2018.05.19版

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、qrqc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

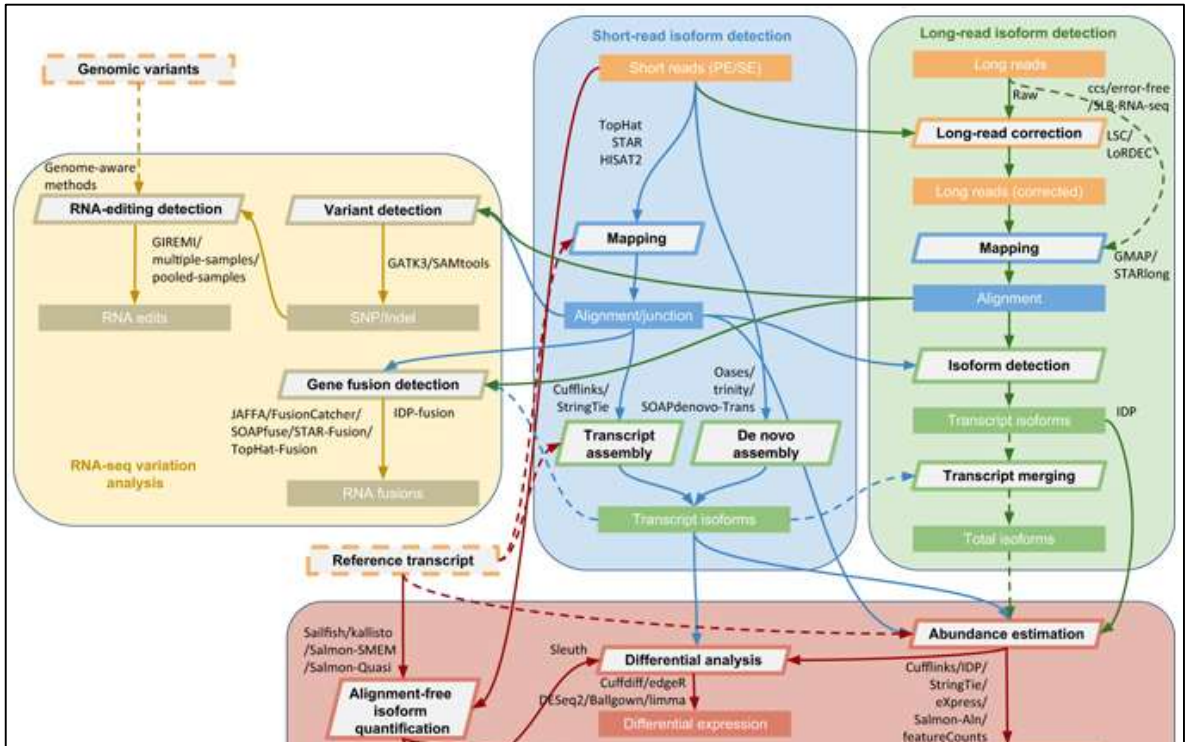
全体像のおさらい

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）

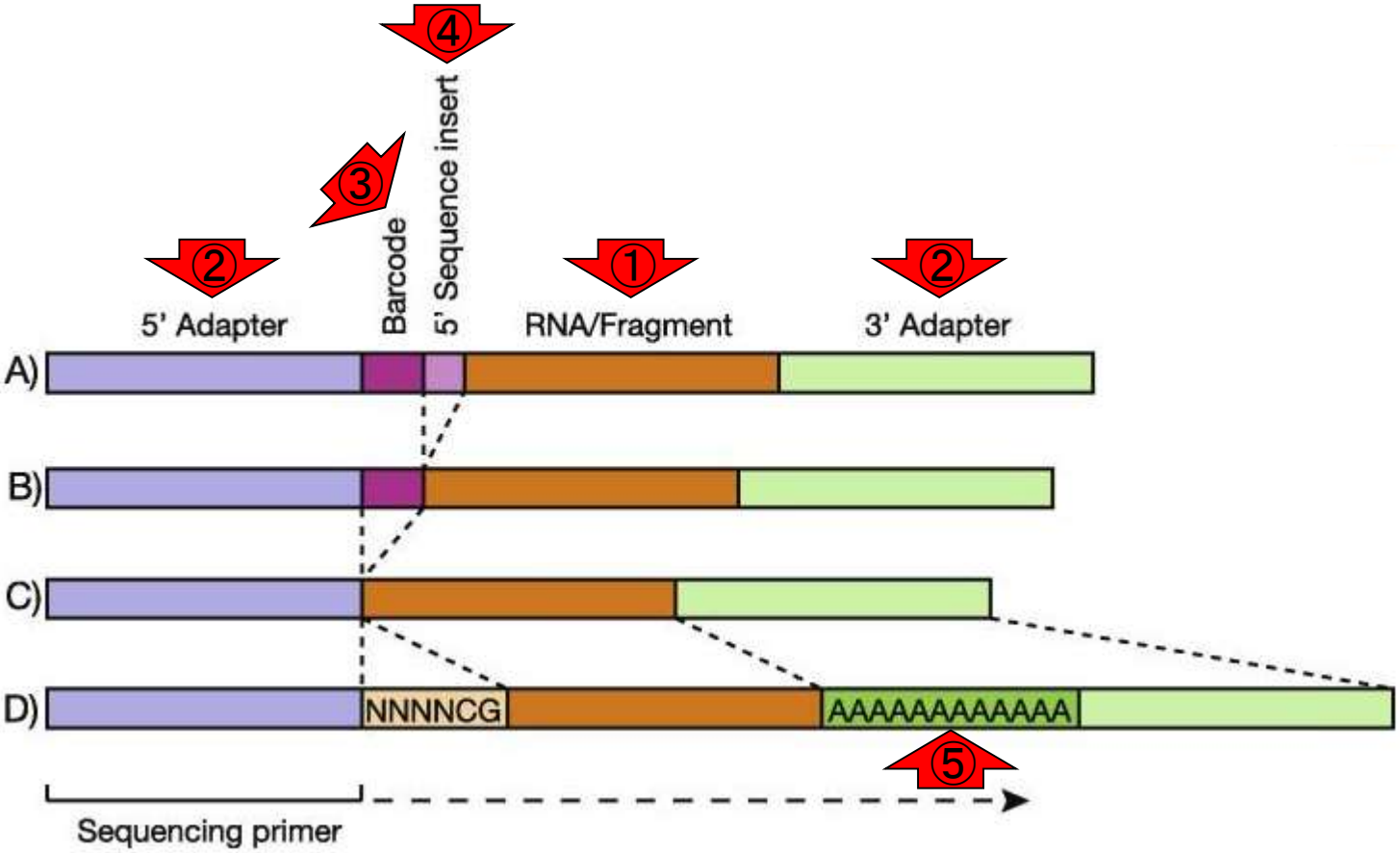
前処理（preprocessing） or Quality Control （QC） ①



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

NGSリードの模式図

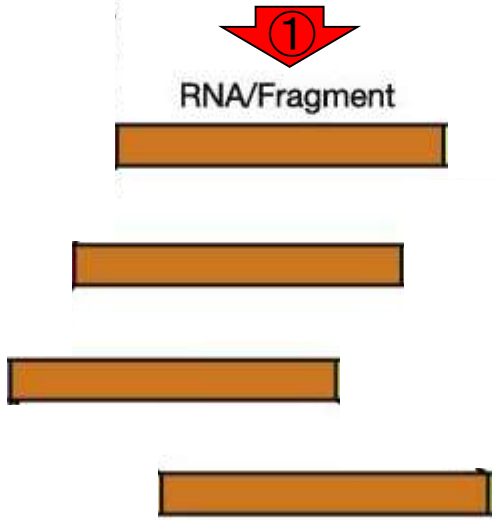
実験デザインにもよるが、RNA-seqのNGSリードには、目的の①RNA断片配列以外に、②アダプター配列、③バーコード配列、④インサート配列、⑤ポリA配列などが含まれる



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

QCの目的

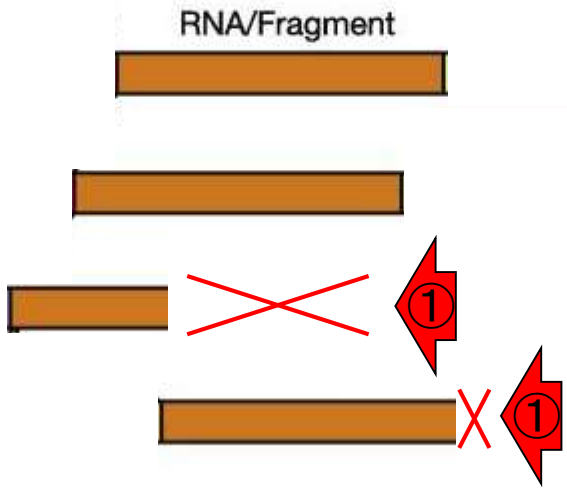
QCの目的は、基本的にはこんな感じにして、①目的の塩基配列のみにすること



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

QCの目的

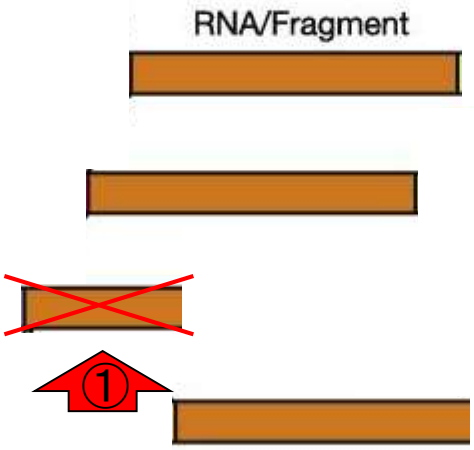
その上でさらに、クオリティスコアが設定した閾値以下の領域をトリミングしたりして…



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

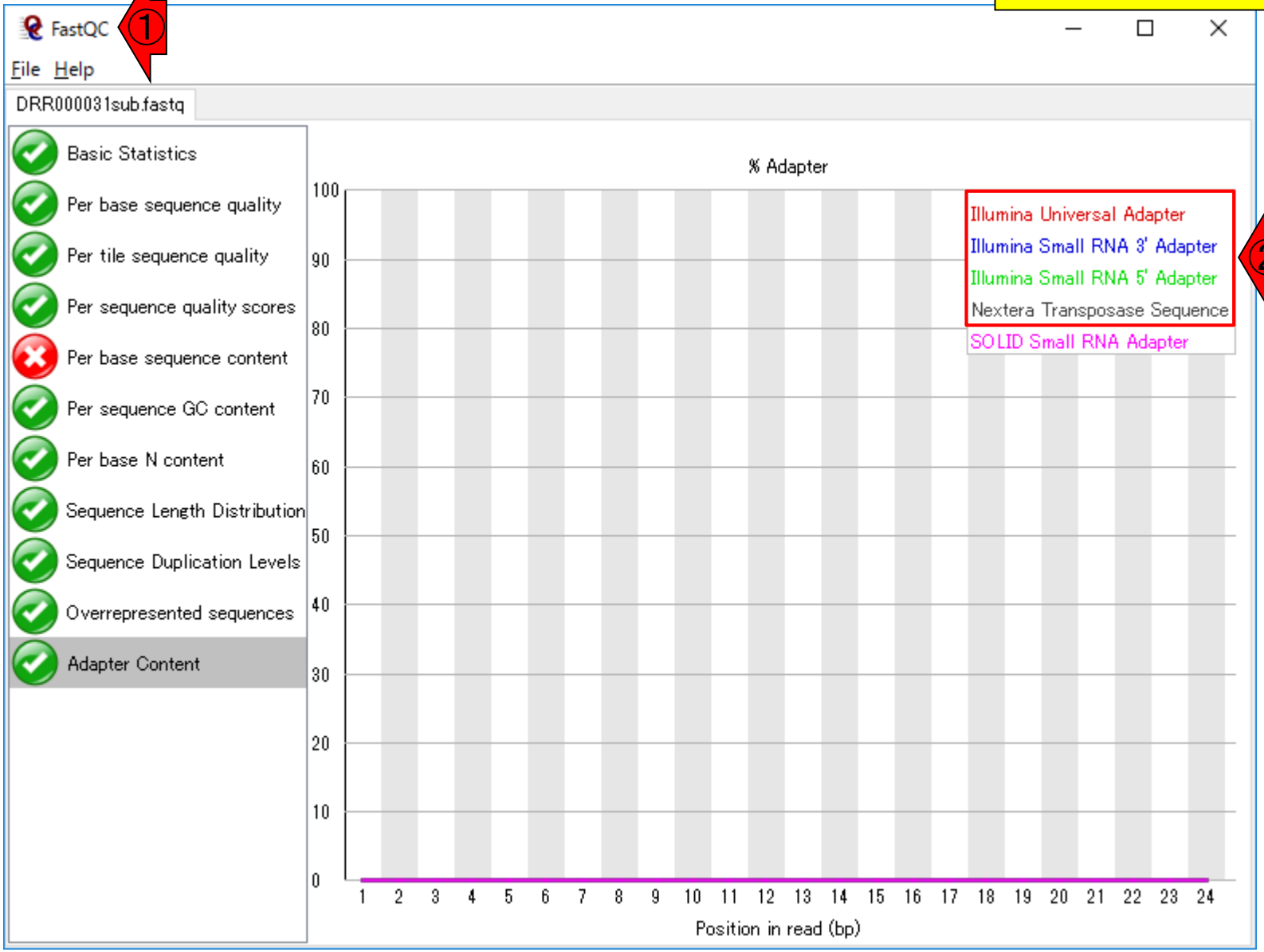


②

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。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

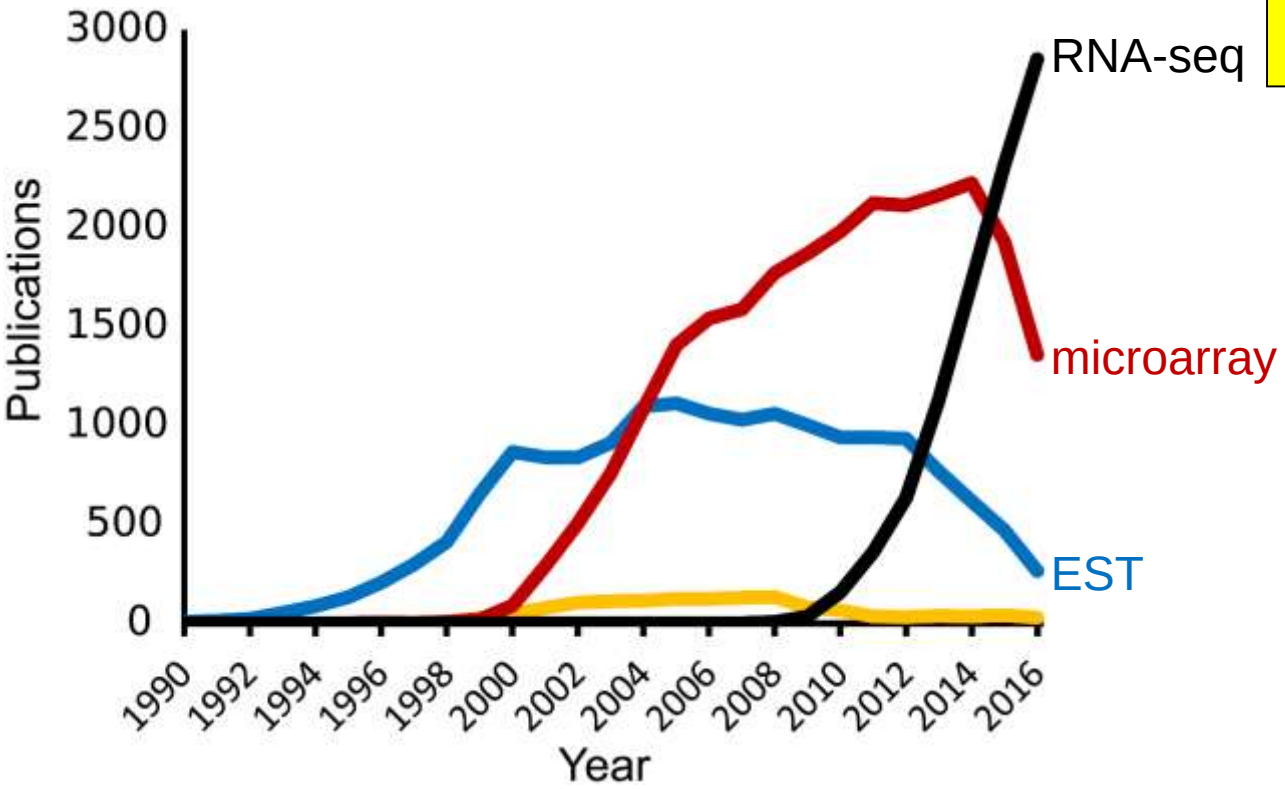
NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

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

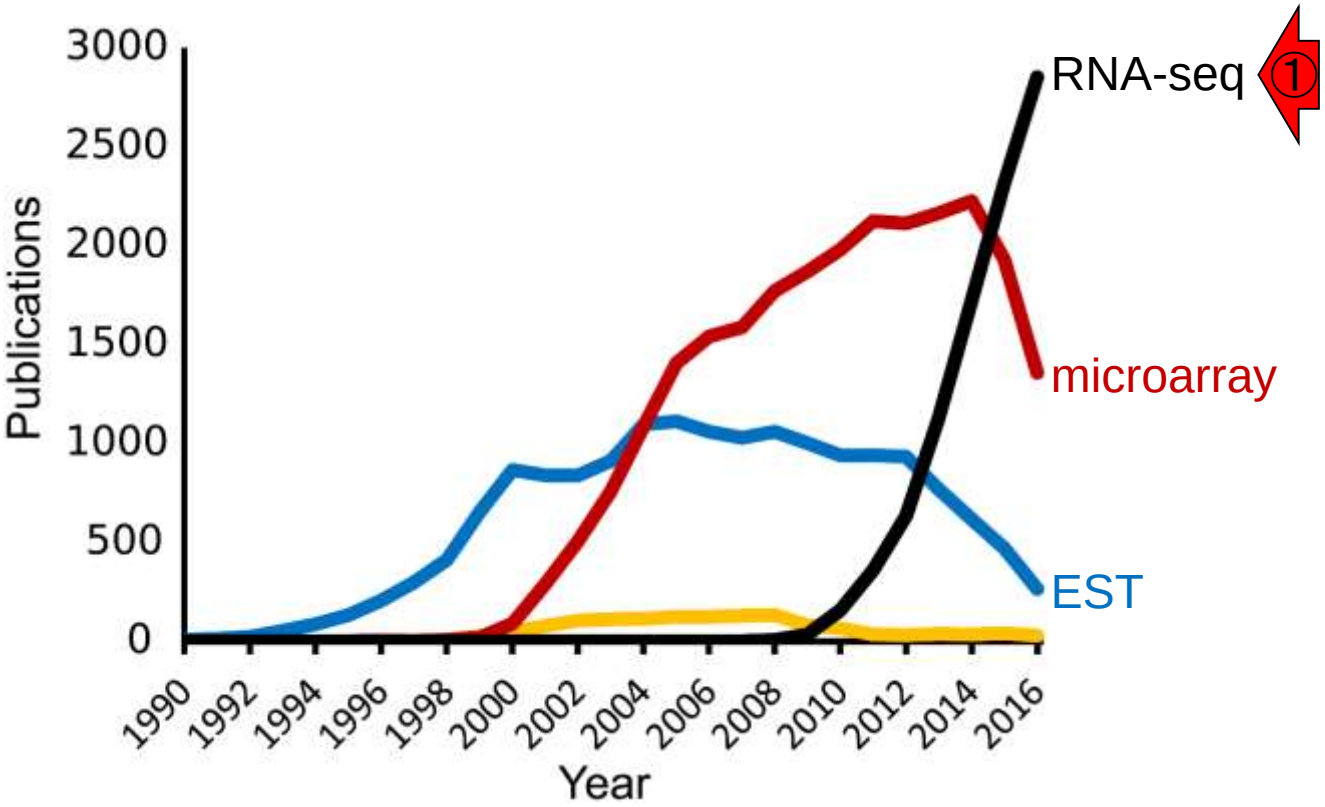
publishされた論文総数は、microarrayのほうがおそらくまだ多い。また、RNA-seqデータ解析についてもmicroarrayの経験が生かされている。①右肩上がりのRNA-seqと②右肩下がりのmicroarrayの論文数の推移を鑑み、平成30年度の本科目の講義内容をRNA-seqに据えることとしました



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

①RNA-seq論文のほとんどはIlluminaデータ



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

参考

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クエンサーも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クエ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

①最近の総説の、②ここのあたり。③RNA-seqデータ解析は4ステップからなり…

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 accordi[illegible]e experimental design and goals. ③ 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 experim[illegible]design [illegible]als. 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 com[illegible]d-line interface, eith[illegible]a Unix envi[illegible]ment or within the R/Bioconductor statistical environment [90].

①

②

③

④

⑤

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.

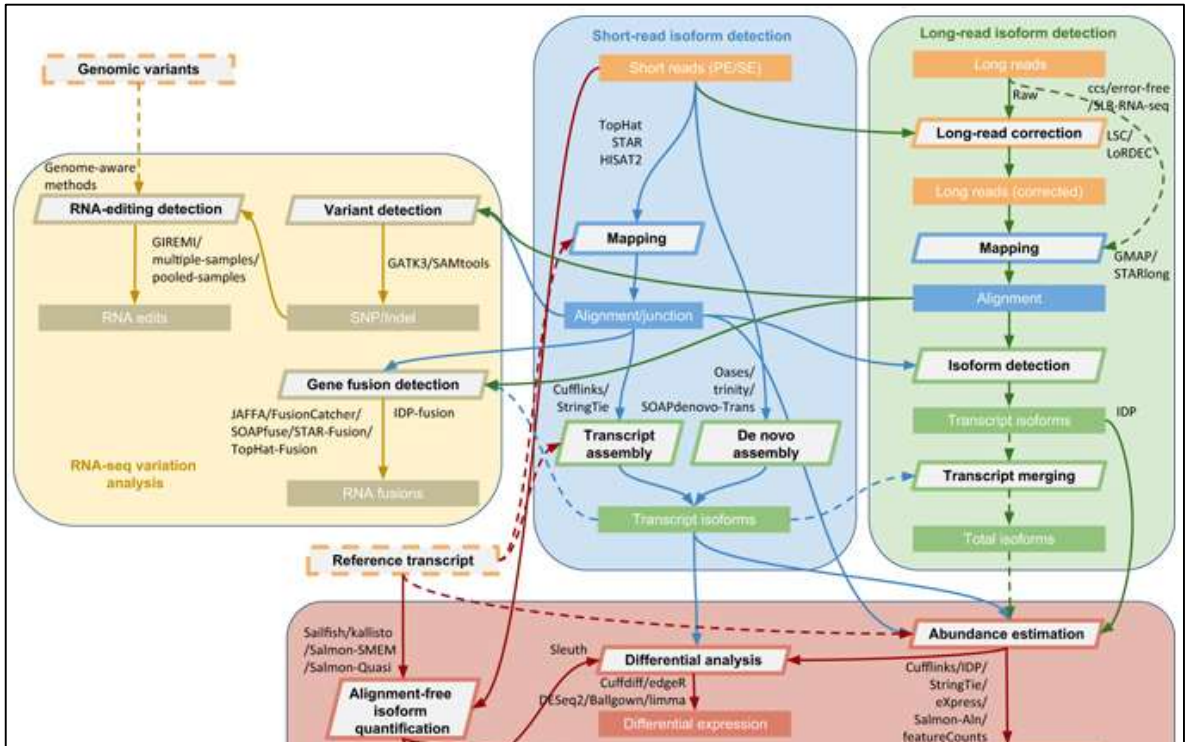
全体像のおさらい

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）

異なる論文間の対応関係

全体像のおさらい

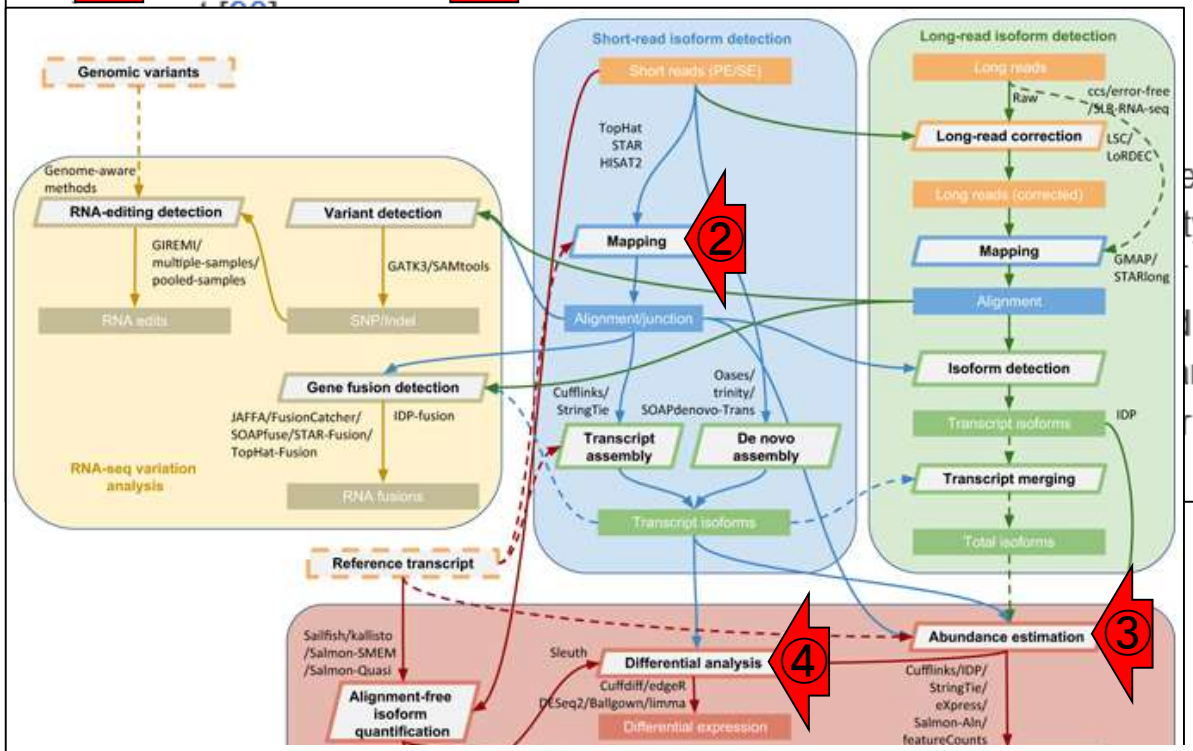
RNA-Seq data analysis

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 ② als. 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 com d-line interface, eith a Unix environment or within the R/Bioconductor statistical

③ ④

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



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

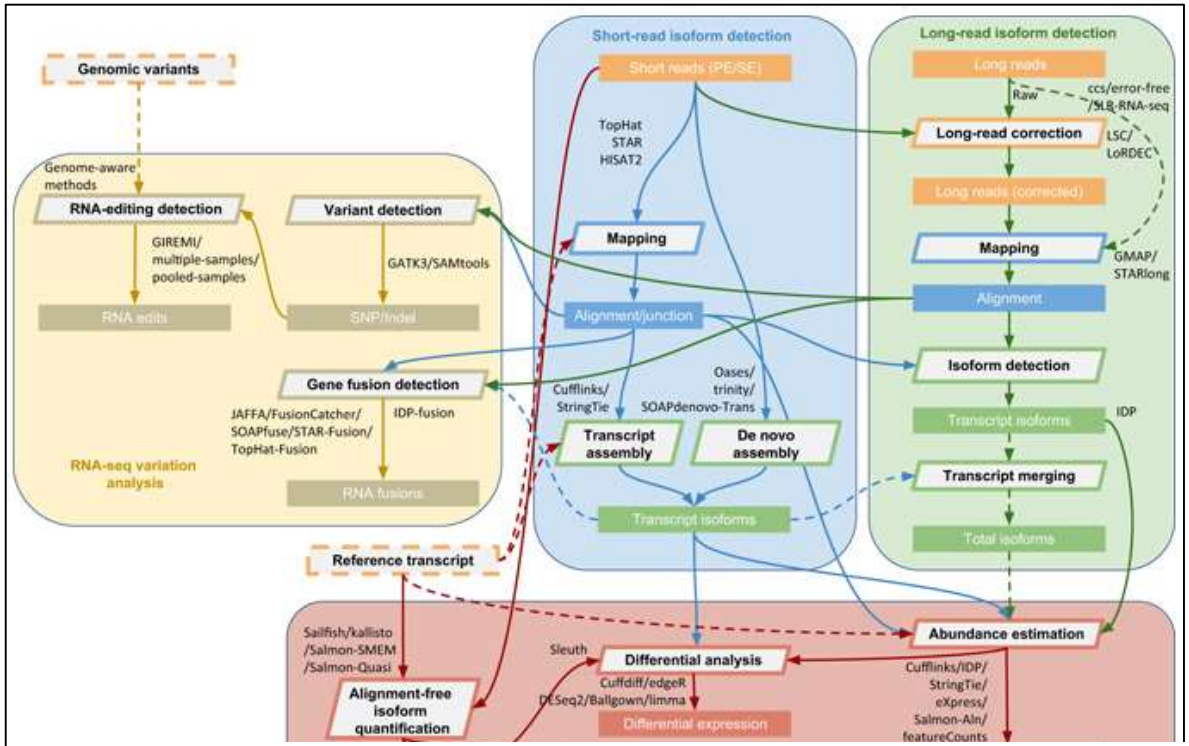
全体像のおさらい

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、qrqc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]. ②
Several options exist for sequence quality analysis, including the FastQC and FaQCs software packages [91][92]. Abnormalities identified may be removed by tri ③ or tagged for special treatment during later processes.

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

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、みたいな使い方をする。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].

③

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.

② ①

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

FastQCとFaQCsの関係

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の入力データとして用いる

NGSリードデータ（SRAファイル）

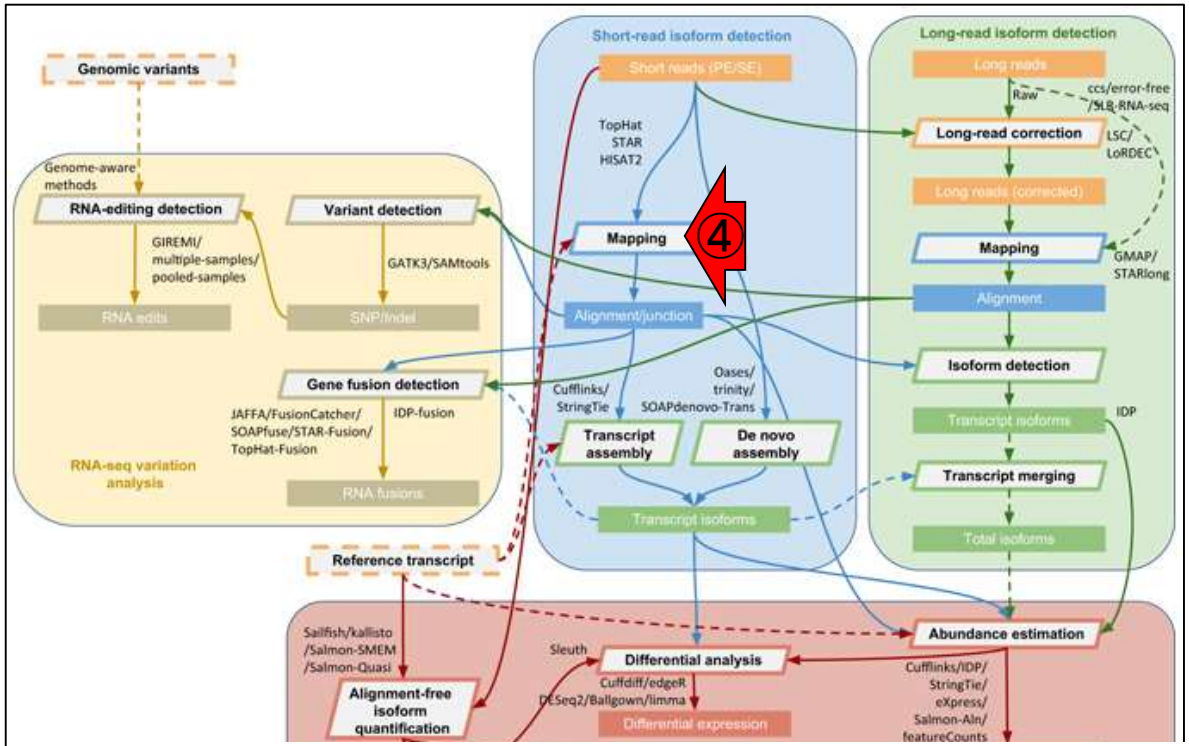
⇩

NGSリードデータ（FASTQファイル）

⇩

前処理（preprocessing） or Quality Control （QC）

⇩



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

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）

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R
トール済みであるとい
的にまとめた書籍も

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 | 講習会

- 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 modif 1 201
- 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 201 /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
- 書籍 | 日本乳酸菌学
- 書籍 | 日本乳酸菌学
- 書籍 | 日本乳酸菌学
- 書籍 | 日本乳酸菌学
- イントロ | 一般 | ラン
- イントロ | 一般 | 任意
- イントロ | 一般 | 任意

① ②

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述(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 ですがお知らせ願います。

- 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(201 12.11版; 約24MB; 推奨)
  - Macintosh用 着手)

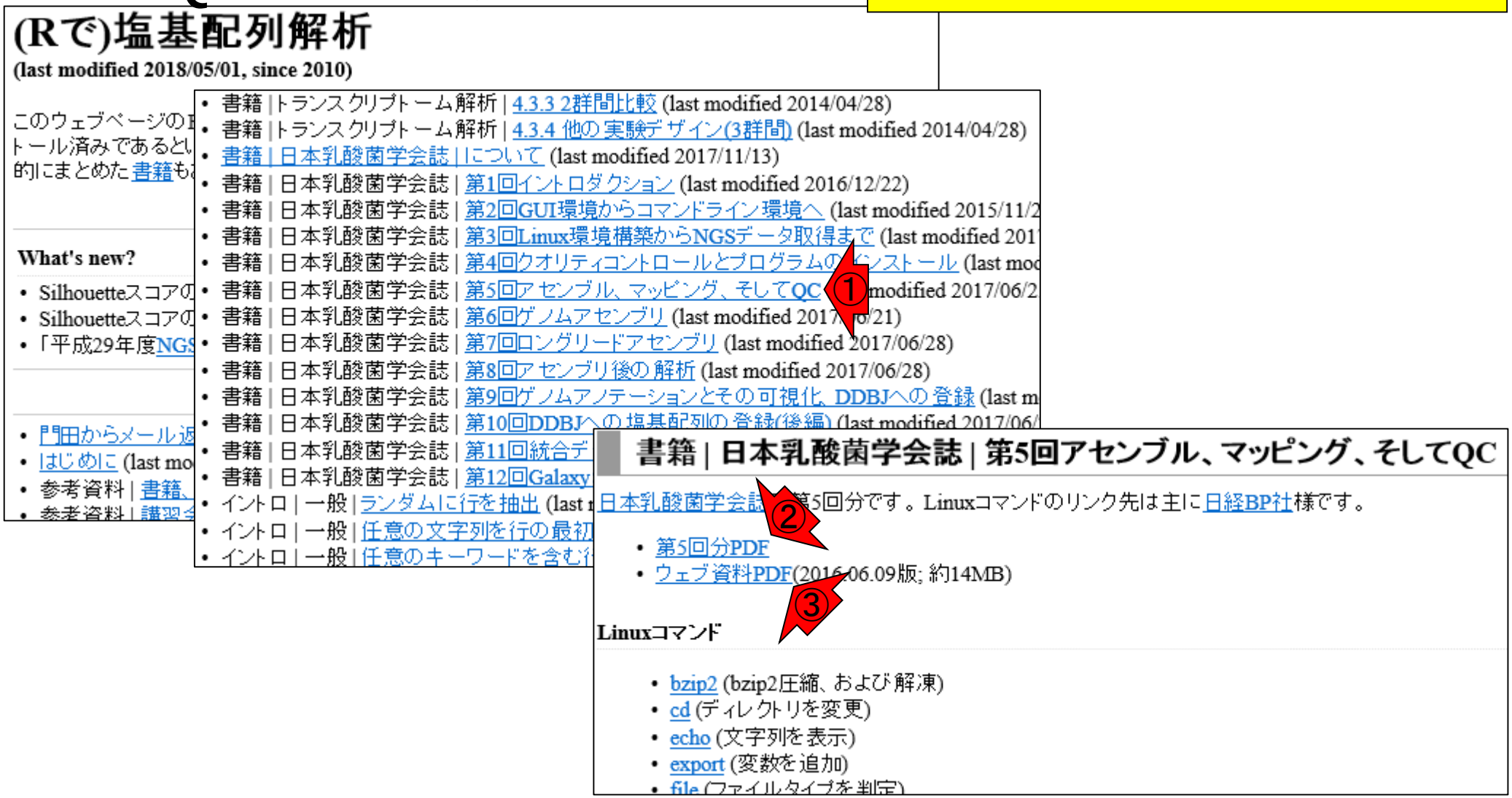
③

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)

