NGS解析基礎

ITのチカラで研究を支援

amelieff

アメリエフ株式会社

講 義 内 容

- ファイル形式
- データの可視化
- データのクオリティチェック
- マッピング
- アセンブル

資料の見方

```
$ pwd
```

※実際に入力するコマンドを黄色い四角の中に示します

ファイル形式

- NGS解析でよく使われるファイル形式

ファイル形式	サンプルデータの場所
fastq	/home/ユーザ名/Desktop/amelieff/1K_ERR038793_1.fastq
bam/sam	/home/ユーザ名/Desktop/amelieff/1K_ERR038793.bam
vcf	/home/ユーザ名/Desktop/amelieff/1K_ERR038793_sort.vcf
bed	/home/ユーザ名/Desktop/amelieff/1K_ERR038793.bed （講義中に作成）
fasta	/home/ユーザ名/Desktop/amelieff/Scerevisiae/WholeGenomeFasta/genome.fa

ファイル形式 | fastq

- シーケンサから出力されるリード情報

```
$ less 1K_ERR038793_1.fastq
```

4行で1リード

```
@ERR038793.1 HS19_6178:5:1208:12689:35298#1 length=100
GGACAAGGTTACTTCCTAGATGCTATATGTCCCTACGGCCTTGTCTAACACCATCC
+ERR038793.1 HS19_6178:5:1208:12689:35298#1 length=100
D/DDBD@B>DFFEEEEEEEEF@FDEEEBEDBBDDD:AEEE<>CB?FCFF@F?FBFF
          :
```

	必須の情報	オプション
1行め	@から始まる配列ID	付加情報
2行め	リードの塩基配列	
3行め	+	配列ID、または1行めと同じ情報
4行め	リードのクオリティ	

ファイル形式｜fastq

- fastqのクオリティは、「記号のASCIIコード－33」と対応する。

（例）クオリティ値：**%** → 37－33＝**4**

ASCIIコード表

33:!	34:"	35:#	36:$	37:%	38:&	39:'	40:(
41:)	42:*	43:+	44:,	45:-	46:.	47:/	48:0
49:1	50:2	51:3	52:4	53:5	54:6	55:7	56:8
57:9	58::	59:;	60:<	61:=	62:>	63:?	64:@
65:A	66:B	67:C	68:D	69:E	70:F	71:G	72:H
73:I	74:J	75:K	76:L	77:M	78:N	79:O	80:P
81:Q	82:R	83:S	84:T	85:U	86:V	87:W	88:X
89:Y	90:Z	91:[	92:\	93:]	94:^	95:_	96:`
97:a	98:b	99:c	100:d	101:e	102:f	103:g	104:h
105:i	106:j	107:k	108:l	109:m	110:n	111:o	112:p
113:q	114:r	115:s	116:t	117:u	118:v	119:w	120:x
121:y	122:z	123:{	124:|	125:}	126:~		

ファイル形式 | bam / sam

- リードをゲノムにマッピングしたアライメント情報
  - **sam**: テキストデータ
  - **bam**: 圧縮したsam。コンピュータが扱いやすいバイナリデータ
- 相互変換には主に**samtools**というソフトを用いる

■ samからbam

入力がsam、出力がbam

```
samtools view -Sb sam > bam
```

■ bamからsam

ヘッダ付で出力

```
samtools view -h bam > sam
```

```
$ samtools view -h 1K_ERR038793.bam > 1K_ERR038793.sam
$ ls
```

ファイル形式 | bam / sam

- samファイルの中身
  - @から始まるヘッダ行と、1行に1リードの情報がタブ区切りで記載されているデータ行からなる

```
$ less 1K_ERR038793.sam
```

ヘッダ行

```
@SQ     SN:I    LN:230218
@SQ     SN:II   LN:813184
@SQ     SN:III  LN:316620
                          :
```

ファイル形式 | bam / sam

- samファイルの中身
  - @から始まるヘッダ行と、1行に1リードの情報がタブ区切りで記載されているデータ行からなる

```
$ less 1K_ERR038793.sam
```

1行で1リード

```
                                        :
ERR038793.1 113 XII 1065143 4   12M4I84M    I   150 0
    AGGGTGTGGTGTGTGGGTATATCTATGTCACCTTATTGCATGCTGGATGGTGTTAG
ACAAGGCCGTAGGGACATATAGCATCTAGGAAGTAACCTTGTCC
    CD;?C@FEFEFFFFFDC8=DA=?>>.EEE=BEEEBEE:EEE:?@FFBF?F@FFCF?
BC><EEEA:DDDBBDEBEEEDF@FEEEEEEEEFFD>B@DBDD/D  NM:i:6
    MD:Z:0T93A1 AS:i:83 XS:i:80 RG:Z:ERR038793  XA:Z:V,-
570330,18S82M,1;
                                        :
```

ファイル形式 | bam / sam

- samファイルの中身
  - 最初の11列は必須である

列	項目	意味	例
1	QNAME	リード名	ERR038793.1
2	FLAG	フラグ	113
3	RNAME	染色体名	XII
4	POS	リードのスタートポジション	1065143
5	MAPQ	マッピングクオリティ	4
6	CIGAR	CIGAR	12M4I84M
:	:	:	:

