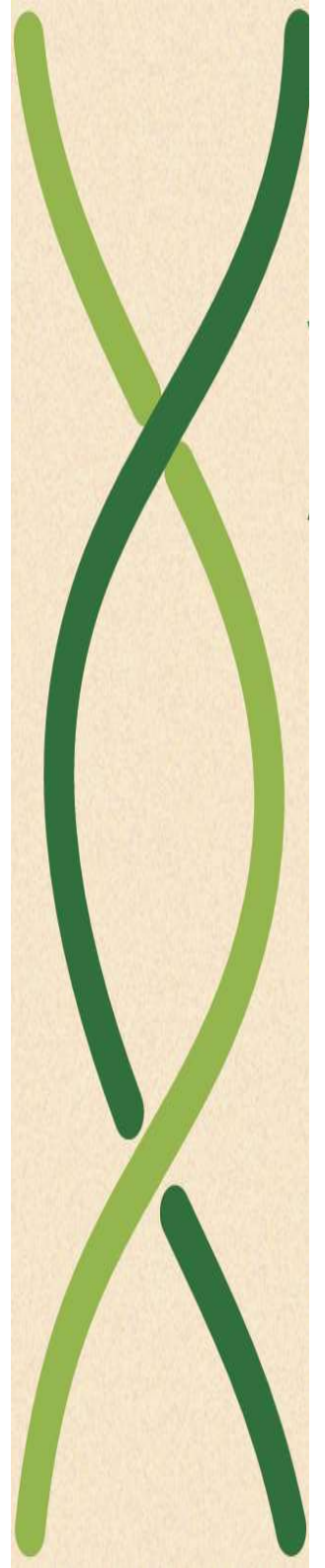


NIBB ゲノムインフォマティクス・トレーニングコース2025
「NGS解析入門」

2025.2.4-2025.2.5

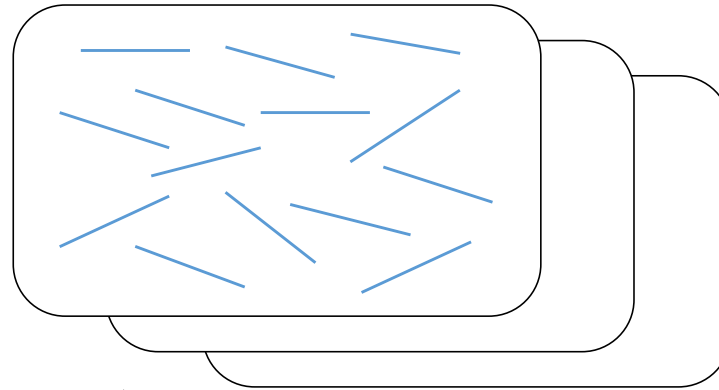
クオリティコントロールと NGS基本ツール

基礎生物学研究所
超階層生物学センター
トランスオミクス解析室
山口勝司



NGSデータ処理の概要

シーケンスリード (DNA/RNA由来)



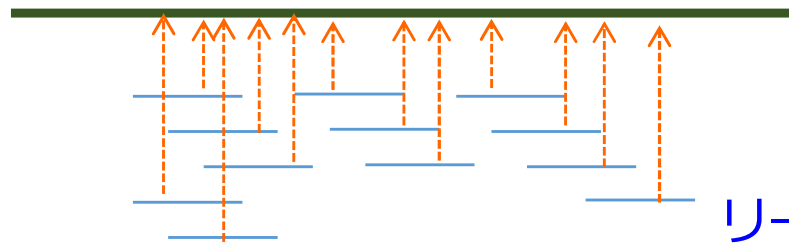
リファレンス配列あり

リファレンス配列なし

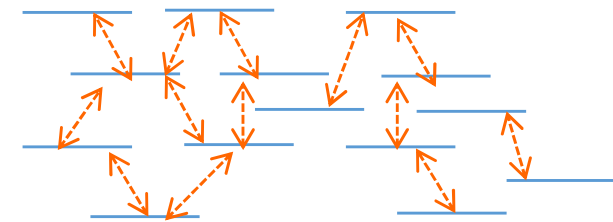
リファレンス配列へのマッピング

デノボ・アセンブル

リファレンス配列



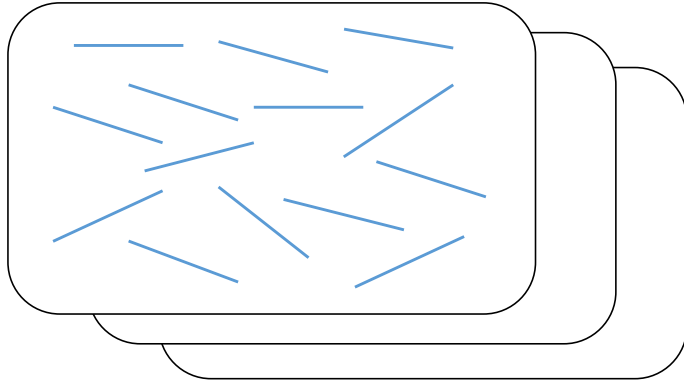
リード配列



アセンブル配列

シーケンスリード (DNA/RNA由来)

FASTQファイル (配列+クオリティ)



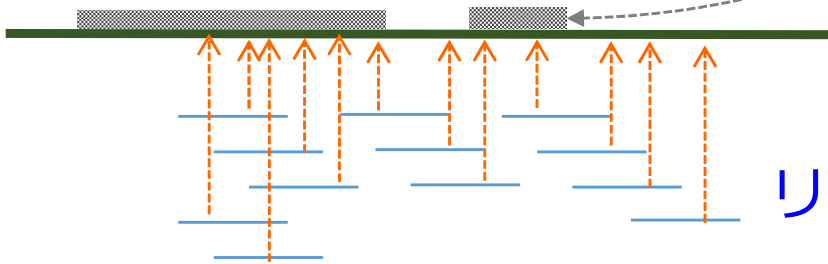
リファレンス配列あり

```
@SRRL151276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCCGGCGGAATAACAAGCTGGTGAAG
+SRRL151276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DFF7DC?FFBF@DFII<DFAAAAFAEBDBDCA?>A?B=>B::
@SRRL151276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGAGTACCAGCATCCTGCGTACAATCAGCAATCCAGTCTCCCC
+SRRL151276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFDFDFHDFHIIIEGHIJJJJGFHGGHGGHGGIIJDGIHHGGGHH
@SRRL151276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTCAAGCTTCGGAACCTGCATTTTCAG
+SRRL151276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFDFDFHFIHIIJGIIJJJJHIIHIIJGIIHFIHIIA@FIHGGIIGI
```

遺伝子アノテーション GFF(GTF)ファイル

```
chr RefSeq start_codon 190 192 1.000+ . gene_id "b0001"; transcript_id "b0001";
chr RefSeq CDS 190 252 1.000+ 0 gene_id "b0001"; transcript_id "b0001";
chr RefSeq stop_codon 253 255 1.000+ . gene_id "b0001"; transcript_id "b0001";
chr RefSeq exon 190 255 1.000+ . gene_id "b0001"; transcript>
```

リファレンス配列へのマッピング



リード配列

ゲノム (リファレンス) 配列

FASTAファイル

xchr

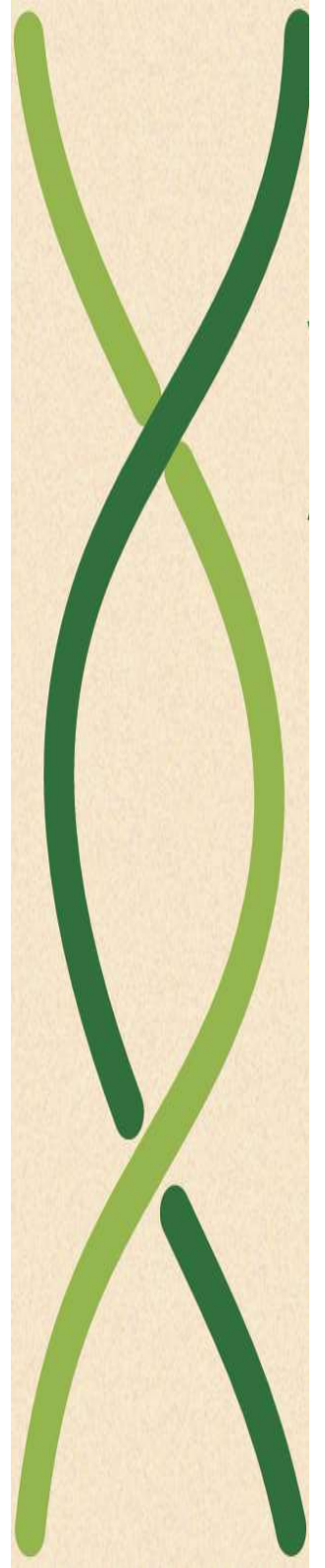
@ID	WI:1.0	SO:unsorted					
@SQ	SN:chr	LN:463675					
@PG	ID:bowtie2	PN:bowtie2	WI:2.2.4		CL:"/big/bin/bowtie2-align		
SRR1515276.40	0 chr	4423609	42	51M	*	0	0
SRR1515276.158	16 chr	501700	42	51M	*	0	0
SRR1515276.212	4 *	0	0	*	*	0	0
SRR1515276.319	0 chr	2922768	42	51M	*	0	0
SRR1515276.367	16 chr	2753873	42	51M	*	0	0
SRR1515276.411	0 chr	3440721	42	51M	*	0	0
SRR1515276.434	0 chr	4198737	42	51M	*	0	0

マッピング結果 SAM ファイル

SAM ファイル

クオリティーコントロール

- Fastqc
- Cutadapt
(Pre-processing tools)



NGSデータ解析におけるクオリティーコントロールの重要性



- ・ 作製したライブラリーに問題はなかったか
 - アダプター配列ばかり . . .
 - コンタミ配列が多い . . .
 - PCR増幅のバイアス産物が多い . . .
 - 短いライブラリーほど増幅されやすい
 - GC率に偏りがあるものは増幅されにくい

