

RNA-seq

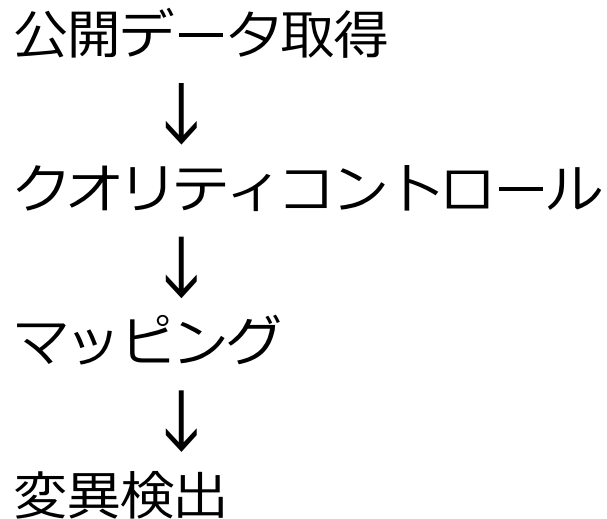
ITのチカラで研究を支援



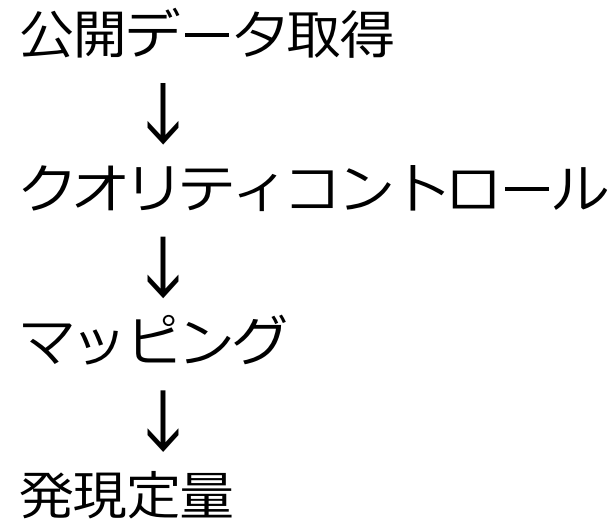
アメリエフ株式会社

本講義の内容

- Reseq解析



- RNA-seq解析



FPKMを算出します。

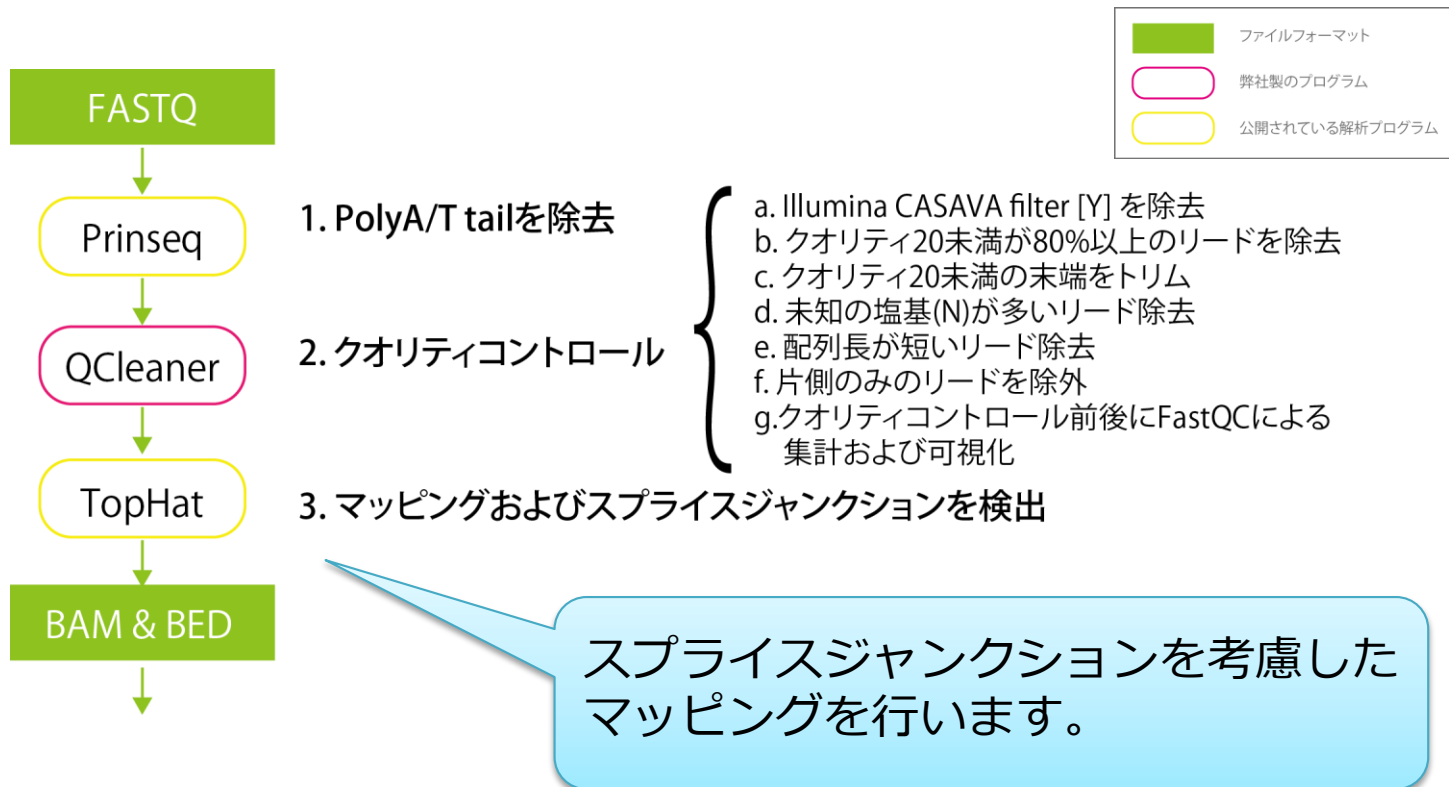
RNA-seqとは

- メッセンジャーRNA (mRNA) をキャプチャして次世代シーケンサーでシーケンシングする手法
- リファレンスがある生物種の場合：
 - 既知遺伝子にマッピングする
 - **リファレンスにマッピングして遺伝子構造を同定する**
- リファレンスがない生物種の場合：
 - アセンブリングして転写物構造を予測し、それに対してマッピングする
 - 近いゲノムのリファレンスにマッピングする

