2019.06.04版

前回の講義資料の自習部分（スライド113以降）を読んでいるという前提でスタートします。まだのヒトは講義が始まるまでにざっと見ておきましょう。

機能ゲノム学第2回

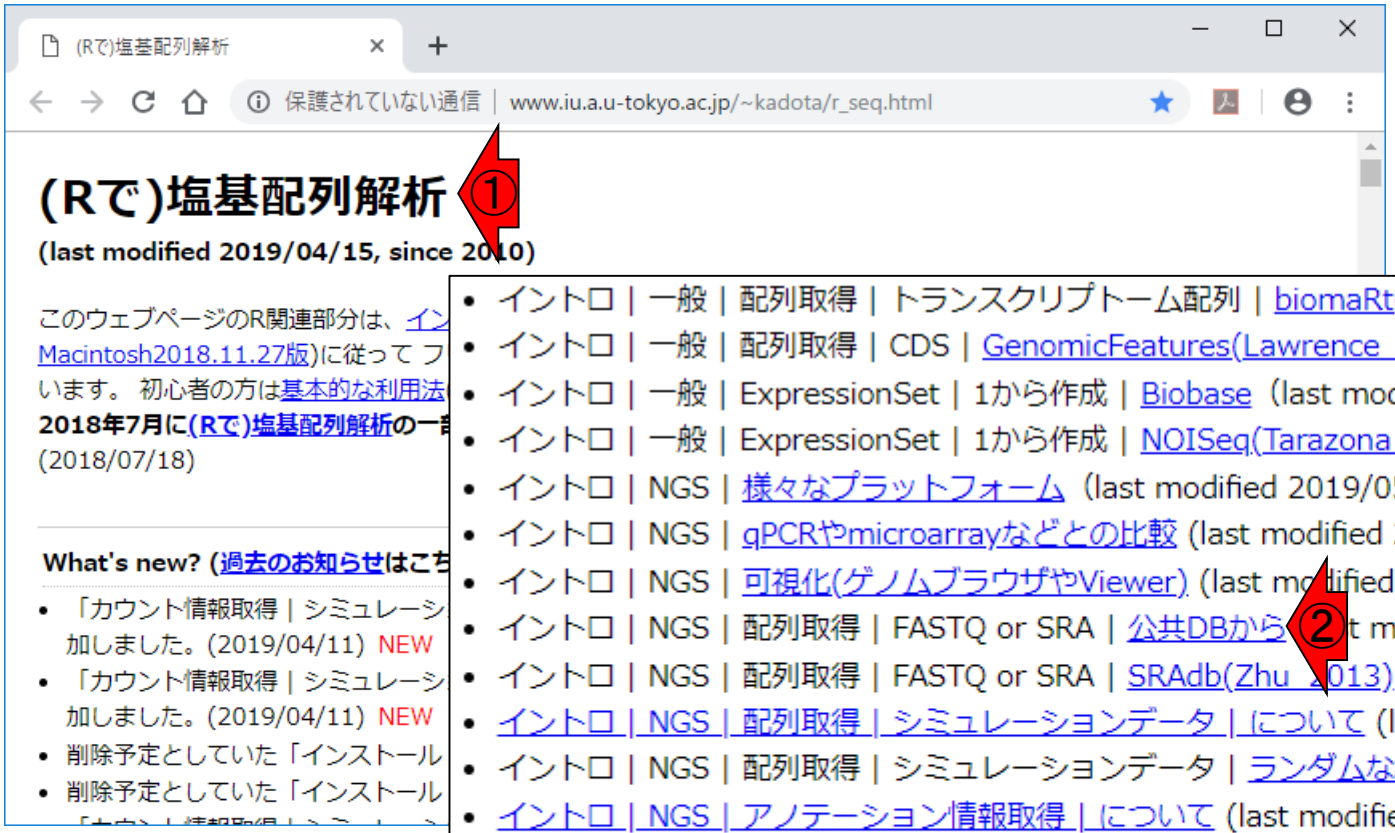
[1]大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
[2]微生物科学イノベーション連携研究機構
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

公共DB

①「(Rで)塩基配列解析」、②公共DBから、をクリック



- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009) (last modified 2015/02/20)
- イントロ | 一般 | 配列取得 | CDS | GenomicFeatures(Lawrence 2013) (last modified 2019/04/19)
- イントロ | 一般 | ExpressionSet | 1から作成 | Biobase (last modified 2018/08/01)
- イントロ | 一般 | ExpressionSet | 1から作成 | NOISeq(Tarazona 2015) (last modified 2018/08/02)
- イントロ | NGS | 様々なプラットフォーム (last modified 2019/05/21) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから ② t modified 2019/02/01)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 2019/02/01)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2019/03/19)
- イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル (last modified 2018/03/29)
- イントロ | NGS | アノテーション情報取得 | refFlat形式ファイル (last modified 2013/09/25)
- イントロ | NGS | アノテーション情報取得 | biomaRt(Durinck 2009) (last modified 2013/09/26)
- イントロ | NGS | アノテーション情報取得 | TxDb | について (last modified 2014/03/28)
- イントロ | NGS | アノテーション情報取得 | TxDb | TxDb.*から (last modified 2015/02/19)

公共DB

NGSデータの公共DBは、①DDBJ SRA（通称DRA）、②NCBI SRA（通称SRA）、③EMBL-EBI ENA（通称ENA）の三極で運用されており、データ共有がなされている。とはいえ、タイムラグは結構あるので注意してください

イントロ | NGS | 配列取得 | FASTQ or SR

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。 マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。 メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、 操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、 submitterが提供してくれている場合は（まだまだ少ないようですが）リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。
データの形式は基本的にSanger typeのFASTQ形式です。 FASTA形式はリードあたり二行（idの行と配列の行）で表現します。 FASTQ形式はリードあたり4行（@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行）で表現します。 FASTQ形式は、Sangerのものがデファクトスタンダード（業界標準）です。 かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、 NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータ[illegible]ていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

① ③

- DDBJ Sequence Read Archive (DRA) : Kodama et al., Nucleic Acids Res., 2018
- EMBL-EBI European Nucleotide Archive (ENA) : [illegible]ribio et al., Nucleic Acids Res., 2017
- NCBI Sequence Read Archive (SRA) : Sayers et al., Nucleic Acids Res., 2019

②

- ArrayExpress : Kolesnikov et al., [illegible]leic Acids Res., 2015
- GEO : Clough and Barrett, Methods Mol Biol., 2016
- DBCLS SRA : Nakazato et al., PLoS One, 2013

公共DB

①DRAと②ENAは、③大元のSRAファイルを入力として、(Linux上で使えるfastq-dumpというプログラムを実行して)④FASTQファイルを作成し、それを提供しています。FASTQ作成の際、クオリティの低いリードを除去するオプションをつけてfastq-dumpを実行しています。結果として、フィルタリング前のウェブ上の数値と異なる(が気にしなくてもよい)というお話。①をクリック

④ ③

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取
レイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロード
全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(pr
にわかることなど、 操作性の点で他を凌駕していると思います。上記でも触れて
らマッピングまでを行うのはエンドユーザレベルでは大変ですが、 submitterが
ようですが) リファレンス配列へのマップ後のデータ、つまりBAM形式ファイル
2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。
データの形式は基本的にSanger typeのFASTQ形式です。 FASTA形式はリードあたり二行 (idの行と配列の行) で表現します。 FASTQ形式はリードあたり4行 (@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行) で表現します。 FASTQ形式は、Sangerのものがデファクトスタンダード (業界標準) です。 かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、 NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータ
ていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

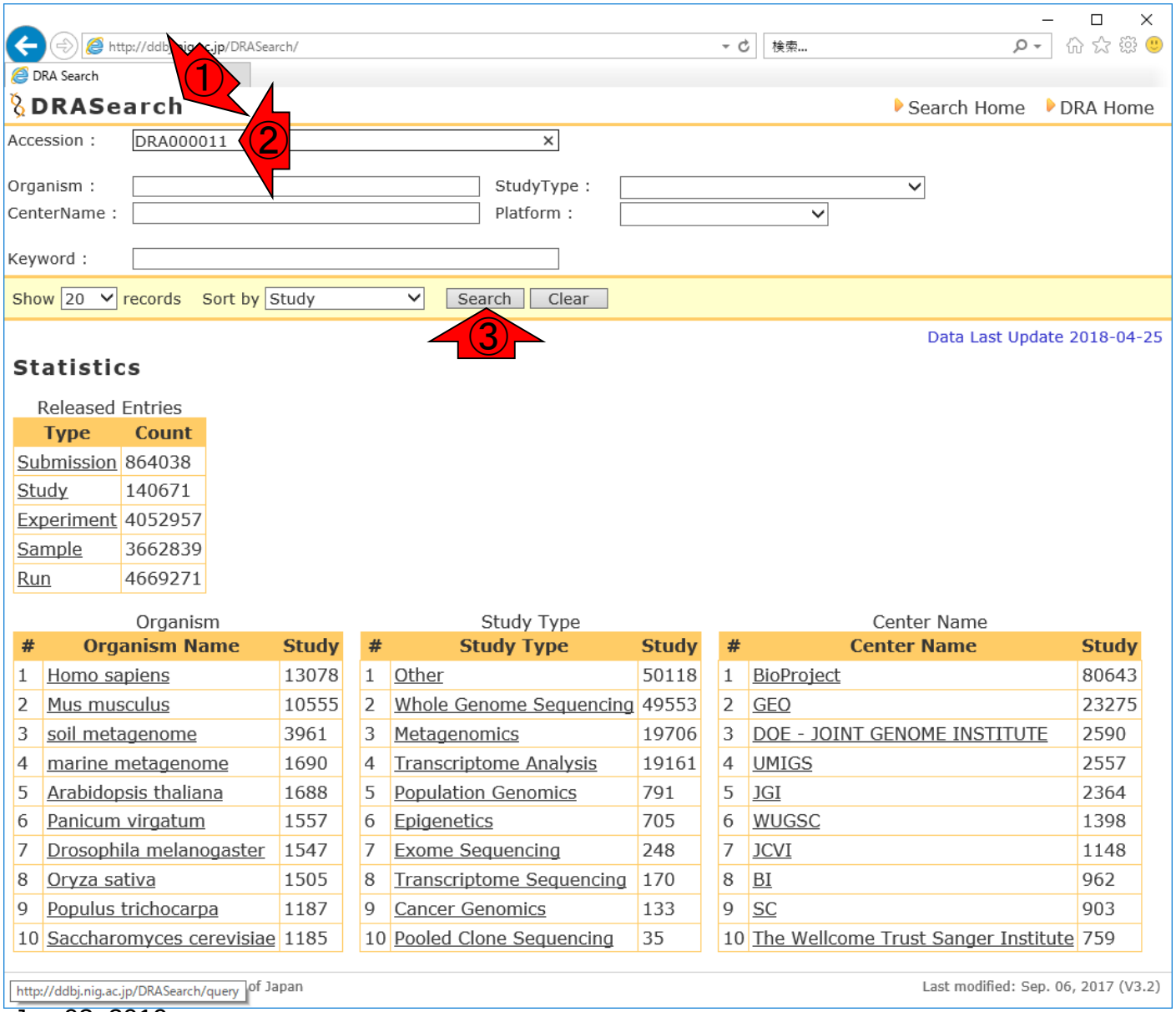
① ②

- DDBJ Sequence Read Archive (DRA) : Kodama et al., Nucleic Acids Res., 2018
- EMBL-EBI European Nucleotide Archive (ENA) : Toribio et al., Nucleic Acids Res., 2017
- NCBI Sequence Read Archive (SRA) : Sayers et al., Nucleic Acids Res., 2019
- ArrayExpress : Kolesnikov et al., Nucleic Acids Res., 2015
- GEO : Clough and Barrett, Methods Mol Biol., 2016
- DBCLS SRA : Nakazato et al., PLoS One, 2013

DRA000011

①DRA Search上で、②DRA000011と打ち込んで、③Search。前回と同じデータです

これは全員やる



Statistics

Released Entries

Type	Count
Submission	864038
Study	140671
Experiment	4052957
Sample	3662839
Run	4669271

Organism

#	Organism Name	Study
1	Homo sapiens	13078
2	Mus musculus	10555
3	soil metagenome	3961
4	marine metagenome	1690
5	Arabidopsis thaliana	1688
6	Panicum virgatum	1557
7	Drosophila melanogaster	1547
8	Oryza sativa	1505
9	Populus trichocarpa	1187
10	Saccharomyces cerevisiae	1185

Study Type

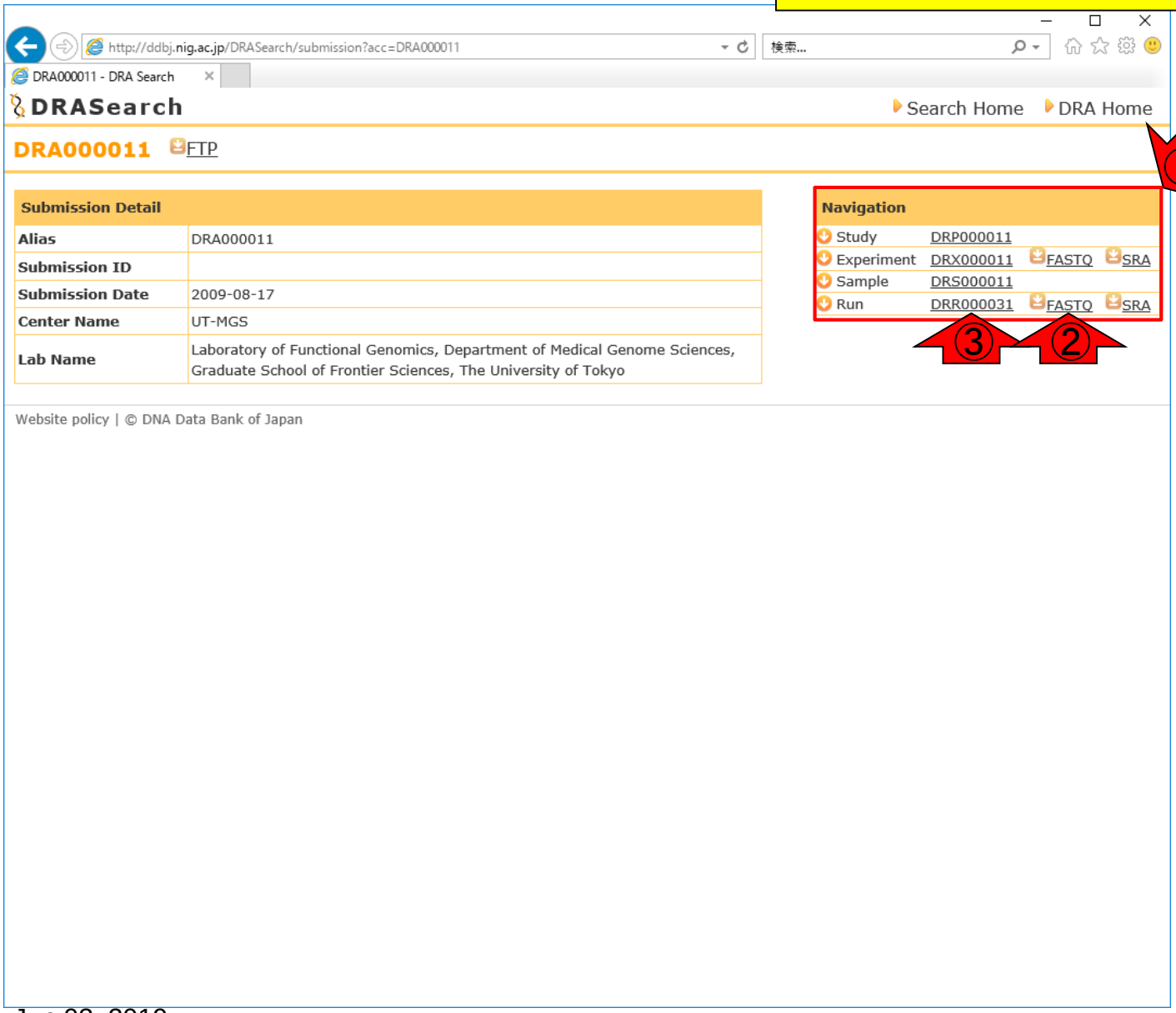
#	Study Type	Study
1	Other	50118
2	Whole Genome Sequencing	49553
3	Metagenomics	19706
4	Transcriptome Analysis	19161
5	Population Genomics	791
6	Epigenetics	705
7	Exome Sequencing	248
8	Transcriptome Sequencing	170
9	Cancer Genomics	133
10	Pooled Clone Sequencing	35

Center Name

#	Center Name	Study
1	BioProject	80643
2	GEO	23275
3	DOE - JOINT GENOME INSTITUTE	2590
4	UMIGS	2557
5	JGI	2364
6	WUGSC	1398
7	JCVI	1148
8	BI	962
9	SC	903
10	The Wellcome Trust Sanger Institute	759

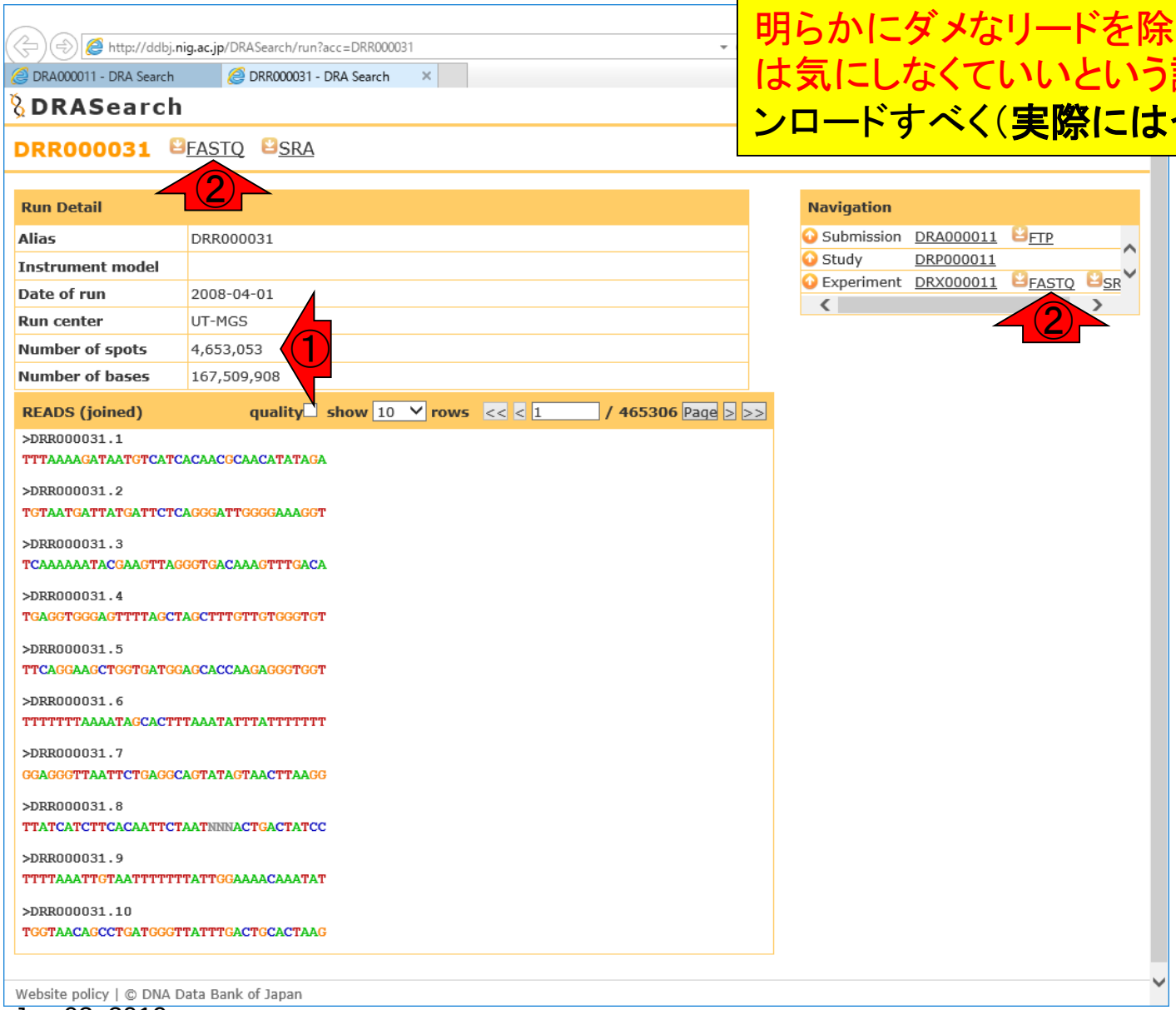
DRA000011

①ここで見られるように、様々な関連ID情報もあります。②でいきなりFASTQファイルをダウンロードできるが、③をクリックして大元のリード数情報を把握しておく



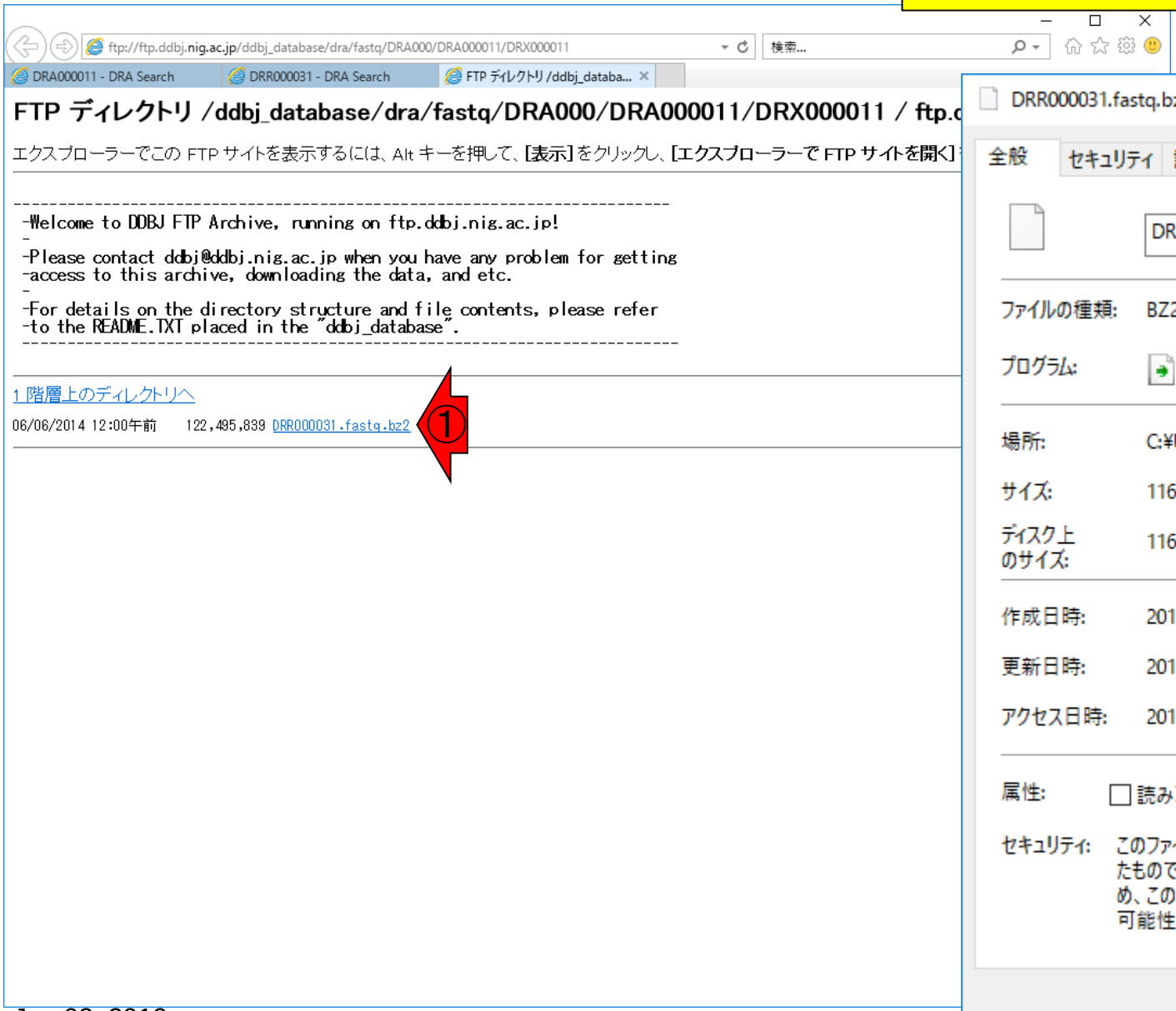
DRR000031

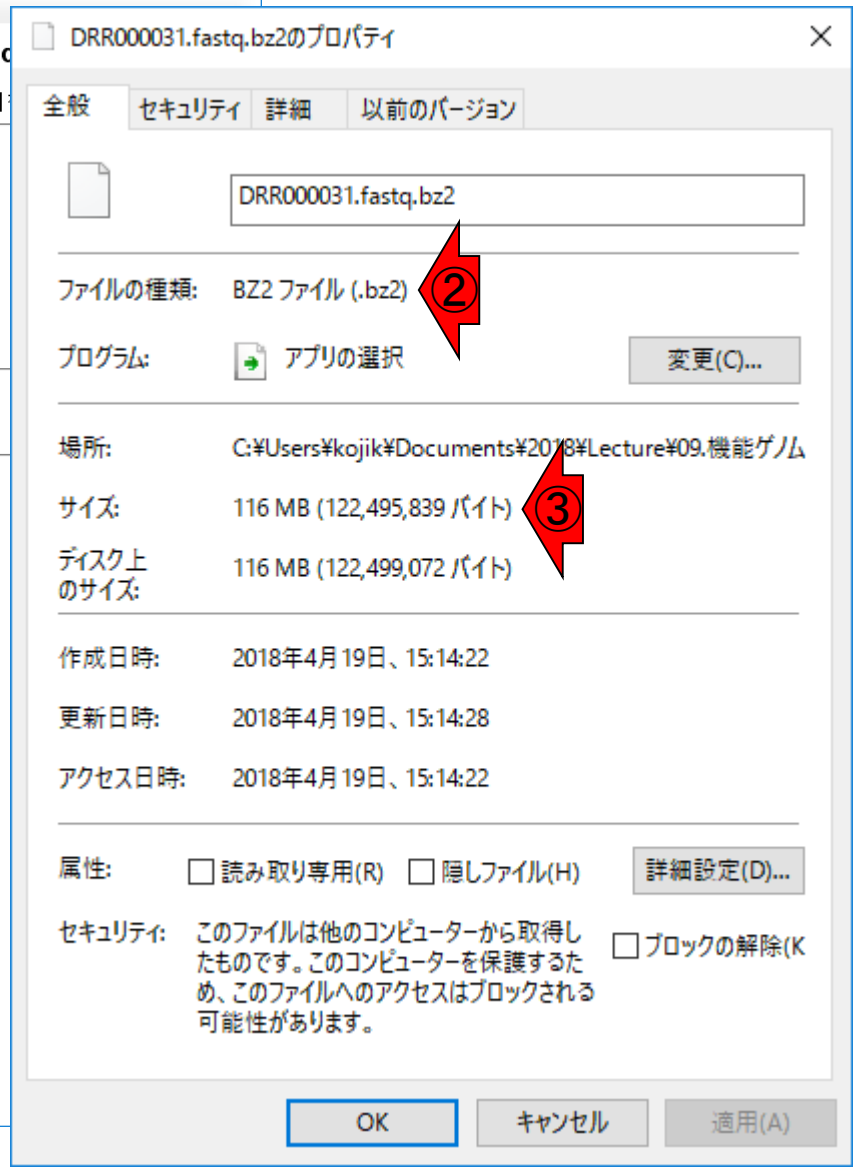
①大元のリード数は4,653,053です。これはSRAファイル中のリード数に相当します。FASTQファイルをダウンロードしてリード数を調べると、①よりも少ない数になります。明らかにダメなリードを除いているので、リード数の違いは気にしなくていいという話です。FASTQファイルをダウンロードすべく（**実際にはやらないで!**）②をクリック



DRR000031

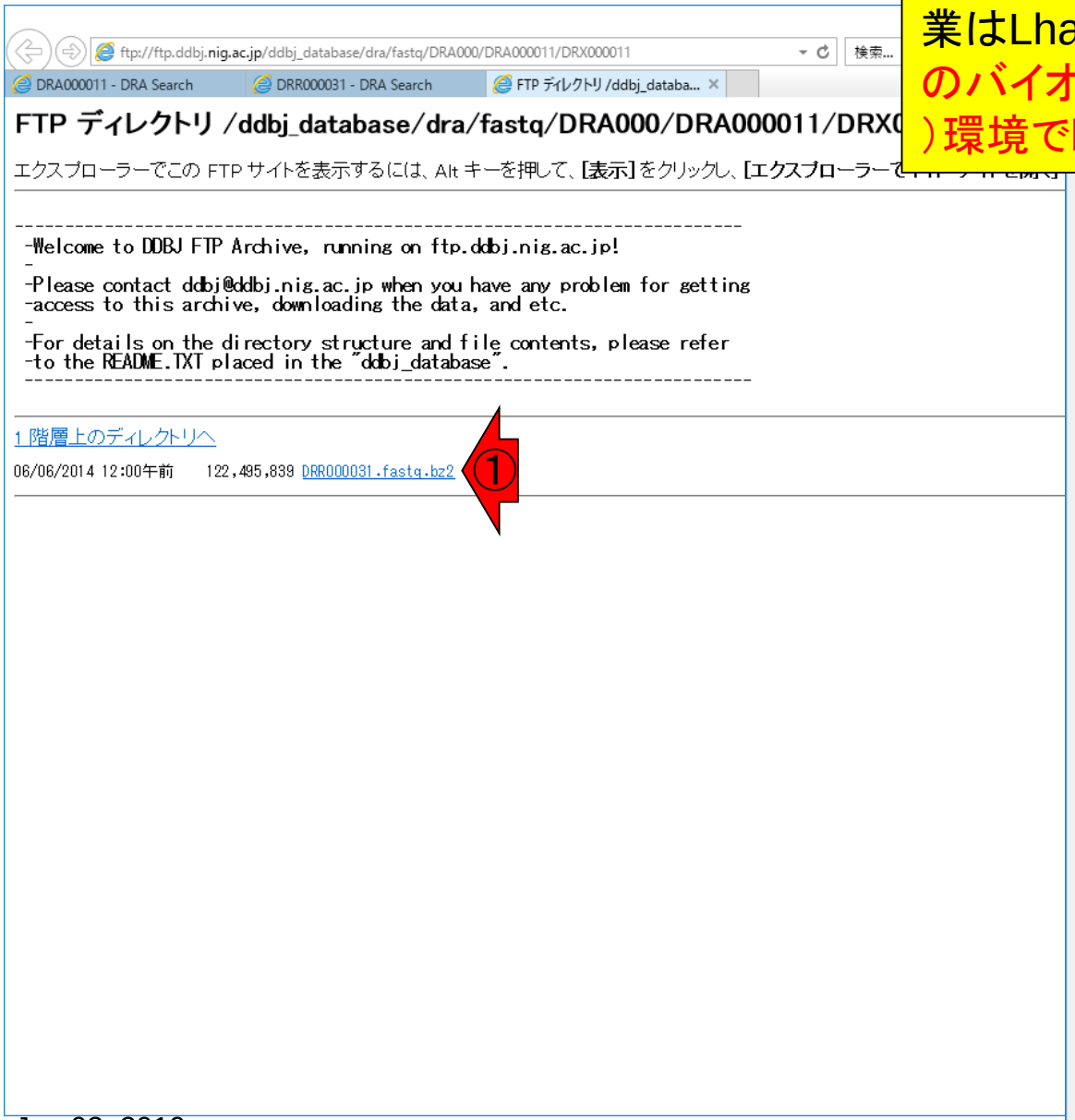
①DRR000031.fastq.bz2をダウンロード（したつもりで実際にはやらない）。②bzip2圧縮状態で、③116MB（122,495,839 bytes）あります

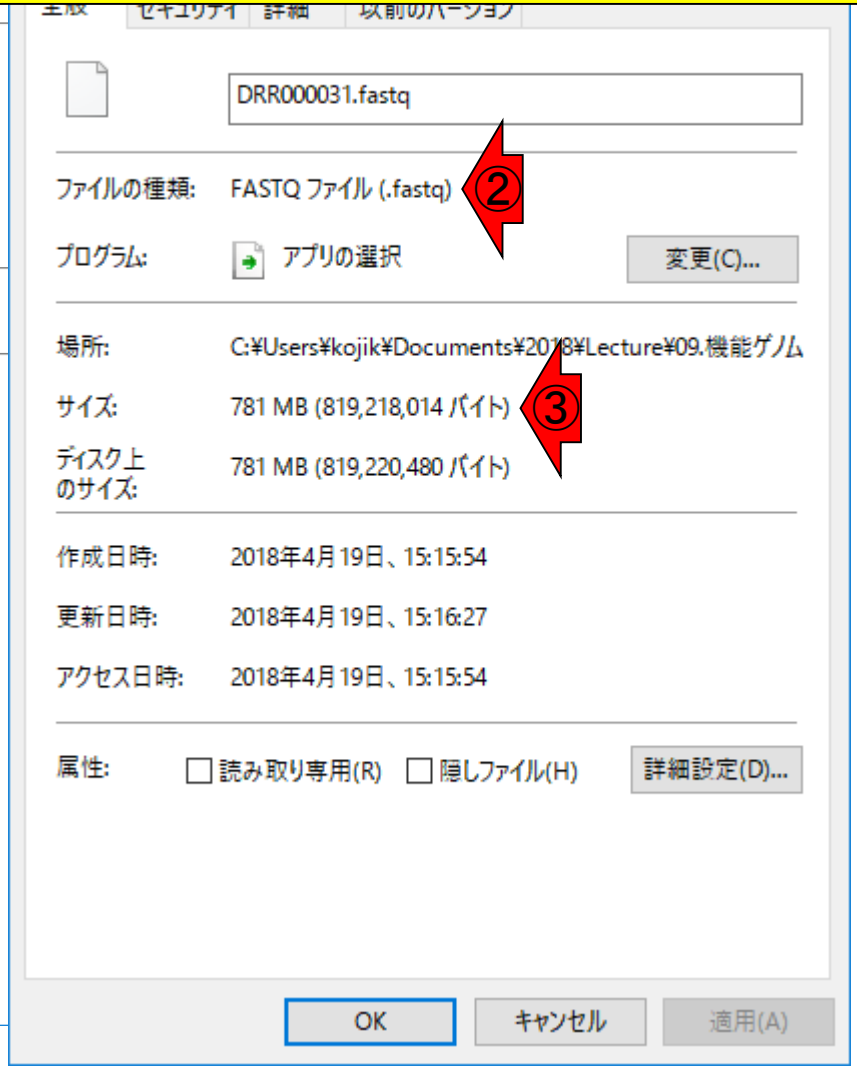




DRR000031

①bzip2圧縮ファイルを解凍して、②FASTQファイル（DRR000031.fastq）にすると、③781 MB（819,218,014 bytes）に膨れ上がります。この解凍作業はLhaplusというフリーソフトで行いましたが、多くのバイオインフォマティシャンはLinux（コマンドライン）環境でbunzipやbzip2コマンドを駆使して行います





Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

スライドを見るだけ

乳酸菌学会誌の…

①NGS連載第3回に、Linux環境下でのNGSデータのダウンロードから解凍のやり方を示しています。②の原稿PDFファイルをメインで読みつつ、実際の作業は③ウェブ資料PDFを参考にするという利用法を想定しています

(Rで)塩基配列解析のサブ

(last modified 2019/04/10, since 2010)

ここは、(Rで)塩基

What's new?

- 「参考資料 | 講
- 乳酸菌学会誌のN
  変更しました。(
- 日本乳酸菌学会
- RNA-seqカウン
- TCC-GUIのオン

- 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 2019/04/05)
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2018/09/03)
- 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 2018/09/03)
- 書籍 | 日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで (last modified 2017/07/02)
- 書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール (last modified
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会

① ② ③

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

日本乳酸菌学会誌の第3回分です。Linuxコマンドのリンク先は主に日経BP社様です。ユーザからの要望を踏まえ、ウェブ資料を更新した。念のため**オリジナル版**も残してはいますが、基本的に**軽量版**をご利用ください(2015.12.07追加)。

- 原稿PDF
- ウェブ資料PDF(**オリジナル版**; 原稿PDFと同じく、HDD150GB、約1.3億の全リードのダウンロード。)
  - Windows用(2015.07.01版; 約21MB; 非推奨)
  - Macintosh用(2015.04.27版; 約23MB)
- ウェブ資料PDF(**軽量版**; 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限してダウンロード。)
  - Windows用(2015.12.07版; 約20MB; 推奨)
  - Macintosh用(着手)

Bio-Linuxの導入と起動:

連載第2回最後のものと同じです。

- Bio-Linux: Field et al., Nat Biotechnol., 2006
- 2014年7月にリリースされたBio-Linux 8のインストール手順の概要説明

スライドを見るだけ

乳酸菌学会誌の…

①NGS連載第3回に、Linux環境下でのNGSデータのダウンロードから解凍のやり方を示しています。②の原稿PDFファイルをメインで読みつつ、実際の作業は③ウェブ資料PDFを参考にするという利用法を想定しています

(Rで)塩基配列解析のサブ

(last modified 2019/04/10, since 2010)

ここは、(Rで)塩基

What's new?

- 「参考資料 | 講
- 乳酸菌学会誌のN
  変更しました。(
- 日本乳酸菌学会
- RNA-seqカウン
- TCC-GUIのオン

- 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 2019/04/05)
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2018/09/03)
- 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 2018/09/03)
- 書籍 | 日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで (last modified 2017/07/02)
- 書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール (last modified
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会

① ② ③

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

日本乳酸菌学会誌の第3回分です。Linuxコマンドのリンク先は主に日経BP社様です。ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のため**オリジナル版**も残してはいますが、基本的に**軽量版**をご利用ください(2015.12.07追加)。

- 原稿PDF
- ウェブ資料PDF(**オリジナル版**; 原稿PDFと同じく、HDD150GB、約1.3億の全リードのダウンロード。)
  - Windows用(2015.07.01版; 約21MB; 非推奨)
  - Macintosh用(2015.04.27版; 約23MB)
- ウェブ資料PDF(**軽量版**; 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限してダウンロード。)
  - Windows用(2015.12.07版; 約20MB; 推奨)
  - Macintosh用(着手)

Bio-Linuxの導入と起動:

連載第2回最後のものと同じです。

- Bio-Linux: Field et al., Nat Biotechnol., 2006
- 2014年7月にリリースされたBio-Linux 8のインストール手順の概要説明

乳酸菌学会誌の…

①NGS連載第3回に、Linux環境下でのNGSデータのダウンロードから解凍のやり方を示しています。②の原稿PDFファイルをメインで読みつつ、実際の作業は③ウェブ資料PDFを参考にするという利用法を想定しています。
④の部分が左上になるようにしたのが次のスライド。

(Rで)塩基配列解析のサブ

(last modified 2019/04/10, since 2010)

ここは、(Rで)塩基

What's new?

- 「参考資料 | 講
- 乳酸菌学会誌のN
  変更しました。(
- 日本乳酸菌学会
- RNA-seqカウン
- TCC-GUIのオン

- 書籍 | トランスクリプトーム解析 | 4.3.3 2群間比較 (last mo
- 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
- 書籍 | 日本乳酸菌学会誌 | について (last modified 2019/04/05)
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2018/09/03)
- 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 2018/09/03)
- 書籍 | 日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで (last modified 2017/07/02)
- 書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール (last modified
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会
- 書籍 | 日本乳酸菌学会

① 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

日本乳酸菌学会誌の第3回分です。Linuxコマンドのリンク先は主に日経BP社様です。ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のため**オリジナル版**も残してはいますが、基本的に**軽量版**をご利用ください(2015.12.07追加)。

- 原稿PDF
- ウェブ資料PDF(**オリジナル版**; 原稿PDFと同じく、HDD150GB、約1.3億の全リードのダウンロード。)
  - Windows用(2015.07.01版; 約21MB; 非推奨)
  - Macintosh用(2015.04.27版; 約23MB)
- ウェブ資料PDF(**軽量版**; 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限してダウンロード。)
  - Windows用(2015.12.07版; 約20MB; 推奨)
  - Macintosh用(着手)

③ ④

Bio-Linuxの導入と起動:

連載第2回最後のものと同じです。

- Bio-Linux: Field et al., Nat Biotechnol., 2006
- 2014年7月にリリースされたBio-Linux 8のインストール手順の概要説明

Bio-Linux

こんな感じです。このあたりを見ながらWindowsまたはMacintosh上で仮想環境構築（Linuxの導入）を行います。とはいえ、Macは「ターミナル」上でLinuxコマンドが普通に使えます。また、WindowsもWindows Subsystem for Linux (WSL)が実用に耐えうるレベルになってきています。WSLは、乳酸菌学会誌NGS連載第11回で述べたBash on Ubuntu on Windows（ベータ版）の正式リリース版という位置づけです。①2019年冬に、コマンドプロンプトやWSLを統合した「Windows Terminal」というものがリリースされる予定だそうです。

Bio-Linuxの導入と起動：

連載第2回最後のものと同じです。

- Bio-Linux：Field et al., Nat Biotechnol., 2006

- 2014年7月にリリースされたBio-Linux 8のインストール
  Bio-Linuxが使えるようにするには、大まかに以下の4つ
  1. 仮想化ソフト(ここではVirtualBox)のインストール
  2. VirtualBox Extension Packのインストール
  3. 仮想マシン(Virtual Machine)の新規作成。Windows/Macintoshの既存マシン(ホストOS)中に新たなLinux環境(ゲストOS)を作成することに相当するため、 既存マシンの空き容量からどの程度をLinux環境に割り振るかを指定するところ。この段階ではまだLinux OSはインストールされていない。
  4. Bio-Linux 8のインストール。isoという拡張子がついた4GB弱のファイルのダウンロードとインストール。日本語ではなく英語で行うのがポイント。 比較的シンプルなovaという拡張子がついたファイルからのインストール法も示しました(2015/11/25追加)。

- 1. VirtualBox、および2. Extension Packのインストール手順：
  - Windows用(2018.09.03版; 約1MB)
  - Macintosh用(2015.11.18版; 約8MB)
- 3. 仮想マシンの作成、および4. Bio-Linux 8の**iso**ファイルからのインストール手順：
  - Windows用(2015.11.19版; 約6MB)
  - Macintosh用(2015.11.19版; 約5MB)
- Bio-Linux 8の**ova**ファイルからのインストール手順(2015/11/25追加)：
  - Windows用(2015.11.24版; 約2MB)
  - Macintosh用(2015.11.25版; 約4MB)

https://www.itmedia.co.jp/news/articles/1905/07/news080.html

Bio-Linux

Bio-Linuxの導入と起動：①

連載第2回最後のものと同じです。

- Bio-Linux：Field et al., Nat Biotechnol., 2006

- 2014年7月にリリースされたBio-Linux 8のインストール
  Bio-Linuxが使えるようにするには、大まかに以下の4つ
  1. 仮想化ソフト(ここではVirtualBox)のインストール
  2. VirtualBox Extension Packのインストール
  3. 仮想マシン(Virtual Machine)の新規作成。WindowsやMacintoshの既存マシン(ホストOS)中に新たなLinux環境(ゲストOS)を作成することに相当するため、 既存マシンの空き容量からどの程度をLinux環境に割り振るかを指定するところ。この段階ではまだLinux OSはインストールされていない。
  4. Bio-Linux 8のインストール。isoという拡張子がついた4GB弱のファイルのダウンロードとインストール。日本語ではなく英語で行うのがポイント。 比較的シンプルなovaという拡張子がついたファイルからのインストール法も示しました(2015/11/25追加)。

- 1. VirtualBox、および2. Extension Packのインストール手順：
  - Windows用(2018.09.03版; 約1MB)
  - Macintosh用(2015.11.18版; 約8MB)
- 3. 仮想マシンの作成、および4. Bio-Linux 8の**iso**ファイルからのインストール手順：
  - Windows用(2015.11.19版; 約6MB)
  - Macintosh用(2015.11.19版; 約5MB)
- Bio-Linux 8の**ova**ファイルからのインストール手順(2015/11/25追加)：
  - Windows用(2015.11.24版; 約2MB)
  - Macintosh用(2015.11.25版; 約4MB)

連載第3回の①Bio-Linuxの詳細については、(興味をもったヒトは)各自で第3回の原稿を読んで自習してください。大まかには「Bio-Linuxは、Linux上で動く一通りのプログラム群がインストールされているので導入に時間もかかって大変だが、全受講生のレベルが一定以上あればMacもWinも同じ見栄えなので教える側も楽だし教わる側もOSの違いに惑わされなくてよいもの」です。但し、2014年以降アップデートされていないようなので、out-of-dateになりつつある…

Bio-Linux

連載第3回の①Bio-Linuxの導入の説明として、②がありますが、この部分のアップデート版に相当するのが…

Bio-Linuxの導入と起動： ①

連載第2回最後のものと同じです。

- Bio-Linux：Field et al., Nat Biotechnol., 2006

- 2014年7月にリリースされたBio-Linux 8のインストール手順の概要説明
  Bio-Linuxが使えるようにするには、大まかに以下の4つ手順を行う必要があります。
  1. 仮想化ソフト(ここではVirtualBox)のインストール。
  2. VirtualBox Extension Packのインストール
  3. 仮想マシン(Virtual Machine)の新規作成。WindowsやMacintoshの既存マシン(ホストOS)中に新たなLinux環境(ゲストOS)を作成することに相当するため、 既存マシンの空き容量からどの程度をLinux環境に割り振るかを指定するところ。この段階ではまだLinux OSはインストールされていない。
  4. Bio-Linux 8のインストール。isoという拡張子がついた4GB弱のファイルのダウンロードとインストール。日本語ではなく英語で行うのがポイント。 比較的シンプルなovaという拡張子がついたファイルからのインストール法も示しました(2015/11/25追加)。

- 1. VirtualBox、および2. Extension Packのインストール手順：
  - Windows用(2018.09.03版; 約1MB)
  - Macintosh用(2015.11.18版; 約8MB)
- 3. 仮想マシンの作成、および4. Bio-Linux 8の**iso**ファイルからのインストール手順：
  - Windows用(2015.11.19版; 約6MB)
  - Macintosh用(2015.11.19版; 約5MB)
- Bio-Linux 8の**ova**ファイルからのインストール手順(2015/11/25追加)：
  - Windows用(2015.11.24版; 約2MB)
  - Macintosh用(2015.11.25版; 約4MB)

②

Bio-Linux

連載第3回の①Bio-Linuxの導入の説明として、②がありますが、この部分のアップデート版に相当するのが、③第6回の、④です。Bio-Linuxを導入してみたいヒトにとっては、多少は参考になるかと思います。但し、2年以上動作確認しておりませんので、うまくいかないかもしれません。

Bio-Linuxの導入と起動： ①

連載第2回最後のものと同じです。

- Bio-Linux：Field et al., Nat Biotechnol.
- 2014年7月にリリースされたBio-Linux 8
  Bio-Linuxが使えるようにするには、大ま
  1. 仮想化ソフト(ここではVirtualBox)
  2. VirtualBox Extension Packのイン
  3. 仮想マシン(Virtual Machine)の新規
     なLinux環境(ゲストOS)を作成する
     環境に割り振るかを指定するところ
  4. Bio-Linux 8のインストール。isoと
     ル。日本語ではなく英語で行うのが
     からのインストール法も示しました
- 1. VirtualBox、および2. Extension Pa
  - Windows用(2018.09.03版; 約1
  - Macintosh用(2015.11.18版; 約8M
- 3. 仮想マシンの作成、および4. Bio-Linux 8の**iso**ファイルからのインストール手順：
  - Windows用(2015.11.19版; 約6MB)
  - Macintosh用(2015.11.19版; 約5MB)
- Bio-Linux 8の**ova**ファイルからのインストール手順(2015/11/25追加)：
  - Windows用(2015.11.24版; 約2MB)
  - Macintosh用(2015.11.25版; 約4MB)

②

書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ ③

日本乳酸菌学会誌の第6回分です。 Linuxコマンドのリンク先は主に日経BP社様です。 原稿PDFの「はじめに」項目のところにtypoがあります。誤：するであれば、正：する"の"であれば、ですねm(_ _)m。 また、「配列長によるフィルタリング」項目の最後の文章「これらについては、第7回で詳述する予定である。」についてですが、これは「これらについては、"第8回以降"で詳述する予定である。」と読み替えてください。第7回ドラフト原稿作成時点で、これらの内容は含まれていないためです（2016年4月23日追加）。

- 原稿PDF
- ウェブ資料PDF
  - Windows用(2016.06.17版; 約20MB)
  - Macintosh用

④

- (共有フォルダ設定情報を含む)連載第3回終了時点以降の**ova**ファイルからのインストール手順：
  - Windows用(2015.12.28版; 約2MB)
  - Macintosh用(2015.12.28版; 約2MB)

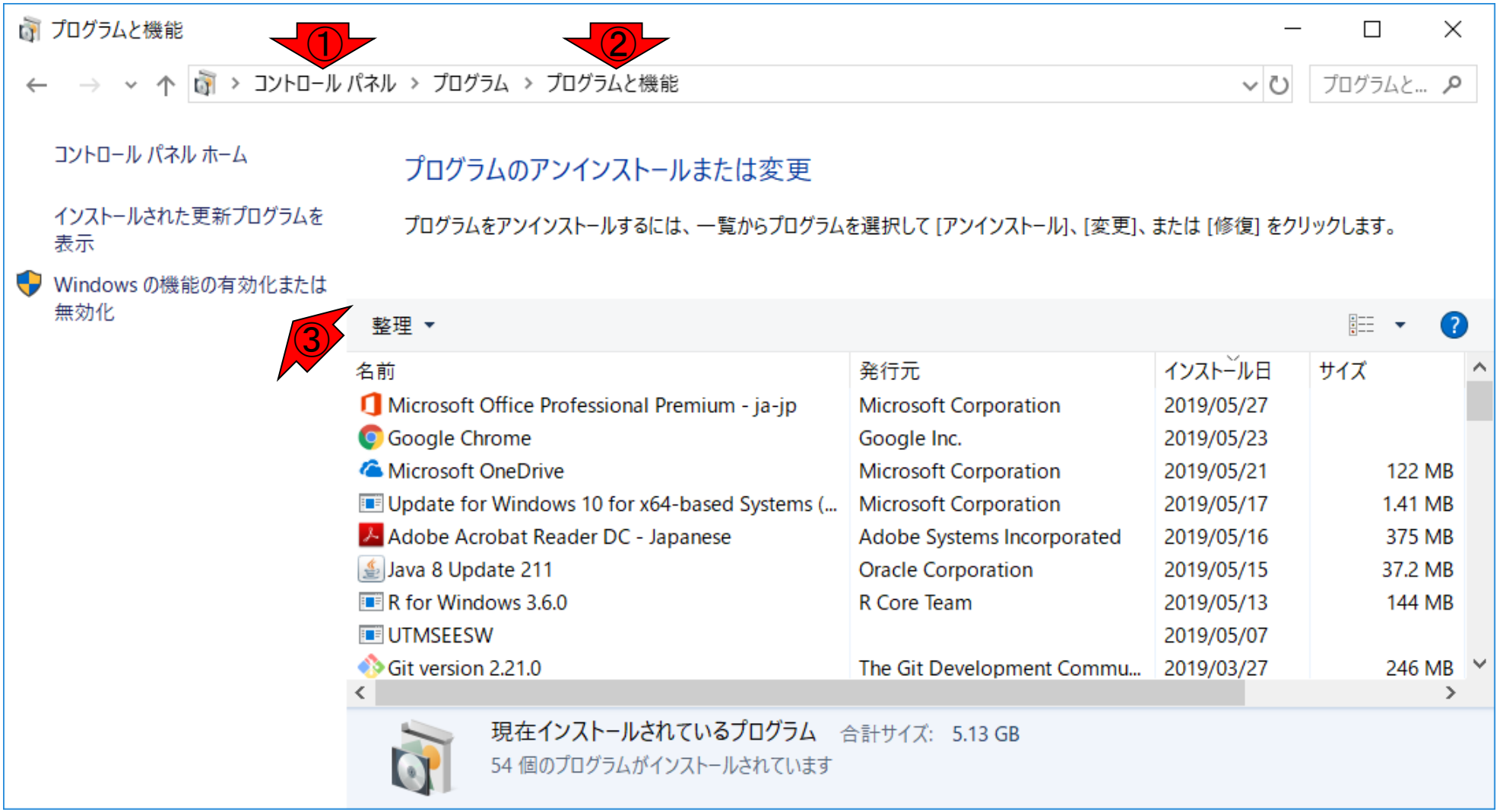
Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

参考

WSL機能の有効化

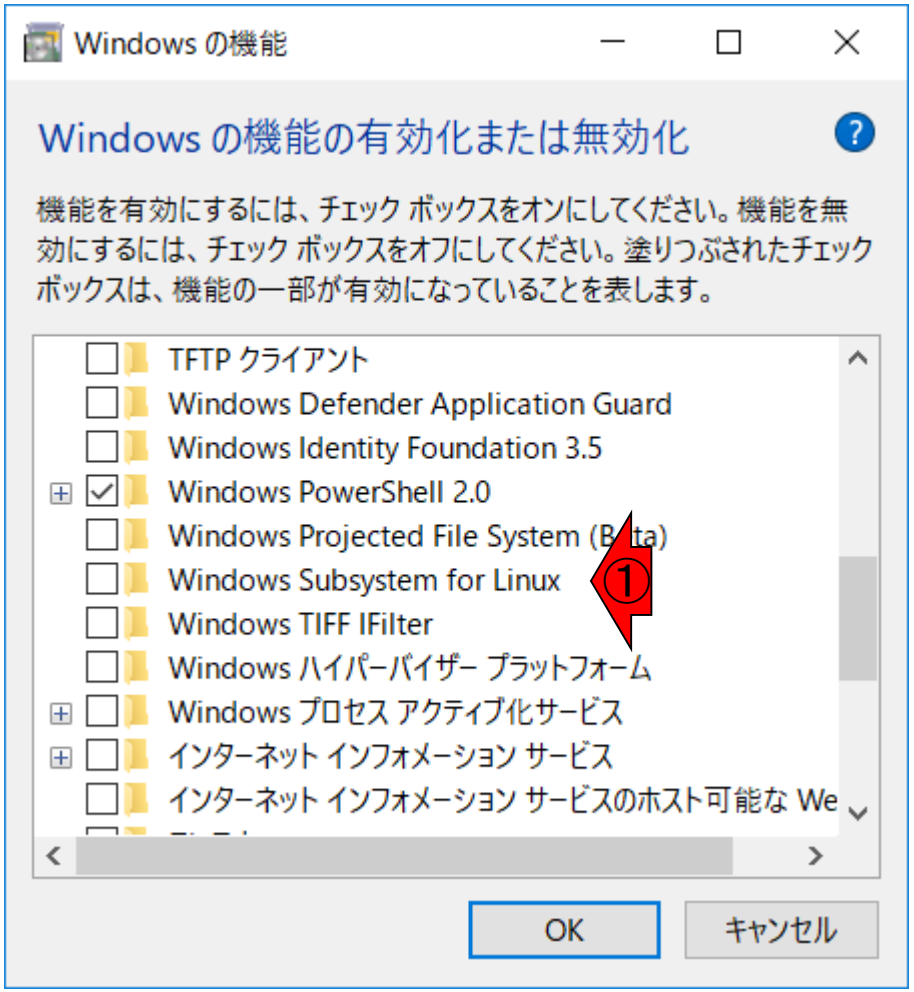
前処理として、①コントロールパネル、②プログラムと機能、③Windowsの機能の有効化または無効化。

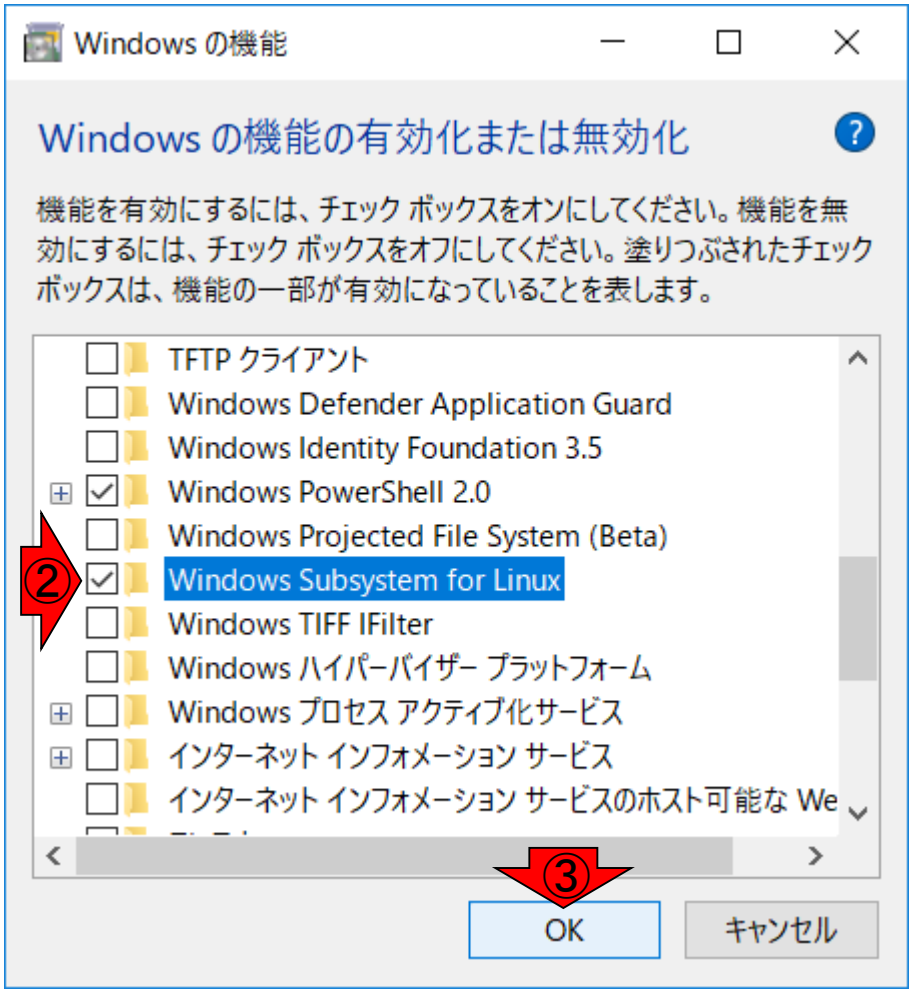


参考

WSL機能の有効化

Windowsの機能、が開く。デフォルトは①にチェックが入っていないので、②チェックを入れて、③OK。

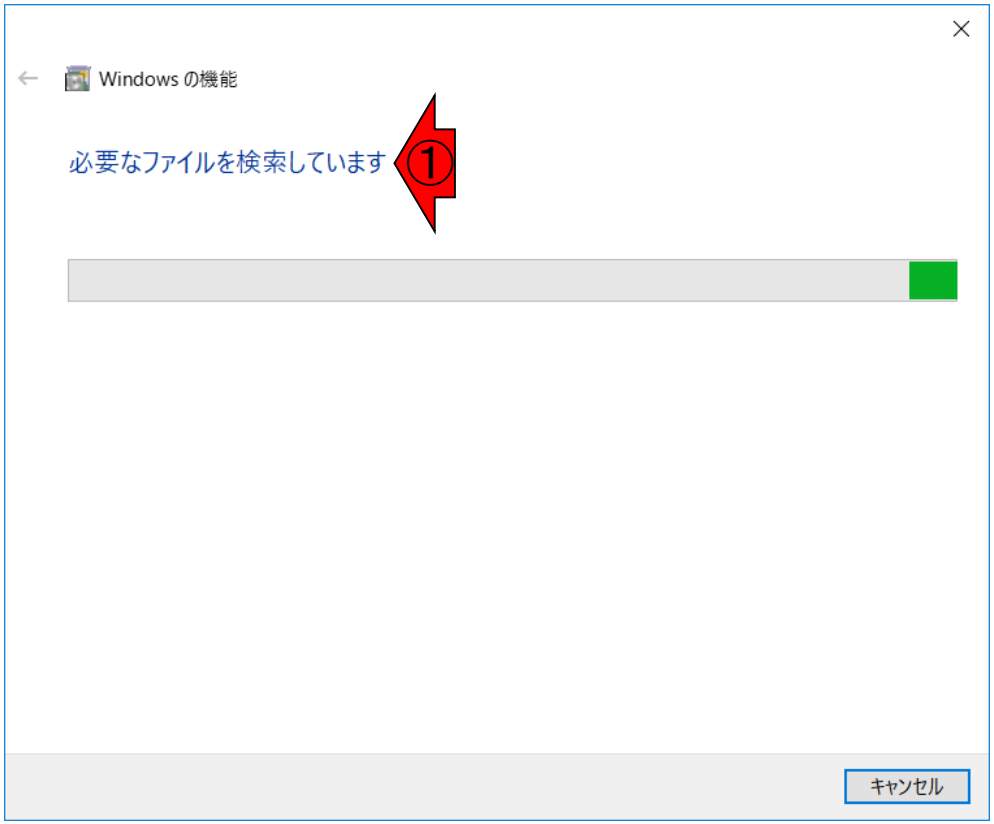


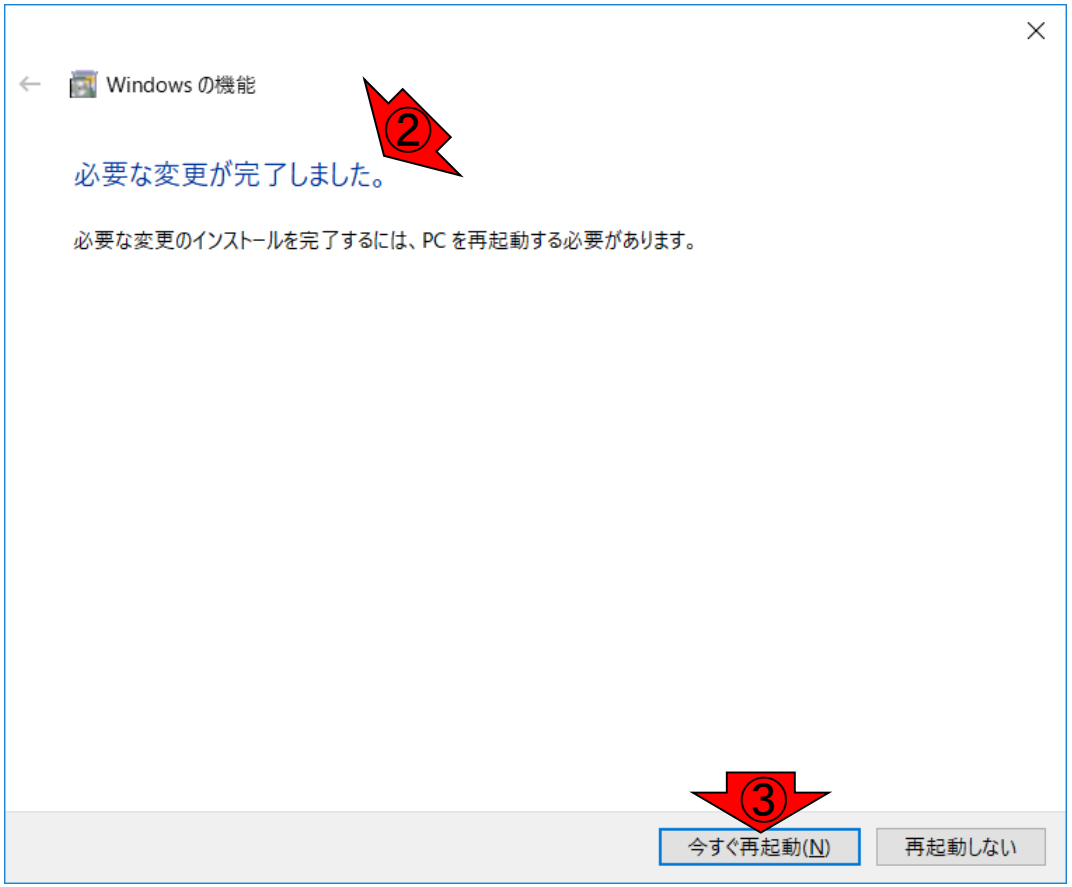


参考

WSL機能の有効化

すると、①のような感じになって、数十秒後に②のようになります。③再起動。

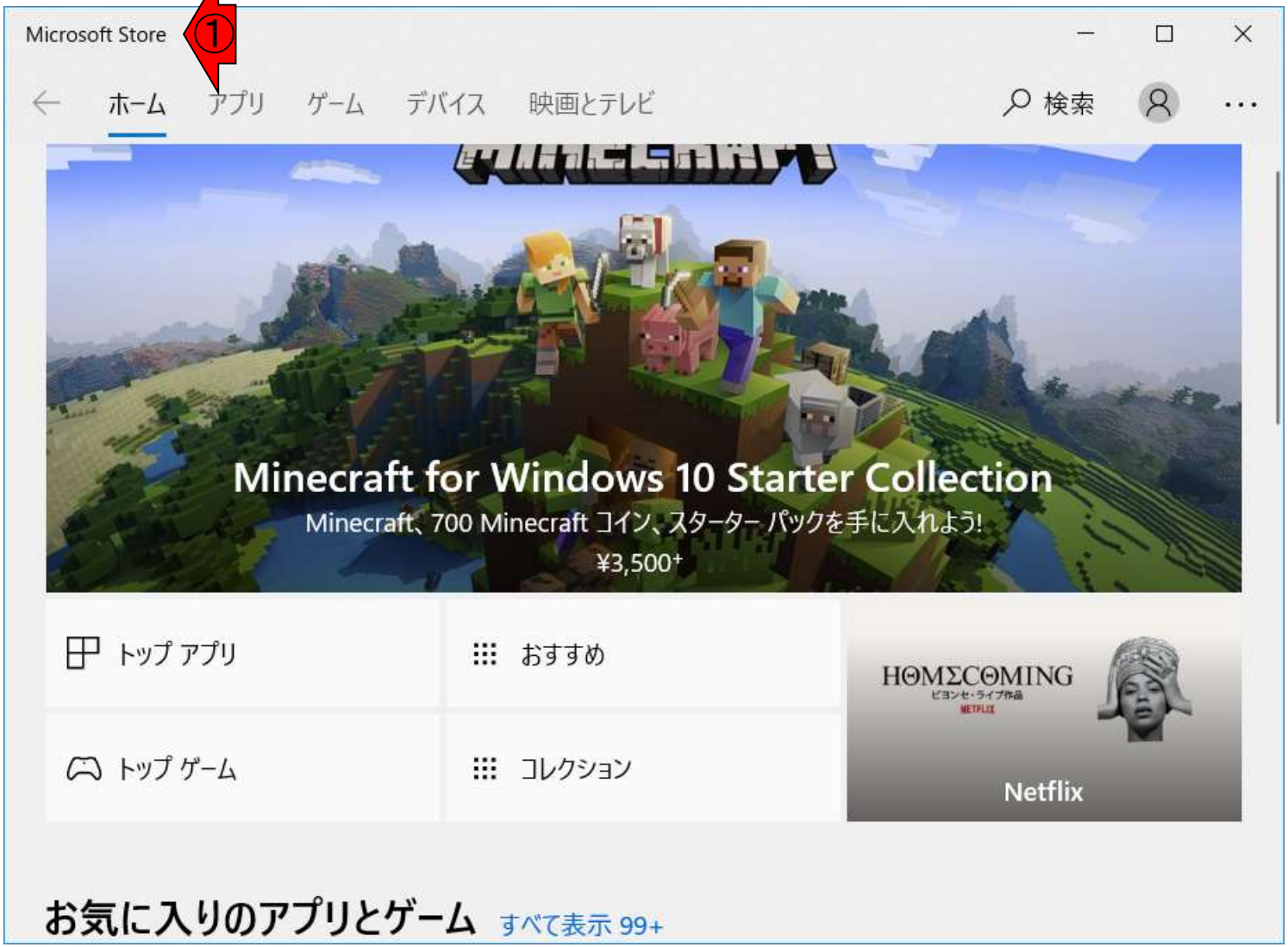




参考

WSLインストール

再起動後に、①Microsoft Storeを開く（Windows画面の左下にある検索窓で「ストア」で検索するとヒットします）。



参考

WSLインストール

①Microsoft Storeを開く（Windowsの左下にある検索窓で「ストア」で検索するとヒットします）。さらにこの中の検索窓上で、②WSLと入力するとこんな感じになるので、③を選択します。貸与PCはインストール済みです。