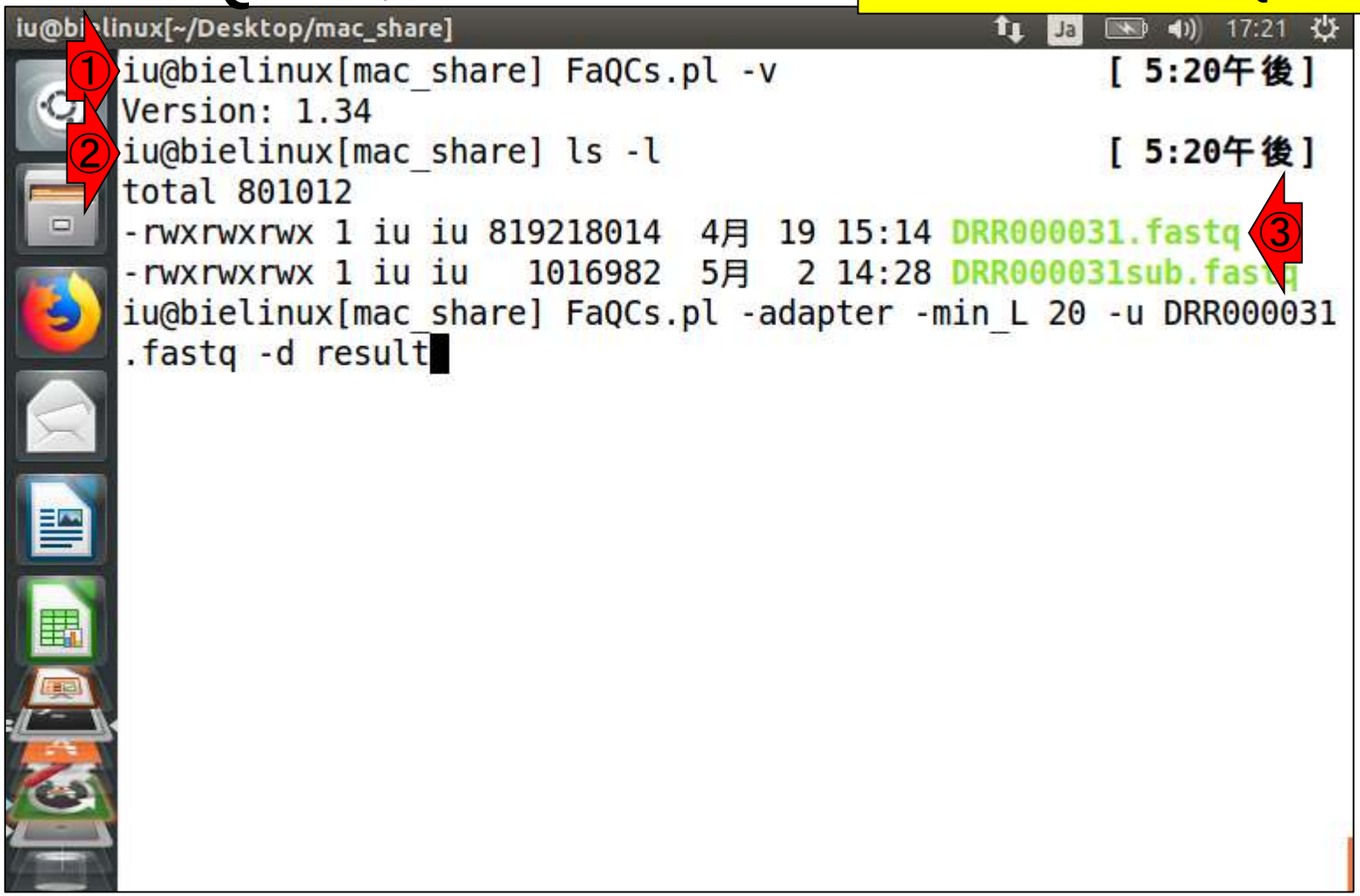
FaQCsの利用

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



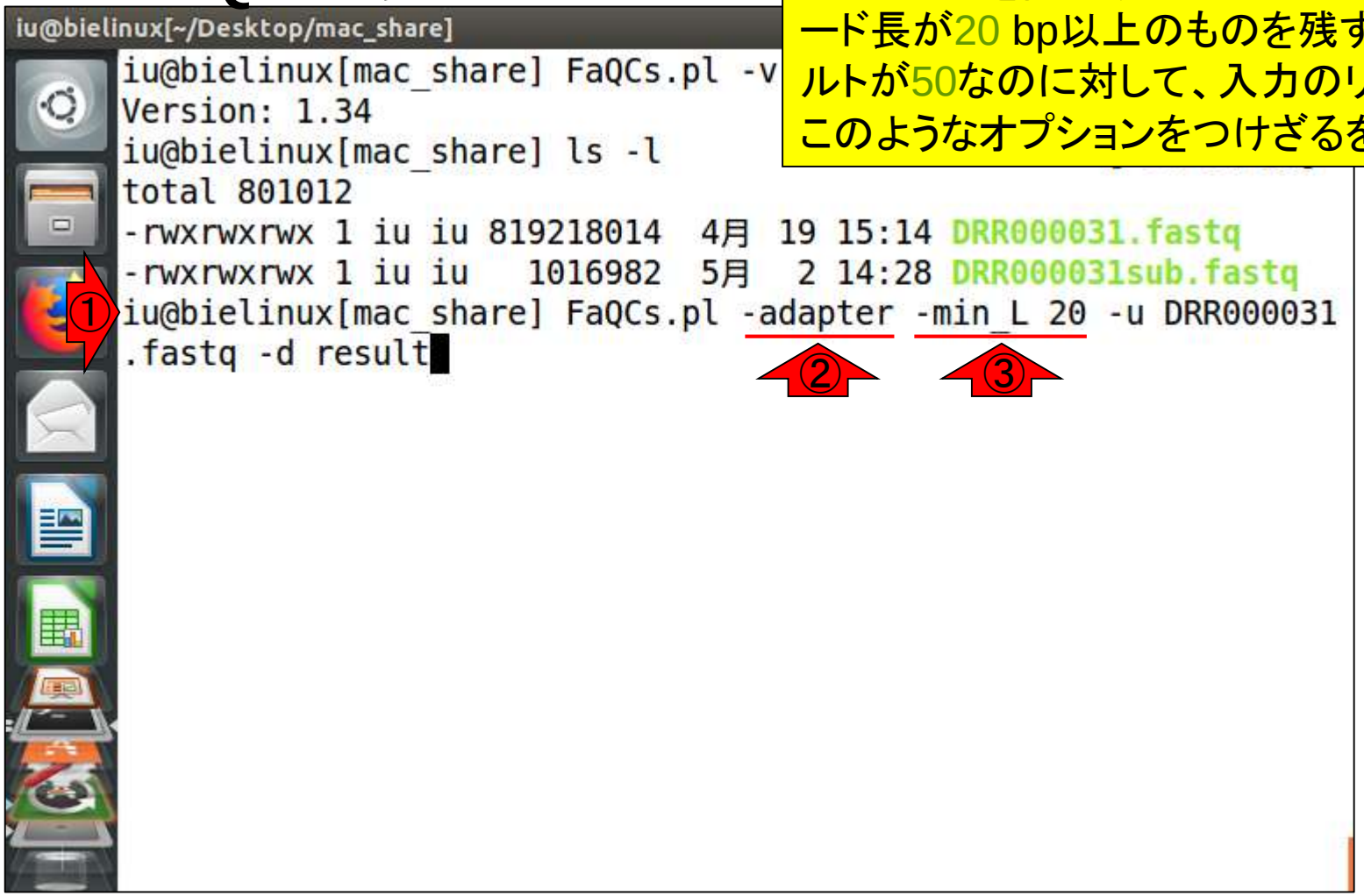
FaQCs実行

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



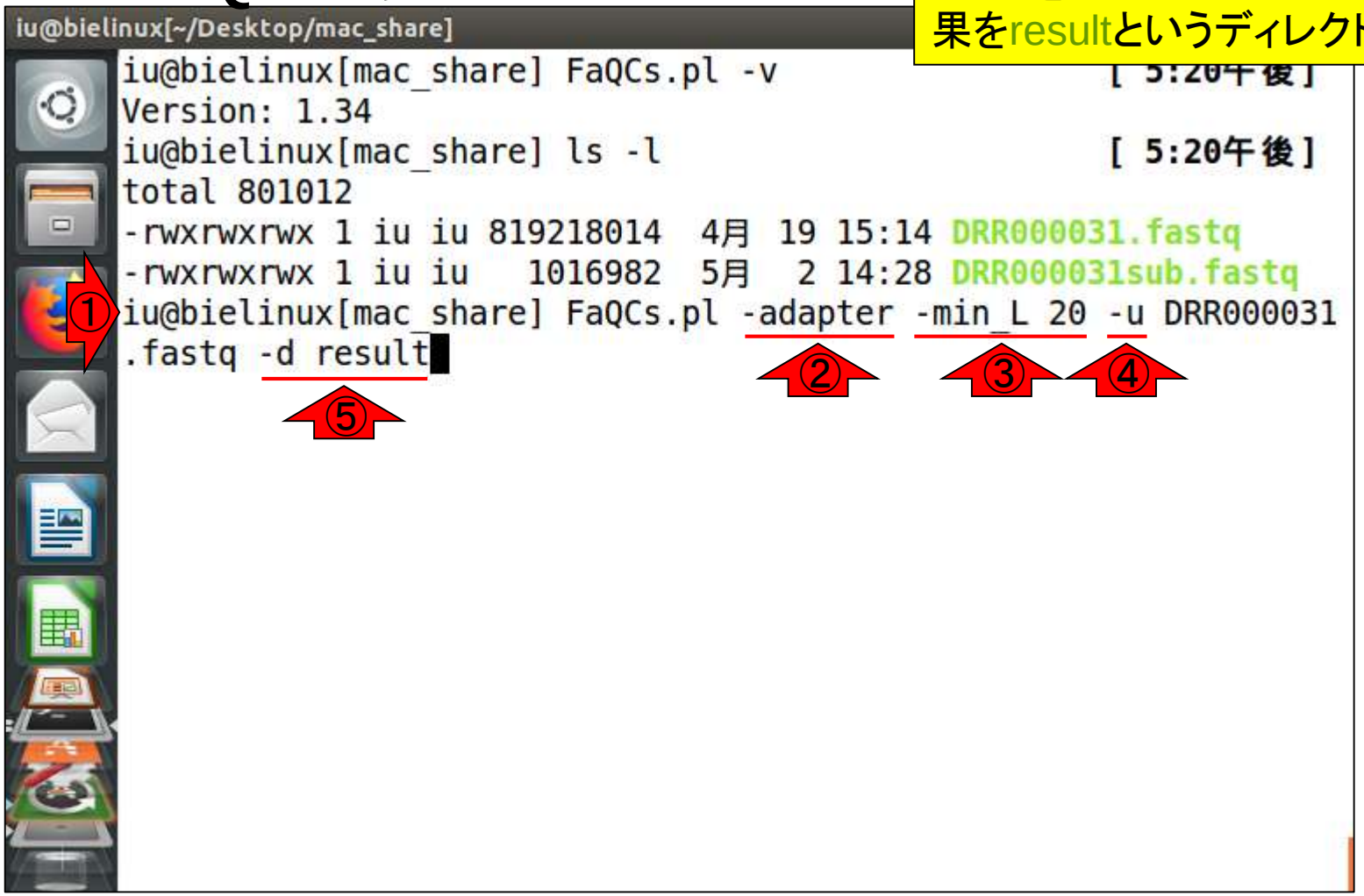
FaQCs実行

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



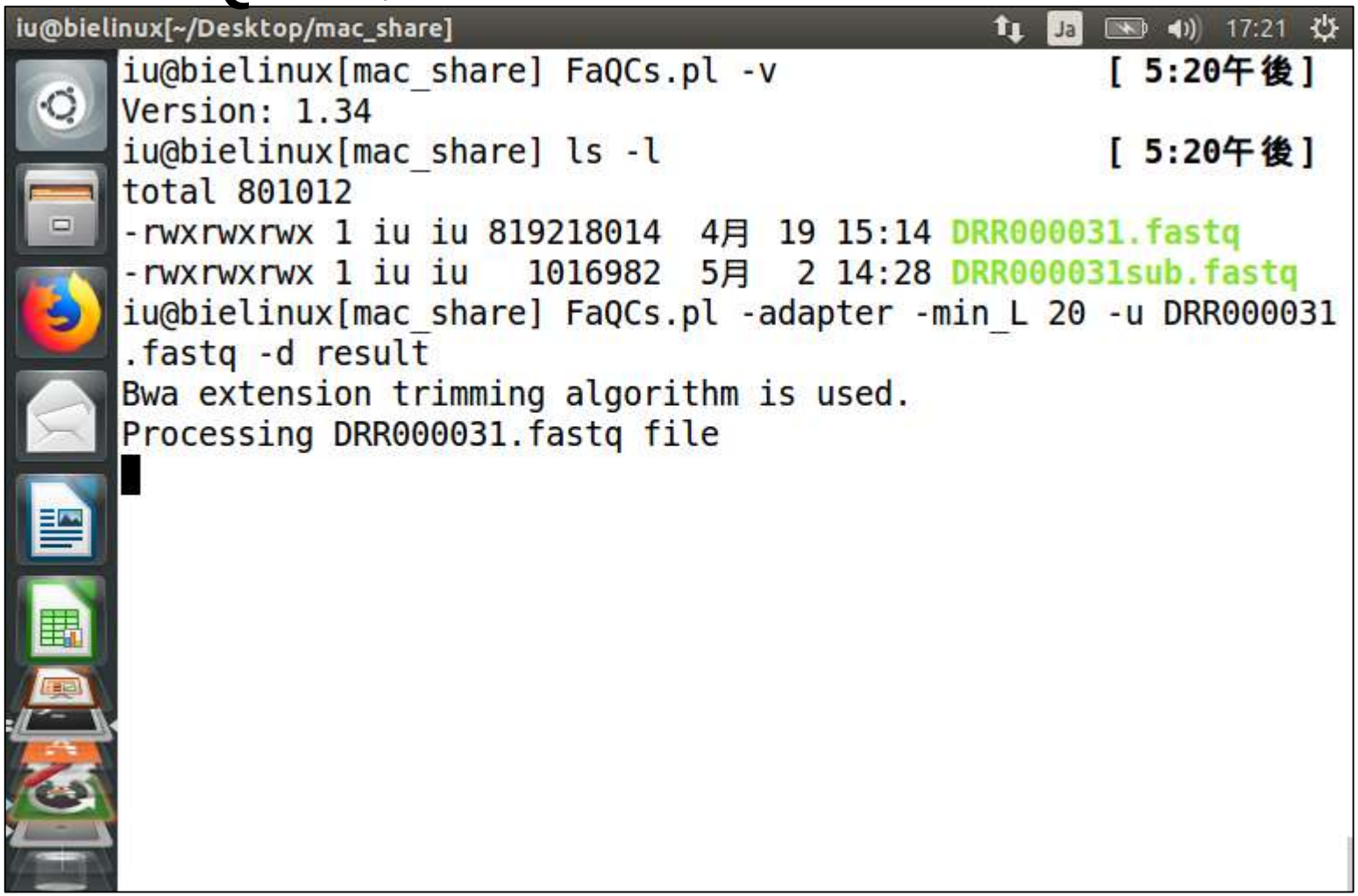
FaQCs実行

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



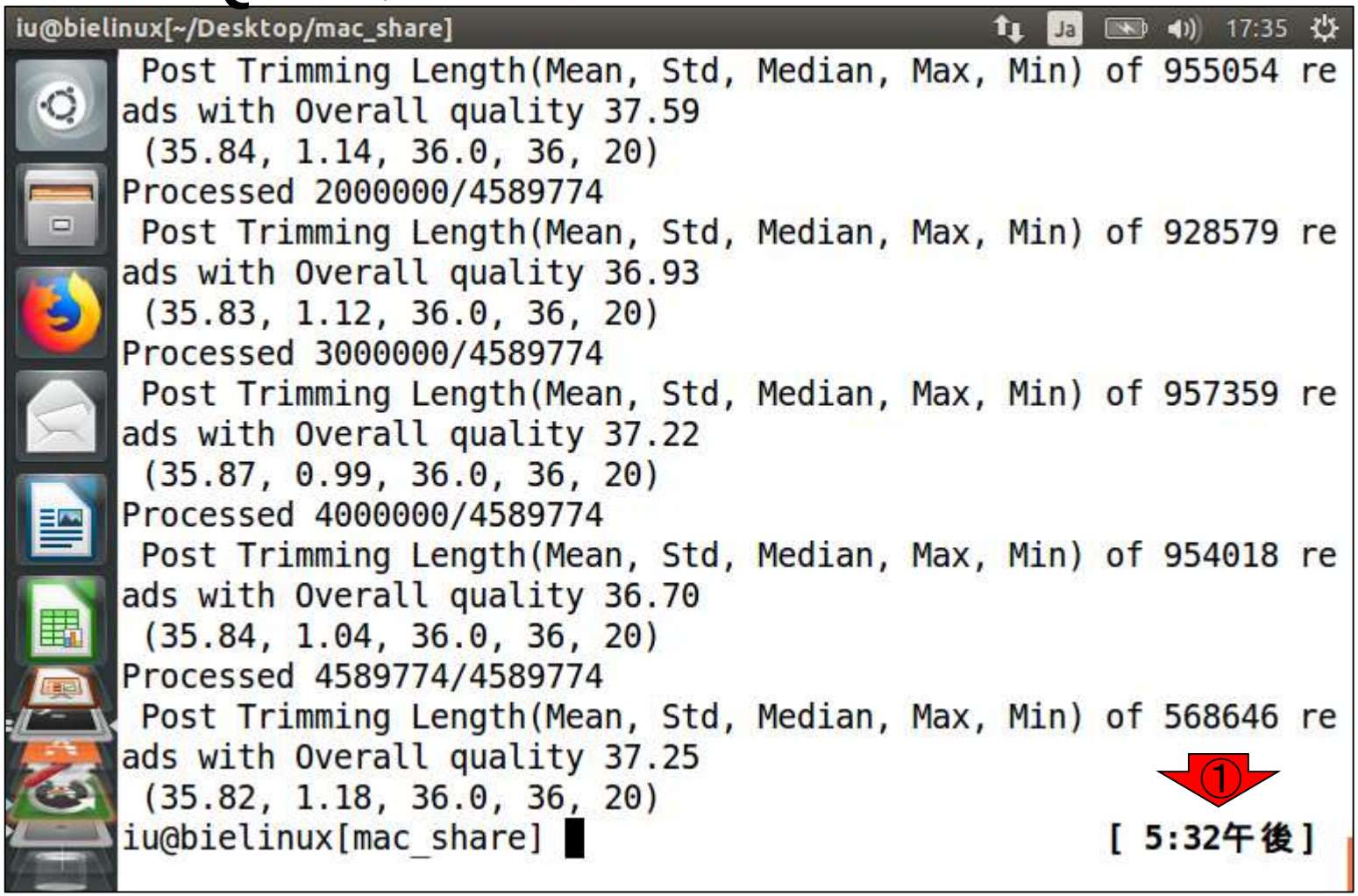
リターンキーを押した直後

FaQCs実行中



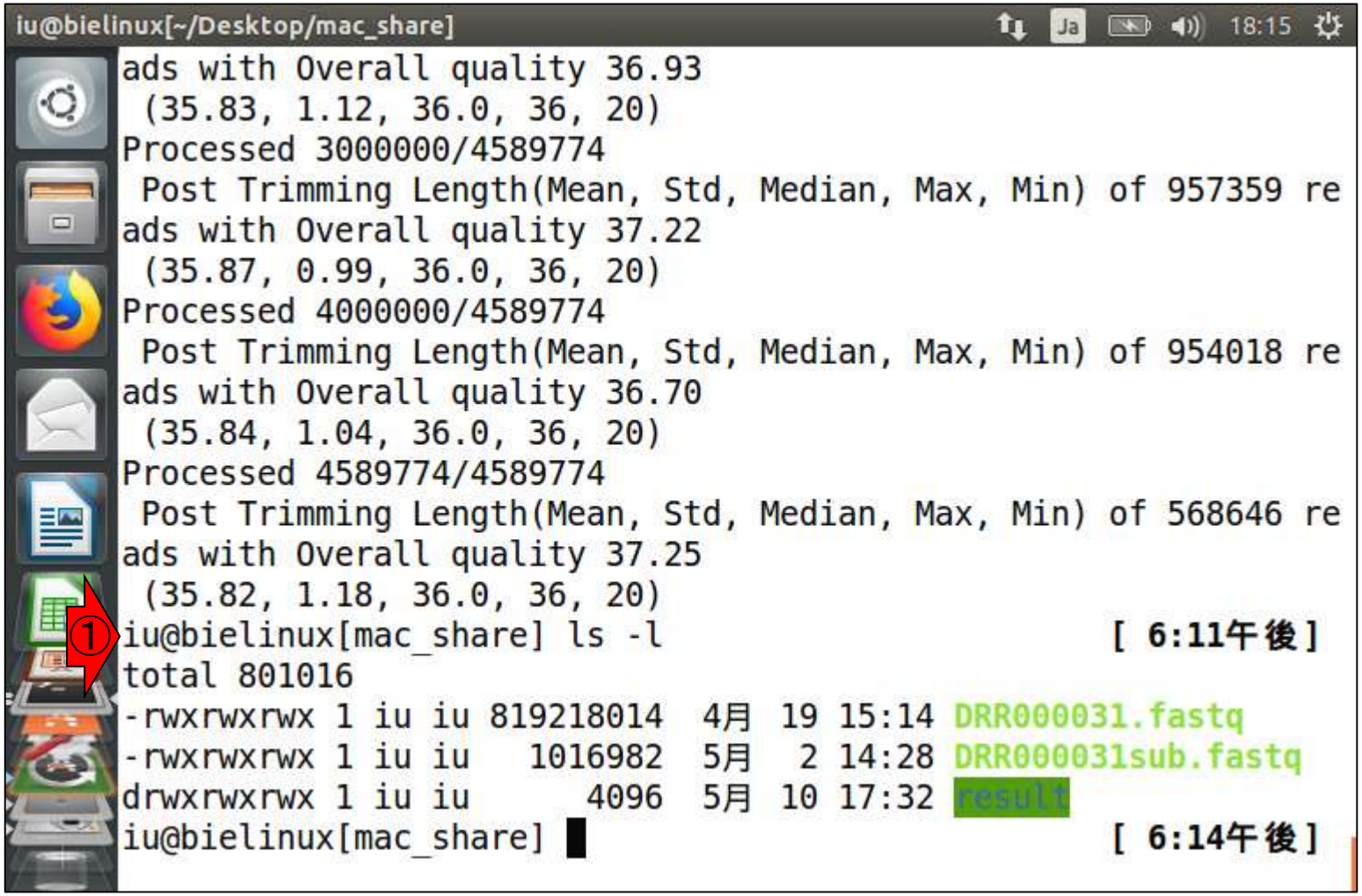
FaQCs実行完了

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



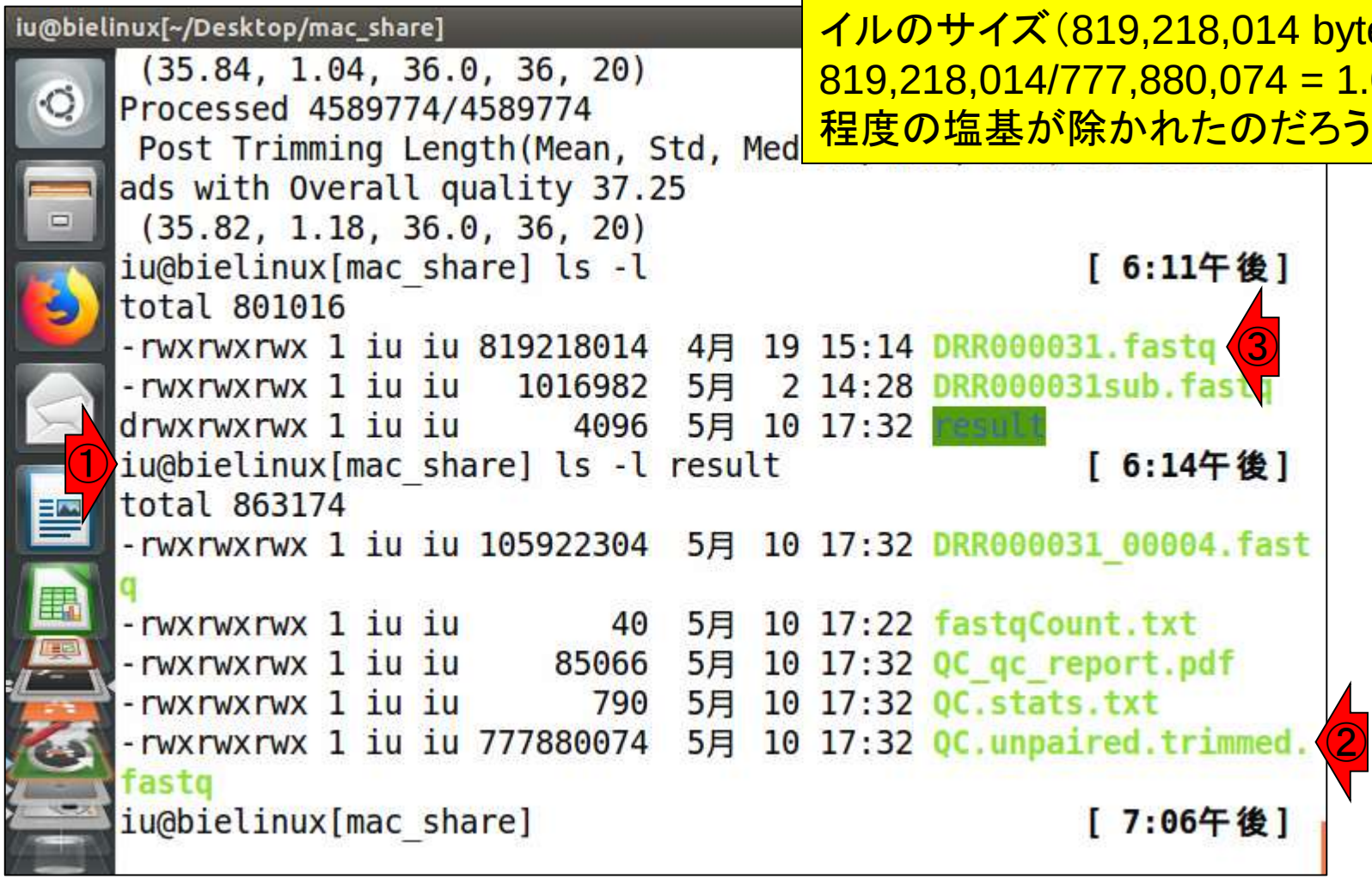
ls -lで確認

①ls –l。確かに②resultディレクトリが作成されています



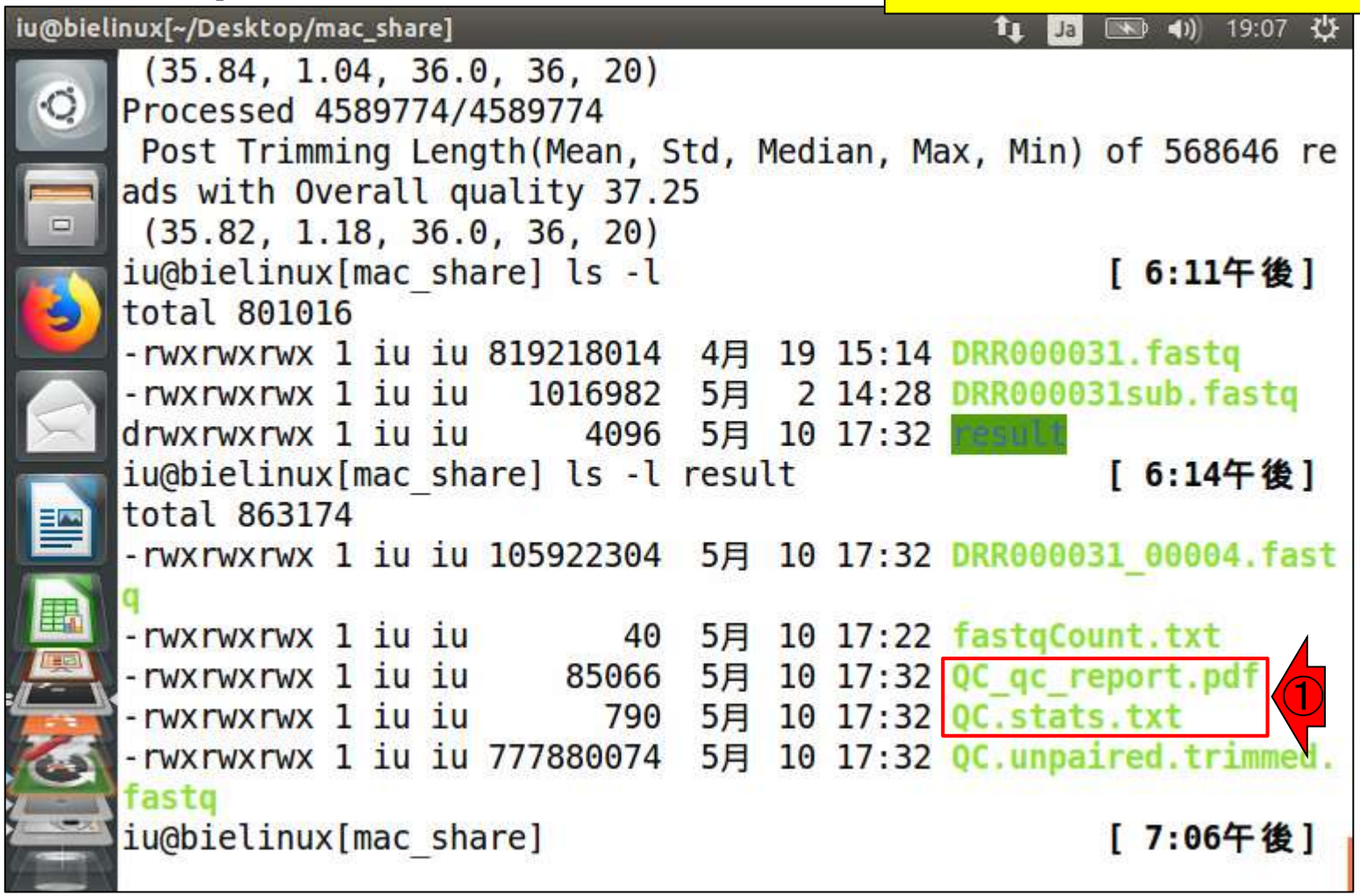
ls -lで確認

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ろうと解釈する



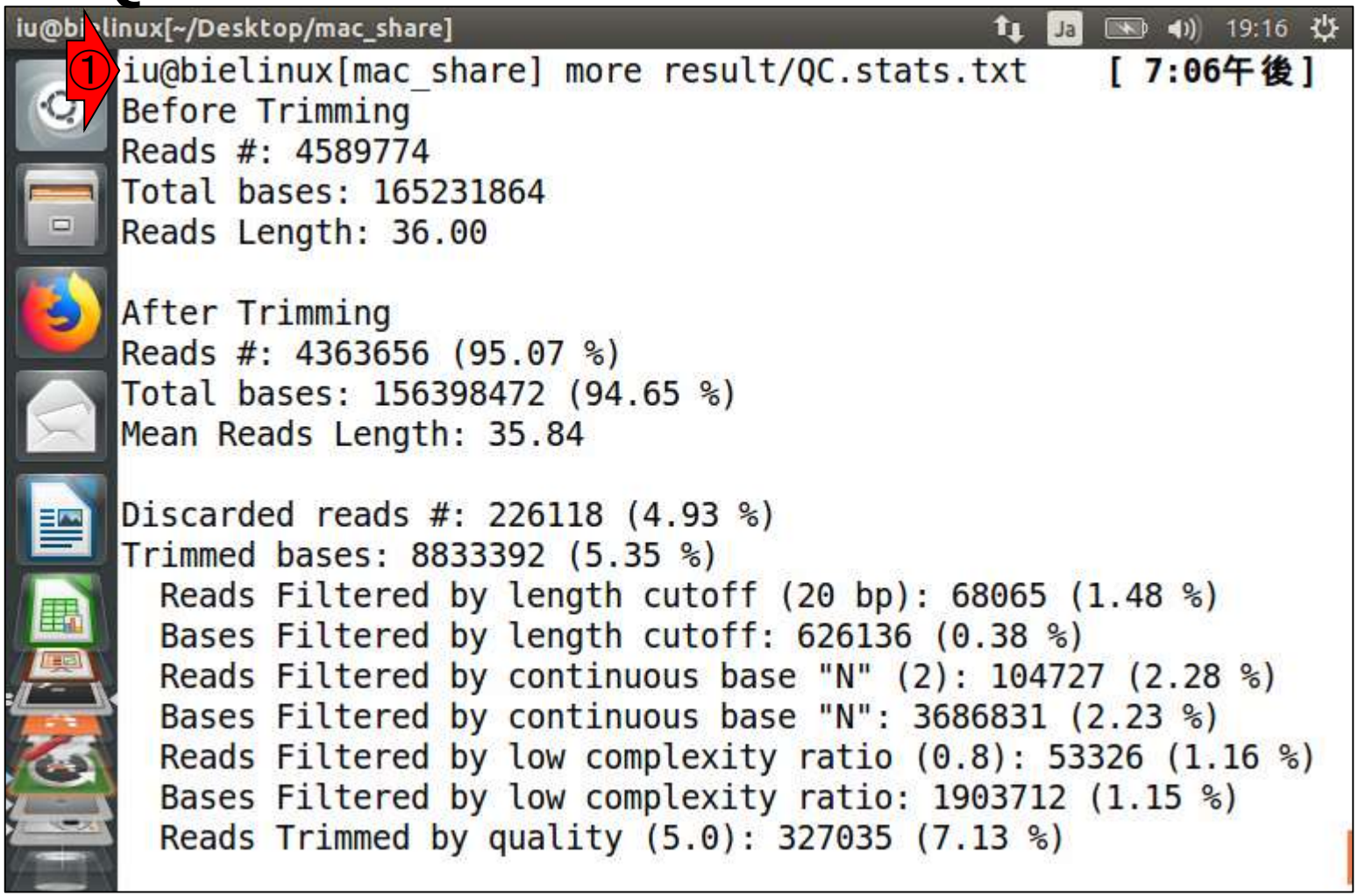
結果の詳細は…

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



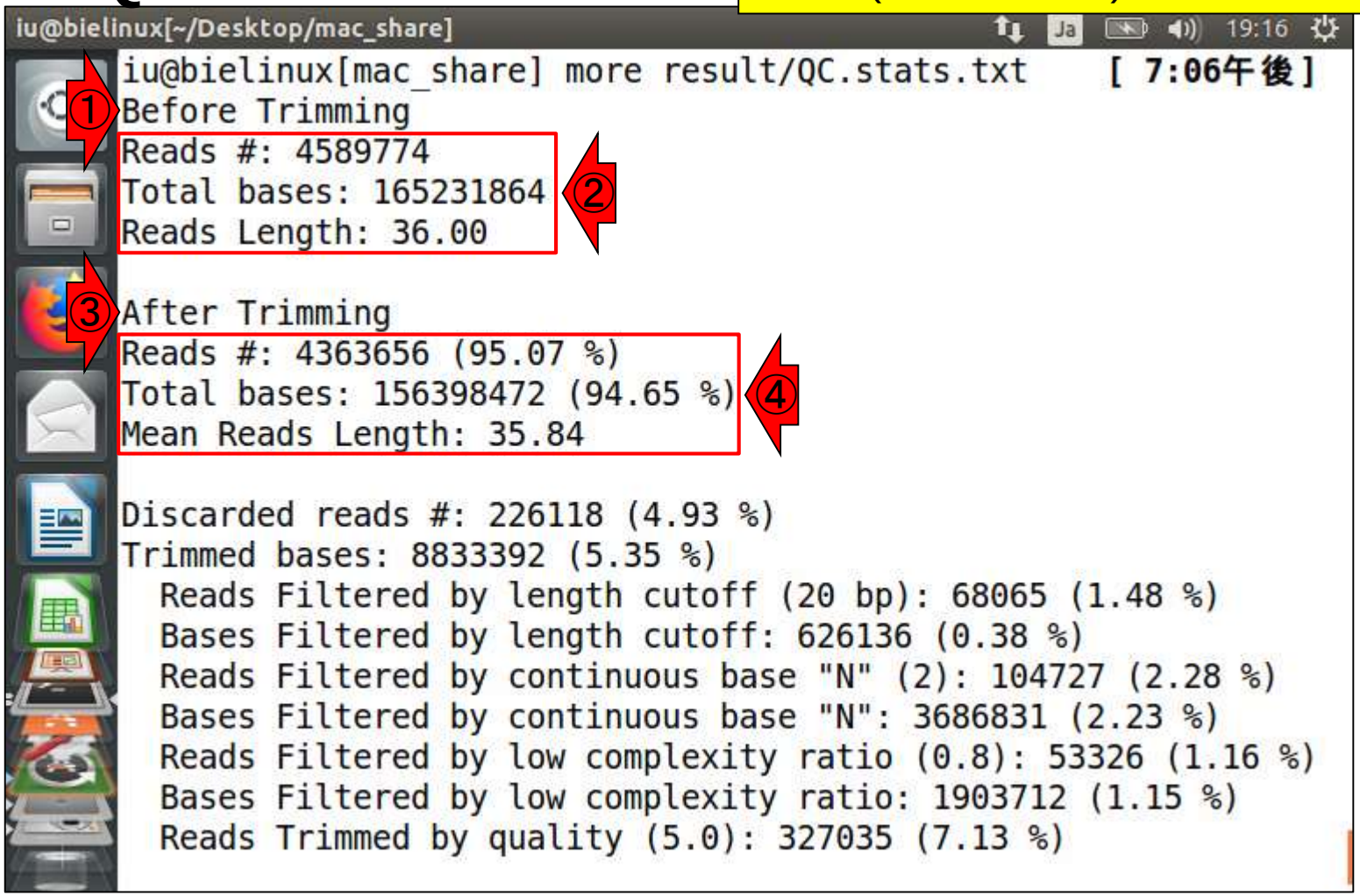
①more result/QC.stats.txt実行結果

QC.stats.txt



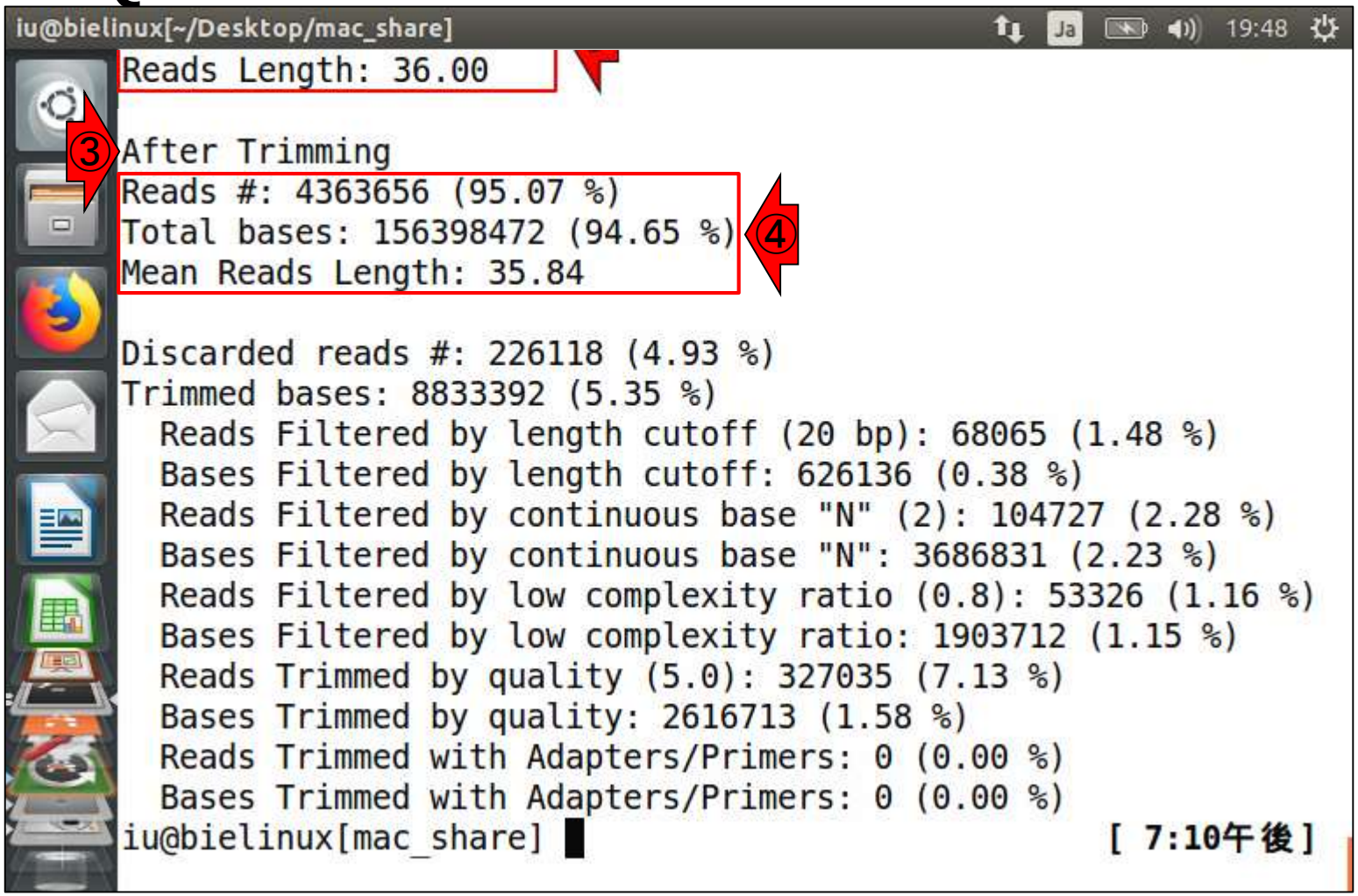
QC.stats.txt

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



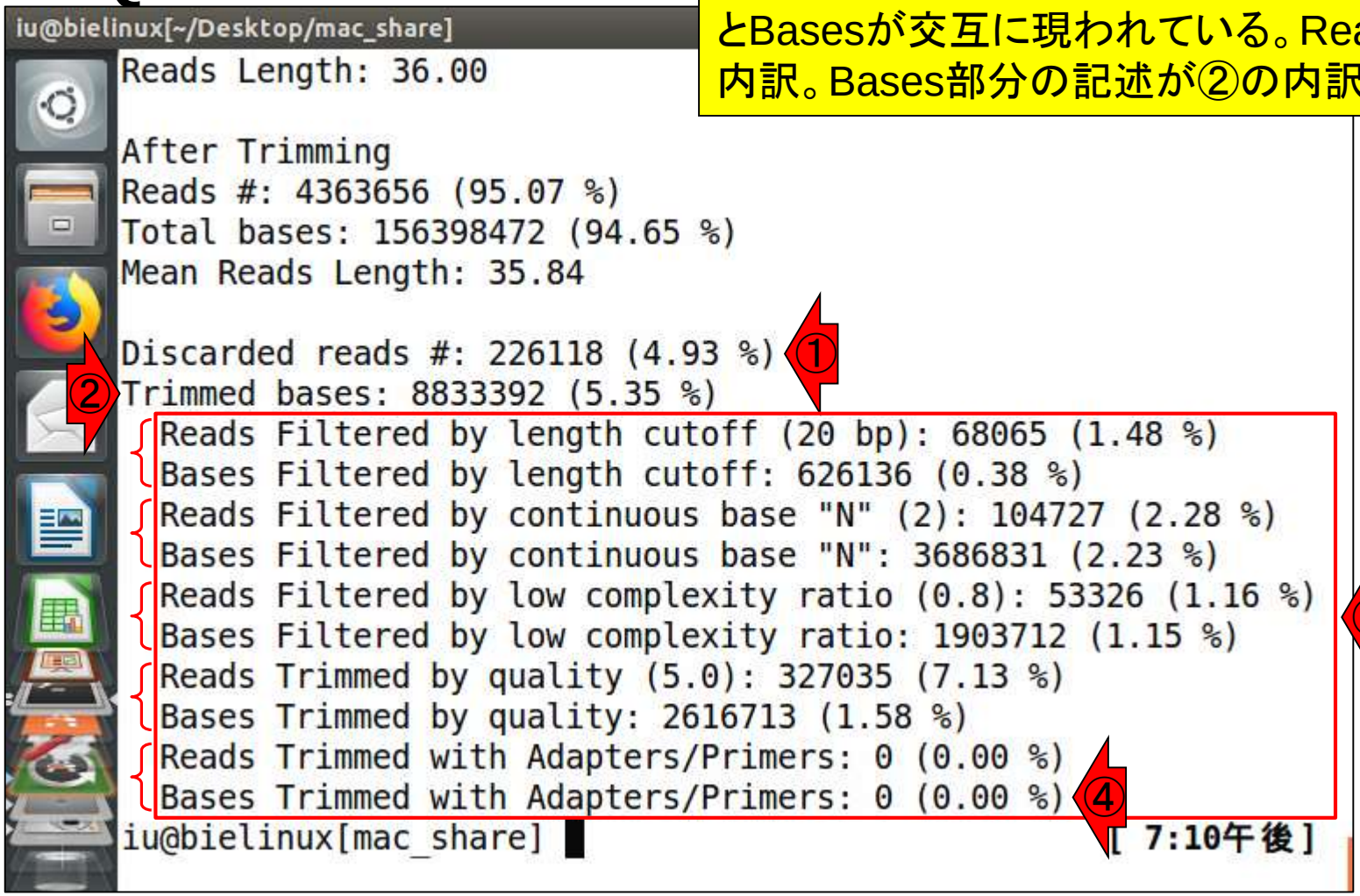
最終行のところまで表示させたところ

QC.stats.txt



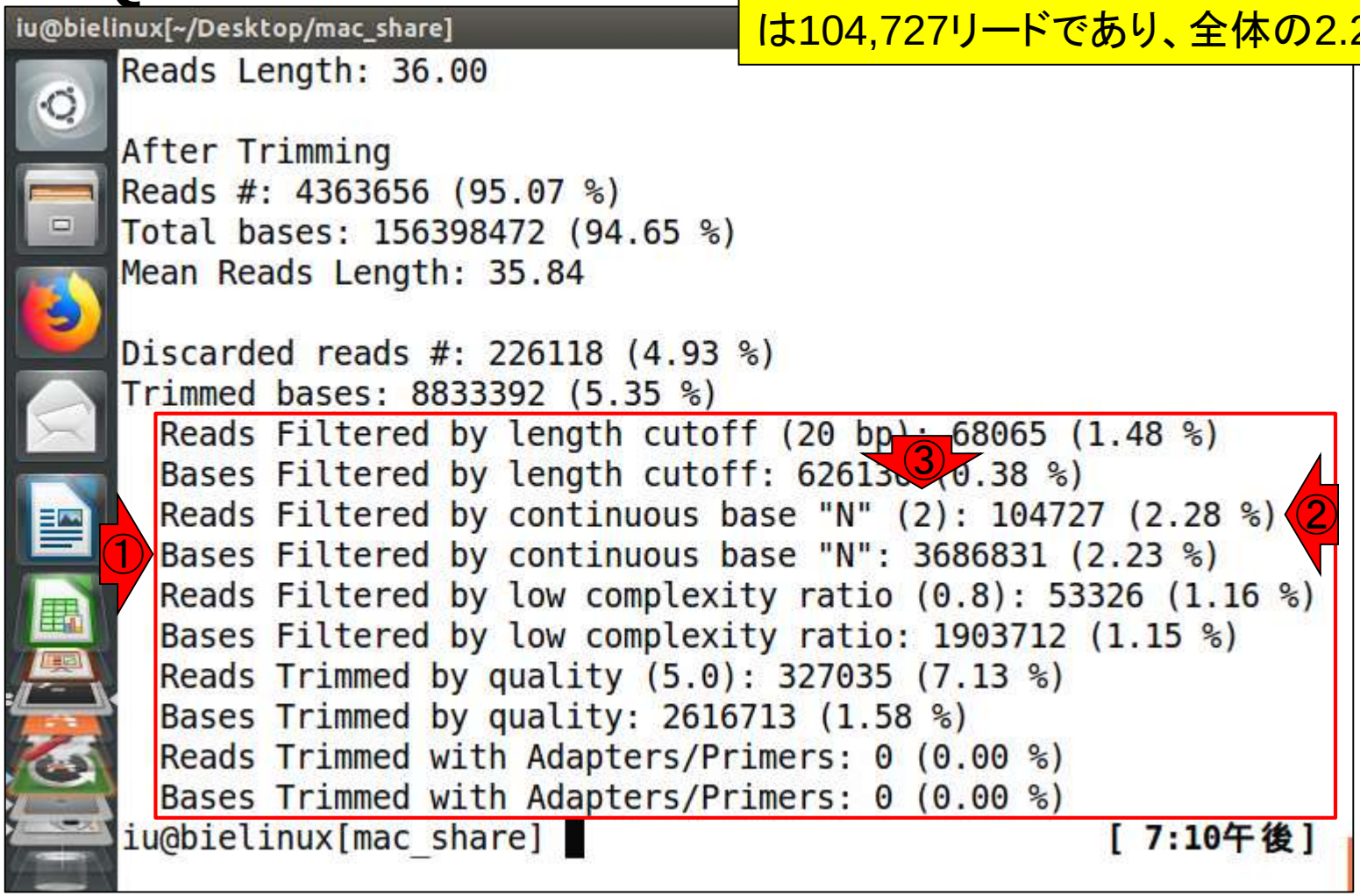
QC.stats.txt

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



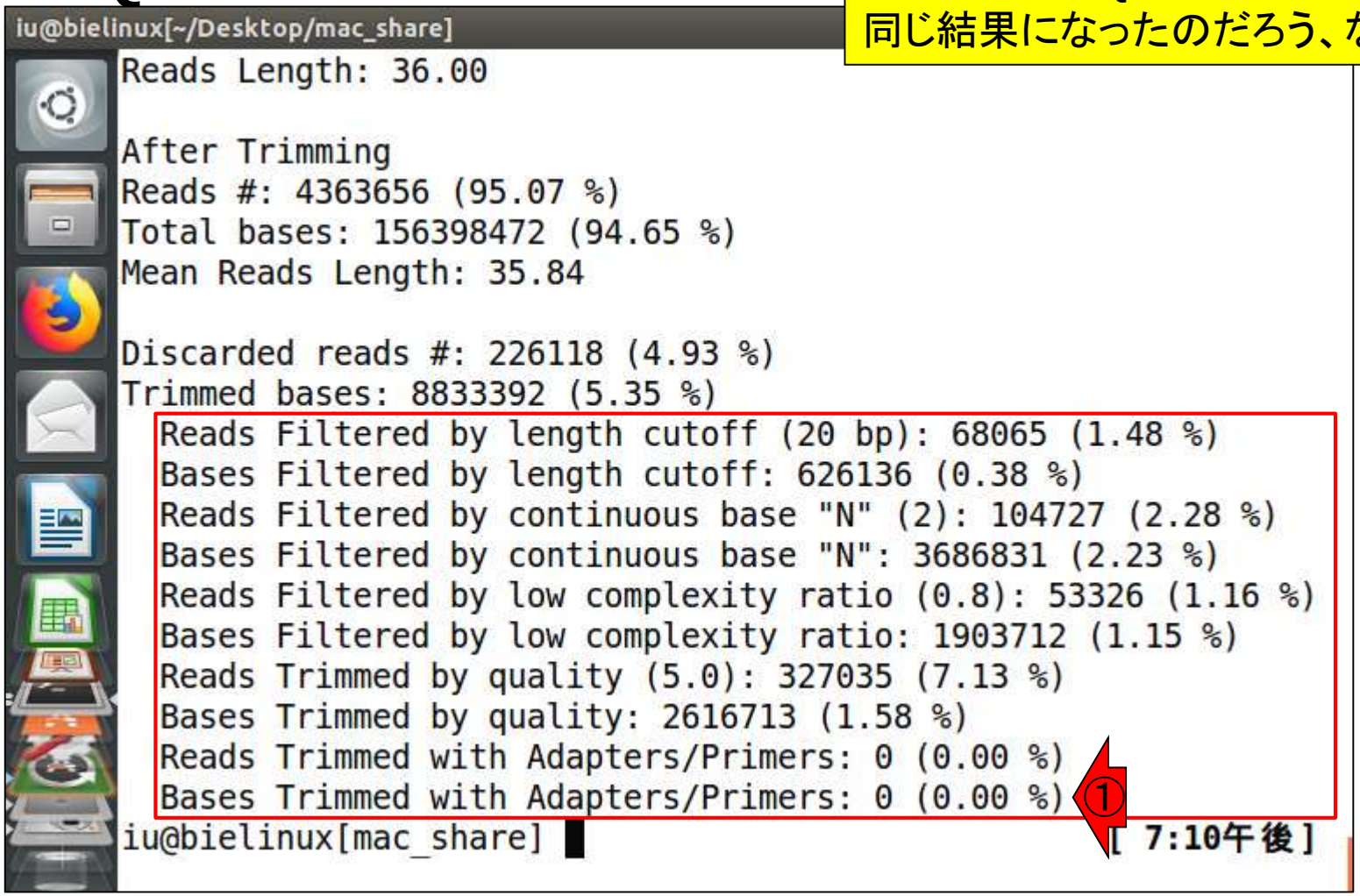
QC.stats.txt

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



QC.stats.txt

また、①の最終行では、Adapters/Primers由来のBasesは一つもなかったということを示している。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

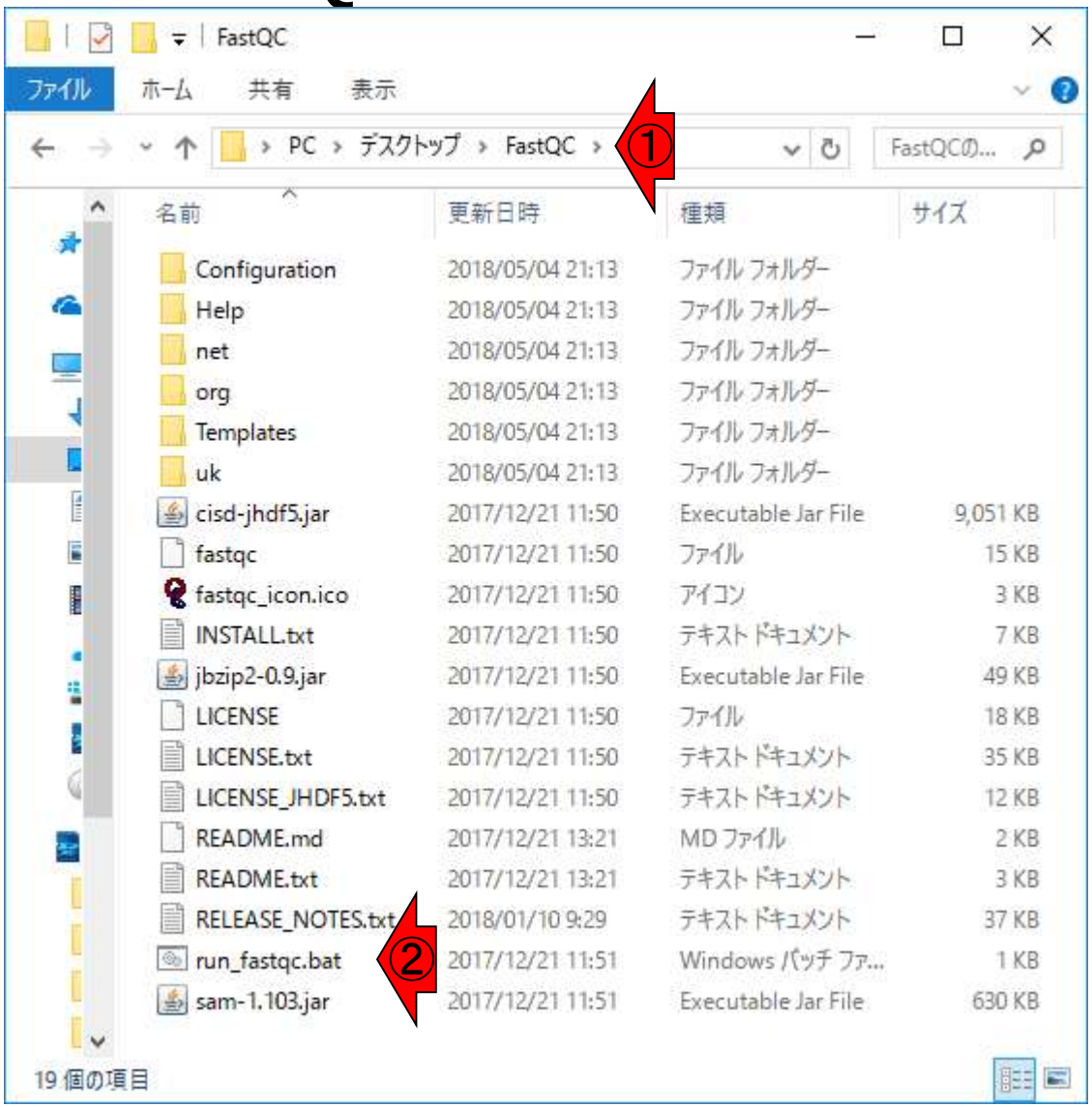


Contents

- 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、qrqc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

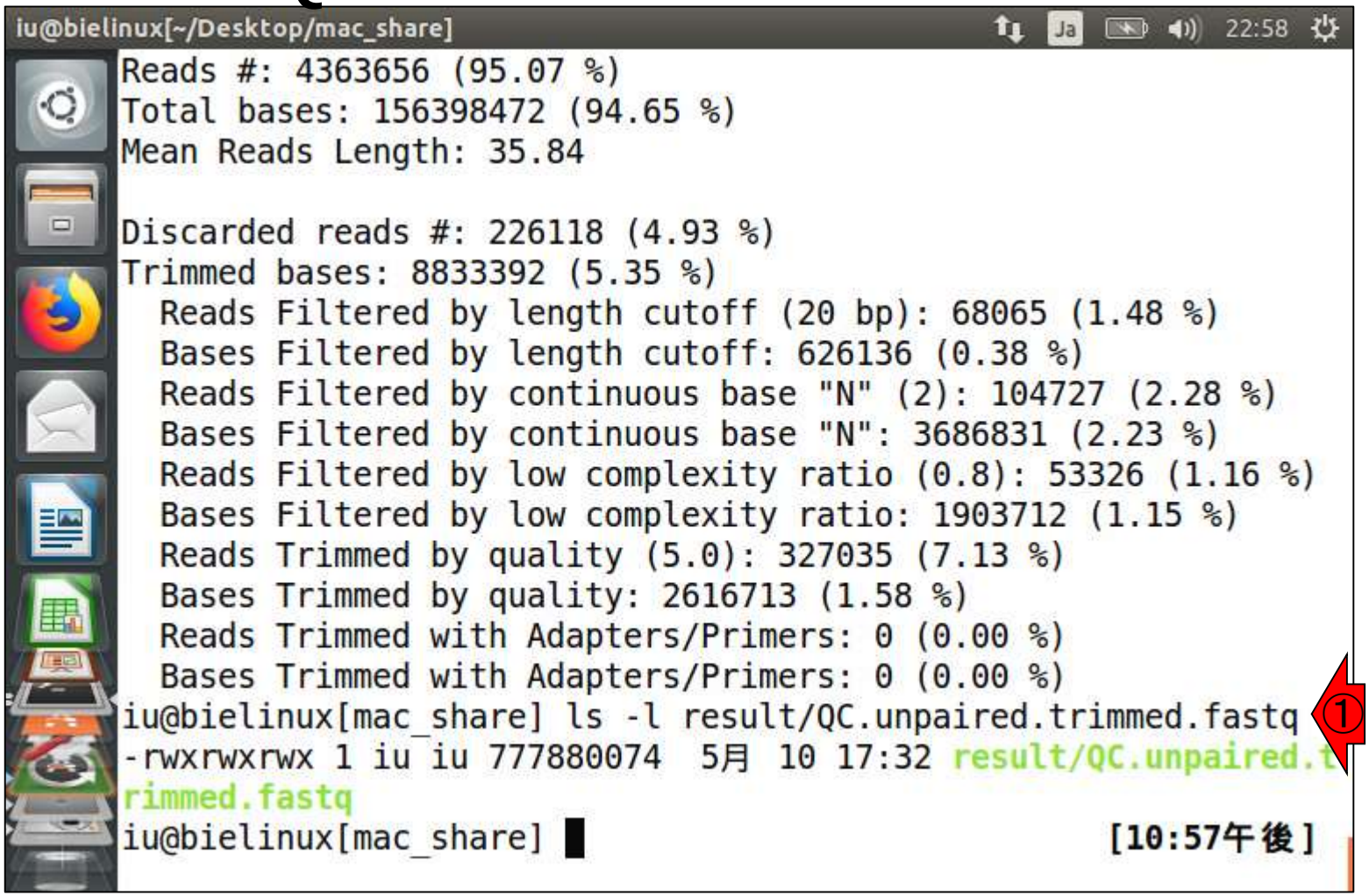
FastQCを起動

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



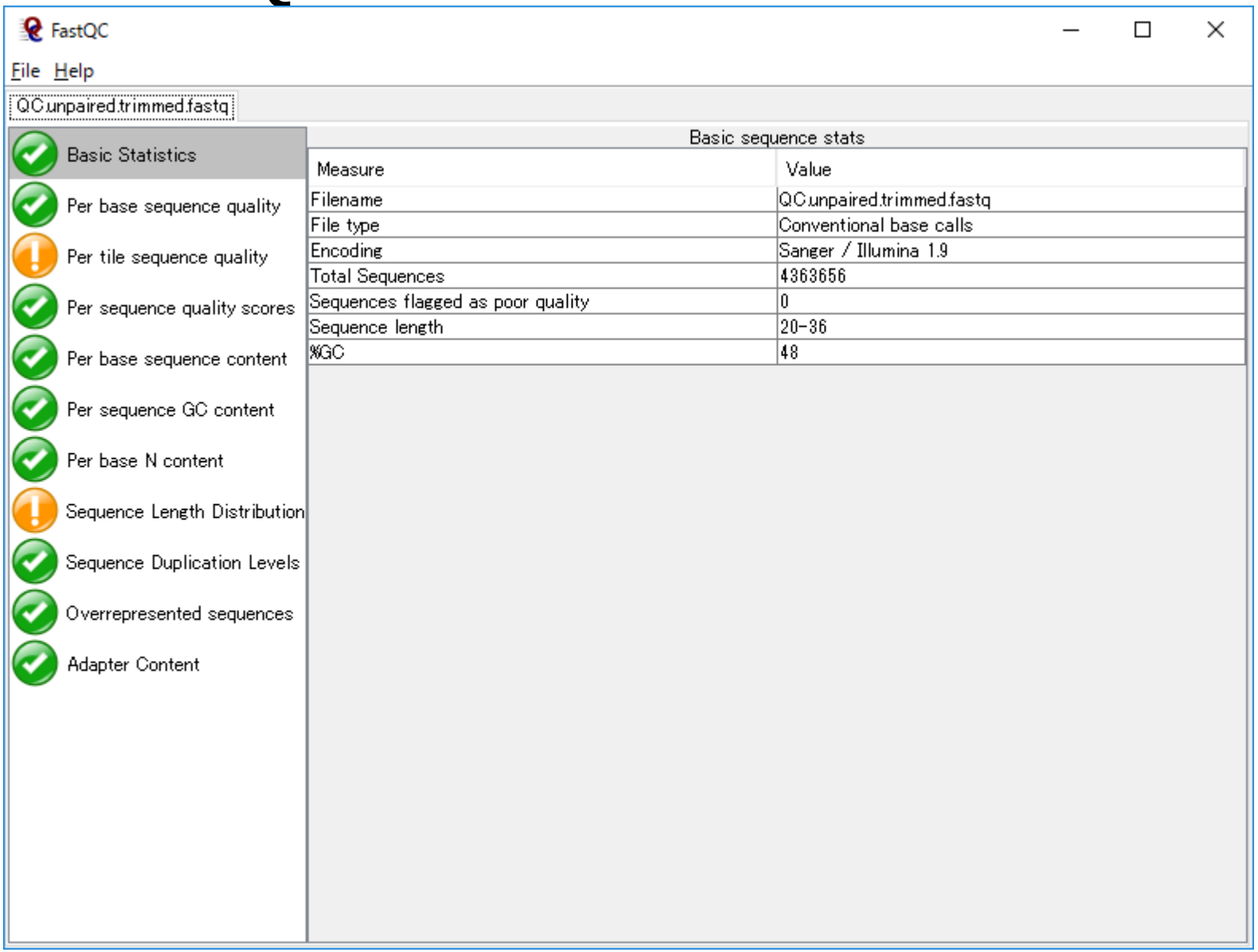
入力ファイルは、①FaQCsの実行結果ファイル

FastQCの入力ファイル



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…



①これです

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) ①
   - ShortRead : Morgan et al., Bioinformatics, 2
   - 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

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

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, 2
   - 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)

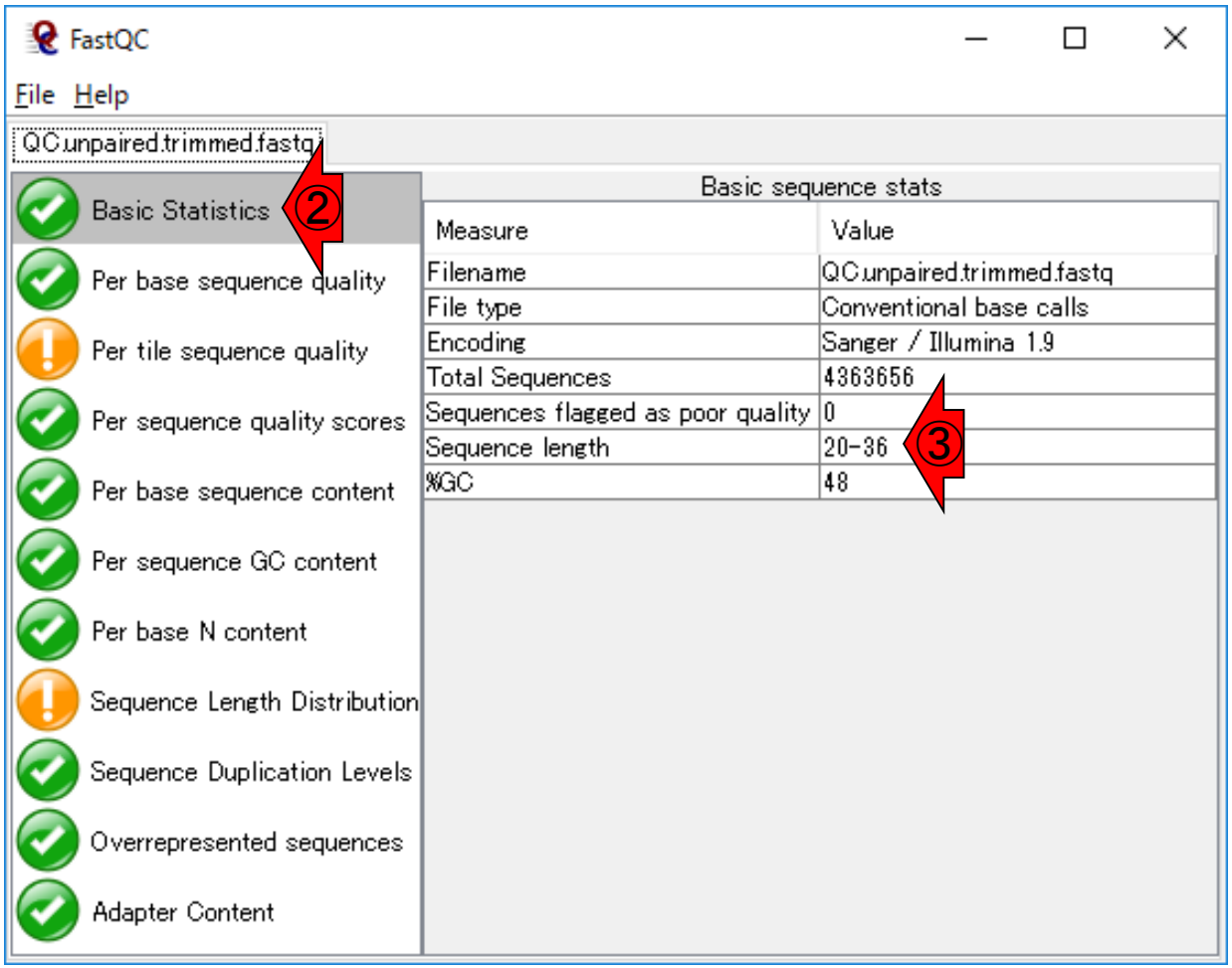
課題：論点1

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）

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, 2
   - 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)

①

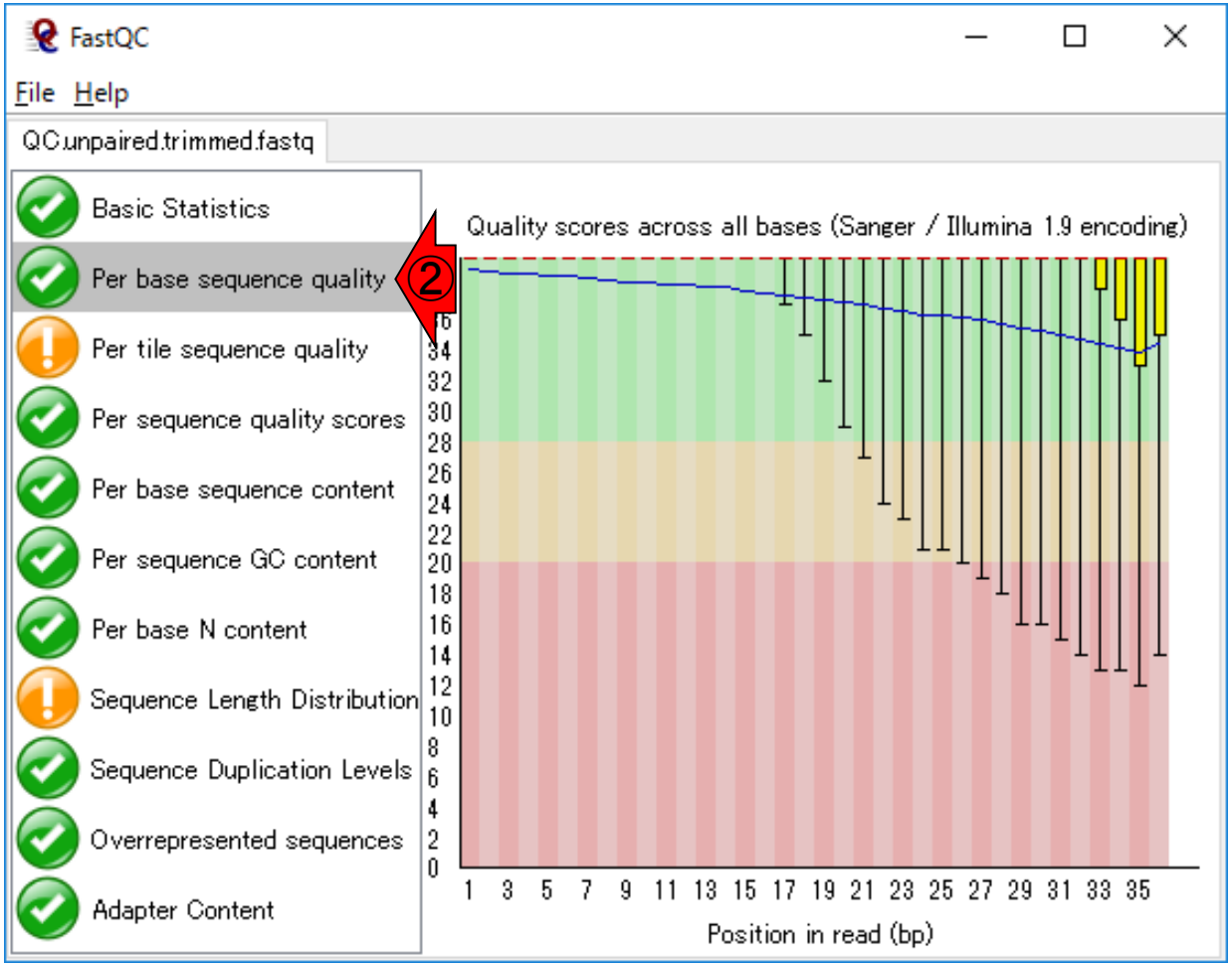


課題：論点2

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

講義日程（平成30年度）

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) ①
   - ShortRead : Morgan et al., Bioinformatics, 2
   - 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)

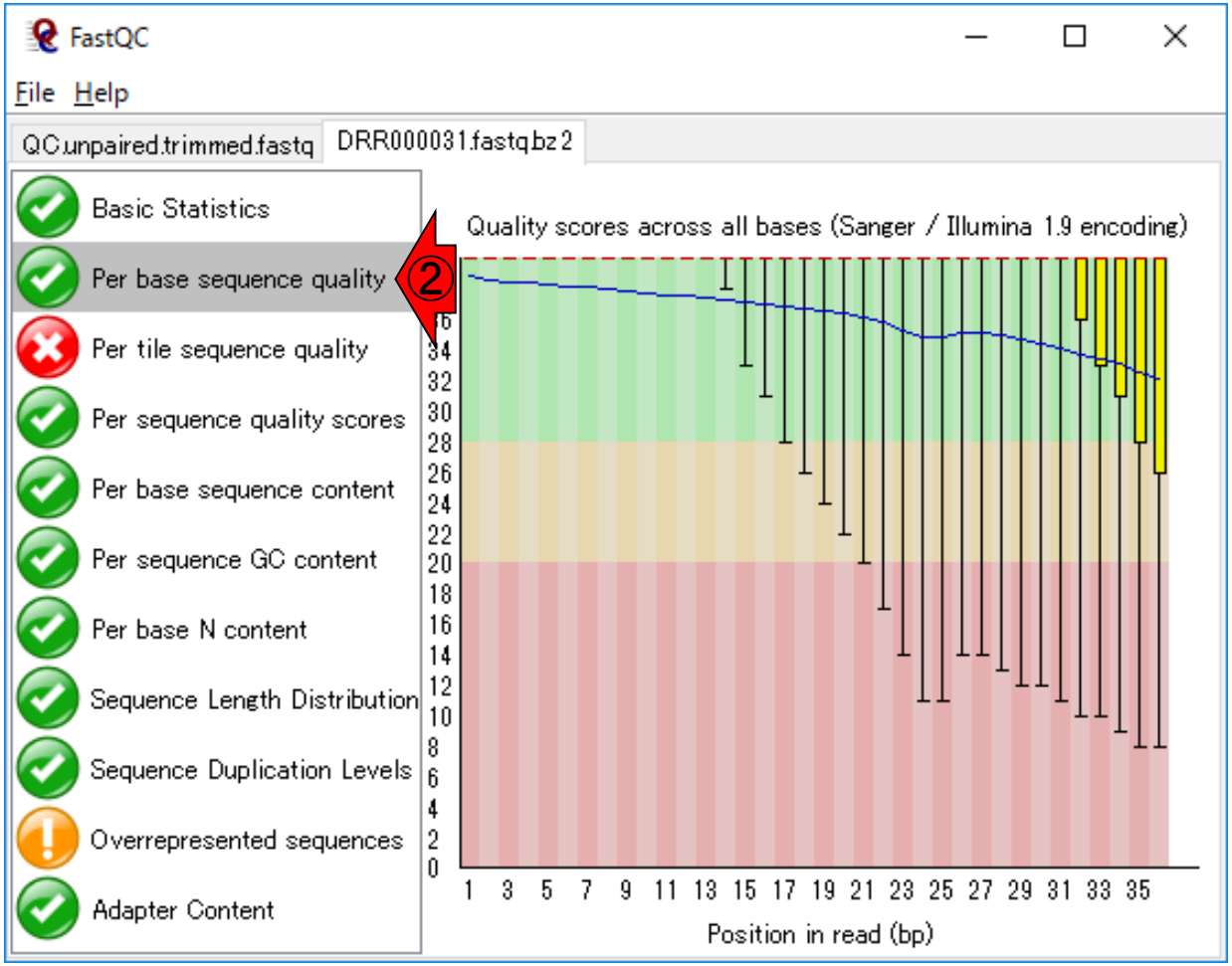


課題：論点2

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）