→問題点を検出・検証する手段

- ・ 得られるdataのクオリティーは同一ではない
 - シーケンサーの調子
 - 作製ライブラリーのサイズ分布
 - シーケンサー間の性能差

→可能な範囲でクオリティーを揃える手段

シーケンスのクオリティ-checkツール FASTQC



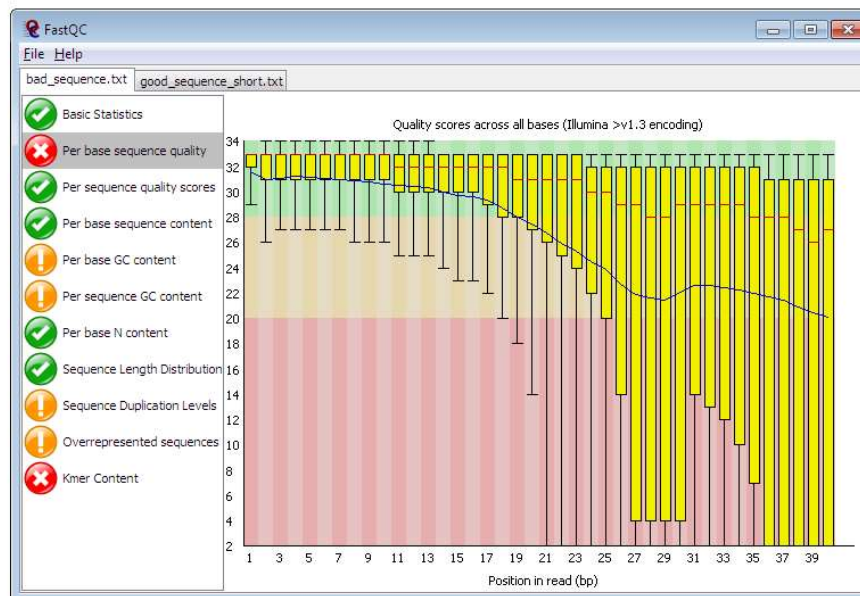
Babraham Bioinformatics

[About](#) | [People](#) | [Services](#) | [Projects](#) | [Training](#) | [Publications](#)

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)



Documentation

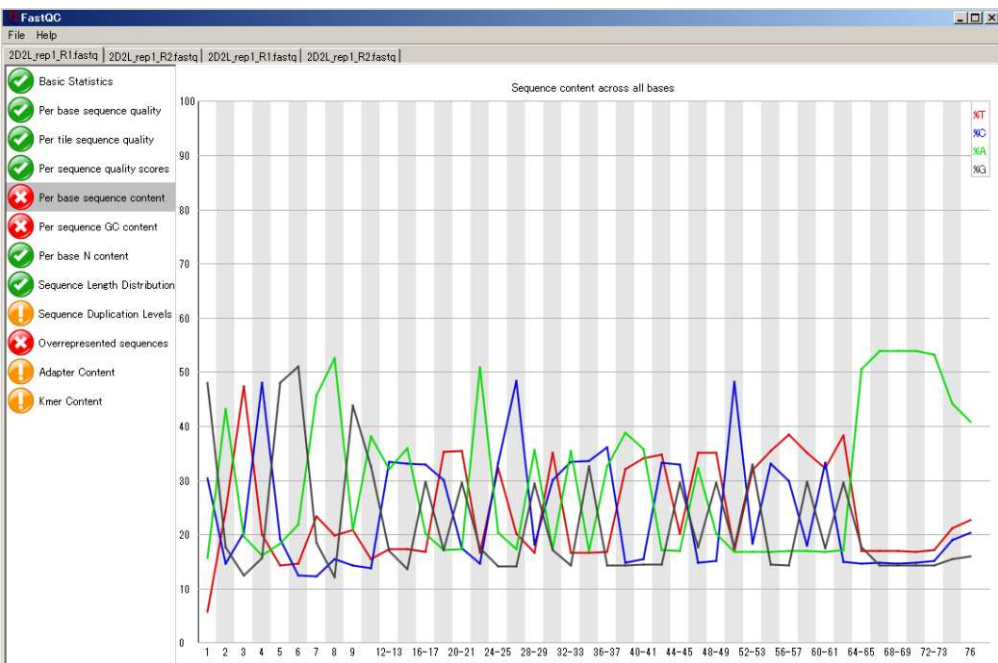
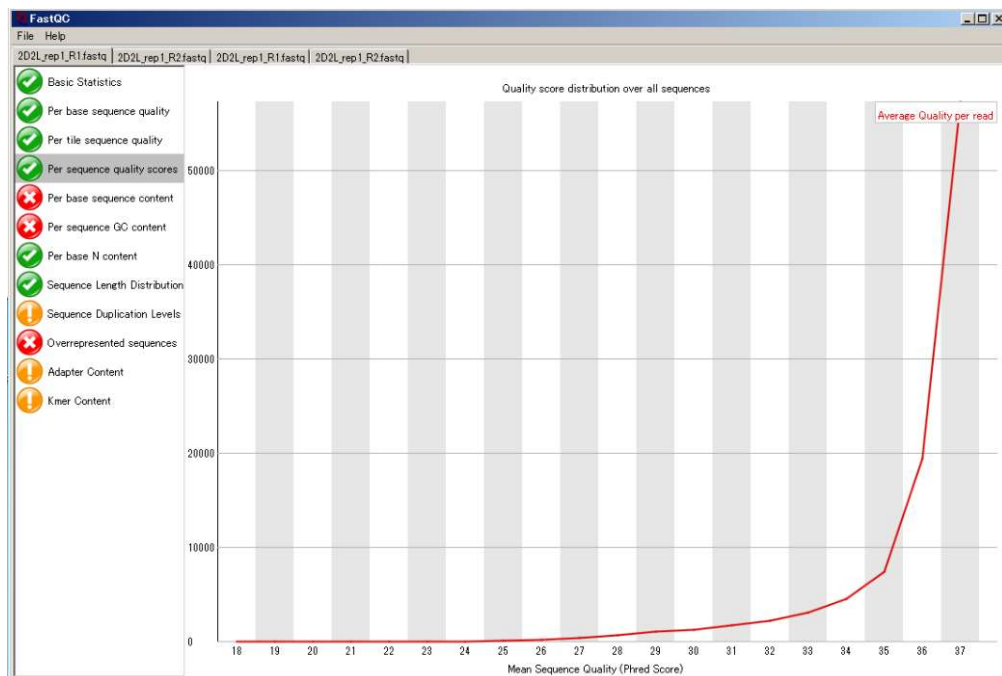
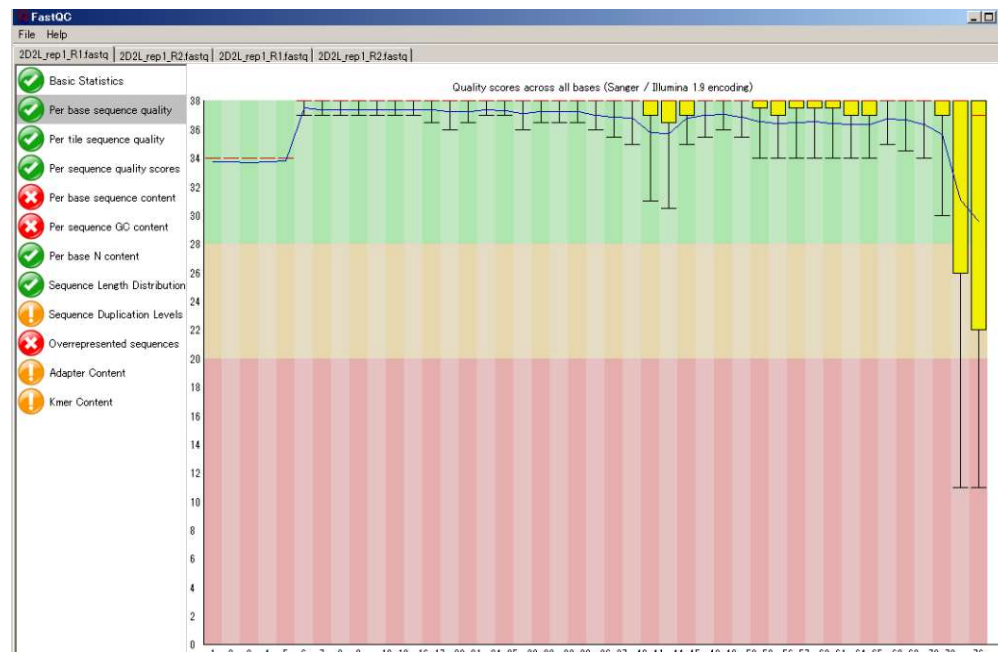
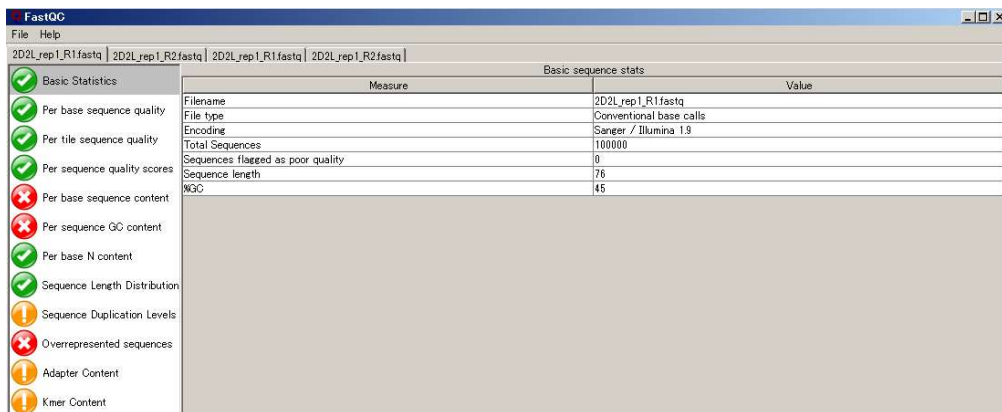
A [copy of the FastQC documentation](#) is available for you to try before you buy (well download..).

