次世代シークエンサ実習II

ITのチカラで研究を支援

amelieff

アメリエフ株式会社

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

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

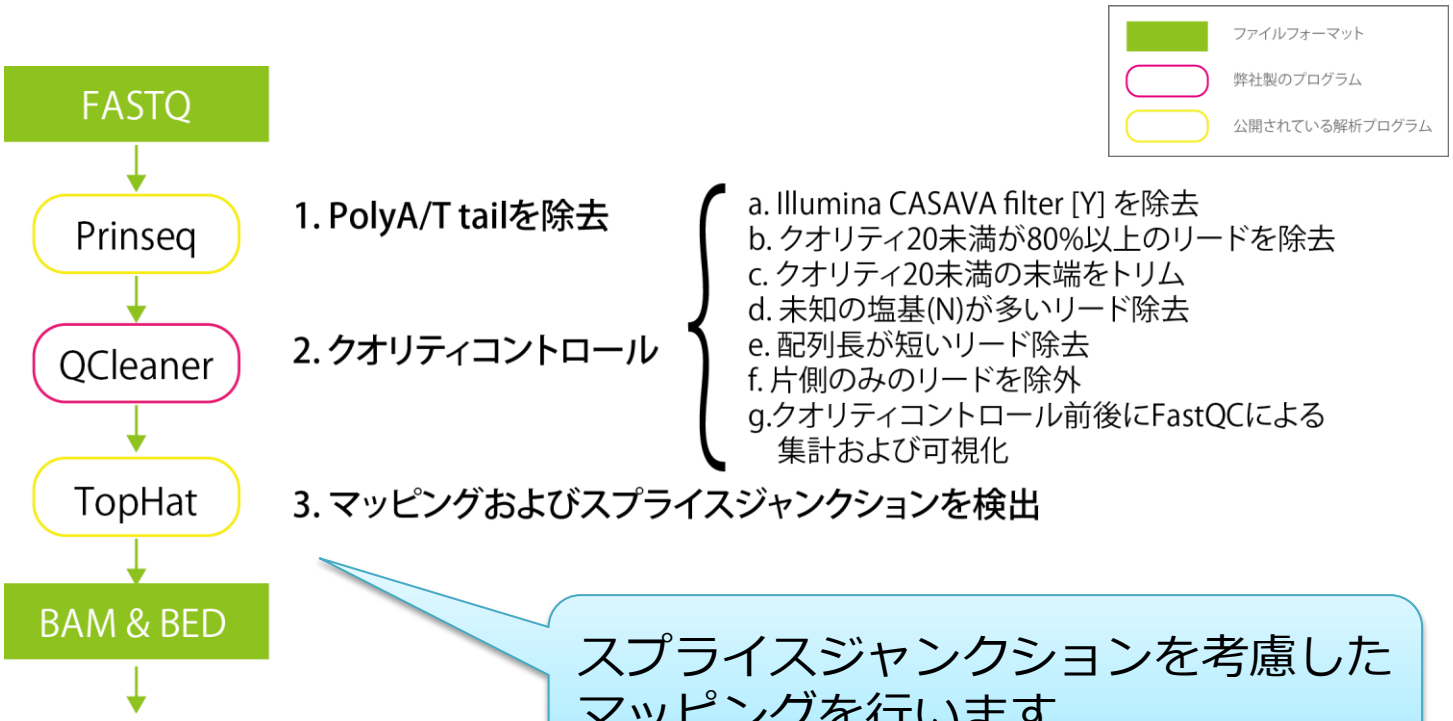
- リファレンスがある生物種の場合：
  - 既知遺伝子にマッピングする
  - **リファレンスにマッピングして遺伝子構造を同定する**