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) ①
   - 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)



課題：論点3

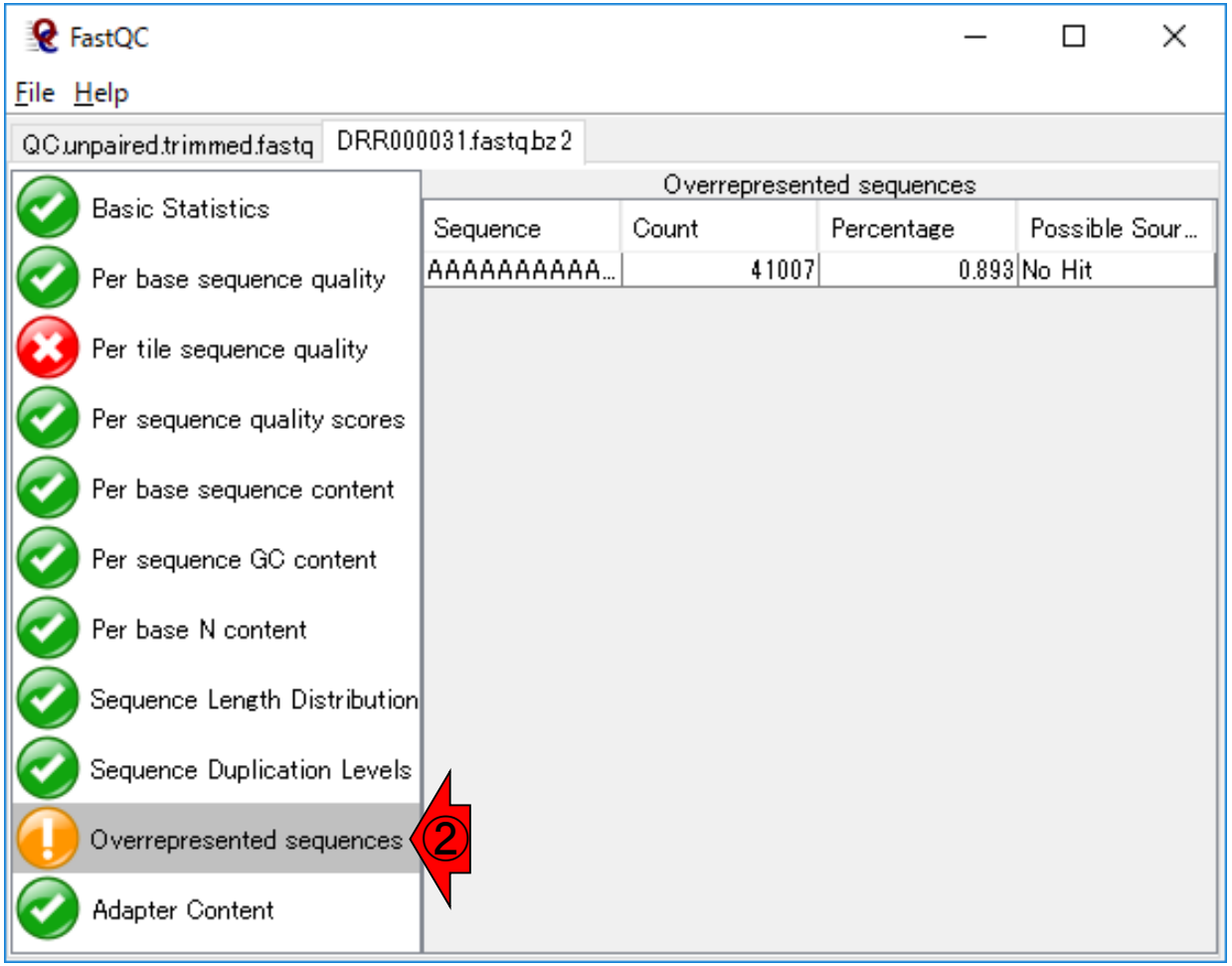
①FaQCs実行前の、②Overrepresented sequencesではpoly Aが見えていましたが…。③FaQCs実行後はどうなっているかなど…

講義日程（平成30年度）

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, 2
   - 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)

①

③



②

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、qrqc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）

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., 20
   - Rsubread(Windows版なし) : Liao et al., Nucleic Acids Res., 20
   - report.html(qrqcを用いた

①

③

Bioinformatics. 2009 Oct 1; 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 M[1], Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R.

Author information

Abstract

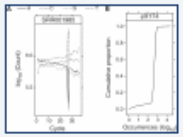
ShortRead is a package for input, quality assessment, manipulation and output of high-throughput sequencing data. ShortRead is provided in the R and Bioconductor environments, allowing ready access to additional facilities for advanced statistical analysis, data transformation, visualization and integration 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 Bioconduct [Bioinformatics. 2014]

QuasR: quantification and annotation [Bioinformatics. 2015]

R453Plus1Toolbox: an R/Biocondu [Bioinformatics. 2011]

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

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

See reviews...

See all...

②

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)

参考

ShortRead

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）

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 Con
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 Genor
   - 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)
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 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
   - ①
   - QC…ired.trimmed_fastqc.html(課題
   - ShortRead : Morgan et al., Bioinformat
   - Bioconductor : Gentleman et al., Genor
   - Rsubread(Windows版なし) : Liao et al., N
   - report.html(qrqcを用いた(

http://bioconductor.org/pa

Bioconductor - ShortRead

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help | Developers | About ⑤

Home » Bioconductor 3.7 » Software Packages » ShortRead

ShortRead ②③

platforms all | downloads top 5% | posts 5 / 1 / 2 / 3 | in Bioc 9.5 years
build warnin ④

DOI: 10.18129/B9.bioc.ShortRead

FASTQ input and manipulation

Bioconductor version: Release (3.7)

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.

Author: Martin Morgan, Michael Lawrence, Simon Anders

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

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

Morgan M, Anders S, Lawrence M, Aboyoun 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.org10.1093/bioinformatics/btp450.