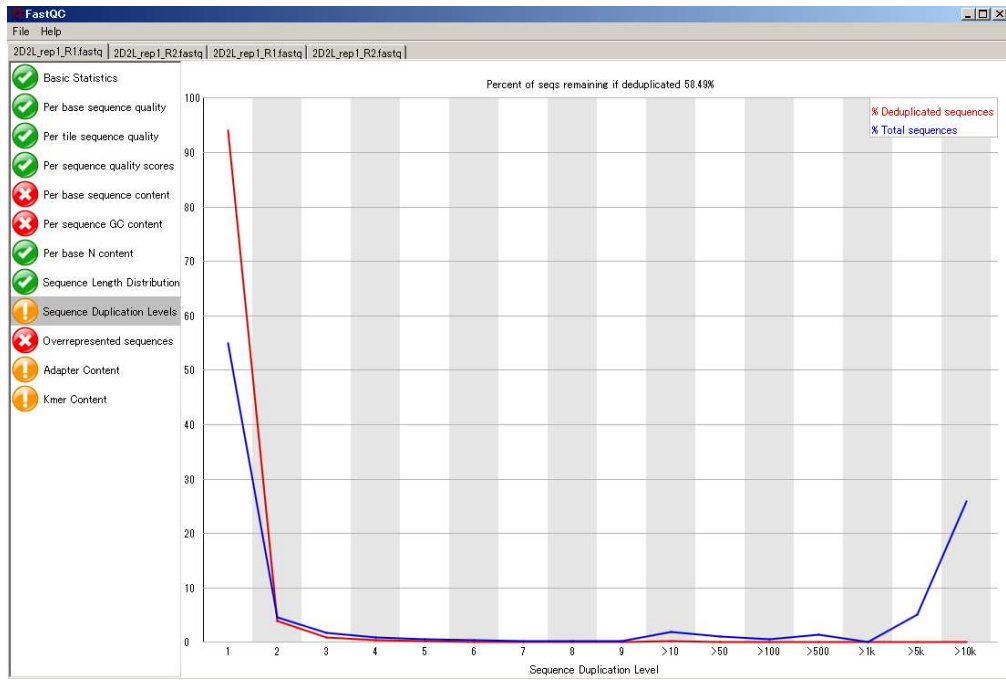
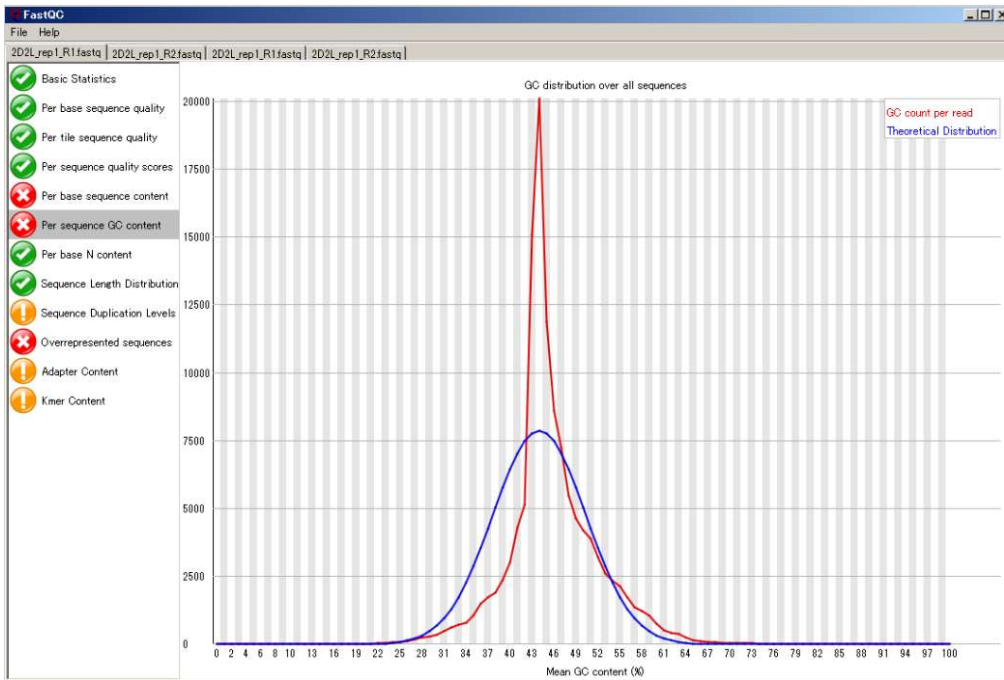
Example Reports

- [Good Illumina Data](#)
- [Bad Illumina Data](#)
- [Adapter dimer contaminated run](#)
- [Small RNA with read-through adapter](#)
- [Reduced Representation BS-Seq](#)
- [PacBio](#)
- [454](#)

Version 0.12.1

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>





FastQC

File Help

2D2L_rep1_R1.fastq | 2D2L_rep1_R2.fastq | 2D2L_rep1_R1.fastq | 2D2L_rep1_R2.fastq