本日の内容

- 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

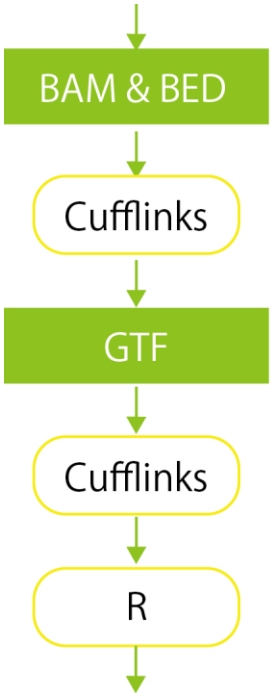
ＲＮＡ-ｓｅｑ解析：パイプライン

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



4. 遺伝子構造を予測／新規の転写産物を推定／FPKMを算出

5. グループ間比較

今日は一部のコマンドを実行します。

ヒートマップ、GO解析、パスウェイ解析　など様々。

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アントの探索

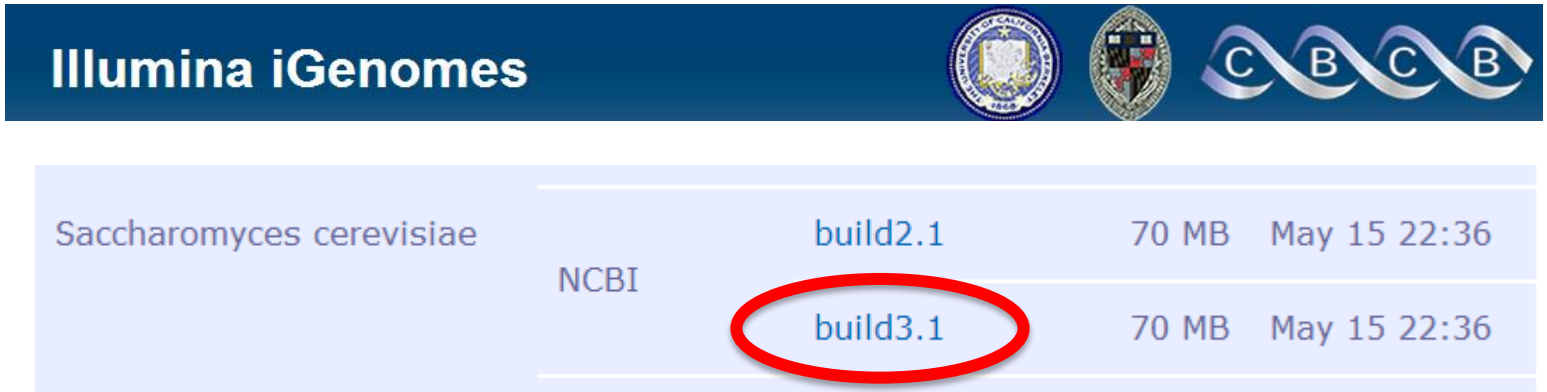
RNA-seqがマイクロアレイと比較して優れている点

- 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



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/admin1409/genome/Saccharomyces_cerevis
iae/NCBI/build3.1
$ ll Sequence
```