参考

こんな感じになります。①Ubuntuを選択。

WSLインストール

参考

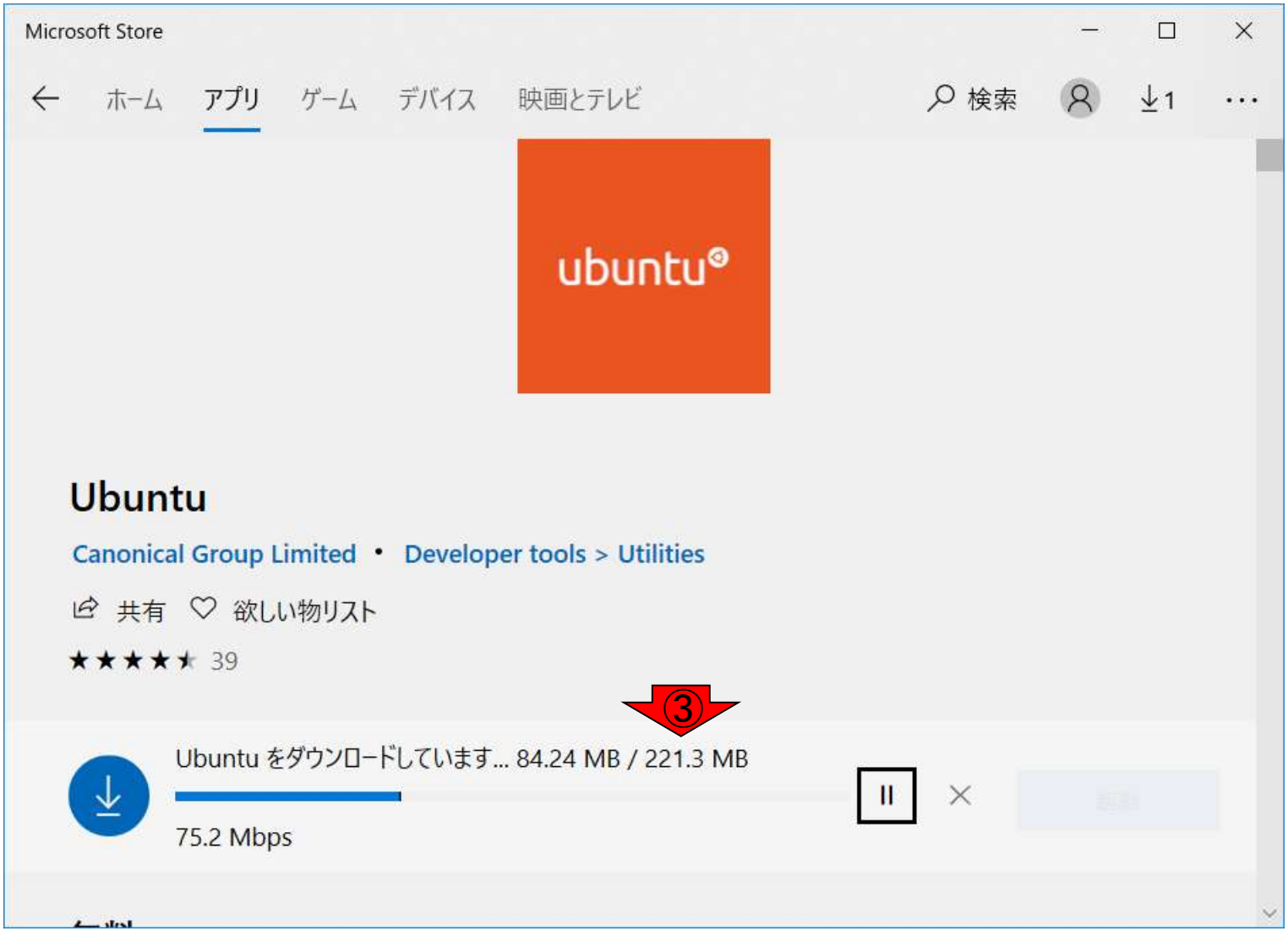
こんな感じになります。①Ubuntuを選択。②入手。

WSLインストール

参考

WSLインストール

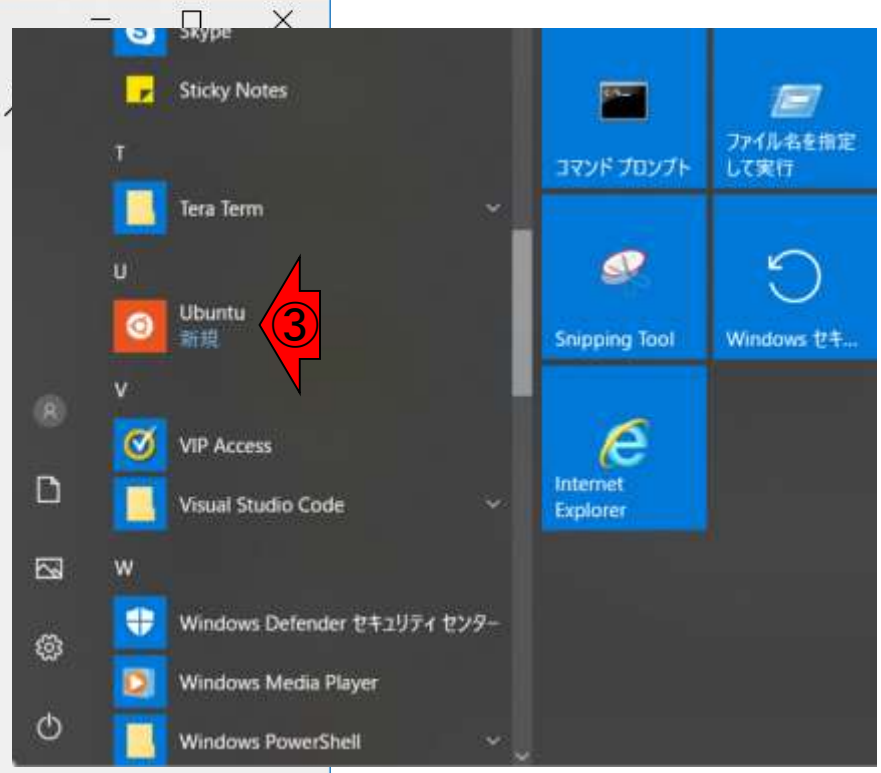
こんな感じになります。①Ubuntuを選択。②入手。ダウンロード中です。③200MB超なのでそこそこデカいです。



参考

WSLインストール

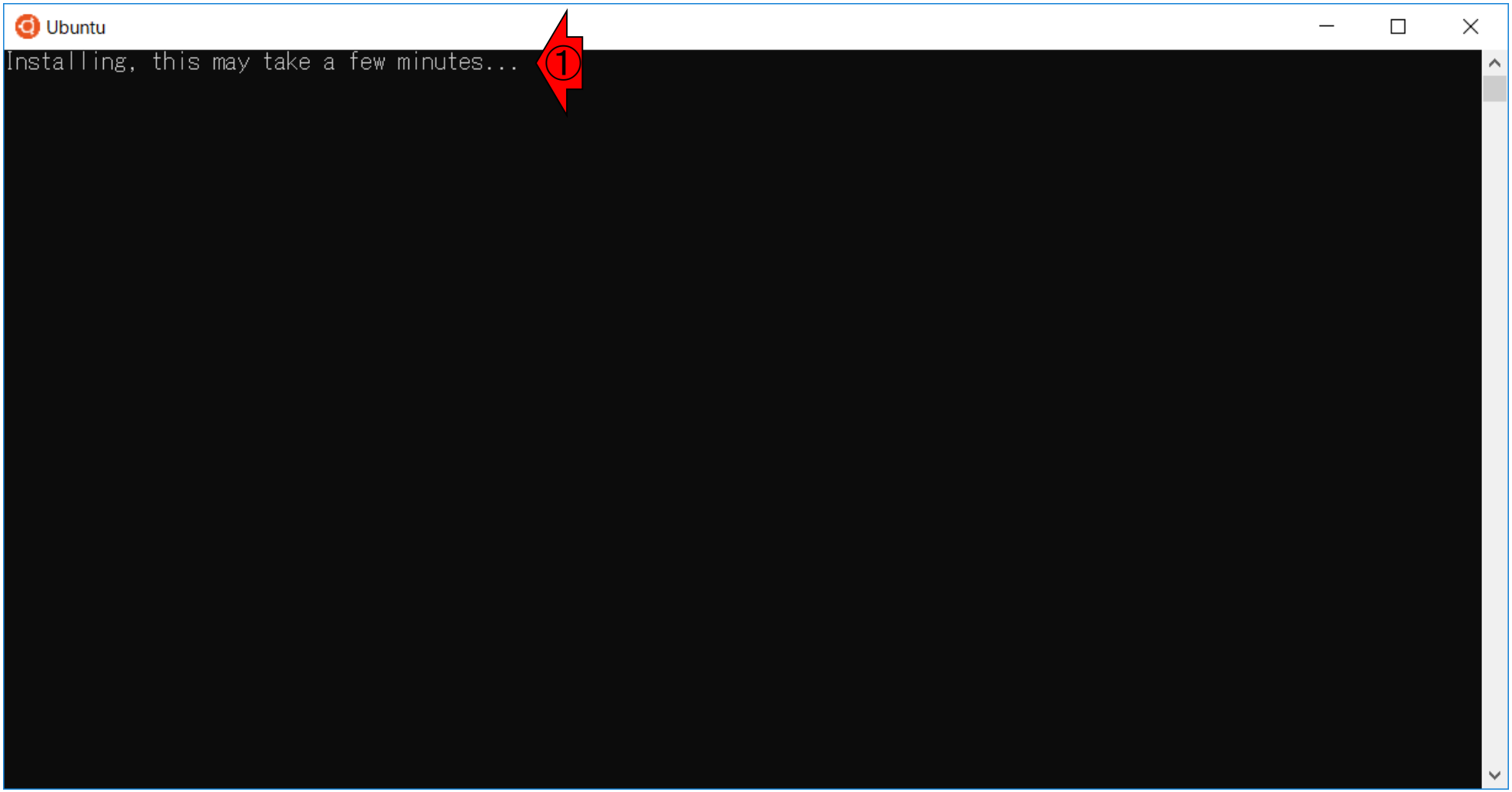
①無事インストール完了。②起動。後で起動したい場合は、スタートメニューから③Ubuntuを起動。つまり、WSLといいつつ、実際にはUbuntuだということ。



参考

こんな感じになります。①数分かかると書かれてます。

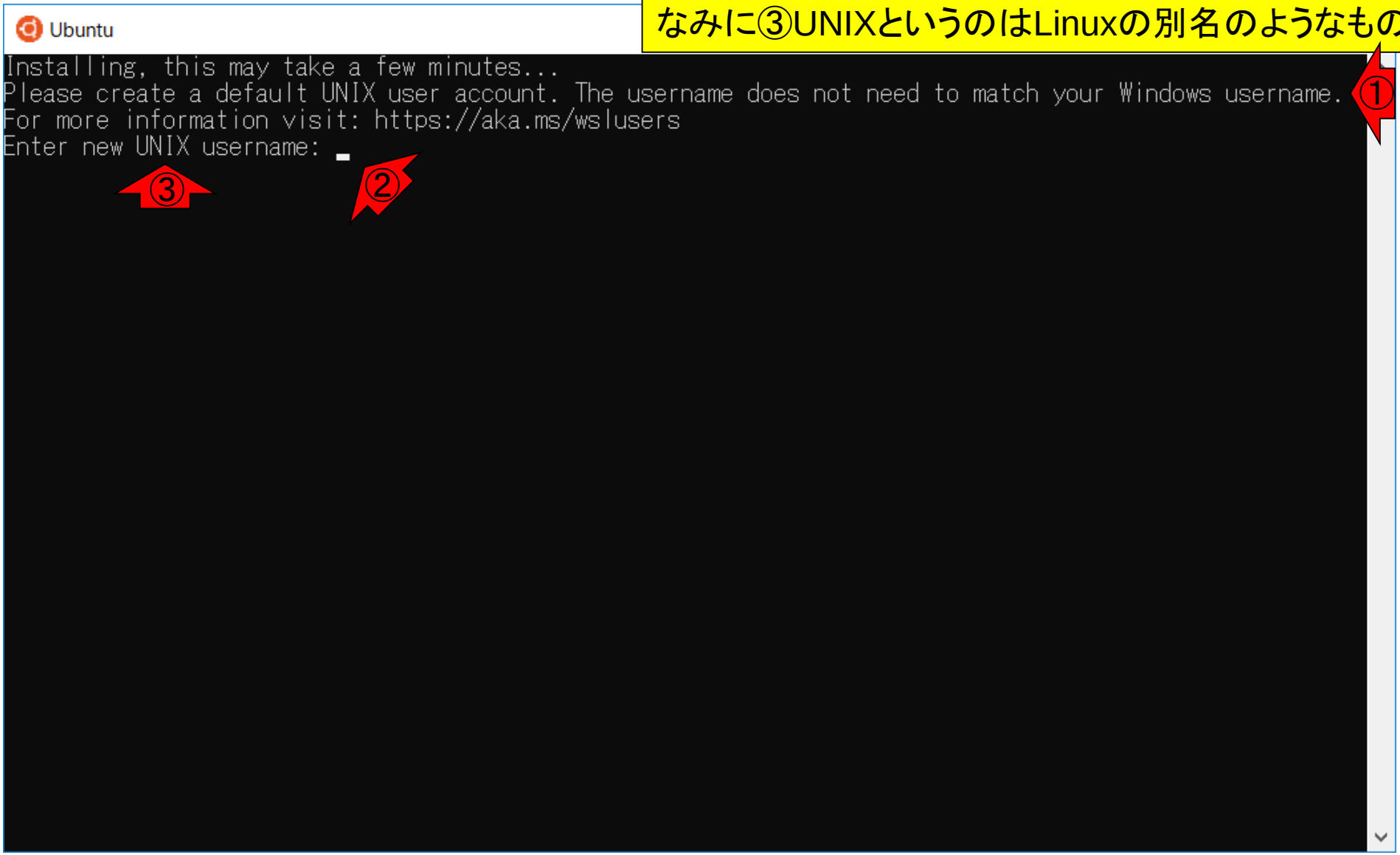
Ubuntu起動



参考

ユーザ名の作成

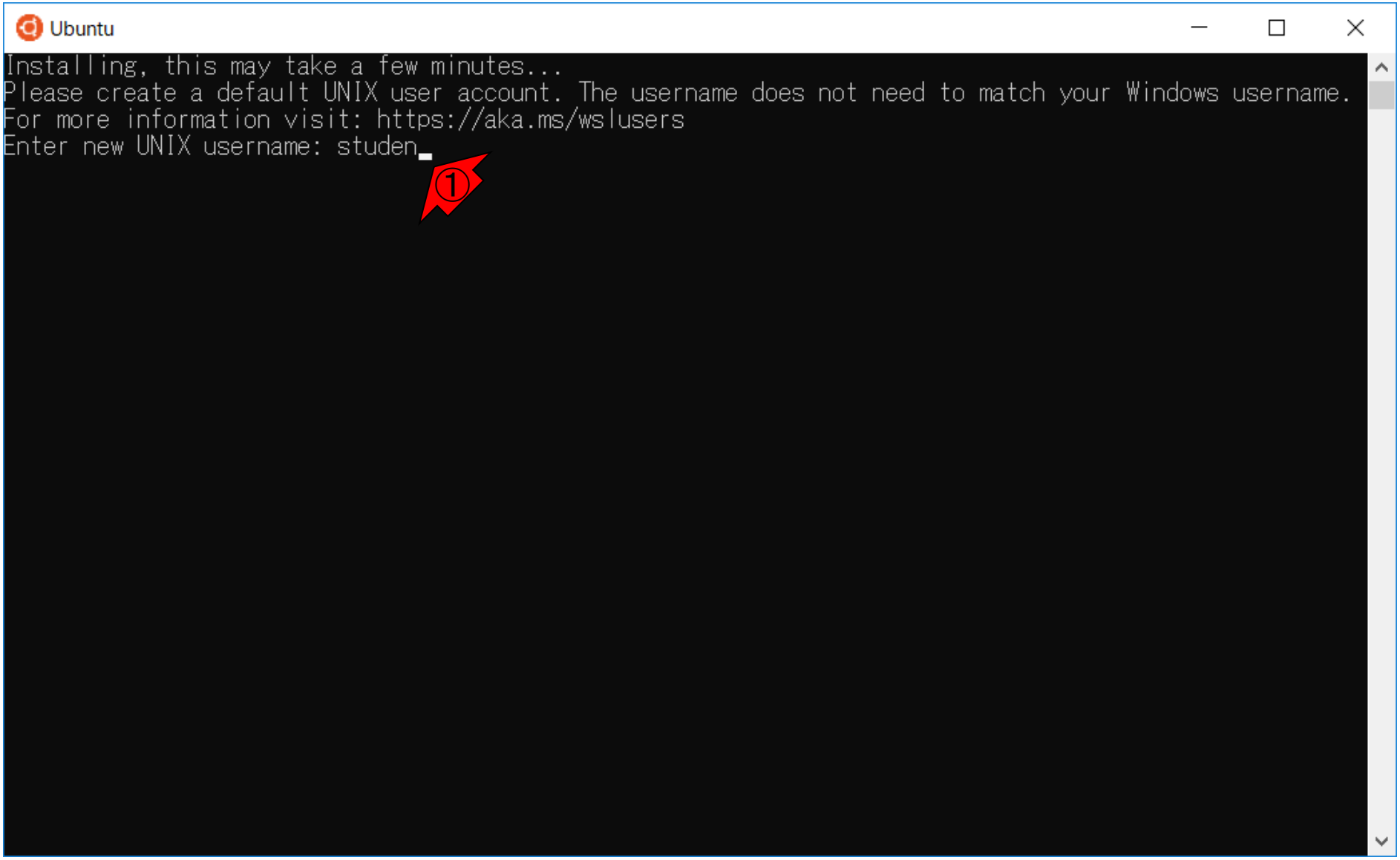
数分後にこんな感じになります。①今ログイン中のWindowsのユーザ名と同じである必要はないので、②新しいユーザ名を入力してね、とリクエストされています。ちなみに③UNIXというのはLinuxの別名のようなものです。



参考

ここでは、①studenとして、リターンキーを押します。

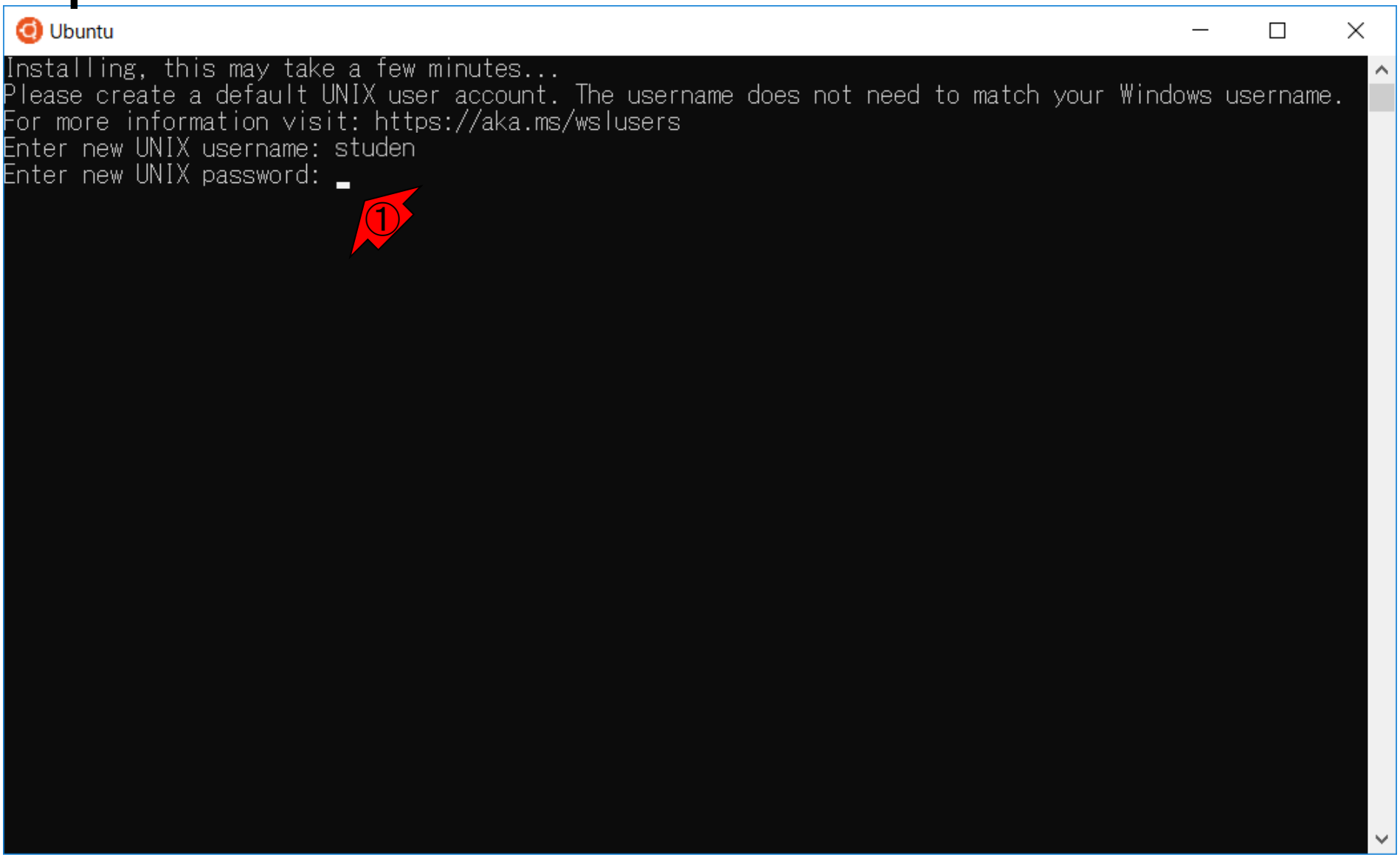
ユーザ名の作成



参考

p/wの作成

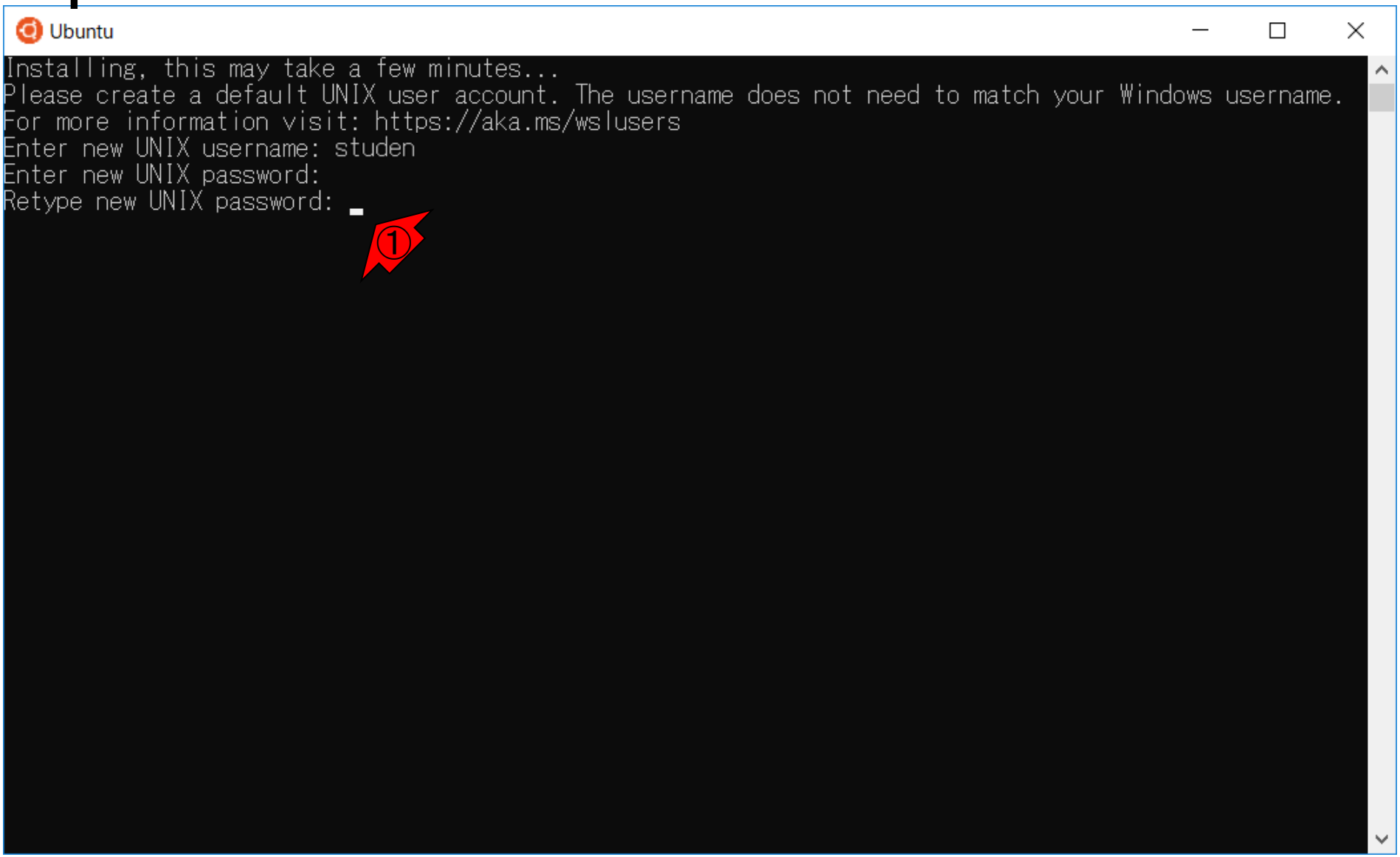
①今度はパスワードの作成を要求されますので、言われるがままにテキトー（ただしそれを忘れず）に作成します。



参考

再確認されますので、①打ち込みます。

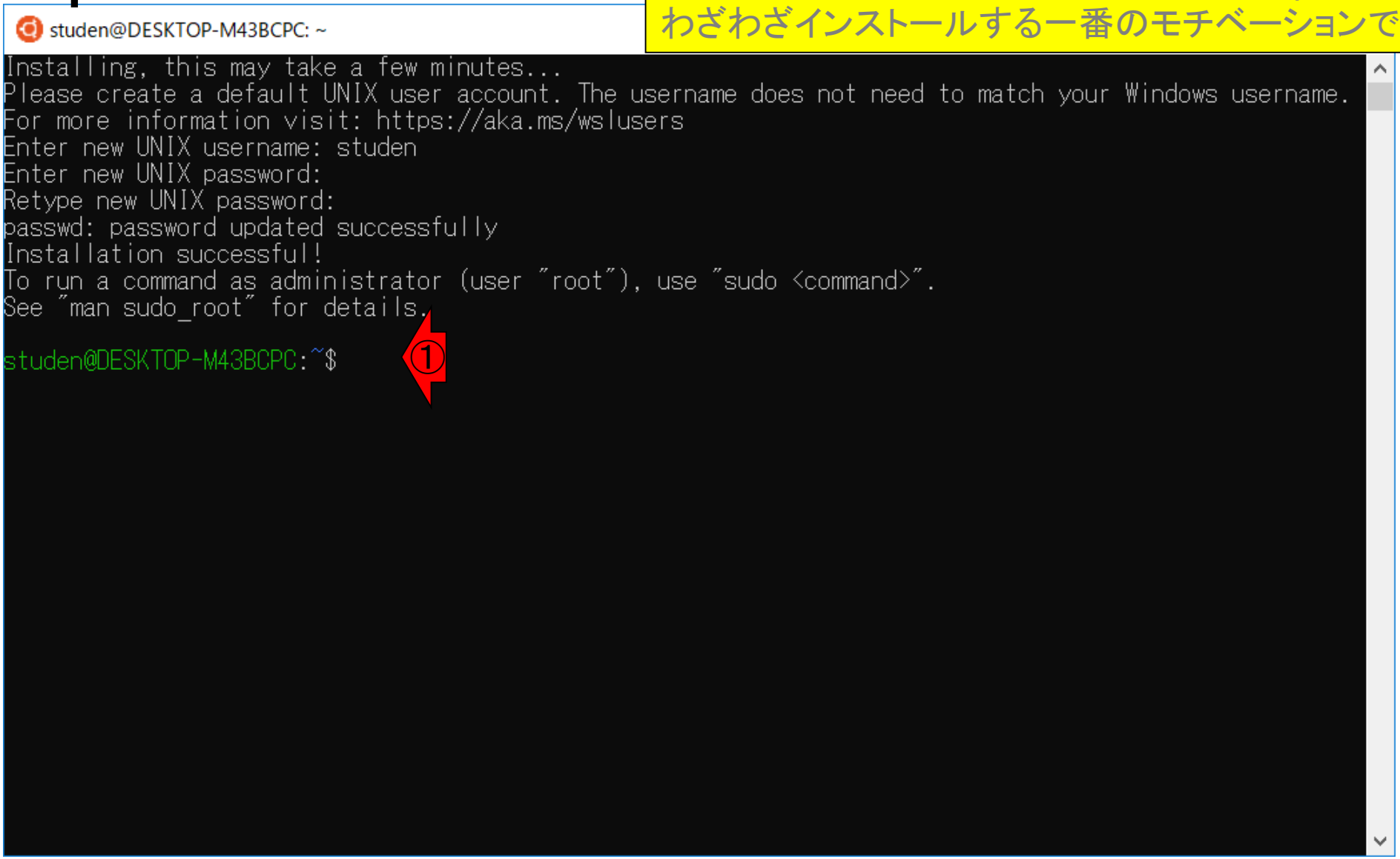
p/wの再確認



参考

p/wの再確認

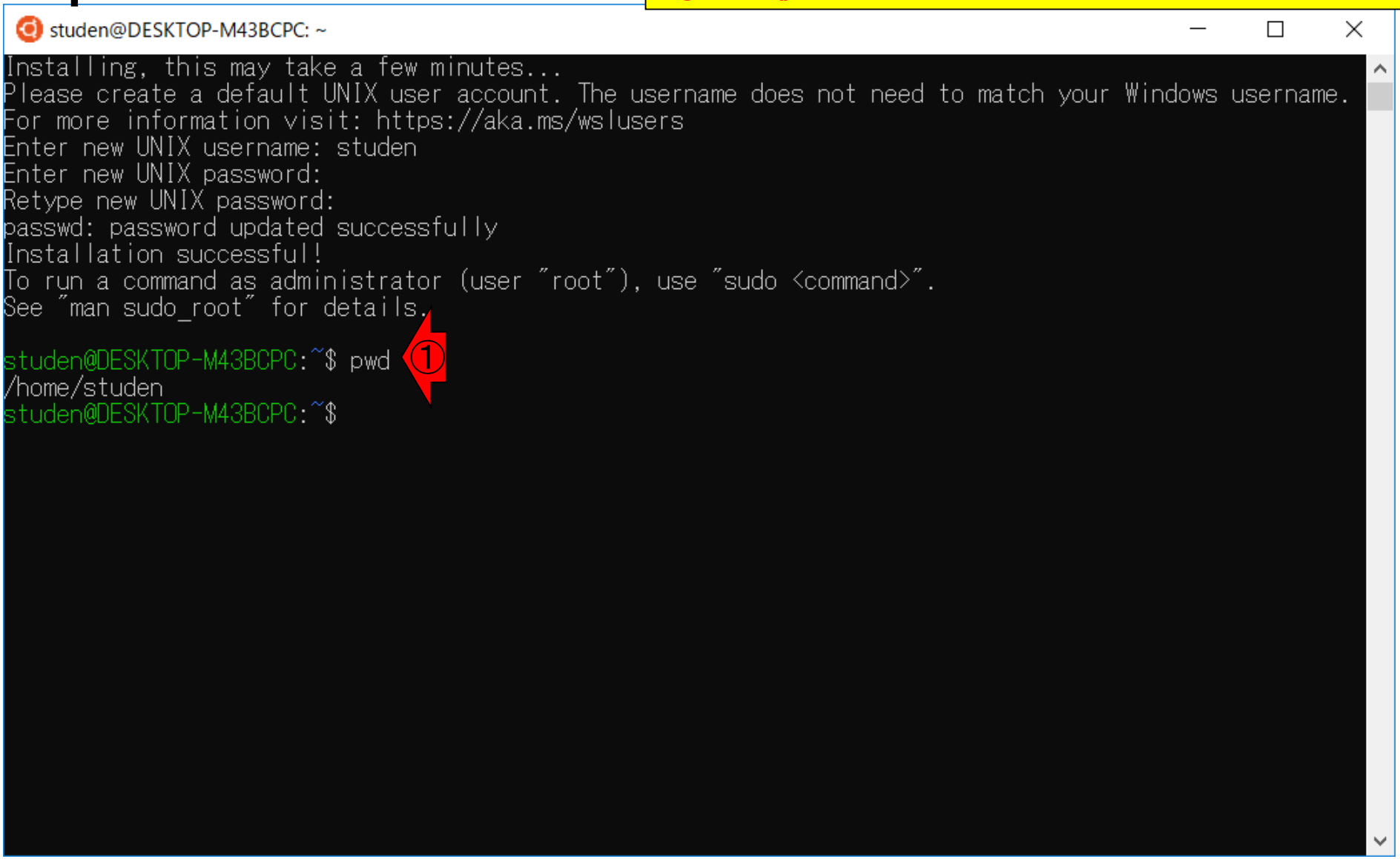
こんな感じになります。①が、Linuxのコマンド入力待ち状態です。「コマンドプロンプト」と同じようなものですが、Linuxのコマンドを利用できる点がUbuntu (つまりWSL)をわざわざインストールする一番のモチベーションです。



参考

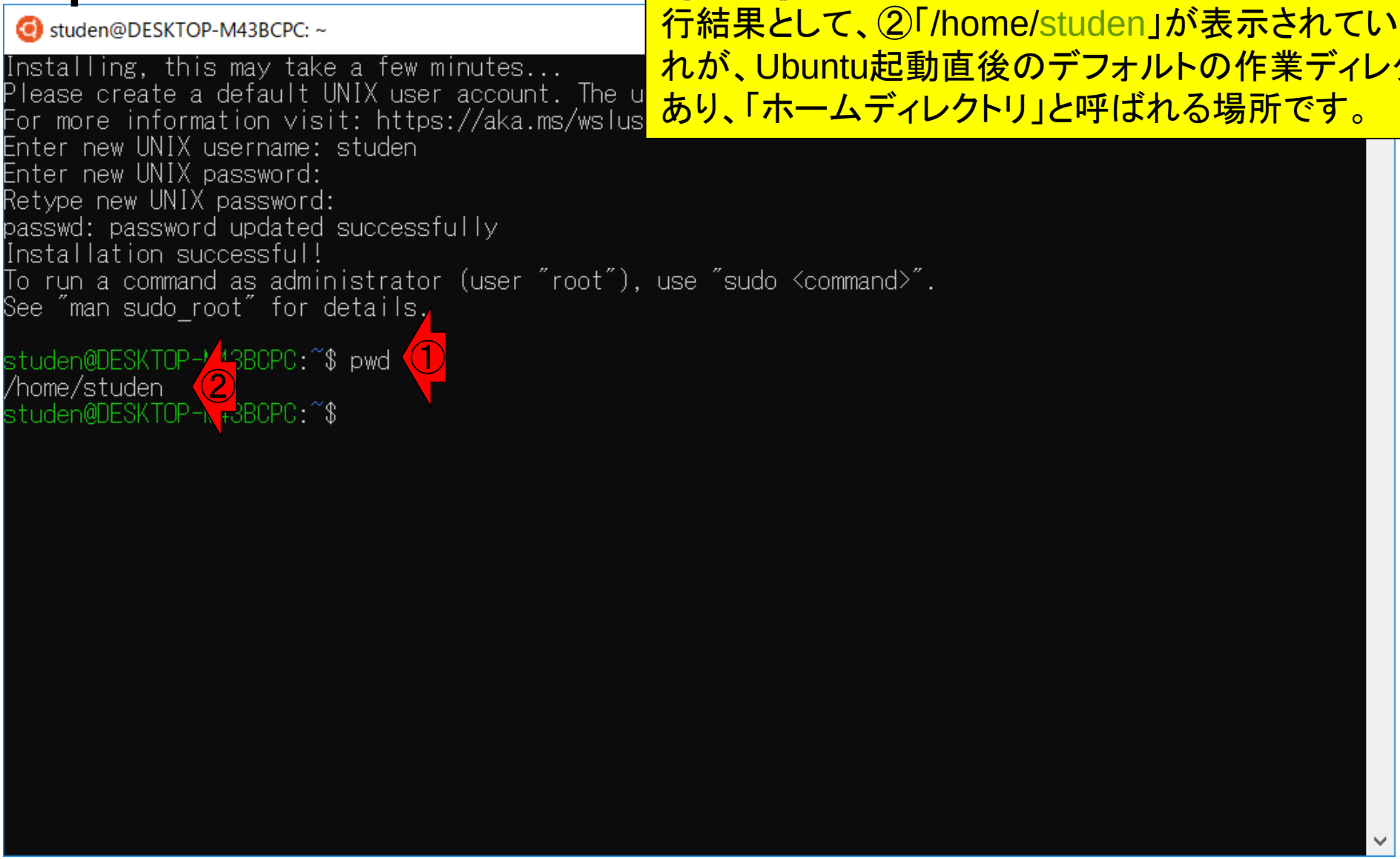
pwd

①pwdと打ってリターン。これは「print working directory（作業ディレクトリの表示）」を行うLinuxコマンドです。R上で「getwd()」を打ち込んでいるのと同じだと思えばよい。



参考

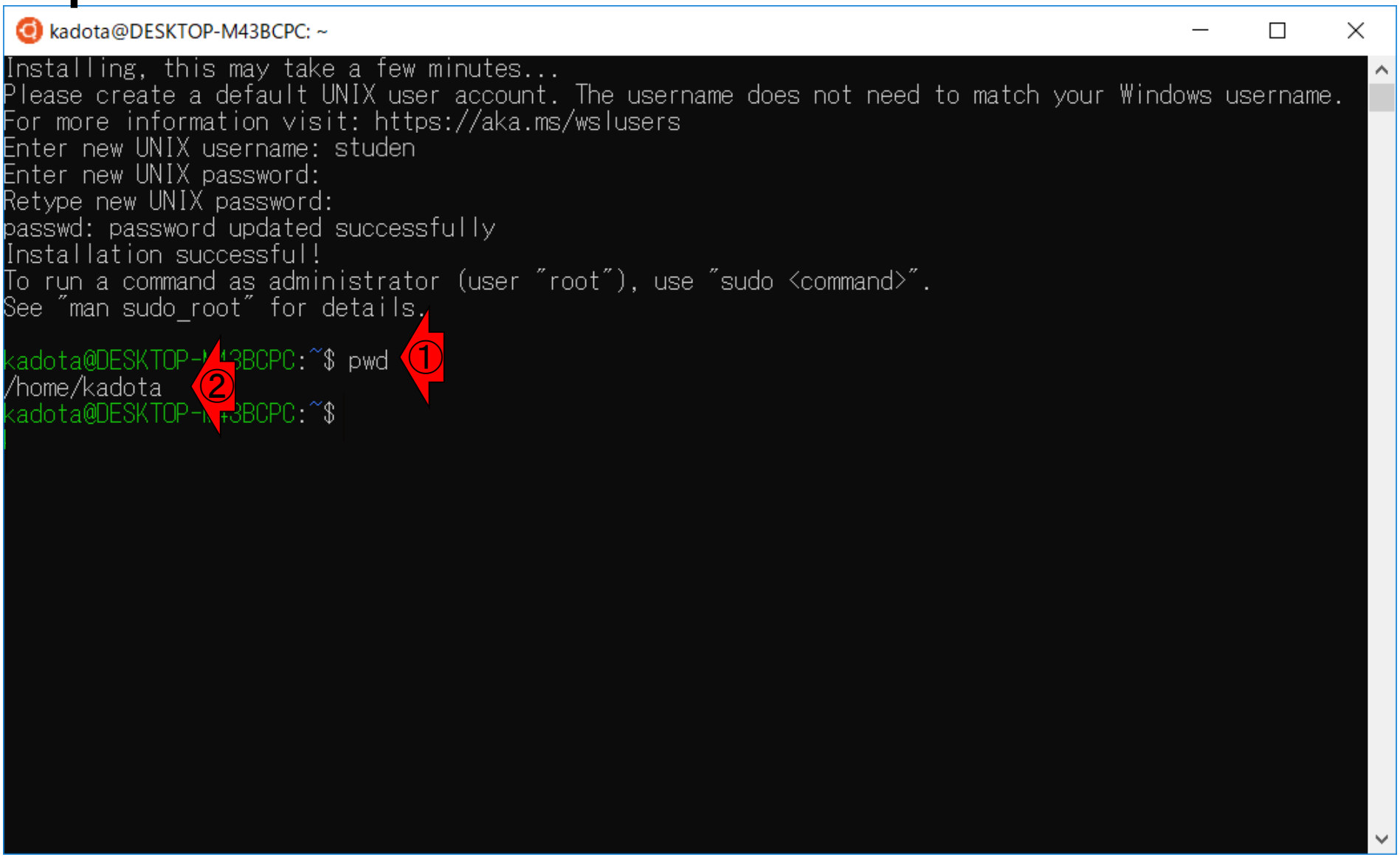
pwd



①pwdと打ってリターン。これは「print working directory（作業ディレクトリの表示）」を行うLinuxコマンドです。R上で「getwd()」を打ち込んでいるのと同じだと思えばよい。実行結果として、②「/home/studen」が表示されています。これが、Ubuntu起動直後のデフォルトの作業ディレクトリであり、「ホームディレクトリ」と呼ばれる場所です。

pwd

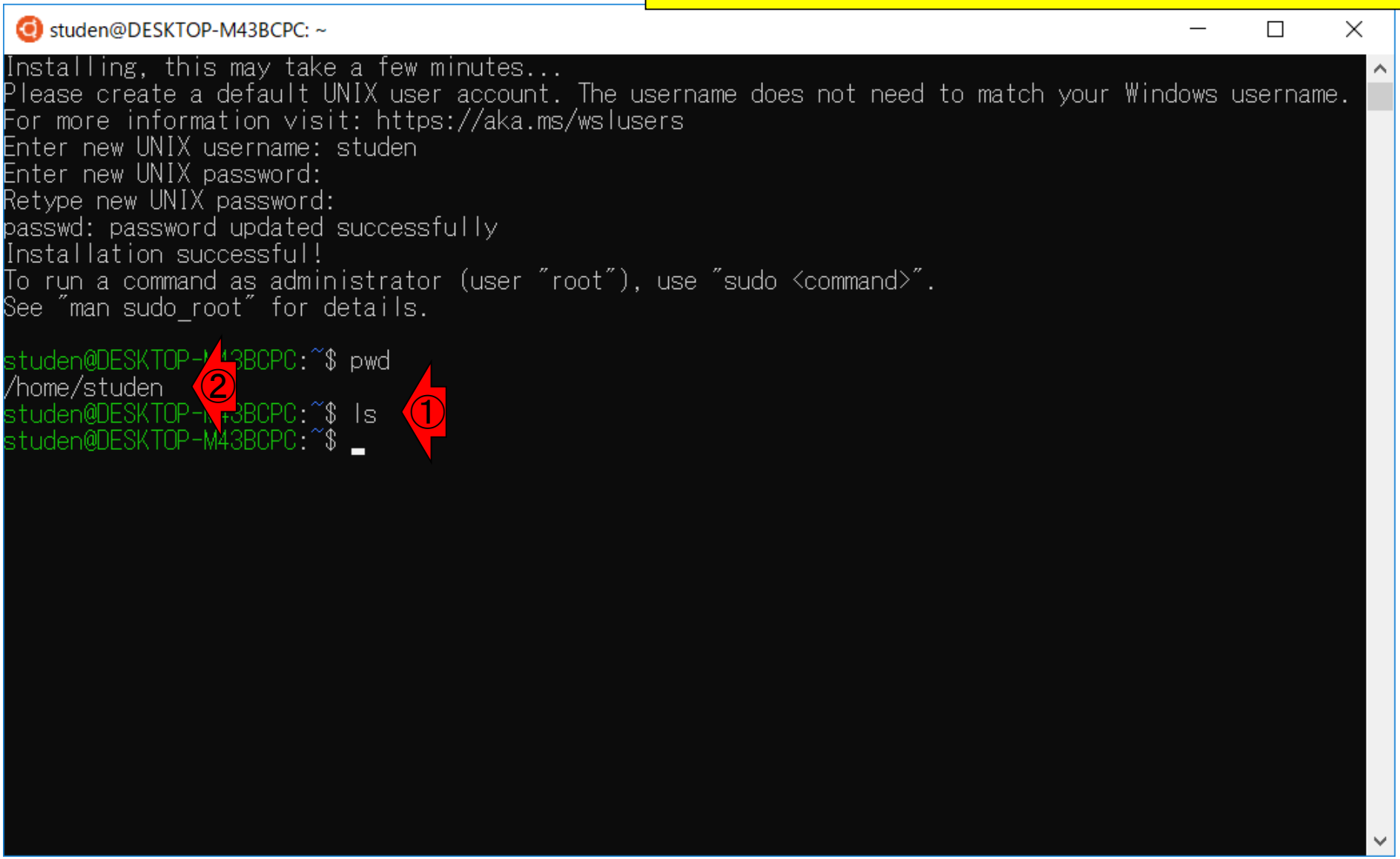
ちなみに、ユーザ名をkadotaとすると、②の部分が「/home/kadota」のようになります。



参考

ls

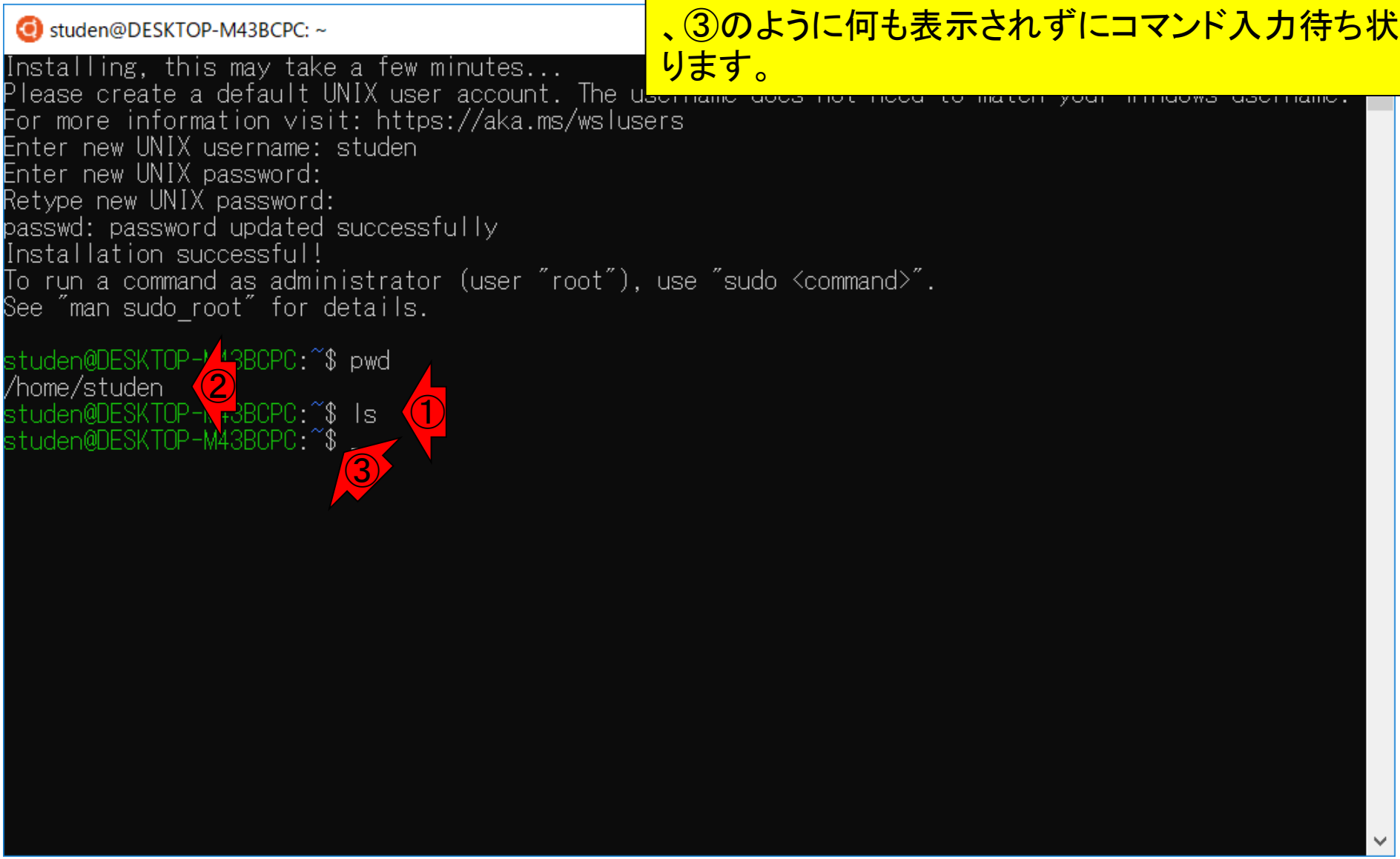
①ls（えるえす）と打ってリターン。これはR上で「list.files()」を実行するのと同じです。つまり、②現在の作業ディレクトリ上にあるものを表示しているのです。



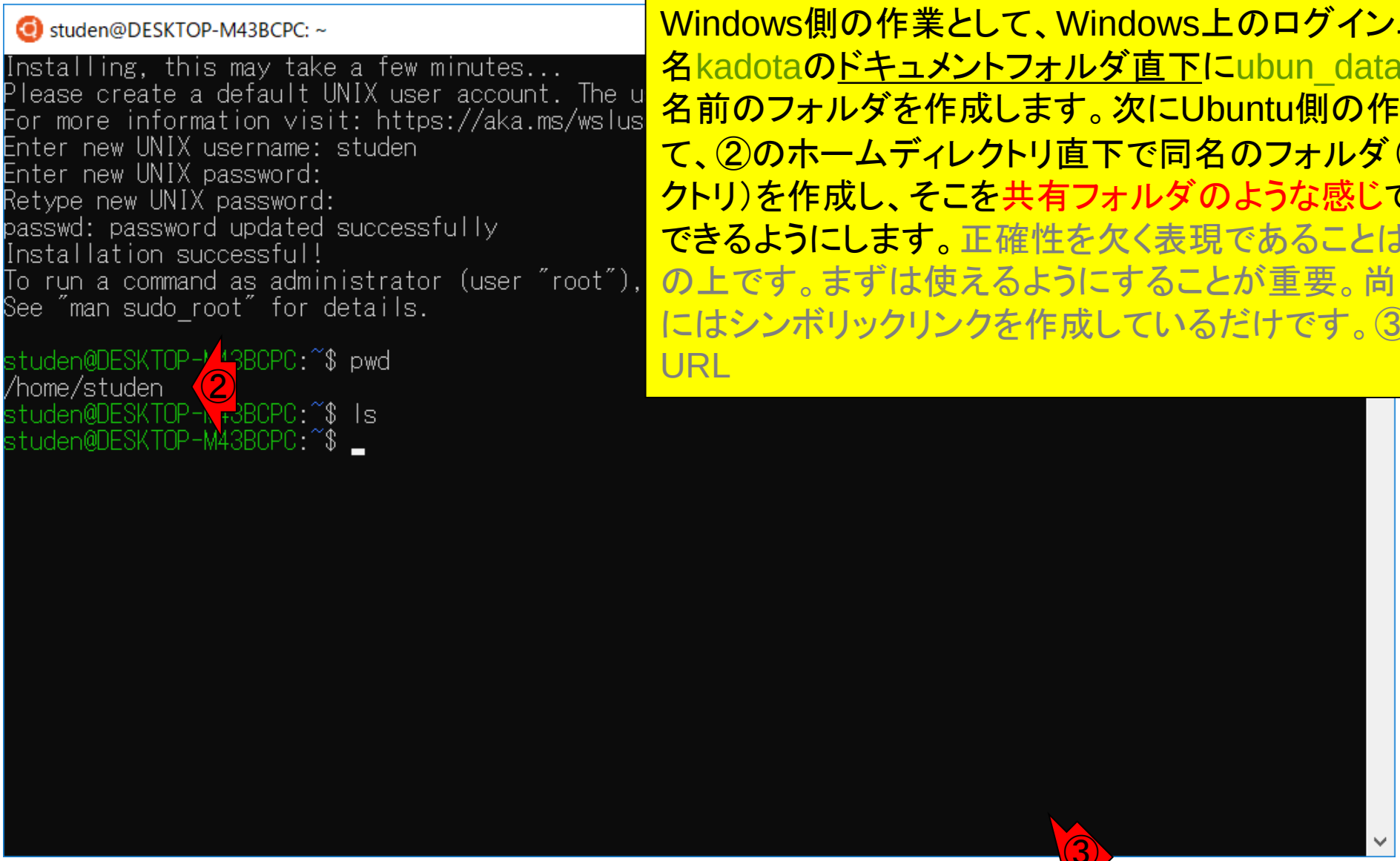
参考

ls

①ls（えるえす）と打ってリターン。これはR上で「list.files()」を実行するのと同じです。つまり、②現在の作業ディレクトリ上にあるものを表示しているのです。何もない場合は、③のように何も表示されずにコマンド入力待ち状態になります。



/home/studen?