Basic Statistics

Per base sequence quality

Per tile sequence quality

Per sequence quality scores

Per base sequence content

Per sequence GC content

Per base N content

Sequence Length Distribution

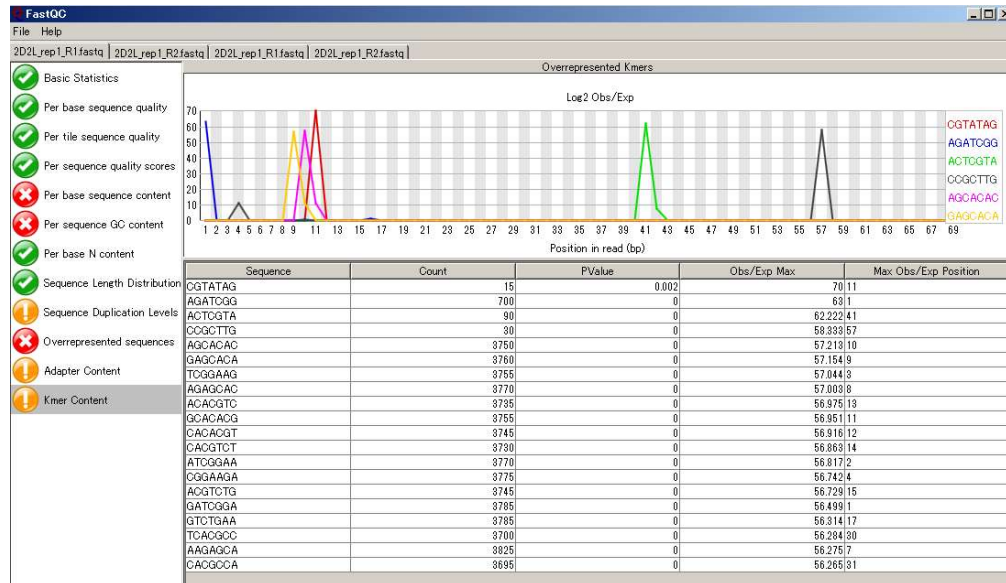
Sequence Duplication Levels

Overrepresented sequences

Adapter Content

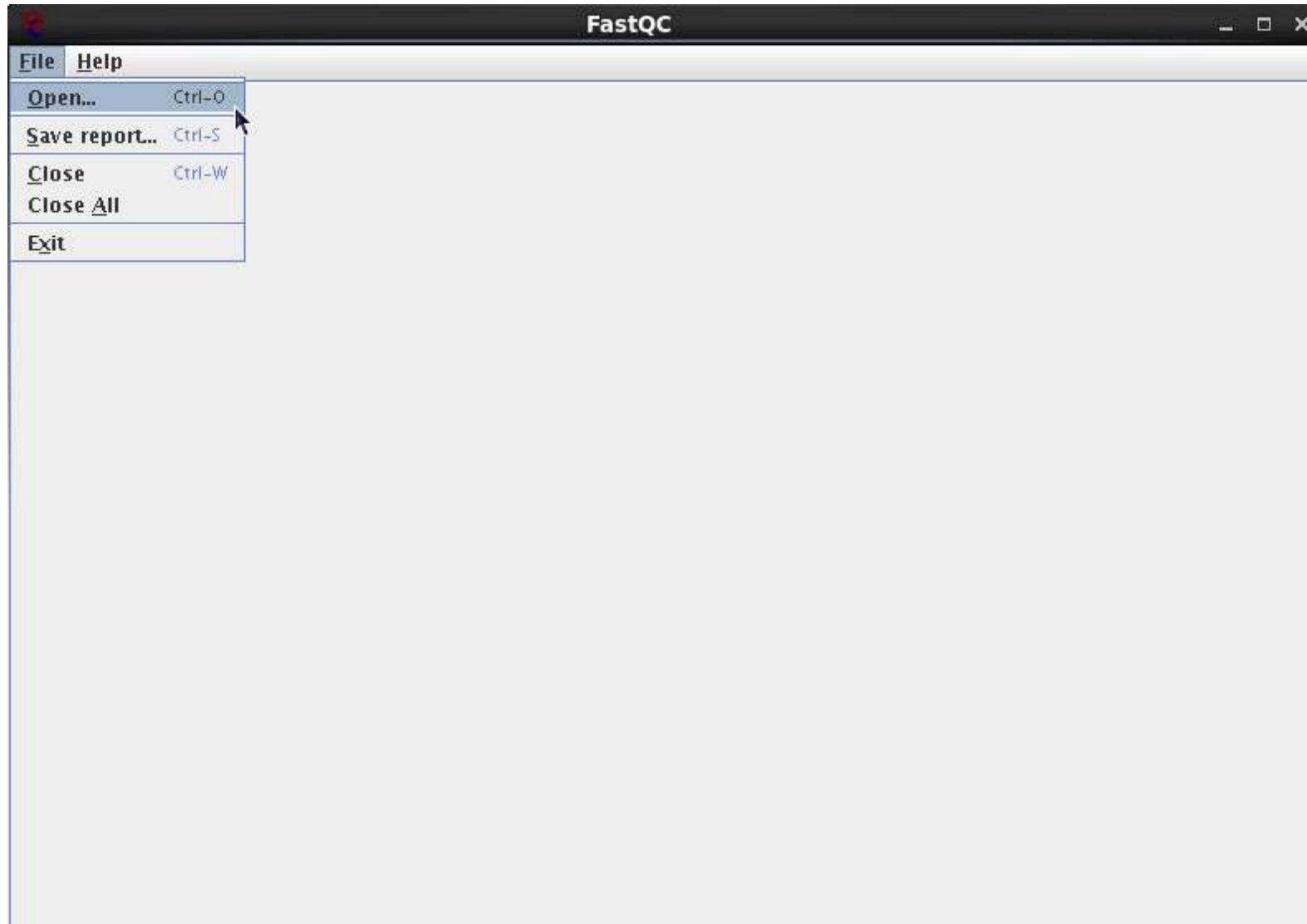
Kmer Content

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCG...	25945	0.100%	TruSeq Adapter, Index: 6 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCG...	5147	0.020%	TruSeq Adapter, Index: 6 (100% over 49bp)
GATCGGAAGAGCACACGTCTGAACTCG...	784	0.003%	TruSeq Adapter, Index: 6 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCG...	617	0.002%	TruSeq Adapter, Index: 6 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCG...	242	0.001%	TruSeq Adapter, Index: 6 (98% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCG...	172	0.001%	TruSeq Adapter, Index: 6 (97% over 49bp)
GATCGGAAGAGCACACGTCTGAACTCG...	111	0.000%	TruSeq Adapter, Index: 6 (98% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCG...	102	0.000%	TruSeq Adapter, Index: 6 (97% over 49bp)




FASTQC使用法 GUI (グラフィカルユーザーインターフェース)

GUI java jdkを予めインストールしておく必要がある。



FASTQC使用法 CUI (コマンドユーザーインターフェース)



```
$ 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
```

gzファイルなら
--extract
の記載

実習 1 FASTQC

実習用ディレクトリ ~/gitc /data/5_ngs に移動して 中を見る

- read結果 2D2L_rep1_R1.fastq

のファイルがあることを確認

これをFASTQCに読み込ませて、クオリティーを確認しよう

コマンドラインから

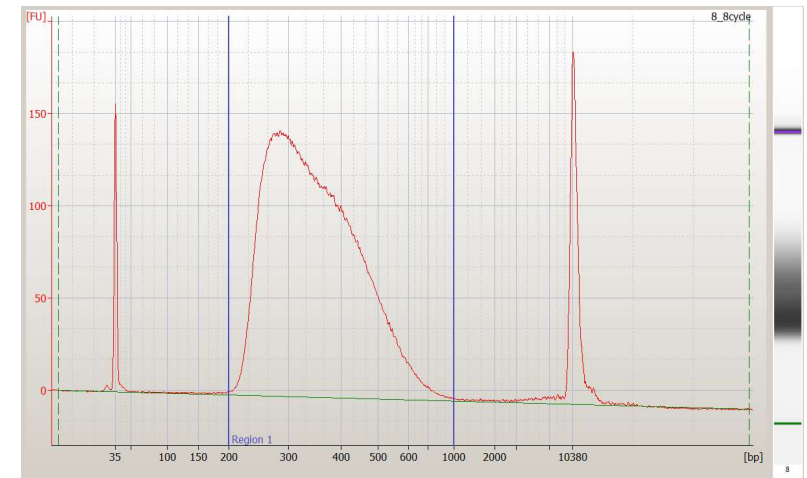
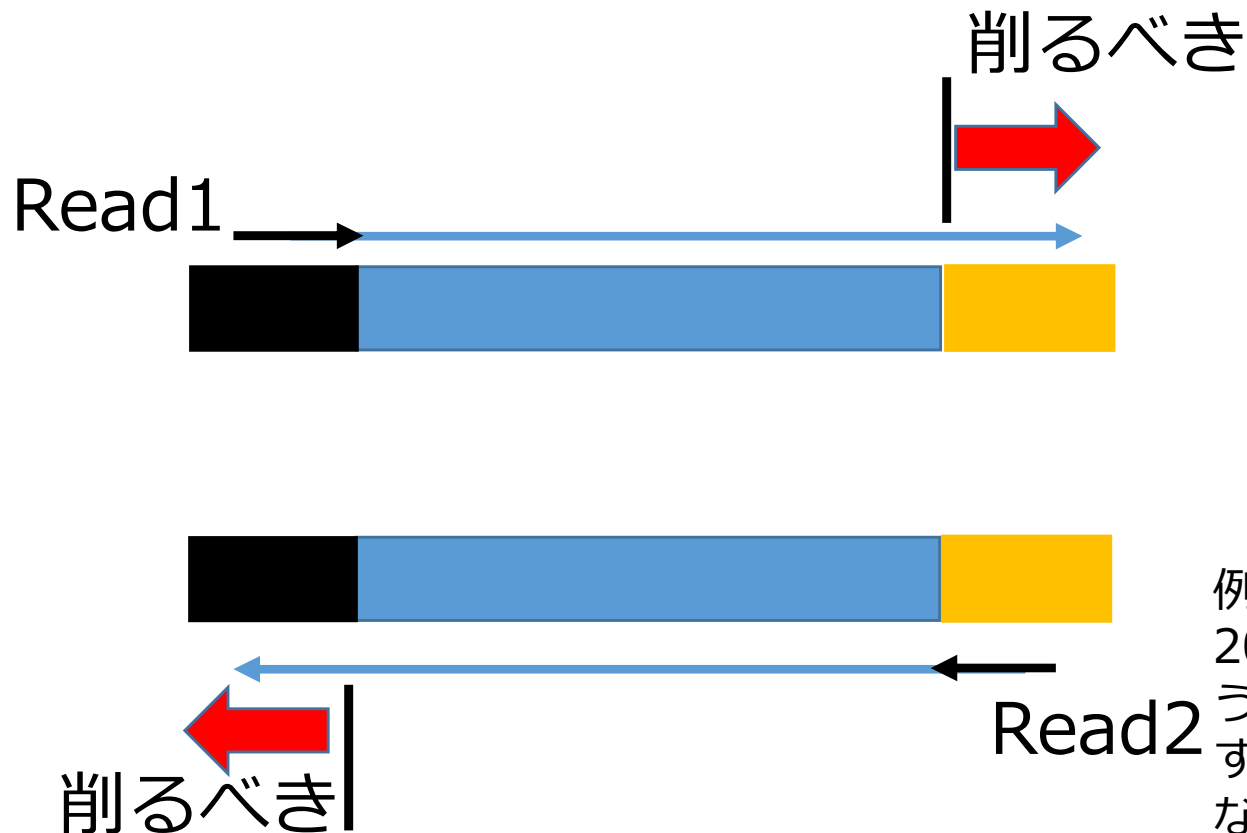
> **fastqc**

と入力してfastqcが起動したことを確認

ここからはGUIでの作業

NGSデータのPre-processingの必要性

- ・余計な配列（アダプター配列）がリファレンス配列へのmappingに影響
- ・余計な配列（アダプター配列）がゲノム配列と誤認されうる



例) イルミナ Truseq ライブラリーの場合、200base くらいの長さから存在する。うち両端にアダプター 63base ずつすなわち 75base 程度しか insert 配列がないライブラリーが存在することになる。

Pre-processing tools

- Cutadapt
- Trimmomatic
- fastp
- etc

- adapter配列を除去
- 一定クオリティー以下の部位を除去
- 任意の配列部位を除去

生データを処理することで、アダプター配列を除去し、一定のクオリティーを確保したデータとなる

The screenshot shows the Cutadapt user guide documentation page. The left sidebar contains a navigation menu with the following items: 'Basic usage' (expanded), 'Read processing stages', 'Adapter types', 'Adapter-trimming parameters', 'Specifying adapter sequences', 'Modifying reads', 'Filtering reads', 'Trimming paired-end reads', 'Multiple adapters', 'Illumina TruSeq', 'Dealing with N bases', and 'Cutadapt's output'. The main content area is titled 'User guide' and 'Basic usage'. It explains how to trim a 3' adapter using the command-line tool Cutadapt. The basic command shown is: `cutadapt -a AACCGGTT -o output.fastq input.fastq`. It also shows the command for compressed files: `cutadapt -a AACCGGTT -o output.fastq.gz input.fastq.gz`. The text explains that the adapter sequence is given with the `-a` option and that reads are read from the input file and written to the output file.