本日の内容

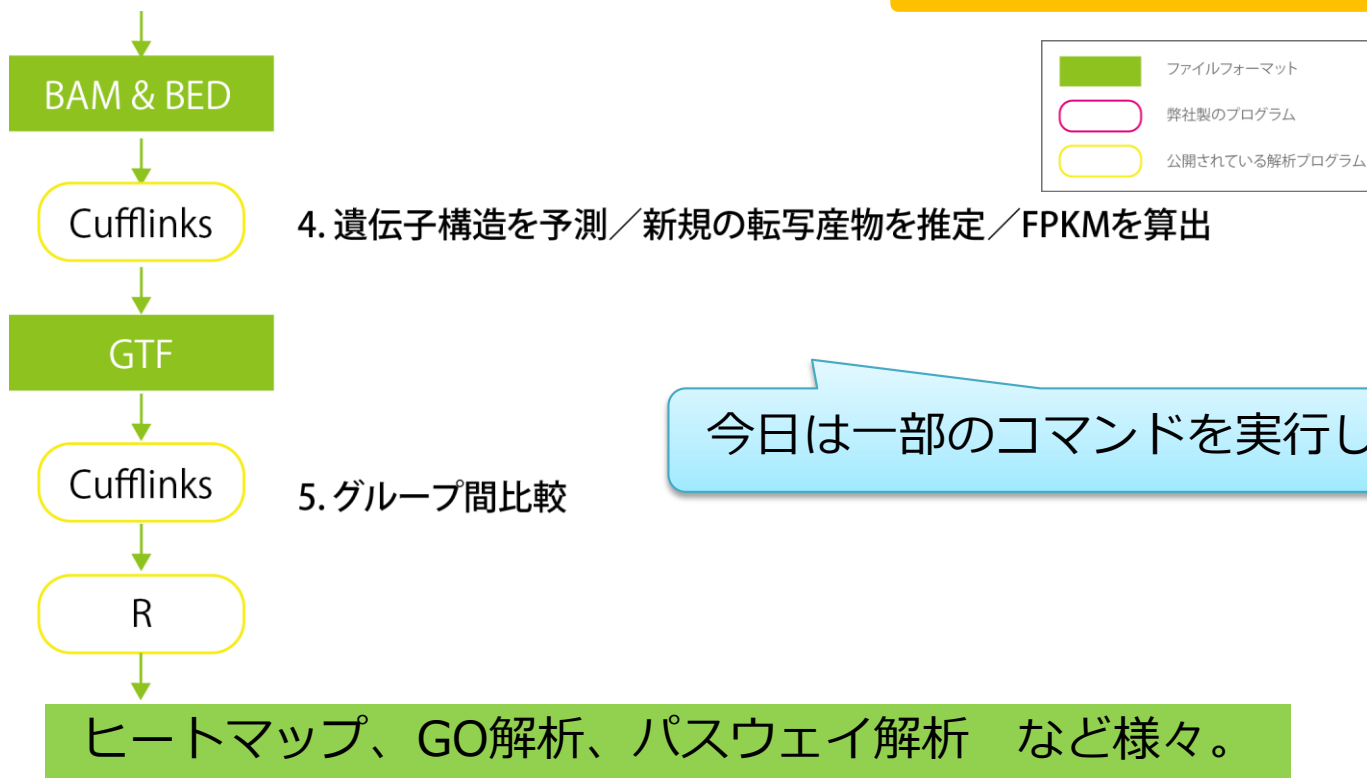
RNA-seq 解析：パイプライン

データ取得 → クオリティコントロール → マッピング → 発現定量



RNA-seq 解析：パイプライン

データ取得 → クオリティコントロール → マッピング → 発現定量



RNA-seq 解析でできること

- 発現量の定量・比較
- 新規転写物・新規スプライシングバリエーションの探索

RNA-seqがマイクロアレイと比較して優れている点

- 新規転写物や融合遺伝子が検出可
- SNV・small Indelも検出可
- プローブの設計を必要としない（非モデル生物にも対応可）

RNA-seq 解析：データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- 酵母のゲノムのリファレンス取得

ILLUMINA iGenomes



Saccharomyces cerevisiae

NCBI

build2.1

70 MB May 15 22:36

build3.1

70 MB May 15 22:36

リファレンス配列のfastaのみではなく、マッピングソフトのインデックスファイルや遺伝子情報ファイルも一緒に圧縮されて公開しています。

TRY!

RNA-seq 解析：データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- 酵母のゲノムのリファレンス配列を確認

```
$ cd /home/ユーザ名/Desktop/amelieff/Scerevisiae  
$ ll
```

```
drwxrwxr-x 2 admin1409 admin1409 4096 Jun  4 01:53 AbundantSequences  
drwxrwxr-x 2 admin1409 admin1409 4096 Apr 11 2012 Bowtie2Index  
drwxrwxr-x 4 admin1409 admin1409 4096 Mar 16 2012 BWAIndex  
drwxrwxr-x 2 admin1409 admin1409 4096 Mar 17 2012 Chromosomes  
drwxrwxr-x 2 admin1409 admin1409 4096 May  9 2013 WholeGenomeFasta
```


TRY!

RNA-seq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- 酵母のゲノムのリファレンス配列を確認

```
$ ll Bowtie2Index
```

```
genome.1.bt2  
genome.2.bt2  
genome.3.bt2  
genome.4.bt2  
genome.rev.1.bt2  
genome.rev.2.bt2
```

今回はこちらのインデックスを使用します。

RNA-seq 解析：データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得 <http://trace.ddbj.nig.ac.jp/dra/index.html>

DDBJ
DNA Data Bank of Japan

Login & Submit | Databases▼ | Japanese | Contact

Google™ Custom Search

Home | Handbook | FAQ | **Search** | Download▼ | Pipeline | About DRA

News

2014-05-13: [New DRA submission system is released.](#) less...

We have released the new DRA submission system. For major changes, please see the [slides](#) and [new handbook](#).

(6th, June, 2014)

