平成28年度NGSハンズオン講習会
RNA-seq解析

2016年7月27日

amelieff

本講義にあたって

- 代表的な解析の流れを紹介します。
  - 論文でよく使用されているツールを使用します。

- コマンドを沢山実行します。
  - タイプミスが心配な方は、コマンド例がありますのでコピーして実行してください。
  - 実行が遅れてもあせらずに、課題や休憩の間に追い付いてください。

本講義の内容

前半パート (講義)

- RNA-seqとは
- RNA-seq解析の流れ
- 公開データの取得
- クオリティコントロール
- マッピング
- 発現定量

後半パート (実習)

- クオリティコントロール
- マッピング
- 発現定量
- 発現比較
- 可視化
- まとめ

前半パート（講義）

RNA-seqとは

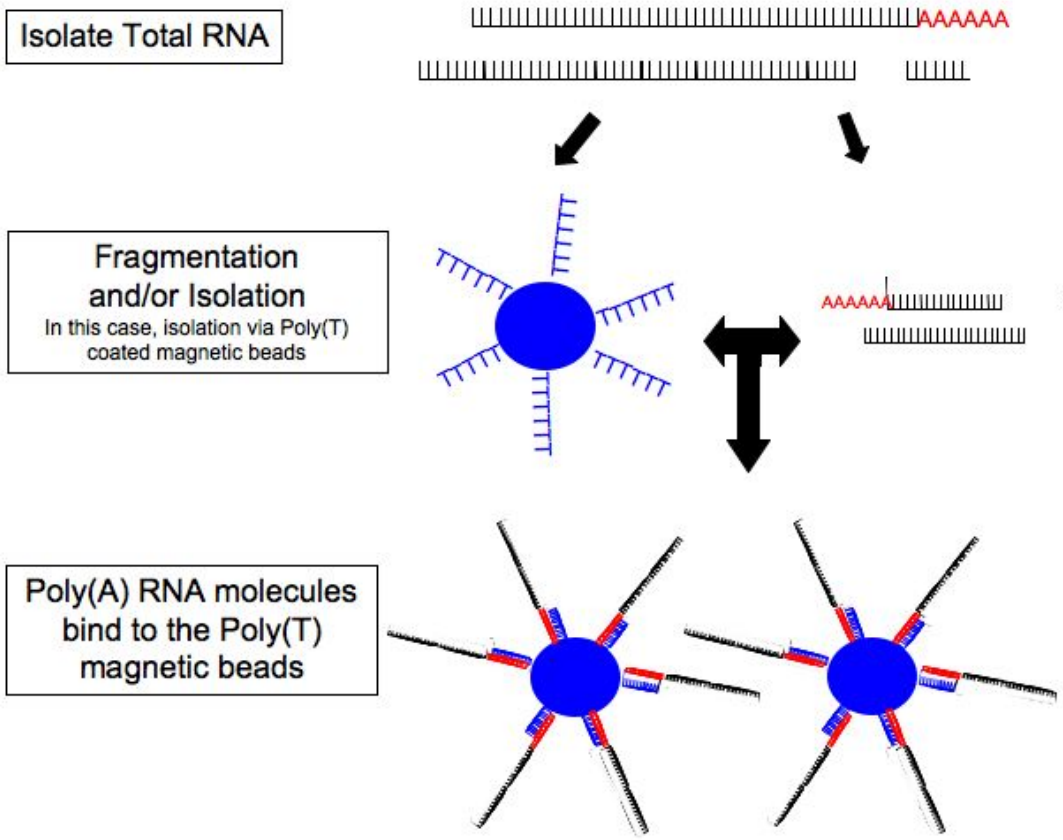
- メッセンジャーRNA（mRNA）をキャプチャして次世代シーケンサーでシーケンシングする手法

- 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発現量を定量する

- 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

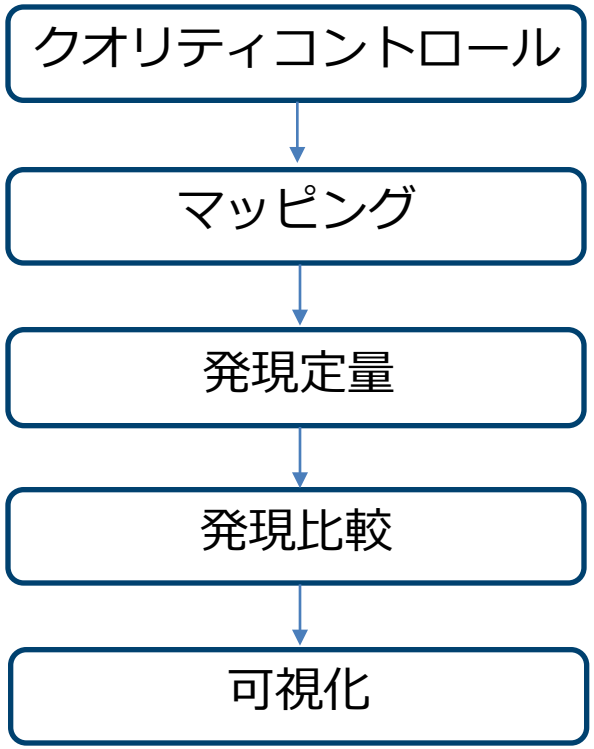
クオリティコントロール

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、リードには、PolyA/T tailが存在する

RNA-seq解析の流れ



解析ソフト例

PRINSEQ, cutadapt, FASTX_Toolkit...

HISAT2, TopHat2, STAR, Bowtie2...

Cufflinks, StringTie, RSEM...

Cuffdiff, DESeq, edgeR...

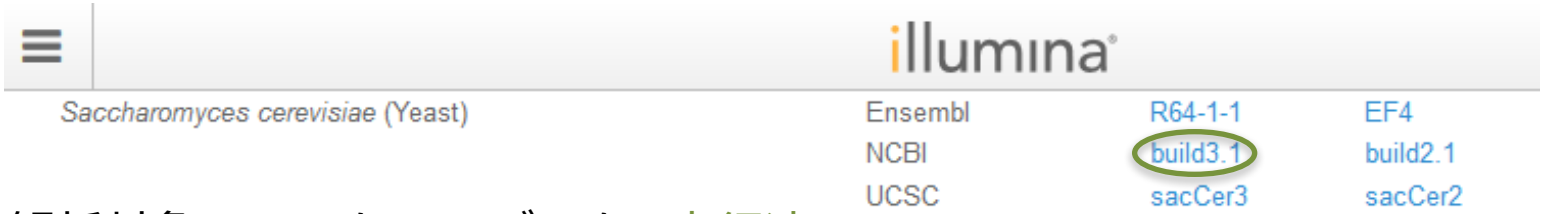
CummeRbund...

- RNA-seq解析の一般的な流れであり、全てのRNA-seqで同一の解析を行うわけではありません。
- 研究の目的やデータに合わせて、最適な解析を行います。

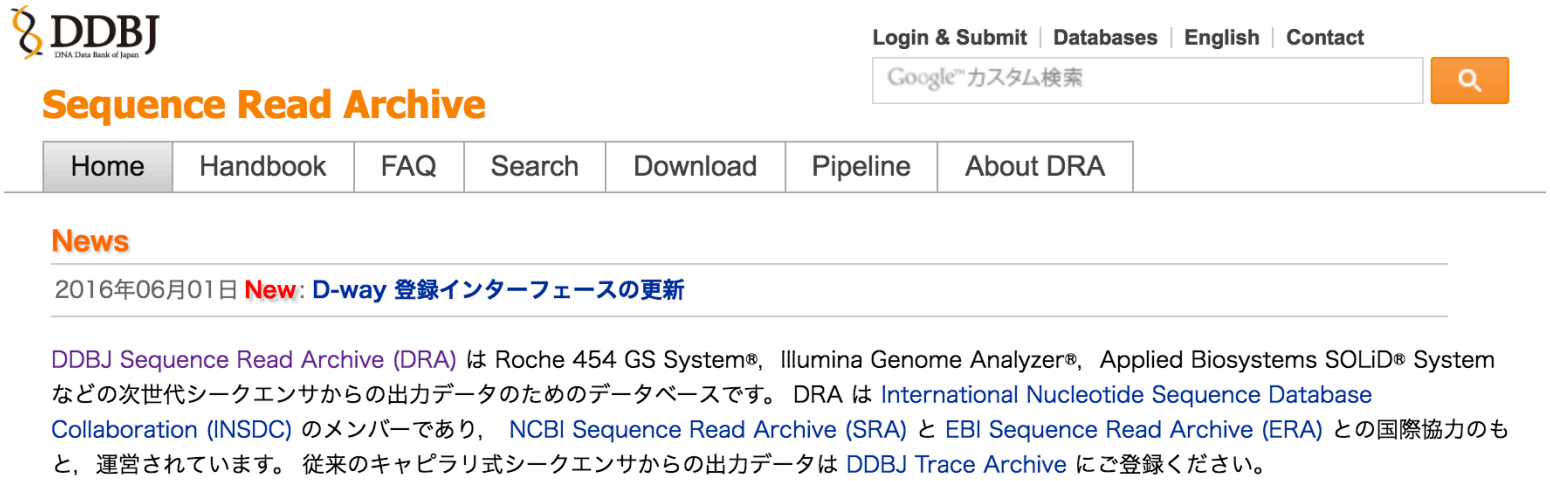
公開データの取得

今回の解析に必要なデータ

- リファレンスゲノム (実行済み)
  - http://support.illumina.com/sequencing/sequencing_software/igenome.html



- 解析対象のシーケンスデータ (実行済み)



公開データの取得

酵母のリファレンスゲノムデータの取得方法

```
$ wget ftp://igenome:G3nom3s4u@ussd-
  ftp.illumina.com/Saccharomyces_cerevisiae/NCBI/build3.1/Saccha
  romyces_cerevisiae_NCBI_build3.1.tar.gz
$ tar zxvf Saccharomyces_cerevisiae_NCBI_build3.1.tar.gz
```

Saccharomyces cerevisiaeのリファレンスゲノムをイルミナのWebページからダウンロードし解凍する (実行済み)。

```
$ ls -l /home/iu/genome/sacCer3/
                                    :
-rwxr-xr-x. 1 iu iu 12400379  5月 23 11:09 genome.fa
                                    :
-rwxr-xr-x. 1 iu iu      462  5月 23 11:09 genome.fa.fai
                                    :
```

/home/iu/genome/sacCer3/に
解凍したファイル（今回使用するデータのみ）を置いてあるので確認する。

公開データの取得

fastaファイルの中身の確認

```
$ less /home/iu/genome/sacCer3/genome.fa
```

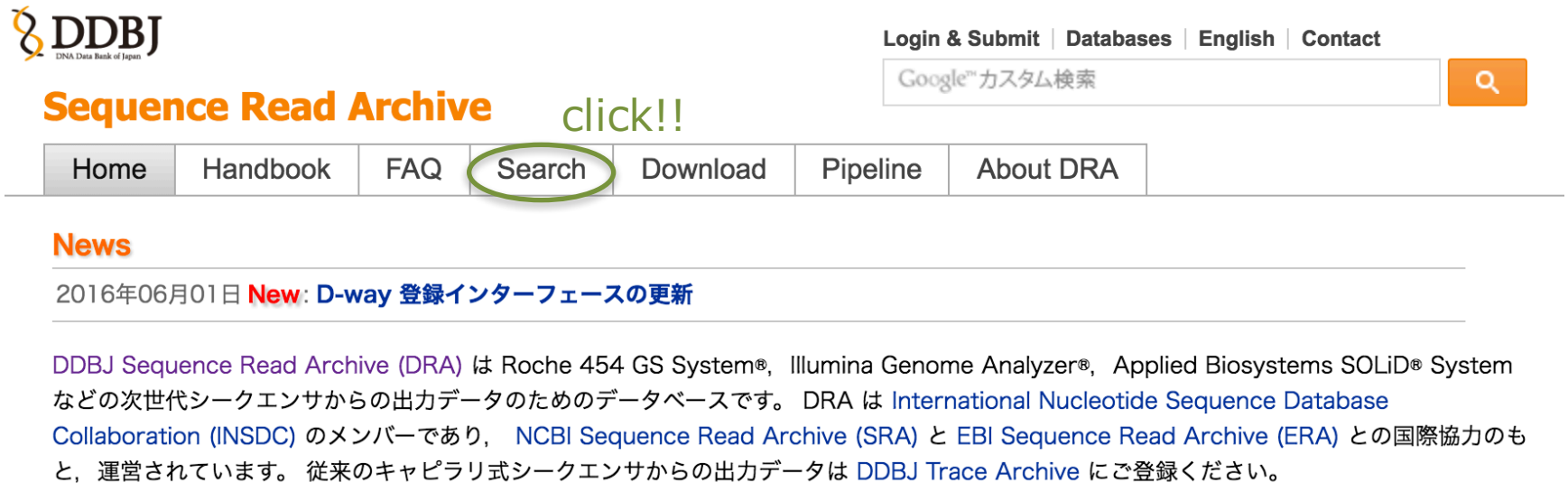
```
>chrI
CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACC
CACACACACACATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTG
GCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTAC
CCTGTCCCATTCAACCATACCACTCCGAACCACCATCCATCCCTCTACTT
ACTACCACTCACCCACCGTTACCCTCCAATTACCCATATCCAACCCACTG
                                                  :
```