Paired end readに対応
(ver. 1.8以降)
片方のreadが非常に
短くしか残らない場合、
pair read両方とも除去
する。

<https://cutadapt.readthedocs.io/en/stable/guide.html>

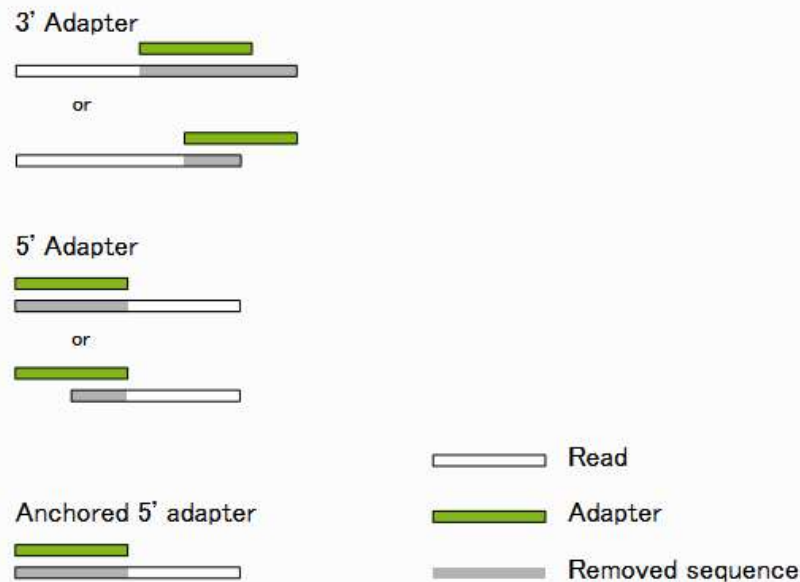
Cutadapt

Removing adapters

Cutadapt supports trimming of multiple types of adapters:

Adapter type	Command-line option
3' adapter	<code>-a ADAPTER</code>
5' adapter	<code>-g ADAPTER</code>
Anchored 3' adapter	<code>-a ADAPTER\$</code>
Anchored 5' adapter	<code>-g ^ADAPTER</code>
5' or 3' (both possible)	<code>-b ADAPTER</code>

Here is an illustration of the allowed adapter locations relative to the read and depending on the adapter type:



Cutしたいアダプター配列の位置関係など詳細に指定可能

fastqファイルはgzip圧縮してあってもよい
fastaファイルも可

最新versionではもっと、細かい条件の指定が可能

- ・ 5'側と3'側でquality cutの条件を別々に指定できる
- ・ polyA, polyTも判断してcutできる


```
$ cutadapt -h  
cutadapt version 5.0
```

用いられるバージョンが確認できる

最新はv5.0

Copyright (C) 2010-2022 Marcel Martin <marcel.martin@scilifelab.se>

cutadapt removes adapter sequences from high-throughput sequencing reads.

Usage:

```
cutadapt -a ADAPTER [options] [-o output.fastq] input.fastq
```

For paired-end reads:

```
cutadapt -a ADAPT1 -A ADAPT2 [options] -o out1.fastq -p out2.fastq in1.fastq in2.fastq
```

その他、有用なパラメータ

-j --cores	使うCPU core数 defaultは1 0を指定しておくとも自動検出
-q --quality-cutoff	クオリティをcutoffするQV値を指定
-m --minimum-length	指定する長さ以下にcutされたものはreadそのものを削除
-O --overlap	指定する配列とのオーバーラップを最小何baseとするか

crude_fastqフォルダーに生シーケンス配列
trim_fastqフォルダーにcutadaptにかけた配列
を用意してあります

Single readの場合

```
$ cutadapt ¥  
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA ¥  
-o hoge_read1.cut.fastq ¥  
hoge_read1.fastq
```

Paired end readの場合

```
$ cutadapt ¥  
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA ¥  
-A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT ¥  
-o hoge_read1.cut.fastq ¥  
-p hoge_read2.cut.fastq ¥  
hoge_read1.fastq ¥  
hoge_read2.fastq
```

シーケンスライブラリーの構造やアダプター配列もしっかり把握しておく。例えばイルミナなら以下各種kitごとにアダプター配列やcutすべきアダプター配列が記述されている。



TruSeq Single Indexes

A-tailing is performed before adapter ligation. For example, the additional A base is in parentheses in the i7 adapter, as follows.

Index 1 (i7) Adapters

(A)GATCGGAAGAGCACACGTCTGAACTCCAGTCAC([i7])ATCTCGTATGCCGCTCTTCTGCTTG

▼ Adapter Trimming

The following sequences are used for adapter trimming.

Read 1

AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

Read 2

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

▶ TruSeq Universal Adapter

▶ DNA and RNA Index Adapters

<https://support-docs.illumina.com/SHARE/AdapterSequences/Content/SHARE/FrontPages/AdapterSeq.htm>

<https://support-docs.illumina.com/SHARE/AdapterSequences/Content/SHARE/AdapterSeq/TruSeq/SingleIndexes.htm>

実習 2 cutadapt

実習用ディレクトリ `~/gitc/data/5_ngs` に移動して `ls` で中を見る

```
$ cd ~/gitc/data/5_ngs
$ ls
```

- read結果 2D2L_rep1_R1.fastq

adapterがどの程度残っているか概算してみる

```
$ cat 2D2L_rep1_R1.fastq|grep 'AGATCGGAAGAGCAC' |wc
$ cat 2D2L_rep1_R1.fastq|wc
```

実際にcutadaptにかけて見よう

```
$ cutadapt ¥
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA ¥
-o 2D2L_rep1_R1.fastq.cut.fastq ¥
2D2L_rep1_R1.fastq
```

発展) 圧縮されたQV値

fastqファイルに記載されるQV値は、近年のシーケンサーではストレージ容量削減の為に、圧縮効率の良い圧縮されたqv値が用いられている。

この場合、quality valueの階調数が減らされているので、それを考慮して-q値を指定すべき。

圧縮されたQV値で表記されたfastq

```
@E00441:177:HHNWCCXY:6:1101:13433:25464 1:N:0:GTGTATTA
AATGTGCGTTTGTGTTGGGATAGGACATTTGTCAGCTACGCGCCGGCTCTCTGTGAAGTAATTGGTTGAATGAATAAA
+
AAFFFFJJJFF-AFAF7A<<FJJF<AFJJJJJ7<FJJJJF-7F-7FJJFFJ-A-FFAFJF-FFJJA-FJAFJJ<AAFA
```

-	12	A	32	数が限定されている -q 31と28を比較しても 同じことになる
7	22	F	37	
<	27	J	41	

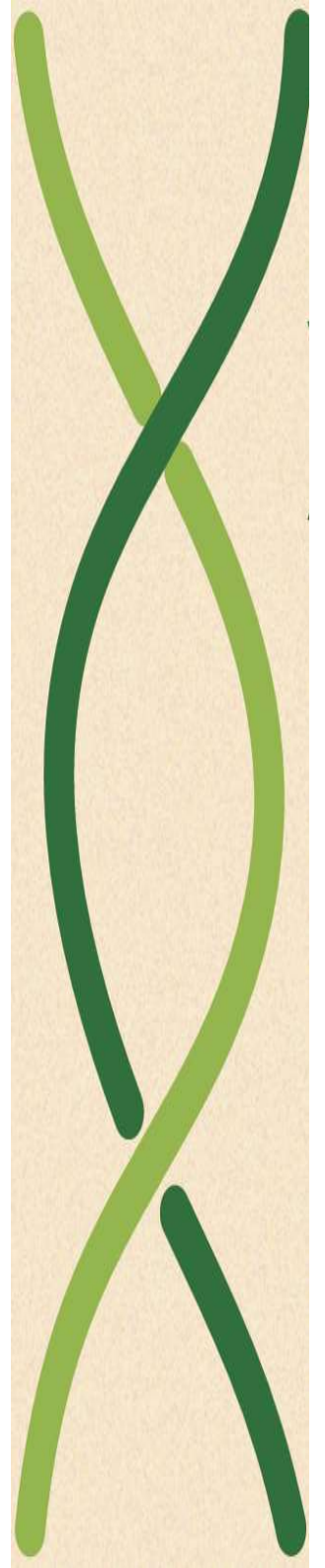


Illumina White paper

Reducing Whole-Genome Data Storage Footprintより

NGS基本ツール

- Seqkit
- Bowtie2
- SAMtools



SeqKitを使って見よう

fasta/fastqに関する様々な操作が可能なツール



RESEARCH ARTICLE

OPEN ACCESS

Citation: Shen W, Le S, Li Y, Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE 11(10): e0163962. doi:10.1371/journal.pone.0163962

Editor: Quan Zou, Tianjin University, CHINA

Received: May 23, 2016

Accepted: September 16, 2016

Published: October 5, 2016

SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation

Wei Shen¹, Shuai Le¹, Yan Li^{2*}, Fuquan Hu^{1*}

1 Department of Microbiology, College of Basic Medical Sciences, Third Military Medical University, 30# Gaotanyan St., Shapingba District, Chongqing, China, **2** Medical Research Center, Southwest hospital, Third Military Medical University, 29# Gaotanyan St., Shapingba District, Chongqing, China

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163962>

SeqKitで出来ること、類似ツールとの比較

Features comparison

Categories	Features	seqkit	fasta_utilities	fastx_toolkit	pyfaidx	seqmagick	seqtk
Formats support	Multi-line FASTA	Yes	Yes	--	Yes	Yes	Yes
	FASTQ	Yes	Yes	Yes	--	Yes	Yes
	Multi-line FASTQ	Yes	Yes	--	--	Yes	Yes
	Validating sequences	Yes	--	Yes	Yes	--	--
	Supporting RNA	Yes	Yes	--	--	Yes	Yes
Functions	Searching by motifs	Yes	Yes	--	--	Yes	--
	Sampling	Yes	--	--	--	Yes	Yes
	Extracting sub-sequence	Yes	Yes	--	Yes	Yes	Yes
	Removing duplicates	Yes	--	--	--	Partly	--
	Splitting	Yes	Yes	--	Partly	--	--
	Splitting by seq	Yes	--	Yes	Yes	--	--
	Shuffling	Yes	--	--	--	--	--
	Sorting	Yes	Yes	--	--	Yes	--
	Locating motifs	Yes	--	--	--	--	--
	Common sequences	Yes	--	--	--	--	--
	Cleaning bases	Yes	Yes	Yes	Yes	--	--
	Transcription	Yes	Yes	Yes	Yes	Yes	Yes
	Translation	--	Yes	Yes	Yes	Yes	--
	Filtering by size	Yes	Yes	--	Yes	Yes	--
	Renaming header	Yes	Yes	--	--	Yes	Yes
Other features	Cross-platform	Yes	Partly	Partly	Yes	Yes	Yes
	Reading STDIN	Yes	Yes	Yes	--	Yes	Yes
	Reading gzipped file	Yes	Yes	--	--	Yes	Yes
	Writing gzip file	Yes	--	--	--	Yes	--

