

機能ゲノム学 第3回

¹大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム

²微生物科学イノベーション連携研究機構

門田幸二(かどた こうじ)

kadota@iu.a.u-tokyo.ac.jp

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

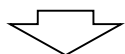
Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

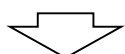
全体像のおさらい

①QCには、FastQCによるクオリティチェック、フィルタリングやトリミングの作業が含まれる

NGSリードデータ(SRAファイル)

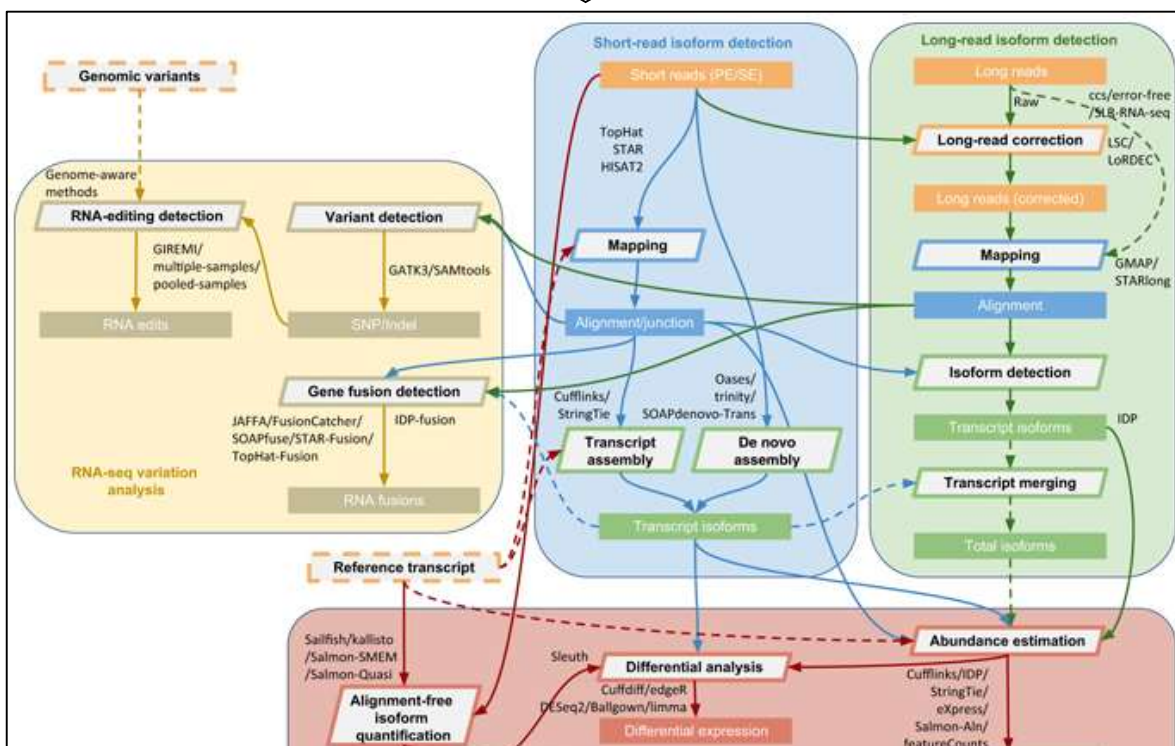


NGSリードデータ(FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)

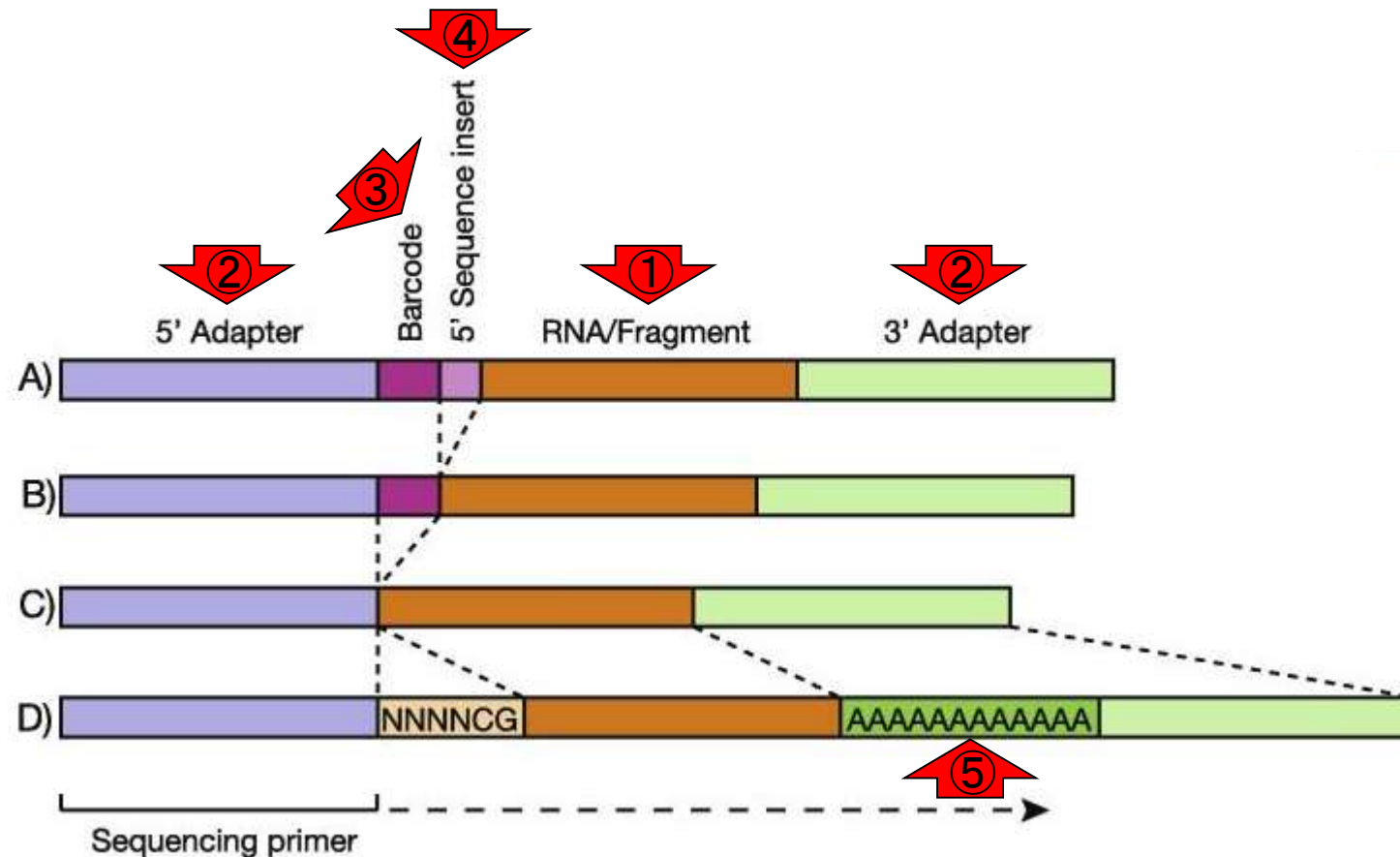
①



RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

NGSリードの模式図

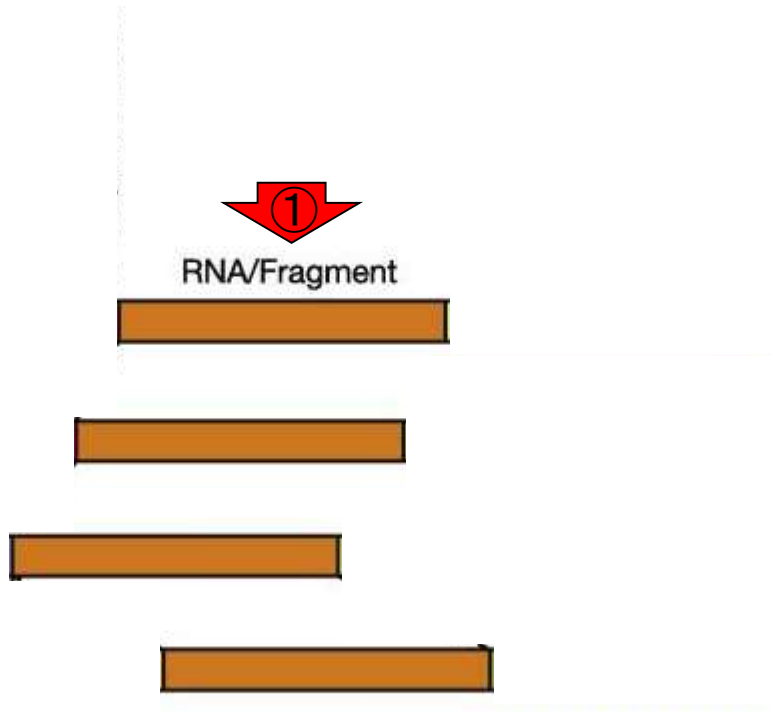
実験デザインにもよるが、RNA-seqのNGSリードには、目的の①RNA断片配列以外に、②アダプター配列、③バーコード配列、④インサート配列、⑤ポリA配列などが含まれる



Kraken (Davis et al., Methods, 63: 41-49, 2013)の図2

QCの目的

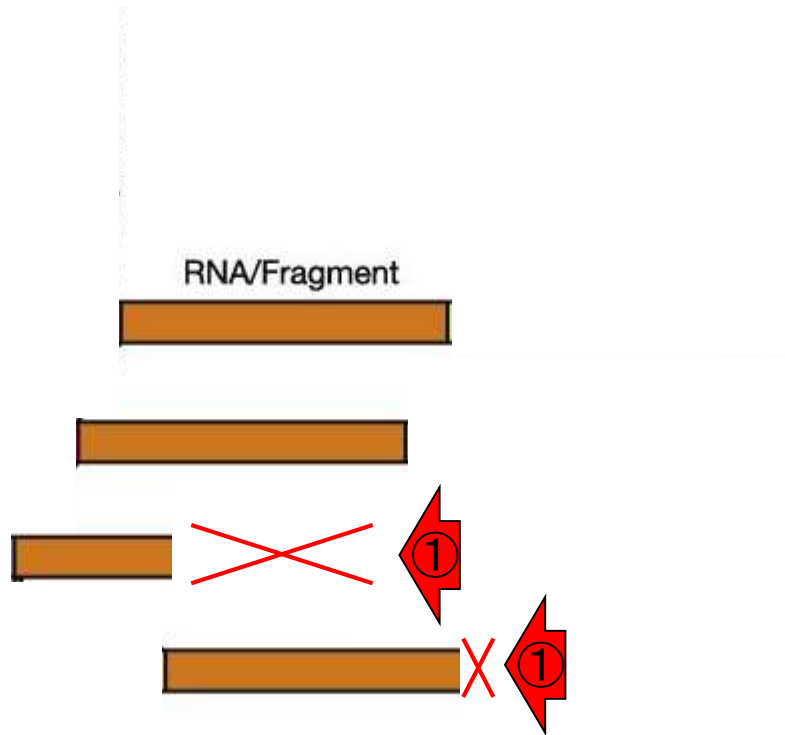
QCの目的は、基本的にはこんな感じにして、①目的の塩基配列のみ



Kraken (Davis et al., Methods, **63**: 41–49, 2013)の図2を改変

その上でさらに、クオリティスコアが設定した閾値以下の領域をトリミングしたりして…

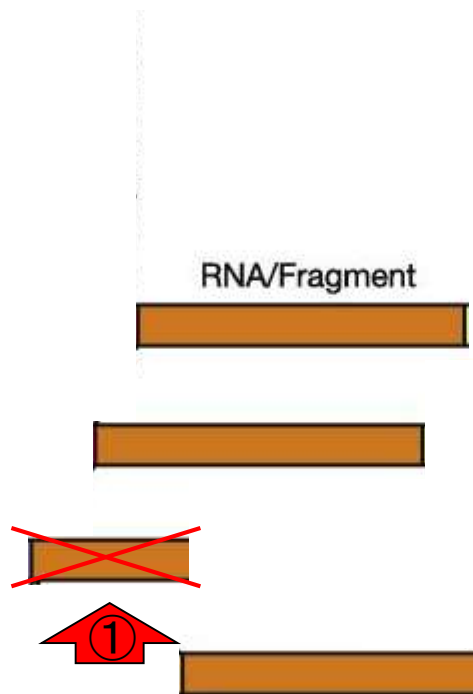
QCの目的



Kraken (Davis et al., Methods, **63**: 41–49, 2013)の図2を改変

QCの目的

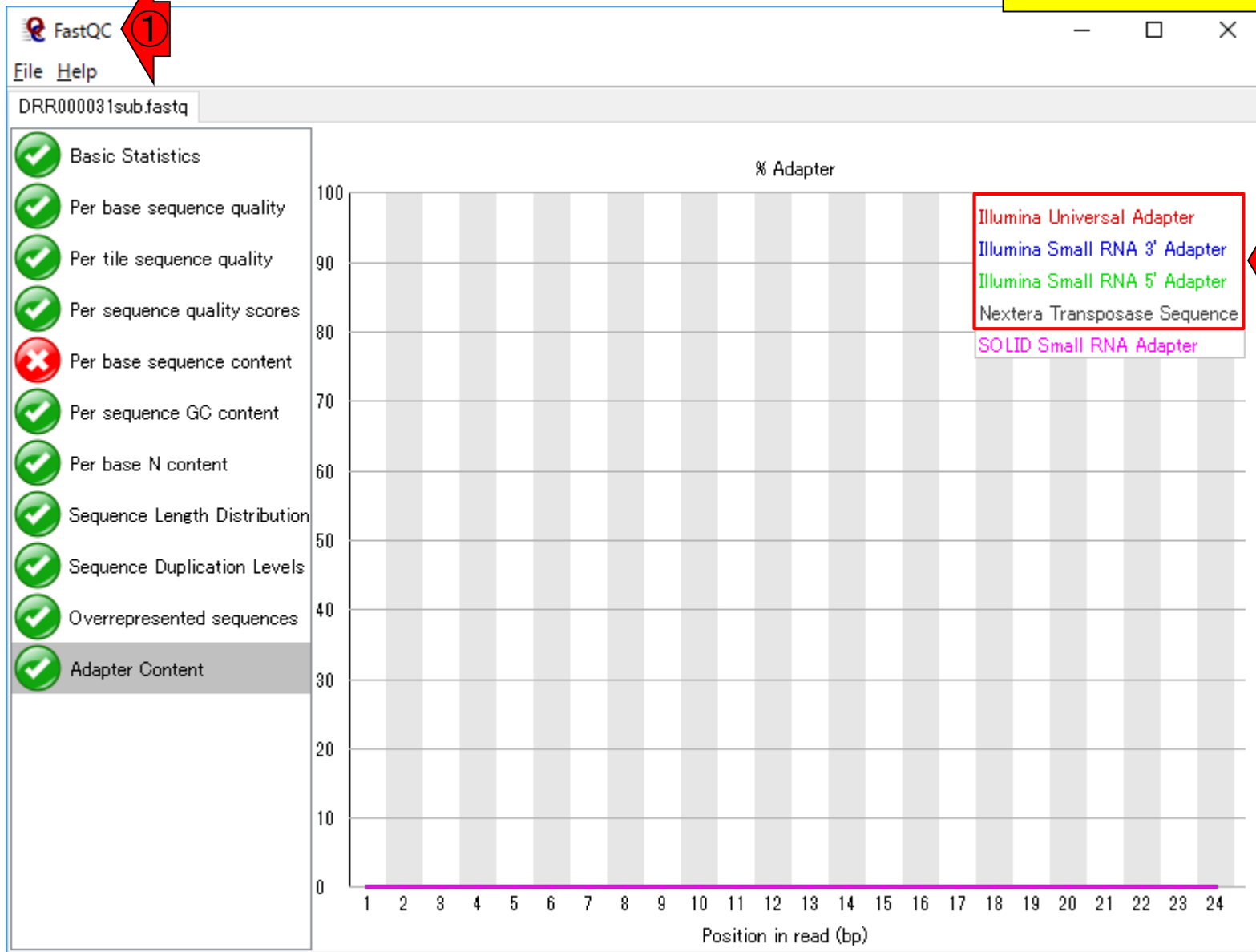
①一定の長さ未満のリードは捨てて (filter out)、残りのものを出力するのが、②KrakenなどのQC用プログラム



Kraken (Davis et al., Methods, **63**: 41–49, 2013)の図2を改変




多くのプログラムが...

①FastQCしかり、RNA-QC-chainやKrakenなど多くのプログラムが②IlluminaのNGSデータに対応しています



RNA-seqとNGS機器

①NGS機器(プラットフォーム)ごとの②RNA-seqに用いられたラン数(i.e., 実験数)。③Illuminaのデータが圧倒的に多いことがわかる



Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t002>

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)のTable 2

RNA-seqとNGS機器

Illuminaは、①歴史が古く、③産出データ量(リード数)も多く、②配列決定精度も高かったので、④このような結果になったのでしょうか。発現解析分野でのマイクロアレイからの移行を後押しするのに十分な条件が整っていたともいえる

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

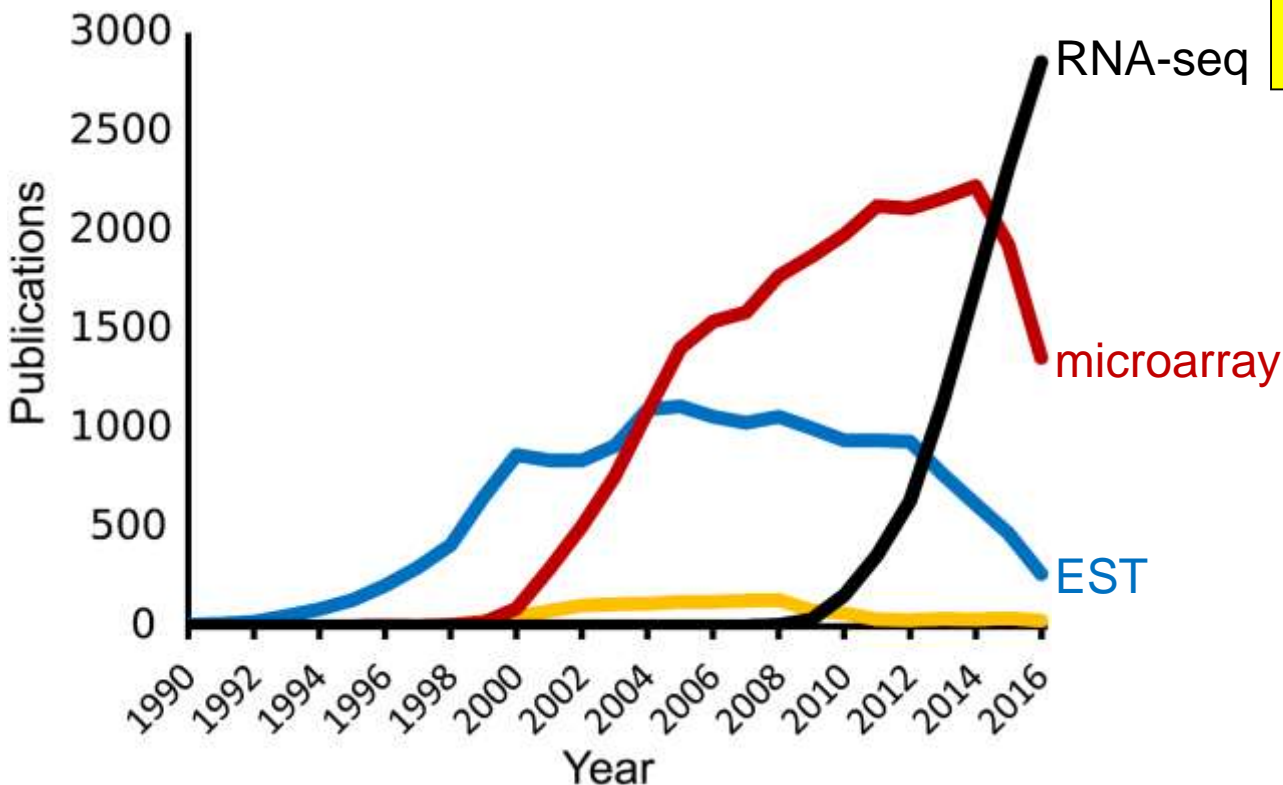
NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t002>

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)のTable 2

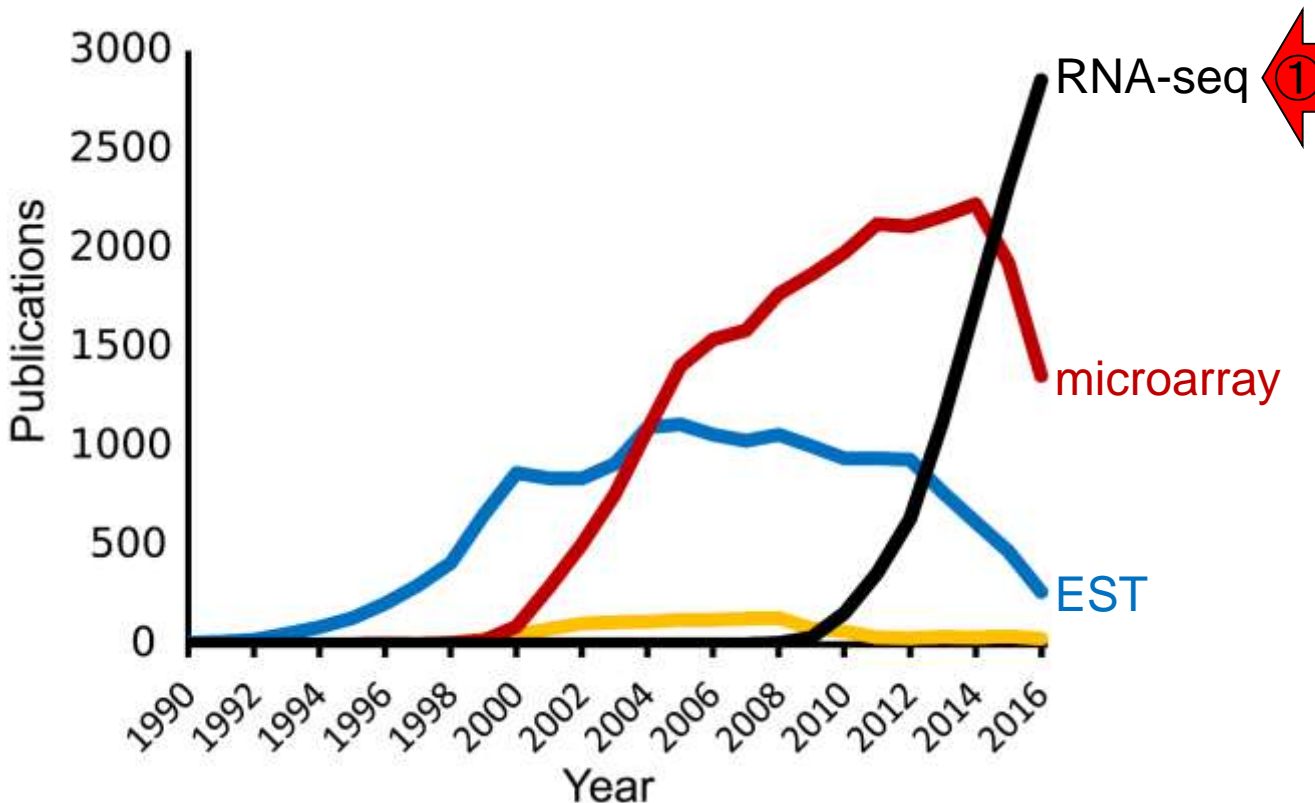
テクノロジーの栄枯盛衰

publishされた論文総数は、microarrayのほうがおそらくまだ多い。また、RNA-seqデータ解析についてもmicroarrayの経験が生かされている。
①右肩上がりのRNA-seqと②右肩下がりのmicroarrayの論文数の推移を鑑み、平成30年度の本科目の講義内容をRNA-seqに据えることとしました



最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)の図1を少し改変

テクノロジーの栄枯盛衰



最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)の図1を少し改変

RNA-seqとNGS機器

①この表には載っていないが、Oxford Nanopore社のMinIONに代表されるナノポアシーケンサーも2017年頃より本格的に普及してきた。RNA-seqとの関連でいえば、cDNAへの変換変換(逆転写)・PCR増幅・サイズ選択も不要であり、RNAそのものを直接シーケンス可能な点が注目を集めている(第1回のおさらい)。

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160



NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t002>

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)のTable 2

RNA-seqデータ解析

RNA-Seq data analysis ②

RNA-Seq experiments generate a large volume of raw sequence reads, which have to be processed to yield useful information. Data analysis usually requires a combination of [bioinformatics software](#) tools that vary according to the experimental design and goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a [command-line interface](#), either in a [Unix](#) environment or within the [R/Bioconductor](#) statistical environment [90].

Quality control.

Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs ([k-mers](#)), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the FastQC and FaQCs software packages [91][92]. Abnormalities identified may be removed by trimming or tagged for special treatment during later processes.



① 最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)

RNA-seqデータ解析

①QC、②アラインメント(マッピングのこと)、③定量化(カウントデータ取得のこと)、④発現変動解析、からなる。もちろんこれは発現変動解析がゴールの場合であり、ステップの分け方なども⑤の論文著者の主観

RNA-Seq data analysis

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005

processed to yield useful information. Data analysis usually requires a combination of **bioinformatics software** tools that vary according to the experimental design goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a **command-line interface**, either on a **Unix** environment or within the **R/Bioconductor** statistical environment [90].

Quality control.

Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs (**k-mers**), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the FastQC and FaQCs software packages [91][92]. Abnormalities identified may be removed by trimming or tagged for special treatment during later processes.