ファイル形式 | bam / sam

- samファイルの中身

列	項目	意味	例
:	:	:	:
7	RNEXT	ペアリードがある染色体名	I
8	PNEXT	ペアリードのスタート位置	150
9	TLEN	ペア間の距離＋各リード長	0
10	SEQ	リード配列	AGGGTGTGGTGTGTGGGTATATCTATGTCACCTTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGTAACCTTGTCC
11	QUAL	リードクオリティ	CD;?C@FEFEFFFFFDC8=DA=?>>.EEE=BEEEBEE:EEE:?@FFBF?F@FFCF?BC><EEEA:DDDBBDEBEEEDF@FEEEEEEEEFFD>B@DBDD/D
:	:	:	:

ファイル形式 | vcf

- 変異の情報
  - # で始まるヘッダ行と、1行に1つの変異の情報がタブ区切りで記載されているデータ行から成る

```
$ less 1K_ERR038793_sort.vcf
```

ヘッダ行

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic
depths for the ref and alt alleles in the order listed">
                                    :
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count
in genotypes, for each ALT allele, in the same order as
listed">
                                    :
##reference=file:///home/genome/genome.fa
#CHROM  POS ID  REF ALT QUALFILTER  INFOFORMAT  ERR038793
                                    :
```

ファイル形式 | vcf

- 変異の情報
  - # で始まるヘッダ行と、1行に1つの変異の情報がタブ区切りで記載されているデータ行から成る

```
$ less 1K_ERR038793_sort.vcf
```

1行で1変異

```
                                    :
                                    :
I    111 .    C    T    105.93   .
     AC=1;AF=0.50;AN=2;BaseQRankSum=0.729;DP=9;Dels=0.00;FS
=0.000;HRun=1;HaplotypeScore=0.0000;MQ=59.16;MQ0=0;MQRankS
um=-1.159;QD=11.77;ReadPosRankSum=-0.361;SB=-0.01
     GT:AD:DP:GQ:PL  0/1:5,4:9:99:136,0,173
                                    :
                                    :
```

ファイル形式 | vcf

- 変異の情報
  - # で始まるヘッダ行と、1行に1つの変異の情報がタブ区切りで記載されているデータ行から成る

列	項目	説明	例
1	#CHROM	変異がある染色体名	I
2	POS	変異のポジション	111
3	ID	rsID、COSMIC IDなど	.
4	REF	該当ポジションにおけるリファレンスゲノムのアリル	C
5	ALT	変異のアリル	T
:	:	:	:

ファイル形式 | vcf

- 変異の情報

列	項目	説明	例
:	:	:	:
6	QUAL	変異のクオリティ	105.93
7	FILTER	変異検出ソフトが変異につける変異のクオリティ	.
8	INFO	検出ソフトやアノテーションソフトが変異につける変異の情報やアノテーション。記述は自由	AC=1;AF=0.50;AN=2;BaseQRankSum=0.729;DP=9;Dels=0.00;FS=0.000;HRun=1;HaplotypeScore=0.0000;MQ=59.16;MQ0=0;MQRankSum=-1.159;QD=11.77;ReadPosRankSum=-0.361;SB=-0.01
9	FORMAT	以降の列に記載されるサンプルごとの変異情報の書式説明	GT:AD:DP:GQ:PL
:	サンプル列	変異の情報。書式はFORMATに従う	0/1:5,4:9:99:136,0,173

ファイル形式 | bed

- ゲノム上の領域の情報
  - エクソームシーケンスなどのターゲットシーケンスで解析範囲を指定するために用いられるほか、ChIP-seqで検出されたピークを示すのに用いる
  - 例としてbamファイルをbedファイルに変換した場合

```
$ bamToBed -i 1K_ERR038793.bam > 1K_ERR038793.bed
$ less 1K_ERR038793.bed
```

```
XII 1065142 1065238 ERR038793.1/1   4   -
I   149 248 ERR038793.1/2   60  -
XIII    923961  924028  ERR038793.2/1   40  +
                        :
```

ファイル形式 | bed

- ゲノム上の領域の情報
  - エクソームシーケンスなどのターゲットシーケンスで解析範囲を指定するために用いられるほか、ChIP-seqで検出されたピークを示すのに用いる

列		項目	説明	例
1	必須	chrom	染色体	XII
2		chromStart	開始ポジション。最初の塩基は0	1065142
3		chromEnd	終了ポジション	1065238
4	オプション	name	遺伝子名や任意の文字列	ERR038793.1/1
5		score	0-1000までの数値	4
6		strand	順鎖なら＋、逆鎖なら－	-
:		:	:	:

ファイル形式 | fasta

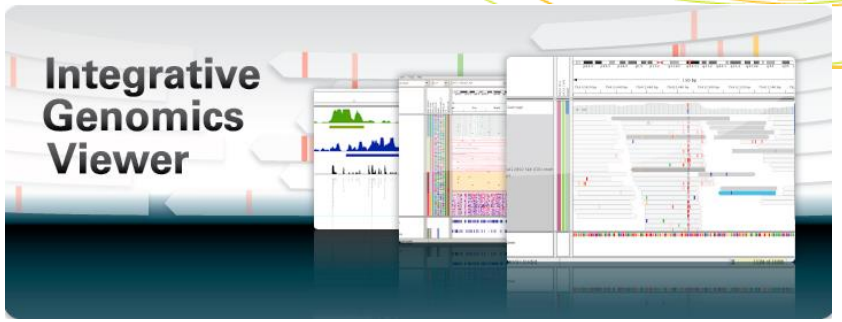
- NGS解析以外でもよく使われる、塩基配列やアミノ酸配列の情報。
  ここではリファレンスゲノム配列のfastaについて説明する
  - 拡張子が統一されておらず、.fa、.fasta、.fna、.fasなどが使われていることがあるが、中身は同じ

- 1行めは「>」で始まるヘッダ、2行めから配列