過去掲載

<https://bioinf.shenwei.me/seqkit>

類似のツールと比較して、
より多くのコマンドが利用でき、
高速である。

gz圧縮にも対応している。

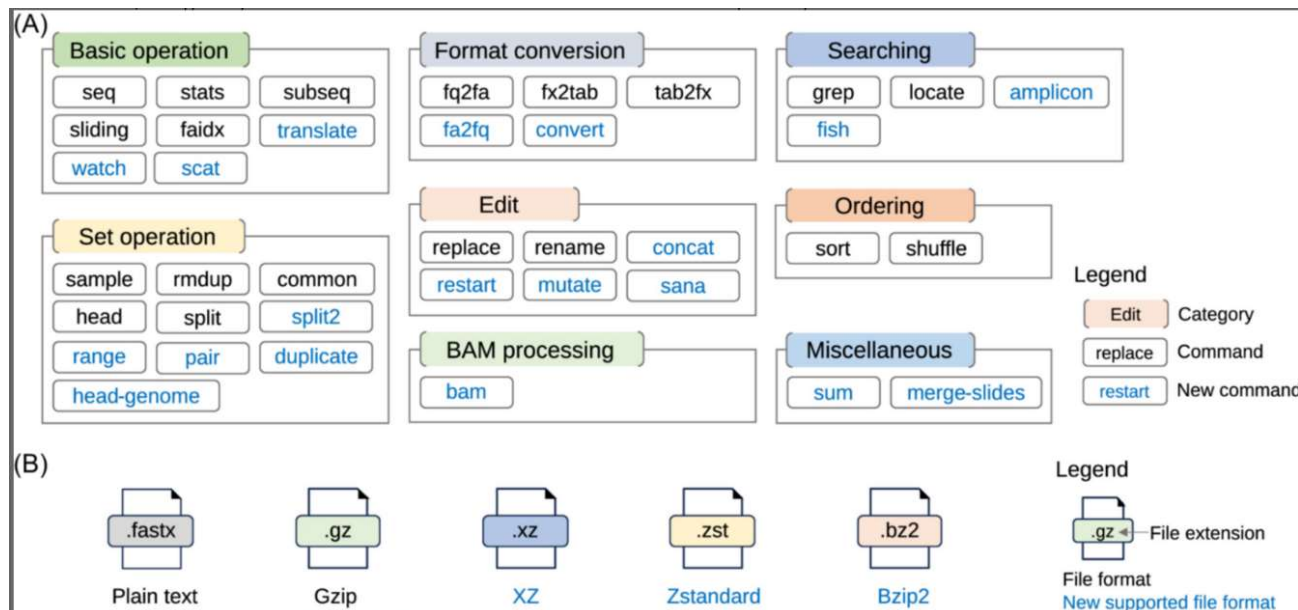
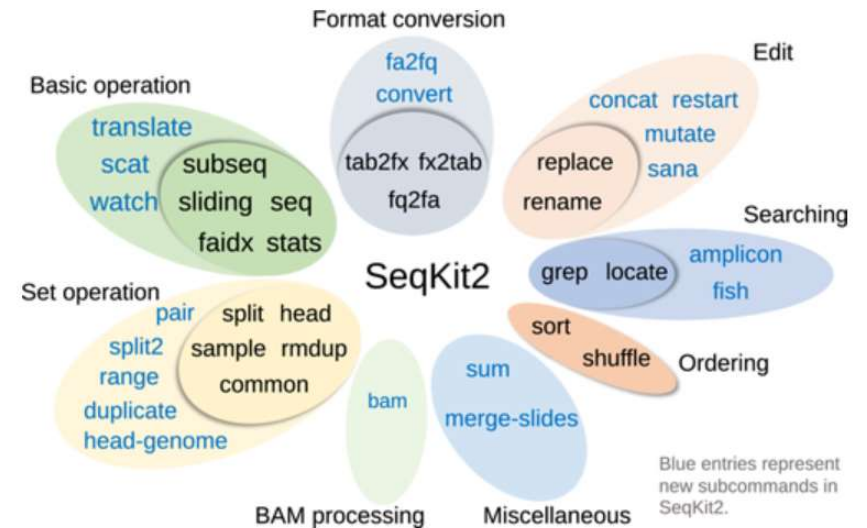
SeqKit2: A Swiss army knife for sequence and alignment processing

Wei Shen , Botond Sipos, Liuyang Zhao

First published: 05 April 2024 | <https://doi.org/10.1002/imt2.191> | Citations: 1

Seqkit2として論文も出ている

Graphical Abstract



最新はv2.9

```

$seqkit -h
SeqKit -- a cross-platform and ultrafast toolkit for FASTA/Q file manipulation
Version: 2.9.0
Author: Wei Shen <shenwei356@gmail.com>
Documents : http://bioinf.shenwei.me/seqkit
Source code: https://github.com/shenwei356/seqkit
:
:
Usage:
  seqkit [command]

Commands for Basic Operation:
  faidx      create the FASTA index file and extract subsequences
  scat       real time recursive concatenation and streaming of fastx files
  seq        transform sequences (extract ID, filter by length, remove gaps, reverse complement...)
  sliding    extract subsequences in sliding windows
  stats      simple statistics of FASTA/Q files
  subseq     get subsequences by region/gtf/bed, including flanking sequences
  translate  translate DNA/RNA to protein sequence (supporting ambiguous bases)
  watch      monitoring and online histograms of sequence features

Commands for Format Conversion:
  convert    convert FASTQ quality encoding between Sanger, Solexa and Illumina
  fa2fq      retrieve corresponding FASTQ records by a FASTA file
  fq2fa      convert FASTQ to FASTA
  fx2tab     convert FASTA/Q to tabular format (and length, GC content, average quality...)
  tab2fx     convert tabular format to FASTA/Q format

Commands for Searching:
  amplicon   extract amplicon (or specific region around it) via primer(s)
  fish       look for short sequences in larger sequences using local alignment
  grep       search sequences by ID/name/sequence/motifs, mismatch allowed
  locate     locate subsequences/motifs, mismatch allowed

Commands for Set Operation:
  common     find common/shared sequences of multiple files by id/name/sequence
  duplicate  duplicate sequences N times
  head       print first N FASTA/Q records
  head-genome print sequences of the first genome with common prefixes in name
  pair       match up paired-end reads from two fastq files
  range      print FASTA/Q records in a range (start:end)
  rmdup      remove duplicated sequences by ID/name/sequence
  sample     sample sequences by number or proportion
  split      split sequences into files by id/seq region/size/parts (mainly for FASTA)
  split2     split sequences into files by size/parts (FASTA, PE/SE FASTQ)

Commands for Edit:
  concat     concatenate sequences with the same ID from multiple files
  mutate     edit sequence (point mutation, insertion, deletion)
  rename     rename duplicated IDs
  replace    replace name/sequence by regular expression
  restart    reset start position for circular genome
  sana       sanitize broken single line FASTQ files

Commands for Ordering:
  shuffle    shuffle sequences
  sort       sort sequences by id/name/sequence/length

Commands for BAM Processing:
  bam        monitoring and online histograms of BAM record features

Commands for Miscellaneous:
  merge-slides merge sliding windows generated from seqkit sliding
  sum        compute message digest for all sequences in FASTA/Q files
:
:

```

seqkitと打つとサブコマンドリストが出る
サブコマンドリストまで打って-hで、
その使い方や説明が表示される

Seqkitコマンド例

fastq/fastqファイルのstatisticsを見る

```
$ seqkit stats hoge.fastq
```

fastqファイルからfastaファイルに変換する

```
$ seqkit fq2fa hoge.fastq > hoge.fasta
```

fastq/fastqファイルを複数のファイルに分割する

```
seqkit split -p 2 hoge.fastq
```

fastq/fastqファイルから一部のsamplingする

-nで指定する数は厳密なsampling数とは合致しないので注意

```
$ seqkit sample -n 100 hoge.fastq > hoge_100.fastq
```

fastq/fastqファイルのread順番をシャッフルする


```
$ seqkit shuffle hoge.fastq > shf_hoge.fastq
```


fastq/fastqファイルの上から単純にsamplingする


```
$ seqkit head -n 100 hoge.fastq > hoge_head100.fastq
```


Bowtieを使って見よう

- Burrows-Wheeler 変換に基づくインデックスを利用したショートリードのマッピングプログラム
- BowtieとBowtie2がある。後者はギャップを考慮した検索を行い、感度がより高い。また、検索の方針が単純化されて分かりやすくなるなど、多くの点で改良されている。
- シーケンスのリード長が長い（50bp以上）時はBowtie2の方が一般に検索効率がよく、精度も高い。リード長が短い（50bp未満）時はBowtieの方が検索効率または精度がいい場合もある。

**Bowtie**
An ultrafast memory-efficient short read aligner

JOHNS HOPKINS
UNIVERSITY

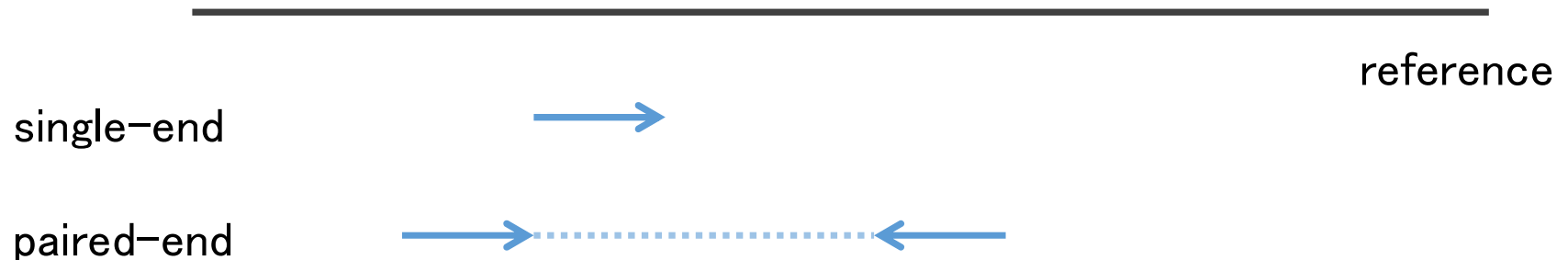
**Bowtie 2**
Fast and sensitive read alignment