全体像のおさらい

①QCには、FastQCによるクオリティチェック、フィルタリングやトリミングの作業が含まれる

NGSリードデータ(SRAファイル)

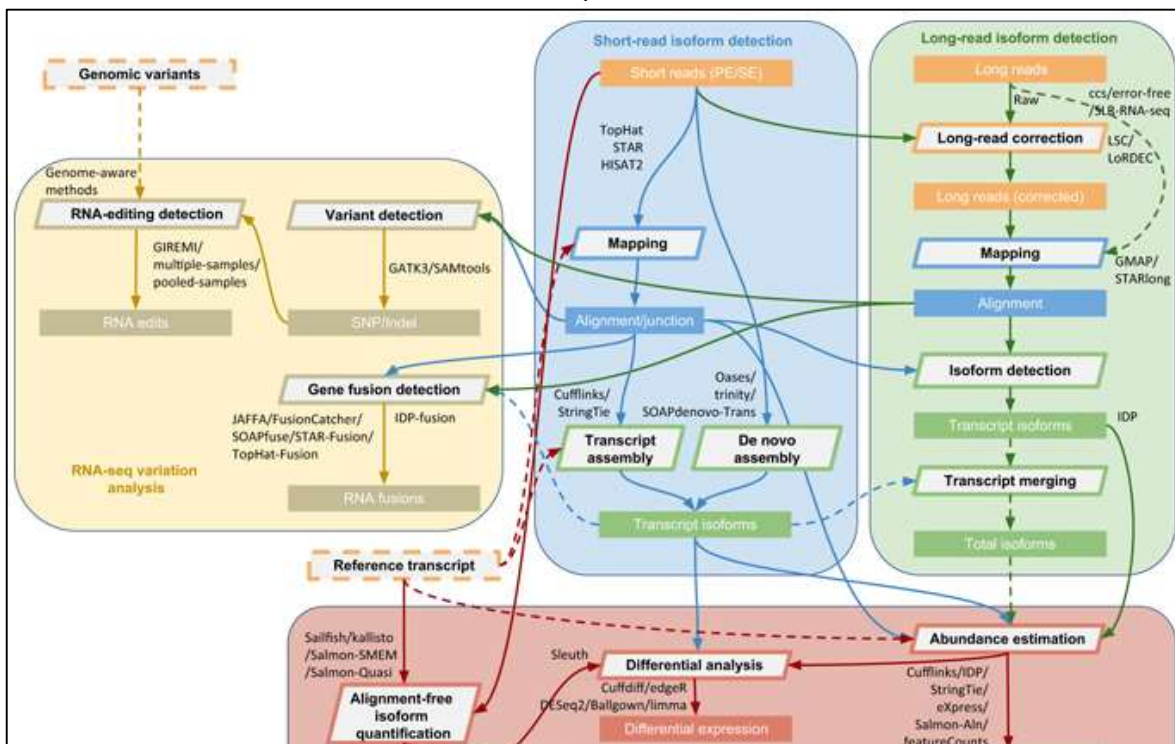


NGSリードデータ(FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)

①



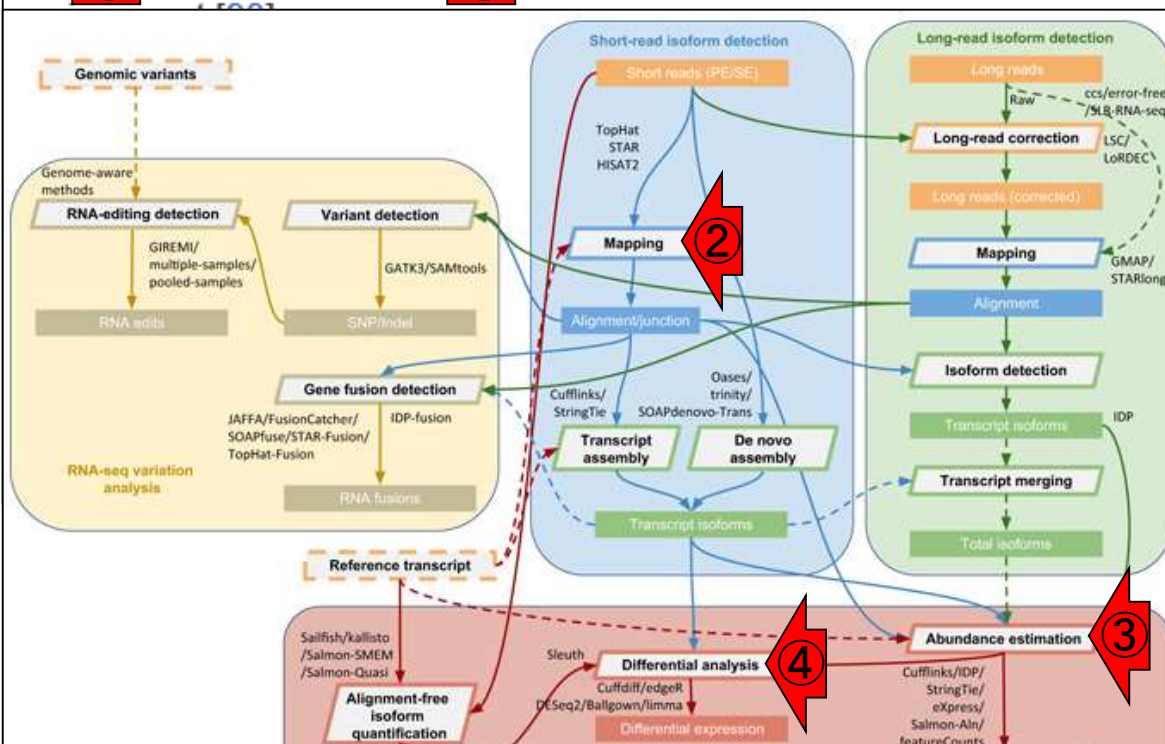
RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

全体像のおさらい

RNA-Seq data analysis

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)

processed to yield useful information. Data analysis usually requires a combination of **bioinformatics software** tools that vary according to the experimental design. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a command-line interface, either in a Unix environment or within the R/Bioconductor statistical



ence needs to be
ty scores for base
representation of
duplication rate [85].
and FaQCs software
tagged for special

RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

全体像のおさらい

①QCには、FastQCによるクオリティチェック、フィルタリングやトリミングの作業が含まれる

NGSリードデータ(SRAファイル)

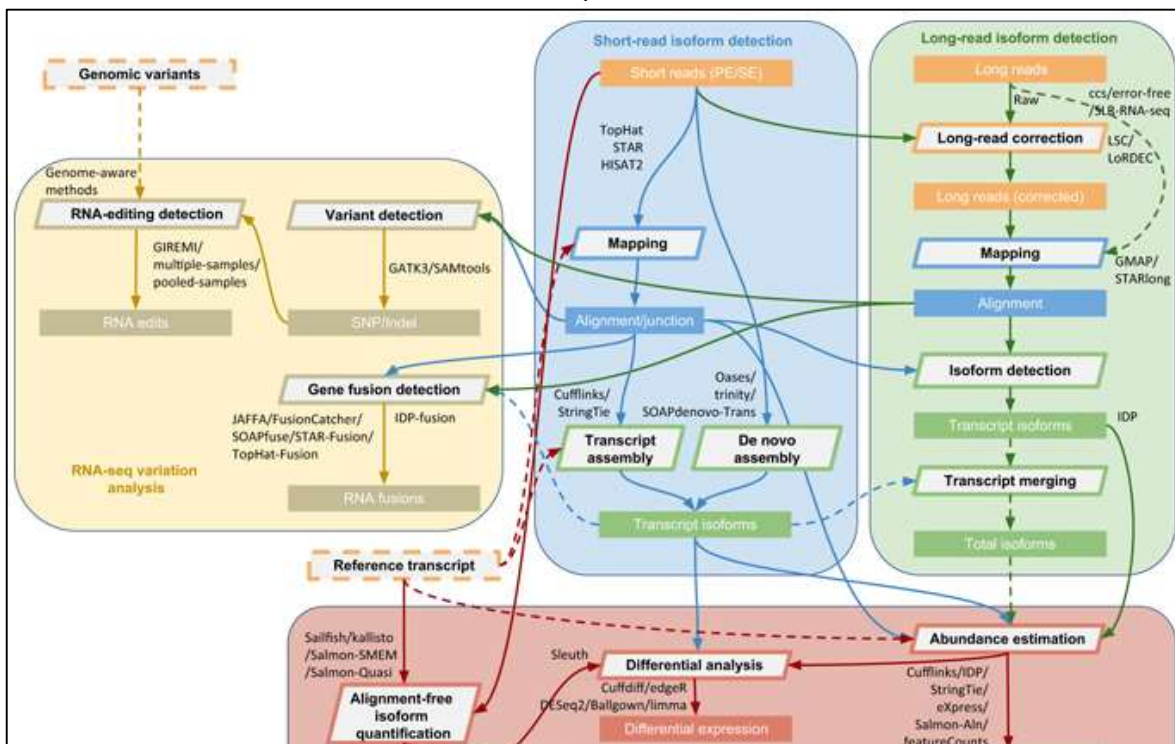


NGSリードデータ(FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)

①



RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

FastQCの位置づけ

FastQCで行うクオリティチェックに関する記述は、①QCのところの、②のあたり。③FastQCは、フィルタリングやトリミングの実行前後に行うことで、うまくフィルタリングできているかなどを確認する

RNA-Seq data analysis

RNA-Seq experiments generate a large volume of raw sequence reads, which have to be processed to yield useful information. Data analysis usually requires a combination of [bioinformatics software](#) tools that vary according to the experimental design and goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a [command-line interface](#), either in a [Unix](#) environment or within the [R/Bioconductor](#) statistical environment [90].

Quality control. ①

Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs ([k-mers](#)), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the [FastQC](#) and [FaQCs](#) software packages [91][92]. Abnormalities identified may be removed by [trimming](#) or tagged for special treatment during later processes. ②

③

FaQCs

①FaQCsは、フィルタリングやトリミングを実行するプログラム。それゆえ、例えば①FaQCsを実行した結果のFASTQファイルをさらに②FastQCにかけることで、アダプター配列のトリミングなどがうまくできているかを確認する、みたいな使い方をします。①FaQCsは、③Linux環境で動くプログラムです

RNA-Seq data analysis

RNA-Seq experiments generate a large volume of raw sequence reads, which have to be processed to yield useful information. Data analysis usually requires a combination of [bioinformatics software](#) tools that vary according to the experimental design and goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a [command-line interface](#), either in a [Unix environment](#) or within the [R/Bioconductor](#) statistical environment [90].



Quality control.

Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs ([k-mers](#)), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the [FastQC](#) and [FaQCs](#) software packages [91][92]. Abnormalities identified may be removed by [trimming](#) or [tagging](#) for special treatment during later processes.



FastQCとFaQCsの関係

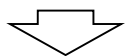
NGSリードデータ(SRAファイル)



NGSリードデータ(FASTQファイル)

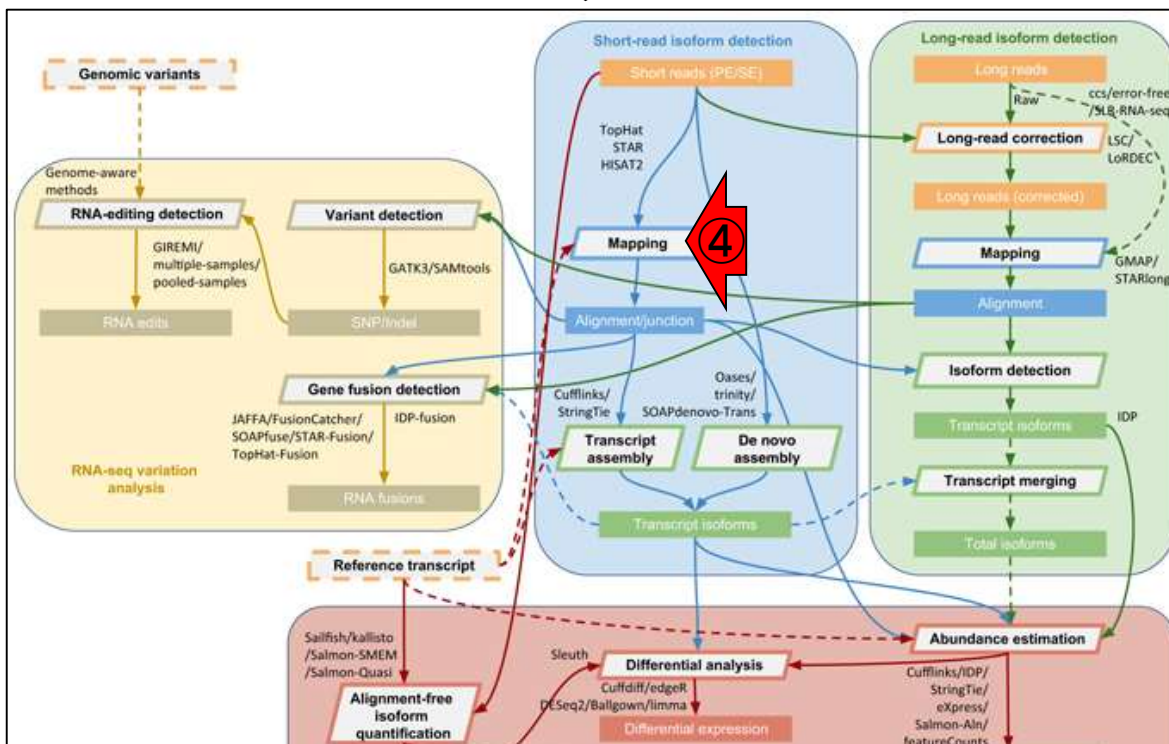


前処理(preprocessing) or Quality Control (QC)



QCの枠組みの中で、①FastQCでデータの全体像を把握し、②FaQCsを実行して、③その結果をFastQCで確認し、④Mappingの入カデータとして用いる

- ①FastQC(クオリティチェック)
- ②FaQCs(フィルタリングやトリミング)
- ③FastQC(クオリティチェック)



RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

FaQCsのインストール

FaQCsのインストールは、①NGS連載第4回に解説あり。②原稿PDF中のp131、③ウェブ資料PDF中のW15-2以降です。もちろん、それ以前の説明などを理解していないと、その部分だけを読んでも理解するのは困難

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページのインストール済みであるとのためにまとめた書籍も

What's new?

- Silhouetteスコアの
- Silhouetteスコアの
- 「平成29年度NGS

- 門田からメール返
- はじめに (last mo
- 参考資料 | 書籍
- 参考資料 | 講習

- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2017/11/13)
- 書籍 | 日本乳酸菌学会誌 | [第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | 日本乳酸菌学会誌 | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2016/11/2)
- 書籍 | 日本乳酸菌学会誌 | [第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | [第4回クオリティコントロールとプログラムのインストール](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | [第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | [第7回ロングリードアセンブリ](#) (last modified 2017/06/28)

書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール

日本乳酸菌学会誌の第4回分です。Linuxコマンドのリンク先は主に日経BP社様です。ウェブ資料は、共有フォルダ関連の記事(W4-5から4-8周辺)を(8/14, 8/23に引き続いて)2015年9月18日にアップデートしました。これでもまだ設定がリセットされるという方はお知らせ願います。

- [原稿PDF](#)
- ウェブ資料PDF(オリジナル版; 原稿PDFと同じく、HDD150GB、約1.3億の全リードファイルからスタート。)
 - [Windows用](#)(2015.09.18版; 約24MB)
 - [Macintosh用](#)(未着手)
- ウェブ資料PDF(軽量版; 原稿PDFと異なり、HDD100GB、最初の150万リードのみのファイルからスタート。)
 - [Windows用](#)(2015.12.11版; 約24MB; 推奨)
 - [Macintosh用](#)(未着手)

Linuxコマンド

- [apt-get](#) (パッケージのインストールやアップデート)
- [cd](#) (ディレクトリを変更)
- [chmod](#) (ファイルやディレクトリの権限変更)

FaQCsの利用

FaQCsの利用例は、①NGS連載第5回にあります。②原稿PDF中のp194、③ウェブ資料PDF中のW1-1にもあります

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページの目次は、ツール済みであるとのためにまとめた書籍も

What's new?

- Silhouetteスコアの
- Silhouetteスコアの
- 「平成29年度NGS

- 門田からメール返
- はじめに (last mo
- 参考資料 | 書籍
- 参考資料 | 講習

- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2017/11/13)
- 書籍 | 日本乳酸菌学会誌 | [第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | 日本乳酸菌学会誌 | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/2)
- 書籍 | 日本乳酸菌学会誌 | [第3回Linux環境構築からNGSデータ取得まで](#) (last modified 201
- 書籍 | 日本乳酸菌学会誌 | [第4回クオリティコントロールとプログラムのインストール](#) (last mod
- 書籍 | 日本乳酸菌学会誌 | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/2
- 書籍 | 日本乳酸菌学会誌 | [第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | [第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last m
- 書籍 | 日本乳酸菌学会誌 | [第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/
- 書籍 | 日本乳酸菌学会誌 | [第11回統合デ](#)
- 書籍 | 日本乳酸菌学会誌 | [第12回Galaxy](#)
- イントロ | 一般 | [ランダムに行を抽出](#) (last m
- イントロ | 一般 | [任意の文字列を行の最初](#)
- イントロ | 一般 | [任意のキーワードを含む](#)

書籍 | 日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC

日本乳酸菌学会誌 | 第5回分です。Linuxコマンドのリンク先は主に日経BP社様です。

- [第5回分PDF](#)
- [ウェブ資料PDF](#)(2016-06.09版; 約14MB)

Linuxコマンド

- [bzip2](#) (bzip2圧縮、および解凍)
- [cd](#) (ディレクトリを変更)
- [echo](#) (文字列を表示)
- [export](#) (変数を追加)
- [file](#) (ファイルタイプを判定)

FaQCs実行

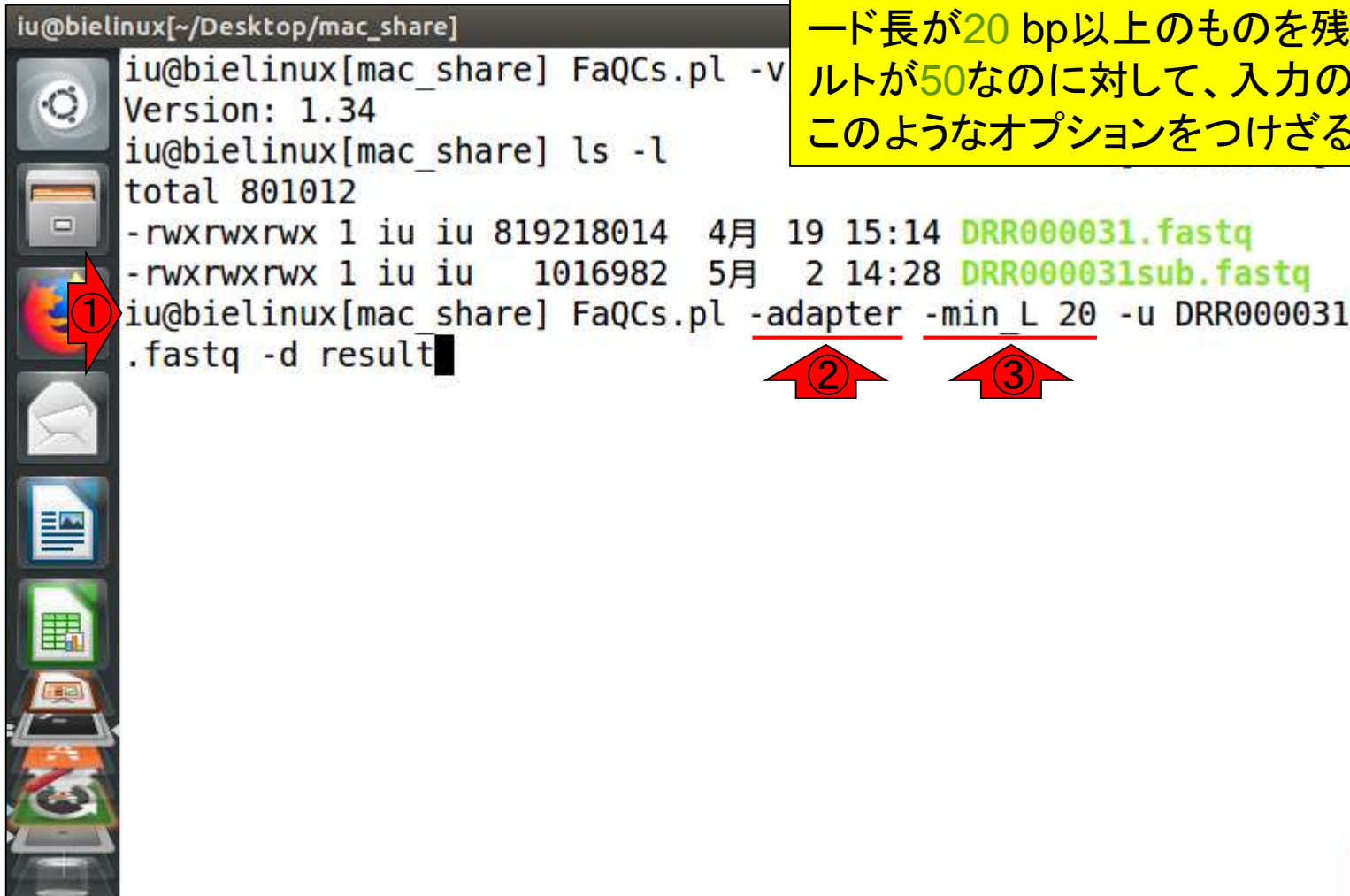
①FaQCs.pl -vは、バージョンの確認。②lsで入力ファイルの確認。入力はgzip圧縮ファイルでもOKだが、ここでは③のFASTQファイルを入力として与えます

```
iu@bielinux[~/Desktop/mac_share]
① iu@bielinux[mac_share] FaQCs.pl -v [ 5:20午後 ]
Version: 1.34
② iu@bielinux[mac_share] ls -l [ 5:20午後 ]
total 801012
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] FaQCs.pl -adapter -min_L 20 -u DRR000031
.fastq -d result
```

FaQCs実行

ここでは、NGS連載第5回W1-1とは一部違って、①のような実行コマンドとした。②の部分は不変で、Illuminaのアダプターを除去するオプション。③は様々な処理後のリード長が20 bp以上のものを残すというオプション。デフォルトが50なのに対して、入力のリード長が36 bpなので、このようなオプションをつけざるを得なかったのです

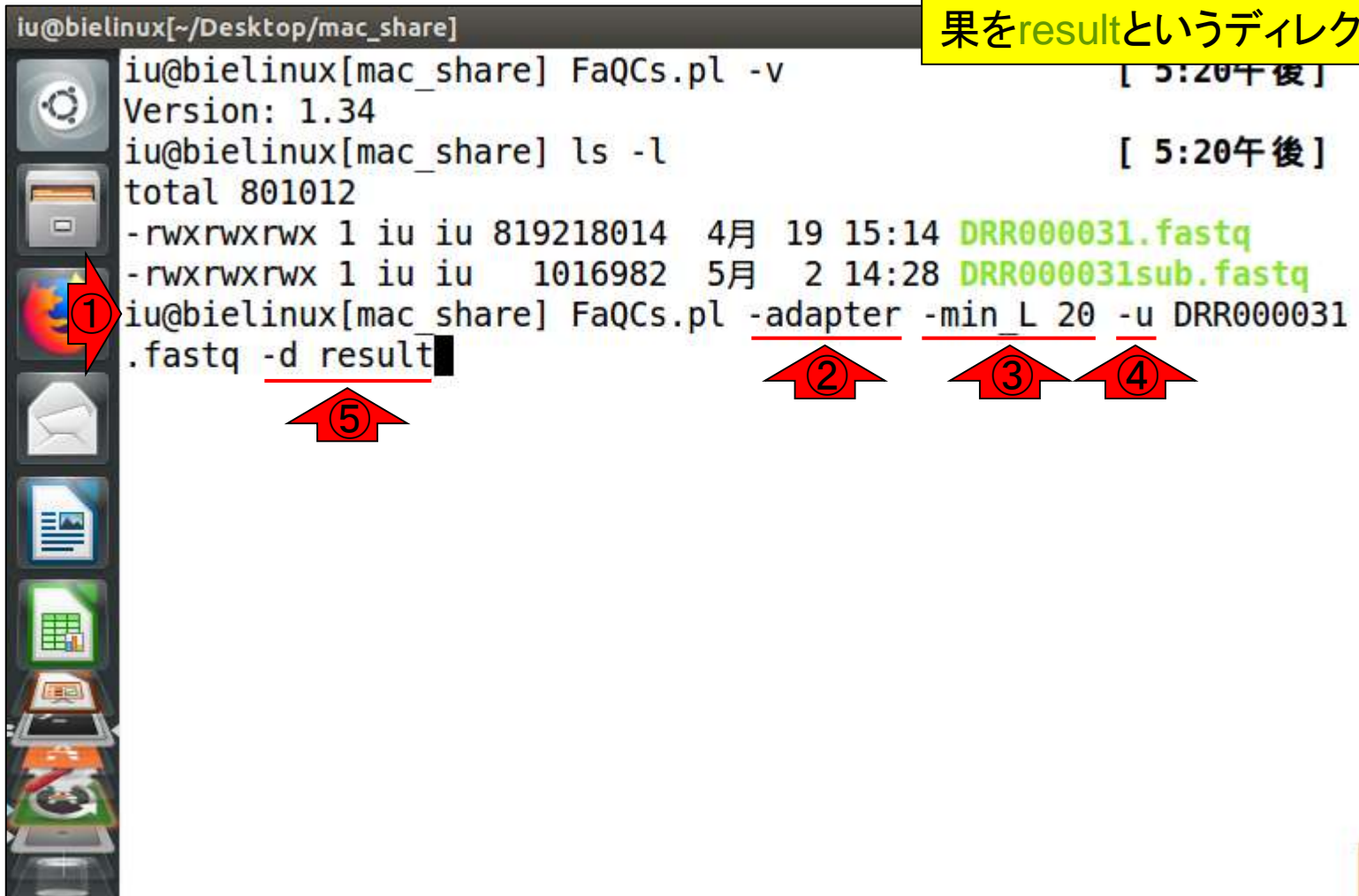
```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] FaQCs.pl -v
Version: 1.34
iu@bielinux[mac_share] ls -l
total 801012
-rwxrwxrwx 1 iu iu 819218014  4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu  1016982  5月  2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] FaQCs.pl -adapter -min_L 20 -u DRR000031
.fastq -d result
```



FaQCs実行

④は、今回の入力がpaired-endではないからです。Unpaired read(single-end)の場合は-uオプションを付けなければなりません。⑤は、実行結果をresultというディレクトリに格納せよ、です

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] FaQCs.pl -v
Version: 1.34
iu@bielinux[mac_share] ls -l
total 801012
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] FaQCs.pl -adapter -min_L 20 -u DRR000031
.fastq -d result
```



FaQCs実行中

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] FaQCs.pl -v [ 5:20午後 ]
Version: 1.34
iu@bielinux[mac_share] ls -l [ 5:20午後 ]
total 801012
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] FaQCs.pl -adapter -min_L 20 -u DRR000031
.fastq -d result
Bwa extension trimming algorithm is used.
Processing DRR000031.fastq file
```


FaQCs実行完了

このときは、(リターンキーを押したのが17:20なので)①約12分かかりました

```
iu@bielinux[~/Desktop/mac_share] 17:35
Post Trimming Length(Mean, Std, Median, Max, Min) of 955054 re
ads with Overall quality 37.59
(35.84, 1.14, 36.0, 36, 20)
Processed 2000000/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 928579 re
ads with Overall quality 36.93
(35.83, 1.12, 36.0, 36, 20)
Processed 3000000/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 957359 re
ads with Overall quality 37.22
(35.87, 0.99, 36.0, 36, 20)
Processed 4000000/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 954018 re
ads with Overall quality 36.70
(35.84, 1.04, 36.0, 36, 20)
Processed 4589774/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 568646 re
ads with Overall quality 37.25
(35.82, 1.18, 36.0, 36, 20)
iu@bielinux[mac_share] [ 5:32午後 ]
```



[5:32午後]

①ls -l。確かに②resultディレクトリが作成されています

ls -lで確認

```
iu@bielinux[~/Desktop/mac_share]
ads with Overall quality 36.93
(35.83, 1.12, 36.0, 36, 20)
Processed 3000000/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 957359 re
ads with Overall quality 37.22
(35.87, 0.99, 36.0, 36, 20)
Processed 4000000/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 954018 re
ads with Overall quality 36.70
(35.84, 1.04, 36.0, 36, 20)
Processed 4589774/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 568646 re
ads with Overall quality 37.25
(35.82, 1.18, 36.0, 36, 20)
iu@bielinux[mac_share] ls -l [ 6:11午後 ]
total 801016
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] [ 6:14午後 ]
```



ls -lで確認

①resultディレクトリ内をls -l。single-endの場合の出力ファイルは、②のQC.unpaired.trimmed.fastqです。このファイルサイズ(777,880,074 bytes)と、③入力ファイルのサイズ(819,218,014 bytes)の関係から、 $819,218,014 / 777,880,074 = 1.053142$ となり、約5%程度の塩基が除かれたのだらうと解釈する

```
iu@bielinux[~/Desktop/mac_share]
(35.84, 1.04, 36.0, 36, 20)
Processed 4589774/4589774
Post Trimming Length(Mean, Std, Med
ads with Overall quality 37.25
(35.82, 1.18, 36.0, 36, 20)
iu@bielinux[mac_share] ls -l
total 801016
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] ls -l result
total 863174
-rwxrwxrwx 1 iu iu 105922304 5月 10 17:32 DRR000031_00004.fast
q
-rwxrwxrwx 1 iu iu 40 5月 10 17:22 fastqCount.txt
-rwxrwxrwx 1 iu iu 85066 5月 10 17:32 QC_qc_report.pdf
-rwxrwxrwx 1 iu iu 790 5月 10 17:32 QC.stats.txt
-rwxrwxrwx 1 iu iu 777880074 5月 10 17:32 QC.unpaired.trimmed.
fastq
iu@bielinux[mac_share]
```

[6:11午後]



[6:14午後]



[7:06午後]



結果の詳細は...