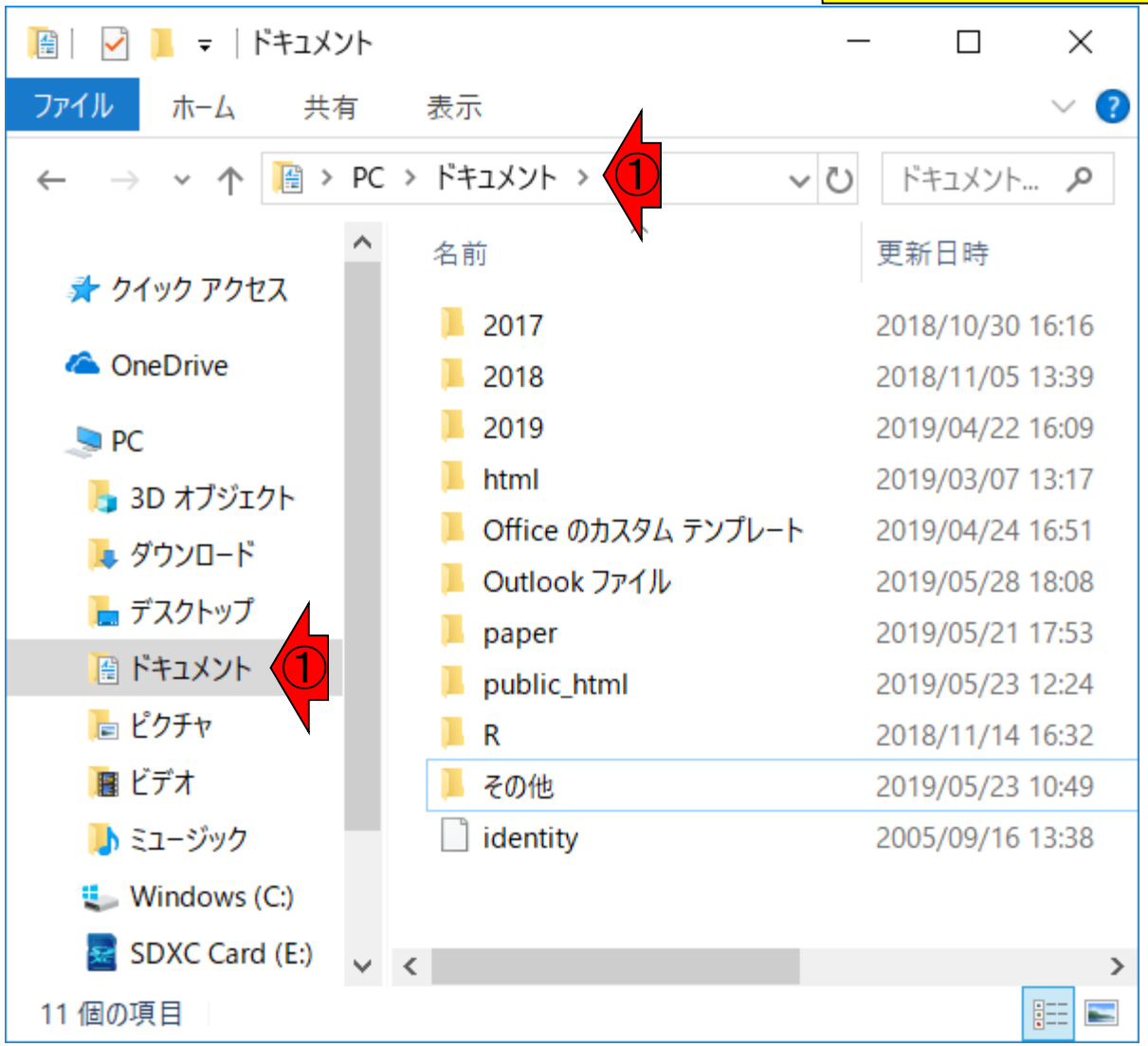
②この場所は一体どこにあるんだ？とか、この場所を分かりやすくアクセスしやすい場所に変更したい、が最初の希望です。そのやり方までを示します。ここではまずWindows側の作業として、Windows上のログインユーザ名kadotaのドキュメントフォルダ直下にubun_dataという名前のフォルダを作成します。次にUbuntu側の作業として、②のホームディレクトリ直下で同名のフォルダ（ディレクトリ）を作成し、そこを共有フォルダのような感じで利用できるようにします。正確性を欠く表現であることは承知の上です。まずは使えるようにすることが重要。尚、正確にはシンボリックリンクを作成しているだけです。③参考URL

https://fuchiaz.com/wsl-symbolic-link/

参考

Win側の作業

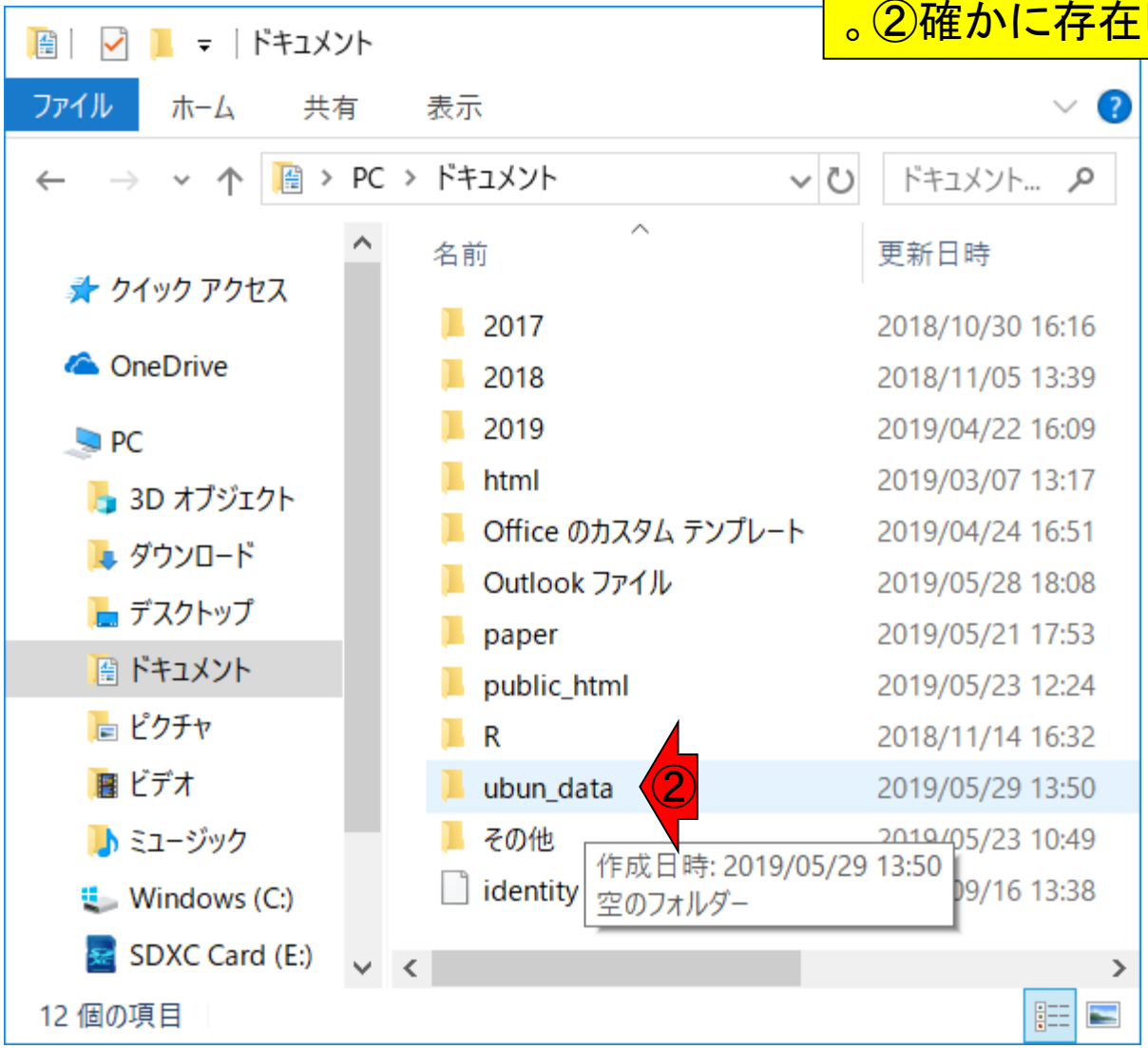
最初のWindows側の作業を行っていきます。Windows上のログインユーザ名kadotaの①ドキュメントフォルダ直下にubun_dataという名前のフォルダを作成します。作成前



参考

Win側の作業

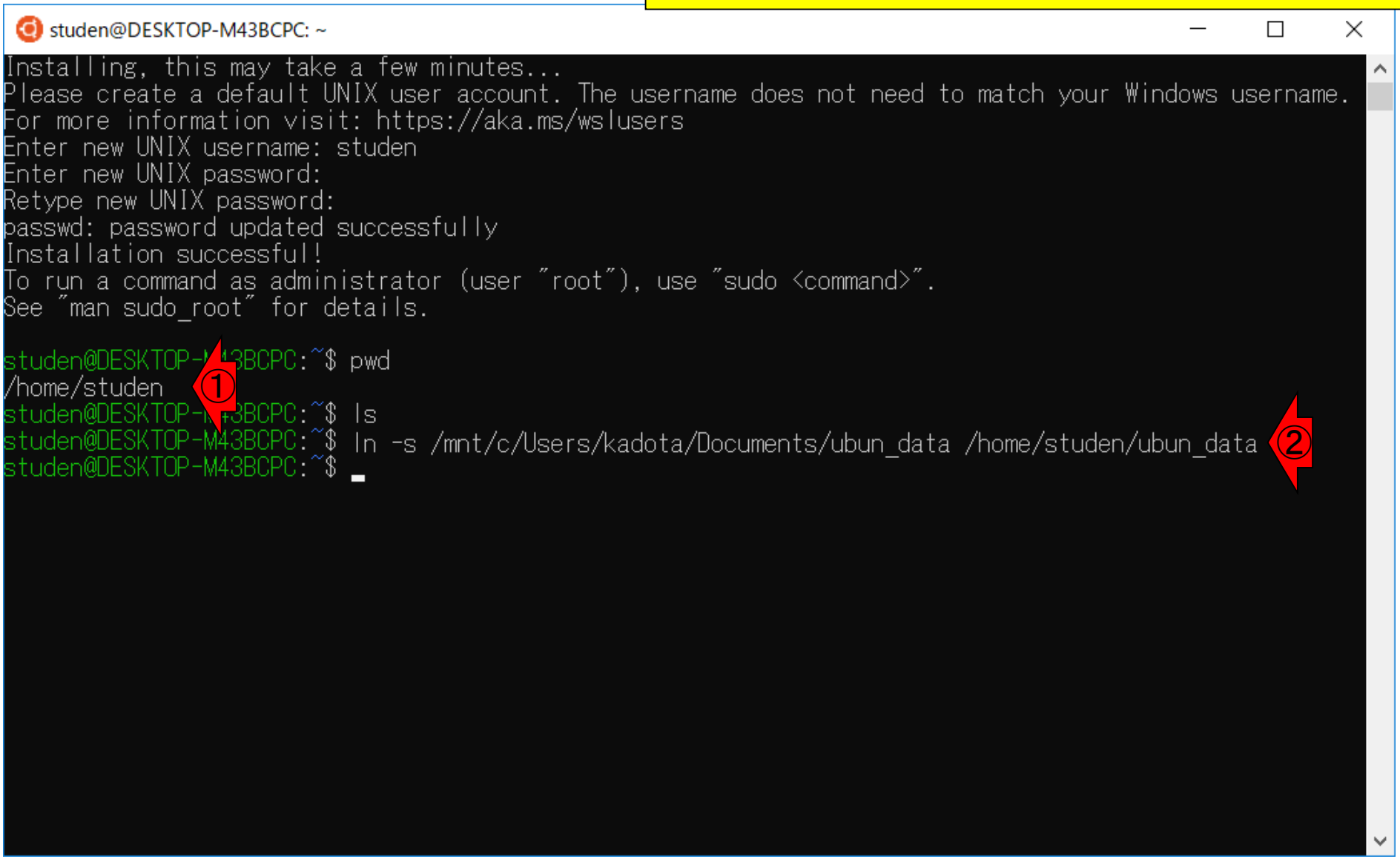
最初のWindows側の作業を行っていきます。Windows上のログインユーザ名kadotaの①ドキュメントフォルダ直下にubun_dataという名前のフォルダを作成します。作成後。②確かに存在します。作成直後なので中身は空です。



参考

Ubuntu側の作業

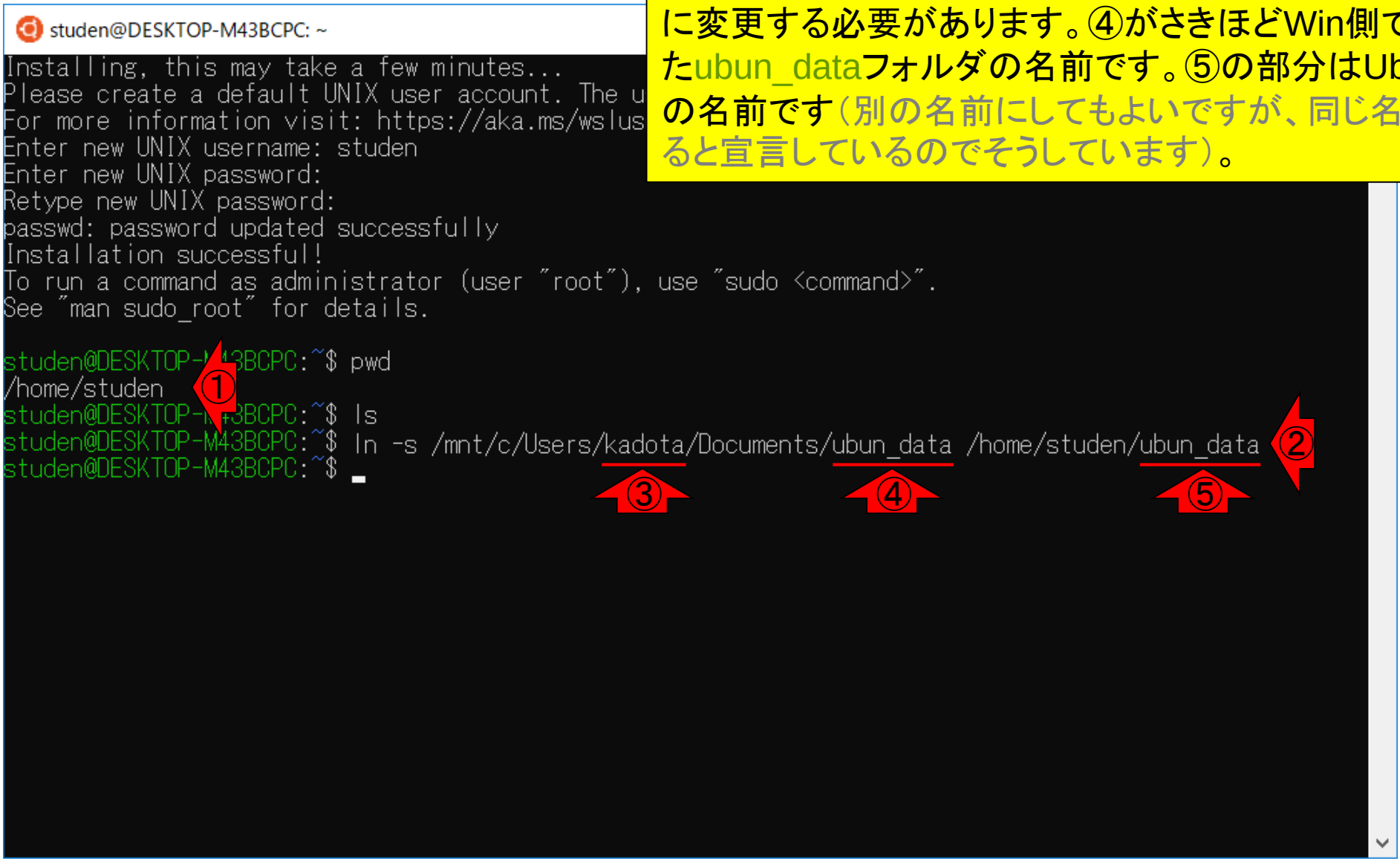
次にUbuntu側の作業として、②のような長いコマンドを打ち込みます。次のスライドにも説明があるので、実行前に必ず読んで必要な箇所を変更してください。



参考

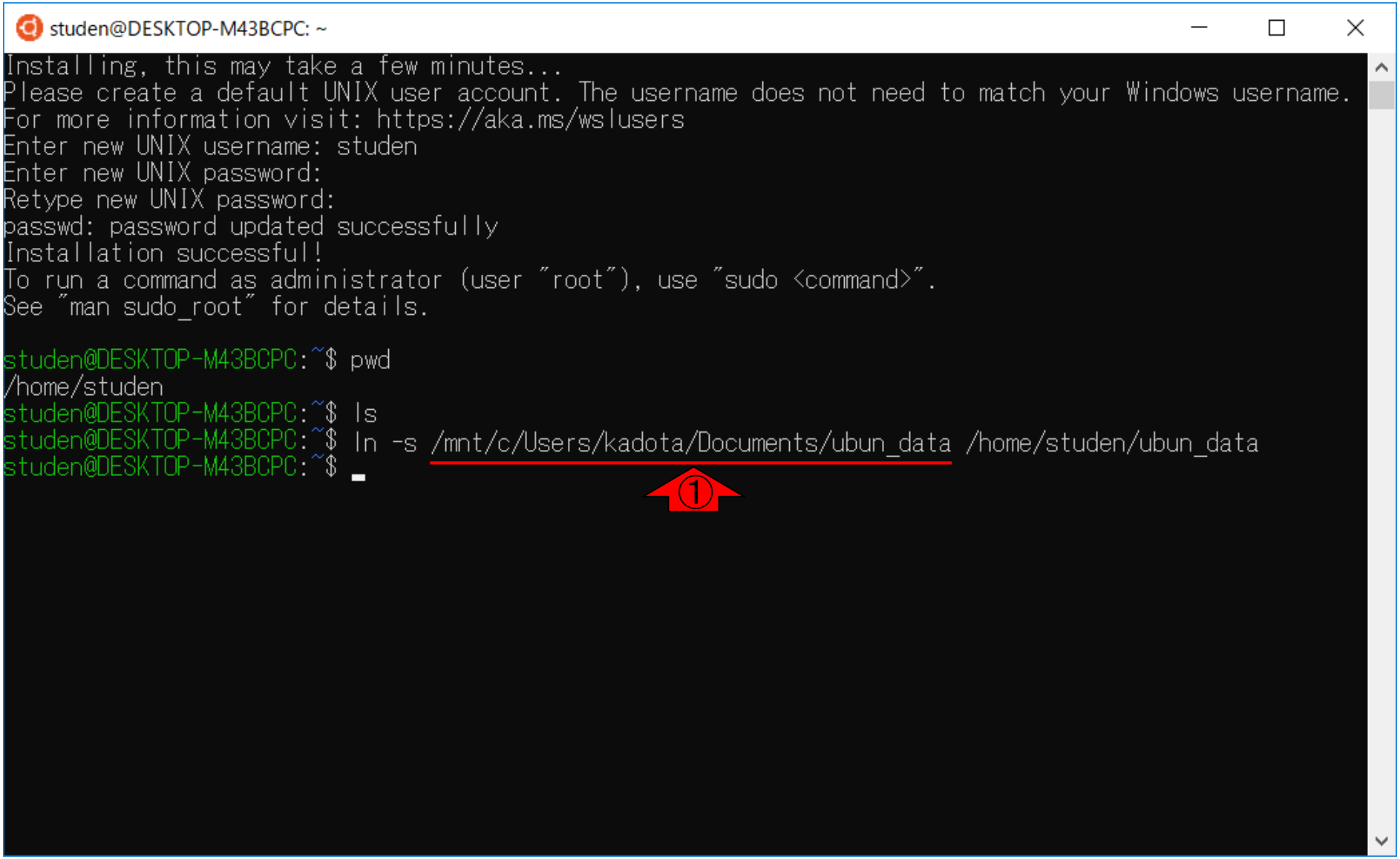
コマンドの解説

次にUbuntu側の作業として、②のような長いコマンドを打ち込みます。注意点としては、③の部分はWindowsのログインユーザ名（私の場合はkadota）なので、各自のものに変更する必要があります。④がさきほどWin側で作成したubun_dataフォルダの名前です。⑤の部分はUbuntu側の名前です（別の名前にしてもよいですが、同じ名前にすると宣言しているのでそうしています）。



参考　①の部分全体で…

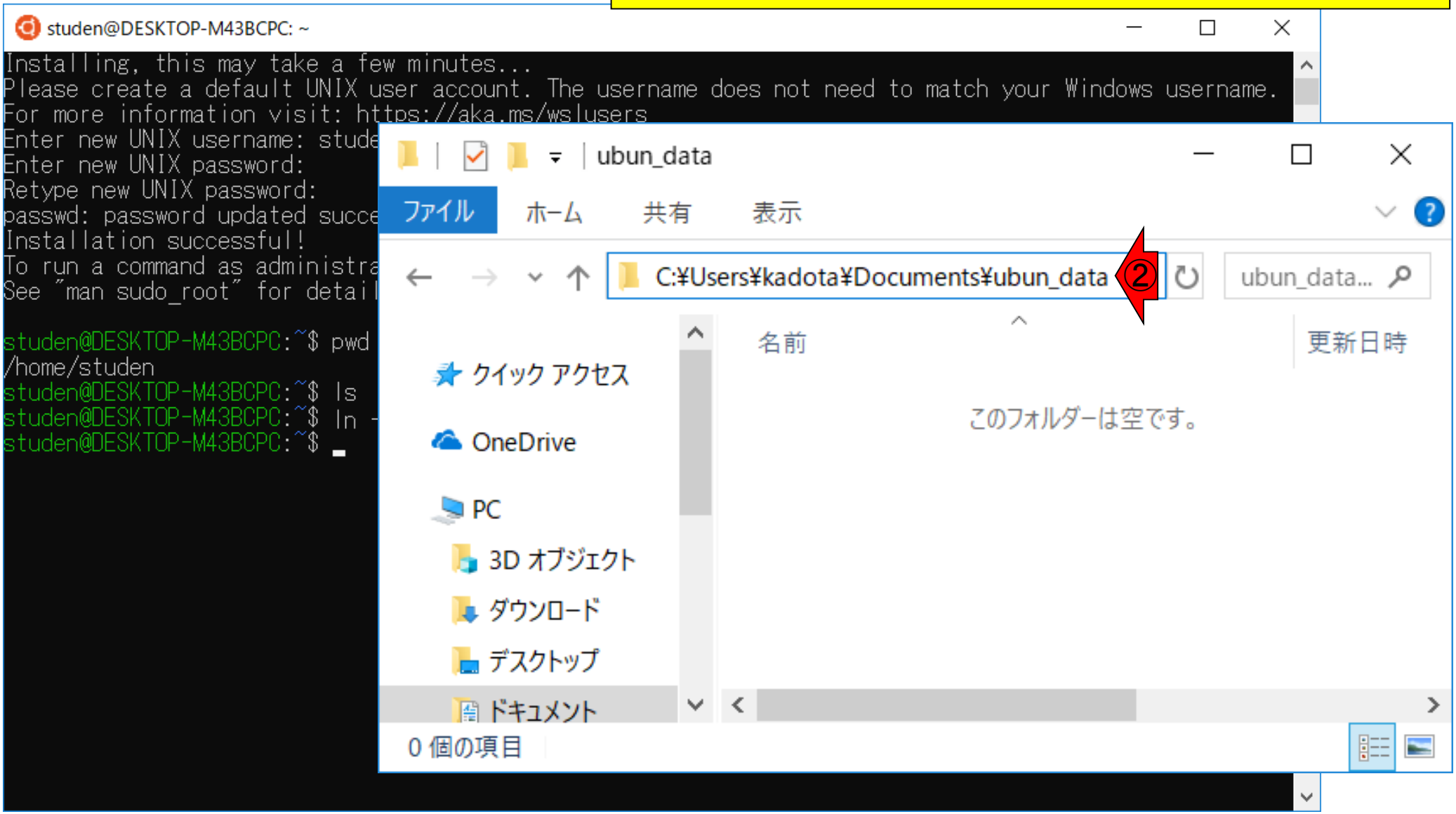
コマンドの解説



参考

コマンドの解説

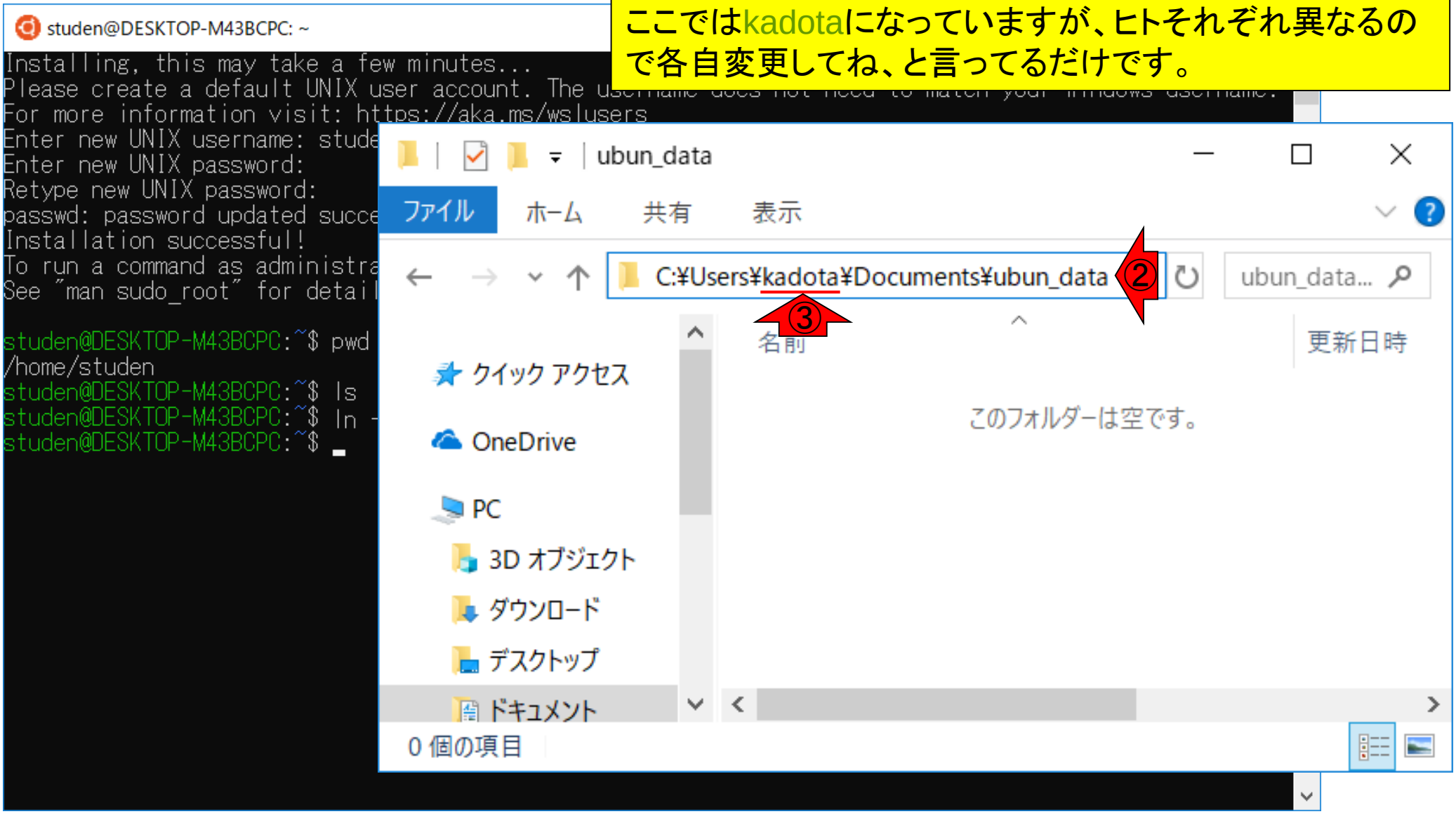
①の部分全体で、さきほどWin側で作成したubun_dataフォルダの②「完全パス（full pass）」情報を指し示していることに相当します。



参考

コマンドの解説

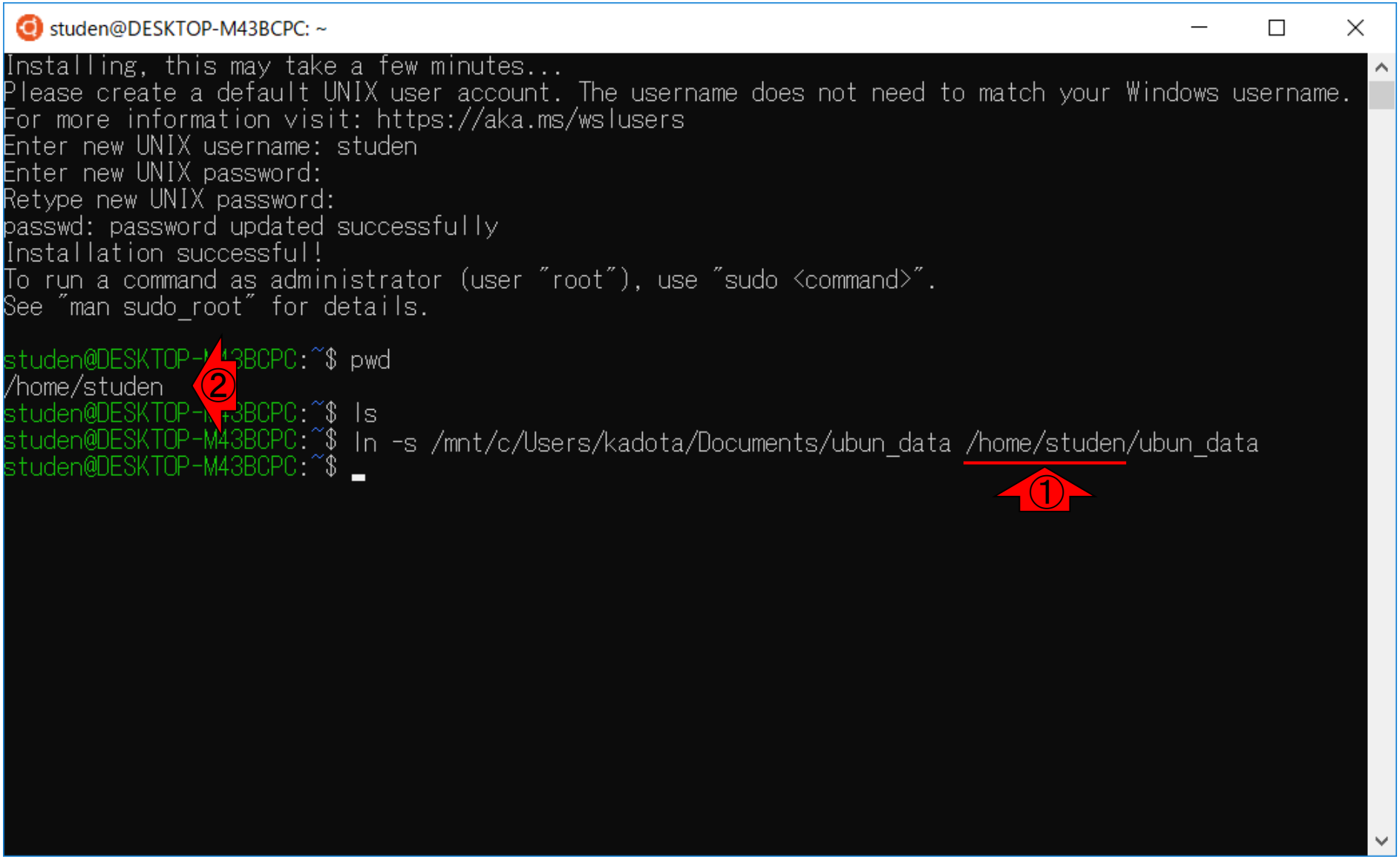
①の部分全体で、さきほどWin側で作成したubun_dataフォルダの②「完全パス（full pass）」情報を指し示していることに相当します。③のWindowsのログインユーザ名は、ここではkadotaになっていますが、ヒトそれぞれ異なるので各自変更してね、と言ってるだけです。



参考

ちなみに、①の部分は、②と同じです。

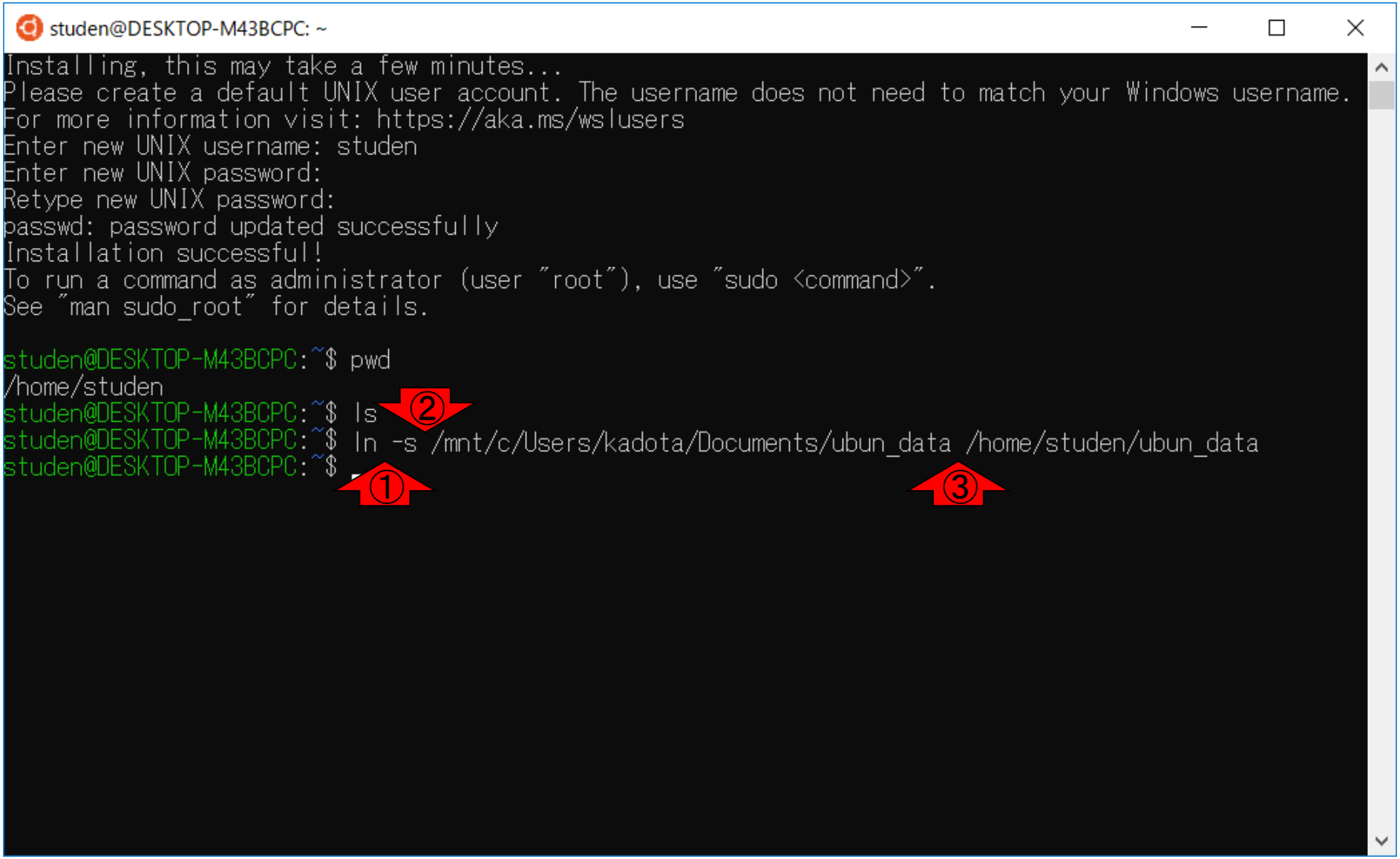
コマンドの解説



参考

ちなみに、①から③の計3カ所に「スペース」があります。

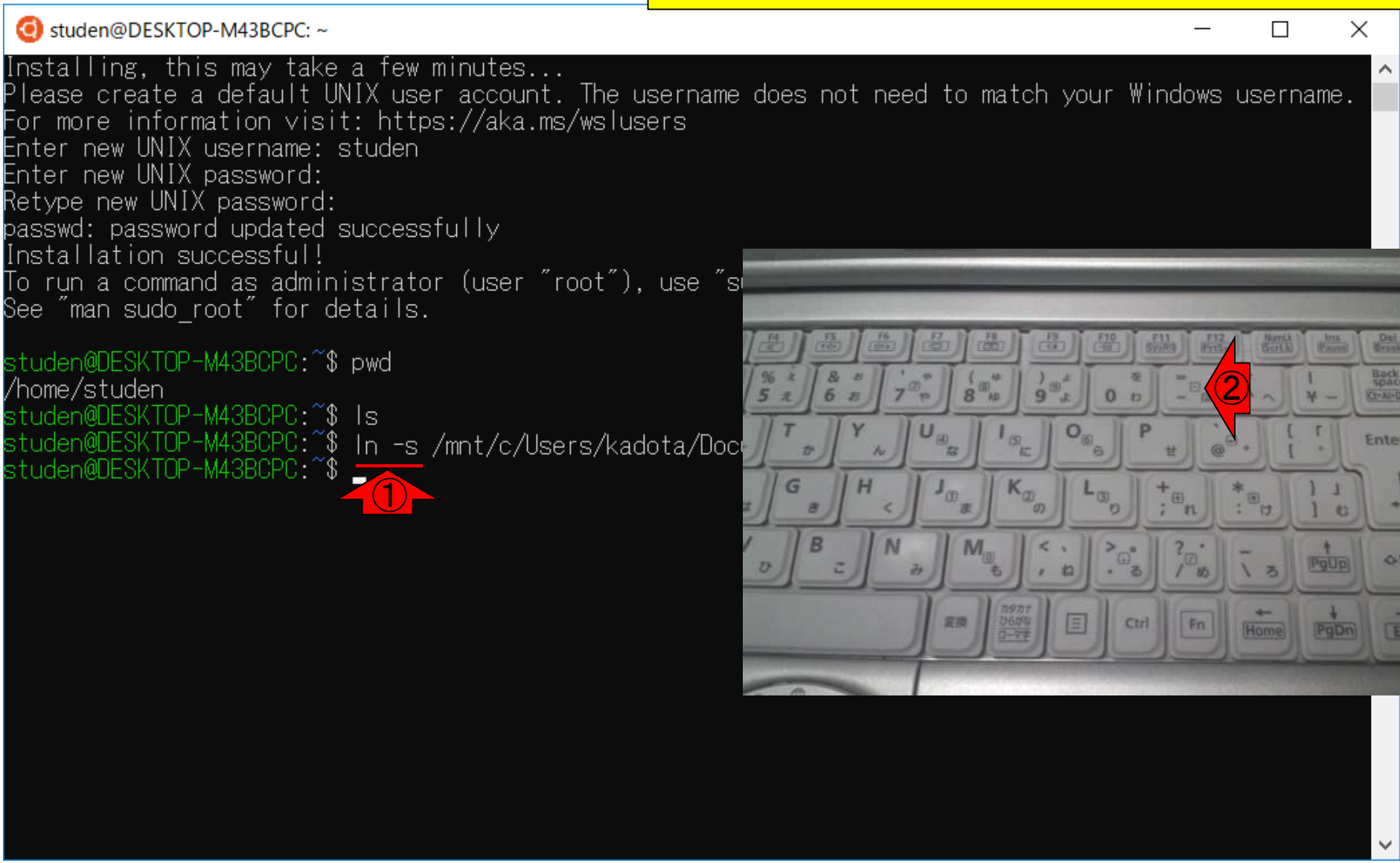
コマンドの解説



参考

コマンドの解説

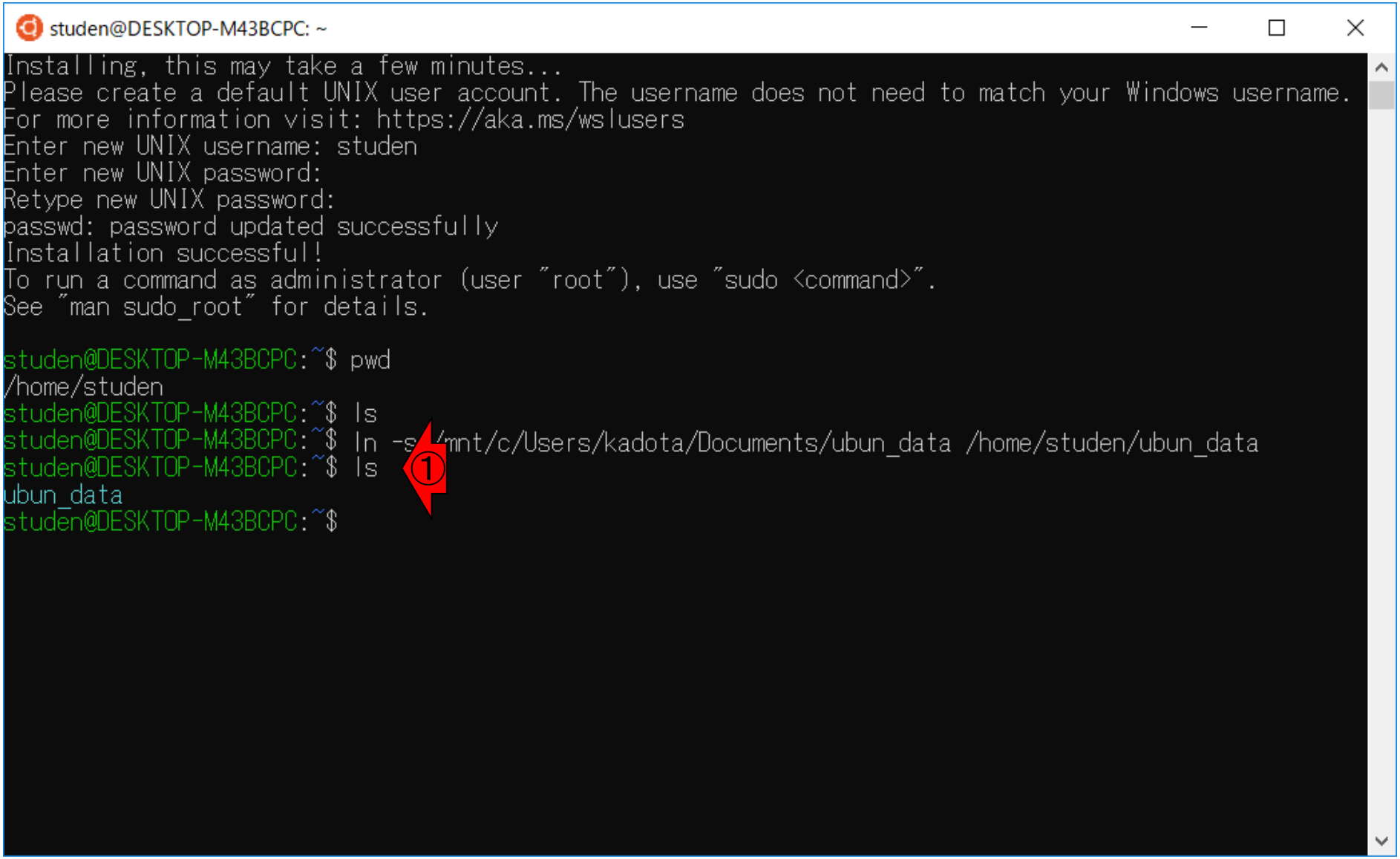
ちなみに、①の部分を読み上げると、「えるえぬ、スペース、ハイフン、えす」です。ハイフンは-のことで、②のキーで出します。



参考

①lsコマンドを再度実行すると、こんな感じになります。

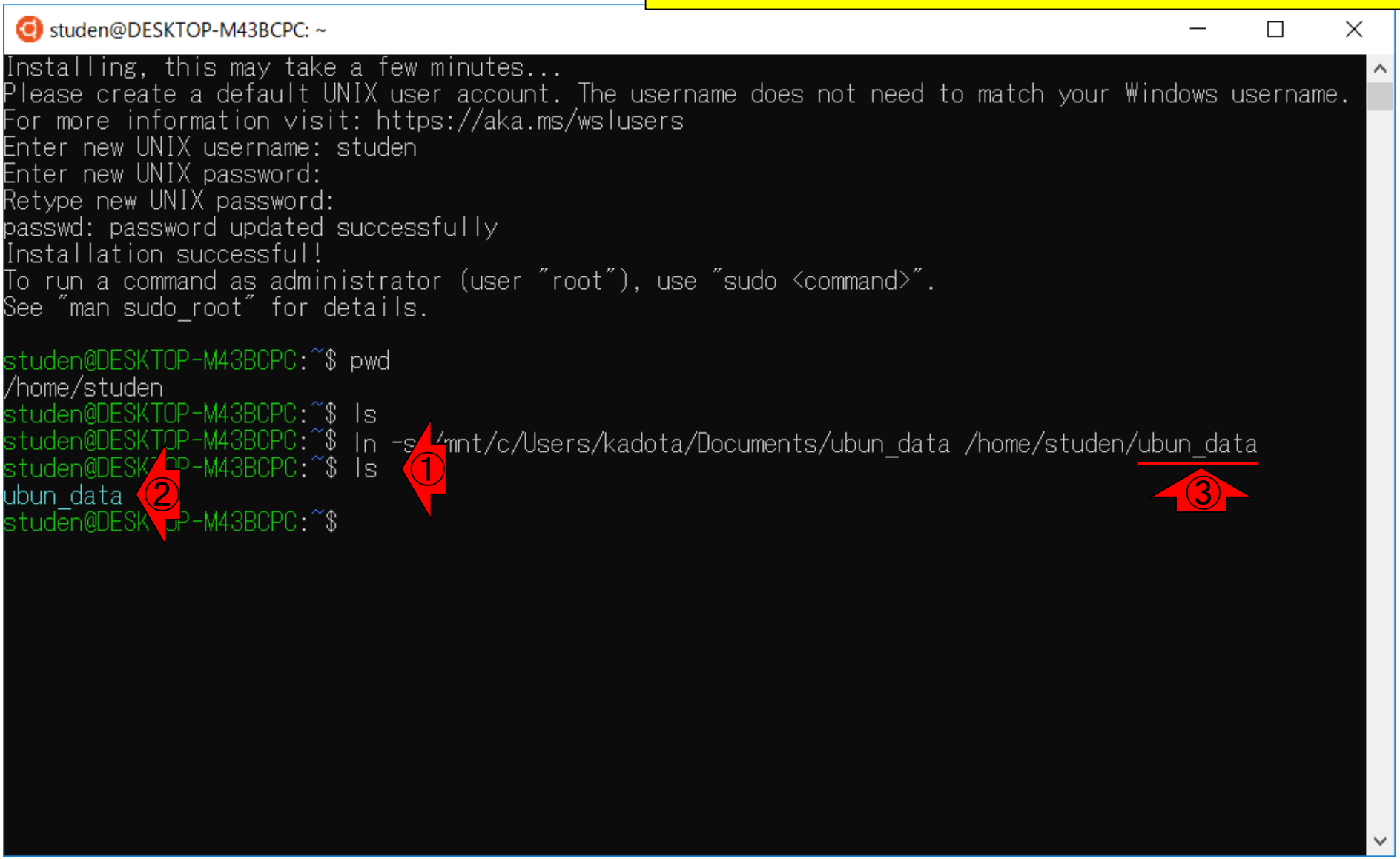
lsで確認



参考

lsで確認

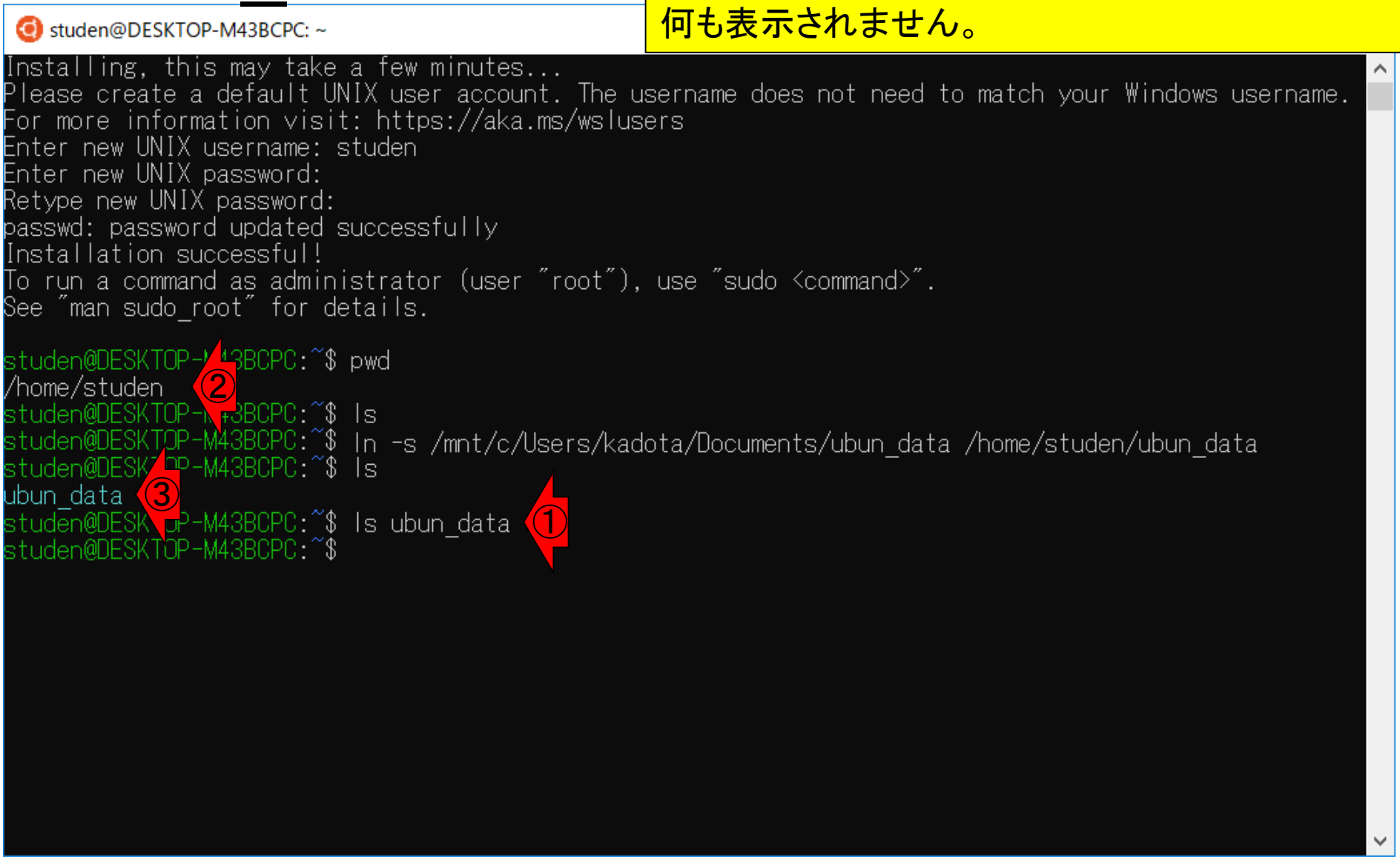
①lsコマンドを再度実行すると、こんな感じになります。②で見えているubun_dataが、③で指定したものに対応します。



参考

ubun_data確認

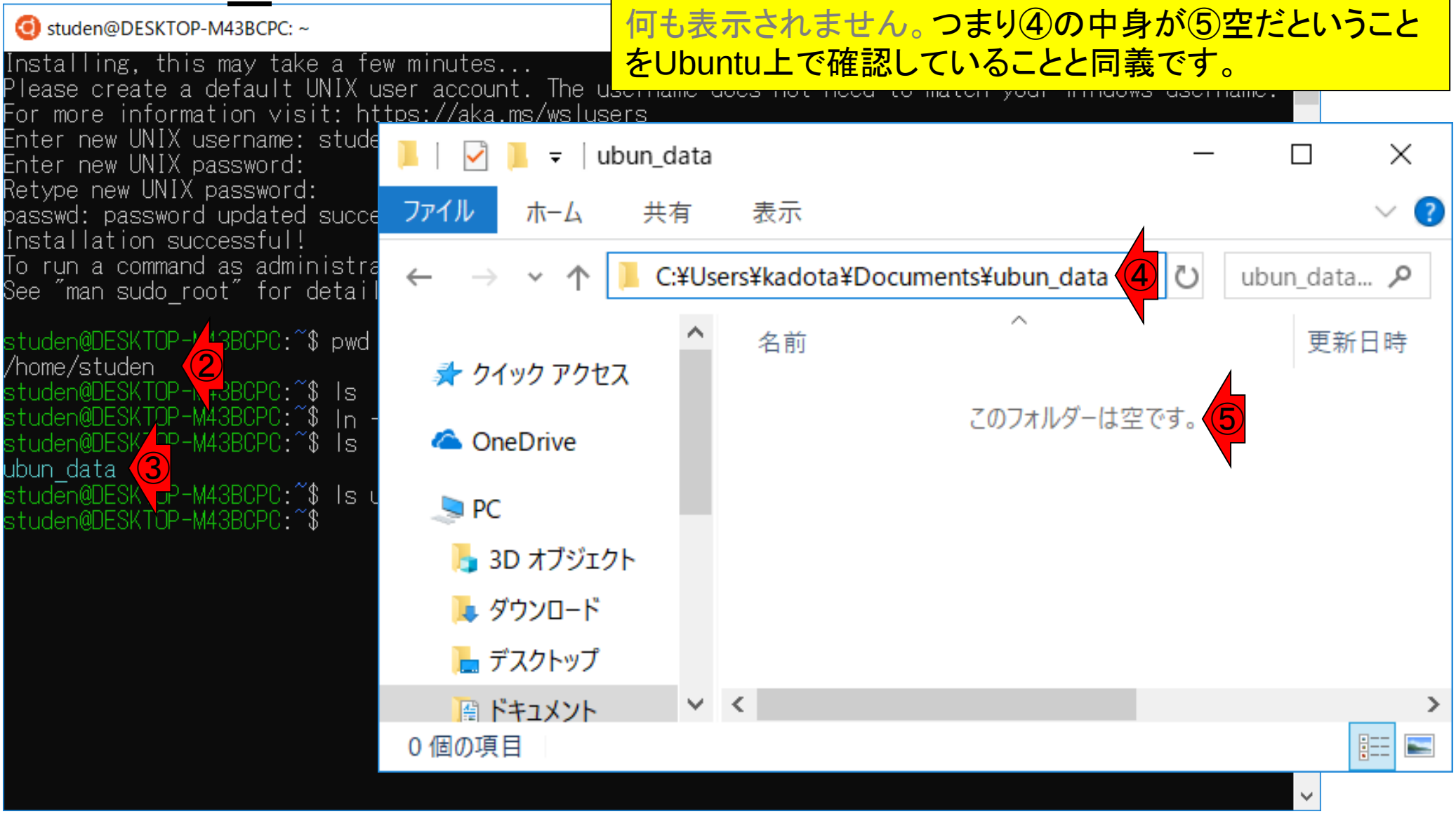
①「ls ubun_data」を実行。これは、②現在の作業ディレクトリ上で、③で見えているubun_dataディレクトリの中身を表示させていることに相当します。中身は何もないので、何も表示されません。



参考

ubun_data確認

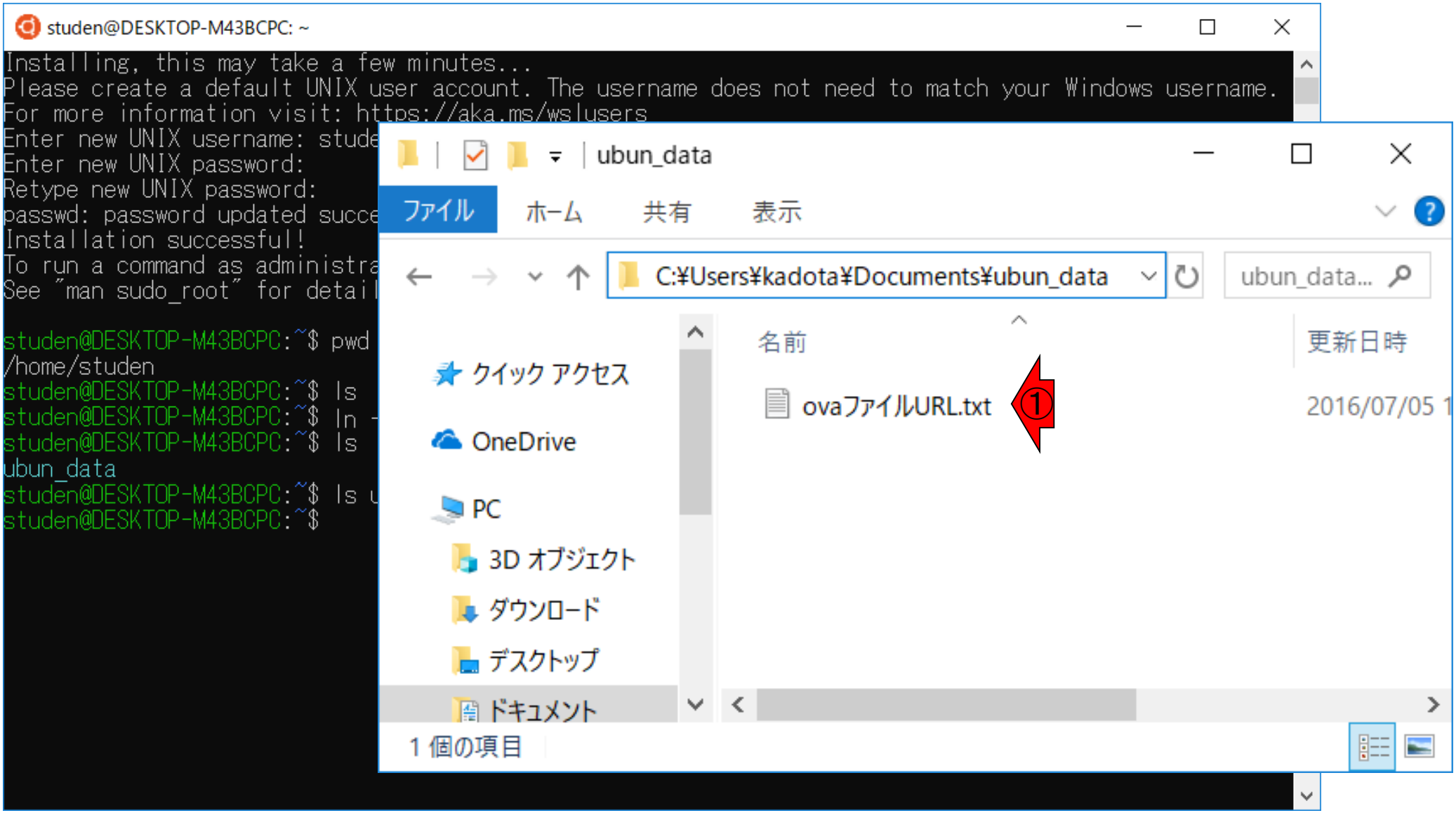
①「ls ubun_data」を実行。これは、②現在の作業ディレクトリ上で、③で見えているubun_dataディレクトリの中身を表示させていることに相当します。中身は何もないので、何も表示されません。つまり④の中身が⑤空だということをUbuntu上で確認していることと同義です。



参考

従って、①に何か手持ちのファイルを置いておき…

共有?!確認



参考

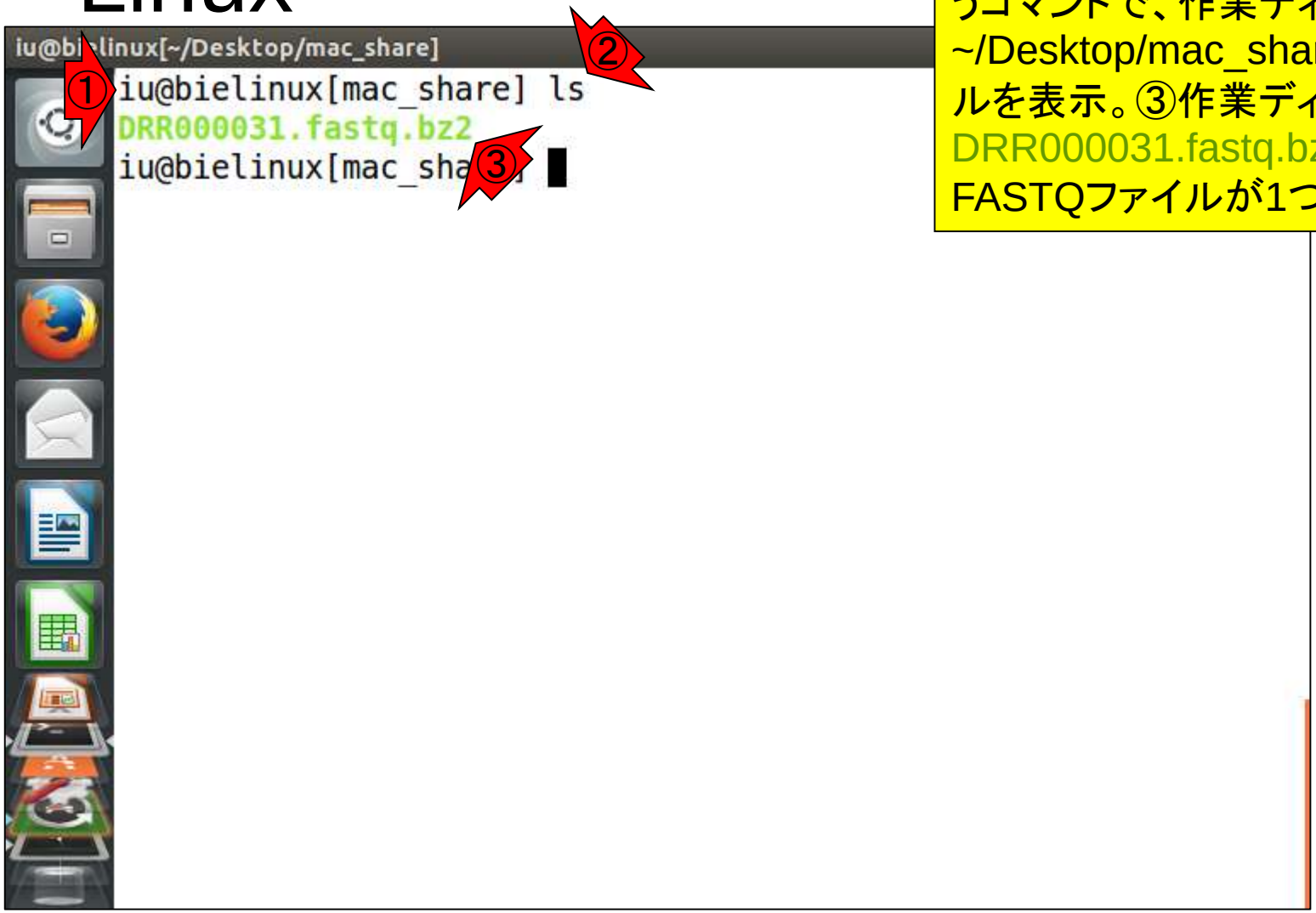
共有?!確認

従って、①に何か手持ちのファイルを置いておき、もう一度②「ls ubun_data」を実行。同じものが見えることでしょう。ちなみに、②のコマンドはわざわざ打ち込まなくても、「上矢印キー」を1回押せば、直前に打ったコマンドが表示されるので、それを利用すると楽です。このあたりはRと同じですね。

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

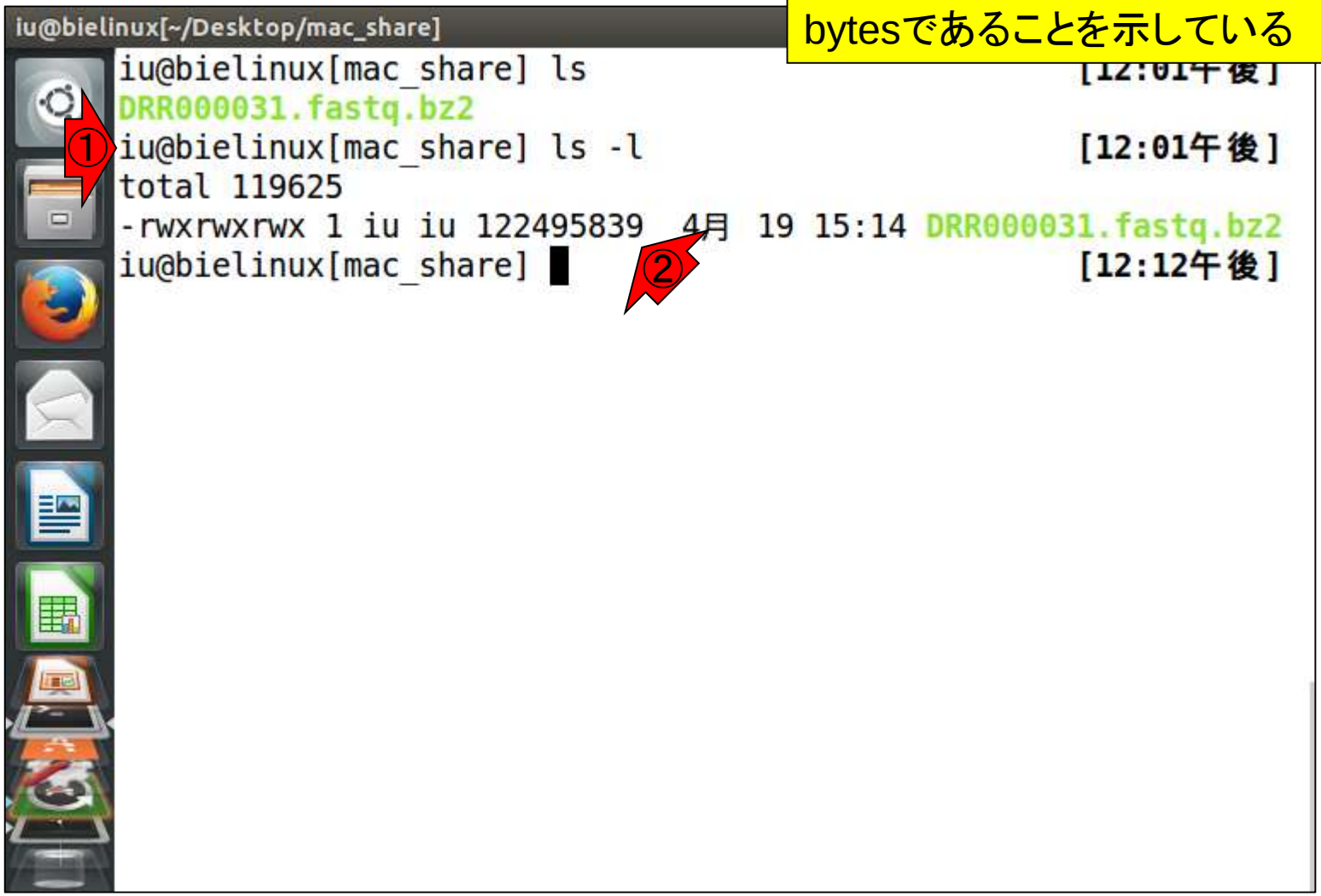
Linux



これは、Bio-Linuxのターミナル画面です。細かい点はすっ飛ばします。①ls（えるえす）というコマンドで、作業ディレクトリ（この場合は~/Desktop/mac_shareという場所）中のファイルを表示。③作業ディレクトリ中に、DRR000031.fastq.bz2というbzip2圧縮FASTQファイルが1つだけあることがわかる