1行目：　　コンティグ名
2行目以降：　実際の配列情報

※「q」で閲覧を終了する

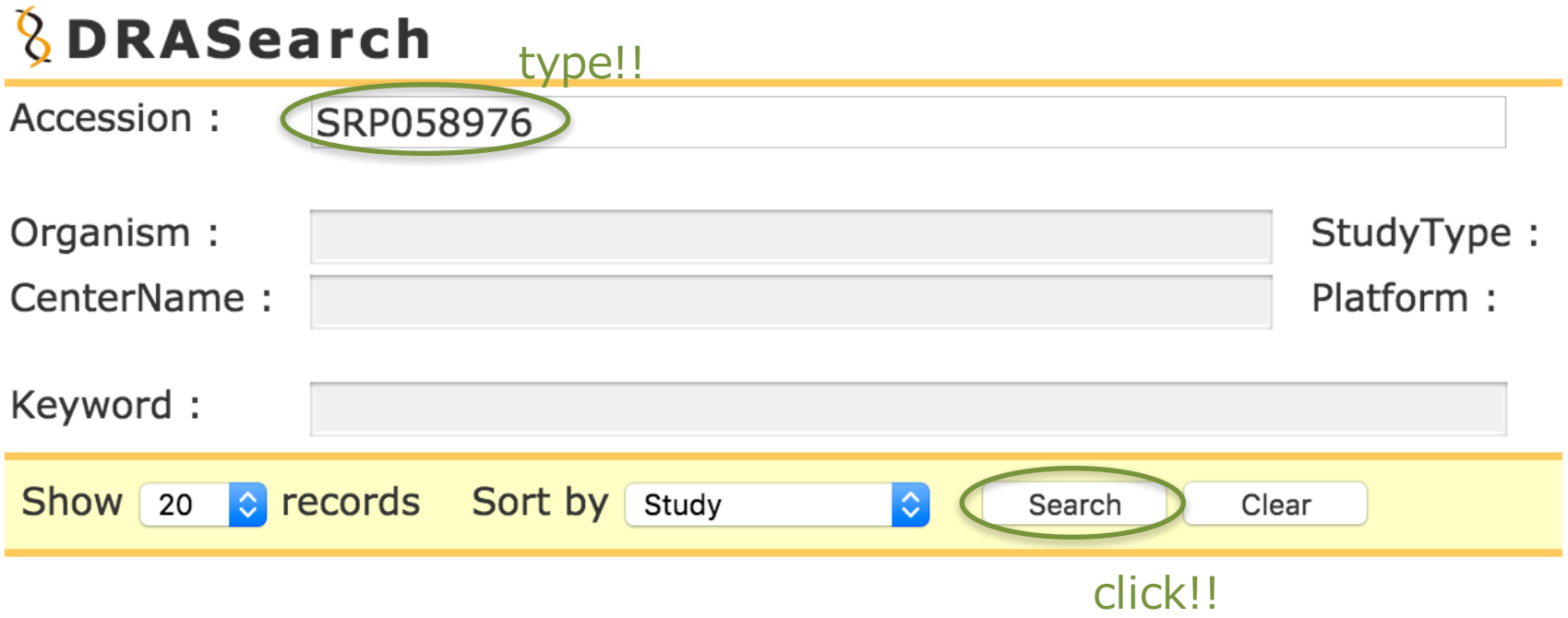
公開データの取得

解析対象のシーケンスデータの取得方法 ①
http://trace.ddbj.nig.ac.jp/dra/index.htmlへアクセスする。



公開データの取得

解析対象のシーケンスデータの取得方法 ②
SRP058976を検索する。



公開データの取得

解析対象のシーケンスデータの取得方法 ③
研究概要を確認する。

DRASearch　Send Feedback　Search Home　DRA Home

SRP058976

Study Detail	
Title	RNA Proximity Ligation to Resolve Intramolecular RNA Structures in situ
Study Type	Other
Abstract	Proof-of-concept of a new method involving the limited digestion and subsequent ligation of intramolecular RNA structures in situ followed by deep sequencing Overall design: Proof-of-concept of RPL in S. cerevisiae and H. sapiens tissue culture
Description	
Center Name	GEO

Navigation			
Submission	SRA271046	FTP	
Experiment	SRX1046447	FASTQ	SRA
	SRX1046448	FASTQ	SRA
	SRX1046449	FASTQ	SRA
	SRX1046450	FASTQ	SRA
	SRX1046451	FASTQ	SRA
click!!	SRX1046452	FASTQ	SRA
	SRX1046453	FASTQ	SRA
	SRX1046454	FASTQ	SRA
	SRX1046455	FASTQ	SRA
	SRX1046456	FASTQ	SRA
	SRX1046457	FASTQ	SRA
	SRX1046458	FASTQ	SRA
	SRX1046459	FASTQ	SRA

公開データの取得

解析対象のシーケンスデータの取得方法 ④
実験詳細を確認する。

DRASearch　Send Feedback　Search Home　DRA Home

SRX1046452　FASTQ　SRA

Experiment Detail	
Title	GSM1701707: RPL_yeast_imidazole_ligase; Saccharomyces cerevisiae; OTHER
Design Description	
Organism	Saccharomyces cerevisiae

Library Description	
Name	
Strategy	OTHER
Source	TRANSCRIPTOMIC
Selection	other
Layout	PAIRED
Orientation	
Nominal Length	
Nominal Sdev	
Construction Protocol	RNA was extracted using TRIZol and Zymo DirectZOL columns Illumina TruSeq

Navigation		
Submission	SRA271046	FTP
Study	SRP058976	
Sample	SRS951223	
Run	SRR2048224	FASTQ SRA

ここからダウンロード可能

公開データの取得

解析対象のシーケンスデータの取得方法 ⑤
データをダウンロードする(実行済み) 。

```
$ mkdir -p rnaseq/data
```

解析を行うrnaseqディレクトリとdataディレクトリを作成する。

```
$ cd rnaseq/data/
$ wget ¥
  ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX104/SRX1046452/SRR2048224/SRR2048224.sra
```

SRR2048225、SRR2048228、SRR2048229についても同様にダウンロード

- SRR2048224 : RPL_yeast_imidazole_ligase
- SRR2048225 : RPL_yeast_imidazole_ligase_rep2
- SRR2048228 : RPL_yeast_imidazole_noligase
- SRR2048229 : RPL_yeast_imidazole_noligase_rep2

公開データの取得

解析対象のシーケンスデータの取得方法 ⑥

SRAデータをFASTQファイルへ変換する(実行済み) 。

```
$ fastq-dump --split-files SRR2048229.sra
```

fastq-dumpコマンドは、NCBI SRA toolkit をインストールすると利用できる。

```
$ head -40000 SRR2048224_1.fastq > 10K_SRR2048224_1.fastq
$ head -40000 SRR2048224_2.fastq > 10K_SRR2048224_2.fastq
```

先頭1万リードを抽出する(実行済み) 。SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

```
$ ls data
10K_SRR52048224_1.fastq 10K_SRR52048228_1.fastq
10K_SRR52048224_2.fastq 10K_SRR52048228_2.fastq
10K_SRR52048225_1.fastq 10K_SRR52048229_1.fastq
10K_SRR52048225_2.fastq 10K_SRR52048229_2.fastq
```

公開データの取得

解析対象のシーケンスデータの取得方法 ⑦

シーケンスデータを確認する。

```
$ less data/10K_SRR2048224_1.fastq
```

```
@SRR2048224.1 NS500272:29:H2KGHBGXX:1:11101:4753:1025 length=80
NTGGTNCCGAAGCTCCCACTTATTCTACACCCTCTATGTCTCTTCACAATGTCAAACTAGAGTC
AAGCTCAACAGGGTCT
+SRR2048224.1 NS500272:29:H2KGHBGXX:1:11101:4753:1025 length=80
#AAAA#FFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFFFF
FAFFFFF.FFFFFFFF
                                        :
```

fastqファイルの中身を表示する。

1行目：　@配列IDと付加情報
2行目：　塩基配列
3行目：　+配列IDと付加情報
4行目：　クオリティ

※ fastqファイルは1リードあたり4行で表記される。

公開データの取得

解析対象のシーケンスデータの取得方法 ⑧
リード数を確認する。

```
$ wc -l data/10K_SRR2048224_1.fastq
40000 data/10K_SRR2048224_1.fastq
```

40,000行が抽出されていることを確認する。

```
$ wc -l data/10K_SRR2048224_2.fastq
40000 data/SRR2048224_2.fastq
```

1リードは4行なので、リード数は 40,000 / 4 = 1万リードである。

```
$ wc -l data/*
```

ワイルドカード（＊）でまとめて確認できる。

クオリティコントロール

FastQC：シーケンスクオリティチェックソフトウェア

```
$ fastqc -v
```

```
FastQC v0.10.1
```

バージョンを確認する (最新版はv0.11.5)。

```
$ fastqc -h
```

```
      FastQC - A high throughput sequence QC analysis tool

SYNOPSIS

        fastqc seqfile1 seqfile2 .. seqfileN

    fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
           [-c contaminant file] seqfile1 .. seqfileN
                                 :
```

.fastq以外に.samや.bamも指定可能、複数ファイルの指定も可能である。

クオリティコントロール

FastQCの実行

```
$ mkdir fastqc_res
$ fastqc -o fastqc_res -f fastq --nogroup ¥
  data/10K_SRR2048224_1.fastq data/10K_SRR2048224_2.fastq
$ ls fastqc_res
```

```
10K_SRR2048224_1_fastqc      10K_SRR2048224_2_fastqc
10K_SRR2048224_1_fastqc.zip  10K_SRR2048224_2_fastqc.zip
```

解析結果のhtmlファイルをブラウザ (firefox)で確認する。

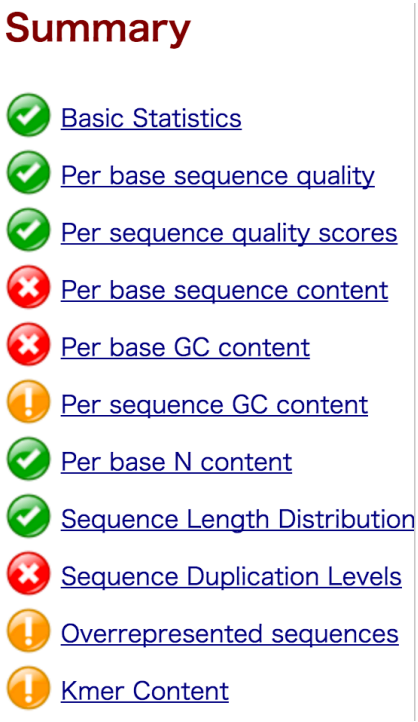
```
$ firefox ¥
 fastqc_res/10K_SRR2048224_1_fastqc/fastqc_report.html ¥
 fastqc_res/10K_SRR2048224_2_fastqc/fastqc_report.html
```

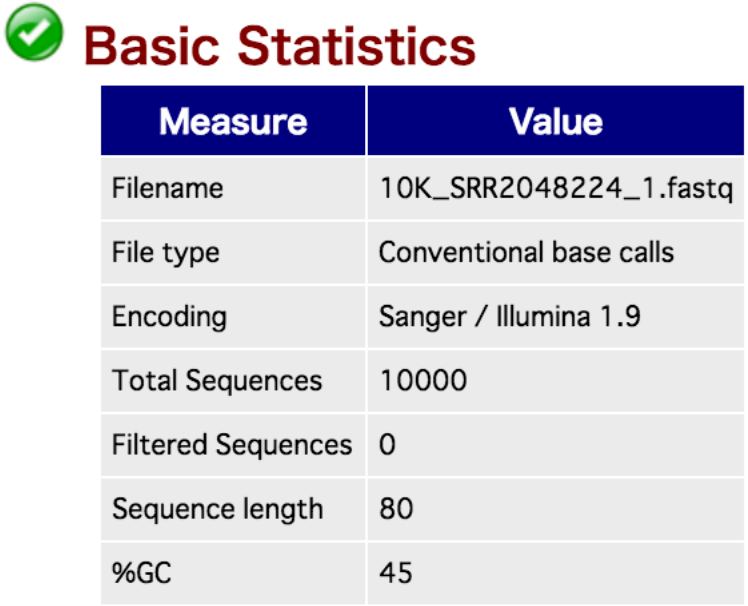
WEBブラウザ上で、クオリティチェックの解析結果が確認できる。

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

クオリティコントロール

FastQCの結果確認 ①





Basic Statistics

Measure	Value
Filename	10K_SRR2048224_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10000
Filtered Sequences	0
Sequence length	80
%GC	45

Basic Statistics
ファイルの基本的な情報。
ファイルタイプや、リード数、リード長などの情報が表示される。
ここではwarning, failureは出ない。