①これらのファイルに書かれています。pdf形式かtxt形式かの違いだけで、中身は同じ。QC.stats.txtをmoreコマンドで眺めたのが次のスライド

```
iu@bielinux[~/Desktop/mac_share]
(35.84, 1.04, 36.0, 36, 20)
Processed 4589774/4589774
Post Trimming Length(Mean, Std, Median, Max, Min) of 568646 re
ads with Overall quality 37.25
(35.82, 1.18, 36.0, 36, 20)
iu@bielinux[mac_share] ls -l                                [ 6:11午後 ]
total 801016
-rwxrwxrwx 1 iu iu 819218014  4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu  1016982  5月  2 14:28 DRR000031sub.fastq
drwxrwxrwx 1 iu iu    4096  5月 10 17:32 result
iu@bielinux[mac_share] ls -l result                        [ 6:14午後 ]
total 863174
-rwxrwxrwx 1 iu iu 105922304  5月 10 17:32 DRR000031_00004.fast
q
-rwxrwxrwx 1 iu iu    40  5月 10 17:22 fastqCount.txt
-rwxrwxrwx 1 iu iu  85066  5月 10 17:32 QC_qc_report.pdf
-rwxrwxrwx 1 iu iu    790  5月 10 17:32 QC.stats.txt
-rwxrwxrwx 1 iu iu 777880074  5月 10 17:32 QC.unpaired.trimmed.
fastq
iu@bielinux[mac_share]                                    [ 7:06午後 ]
```



QC.stats.txt

```
iu@bielinux[~/Desktop/mac_share] more result/QC.stats.txt [ 7:06午後 ]
Before Trimming
Reads #: 4589774
Total bases: 165231864
Reads Length: 36.00

After Trimming
Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84

Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)
  Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
  Bases Filtered by length cutoff: 626136 (0.38 %)
  Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
  Bases Filtered by continuous base "N": 3686831 (2.23 %)
  Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
  Bases Filtered by low complexity ratio: 1903712 (1.15 %)
  Reads Trimmed by quality (5.0): 327035 (7.13 %)
```

QC.stats.txt

①トリム前は②こんな状態で、③トリム後は④こんな状態。
例えば、トリム後のリード数(4,363,656個)は、トリム前のリード数(4,589,774個)の95.07%になっているなどと読み解く

```
iu@bielinux[~/Desktop/mac_share] more result/QC.stats.txt [ 7:06午後 ]
① Before Trimming
Reads #: 4589774
Total bases: 165231864
Reads Length: 36.00
②
③ After Trimming
Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84
④
Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)
  Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
  Bases Filtered by length cutoff: 626136 (0.38 %)
  Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
  Bases Filtered by continuous base "N": 3686831 (2.23 %)
  Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
  Bases Filtered by low complexity ratio: 1903712 (1.15 %)
  Reads Trimmed by quality (5.0): 327035 (7.13 %)
```


QC.stats.txt

```
iu@bielinux[~/Desktop/mac_share]
Reads Length: 36.00
After Trimming
Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84
Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)
  Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
  Bases Filtered by length cutoff: 626136 (0.38 %)
  Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
  Bases Filtered by continuous base "N": 3686831 (2.23 %)
  Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
  Bases Filtered by low complexity ratio: 1903712 (1.15 %)
  Reads Trimmed by quality (5.0): 327035 (7.13 %)
  Bases Trimmed by quality: 2616713 (1.58 %)
  Reads Trimmed with Adapters/Primers: 0 (0.00 %)
  Bases Trimmed with Adapters/Primers: 0 (0.00 %)
iu@bielinux[mac_share] [ 7:10午後 ]
```

QC.stats.txt

①リードごと除去されたのは226,118個で、全リードの4.93%。
②トリムされた塩基数は8,833,392個で、全体の5.35%などということがわかります。③は、①と②の内訳。左側にReadsとBasesが交互に現われている。Reads部分の記述が①の内訳。Bases部分の記述が②の内訳に相当するのであろう

iu@bielinux[~/Desktop/mac_share]

Reads Length: 36.00

After Trimming

Reads #: 4363656 (95.07 %)

Total bases: 156398472 (94.65 %)

Mean Reads Length: 35.84

Discarded reads #: 226118 (4.93 %) ①

Trimmed bases: 8833392 (5.35 %) ②

{ Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
{ Bases Filtered by length cutoff: 626136 (0.38 %)
{ Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
{ Bases Filtered by continuous base "N": 3686831 (2.23 %)
{ Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
{ Bases Filtered by low complexity ratio: 1903712 (1.15 %)
{ Reads Trimmed by quality (5.0): 327035 (7.13 %)
{ Bases Trimmed by quality: 2616713 (1.58 %)
{ Reads Trimmed with Adapters/Primers: 0 (0.00 %)
{ Bases Trimmed with Adapters/Primers: 0 (0.00 %) ④

iu@bielinux[mac_share]

[7:10午後]

QC.stats.txt

例えば、①は連続するNによってフィルタリングされた塩基数が3686831個で、全体の2.23%。②の記述より、連続するNの数はおそらく③2個で、2個以上連続するNを含むのは104,727リードであり、全体の2.28%ということなのだろう

iu@bielinux[~/Desktop/mac_share]
Reads Length: 36.00

After Trimming

Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84

Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)

Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
Bases Filtered by length cutoff: 626138 (0.38 %)
Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
Bases Filtered by continuous base "N": 3686831 (2.23 %)
Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
Bases Filtered by low complexity ratio: 1903712 (1.15 %)
Reads Trimmed by quality (5.0): 327035 (7.13 %)
Bases Trimmed by quality: 2616713 (1.58 %)
Reads Trimmed with Adapters/Primers: 0 (0.00 %)
Bases Trimmed with Adapters/Primers: 0 (0.00 %)

iu@bielinux[mac_share]

[7:10午後]

QC.stats.txt

また、①の最終行では、Adapters/Primers由来のBasesは一つもなかったことを示している。この結果から、FaQCs実行時に`-adapter`はつけなくても同じ結果になったのだろう、などということがわかる

```
iu@bielinux[~/Desktop/mac_share]
Reads Length: 36.00
```

```
After Trimming
```

```
Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84
```

```
Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)
```

```
Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
Bases Filtered by length cutoff: 626136 (0.38 %)
Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
Bases Filtered by continuous base "N": 3686831 (2.23 %)
Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
Bases Filtered by low complexity ratio: 1903712 (1.15 %)
Reads Trimmed by quality (5.0): 327035 (7.13 %)
Bases Trimmed by quality: 2616713 (1.58 %)
Reads Trimmed with Adapters/Primers: 0 (0.00 %)
Bases Trimmed with Adapters/Primers: 0 (0.00 %)
```

```
iu@bielinux[mac_share] █
```

```
[ 7:10午後 ]
```

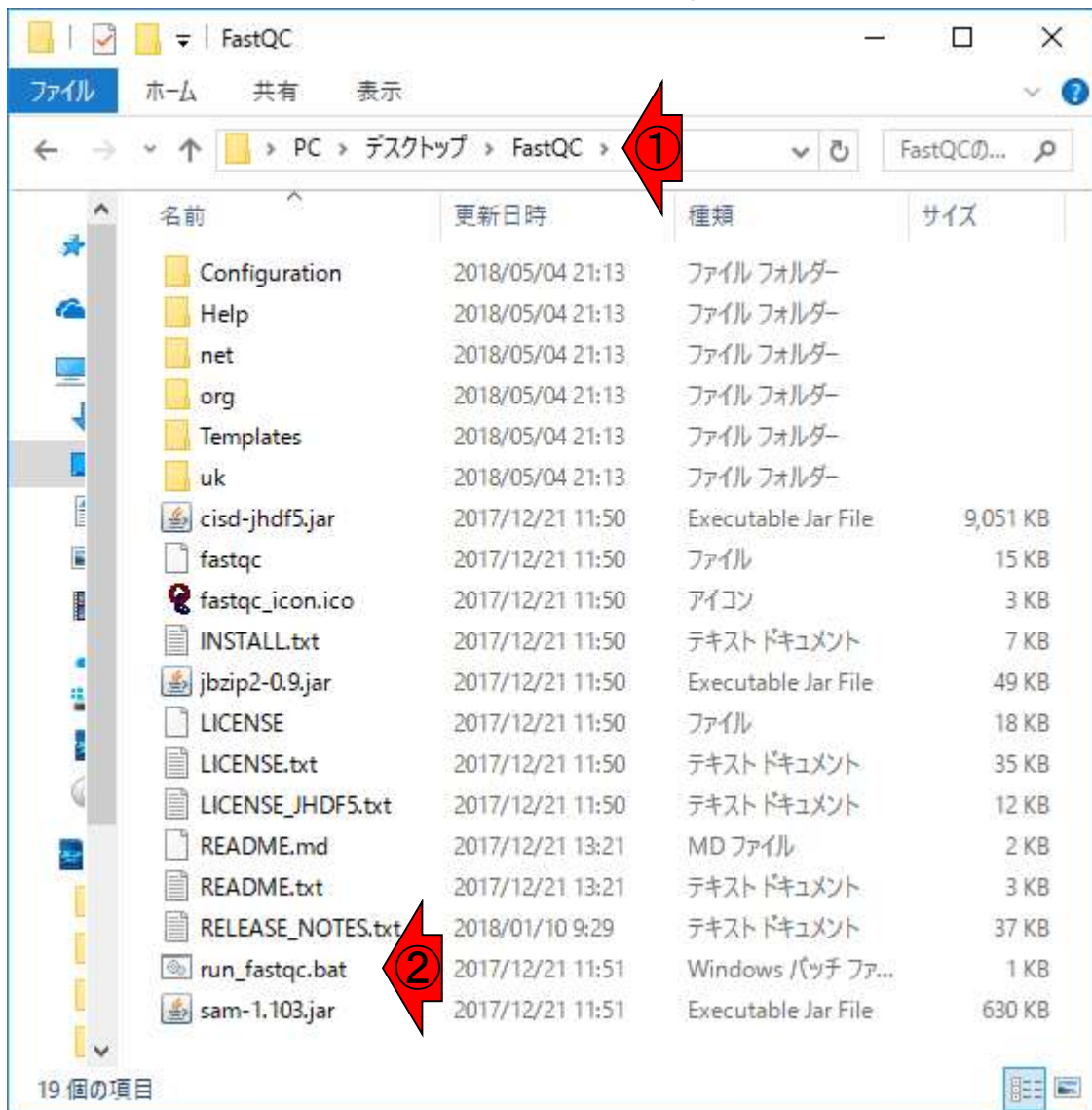


Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題(FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング(アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング(内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

FastQCを起動

①デスクトップ上にあるFastQCフォルダ内の
、②run_fastqc.batをダブルクリックして起動



FastQCの入力ファイル

```
iu@bielinux[~/Desktop/mac_share]
Reads #: 4363656 (95.07 %)
Total bases: 156398472 (94.65 %)
Mean Reads Length: 35.84

Discarded reads #: 226118 (4.93 %)
Trimmed bases: 8833392 (5.35 %)
  Reads Filtered by length cutoff (20 bp): 68065 (1.48 %)
  Bases Filtered by length cutoff: 626136 (0.38 %)
  Reads Filtered by continuous base "N" (2): 104727 (2.28 %)
  Bases Filtered by continuous base "N": 3686831 (2.23 %)
  Reads Filtered by low complexity ratio (0.8): 53326 (1.16 %)
  Bases Filtered by low complexity ratio: 1903712 (1.15 %)
  Reads Trimmed by quality (5.0): 327035 (7.13 %)
  Bases Trimmed by quality: 2616713 (1.58 %)
  Reads Trimmed with Adapters/Primers: 0 (0.00 %)
  Bases Trimmed with Adapters/Primers: 0 (0.00 %)
iu@bielinux[mac_share] ls -l result/QC.unpaired.trimmed.fastq
-rwxrwxrwx 1 iu iu 777880074  5月 10 17:32 result/QC.unpaired.t
rimmed.fastq
iu@bielinux[mac_share] █ [10:57午後]
```



FastQC実行結果

こんな感じになります。これをhtmlファイルとして保存したものが…

The screenshot shows the FastQC application window. The title bar reads 'FastQC'. The menu bar contains 'File' and 'Help'. The main window title is 'QC.unpaired.trimmed.fastq'. On the left, a sidebar lists various analysis modules, each with a status icon (green checkmark for success, orange exclamation mark for warning). The 'Basic Statistics' module is selected, and its results are displayed in a table titled 'Basic sequence stats'.

Measure	Value
Filename	QC.unpaired.trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4363656
Sequences flagged as poor quality	0
Sequence length	20-36
%GC	48

Left sidebar items:

- Basic Statistics (Green checkmark)
- Per base sequence quality (Green checkmark)
- Per tile sequence quality (Orange exclamation mark)
- Per sequence quality scores (Green checkmark)
- Per base sequence content (Green checkmark)
- Per sequence GC content (Green checkmark)
- Per base N content (Green checkmark)
- Sequence Length Distribution (Orange exclamation mark)
- Sequence Duplication Levels (Green checkmark)
- Overrepresented sequences (Green checkmark)
- Adapter Content (Green checkmark)

QC.unpaired.trimmed_fastqc.html

講義日程 (平成30年度)

1. 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
3. 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用) 
ShortRead : Morgan et al., Bioinformatics, 2011
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)

課題

講義日程 (平成30年度)

1. 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用) 
report.html(qrqcを用いたQC結果)
3. 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用) 
ShortRead : Morgan et al., Bioinformatics, 2011
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)

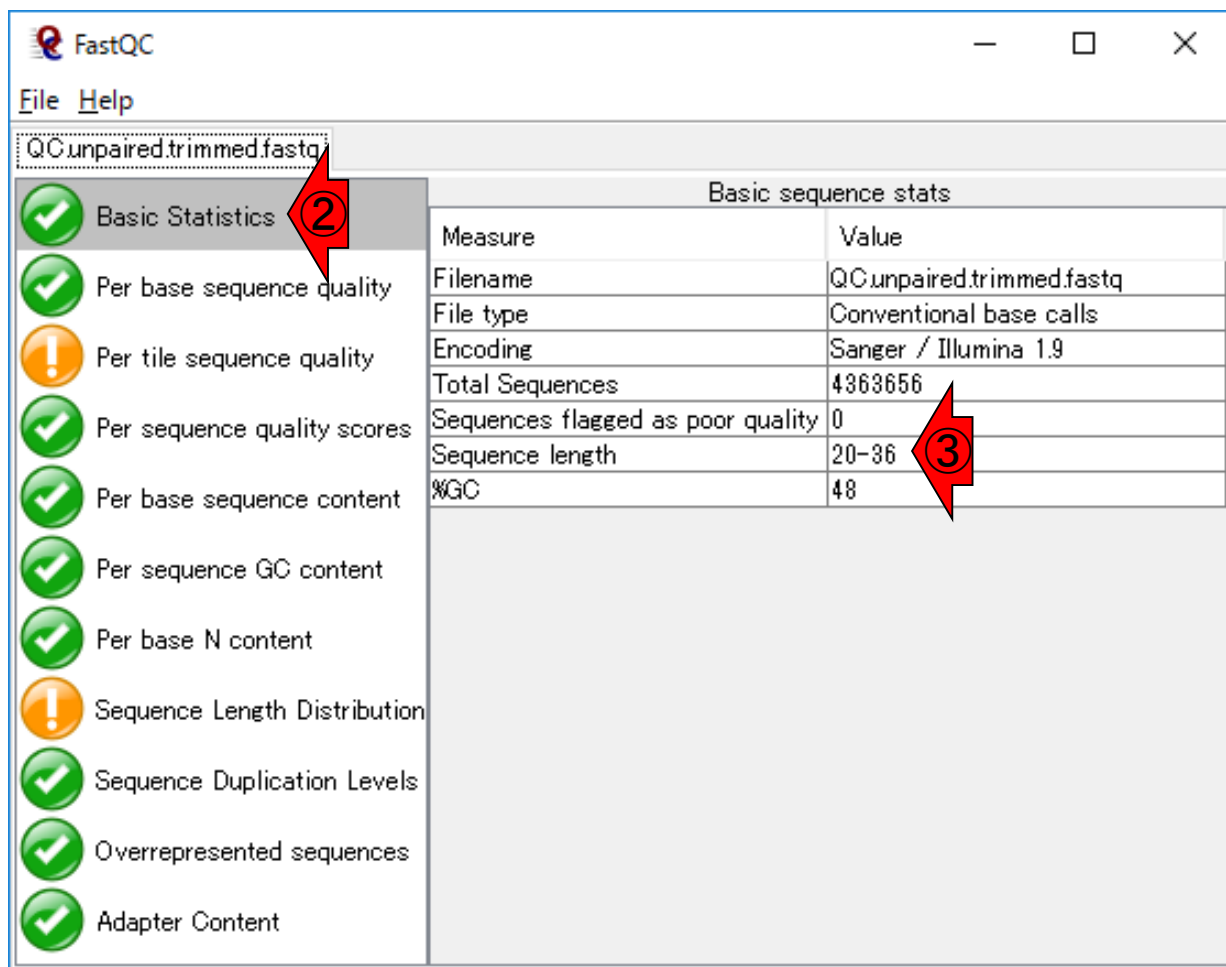
①FaQCs実行後のFastQC実行結果について、
②FaQCs実行前のFastQC実行結果と対比させて考察せよ。フィルタリングやトリミングを行うFaQCsプログラムがうまく機能しているかどうか?が主な論点となると思います

課題：論点1

例えば、①FaQCs実行後の、②Basic Statisticsの、③Sequence lengthの範囲(20-36)と、FaQCs実行時に与えたオプションとの関係も議論可能かと思います

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrhcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用) ①
ShortRead : Morgan et al., Bioinformatics, 2012
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrhcを用いたQC結果)



FastQC

File Help

QC.unpaired.trimmed.fastq

Basic sequence stats	
Measure	Value
Filename	QC.unpaired.trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4363656
Sequences flagged as poor quality	0
Sequence length	20-36 ③
%GC	48

① Basic Statistics ②

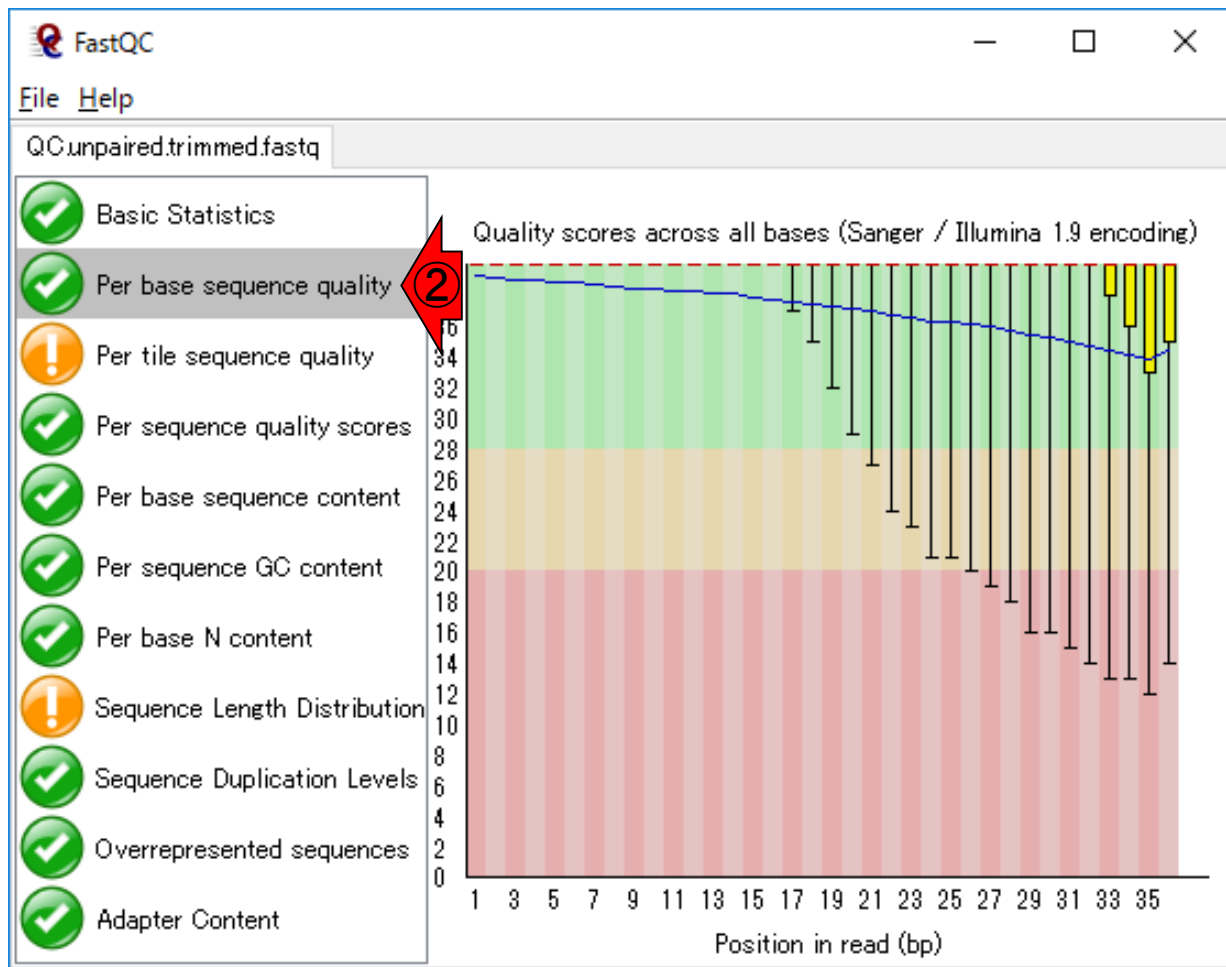
③

課題：論点2

また、①FaQCs実行後の、②Per base sequence qualityの分布と…

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用) ①
ShortRead : Morgan et al., Bioinformatics, 2004
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)

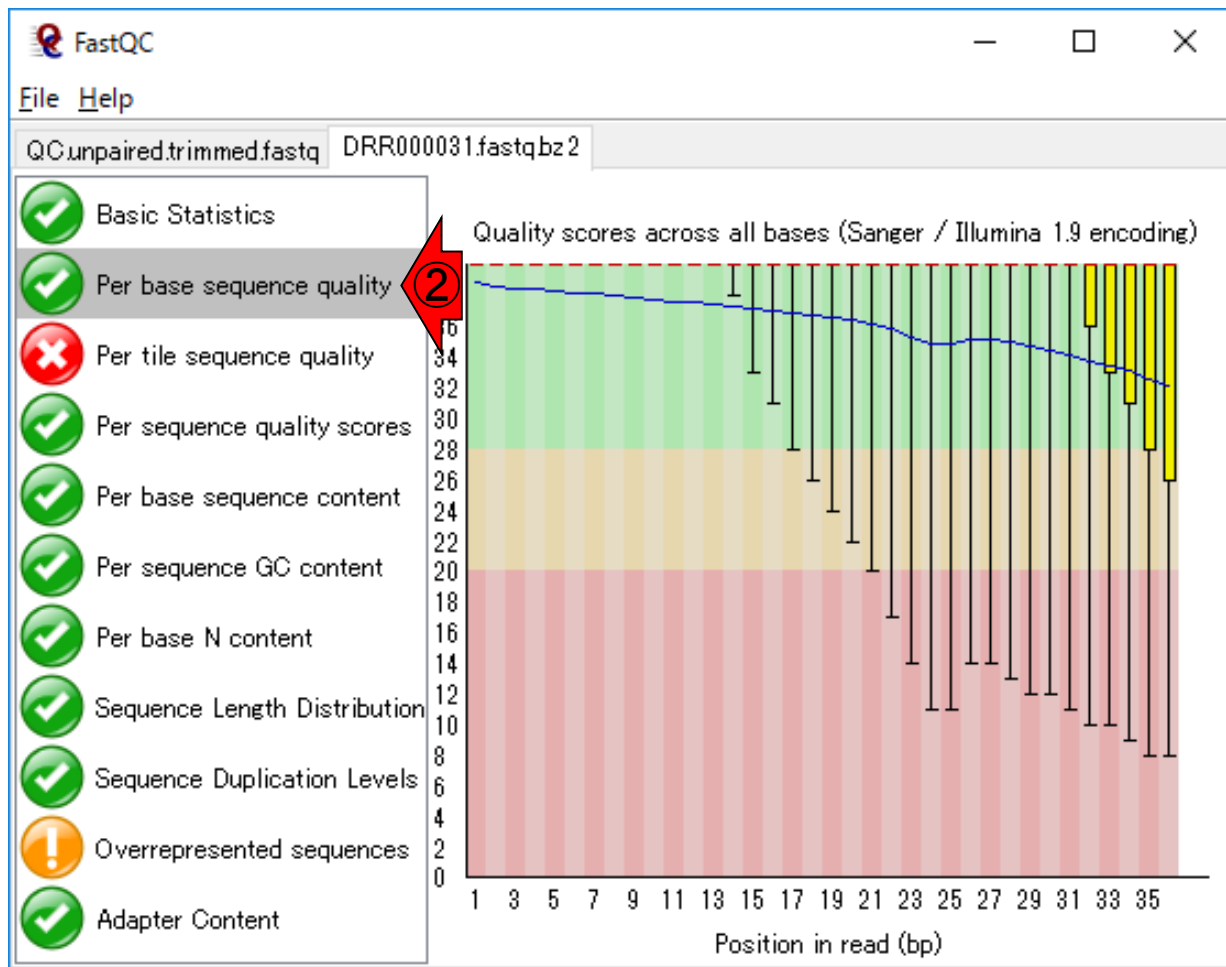


課題：論点2

また、①FaQCs実行前の、②Per base sequence qualityの分布の違いも分かりやすい評価基準

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果) **①**
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用)
ShortRead : Morgan et al., Bioinformatics, 2009
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)



課題：論点3

①FaQCs実行前の、②Overrepresented sequencesではpoly Aが見えていたが…。③FaQCs実行後はどうなっているかなど…

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用) **①**
report.html(qrqcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用) **③**
ShortRead : Morgan et al., Bioinformatics, 2011
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)

FastQC

File Help

QCunpaired.trimmed.fastq DRR000031.fastqbz 2

Overrepresented sequences			
Sequence	Count	Percentage	Possible Sour...
AAAAAAAAA...	41007	0.893	No Hit

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 ! **②**

Adapter Content ✓

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

ShortReadパッケージ

①ShortReadの原著論文のPubMedサイト。
ShortReadパッケージは、②100回以上引用
されている、③2009年の原著論文があります

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用)
ShortRead : Morgan et al., Bioinformatics, 2009
Bioconductor : Gentleman et al., Genome Biol., 2005
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2014
report.html(qrqcを用いた)

Bioinformatics. 2009 Oct 16; 25(19):2607-8. doi: 10.1093/bioinformatics/btp450. Epub 2009 Aug 3.

ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data.

Morgan M¹, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R.

Author information

Abstract

ShortRead is a package for input, quality assessment, manipulation and output of high-throughput sequencing data. ShortRead is provided in the R and Bioconductor environments, allowing ready access to additional facilities for advanced statistical analysis, data transformation, visualization and integration with diverse genomic resources.

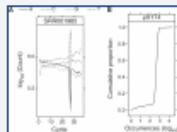
AVAILABILITY AND IMPLEMENTATION: This package is implemented in R and available at the Bioconductor web site; the package contains a 'vignette' outlining typical work flows.

PMID: 19654119 PMCID: PMC2752612 DOI: 10.1093/bioinformatics/btp450

[Indexed for MEDLINE] [Free PMC Article](#)



Images from this publication. [See all images \(1\)](#) [Free text](#)



Publication types, MeSH terms, Grant support

LinkOut - more resources

OXFORD
ACADEMIC

PMC **FREE**
Full text

Save items

☆ Add to Favorites

Similar articles

VariantAnnotation: a
Bioconductor [Bioinformatics. 2014]

QuasR: quantification and
annotation [Bioinformatics. 2015]

R453Plus1Toolbox: an
R/Biocond [Bioinformatics. 2011]

Review Computational solutions
for omics [Nat Rev Genet. 2013]

Review Orchestrating high-
throughput q [Nat Methods. 2015]

See reviews...

See all...

Cited by over 100
PubMed Central articles

Imputation from SNP chip to
seq [J Anim Sci Biotechnol. 2018]

Genetic signatures of microbial
[Proc Natl Acad Sci U S A. 2018]

Comprehensive genomic
analysis [BMC Med Genet. 2018]

See all...

ShortRead(Morgan et al., Bioinformatics, 25: 2607–8, 2009)

ShortRead

参考

①ShortReadのBioconductorのサイト。②ShortReadのダウンロード数は、③Bioconductor提供パッケージ群のうち上位5%に入る(ほどよく利用されている)。大抵のパッケージは、どのプラットフォーム(Windows, Macintosh, Linux)でも動く。実際、④ShortReadもここがallになっている。もしsomeとかになっていれば、⑤ページ下部に移動し…

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNAcocktail : Sahraeian et al., Nat Commun, 2018
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2018
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNAcocktail : Sahraeian et al., Nat Commun, 2018
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2018
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC_stats_trimmed_fastqc.html(課題用)
ShortRead : Morgan et al., Bioinformatics, 2004
Bioconductor : Gentleman et al., Genome Biology, 2004
Rsubread(Windows版なし) : Liao et al., Nat Methods, 2014
report.html(qrqcを用いたQC結果)

The screenshot shows the Bioconductor website for the ShortRead package. A navigation bar at the top includes links for Home, Install, Help, Developers, and About. The breadcrumb trail is Home » Bioconductor 3.7 » Software Packages » ShortRead. The package name 'ShortRead' is prominently displayed. Below it, a row of buttons shows 'platforms all', 'downloads top 5%', 'posts 5 / 1 / 2 / 3', and 'in Bioc 9.5 years'. A 'build warning' button is also visible. The DOI is 10.18129/B9.bioc.ShortRead. The main heading is 'FASTQ input and manipulation'. The Bioconductor version is Release (3.7). The description states: 'This package implements sampling, iteration, and input of FASTQ files. The package includes functions for filtering and trimming reads, and for generating a quality assessment report. Data are represented as DNASTringSet-derived objects, and easily manipulated for a diversity of purposes. The package also contains legacy support for early single-end, ungapped alignment formats.' The author is listed as Martin Morgan, Michael Lawrence, and Simon Anders. The maintainer is the Bioconductor Package Maintainer. The citation is: Morgan M, Anders S, Lawrence M, Aboyou P, Pagès H, Gentleman R (2009). "ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data." *Bioinformatics*, 25, 2607-2608. doi: 10.1093/bioinformatics/btp450, http://dx.doi.org/10.1093/bioinformatics/btp450. On the right side, there are sections for 'Documentation' and 'Support'. A red arrow labeled '5' points to the 'Support' section.

Bioconductor(Gentleman et al., Genome Biol., 5: R80, 2004)

①このあたりまで移動して…、②Windows or Macのどちらが使えないかを見る必要があります。あくまでも直前のスライドでplatformsがsomeになっている場合の話。someとなっている実例はRsubreadパッケージ

ShortRead

講義日程 (平成30年度)

- 平成30年05月08日
 講義資料PDF
 .gff3ファイル (約1.3MB)
 .faファイル (約2.2MB)
 (Rで)塩基配列解析
 (Rで)マイクロアレイデータ解析
 plasmid1.gff3(課題用)
 plasmid2.gff3(課題用)
 de Lannoy et al., F1000Res., 2017
 Garalde et al., Nat Methods, 2018
 RNAcocktail : Sahraeian et al., Nat Con
- 平成30年05月15日
 講義資料PDF(約5MB; 2018.05.11版)
 (Rで)塩基配列解析
 DRR000031sub.fastq
 RNA-QC-chain : Zhou et al., BMC Genom
 Biostar : Parnell et al., PLoS Comput Bi
 FastQC
 DRR000031sub_fastqc.html
 DRR000031_fastqc.html(課題用)
 report.html(qrqcを用いたQC結果)
- 平成30年05月22日
 講義資料PDF(約5MB; 2018.05.17版)
 (Rで)塩基配列解析
 RNAcocktail : Sahraeian et al., Nat Con
 Kraken : Davis et al., Methods, 2013
 Lowe et al., PLoS Comput Biol., 2017
 FaQCs : Lo and Chain, BMC Bioinformat
 FaQCs実行結果のQC.stats.txt
 FaQCs実行結果のQC_qc_report.pdf
 FastQC
 QC.unpaired.trimmed_fastqc.html(課題
 ShortRead : Morgan et al., Bioinformat
 Bioconductor : Gentleman et al., Genom
 Rsubread(Windows版なし) : Liao et al., M
 report.html(qrqcを用いた

http://bioconductor.org/packages/release/bioc/html/ShortRead.html

Bioconductor - ShortRead

Depends: [BiocGenerics](#)(>= 0.23.3), [BiocParallel](#), [Biostrings](#)(>= 2.47.6), [Rsamtools](#)(>= 1.31.2), [GenomicAlignments](#)(>= 1.15.6)

Imports: [Biobase](#), [S4Vectors](#)(>= 0.17.25), [IRanges](#)(>= 2.13.12), [GenomeInfoDb](#)(>= 1.15.2), [GenomicRanges](#)(>= 1.31.8), [hwriter](#), [methods](#), [zlibbioc](#), [lattice](#), [latticeExtra](#)

LinkingTo: [S4Vectors](#), [IRanges](#), [XVector](#), [Biostrings](#)

Suggests: [BiocStyle](#), [RUnit](#), [biomaRt](#), [GenomicFeatures](#), [yeastNagalakshmi](#)

SystemRequirements

Enhances

URL

Depends On Me: [chipseq](#), [EatonEtAlChIPseq](#), [EDASeg](#), [esATAC](#), [girafe](#), [HTSeqGenie](#), [OTUbase](#), [Rqc](#), [rSFFreader](#), [segmentSeq](#), [sequencing](#), [systemPipeR](#)

Imports Me: [amplican](#), [ArrayExpressHTS](#), [basecallQC](#), [BEAT](#), [chipseq](#), [ChIPseqR](#), [ChIPsim](#), [dada2](#), [easyRNASeq](#), [GOTHIC](#), [IONiseR](#), [MACPET](#), [nucleR](#), [QuasR](#), [R453Plus1Toolbox](#), [RSVSim](#)

Suggests Me: [BiocParallel](#), [CSAR](#), [DBChIP](#), [GenomicAlignments](#), [Genominator](#), [HiCDataLymphoblast](#), [PICS](#), [PING](#), [Repitools](#), [RnaSeqTutorial](#), [Rsamtools](#), [S4Vectors](#), [yeastRNASeq](#)

[Build Report](#)

Package Archives

Follow [Installation](#) instructions to use this package in your R session.

Source Package	ShortRead_1.38.0.tar.gz
Windows Binary	ShortRead_1.38.0.zip (32- & 64-bit)
Mac OS X 10.11 (El Capitan)	ShortRead_1.38.0.tgz
Source Repository	git clone https://git.bioconductor.org/packages/ShortRead
Source Repository (Developer Access)	git clone git@git.bioconductor.org:packages/ShortRead
Package Short Url	http://bioconductor.org/packages/ShortRead/
Package Downloads Report	Download Stats

Contact us: support.bioconductor.org
 Copyright © 2003 - 2018, Bioconductor

Bioconductor
 OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Bioconductor(Gentleman et al., Genome Biol., 5: R80, 2004)

Rsubread

①RsubreadのBioconductorのサイト。確かに② platformsがsomeになっている。③ページ下部に移動

講義日程 (平成30年度)

- 平成30年05月08日
 講義資料PDF
[.gff3ファイル](#) (約1.3MB)
[.faファイル](#) (約2.2MB)
 (Rで)塩基配列解析
 (Rで)マイクロアレイデータ解析
[plasmid1.gff3](#)(課題用)
[plasmid2.gff3](#)(課題用)
[de Lannoy et al., F1000Res., 2017](#)
[Garalde et al., Nat Methods, 2018](#)
[RNACocktail : Sahraeian et al., Nat Con](#)
- 平成30年05月15日
 講義資料PDF(約5MB; 2018.05.11版)
 (Rで)塩基配列解析
[DRR000031sub.fastq](#)
[RNA-QC-chain : Zhou et al., BMC Genom](#)
[Biostar : Parnell et al., PLoS Comput Bi](#)
[FastQC](#)
[DRR000031sub_fastqc.html](#)
[DRR000031_fastqc.html](#)(課題用)
[report.html](#)([qrqc](#)を用いたQC結果)
- 平成30年05月22日
 講義資料PDF(約5MB; 2018.05.17版)
 (Rで)塩基配列解析
[RNACocktail : Sahraeian et al., Nat Con](#)
[Kraken : Davis et al., Methods, 2013](#)
[Lowe et al., PLoS Comput Biol., 2017](#)
[FaQCs : Lo and Chain, BMC Bioinformat](#)
[FaQCs実行結果のQC.stats.txt](#)
[FaQCs実行結果のQC_qc_report.pdf](#)
[FastQC](#)
[QC.unpaired.trimmed_fastqc.html](#)(課題
[Shad : Morgan et al., Bioinformat](#)
[Bioconductor : Gentleman et al., Genom](#)
[Rsubread\(Windows版なし\) : Liao et al., N](#)
[report.html](#)([qrqc](#)を用いたQC結果)

http://bioconductor.org/packages/release/bioc/html/Rsubread.html

Bioconductor - Rsubread

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Home » Bioconductor 3.7 » Software Packages » Rsubread

Rsubread

platforms some downloads top 5% posts 8 / 1 / 2 / 3 in Bioc 7 years
 build ok

DOI: [10.18129/B9.bioc.Rsubread](https://doi.org/10.18129/B9.bioc.Rsubread)

Subread sequence alignment for R

Bioconductor version: Release (3.7)

Provides powerful and easy-to-use tools for analyzing next-gen sequencing read data. Includes quality assessment of sequence reads, read alignment, read summarization, exon-exon junction detection, fusion detection, detection of short and long indels, absolute expression calling and SNP calling. Can be used with reads generated from any of the major sequencing platforms including Illumina GA/HiSeq/MiSeq, Roche GS-FLX, ABI SOLiD and LifeTech Ion PGM/Proton sequencers.

Author: Wei Shi and Yang Liao with contributions from Gordon Smyth, Jenny Dai and Timothy Triche, Jr.

Maintainer: Wei Shi <shi at wehi.edu.au>, Yang Liao <liao at wehi.edu.au> and Gordon K Smyth <smyth at wehi.EDU.AU>

Citation (from within R, enter `citation("Rsubread")`):

Liao Y, Smyth GK, Shi W (2013). "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote." *Nucleic Acids Research*, **41**, e108.

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

Subread(Liao et al., Nucleic Acids Res., 41: e108, 2013)

Rsubread

①このあたりまで移動。②RsubreadパッケージはWindowsでは使えない。こういう場合は、一見パッケージのインストールがうまくいったように見えても、実際にはインストールできてませんので注意してください!

講義日程 (平成30年度)

- 平成30年05月08日
 講義資料PDF
 .gff3ファイル (約1.3MB)
 .faファイル (約2.2MB)
 (Rで)塩基配列解析
 (Rで)マイクロアレイデータ解析
 plasmid1.gff3(課題用)
 plasmid2.gff3(課題用)
 de Lannoy et al., F1000Res., 2017
 Garalde et al., Nat Methods, 2018
 RNACocktail : Sahraeian et al., Nat Con
- 平成30年05月15日
 講義資料PDF(約5MB; 2018.05.11版)
 (Rで)塩基配列解析
 DRR000031sub.fastq
 RNA-QC-chain : Zhou et al., BMC Genom
 Biostar : Parnell et al., PLoS Comput Bi
 FastQC
 DRR000031sub_fastqc.html
 DRR000031_fastqc.html(課題用)
 report.html(qrqcを用いたQC結果)
- 平成30年05月22日
 講義資料PDF(約5MB; 2018.05.17版)
 (Rで)塩基配列解析
 RNACocktail : Sahraeian et al., Nat Con
 Kraken : Davis et al., Methods, 2013
 Lowe et al., PLoS Comput Biol., 2017
 FaQCs : Lo and Chain, BMC Bioinformat
 FaQCs実行結果のQC.stats.txt
 FaQCs実行結果のQC_qc_report.pdf
 FastQC
 QC.unpaired.trimmed_fastqc.html(課題
 ShortRead : Morgan et al., Bioinformat
 Bioconductor : Gentleman et al., Genom
 Rsubread(Windows版なし) : Liao et al., M
 report.html(qrqcを用いたQC結果)

http://bioconductor.org/packages/release/bioc/html/Rsubread.html

Bioconductor - Rsubread

biocViews [Alignment](#), [ChIP-seq](#), [Gene Expression](#), [Gene Regulation](#), [Genetic Variability](#), [Genetics](#), [Genome Annotation](#), [Preprocessing](#), [Quality Control](#), [RNA-seq](#), [SNP](#), [Sequence Matching](#), [Sequencing](#), [Software](#)

Version	1.30.0
In Bioconductor since	Bioc 2.8 (R-2.13) (7 years)
License	GPL-3
Depends	
Imports	
LinkingTo	
Suggests	
System Requirements	
Enhances	
URL	http://bioconductor.org/packages/release/bioc/html/Rsubread.html
Depends On Me	chipseqDB , samExploreR
Imports Me	dupRadar
Suggests Me	singleCellTK

[Build Report](#)

Package Archives

Follow [Installation](#) instructions to use this package in your R session.

Source Package	Rsubread_1.30.0.tar.gz
Windows Binary	
Mac OS X 10.11 (El Capitan)	Rsubread_1.30.0.tgz
Source Repository	git clone https://git.bioconductor.org/packages/Rsubread
Source Repository (Developer Access)	git clone git@git.bioconductor.org:packages/Rsubread
Package Short Url	http://bioconductor.org/packages/Rsubread/
Package Downloads Report	Download Stats

Contact us: support.bioconductor.org
 Copyright © 2003 - 2018, Bioconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Subread(Liao et al., Nucleic Acids Res., 41: e108, 2013)

ShortRead

ShortReadパッケージは、赤枠内の多くの項目で利用しています。①クオリティスコアでのフィルタリングをやってみましょう

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページのR関連部分は、[インストール](#)についてツール済みであるという前提で記述しています。初心者的にまとめた[書籍](#)もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱論文([Zhao et al.](#))
- Silhouetteスコアの新たな使い道提唱論文([Zhao et al.](#))
- 「平成29年度NGSハンズオン講習会」の[動画](#)が公開さ

- [門田からメール返信をもらえない場合は](#) (last modified 2015/03/31)
- [はじめに](#) (last modified 2015/03/31)
- 参考資料 | [書籍](#)、[学会誌](#) (last modified 2017/11/13)
- 参考資料 | [講習会](#)、[講義](#)、[講演資料](#) (last modified 2018/05/01)

- イントロ | ファイル形式の変換 | [qseq --> Sanger FASTQ](#) (last modified 2013/08/19)
- 前処理 | [クオリティコントロール](#) | [について](#) (last modified 2018/05/01) **NEW**
- 前処理 | クオリティチェック | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/15)
- 前処理 | クオリティチェック | [qrc](#) (last modified 2014/07/17)
- 前処理 | クオリティチェック | [PHREDスコアに変換](#) (last modified 2018/05/06) **NEW**
- 前処理 | クオリティチェック | [配列長分布を調べる](#) (last modified 2015/06/22)
- 前処理 | クオリティチェック | Overrepresented sequences | [ShortRead\(Morgan 2009\)](#) (last modified 2015/07/29)
- 前処理 | トリミング | ポリA配列除去 | [ShortRead\(Morgan 2009\)](#) (last modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/26)推奨
- 前処理 | トリミング | アダプター配列除去(基礎) | [girafe\(Toedling 2010\)](#) (last modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(基礎) | [ShortRead\(Morgan 2009\)](#) (last modified 2014/06/21)
- 前処理 | トリミング | アダプター配列除去(応用) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/26)推奨
- 前処理 | トリミング | アダプター配列除去(応用) | [ShortRead\(Morgan 2009\)](#) (last modified 2015/09/12)
- 前処理 | トリミング | [指定した末端塩基数だけ除去](#) (last modified 2017/11/08)
- 前処理 | フィルタリング | [PHREDスコアが低い塩基をNに置換](#) (last modified 2014/03/03)
- 前処理 | フィルタリング | [PHREDスコアが低い配列\(リード\)を除去](#) **①** (last modified 2014/08/27)
- 前処理 | フィルタリング | [ACGTのみからなる配列を抽出](#) (last modified 2015/09/12)
- 前処理 | フィルタリング | [ACGT以外のcharacter "-"をNに変換](#) (last modified 2013/06/18)
- 前処理 | フィルタリング | [ACGT以外の文字数が閾値以下の配列を抽出](#) (last modified 2015/09/12)
- 前処理 | フィルタリング | [重複のない配列セットを作成](#) (last modified 2013/06/18)
- 前処理 | フィルタリング | [指定した長さ以上の配列を抽出](#) (last modified 2016/02/08)
- 前処理 | フィルタリング | [任意のリード\(サブセット\)を抽出](#) (last modified 2014/08/21)
- 前処理 | フィルタリング | [指定した長さの範囲の配列を抽出](#) (last modified 2015/02/26)
- 前処理 | フィルタリング | [任意のIDを含む配列を抽出](#) (last modified 2013/06/18)
- 前処理 | フィルタリング | [Illuminaのpass filtering](#) (last modified 2013/06/19)
- 前処理 | フィルタリング | [GFF/GTF形式ファイル](#) (last modified 2013/10/10)
- 前処理 | フィルタリング | 組合せ | [ACGTのみ & 指定した長さの範囲の配列](#) (last modified 2015/09/12)
- 前処理 | フィルタリング | paired-end | 配列長とN数 | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/26)
- 前処理 | フィルタリング | paired-end | 共通リード抽出 | [ShortRead\(Morgan 2009\)](#) (last modified 2015/06/26)
- [アセンブル](#) | [について](#) (last modified 2014/06/20)

ShortRead(Morgan et al., Bioinformatics, 25: 2607–8, 2009)

Rでフィルタリング

①ここでShortReadパッケージをロードしている。②クオリティスコアの閾値が20未満のものが、③(リード長×0.1)個以上あるリードを除去する場合のやり方です

前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去

Sanger FASTQ形式ファイルを読み込んで、PHREDスコアが低いリードを除去するやり方を紹介します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです (Bullard et al. 2010)。PHREDスコアが20未満のものがリード長に占める割合が0.1以上のリードを除去するやり方です。(例題のファイル中のリードは全て35bpのリードである。その10%以上ということで実質的にPHREDスコアが閾値未満のものが4塩基以上あるリードはダメということ) writeFastq関数のデフォルトオプションはcompress=Tで、gzip圧縮ファイルを出力します。ここではcompress=Fとして非圧縮ファイルを出力しています。

```
in_f <- "SRR037439.fastq" #入力ファイル名を指定してin_fに格納
out_f <- "hoge.fastq" #出力ファイル名を指定してout_fに格納
param1 <- 20 #PHREDスコアの閾値を指定
param2 <- 0.1 #指定した閾値未満のものが配列長に占める割合を指定

#必要なパッケージをロード
library(ShortRead) #パッケージの読み込み

#入力ファイルの読み込み
fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
sread(fastq) #配列情報を表示

#本番
hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality scoreをPHRED scoreに変換し、
obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件を満たすかどうかを判定した結果を
fastq <- fastq[obj] #objがTRUEとなる要素のみ抽出した結果をfastqに格納
sread(fastq) #配列情報を表示

#ファイルに保存
writeFastq(fastq, out_f, compress=F) #fastqの中身を指定したファイル名で保存
```

Rでフィルタリング

①6,000リードからなるDRR000031sub.fastqをデスクトップに保存したのち、②の部分を変更してコピペ実行してみましょう

前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去

Sanger FASTQ形式ファイルを読み込んで、PHREDスコアが低いリードを除去するやり方を紹介します。「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard 2010)。PHREDスコアが20未満のものがリード長に占める割合が0.1以上のリードを除去するやり方です。ファイル中のリードは全て35bpのリードである。その10%以上ということを実質的にPHREDスコアが閾値が4塩基以上あるリードはダメということ) writeFastq関数のデフォルトオプションはcompress=Tで、gzip圧縮出力します。ここではcompress=Fとして非圧縮ファイルを出力しています。

```
in_f <- "DRR000031sub.fastq" ②
out_f <- "hoge1.fastq"
param1 <- 20
param2 <- 0.1

#必要なパッケージをロード
library(ShortRead)

#パッケージの読み込み

#入力ファイルの読み込み
fastq <- readFastq(in_f)
sread(fastq)

#本番
hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality scoreをPHRED score
obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件を満たすかどうかを判定
fastq <- fastq[obj]
sread(fastq)

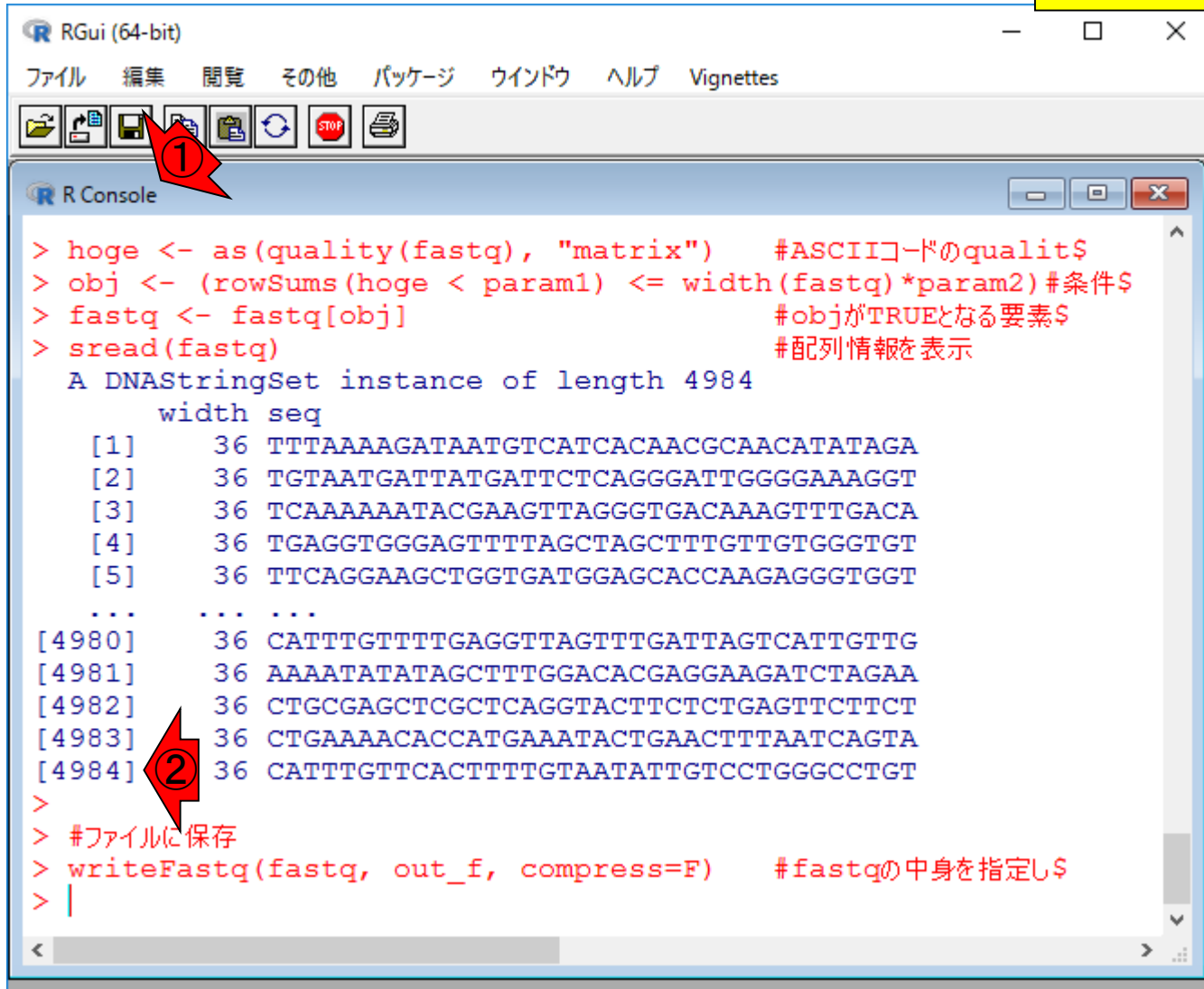
#ファイルに保存
writeFastq(fastq, out_f, compress=F) #fastqの中身を指定したファイル名で保存
```

