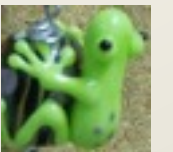


次世代シーケンサー（NGS）解析・ 実践編：目的別データ解析

仲里 猛留

NAKAZATO, Takeru

@chalkless

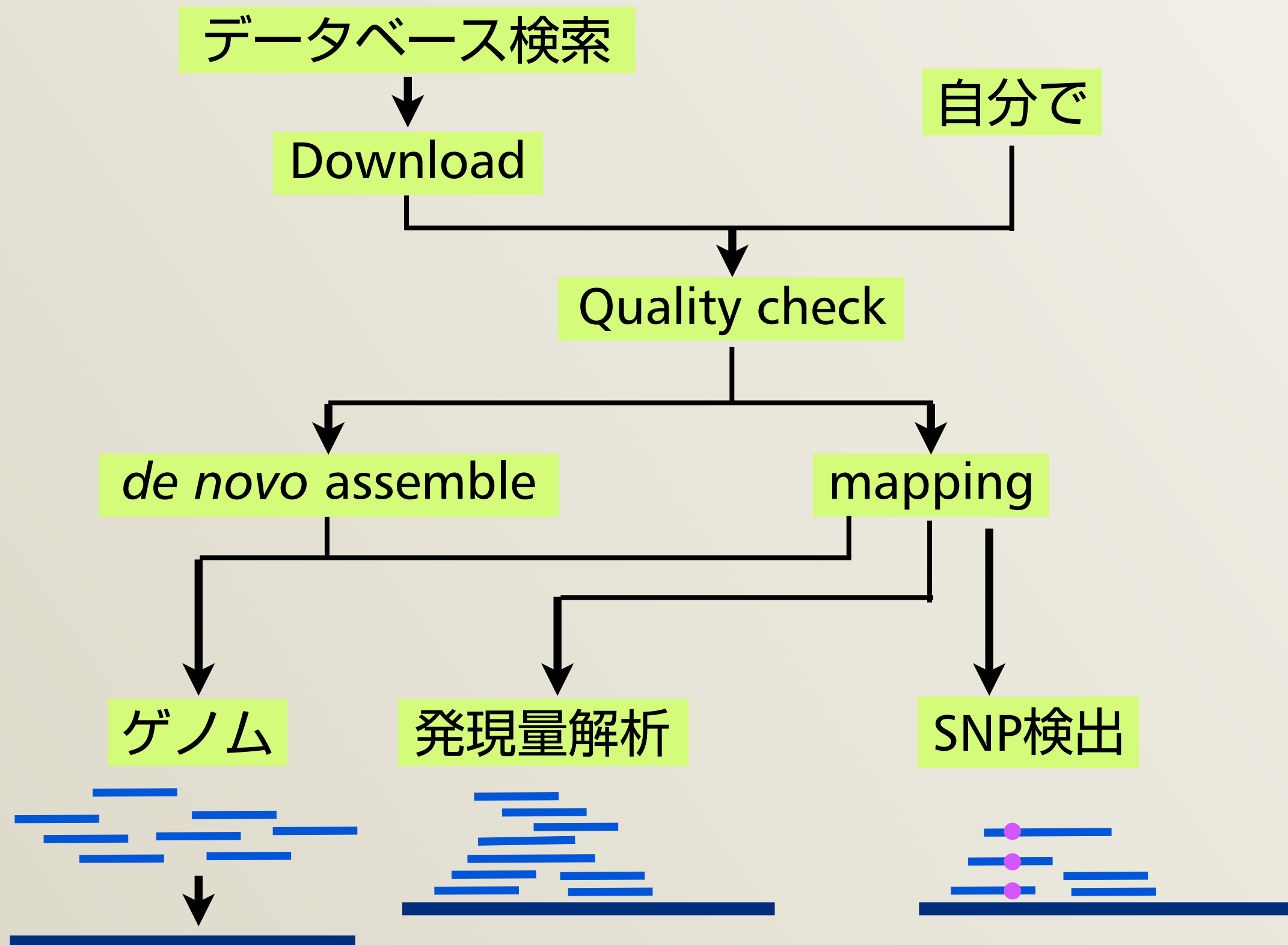


情報・システム研究機構 データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター

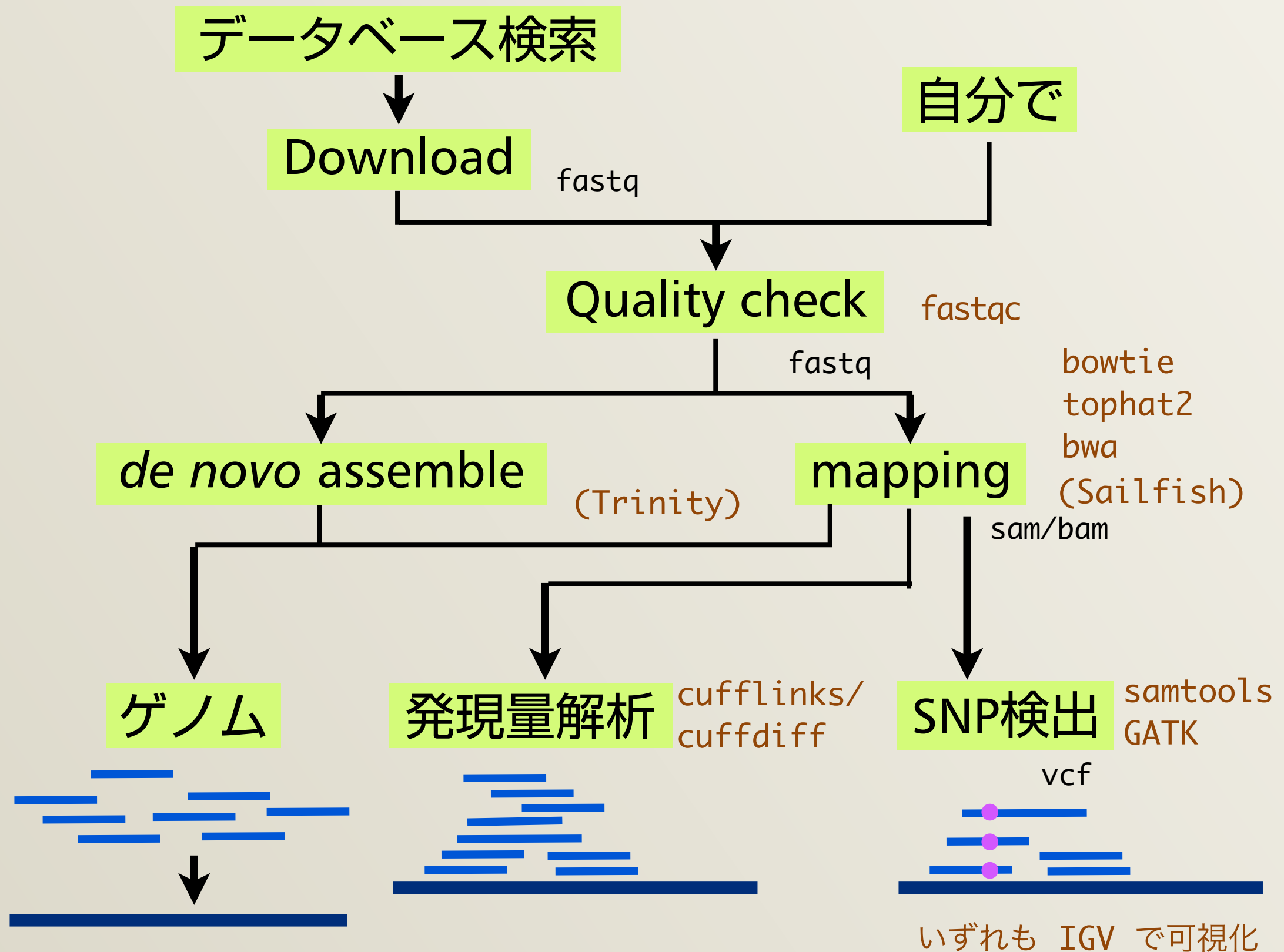
Database Center for Life Science (DBCLS),
Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)