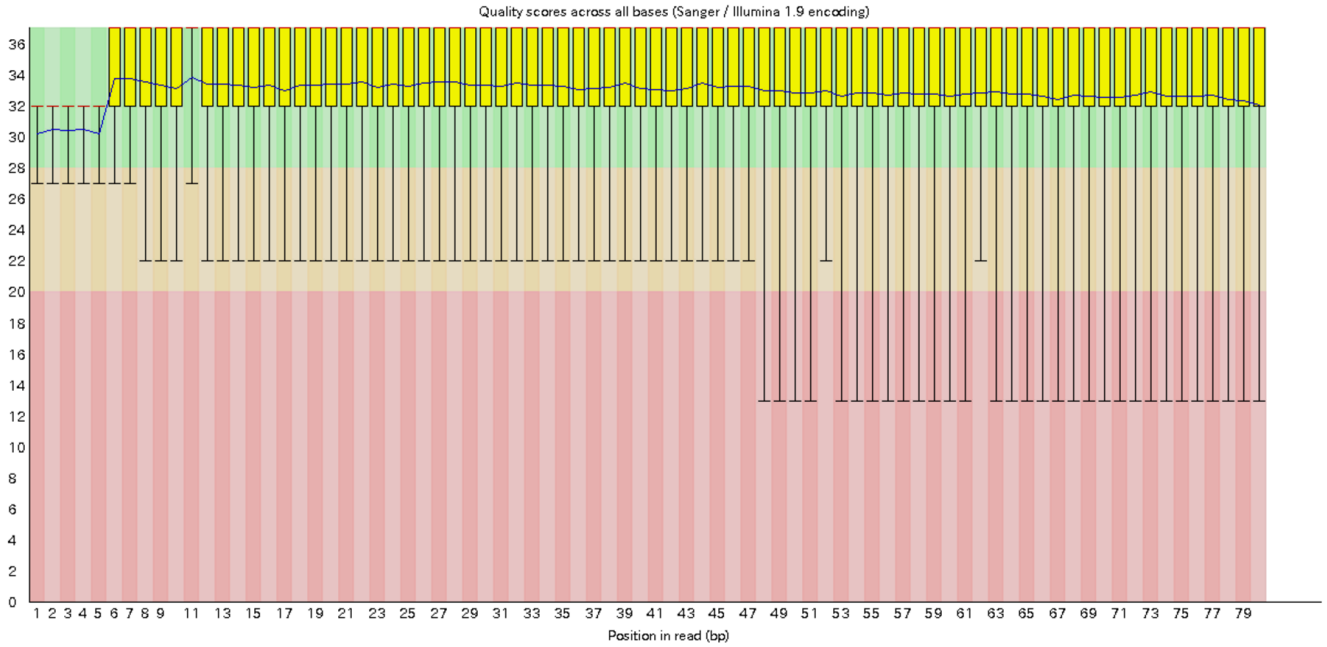
クオリティコントロール

FastQCの結果確認 ②

Per Base Sequence Quality

横軸はリード長、縦軸はquality valueを表す。
リードの位置における全体のクオリティの中央値や平均を確認できる。赤線は中央値、青線は平均値、黄色のボックスは25%～75%の領域を表す。上下に伸びた黒いバーが10%～90%の領域を意味する。

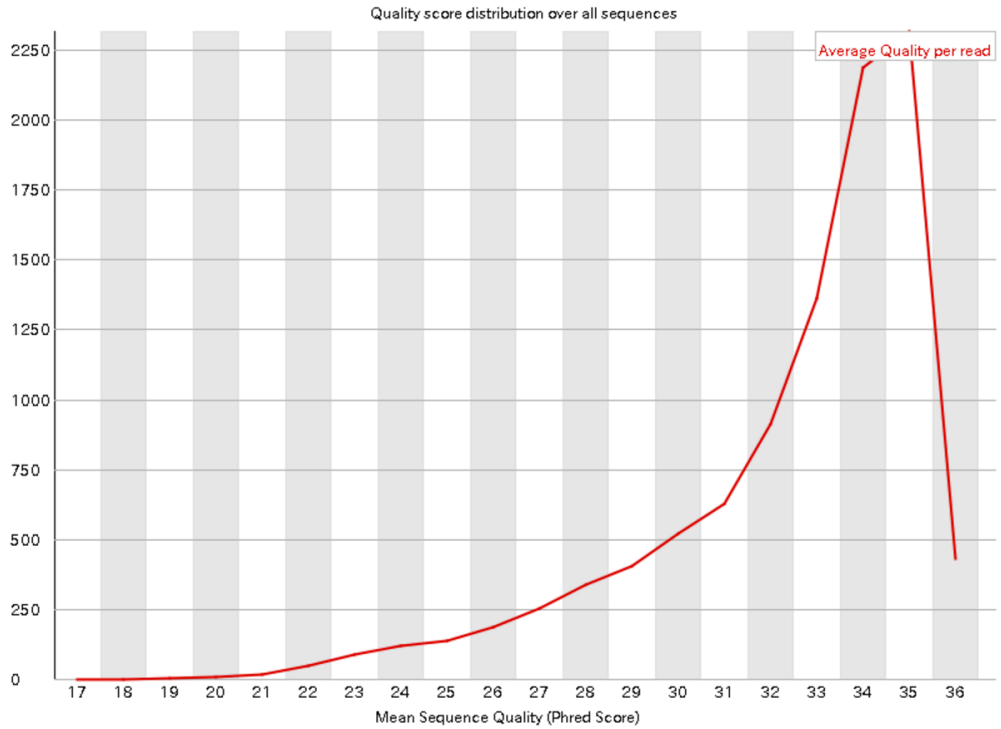
Per base sequence quality



クオリティコントロール

FastQCの結果確認 ③

Per sequence quality scores



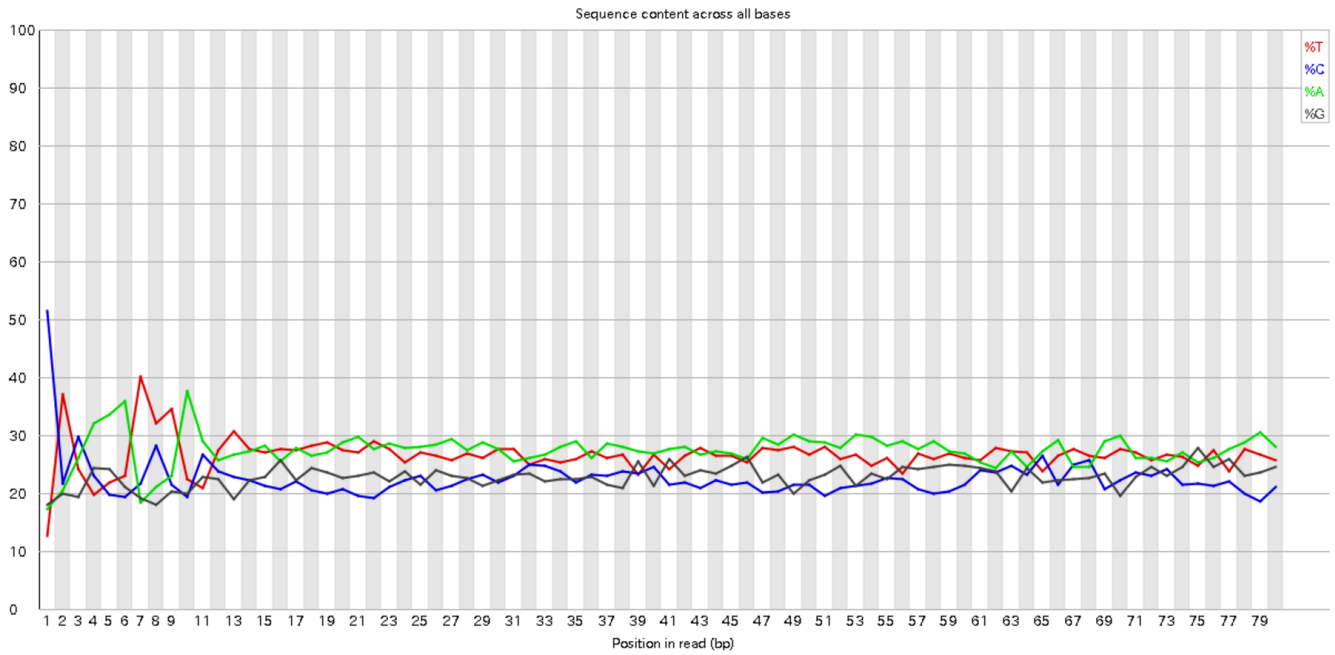
Per Sequence Quality Scores
縦軸がリード数、横軸がPhred quality scoreの平均値。

クオリティコントロール

FastQCの結果確認 ④

Per Base Sequence Content
リードにおける位置での各塩基の割合を示す。
いずれかの位置で、AとTの割合の差、もしくはGとCの割合の差が10%以上だとwarning,20%以上でfailureとなる。

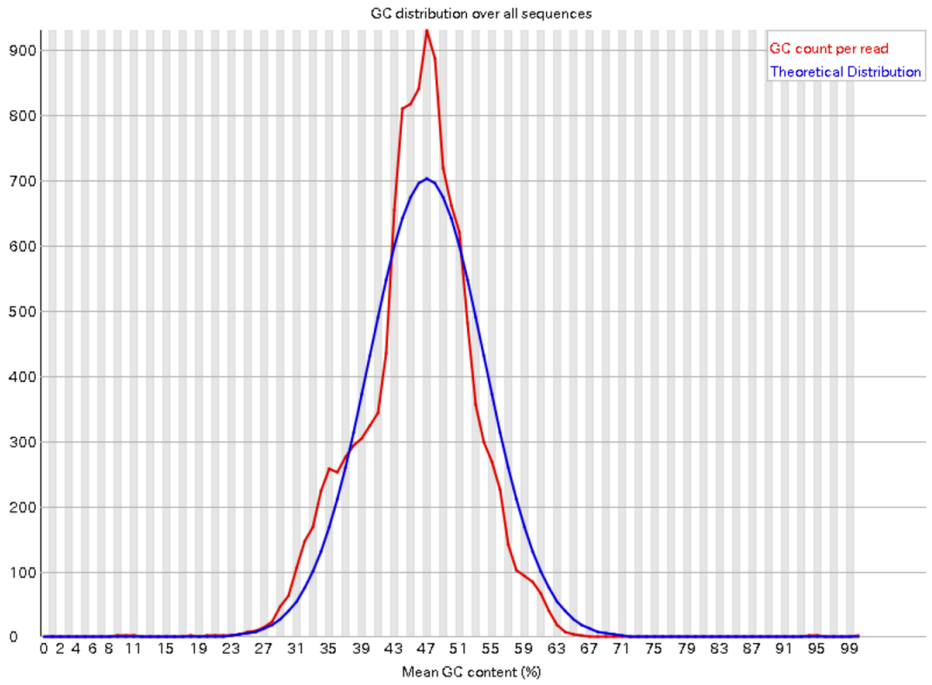
Per base sequence content



クオリティコントロール

FastQCの結果確認 ⑤

Per sequence GC content



Per Base GC Content
リードにおける位置でのGC含量を表す。
いずれかの位置で、全体でのGC含量の平均値より5%以上の差が開くとwarning, 10%でfailureとなる。