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

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

参考

ShortRead

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

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 Con
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 Genor
   - 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)
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 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., Bioinformati
   - Bioconductor : Gentleman et al., Genor
   - Rsubread(Windows版なし) : Liao et al., N
   - report.html(qrqcを用いた(

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

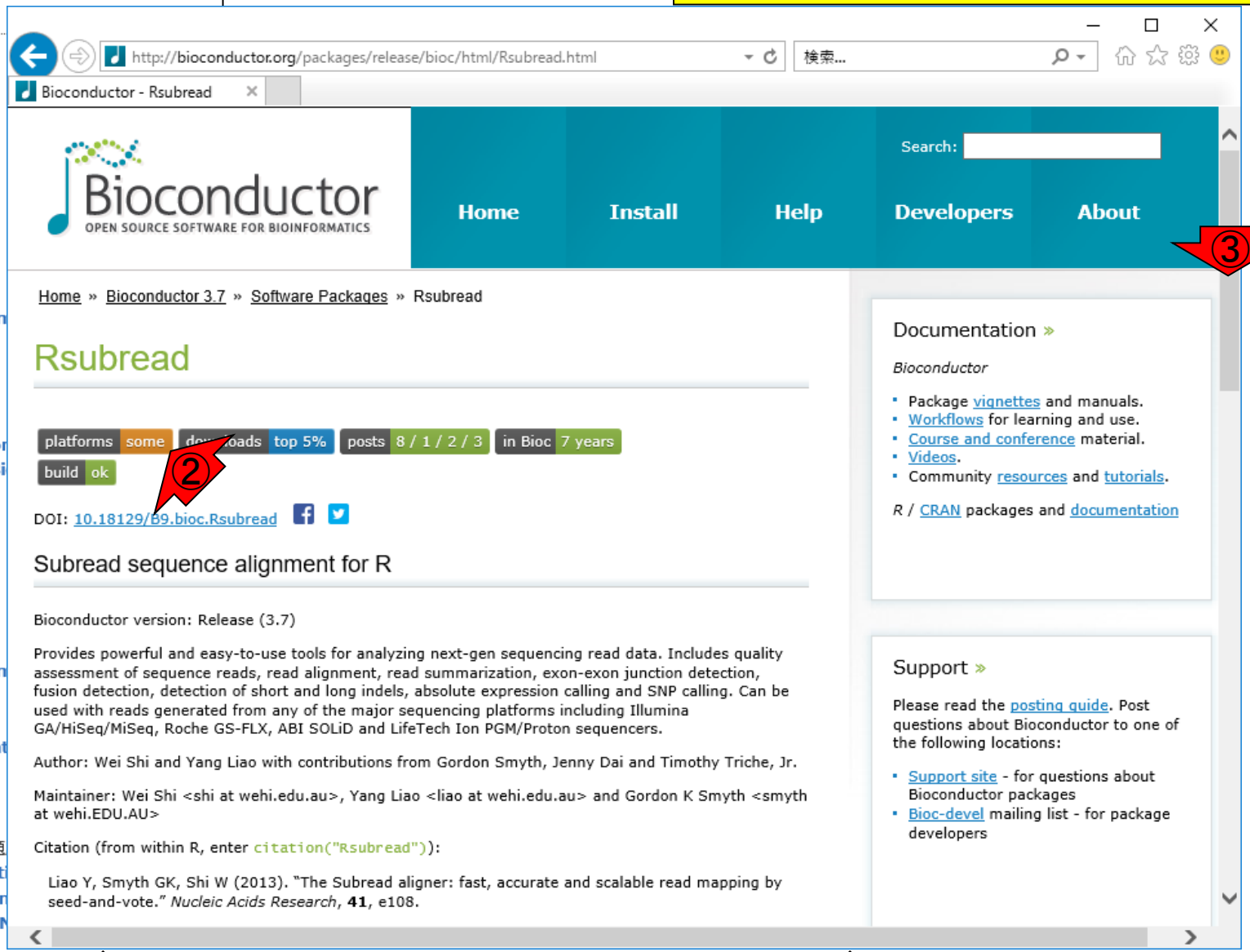
Rsubread

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）

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 Con
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 Genor
   - 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)
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 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
   - Sh…ad : Morgan et al., Bioinformati
   - Bio…uctor : Gentleman et al., Genor
   - Rsubread(Windows版なし) : Liao et al., N
   - report.html(qrqcを用いたQC結果)

①



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

参考

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）

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 Con
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 Genor
   - 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)
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 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., Bioinformati
   - Bioconductor : Gentleman et al., Genor
   - Rsubread(Windows版なし) : Liao et al., N
   - report.html(qrqcを用いたQC結果)

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||ついて
トール済みであるという前提で記述しています。初心者の
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 modifie
- 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 20

- 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 | 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 (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)。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%以上ということで実質的にPHREDスコアが閾値未が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。

②

```
in_f <- "DRR000031sub.fastq"              #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.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

#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
obj <- (rowSums(hoge < param1) <= width(fastq)*param2)#条件を満たすかどうかを判定
fastq <- fastq[obj]                       #objがTRUEとなる要素のみ抽出した結果をfast
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
```

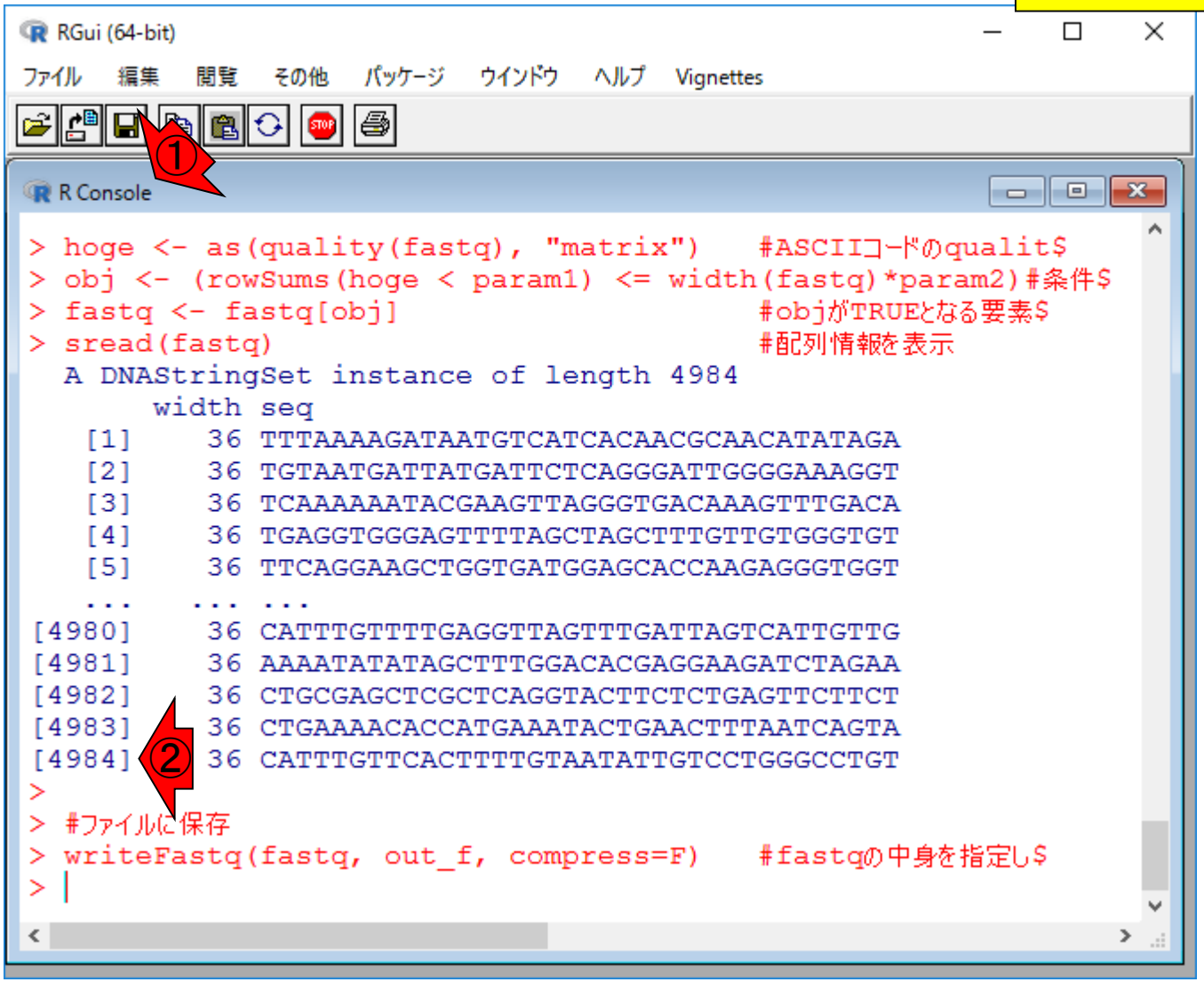
①

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)

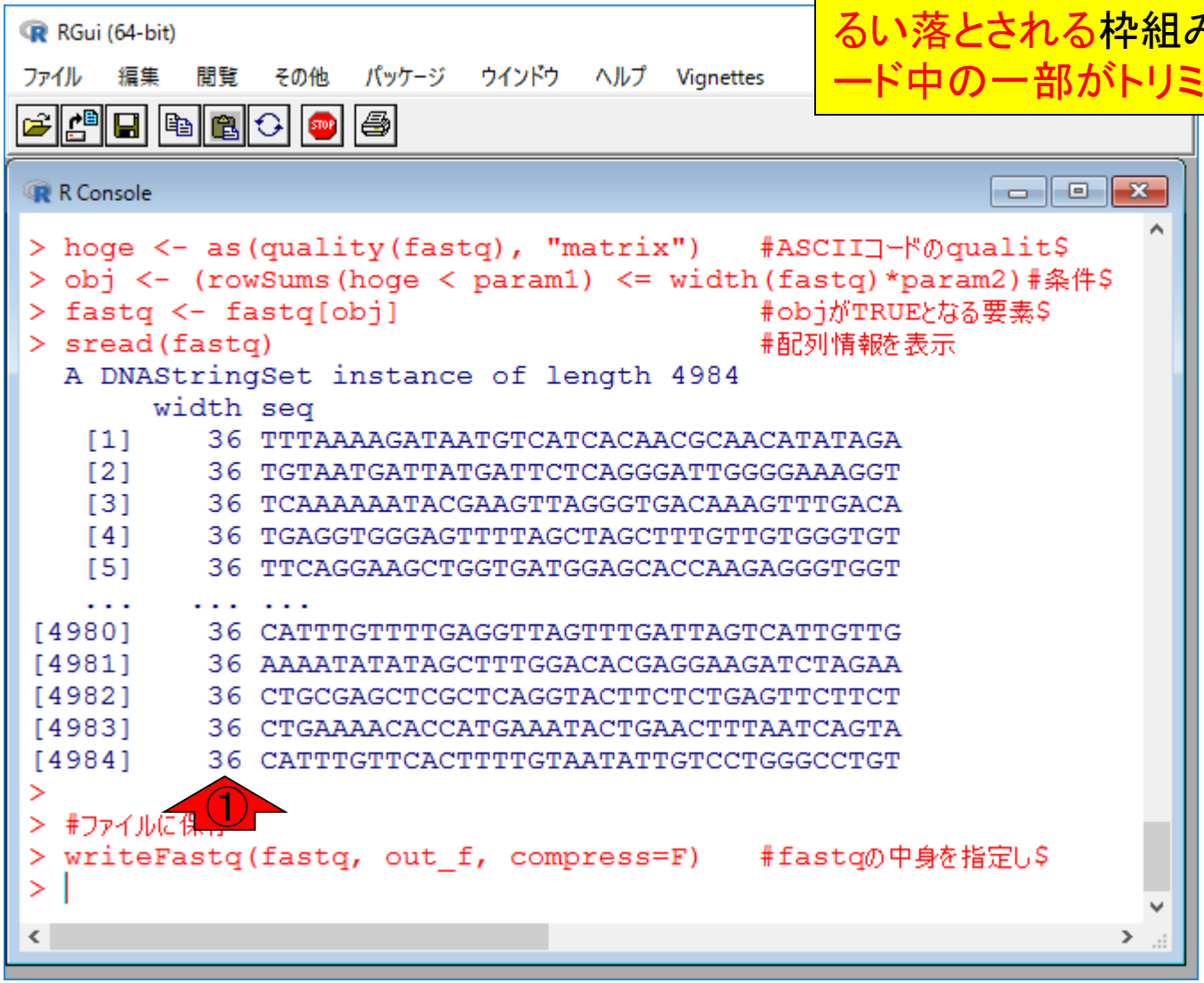
R実行結果画面

コピペ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



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



①出力ファイル中のリード数は、②4,984個

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

1. サンプルデータの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
2010)。PHREDスコアが20未満のものがリード長に占め
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
out_f <- "hoge1.fastq"                          #出
param1 <- 20                                    #PH
param2 <- 0.1                                   #指

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

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

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

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

①

RGui (64-bit)

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

R Console

```
> 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 TTTAAAAGATAATGTCATCACAACGCAACATATAGA
    [2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
    [3]    36 TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
    [4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
    [5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
    ...   ... ...
[4980]    36 CATTTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
[4981]    36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
[4982]    36 CTGCGAGCTCGCTCAGGTACTTCTCTGAGTTCTTCT
[4983]    36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
[4984]    36 CATTTGTTCACTTTTGTAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F)    #fastqの中身を指定し$
>
```

②

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、qrqc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qrqc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

1. サンプルデータの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
2010)。PHREDスコアが20未満のものがリード長に占め
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
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
out_f <- "hoge1.fastq"                          #出
param1 <- 20                                    #PH
param2 <- 0.1                                   #指

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

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

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

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

② (hoge1.fastq)

RGui (64-bit)

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

R Console

```
> 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 TTTAAAAGATAATGTCATCACAACGCAACATATAGA
    [2]    36 TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
    [3]    36 TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
    [4]    36 TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
    [5]    36 TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
    ...   ... ...
[4980]    36 CATTTGTTTTGAGGTTAGTTTGATTAGTCATTGTTG
[4981]    36 AAAATATATAGCTTTGGACACGAGGAAGATCTAGAA
[4982]    36 CTGCGAGCTCGCTCAGGTACTTCTCTGAGTTCTTCT
[4983]    36 CTGAAAACACCATGAAATACTGAACTTTAATCAGTA
[4984]    36 CATTTGTTCACTTTTGTAATATTGTCCTGGGCCTGT
>
> #ファイルに保存
> writeFastq(fastq, out_f, compress=F)    #fastqの中身を指定し$
>
```

① (4,984リード)

Rでクオリティチ

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

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
2010)。PHREDスコアが20未満のものがリード長に占め
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
out_f <- "hoge1.fastq"                      #出
param1 <- 20                                #PH
param2 <- 0.1                               #指

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

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

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

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

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケ

R Console

```
> hoge <- as(quality(f
> obj <- (rowSums(hoge
> fastq <- fastq[obj]
> sread(fastq)
  A DNAStringSet insta
        width seq
    [1]     36 TTTAAAAGA
    [2]     36 TGTAATGAT
    [3]     36 TCAAAAAAT
    [4]     36 TGAGGTGGC
    [5]     36 TTCAGGAAC
    ...    ... ...
[4980]     36 CATTTGTTT
[4981]     36 AAAATATAT
[4982]     36 CTGCGAGCT
[4983]     36 CTGAAAACA
[4984]     36 CATTTGTTC
>
> #ファイルに保存
> writeFastq(fastq, ou
>
```

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
   - ① 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)
   - ② 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

qrqcを実行

①qrqc。②の入力ファイル名部分をhoge1.fastqに変更して実行

(Rで)塩基配列解析

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

- イントロ | ファイル形式の変換 | qseq --> FASTA (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ (last modified 2013/06/17)
- 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 | 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 | 配列長分布を調べる
- 前処理 | クオリティチェック | Overrepr
- 前処理 | トリミング | ポリA配列除去
- 前処理 | トリミング | アダプター配列除
- 前処理 | トリミング | アダプター配列除
- 前処理 | トリミング | アダプター配列除
- 前処理 | トリミング | アダプター配列除
- 前処理 | トリミング | アダプター配列除
- 前処理 | トリミング | 指定した末端塩
- 前処理 | フィルタリング | PHREDスコ
- 前処理 | フィルタリング | PHREDスコ
- 前処理 | フィルタリング | ACGTのみ
- 前処理 | フィルタリング | ACGT以外

①

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。

②

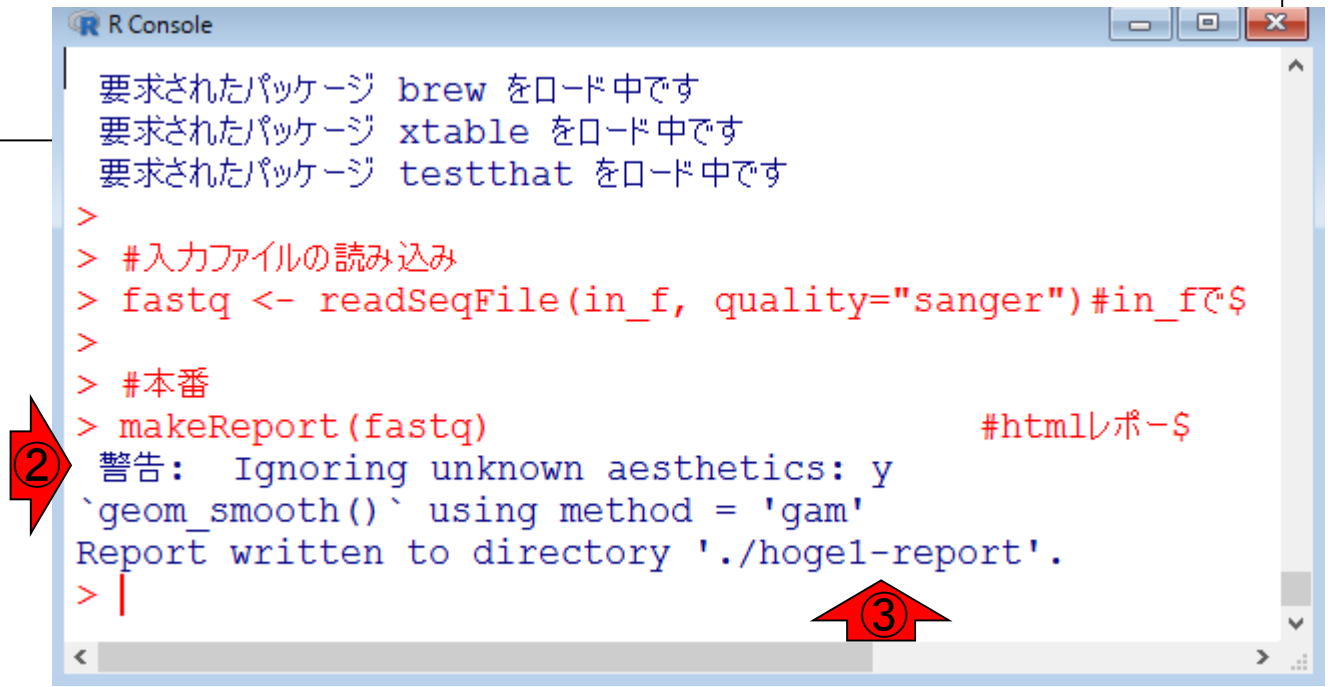
```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納

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

qrqcを実行

①入力ファイル名部分を変更してコピペ実行。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 <- "hoge1.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(qrqc)                            #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)
```

①



qrqc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

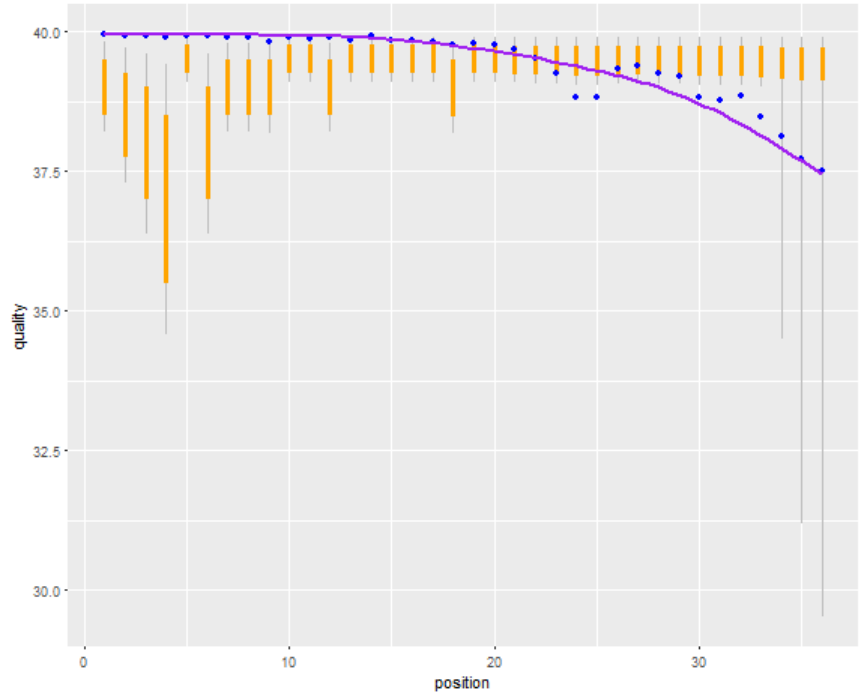
3. 平成30年05月22日
   - 講義資料PDF(約5MB; 2018.05.17版)
   - (Rで)塩基配列解析
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
   - Kraken : Davis et al., Methods, 2013
   - Lowe et al., PLoS Comput Biol., 2017
   - FaQCs : Lo and Chain, BMC Bioinformatics, 2014
   - FaQCs実行結果のQC.stats.txt
   - FaQCs実行結果のQC_qc_report.pdf
   - 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
   - ① 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

General Information

File: hoge1.fastq
Type: FASTQ
Sequence Length Range: 36 to 36
Total Sequences: 4984
Unique Sequences: 4453

③

② **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、qrqc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

①のQCが終わったので…

全体像のおさらい

NGSリードデータ（SRAファイル）

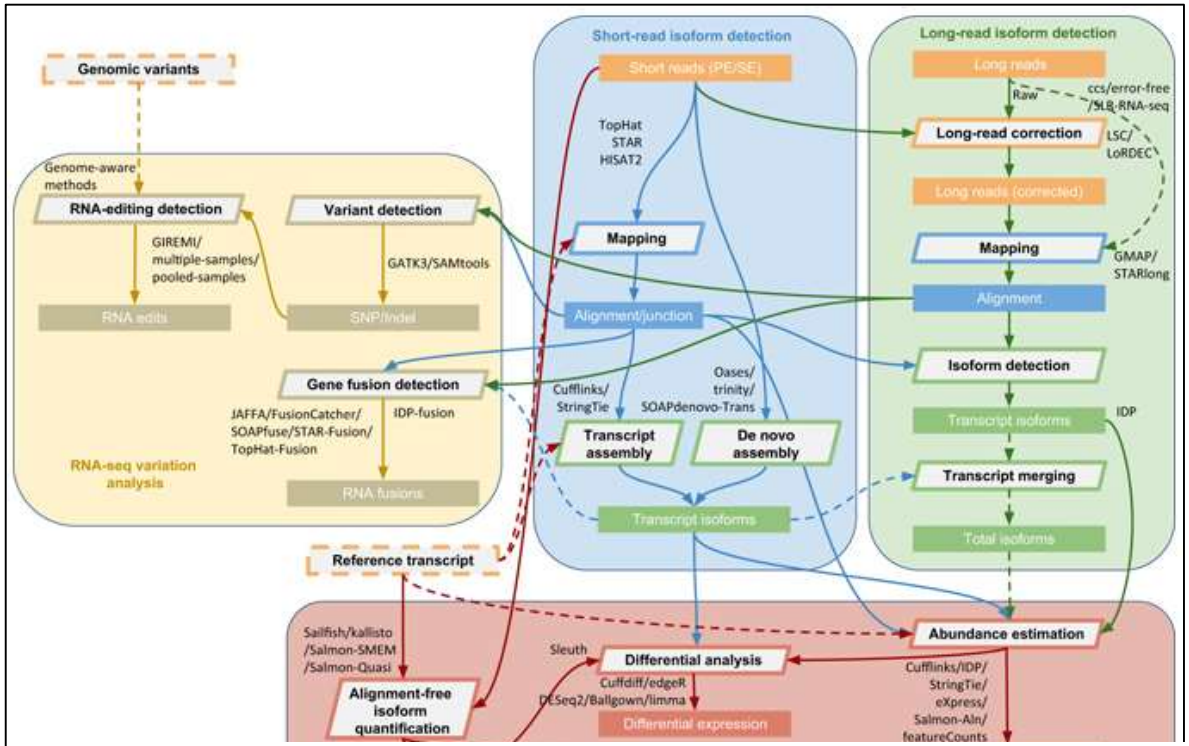
⇩

NGSリードデータ（FASTQファイル）

⇩

前処理（preprocessing） or Quality Control （QC） ①

⇩



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

全体像のおさらい

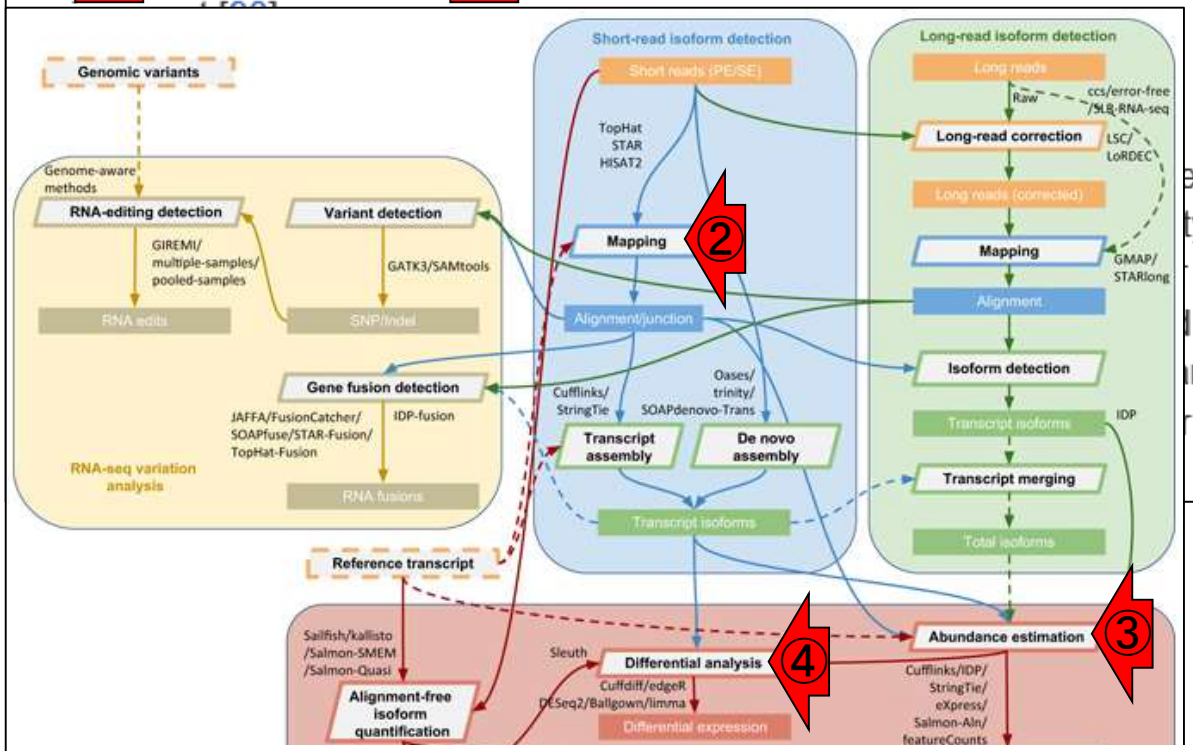
①のQCが終わったので、次は
②マッピング（アラインメント）

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 experim… design …als. 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 com…d-line interface, eith… a Unix environment or within the R/Bioconductor statistical

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

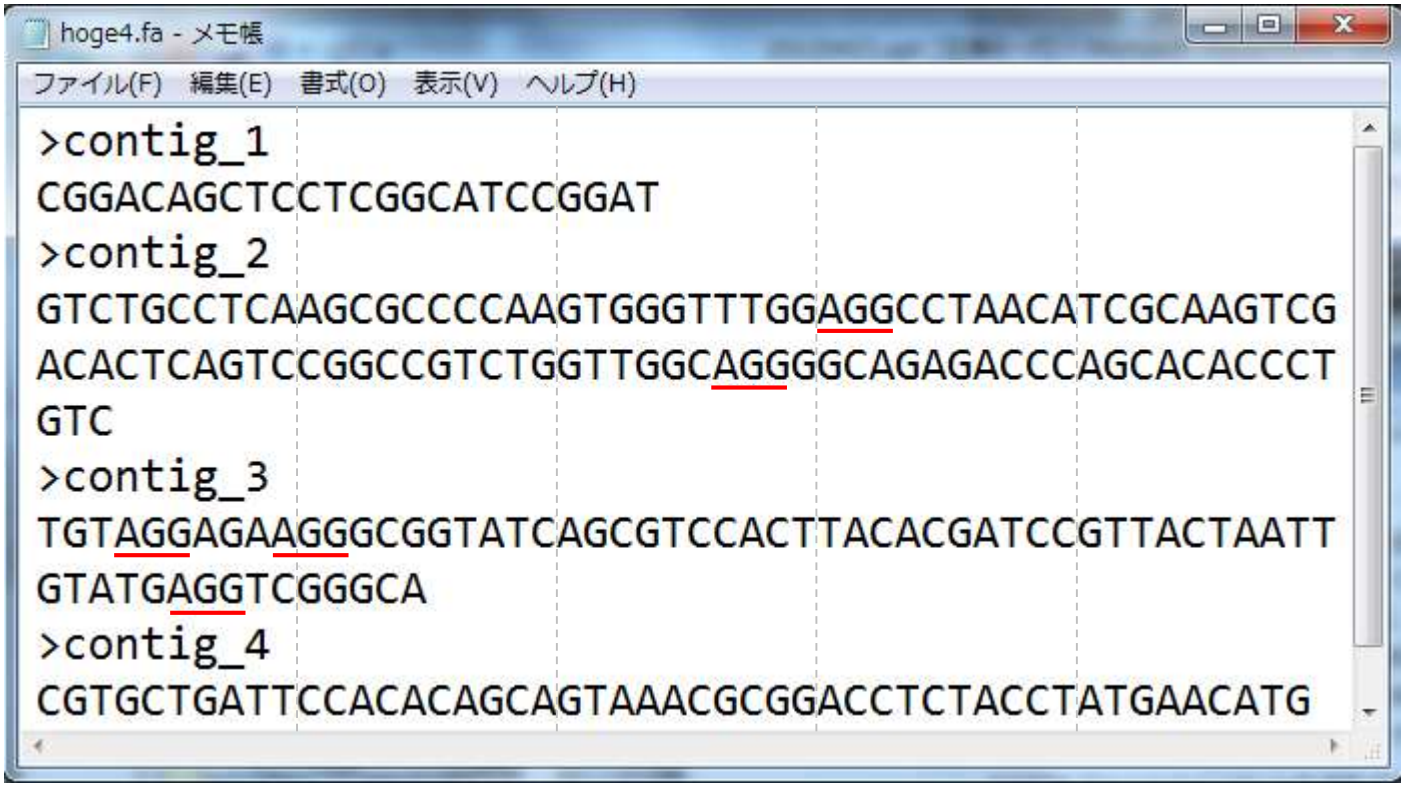


RNACocktail（Sahraeian et al., Nat Commun., **8**: 59, 2017）

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

- 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”



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年となってはいるが、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 (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(Gaidat
- マッピング｜single-end｜ゲノム｜basic aligner(応用)｜QuasR(Gaidat
- マッピング｜single-end｜ゲノム｜splice-aware aligner｜QuasR(Gaida
- マッピング｜paired-end｜ゲノム｜basic aligner(基礎)｜QuasR(Gaidat
- マッピング｜paired-end｜ゲノム｜basic aligner(応用)｜QuasR(Gaidat
- マッピング｜paired-end｜トランスクリプトーム｜basic aligner(基礎)｜
- マッピング｜paired-end｜トランスクリプトーム｜basic aligner(応用)｜
- 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(Morg
- 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(Gaidat
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidat
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaida
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidat
- マッピング | paired-end | ゲノム | basic aligner(応用) | QuasR(Gaidat
- マッピング | paired-end | トランスクリプトーム | basic aligner(基礎) |
- マッピング | paired-end | トランスクリプトーム | basic aligner(応用) |
- 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

②　①

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

③

マッピング

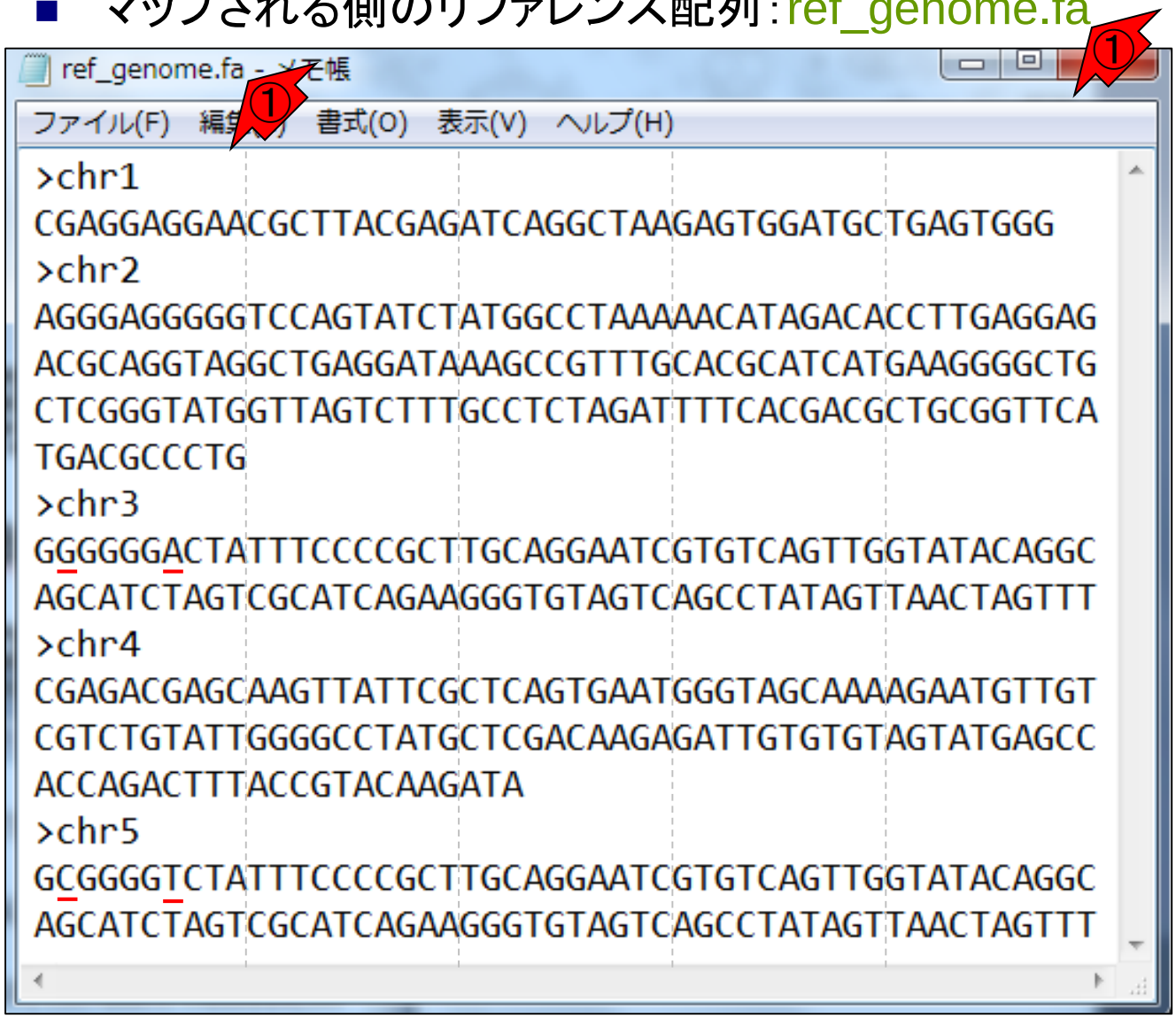
オプションは、デフォルトである程度よきに計らってくれるが…実際の挙動を完全に把握できる状況で様々なオプションを試したい

- 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か所ということを意味し、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

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

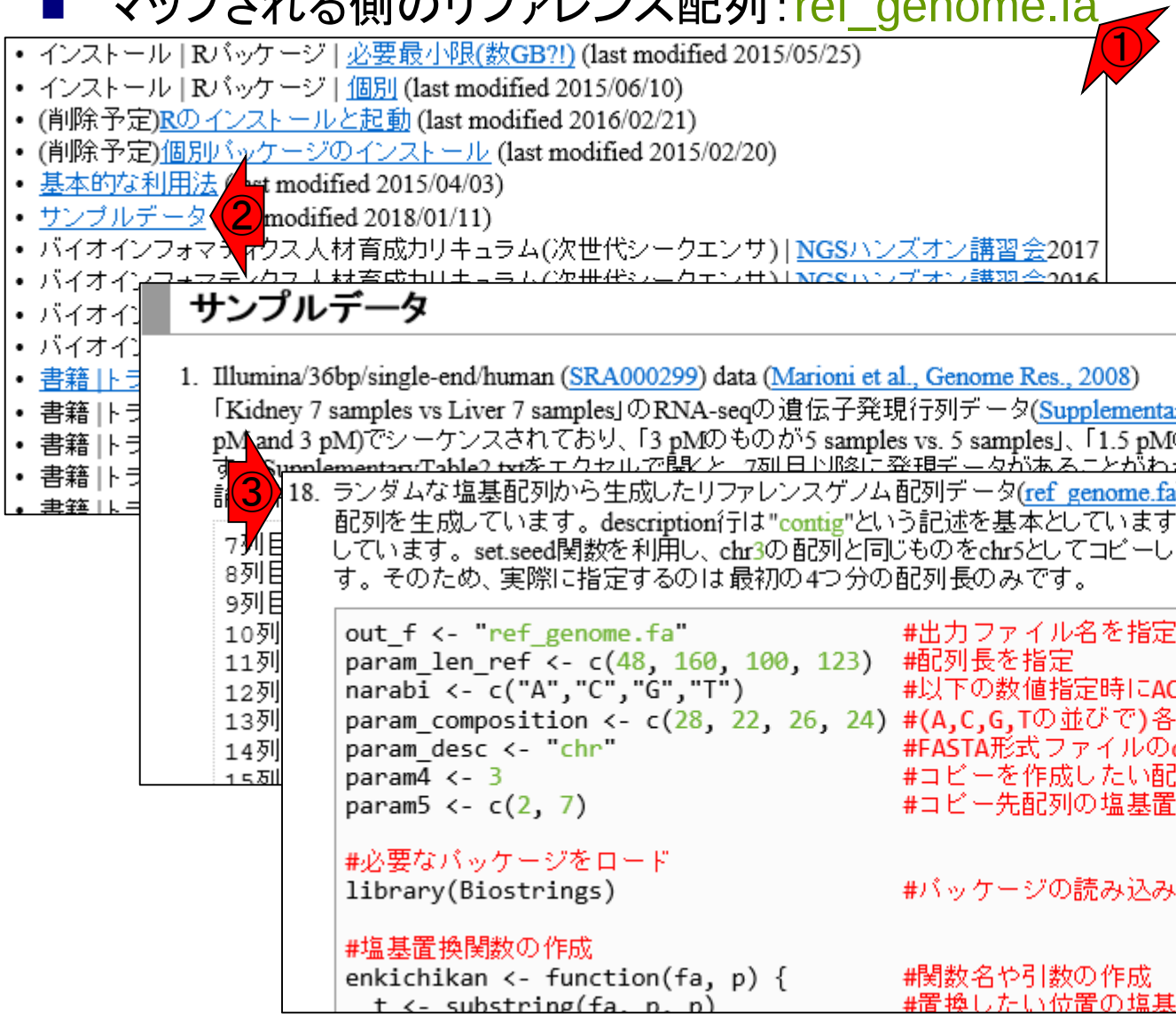
- マップされる側のリファレンス配列：ref_genome.fa



マップされる側

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

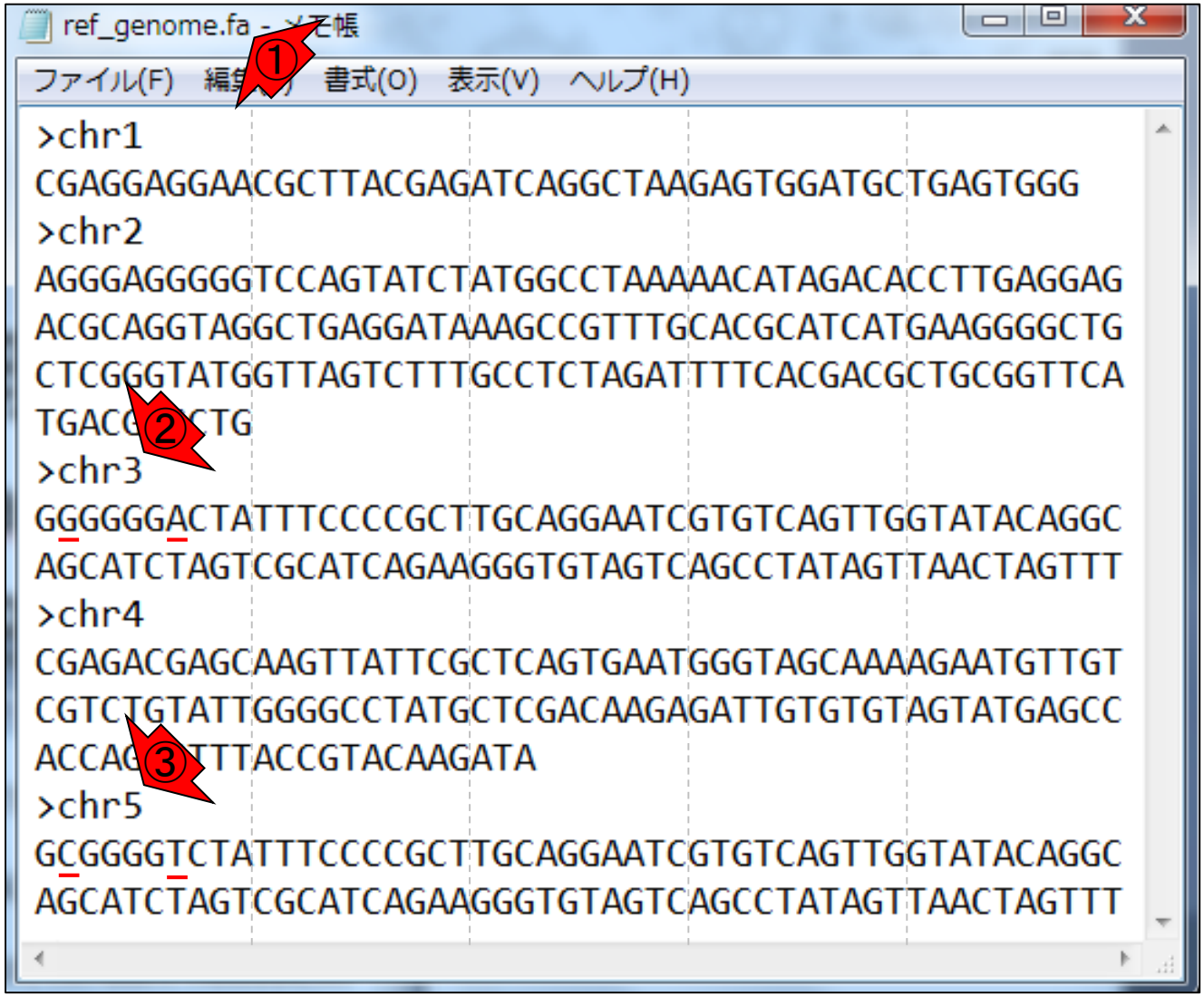
- マップされる側のリファレンス配列：ref_genome.fa



①ref_genome.faの説明。②chr3と③chr5の違いは…

マップされる側

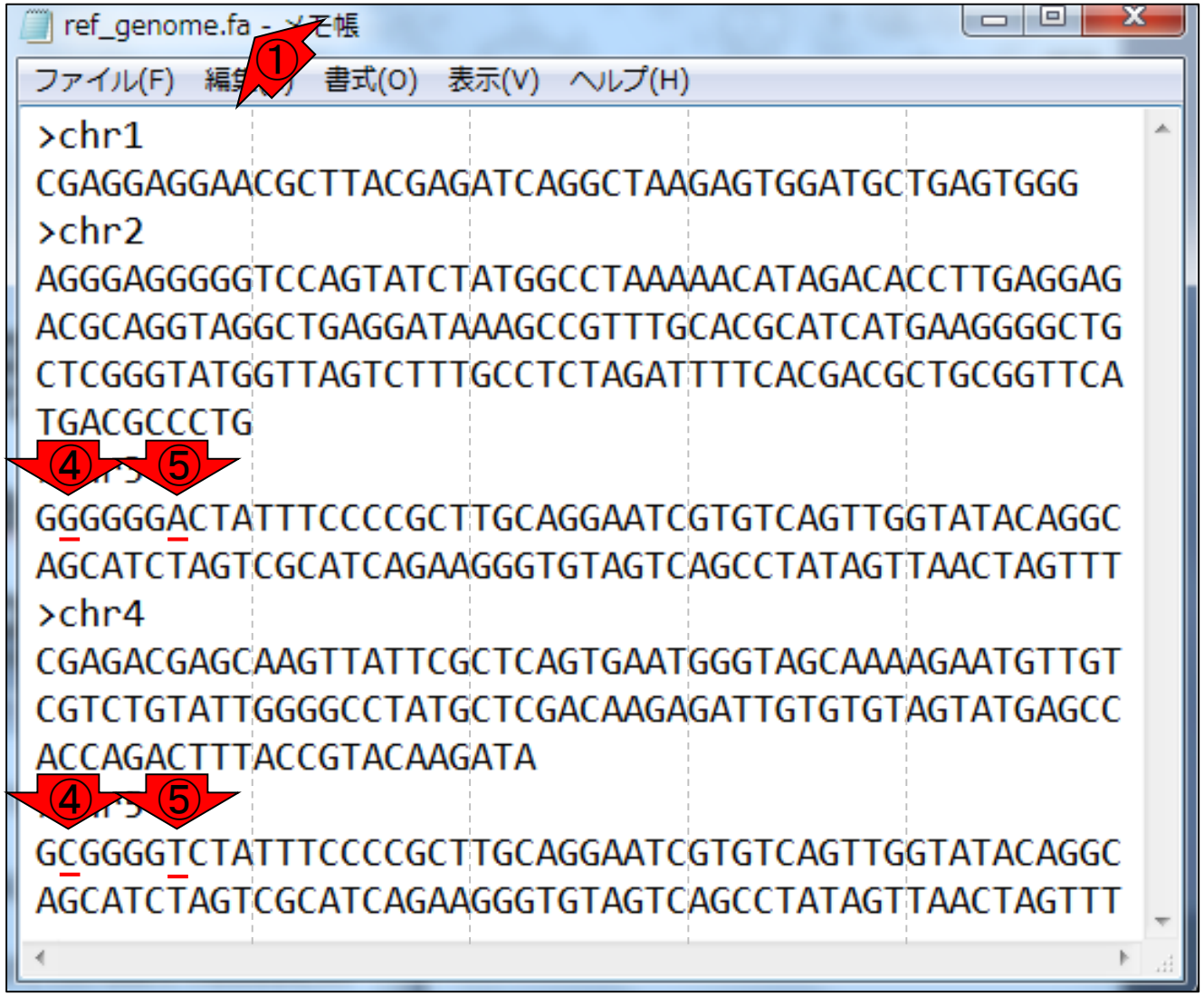
- マップされる側のリファレンス配列：ref_genome.fa



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



マップする側

マップする側のリードファイルは①sample_RNAseq1.fa。許容するミスマッチ数による違いや、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
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