JOHNS HOPKINS
UNIVERSITY

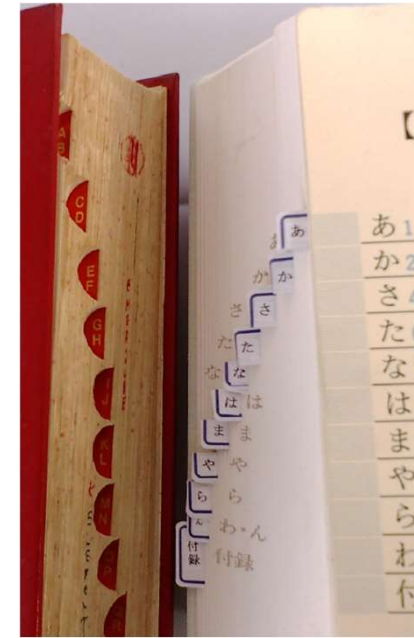
リファレンス配列へのマッピング

Bowtie, BWA, SOAP など

- 長大なリファレンス配列に、大量の短いリード配列を若干のミスマッチを許して照合する
- リファレンス配列に対して、あらかじめ全文検索インデックスを作成することにより高速に検索を行う
- paired-end read に対応。



インデックスとは



辞書における
インデックスタブ

- 索引、目次、見出し
- ファイルのどの辺りに何が書いてあるかの指標
- インデックスを作成すると別ファイルができるのは、分厚い本の「別冊目次」ができるイメージ
- 欲しい情報を探すのにファイル（本）を先頭から総ナメして探さなくてもよい

リファレンス配列のインデックスを作成



bowtie2-build リファレンス配列ファイル インデックス名

- 実行すると、インデックスとして、
 - ✓ インデックス名.n.bt2 (n=1-4)
 - ✓ インデックス名.rev.m.bt2 (m=1-2)の、計6つのファイルが作成される
- 配列ファイルはカンマで区切って複数を指定可能

実習 3 bowtie2-build

実習用ディレクトリ ~/gitc/data/5_ngs に移動して ls で中を見る

```
$ cd  
~/gitc/data/5_ngs  
$ ls
```

リファレンス用ゲノムデータ（FASTA形式） **ecoli_genome.fa**

- bowtie2用インデックスの作成（インデックス名：**eco**）

```
$ bowtie2-build ecoli_genome.fa eco
```

- インデックスから元の配列データを再構築

```
$ bowtie2-inspect eco | less
```

NGSマッピングの実行 bowtie2



- マッピングの実行

- ✓ single-end read の場合

```
bowtie2 -x インデックス名 -U リードファイル -S 出力ファイル
```

- ✓ paired-end read の場合

```
bowtie2 -x インデックス名 -1 リードファイル1  
-2 リードファイル2 -S 出力ファイル
```

(実際は改行せずに1行で打つ)

- リードファイルはカンマ区切りで複数を指定可能

実習 4 bowtie2

- リード配列 (FASTQ 形式, single-end read)
ecoli.fastq

リファレンス配列のインデックス名 (実習4で作ったもの)
eco

- bowtie2の実行

```
$ bowtie2 -x eco -U ecoli.fastq -S eco_bowtie2.sam
```

マッピング結果：SAMフォーマットファイル

```
$ less -S eco_bowtie2.sam
```

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 chr1 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 chr1 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 chr1 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 chr1 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 chr1 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 chr1 37 30 M = 7 -39 CAGCGCCAT *
```

テンプレート名

フラグ

マップ結果

アライメント
(CIGAR)

対となるリード
の位置情報

リードの配列

オプション

Bowtie2: その他のオプション



- **-h** ヘルプを表示する
- **-a** 全てのアライメントを表示する
- **-p** 整数 指定した数のCPUコアを使って実行する
- **-f** リードがFASTA形式のファイルである
- 他、Bowtie2 マニュアル詳細
<http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>

Samtoolsを使って見よう



Samtools

Home Download ▾ Workflows ▾ Documentation ▾ Support ▾


Samtools

SAM<->BAM等の変換、データのソート、検索付け、
特定readの抽出、統計情報収集などができる

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

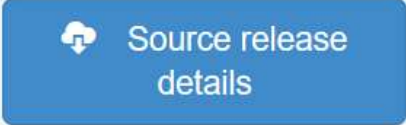
Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data


Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.



Download

Source code releases can be downloaded from [GitHub](#) or [Sourceforge](#):






Workflows


We have described some standard workflows using Samtools:

- WGS/WES Mapping to Variant Calls
- Using CRAM within Samtools



Documentation

- Manuals
- HowTos
- Specifications
- Duplicate Marking
- Zlib Benchmarks
- CRAM Benchmarks
- Publications



Support

- Mailing Lists
- HTSlib issues
- BCFtools issues
- Samtools issues

<http://www.htslib.org/>

現行の最新はv1.21

バージョンによってオプションの与え方が変わっているコマンドに注意

NGSデータを扱うための最も基盤となるツール

Samtoolsの起動

\$ samtools

```
$ samtools
```

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.21 (using htslib 1.21)
```

```
Usage:  samtools <command> [options]
```

```
Commands:
```

```
-- Indexing
dict          create a sequence dictionary file
faidx         index/extract FASTA
fqidx         index/extract FASTQ
index         index alignment

-- Editing
calmd         recalculate MD/NM tags and '=' bases
fixmate       fix mate information
reheader      replace BAM header
targetcut     cut fosmid regions (for fosmid pool only)
addreplacerg  adds or replaces RG tags
markdup       mark duplicates
ampliconclip  clip oligos from the end of reads

-- File operations
collate       shuffle and group alignments by name
cat           concatenate BAMs
consensus     produce a consensus Pileup/FASTA/FASTQ
merge         merge sorted alignments
mpileup       multi-way pileup
sort          sort alignment file
split         splits a file by read group
quickcheck    quickly check if SAM/BAM/CRAM file appears intact
fastq         converts a BAM to a FASTQ
fasta         converts a BAM to a FASTA
import        Converts FASTA or FASTQ files to SAM/BAM/CRAM
reference     Generates a reference from aligned data
reset         Reverts aligner changes in reads

-- Statistics
bedcov        read depth per BED region
coverage      alignment depth and percent coverage
depth         compute the depth
flagstat      simple stats
idxstats      BAM index stats
cram-size     list CRAM Content-ID and Data-Series sizes
phase         phase heterozygotes
stats         generate stats (former bamcheck)
ampliconstats generate amplicon specific stats

-- Viewing
flags         explain BAM flags
head          header viewer
tview         text alignment viewer
view          SAM<->BAM<->CRAM conversion
depad         convert padded BAM to unpadded BAM
samples       list the samples in a set of SAM/BAM/CRAM files

-- Misc
help [cmd]    display this help message or help for [cmd]
version       detailed version information
```

オプション/引数
なしで起動すると
Samtools の基本的な
使い方が表示される

以降、実習しながら進めます
(実習5)

Samtools の起動: コマンド簡易マニュアル

基本的な使い方: `$ samtools command options`

`$ samtools view`

コマンドを付けてオプション無しで実行すると
そのコマンドのマニュアルが表示される

```
Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]
```

Output options:

<code>-b, --bam</code>	Output BAM
<code>-C, --cram</code>	Output CRAM (requires -T)
<code>-l, --fast</code>	Use fast BAM compression (and default to --bam)
<code>-u, --uncompressed</code>	Uncompressed BAM output (and default to --bam)
<code>-h, --with-header</code>	Include header in SAM output
<code>-H, --header-only</code>	Print SAM header only (no alignments)
<code>--no-header</code>	Print SAM alignment records only [default]
<code>-c, --count</code>	Print only the count of matching records
<code>-o, --output FILE</code>	Write output to FILE [standard output]
<code>-U, --unoutput FILE, --output-unselected FILE</code>	Output reads not selected by filters to FILE
<code>-p, --unmap</code>	Set flag to UNMAP on reads not selected then write to output file.
<code>-P, --fetch-pairs</code>	Retrieve complete pairs even when outside of region

Input options:

<code>-t, --fai-reference FILE</code>	FILE listing reference names and lengths
<code>-M, --use-index</code>	Use index and multi-region iterator for regions
<code>--region[s]-file FILE</code>	Use index to include only reads overlapping FILE
<code>-X, --customized-index</code>	Expect extra index file argument after <in.bam>
	:
	:

<http://www.htslib.org/doc/samtools.html> を参照

SAM/BAM変換

samtools view *options...*

- SAMファイルからBAMファイルの作成

```
$ samtools view -bS eco_bowtie2.sam -o eco_bowtie2.bam
```

- BAMをSAMに変換して less コマンドで表示

```
$ samtools view eco_bowtie2.bam | less
```

- BAMファイルを less で読もうとすると... ?

```
$ less eco_bowtie2.bam
```

- SAMファイルに比べてBAMファイルのサイズは？

```
$ ls -l eco_bowtie2.*
```

Samtoolsによるsort

samtools sort *options...*

```
$ samtools sort
Usage: samtools sort [options...] [in.bam]
Options:
-l INT      Set compression level, from 0 (uncompressed) to 9 (best)
-u          Output uncompressed data (equivalent to -l 0)
-m INT      Set maximum memory per thread; suffix K/M/G recognized [768M]
-M          Use minimiser for clustering unaligned/unplaced reads
-R          Do not use reverse strand (only compatible with -M)
-K INT      Kmer size to use for minimiser [20]
-I FILE     Order minimisers by their position in FILE FASTA
-w INT      Window size for minimiser indexing via -I ref.fa [100]
-H          Squash homopolymers when computing minimiser
-n          Sort by read name (natural): cannot be used with samtools index
-N          Sort by read name (ASCII): cannot be used with samtools index
-t TAG      Sort by value of TAG. Uses position as secondary index (or read name if -n is set)
-o FILE     Write final output to FILE rather than standard output
-T PREFIX   Write temporary files to PREFIX.nnnn.bam
  --no-PG      Do not add a PG line
  --template-coordinate Sort by template-coordinate
  --input-fmt-option OPT[=VAL]
                  Specify a single input file format option in the form
                  of OPTION or OPTION=VALUE
-O, --output-fmt FORMAT[,OPT[=VAL]]...
                  Specify output format (SAM, BAM, CRAM)
  --output-fmt-option OPT[=VAL]
                  Specify a single output file format option in the form
                  of OPTION or OPTION=VALUE
  --reference FILE
                  Reference sequence FASTA FILE [null]
-@, --threads INT
                  Number of additional threads to use [0]
  --write-index
                  Automatically index the output files [off]
  --verbosity INT
                  Set level of verbosity
```