```
$ less /home/ユーザ名/Desktop/amelieff/Scerevisiae/WholeGenomeFasta/genome.fa
```

```
>I
CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACCCACACACACACATCCTAACA
CTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTC
CACTCGTTACCCTGTCCCATTCAACCATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTC
```

データの可視化

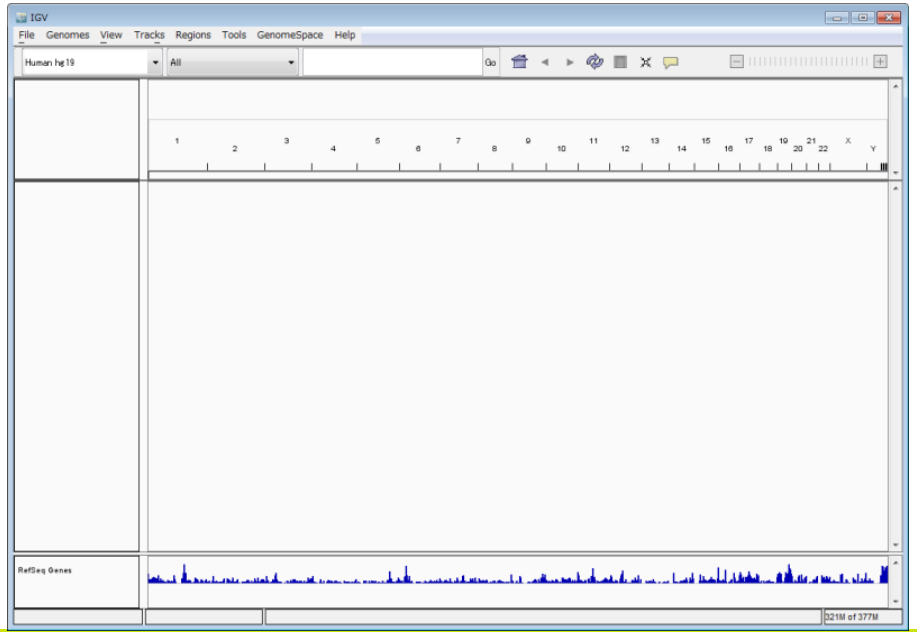


- Integrative Genomics Viewer（IGV）
  - 米 Broad Instituteが開発したゲノムブラウザ
  - GUIで直感的な操作が行える
  - bam、bed、vcfなどのファイル形式に対応（可視化できる形式一覧は http://www.broadinstitute.org/software/igv/FileFormats）
  - Windows、MacOS、LinuxのいずれのOSでも動作する
  - クローズドな環境で使用でき、セキュリティ上安全

データの可視化

- IGVの起動

```
$ igv.sh
```



データの可視化｜インデックスの作成

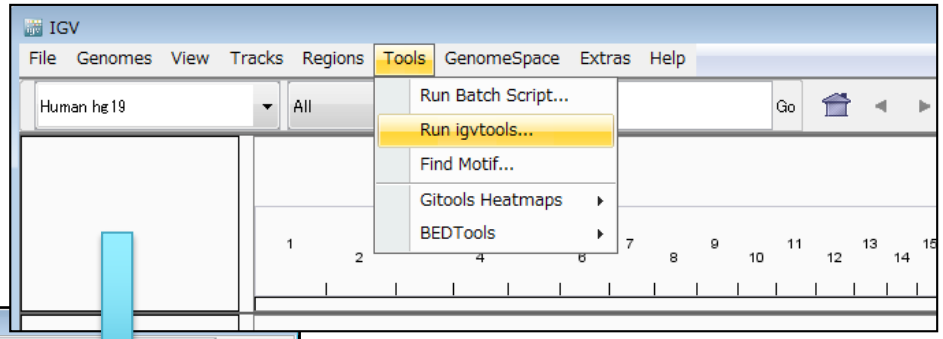
- サイズの大きなデータを高速に扱うため、サイズの大きなファイルにはインデックス（目次）ファイルが必要なことが多い
  - bamファイル

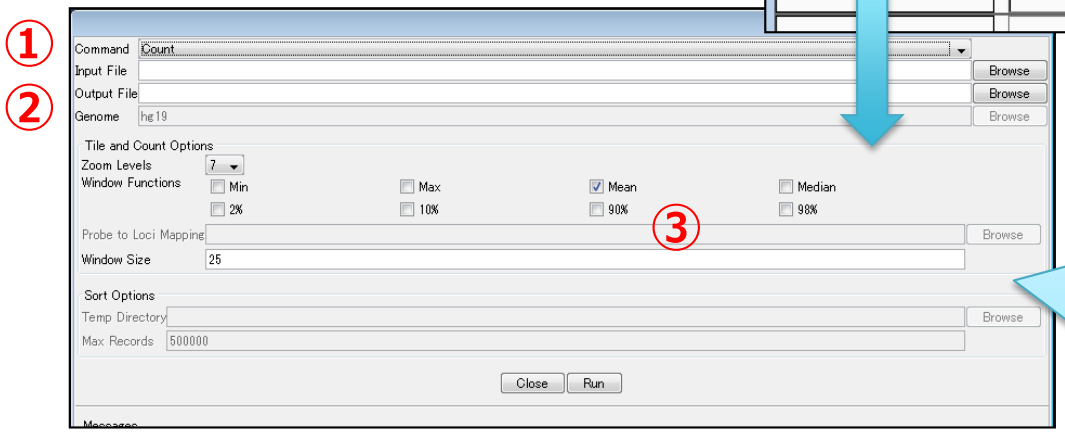
インデックス作成前にソートが必要