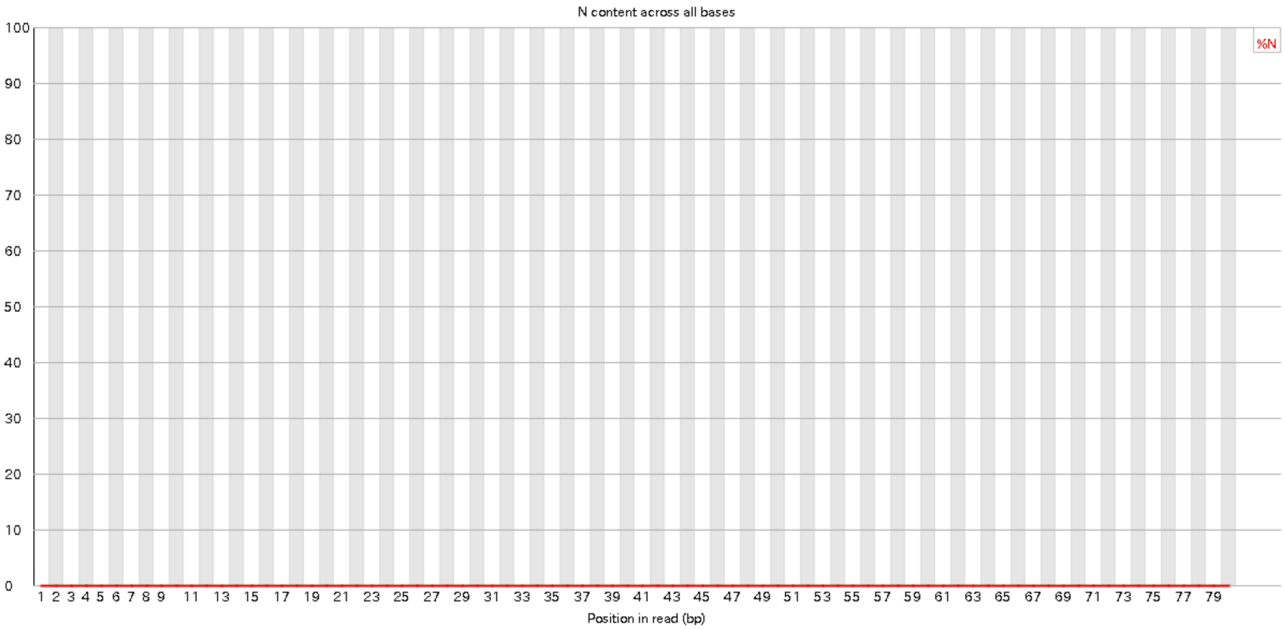
クオリティコントロール

FastQCの結果確認 ⑥

Per Base N Content

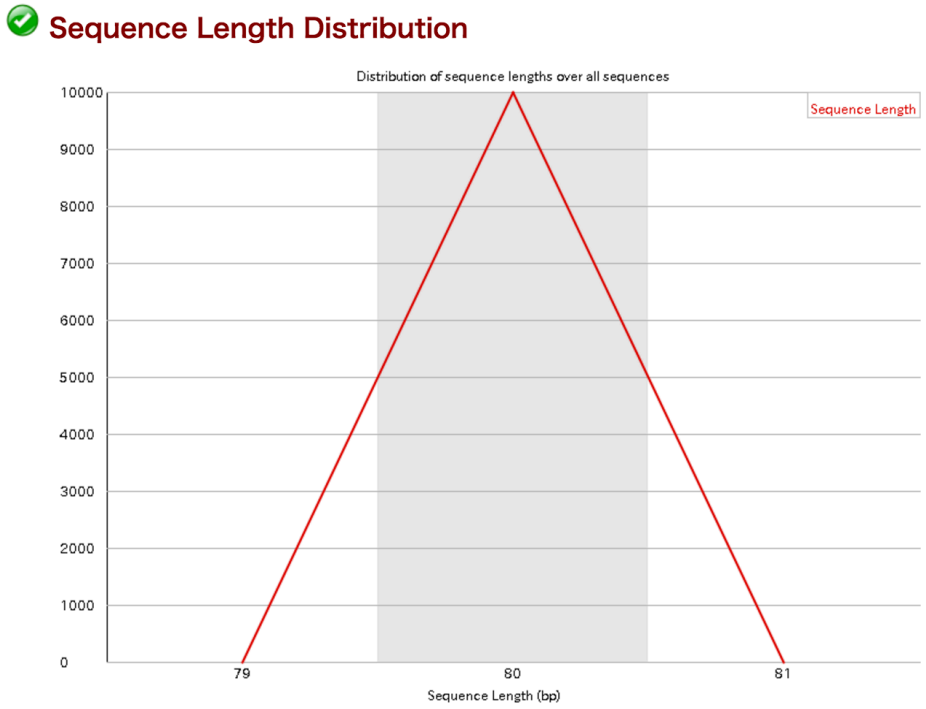
“N”はシーケンサーの問題でATGCいずれの塩基にも決定出来なかった場合に記述される。リードのいずれかの位置で5%以上Nが存在するとwarning, 20%以上でfailureとなる。

Per base N content



クオリティコントロール

FastQCの結果確認 ⑦

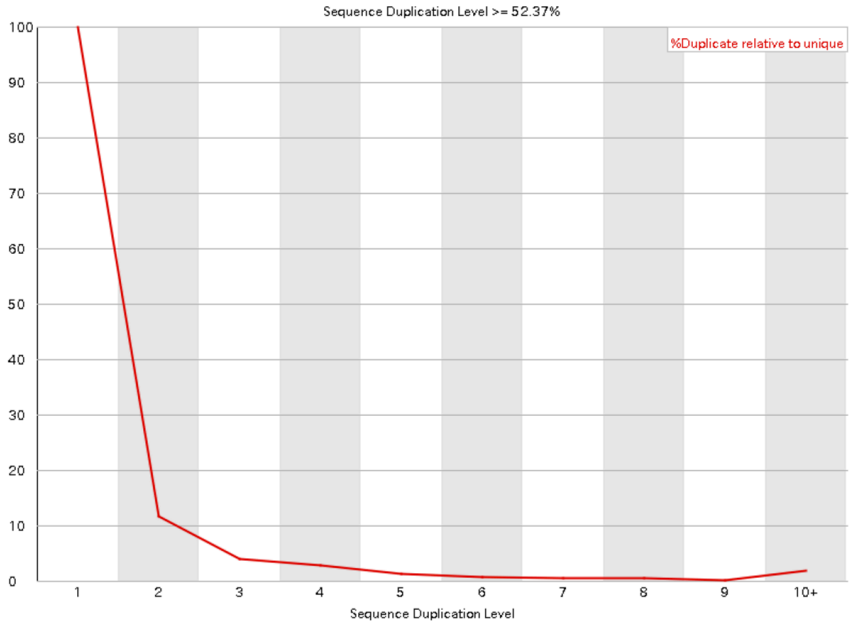


Sequence Length Distribution
リード長の全体の分布。
全てのリードの長さが同じであることを前提としており、一定でなければwarning、ゼロのものが含まれているとfailureになる。

クオリティコントロール

FastQCの結果確認 ⑧

Sequence Duplication Levels



Sequence Duplication Levels
リードの重複レベルを見ている。
1～10はそれぞれ重複のレベルで、全体の20%以上がユニークでないものだとwarning, 50%以上がユニークでないとfailureとなる。

クオリティコントロール

FastQCの結果確認 ⑨

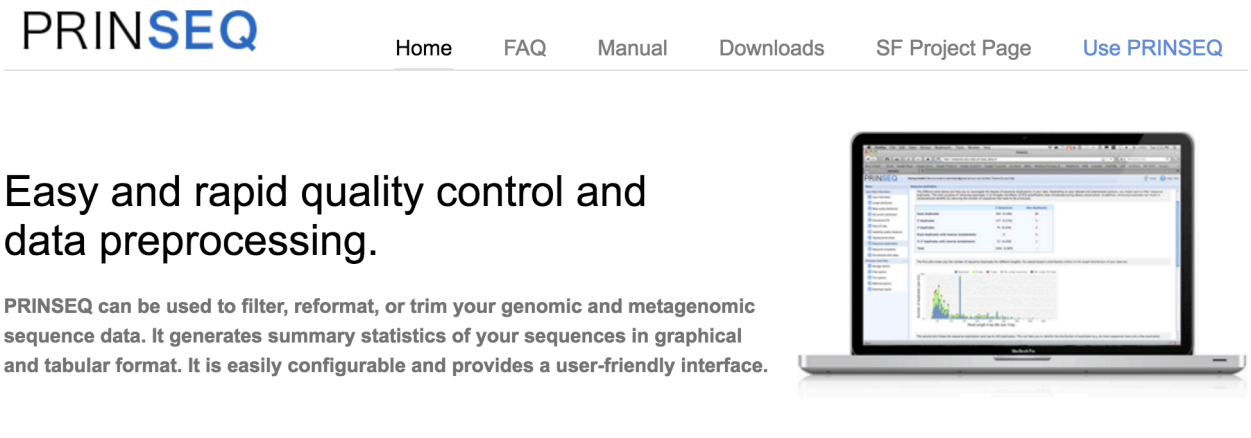
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTCAAACTTCCATCGGCTTGAAACCGATAGTCCCTCTAAGAAGTGGATAA	89	0.89	No Hit
CTAACGTCTATGCGAGTGTTTGGGTGTAAAACCCATACGCGTAATGAAAG	43	0.43	No Hit
CTAACTTTCGTTCTTGATTAATGAAAACGTCCTTGGCAAATGCTTTCGCA	42	0.42	No Hit
CAGGTCCAGACACAATAAGGATTGACAGATTGAGAGCTCTTTCTTGATTT	40	0.4	No Hit
CAGAAAGTGATGTTGACGCAATGTGATTTCTGCCCAGTGCTCTGAATGTC	37	0.37	No Hit
CTCTTTTCAAAGTTCTTTTCATCTTTCCATCACTGTACTTGTTCGCTATC	36	0.36	No Hit
GCTGAACTTAAGCATATCAATAAGCGGAGGAAAAGAAACCAACCGGGATT	36	0.36	No Hit
ACCAGGTCCAGACACAATAAGGATTGACAGATTGAGAGCTCTTTCTTGAT	33	0.33	No Hit
CTCACCAGGTCCAGACACAATAAGGATTGACAGATTGAGAGCTCTTTCTT	32	0.32	No Hit
CCAGAACCCAAAGACTTTGATTTCTCGTAAGGTGCCGAGTGGGTCATTAA	31	0.31	No Hit
CTTCCCTTTCAACAATTTCACGTACTTTTTCACTCTCTTTTCAAAGTTCT	29	0.29	No Hit
CCCTGTGGTAACTTTTCTGGCACCTCTAGCCTCAAATTCCGAGGGACTAA	29	0.29	No Hit
CTAAGGGGGGCTCATGGAGAACAGAAATCTCCAGTAGAACAAAAGGGTAA	28	0.27999999999999997	No Hit
CTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAG	27	0.27	No Hit
CTCTCTTTTCAAAGTTCTTTTCATCTTTCCATCACTGTACTTGTTCGCTA	26	0.26	No Hit
CTGAACTTAAGCATATCAATAAGCGGAGGAAAAGAAACCAACCGGGATTG	25	0.25	No Hit
CTCAAACAGGCATGCCCCCTGGAATACCAAGGGGCGCAATGTGCGTTCAA	24	0.24	No Hit

Overrepresented Sequences
重複している配列とその割合を表す。特定の配列が全リードの0.1%を超えるとwarning、1%を超えるとfailureとなる。

クオリティコントロール

PRINSEQ：クオリティコントロールソフトウェア



http://prinseq.sourceforge.net

機能

- PolyA/T tailの除去
- クオリティが低いリード末端のトリミング
- 配列長が短いリードの除去
- 片側のみのリードの除去

クオリティコントロール

PRINSEQのインストール(実行済み)

```
$ wget ¥
  https://sourceforge.net/projects/prinseq/files/standalone/prin
  seq-lite-0.20.4.tar.gz  #ソースコードをダウンロード
$ tar zxvf prinseq-lite-0.20.4.tar.gz  #展開
$ cd prinseq-lite-0.20.4  #プログラムの入っているディレクトリに移動
$ chmod +x prinseq-lite.pl  #実行権限を付与
$ ln –s /path/to/prinseq-lite-0.20.4/prinseq-lite.pl ¥
  /usr/local/bin  #PATHを通す
```

PRINSEQの使い方の確認

```
$ prinseq-lite.pl -h
```

マッピング

HISAT2：マッピングソフトウェア

http://ccb.jhu.edu/software/hisat2/index.shtml

特徴

- スプライシングを考慮してゲノム配列にマッピングする
- TopHat2よりも精度・速度ともに向上している
- メモリ消費量が少ない

Table 1 | Sensitivity and precision of leading spliced aligners

Program	No. of splice sites reported	No. of true splice sites reported	Sensitivity (%)	Precision (%)
HISATx1	91,904	85,546	97.3	93.1
HISATx2	90,331	85,603	97.3	94.8
HISAT	90,300	85,587	97.3	94.8
STAR	95,892	84,678	96.3	88.3
STARx2	92,254	84,734	96.3	91.8
GSNAP	92,547	85,598	97.3	92.5
OLego	86,779	82,879	94.2	95.5
TopHat2	96,474	79,705	90.6	82.6

Table 2 | Run times and memory usage for HISAT and other spliced aligners

Program	Run time (min)	Memory usage (GB)
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
OLego	989.5	3.7
TopHat2	1,170	4.3

Kim *et al.*, *Nature Methods*, 2015

マッピング

HISAT2のインストール(実行済み)

```
$ wget ¥
  ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/downloads/hisat2-
  2.0.4-Linux_x86_64.zip  ＃ソースコードをダウンロード
$ unzip hisat2-2.0.4-Linux_x86_64.zip  ＃展開
$ ln –s /path/to/hisat2-2.0.4/hisat2 /usr/local/bin  ＃PATHを通す
```

マッピング

HISAT2のIndexファイルのダウンロード(実行済み)

```
$ wget ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/data/sc3.tar.gz
$ tar zxvf sc3.tar.gz
$ ls /home/iu/genome/sacCer3/Hisat2Index/
genome.1.ht2  genome.3.ht2  genome.5.ht2  genome.7.ht2
genome.2.ht2  genome.4.ht2  genome.6.ht2  genome.8.ht2
```

HISAT2の使い方の確認

```
$ hisat2 -h
HISAT2 version 2.0.4 by Daehwan Kim (infphilo@gmail.com,
www.ccb.jhu.edu/people/infphilo)
Usage:hisat2 [options]* -x <ht2-idx> {-1 <m1> -2 <m2> | -U <r> |
--sra-acc <SRA accession number>} [-S <sam>]
                                    :
```

発現定量

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る

- 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

発現定量

Cufflinks：遺伝子発現解析ソフトウェア

http://cole-trapnell-lab.github.io/cufflinks

cufflinksの使い方の確認

```
$ cufflinks
```

```
cufflinks v2.2.1
linked against Boost version 105400
-----------------------------
Usage:   cufflinks [options] <hits.sam>
General Options:
  -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
  -M/--mask-file   ignore all alignment within transcripts in this file
                                  :
```