For submissions with status "new" which had been created before 12th, May, 2014, addition or deletion of metadata objects could cause errors. It is recommended that download metadata as a tab-delimited text file and upload it into a newly created submission.

DDBJのSequence Read Archive → Search

DDBJ Seq... ng machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others. DRA is a member of the International Nucleotide Sequence Database Collaboration (INSDC) and archiving the data in a close collaboration with NCBI Sequence Read Archive (SRA) and EBI Sequence Read Archive (ERA). Please submit the trace data from conventional capillary sequencers to DDBJ Trace Archive.

RNA-seq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得

 DRASearch

Accession :

Organism :

StudyType :

CenterName :

Platform :

Keyword :

Show records

Sort by

Accessionに「SRR518891」と入力 → Search

RNA-seq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得

今回は1サンプルで実行しますが、発現比較する場合には複数サンプル必要で、replicateも多いほうが良いです。

DRASearch

Send Feedback Search

SRR518891 FASTQ SRA

ここからダウンロード

Run Detail

Alias	GSM956493_r1
Instrument model	
Date of run	
Run center	
Number of spots	9,350,778
Number of bases	1,963,663,380

Navigation

Submission	SRA055683
Study	SRP021137
Experiment	SRX157933
Sample	SRR157933.1

実験の詳細

READS (joined) quality show 10 rows << < 1 / 935078 Page > >>

```
>SRR518891.1
NACTGTTAACAAATATATAACAATTGGGATTTAGTAAAAAAAAAAAAAAAAAGGGAGGGGGCGGCTATATCCCTCTGAG
CAAAACCAAAAAAAAAATTTTCTTTCACTGTTTGATATAGTGTAAAGCGAATGACAGAAATTAAATTTCTTGGTATTGC
TCAGAGTGATATA
```

NavigationエリアのExperiment → 「SRX157933」をクリック

RNA-seq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得

Experiment Detail	
Title	GSM956493: ypd_bio1_lib1; Saccharomy
Design Description	
Organism	Saccharomyces cerevisiae
Library Description	
Name	GSM956493: ypd_bio1_lib1
Strategy	OTHER
Source	TRANSCRIPTOMIC
Selection	other
Layout	PAIRED