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）

amazonstudent 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

単行本
¥ 5,832

今すぐお読みいただけます: **無料アプリ**

¥ 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)

amazonstudent

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)。



この画像を表示

NGSデータ解析を丁寧に解説。Kindle版あり

詳細な解析をひととおり知りたい

The screenshot shows the NBDC (National Bioscience Database Center) website. The header includes the NBDC logo and the text "バイオサイエンスデータベースセンター". Below the header is a navigation menu with links like "ホーム", "NBDCについて", "研究開発", "公募情報", "採用情報", "イベント", "人材支援", "アクセス", and "リンク". The main content area is titled "H28年度 NGSハンズオン講習会カリキュラム". It includes a sub-header "H28年度日程・講義資料・動画等" and a link "カリキュラム (PDF: 72KB)". A table lists the workshop schedule and topics.

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(統合TV)
7月19日 (火)	10:30-18:15	はじめに (講習会参加者必読) PC環境の構築	Bio-Linux8とRのインストール状況確認	・ Bio-Linux8 (第2部および3部で利用するovaファイル) の導入確認 ・ 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 ・ 講師指定の事前学習内容の再確認 ・ 講習会期間中に貸与されるノートPCを用いた各種動作確認	主催・共催機関	講義資料 (PDF: 4MB)
7月20日 (水)	10:30-18:15	第1部 統計解析 (農学生命情報科学特論I)	ゲノム解析、塩基配列解析	・ NGS解析手段、ウェブツール (DDBJ Pipeline) との連携 ・ k-mer解析 (k個の連続塩基に基づく各種解析) の基礎と応用 ・ 塩基ごとの出現頻度解析 (k=1)、2連続塩基の出現頻度解析 (k=2) ・ 塩基配列解析を行うための基本スキルの復習	門田 幸二 (東京大学)	講義資料 (PDF: 7.3MB) 解析データ (ZIP: 2.2MB) 統合TV