-g：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ができる。

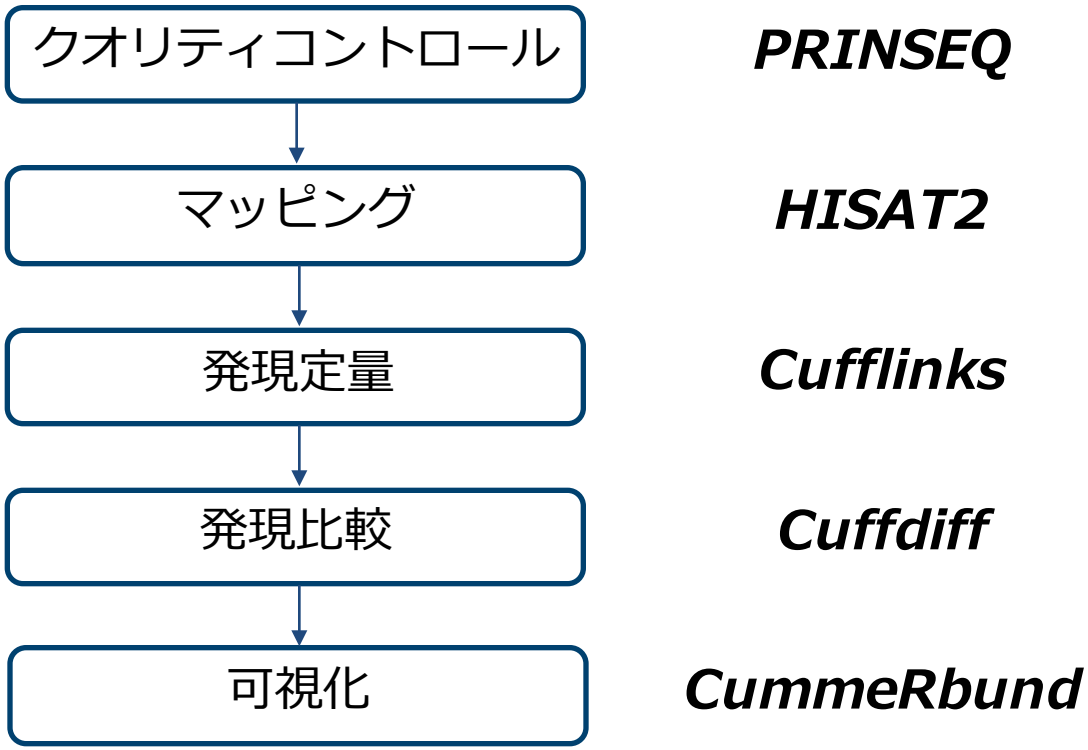
その他のRNA-seq解析

- リファレンスがある生物種の場合：
  - SNP・InDel検出（GATK, freebays）
  - 融合遺伝子検出（Chimerascan, TopHat-Fusion）
  - circular RNA検出（DCC, CIRI）
  - シングルセルRNA解析 (Cell Ranger (10x Genomics))
- リファレンスがない生物種の場合：
  - De novoアセンブリ（Trinity）

後半パート（実習）

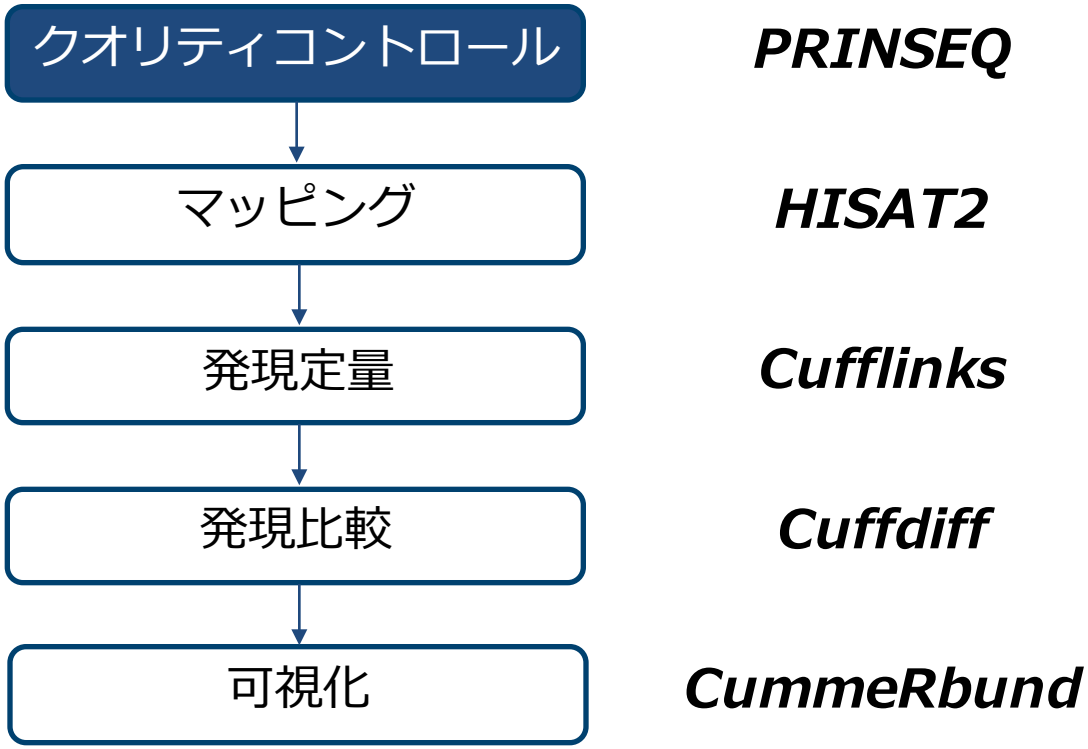
RNA-seq解析の実行

本講義でご紹介するパイプライン



RNA-seq解析の実行

本講義でご紹介するパイプライン



クオリティコントロール

PRINSEQによるクオリティコントロール

```
$ mkdir 1_qc
$ prinseq-lite.pl -fastq data/10K_SRR2048224_1.fastq ¥
  -fastq2 data/10K_SRR2048224_2.fastq ¥
  -out_good 1_qc/10K_SRR2048224.notail ¥
  -out_bad null -out_format 3 -trim_left 5 -trim_tail_right 5 ¥
  -trim_qual_right 30 -ns_max_p 20 -min_len 30
```

-fastq	入力のFASTQファイル
-fastq2	入力のFASTQファイル（ペアエンドの場合）
-out_good	フィルターを通過したリードの名前
-out_bad	フィルタリングされたリードの名前（nullは出力しない）
-out_format	1 (FASTA only), 2 (FASTA and QUAL), 3 (FASTQ), 4 (FASTQ and FASTA), or 5 (FASTQ, FASTA and QUAL)

クオリティコントロール

PRINSEQによるクオリティコントロール

```
$ mkdir 1_qc
$ prinseq-lite.pl -fastq data/10K_SRR2048224_1.fastq ¥
  -fastq2 data/10K_SRR2048224_2.fastq ¥
  -out_good 1_qc/10K_SRR2048224.notail ¥
  -out_bad null -out_format 3 -trim_left 5 -trim_tail_right 5 ¥
  -trim_qual_right 30 -ns_max_p 20 -min_len 30
```

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

-trim_tail_right	3’末端のポリテールが５以上の末端を除去
-trim_qual_right	3’末端からクオリティ30以下の塩基を除去
-ns_max_p	未知の塩基(N)が多いリード除去 (20%以上)
-min_len	配列長が短いリード除去 (30bp以下)

PRINSEQは極めて多機能なソフトウェアであり、クオリティチェックからトリミング、フィルタリングまで様々なプロセスが可能

クオリティコントロール

FastQCの実行

```
$ fastqc -o fastqc_res -f fastq --nogroup ¥
  1_qc/10K_SRR2048224.notail_1.fastq ¥
  1_qc/10K_SRR2048224.notail_2.fastq
```

解析結果のhtmlファイルをブラウザ (firefox)で確認する。

```
$ firefox fastqc_res/10K_SRR2048224.notail_1_fastqc.html
```

Basic Statistics

Measure	Value
Filename	10K_SRR2048224_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10000
Filtered Sequences	0
Sequence length	80
%GC	45

クリーニング前

Basic Statistics

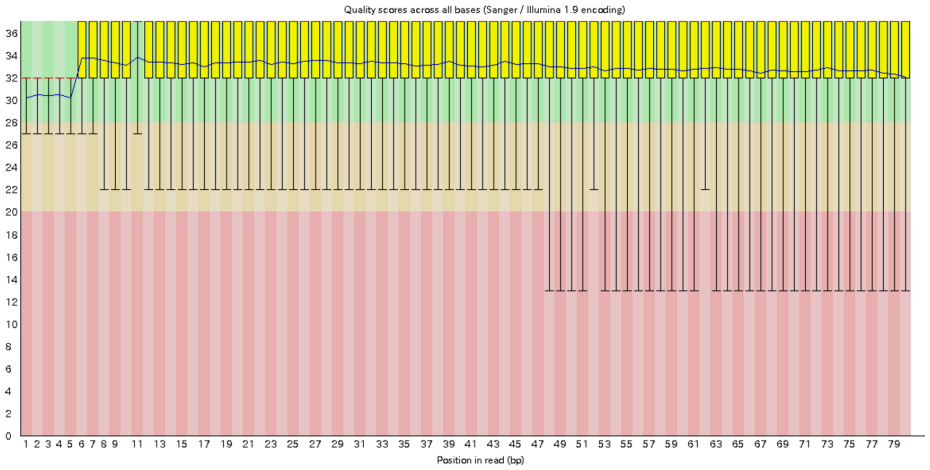
Measure	Value
Filename	10K_SRR2048224.notail_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9992
Filtered Sequences	0
Sequence length	46-75
%GC	45