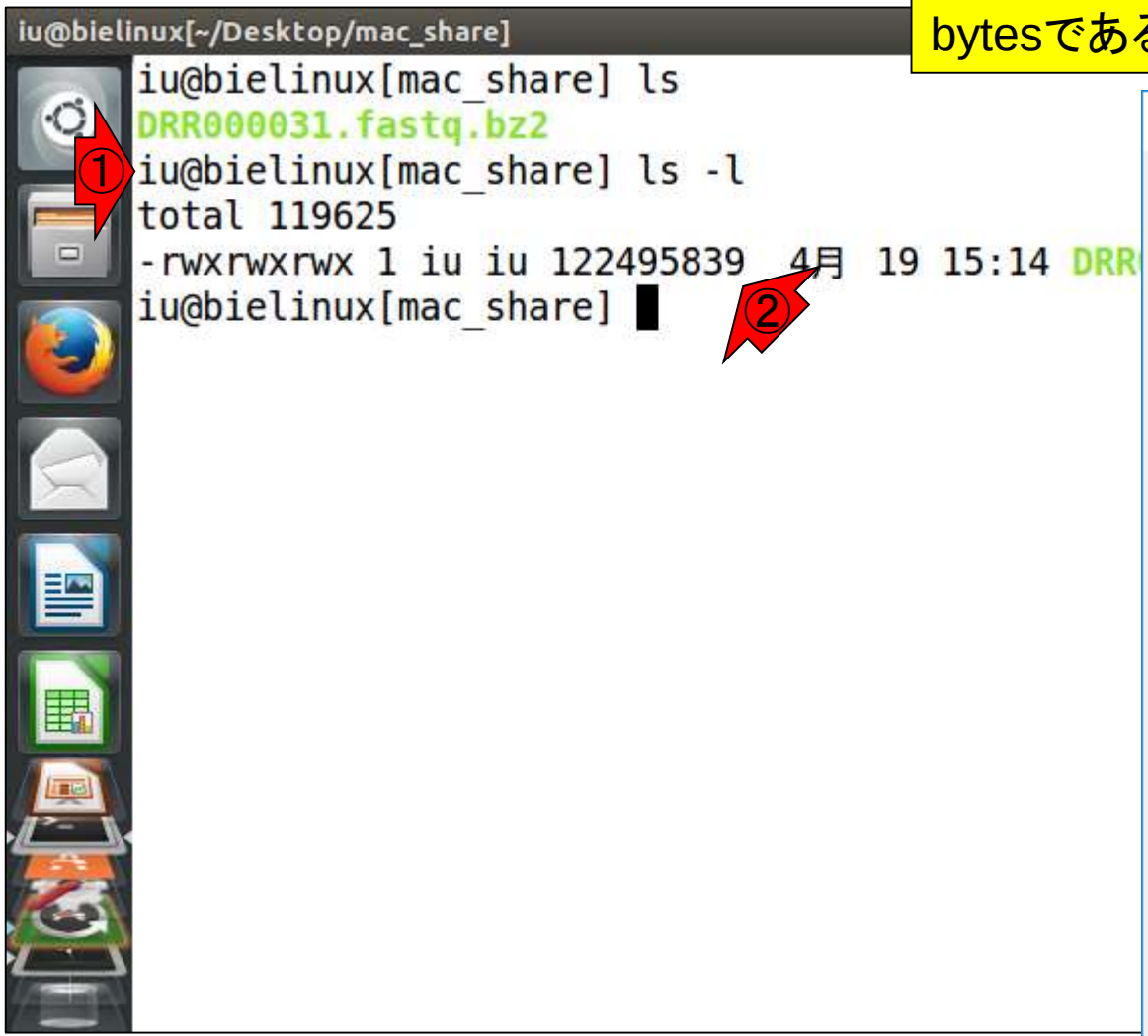
Linux

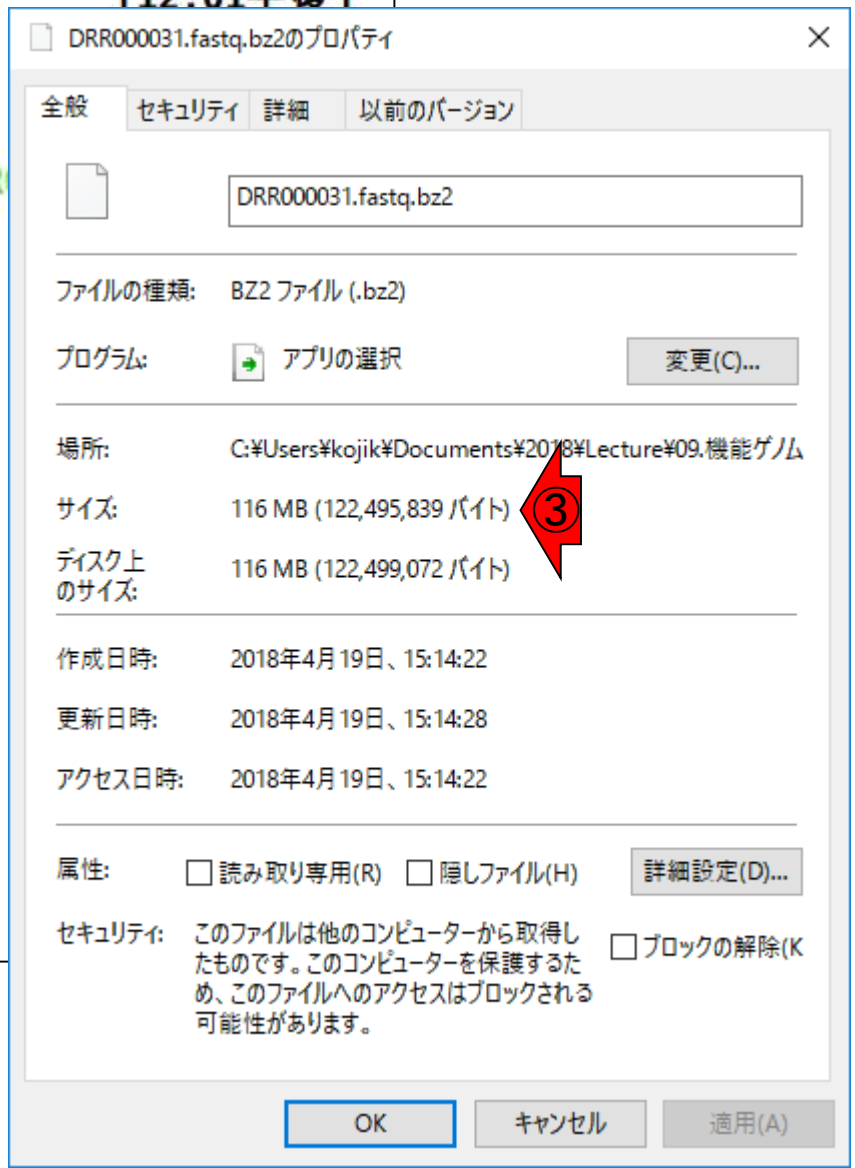
①ls -l（えるえす、スペース、ハイフンえる）として、lsコマンドに詳細情報(long情報)を表示させるオプションをつけて実行した結果。②がファイルサイズで122,495,839 bytesであることを示している



Linux

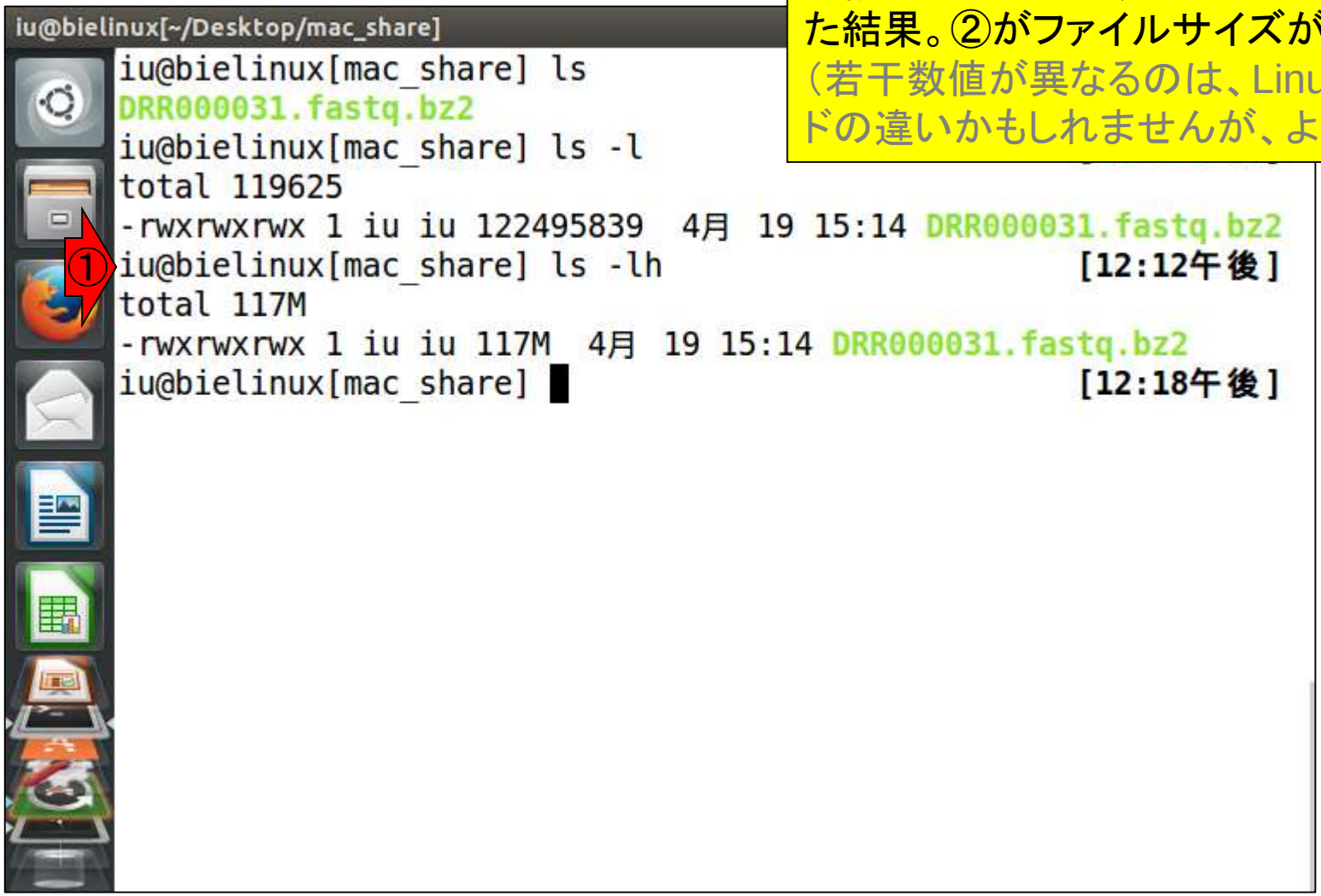
①ls -l（えるえす、スペース、ハイフンえる）として、lsコマンドに詳細情報(long情報)を表示させるオプションをつけて実行した結果。②がファイルサイズで122,495,839 bytesであることを示している。③と同じ数値ですね





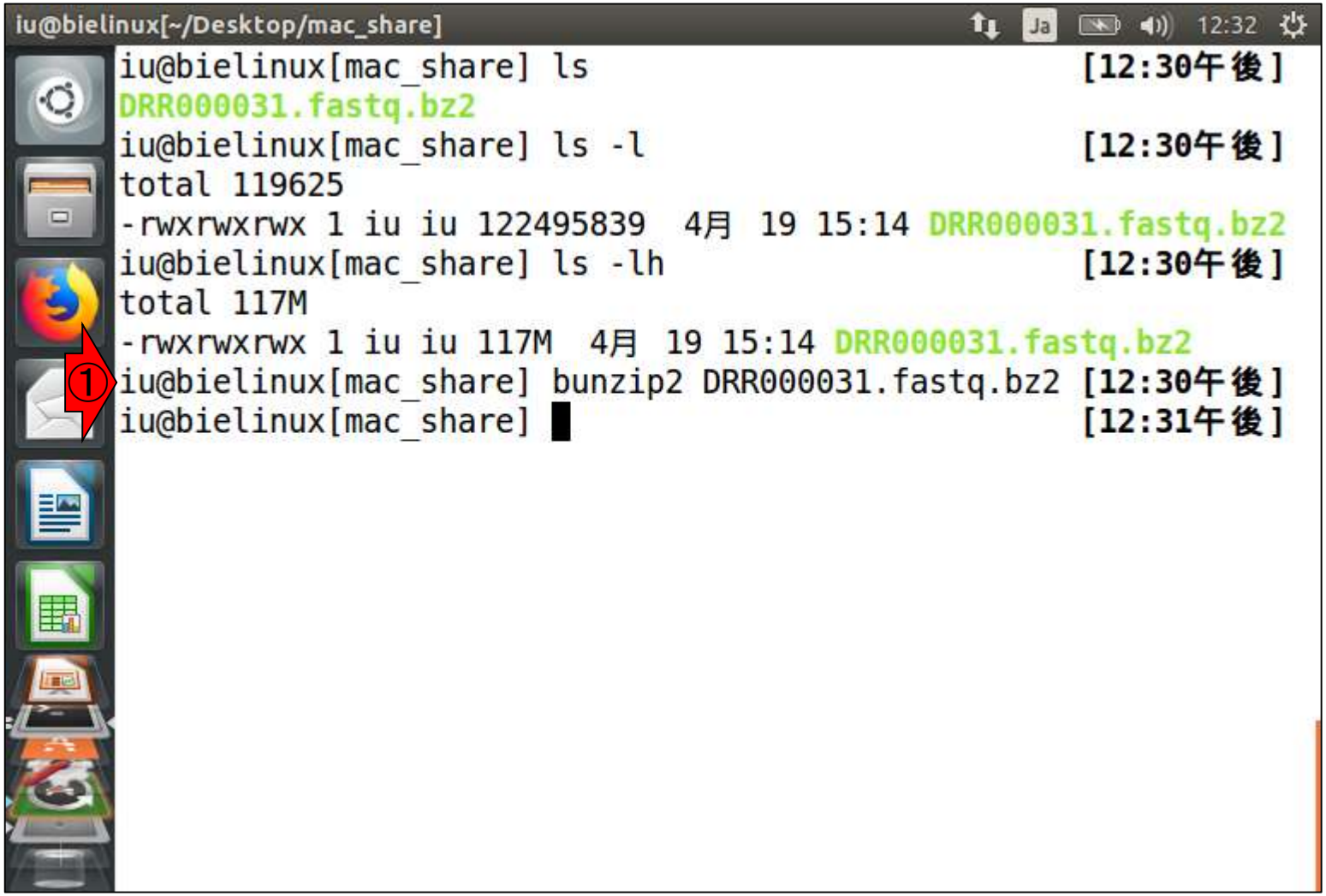
Linux

①ls -lh（えるえす、スペース、ハイフンえるえいち）として、lsコマンドに詳細情報(long情報)をヒト(human)が判読しやすい形で表示させるオプションをつけて実行した結果。②がファイルサイズが117MBとなっています。（若干数値が異なるのは、LinuxとWindowsの改行コードの違いかもしれませんが、よくわかりません）



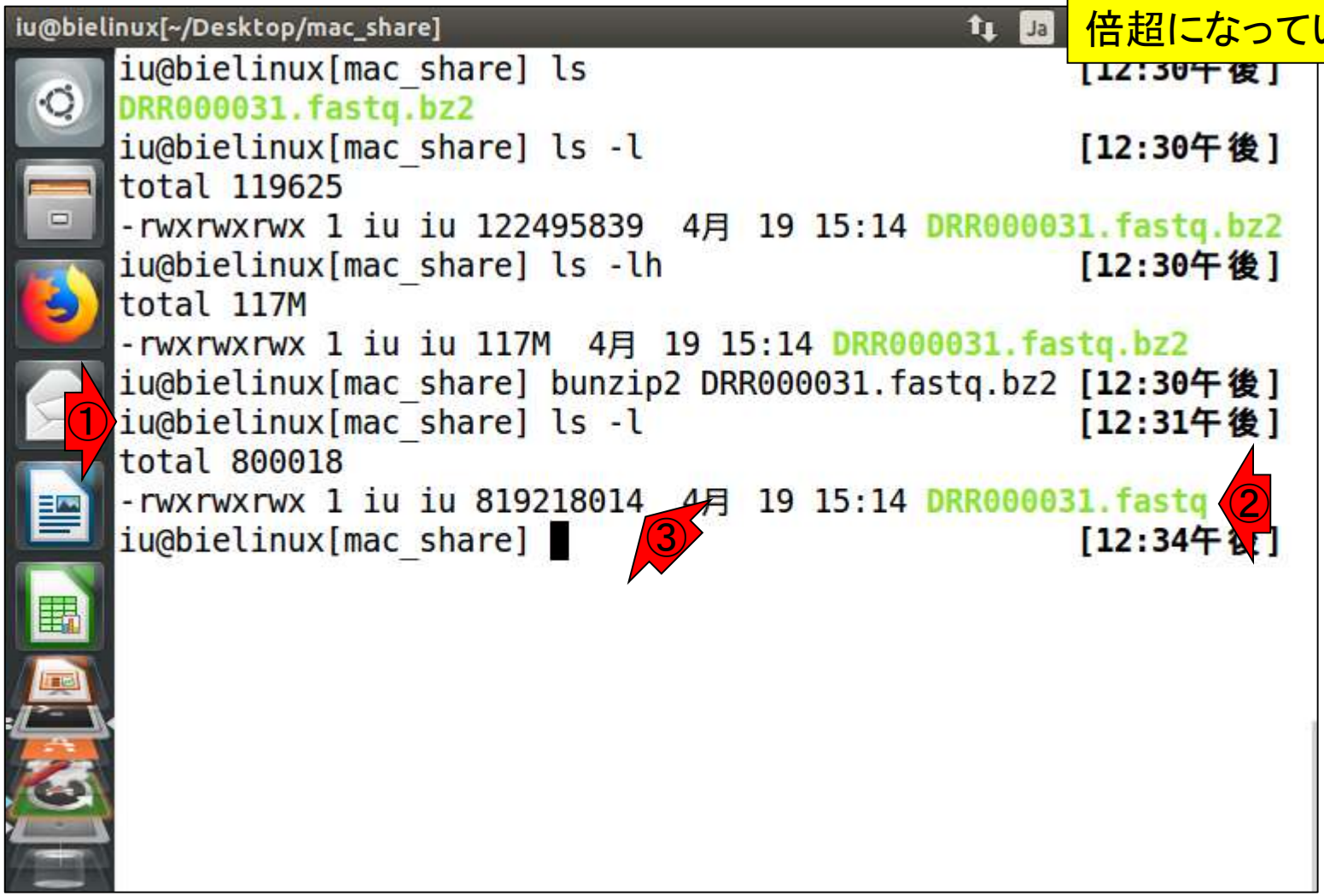
Linux

①bunzip2コマンドで.bz2という拡張子がついた圧縮ファイルを解凍



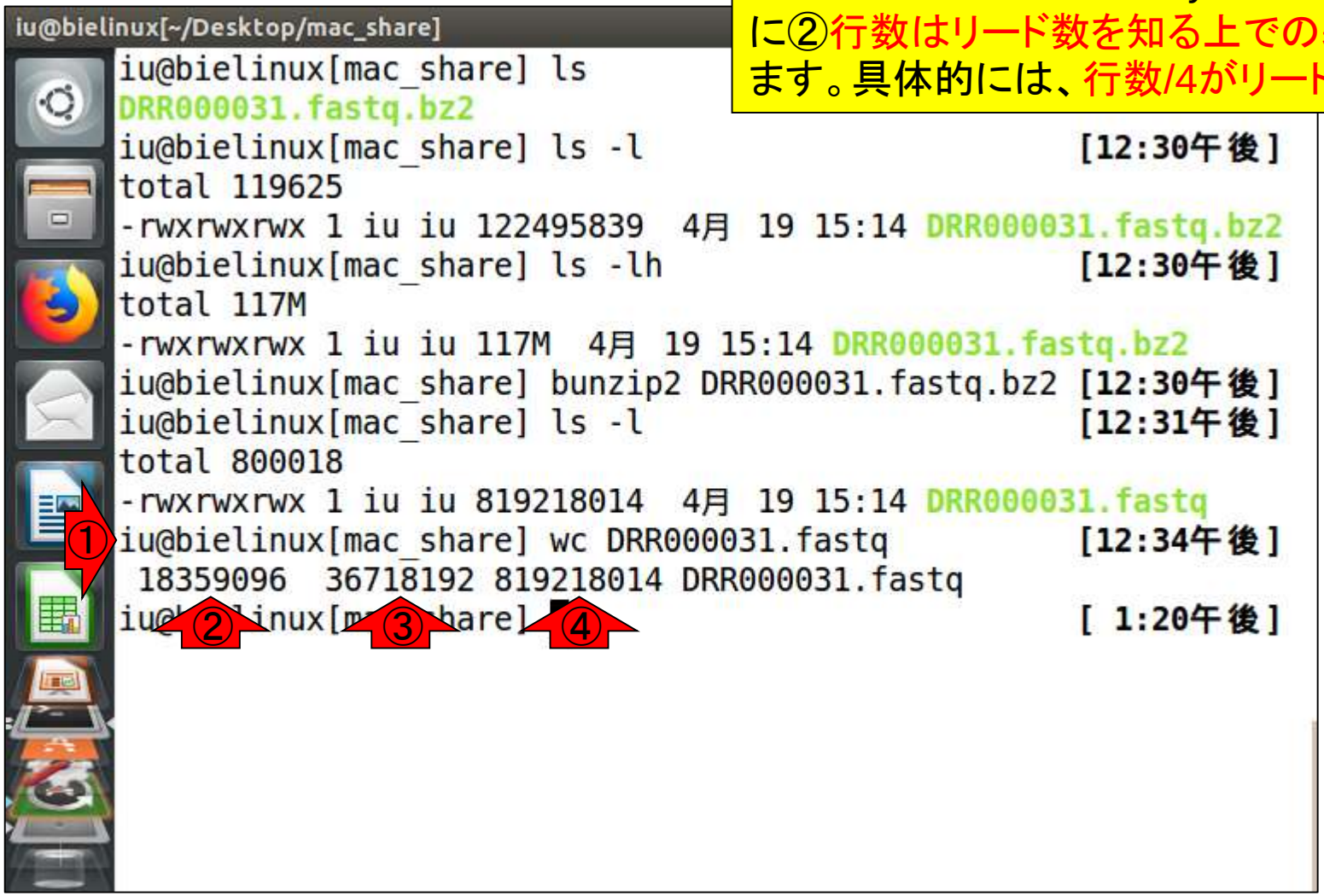
Linux

①ls -l（えるえす、スペース、ハイフンえる）実行結果。②確かに解凍され、③ファイルサイズも819,218,014 bytesと6倍超になっていることがわかる



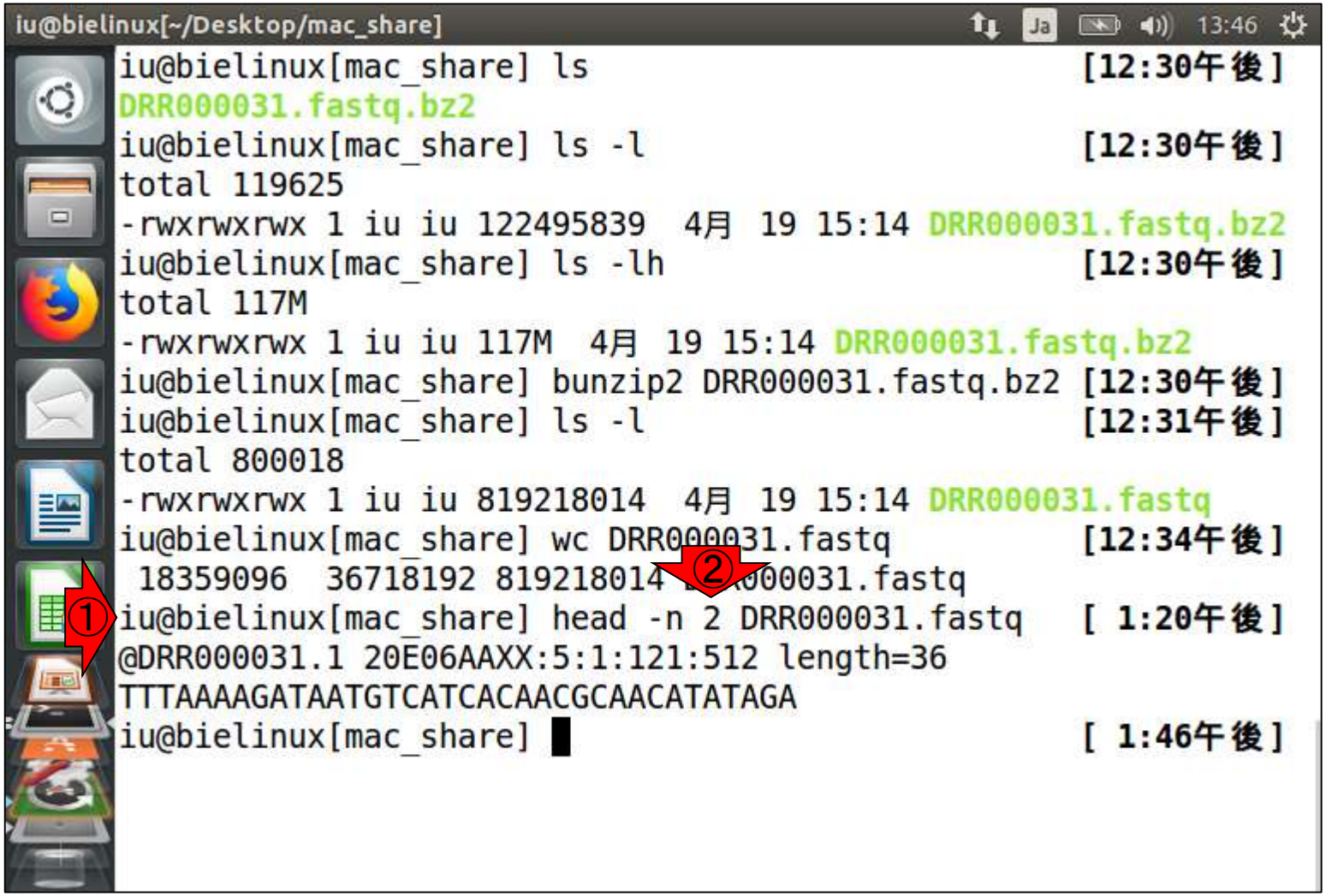
Linux

①wcは行数、単語数、ファイルサイズを表示するコマンド。DRR000031.fastqは②18,359,096 lines、③36,718,192 words、④819,218,014 bytesであることがわかります。特に②行数はリード数を知る上での基礎情報としてよく使います。具体的には、行数/4がリード数になります



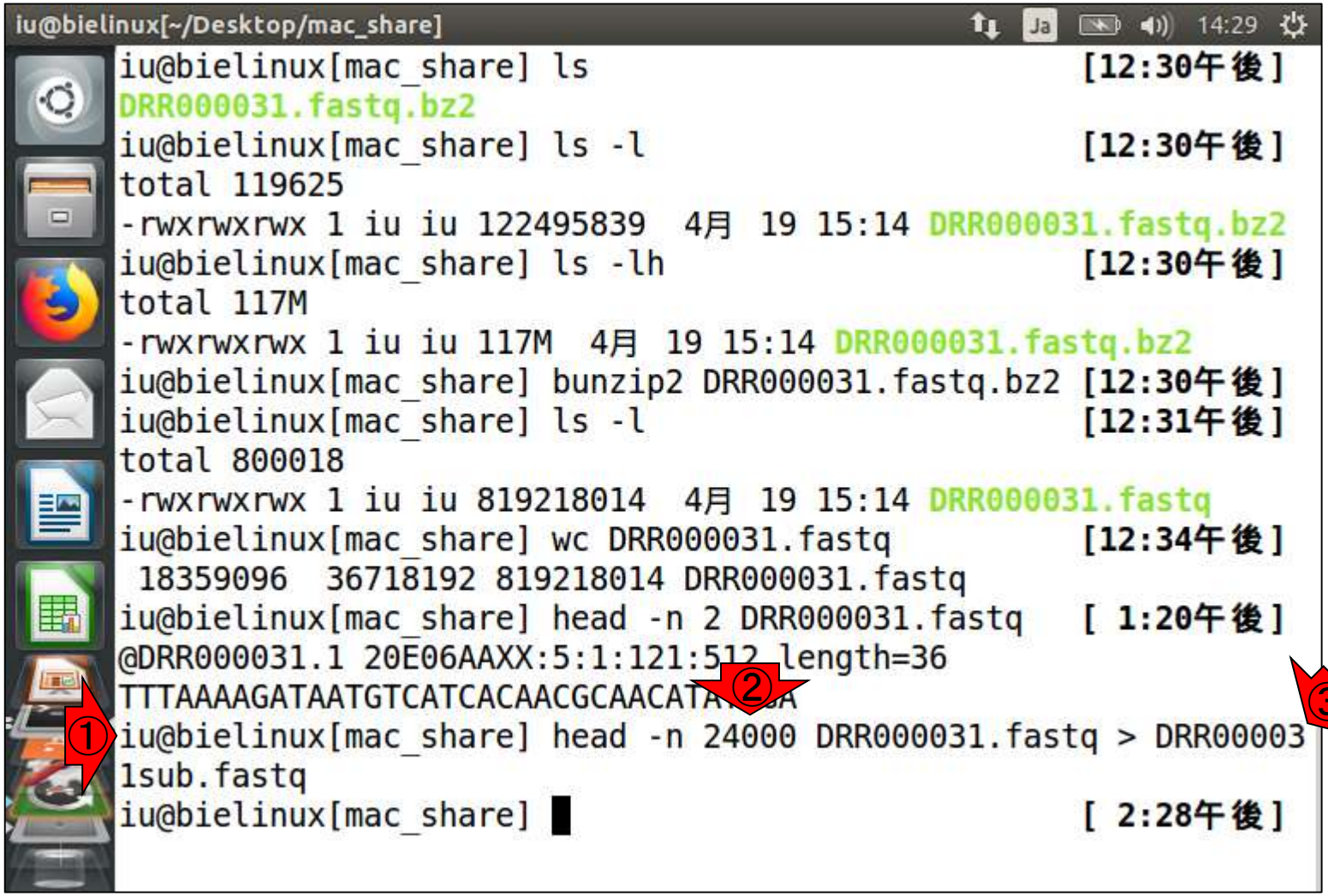
Linux

①headは、ファイルの先頭から-nで指定した行数分だけを表示するコマンド。この場合は、②最初の2行分だけを表示



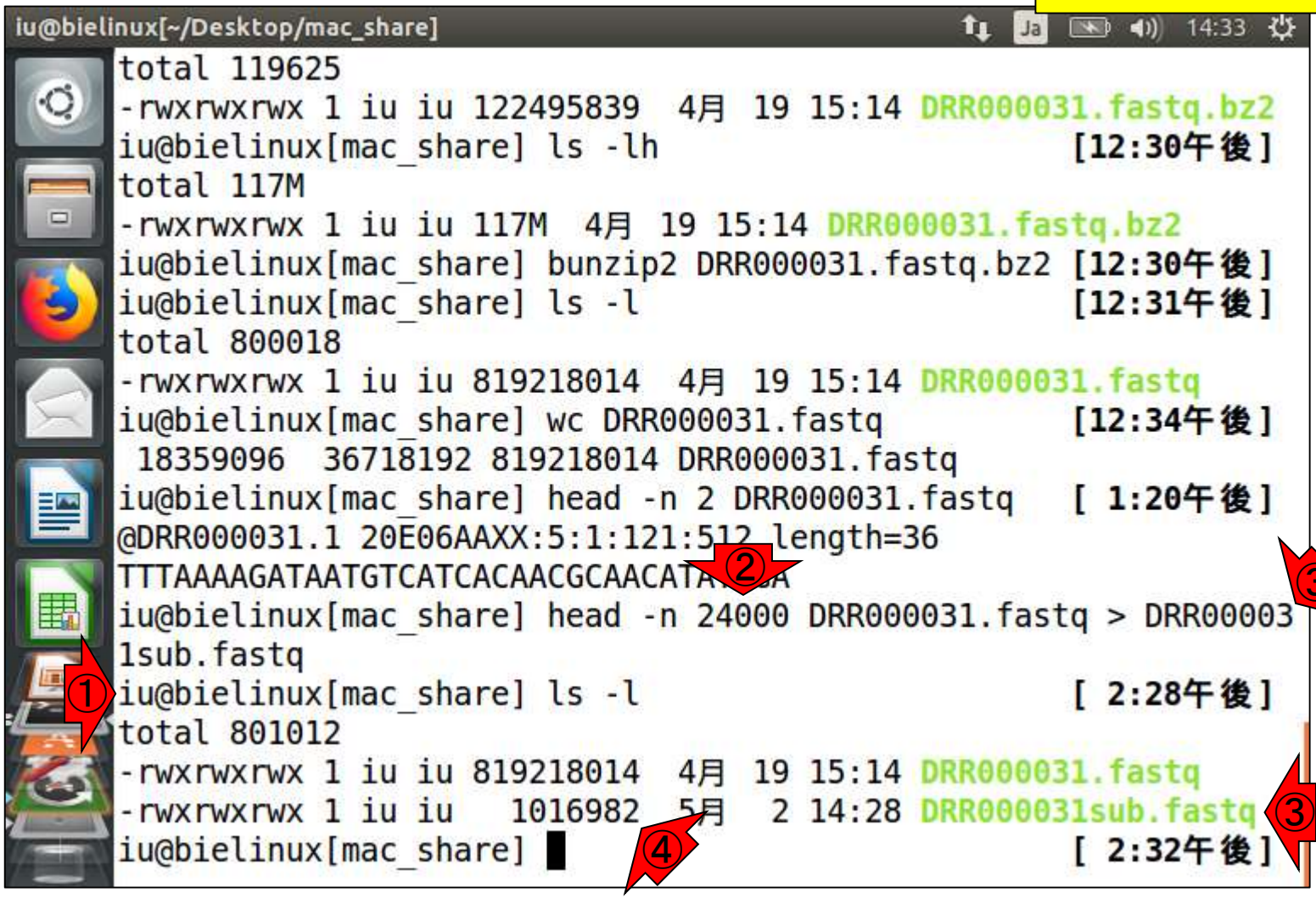
Linux

①同じノリで、②最初の24000行分だけを表示させるのではなく、③DRR000031sub.fastqという名前のファイルに出力



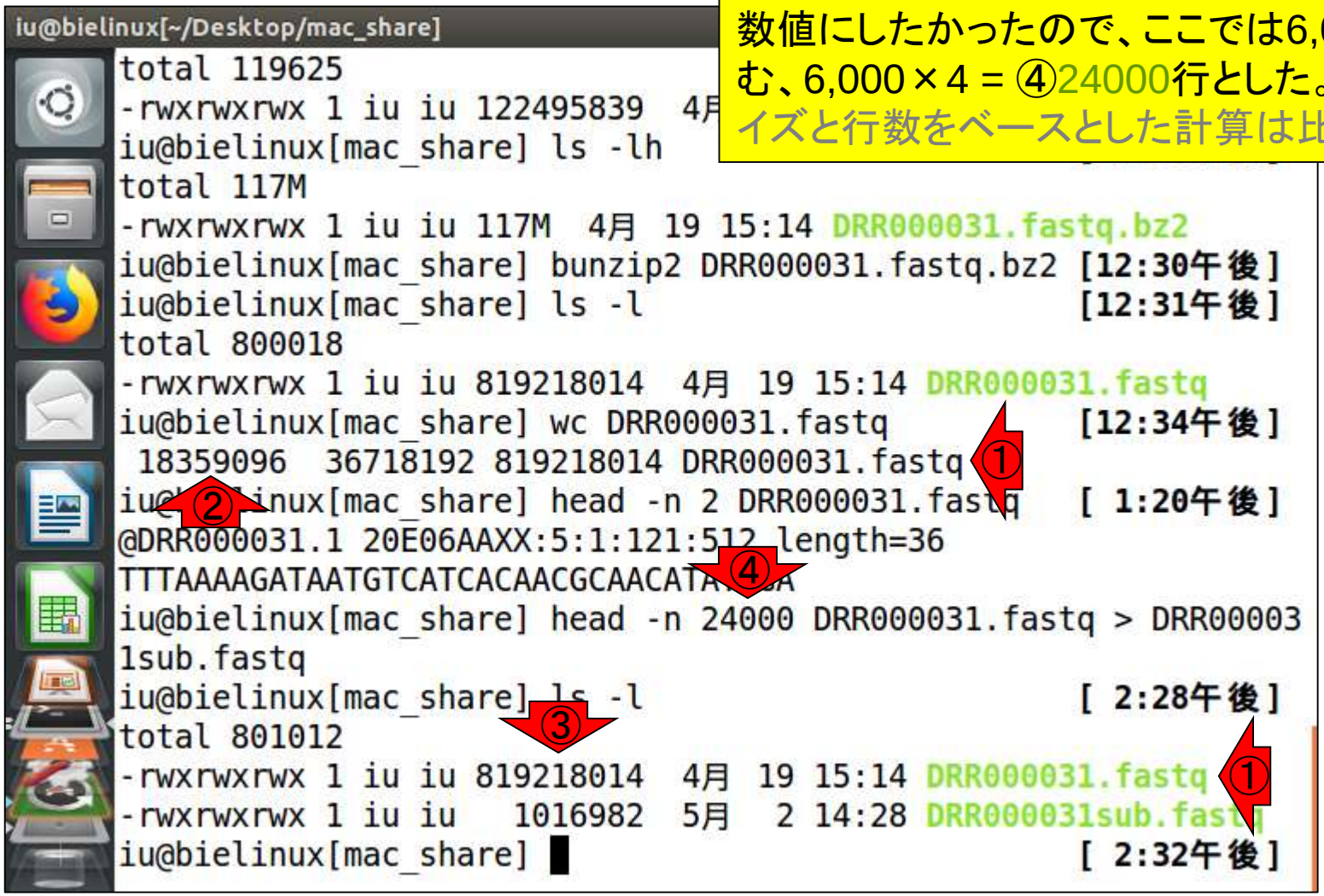
Linux

①ls -lで確認。②最初の24000行としたのは、③DRR000031sub.fastqのファイルサイズを④約1MBにしたかったから



Linux

①DRR000031.fastqは、②18,359,096行で、③約800MB。それゆえ、18,359,096 / 800 = 22,948.87行分程度を抽出すれば、約1MBのファイルとなる。リード数をキリのいい数値にしたかったので、ここでは6,000リード分の情報を含む、6,000×4 = ④24000行とした。こんな感じのファイルサイズと行数をベースとした計算は比較的よく行います



DRR000031sub.fastq

①と②は同じもの（約1MBのFASTQファイル）です。②をダウンロードして、ワードパッドなど手持ちのエディタで開いておくとよい

```
iu@bielinux[~/Desktop/mac_share]
total 119625
-rwxrwxrwx 1 iu iu 122495839  4月  19 15:14 DRR000031
iu@bielinux[mac_share] ls -lh                          [
total 117M
-rwxrwxrwx 1 iu iu 117M  4月  19 15:14 DRR000031.fast
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [
iu@bielinux[mac_share] ls -l                           [
total 800018
-rwxrwxrwx 1 iu iu 819218014  4月  19 15:14 DRR000031
iu@bielinux[mac_share] wc DRR000031.fastq              [
 18359096  36718192 819218014 DRR000031.fastq
iu@bielinux[mac_share] head -n 2 DRR000031.fastq       [
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
iu@bielinux[mac_share] head -n 24000 DRR000031.fastq
1sub.fastq
iu@bielinux[mac_share] ls -l                           [
total 801012
-rwxrwxrwx 1 iu iu 819218014  4月  19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu   1016982  5月   2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share]                                 [ 2:32午後 ]
```

講義日程（2019年度）

1. 2019年05月27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)
   - Maser : Kinjo et al., Database (Oxford), 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
   - 講義資料PDF
   - (Rで)塩基配列解析
   - DRR000031sub.fastq ②
   - RNA-QC-chain : Zhou et al., BMC Genomics, 2018
   - Biostar : Parnell et al., PLoS Comput Biol., 2011
   - FastQC
   - DRR000031sub_fastqc.html
   - DRR000031_fastqc.html(課題用)
   - report.html(qrqcを用いたQC結果)

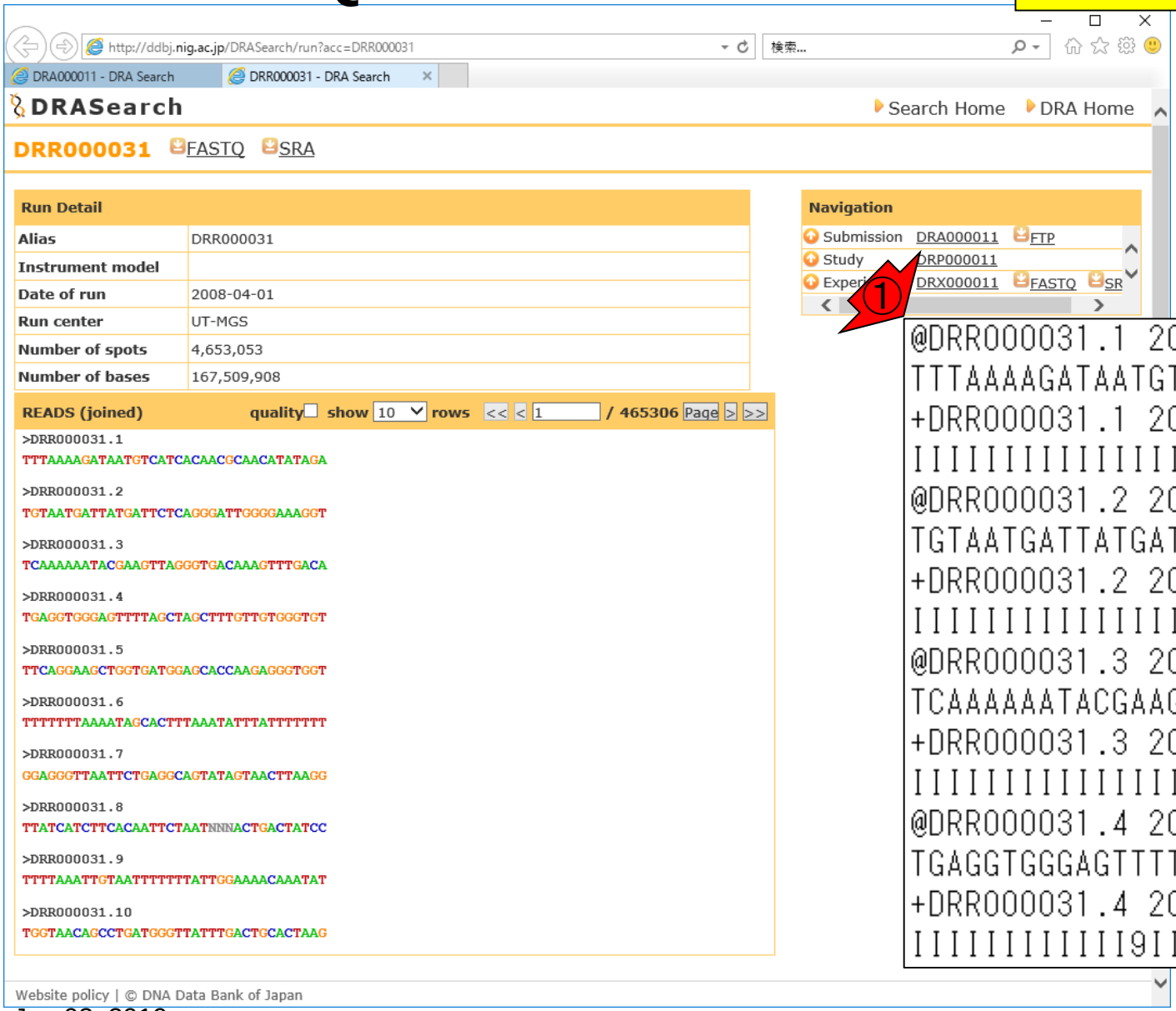
①

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

FASTQの説明

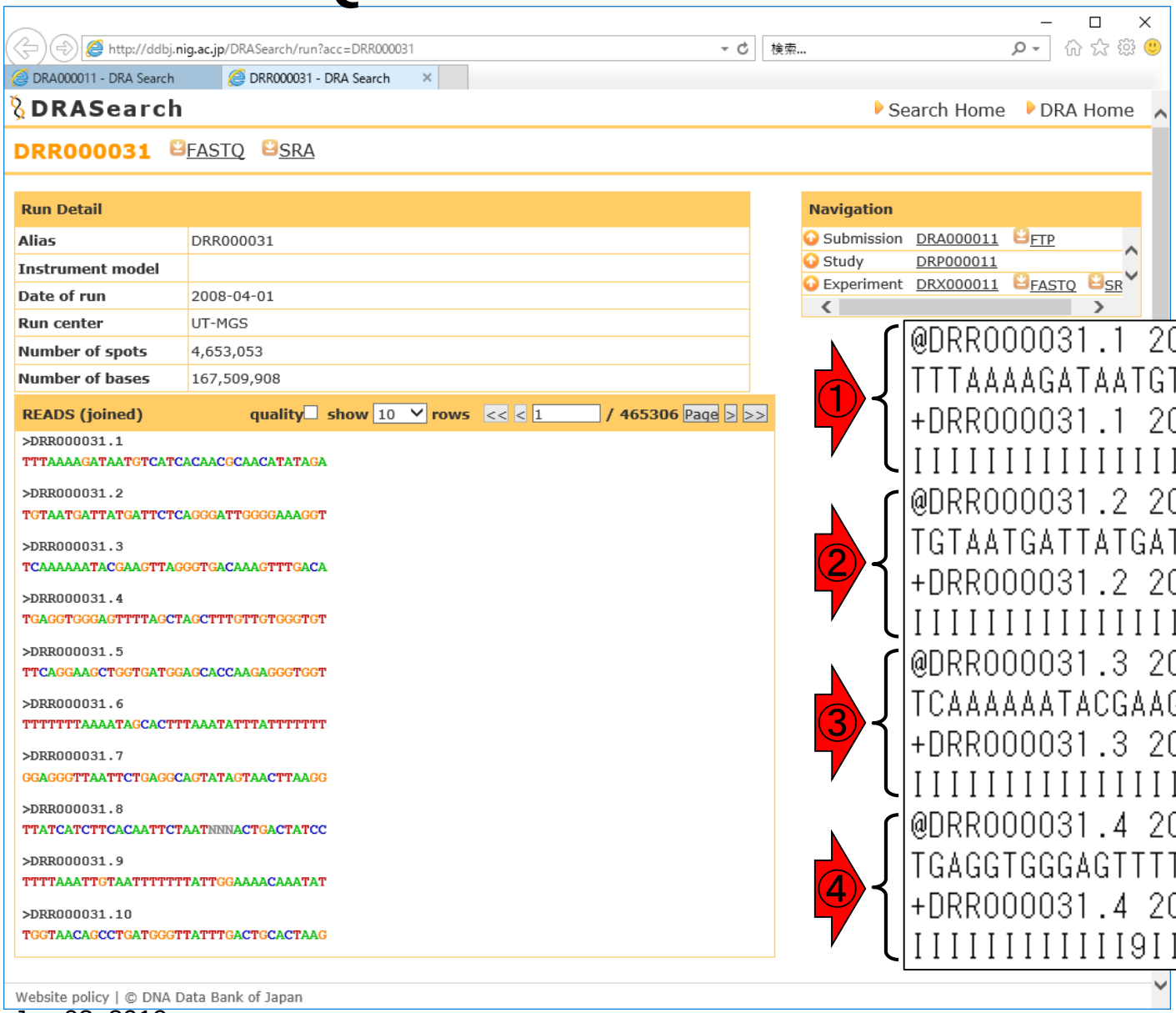
①FASTQファイル（DRR000031.fastq）の最初の16行分を表示。FASTQ形式は、1つのリード情報を4行で表す




```
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
+DRR000031.1 20E06AAXX:5:1:121:512 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.2 20E06AAXX:5:1:120:207 length=36
TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
+DRR000031.2 20E06AAXX:5:1:120:207 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.3 20E06AAXX:5:1:122:269 length=36
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
+DRR000031.3 20E06AAXX:5:1:122:269 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.4 20E06AAXX:5:1:234:114 length=36
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
+DRR000031.4 20E06AAXX:5:1:234:114 length=36
IIIIIIIIIII9IIII.III'IIII7IIII&IDII
```

FASTQの説明

①1番目、②2番目、③3番目、④4番目のリード、といった具合です



①
```
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
+DRR000031.1 20E06AAXX:5:1:121:512 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

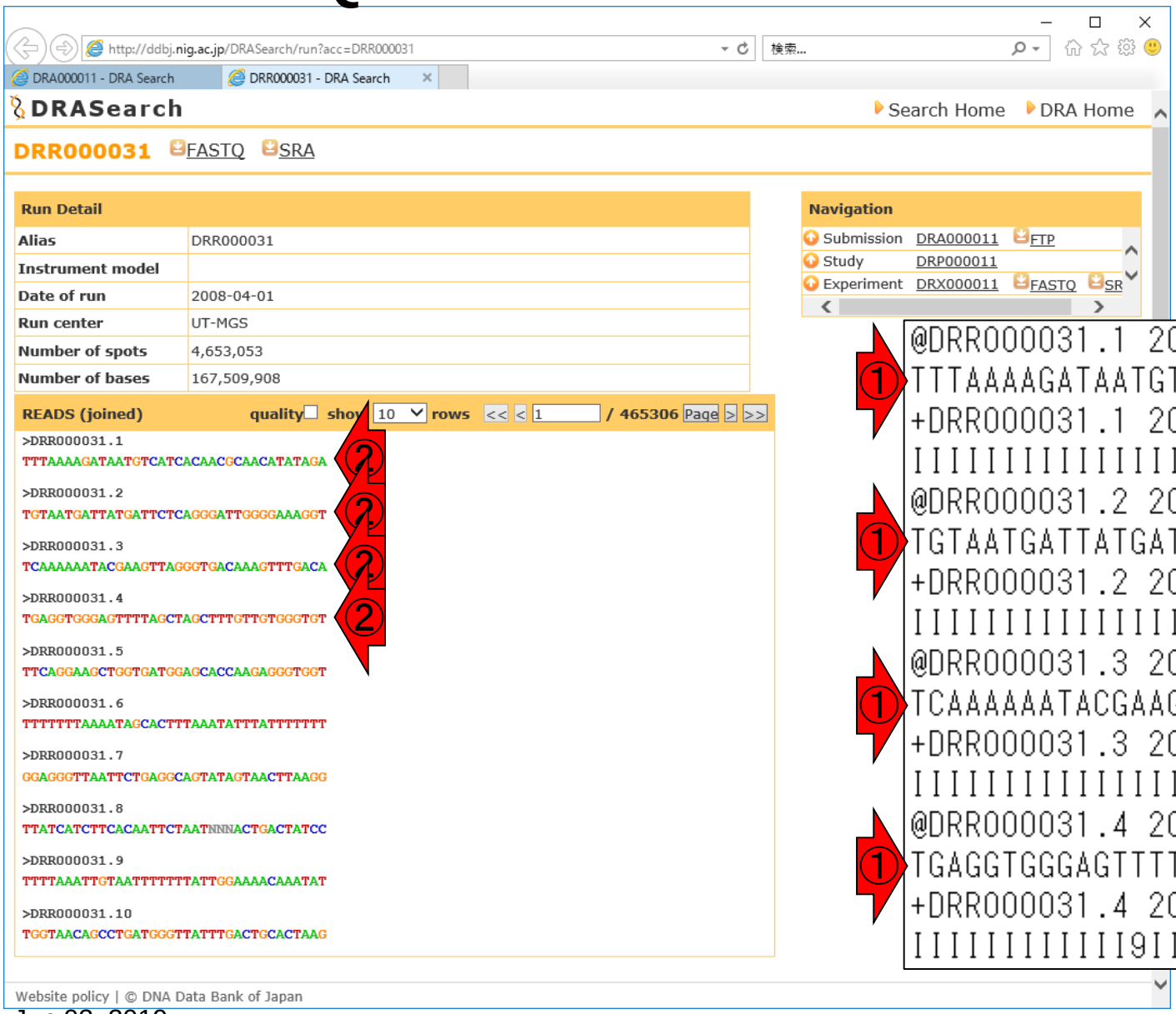
②
```
@DRR000031.2 20E06AAXX:5:1:120:207 length=36
TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
+DRR000031.2 20E06AAXX:5:1:120:207 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

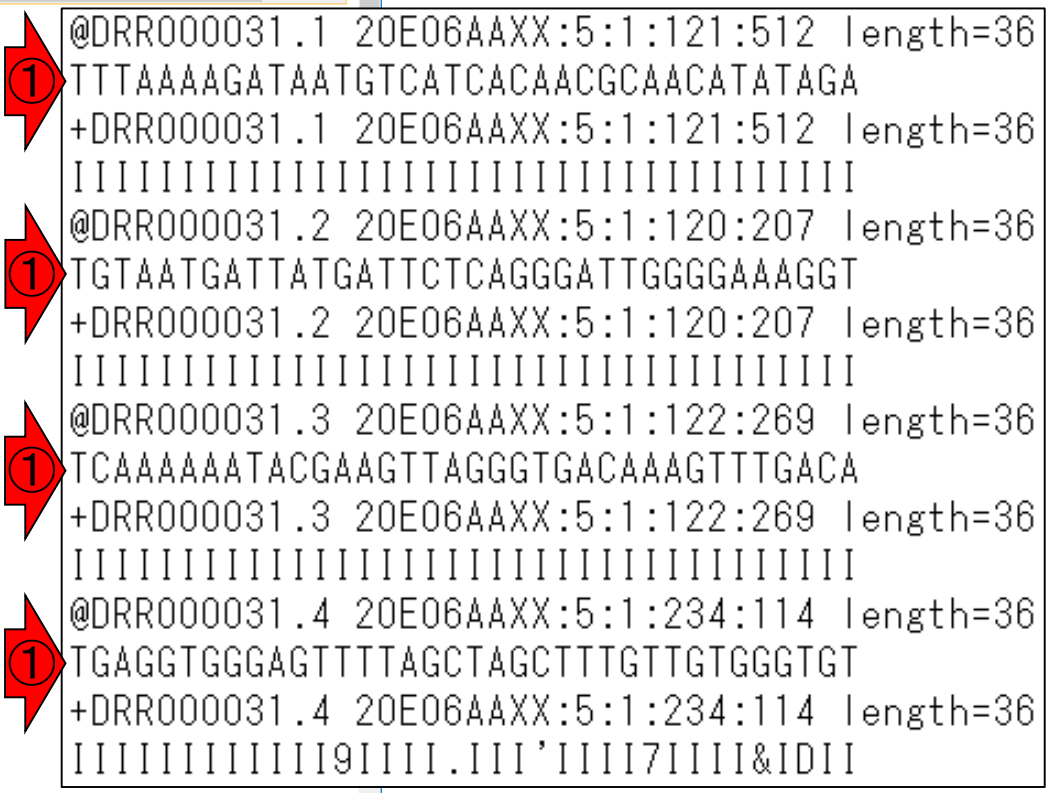
③
```
@DRR000031.3 20E06AAXX:5:1:122:269 length=36
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
+DRR000031.3 20E06AAXX:5:1:122:269 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

④
```
@DRR000031.4 20E06AAXX:5:1:234:114 length=36
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
+DRR000031.4 20E06AAXX:5:1:234:114 length=36
IIIIIIIIIII9IIII.III'IIII7IIII&IDII
```

FASTQの説明

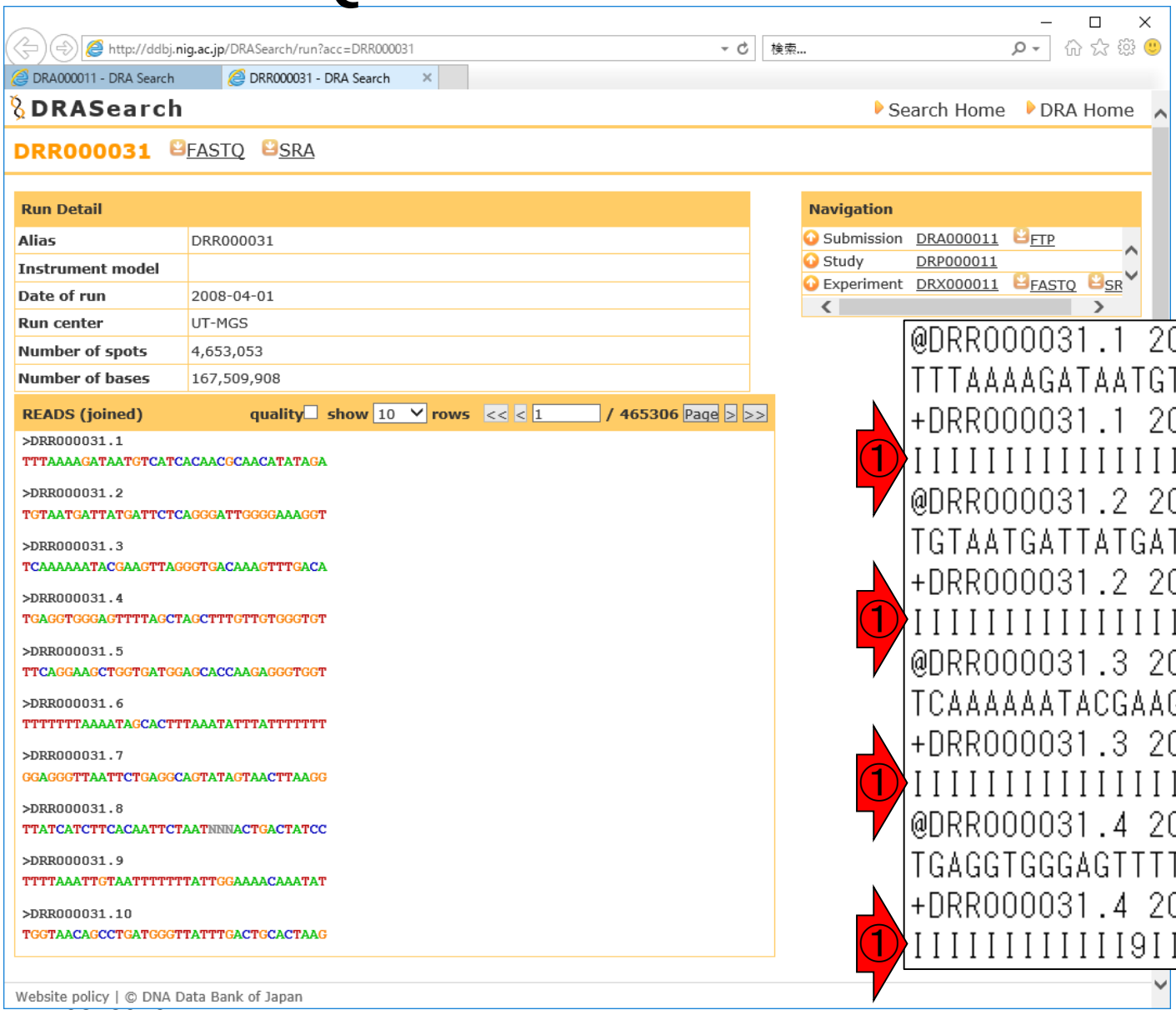
①これらがリードの塩基配列情報で、②これらが対応するものです





①これらがリードのクオリティ情報です

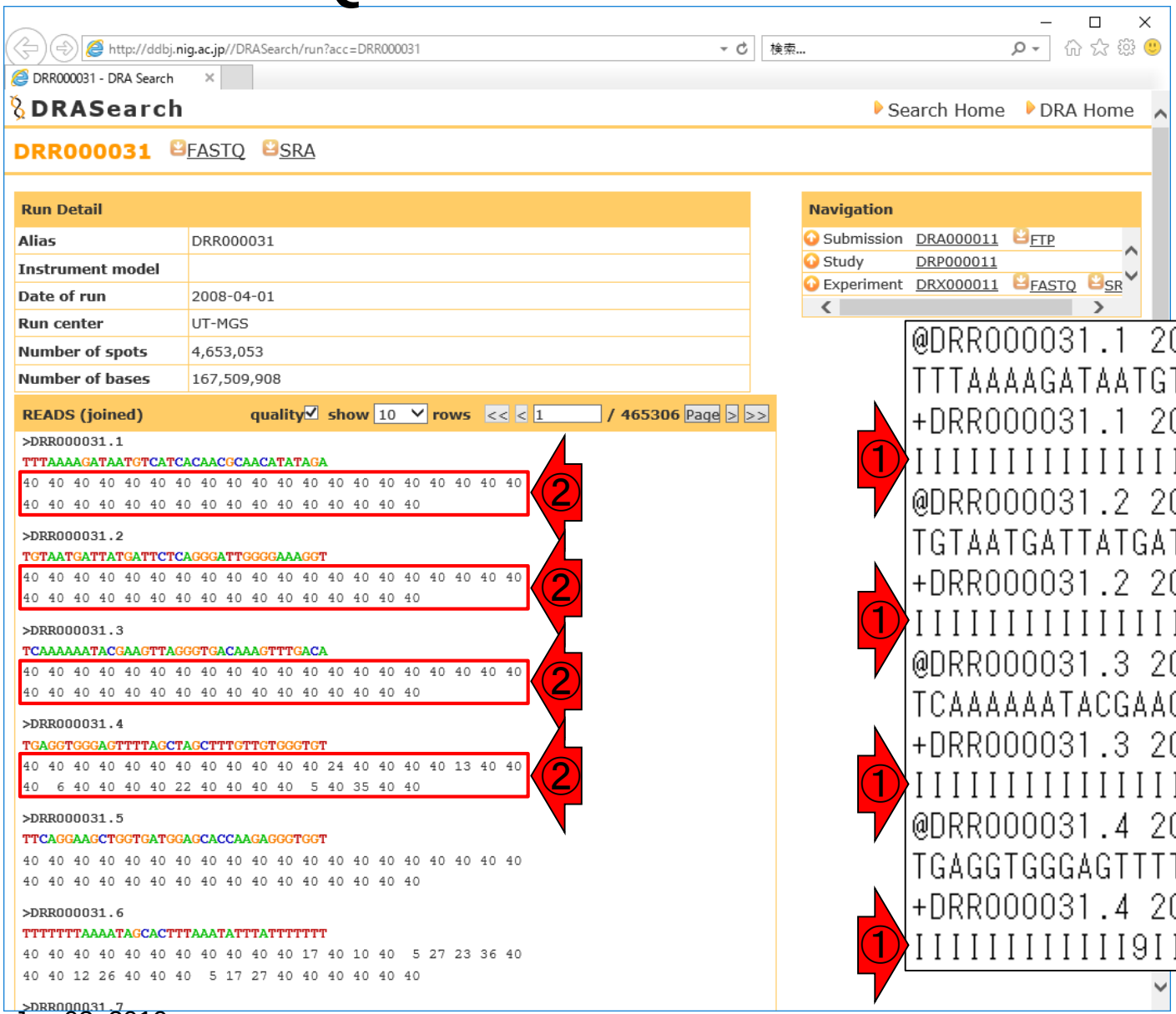
FASTQの説明

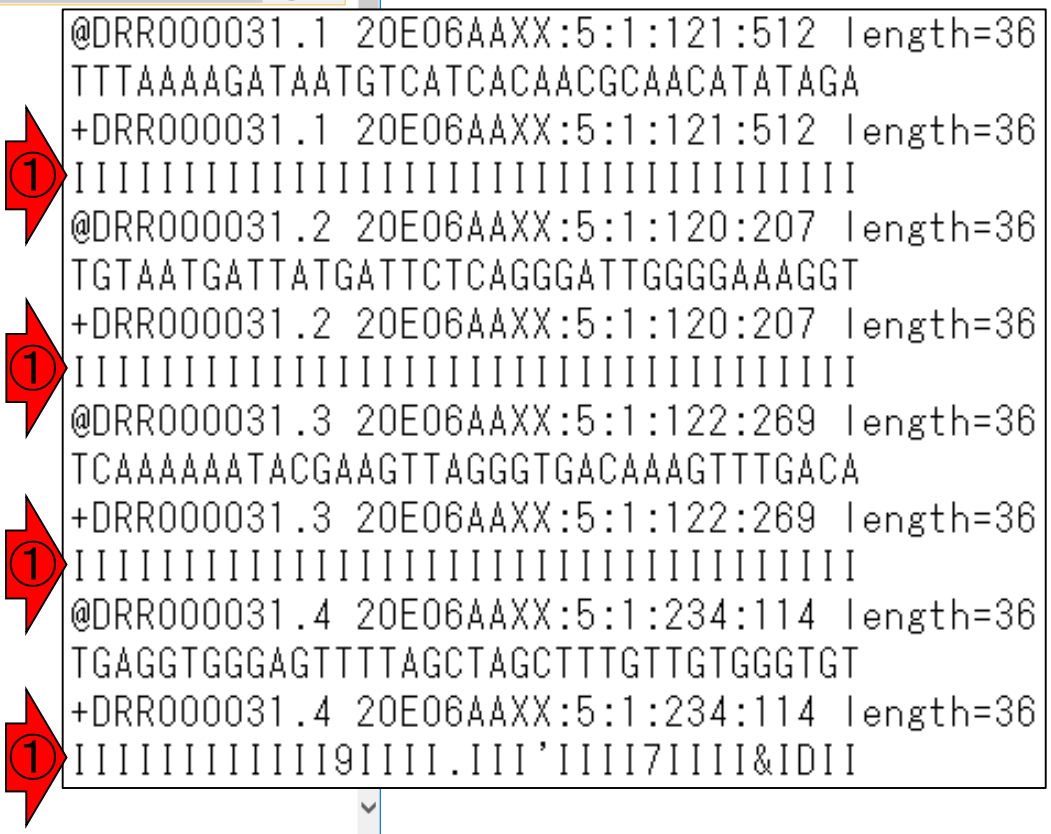