NGSハンズオン講習会

JST-NBDC + 東大アグリバイオ

このところ毎年やっています
統合TVで録画・公開済

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

※ 「NGS 講習会」 でググれ

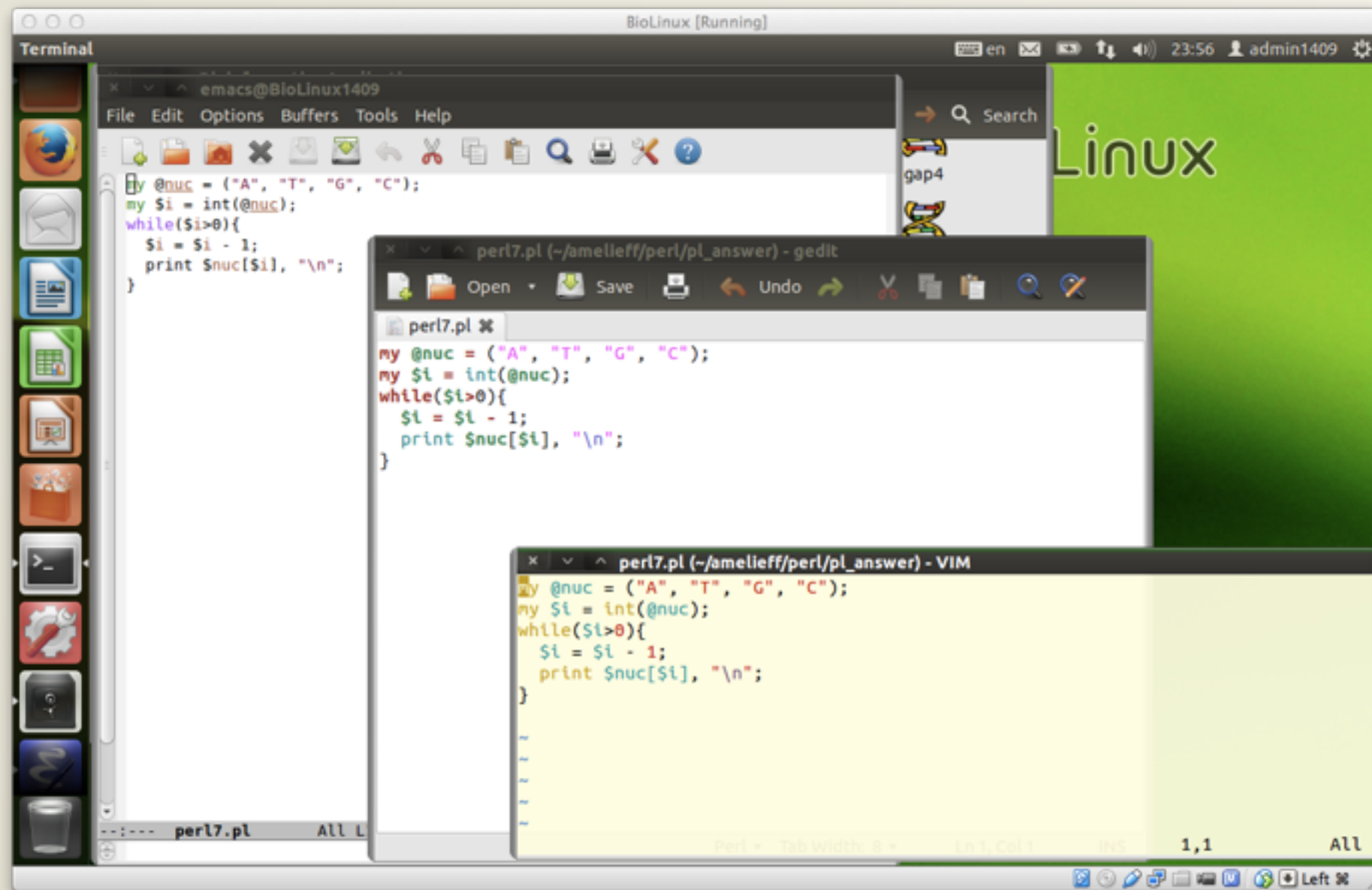
解析について詳細な情報を探したい



門田さん (東大アグリバイオ)

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

ツールをひとつと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

解析環境・ウェブベース

The screenshot shows the DDBJ Read Annotation Pipeline web interface. The page title is "Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE". The interface includes a navigation bar with steps: Select Query Files, Select Tools (current), Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. A sidebar on the left contains sections for ACCOUNT (login ID [nakazato], Logout, Change password), ANALYSIS (Data setup, DRA Start, FTP upload, HTTP upload, DRA Import, Preprocessing Start, step-1, step-2, Workflow), JOB STATUS (step1, step2), and HELP.

Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

Reference Genome Mapping

	Tool	Help	Version	Input data			Evaluation			Analysis		Output format			Comment
				Base space	Color space	Paired end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	
<input type="checkbox"/>	bowtie		0.6.1	✓		✓	✓	✓	✓				✓		
<input type="checkbox"/>	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓			✓		
<input type="checkbox"/>	TopHat		1.0.11	✓		✓	✓	✓	✓				✓		
<input type="checkbox"/>	Bowtie2		2.2.6	✓	✓	✓	✓	✓	✓	✓			✓	For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.	
<input type="checkbox"/>	TopHat2		2.1.0	✓		✓	✓	✓	✓				✓		

☐ **de novo Assembly**
Total limit = 22 Gbp

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
<input type="checkbox"/>	SOAPdenovo		2.04-r240	✓		✓		
<input type="checkbox"/>	ABySS		1.3.2	✓		✓		Maximum K-mer value is 64.