```
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 Sequence/Bowtie2Index
```

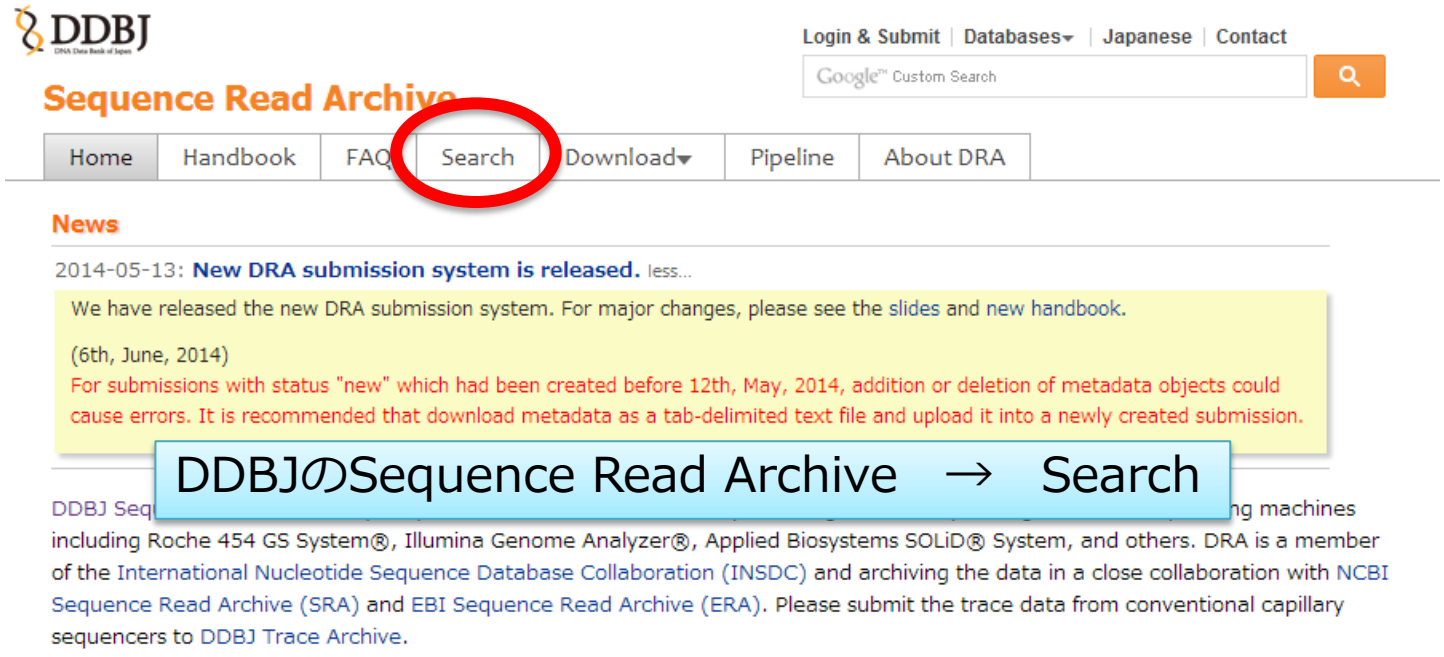
```
genome.1.bt2
genome.2.bt2
genome.3.bt2
genome.4.bt2
genome.fa -> ../WholeGenomeFasta/genome.fa
genome.rev.1.bt2
genome.rev.2.bt2
```

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

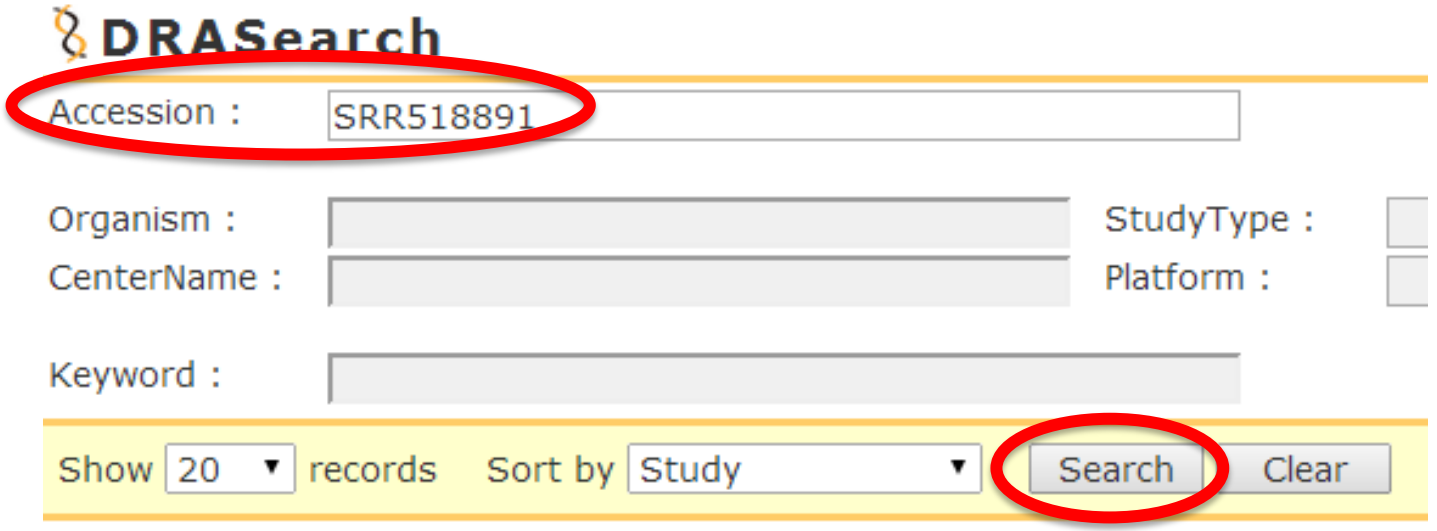
- シーケンスデータ取得　http://trace.ddbj.nig.ac.jp/dra/index.html



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