```
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
+DRR000031.1 20E06AAXX:5:1:121:512 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.2 20E06AAXX:5:1:120:207 length=36
TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
+DRR000031.2 20E06AAXX:5:1:120:207 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.3 20E06AAXX:5:1:122:269 length=36
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
+DRR000031.3 20E06AAXX:5:1:122:269 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.4 20E06AAXX:5:1:234:114 length=36
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
+DRR000031.4 20E06AAXX:5:1:234:114 length=36
IIIIIIIIIII9IIII.III'IIII7IIII&IDII
```

FASTQの説明

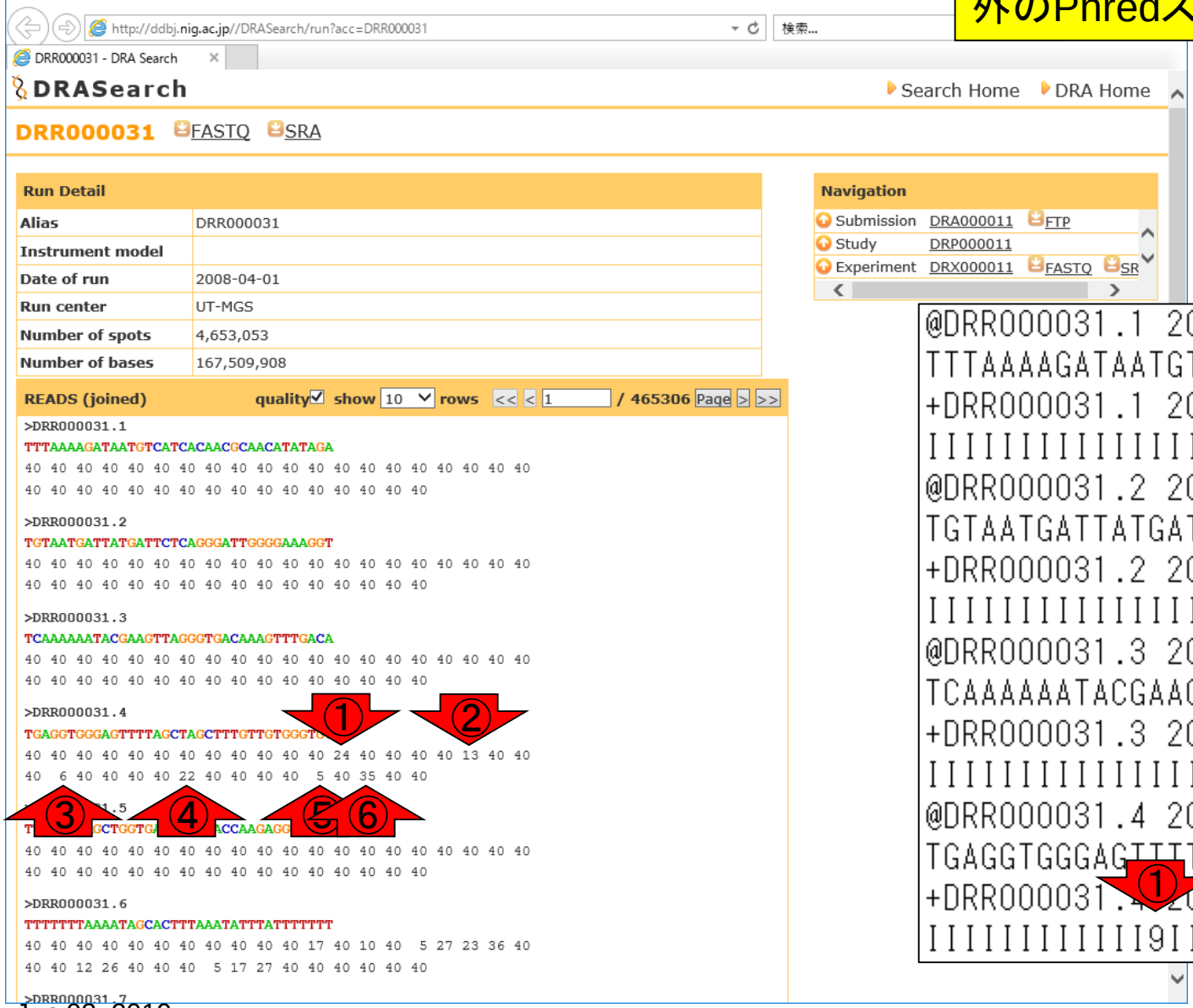
②対応するクオリティスコア情報（Phredスコア）です





FASTQの説明

全体を眺めて、Phredスコア40が、一文字表記のクオリティ情報I（←あい）に相当することがわかる。①～⑥のI以外のところが、40以外のPhredスコアになっていることもわかる

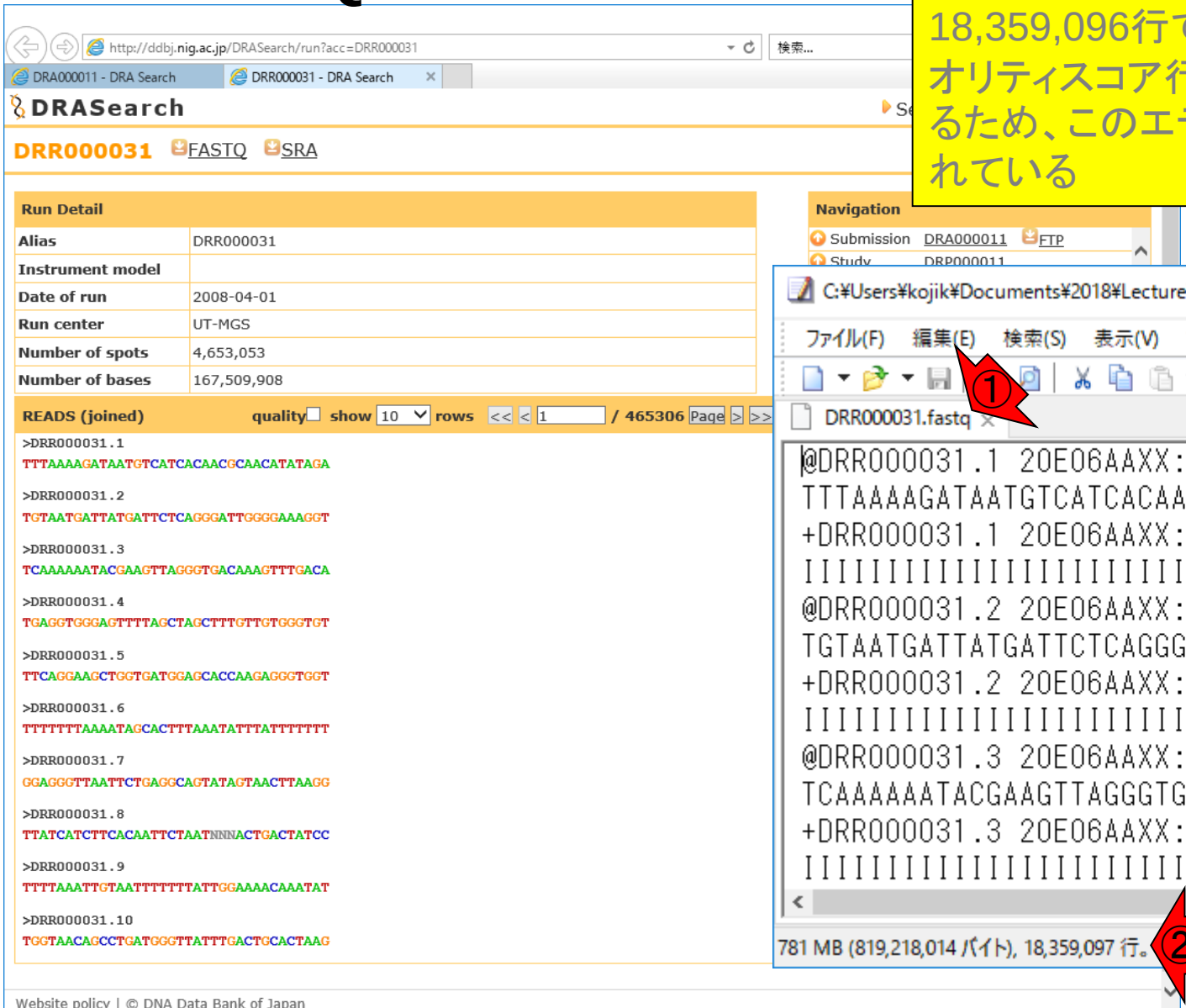


Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

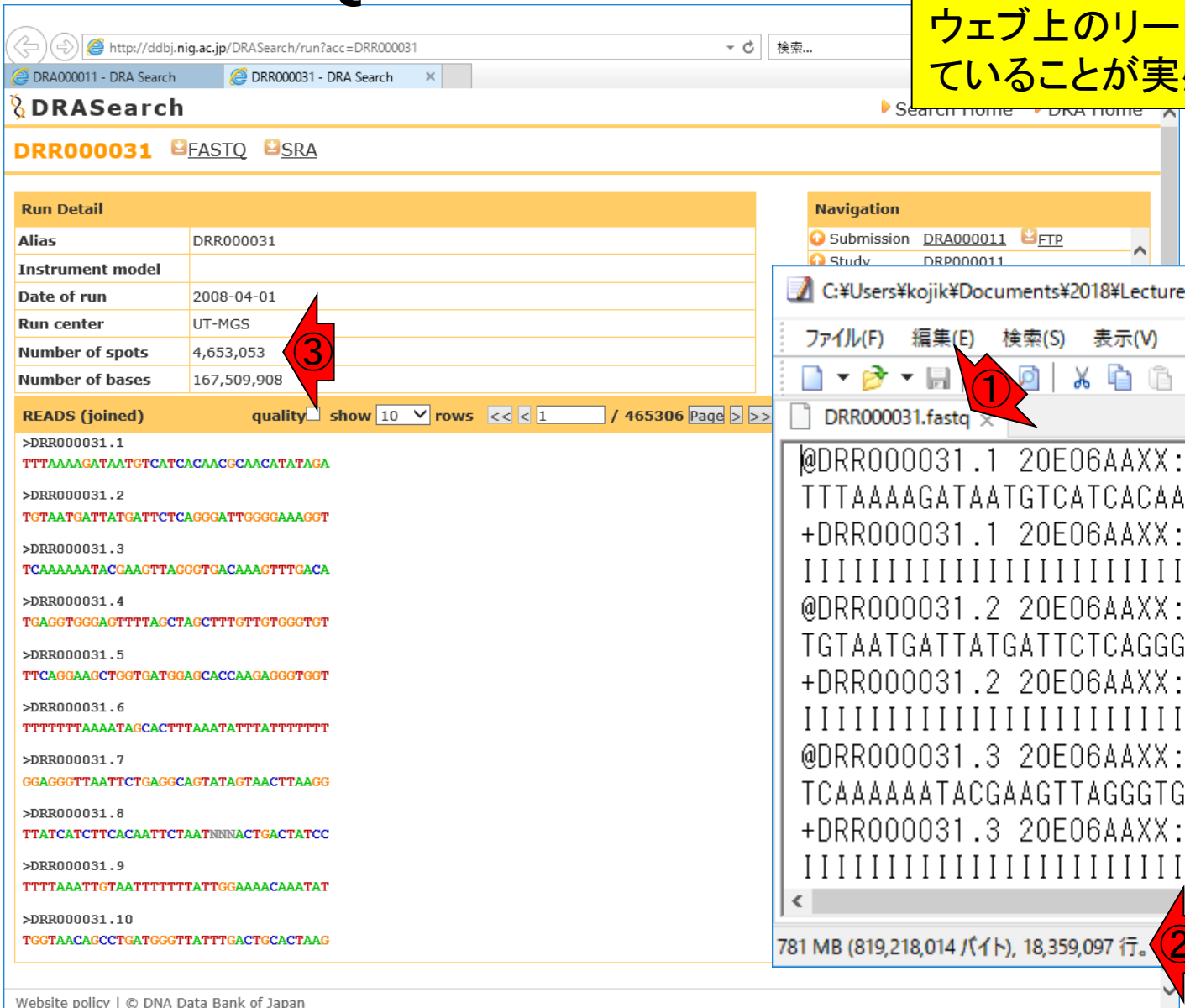
FASTQのリード数

①FASTQファイル（DRR000031.fastq）の、②行数は18,359,097。この行数を4で割った4,589,774がFASTQファイル中のリード数。18,359,096行であって欲しかったが、最後のクオリティスコア行のところに改行が含まれているため、このエディタでは1行余分にカウントされている



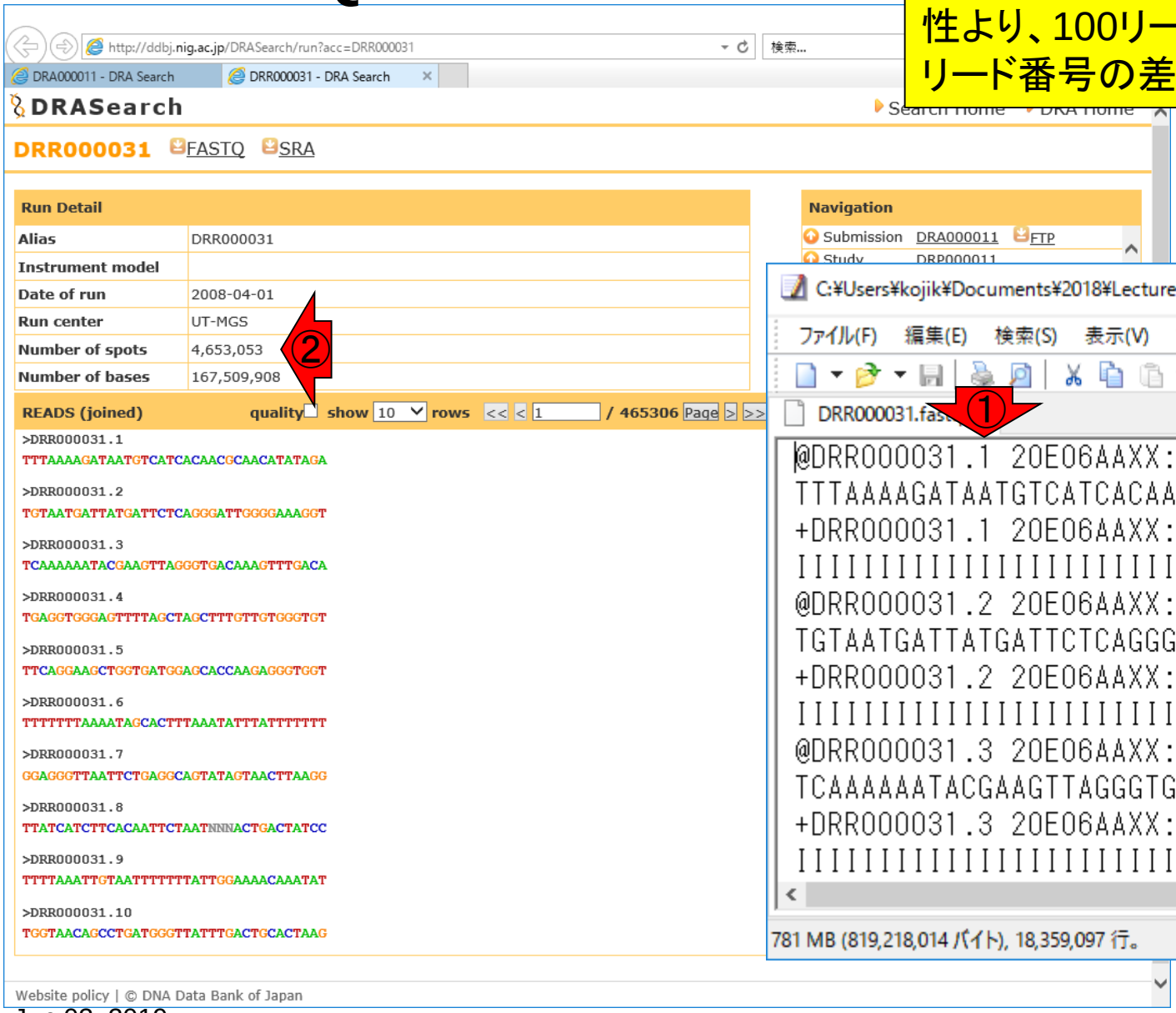
FASTQのリード数

①FASTQファイル（DRR000031.fastq）の、②行数は18,359,097。この行数を4で割った4,589,774がFASTQファイル中のリード数。③ウェブ上のリード数情報（4,653,053）と異なっていることが実感できる



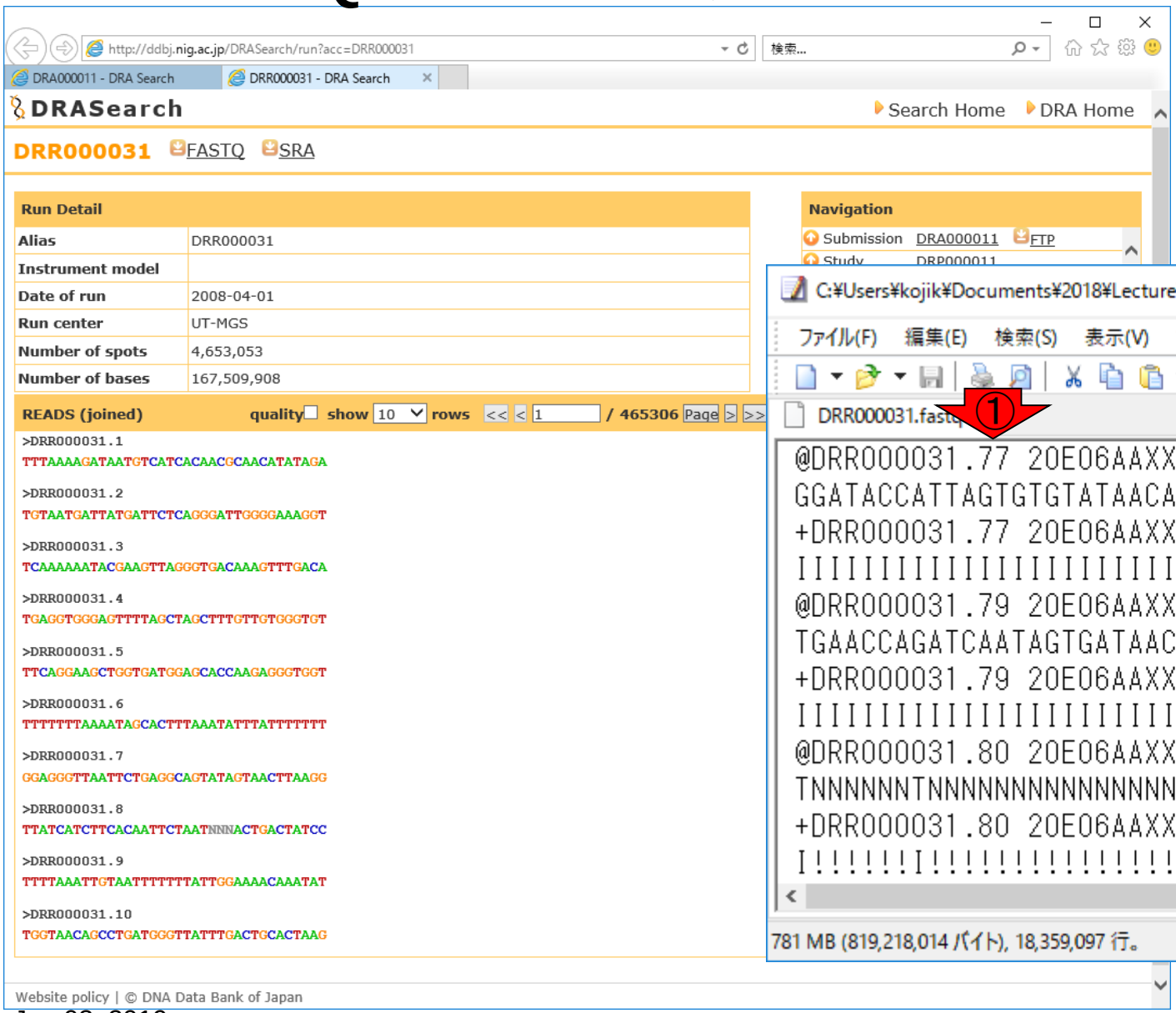
FASTQのリード数

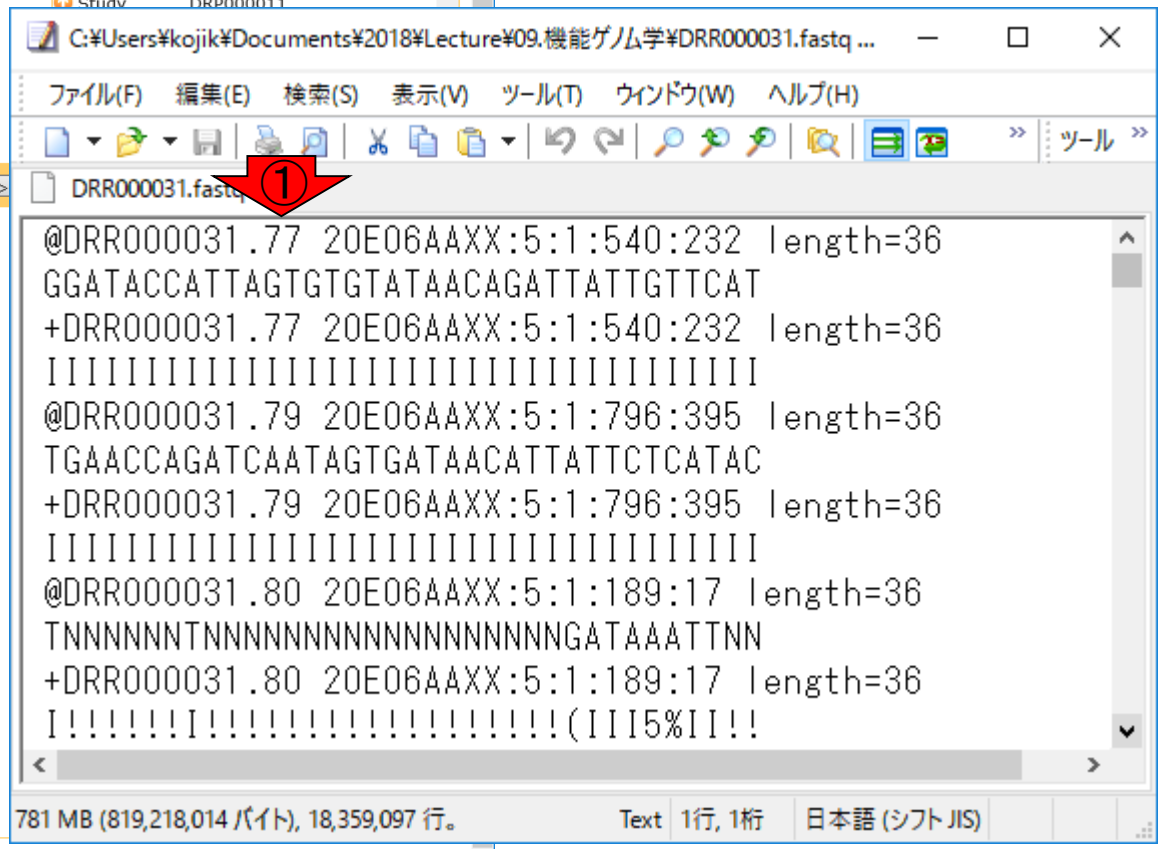
①この部分がリード番号に相当する。FASTQファイル中の4,589,774リードと、②ウェブ上の4,653,053リードの差分（63,279リード）の関係性より、100リードあたり1個程度は①の部分のリード番号の差分が1ではないのだろうと想像



FASTQのリード数

ず～っとページ下部を見ていくと、①リード番号78が抜けているのがわかる





FASTQのリード数

①リード番号80も相当クオリティが低いことがわかります。リード塩基配列がNだらけだからです。ということは、リード番号78は全くお話にならないくらいクオリティが低いのでしょう。それをこの後確認していきます

http://ddbj.nig.ac.jp/DRASearch/run?acc=DRR000031

DRA000011 - DRA Search | DRR000031 - DRA Search

DRASearch Search Home | DRA Home

DRR000031 FASTQ SRA

Run Detail	
Alias	DRR000031
Instrument model	
Date of run	2008-04-01
Run center	UT-MGS
Number of spots	4,653,053
Number of bases	167,509,908

Navigation
Submission DRA000011 FTP
Study DRP000011

READS (joined) quality show 10 rows << < 1 / 465306 Page > >>

```
>DRR000031.1
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
>DRR000031.2
TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
>DRR000031.3
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
>DRR000031.5
TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
>DRR000031.6
TTTTTTTAAAATAGCACTTTAAATATTTATTTTTTT
>DRR000031.7
GGAGGGTTAATTCTGAGGCAGTATAGTAACTTAAGG
>DRR000031.8
TTATCATCTTCACAATTCTAATNNNACTGACTATCC
>DRR000031.9
TTTTAAATTGTAATTTTTTTATTGGAAAACAAATAT
>DRR000031.10
TGGTAACAGCCTGATGGGTTATTTGACTGCACTAAG
```

Website policy | © DNA Data Bank of Japan

C:¥Users¥kojik¥Documents¥2018¥Lecture¥09.機能ゲノム学¥DRR000031.fastq ...

ファイル(F) 編集(E) 検索(S) 表示(V) ツール(T) ウィンドウ(W) ヘルプ(H)

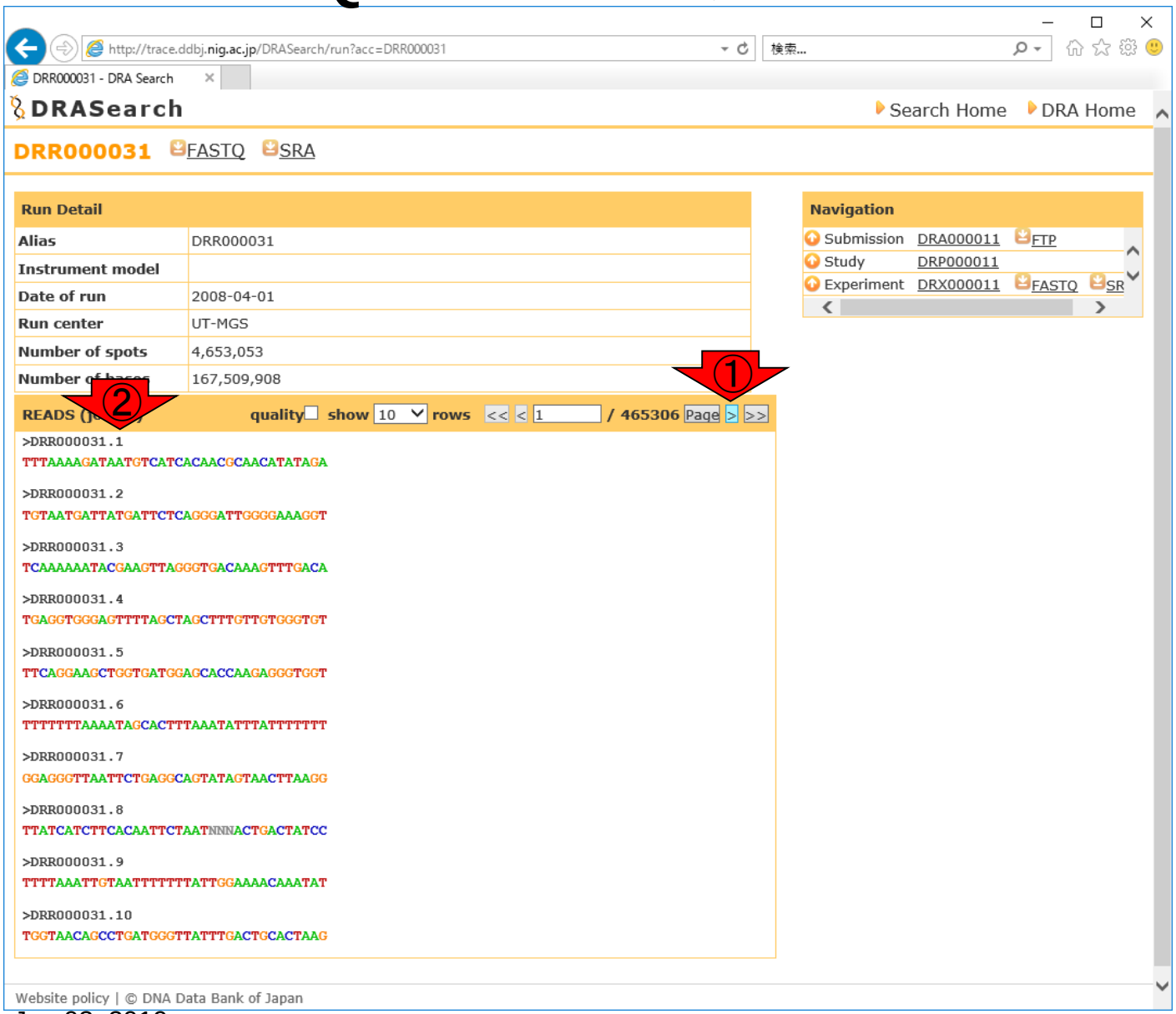
DRR000031.fastq