```
$ ls
1K_ERR038793.bam
$ samtools sort 1K_ERR038793.bam 1K_ERR038793_sort
$ ls
1K_ERR038793.bam          1K_ERR038793_sort.bam
$ samtools index 1K_ERR038793_sort.bam
$ ls
1K_ERR038793.bam          1K_ERR038793_sort.bam
1K_ERR038793_sort.bam.bai
```

データの可視化｜インデックスの作成

- サイズの大きなデータを高速に扱うため、サイズの大きなファイルにはインデックス（目次）ファイルが必要なことが多い
  - vcf・bedファイル
    - igvtoolsを起動する

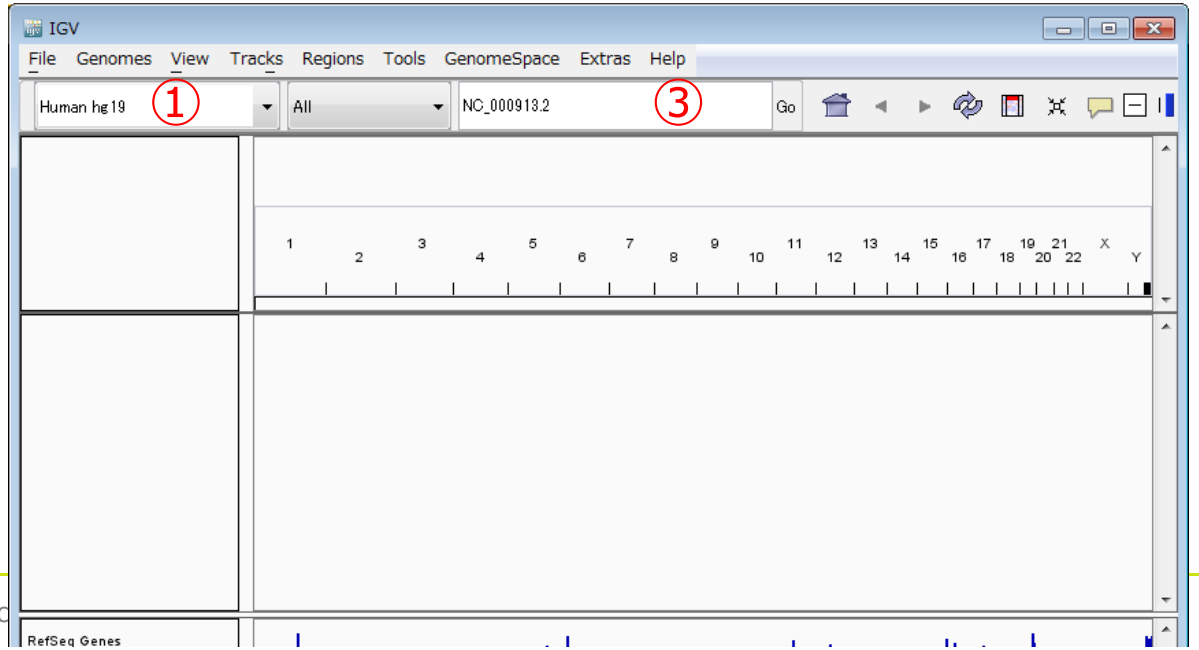




① Commandを「index」
② Input Fileを選択
③ Run
（実行完了のメッセージなどは出ません）

データの可視化

1. リファレンスゲノムを選択する
2. 可視化したいファイルを選択する
   – 「File」 > 「Load from File」からファイルを選択する
3. 詳細に見たい領域を選択する



データのクオリティチェック

- FastQC：fastqまたはbamのクオリティを確認するソフトウェア
  - fastqファイル1つに対して実行する

```
$ ls
```

```
1K_ERR038793_1.fastq
```

```
$ fastqc -f fastq 1K_ERR038793_1.fastq
```

```
Started analysis of 1K_ERR038793_1.fastq
Approx 5% complete for 1K_ERR038793_1.fastq
Approx 10% complete for 1K_ERR038793_1.fastq
                          :
                          :
Approx 100% complete for 1K_ERR038793_1.fastq
Analysis complete for 1K_ERR038793_1.fastq
```

データのクオリティチェック

- FastQC
  - クオリティチェックのレポートがあるディレクトリと、ディレクトリの圧縮ファイルが生成される