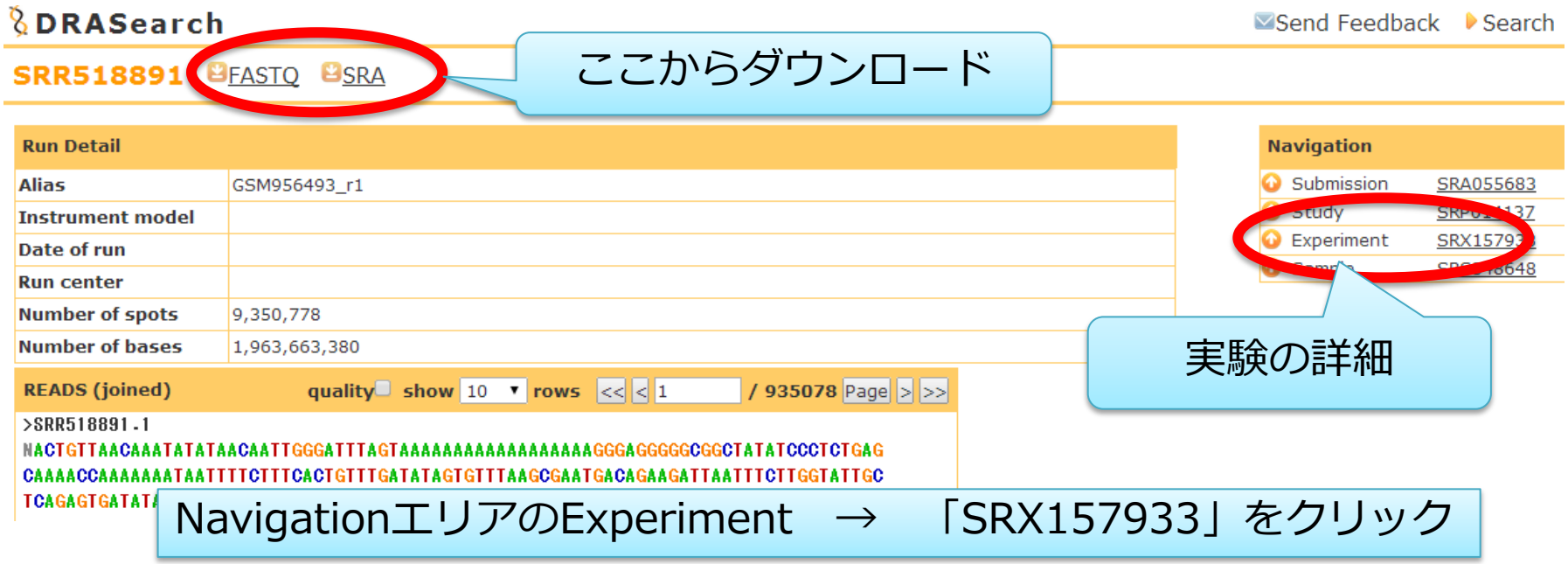
Accessionに「 SRR518891 」と入力　→　Search

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。



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/SR
A055683/SRX157933/SRR518891_1.fastq.bz2
$ wget
ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA055/SR
A055683/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/admin1409/amelieff/ngs
$ 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
```