他にも、シーケンサのプラットフォームやリード長などの情報も記載されています。

転写産物

RNA-seq 解析：データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得 (実行済み)

ダウンロードします。

```
$ wget  
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA055/SRA055683  
/SRX157933/SRR518891_1.fastq.bz2  
$ wget  
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA055/SRA055683  
/SRX157933/SRR518891_2.fastq.bz2
```

RNA-seq 解析：データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータ取得 (実行済み)

解凍して、先頭1000リードを抽出します。

```
$ bunzip2 SRR518891_1.fastq.bz2
$ bunzip2 SRR518891_2.fastq.bz2

$ head -4000 SRR518891_1.fastq > 1K_SRR518891_1.fastq
$ head -4000 SRR518891_2.fastq > 1K_SRR518891_2.fastq
```

TRY!

RNA-seq 解析 : データ

データ取得 → クオリティコントロール → マッピング → 発現定量

- シーケンスデータを確認

```
$ cd /home/ユーザ名/Desktop/amelieff  
$ ll
```

```
-rw-rw-r-- 1 admin1409 admin1409 346770 Dec 3 2013 1K_SRR518891_1.fastq  
-rw-rw-r-- 1 admin1409 admin1409 346770 Dec 3 2013 1K_SRR518891_2.fastq
```


TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量

- シーケンスデータのクオリティを確認

FastQCを実行します。

```
$ mkdir rnaseq
$ fastqc -o rnaseq -f fastq 1K_SRR518891_1.fastq 1K_SRR518891_2.fastq
```

fastqc_report.htmlを、ウェブブラウザで開きます。

```
$ firefox rnaseq/1K_SRR518891_1_fastqc/fastqc_report.html
$ firefox rnaseq/1K_SRR518891_2_fastqc/fastqc_report.html
```

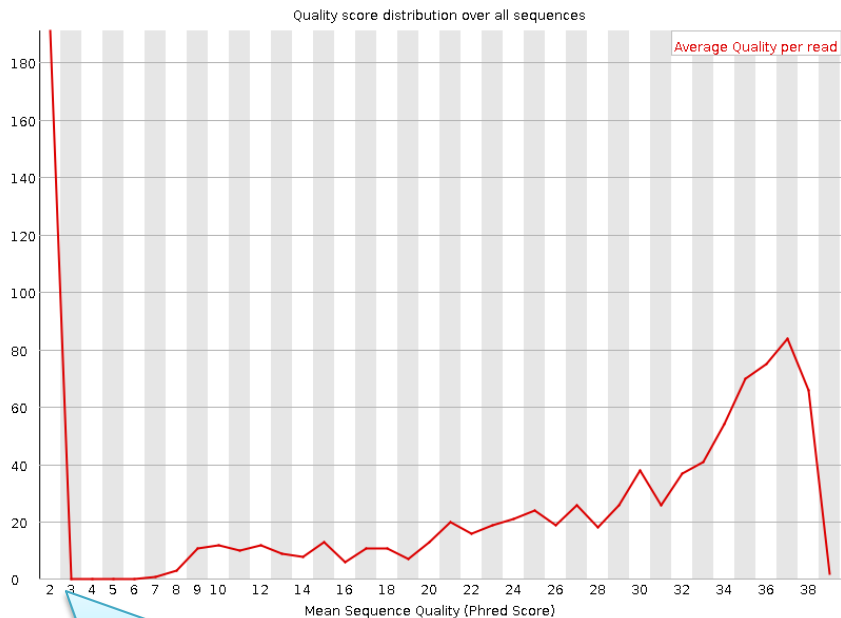
以降の解析は、片側のリードのみ使用します。

TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量

Per sequence quality scores



最初の1塩基

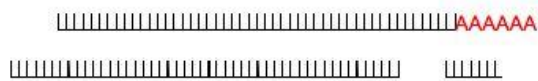
Overrepresented sequences

Sequence	Count
CACTGTTATTGCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTTTT	3
CACTGTTTCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTTTT	3
NACTGTTCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTTTT	2