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
de Lannoy et al., F1000Res., 2017
Garalde et al., Nat Methods, 2018
RNACocktail : Sahraeian et al., Nat Commun., 2017
- 平成30年05月15日
講義資料PDF(約5MB; 2018.05.11版)
(Rで)塩基配列解析
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
- 平成30年05月22日
講義資料PDF(約5MB; 2018.05.17版)
(Rで)塩基配列解析
RNACocktail : Sahraeian et al., Nat Commun., 2017
Kraken : Davis et al., Methods, 2013
Lowe et al., PLoS Comput Biol., 2017
FaQCs : Lo and Chain, BMC Bioinformatics, 2014
FaQCs実行結果のQC.stats.txt
FaQCs実行結果のQC_qc_report.pdf
FastQC
QC.unpaired.trimmed_fastqc.html(課題用)
ShortRead : Morgan et al., Bioinformatics, 2009
Bioconductor : Gentleman et al., Genome Biol., 2004
Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013
report.html(qrqcを用いたQC結果)

R実行結果画面

コピー実行が無事終了すると、①こんな感じのRコンソール画面になります。②リード数が6,000から4,984になっていることがわかる



```
> hoge <- as(quality(fastq), "matrix") #ASCIIコードのqualit$
> obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
> fastq <- fastq[obj] #objがTRUEとなる要素$
> sread(fastq) #配列情報を表示
A DNAStringSet instance of length 4984
      width seq
 [1]    36 TTTAAAAGATAATGTCATCACAAACGCAACATATAGA
 [2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
 [3]    36 TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
 [4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
 [5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
 ...    ...
[4980]   36 CATTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
[4981]   36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
[4982]   36 CTGCGAGCTCGCTCAGGTA CTCTGAGTTCTTCT
[4983]   36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
[4984]   36 CATTGTTCACTTTTGTAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
> |
```

R実行結果画面

また、ここでは最初と最後の計10リード分しか見えては
いないが、①リード長が36 bpに揃っていることがわかる
。この結果から、指定した閾値を満たさないリードごとふ
るい落とされる枠組みであることを思いだす。つまり、リ
ード中の一部がトリミングされるわけではないということ

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes



R Console

```
> hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality$
> obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
> fastq <- fastq[obj] #objがTRUEとなる要素$
> sread(fastq) #配列情報を表示
A DNAStringSet instance of length 4984
      width seq
 [1]    36 TTTAAAAGATAATGTCATCACAAACGCAACATATAGA
 [2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
 [3]    36 TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
 [4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
 [5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
 ...
 [4980]  36 CATTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
 [4981]  36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
 [4982]  36 CTGCGAGCTCGCTCAGGTA CTCTCTGAGTTCTTCT
 [4983]  36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
 [4984]  36 CATTGTTCACTTTTGTAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
> |
```


出力はhoge1.fastq

前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去

Sanger FASTQ形式ファイルを読み込んで、PHREDスコアが低いリードを除去する仕組みを紹介します。

「ファイル」-「ディレクトリの変更」で解析したいファイルを設定

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)

SRR037439から得られるFASTQファイルの最初の2,000行(2010)。PHREDスコアが20未満のものがリード長に占める割合は約10%です。ファイル中のリードは全て35bpのリードである。その10%が4塩基以上あるリードはダメということ) writeFastq関数の出力します。ここではcompress=Fとして非圧縮ファイルを作成

```
in_f <- "DRR000031sub.fastq" #入力ファイル
out_f <- "hoge1.fastq" #出力ファイル
param1 <- 20 #PHREDスコアの閾値
param2 <- 0.1 #指定するリード長
```

```
#必要なパッケージをロード
library(ShortRead) #パッケージ
```

```
#入力ファイルの読み込み
fastq <- readFastq(in_f) #input
sread(fastq) #配列
```

```
#本番
hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality$
obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
fastq <- fastq[obj] #objがTRUEとなる要素$
sread(fastq) #配列情報を表示
```

```
#ファイルに保存
writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
```

```
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes

R Console
> hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality$
> obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
> fastq <- fastq[obj] #objがTRUEとなる要素$
> sread(fastq) #配列情報を表示
A DNAStrngSet instance of length 4984
      width seq
[1]    36 TTTAAAAGATAATGTCATCACAAACGCAACATATAGA
[2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
[3]    36 TCAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
[4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
[5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
...    ...
[4980]  36 CATTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
[4981]  36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
[4982]  36 CTGCGAGCTCGCTCAGGTA CTCTCTGAGTTCTTCT
[4983]  36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
[4984]  36 CATTGTTCACTTTTGTAAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
> |
```

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、`qrrc`でクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

Rでクオリティチェック

①4,984リードからなる②hoge1.fastq
を入力としてqracを実行する

前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去

Sanger FASTQ形式ファイルを読み込んで、PHREDスコアが低いリードを除去する仕組みを紹介します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを設置

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)

SRR037439から得られるFASTQファイルの最初の2,000行(2010)。PHREDスコアが20未満のものがリード長に占める割合は約10%です。ファイル中のリードは全て35bpのリードである。その10%が4塩基以上あるリードはダメということ) writeFastq関数の出力します。ここではcompress=Fとして非圧縮ファイルを作成します。

```
in_f <- "DRR000031sub.fastq" #入力ファイル名
out_f <- "hoge1.fastq" #出力ファイル名
param1 <- 20 #PHREDスコアの閾値
param2 <- 0.1 #指定するリード長
```

```
#必要なパッケージをロード
library(ShortRead) #パッケージのロード
```

```
#入力ファイルの読み込み
fastq <- readFastq(in_f) #入力ファイルの読み込み
sread(fastq) #配列情報の取得
```

```
#本番
hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality$
obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
fastq <- fastq[obj] #objがTRUEとなる要素$
sread(fastq) #配列情報を表示
```

```
#ファイルに保存
writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
```

```
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes

R Console
> hoge <- as(quality(fastq), "matrix") #ASCIIコードのquality$
> obj <- (rowSums(hoge < param1) <= width(fastq)*param2) #条件$
> fastq <- fastq[obj] #objがTRUEとなる要素$
> sread(fastq) #配列情報を表示
A DNAStrngSet instance of length 4984
      width seq
[1]    36 TTTAAAAGATAATGTCATCACACGCAACATATAGA
[2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
[3]    36 TCAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
[4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
[5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
...    ...
[4980] 36 CATTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
[4981] 36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
[4982] 36 CTGCGAGCTCGCTCAGGTA CTCTGAGTTCTTCT
[4983] 36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
[4984] 36 CATTGTTCACTTTTGTAAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F) #fastqの中身を指定し$
> |
```

Rでクオリティ

前処理 | フィルタリング | PHREDスコア

Sanger FASTQ形式ファイルを読み込んで、PHREDスコアが「ファイル」-「ディレクトリの変更」で解析したいファイルを置

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fa)

SRR037439から得られるFASTQファイルの最初の2,000行(2010)。PHREDスコアが20未満のものがリード長に占めるファイル中のリードは全て35bpのリードである。その10%が4塩基以上あるリードはダメということ) writeFastq関数の出力します。ここではcompress=Fとして非圧縮ファイルを

```
in_f <- "DRR000031sub.fastq" #入力ファイル名
out_f <- "hoge1.fastq" #出力ファイル名
param1 <- 20 #PHREDスコア閾値
param2 <- 0.1 #指
```

```
#必要なパッケージをロード
library(ShortRead) #パッケージ
```

```
#入力ファイルの読み込み
fastq <- readFastq(in_f) #input
sread(fastq) #配列
```

```
#本番
hoge <- as(quality(fastq), "matrix") #AS
obj <- (rowSums(hoge < param1) <= width(fa #ob
fastq <- fastq[obj] #ob
sread(fastq) #配列
```

```
#ファイルに保存
writeFastq(fastq, out_f, compress=F) #fa
```

①フィルタリング前の6,000リードからなるDRR000031sub.fastqの、②qrqc実行結果(report.html)と比較することを通じてRでも似たようなフィルタリングができることを学んでもらう。このような単機能のコード実行結果と比較することで、(この場合はFaQCs)プログラムがきちんと動いているかの大まかな確認を行ったりします

```
R GUI (64-bit)
ファイル 編集 閲覧 その他 パック
[Icons]
R Console
> hoge <- as(quality(fastq), "matrix")
> obj <- (rowSums(hoge < param1) <= width(fa
> fastq <- fastq[obj]
> sread(fastq)
A DNASringSet instance
width seq
[1] 36 TTTAAAAGA
[2] 36 TGTAATGAT
[3] 36 TCAAAAAAT
[4] 36 TGAGGTGGC
[5] 36 TTCAGGAAC
... ..
[4980] 36 CATTGTTC
[4981] 36 AAAATATA
[4982] 36 CTGCGAGCT
[4983] 36 CTGAAAACA
[4984] 36 CATTGTTC
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F)
> |
```

講義日程 (平成30年度)	
1. 平成30年05月08日	講義資料PDF .gff3ファイル (約1.3MB) .faファイル (約2.2MB) (Rで)塩基配列解析 (Rで)マイクロアレイデータ解析 plasmid1.gff3(課題用) plasmid2.gff3(課題用) de Lannoy et al., F1000Res., 2017 Garalde et al., Nat Methods, 2018 RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 平成30年05月15日	講義資料PDF(約5MB; 2018.05.11版) (Rで)塩基配列解析 DRR000031sub.fastq RNA-QC-chain : Zhou et al., BMC Genomics, 2018 Biostar : Parnell et al., PLoS Comput Biol., 2011 FastQC DRR000031sub_fastqc.html DRR000031_fastqc.html(課題用) report.html(qrhcを用いたQC結果)
3. 平成30年05月22日	講義資料PDF(約5MB; 2018.05.17版) (Rで)塩基配列解析 RNACocktail : Sahraeian et al., Nat Commun., 2017 Kraken : Davis et al., Methods, 2013



①qrqc。②の入力ファイル名部分をhoge1.fastqに変更して実行

qrqcを実行

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページは、
ツール済みで
的にまとめた

- インポート | ファイル形式の変換 | [qseq -> FASTA](#) (last modified 2013/06/17)
- インポート | ファイル形式の変換 | [qseq -> Illumina FASTQ](#) (last modified 2013/06/17)
- インポート | ファイル形式の変換 | [qseq -> Sanger FASTQ](#) (last modified 2013/08/19)
- [前処理 | クオリティコントロール](#) | [について](#) (last modified 2018/05/01) **NEW**
- 前処理 | クオリティチェック | [QuasR \(Gaidatzis 2015\)](#) (last modified 2015/06/15)
- 前処理 | クオリティチェック | [qrqc](#) (last modified 2014/07/17)
- 前処理 | クオリティチェック | [PHREDスコアに変換](#) (last modified 2018/05/06) **NEW**
- 前処理 | クオリティチェック | [配列長を調整する](#) (last modified 2018/05/06)



前処理 | クオリティチェック | qrqc

[FastQC](#)のR版のようなものです。Sanger FASTQ形式ファイルを読み込んで、positionごとの「クオリティスコア (quality score)」、「どんな塩基が使われているのか(base frequency and base proportion)」、「リード長の分布」、「GC含量」、「htmlレポート」などを出力してくれます。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

[SRR037439](#)から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです ([Bullard et al., 2010](#))。下記を実行すると「SRR037439-report」という名前のフォルダが作成されます。中にあるreport.htmlをダブルクリックするとhtmlレポートを見ることができます。



```

in_f <- "SRR037439.fastq" #入力ファイル名を指定してin_fに格納
#必要なパッケージをロード
library(qrhc) #パッケージの読み込み

#入力ファイルの読み込み
fastq <- readSeqFile(in_f, quality="sanger")#in_fで指定したファイルの読み込み

#本番
makeReport(fastq) #htmlレポートの作成

```

What's new

- Silhouette
- Silhouette
- 「平成29年

- [門田から](#)
- [はじめに](#)
- [参考資料](#)
- [参考資料](#)

qrqcを実行

①入力ファイル名部分を変更してコピー実行。②警告メッセージが出ていますがよくわかりません。結果は③というフォルダ中のreport.htmlというファイルに吐き出されます

```
in_f <- "hogel.fastq" ①
#必要なパッケージをロード
library(qrhc)          #パッケージの読み込み

#入力ファイルの読み込み
fastq <- readSeqFile(in_f, quality="sanger")#in_fで指定したファイルの読み込み

#本番
makeReport(fastq)
```

#入力ファイル名を指定してin_fに格納

#パッケージの読み込み

```
R Console
要求されたパッケージ brew をロード中です
要求されたパッケージ xtable をロード中です
要求されたパッケージ testthat をロード中です
>
> #入力ファイルの読み込み
> fastq <- readSeqFile(in_f, quality="sanger")#in_fで$
>
> #本番
> makeReport(fastq) #htmlレポ-$
警告: Ignoring unknown aesthetics: y
`geom_smooth()` using method = 'gam'
Report written to directory './hogel-report'.
> | ③
```



qrrc実行結果

①report.htmlファイルはこちらにもあります。②のグラフが、FastQCのPer base sequence qualityに相当することがわかる。③この部分の数値は実行ごとにコロコロ変わるようです

3. 平成30年05月22日

講義資料PDF(約5MB; 2018.05.17版)

(Rで)塩基配列解析

RNAcocktail : Sahraeian et al., Nat Commun., 2017

Kraken : Davis et al., Methods, 2013

Lowe et al., PLoS Comput Biol., 2017

FaQCs : Lo and Chain, BMC Bioinformatics, 2014

FaQCs実行結果のQC.stats.txt

FaQCs実行結果のQC_qc_report.pdf

FastQC

QC.unpaired.trimmed_fastqc.html(課題用)

ShortRead : Morgan et al., Bioinformatics, 2009

Bioconductor : Gentleman et al., Genome Biol., 2004

Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013

report.html(qrrcを用いたQC結果)

QuasR(Windows版あり) : Gaidatzis et al., Bioinformatics, 2015

Bowtie : Langmead et al., Genome Biol., 2009

Bowtie2 : Langmead and Salzberg, Nat. Methods, 2012

sample_RNAseq1.sam(Bowtie2の実行結果SAMファイル)

Sequence Alignment/Map Format Specification

General information

File: hoge1.fastq

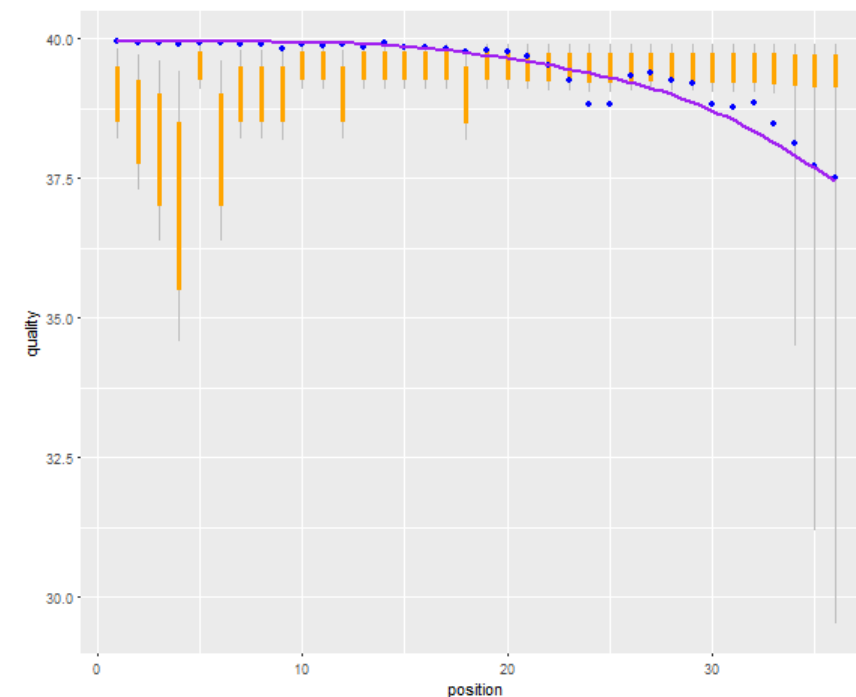
Type: FASTQ

Sequence Length Range: 36 to 36

Total Sequences: 4984

Unique Sequences: 4453

Quality by Position



Grey lines: 10% and 90% quantiles

Orange lines: 25% and 75% quartiles

Blue point: median

Green dash: mean

Purple line: lowess curve

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

全体像のおさらい

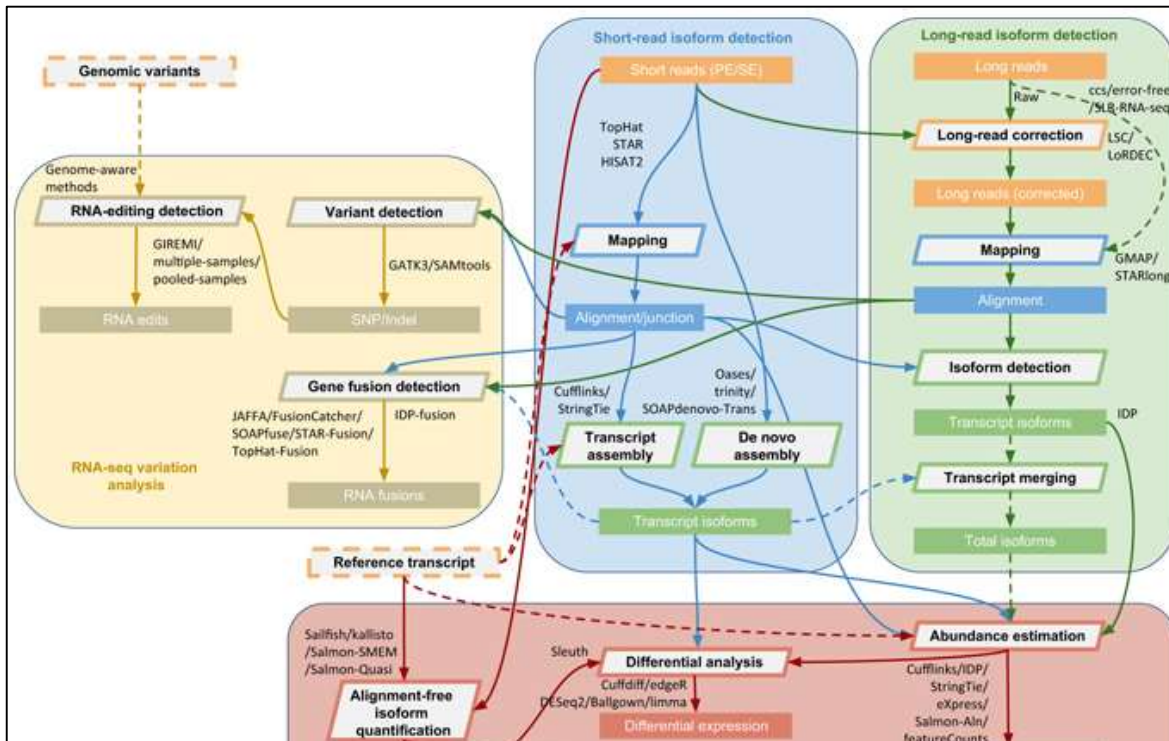
NGSリードデータ(SRAファイル)



NGSリードデータ(FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)



RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

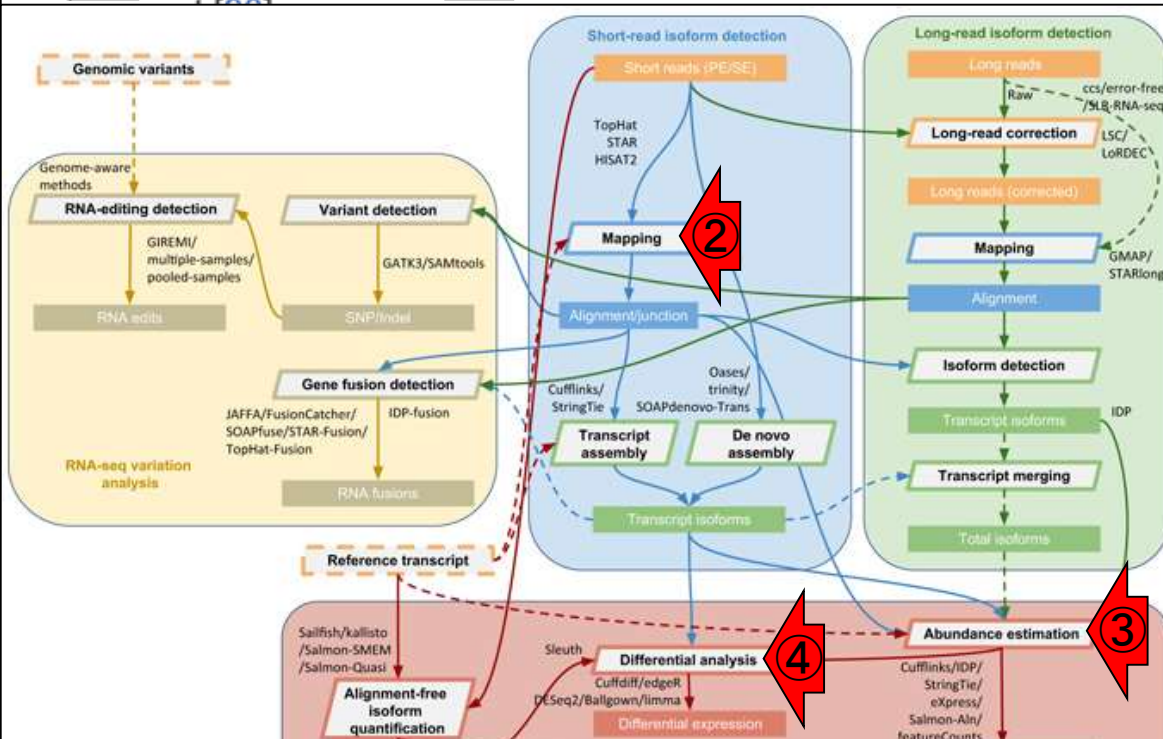
全体像のおさらい

①のQCが終わったので、次は
②マッピング(アラインメント)

RNA-Seq data analysis

最近の総説 (Lowe et al., PLoS Comput. Biol., 13: e1005457, 2017)

processed to yield useful information. Data analysis usually requires a combination of **bioinformatics software** tools that vary according to the experimental design goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a command-line interface, either in a Unix environment or within the R/Bioconductor statistical



...ence needs to be
...ty scores for base
... representation of
... duplication rate [85].
... and FaQCs software
... tagged for special

RNACocktail (Sahraeian et al., Nat Commun., 8: 59, 2017)

マッピング = 大量高速文字列検索

- マップされる側のリファレンス配列: `hoge4.fa`
- マップする側のRNA-seqリードデータ: "AGG"

マッピングプログラムの出力:
(どのリードが)リファレンス配列上のどの位置から転写されたものかという座標情報

```
hoge4.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>contig_1
CGGACAGCTCCTCGGCATCCGGAT
>contig_2
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCCTAACATCGCAAGTCG
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCCT
GTC
>contig_3
TGTAGGAGAAGGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT
GTATGAGGTCGGGCA
>contig_4
CGTGCTGATTCCACACAGCAGTAAACGCGGACCTCTACCTATGAACATG
```

出力ファイル

	start	end
contig_2	31	33
contig_2	77	79
contig_3	4	6
contig_3	10	12
contig_3	56	58

マッピング

①「マッピング | について」の最終更新日は②2018年となっはいるが、きっちり調査していたのは2014年頃まで。情報としては相当古い

- 前処理 | フィルタリング | paired-end | 共通リード抽出 | [ShortRead\(Morgan 2009\)](#) (last modified 2015/06/26)
- [アセンブル | について](#) (last modified 2014/06/20)
- アセンブル | [ゲノム用](#) (last modified 2016/03/24)
- アセンブル | [トランスクリプトーム\(転写物\)用](#) (last modified 2016/06/20)
- [マッピング | について](#) ① (last modified 2018/05/12) ②
- マッピング | [basic aligner](#) (last modified 2014/08/08)
- マッピング | [splice-aware aligner](#) (last modified 2016/04/07)
- マッピング | [Bisulfite sequencing用](#) (last modified 2014/07/09)
- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
- マッピング | [基礎](#) (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | paired-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | paired-end | トランスクリプトーム | basic aligner(基礎) | [QuasR\(Gaidatzis et al., 2015\)](#)
- マッピング | paired-end | トランスクリプトーム | basic aligner(応用) | [QuasR\(Gaidatzis et al., 2015\)](#)
- [マップ後 | について](#) (last modified 2013/06/19)
- [マップ後 | 出力ファイル形式について](#) (last modified 2013/11/05)