①

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
```

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(次世代シークエンサ) | NGSハンズオン講習会2017

① ②

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
```

TGAGGAG

AAAACAT

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。詳細な情報は以下の通りです(原著論

③

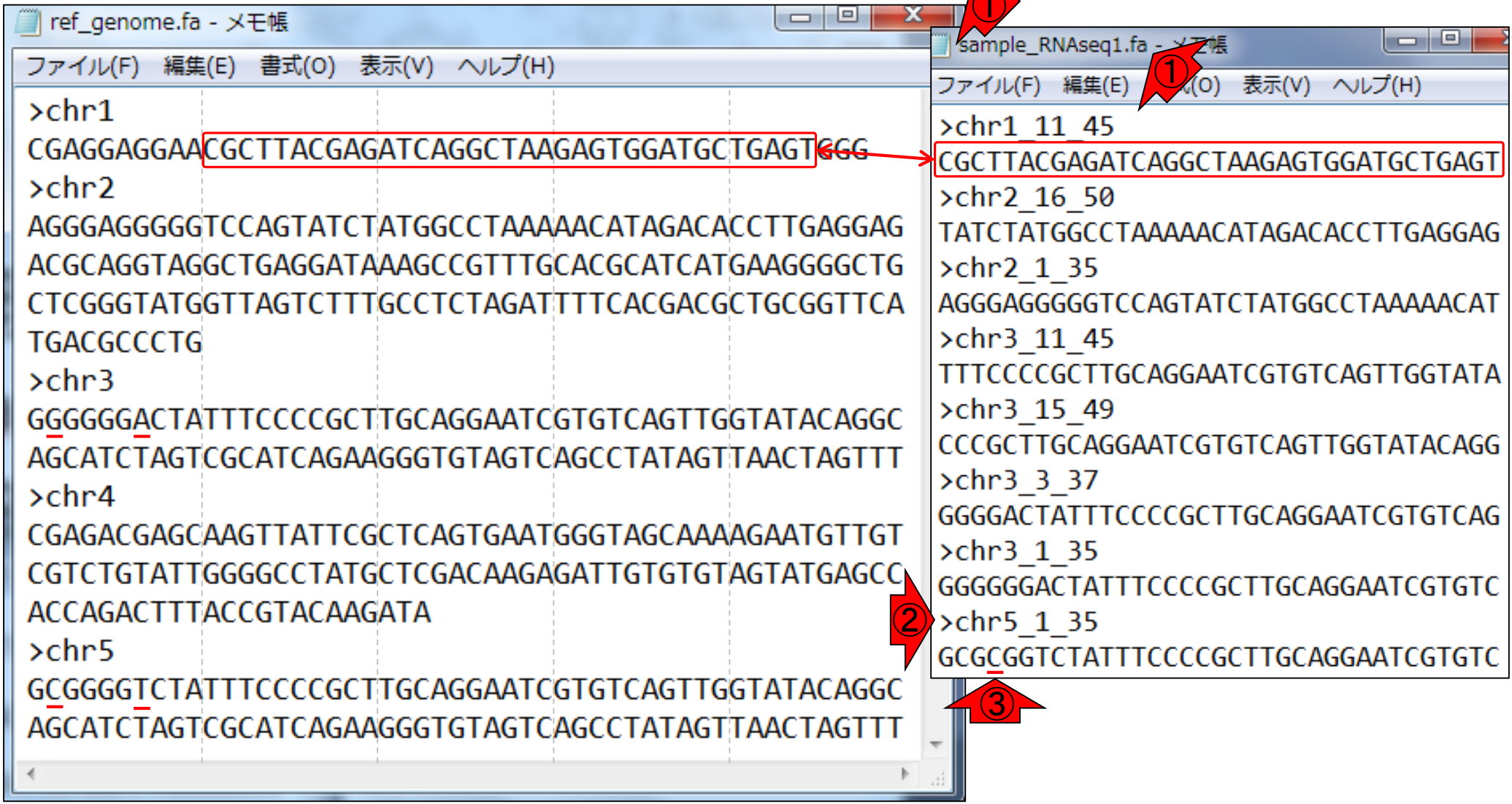
19. 上記リファレンスゲノム配列データ(ref_genome.fa)に対してbasic alignerでマッピングする際の動作確認用RNA-seqデータ(sample_RNAseq1.fa)とそのgzip圧縮ファイル(sample_RNAseq1.fa.gz)。リファレンス配列を読み込んで、list_sub3.txtで与えた部分配列を抽出したものです。どこに置換を入れているかがわかっているので、basic alignerで許容するミスマッチ数を変えてマップされるorされないの確認ができます。DNAStringSetオブジェクトを入力として塩基置換を行うDNAString_chartr関数を用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

マップする側

①sample_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



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、qrqc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 | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

③

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(uniquely 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が置いてあるディレクトリに移動し以下をコピペ。

② ①

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 | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

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(uniquely 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ペ。

②　①　③

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):

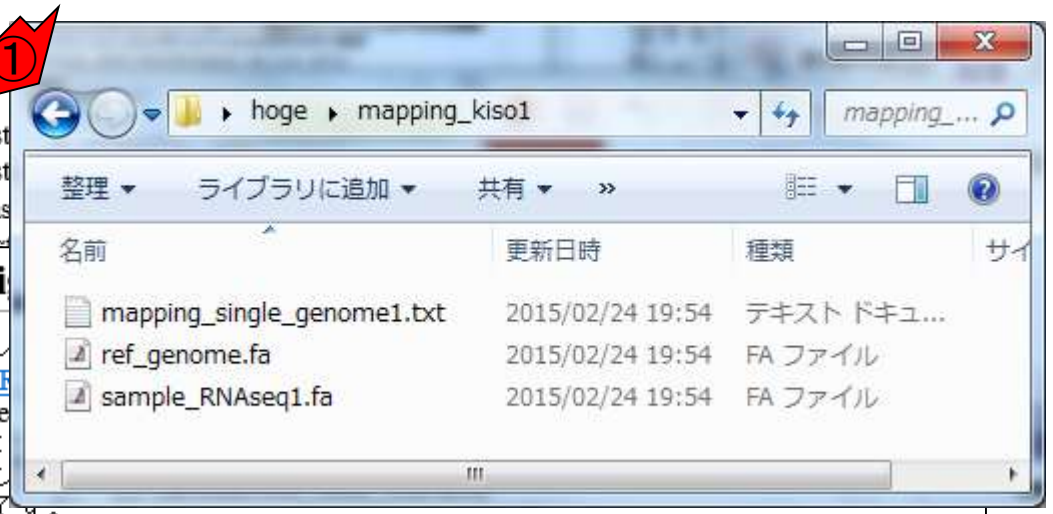
オプションを"-m 1 --best --strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
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

①のような感じにして、②R上でも確認しましょう

- 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
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (las
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last
- マッピング | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ



マッピング | single-end | ゲノム | basic ali

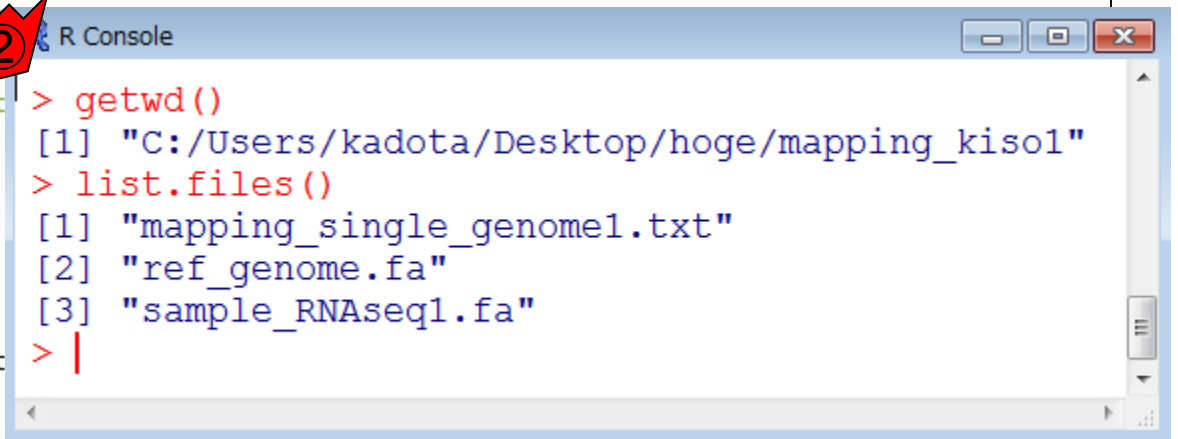
QuasRパッケージを用いてsingle-end RNA-seqデータのリファレ
あるBowtie (Langmead et al., Genome Biol., 2009)を実装した R
されるリードの取り扱い(uniquely mapped reads or multi-mappe
どの様々なオプションを利用可能ですが、「基礎」のところでは
て挙動を確認したり、複数のRNA-seqファイルを一度にマッピン
"*.bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「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ペ。

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.tx
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best --strat

#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, alignment
time_e <- proc.time()
time_e - time_s
out
alignmentStats(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
#マッピング結果(alignment statistics)の表示。seglength：リファレンス
```

①例題1の入力は、②リファレンス配列のファイルと…

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 | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

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(uniquely 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ペ。

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。

②

```
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 | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

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
あるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内
されるリードの取り扱い(uniquely mapped reads or multi-mapped reads)を"-m"オプショ
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
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
"*.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ペ。

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。

②

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	namae

①

```
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
```

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) (la
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (la
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (l
- 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 | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

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
あるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内
されるリードの取り扱い(uniquely mapped reads or multi-mapped reads)を"-m"オプショ
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
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
"*.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ペ。

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。

②

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	namae

①
③

```
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
```

コピペ実行

コピペ実行時の途中経過。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(uniquely 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)
```

①

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
```

途中経過

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

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(uniquely 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_ time_s
out
alignStats(out)
```

①

②

R Console

```
> out                                          #マッピングに用いたパラ$
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

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

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

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(uniquely 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_ time_s
out
alignStats(out)
```

①

②

R Console

```
> out                                          #マッピングに用いたパラ$
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
```

②

③

④

⑤

途中経過

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

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

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パッケージをされるリードの取り扱い(uniquely mapped reads or multi-mapped reads)を"-m"オプ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。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"*.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_
out
alignStats(out)
```

①

②

R Console

```
> out                                    #マッピングに用いたパラ$
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
```

②

③

④

⑤

⑥

⑦

①コードの下部を表示。次は②alignmentStats

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(uniquely 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ペ。

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])#Quqlity Controlレポートのpdfファイル名を作成した結果をo
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
}
```

②

①

①alignmentStatsの結果はこんな感じ

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(uniquely 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よいです。

「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="
}
```

①

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
namae: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)
```

①

マッピング結果の概要

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

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

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(uniquely 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よいです。

「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="
}
```

① ②

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
namae: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)
```

① ② ③ ④

こんな感じになります

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(uniquely 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よいです

「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

```
> 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$
+ }
>
```

①list.files() 。②の3つ以外全てが出力ファイル

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(uniquely 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よいです。

「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
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="
}
```

①

②

②

②

R Console

```
+   write.table(tmp, out_f, sep="\t", append=F, quote=F, row.nam$
+ }
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> 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"
>
```

②

主なマッピング結果ファイルは

①拡張子が.bamの、通称BAMファイルと呼ばれるものです

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(uniquely 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よいです。

「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.nam$
+ }
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> 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

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、qrqc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

実用上はBAM形式、視覚上はBED形式

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)

  …

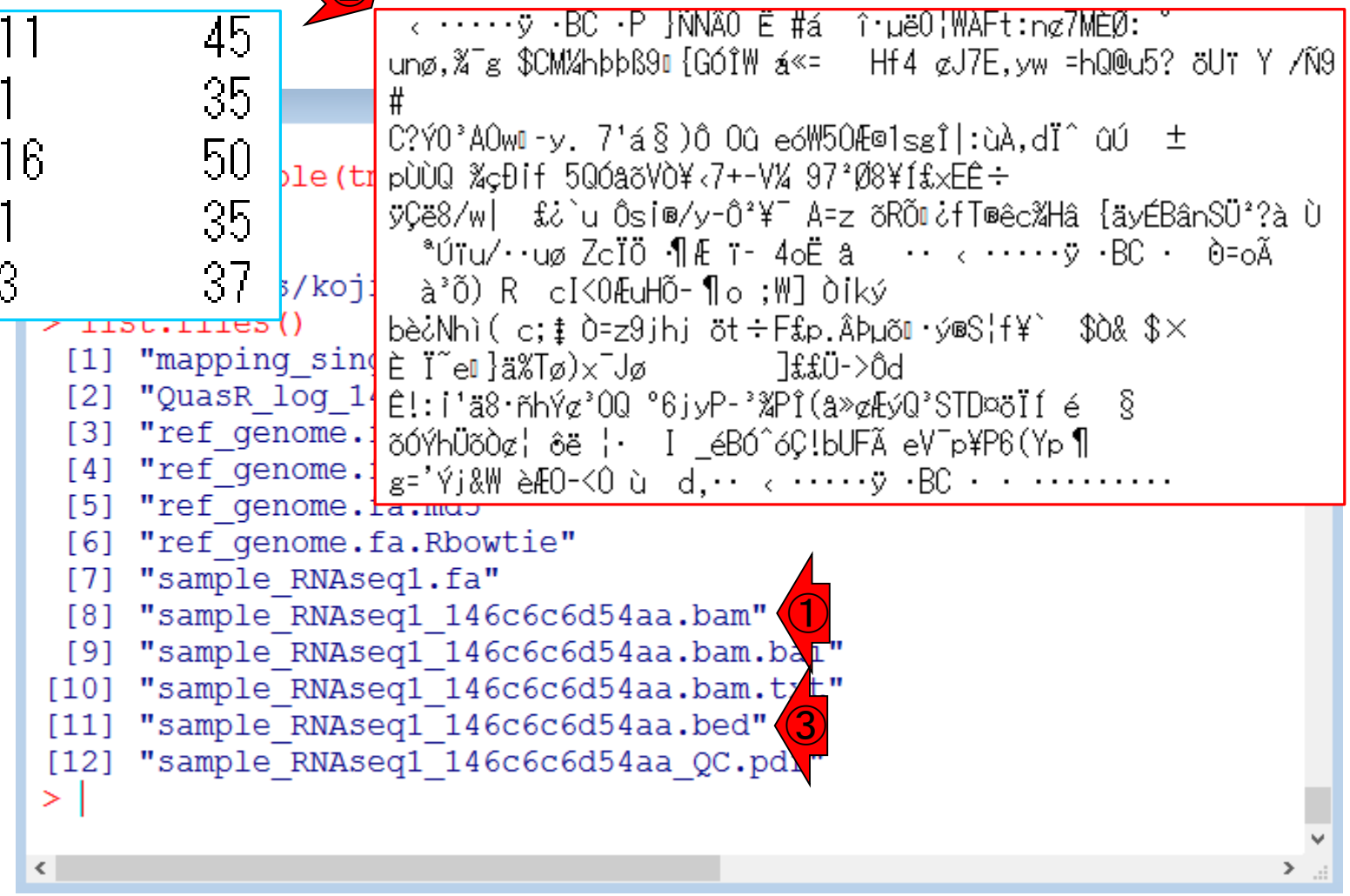
出力ファイル形式

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

④

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

②



ちなみに①BEDファイルは、②で作成しています

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 (La…ead 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扱い(uniquely 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…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…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。

「ファイル」ー「ディレ…ディレクトリに移動し以下をコピペ。

1.サンプルデータ1…f_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1…q1.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"…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…。

①

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

```
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])#Quqlity Controlレポートのpdfファイル名を作成した結果をo
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
  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
}
```

②

使用オプションと結果の解釈

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

- “-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)

```
>chr1
CGAGGAGGA                                   TGGATGCTGAGTGGG
>chr2
AGGGAGGGC                                   ATAGACACCTTGAGGAG
ACGCAGGTA                                   GCATCATGAAGGGGCTG
CTCGGGTAT                                   CACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

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

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
```

使用オプションと結果の解釈

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

- “-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
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
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
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

使用オプションと結果の解釈

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

- “-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
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
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)