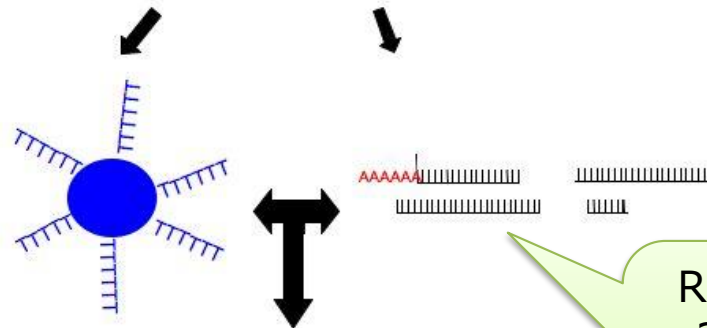
PolyA/T tail が存在

応用) PolyA/T tailの混入

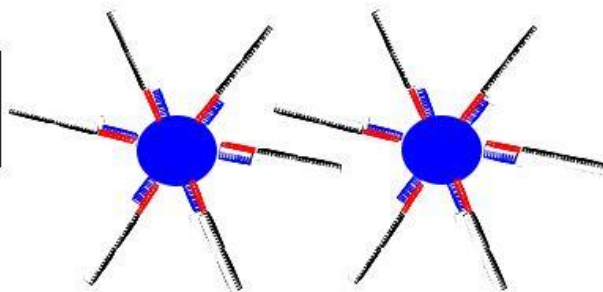
Isolate Total RNA



Fragmentation and/or Isolation
In this case, isolation via Poly(T) coated magnetic beads



Poly(A) RNA molecules bind to the Poly(T) magnetic beads



RNA-seq (mRNA) では
3'末端にPolyA/T tailが
ついている転写物を
シーケンシングするため

TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量

- 最初の 1 塩基を削除

fastx_trimmer の使用方法を確認する

```
$ fastx_trimmer -h
```

```
usage: fastx_trimmer [-h] [-f N] [-l N] [-t N] [-m MINLEN] [-z] [-v] [-i INFILE] [-o OUTFILE]
```

```
[-h]           = This helpful help screen.  
[-f N]        = First base to keep. Default is 1 (=first base).
```

TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → クオリティコントロール → マッピング → 発現定量

- 最初の 1 塩基を削除

2塩基目から使う

```
$ cd rnaseq
```

```
$ fastx_trimmer -f 2 -i ../1K_SRR518891_1.fastq -o 1K_SRR518891_1_s.fastq -Q33
```

TRY!

RNA-seq 解析 : クオリティコントロール

データ取得 → クオリティコントロール → マッピング → 発現定量

- PolyA/T tailを除去

3'端にAを5連続以上含むリード数がどのくらいあるか調べる

```
$ grep "AAAAA$" 1K_SRR518891_1_s.fastq | wc -l
```

```
39
```

fastx_clipperの使用方法を確認する

```
$ fastx_clipper -h
```

TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → クオリティコントロール → マッピング → 発現定量

- PolyA/T tailを除去

PolyA/T tailを除去する

```
$ fastx_clipper -a AAAAA -i 1K_SRR518891_1_s.fastq  
-o 1K_SRR518891_1_s_notail.fastq -Q 33
```

Prinseqなど、各リードのPolyA/Tの数に合わせて除去するソフトもあります。

RNA-seq 解析 : クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量

- クオリティの低いリードを除外

3'末端からクオリティ20未満の塩基をトリミングし、長さが30塩基未満になったリードを破棄する。

さらに、80%以上の塩基がクオリティ20以上のリードのみを抽出する。

```
$ fastq_quality_trimmer -t 20 -l 30 -Q 33  
-i 1K_SRR518891_1_s_notail.fastq | fastq_quality_filter  
-q 20 -p 80 -Q 33 -o 1K_SRR518891_1_clean.fastq
```


TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量

- クリーニング結果の確認

FastQCを実行します。

```
$ fastqc -f fastq 1K_SRR518891_1_clean.fastq
```

fastqc_report.htmlを、ウェブブラウザで開きます。

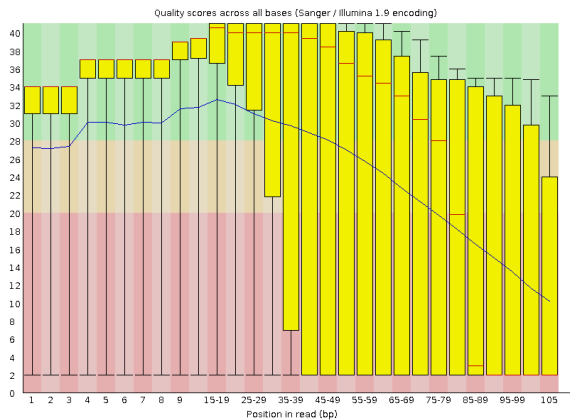
```
$ firefox 1K_SRR518891_1_clean_fastqc/fastqc_report.html
```

クリーニング前後のリード数と、クオリティの変化を確認してください。

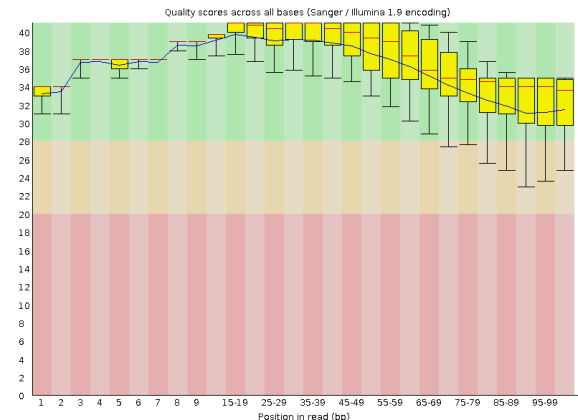
TRY!