マッピング | について NEW

リファレンス配列にマッピングを行うプログラム達です。basic aligner (unspliced aligner)はsplice-aware aligner (spliced aligner)内部で使われていたりします。

R用:

- [Rsubread](#)(Windows版なし): [Liao et al., Nucleic Acids Res., 2013](#)
- [QuasR](#)(Windows版あり): [Gaidatzis et al., Bioinformatics, 2015](#)
- [HTSeqGenie](#)(Windows版なし): [Wu et al., Methods Mol Biol., 2016](#)

R以外(basic aligner; unspliced aligner):

- [SSAHA2](#): [Ning et al., Genome Res., 2001](#)
- [RMAP](#): [Smith et al., BMC Bioinformatics, 2008](#)
- [MAQ](#): [Li et al., Genome Res., 2008](#)
- [PASS](#): [Campagna et al., Bioinformatics, 2009](#)
- [MOM](#): [Eaves and Gao, Bioinformatics, 2009](#)
- [Bowtie](#): [Langmead et al., Genome Biol., 2009](#)
- [BWA](#): [Li and Durbin, Bioinformatics, 2009](#)(BWA-shortの論文)
- [SHRiMP](#): [Rumble et al., PLoS Comput Biol., 2009](#)

マッピング

①Rパッケージとして提供されているものは圧倒的に少ないですが、②QuasRパッケージのおかげでWindows OS上でもマッピングを気軽に利用できるようになりました。これは内部的に③Bowtie プログラムを利用しています

- 前処理 | フィルタリング | paired-end | 共通リード抽出 | [ShortRead\(More\)](#)
- [アセンブル | について](#) (last modified 2014/06/20)
- アセンブル | [ゲノム用](#) (last modified 2016/03/24)
- アセンブル | [トランスクリプトーム\(転写物\)用](#) (last modified 2016/06/20)
- [マッピング | について](#) (last modified 2018/05/12) **NEW**
- マッピング | [basic aligner](#) (last modified 2014/08/08)
- マッピング | [splice-aware aligner](#) (last modified 2016/04/07)
- マッピング | [Bisulfite sequencing用](#) (last modified 2014/07/09)
- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
- マッピング | [基礎](#) (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis\)](#)
- マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis\)](#)
- マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis\)](#)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis\)](#)
- マッピング | paired-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis\)](#)
- マッピング | paired-end | トランスクリプトーム | basic aligner(基礎) | [QuasR\(Gaidatzis\)](#)
- マッピング | paired-end | トランスクリプトーム | basic aligner(応用) | [QuasR\(Gaidatzis\)](#)
- [マップ後 | について](#) (last modified 2013/06/19)
- [マップ後 | 出力ファイル形式について](#) (last modified 2013/11/05)

マッピング | について **NEW**

リファレンス配列にマッピングを行うプログラム達です。basic aligner (unspliced aligner)はsplice-aware aligner (spliced aligner)内部で使われていたりします。

R用:

- [Rsubread](#)(Windows版なし): [Liao et al., Nucleic Acids Res., 2013](#)
- [QuasR](#)(Windows版あり): [Gaidatzis et al., Bioinformatics, 2015](#)
- [HTSeqGenie](#)(Windows版なし): [Wu et al., Methods Mol Biol., 2016](#)

R以外(basic aligner; unspliced aligner):

- [SSAHA2](#): [Ning et al., Genome Res., 2001](#)
- [RMAP](#): [Smith et al., BMC Bioinformatics, 2008](#)
- [MAQ](#): [Li et al., Genome Res., 2008](#)
- [PASS](#): [Campagna et al., Bioinformatics, 2009](#)
- [MOM](#): [Eaves and Gao, Bioinformatics, 2009](#)
- [Bowtie](#): [Langmead et al., Genome Biol., 2009](#)
- [BWA](#): [Li and Durbin, Bioinformatics, 2009](#)(BWA-shortの論文)
- [SHRiMP](#): [Rumble et al., PLoS Comput Biol., 2009](#)

マッピング

オプションは、デフォルトである程度よきに計らってくれるが...実際の挙動を完全に把握できる状況で様々なオプションを試したい

■ 内部的にBowtieを利用

- マッピング時に多くのオプションを指定可能
- “-v”: 許容するミスマッチ数を指定するオプション。“-v 0”は、リードがリファレンスに完全一致するもののみレポート。“-v 2”は、2塩基ミスマッチまで許容してマップされうる場所を探索。
- “-m”: 出力するリード条件を指定するオプション。“-m 1”は、複数個所にマップされるリードを除外して、1か所にのみマップされたリードをレポート。“-m 3”は、合計3か所にマップされるリードまでをレポート。
- “--best --strata”: 最も少ないミスマッチ数でマップされるもののみ出力する、という意思表示。これをつけずに“-v 2 -m 1”などと指定すると、たとえ完全一致(ミスマッチ数0)で1か所にのみマップされるリードがあったとしても、どこか別の場所で1塩基ミスマッチでマップされる個所があれば、マップされうる場所が2か所ということの意味し、そのリードは出力されなくなる。それを防ぐのが主な目的
- ...

マップされる側のリファレンス配列は、①
ref_genome.fa。後でダウンロードできます

マップされる側

- マップされる側のリファレンス配列: ref_genome.fa

```
ref_genome.fa - ノート帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```


①ref_genome.faは、②サンプルデータの、③例題18のコピペで作成しています

マップされる側

■ マップされる側のリファレンス配列: ref_genome.fa

- インストール | Rパッケージ | [必要最小限\(数GB?\)](#) (last modified 2015/05/25)
- インストール | Rパッケージ | [個別](#) (last modified 2015/06/10)
- (削除予定)[Rのインストールと起動](#) (last modified 2016/02/21)
- (削除予定)[個別パッケージのインストール](#) (last modified 2015/02/20)
- [基本的な利用法](#) (last modified 2015/04/03)
- [サンプルデータ](#) (last modified 2018/01/11)
- バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | [NGSハンズオン講習会2017](#)
- バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | [NGSハンズオン講習会2016](#)



サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 2008)
 「Kidney 7 samples vs Liver 7 samples」のRNA-seqの遺伝子発現行列データ(SupplementaryTable2.txt)です。サンプルは二つの濃度(1.5 pM and 3 pM)でシーケンスされており、「3 pMのものが5 samples vs. 5 samples」、「1.5 pMのものが2 samples vs. 2 samples」という構成です。SupplementaryTable2.txtをエクセルで開くと7列目以降に発現データがあることがわかります。詳細な情報はこちらの通りです(直葉



18. ランダムな塩基配列から生成したリファレンスゲノム配列データ(ref_genome.fa)。48, 160, 100, 123, 100 bpの配列長をもつ、計5つの塩基配列を生成しています。description行は"contig"という記述を基本としています。塩基の存在比はAが28%, Cが22%, Gが26%, Tが24%にしています。set.seed関数を利用し、chr3の配列と同じものをchr5としてコピーして作成したのち、2番目と7番目の塩基置換を行っています。そのため、実際に指定するのは最初の4つ分の配列長のみです。

```

out_f <- "ref_genome.fa" #出力ファイル名を指定してout_fに格納
param_len_ref <- c(48, 160, 100, 123) #配列長を指定
narabi <- c("A", "C", "G", "T") #以下の数値指定時にACGTの並びを間違えないようにするために表示(内部的)
param_composition <- c(28, 22, 26, 24) #(A,C,G,Tの並びで)各塩基の存在比率を指定
param_desc <- "chr" #FASTA形式ファイルのdescription行に記述する内容
param4 <- 3 #コピーを作成したい配列番号を指定
param5 <- c(2, 7) #コピー先配列の塩基置換したい位置を指定

#必要なパッケージをロード
library(Biostrings) #パッケージの読み込み

#塩基置換関数の作成
enkichikan <- function(fa, p) { #関数名や引数の作成
  t <- substring(fa, p, p) #置換したい位置の塩基を取り出す

```

マップされる側

- マップされる側のリファレンス配列: `ref_genome.fa`

```

ref_genome.fa - ノート
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACCGCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAAGTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
  
```

マップされる側

- マップされる側のリファレンス配列:

①ref_genome.faの説明。②chr3と③chr5の違いは、④2番目と⑤7番目の塩基のみ。“-m”オプションの違いの把握が可能。“-m 1”は、複数個所にマップされるリードを除外して、1か所にのみマップされたリードをレポート

```
ref_genome.fa - ノート
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```

マップする側

マップする側のリードファイルは①sample_RNAseq1.fa。許容する mismatches 数による違いや、マップされるべき場所が完全に把握できるように、リードの description 行に記述されている

- マップする側の RNA-seq データ: sample_RNAseq1.fa

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT

sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1 11 45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```


マップする側

① sample_RNAseq1.faは、②サンプルデータの、③例題19のコピペで作成しています。実際のマッピング時の直前にダウンロードします

■ マップする側のRNA-seqデータ: sample_RNAseq1.fa

- インストール | Rパッケージ | [必要最小限\(数GB?\)](#) (last modified 2015/05/25)
- インストール | Rパッケージ | [個別](#) (last modified 2015/06/10)
- (削除予定)[Rのインストールと起動](#) (last modified 2016/02/21)
- (削除予定)[個別パッケージのインストール](#) (last modified 2015/02/20)
- [基本的な利用法](#) (last modified 2015/04/03)
- [サンプルデータ](#) (last modified 2018/01/11)
- バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | [NGSハンズオン講習会2017](#)
- バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ) | [NGSハンズオン講習会2016](#)

```
sample_RNAseq1.fa - ノート帳
ファイル(F) 編集(E) 検索(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
```

サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 2008)

「Kidney 7 samples vs Liver 7 samples」のRNA-seqの遺伝子発現行列データ(SupplementaryTable2.txt)です。サンプルは二つの濃度(1.5 pM and 3 pM)でシーケンスされており、「3 pMのものが5 samples vs. 5 samples」、「1.5 pMのものが2 samples vs. 2 samples」という構成です。SupplementaryTable2.txtをエクセルで開くと7列目以降に発現データがあることがわかります。詳細な情報はこちらの通りです。

19. 上記リファレンスゲノム配列データ(ref_genome.fa)に対してbasic alignerでマッピングする際の動作確認用RNA-seqデータ(sample_RNAseq1.fa)とそのgzip圧縮ファイル(sample_RNAseq1.fa.gz)。リファレンス配列を読み込んで、list_sub3.txtで与えた部分配列を抽出したものです。どこに置換を入れているかがわかっているので、basic alignerで許容するミスマッチ数を変えてマップされる or されないの確認ができます。DNAStrngSetオブジェクトを入力として塩基置換を行うDNAStrng_chartr関数を用いて、最後のリードのみ4番目の塩基にミスマッチを入れています。

```
in_f1 <- "ref_genome.fa" #入力ファイル名を指定してin_f1に格納(multi-FASTAファイル)
in_f2 <- "list_sub3.txt" #入力ファイル名を指定してin_f2に格納(リストファイル)
out_f <- "sample_RNAseq1.fa" #出力ファイル名を指定してout_fに格納
param <- 4 #塩基置換したい位置を指定
```

```
#必要なパッケージをロード
library(Biostrings) #パッケージの読み込み
```

```
#塩基置換関数の作成
DNAStrng_chartr <- function(fa, p) { #関数名や引数の作成
  str_list <- as.character(fa) #文字列に変更
  t <- substring(str_list, p, p) #置換したい位置の塩基を取り出す
  t <- chartr("CGAT" "GCTA" t) #置換後の塩基を作成
```


マップする側

① sample_RNAseq1.faは、全部で8リードからなる。このうち、②最後のリード(chr5_1_35)のみ、③4番目の塩基を変えている

- マップする側のRNA-seqデータ: sample_RNAseq1.fa

The image shows two Notepad++ windows side-by-side. The left window is titled 'ref_genome.fa - メモ帳' and displays the reference genome sequence. The right window is titled 'sample_RNAseq1.fa - メモ帳' and displays the sample RNA-seq data. Red boxes and arrows highlight specific differences between the two files.

Window 1: ref_genome.fa

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```

Window 2: sample_RNAseq1.fa

```
>chr1 11 45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

Annotations:

- ①: Red arrows pointing to the window titles.
- ②: A red box around the sequence 'CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT' in the sample window, with an arrow pointing to the corresponding sequence in the reference window.
- ③: A red box around the sequence 'GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC' in the sample window, with an arrow pointing to the corresponding sequence in the reference window.

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

マッピング

実際に、①リファレンス配列ファイル(ref_genome.fa)と、②仮想RNA-seqリードファイル(sample_RNAseq1.fa)を用いてマッピングを行ってみましょう。③basic aligner(応用)の、④例題1です

- マッピング | (ESTレベルの長さの)contig (last modified 2014/06/24)
- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis 2015) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis 2015) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis 2015) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis 2015) (last modified 2016/02/11)

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis 2015)

QuasRパッケージを用いて single-end RNA-seq データのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいフォルダ②置いてあるディレクトリ①に移動し以下をコピー。

④ 1. サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0"#マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR) #パッケージの読み込み
library(GenomicAlignments) #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time() #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping)#マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment_statistics)の表示。seqlength: リファレンス
```

マッピング

デスクトップのhogeフォルダ内にmapping_kiso1というフォルダを作成し、そこに①②③の3つのファイルを保存しましょう。

- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
- マッピング | [基礎](#) (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2016/02/11)

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいフォルダ②置いてあるディレクトリ①に移動し以下をコピー。

1. サンプルデータ18,19のRNA-seqデータ(sample RNAseq1.fa)のref genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0" #マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR) #パッケージの読み込み
library(GenomicAlignments) #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time() #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping) #マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment_statistics)の表示。seqlength: リファレンス
```

マッピング

- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
 - マッピング | [基礎](#) (last modified 2013/06/19)
 - マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/02/24)
 - マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/02/24)
 - マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/02/24)
 - マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/02/24)
 - マッピング | paired-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/02/24)
- マッピング | single-end | ゲノム | basic aligner**
- QuasRパッケージを用いて single-end RNA-seq データのリファレンスゲノムへマッピングする。あるBowtie (Langmead et al., Genome Biol., 2009)を実装したBWA (Li and Durbin, Bioinformatics, 2009)のライブラリが提供されるリードの取り扱い (uniquely mapped reads or multi-mapped reads) などの様々なオプションを利用可能ですが、「基礎」のところでは基本的な動作を確認したり、複数のRNA-seqファイルを一括でマッピングするためのオプションを確認したい。*.bedの3つです。それ以外のファイルは基本無視で大丈夫です。
- 「ファイル」→「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. サンプルデータ18,19のRNA-seqデータ(sample RNAseq1.fa)のref genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt"
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best --strata -v 0"

#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)

#本番(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, alignmentParameters=param_mapping)
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/mapping_kiso1"
> list.files()
[1] "mapping_single_genome1.txt"
[2] "ref_genome.fa"
[3] "sample_RNAseq1.fa"
> |
```

#マッピングに用いたパラメータや入力ファイルの情報などを表示
#マッピング結果(alignment_statistics)の表示。seqlength: リファレンス

入力ファイル

- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
- マッピング | [基礎](#) (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2016/02/11)

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

① 1. サンプルデータ18,19のRNA-seqデータ(sample RNAseq1.fa)のref genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```

in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0" #マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR) #パッケージの読み込み
library(GenomicAlignments) #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time() #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping) #マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment_statistics)の表示。seqlength: リファレンス
  
```

リストファイル

①マップする側のリードファイル名のリストファイル。これは、複数のRNA-seqサンプルを実行できるようにするため。②がリストファイルの中身

- マッピング | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24)
- マッピング | [基礎](#) (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | [QuasR\(Gaidatzis 2015\)](#) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | [QuasR\(Gaidatzis 2015\)](#) (last modified 2016/02/11)

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seq データのリファレンスゲノム配列クローンであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージをインストールするリードの取り扱い (uniquely mapped reads or multi-mapped reads)を"-m"オプションの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんので挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示した "*.bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. サンプルデータ18,19のRNA-seqデータ(sample RNAseq1.fa)のref genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0" #マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR) #パッケージの読み込み
library(GenomicAlignments) #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time() #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping) #マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment_statistics)の表示。seqlength: リファレンス
```

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	namea

Botwieオプション

①マップする側のリードファイル名のリストファイル。これは、複数のRNA-seqサンプルを実行できるようにするため。②がリストファイルの中身。③がマッピングプログラムであるBowtie実行時に指定しているオプション。許容するミスマッチ数は0個(-v 0)、1か所にマップされるリードのみ出力(-m 1)です

- マッピング | (ESTレベルの長さの)contig (last modified 2014/06/24)
- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マップ後 | 出力 | 基礎
- マップ後 | 出力 | 応用
- マップ後 | 出力 | 応用
- マップ後 | 出力 | 応用
- マップ後 | 出力 | 応用
- マップ後 | 出力 | 応用

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seq データのリファレンスゲノム配列に対してあるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージをインストールされるリードの取り扱い (uniquely mapped reads or multi-mapped reads)を"-m"オプションなどの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんので挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。 "*.bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	namea

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。
1. サンプルデータ18,19のRNA-seqデータ(sample RNAseq1.fa)のref genome.faへのマッピングの場合(mapping single genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0" #マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR) #パッケージの読み込み
library(GenomicAlignments) #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time() #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping) #マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment_statistics)の表示。seqlength: リファレンス
```

コピー実行

コピー実行時の途中経過。R Console画面は、マッピングのメインである①qAlign関数実行部分

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 -best -strata -v 0"とした("chr3_11_45"と"chr3_15_49")はchr5にもではない(4番目の塩基にミスマッチをいれ

```
in_f1 <- "mapping_single_genome
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best -
```

```
#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)
```

```
#本番(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
R Console
> time_s <- proc.time() #計算時間を計測するため
> out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping) # $
Creating .fai file for: C:/Users/kojik/Desktop/hoge/mapping_kiso$
alignment files missing - need to:
  create alignment index for the genome
  create 1 genomic alignment(s)
will start in ..9s..8s..7s..6s..5s..4s..3s..2s..1s
Creating an Rbowtie index for C:/Users/kojik/Desktop/hoge/mappin$
Finished creating index
Testing the compute nodes...OK
Loading QuasR on the compute nodes...OK
Available cores:
nodeNames
DESKTOP-3J8LKP8
1
Performing genomic alignments for 1 samples. See progress in the$
C:/Users/kojik/Desktop/hoge/mapping_kiso1\QuasR_log_146c2f3d4e64$
Genomic alignments have been created successfully
```


①qAlign関数実行結果を格納したoutオブジェクトを、②確認している部分

途中経過

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 --best --strata -v 0"とした("chr3_11_45"と"chr3_15_49")はchr5にもではない(4番目の塩基にミスマッチをいれ

```
in_f1 <- "mapping_single_genome
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best -
```

```
#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)
```

```
#本番(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_s
out
alignerStats(out)
```

```
R Console
> out #マッピングに用いたパラメータ
Project: qProject
Options: maxHits: 1
         paired: no
         splicedAlignment: FALSE
         bisulfite: no
         snpFile: none
Aligner: Rbowtie v1.18.0 (parameters: -m 1 --best --strata -$)
Genome: C:/Users/kojik/Desktop/hoge/.../ref_genome.fa (file)

Reads: 1 file, 1 sample (fasta format):
1. sample_RNAseq1.fa namae

Genome alignments: directory: same as reads
1. sample_RNAseq1_146c6c6d54aa.bam

Aux. alignments: none
```


途中経過

マッピング | single-end | ゲノム | basic aligner(応用) | C

①qAlign関数実行結果を格納したoutオブジェクトを、②確認している部分。③使用したマッピングプログラム、④マップされる側のリファレンス配列、⑤マップする側のリードファイル情報などが見られます

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 --best --strata -v 0"とした("chr3_11_45"と"chr3_15_49")はchr5にもではない(4番目の塩基にミスマッチをいれ

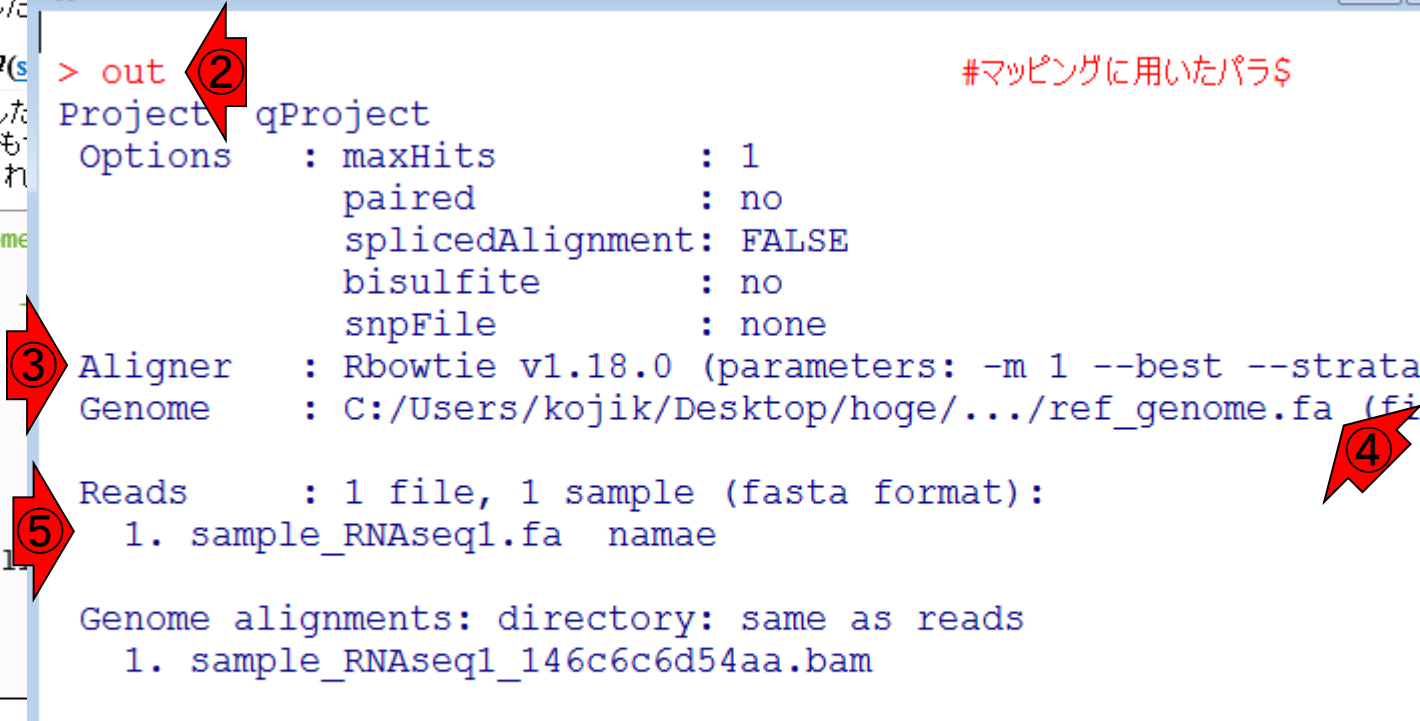
```
in_f1 <- "mapping_single_genome
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best
```

```
#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)
```

```
#本番(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, al
time_e <- proc.time()
time_s
time_e
out
alignerStats(out)
```

```
R Console
> out
Project qProject
Options : maxHits : 1
         paired   : no
         splicedAlignment: FALSE
         bisulfite  : no
         snpFile   : none
Aligner  : Rbowtie v1.18.0 (parameters: -m 1 --best --strata -$
Genome   : C:/Users/kojik/Desktop/hoge/.../ref_genome.fa (file)
Reads    : 1 file, 1 sample (fasta format):
          1. sample_RNAseq1.fa  namae
Genome alignments: directory: same as reads
                  1. sample_RNAseq1_146c6c6d54aa.bam
Aux. alignments: none
```

#マッピングに用いたパラメータ



途中経過

マッピング | single-end | ゲノム | basic aligner(応用) | C

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行う。あるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを用いてマッピングされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定する。様々なオプションを利用可能ですが、「基礎」のところではやり方を示しません。挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示す".bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ

オプションを"-m 1 -best --strata -v 0"としたマッピング ("chr3_11_45"と"chr3_15_49")はchr5にもマッピングされる(4番目の塩基にミスマッチをいれ

```
in_f1 <- "mapping_single_genome
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best
```

```
#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)
```

```
#本番(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, aligner="Rbowtie",
time_e <- proc.time()
time_s
out
alignerStats(out)
```

①qAlign関数実行結果を格納したoutオブジェクトを、②確認している部分。③使用したマッピングプログラム、④マップされる側のリファレンス配列、⑤マップする側のリードファイル情報などが見られます。⑥がBAM形式のマッピング結果ファイルです。⑦の赤下線部分は乱数で発生させているので、ヒトそれぞれです

```
R Console
> out
Project: qProject
Options : maxHits : 1
         paired   : no
         splicedAlignment: FALSE
         bisulfite  : no
         snpFile   : none
Aligner  : Rbowtie v1.18.0 (parameters: -m 1 --best --strata -v 0)
Genome   : C:/Users/kojik/Desktop/hoge/.../ref_genome.fa (file)
Reads    : 1 file, 1 sample (fasta format):
           1. sample_RNAseq1.fa  nameae
Genome alignments: directory: same as reads
                   1. sample_RNAseq1_146c6c6d54aa.bam
Aux. alignments: none
```

マッピング結果の概要

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いて single-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 -best -strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```

out <- qAlign(in_f1, in_f2, alignmentParameter=param mapping)#マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment statistics)の表示。seqlength: リファレンス

#ファイルに保存(QCレポート用のpdfファイル作成)
out_f <- sub(".bam", "_QC.pdf", out@alignments[,1])#Ququality Controlレポートのpdfファイル名を作成した結果をc
qQCReport(out, pdfFilename=out_f) #QCレポート結果をファイルに保存
out_f #ファイル名を表示してるだけです

#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1] #ファイル名(in_f1の1列目に相当)をtmpfnameとして取り扱いたいだけです
for(i in 1:length(tmpfname)){ #サンプル数(ファイル数)分だけループを回す
  hoge <- readGAlignments(tmpfname[i]) #BAM形式ファイルを読み込んだ結果をhogeに格納(これはGAlignmentsオブジ
  hoge <- as.data.frame(hoge) #データフレーム形式に変換
  tmp <- hoge[, c("seqnames", "start", "end")]#必要な列の情報のみ抽出した結果をtmpに格納
  out_f <- sub(".bam", ".bed", tmpfname[i])#BED形式ファイル名を作成した結果をout_fに格納
  out_f #ファイル名を表示してるだけです
  write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F)#tmpの中身を指定したファイ
}

```

マッピング結果の概要

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 -best -strata -v 0"とした
("chr3_11_45"と"chr3_15_49")はchr5にも
ではない(4番目の塩基にミスマッチをいれ

```
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
#ファイルに保存(QCレポート用のpdf)
out_f <- sub(".bam", "_QC.pdf",
qQCReport(out, pdfFilename=out_
out_f
```

```
#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1]
for(i in 1:length(tmpfname)){
  hoge <- readGAlignments(tmpfr
  hoge <- as.data.frame(hoge)
  tmp <- hoge[, c("seqnames", "s
  out_f <- sub(".bam", ".bed",
  out_f
  write.table(tmp, out_f, sep="
}
```

```
Genome alignments: directory: same as reads
```

```
1. sample_RNAseq1_146c6c6d54aa.bam
```

```
Aux. alignments: none
```

```
> alignmentStats(out)
```

```
#マッピング結果 (alignme$
```

```
          seqlength mapped unmapped
name:genome      531         5         3
```

```
> #ファイルに保存(QCレポート用のpdfファイル作成)
```

```
> out_f <- sub(".bam", "_QC.pdf", out@alignments[,1])#Quqlity Co$
```

```
> qQCReport(out, pdfFilename=out_f) #QCレポート結果をファイ$
```

```
collecting quality control data
```

```
creating QC plots
```

```
> out_f
```

```
#ファイル名を表示してる$
```

```
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1/sample_RNAseq1_1$
```

```
>
```

```
> #ファイルに保存(BED形式ファイル)
```

マッピング結果の概要

マッピング | single-end | ゲノム | basic aligner(応用) | Q

マップされる側のリファレンス配列ファイル(`ref_genome.fa`)のゲノムサイズに相当する総塩基数は②531 bp。③マップされたリード数は5、④マップされなかったリード数は3。こんな感じでどの程度のリード数がマップされたかの情報を把握する

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのあるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 -best -strata -v 0"とした("chr3_11_45"と"chr3_15_49")はchr5にもではない(4番目の塩基にミスマッチをいれ

```
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
#ファイルに保存(QCレポート用のpdf
out_f <- sub(".bam", "_QC.pdf",
qQCReport(out, pdfFilename=out_
out_f
```

```
#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1]
for(i in 1:length(tmpfname)){
  hoge <- readGAlignments(tmpfr
  hoge <- as.data.frame(hoge)
  tmp <- hoge[, c("seqnames", "s
  out_f <- sub(".bam", ".bed",
  out_f
  write.table(tmp, out_f, sep="
}
```

```
R Console
Genome alignments: directory: same as reads
  1. sample_RNAseq1_146c6c6d54aa.bam

Aux. alignments: none

> alignmentStats(out)
#マッピング結果 (alignme$
      seqlength mapped unmapped
name:genome      531      5      3
> #ファイルに保存(QCレポート用のpdfファイル作成)
> out_f <- sub(".bam", "_QC.pdf", out@alignments[,1])#Quqlity Co$
> qQCReport(out, pdfFilename=out_f) #QCレポート結果をファイ$
collecting quality control data
creating QC plots
> out_f #ファイル名を表示してる$
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1\\sample_RNAseq1_1$
> #ファイルに保存(BED形式ファイル)
```


無事終了したら...