```
@DRR000031.77 20E06AAXX:5:1:540:232 length=36
GGATACCATTAGTGTGTATAACAGATTATTGTTCAT
+DRR000031.77 20E06AAXX:5:1:540:232 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.79 20E06AAXX:5:1:796:395 length=36
TGAACCAGATCAATAGTGATAACATTATTCTCATAC
+DRR000031.79 20E06AAXX:5:1:796:395 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.80 20E06AAXX:5:1:189:17 length=36
TNNNNNNTNNNNNNNNNNNNNNNNNGATAAATTNN
+DRR000031.80 20E06AAXX:5:1:189:17 length=36
I!!!!!!I!!!!!!!!!!!!!!!!!!(III5%II!!
```

① 781 MB (819,218,014 バイト), 18,359,097 行。 Text 1行, 1桁 日本語 (シフト JIS)

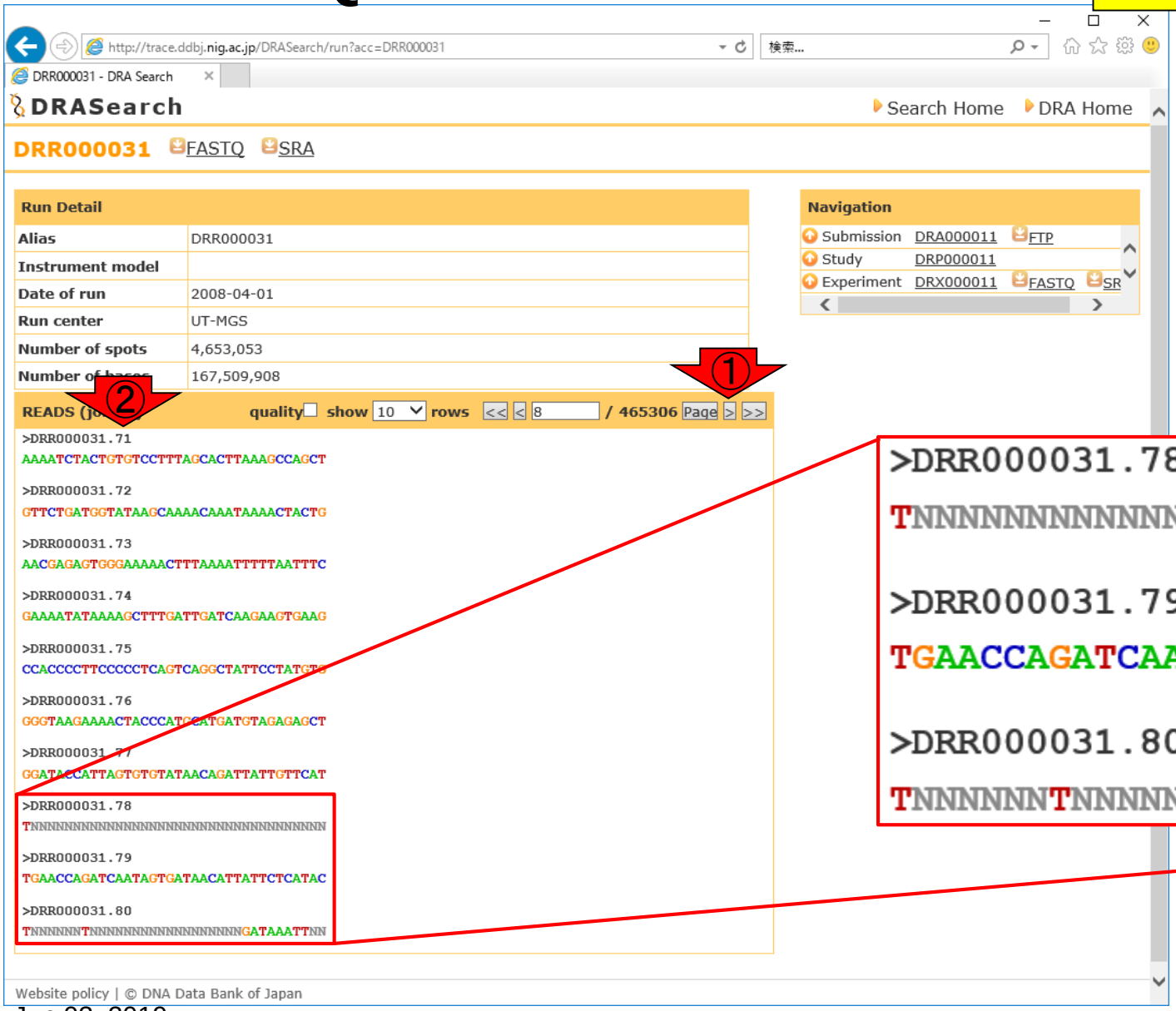
FASTQのリード数

①をガスガス押して、②のところがリード番号78を含むところまで移動



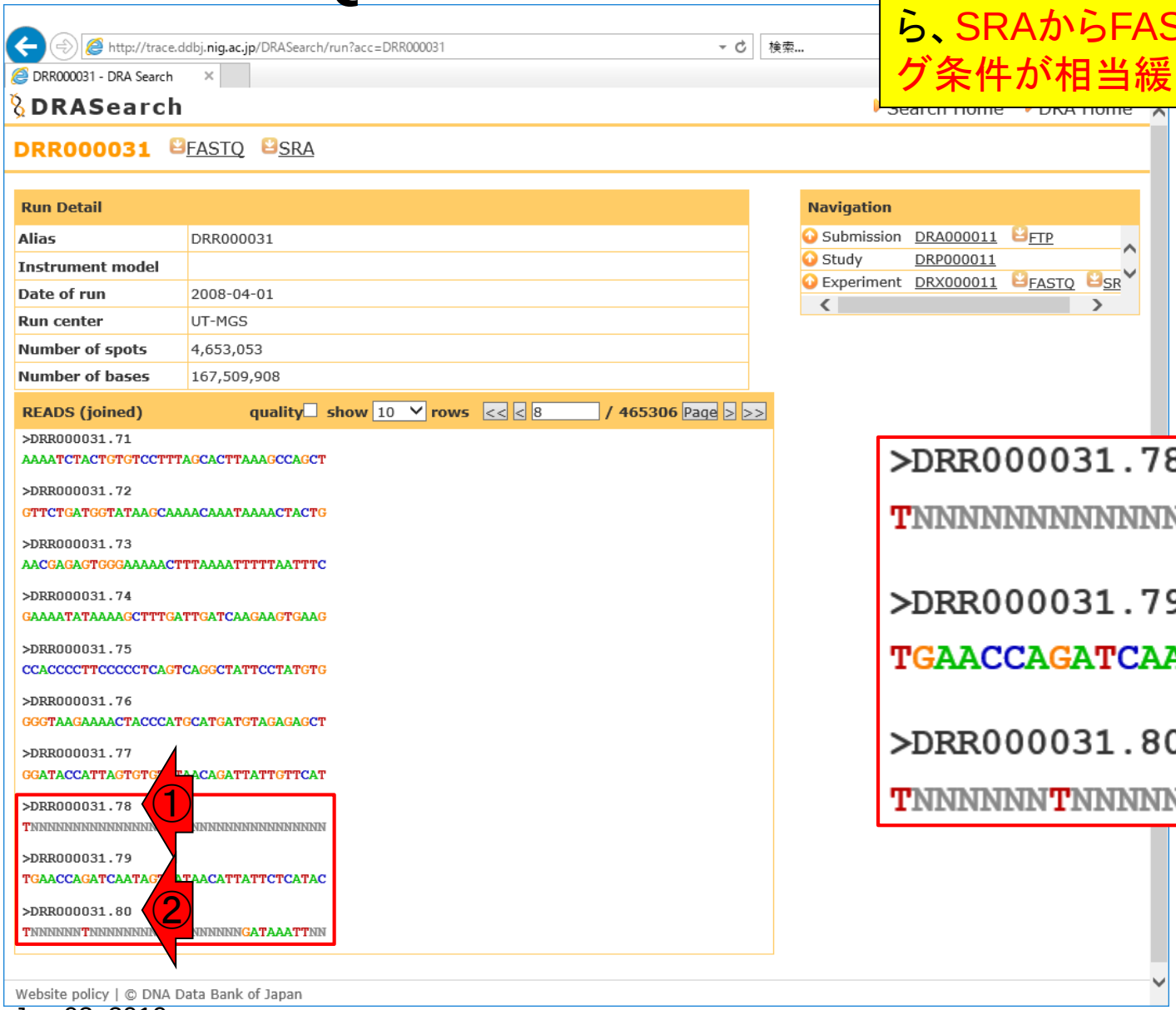
FASTQのリード数

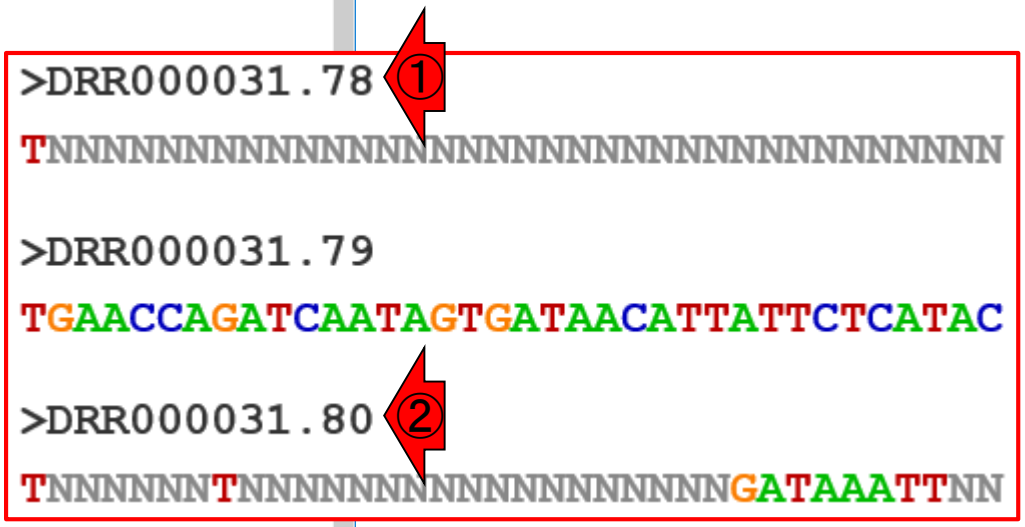
①をガスガス押して、②のところがリード番号78を含むところまで移動できました



FASTQのリード数

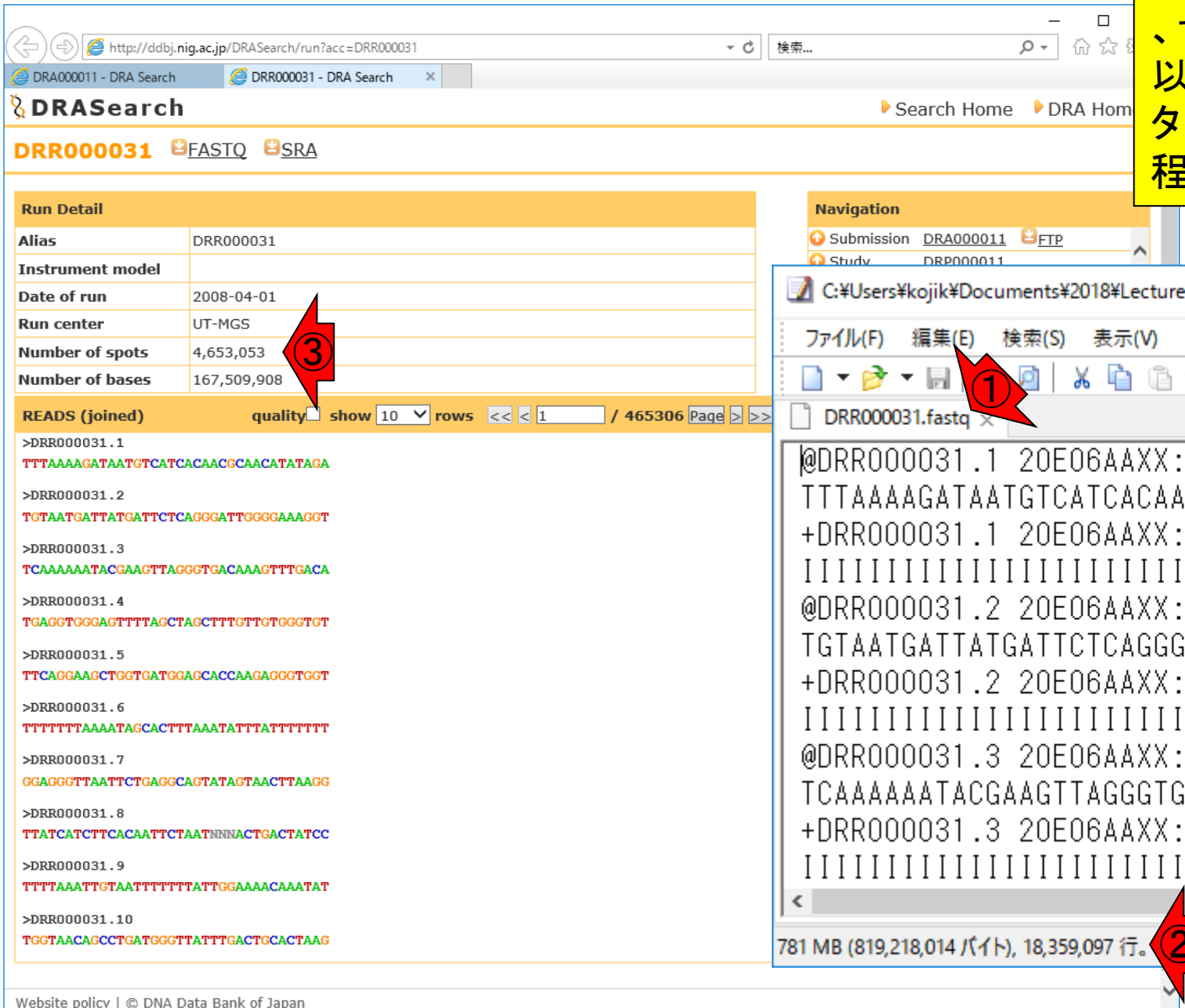
①リード番号78はNだらけです。確かに使い物にならないリードであることが分かります。その一方で、②リード番号80が生き残っていることから、SRAからFASTQを作成する際のフィルタリング条件が相当緩いのだろうということがわかる





これはIlluminaデータの話

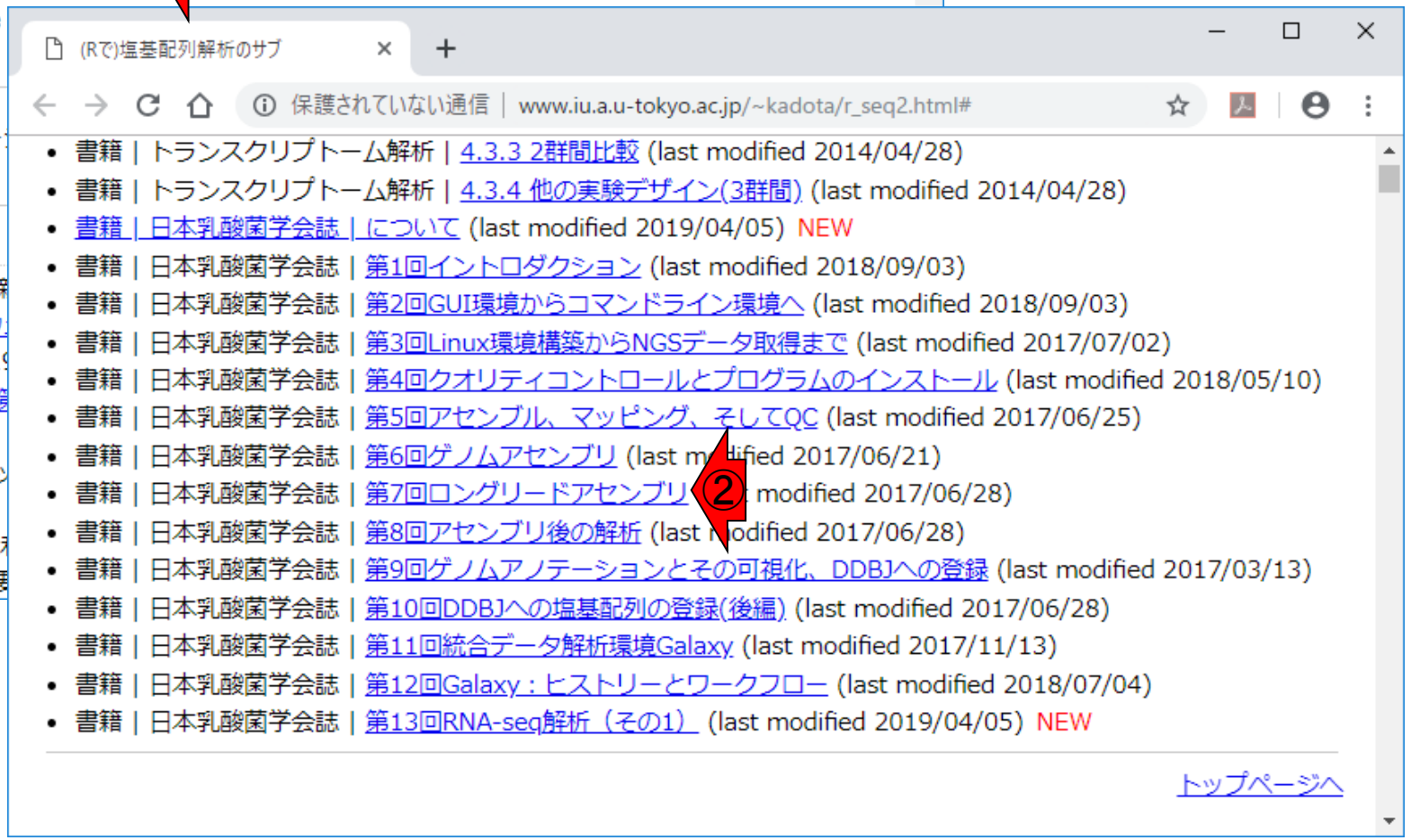
ウェブ上のリード数（4,653,053）とFASTQファイル中のリード数（4,589,774）の違いはあるものの、せいぜい総リード数の数%程度以内の違いです。Illuminaのデータはクオリティが高いので、数%程度しかfilter outされないのです



参考

ロングリードデータの場合

①サブのほうです。②はPacBioのゲノムデータの話ですが、Illuminaとはデータの性質も異なるため、アプローチも大きく異なる。例えば、公共DBからFASTQファイルをダウンロードする意味はほぼないなど…。詳細については②をご覧ください



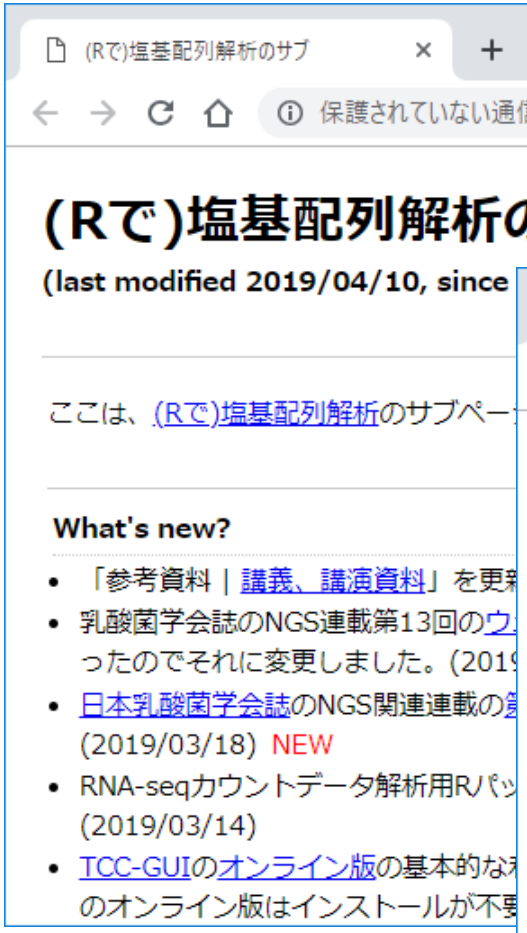
Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control (QC)
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

参考

手軽な解析手段

Linuxコマンドを覚える必要がなく、ウェブ上で大容量メモリをタダで使わせてもらって*de novo*アセンブリやマッピングプログラムを利用可能なウェブツールは存在します。①NGS連載第6回後半と第7回ではDDBJ Pipelineを紹介しました。しかし2019年2月をもってサービスが終了しました。代替ツールとして、例えば前回（2019年5月27日）の講義でも述べたMaserが利用可能です。

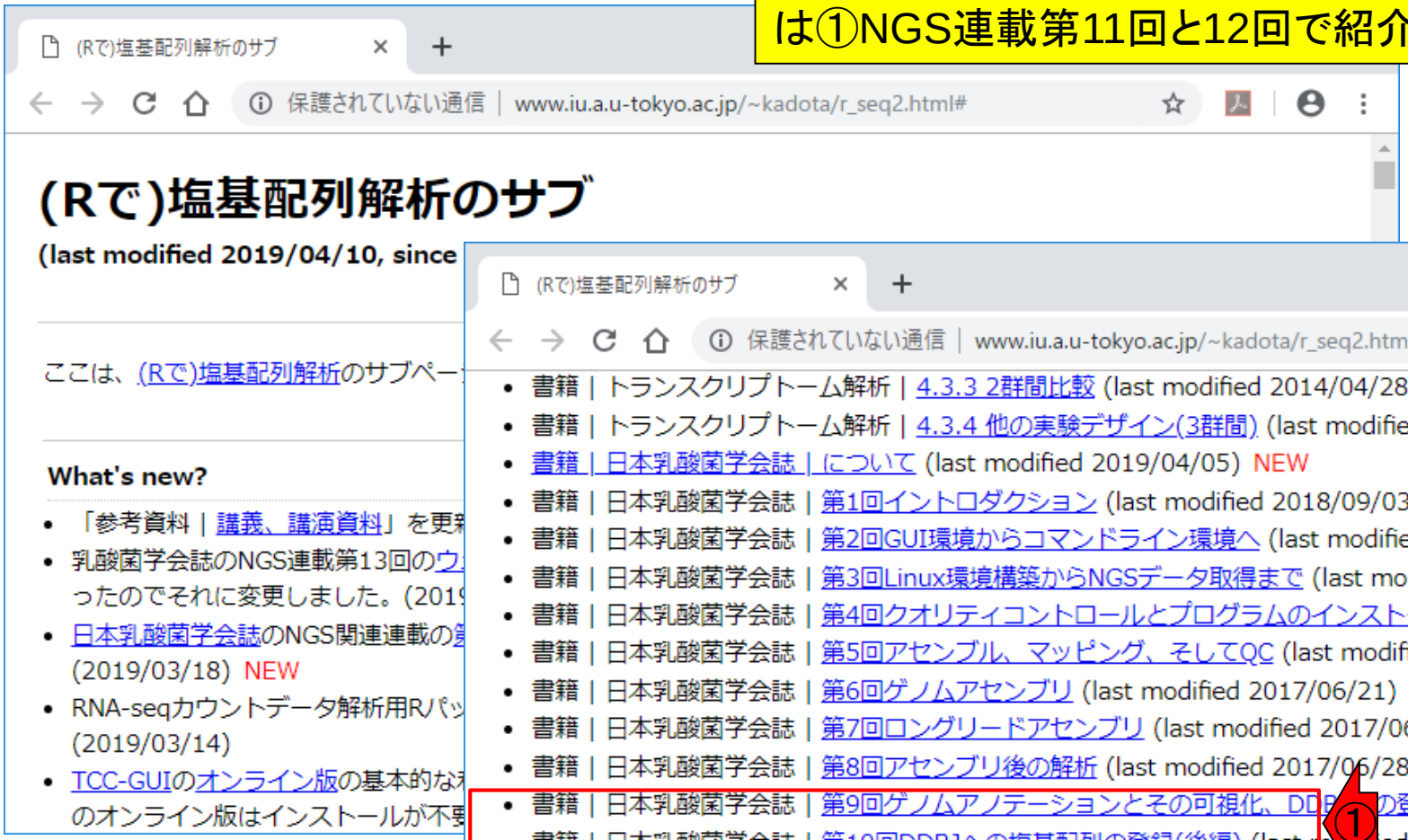


DDBJ Pipeline (Nagasaki et al., DNA Res., 20: 383–390, 2013)

参考

手軽な解析手段

Linuxコマンドを覚える必要がなく、ウェブ上で大容量メモリをタダで使わせてもらってde novoアセンブリやマッピングプログラムを利用可能なウェブツールは他にもあります。それは①NGS連載第11回と12回で紹介している②Galaxyです



Galaxy(Giardine et al., Genome Res., 15(10): 1451–1455, 2005)

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control (QC)
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

公共DB

②EMBL-EBI ENA上で、DRR000031を眺める。
ENAはブラウザによっては見られないという話と、様々なID間の関係性が分かりやすいという話です。①から辿らずに「ENA EBI」で直接ググってもよい

(Rで)塩基配列解析

(last modified 2019/05/24, since 2010)

このウェブページの
Macintosh2018.1
ています。 初心者
2018年7月に(Rて
(2018/07/18)

What's new? (
- 「解析 | 発現変
- 「正規化 | サン

- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2019/02/01)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu, 2013) (last modified 2019/02/01)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG

①

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから

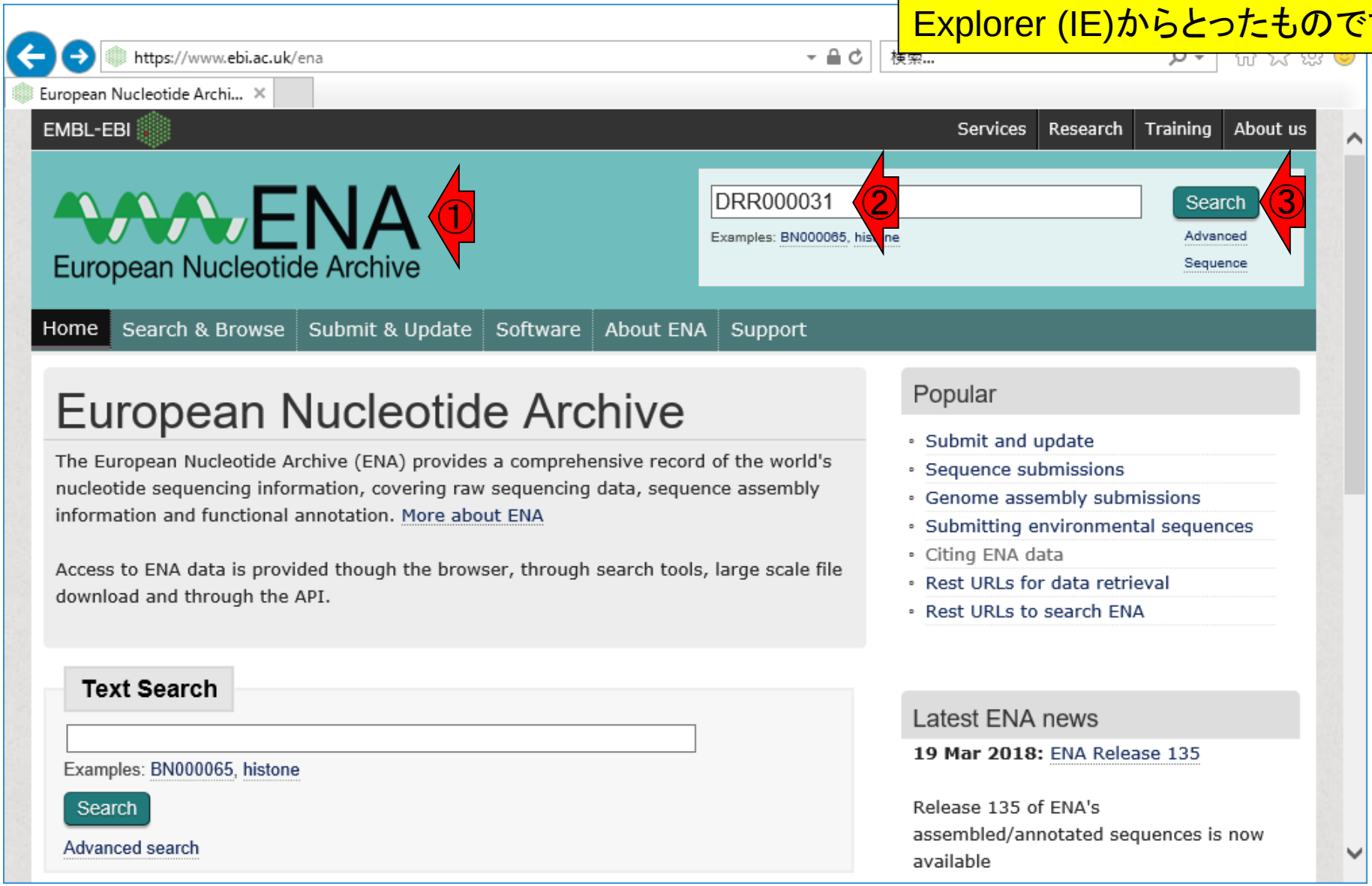
次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。 マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。 メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、 操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、 submitterが提供してくれている場合は（まだまだ少ないようですが）リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。
2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。
データの形式は基本的にSanger typeのFASTQ形式です。 FASTA形式はリードあたり二行（idの行と配列の行）で表現します。 FASTQ形式はリードあたり4行（@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行）で表現します。 FASTQ形式は、Sangerのものがデファクトスタンダード（業界標準）です。 かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、 NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

- DDBJ Sequence Read Archive (DRA) : Kodama et al., Nucleic Acids Res., 2018
- EMBL-EBI European Nucleotide Archive (ENA) : Leinonen et al., Nucleic Acids Res., 2017
- NCBI Sequence Read Archive (SRA) : Sayers et al., Nucleic Acids Res., 2019
- ArrayExpress : Kolesnikov et al., Nucleic Acids Res., 2015
- GEO : Clough and Barrett, Methods Mol Biol., 2016
- DBCLS SRA : Nakazato et al., PLoS One, 2013

②

ENAをブラウザIEでみる

①EMBL-EBI ENA上で、②DRR000031を眺めるべく、③Search。このスクリーンショットはウェブブラウザInternet Explorer (IE)からとったものです

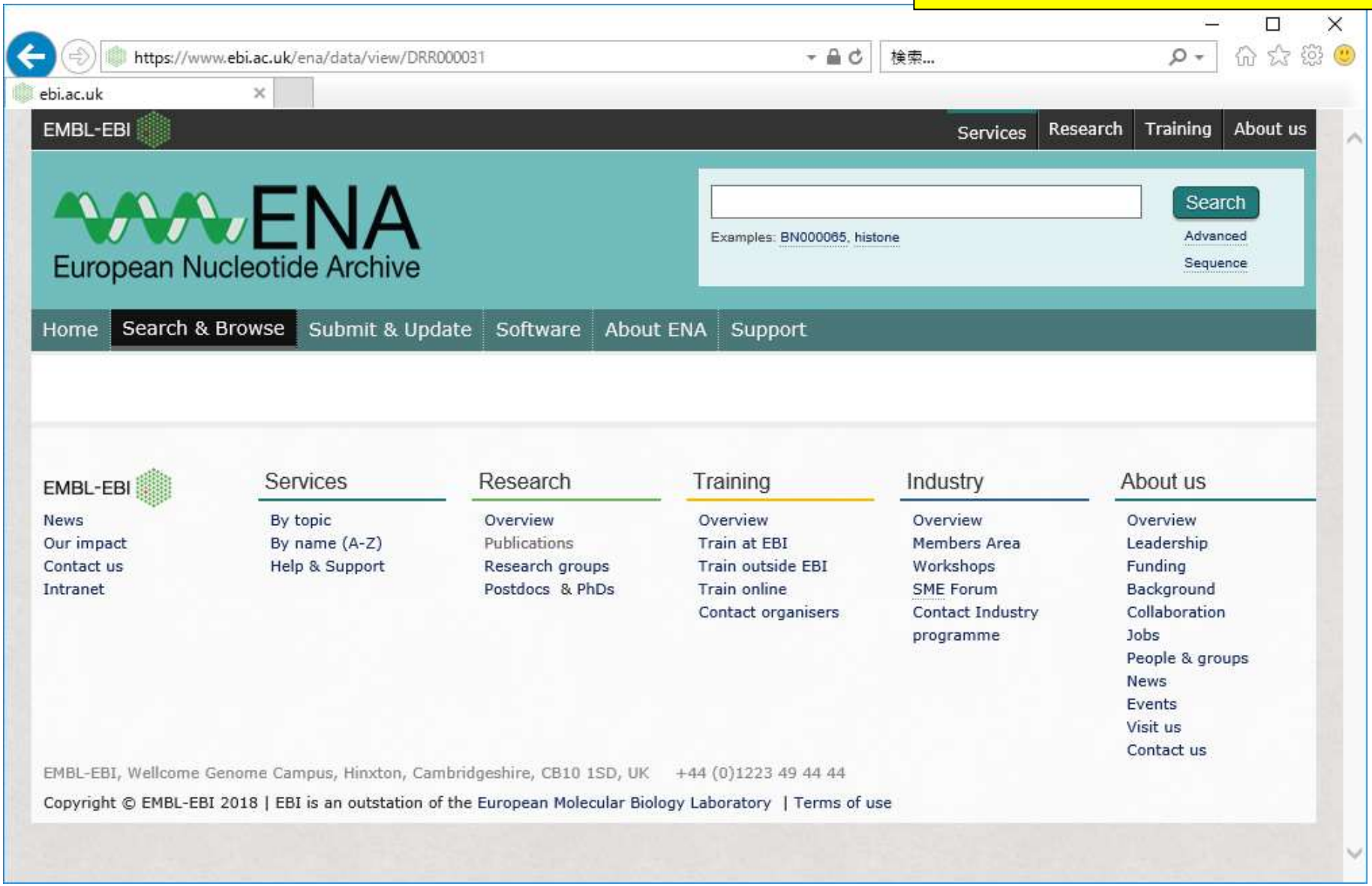


ENAをブラウザIEでみる

こんな感じになります。①でも②でもどちらでもいいが、とりあえず目的の②を押す

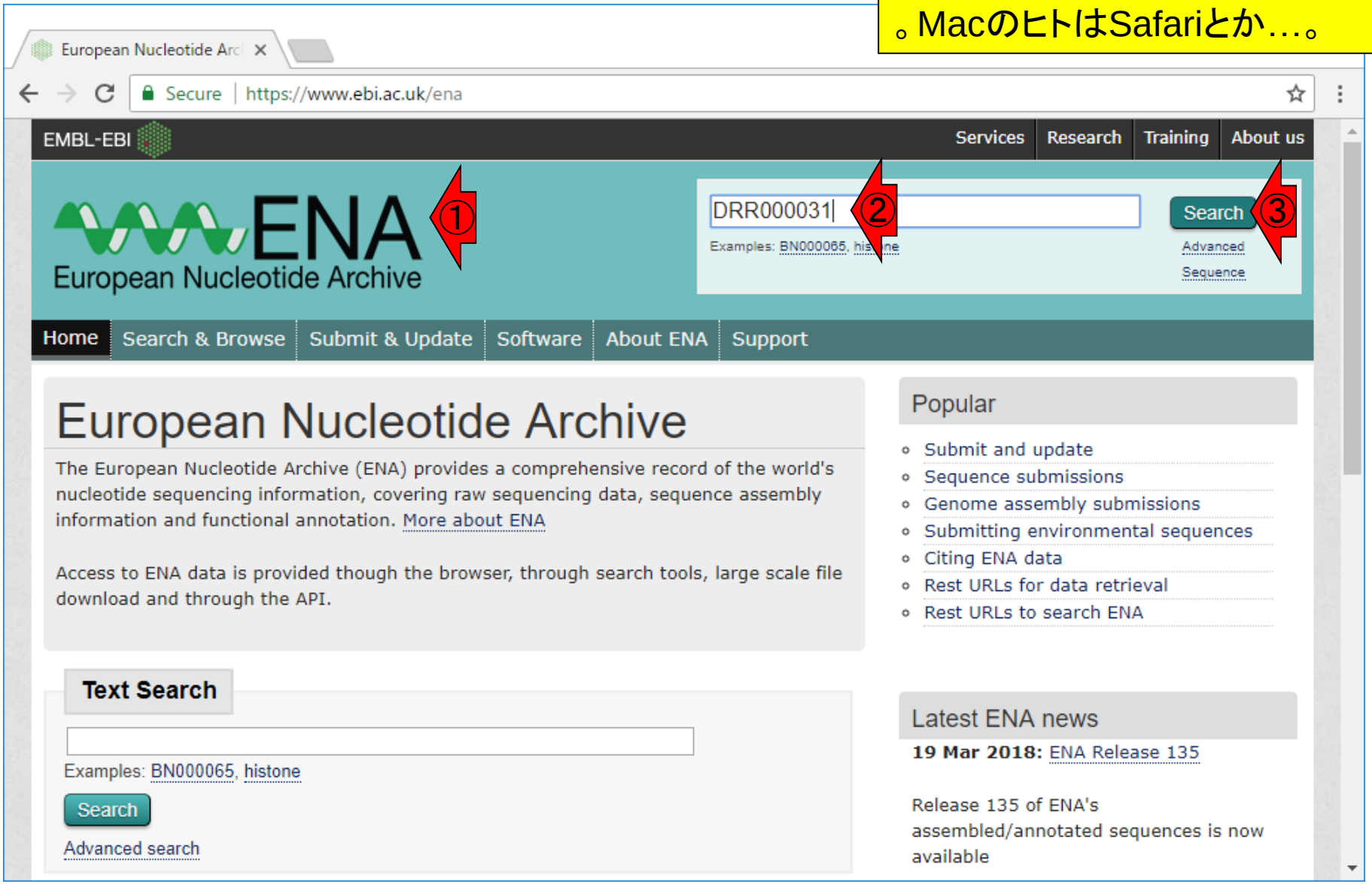
ENAをブラウザIEでみる

こんな感じになって、まともに表示されません。結論としては、ENA利用時はIEをブラウザとして使ってはいけない、です



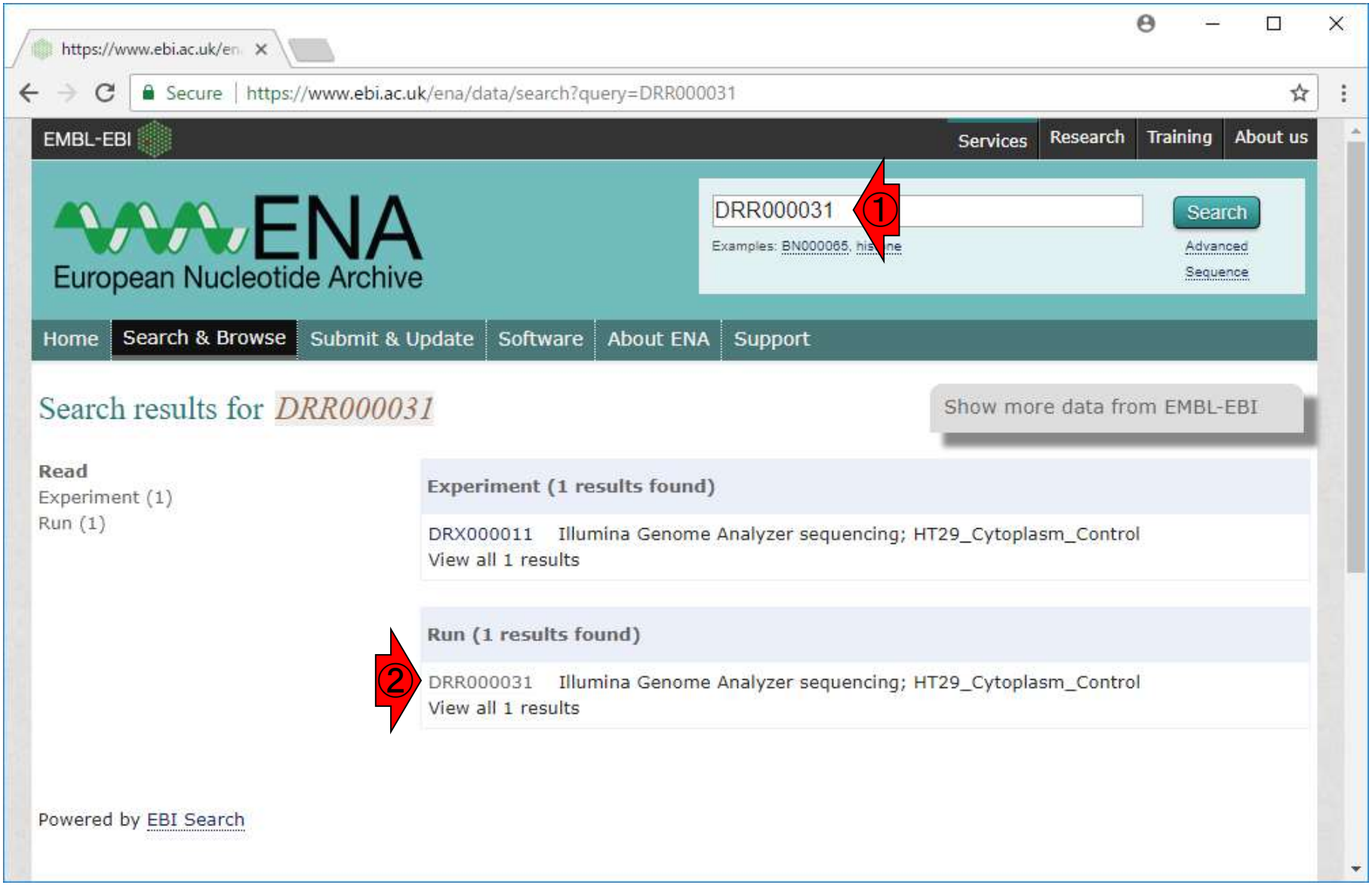
ENAをChromeでみる

①EMBL-EBI ENA上で、②DRR000031を眺めるべく、③Search。Chrome以外のブラウザがあれば是非試してみてください。MacのヒトはSafariとか…。



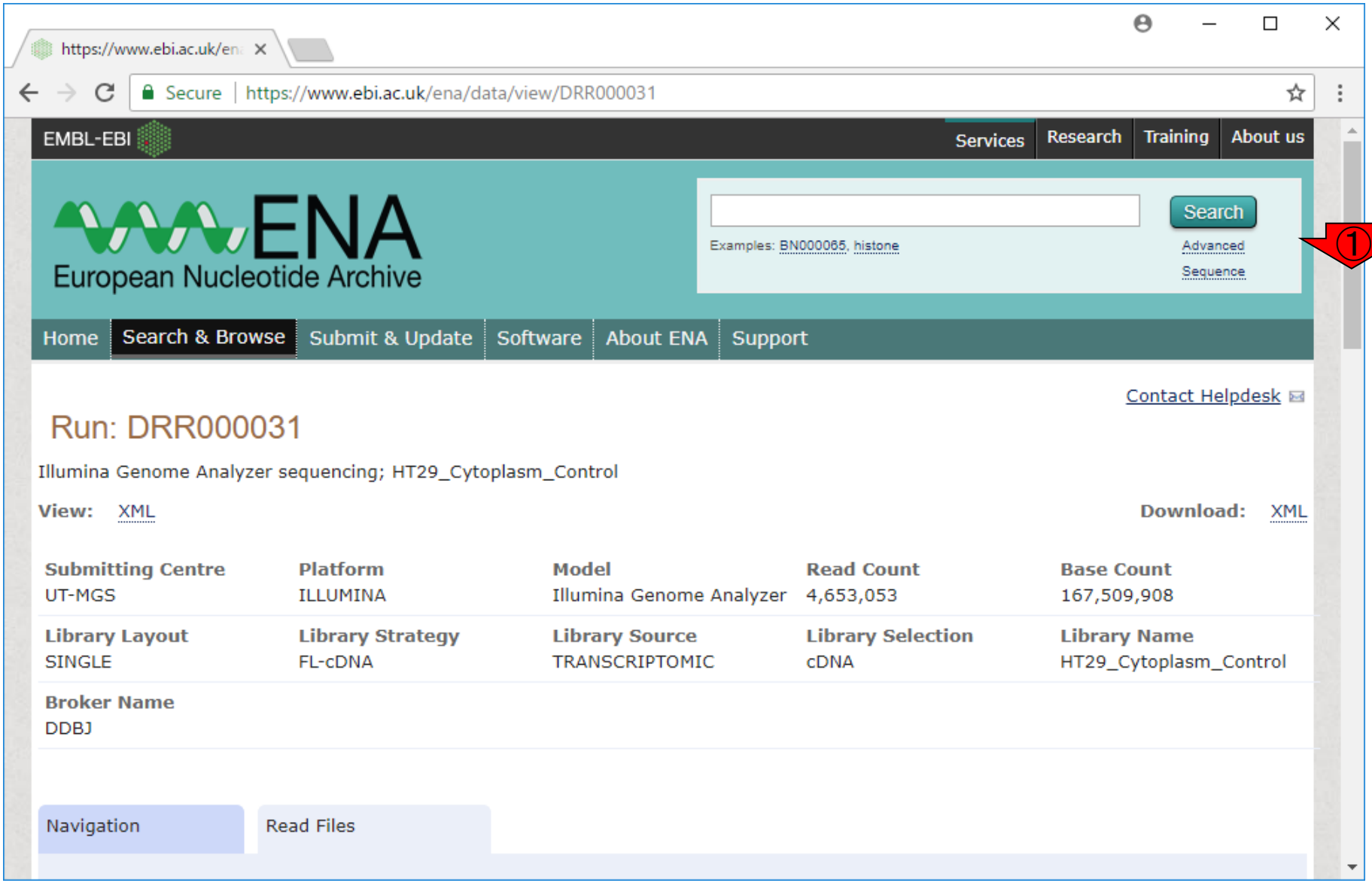
ENAをChromeでみる

ここまではIEと同じです。①目的のDRR000031と同じ、②を押すと…



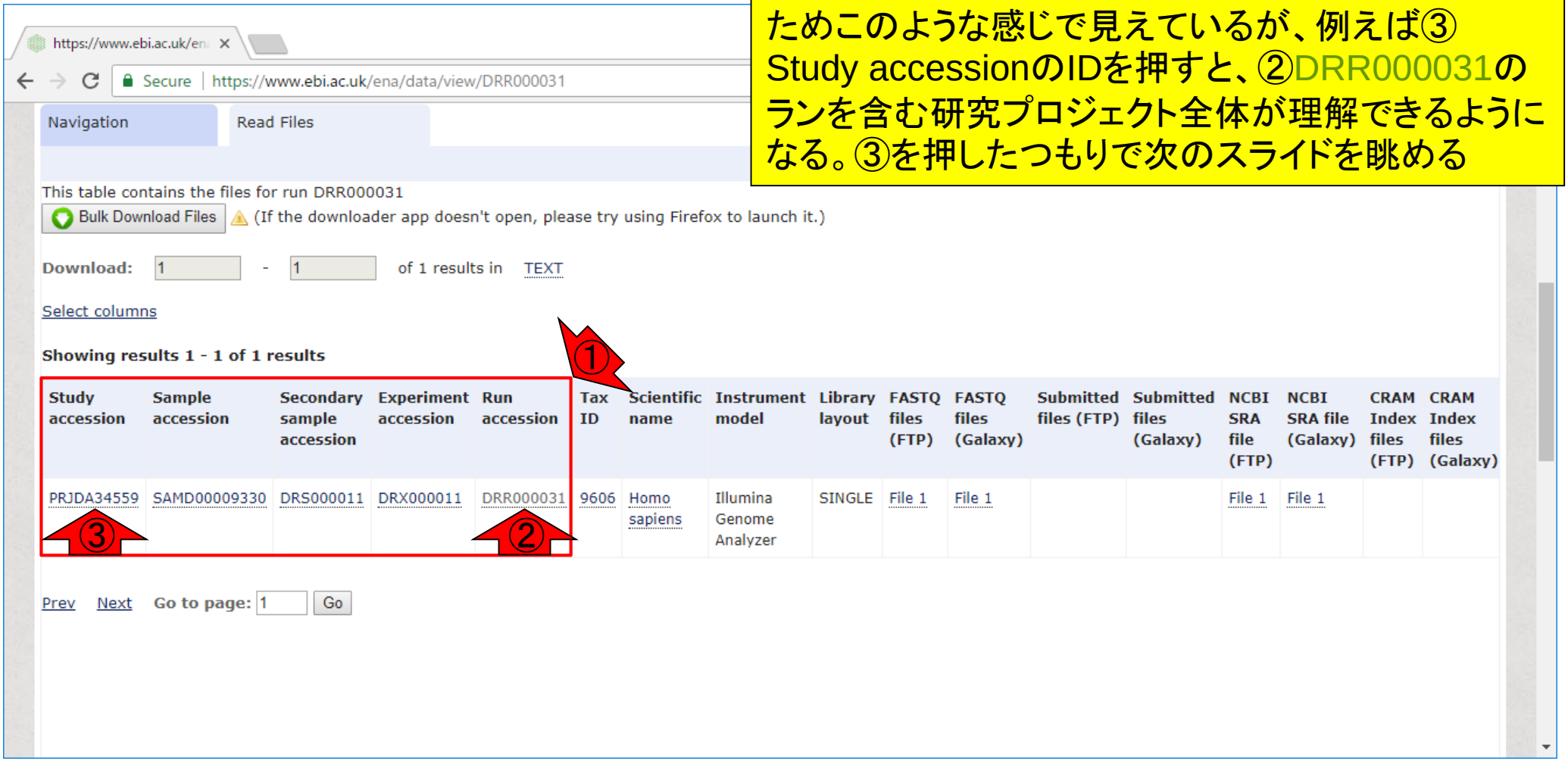
ENAをChromeでみる

IEとは違った結果になることがわかります。①ページ下部に移動



ENAをChromeでみ

こんな感じに見えます。全体的に小さいため見辛いですが…①赤枠のあたりが様々なIDの関係性が一望できるところ。今回は、②DRR000031で検索したためこのような感じで見えているが、例えば③Study accessionのIDを押すと、②DRR000031のランを含む研究プロジェクト全体が理解できるようになる。③を押したつもりで次のスライドを眺める

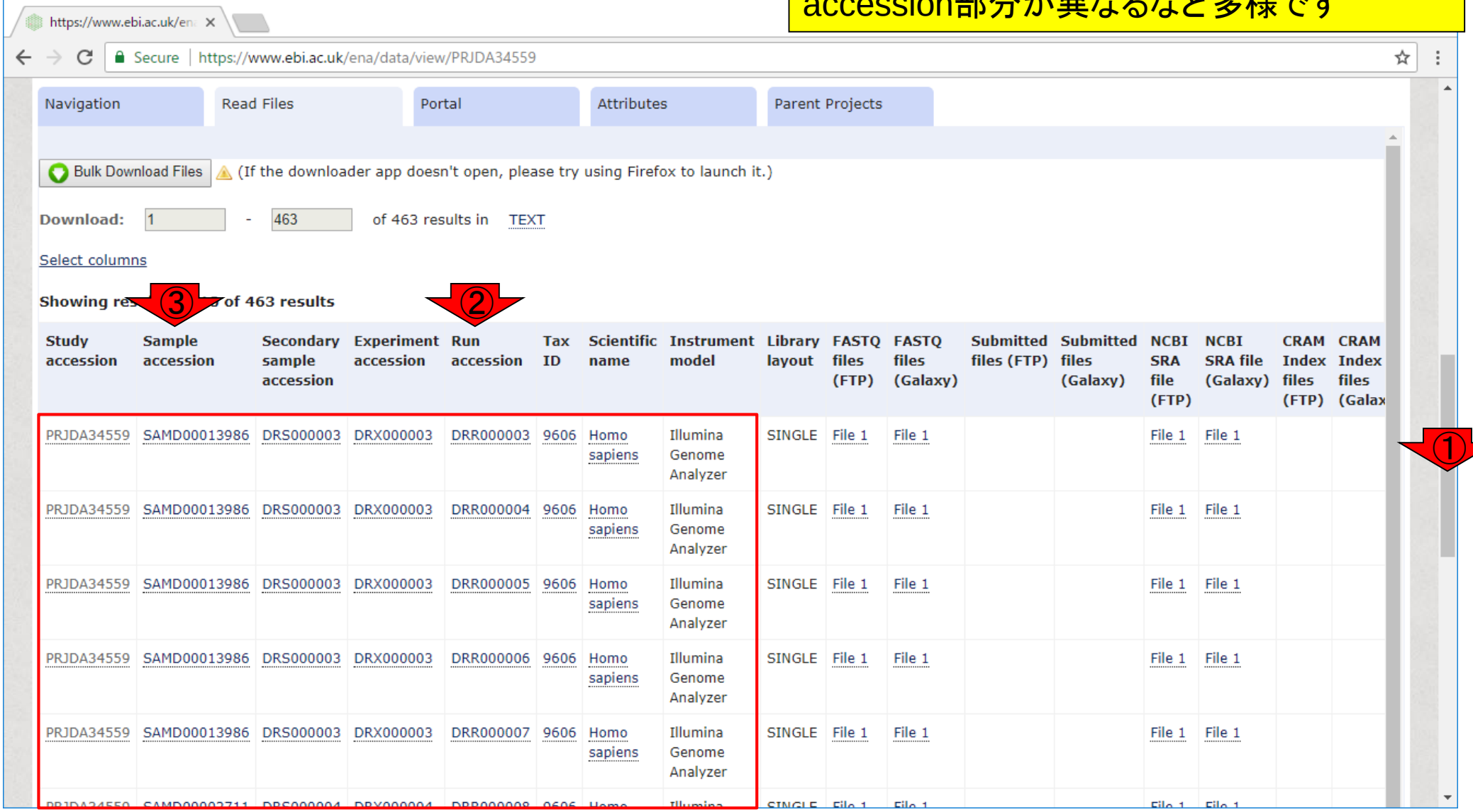


ENAをChromeでみ

①PRJDA34559は、②の組織によって行われた③ヒトのトランスクリプトームプロジェクト全体のIDである。今回の対象であるDRR000031もこの枠組みで得られたデータである。④ページ下部に移動

ENAをChromeでみる

①このあたりまで移動。この画面上の赤枠内では、数値が異なるのが②Run accession部分のみであるが、ものによっては③Sample accession部分が異なるなど多様です



Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galax
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000003	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

ENAをChromeでみ

①が用いたNGS機器情報。通常、②はsingle-endデータであることを表すSINGLEか、paired-endデータであることを表すPAIREDかのどちらかです

https://www.ebi.ac.uk/en

Secure | https://www.ebi.ac.uk/ena/data/view/PRJDA34559

Navigation | Read Files | Portal | Attributes | Parent Projects

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 463 of 463 results in TEXT

Select columns

Showing results 1 - 10 of 463 results

①　②

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galax
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000003	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

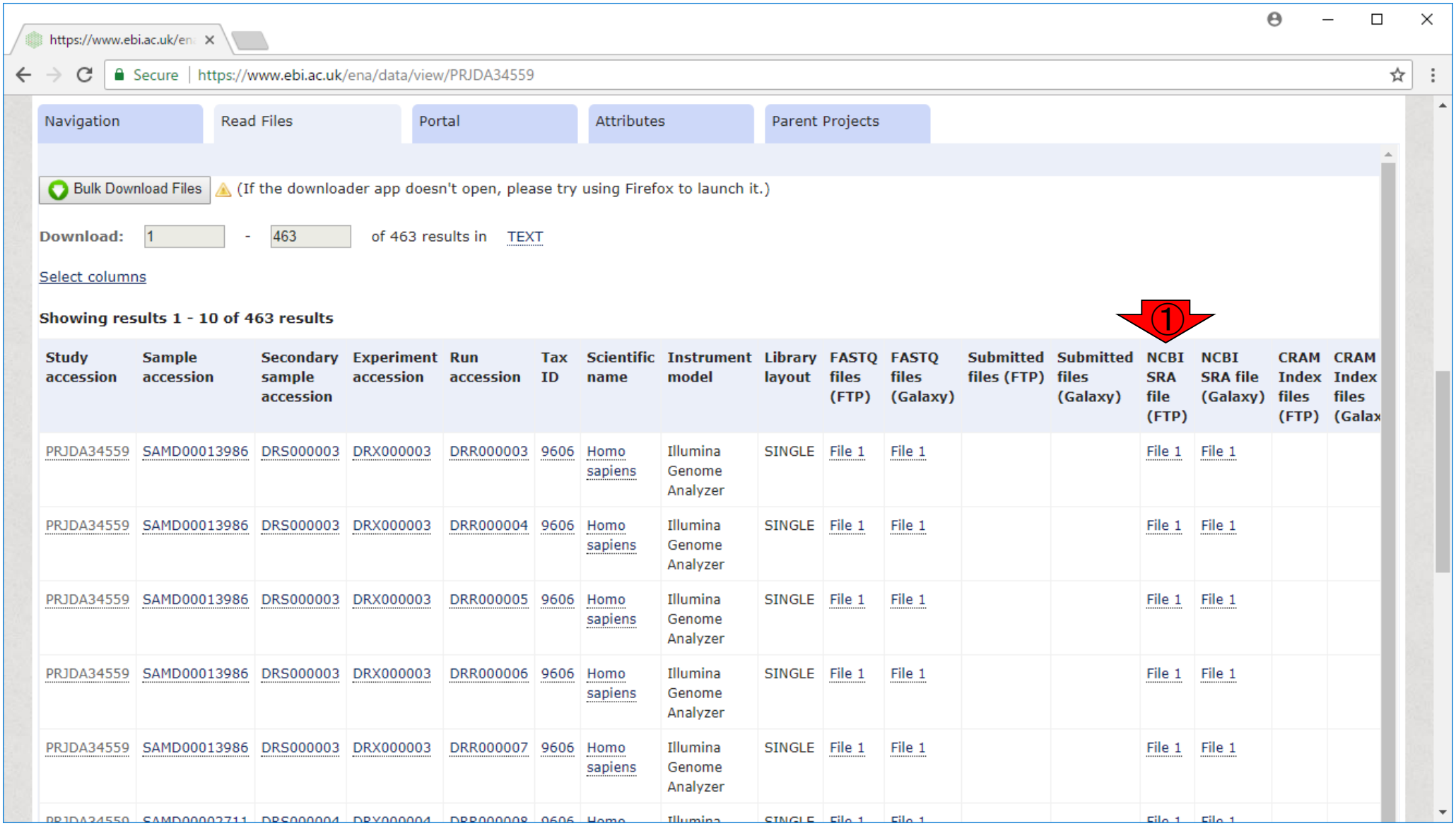
ENAをChromeでみ

①はFASTQファイルをダウンロードするところ。このデータはsingle-endなので、ダウンロードのリンク先としてFile 1のみが提供されている。paired-endの場合だとFile 1直下にFile 2が見られます

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galax
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000003	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

①SRAファイルをダウンロードしたい場合はこちら

ENAをChromeでみる



Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galax
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000003	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

①は、ウェブツールGalaxyとの連携用

ENAをChromeでみる

https://www.ebi.ac.uk/en

Secure | https://www.ebi.ac.uk/ena/data/view/PRJDA34559

Navigation | Read Files | Portal | Attributes | Parent Projects

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 463 of 463 results in TEXT

Select columns

Showing results 1 - 10 of 463 results

①

①

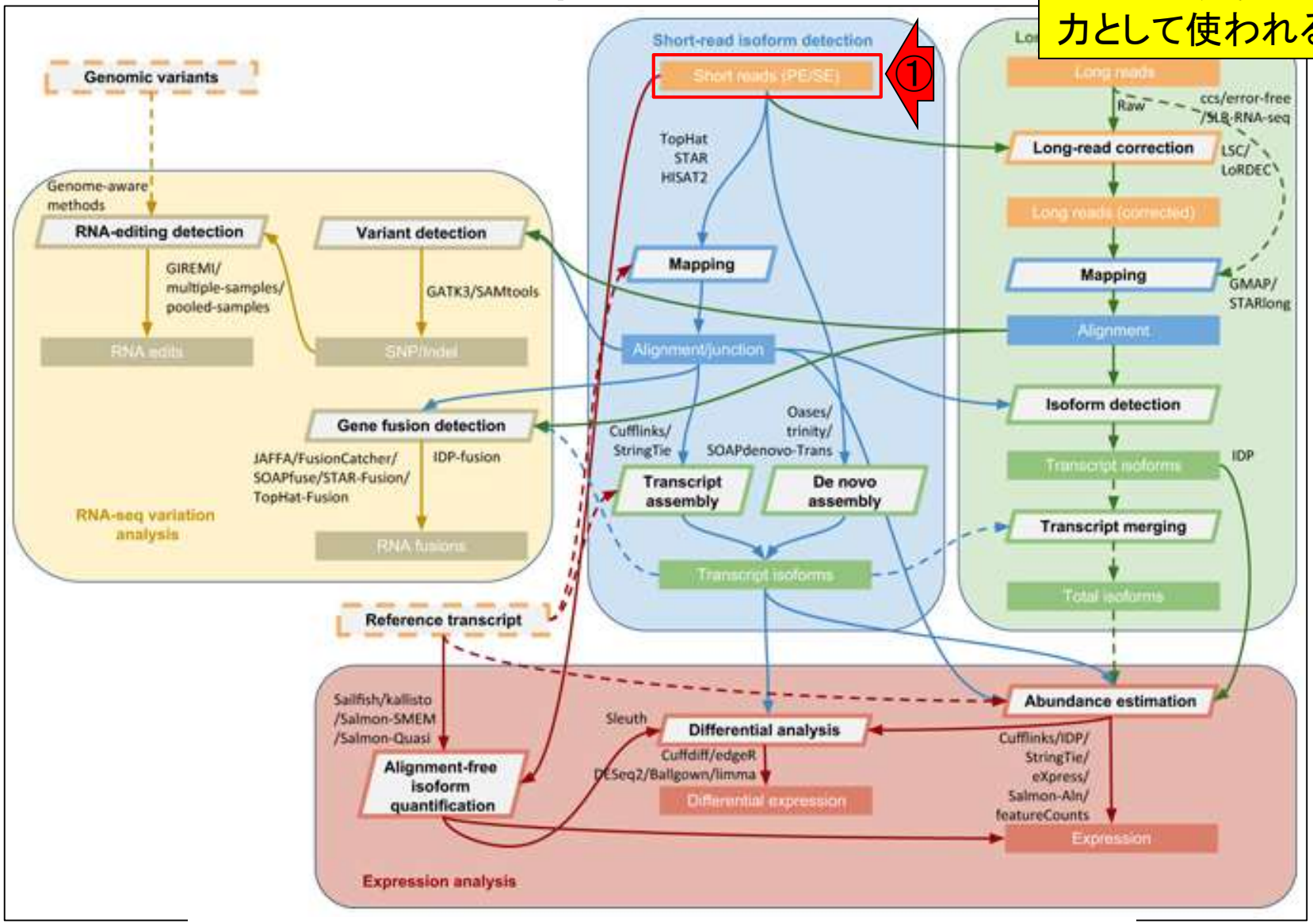
Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galax
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000003	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS000003	DRX000003	DRR000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control (QC)
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

RNACocktail論文の図1

①は前処理後のリードデータ。公共DBからダウンロード後のFASTQファイルが、そのままデータ解析の入力として使われることはほぼない



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

前処理（preprocessing）

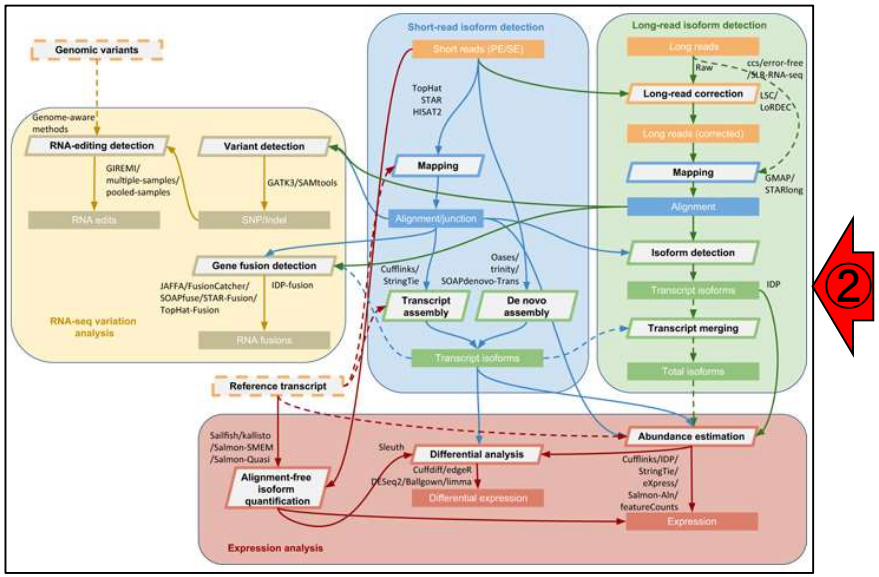
①この部分が、②これ以降の下流部分の解析（downstream analysis）に大きく影響するので重要です。赤枠部分は、マイクロアレイでは③前処理と呼ばれ、RNA-seqでは④QCと呼ばれることが多い

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

③ ④



②

前処理（preproce

一言でQCとはいっても、内部で行う処理は様々。①rRNAやtRNAは、通常RNA-seqの解析対象外。②は、転写物ごとに発現量が異なるという事実をゲノムアセンブリ側から述べた表現。ゲノム配列の場合はcoverageがほぼ一定であるのに対し、高発現転写物のcoverageは非常に高く、低発現の場合はcoverageが低くなるということ。

NGSリードデータ（SRAファイル）

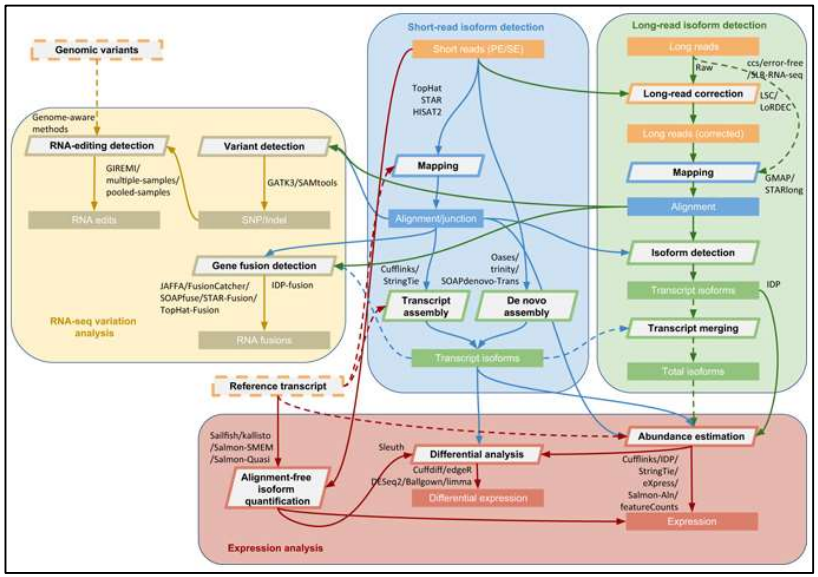
⇩

NGSリードデータ（FASTQファイル）

⇩

前処理（preprocessing） or Quality Control (QC)

⇩



～NGSデータ**共通**の事柄～
- クオリティの低いリード
- 解析対象外の生物種に由来するリード
- アダプター配列

～RNA-seqデータ**特有**の事柄～
- リボソームRNA（rRNA）やtRNAの混入 ①
- RNA degradation（RNA分解）
- 多様なリードカバレッジ（varied read coverage） ②

RNA-QC-chain

①RNA-QC-chainという比較的最近のプログラムの原著論文（のAbstract）から最近の動向を知る。まず、②QCは今でも必須であることがわかる

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications based on next-generation sequencing technology. However, raw RNA-Seq data may have quality issues, which can significantly distort analytical results and lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous quality control (QC) procedures before downstream analysis. Currently, an accurate and complete QC of RNA-Seq data requires of a suite of different QC tools used consecutively, which is inefficient in terms of usability, running time, file usage, and interpretability of the results.

RESULTS: We developed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RNA-QC-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNAs) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such as read number, alignment coverage, sequencing depth and pair-end read mapping information). This package was developed based on our previously reported tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with extensions specifically designed for RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data, such as RNA sequence trimming, automatic rRNA detection and automatic contaminating species identification. The three QC steps can run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive package with high flexibility and usability. Moreover, parallel computing and optimizations are embedded in most of the QC procedures, providing a superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

②

①

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

RNA-QC-chain

①RNA-QC-chainプログラム開発のモチベーションに関する記載。現状ではRNA-seqデータのQC作業は複数のQCプログラムを使わないといけないので大変だ。なので②早くて使いやすいツールを開発したよ、という論文