RNA-seq 解析：クオリティコントロール

データ取得 → **クオリティコントロール** → マッピング → 発現定量



クリーニング前



クリーニング後

**サンプルや調整方法、シーケンサの特徴にあわせて
クリーニング項目や条件を工夫しています。**

応用) アダプタ配列を除外するソフト

- ① cutadapt
- ② FastX-Toolkit (fastxclipper)
- ③ tagcleaner

指定した配列はどのソフトでも除ける。

fastx_clipperは、部分配列もかなり除けたが、リード数も1/10以下に減るためアダプタ以外の配列も除いている可能性がある。

tagcleanerは、一度に1アダプタ配列しか指定できない。

```
$ cutadapt -b TCTCGTATGCCGTCTTC -b CTACAGTCCGACGA  
-m 10 -n 2 $FILE.fastq 1> $FILE_cutadapt.fastq
```

※オプション

m	これより短くなったものは破棄
n	同リードへのアダプタ出現回数
O	マッチ領域の最少長
e	「エラー塩基数/マッチ領域長」の最大

TRY!

RNA-seq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 発現定量

- TopHatの使い方を確認

```
$ tophat
```

```
Usage:
tophat [options] <bowtie_index> <reads1[,reads2,...]> [reads1[,reads2,...]]
                                     [quals1[,quals2,...]] [quals1[,quals2,...]]
```

スプライシングを考慮して、マッピングするため、既知の遺伝子情報を使用することもできます。

```
-G/--GTF <filename> (GTF/GFF with known transcripts)
```

```
-g/--max-multihits <int> default: 20
```

TRY!

RNA-seq 解析 : マッピング

データ取得 → クオリティコントロール → **マッピング** → 発現定量

- マッピング ※今回はリード数が少ないため、マッピング基準を緩めています。

```
$ tophat -o 1K_SRR518891 -g 3 -N 10 --read-edit-dist 10  
--read-gap-length 10  
/home/ユーザ名/Desktop/amelieff/Scerevisiae/Bowtie2Index/genome  
1K_SRR518891_1_clean.fastq
```

```
$ ls 1K_SRR518891
```

BAMとインデックス、
BEDなどが作成されます。

```
accepted_hits.bam  deletions.bed  junctions.bed  prep_reads.info  
align_summary.txt  insertions.bed  logs           unmapped.bam
```

-N/--read-mismatches Final read alignments having more than these many mismatches are discarded.
--read-edit-dist Final read alignments having more than these many edit distance are discarded.
--read-gap-length Final read alignments having more than these many total length of gaps are discarded.

TRY!

RNA-seq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 発現定量

- マッピング率

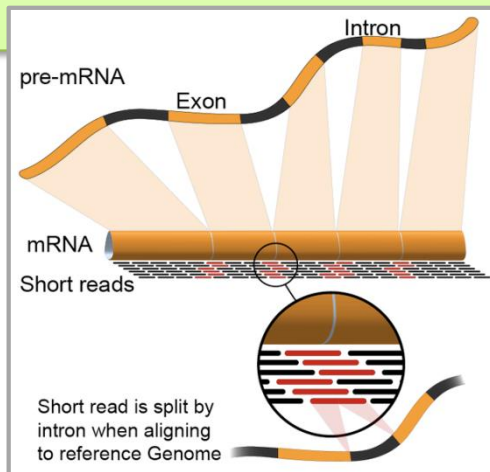
```
$ less 1K_SRR518891/align_summary.txt
```

```
Reads:
  Input      :      752
  Mapped     :      79 (10.5% of input)
  of these:  11 (13.9%) have multiple alignments (0 have >3)
10.5% overall read mapping rate.
```