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 -best -strata -v 0"とした
("chr3_11_45"と"chr3_15_49")はchr5にも
ではない(4番目の塩基にミスマッチをいれ

```
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
#ファイルに保存(QCレポート用のpdf)
out_f <- sub(".bam", "_QC.pdf",
qQCReport(out, pdfFilename=out_
out_f
```

```
#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1]
for(i in 1:length(tmpfname)){
  hoge <- readGAlignments(tmpfr
  hoge <- as.data.frame(hoge)
  tmp <- hoge[, c("seqnames", "s
  out_f <- sub(".bam", ".bed",
  out_f
  write.table(tmp, out_f, sep="
}
```

```
> out_f <- sub(".bam", "_QC.pdf", out@alignments[,1])#Quqlity Co$
> qQCReport(out, pdfFilename=out_f) #QCレポート結果をファイ$
collecting quality control data
creating QC plots
> out_f #ファイル名を表示してる$
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1/sample_RNAseq1_1$
>
> #ファイルに保存(BED形式ファイル)
> tmpfname <- out@alignments[,1] #ファイル名(in_f1の1列)$
> for(i in 1:length(tmpfname)){ #サンプル数(ファイル数)$
+ hoge <- readGAlignments(tmpfname[i]) #BAM形式ファイルを読み$
+ hoge <- as.data.frame(hoge) #データフレーム形式に変$
+ tmp <- hoge[, c("seqnames", "start", "end")] #必要な列の情報の$
+ out_f <- sub(".bam", ".bed", tmpfname[i]) #BED形式ファイル名$
+ out_f #ファイル名を表示してる$
+ write.table(tmp, out_f, sep="\t", append=F, quote=F, row.nam$
+ }
> |
```

出力ファイルの確認

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniuely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*.bam", "*_QC.pdf", "*.bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s

オプションを"-m 1 -best -strata -v 0"とした
("chr3_11_45"と"chr3_15_49")はchr5にも
ではない(4番目の塩基にミスマッチをいれ

```
out <- qAlign(in_f1, in_f2,
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
#ファイルに保存(QCレポート用のpdf)
out_f <- sub(".bam", "_QC.pdf",
qQCReport(out, pdfFilename=out_f,
out_f
```

```
#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1]
for(i in 1:length(tmpfname)){
hoge <- readGAlignments(tmpfname[i])
hoge <- as.data.frame(hoge)
tmp <- hoge[, c("seqnames", "start", "end", "strand")]
out_f <- sub(".bam", ".bed", tmpfname[i], tmp)
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)
}
```

```
R Console
+ write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)
+ }
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kis01"
> list.files()
[1] "mapping_single_genome1.txt"
[2] "QuasR_log_146c2f3d4e64.txt"
[3] "ref_genome.fa"
[4] "ref_genome.fa.fai"
[5] "ref_genome.fa.md5"
[6] "ref_genome.fa.Rbowtie"
[7] "sample_RNAseq1.fa"
[8] "sample_RNAseq1_146c6c6d54aa.bam"
[9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
> |
```

①拡張子が.bamの、通称BAMファイルと呼ばれるものです

主なマッピング結果ファイルは

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniqely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*_bam", "*_QC.pdf", "*_bed"の3つです。それ以外のファイルは基本無視で大丈夫です

「ファイル」-「ディレクトリの変更」で解析した

1. サンプルデータ18,19のRNA-seqデータ(s) オプションを"-m 1 -best -strata -v 0"とした ("chr3_11_45"と"chr3_15_49")はchr5にも ではない(4番目の塩基にミスマッチをいれ

```
out <- qAlign(in_f1, in_f2, ali
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)
```

```
#ファイルに保存(QCレポート用のpdf)
out_f <- sub(".bam", "_QC.pdf",
qQCReport(out, pdfFilename=out_
out_f
```

```
#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1]
for(i in 1:length(tmpfname)){
  hoge <- readGAlignments(tmpfr
  hoge <- as.data.frame(hoge)
  tmp <- hoge[, c("seqnames", "s
  out_f <- sub(".bam", ".bed",
  out_f
  write.table(tmp, out_f, sep="
}
```

```
R Console
+ write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names
+ )
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kis01"
> list.files()
[1] "mapping_single_genome1.txt"
[2] "QuasR_log_146c2f3d4e64.txt"
[3] "ref_genome.fa"
[4] "ref_genome.fa.fai"
[5] "ref_genome.fa.md5"
[6] "ref_genome.fa.Rbowtie"
[7] "sample_RNAseq1.fa"
[8] "sample_RNAseq1_146c6c6d54aa.bam"
[9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
> |
```



Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

出力ファイル形式

- ゲノム上のどの位置にどのリードがマッピングされたか(トランスクリプトームの場合どの転写物配列上のどの位置にどのリードがマッピングされたか)を表す出力ファイル形式は複数あります。
 - SAM (Sequence Alignment/Map) format
 - SAMtools (Li et al., *Bioinformatics*, **25**: 2078-2079, 2009)
 - **BAM** (Binary Alignment/Map) format
 - SAMtools (Li et al., *Bioinformatics*, **25**: 2078-2079, 2009)
 - **BED** (Browser Extensible Data) format
 - BEDtools (Quinlan et al., *Bioinformatics*, **26**: 841-842, 2010)
 - ...

出力ファイル形式

①BAMファイルの②中身と、③BEDファイルの④中身。④BEDファイルの最小限の情報は、リードIDを含まないことがわかります。④こんな感じの情報が、②の中に含まれているということさえわかれば、わざわざBEDファイルを作成する必要はありません



chr1	11	45
chr2	1	35
chr2	16	50
chr3	1	35
chr3	3	37



```
> list.files()
[1] "mapping_single.bam"
[2] "QuasR_log_1.txt"
[3] "ref_genome.fa"
[4] "ref_genome.fa.idx"
[5] "ref_genome.fa.mds"
[6] "ref_genome.fa.Rbowtie"
[7] "sample_RNAseq1.fa"
[8] "sample_RNAseq1_146c6c6d54aa.bam"
[9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
> |
```



```
< .....ÿ·BC·P]NNAO É#á î·uë0!WAFt:nz7MEØ: °
unø,%_g $CM%hpbB9 [GÓIW á<= Hf4 eJ7E,yw =hQ@u5? öUî Y /Ñ9
#
C?Y0³A0w -y. 7'á §)Ô 0ú eóW50Æ@lsgÎ|:ùÀ,dĭ^ 0Ú ±
pÙÙQ %çDif 5QÓaöV0¥<7+-V% 97²Ø8¥f&xEE÷
ÿÇè8/w| Æì`u Ôsi@/y-Ô²¥- A=z öRÖ ïfT@éc%Hà {äyÉBánSÜ?²à Ù
*Útu/··uø ZcİÖ ·¶Æ ĩ- 4oË á ·· < .....ÿ·BC· ò=oÃ
à³Ö) R cI<0ÆuHÖ- ¶o ;W] òiký
bèzNhi( c;‡ ò=z9jhj ö÷F&p.Âpüö·ý@S|f¥` $0& $×
È ĩ~e ]ä%Tø)x-Jø ]&Ü->ôd
É!:i'ä8·nhÝç³0Q °6jyP-³%PI(a>çÆýQ³STD=öïí é §
öÓYhÜöðç! öë |· I_éBÓ^óÇ!bUFÄ eV~p¥P6(Yp ¶
g='Ýj&W èÆO-<0 ù d,·· < .....ÿ·BC· ······
```

出力ファイル形式

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装したRbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの一意に(unicely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションで挙動を確認したりできます。出力ファイルは、"*.bam", "*_QC.pdf", "ファイル"-「ディレ

chr1	11	45
chr2	1	35
chr2	16	50
chr3	1	35
chr3	3	37

を示しませんでした。ここでは、マッピングのオプションをいくつか変更するやり方を示します。尚、出力ファイルは、"*.bam", "*_QC.pdf", ディレクトリに移動し以下をコピー。

f genome.faへのマッピングの場合(mapping single genomel.txt): q1.faでマップされないのは計3リードです。2リード オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致。

1. サンプルデータ1
オプションを"-m 1" ("chr3_11_45"と"chr3_16_50"ではない(4番目の

```

out <- qAlign(in_r1, in_r2, alignmentparameter=param mapping)#マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time() #計算時間を計測するため
time_e - time_s #計算時間を表示(一番右側の数字。単位はsecond)
out #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out) #マッピング結果(alignment statistics)の表示。seqlength: リファレンス

#ファイルに保存(QCレポート用のpdfファイル作成)
out_f <- sub(".bam", "_QC.pdf", out@alignments[,1])#Ququality Controlレポートのpdfファイル名を作成した結果をc
qQCReport(out, pdffilename=out_f) #QCレポート結果をファイルに保存
out_f #ファイル名を表示してるだけです
    
```

```

#ファイルに保存(BED形式ファイル)
tmpfname <- out@alignments[,1] #ファイル名(in_f1の1列目に相当)をtmpfnameとして取り扱いたいだけです
for(i in 1:length(tmpfname)){ #サンプル数(ファイル数)分だけループを回す
  hoge <- readGAlignments(tmpfname[i]) #BAM形式ファイルを読み込んだ結果をhogeに格納(これはGAlignmentsオブジ:
  hoge <- as.data.frame(hoge) #データフレーム形式に変換
  tmp <- hoge[, c("seqnames", "start", "end")]#必要な列の情報のみ抽出した結果をtmpに格納
  out_f <- sub(".bam", ".bed", tmpfname[i])#BED形式ファイル名を作成した結果をout_fに格納
  out_f #ファイル名を表示してるだけです
  write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F)#tmpの中身を指定したファイルに保存
}
    
```

計8リードのうち、マップされなかったのは赤枠の3リード

使用オプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0ミスマッチで1か所にのみマップされるリードを出力

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGA chr1 11 45
>chr2
AGGGAGGGC chr2 1 35 TGGATGCTGAGTGGG
ACGCAGGTA chr2 16 50
CTCGGGTAT chr3 1 35 ATAGACACCTTGAGGAG
TGACGCCCTG chr3 3 37 GCATCATGAAGGGGCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

完全一致でも複数個所にマップされるために落とされたのは2リード

使用オプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0 mismatchesで1か所へのみマップされるリードを出力

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```


1か所にのみマップされるがミスマッチのため落とされたのは1リード

使用オプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0ミスマッチで1か所にのみマップされるリードを出力

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTC

Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

ディレクトリの作成

①現在作業中のディレクトリ(カレントディレクトリ)をpwdコマンドで確認。②その中身をls -lで確認

```
iu@bielinux[~/Desktop/mac_share]
① iu@bielinux[mac_share] pwd [ 2:59午後 ]
/home/iu/Desktop/mac_share
② iu@bielinux[mac_share] ls -l [ 2:59午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] [ 2:59午後 ]
```

ディレクトリの作成

①ディレクトリ作成コマンドmkdirで
mapping_kiso1というディレクトリを作成

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] pwd [ 2:59午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l [ 2:59午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] mkdir mapping_kiso1 [ 2:59午後 ]
iu@bielinux[mac_share] █ [ 3:05午後 ]
```



ディレクトリの作成

①lsで確認。確かに②mapping_kiso1というディレクトリが作られています

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] pwd [ 2:59午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l [ 2:59午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] mkdir mapping_kiso1 [ 2:59午後 ]
iu@bielinux[mac_share] ls -l [ 3:05午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 0 5月 14 15:05 mapping_kiso1
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] [ 3:07午後 ]
```



mapping_kiso1に移動

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mac_share] pwd [ 2:59午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l [ 2:59午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] mkdir mapping_kiso1 [ 2:59午後 ]
iu@bielinux[mac_share] ls -l [ 3:05午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 0 5月 14 15:05 mapping_kiso1
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
① iu@bielinux[mac_share] cd mapping_kiso1 [ 3:07午後 ]
iu@bielinux[mapping_kiso1] [ 3:14午後 ]
```


①pwdで作業ディレクトリを確認(カレントディレクトリを表示)。②確かに移動できていることがわかります

pwdで確認

```

iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mac_share] pwd [ 2:59午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l [ 2:59午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] mkdir mapping_kiso1 [ 2:59午後 ]
iu@bielinux[mac_share] ls -l [ 3:05午後 ]
total 801638
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
-rwxrwxrwx 1 iu iu 636707 5月 11 13:29 hoge1.fastq
drwxrwxrwx 1 iu iu 0 5月 14 15:05 mapping_kiso1
drwxrwxrwx 1 iu iu 4096 5月 10 17:32 result
iu@bielinux[mac_share] cd mapping_kiso1 [ 3:07午後 ]
iu@bielinux[mapping_kiso1] pwd [ 3:14午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] [ 3:15午後 ]

```



wgetでダウンロード

①pwdと②ls。③wgetコマンドを用いて、
まずはマップされる側のリファレンス配列
ファイル(ref_genome.fa)をダウンロード

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
① iu@bielinux[mapping_kiso1] pwd [ 3:28午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
② iu@bielinux[mapping_kiso1] ls -l [ 3:28午後 ]
total 0
③ iu@bielinux[mapping_kiso1] wget http://www.iu.a.u-tokyo.ac.jp/~
kadota/R_seq/ref_genome.fa
```

wgetでダウンロード

③ wgetを実行すると、赤枠のような感じでダウンロードの途中経過が表示されます。無事ダウンロードが終わると、④のような感じでsavedというメッセージが表示されて、⑤コマンド入力待ち状態になります

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
② iu@bielinux[mapping_kiso1] ls -l
total 0
③ iu@bielinux[mapping_kiso1] wget http://www.iu.a.u-tokyo.ac.jp/~
kadota/R_seq/ref_genome.fa
--2018-05-14 15:31:20-- http://www.iu.a.u-tokyo.ac.jp/~kadota/
R_seq/ref_genome.fa
Resolving www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)... 13
3.11.224.25
Connecting to www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)|1
33.11.224.25|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 590 [text/plain]
Saving to: 'ref_genome.fa'

100%[=====>] 590          ---K/s   in 0s

2018-05-14 15:31:25 (39.7 MB/s) - 'ref_genome.fa' saved [590/59
0]
④
iu@bielinux[mapping_kiso1] █ [ 3:31午後 ]
⑤
```


clearで画面リフレッシュ

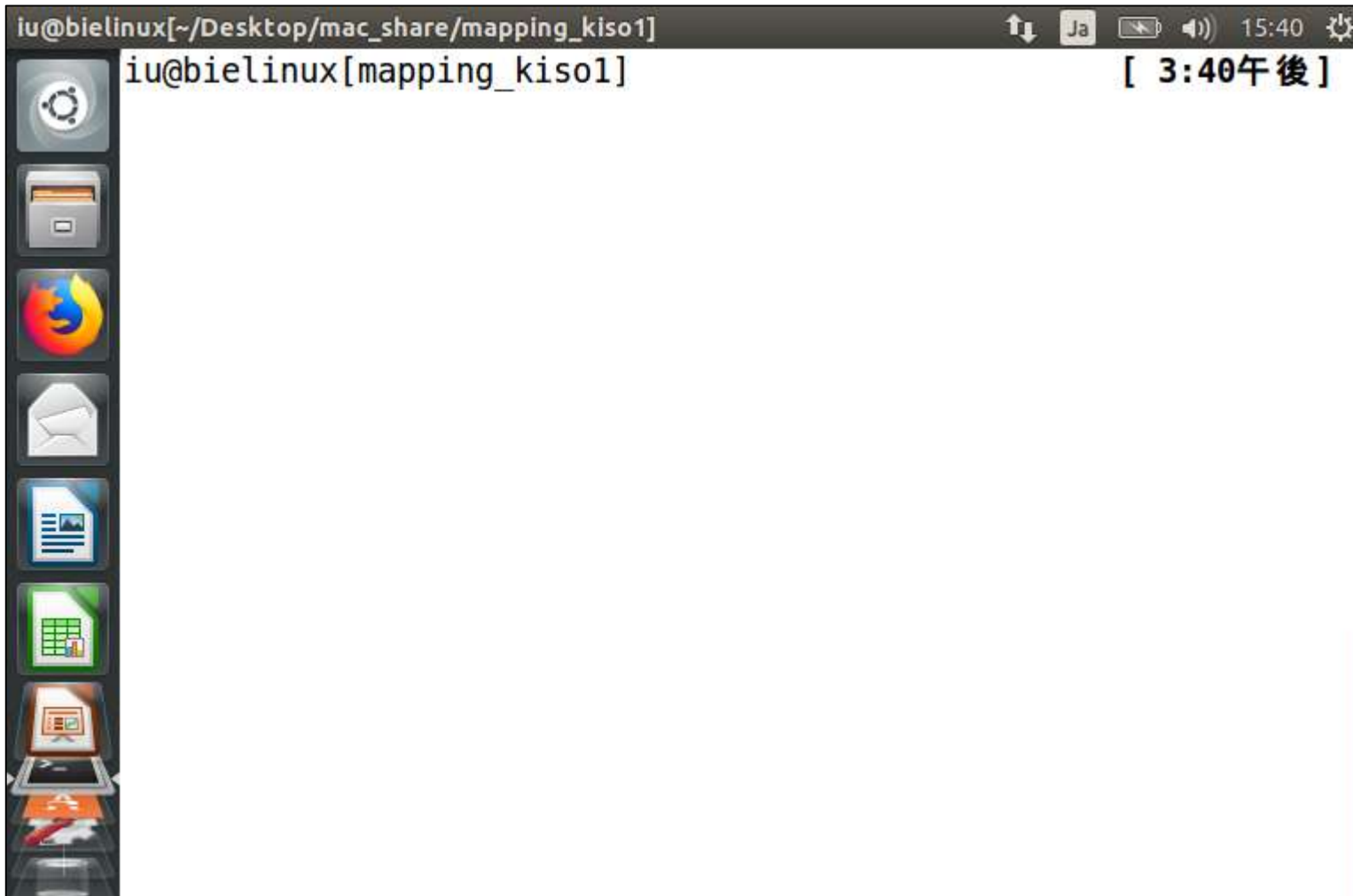
```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1] [ 3:28午後 ]
iu@bielinux[mapping_kiso1] ls -l
total 0
iu@bielinux[mapping_kiso1] wget http://www.iu.a.u-tokyo.ac.jp/~
kadota/R_seq/ref_genome.fa
--2018-05-14 15:31:20-- http://www.iu.a.u-tokyo.ac.jp/~kadota/
R_seq/ref_genome.fa
Resolving www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)... 13
3.11.224.25
Connecting to www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)|1
33.11.224.25|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 590 [text/plain]
Saving to: 'ref_genome.fa'

100%[=====>] 590          ---K/s   in 0s

2018-05-14 15:31:25 (39.7 MB/s) - 'ref_genome.fa' saved [590/59
0]

① iu@bielinux[mapping_kiso1] clear [ 3:31午後 ]
```

clearで画面リフレッシュ



lsで確認

- ①lsで確認。確かにwgetでダウンロードした
- ②ref_genome.faが存在することがわかります

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:40午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:40午後 ]
total 1
-rwxrwxrwx 1 iu iu 590  9月 29  2013 ref_genome.fa [ 3:40午後 ]
iu@bielinux[mapping_kiso1] █
```

①headは、最初の10行分(デフォルトオプション)を表示するコマンド

headで確認

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:40午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:40午後 ]
total 1
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
iu@bielinux[mapping_kiso1] head ref_genome.fa [ 3:40午後 ]
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
iu@bielinux[mapping_kiso1] [ 3:45午後 ]
```

ダウンロード2

②マップする側の仮想RNA-seqファイル(sample_RNAseq1.fa)をダウンロード。wget実行時に-qオプションをつけることで途中経過を表示させない(quiet)ようにすることができる

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
total 1
-rwxrwxrwx 1 iu iu 590  9月 29  2013 ref_genome.fa
iu@bielinux[mapping_kiso1] wget -q http://www.iu.a.u-tokyo.ac.jp/~kadota/R_seq/sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] █
```



①lsで確認。②ダウンロードできているようですね

lsで確認

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:56午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:56午後 ]
total 1
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
iu@bielinux[mapping_kiso1] wget -q http://www.iu.a.u-tokyo.ac.jp/~kadota/R_seq/sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] ls -l [ 3:56午後 ]
total 2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] █ [ 3:57午後 ]
```



headの-nオプション

①head -n 4で最初の4行分を表示。確かにsample_RNAseq1.faの中身は(少なくとも最初の4行分)は妥当ですね

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:56午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:56午後 ]
total 1
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
iu@bielinux[mapping_kiso1] wget -q http://www.iu.a.u-tokyo.ac.jp/~kadota/R_seq/sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] ls -l [ 3:56午後 ]
total 2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] head -n 4 sample_RNAseq1.fa [ 4:12午後 ]
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
```



Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

Bowtie

QuasRパッケージ内部で用いられているのは、①bowtieプログラムだが…

Genome Biol. 2009;10(3):R25. doi: 10.1186/gb-2009-10-3-r25. Epub 2009 Mar 4.

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.

Langmead B¹, Trapnell C, Pop M, Salzberg SL.

⊕ Author information

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source (<http://bowtie.cbcb.umd.edu>).



Comment in

The need for speed. [*Genome Biol.* 2009]

PMID: 19261174 PMCID: [PMC2690996](https://pubmed.ncbi.nlm.nih.gov/PMC2690996/) DOI: [10.1186/gb-2009-10-3-r25](https://doi.org/10.1186/gb-2009-10-3-r25)

Bowtie (Langmead et al., *Genome Biol.*, 10: R25, 2009)

Bowtie2

Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923.

Fast gapped-read alignment with Bowtie 2.

Langmead B¹, Salzberg SL.

 Author information

Abstract

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy.