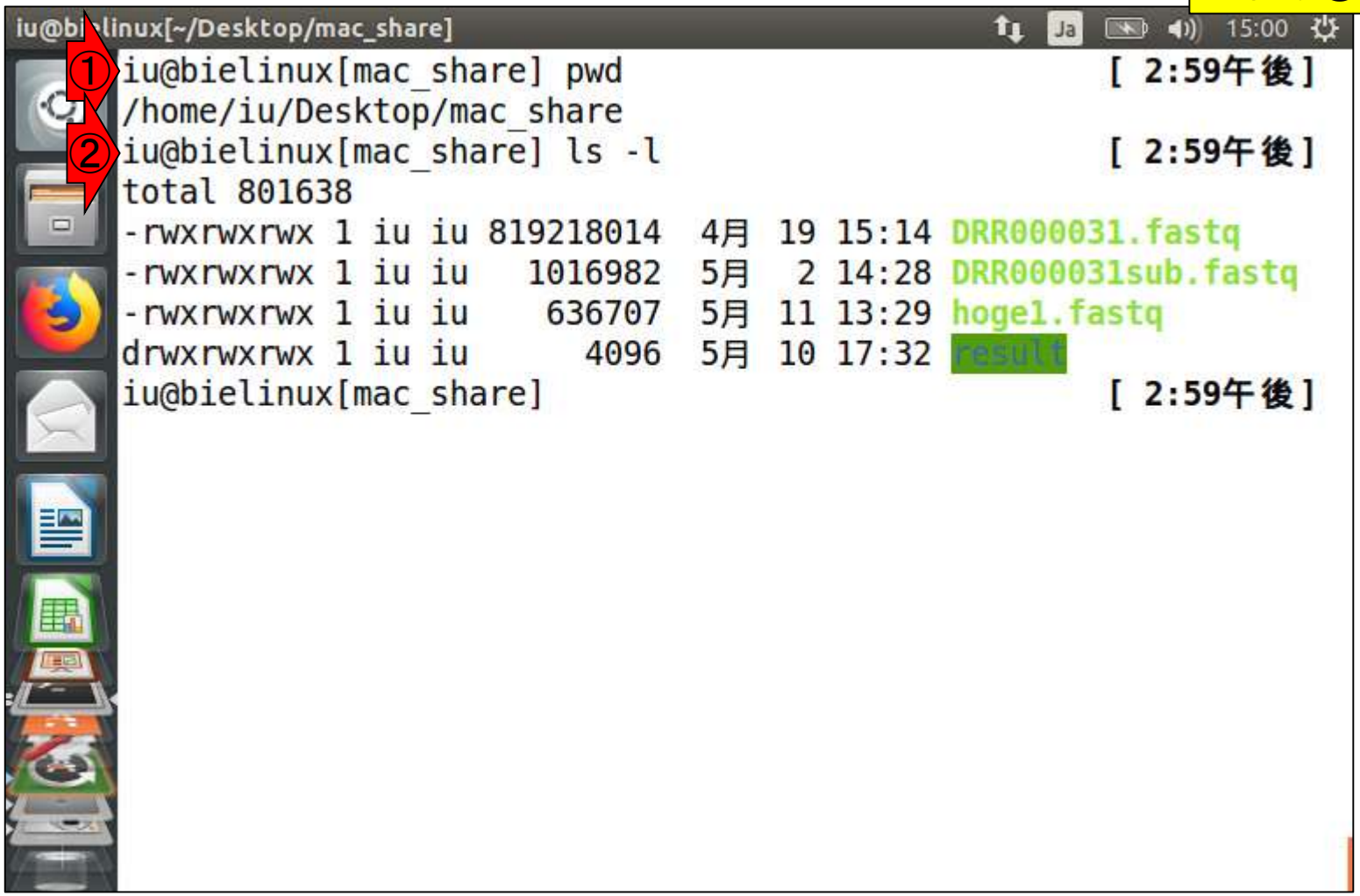
```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

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、qrqc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

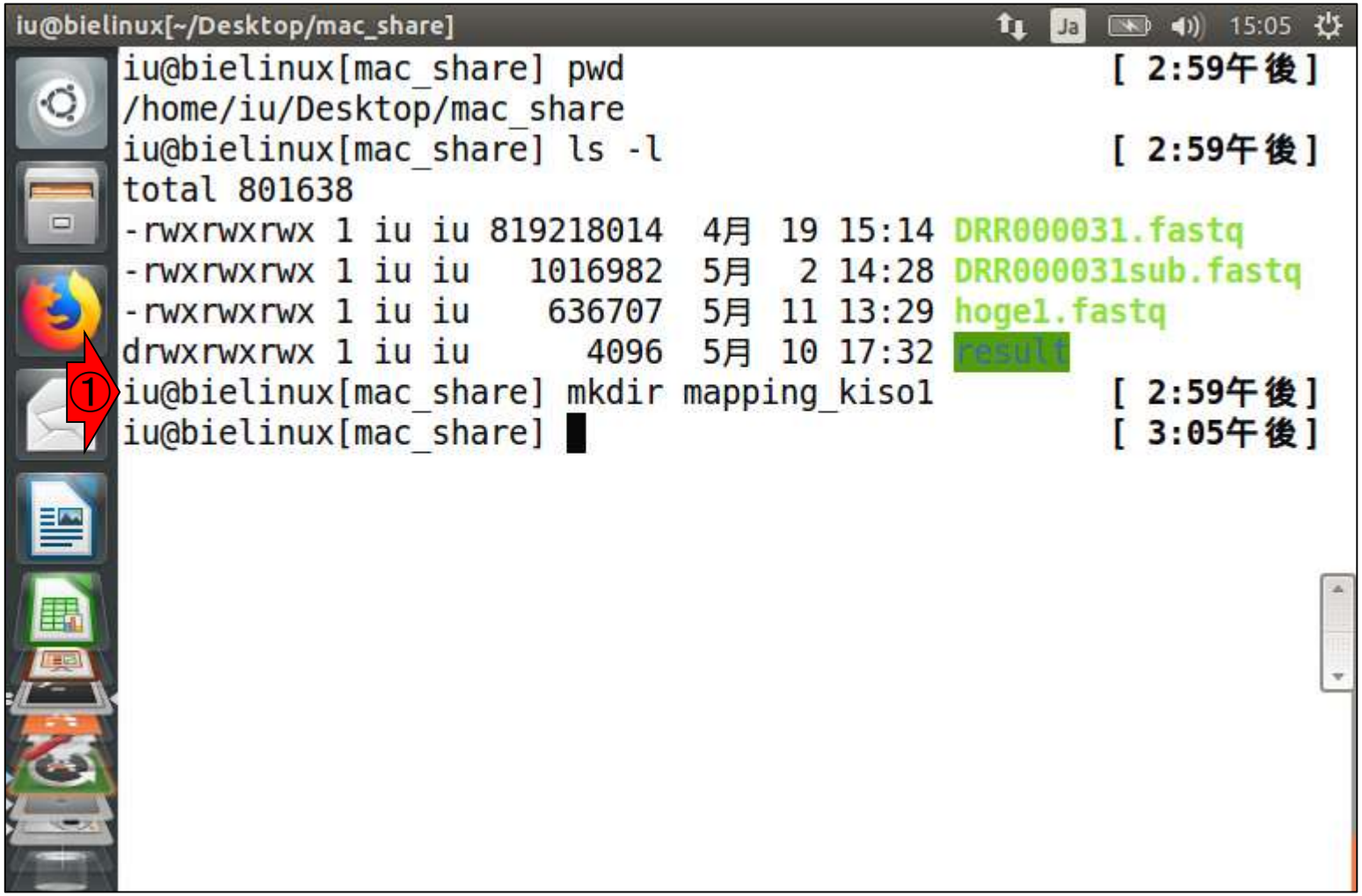
ディレクトリの作成

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



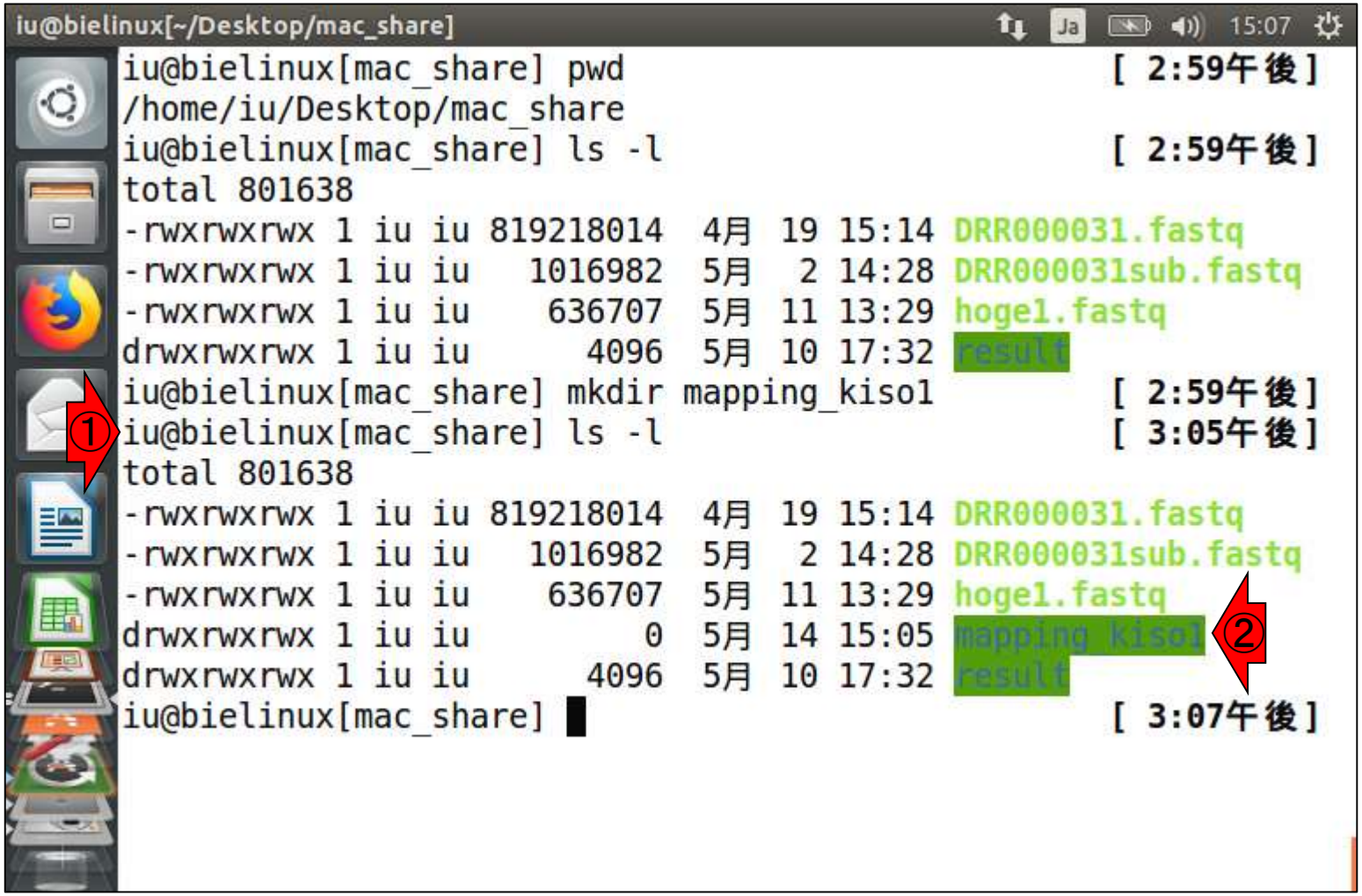
ディレクトリの作成

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



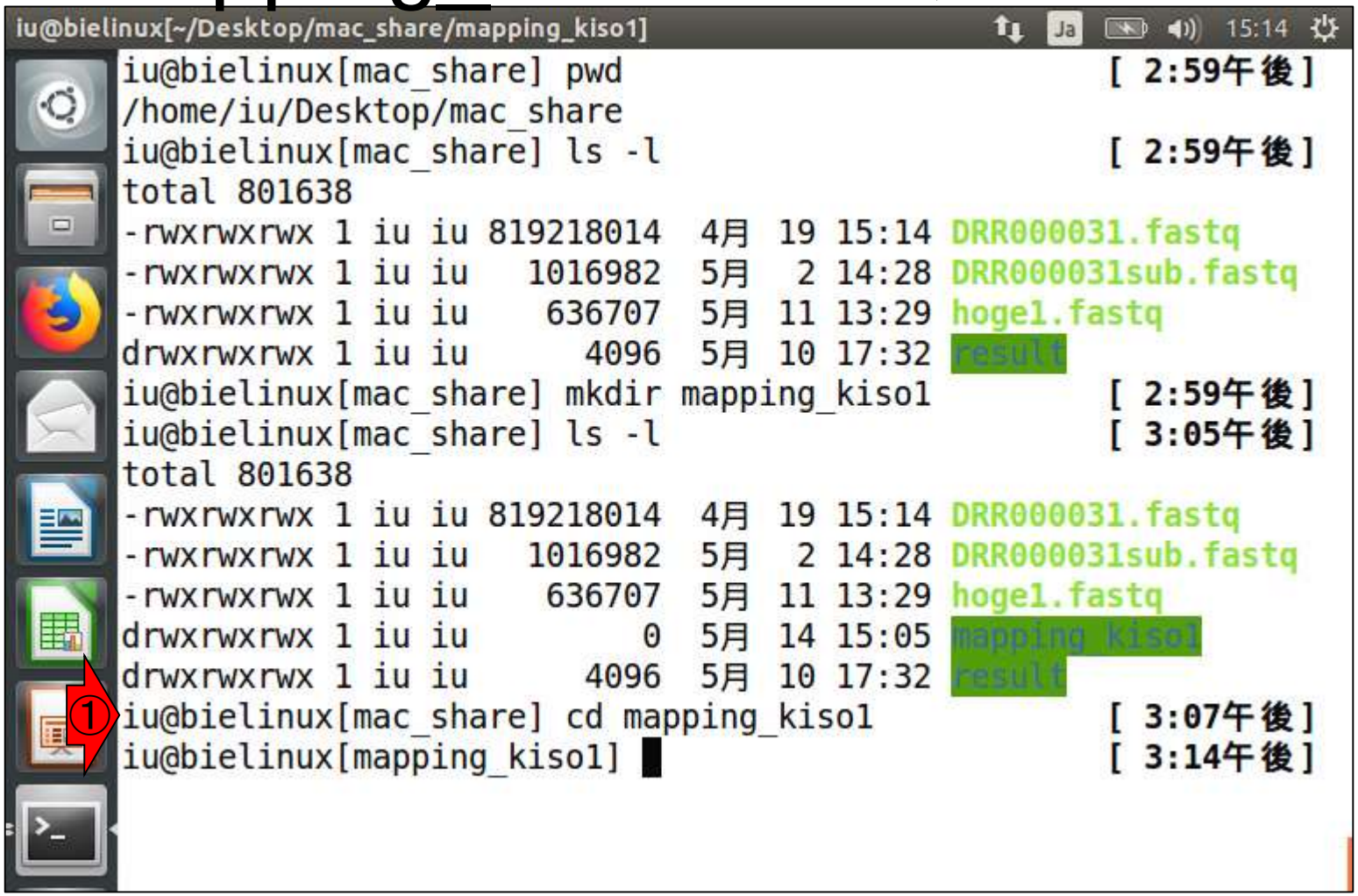
ディレクトリの作成

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



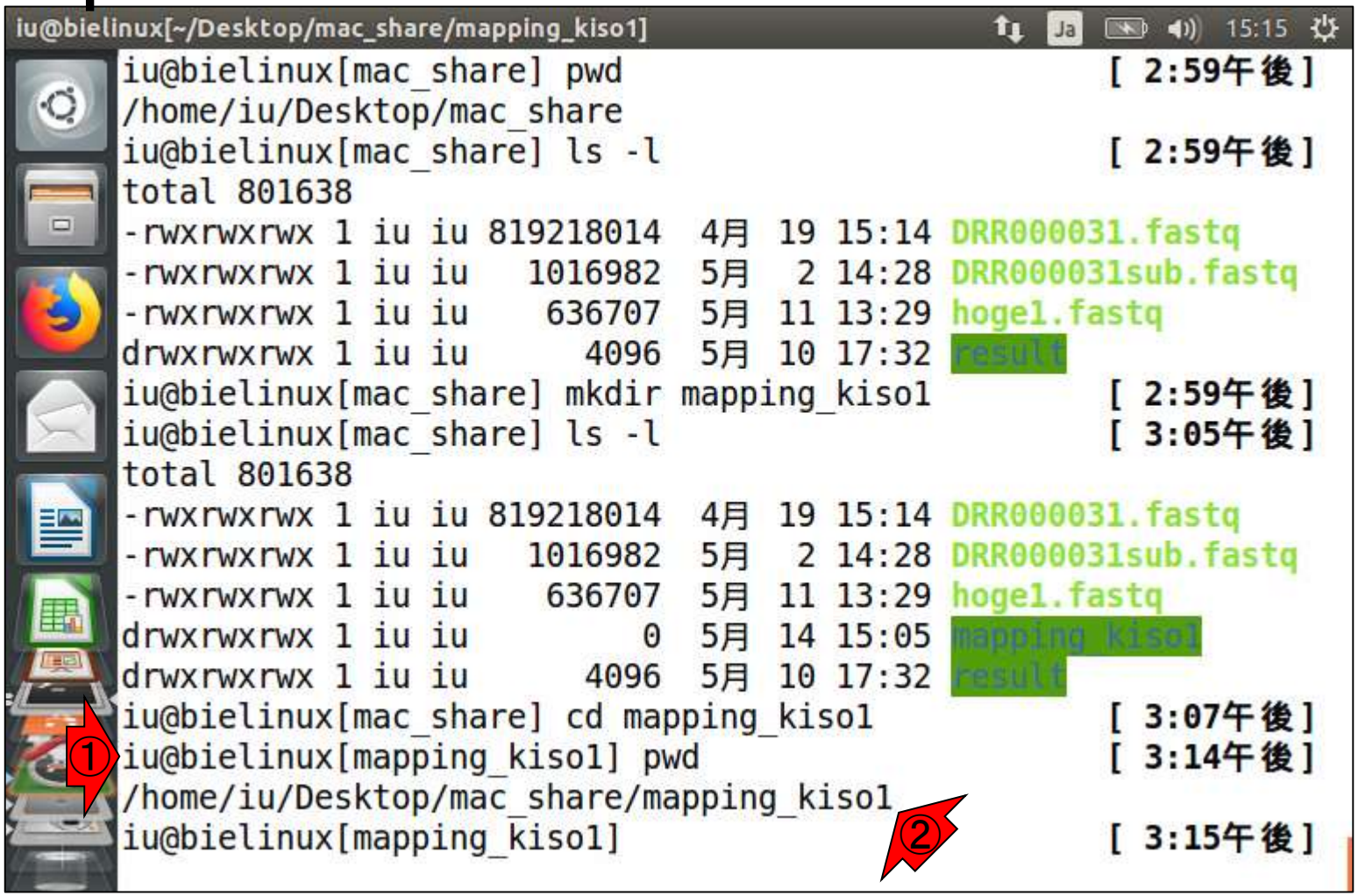
①cdコマンドでmapping_kiso1ディレクトリに移動

mapping_kiso1に移動



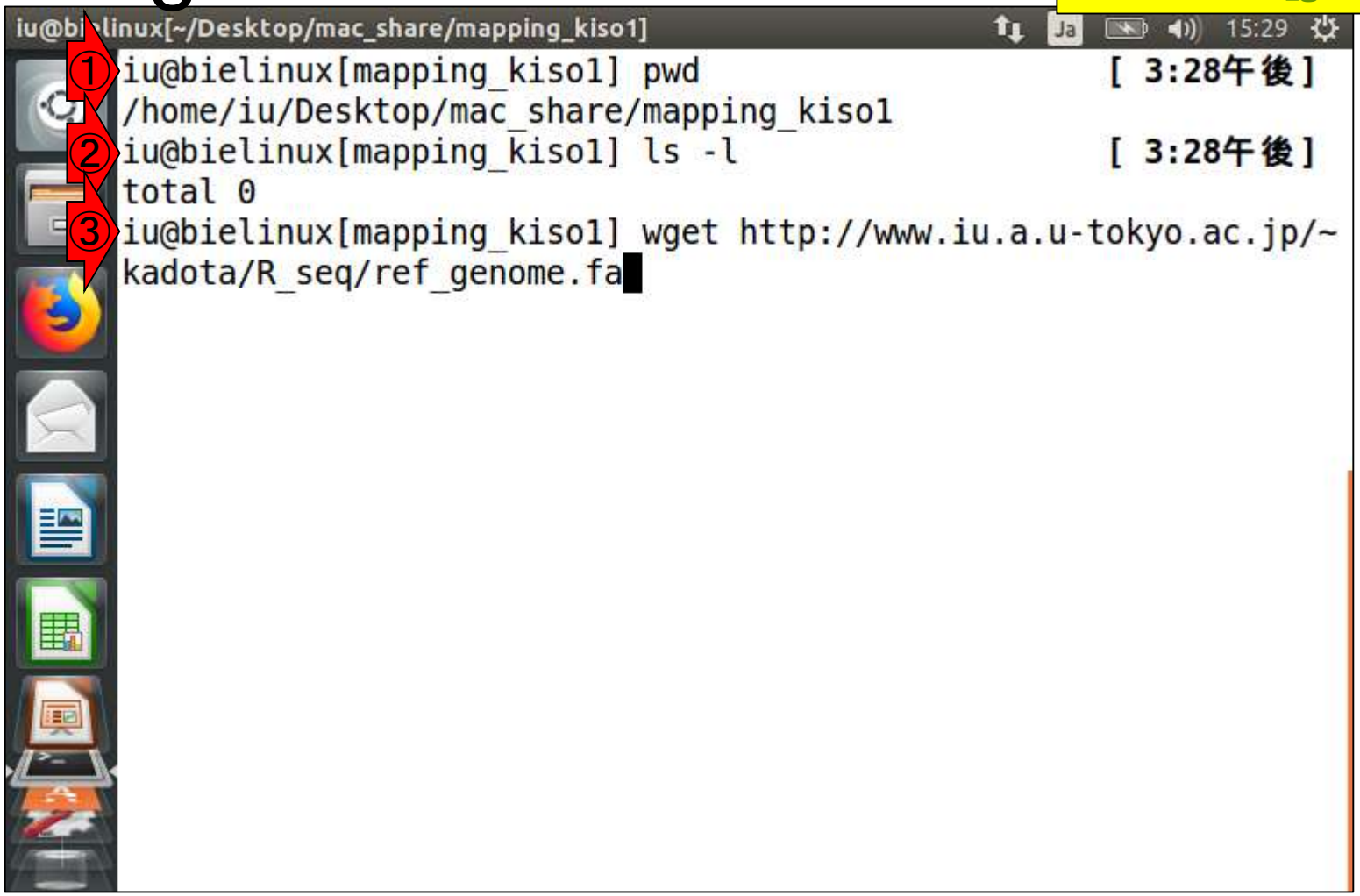
pwdで確認

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



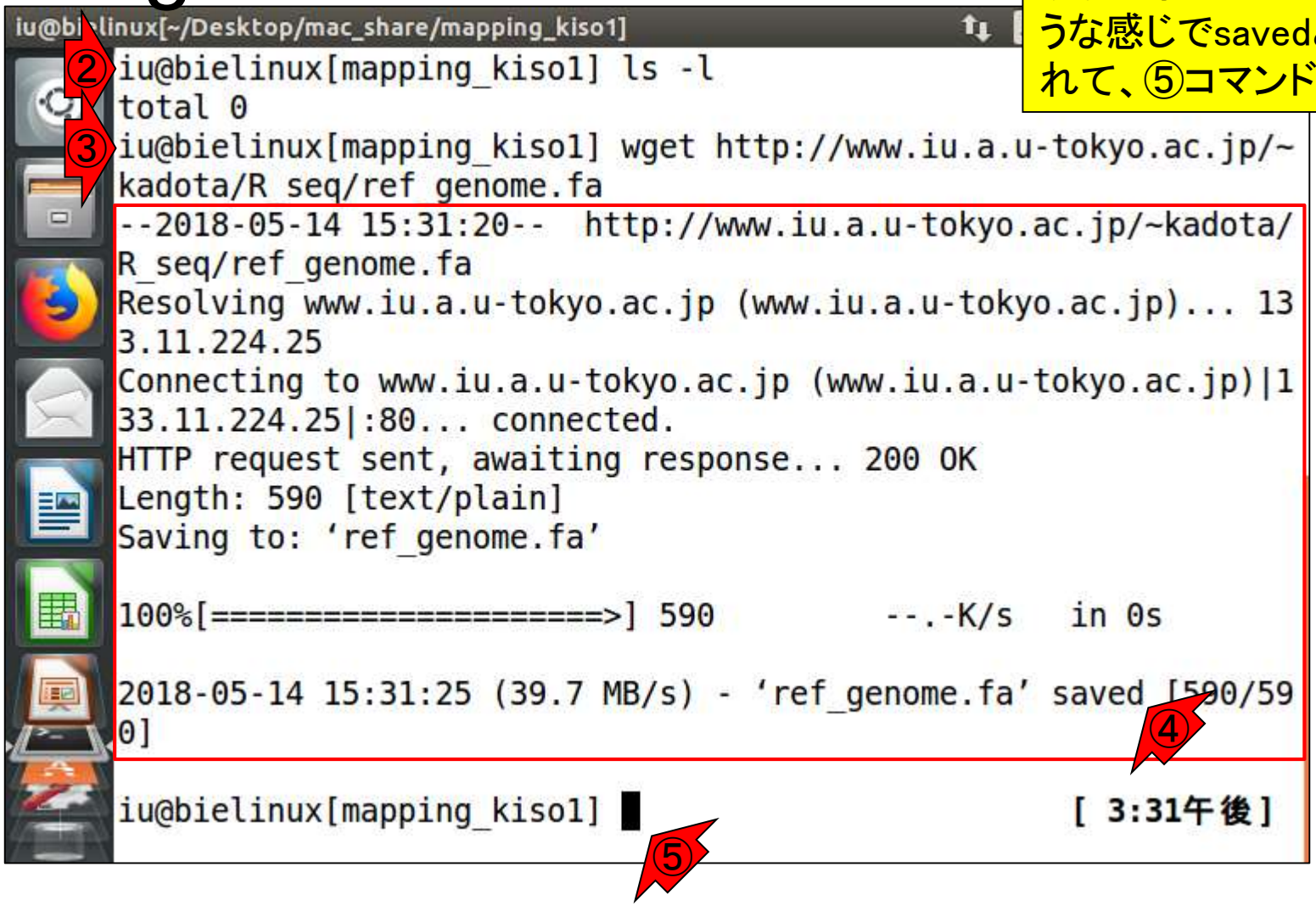
wgetでダウンロード

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



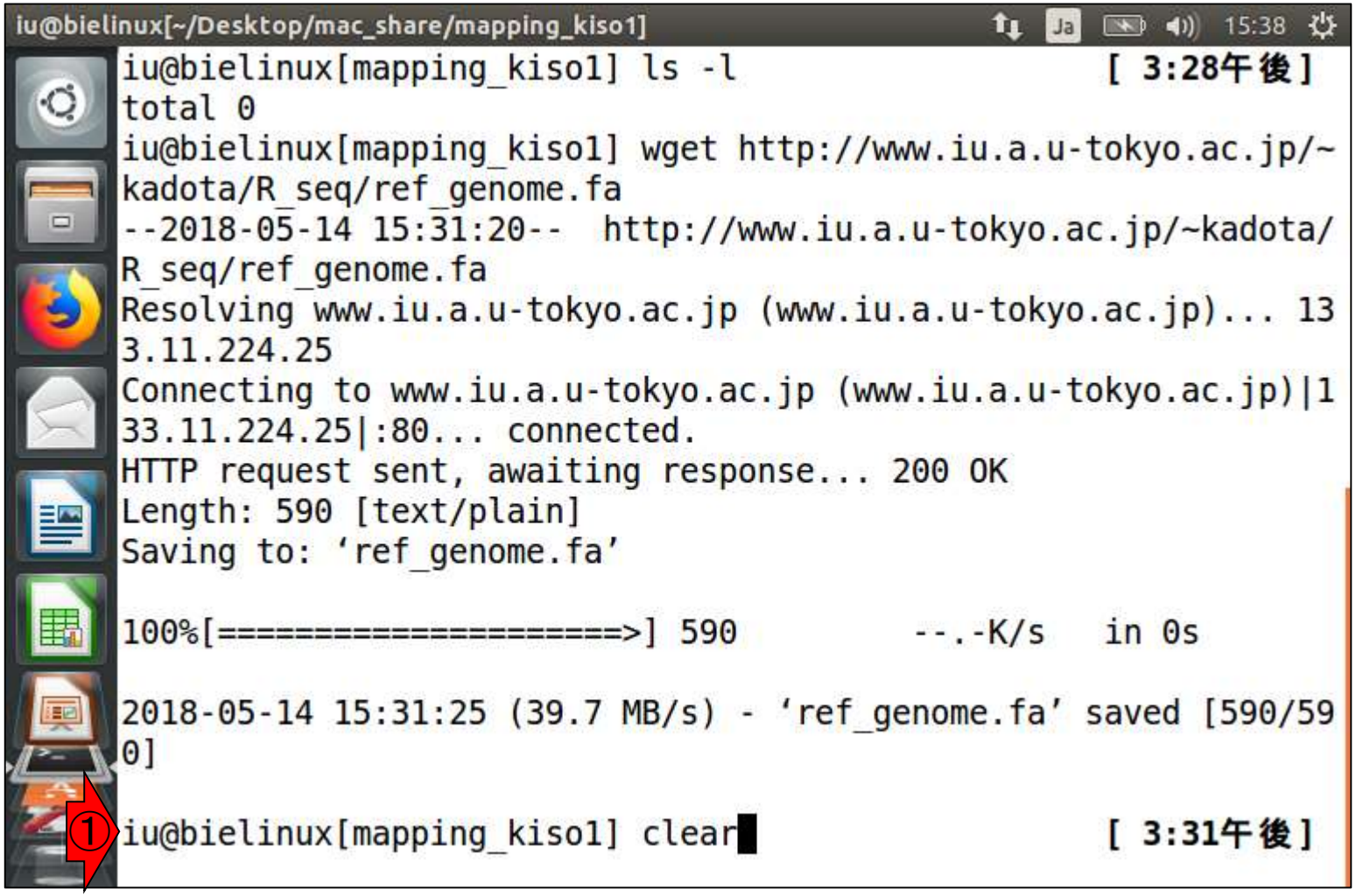
wgetでダウンロード

③wget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



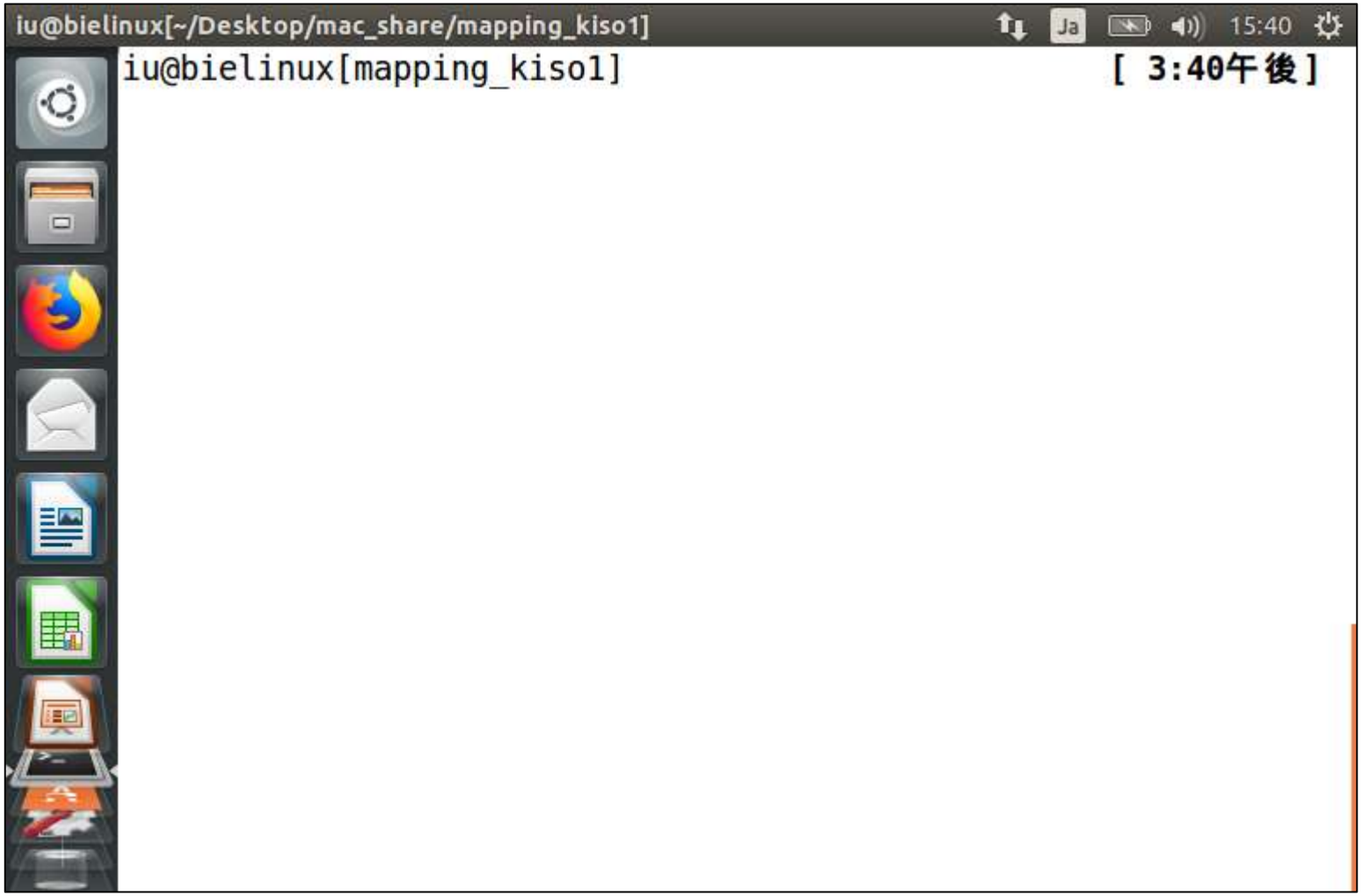
①clearコマンドを入力すると…

clearで画面リフレッシュ



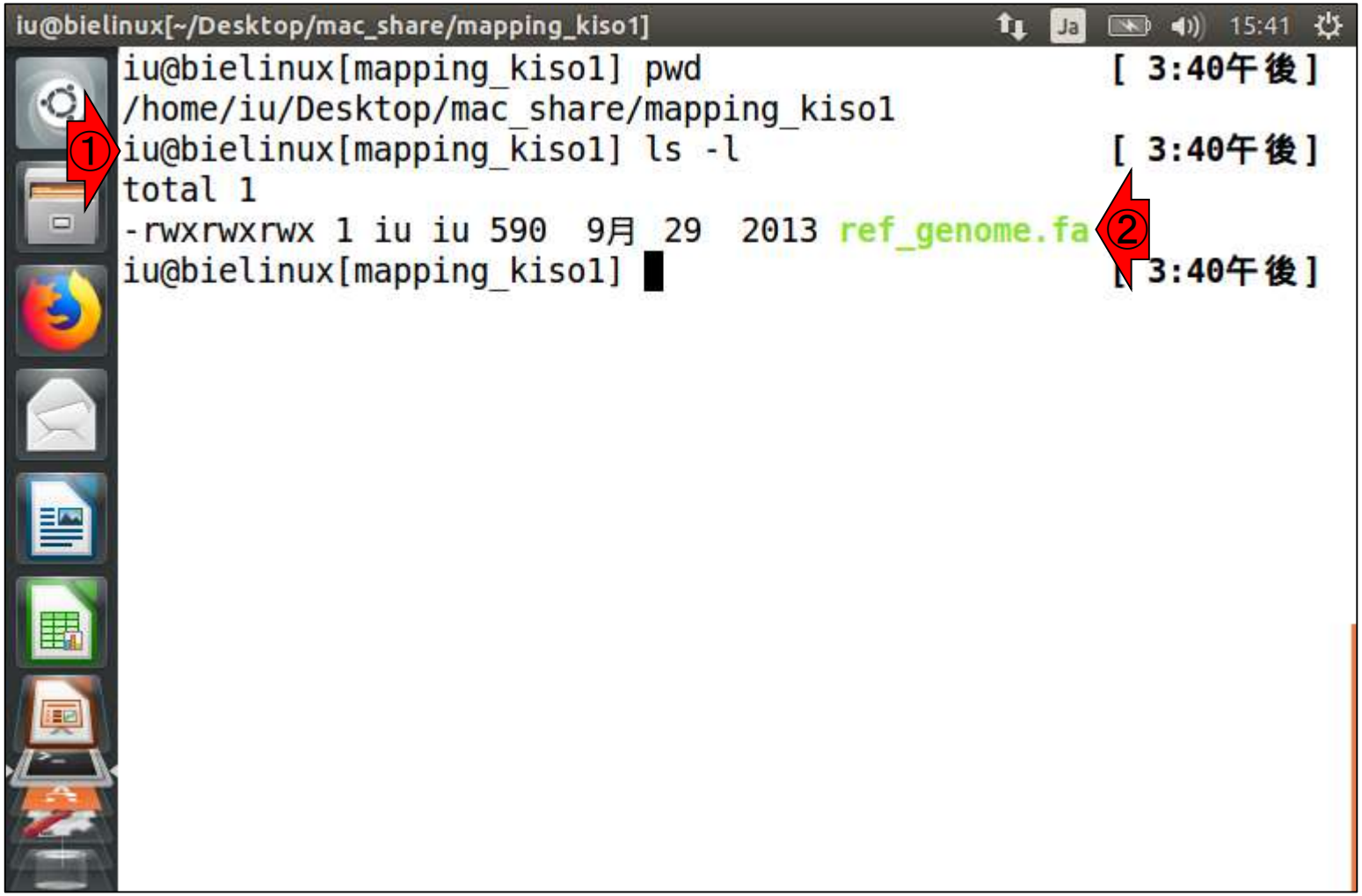
こんな感じで画面がリフレッシュされます

clearで画面リフレッシュ



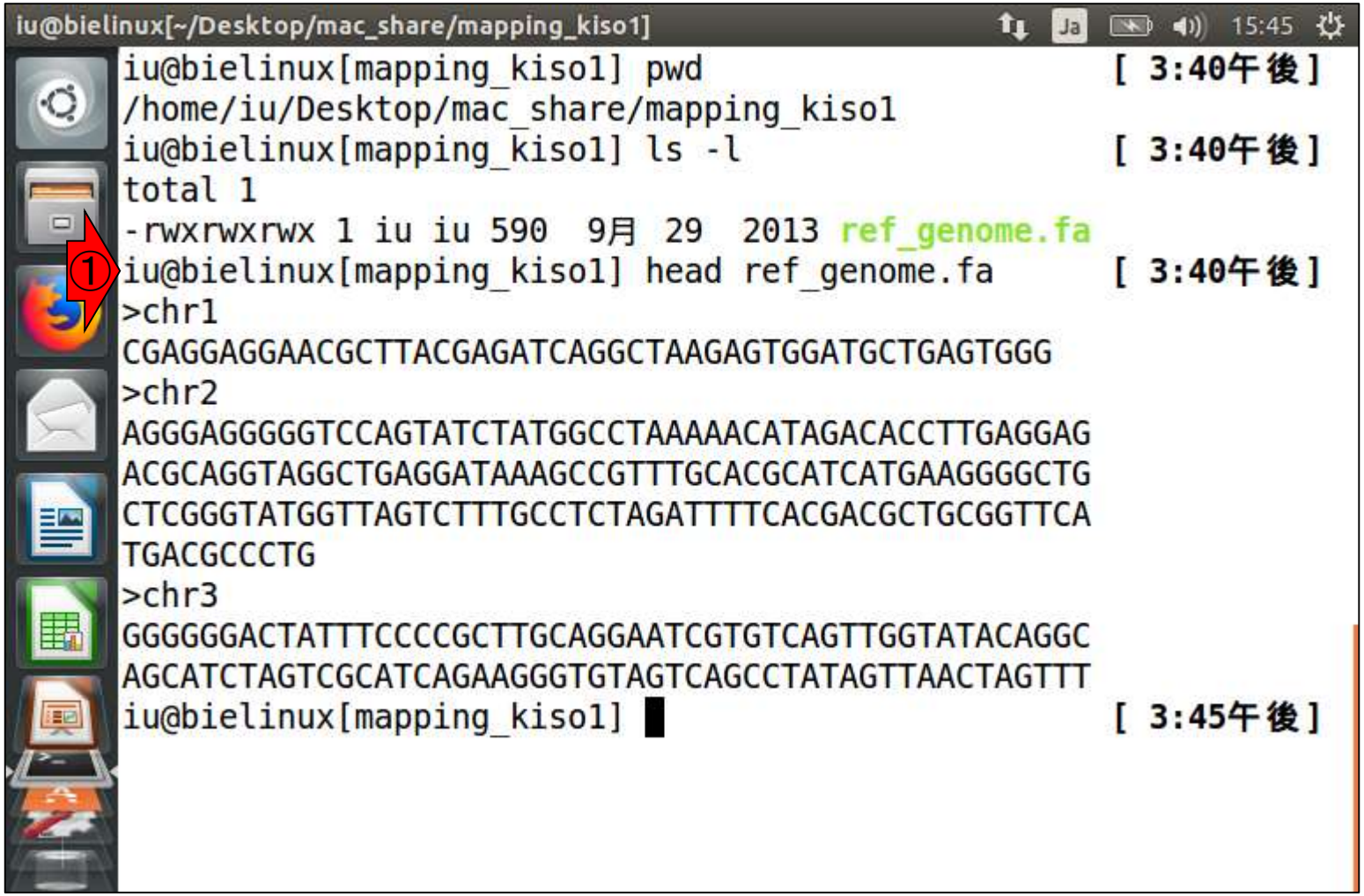
lsで確認

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



headで確認

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

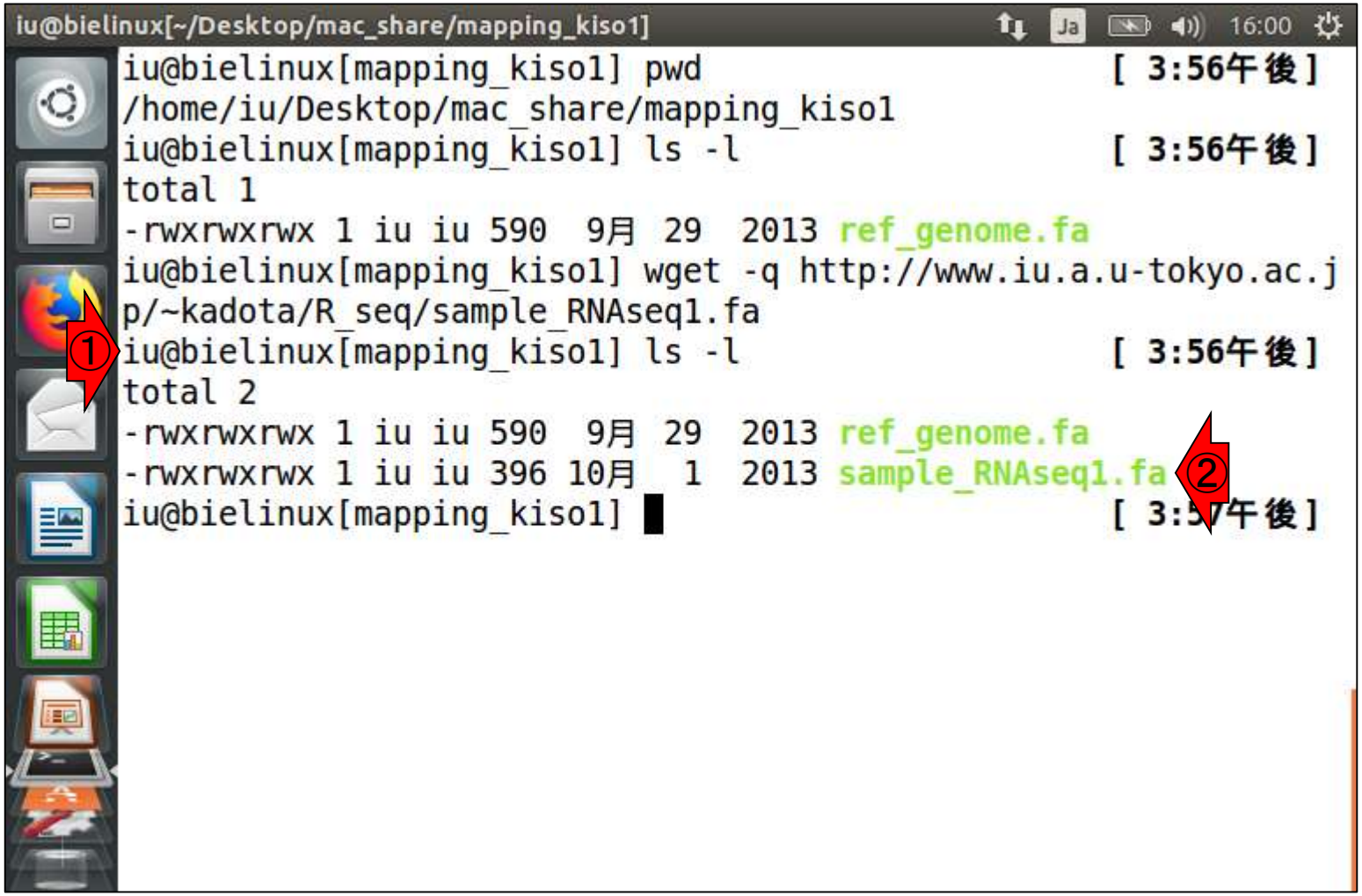


ダウンロード2

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

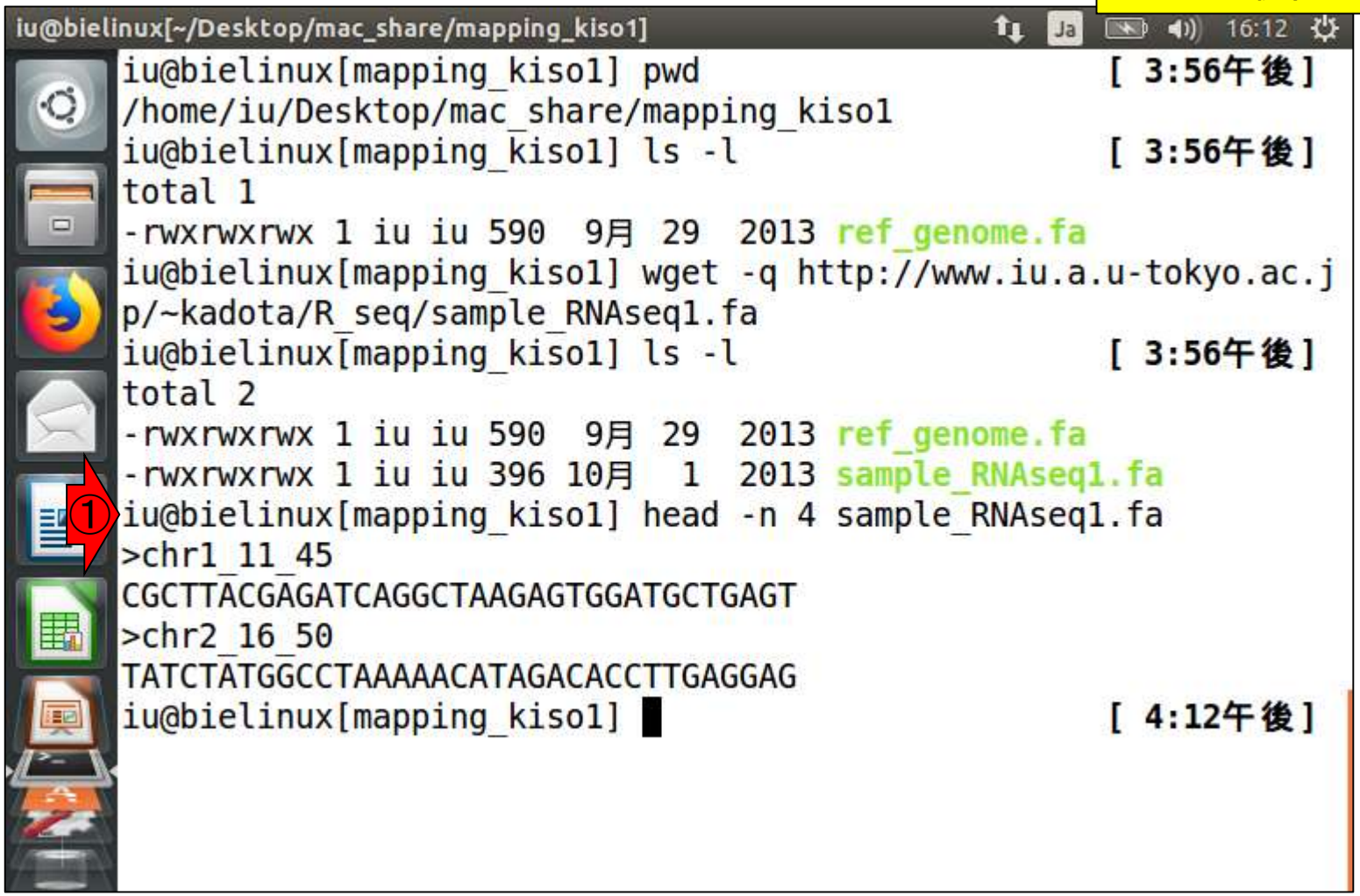
lsで確認

①lsで確認。②ダウンロードできているようですね



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



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、qrqc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 B[1], 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).

①

Comment in

The need for speed. [Genome Biol. 2009]

PMID: 19261174 PMCID: PMC2690996 DOI: 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 B[1], 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.

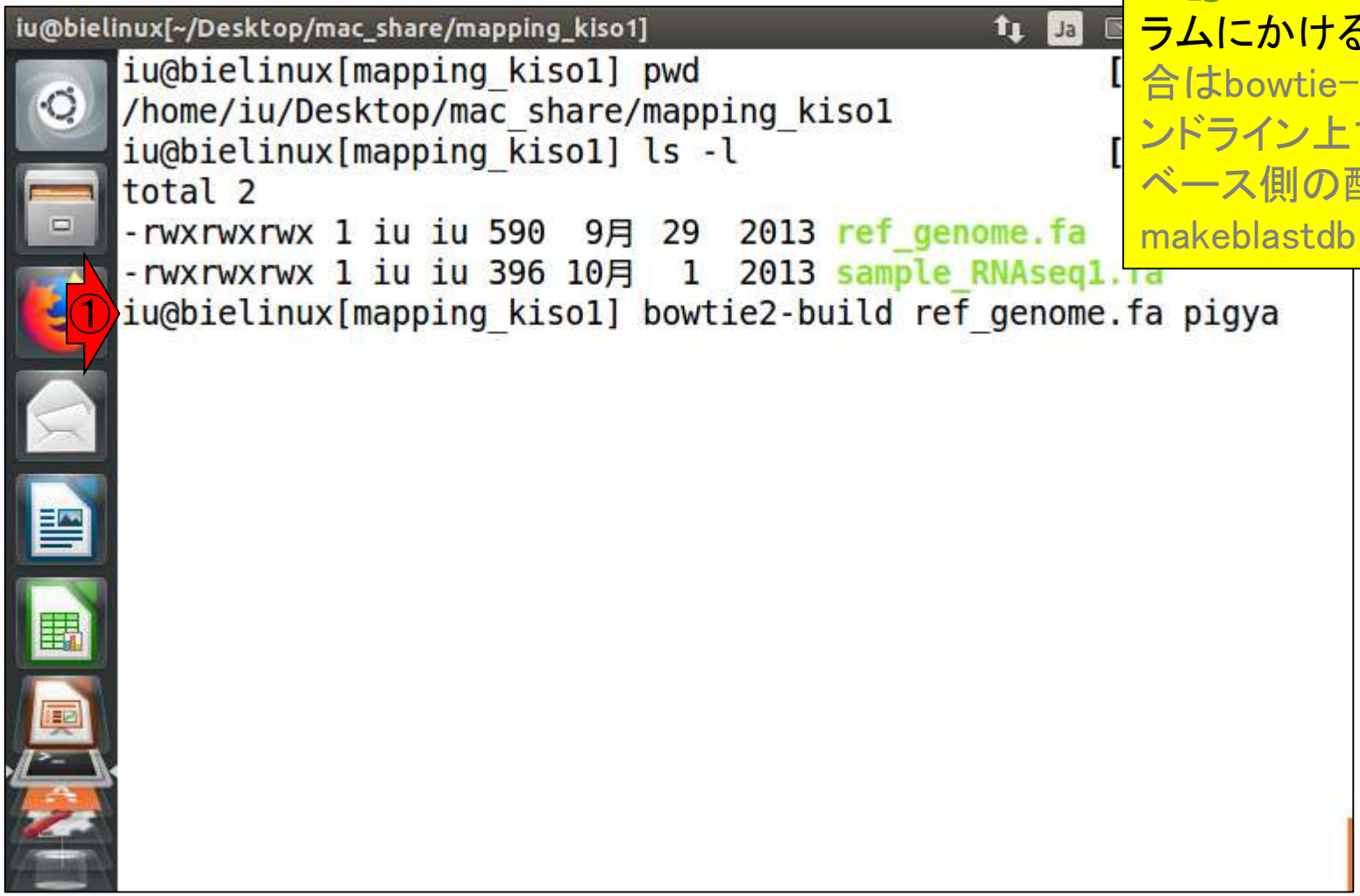
PMID: 22388286 PMCID: PMC3322381 DOI: 10.1038/nmeth.1923

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ょうがないので、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

Bowtie 2 (Langmead and Salzberg, Nat Methods, **9**: 357–9, 2012)

