースアンケート用紙  
記名回答者のみご利用ください。無記名の方は紙ベースのもので提出お願いします。メールの添付でkadota@iu.a.u-tokyo.ac.jp宛てに、件名は「NGS速習コースアンケート」をお願いします。
  - PDF版([questionnaire\\_2014.pdf](#))
  - Microsoft Word版([questionnaire\\_2014.docx](#))
  - Microsoft Word 97-2003版([questionnaire\\_2014.doc](#))

計算機環境構築(Linux系):

Linux環境構築手順は大まかに3つの手順からなります。最低限、VirtualBoxのインストールができていればOKです。VirtualBoxのバージョンは2014年9月の実習では4.3.12以前のものを想定しています。イメージファイルは、初日に速習コース会場にて、USBメモリで持ち込みPCにコピーすることも可能です。また、何らかの理由により持ち込みPCにインストールできなかったとしても主催機関であるアグリバイオインフォマティクス所有のインストール済みのノートPC(60台程度あり)をすぐに貸与可能です。

インストール手順(Windows用): [install\\_NGSsokushu\\_windows.pdf](#) (2014.08.22版; 約6MB; 西岡 輔 氏作成)

インストール手順(Macintosh用): [install\\_NGSsokushu\\_macintosh.pdf](#) (2014.08.26版; 約3MB; 孫 建強 氏作成)

トラブルシューティング(ファイルシステム関連): [troubleshooting\\_NTFS.pdf](#) (2014.08.14版; 約1MB; 西岡 輔 氏作成)

トラブルシューティング(BIOS関連): [troubleshooting\\_BIOS.pdf](#) (2014.08.15版; 約2MB; 西岡 輔 氏作成)

[トップページへ](#)

2週間分の  
基礎から実践までの講義

by NBDC+東大アグリバイオ

<http://bit.ly/ngs2014>

※ 統合TVに録画があります



# 初心者向け次世代DRY解析本発刊

次世代シーケンサー  
**DRY**  
解析教本 ● 清水厚志／坊農秀雅



- 10月8日 日本癌学会学術総会にて先行発売
- 10月14日 日本人類遺伝学会大会にて先行発売
- 10月15日 全国発売

## これからは生命科学者がデータ解析