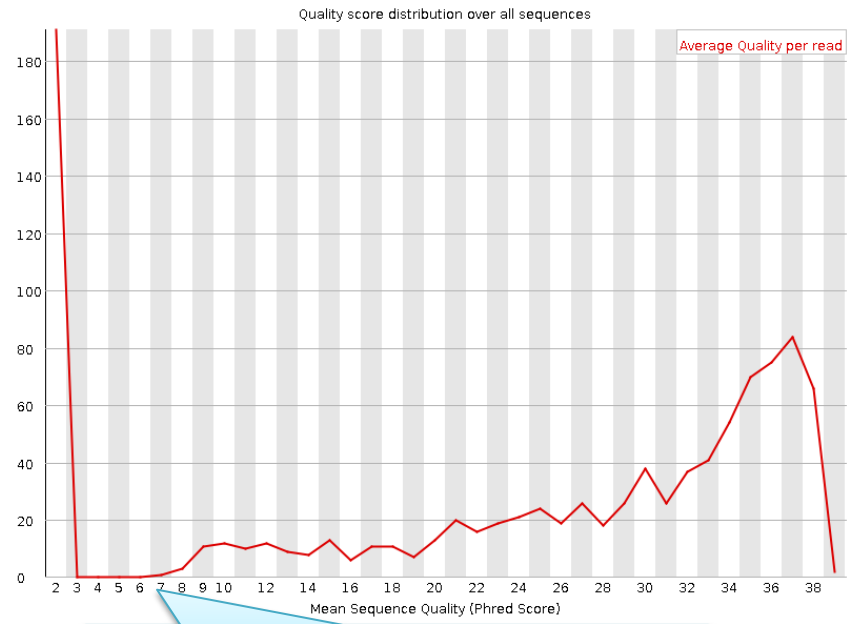
以降の解析は、片側のリードのみ使用します。

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



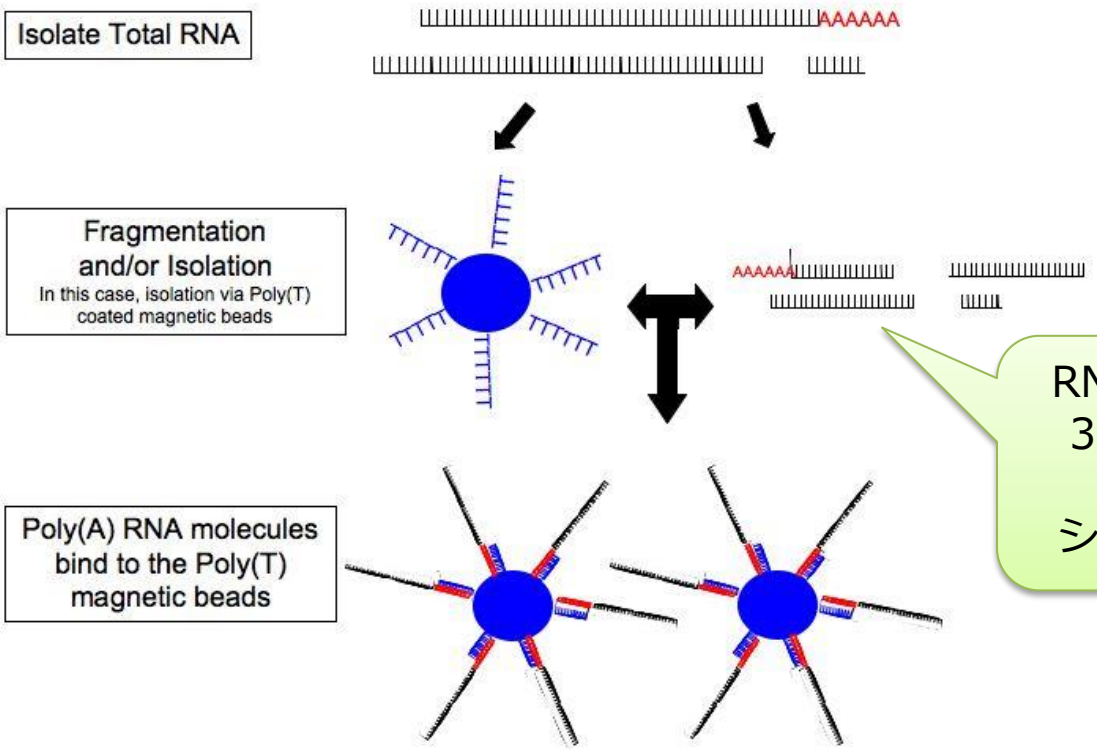
最初の１塩基

Overrepresented sequences

Sequence	Count
CACTGTTATTGCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTT	3
CACTGTTTGCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTTTT	3
NACTGTTCTCAGAGTGATATAGCGGCCGCCTCCACTTTTTTTTTTTTT	2

PolyA/T tail が存在

応用） PolyA/T tailの混入



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

１塩基目を削除する

```
$ 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、各シーケンスのPolyA/Tの数に合わせて除去するソフトもあります。

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

- 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ー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。

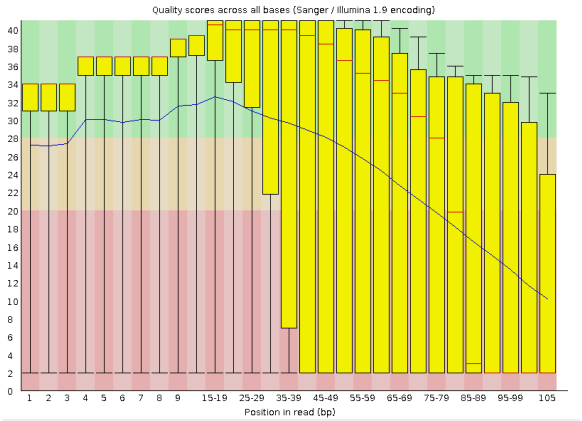
```
$ firefox 1K_SRR518891_1_clean_fastqc/fastqc_report.html
```

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。

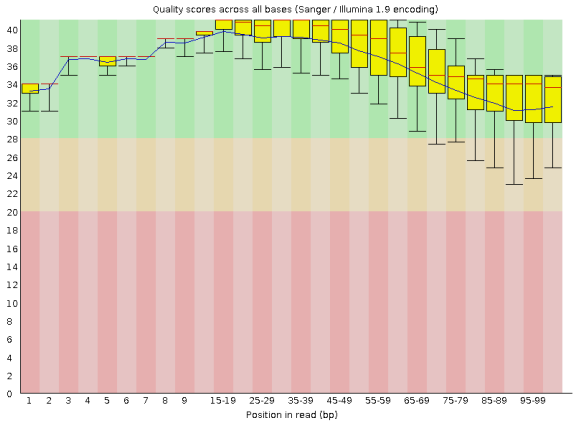
TRY!

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/admin1409/genome/Saccharomyces_cerevisiae/NCBI
/build3.1/Sequence/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

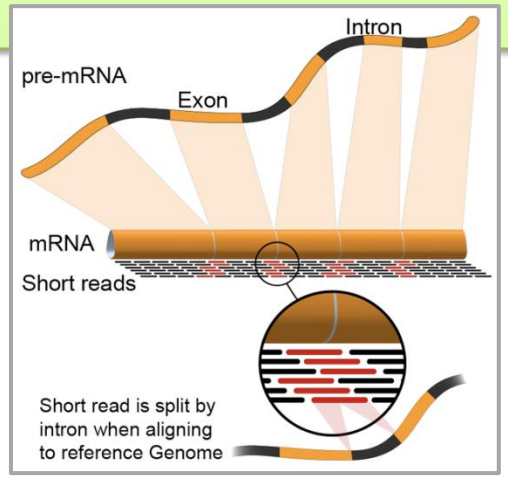
```
$ less 1K_SRR518891/align_summary.txt
```