クリーニング後

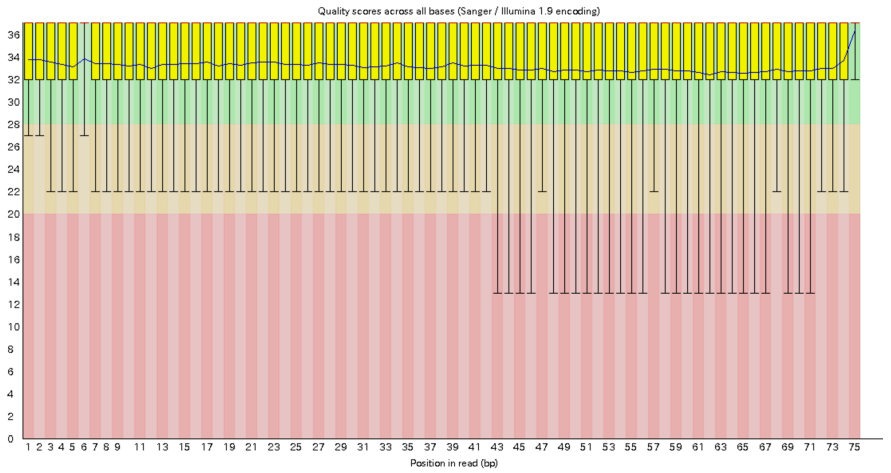
クオリティコントロール

リードクオリティの確認

Per base sequence quality



Per base sequence quality

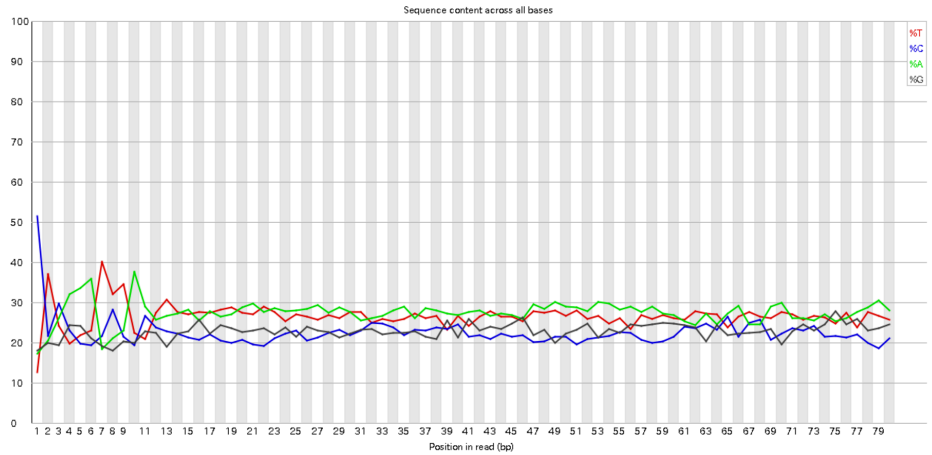


クリーニング前 → クリーニング後

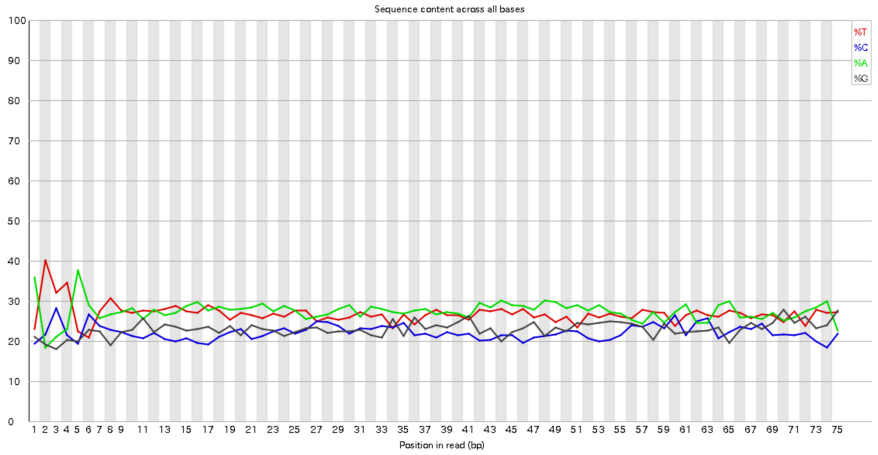
クオリティコントロール

各塩基の含有率の確認

Per base sequence content



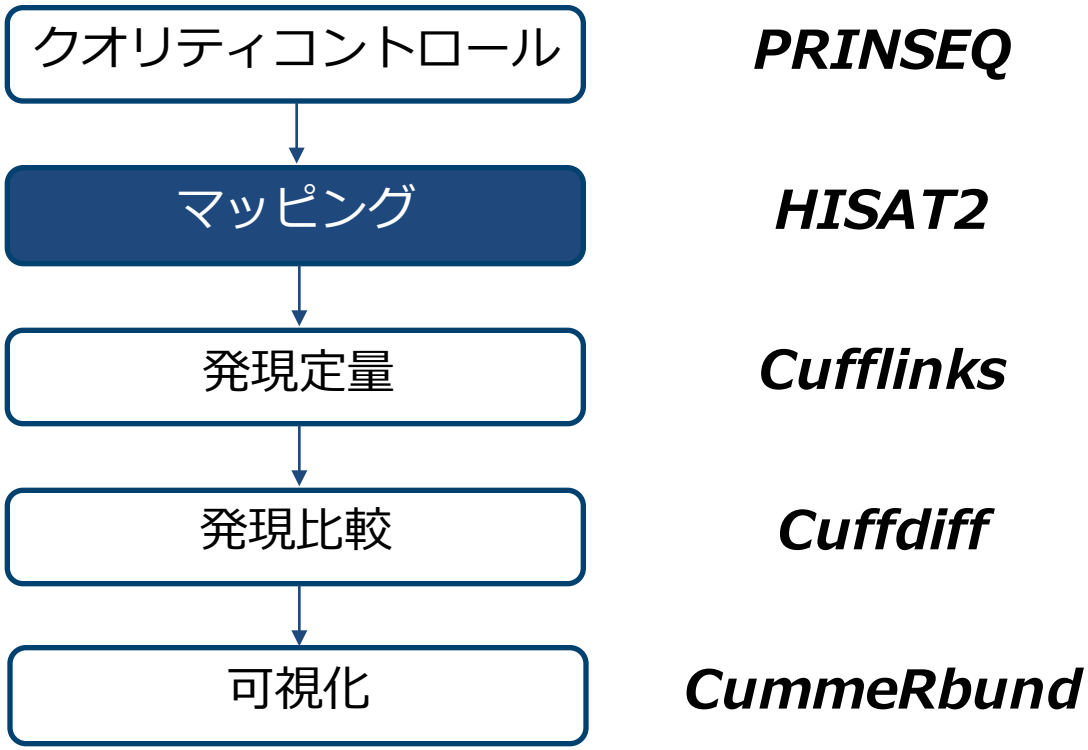
Per base sequence content



クリーニング前 → クリーニング後

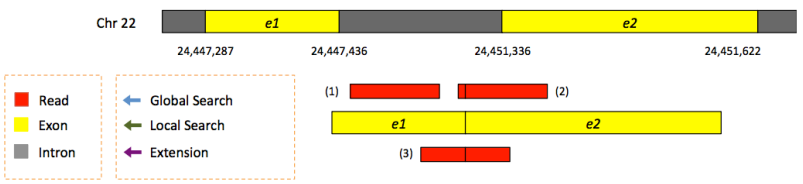
RNA-seq解析の実行

本講義でご紹介するパイプライン



マッピング

HISATのアルゴリズム



特徴

- スプライシングを考慮してゲノム配列にマッピングする
- hierarchical indexingを用いることで高速で高感度のアライメントが可能

Kim *et al., Nature Methods,* 2015

マッピング

```
$ mkdir 2_mapping
$ hisat2 -x /home/iu/genome/sacCer3/Hisat2Index/genome --dta ¥
  --dta-cufflinks -1 1_qc/10K_SRR2048224.notail_1.fastq ¥
  -2 1_qc/10K_SRR2048224.notail_2.fastq ¥
  -S 2_mapping/10K_SRR2048224.sam
```

--dta	マッピング結果からアセンブリを行う
--dta-cufflinks	cufflinksのためのアセンブリを行う
-S	SAMファイルに書き出す名前

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

マッピング

```
$ mkdir 2_mapping
$ hisat2 -x /home/iu/genome/sacCer3/Hisat2Index/genome --dta ¥
  --dta-cufflinks -1 1_qc/10K_SRR2048224.notail_1.fastq ¥
  -2 1_qc/10K_SRR2048224.notail_2.fastq ¥
  -S 2_mapping/10K_SRR2048224.sam
```

```
9992 reads; of these:
  9992 (100.00%) were paired; of these:
    2174 (21.76%) aligned concordantly 0 times
    193 (1.93%) aligned concordantly exactly 1 time
    7625 (76.31%) aligned concordantly >1 times
    ----
    2174 pairs aligned concordantly 0 times; of these:
      3 (0.14%) aligned discordantly 1 time
    ----
    2171 pairs aligned 0 times concordantly or discordantly; of these:
      4342 mates make up the pairs; of these:
        2925 (67.37%) aligned 0 times
        99 (2.28%) aligned exactly 1 time
        1318 (30.35%) aligned >1 times
85.36% overall alignment rate
```

マッピング

SAMファイルをBAMファイルに変換

```
$ samtools view -b 2_mapping/10K_SRR2048224.sam ¥
  > 2_mapping/10K_SRR2048224.bam
$ ls -lh
```

```
-rw-rw-r-- 1 iu iu 1.7M  5月 31 13:52 2016 10K_SRR2048224.bam
-rw-rw-r-- 1 iu iu  13M  5月 31 14:54 2016 10K_SRR2048224.sam
```

13MのSAMファイルが1.7Mのバイナリファイルに変換される。

BAMファイルをソート

```
$ samtools sort 2_mapping/10K_SRR2048224.bam ¥
  -o 2_mapping/10K_SRR2048224.sorted.bam
$ ls
```

```
10K_SRR2048224.bam   10K_SRR2048224.sorted.bam
10K_SRR2048224.sam
```

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

マッピング結果の可視化

BAMファイルのインデックスを作成

```
$ samtools index 2_mapping/10K_SRR2048224.sorted.bam
$ ls 2_mapping
```

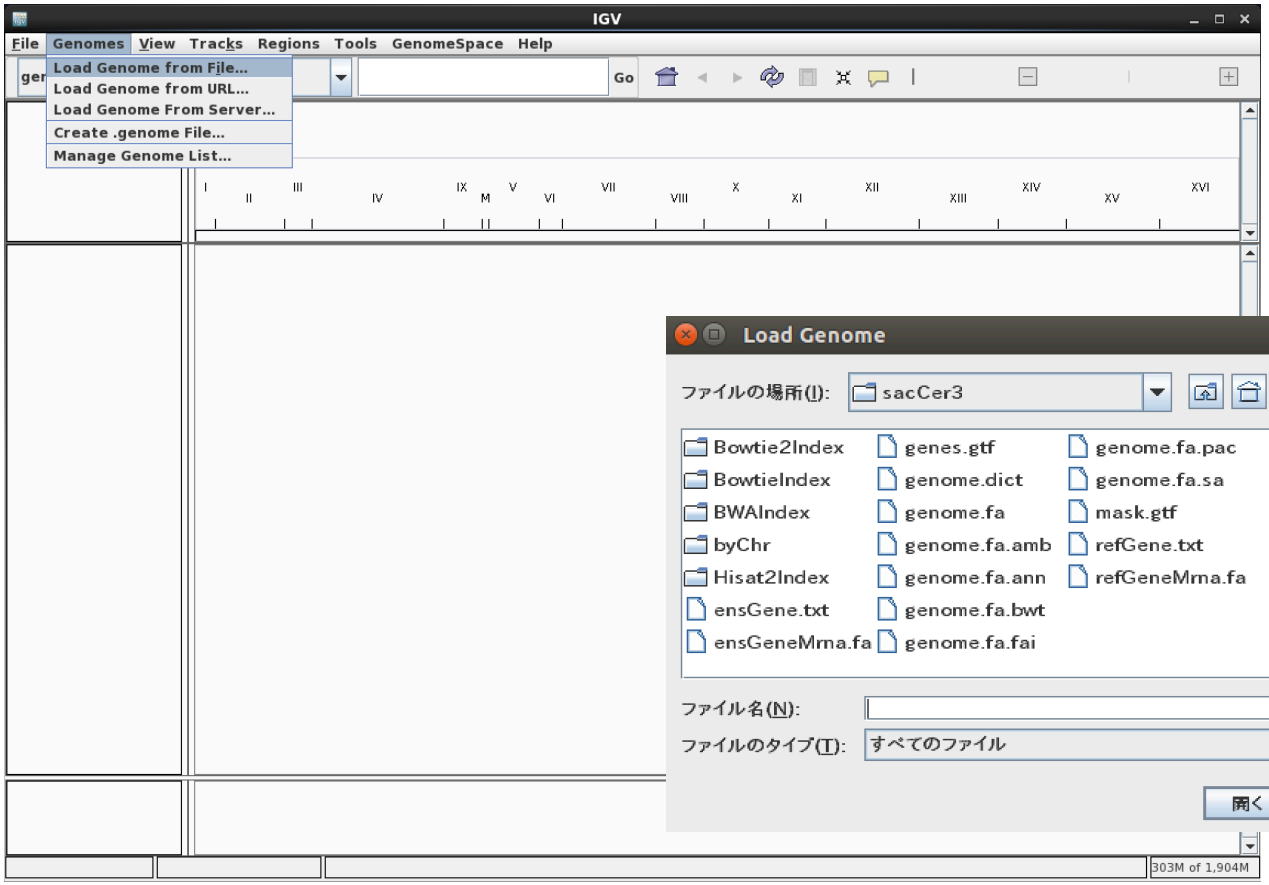
```
10K_SRR2048224.bam
10K_SRR2048224.sam
10K_SRR2048224.sorted.bam
10K_SRR2048224.sorted.bam.bai
             :
```