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications based on next-generation sequencing technology. However, raw RNA-Seq data may have quality issues, which can significantly distort analytical results and lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous quality control (QC) procedures before downstream analysis. Currently, an accurate and complete QC of RNA-Seq data requires of a suite of different QC tools used consecutively, which is inefficient in terms of usability, running time, file usage, and interpretability of the results.

RESULTS: We developed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RNA-QC-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNA) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such as read number, alignment coverage, sequencing depth and pair-end read mapping information). This package was developed based on our previously reported tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with extensions specifically designed for RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data, such as RNA sequence trimming, automatic rRNA detection and automatic contaminating species identification. The three QC steps can run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive package with high flexibility and usability. Moreover, parallel computing and optimizations are embedded in most of the QC procedures, providing a superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

RNA-QC-chain

①の部分が②の事柄に相当し、
③の部分が④の事柄に相当する

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications
technology. However, raw RNA-Seq data may have quality issues, which can sign
lead to erroneous conclusions. Therefore, the raw data must be subjected to vigoro
before downstream analysis. Currently, an accurate and complete QC of RNA-Se
tools used consecutively, which is inefficient in terms of usability, running time, file

RESULTS: We d ed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RNA-QC-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNAs) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such as read number, alignment coverage, sequencing depth and pair-end read mapping information). This package was developed based on our previously reported tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with extensions specifically designed for RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data, such as RNA sequence trimming, automatic rRNA detection and automatic contaminating species identification. The three QC steps can run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive package with high flexibility and usability. Moreover, parallel computing and optimizations are embedded in most of the QC procedures, providing a superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

①

②

②

③

④

～NGSデータ**共通**の事柄～
・クオリティの低いリード
・解析対象外の生物種に由来するリード
・アダプター配列

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

①の部分が②の事柄に相当する

RNA-QC-chain

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications bas
technology. However, raw RNA-Seq data may have quality issues, which can significa
lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous
before downstream analysis. Currently, an accurate and complete QC of RNA-Seq da
tools used consecutively, which is inefficient in terms of usability, running time, file usa

RESULTS: We developed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RN C-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNAs) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such a sequencing depth and pair-end read mapping information). This package was develop tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with RNA-Seq data. It has several features that are not available yet in other QC tools for R sequence trimming, automatic rRNA detection and automatic contaminating species i run either sequentially or independently, enabling RNA-QC-Chain as a comprehensiv usability. Moreover, parallel computing and optimizations are embedded in most of th superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

～NGSデータ**共通**の事柄～
- クオリティの低いリード
- 解析対象外の生物種に由来するリード
- アダプター配列

①

～RNA-seqデータ**特有**の事柄～
- リボソームRNA（rRNA）やtRNAの混入 ②
- RNA degradation（RNA分解）
- 多様なリードカバレッジ（varied read coverage）

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

RNA-QC-chain

特に、例えば①rRNA由来リードや解析対象外生物種由来リードの同定は、②他のQC用プログラムには実装されてなかったそうです。こんな感じで自分のデータの前処理で必要そうなプログラムかどうかを判断し、必須の特徴をもつプログラムを七転八倒しながらインストールして利用するのが一般的

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control fo

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications based on next-generation sequencing technology. However, raw RNA-Seq data may have quality issues, which can significantly distort analytical results and lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous quality control (QC) procedures before downstream analysis. Currently, an accurate and complete QC of RNA-Seq data requires of a suite of different QC tools used consecutively, which is inefficient in terms of usability, running time, file usage, and interpretability of the results.

RESULTS: We developed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RNA-QC-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNAs) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such as read number, alignment coverage, sequencing depth and pair-end read mapping information). This package was developed based on our previously reported tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with extensions specifically designed for RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data, such as RNA sequence trimming, automatic rRNA detection and automatic contaminating species identification. These QC steps can run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive package with high flexibility and usability. Moreover, parallel computing and optimizations are embedded in most of the QC procedures, providing a superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

full textが見られる場所は、①ココ

RNA-QC-chain

BMC Genomics. 2018 Feb 14;19(1):144. doi: 10.1186/s12864-018-4503-6.

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data.

Zhou Q[1], Su X[2,3], Jing G[2], Chen S[4], Ning K[5].

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications based on next-generation sequencing technology. However, raw RNA-Seq data may have quality issues, which can significantly distort analytical results and lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous quality control (QC) procedures before downstream analysis. Currently, an accurate and complete QC of RNA-Seq data requires of a suite of different QC tools used consecutively, which is inefficient in terms of usability, running time, file usage, and interpretability of the results.

RESULTS: We developed a comprehensive, fast and easy-to-use QC pipeline for RNA-Seq data, RNA-QC-Chain, which involves three steps: (1) sequencing-quality assessment and trimming; (2) internal (ribosomal RNAs) and external (reads from foreign species) contamination filtering; (3) alignment statistics reporting (such as read number, alignment coverage, sequencing depth and pair-end read mapping information). This package was developed based on our previously reported tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with extensions specifically designed for RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data, such as RNA sequence trimming, automatic rRNA detection and automatic contaminating species identification. The three QC steps can run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive package with high flexibility and usability. Moreover, parallel computing and optimizations are embedded in most of the QC procedures, providing a superior efficiency. The performance of RNA-QC-Chain has been evaluated with different types of datasets, including an in-house sequencing data, a semi-simulated data, and two real datasets downloaded from public database. Comparisons of RNA-QC-Chain with other QC tools have manifested its superiorities in both function versatility and processing speed.

CONCLUSIONS: We present here a tool, RNA-QC-Chain, which can be used to comprehensively resolve the quality control processes of RNA-Seq data effectively and efficiently.

KEYWORDS: Alignment statistics; Contamination identification; Parallel computing; Quality control; RNA-Seq

PMID: 29444661 PMCID: PMC5813

①

Read free full text at BMC

PMC FREE Full text

Save items

Add to Favorites

Similar articles

QC-Chain: fast and holistic quality control method for next-generati [PLoS One. 2013]

Meta-QC-Chain: comprehensive and fast qualit [Genomics Proteomics Bioinforma...]

Software for pre-processing Illumina next-generation s [Source Code Biol Med. 2014]

Review Standardization and quality management in [Appl Transl Genom. 2016]

Review Prevention, diagnosis and treatment of high-throughp [Mol Ecol. 2014]

See reviews...

See all...

Related information

Articles frequently viewed together

References for this PMC Article

Free in PMC

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

RNA-QC-chain

①ここからプログラムがおかれているサイトなどの情報を知る。Abstractの最後のほうにプログラムのURL情報が掲載されている場合もある。

Software | Open Access

RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data

Qian Zhou †, Xiaoquan Su †, Gongchao Jing, Songlin Chen and Kang Ning

†Contributed equally

BMC Genomics 2018 **19**:144

https://doi.org/10.1186/s12864-018-4503-6