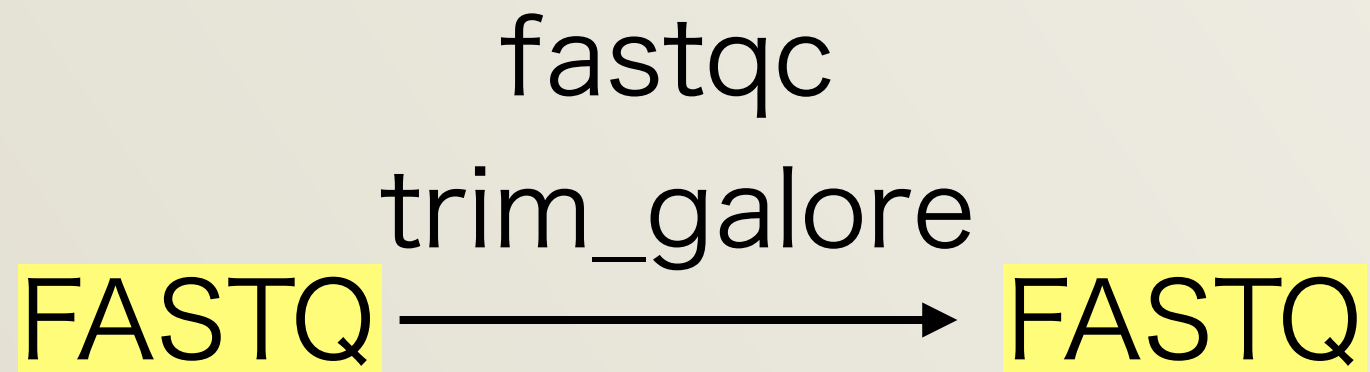
DDBJ Read Annotation Pipeline

<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@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	

[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 ee1 SLC26A6  
GACCCAAAACTGATAGGTGATGTTTCACGTAGTGGC  
CATCGCCTGATAGACGGTTTTTCGCCCTTTGACGTT  
GGAGTCCACGTTCTTTAATAGTGACTCTGAGTAAA ...
```

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

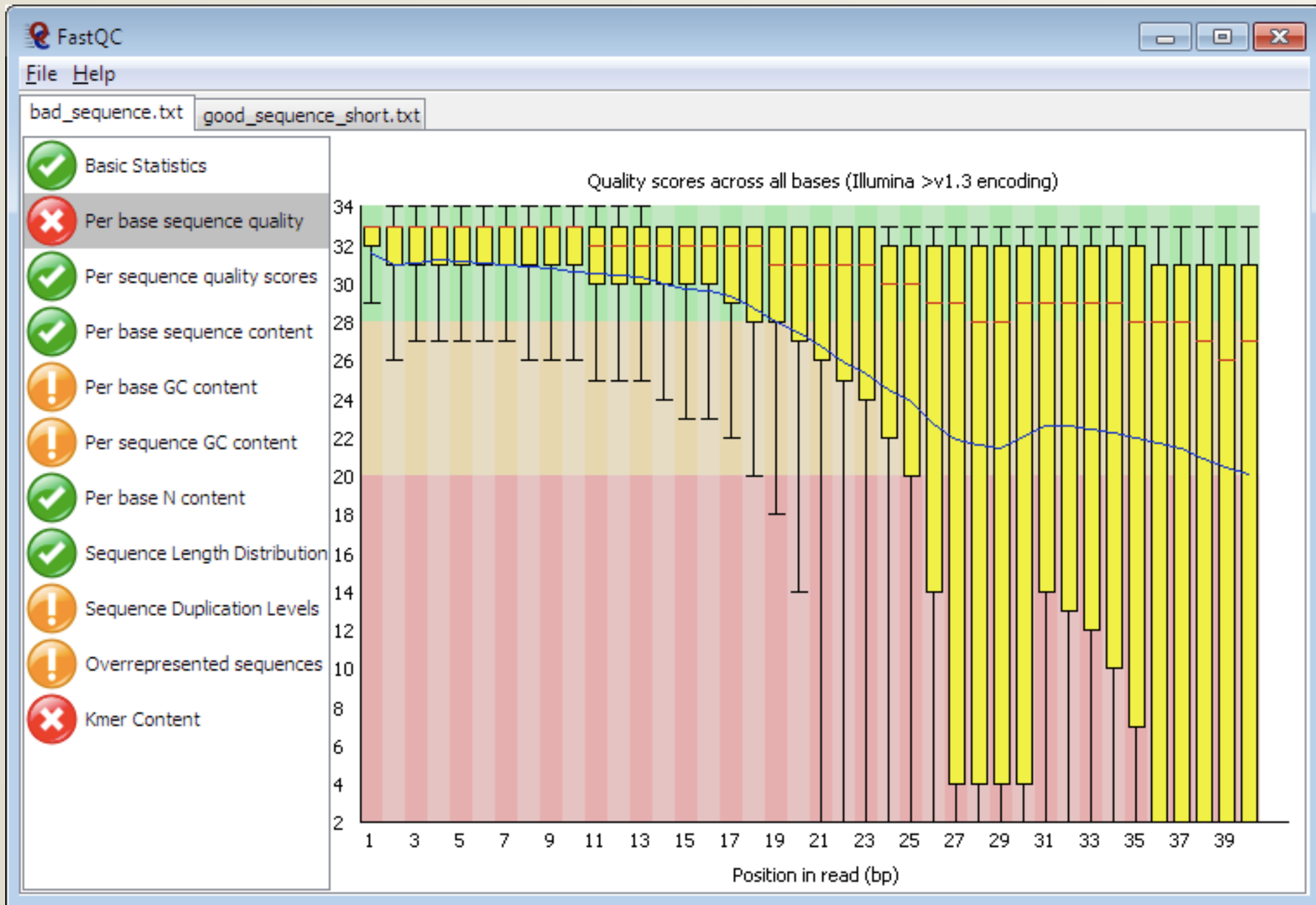
出力先