```
$ ls
```

```
1K_ERR038793_1.fastq        1K_ERR038793_1_fastqc
1K_ERR038793_1_fastqc.zip
```

  - 解析レポート

```
$ cd 1K_ERR038793_1_fastqc
$ ls
```

```
Icons    fastqc_data.txt      summary.txt
Images   fastqc_report.html
```

データのクオリティチェック

- FastQC

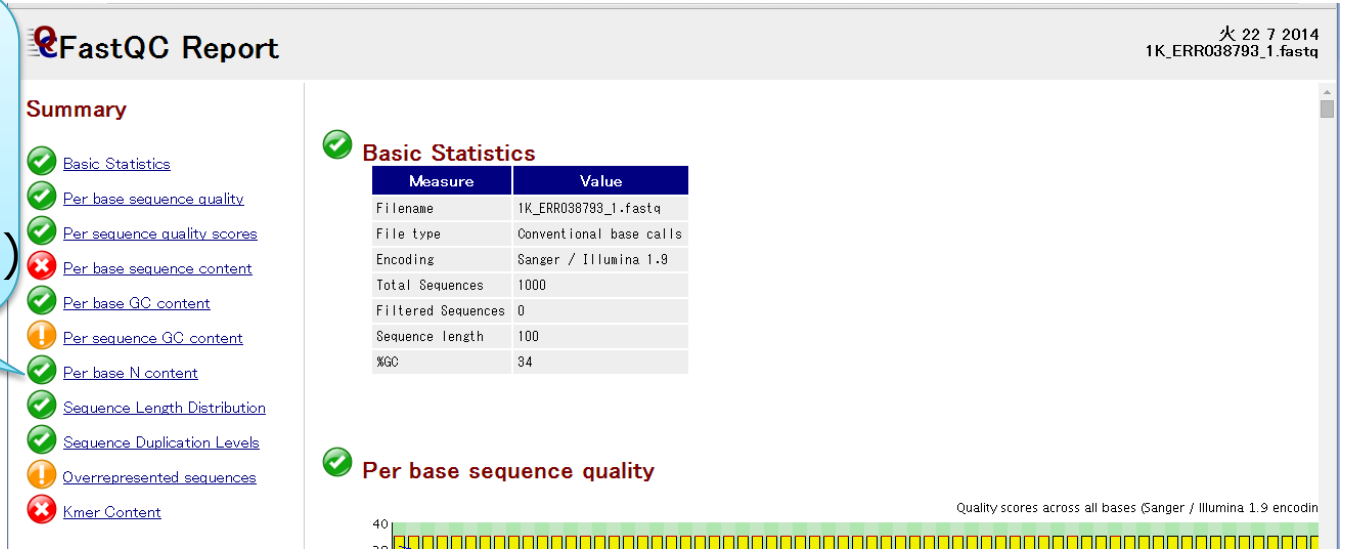
```
$ firefox fastqc_report.html
```

fastqc_report.htmlを、ウェブブラウザで開く

問題なし

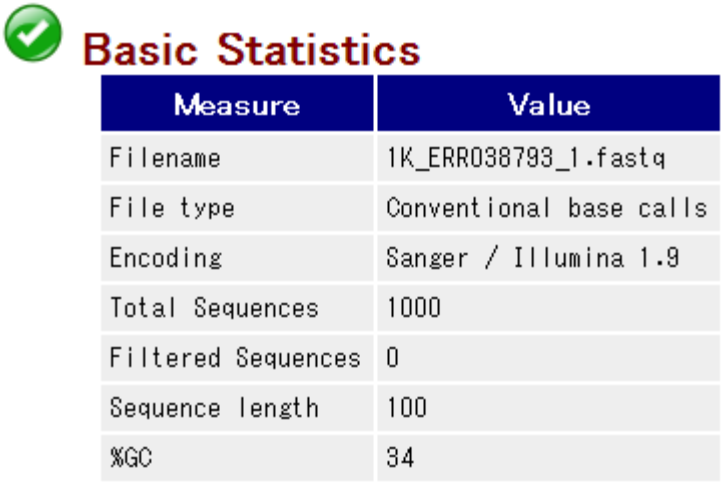
注意 (warning)

問題あり (failure)



データのクオリティチェック

- FastQC



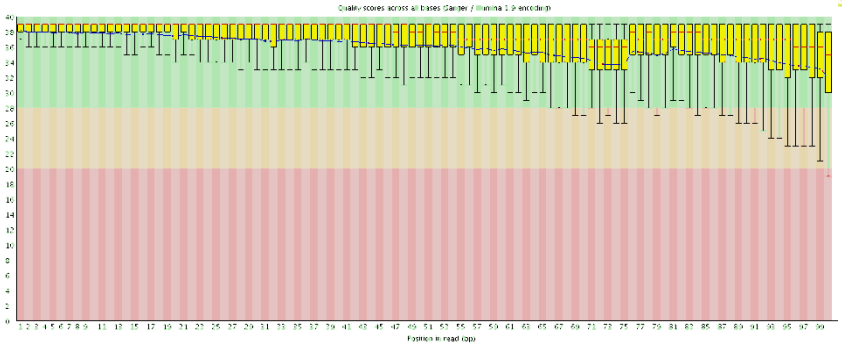
Basic Statistics

Measure	Value
Filename	1K_ERR038793_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Filtered Sequences	0
Sequence length	100
%GC	34

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。

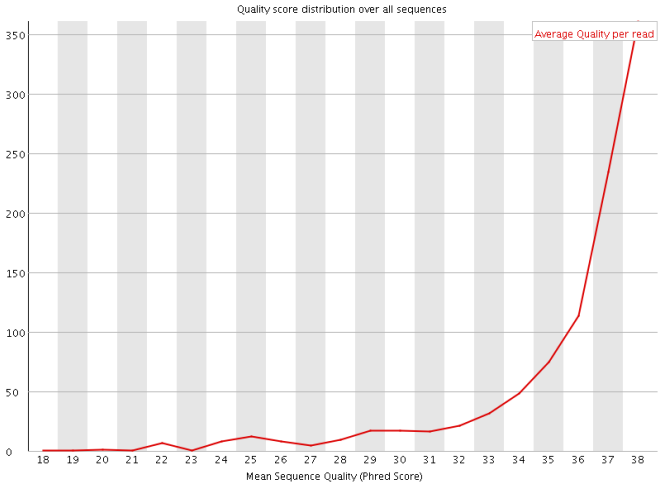
データのクオリティチェック

Per base sequence quality



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。

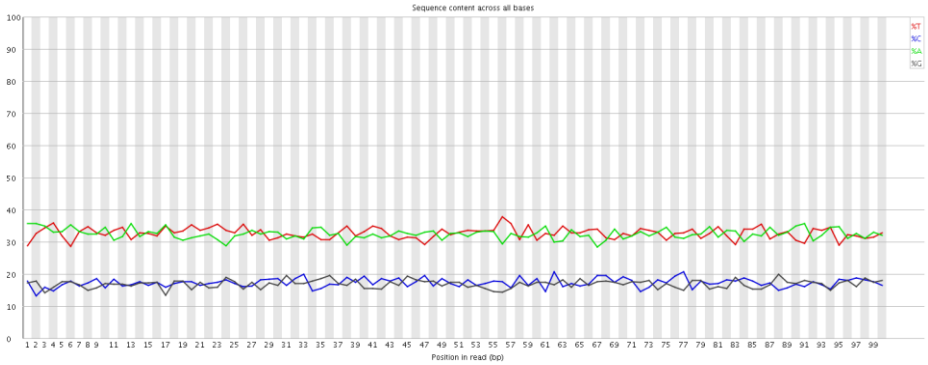
Per sequence quality scores



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