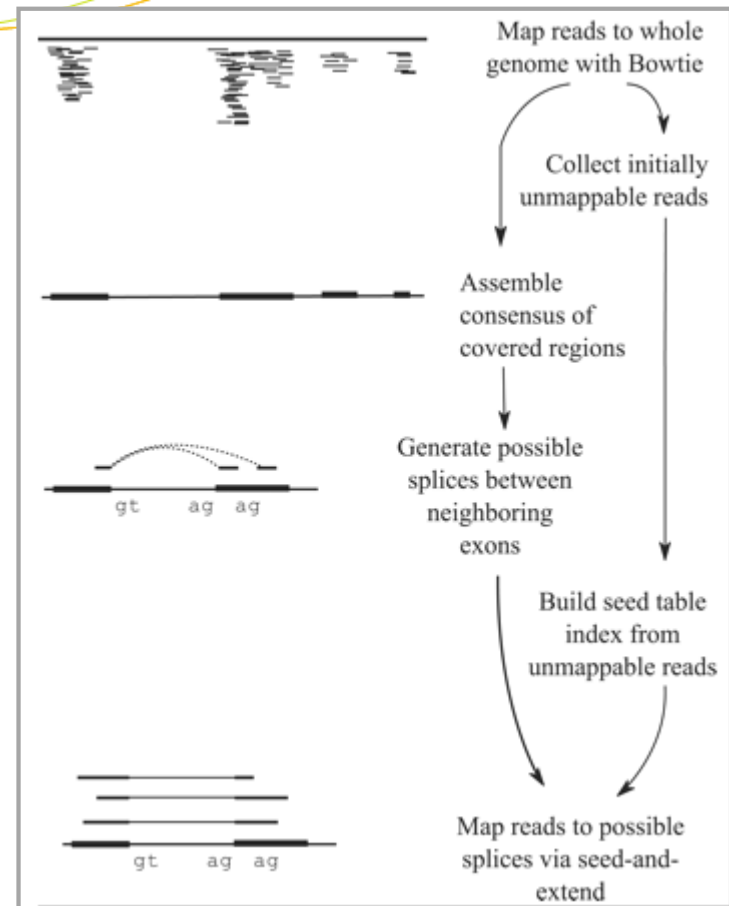
マッピング率

ポイント) TopHatのアルゴリズム

1. リードをペアエンドでリファレンスにマッピングする。
2. マッピングできなかったリードを断片化して、リファレンスにマッピングする。
3. マッピング結果をもとに、転写構造をアセンブリングする。



<http://en.wikipedia.org/wiki/File:RNA-seq-alignment.png>



<http://www.ncbi.nlm.nih.gov/pubmed/19289445>

TRY!

RNA-seq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 発現定量

- マッピング結果の可視化

```
$ samtools index accepted_hits.bam  
$ igv.sh
```

accepted_hits.bamをIGVで表示してください。

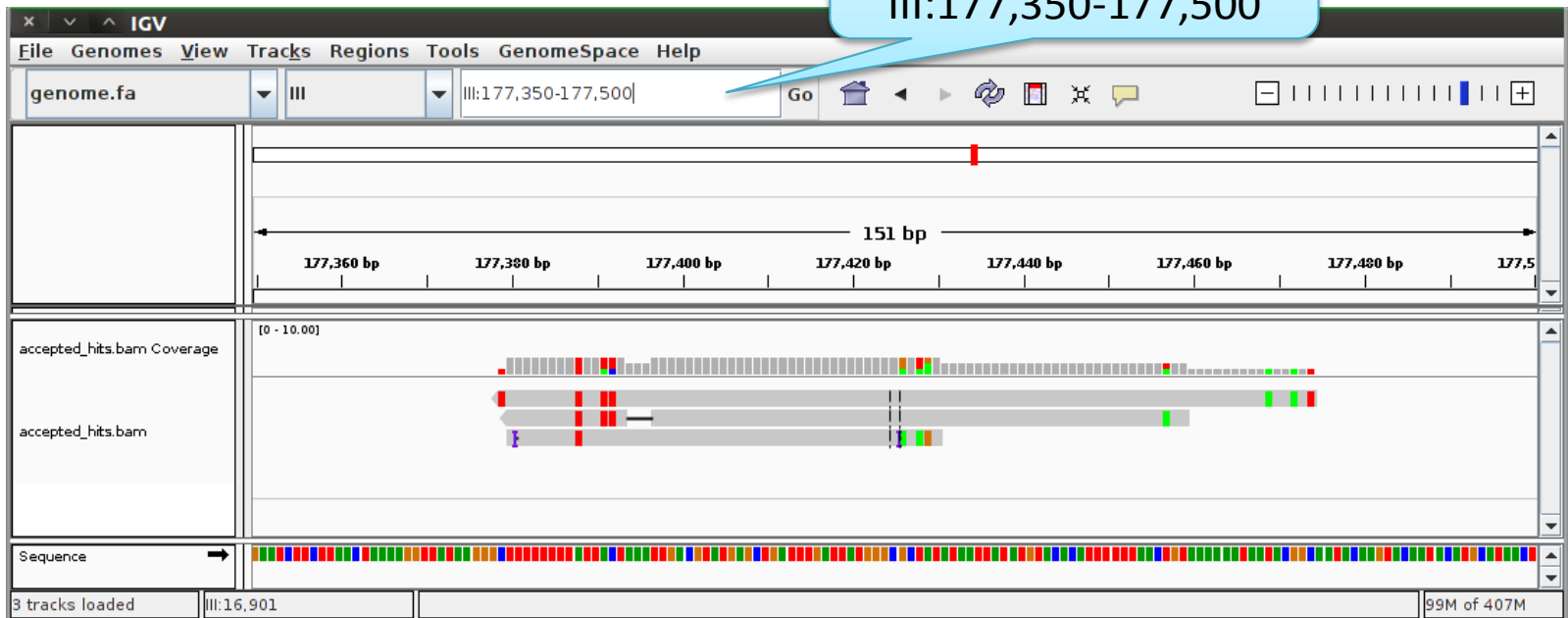
TRY!