```
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

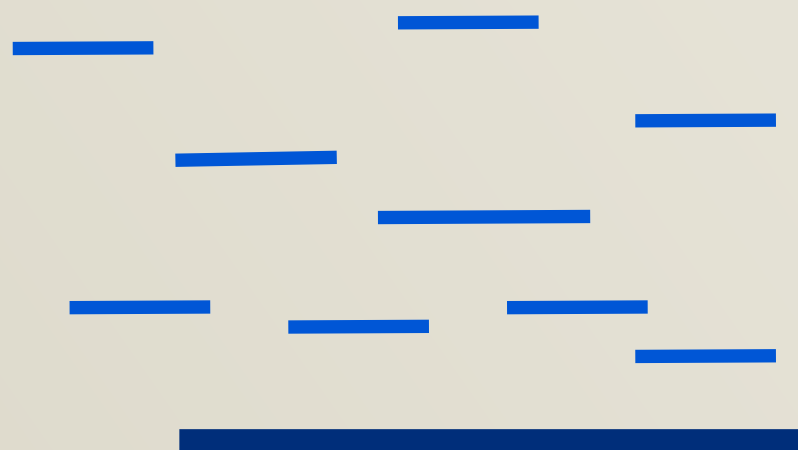


発現解析 (mapping)

マッピングの流れ



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

```
Hsapiens.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	4M1I31M	*	0	0	AAAGCTTCTCACCTGTTCTGCATAGATAATTGCA	?5>7(+2;'1..'+'<