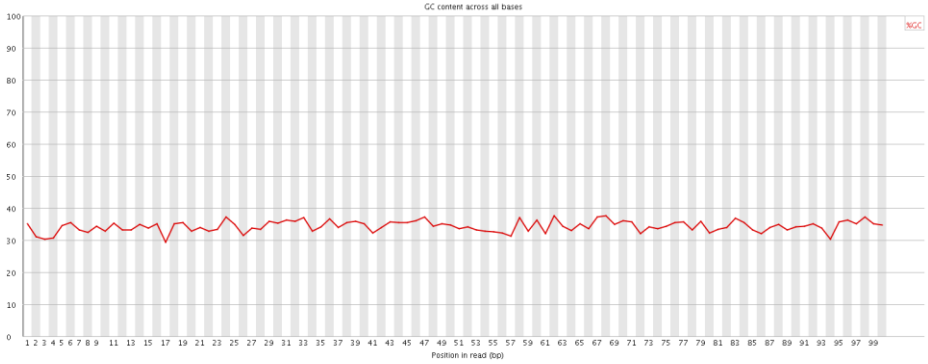
データのクオリティチェック

Per base sequence content



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。

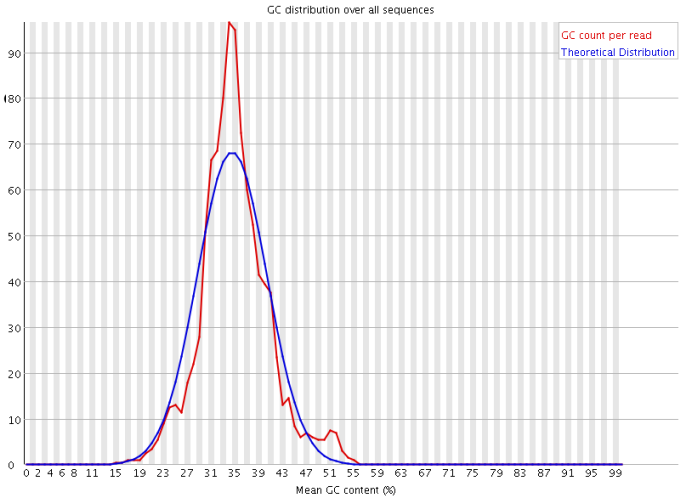
Per base GC content



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。

データのクオリティチェック

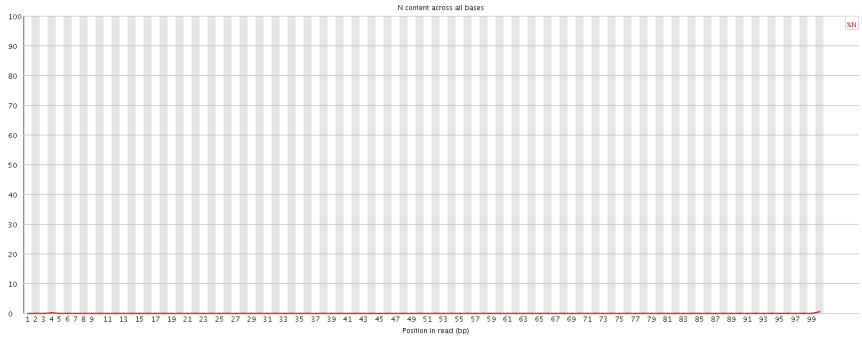
Per sequence GC content



Per Sequence GC Content

各リードにおけるGC含量の平均の分布(赤線)と、理論分布(青線)。
理論分布との偏差の合計が、総リードの15%以上でwarning, 30%以上でfailureとなる。

Per base N content

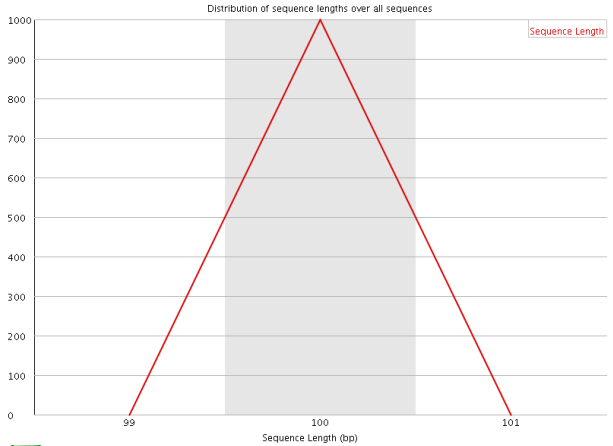


Per Base N Content

“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。

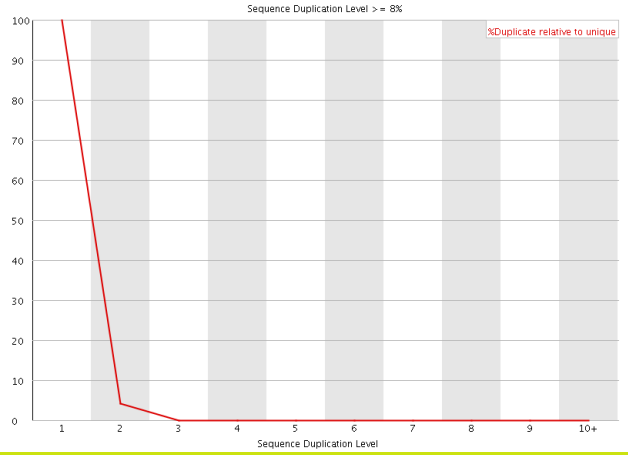
データのクオリティチェック

Sequence Length Distribution



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。

Sequence Duplication Levels



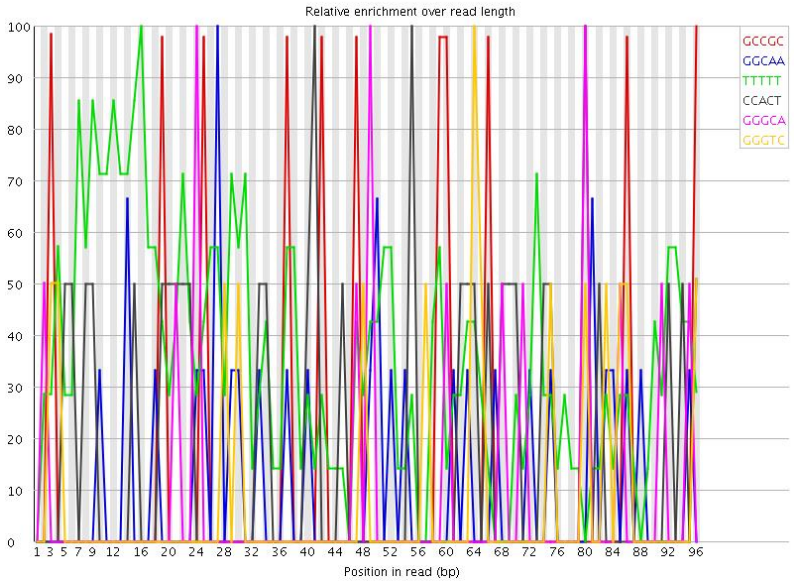
Sequence Duplication Levels
リードの重複レベルを見ている。
1～10はそれぞれ重複のレベルで、全体の20%以上がユニークでないものだとwarning, 50%以上がユニークでないとfailureとなる。

データのクオリティチェック

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGTATTAATATTTCACTGTCTTGATATCGTTATCCCCATCGTAACGTGAA	2	0.2	No Hit