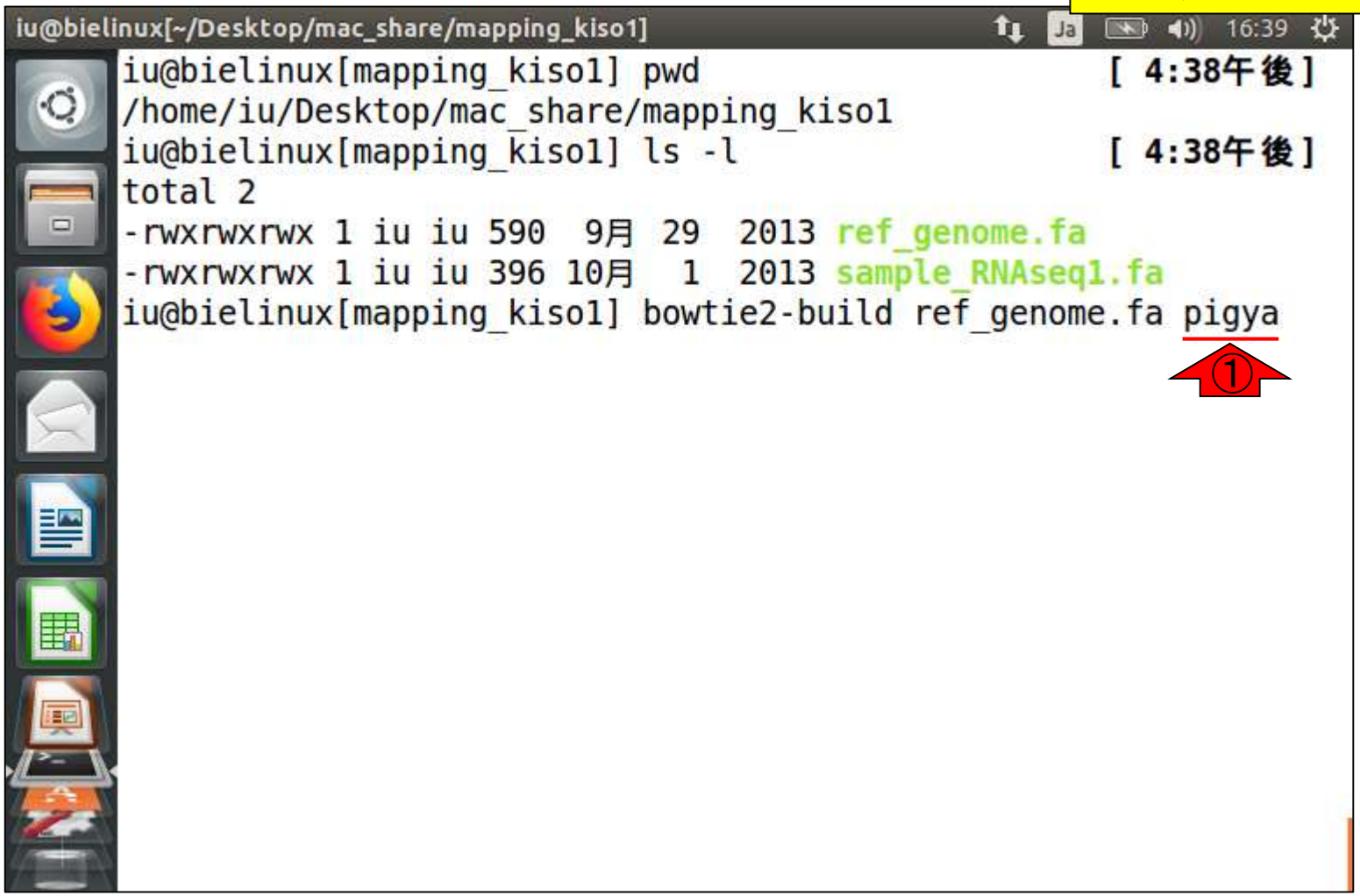
リファレンス配列の前処理

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



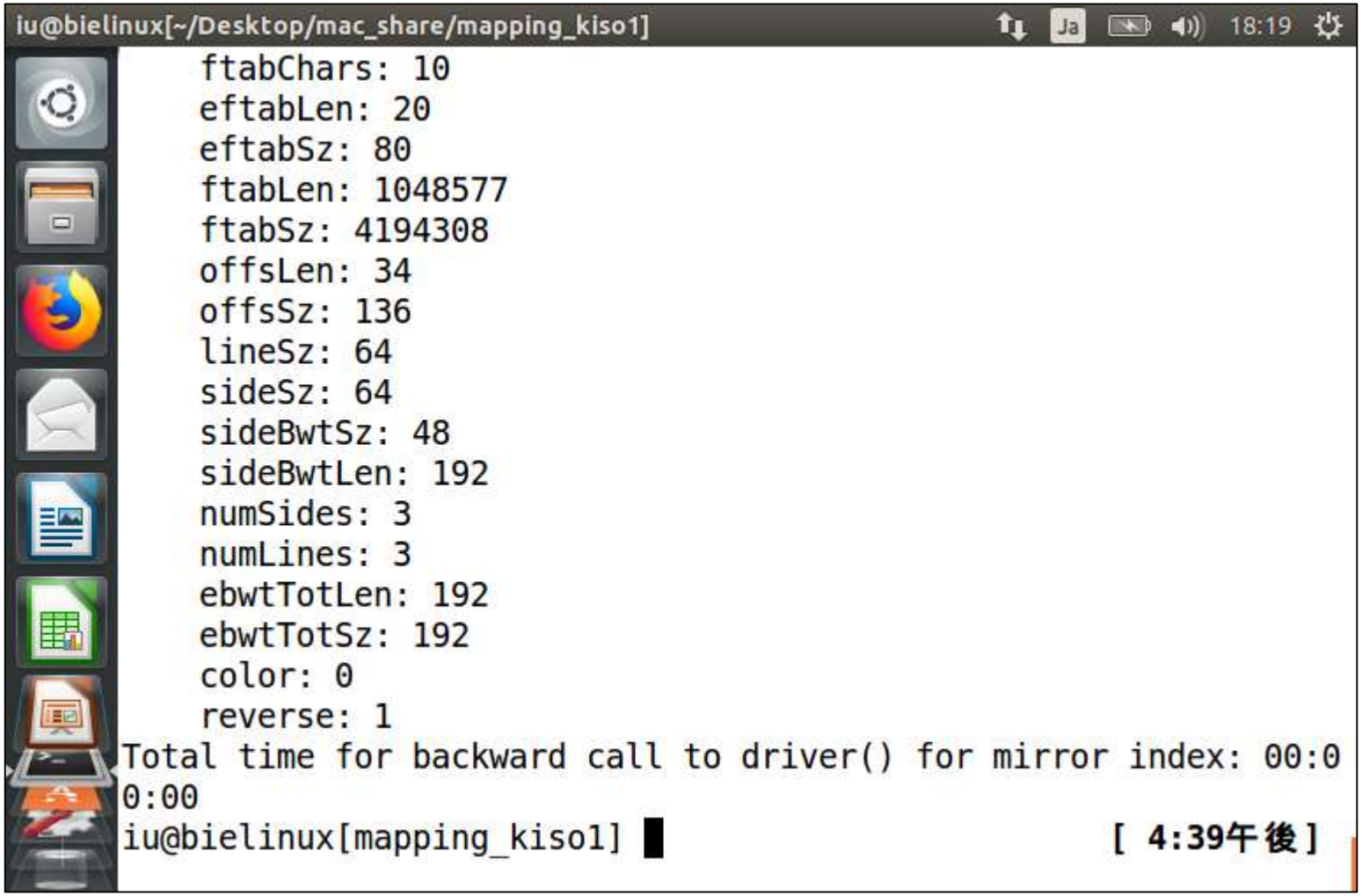
リファレンス配列の前処理

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



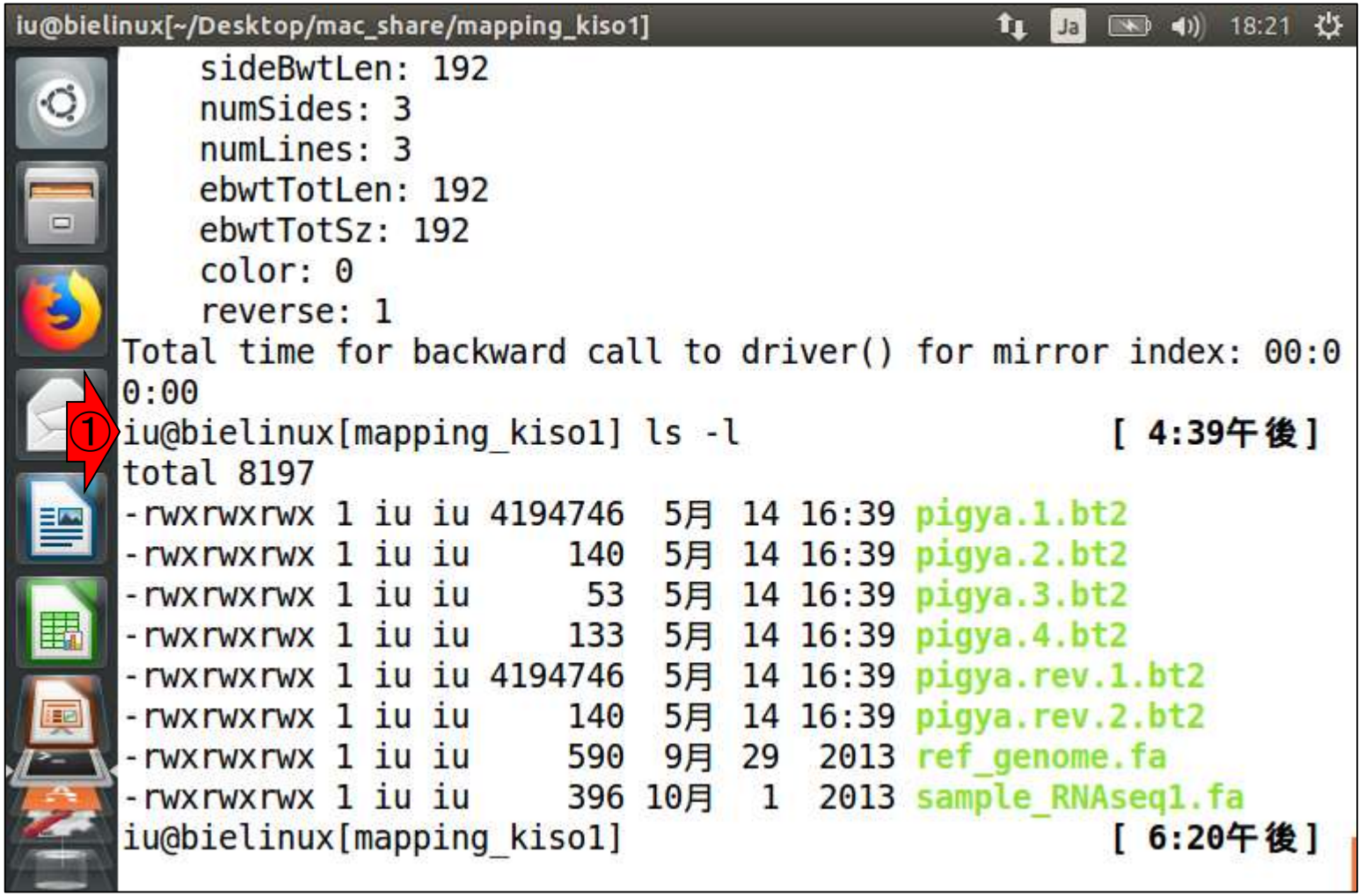
bowtie2-build実行

画面がざっと流れてこんな感じになります。数秒程度



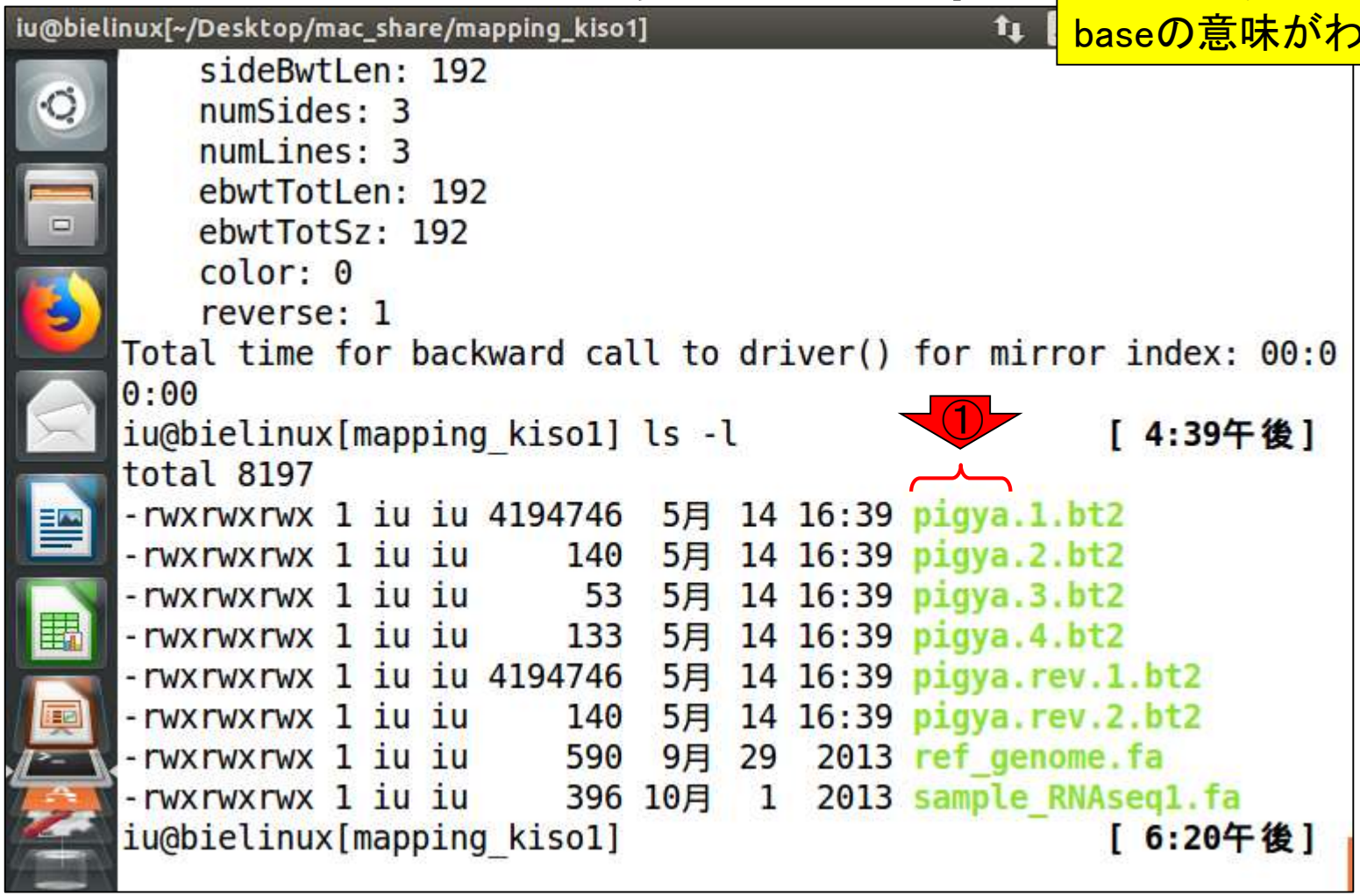
①lsで確認

bowtie2-build実行結果



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

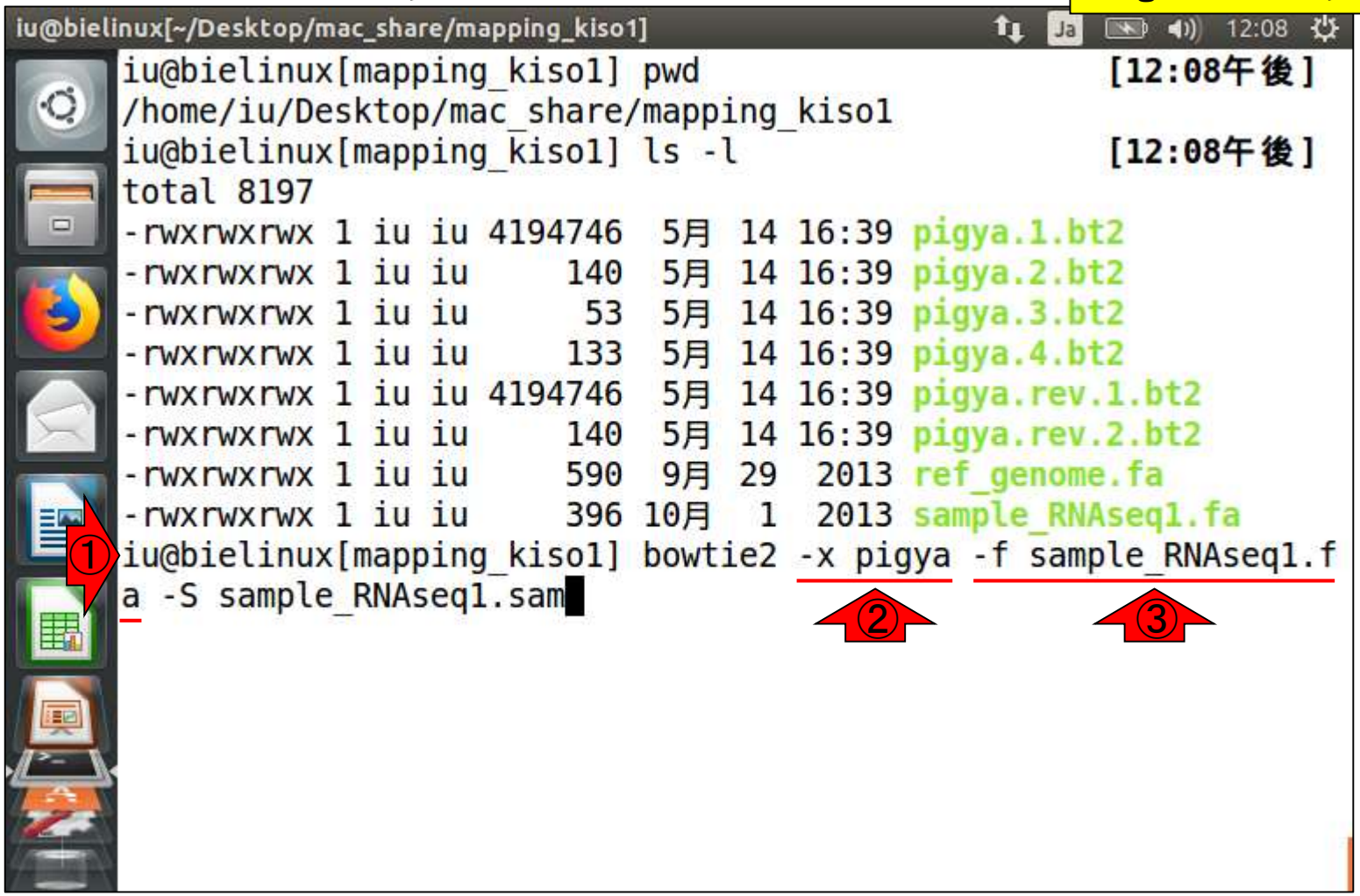


Contents

- 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、qrqc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

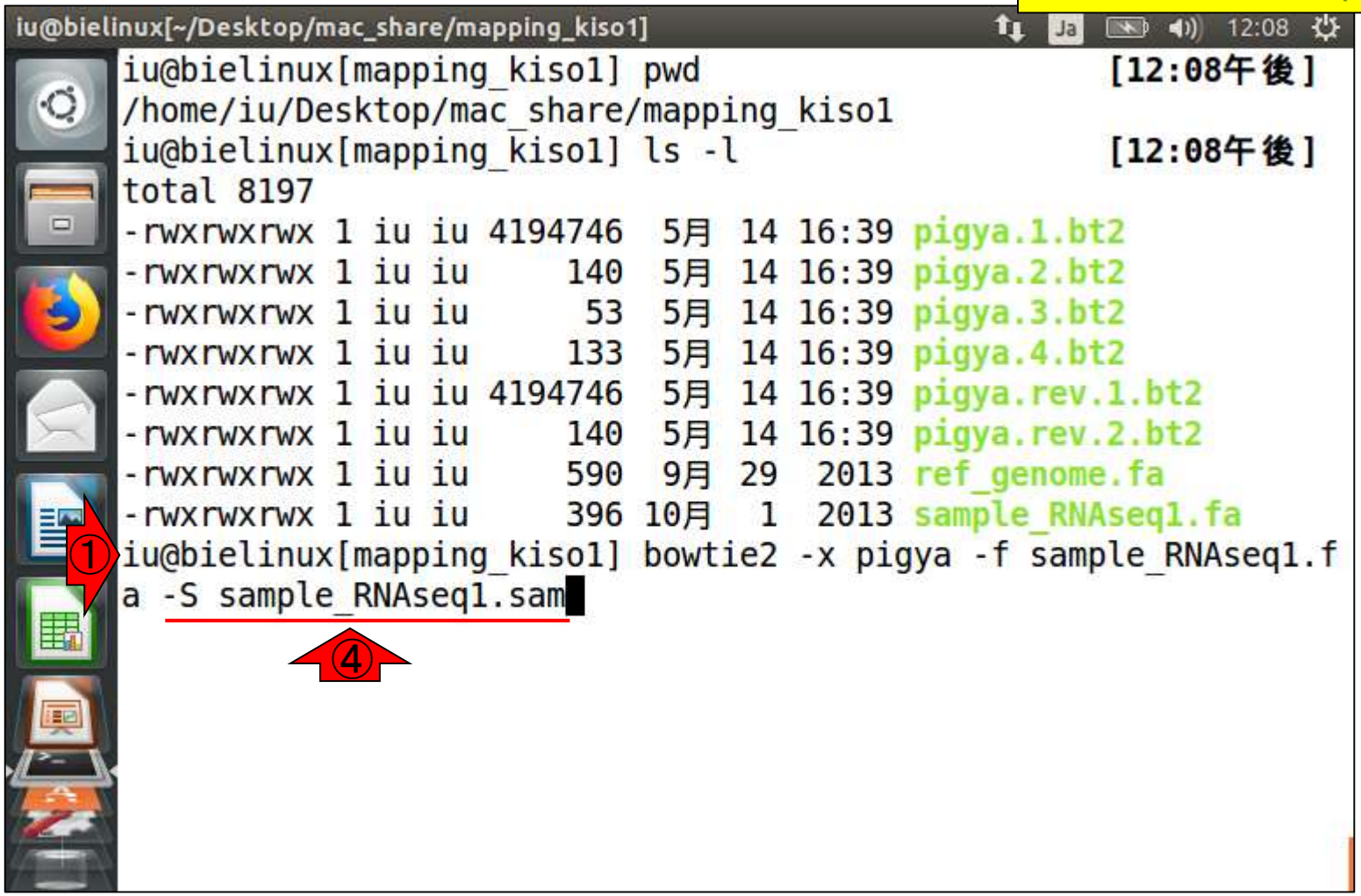
bowtie2実行コマンド

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



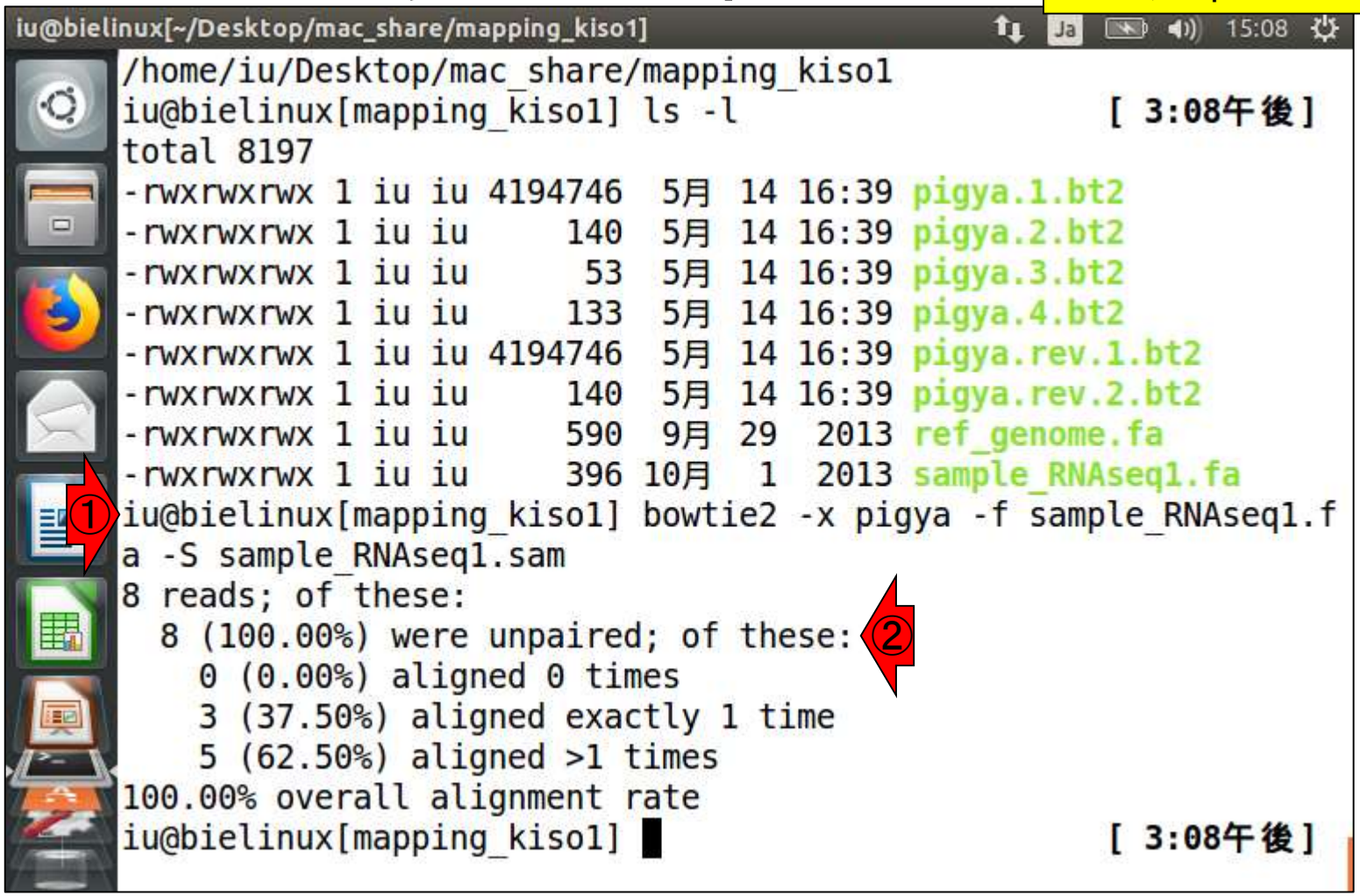
bowtie2実行コマンド

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



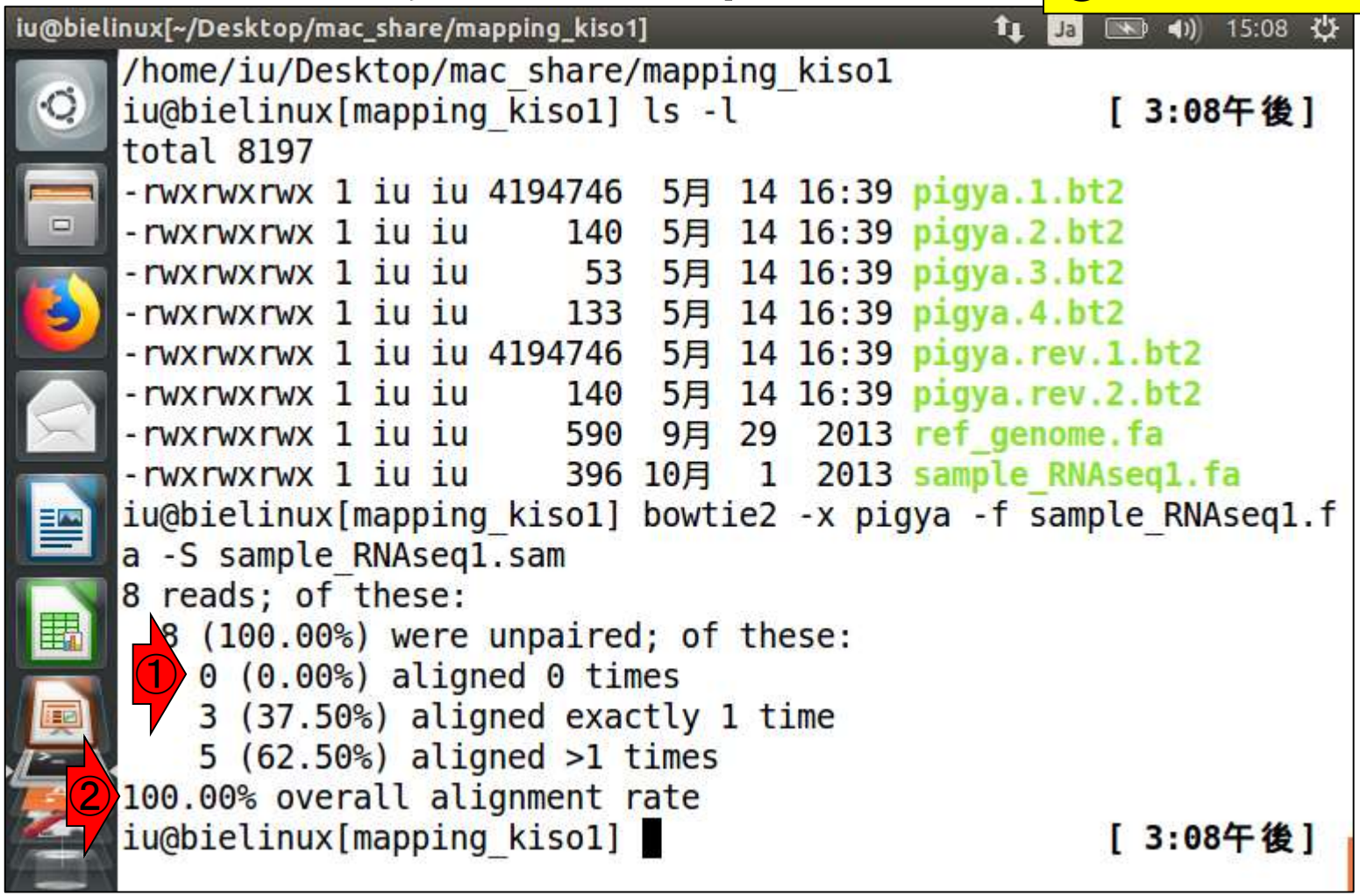
bowtie2実行結果

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



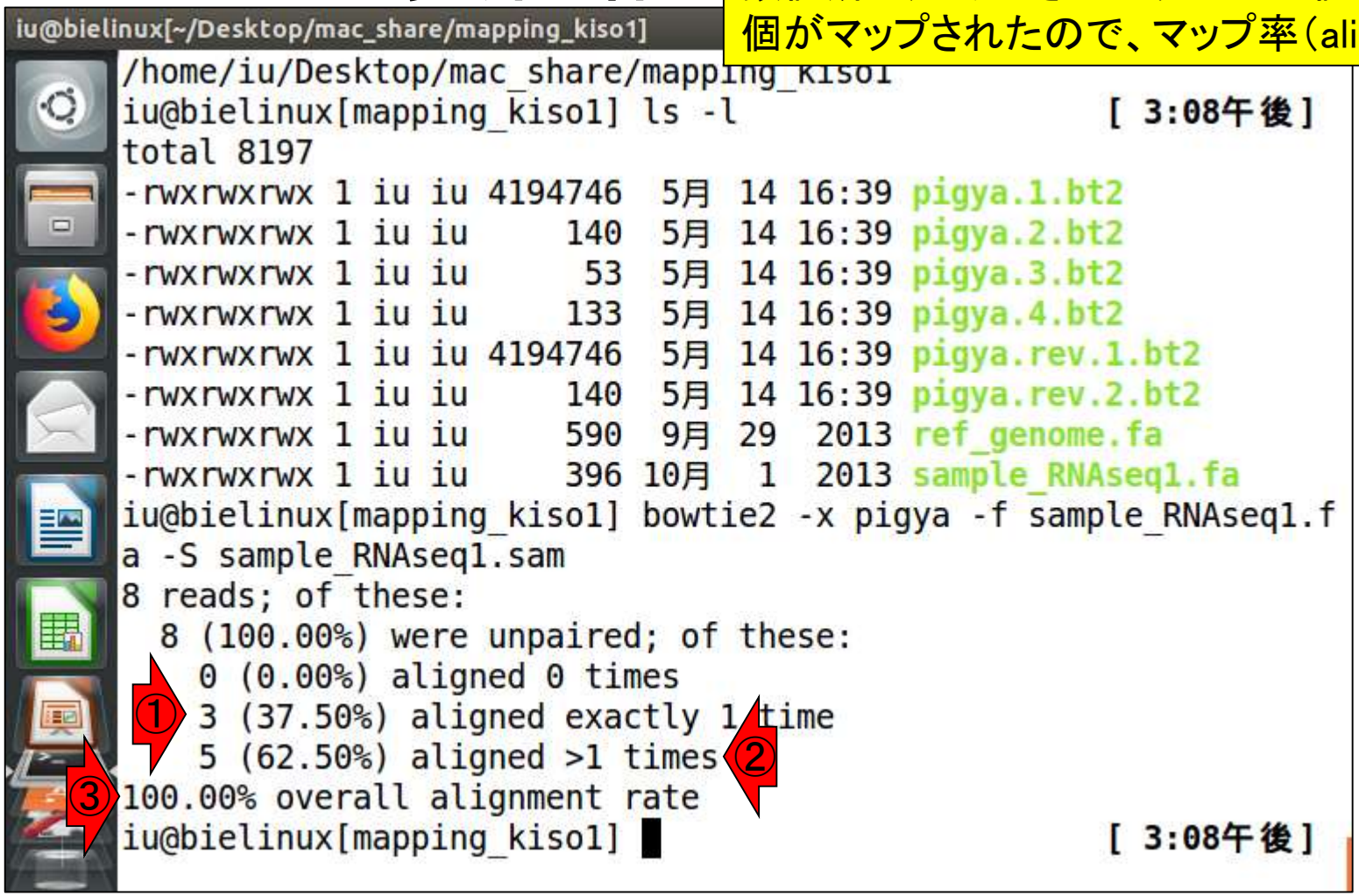
bowtie2実行結果

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



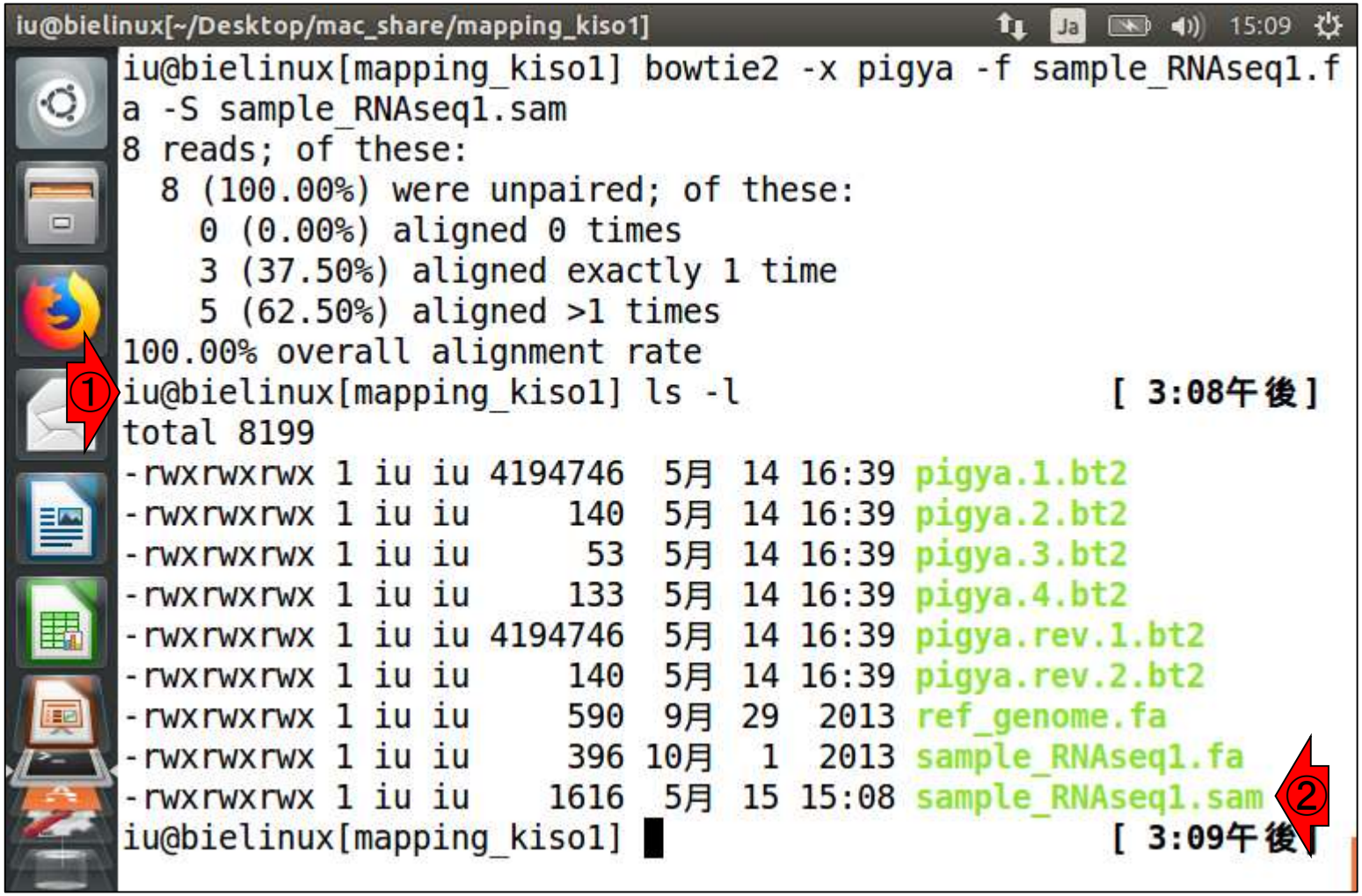
bowtie2実行結果

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%



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

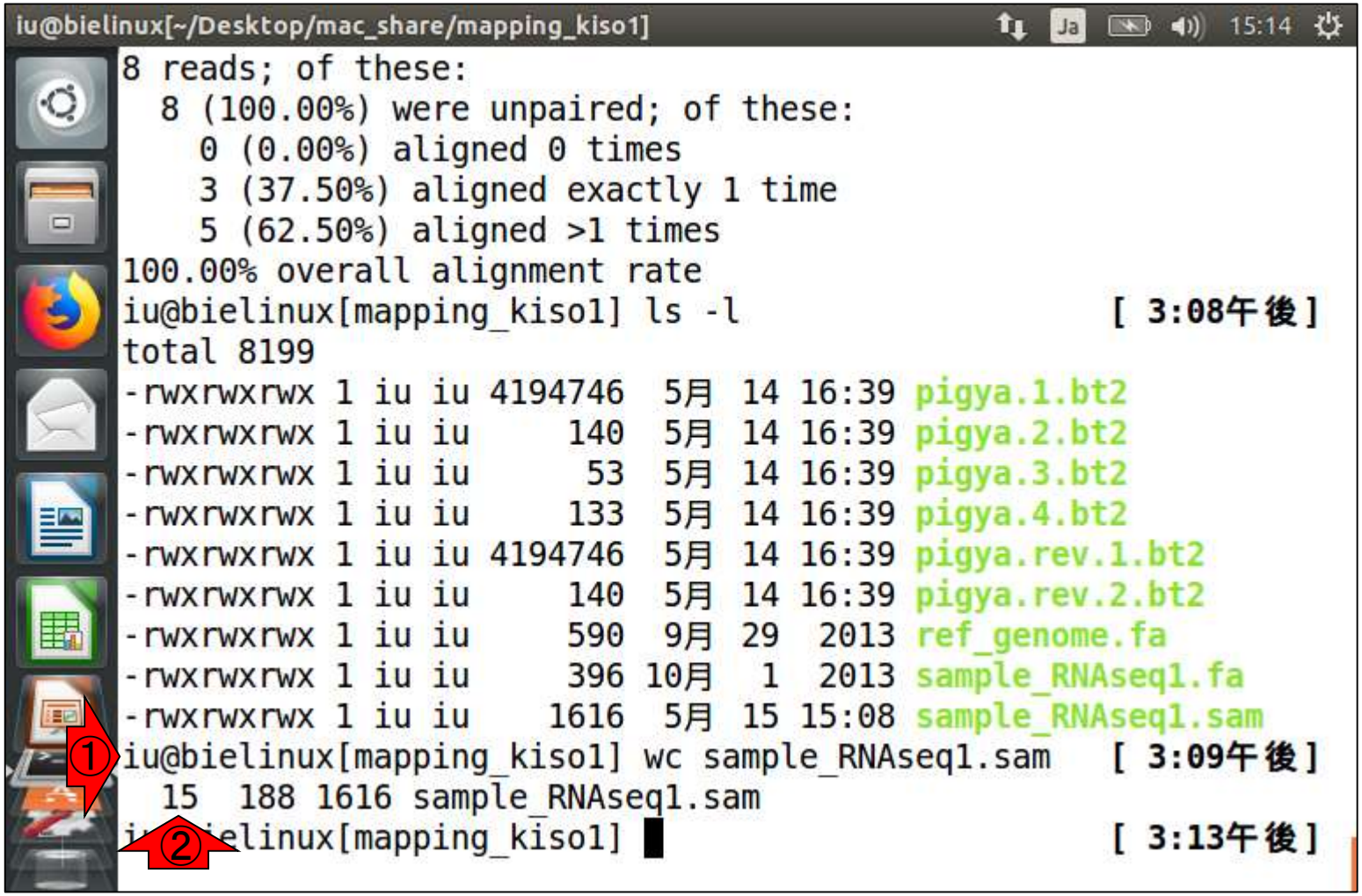


Contents

- 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、qrqc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

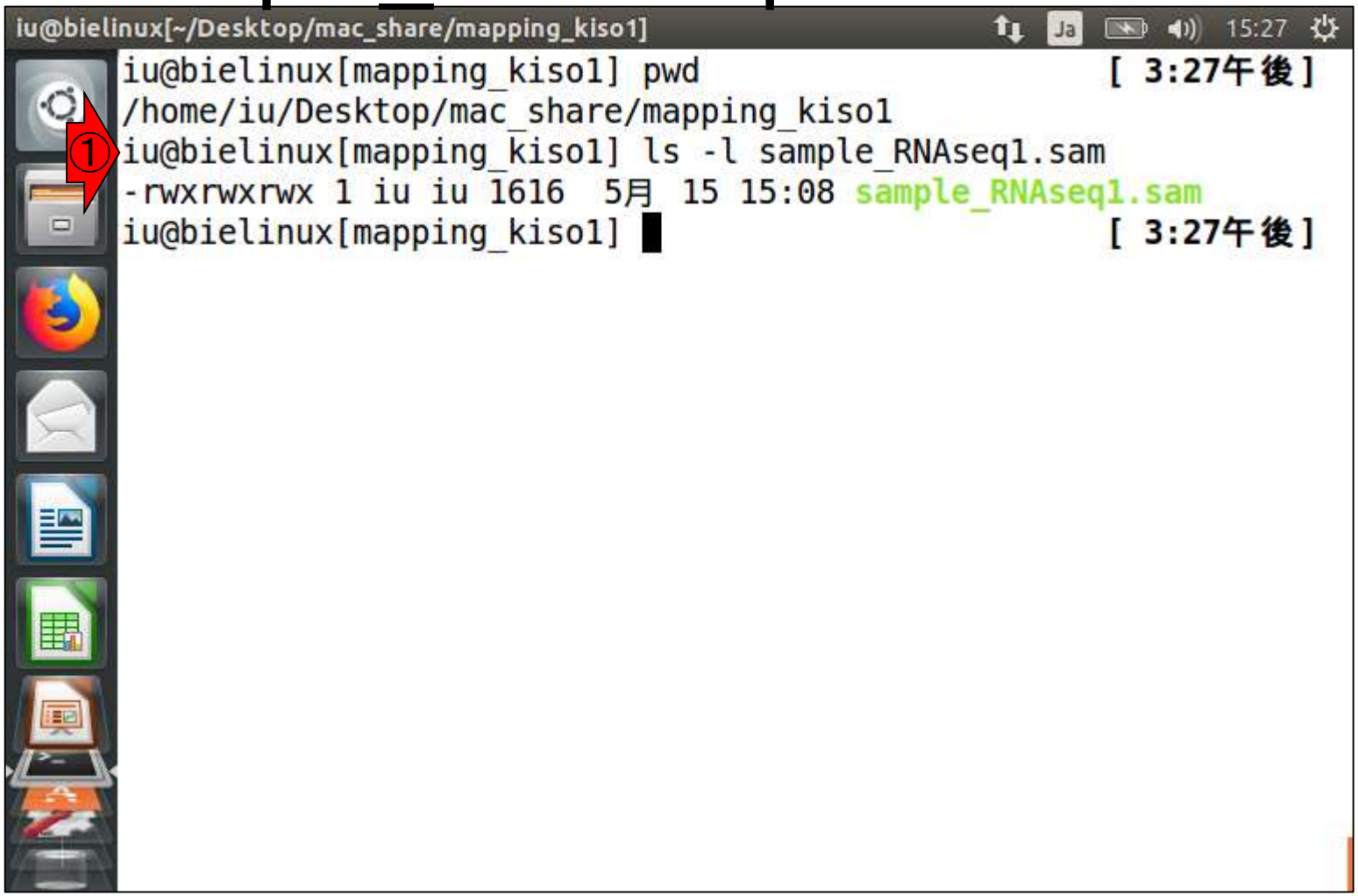
①目的のSAMファイルの行数を確認。②15行ですね

wcで行数を確認



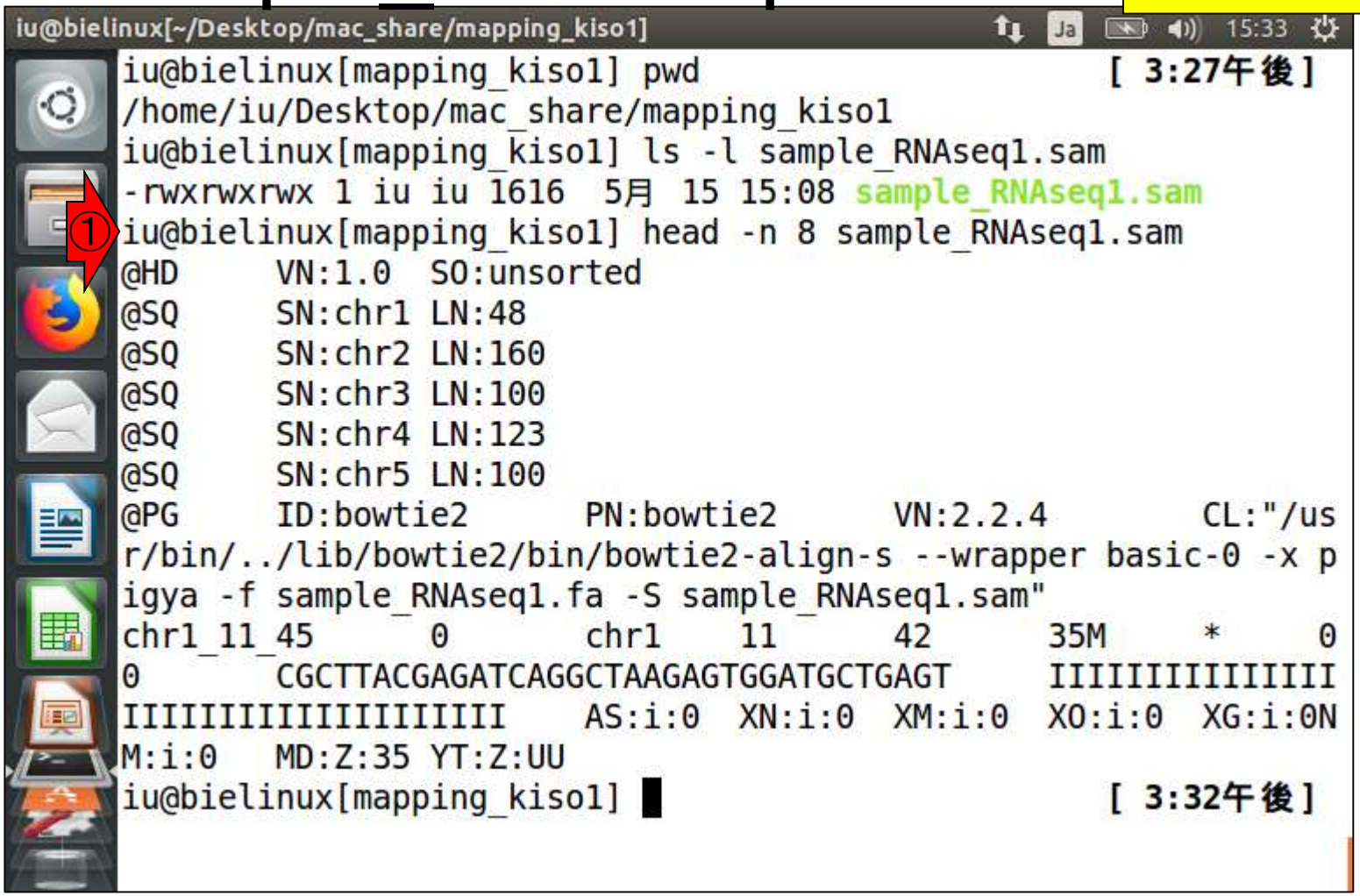
①目的のSAMファイルのみls -l

sample_RNAseq1.sam



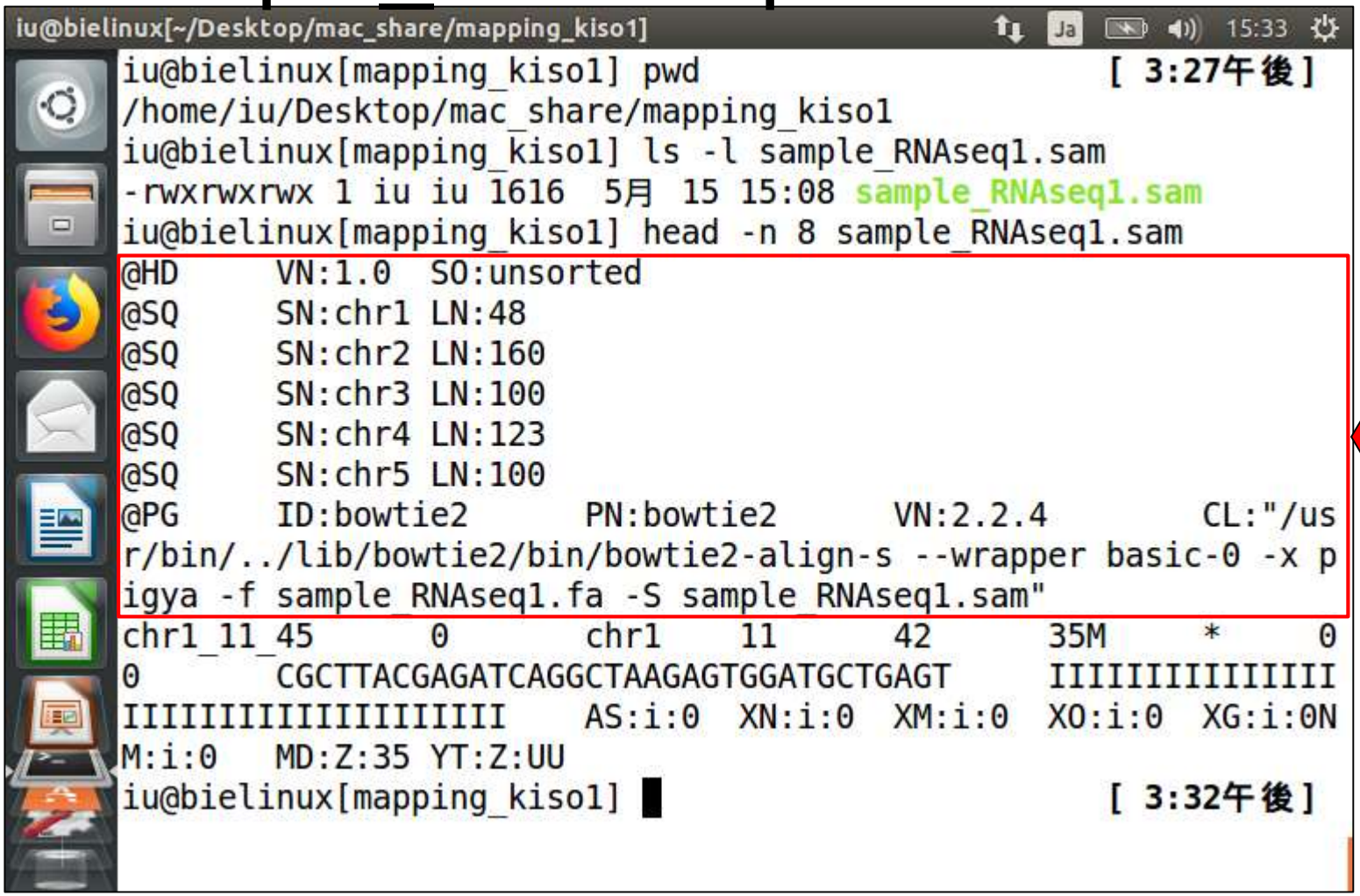
sample_RNAseq1.sam

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行分だっただけです）



sample_RNAseq1.sam

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

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/m
iu@bielinux[mapping_kiso1] l
-rwxrwxrwx 1 iu iu 1616  5月
iu@bielinux[mapping_kiso1] h
@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:b
r/bin/../lib/bowtie2/bin/bow
igya -f sample_RNAseq1.fa -S
chr1_11_45      0       chr1
0       CGCTTACGAGATCAGGCTAA
IIIIIIIIIIIIIIIIIII     AS:i
M:i:0   MD:Z:35 YT:Z:UU
iu@bielinux[mapping_kiso1]
```