SRR445820.29211975	16	chr17	88	42	36M	*	0	0	CCACGACCAACTCCCTGGGCCTGGCACCAGGGAGCT	#####BDB8DACCC
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

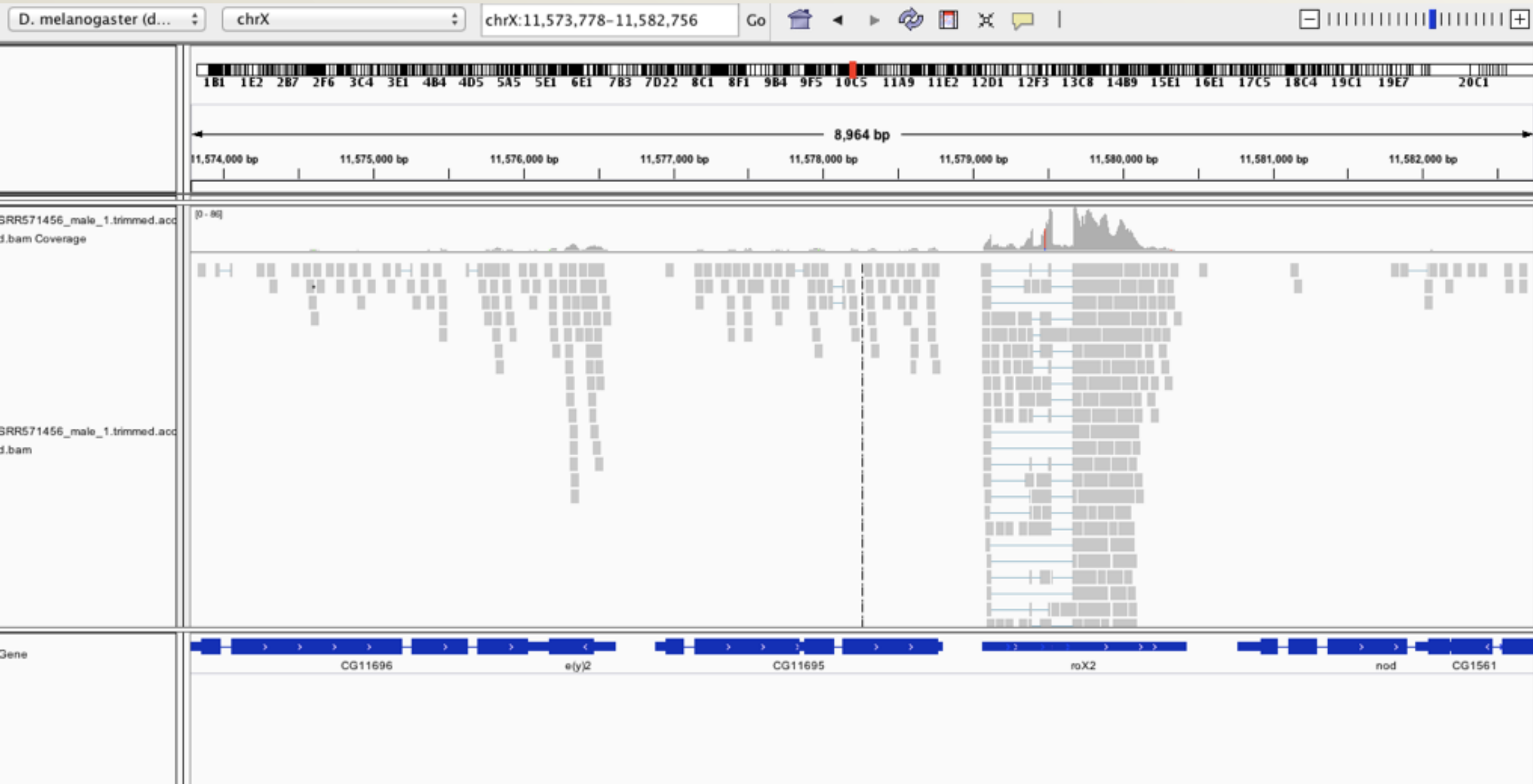
※ その他、マッピングの状況など

※ 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)

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

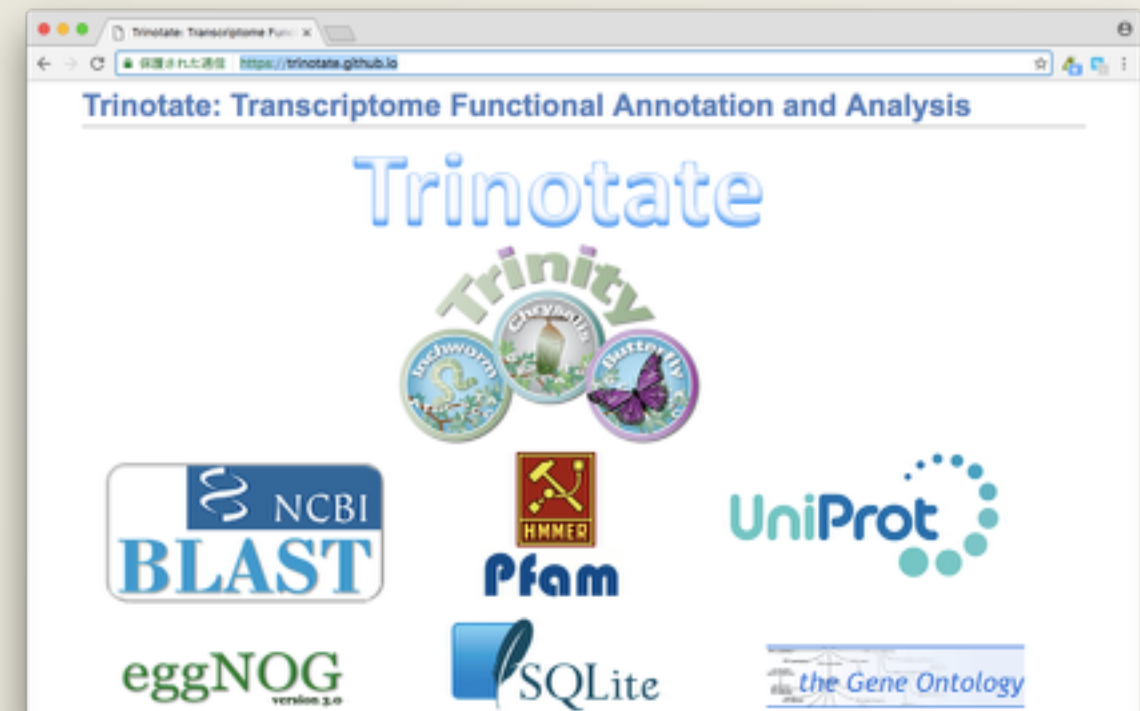
- BLASTで類似性のある遺伝子を検索