GCTTTAAACGGCTTCCGCGGAAGAAATATTTCCATCTCTTGAATTCGTAC	2	0.2	No Hit
CTTTTACACCATATACTAACCACTCAATTTATATACACTTATGCCAATAT	2	0.2	No Hit
CCTGTCCCATTCAACCATACCACTCCGAACCACCATCCATCCCTCTACTT	2	0.2	No Hit
AACCCGCTACGTTGACTACAAGCTCAAACCCGAATACCACATCTGCACGT	2	0.2	No Hit
GTCAATTTCTACTTGCCTCATTAGGGAAAAATTTAATAGCAGTTGTTATA	2	0.2	No Hit
CCATTATGACAAAGTTAAGGAGTTACGCGTGCTACATCACCGTAAAAATT	2	0.2	No Hit

Kmer Content



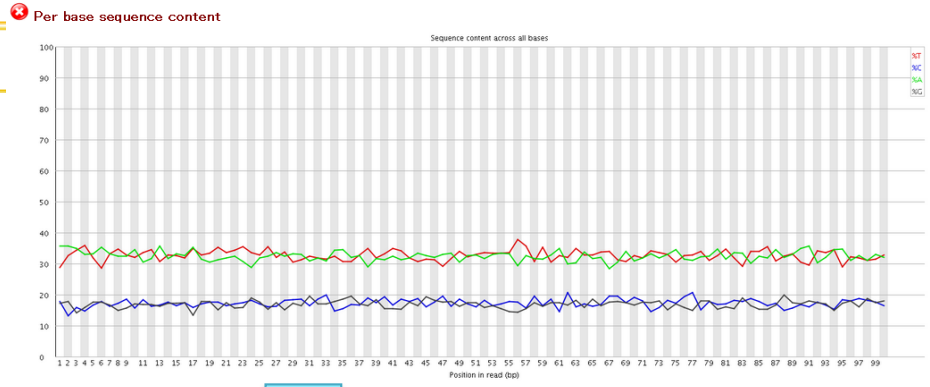
Overrepresented Sequences

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。

K-mer Content

5 bpの任意の配列(5mer)を考えた時、ライブラリに含まれるATGCの割合を元に「実際に観測された値/理論的に観測される期待値」を計算している。
それぞれの任意の配列について、実測が期待値を大きく上回っている時、それはライブラリに配列的な偏りがあると解釈される。
「実測値/期待値」は、リード長全体における計算と、リードのある位置での計算を行い、全体における値が3倍、リードのある位置における値が5倍になるとwarning、リードのある位置における値が10倍になるとfailureとなる。

データのクオリティチェック



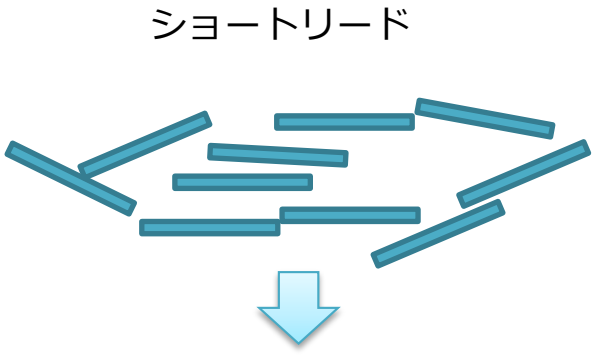
- テキストデータによるレポートも出力される

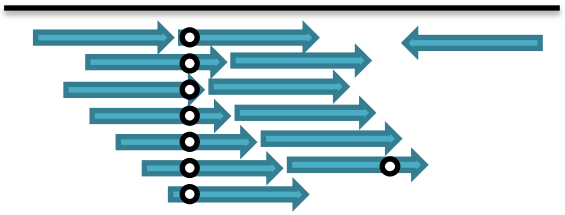
```
$ less fastqc_data.txt
```