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

マッピング結果の可視化

Integrative Genomics Viewer (IGV)を用いた解析結果の確認 ①

```
$ igv.sh
```

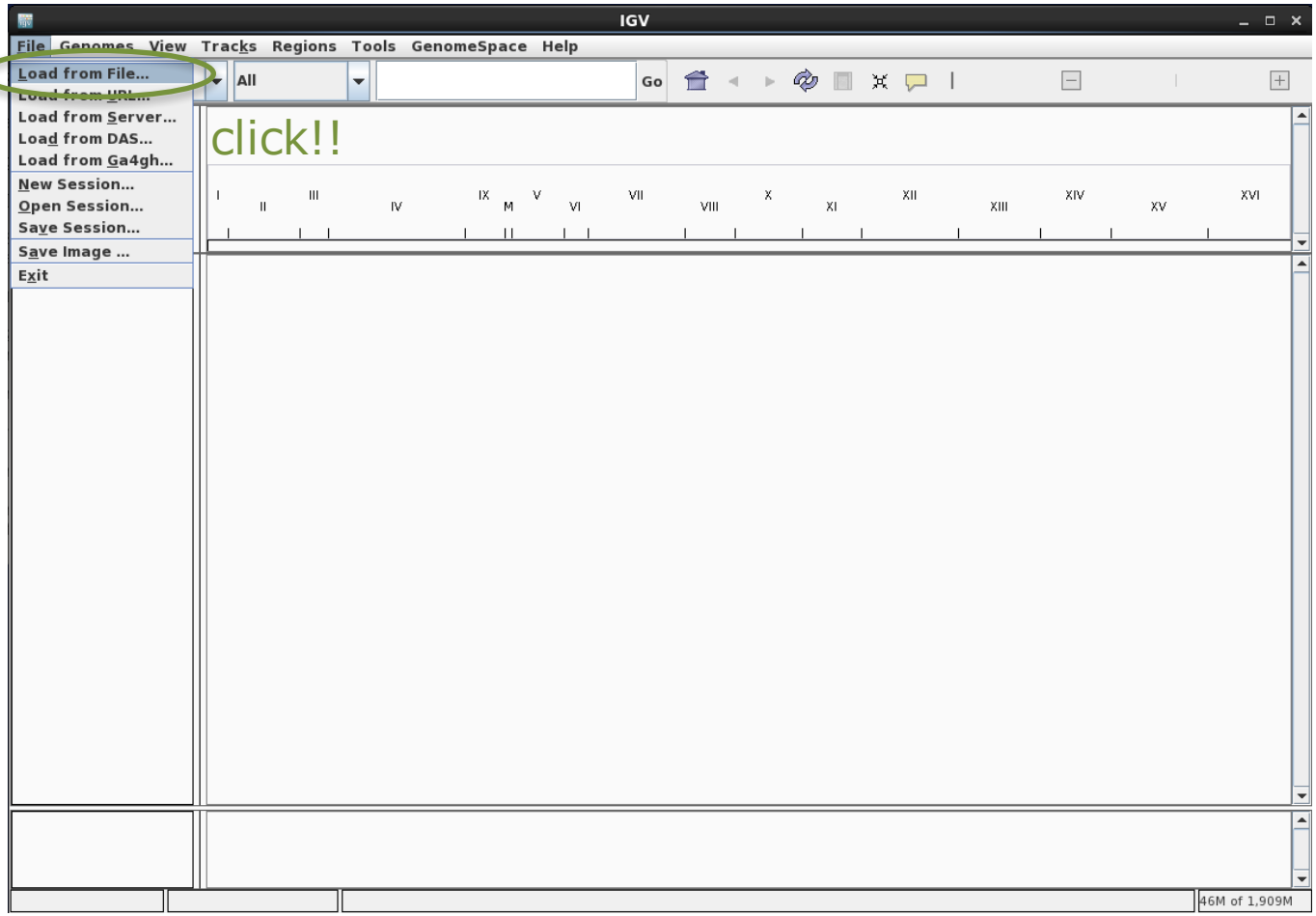


IGVを起動し、Genomesタブから Load Genomes from File...を選択。

/home/iu/genome/sacCer3の下にあるgenome.faを選択し開く。

マッピング結果の可視化

Integrative Genomics Viewer (IGV)を用いた解析結果の確認 ②

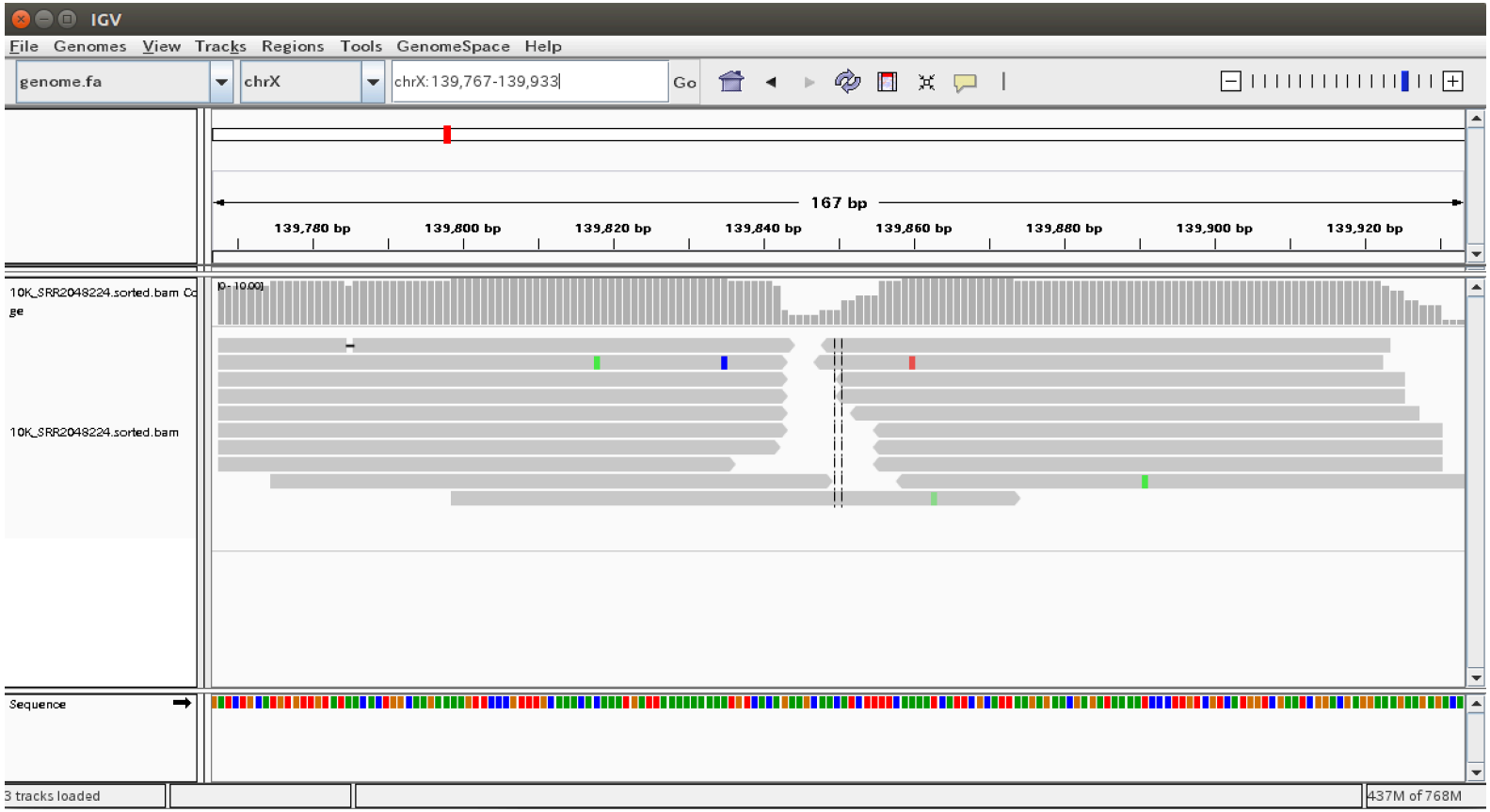


Fileタブから
Load from File...
を選択。

ソート済みの
bamファイルを
選択し開く。

マッピング結果の可視化

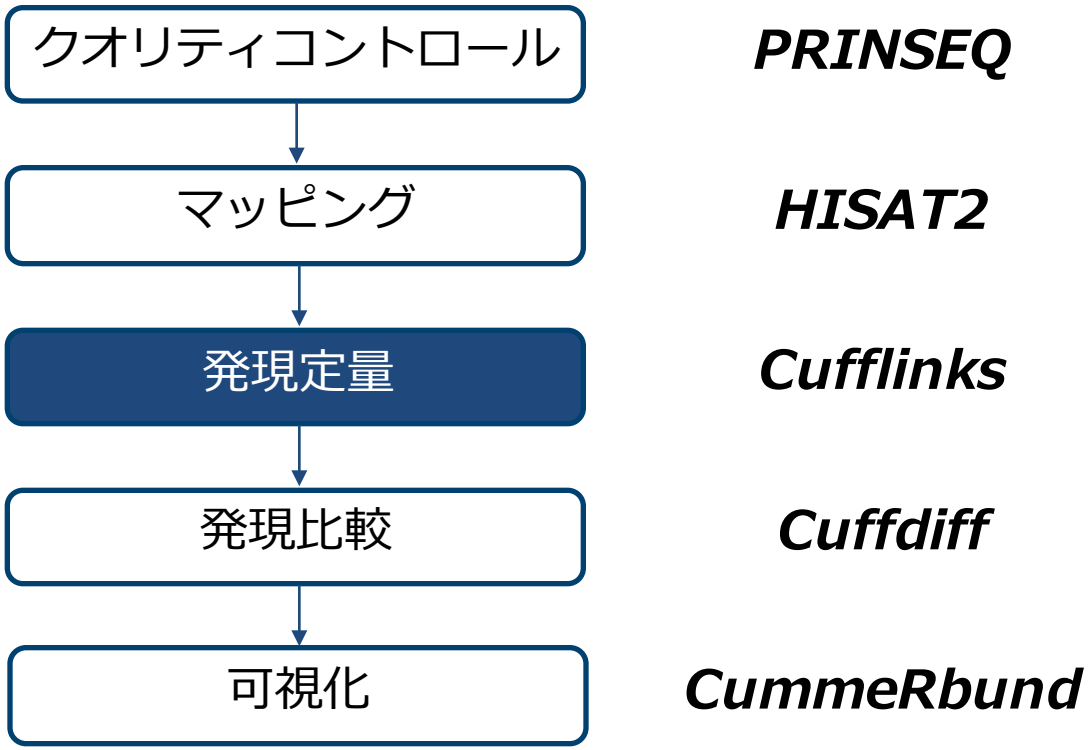
Integrative Genomics Viewer (IGV)を用いた解析結果の確認 ③



サーチウィンドウにchrX:139,767-139,933と入力。

RNA-seq解析の実行

本講義でご紹介するパイプライン



発現定量

cufflinksの実行

```
$ cufflinks -o SRR2048224 --min-frags-per-transfrag 2 ¥
  2_mapping/10K_SRR2048224.sorted.bam
$ ls SRR2048224
```

```
genes.fpkm_tracking    skipped.gtf
isoforms.fpkm_tracking transcripts.gtf
```

genes.fpkm_trackingの確認

```
$ less genes.fpkm_tracking
```

4列目にGene ID、10列目にFPKMが記載されている。

SRR2048225、SRR2048228、SRR2048229についても同様に処理する。

発現定量

transcripts.gtf.txtの作成

```
$ vim transcripts.gtf.txt
```

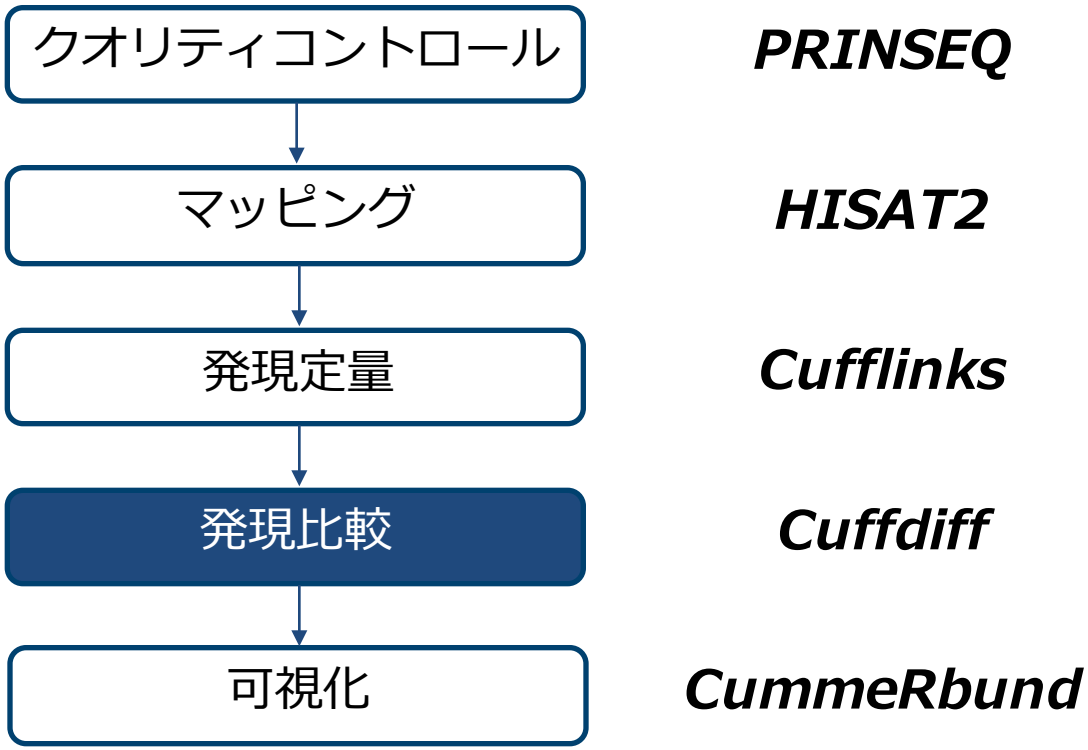
挿入モード（i）で以下を記入

```
SRR2048224/transcripts.gtf
SRR2048225/transcripts.gtf
SRR2048228/transcripts.gtf
SRR2048229/transcripts.gtf
```

挿入モードの終了：エスケープ（ESC）
コマンドモード：コロン（：）
保存：コマンドモードで w + Enter
終了：コマンドモードで q + Enter
保存せずに終了：コマンドモードで q! + Enter

RNA-seq解析の実行

本講義でご紹介するパイプライン



発現比較

cuffmergeの実行

```
$ cuffmerge -o COMPARE ¥
  -g /home/iu/genome/sacCer3/genes.gtf ¥
  -s /home/iu/genome/sacCer3/genome.fa ¥
  transcripts.gtf.txt
```

-o/--output-dir	出力ディレクトリ
-g/--ref-gtf	アノテーション用のgtfファイル
-s/--ref-sequence	リファレンスゲノムFASTAファイル
-p/--num-threads	スレッド数（デフォルト＝１）