RNA-seq 解析 : マッピング

データ取得 → クオリティコントロール → マッピング → 発現定量

- マッピング結果の可視化

Positionの例:
III:177,350-177,500



ポイント)

遺伝子の発現量 ≠ 遺伝子上にマップされたリード数

長い遺伝子ほど、マップされるリードは多くなる（遺伝子間のバイアス）
サンプル量の多いランほど、マップされるリードは多くなる（ラン間のバイアス）

これらのバイアスを補正してから発現量を比較する必要があります

- ・発現量としてよく使われる指標

RPKM (Reads Per Kilobase per Million mapped reads)

FPKM (Fragments Per Kilobase of exon per Million mapped fragments)

どちらも、発現量をエクソン長と全マッピング数で補正した値

$$\text{FPKM} = \text{raw counts} \times \frac{1,000,000}{\text{all reads}} \times \frac{1,000}{\text{gene length}}$$

TRY!

RNA-seq 解析：発現定量

データ取得 → クオリティコントロール → マッピング → **発現定量**

- Cufflinksの使い方を確認

```
$ cufflinks
```

```
cufflinks v2.1.1
```

```
-o/--output-dir      write all output files to this directory
-p/--num-threads     number of threads used during analysis
--seed               value of random number generator seed
-G/--GTF             quantitate against reference transcript annotations
-g/--GTF-guide       use reference transcript annotation to guide assembly
```

アセンブルのガイドとして既知の遺伝子情報を使用することもできます。

TRY!

RNA-seq 解析：発現定量

データ取得 → クオリティコントロール → マッピング → **発現定量**

- 発現量を計算 ※今回はリード数が少ないため、検出基準を緩めています。

```
$ cufflinks --min-frags-per-transfrag 2  
-o 1K_SRR518891 1K_SRR518891/accepted_hits.bam  
  
$ ll -h 1K_SRR518891
```

fpkm_trackingファイル
が作成されます。

```
-rw-rw-r-- 1 admin1409 admin1409 514 Jul 30 11:02 genes.fpkm_tracking  
-rw-rw-r-- 1 admin1409 admin1409 562 Jul 30 11:02 isoforms.fpkm_tracking  
-rw-rw-r-- 1 admin1409 admin1409 0 Jul 30 11:02 skipped.gtf  
-rw-rw-r-- 1 admin1409 admin1409 2.1K Jul 30 11:02 transcripts.gtf
```

--min-frags-per-transfrag minimum number of fragments needed for new transfrags

TRY!

RNA-seq 解析 : 発現定量

データ取得 → クオリティコントロール → マッピング → 発現定量

- 発現量を計算

```
$ less 1K_SRR518891/genes.fpkm_tracking
```

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
CUFF.1	-	CUFF.1	-	III:177378-177474	-	-	1.74964e+07	0	3.76995e+07	OK		
CUFF.2	-	CUFF.2	-	VII:883750-883860	-	-	1.43109e+07	0	2.86287e+07	OK		
CUFF.3	-	CUFF.3	-	XII:370041-370150	-	-	1.10892e+07	0	2.38939e+07	OK		
CUFF.4	-	CUFF.4	-	XIV:302658-302762	-	-	8.74601e+06	0	2.1132e+07	OK		
CUFF.5	-	CUFF.5	-	XIV:415071-415117	-	-	1.16898e+08	0	3.50695e+08	OK		

4列目がGene ID、
10列目がFPKMです。

応用) サンプル間比較

サンプルごとに発現量を計算したあと、サンプルごとに発現している遺伝子が違うため、比較の基準とする遺伝子リストを作成します。

```
$ cuffmerge -o COMPARE -g genes.gtf -s genome.fa  
transcript.gtf.txt
```

```
Group1/S1/transcript.gtf  
Group1/S2/transcript.gtf  
Group2/S3/transcript.gtf  
Group2/S4/transcript.gtf
```

transcript.gtf.txt

各サンプルのcufflinks結果を
羅列したファイル

発現量を比較します

```
$ cuffdiff -o COMPARE -L Group1,Group2 genes.gtf  
Group1/S1/accepted_hits.bam,Group1/S2/accepted_hits.bam  
Group2/S3/accepted_hits.bam,Group2/S4/accepted_hits.bam
```

本講義の内容

- **Reseq解析**

公開データ取得



クオリティコントロール



マッピング



変異検出

- **RNA-seq解析**

公開データ取得



クオリティコントロール



マッピング



発現定量



ご清聴ありがとうございました。