マッピングデータをリファレンス配列上の位置順に並び替える
これをしないとindexを付けられない

BAM ファイルのソート

samtools sort *options...*

```
$ samtools sort eco_bowtie2.bam -o  
eco_bowtie2_sorted.bam
```

- samからの直接変換も可能 (v1.3以降)

```
$ samtools sort eco_bowtie2.sam -o  
eco_bowtie2_sorted.bam
```

- ソートされたBAMファイルをSAMに変換してlessで表示

```
$ samtools view eco_bowtie2_sorted.bam | less
```

- 元のSAMファイルの表示と比較してみよう

```
$ less eco_bowtie2.sam
```

BAMファイルにインデックスを付ける

samtools index options...

- 先にソートされている必要がある
- インデックスは .bai という拡張子付きの別ファイルで生成される。
- 「bamファイル名.bai」が作成されたのを ls コマンドで確認

```
$ samtools index eco_bowtie2_sorted.bam
```

```
$ ls eco_bowtie2_sorted*
```

ここから先はソート & インデックス付与したbamファイルを使う

ソート & インデックス付与したbamファイルを使って

指定した領域内のマッピング結果を表示

```
$ samtools view eco_bowtie2_sorted.bam chr:200-500
```



染色体名：開始位置 - 終了位置

マッピング統計情報収集 1

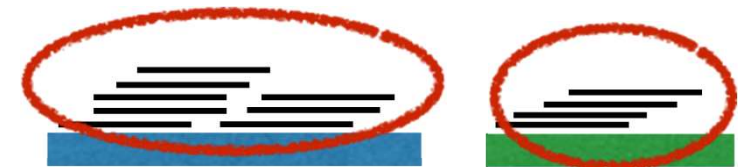
samtools idxstats *options...*

- 染色体毎にマップされたリード数を得る

```
$ samtools idxstats eco_bowtie2_sorted.bam
```

染色体名	染色体配列長	マップされた リード数	片側のみマップさ れたリード数
chr	4639675	326754	0
*	0	0	3364

マップされなかったリード数
染色体名が '*' として表示される



マッピング統計情報収集 2

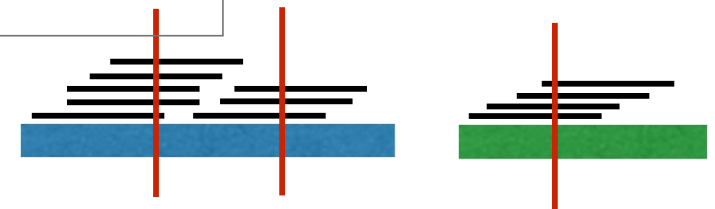
samtools depth *options...*

- 深度（マップされた回数）の統計情報を得る

```
$ samtools depth eco_bowtie2_sorted.bam
```

染色体名	位置	深度（マップされた回数）
------	----	--------------

chr	2753929	1533
chr	2753930	1470
chr	2753931	1446
chr	2753932	1101
chr	2753933	922
chr	2753934	918



紹介したSamtools コマンドまとめ

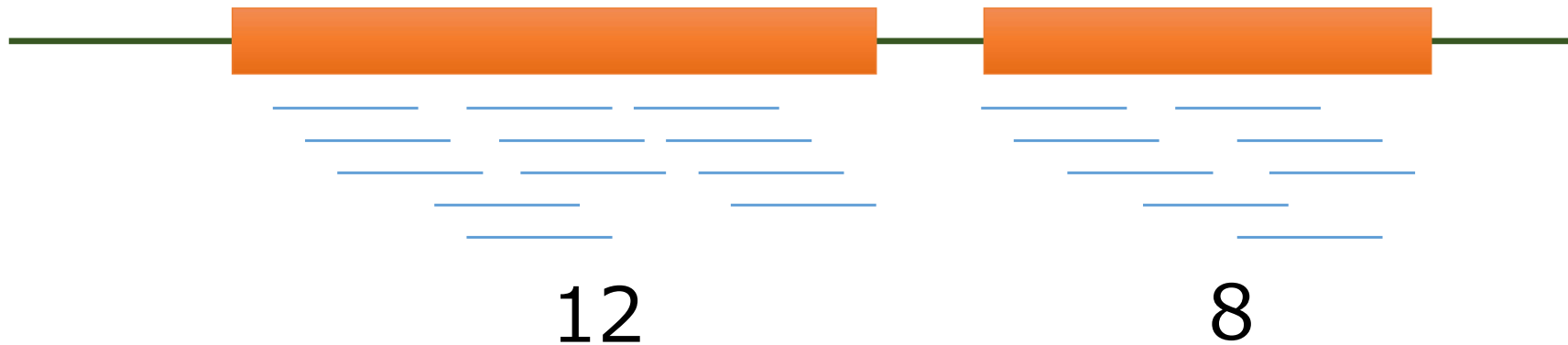


samtools

view	リードを抽出, SAM/BAM変換
sort	ソート
index	.bamのインデックス作成
idxstats	染色体毎のマッピング状況
depth	位置毎のマッピング深度

RNA-Seq解析に向けて

- ゲノム上にマッピングされたリードを遺伝子領域ごとに集めて数をカウント



通常、カウントした数を遺伝子の長さ、およびマップされたリード全体の数で割って標準化する

RPKM (Read Per Kilobase per Million mapped reads)

FPKM (Fragment Per Kilobase per Million mapped reads)

TPM (Transcript Per Million mapped reads)

統計解析を含めた内容は「RNA-seq入門」で

今回紹介したツール・ファイルのまとめ

SeqKit

ゲノム (リファレンス) 配列
FASTAファイル

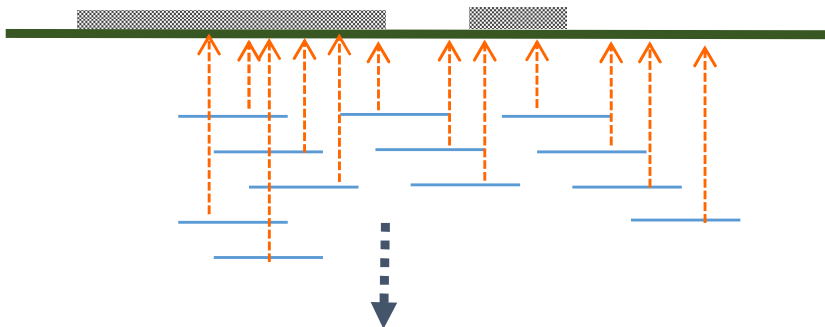
```
>chr
AGCTTTTCATTCGACTGCAACGGGCAATATGCT
CTGTGTGGATTAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATAAA
TTTTATGACTTGGTCAATACTTTAACCAG
TATAGCATAAGCCACAGACAGATAAAATTACAG
AGTACACACATCCATGAACGCAATAGCACACC
```

サンプルリード (ゲノム DNA/RNA)
FASTQファイル (配列+クオリティ)

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGCTGCGCCACCGACCTATGTTCCGCGCAATACAGCTGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@D>DDFF7DC?FFBFF@DFII<DF@AA6AFBBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCACTCCGCTGCAATCAGCAATCCAGTCCCTCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFDPRDPRDFFHIIIEGHHJJJJGHFGHGGHGGUDDGCH+GGGHH
```

インデックス作成 **bowtie2-build**

リファレンス配列へのマッピング **bowtie2**



遺伝子アノテーション GFF(GTF)ファイル

```
chr RefSeq start_codon 190 192 1.000 + . gene_id "b0001"; transcript_id "b0001";
chr RefSeq CDS 190 252 1.000 + 0 gene_id "b0001"; transcript_id "b0001";
chr RefSeq stop_codon 253 255 1.000 + . gene_id "b0001"; transcript_id "b0001";
chr RefSeq exon 190 255 1.000 + . gene_id "b0001"; transcript>
```

遺伝子ごとの集計

**RNA-seq入門
にて**

b0001	11
b0002	117
b0003	36

マッピング結果 SAM ファイル

```
@HD      VN:1.0      SO:unsorted
@SQ      SN:chr      LN:4639675
@PG      ID:bowtie2      PN:bowtie2      VN:2.2.4      CL:"/bio/bin/bowtie2-align
SRR1515276.40 0 chr 4423609 42 51M * 0 0 GGAATTCCTCACTGCCA
SRR1515276.158 16 chr 501700 42 51M * 0 0 ACCGACCGAGTCCAAAG
SRR1515276.212 4 * 0 0 * * 0 0 GCGCGCTTTACCGGTGT
SRR1515276.319 0 chr 2922768 42 51M * 0 0 GCTTAAGTTGATTAAAG
SRR1515276.367 16 chr 2753873 42 51M * 0 0 GCGTGTCGTCGCGACG
SRR1515276.411 0 chr 3440721 42 51M * 0 0 ACCGCATAAATTTCTTGA
SRR1515276.434 0 chr 4198737 42 51M * 0 0 GCGCGTACCGATCTGG
```

samtools

BAMファイル

並べ替え
検索
ゲノムブラウザへ