```
$ 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_9R0SI	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_9R0SI	95.270	148	7	0	3	446	26	173	2.31e-99	294

- hmmerでドメイン検索

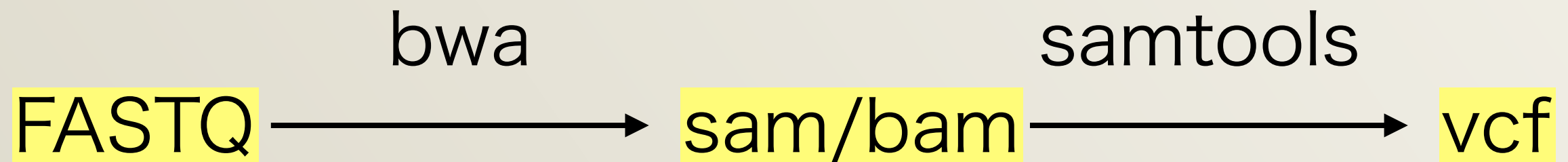
```
$ hmmscan --cpu 8 -domtblout  
TrinotatePFAM.out Pfam-A.hmm  
transdecoder.pep > pfam.log
```



<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

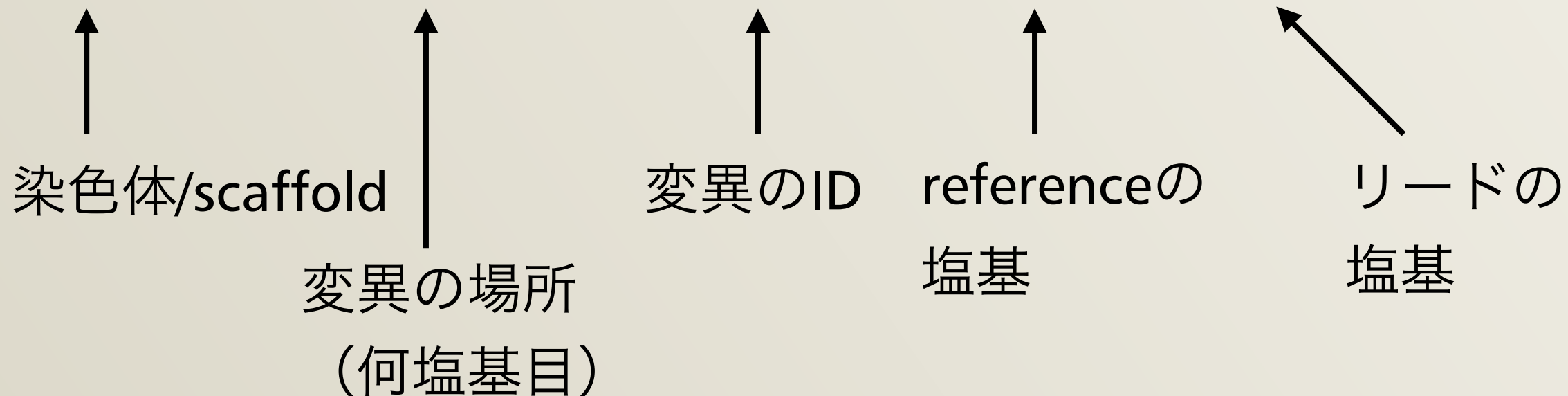
variant call

```
$ samtools mpileup -B -g 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/>

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

<http://microbedb.jp/>