```
Reads:
                Input:        752
             Mapped:           79 (10.5% of input)
          of these:            24 (30.4%) have multiple alignments (0 have >3)
10.5% overall read alignment rate.
```

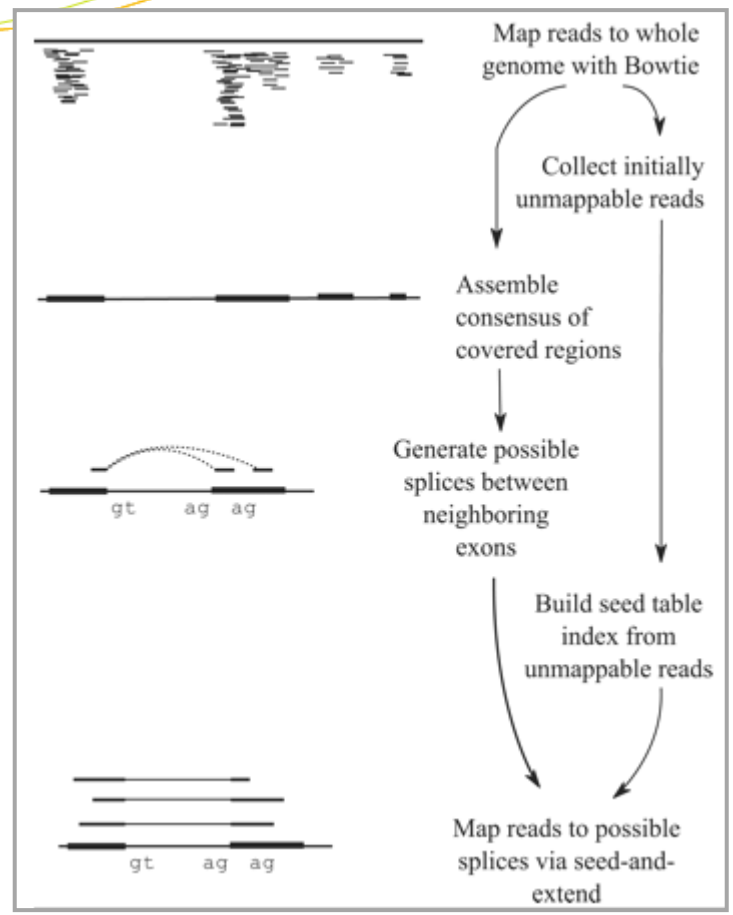
マッピング率

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

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

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

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	137061	1.64474e+06	OK	
CUFF.2	-	CUFF.2	-	-	VII:883750-883860	-	-	1.43109e+07	119617	1.67464e+06	OK	
CUFF.3	-	CUFF.3	-	-	XII:370041-370150	-	-	1.10892e+07	120715	1.44858e+06	OK	
CUFF.4	-	CUFF.4	-	-	XIV:302658-302762	-	-	8.74601e+06	0	1.13866e+06	OK	
CUFF.5	-	CUFF.5	-	-	XIV:415071-415117	-	-	1.16898e+08	0	2.00229e+06	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。