①

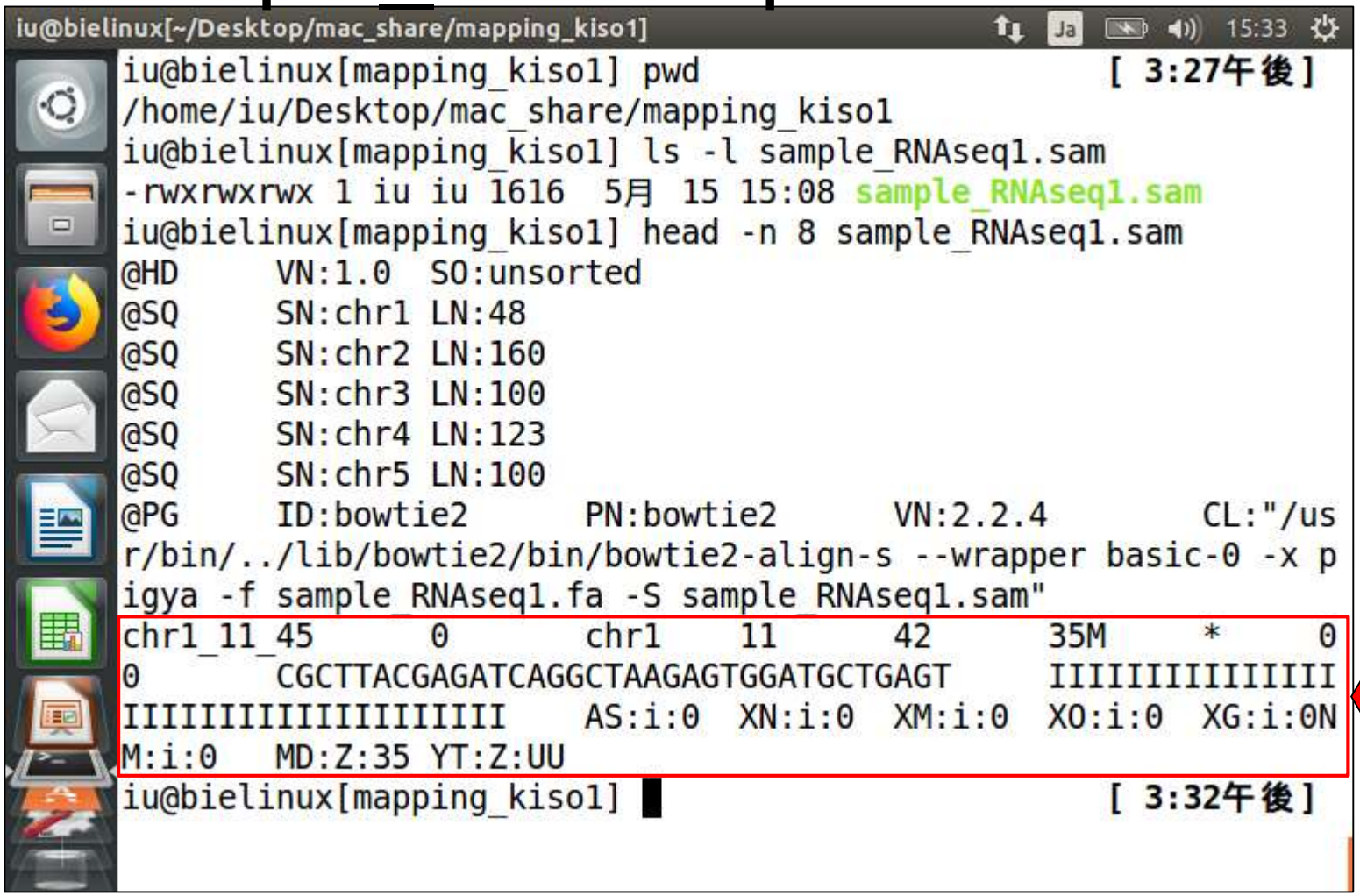
ref_genome.fa - メモ帳

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
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

sample_RNAseq1.sam

①この行（8行目）からがリードごとのマッピング結果情報



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:bowt	PN:bowt	VN:2	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	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	TATCTATGGCCTAAAAACATAGACACCTTGAGGAG	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	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	

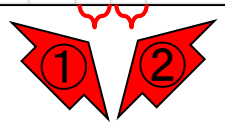
①

②

全体を拡大

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

@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	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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



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

①

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

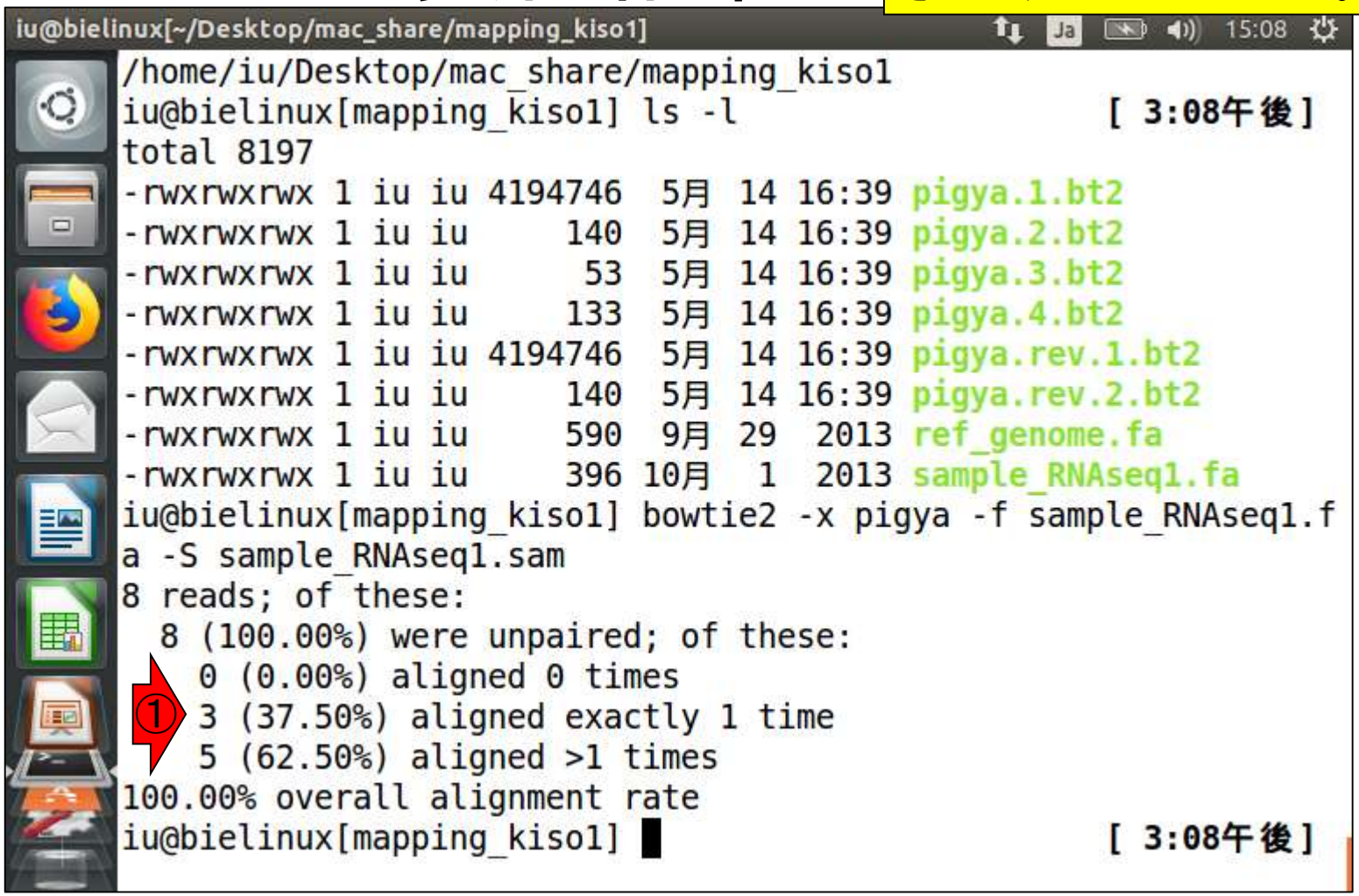
```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

```
@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-alig
chr1_11_45      0  chr1   11  42  35M  *  0  0  CC IIAS:i:0   XN:i:0   XM:i:0
chr2_16_50      0  chr2   16  42  35M  *  0  0  TA IIAS:i:0   XN:i:0   XM:i:0
chr2_1_35       0  chr2    1  42  35M  *  0  0  AC IIAS:i:0   XN:i:0   XM:i:0
chr3_11_45      0  chr5   11   1  35M  *  0  0  TT IIAS:i:0   XS:i:0   XN:i:0
chr3_15_49      0  chr3   15   1  35M  *  0  0  CC IIAS:i:0   XS:i:0   XN:i:0
chr3_3_37       0  chr3    3  31  35M  *  0  0  GC IIAS:i:0   XS:i:-6  XN:i:0
chr3_1_35       0  chr3    1  35  35M  *  0  0  GC IIAS:i:0   XS:i:-12 XN:i:0
chr5_1_35       0  chr5    1  16  35M  *  0  0  GC IIAS:i:-6  XS:i:-18 XN:i:0
```

②

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…



Bowtie2実行結果

①これです。対応するリードは②これ

@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-alig								
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0

①

②

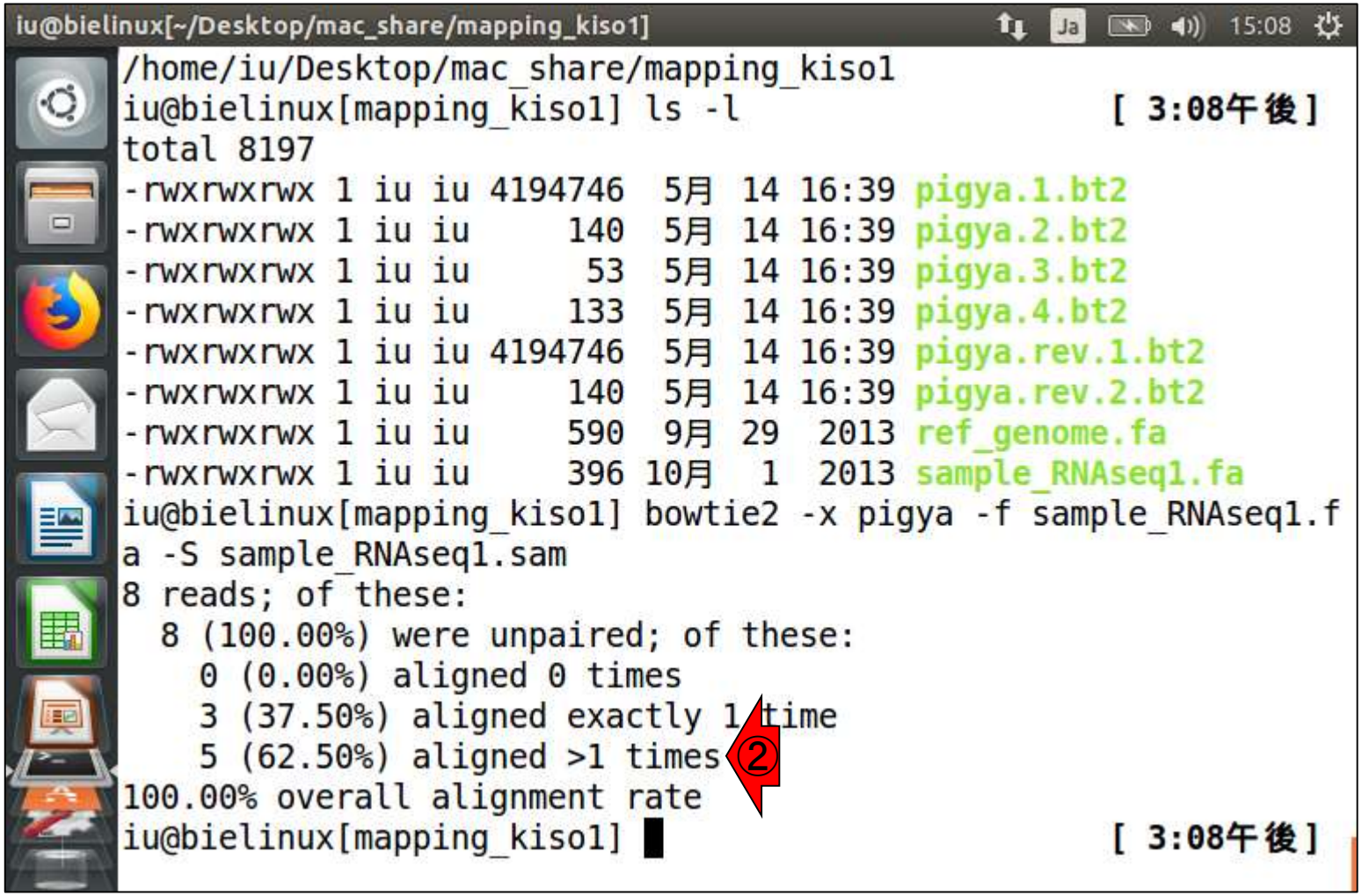
sample_RNAseq1.fa - メモ帳

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。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

①これです。対応するリードは②これ

@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-alig								
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0

①

②

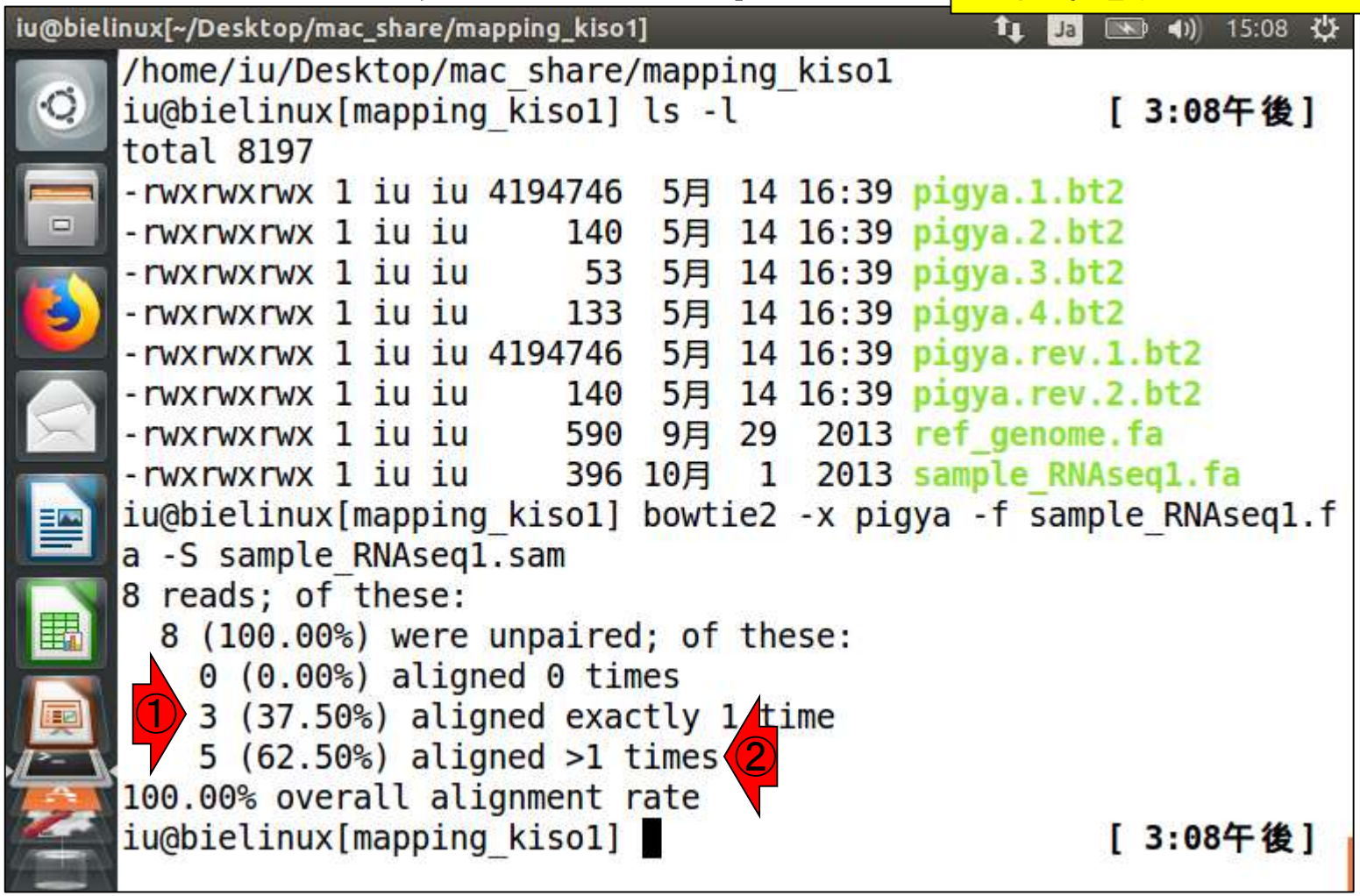
sample_RNAseq1.fa - メモ帳

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

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?



複数個所の結果はXS

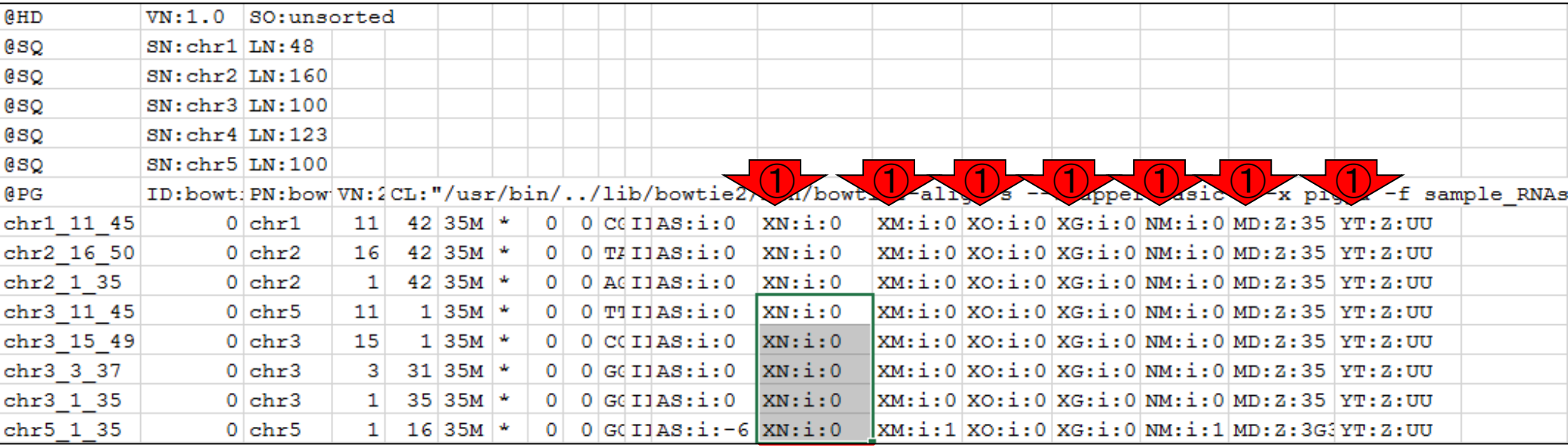
2回以上（複数個所に）マップされたリードの結果は、①XS:という結果が余分に含まれています。もしこの部分を削除して**左詰め**すれば…

@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: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	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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

①

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	CC	II	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	II	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	II	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	II	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_15_49	0	chr3	15	1	35M	*	0	0	CC	II	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_3_37	0	chr3	3	31	35M	*	0	0	GC	II	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_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS: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	GC	II	AS:i:-6	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G3	YT:Z:UU

複数個所の結果はXS

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

@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	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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

①

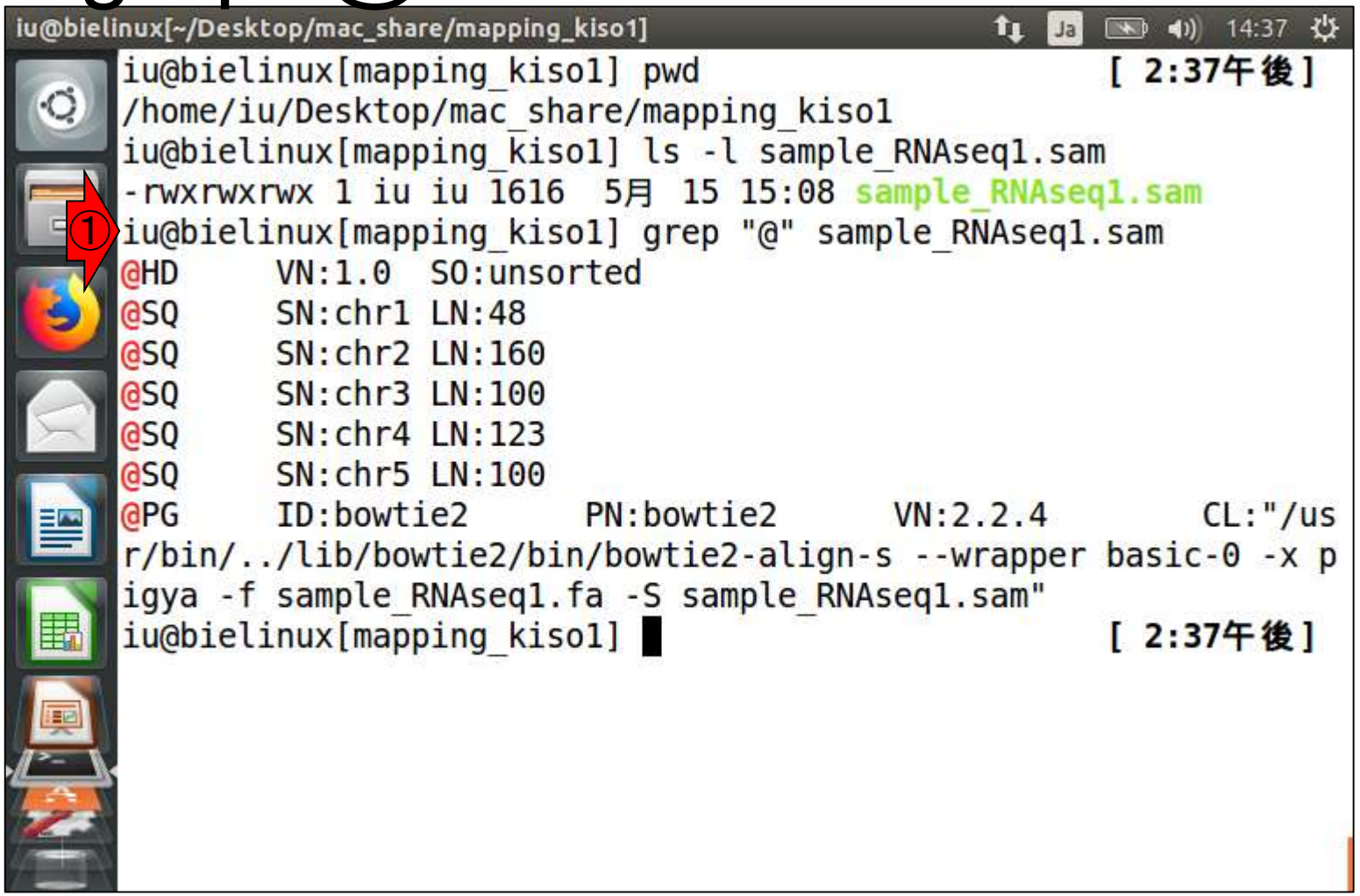
②

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

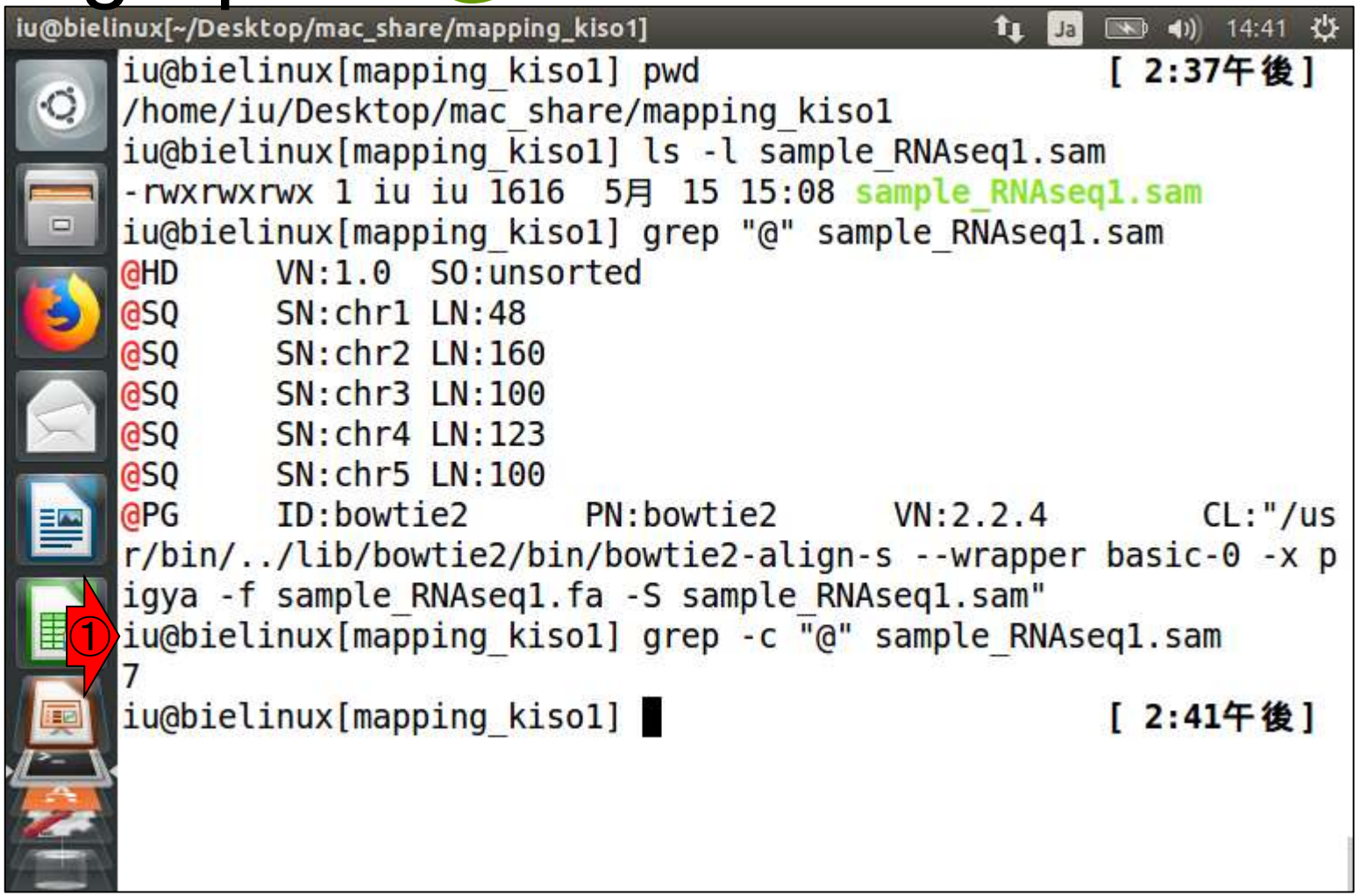
@HD	VN:1.0	SO:unsorted																
@SQ	SN:chr1	LN:48																
@SQ	SN:chr2	LN:160																
@SQ	SN:chr3	LN:100																
@SQ	SN:chr4	LN:123																
@SQ	SN:chr5	LN:100																
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs														
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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]
Ja
14:37
iu@bielinux[mapping_kiso1] pwd
[2:37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
①
iu@bielinux[mapping_kiso1] grep "@" 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 VN:2.2.4 CL:"/us
r/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x p
igya -f sample_RNAseq1.fa -S sample_RNAseq1.sam"
iu@bielinux[mapping_kiso1]
[2:37午後]

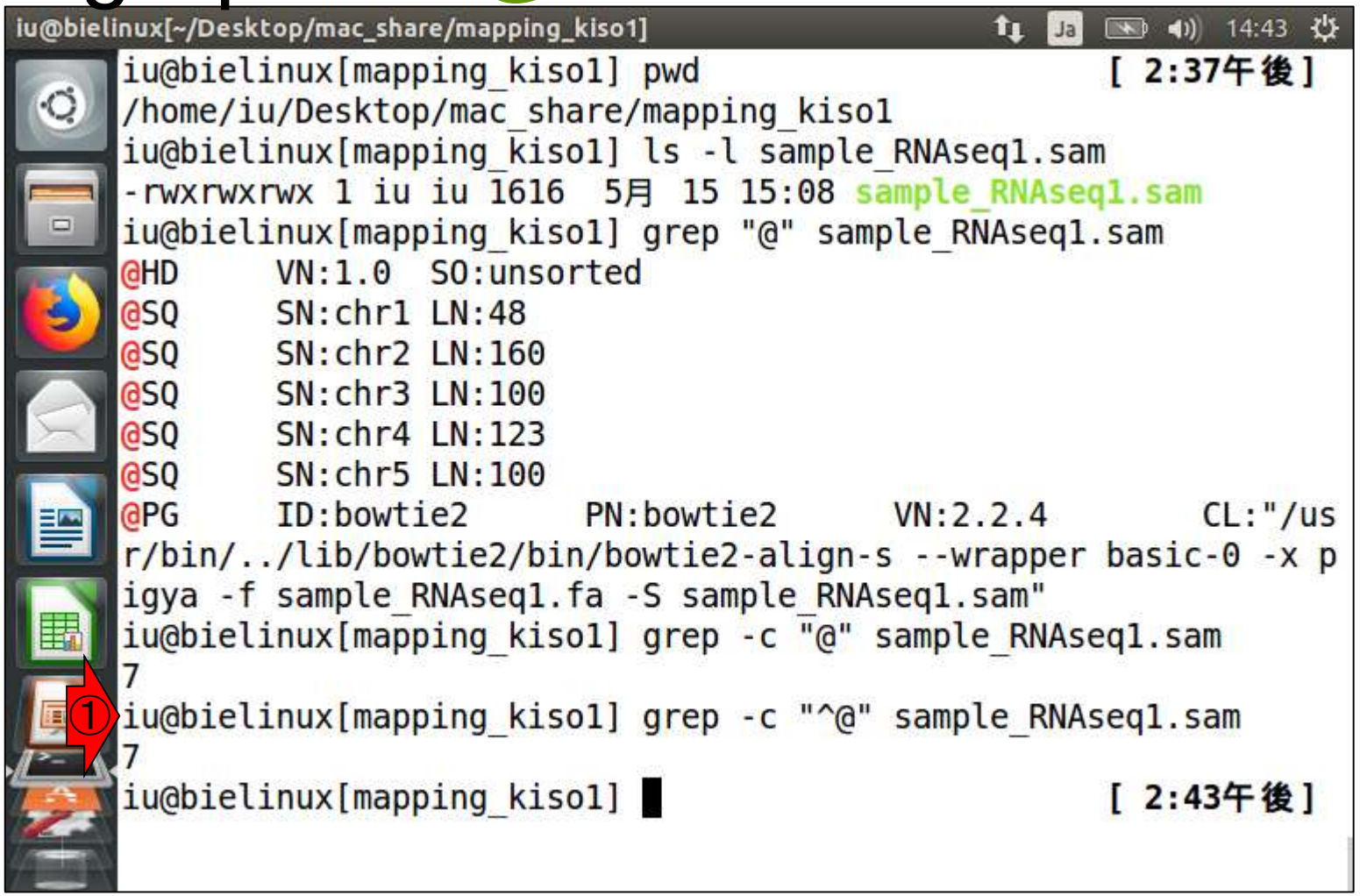
①@を含む行数を表示させたい場合

grep -c “@”



①行頭に@を含む行数を表示させたい場合

grep -c “^@”



この部分がマッピング結果情報のメイン部分

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_RNAs														
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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

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マップングのクオリティ（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:90																	
@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	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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

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

@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	CC	II	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	II	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	II	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	II	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	II	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	II	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	II	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	II	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

①ここに詳細な解説があります

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