PMID: 22388286 PMCID: [PMC3322381](#) DOI: [10.1038/nmeth.1923](#)

①現在主に用いられているのはBowtie 2。bowtieがver. 1でbowtie2がver. 2という位置づけ。bowtieとbowtie2とではオプションの指定方法なども異なる。bowtieのオプションを丁寧に説明してQuasRと同じ結果になることにこだわってもしようがないので、bowtie2で行います。36 bpというこのリード長の場合はbowtieのほうがいいのだと思いますが、2018年現在はこのような長さのリードを扱うほうがむしろ稀なのでその意味でもやはりbowtie2



リファレンス配列の前処理

プログラムを高速に実行するための前処理として、リファレンス配列(`ref_genome.fa`)を`bowtie2-build`プログラムにかける。Bowtieプログラムの場合は`bowtie-build`です。BLASTをコマンドライン上で実行する際に、データベース側の配列の前処理として行う`makeblastdb`と同じようなものです

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
total 2
-rwxrwxrwx 1 iu iu 590  9月 29  2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月  1  2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2-build ref_genome.fa pigya
```



リファレンス配列の前処理

①の部分は任意の文字。ここではpigyaとしたが、ref_genome.faを入力とする場合は、私なら通常はref_genomeとします

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 4:38午後]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 4:38午後]
total 2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2-build ref_genome.fa pigya
```



画面がざっと流れてこんな感じになります。数秒程度

bowtie2-build実行

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
ftabChars: 10
eftabLen: 20
eftabSz: 80
ftabLen: 1048577
ftabSz: 4194308
offsLen: 34
offsSz: 136
lineSz: 64
sideSz: 64
sideBwtSz: 48
sideBwtLen: 192
numSides: 3
numLines: 3
ebwtTotLen: 192
ebwtTotSz: 192
color: 0
reverse: 1
Total time for backward call to driver() for mirror index: 00:00:00
iu@bielinux[mapping_kiso1] [ 4:39午後 ]
```

bowtie2-build実行結果

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
sideBwtLen: 192
numSides: 3
numLines: 3
ebwtTotLen: 192
ebwtTotSz: 192
color: 0
reverse: 1
Total time for backward call to driver() for mirror index: 00:00
0:00
iu@bielinux[mapping_kiso1] ls -l [ 4:39午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] [ 6:20午後 ]
```

bowtie2-build実行結果

①この部分が任意につけたpigyaという文字列。bowtie2-buildプログラムのマニュアルには、basenameと書かれている。baseの意味がわかるのではないだろうか

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
sideBwtLen: 192
numSides: 3
numLines: 3
ebwtTotLen: 192
ebwtTotSz: 192
color: 0
reverse: 1
Total time for backward call to driver() for mirror index: 00:00:00
iu@bielinux[mapping_kiso1] ls -l [ 4:39午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] [ 6:20午後 ]
```



Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

bowtie2実行コマンド

①bowtie2実行コマンド。②マップされる側の情報。③マップする側の情報。-fはsingle-endの場合に指定するオプション

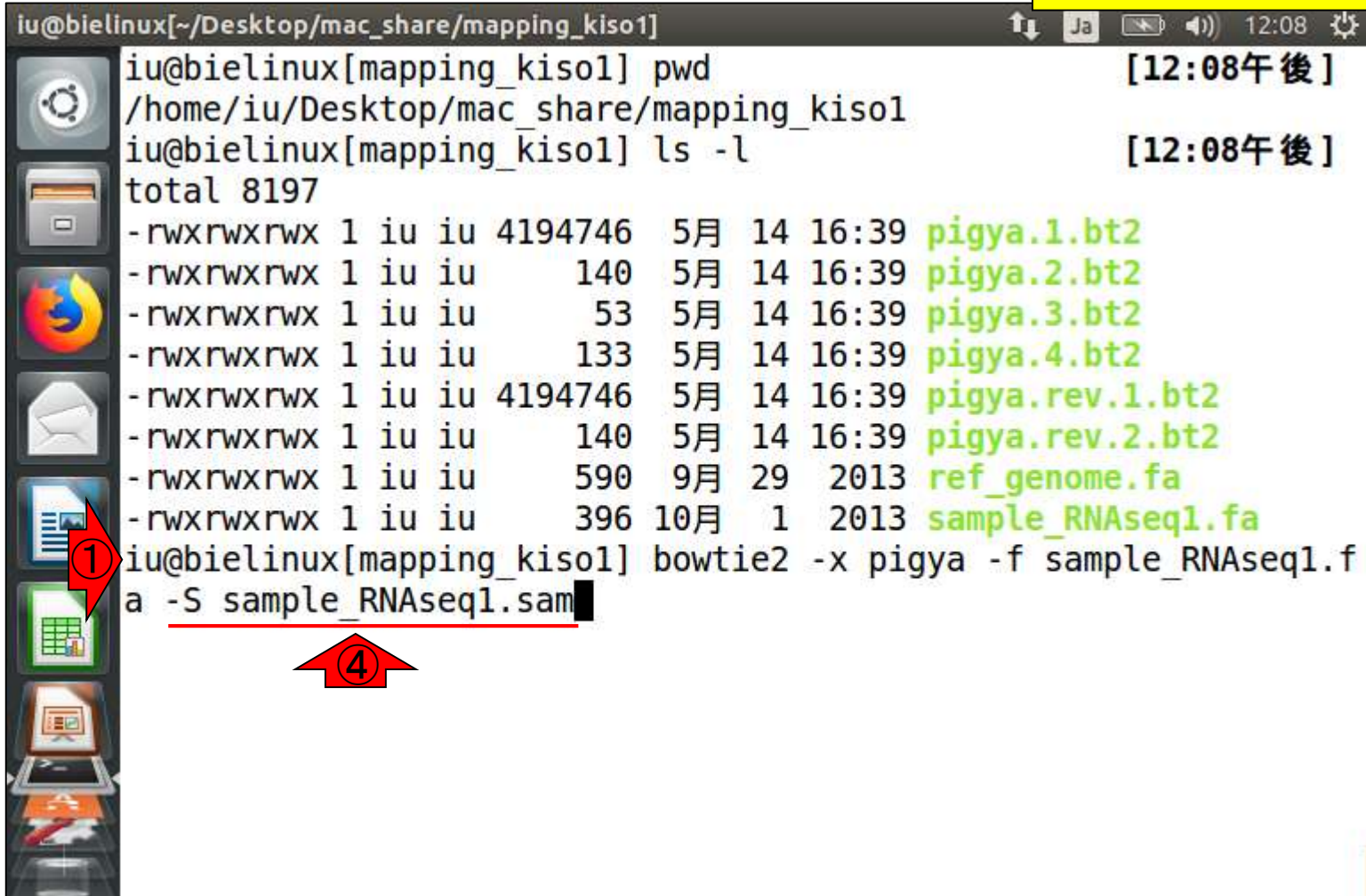
```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
```



bowtie2実行コマンド

④SAM形式の出力ファイル名を指定するところ。out.samなどでもよいが、入力ファイルと対応付けてsample_RNAseq1.samとした

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
```



bowtie2実行結果

①リターンキーを押したところ。一瞬で終わります。③入力は8リードでsingle-endなので、unpairedと書かれているのは妥当

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
① iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
 8 (100.00%) were unpaired; of these:
 0 (0.00%) aligned 0 times
 3 (37.50%) aligned exactly 1 time
 5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] [ 3:08午後 ]
```


bowtie2実行結果

- ①0回マップされたリードは0個という意味なので、マップされなかったリードは0個。
- ②つまり全リードがマップされたということ

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    ① 0 (0.00%) aligned 0 times
      3 (37.50%) aligned exactly 1 time
      5 (62.50%) aligned >1 times
    ② 100.00% overall alignment rate
iu@bielinux[mapping_kiso1] █
```

[3:08午後]

[3:08午後]

bowtie2実行結果

①1回だけマップされたリードは3個(37.50%)。これは1か所
にのみマップされたリードと解釈すればよい。②2回以上(複
数個所に)マップされたリードは5個(62.50%)。③8リード中8
個がマップされたので、マップ率(alignment rate)は100%

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    ① 3 (37.50%) aligned exactly 1 time
    ② 5 (62.50%) aligned >1 times
③ 100.00% overall alignment rate
iu@bielinux[mapping_kiso1] [ 3:08午後 ]
```


lsで確認

①ls -l。②確かに拡張子.sam
のSAMファイルができています

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8199
-rwxrwxrwx 1 iu iu 4194746  5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu    140  5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu    53  5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu   133  5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746  5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu   140  5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu   590  9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu   396 10月  1 2013 sample_RNAseq1.fa
-rwxrwxrwx 1 iu iu  1616  5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] [ 3:09午後 ]
```



Contents

- Quality Control (QC)の続き
 - 全体像のおさらいとQCの位置づけ
 - FastQCとFaQCs、FaQCsの実行
 - FaQCs実行結果ファイルに対してFastQCを実行
 - 課題 (FaQCs実行前後の比較)
 - RでQC: ShortReadでクオリティフィルタリング、qrrqcでクオリティチェック
- マッピング (アラインメント)
 - マップする側とされる側のファイル
 - QuasRでマッピング (内部的にRbowtieパッケージを利用)
 - 出力ファイル形式、使用オプションと結果の解釈
 - Bio-Linux環境でbowtie2を使ってマッピング: 準備、前処理、実行
 - SAMファイルの解説

WCで行数を確認

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
8 reads; of these:
 8 (100.00%) were unpaired; of these:
 0 (0.00%) aligned 0 times
 3 (37.50%) aligned exactly 1 time
 5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8199
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
① iu@bielinux[mapping_kiso1] wc sample_RNAseq1.sam [ 3:09午後 ]
 15 188 1616 sample_RNAseq1.sam
② iu@bielinux[mapping_kiso1] [ 3:13午後 ]
```

sample_RNAseq1.sam

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:27午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
① iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] [ 3:27午後 ]
```

sample_RNAseq1.sam

①SAMファイルの最初の8行分を表示(いくつか試した結果、画面上に表示できるのが8行分だけでした)

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1] [ 3:27午後 ]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616  5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] head -n 8 sample_RNAseq1.sam
@HD      VN:1.0   S0:unsorted
@SQ      SN:chr1  LN:48
@SQ      SN:chr2  LN:160
@SQ      SN:chr3  LN:100
@SQ      SN:chr4  LN:123
@SQ      SN:chr5  LN:100
@PG      ID:bowtie2  PN:bowtie2  VN:2.2.4  CL:"/usr/bin/../../../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
chr1_11_45  0  chr1  11  42  35M  *  0
0  CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT  IIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  AS:i:0  XN:i:0  XM:i:0  X0:i:0  XG:i:0N
M:i:0  MD:Z:35  YT:Z:UU
iu@bielinux[mapping_kiso1] [ 3:32午後 ]
```


①赤枠部分が最初の7行分に相当する。これらは@から始まる行であり、ヘッダー行と呼ばれる

sample_RNAseq1.sam

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd [ 3:27午後 ]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] head -n 8 sample_RNAseq1.sam
@HD VN:1.0 S0:unsorted
@SQ SN:chr1 LN:48
@SQ SN:chr2 LN:160
@SQ SN:chr3 LN:100
@SQ SN:chr4 LN:123
@SQ SN:chr5 LN:100
@PG ID:bowtie2 PN:bowtie2 VN:2.2.4 CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
chr1_11_45 0 chr1 11 42 35M * 0
0 CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT IIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIII AS:i:0 XN:i:0 XM:i:0 X0:i:0 XG:i:0N
M:i:0 MD:Z:35 YT:Z:UU
iu@bielinux[mapping_kiso1] [ 3:32午後 ]
```



sample_RNAseq1.sam

①この場合のSAMファイルの2-6行目は、マップされる側のリファレンス配列の情報が書き込まれている。リファレンスの配列数が増えるとヘッダー部分の行数も増えるのだろう

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l
-rwxrwxrwx 1 iu iu 1616 5月 10 14:00 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] head -n 6 sample_RNAseq1.sam
@HD      VN:1.0  SO:unsorted
@SQ      SN:chr1 LN:48
@SQ      SN:chr2 LN:160
@SQ      SN:chr3 LN:100
@SQ      SN:chr4 LN:123
@SQ      SN:chr5 LN:100
@PG      ID:bowtie2      PN:bowtie2
r/bin/./lib/bowtie2/bin/bowtie2
igya -f sample_RNAseq1.fa -S sample_RNAseq1.sam
chr1_11_45 0 chr1
0 CGCTTACGAGATCAGGCTAA
IIIIIIIIIIIIIIIIIIIIII AS:i:1
M:i:0 MD:Z:35 YT:Z:UU
iu@bielinux[mapping_kiso1]
```

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTAACTAGTTT
```



Excelで見ると...

Excelで全体像を表示させた結果。①がリードの塩基配列情報、②がリードの1文字表記のクオリティスコア情報。マップする側のsample_RNAseq1.faはFASTAファイルでありFASTQであり、クオリティスコアは本来存在しない。しかしSAM形式に合わせるべく、クオリティスコア情報がない場合は、クオリティスコア40に相当するIを与えている

```
@HD VN:1.0 SO:unsorted
@SQ SN:chr1 LN:48
@SQ SN:chr2 LN:160
@SQ SN:chr3 LN:100
@SQ SN:chr4 LN:123
@SQ SN:chr5 LN:100
@PG ID:bowt2 FN:bowt2 VN:2.0 CL: "/usr/bin/./lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
chr1_11_45 0 chr1 11 42 35M * 0 0 CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr2_16_50 0 chr2 16 42 35M * 0 0 TATCTATGGCCATAAAACATAGACACCTTGAGGAG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr2_1_35 0 chr2 1 42 35M * 0 0 AGGGAGGGGGTCCAGTATCTATGGCCATAAAACAT IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr3_11_45 0 chr5 11 1 35M * 0 0 TTTCCCGCTTGCAGGAATCGTGTCTGAGTTGGTATA IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr3_15_49 0 chr3 15 1 35M * 0 0 CCCGCTTGCAGGAATCGTGTCTGAGTTGGTATA CAGG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr3_3_37 0 chr3 3 31 35M * 0 0 GGGGACTATTCCCGCTTGCAGGAATCGTGTCTGAG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XS:i:-6 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr3_1_35 0 chr3 1 35 35M * 0 0 GGGGGGACTATTCCCGCTTGCAGGAATCGTGTCTG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:0 XS:i:-12 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:2:35 YT:2:UU
chr5_1_35 0 chr5 1 16 35M * 0 0 GCCTGGTCTATTCCCGCTTGCAGGAATCGTGTCTG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII AS:i:-6 XS:i:-18 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:2:3031 YT:2:UU
```



Excelで全体像を表示させた結果。リードの①塩基配列情報と②クオリティスコア情報部分の幅を小さくしました

全体を拡大

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:100																	
@PG	ID:bowt:	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNA															
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_16_50	0	chr2	16	42	35M	*	0	0	TAIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_1_35	0	chr2	1	42	35M	*	0	0	ACIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCIIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCIIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCIIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU		



リードの並びは

①入力ファイルのリードの並びと、②出力ファイルのリードの並びは同じですね

@HD	VN:1.0	SO:unsorted																		
@SQ	SN:chr1	LN:48																		
@SQ	SN:chr2	LN:160																		
@SQ	SN:chr3	LN:100																		
@SQ	SN:chr4	LN:123																		
@SQ	SN:chr5	LN:100																		
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align																
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0									
chr2_16_50	0	chr2	16	42	35M	*	0	0	TATIAS:i:0	XN:i:0	XM:i:0									
chr2_1_35	0	chr2	1	42	35M	*	0	0	AGGIIAS:i:0	XN:i:0	XM:i:0									
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTTIIAS:i:0	XS:i:0	XN:i:0									
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCGIIAS:i:0	XS:i:0	XN:i:0									
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCGIIAS:i:0	XS:i:-6	XN:i:0									
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCGIIAS:i:0	XS:i:-12	XN:i:0									
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCGIIAS:i:-6	XS:i:-18	XN:i:0									

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```



Bowtie2実行結果

Bowtie2実行結果のおさらい。①1回だけマップされたリードは3個(37.50%)。これは1か所にのみマップされたリードのことでした。この3リードの結果が…

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    ① 3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] [ 3:08午後 ]
```

Bowtie2実行結果

@HD	VN:1.0	SO:unsorted																		
@SQ	SN:chr1	LN:48																		
@SQ	SN:chr2	LN:160																		
@SQ	SN:chr3	LN:100																		
@SQ	SN:chr4	LN:123																		
@SQ	SN:chr5	LN:100																		
@PG	ID:bowt2	PN:bowt2	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align																
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0									
chr2_16_50	0	chr2	16	42	35M	*	0	0	TATIAS:i:0	XN:i:0	XM:i:0									
chr2_1_35	0	chr2	1	42	35M	*	0	0	AGIIAS:i:0	XN:i:0	XM:i:0									
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTIIAS:i:0	XS:i:0	XN:i:0									
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCIIAS:i:0	XS:i:0	XN:i:0									
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCIIAS:i:0	XS:i:-6	XN:i:0									
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCIIAS:i:0	XS:i:-12	XN:i:0									
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCIIAS:i:-6	XS:i:-18	XN:i:0									

sample_RNAseq1.fa - メモ帳

② ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```

>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC

```


Bowtie2実行結果のおさらい。②2回以上(複数個所に)マップされたリードは5個(62.50%)。この5リードの結果が

Bowtie2実行結果

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] [ 3:08午後 ]
```



Bowtie2実行結果

@HD	VN:1.0	SO:unsorted																		
@SQ	SN:chr1	LN:48																		
@SQ	SN:chr2	LN:160																		
@SQ	SN:chr3	LN:100																		
@SQ	SN:chr4	LN:123																		
@SQ	SN:chr5	LN:100																		
@PG	ID:bowt2	PN:bowt2	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align																
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0									
chr2_16_50	0	chr2	16	42	35M	*	0	0	TATIAS:i:0	XN:i:0	XM:i:0									
chr2_1_35	0	chr2	1	42	35M	*	0	0	AGIIAS:i:0	XN:i:0	XM:i:0									
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTIIAS:i:0	XS:i:0	XN:i:0									
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCIIAS:i:0	XS:i:0	XN:i:0									
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCIAS:i:0	XS:i:-6	XN:i:0									
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCIAS:i:0	XS:i:-12	XN:i:0									
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCIAS:i:-6	XS:i:-18	XN:i:0									



```

sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGTCTATTTCCCGCTTGCAGGAATCGTGTC
    
```

Bowtie2実行結果

①1回だけ(1か所にのみ)マップされたリードと、②2回以上(複数個所に)マップされたリードの結果を、SAMファイルからどう見分けるのか?

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1]
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l [ 3:08午後 ]
total 8197
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2
-rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa
-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    ① 3 (37.50%) aligned exactly 1 time
    ② 5 (62.50%) aligned >1 times
100.00% overall alignment rate
iu@bielinux[mapping_kiso1] [ 3:08午後 ]
```


複数個所の結果はXS

2回以上(複数個所に)マップされたリードの結果は、①XS:という結果が余分に含まれています。もしこの部分を削除して左詰めすれば...

@HD	VN:1.0	SO:unsorted																
@SQ	SN:chr1	LN:48																
@SQ	SN:chr2	LN:160																
@SQ	SN:chr3	LN:100																
@SQ	SN:chr4	LN:123																
@SQ	SN:chr5	LN:100																
@PG	ID:bowt2	PN:bowt2	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNA														
chr1_11_45	0	chr1	11	42	35M	*	0	0	CG	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	IIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	IIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	IIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU



複数個所の結果はXS

こんな具合になって、①ものの見事に同一列内のXN:やXM:などの文字が揃います

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:100																	
@PG	ID:bowt:	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/															
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_16_50	0	chr2	16	42	35M	*	0	0	TIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_1_35	0	chr2	1	42	35M	*	0	0	ACIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCGIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCGIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCGIIAS:i:-6	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G3	YT:Z:UU			



つまり、①赤枠の情報が余分にあるので、②その右側部分がずれたように見えていたということです

複数個所の結果はX

@HD	VN:1.0	SO:unsorted																		
@SQ	SN:chr1	LN:48																		
@SQ	SN:chr2	LN:160																		
@SQ	SN:chr3	LN:100																		
@SQ	SN:chr4	LN:123																		
@SQ	SN:chr5	LN:100																		
@PG	ID:bowt:	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s	--wrapper	basic-0	-x	pigya	-f	sample_RNA										
chr1_11_45	0	chr1	11	42	35M	*	0	0	CG	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU			
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	IIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	IIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	IIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU		



SAMファイルの解説

SAMファイルは、タブ区切りテキスト形式 (tab-delimited text file format) ですので、ヒトの目で判読できます。行頭が@から始まるのがヘッダー部分。この場合は7行ですね

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:100																	
@PG	ID:bowt	PN:bowt	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample RNAs															
chr1_11_45	0	chr1	11	42	35M	*	0	0	CG	I	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	I	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	I	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	I	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	I	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	I	AS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	I	AS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	I	AS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

grep “@”

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1] [ 2:37午後 ]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] grep "@" sample_RNAseq1.sam
@HD      VN:1.0  S0:unsorted
@SQ      SN:chr1 LN:48
@SQ      SN:chr2 LN:160
@SQ      SN:chr3 LN:100
@SQ      SN:chr4 LN:123
@SQ      SN:chr5 LN:100
@PG      ID:bowtie2      PN:bowtie2      VN:2.2.4      CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
iu@bielinux[mapping_kiso1] [ 2:37午後 ]
```


grep -c "@"

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1] [ 2:37午後 ]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616  5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] grep "@" sample_RNAseq1.sam
@HD      VN:1.0   S0:unsorted
@SQ      SN:chr1  LN:48
@SQ      SN:chr2  LN:160
@SQ      SN:chr3  LN:100
@SQ      SN:chr4  LN:123
@SQ      SN:chr5  LN:100
@PG      ID:bowtie2  PN:bowtie2  VN:2.2.4  CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
iu@bielinux[mapping_kiso1] grep -c "@" sample_RNAseq1.sam
7
iu@bielinux[mapping_kiso1] [ 2:41午後 ]
```

grep -c "^@"

```
iu@bielinux[~/Desktop/mac_share/mapping_kiso1] [ 2:37午後 ]
iu@bielinux[mapping_kiso1] pwd
/home/iu/Desktop/mac_share/mapping_kiso1
iu@bielinux[mapping_kiso1] ls -l sample_RNAseq1.sam
-rwxrwxrwx 1 iu iu 1616  5月 15 15:08 sample_RNAseq1.sam
iu@bielinux[mapping_kiso1] grep "@" sample_RNAseq1.sam
@HD      VN:1.0   S0:unsorted
@SQ      SN:chr1  LN:48
@SQ      SN:chr2  LN:160
@SQ      SN:chr3  LN:100
@SQ      SN:chr4  LN:123
@SQ      SN:chr5  LN:100
@PG      ID:bowtie2  PN:bowtie2  VN:2.2.4   CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
iu@bielinux[mapping_kiso1] grep -c "@" sample_RNAseq1.sam
7
iu@bielinux[mapping_kiso1] grep -c "^@" sample_RNAseq1.sam
7
iu@bielinux[mapping_kiso1] [ 2:43午後 ]
```

SAMファイルの解説

@HD	VN:1.0	SO:unsorted															
@SQ	SN:chr1	LN:48															
@SQ	SN:chr2	LN:160															
@SQ	SN:chr3	LN:100															
@SQ	SN:chr4	LN:123															
@SQ	SN:chr5	LN:100															
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAseq													
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGTIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TTTTIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	ACTTIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTTTIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCCTIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCTTIIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCTTIIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCTTIIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

SAMファイルの解説

- ① マップする側のリードID (クエリー名; QNAME)
- ② フラグ情報 (FLAG)。single-endの場合は常に0?!
- ③ マップされる側のリファレンス配列名 (RNAME)
- ④ マップされた領域の左端の位置 (POS)
- ⑤ マッピングのクオリティ (MAPQ)。高いほどよい。Phredスコアと同じような数式 ($-10 \times \log_{10}(p)$) で、 p はそのマップ位置が間違っている確率。例えば、間違っている確率が1% ($p = 0.01$) なら、ここの値は20になる。また、間違っている確率が70%と怪しい結果 ($p = 0.7$) の場合は1.55となる。従って、赤枠のマッピング結果は⑤の値が1なので信頼度が相当低いと解釈する

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:10																	
@PG	ID:bowt2	PN:bowtie2	PP:1.2.1	CM:"	/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s	--wrapper	basic-0	-x	pigya	-f	sample_RNAs								
chr1_11_45	0	chr1	11	42	35M	*	0	0	CG	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	IIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU		
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	IIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	IIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	IIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	IIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU	

SAMファイルの解説

⑥どのようなマッピング状況かを示すCIGAR string。ここでは全て35Mとなっているが、これはリードの35塩基が全て一致(Match)したということを意味している。InsertionやDeletionを意味するIやDなどが見られることもある。

⑦7-9列目は、入力がpaired-endのときに、ペアのもう片方のマッピング結果が記される。このデータはsingle-endなので、*や0で埋められる結果となっている

Header	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8	Field 9	Field 10	Field 11	Field 12	Field 13	Field 14	Field 15	Field 16	Field 17
@HD	VN:1.0	SO:unsorted															
@SQ	SN:chr1	LN:48															
@SQ	SN:chr2	LN:160															
@SQ	SN:chr3	LN:100															
@SQ	SN:chr4	LN:123															
@SQ	SN:chr5	LN:100															
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAS													
chr1_11_45	0	chr1	11	42	35M	*	0	0	CGIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TATIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	ACIIAS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TTIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CCIIAS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GCTIAS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GCTIAS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GCTIAS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU



SAMファイルの解説

講義日程 (平成30年度)

1. 平成30年05月08日

講義資料PDF

.gff3ファイル (約1.3MB)

.faファイル (約2.2MB)

(Rで)塩基配列解析

(Rで)マイクロアレイデータ解析

plasmid1.gff3(課題用)

plasmid2.gff3(課題用)

de Lannoy et al., F1000Res., 2017

Garalde et al., Nat Methods, 2018

RNACocktail : Sahraeian et al., Nat Commun., 2017

2. 平成30年05月15日

講義資料PDF(約5MB; 2018.05.11版)

(Rで)塩基配列解析

DRR000031sub.fastq

RNA-QC-chain : Zhou et al., BMC Genomics, 2018

Biostar : Parnell et al., PLoS Comput Biol., 2011

FastQC

DRR000031sub_fastqc.html

DRR000031_fastqc.html(課題用)

report.html(qrqcを用いたQC結果)

3. 平成30年05月22日

講義資料PDF(約5MB; 2018.05.17版)

(Rで)塩基配列解析

RNACocktail : Sahraeian et al., Nat Commun., 2017

Kraken : Davis et al., Methods, 2013

Lowe et al., PLoS Comput Biol., 2017

FaQCs : Lo and Chain, BMC Bioinformatics, 2014

FaQCs実行結果のQC.stats.txt

FaQCs実行結果のQC_qc_report.pdf

FastQC

QC.unpaired.trimmed_fastqc.html(課題用)

ShortRead : Morgan et al., Bioinformatics, 2009

Bioconductor : Gentleman et al., Genome Biol., 2004

Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013

report.html(qrqcを用いたQC結果)

QuasR(Windows版あり) : Gaidatzis et al., Bioinformatics, 2015

Bowtie : Langmead et al., Genome Biol., 2009

Bowtie2 : Langmead and Salzberg, Nat. Methods, 2012

sample_RNAseq1.sam(Bowtie2の実行結果SAMファイル)

Sequence Alignment/Map Format Specification

4. 平成30年05月29日

3. 平成30年05月22日

講義資料PDF(約5MB; 2018.05.17版)

(Rで)塩基配列解析

RNACocktail : Sahraeian et al., Nat Commun., 2017

Kraken : Davis et al., Methods, 2013

Lowe et al., PLoS Comput Biol., 2017

FaQCs : Lo and Chain, BMC Bioinformatics, 2014

FaQCs実行結果のQC.stats.txt

FaQCs実行結果のQC_qc_report.pdf

FastQC

QC.unpaired.trimmed_fastqc.html(課題用)

ShortRead : Morgan et al., Bioinformatics, 2009

Bioconductor : Gentleman et al., Genome Biol., 2004

Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 2013

report.html(qrqcを用いたQC結果)

QuasR(Windows版あり) : Gaidatzis et al., Bioinformatics, 2015

Bowtie : Langmead et al., Genome Biol., 2009

Bowtie2 : Langmead and Salzberg, Nat. Methods, 2012

sample_RNAseq1.sam(Bowtie2の実行結果SAMファイル)

Sequence Alignment/Map Format Specification

①