```
>>Per base sequence content fail
#Base  G                A                T                C
      1          17.4          35.8          28.9          17.9
      2          17.9          35.9          32.8          13.4
      3          14.4          35.1          34.5            16
      4      16.03206      33.16633      35.97194      14.82966
      5          17.8          33.3            32          16.9
      6          17.7          35.5          28.8            18
      7          16.9          33.3          33.3          16.5
      8          15.1          32.6          34.9          17.4
      9          15.8          32.5            33          18.7
```

マッピング

- シーケンサから得られたリード（DNA配列）を、リファレンスゲノムや転写産物上の類似した配列に対して並べること
  - BLASTのような従来のマッピングソフトは正確だが時間がかかり、NGS解析に向かないため、NGS解析用の高速なマッピングソフトが使われる





マッピング

- 各解析で使われるマッピングソフトの特徴と主なマッピングソフト
  - Reseq：データの大きなゲノムファイルに対して数カ所のミスマッチを許容して高速にマッピングする。BWAやBowtieなど
  - RNA-seq：スプライシングにより生じるギャップを考慮してマッピングする。TopHatなど
  - Methyl-seq：メチル化を考慮してマッピングする。BSMAPなど

アセンブリング

- ゲノムde novoアセンブリングで主に使われるソフト
  - Velvet
  - SOAPdenovo
  - AbySS

- トランスクリプトームde novoアセンブリングで主に使われるソフト
  - Oases
  - SOAPdenovo-Trans
  - Trans-ABySS
  - Trinity