発現比較

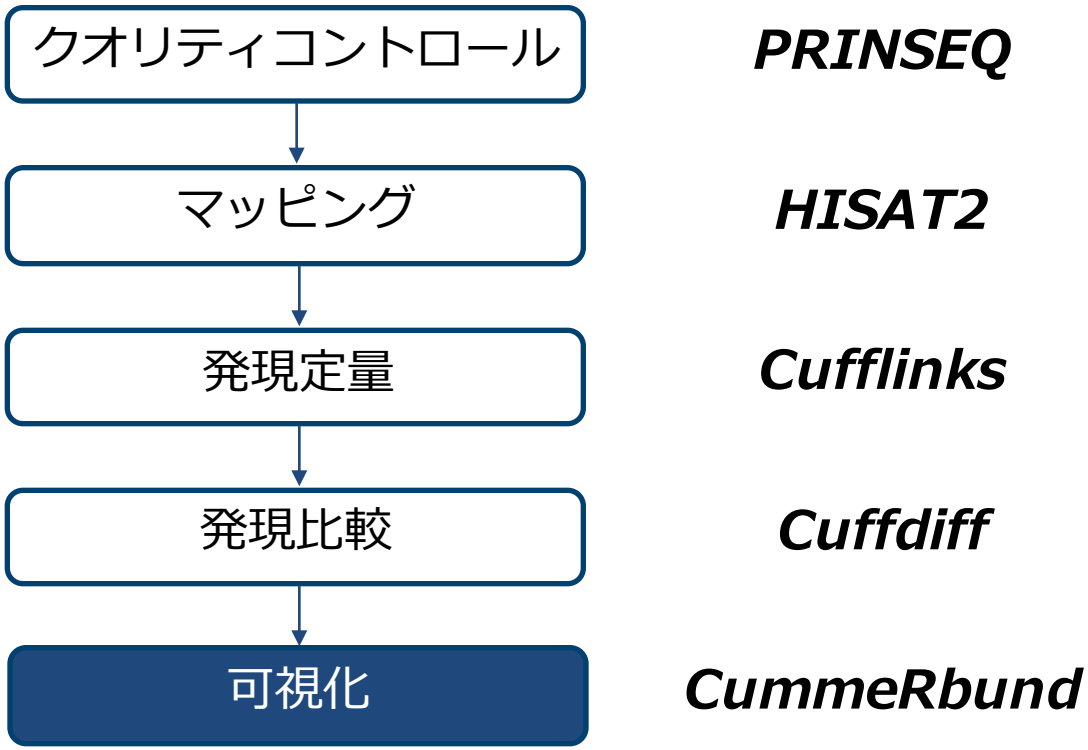
cuffdiffの実行

```
$ cuffdiff -o COMPARE -L Group1,Group2 COMPARE/merged.gtf ¥
  2_mapping/10K_SRR2048224.sorted.bam,¥
  2_mapping/10K_SRR2048225.sorted.bam ¥
  2_mapping/10K_SRR2048228.sorted.bam,¥
  2_mapping/10K_SRR2048229.sorted.bam
```

-o/--output-dir	アノテーション用のgtfファイル
-L/--labels	グループの指定（カンマ区切り）
-p/--num-threads	スレッド数（デフォルト＝１）

RNA-seq解析の実行

本講義でご紹介するパイプライン



可視化

cummeRbundの紹介

http://compbio.mit.edu/cummeRbund/index.html

Cufflinksの結果を用いて可視化を行うRパッケージ

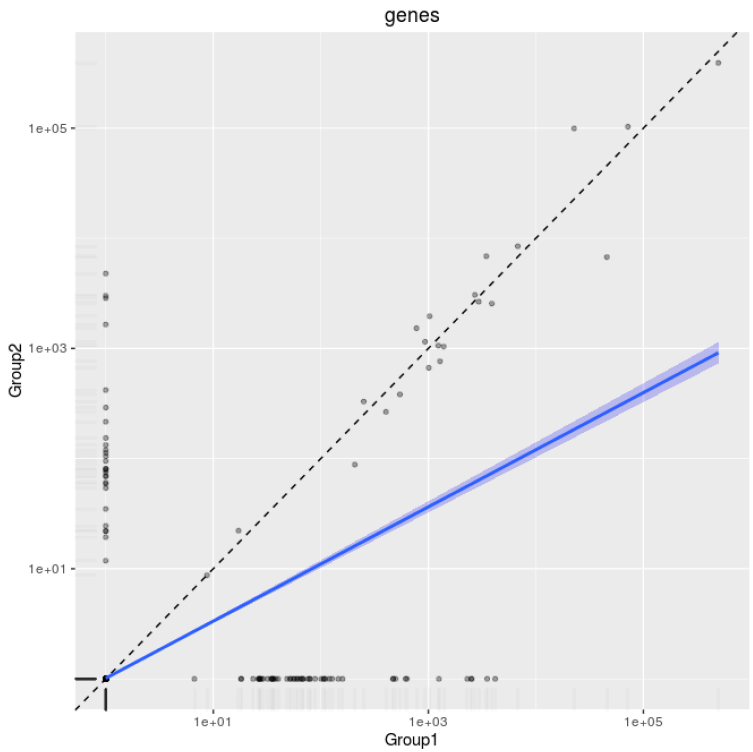
```
$ R   #Cufflinksの実行ディレクトリ（COMPARE）で起動する
> library(cummeRbund)
> cuff <- readCufflinks()
> cuff
```

```
CuffSet instance with:
         2 samples          # サンプル数
         6935 genes         # 遺伝子数
         7077 isoforms      # 転写産物数
         7052 TSS           # 転写開始位置数
         6643 CDS           # コード領域数
         6935 promoters     # プロモーター数
         7052 splicing      # スプライシング領域数
         6534 relCDS        # 調節コード領域
```

可視化

cummeRbundの紹介

```
> s <- csScatter(genes(cuff), "Group1", "Group2", smooth=T)
> s
```



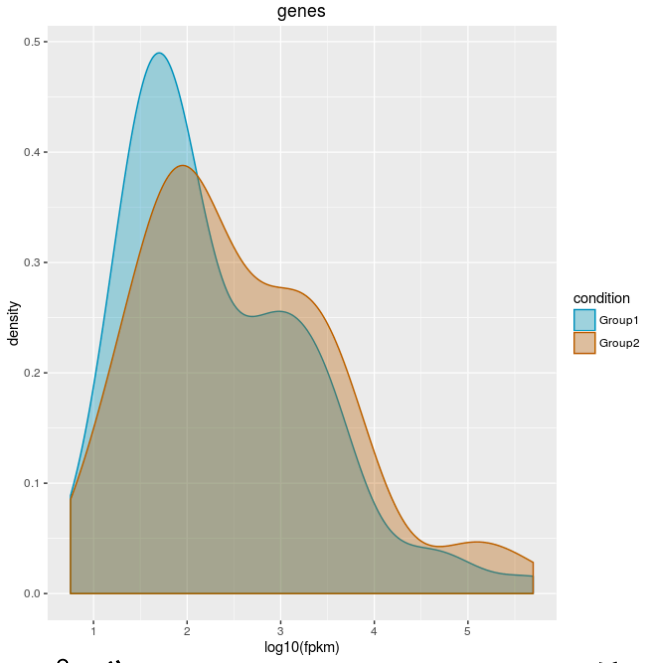
Scatter Plot

グループ間における
遺伝子発現の偏りを示す

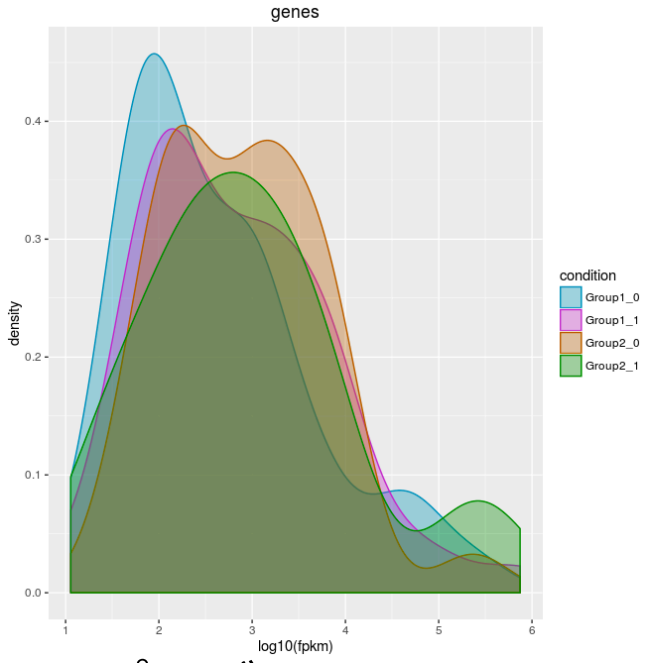
可視化

cummeRbundの紹介

```
> dens <- csDensity(genes(cuff))
> dens
> densRep <- csDensity(genes(cuff),replicates=T)
> densRep
```



グループごとの$\log_{10}$(fpkm)の分布



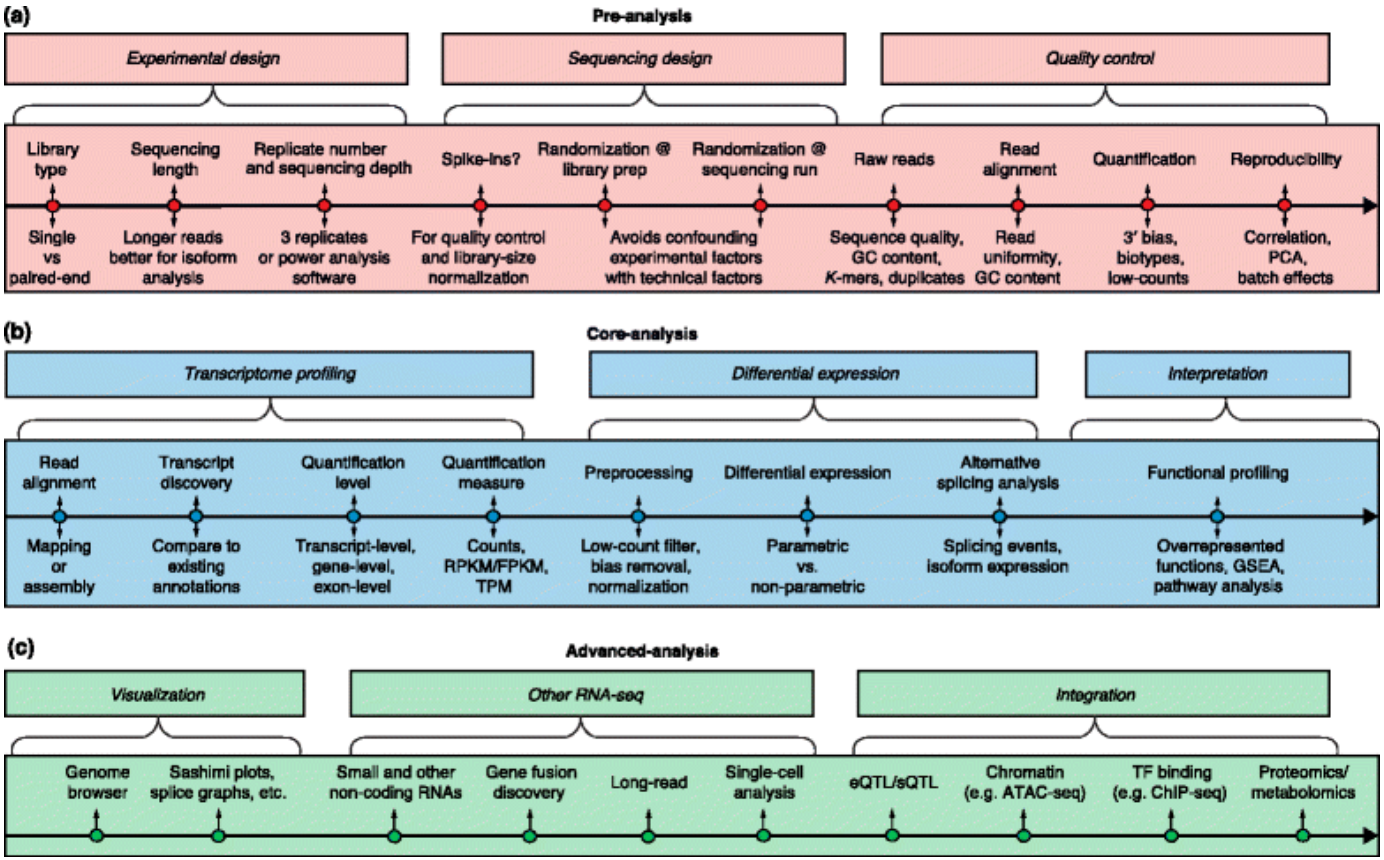
サンプルごとの$\log_{10}$(fpkm)の分布

まとめ

本日行った解析のおさらい。

最後に

本講義でご紹介した流れは、解析方法の一例です。ツールの選択に「正解」はありません。自身のデータに適したツールを選択し、より良い解析手順を確立していってください。



Conesa *et al., Genome Biology,* 2016