Received: 21 September 2017 | Accepted: 28 January 2018 | Published: 14 February 2018

Abstract

Keywords

Background

Results and discussion

Conclusions

Availability and requirements

①

Download PDF

Export citations

Section

Transcriptomic methods

Metrics

Article accesses: 1112

Citations: 0
more information

Altmetric Attention Score: 11

Share This Article

See Updates

Check for updates

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

RNA-QC-chain

Publisherのサイト上でFull textが見られるページ上の、①Availabilityに関する箇所。②動作環境はUNIX/Linux。③プログラミング言語はC++。もしプログラミング言語がJavaと書かれていたら、インストールで失敗することはほぼない。

Availability and requirements ①

Project name: RNA-QC-Chain

Project home page: http://bioinfo.single-cell.cn/rna-qc-chain.html or http://124.16.150.212/rna-qc-chain.html

Operating system(s): Unix/Linux ②

Programming language: C++ ③

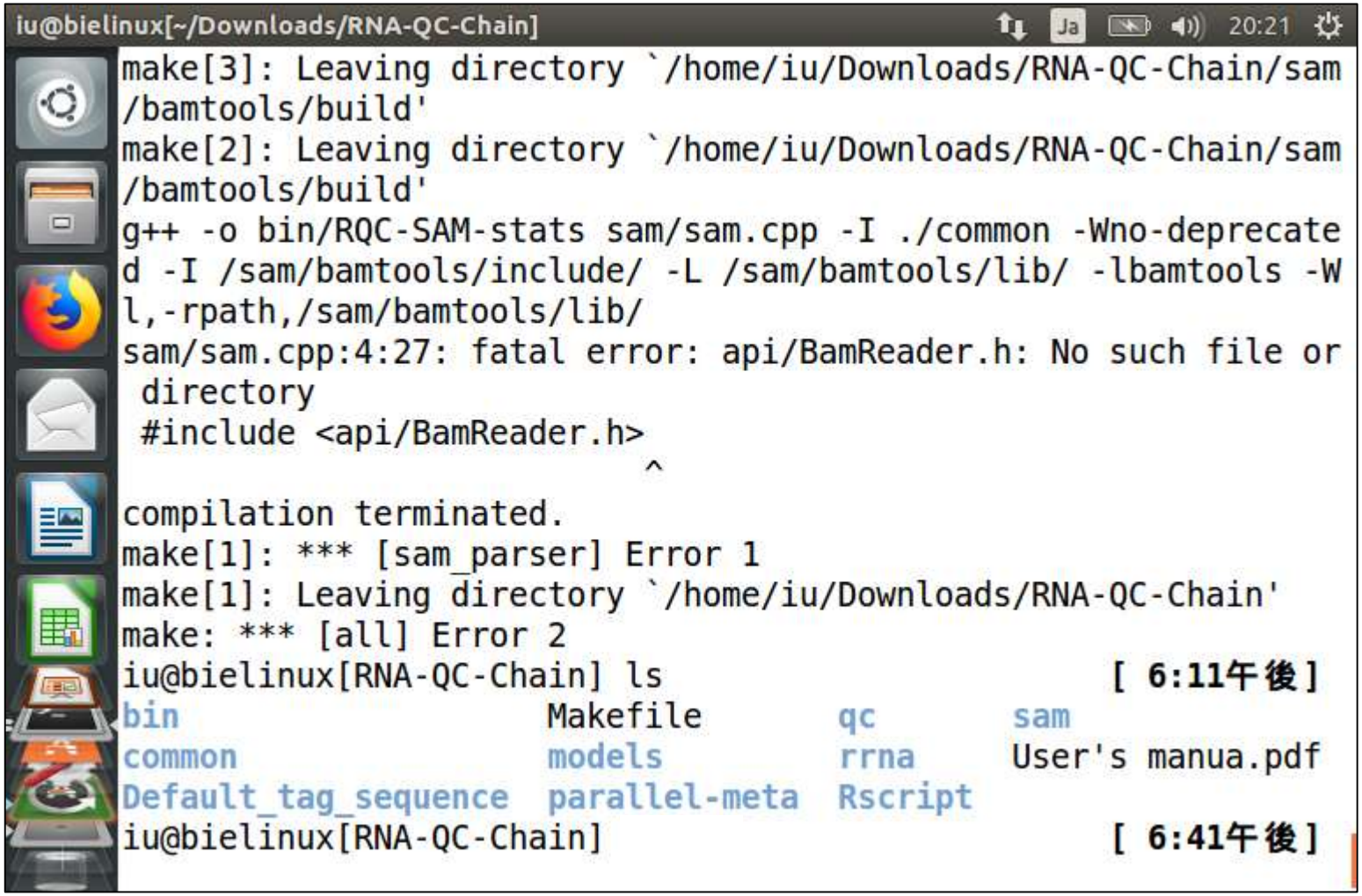
License: GPL-3

Availability: RNA-QC-Chain, including source code, documentation, and examples, is freely available for non-commercial use with no restrictions at http://bioinfo.single-cell.cn/rna-qc-chain.html or http://124.16.150.212/rna-qc-chain.html

RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

インストール失敗例

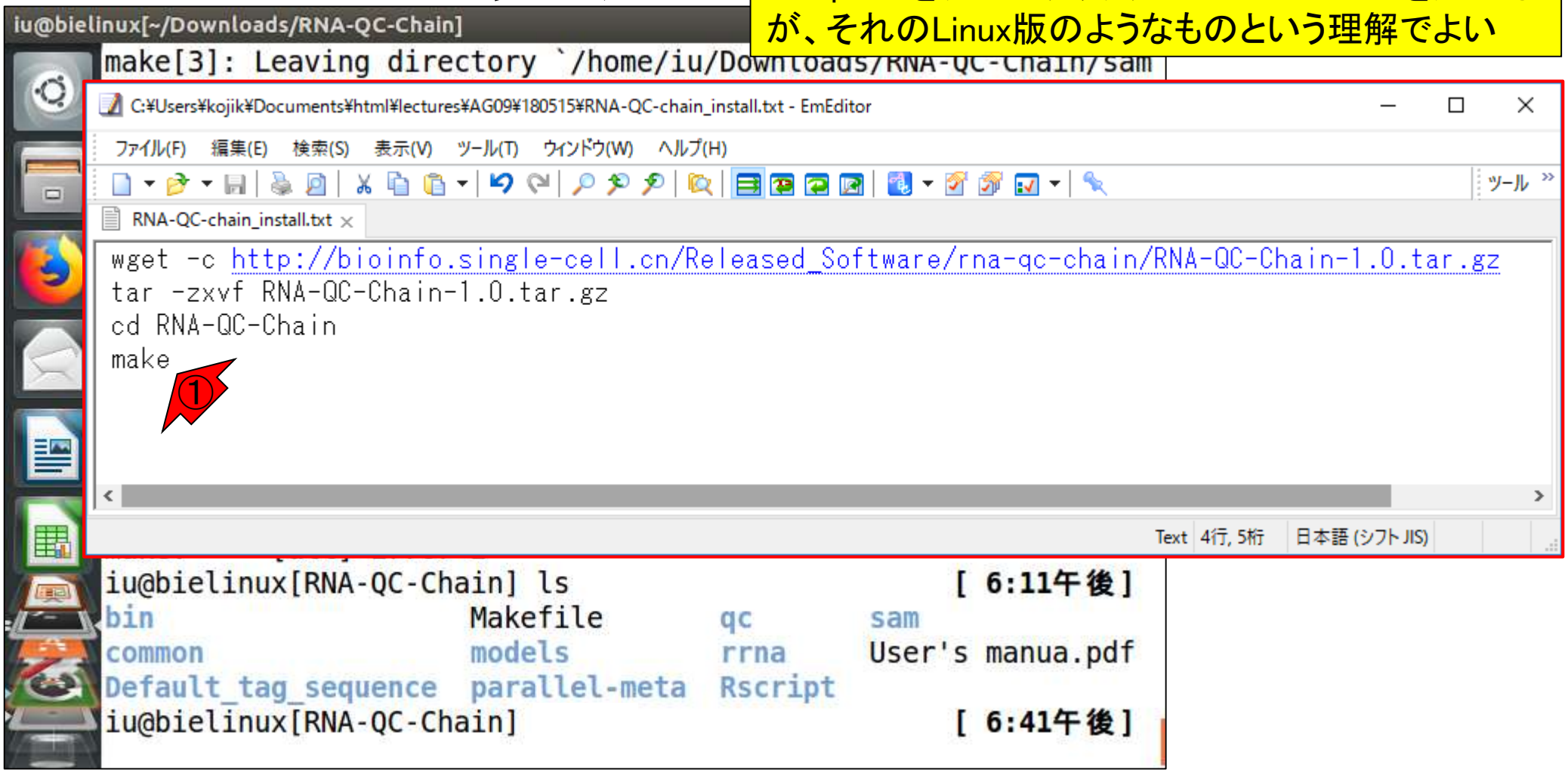
RNA-QC-chainをBio-Linux環境でインストールを試みて失敗したところ



RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

インストール失敗例

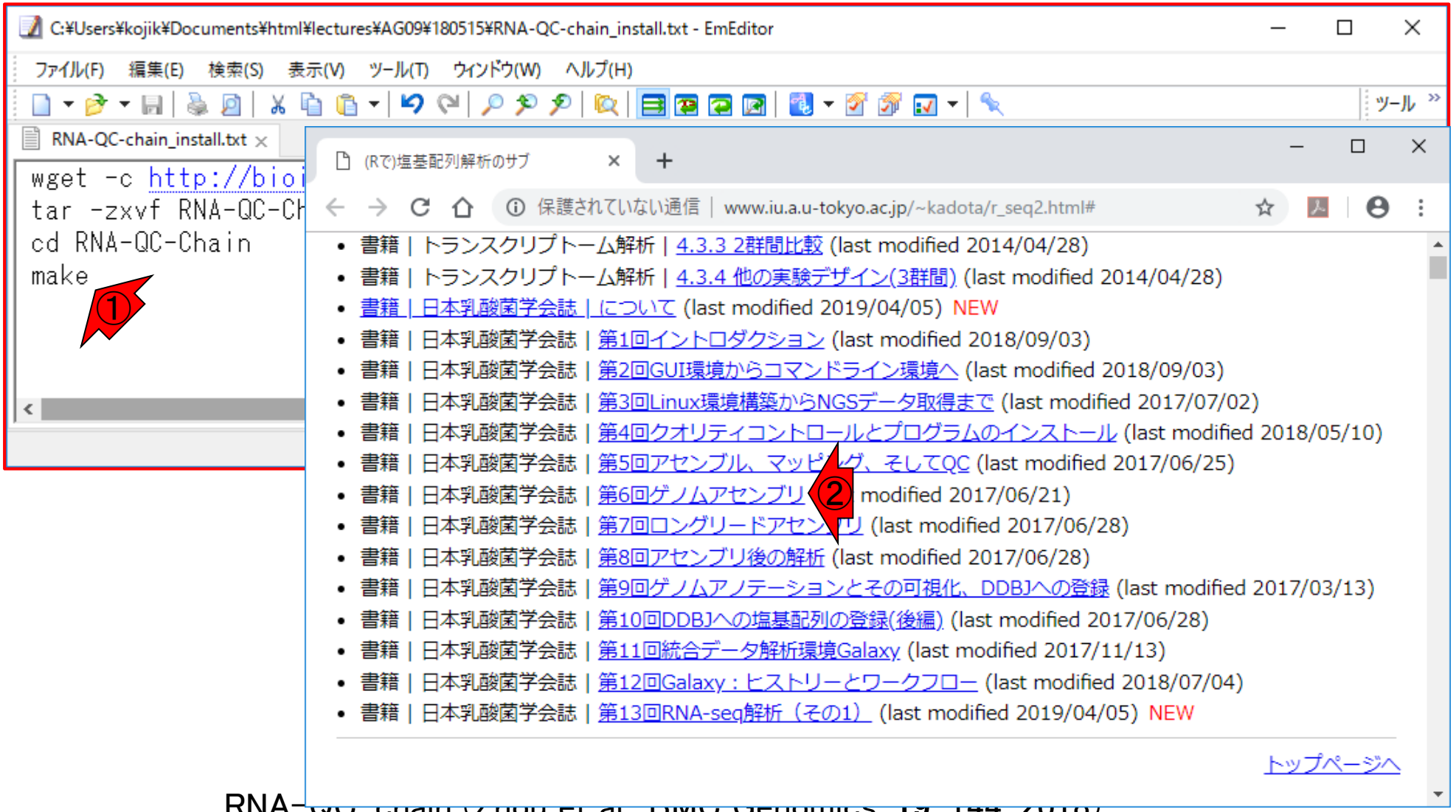
①私がやったのはとりあえずmakeのみ。ギョーカイのヒト以外には難解であろうが、Windowsの場合はsetup.exeをダブルクリックしてインストールを進めるが、それのLinux版のようなものという理解でよい



RNA-QC-chain (Zhou et al., BMC Genomics, **19**: 144, 2018)

make

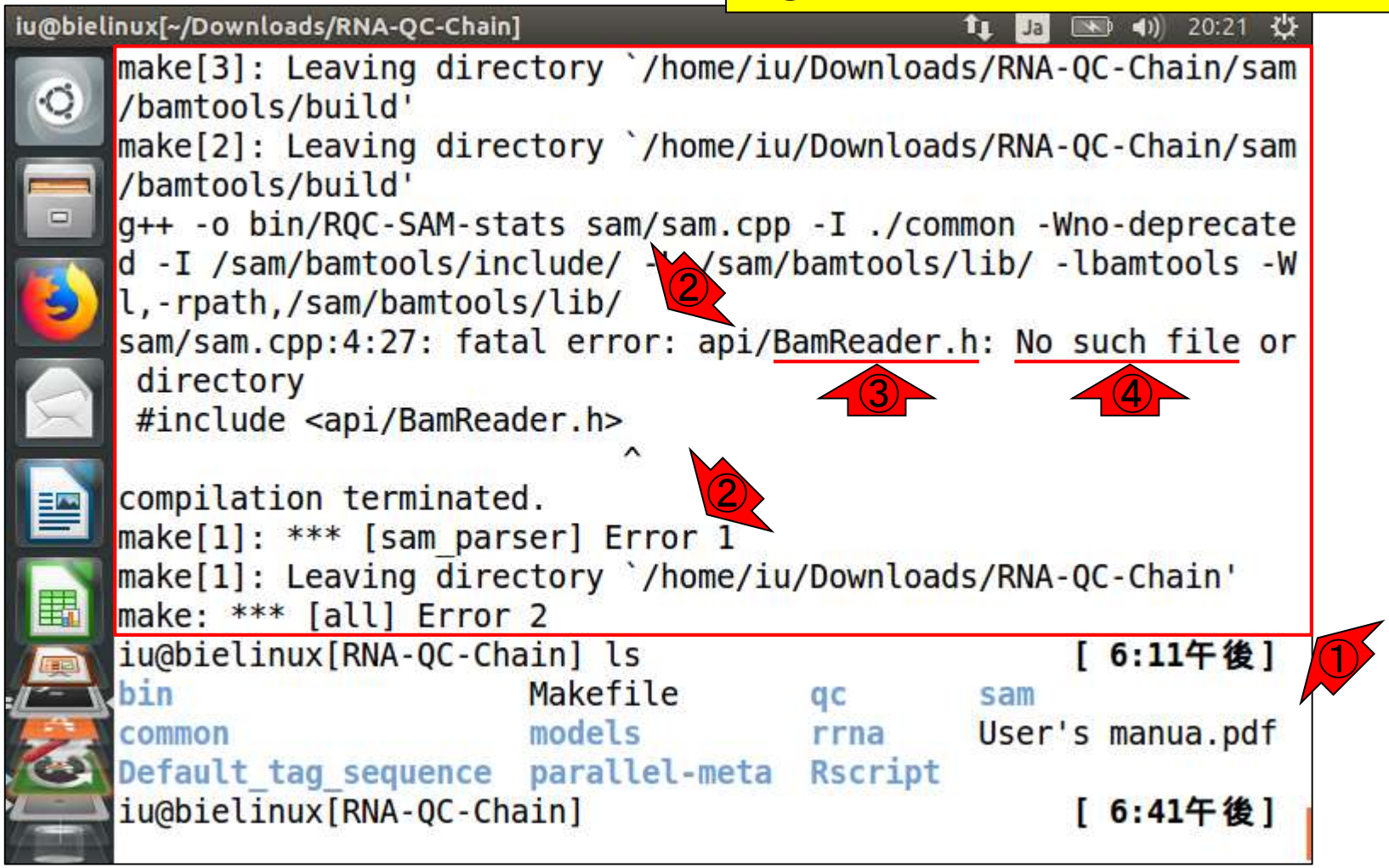
①makeについては、②乳酸菌学会誌のNGS連載第6回中でも使っている（ウェブ資料W9-4とか）



RNA-QC-chain (Zhou et al., BMC Genomics, 19: 144, 2018)

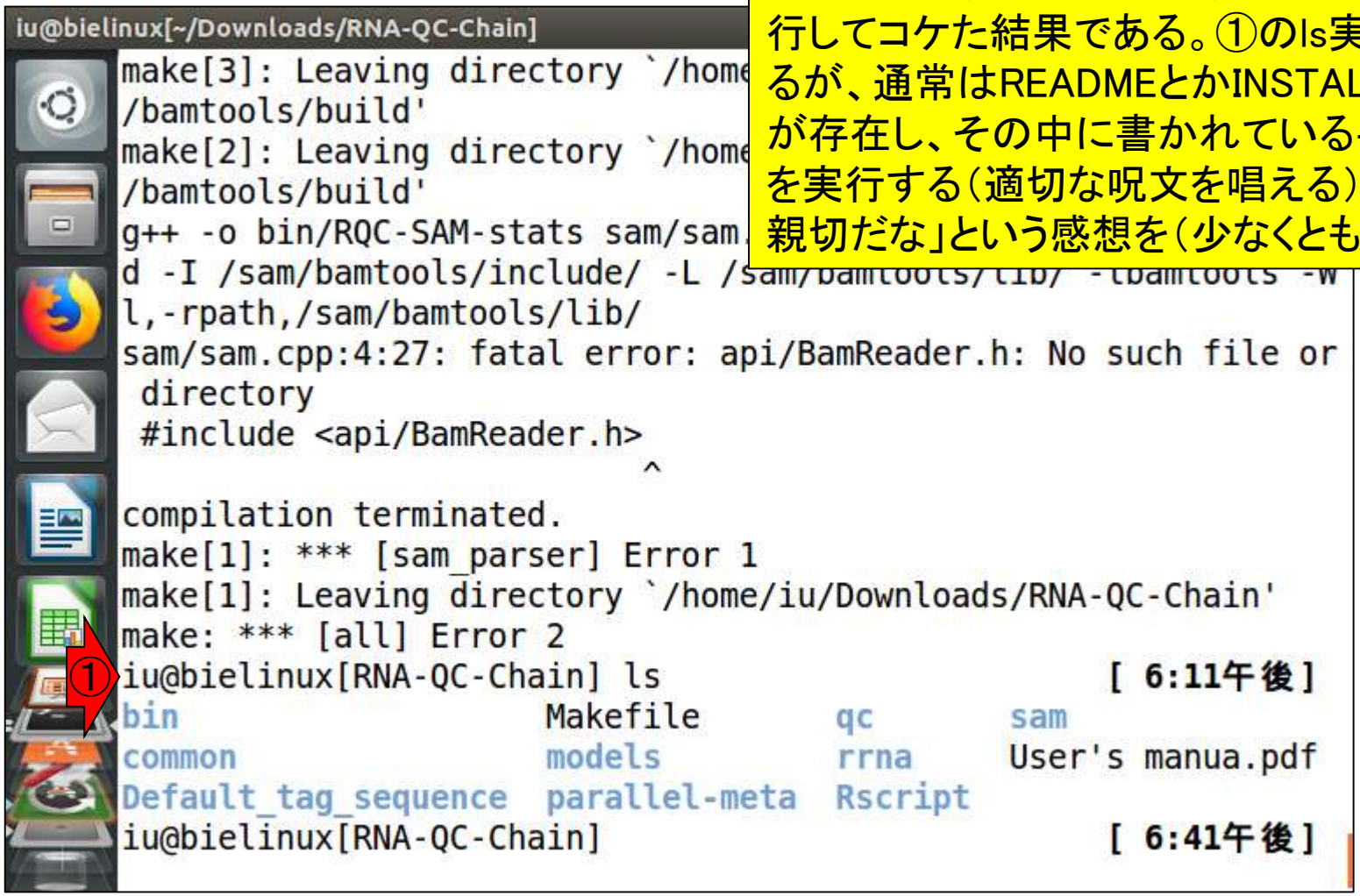
エラーメッセージ

①赤枠部分がmakeコマンドを実行して、②エラーメッセージが出て止まったところ。原因はどうやら、③BamReader.h、④というファイルがない、ことに起因するようだ



エラーメッセージ

今回はいきなりmakeを実行したが、「configure、make、make install」という3つの呪文を順番に唱えるのがおそらくより一般的。どっちにしようか迷ったがとりあえずmakeを実行してコケた結果である。①のls実行結果を眺めてもわかるが、通常はREADMEとかINSTALLという名前のファイルが存在し、その中に書かれている手順を見ながらコマンドを実行する（適切な呪文を唱える）。それがない段階で「不親切だな」という感想を（少なくとも私は）もつ



ググる

ここでは、①「RNA-QC-chain BamReader.h」というキーワードでウェブ検索。②がURL情報から、Biostarのページだと気づく。そして、③あたりで不穏当な記述を見つける

Biostar

①Biostarのリンク先ページ。赤枠の記載内容をみて、(少なくとも私は)同感する。あえてBiostarsではなくBiostarと書いている理由は、この原著論文がBiostarだからです。BiostarsでPubMed検索しても引っかからないので注意

QC-Chain is kind of messy to install, it is a pipeline composed of several tools developed by the same group, but they are not kept in sync. In addition, documentation is also a bit confusing, even though seems to pretty complete. In the end I gave up before using it, I may get back to it if someone says nice things about its results.

Biostar (Parnell et al., PLoS Comput Biol., 7: e1002216, 2011)

Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

前処理（preprocess

①この枠組みには、データの全体像を概観する Quality Checkも含まれる。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。代表的なプログラムは、FastQC

NGSリードデータ（SRAファイル）

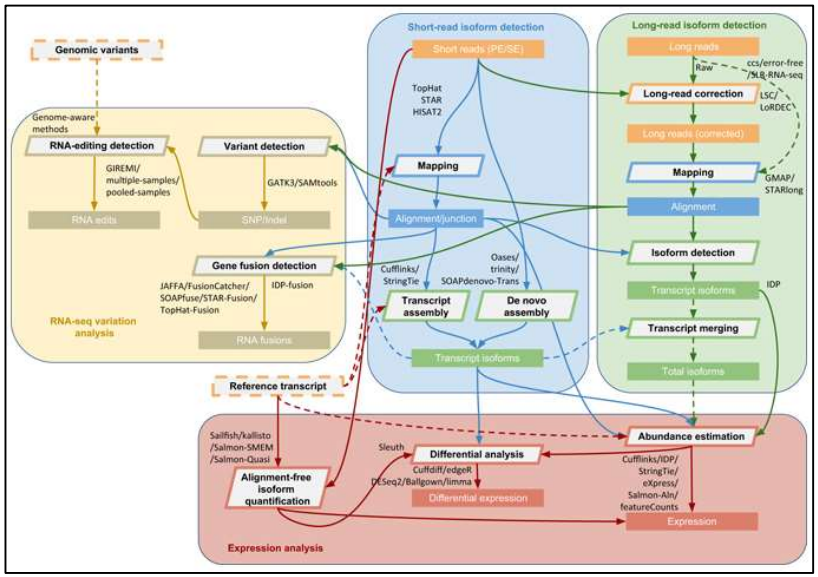
⇩

NGSリードデータ（FASTQファイル）

⇩

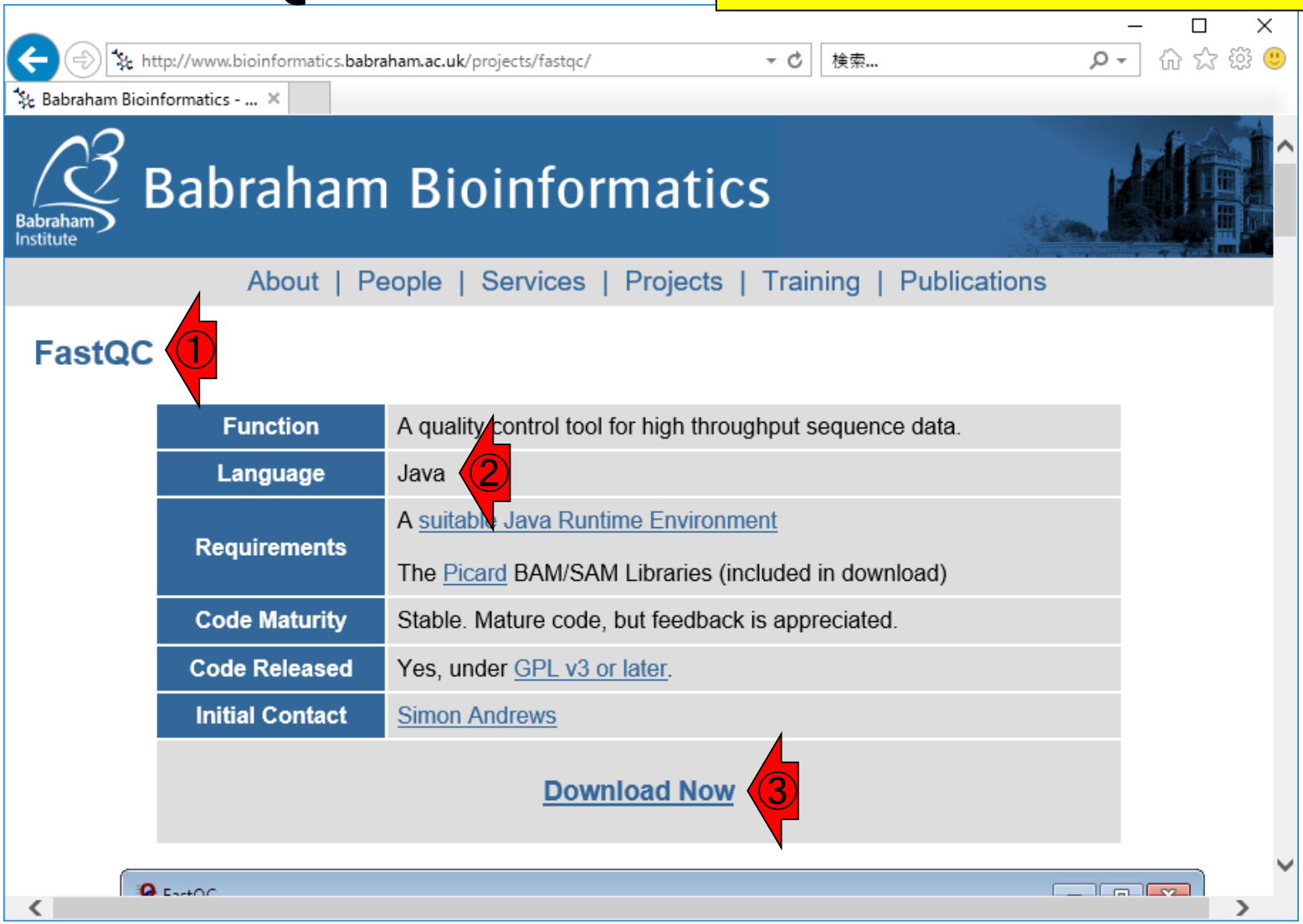
前処理（preprocessing） or Quality Control (QC) ①

⇩



FastQC

①FastQCのサイト。②このプログラムはJavaで書かれているので、特にLinux上でのインストールが非常に簡単（個人の感想です）。③ダウンロードしてインストールしましょう

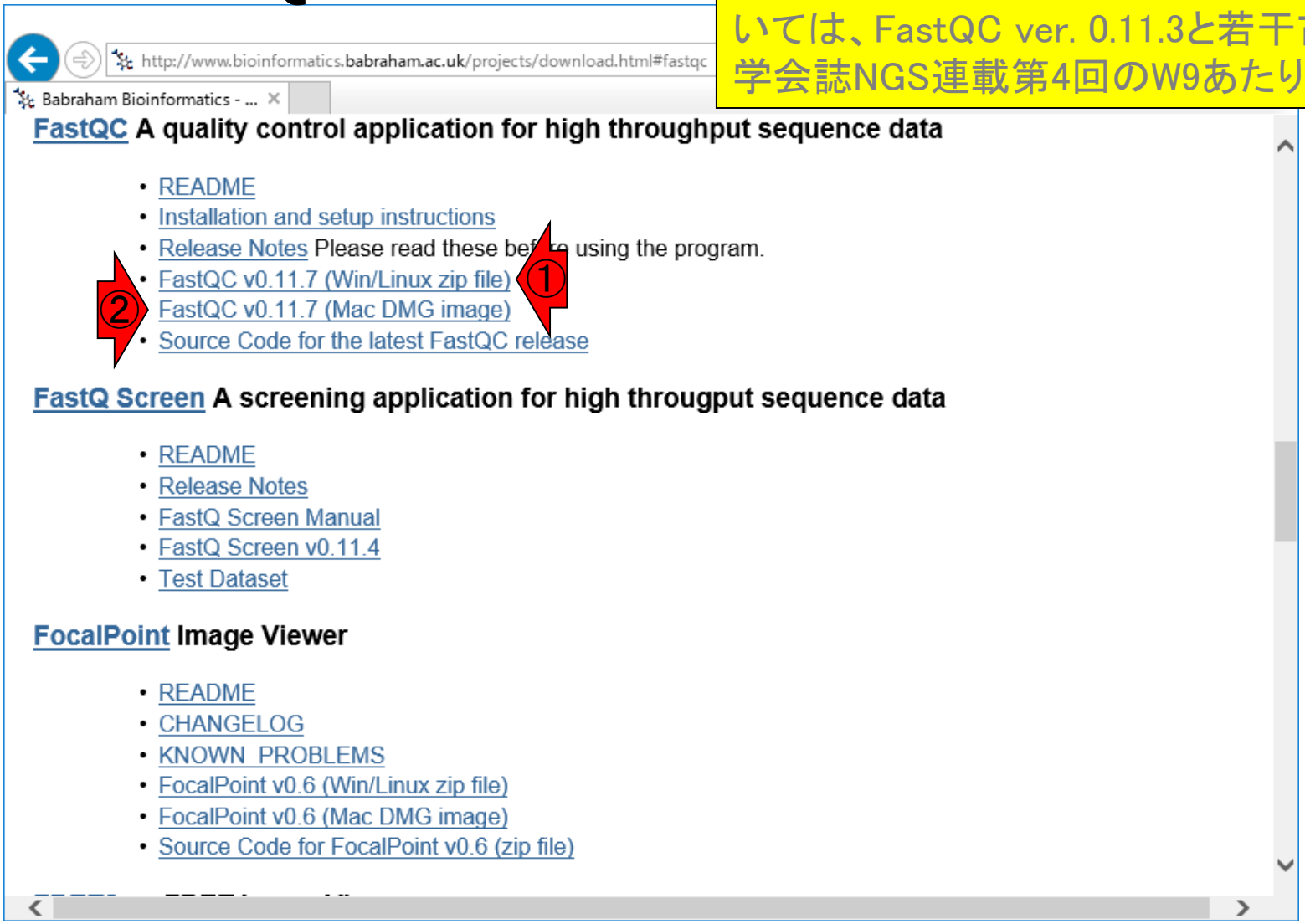


FastQC

こんな感じになります。初めてインストールするヒトにとってはほぼ無関係ですが、ちょこちょこバージョンアップしています。以前インストールしたことのあるヒトは、①のRelease Notesを見て、どの部分が変わったのか（新規機能の追加かバグ修正かなど）をざっとみておくといいでしょう

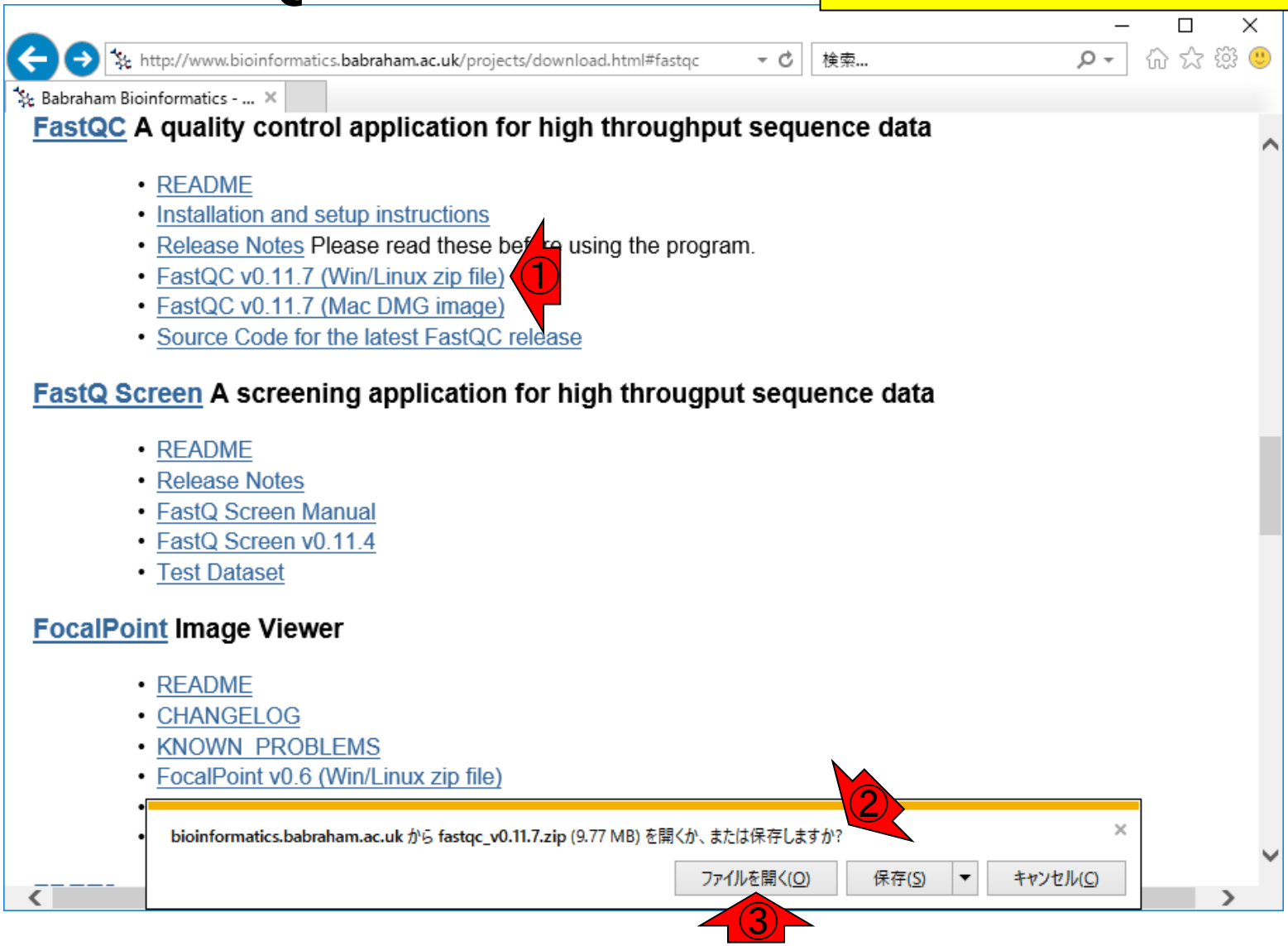
FastQC

①Windowsのヒトはこちら、②Macのヒトはこちら。以降の数枚のスライドはWindowsのインストール手順のスクリーンショットになります。ちなみに、Linux上でのインストールについては、FastQC ver. 0.11.3と若干古いですが日本乳酸菌学会誌NGS連載第4回のW9あたりから解説しています



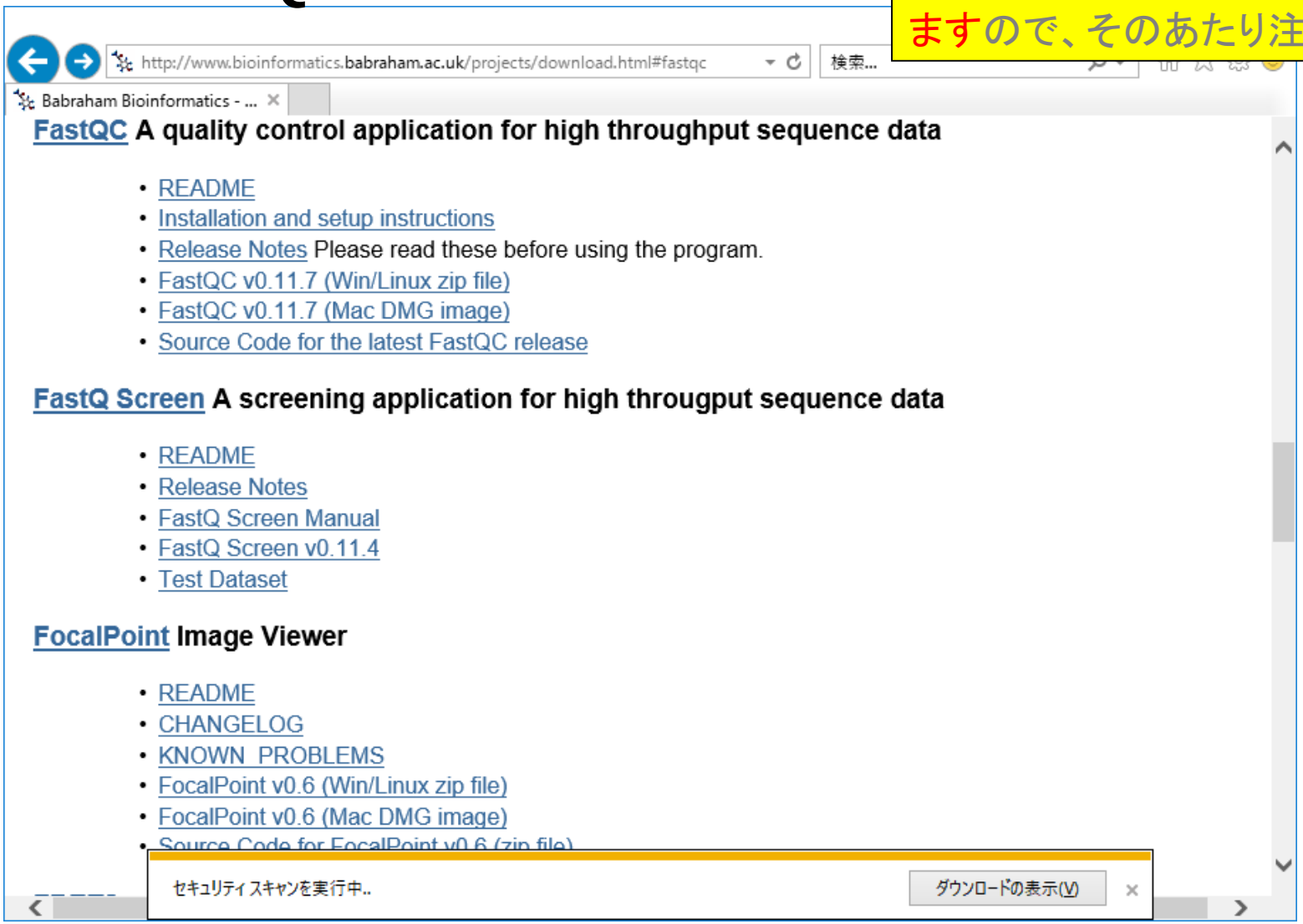
FastQCインストール

①を押すと、②のような感じになる。これ以降は個人所有PCのヒトは好きなやり方でインストールしてもらって構わない。貸与PCのヒトは③ファイルを開く



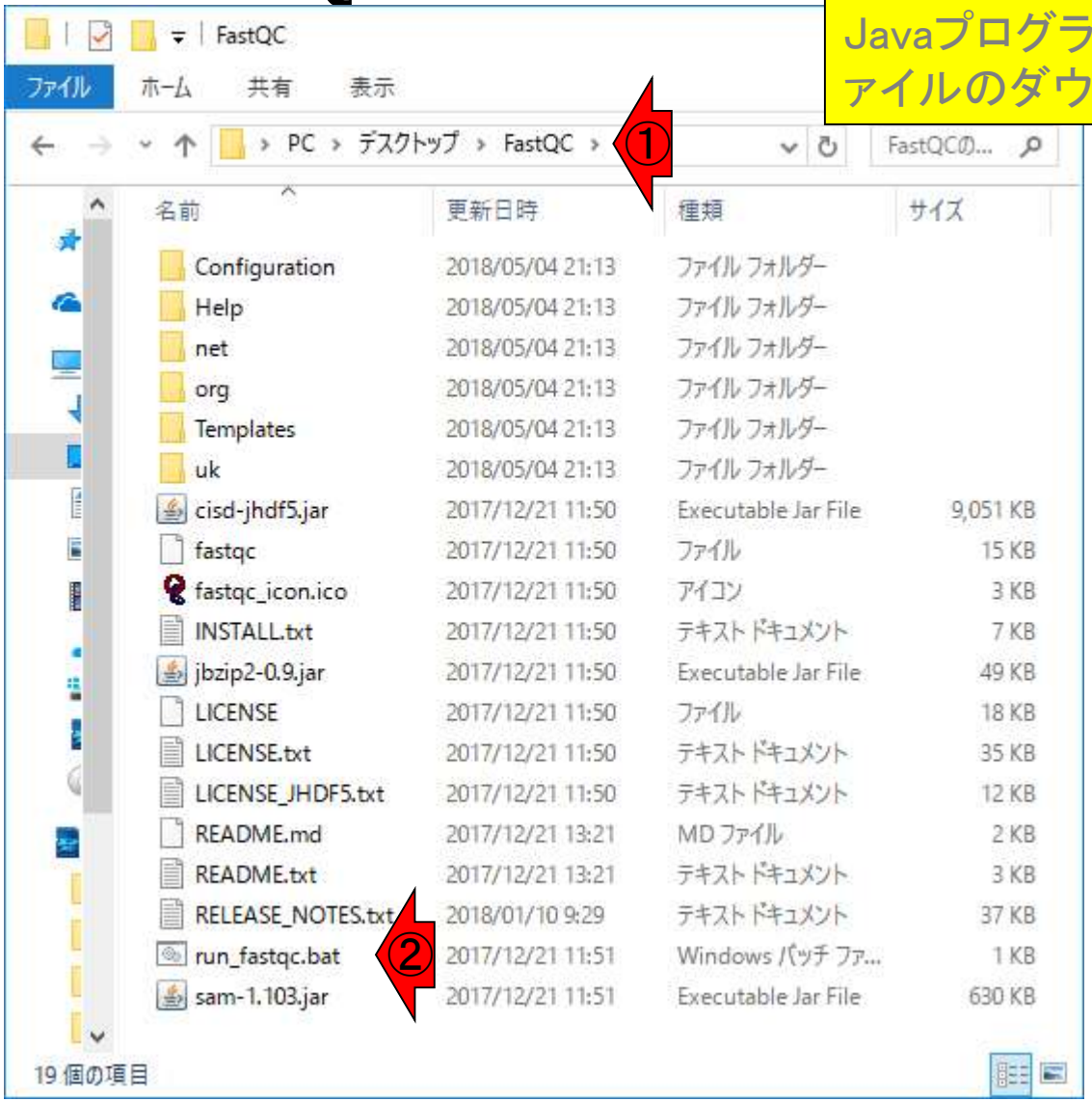
FastQCインストール

前のスライドで「ファイルを開く」を押した直後。
ちなみに、FastQCはJavaプログラムなのでJava本体をインストールしておく必要はもちろんありますので、そのあたり注意しておいてください



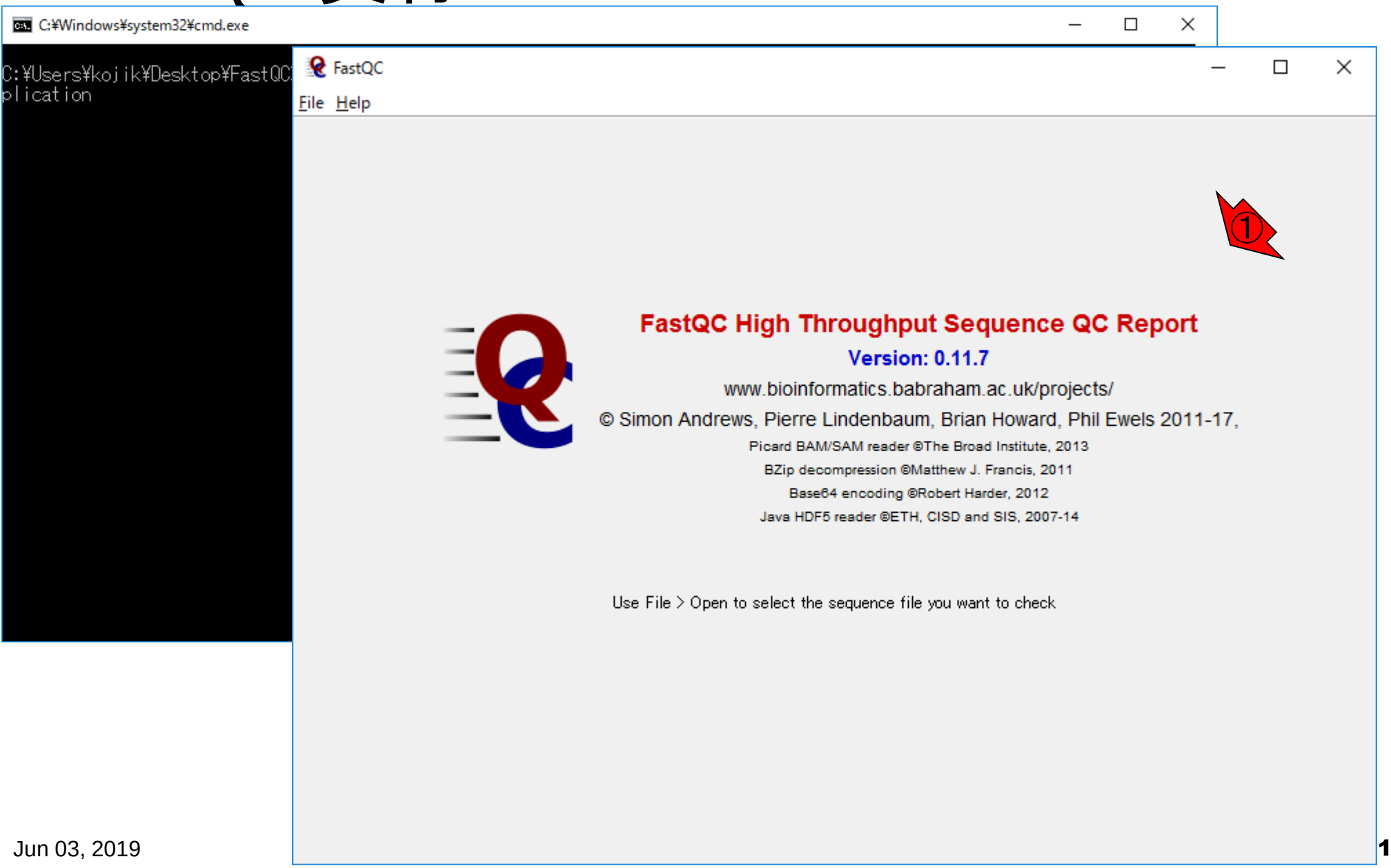
FastQC実行

ほどなくして、こんな感じになります。つまり、①デスクトップ上にFastQCというフォルダが作成されます。Windows上で実行する場合は、②run_fastqc.batをダブルクリック。
Javaプログラムの場合は、インストールというよりは実行ファイルのダウンロード、という理解でよろしいかと思います



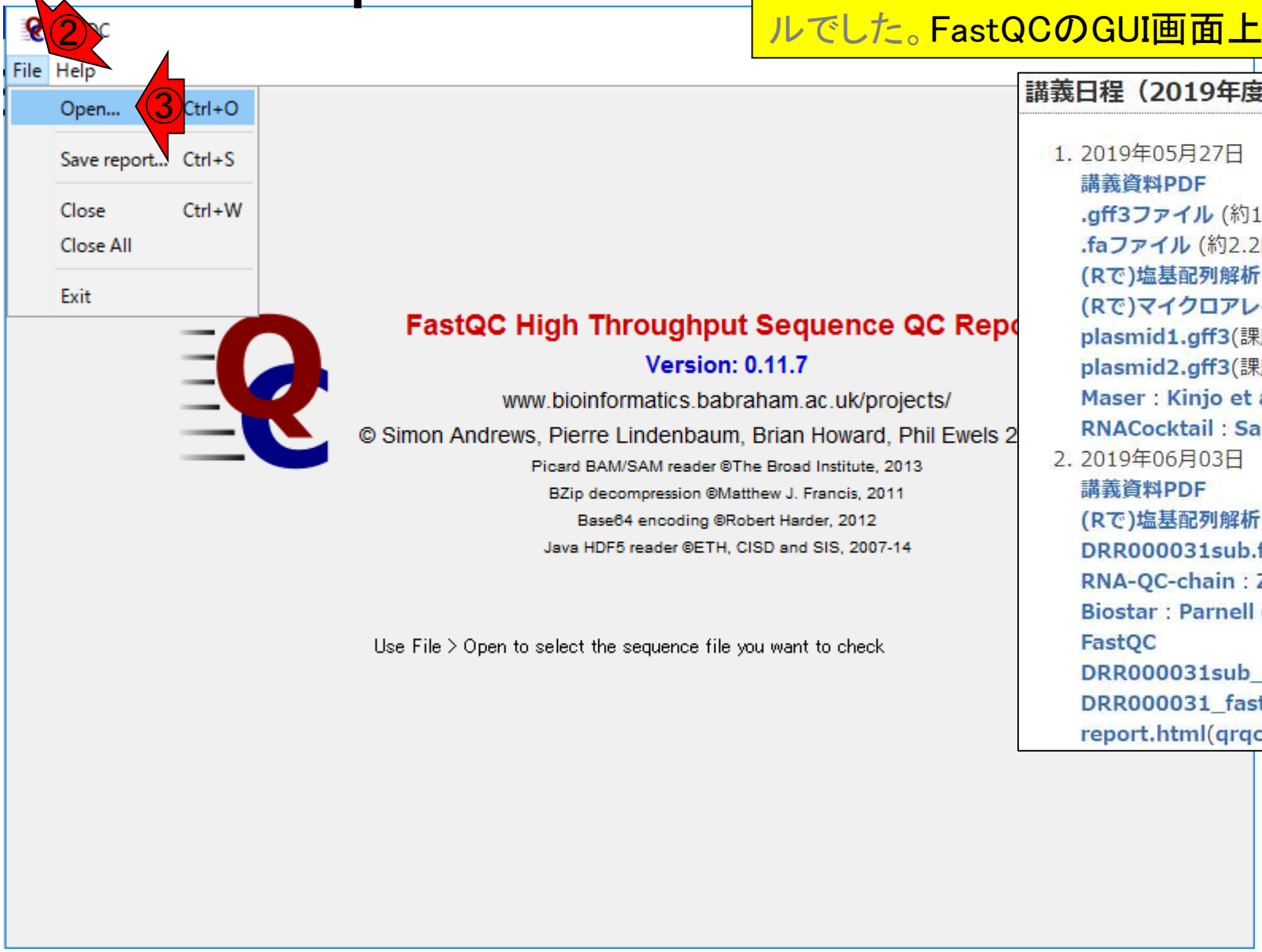
FastQC実行

こんな感じで、2つのウィンドウが立ち上がる。使うのは①こちらです



File - Open

ここでは、①DRR000031sub.fastqをデスクトップにダウンロードしておき、それを入力ファイルとして実行する。ちなみにこれは、6,000リードからなる約1MBのファイルでした。FastQCのGUI画面上で、②File、③Open

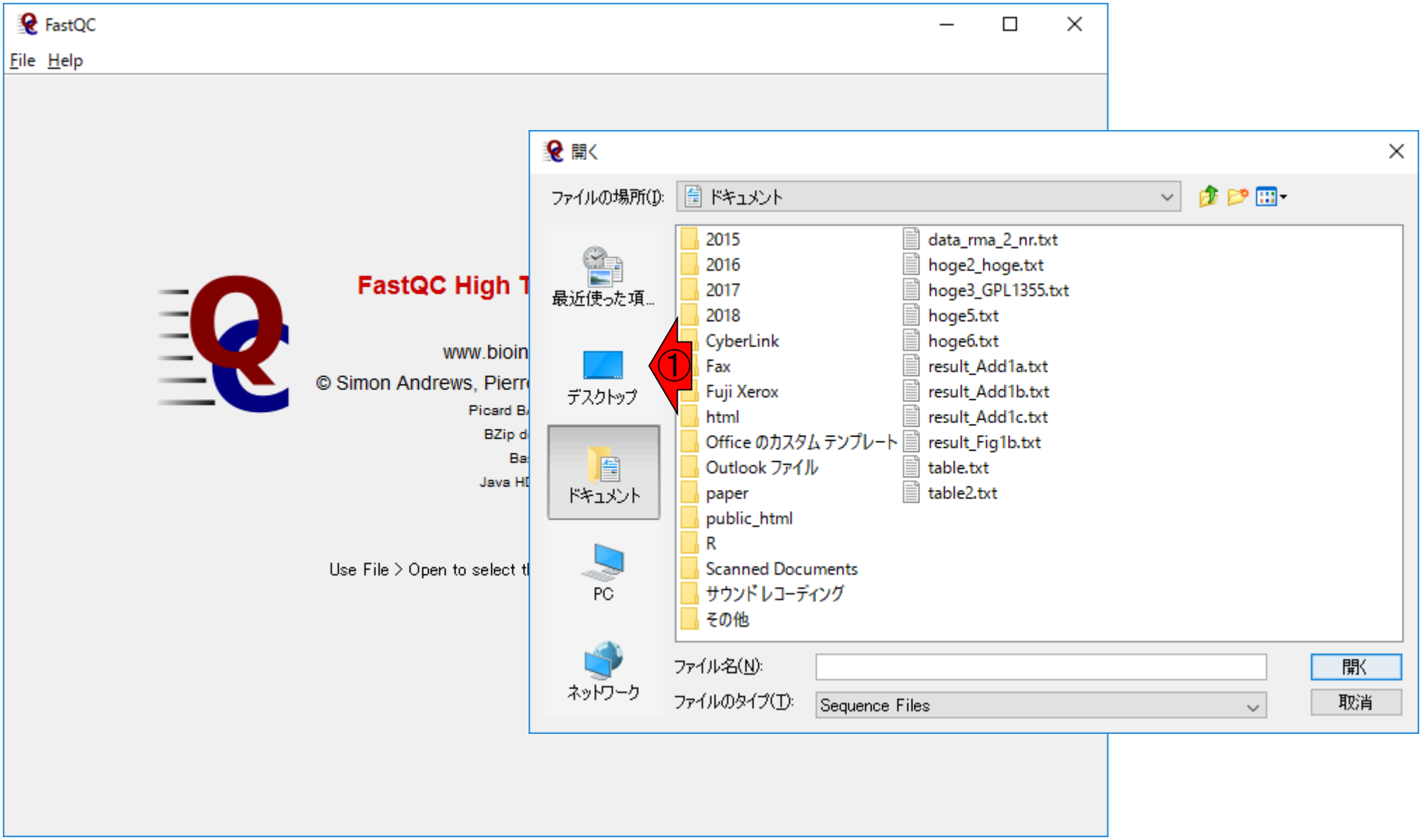


講義日程（2019年度）

1. 2019年05月27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)
   - Maser : Kinjo et al., Database (Oxford), 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
   - 講義資料PDF
   - (Rで)塩基配列解析
   - DRR000031sub.fastq ①
   - RNA-QC-chain : Zhou et al., BMC Genomics, 2018
   - Biostar : Parnell et al., PLoS Comput Biol., 2011
   - FastQC
   - DRR000031sub_fastqc.html
   - DRR000031_fastqc.html(課題用)
   - report.html(qrqcを用いたQC結果)

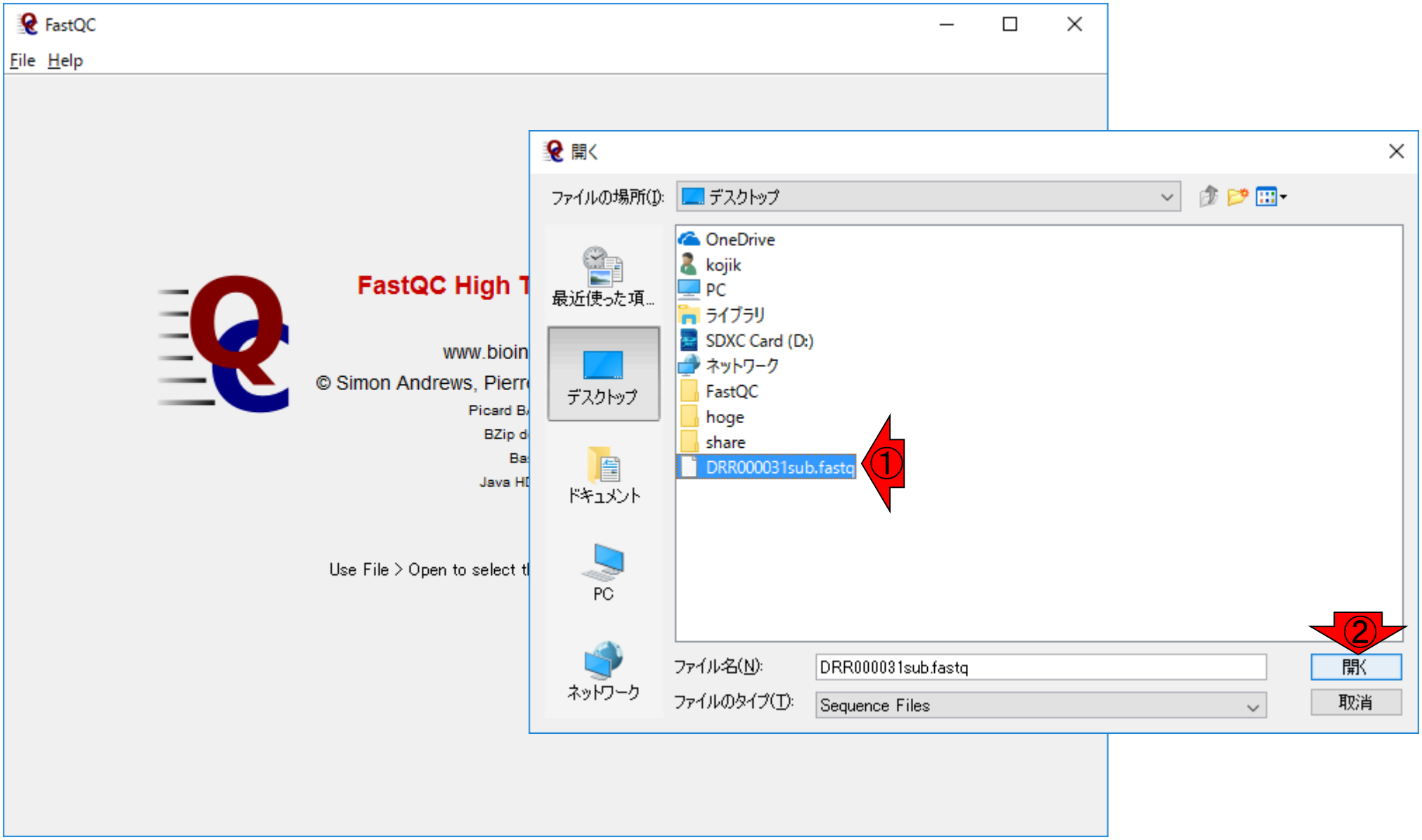
①デスクトップ…

入力ファイルの指定



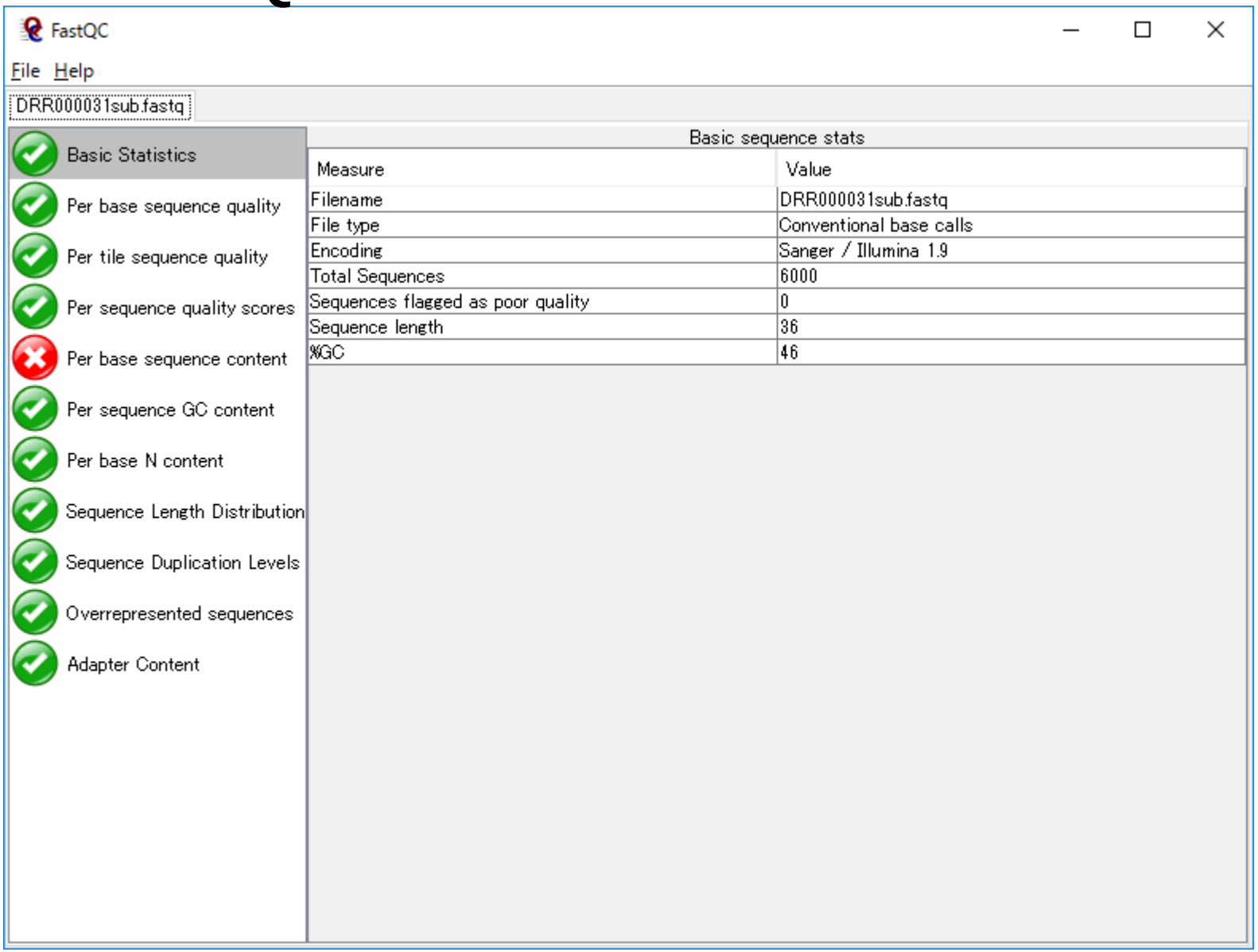
入力ファイルの指定

①解析対象ファイル（DRR000031sub.fastq）を選択して、②開く



FastQC実行結果

プログラムの実行は一瞬で終了し、このような結果が得られます

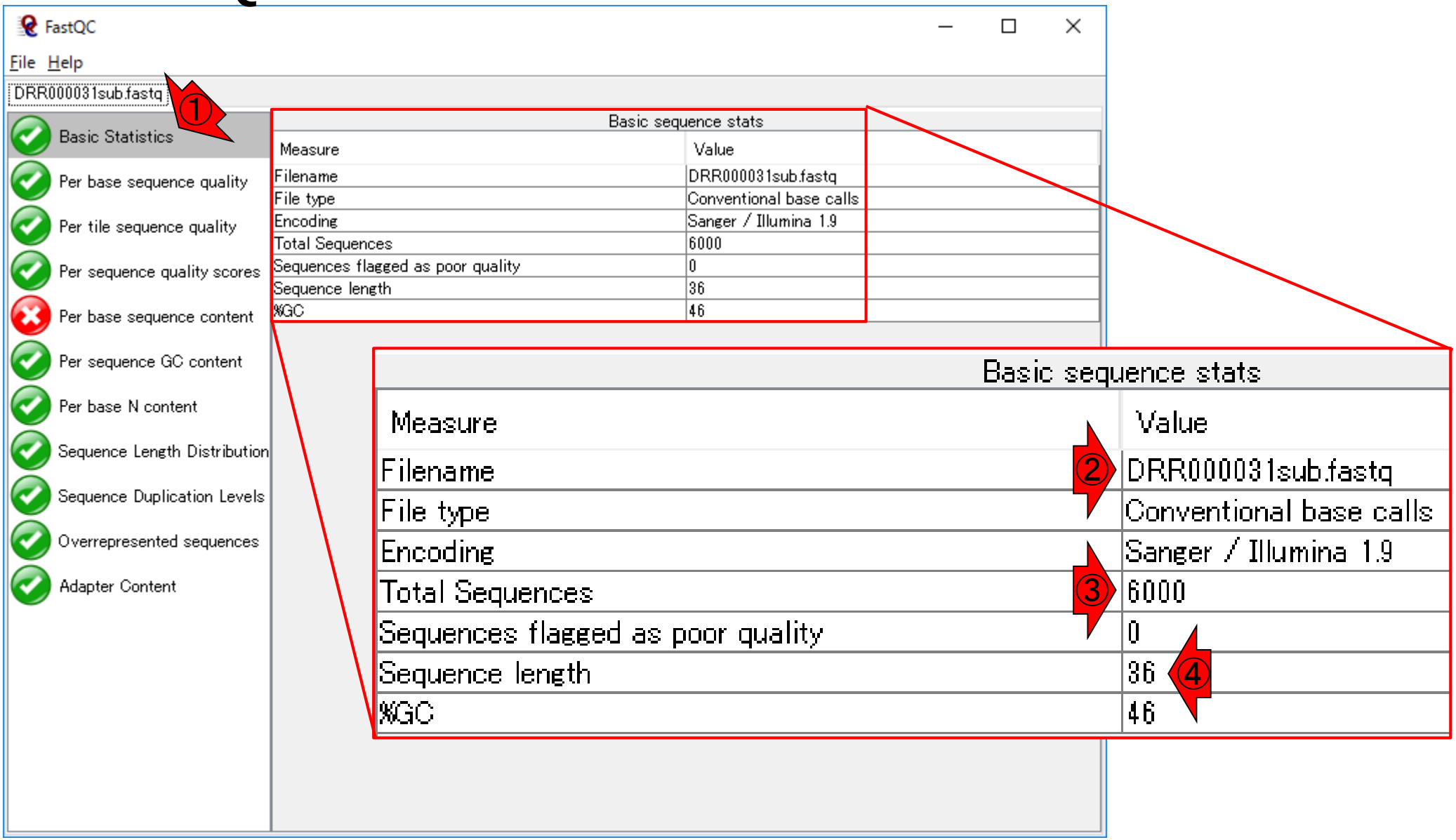


Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

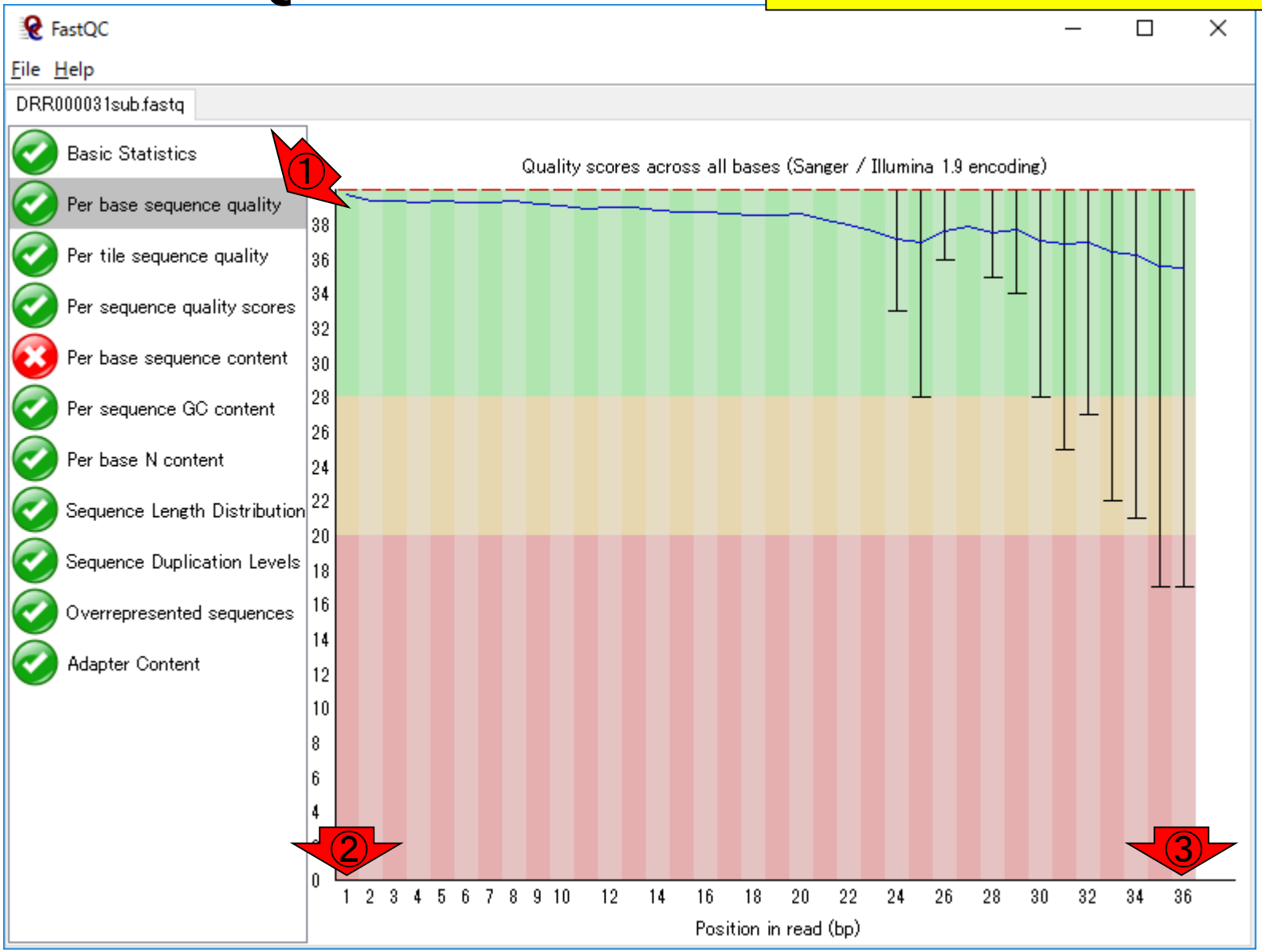
FastQC実行結果

①Basic Statisticsの情報が右側に表示されています。②入力ファイル、③リード数、④配列長



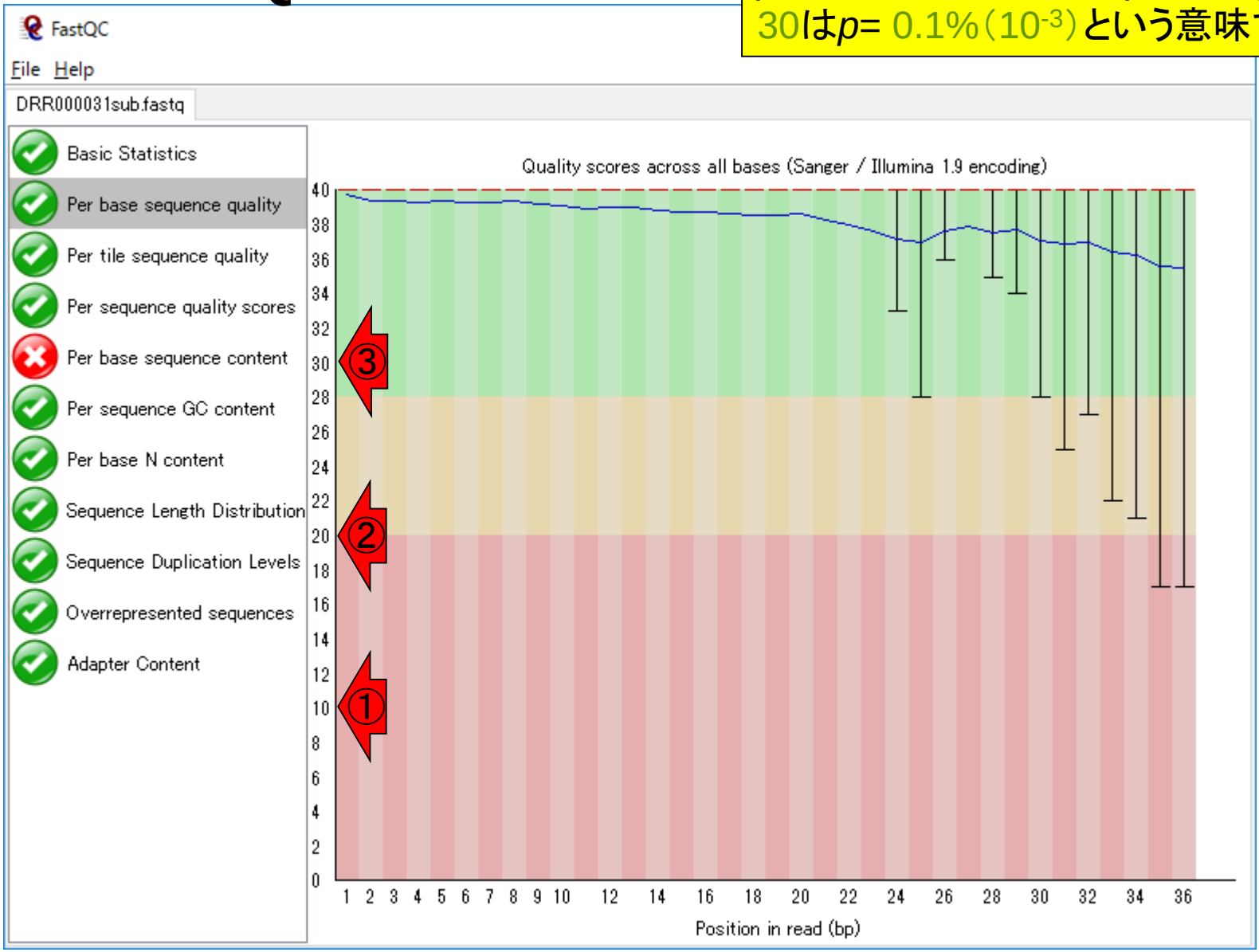
FastQC実行結果

①塩基ごとのクオリティスコア情報。横軸がリード中の塩基ポジション。全部で36 bpしかないので、②1番目から③36番目の塩基位置みたいな感じで読み解く



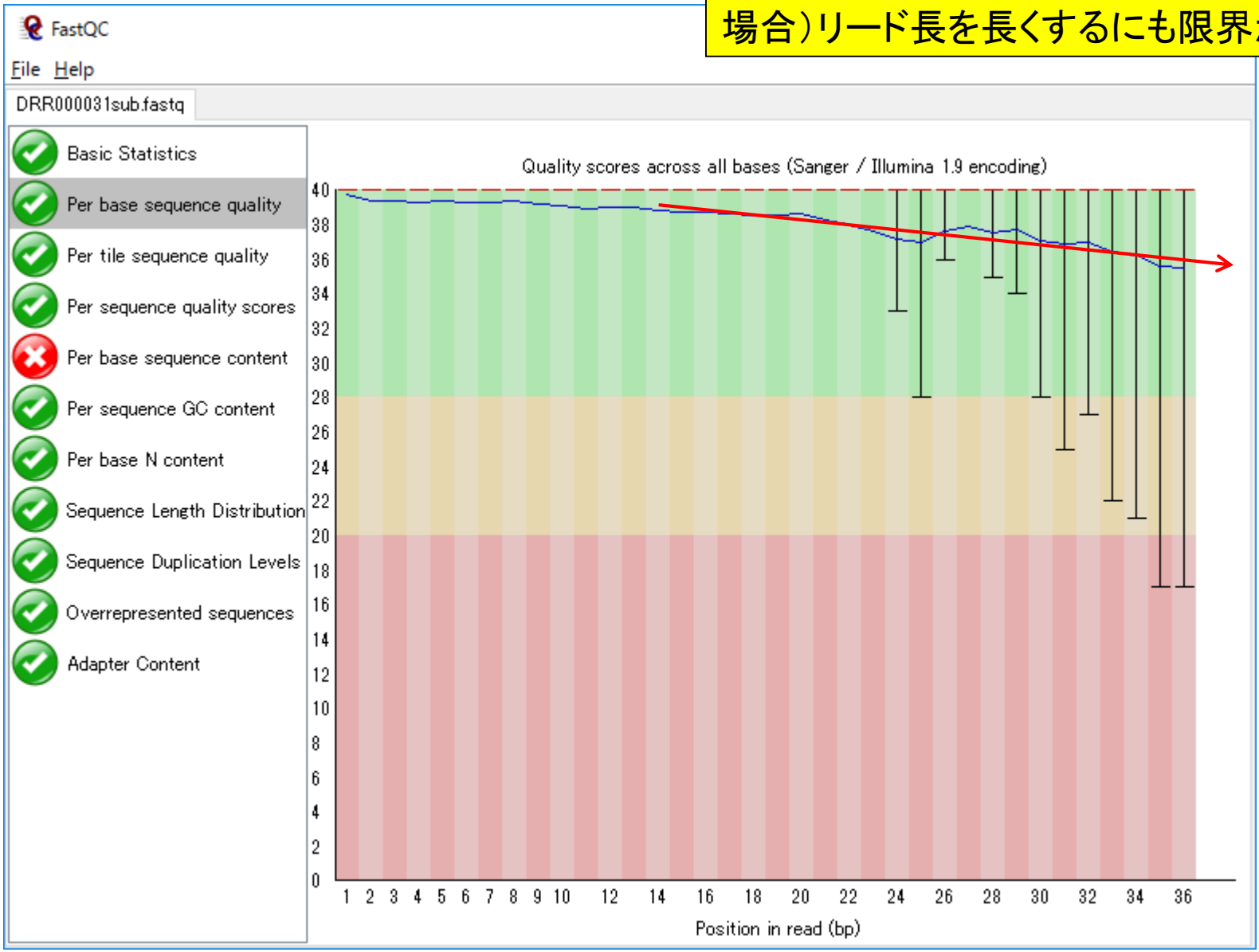
FastQC実行結果

縦軸がクオリティスコアq。高ければ高いほどよい。①q=10はベースコール結果が間違っている確率（エラー率p）が10%（10^{-1}）で、②q = 20はp = 1%（10^{-2}）、③q = 30はp= 0.1%（10^{-3}）という意味でした（前回の講義）



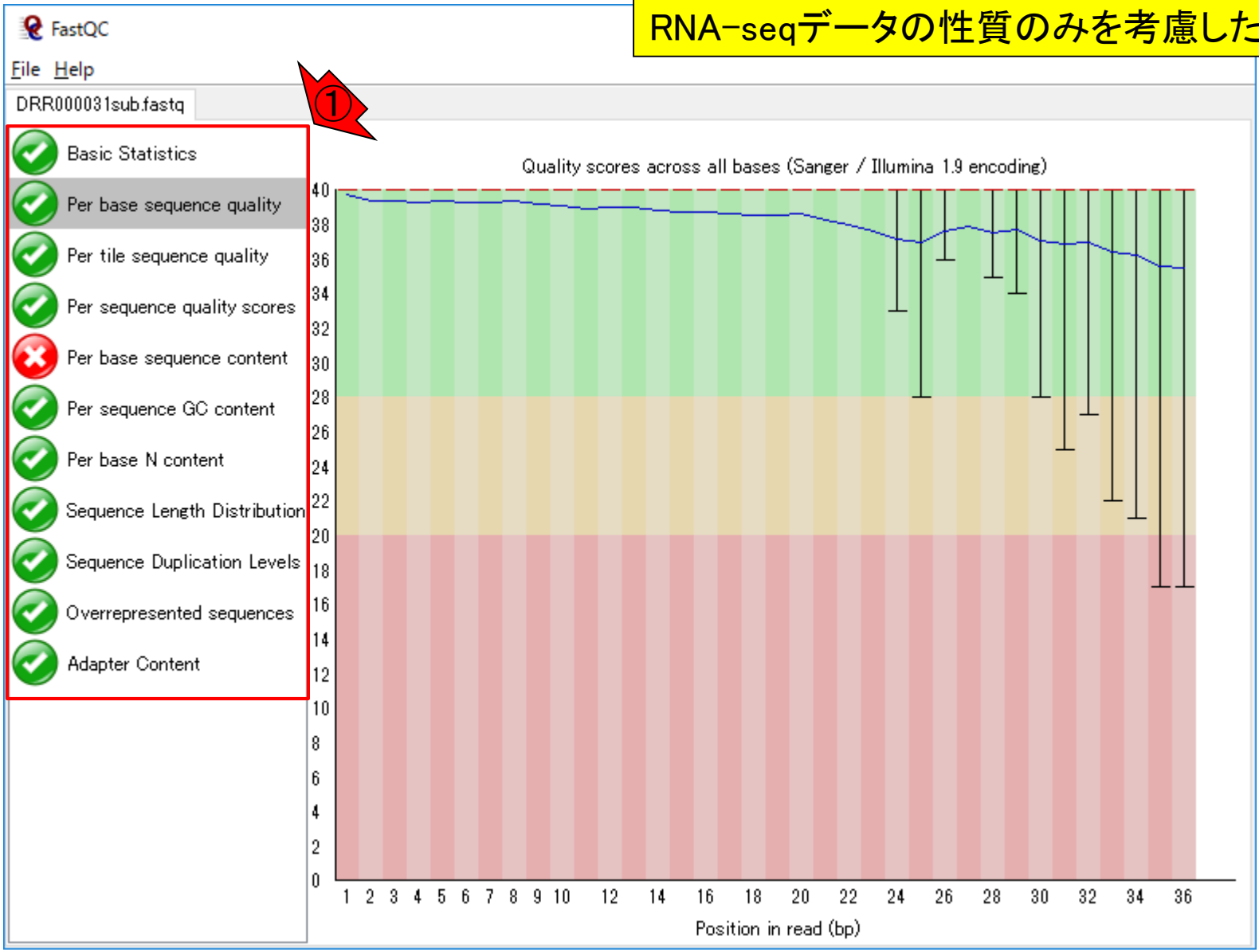
解釈

塩基配列決定精度は、赤矢印で示すようになだらかに右肩下がりになっています。つまり、読み進んでいくにつれて下がっている、と読み解きます。だから（特にIlluminaの場合）リード長を長くするにも限界があるのです。



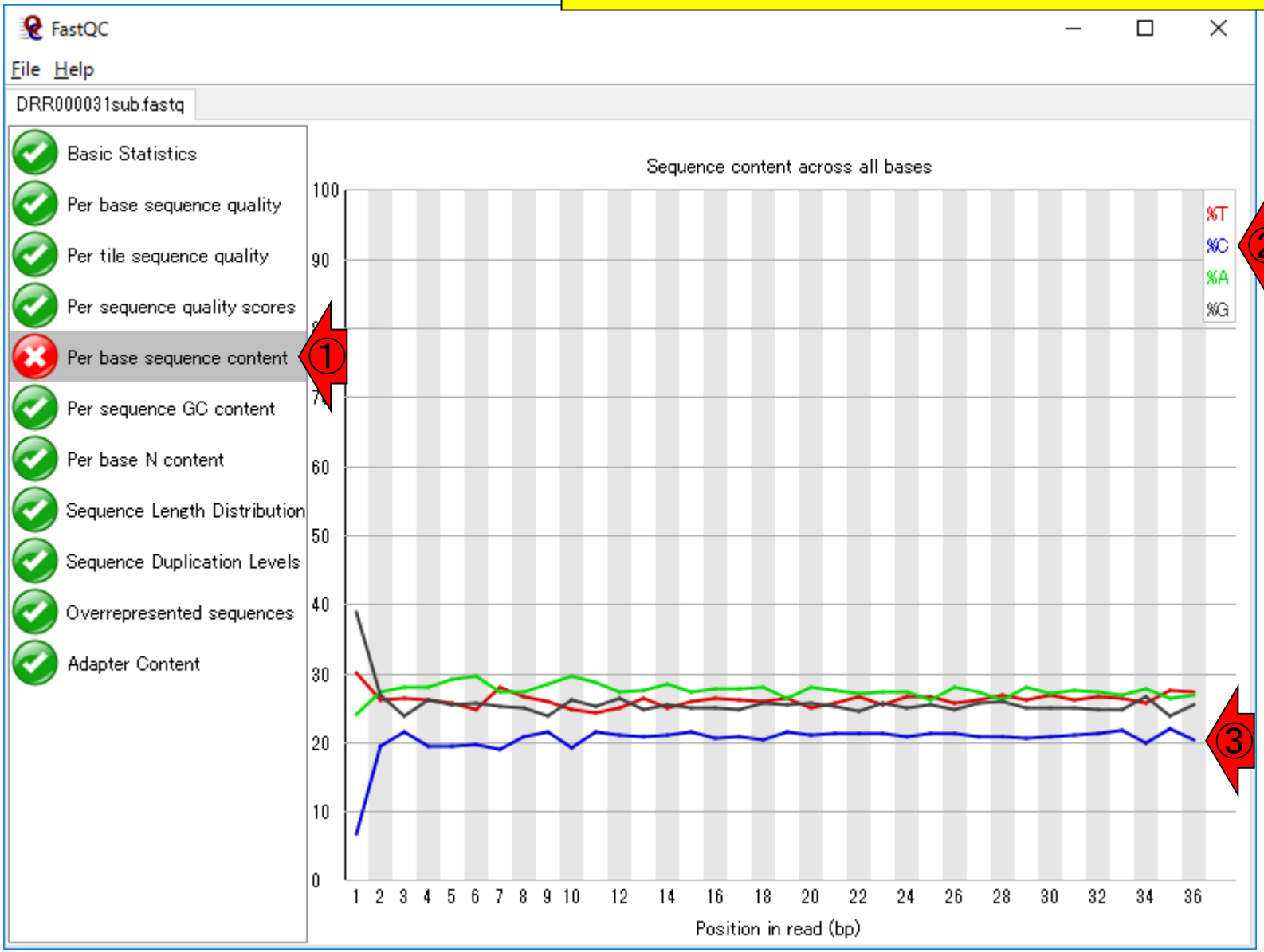
信号機と同じ

赤枠部分は、緑色と赤色が見えていますが、これ以外にも黄色があります。基本的には信号機と同じ解釈でよい（緑は問題ない）が、あくまでもFastQCはNGSデータ全般用であり、RNA-seqデータの性質のみを考慮したものではない点に注意



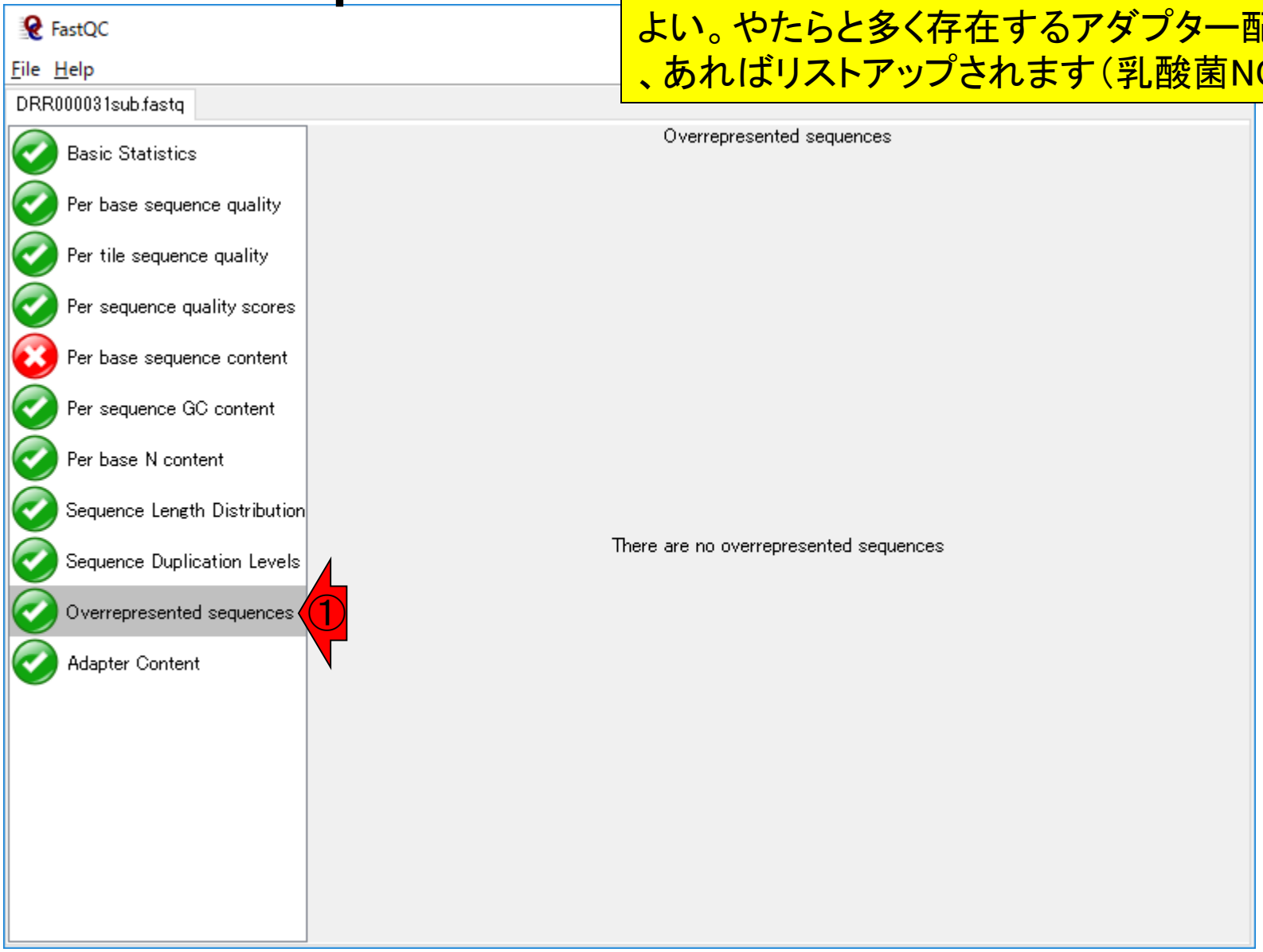
赤色の項目

この入力ファイル（DRR000031sub.fastq）の場合、①Per base sequence contentという項目が赤色だった。おそらく、②Cの、③存在確率のみが低いので、このようなアラートが出ているのであろう



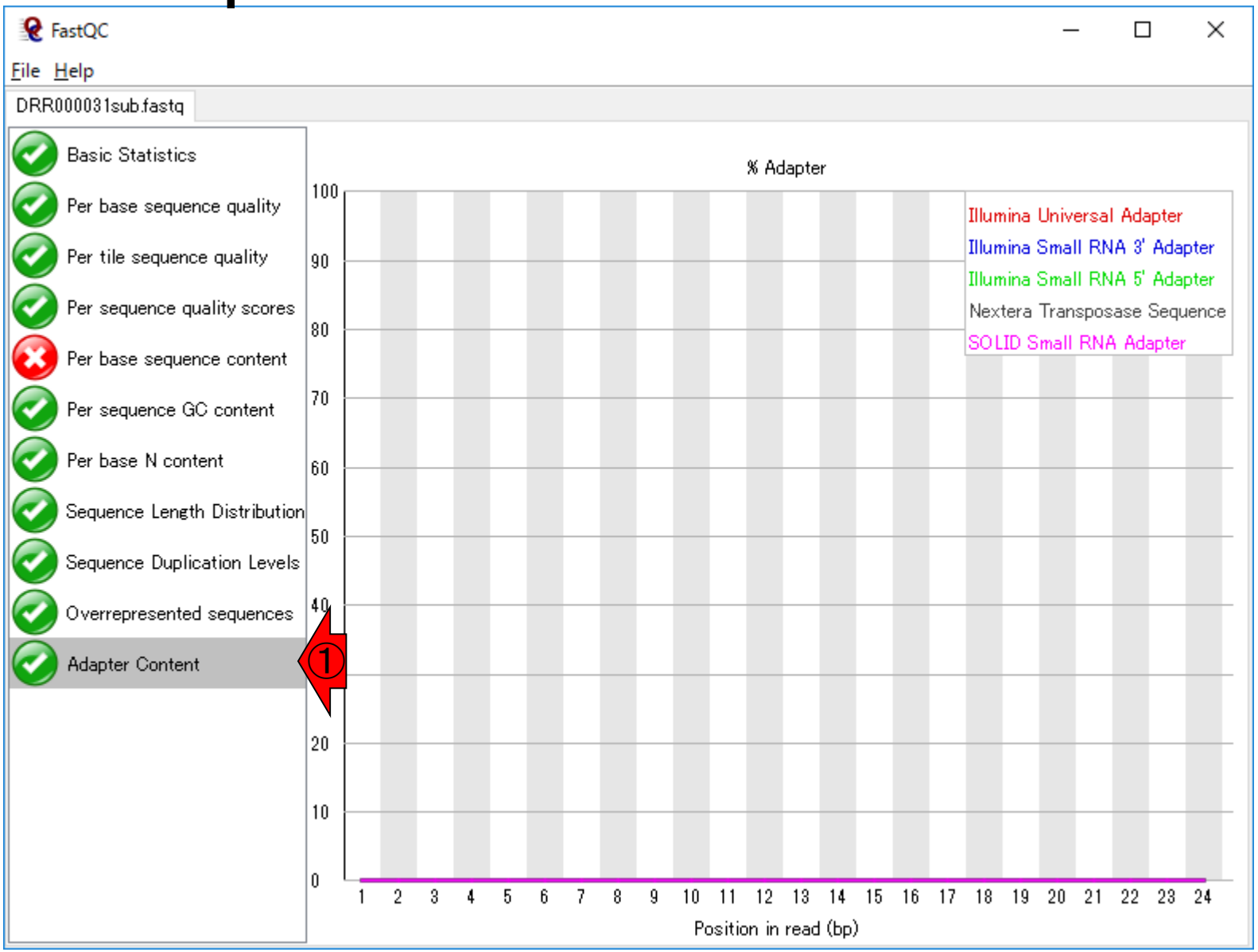
Overrepresent

mRNA-seqの場合は、①ここにポリA由来配列（AAAAAA…）が上位に表示されたりします。Overrepresented sequencesというのは、やたらと多くリード中に存在する部分配列、という理解でよい。やたらと多く存在するアダプター配列由来の部分配列も、あればリストアップされます（乳酸菌NGS連載第4回W8-6）



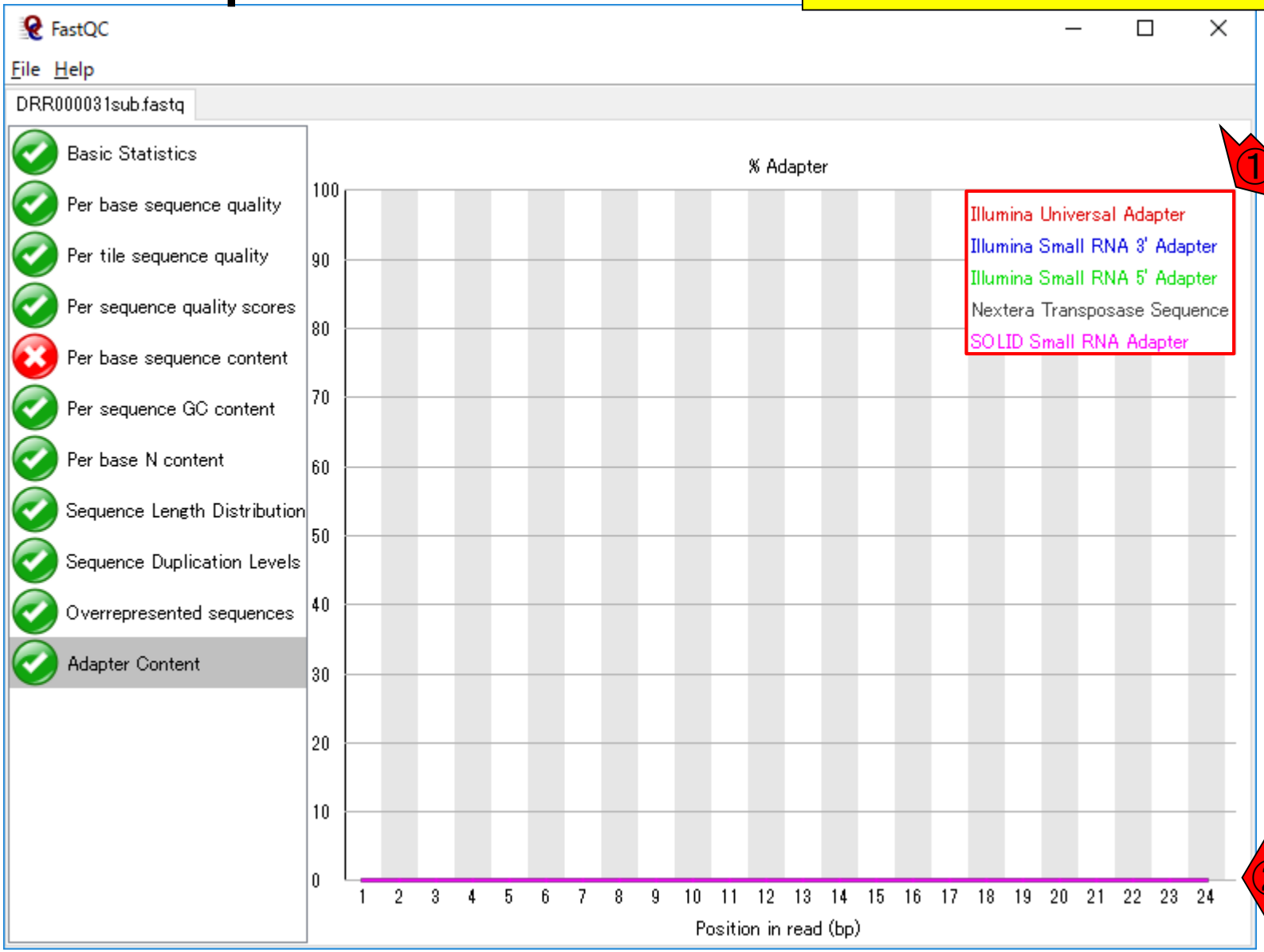
①アダプター配列を含む場所（横軸）と割合（縦軸）

Adapter Content



Adapter Content

このデータの場合は、①既知のアダプター配列が、②含まれていないことがわかる。アダプター配列を含む例は、乳酸菌NGS連載第6回W4-2にあります

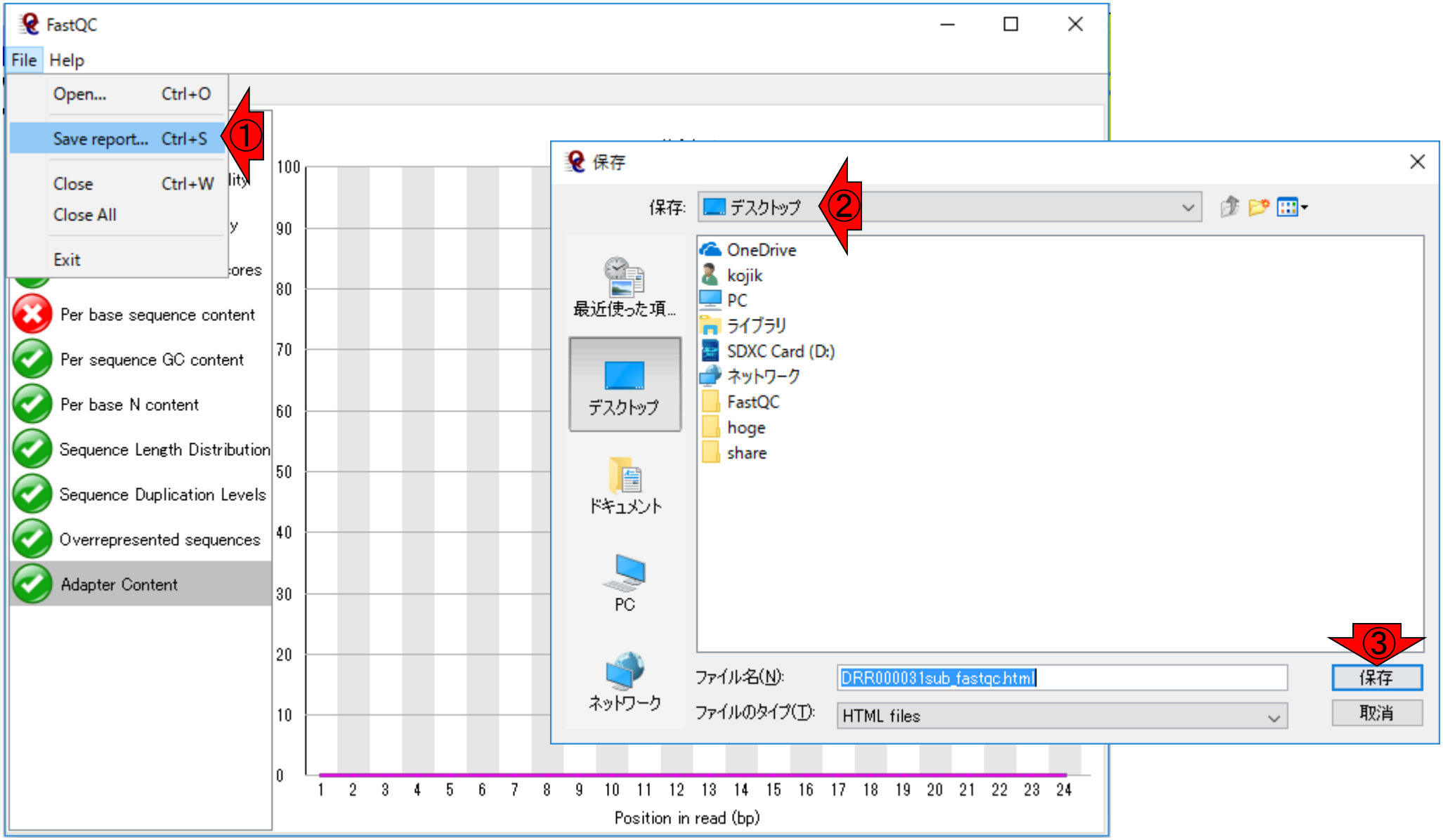


①

②

htmlファイルで保存

①「File – Save report…」、②デスクトップに、③保存して得られたものが…



①DRR000031sub_fastqc.htmlです

DRR000031sub_fastqc.html

講義日程（2019年度）

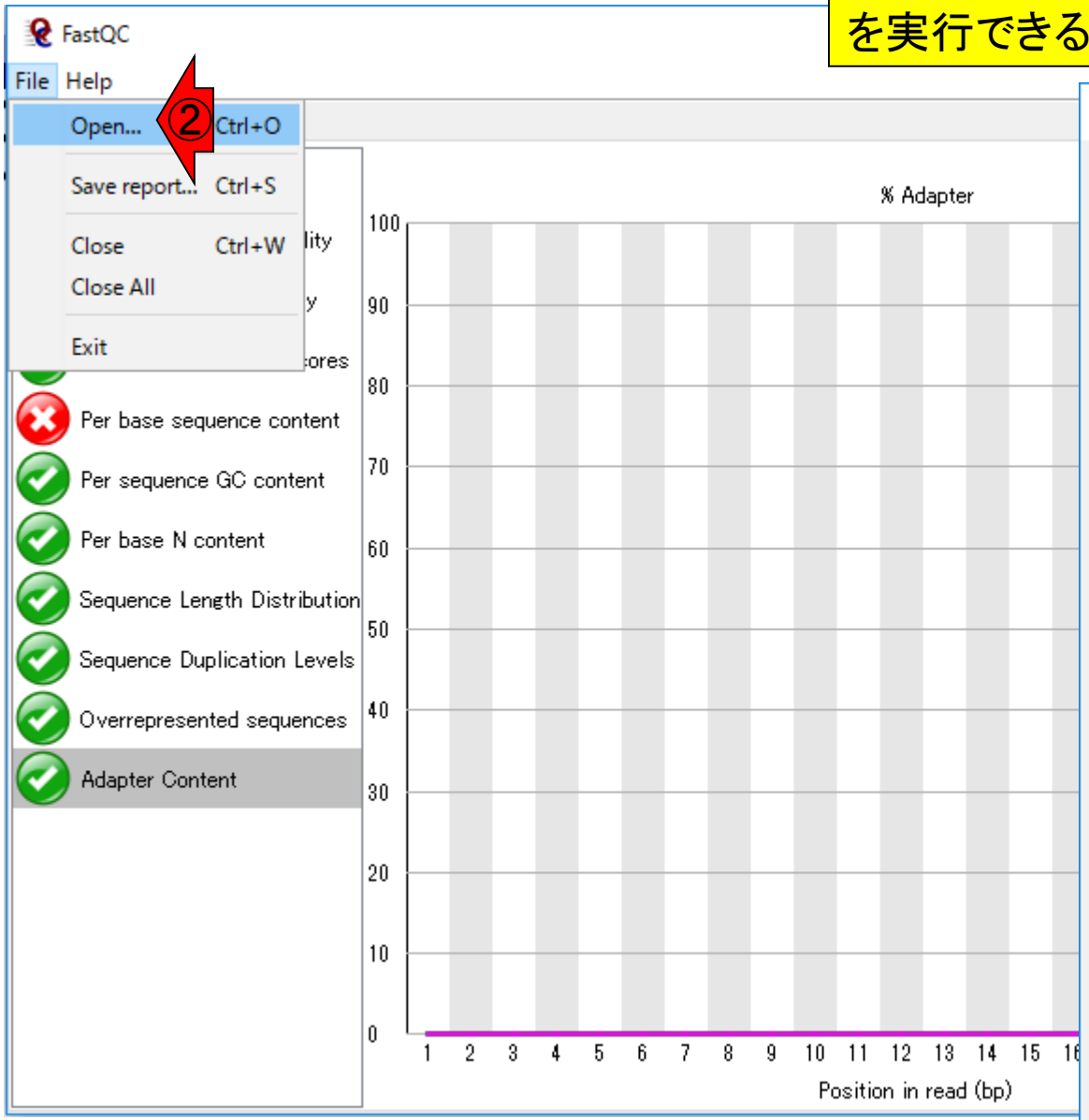
1. 2019年05月27日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)
   - Maser : Kinjo et al., Database (Oxford), 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
   - 講義資料PDF
   - (R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 ①
   - DRR000031_fastqc.html(課題用)
   - report.html(qrqcを用いたQC結果)

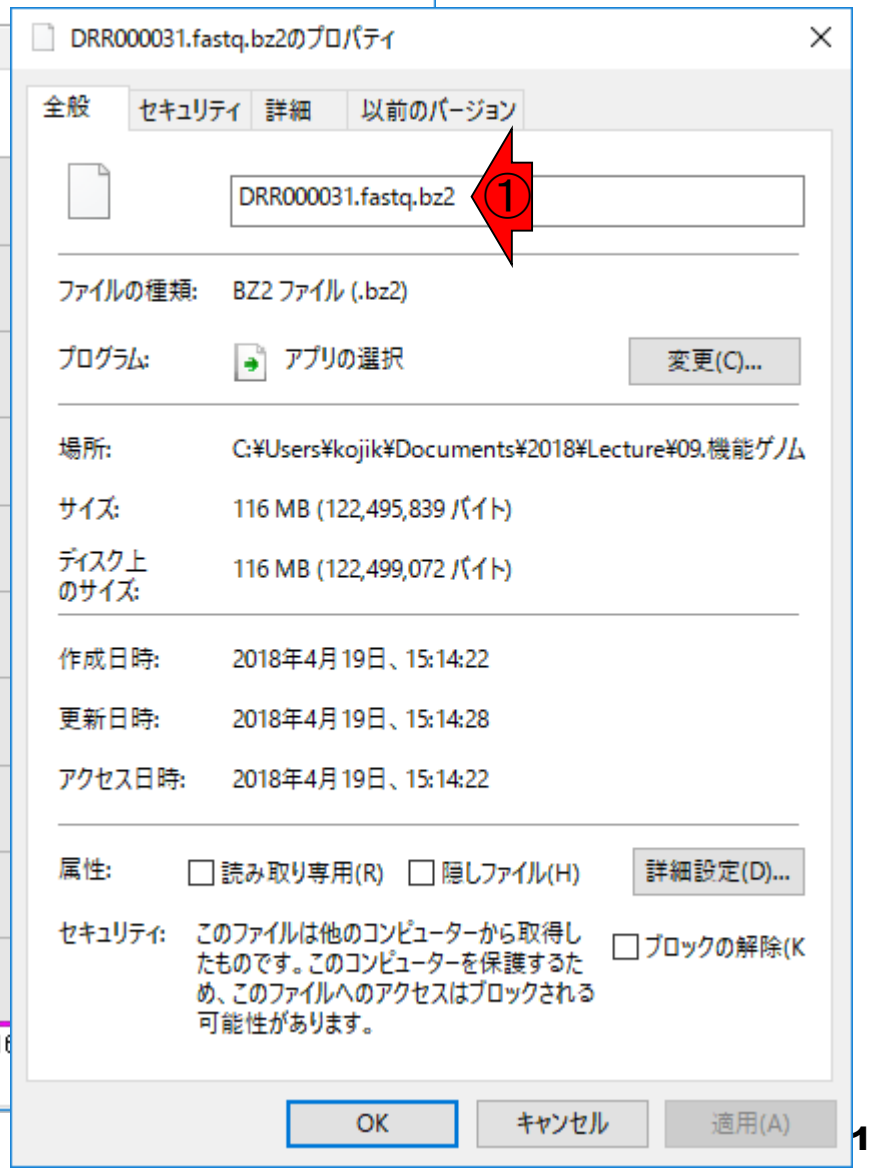
Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

圧縮ファイルも可

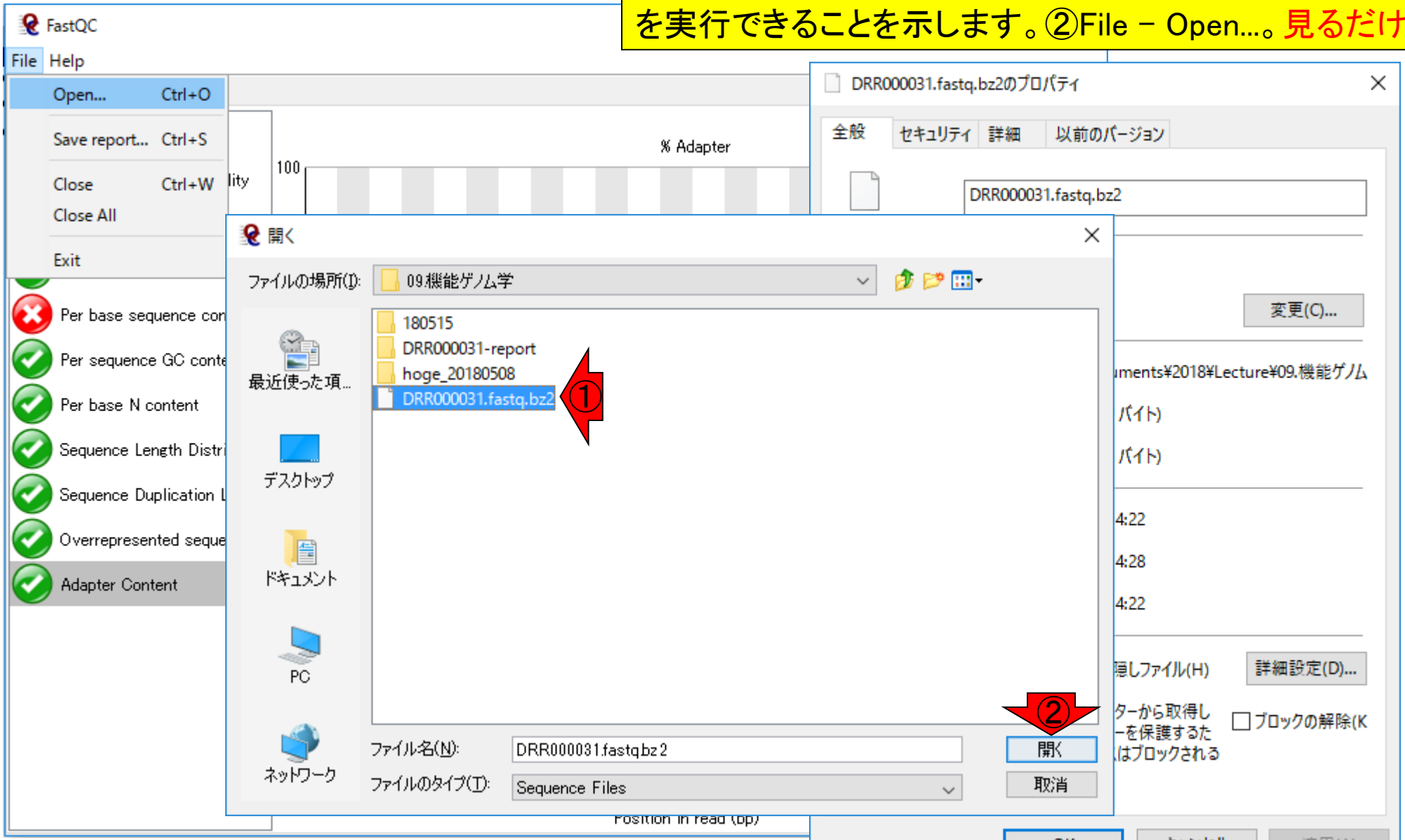
FastQCは、bz2やgzなど圧縮ファイルも入力として受け付けてくれます。ここでは、DRR000031sub.fastqの大元のファイルである①DRR000031.fastq.bz2を入力としてFastQCを実行できることを示します。②File – Open…。見るだけ!





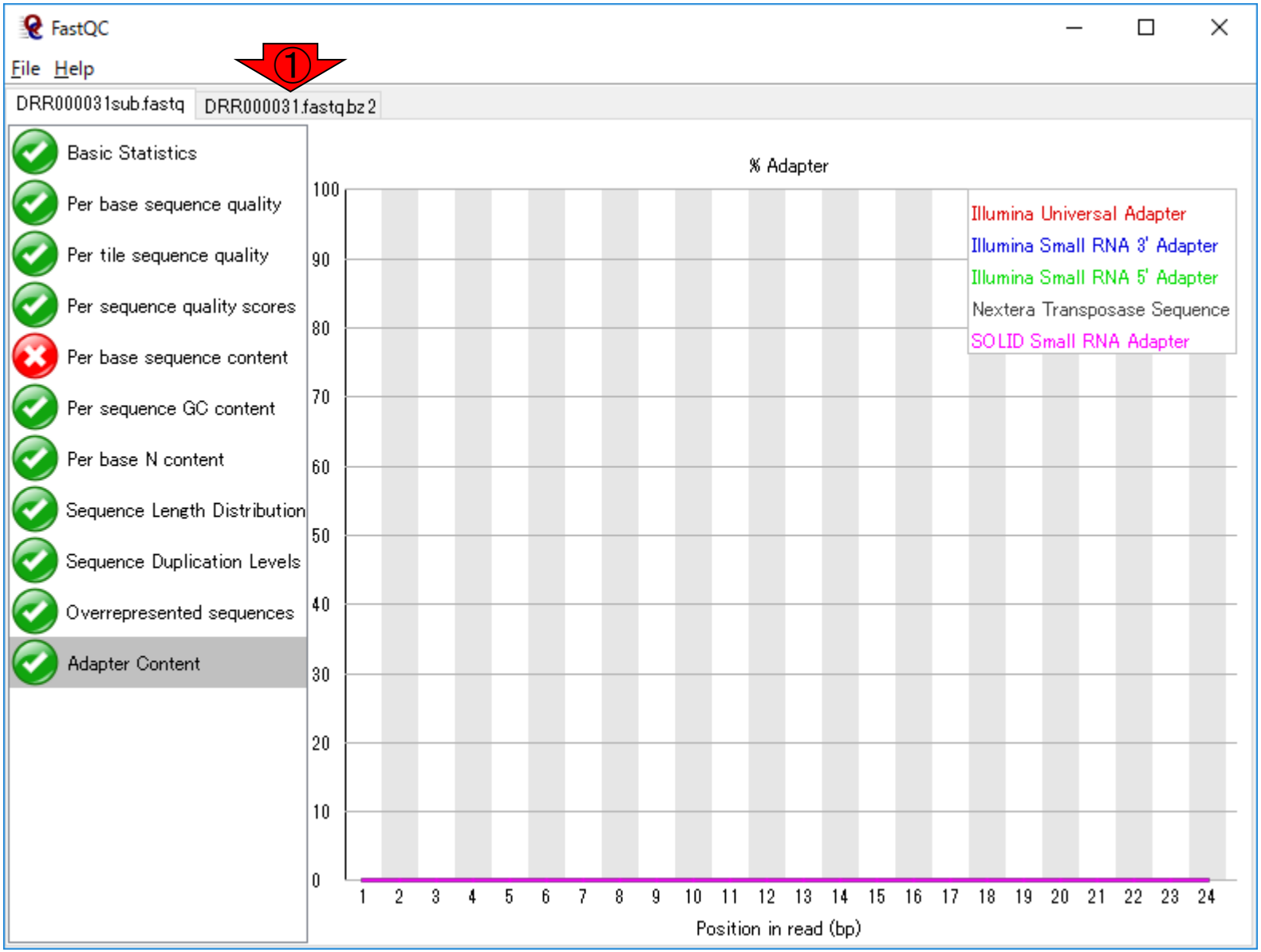
圧縮ファイルも可

FastQCは、bz2やgzなど圧縮ファイルも入力として受け付けてくれます。ここでは、DRR000031sub.fastqの大元のファイルである①DRR000031.fastq.bz2を入力としてFastQCを実行できることを示します。②File – Open…。見るだけ!



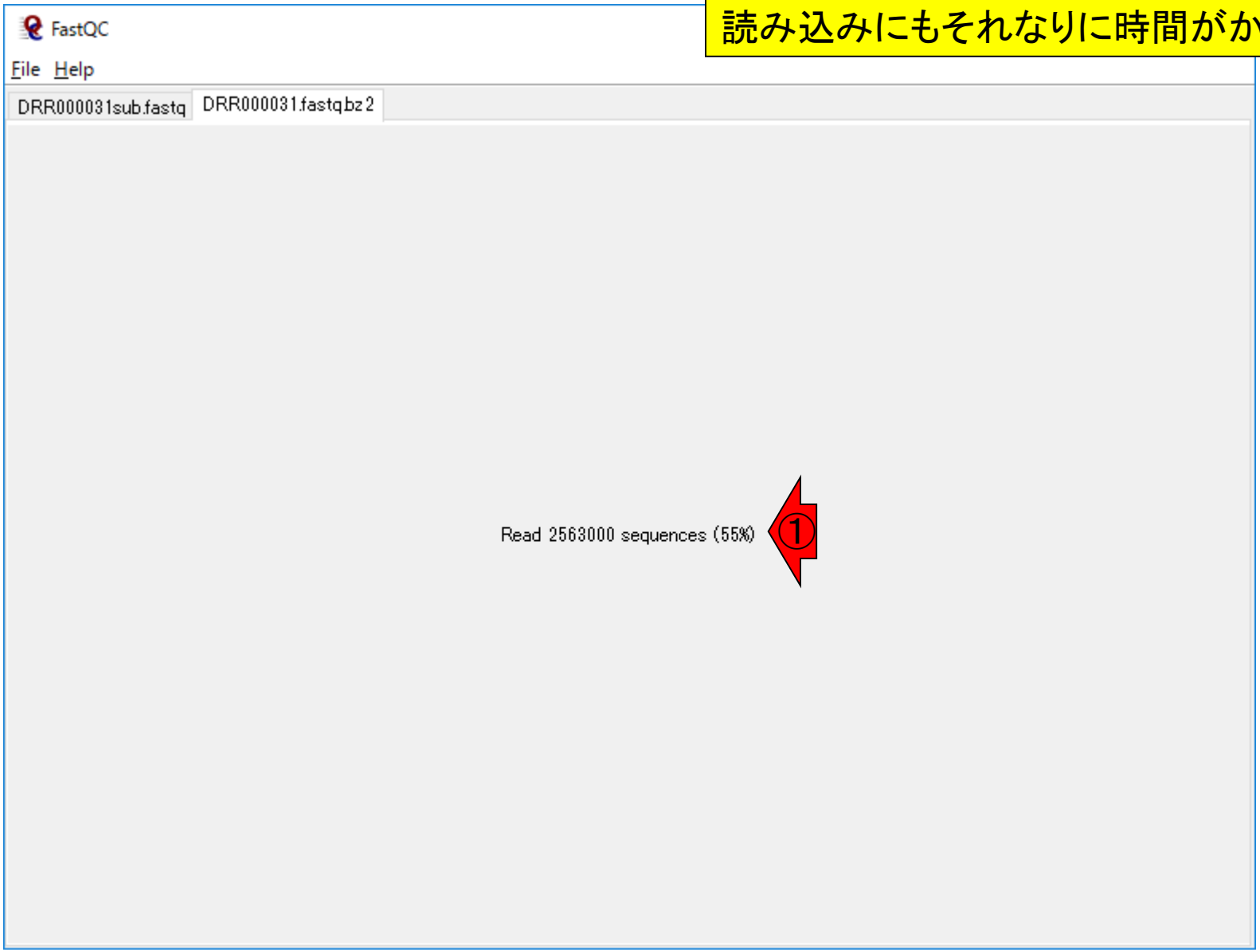
圧縮ファイルも可

一見何も変化ないようですが、①DRR000031.fastq.bz2という新しいタブができています。クリック



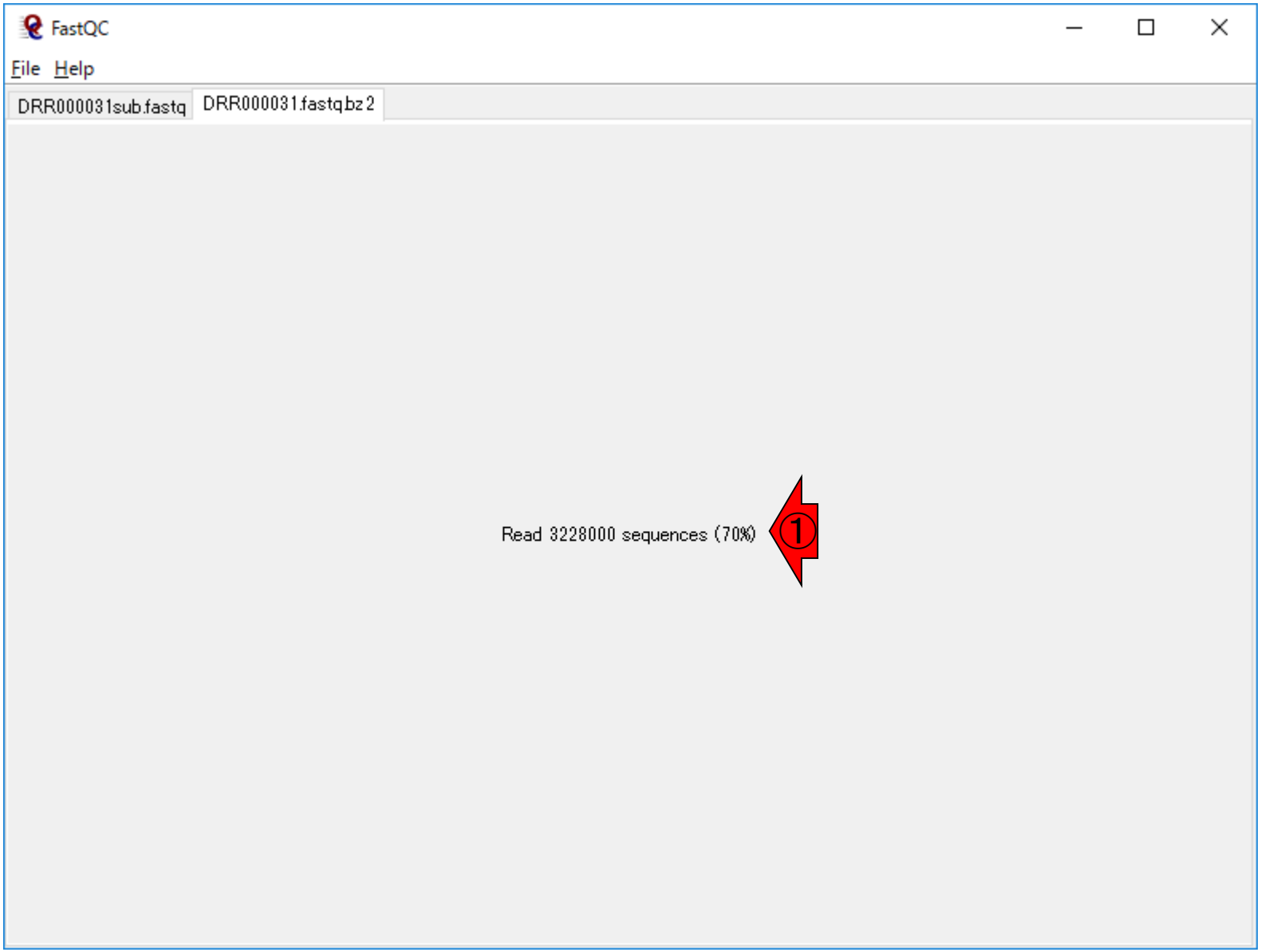
圧縮ファイルも可

一瞬なんじゃこりゃ!となりますが、よく見る①Read 2563000 sequences (55%)と書かれています。確かにこのFASTQファイルは4,589,774リード(約460万)あるので、読み込みにもそれなりに時間がかかるのだろうと納得



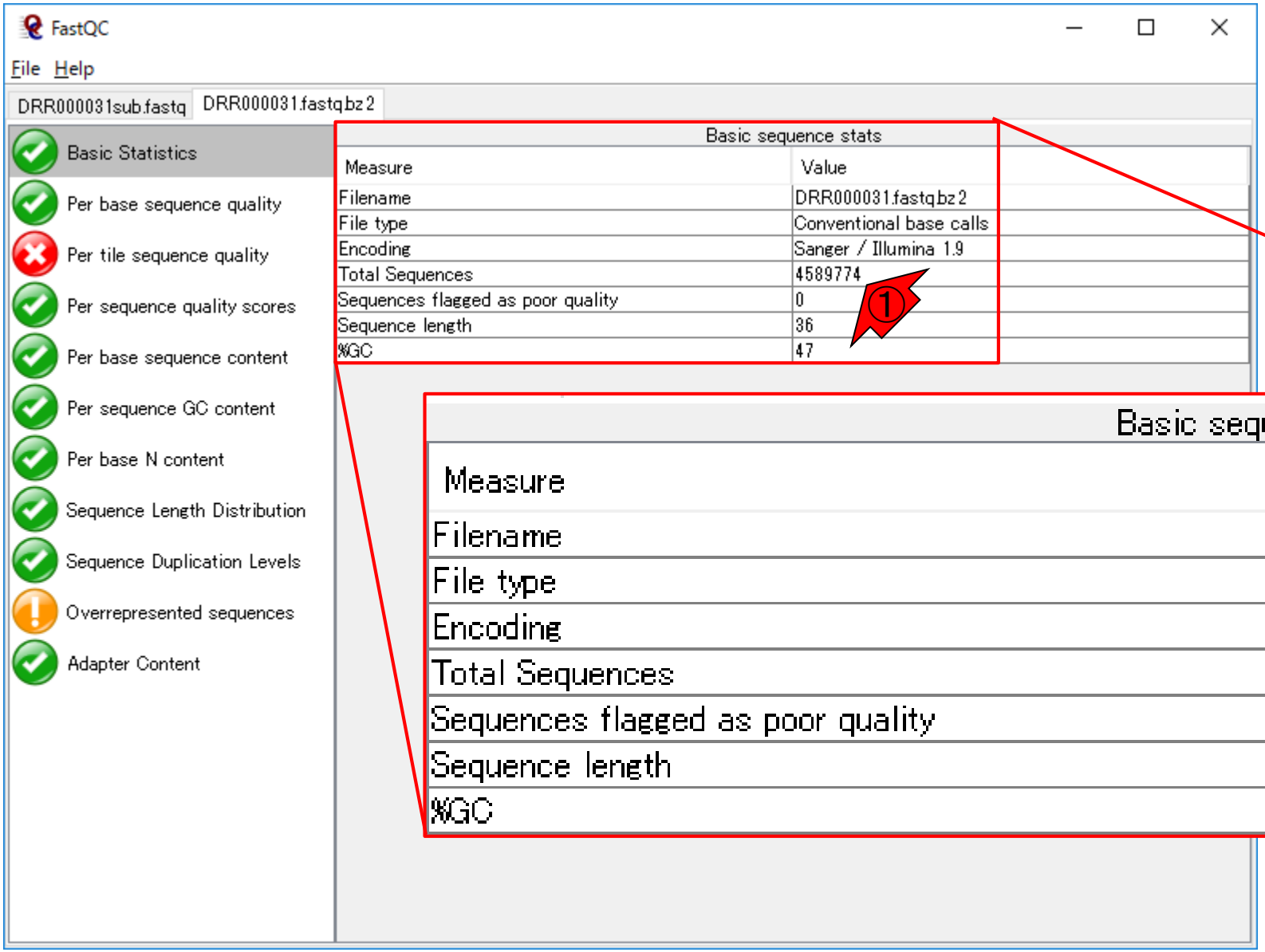
圧縮ファイルも可

さらに数十秒後には、①Read 3228000 sequences (70%)
となっていました。順調に読み込めているようですね



圧縮ファイルも可

このときは、数分程度で完了しました。
①4,589,774リードであり、妥当ですね

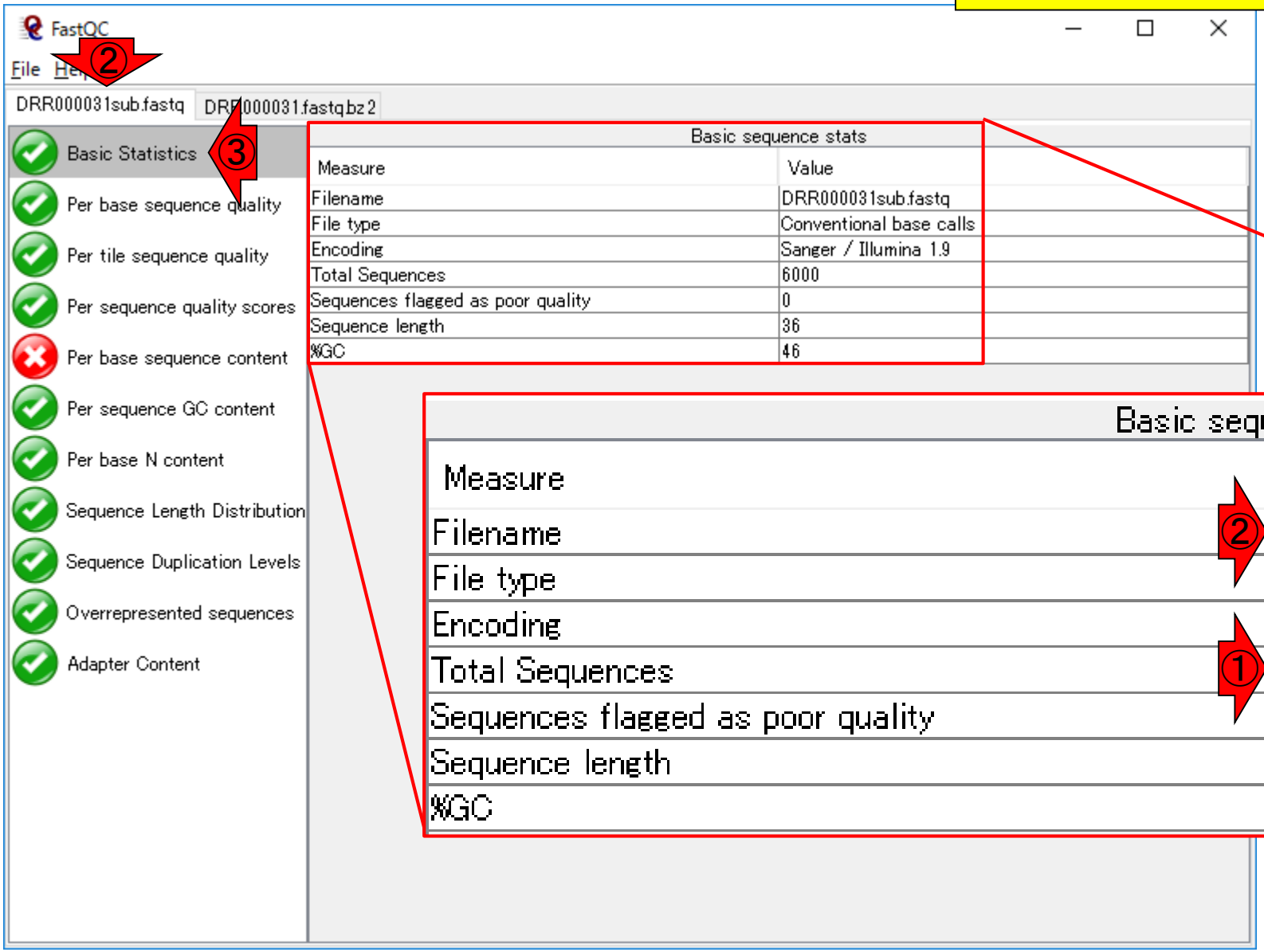


Basic sequence stats

Measure	Value
Filename	DRR000031.fastq.bz2
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4589774
Sequences flagged as poor quality	0
Sequence length	36
%GC	47

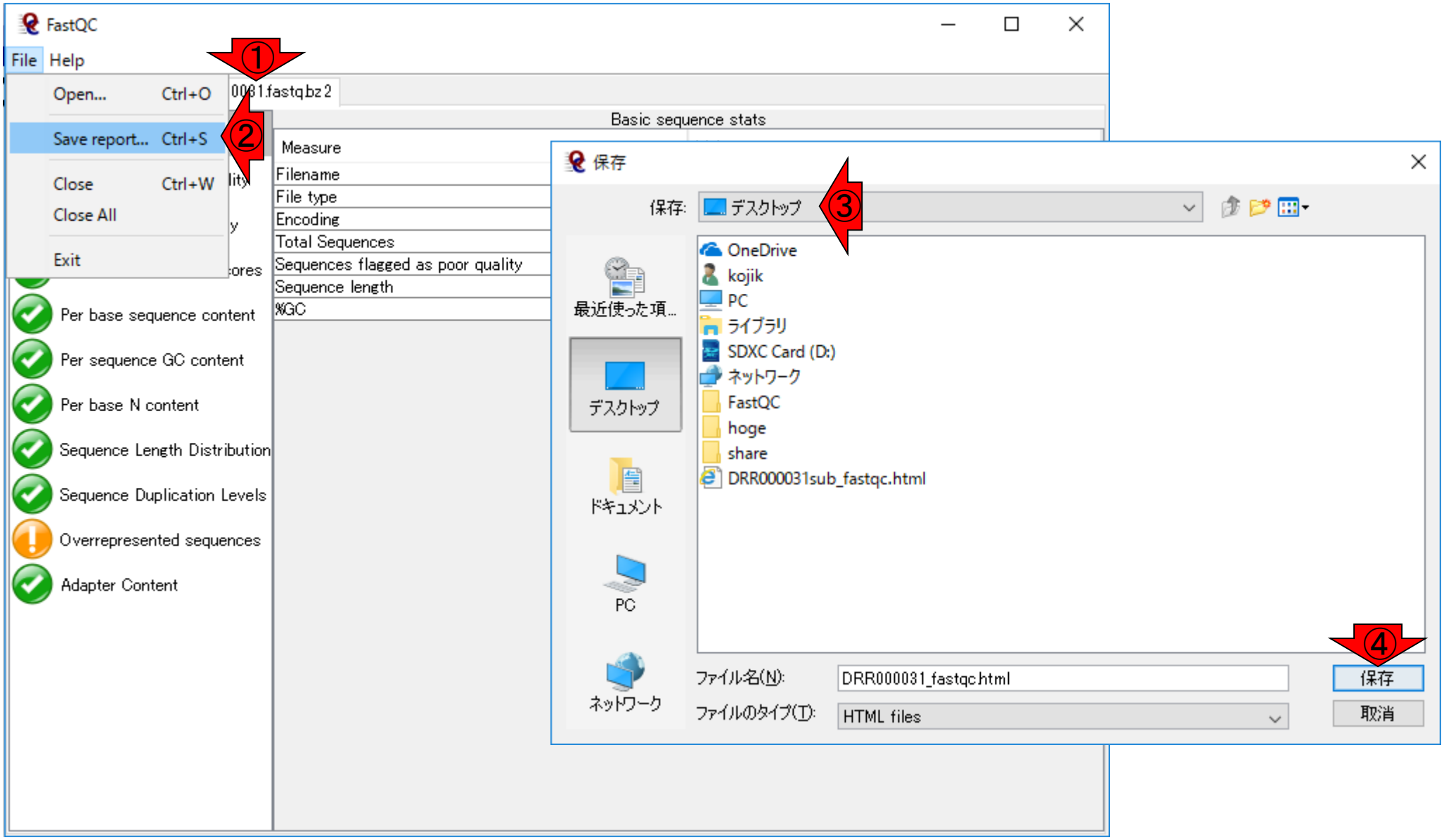
比較（460万 vs. 6千）

①6,000リードの、②DRR000031sub.fastqの結果との、③Basic Statisticsの比較。
④%GCもわずかに異なりますね



htmlファイルで保存

①DRR000031.fastq.bz2を②「File – Save report…」、③デスクトップに、④保存して得られたものが…



①DRR000031_fastqc.htmlです

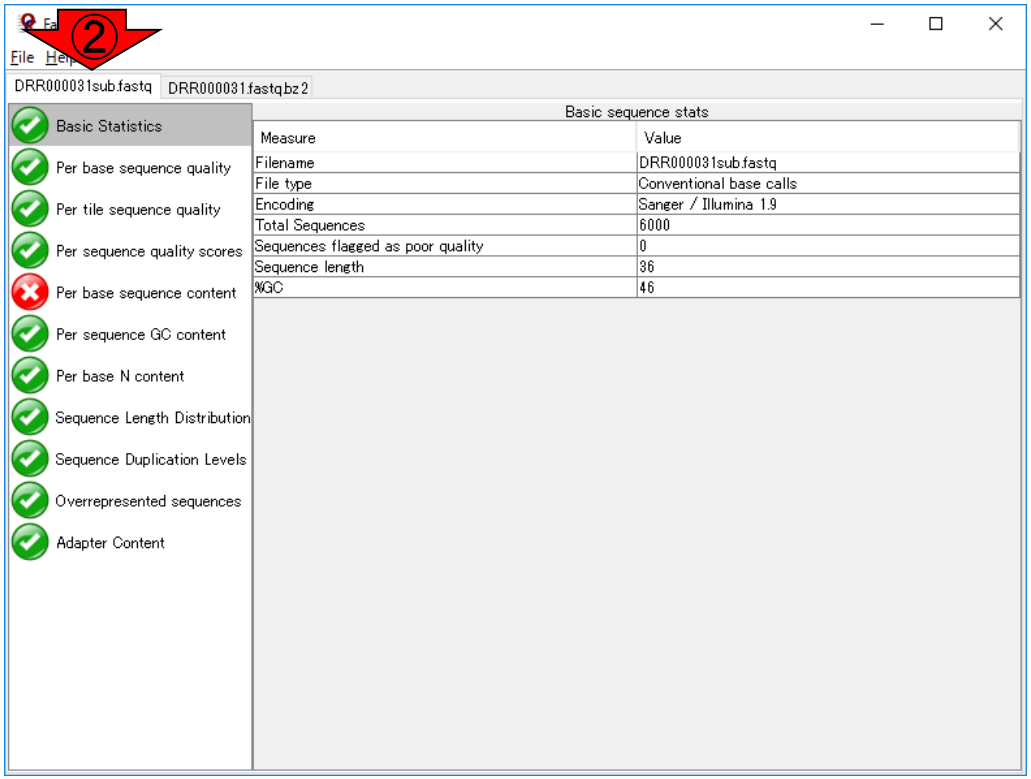
DRR000031_fastqc.html

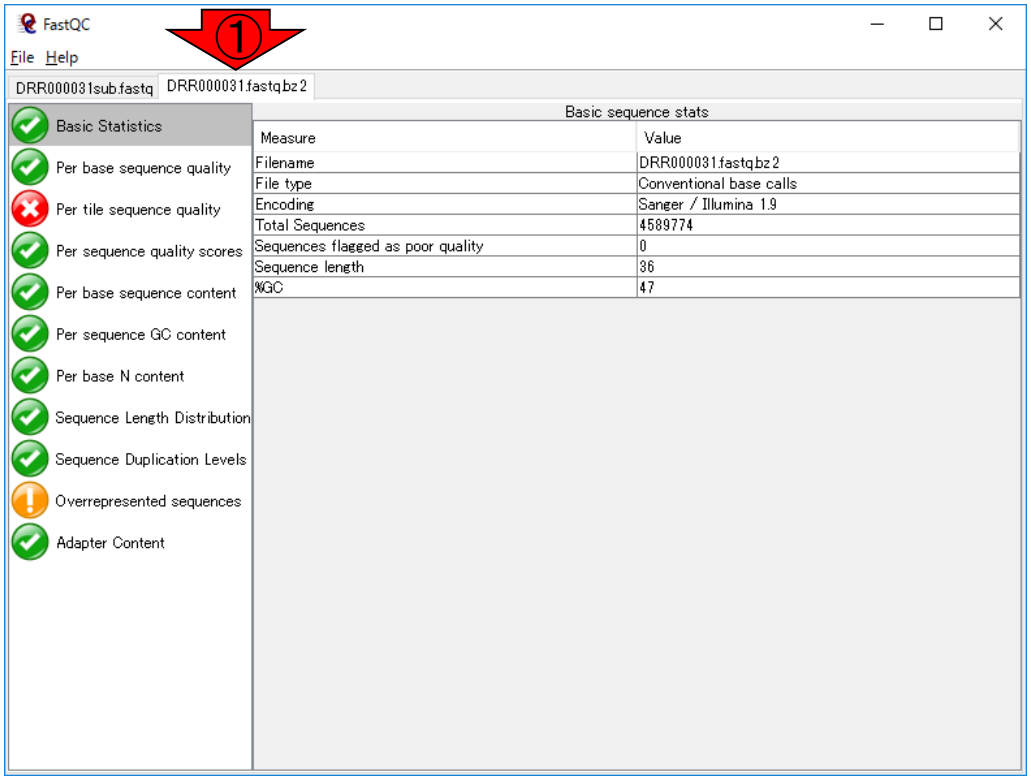
講義日程（2019年度）

1. 2019年05月27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)
   - Maser : Kinjo et al., Database (Oxford), 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
   - 講義資料PDF
   - (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)

課題

①4,589,774リードからなるDRR000031.fastq.bz2のFastQC実行結果について、②6,000リードからなるDRR000031sub.fastqのFastQC実行結果を比較対象として考察せよ

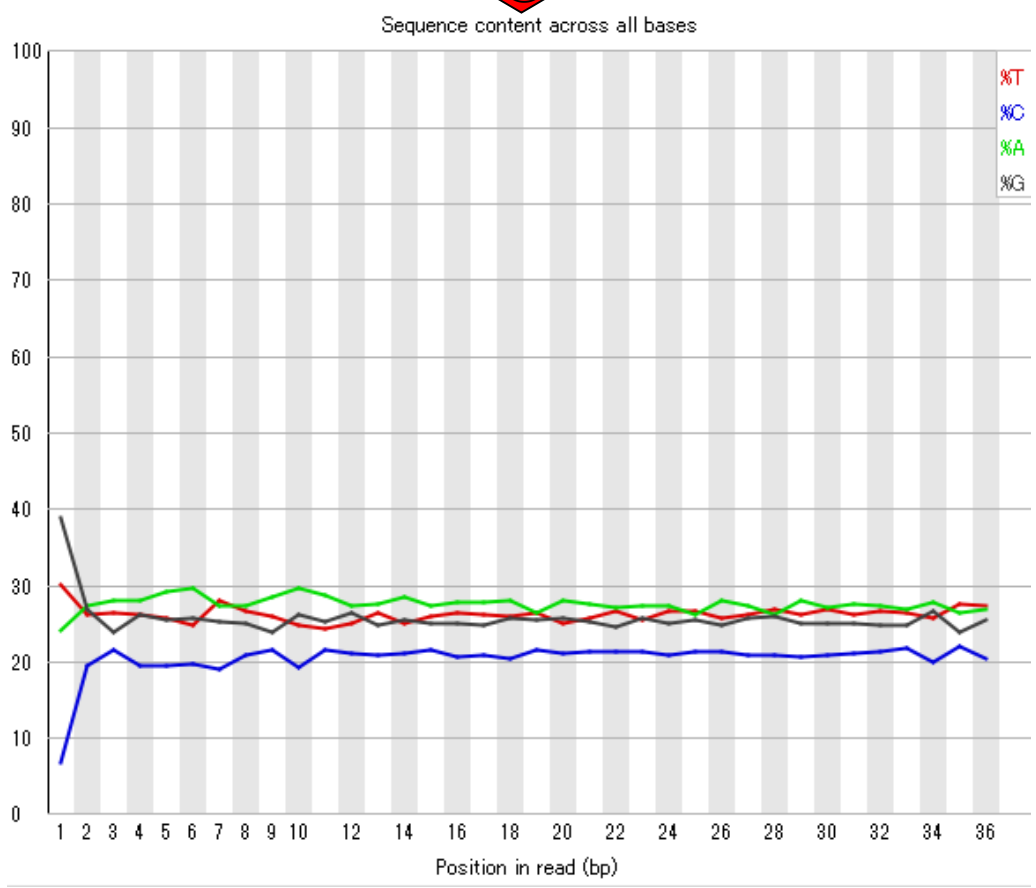




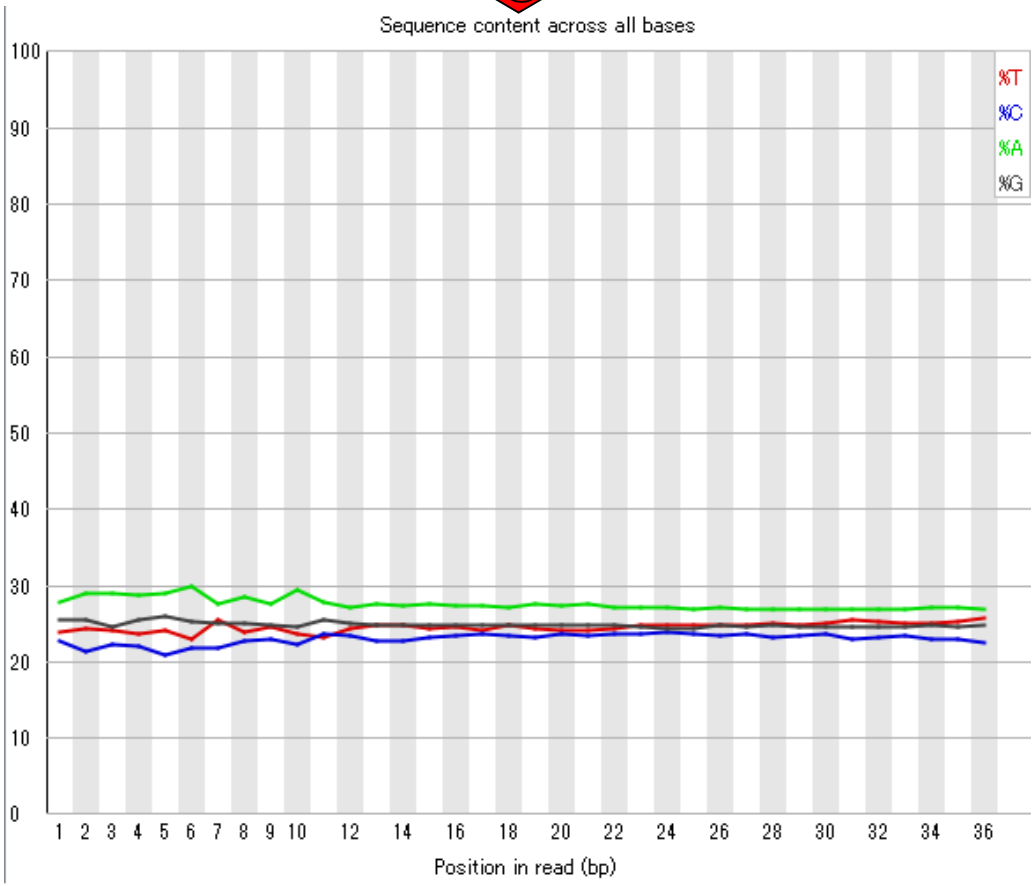
課題：論点1

FastQC実行結果のPer base sequence content。入力ファイルが①4,589,774リードからなるDRR000031.fastq.bz2と、②6,000リードからなるDRR000031sub.fastqの違いについて簡単に考察

②



①



課題：論点2

FastQC実行結果のOverrepresented sequences。入力ファイルが①4,589,774リードからなるDRR000031.fastq.bz2と、②6,000リードからなるDRR000031sub.fastqの違いについて簡単に考察

②

Overrepresented sequences

There are no overrepresented sequences

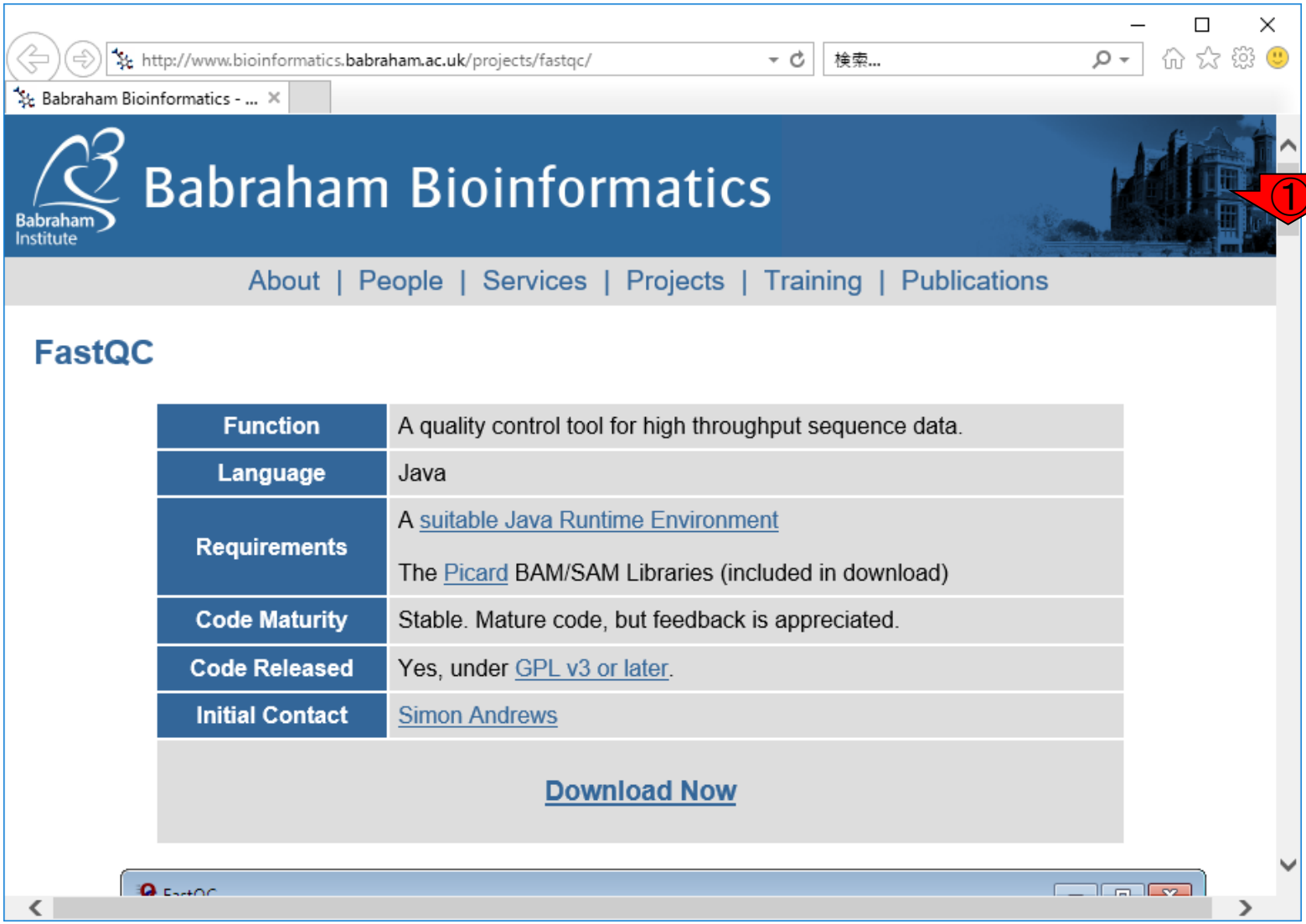
①

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAAAAAAAAAAAAAAAA...	41007	0.893	No Hit

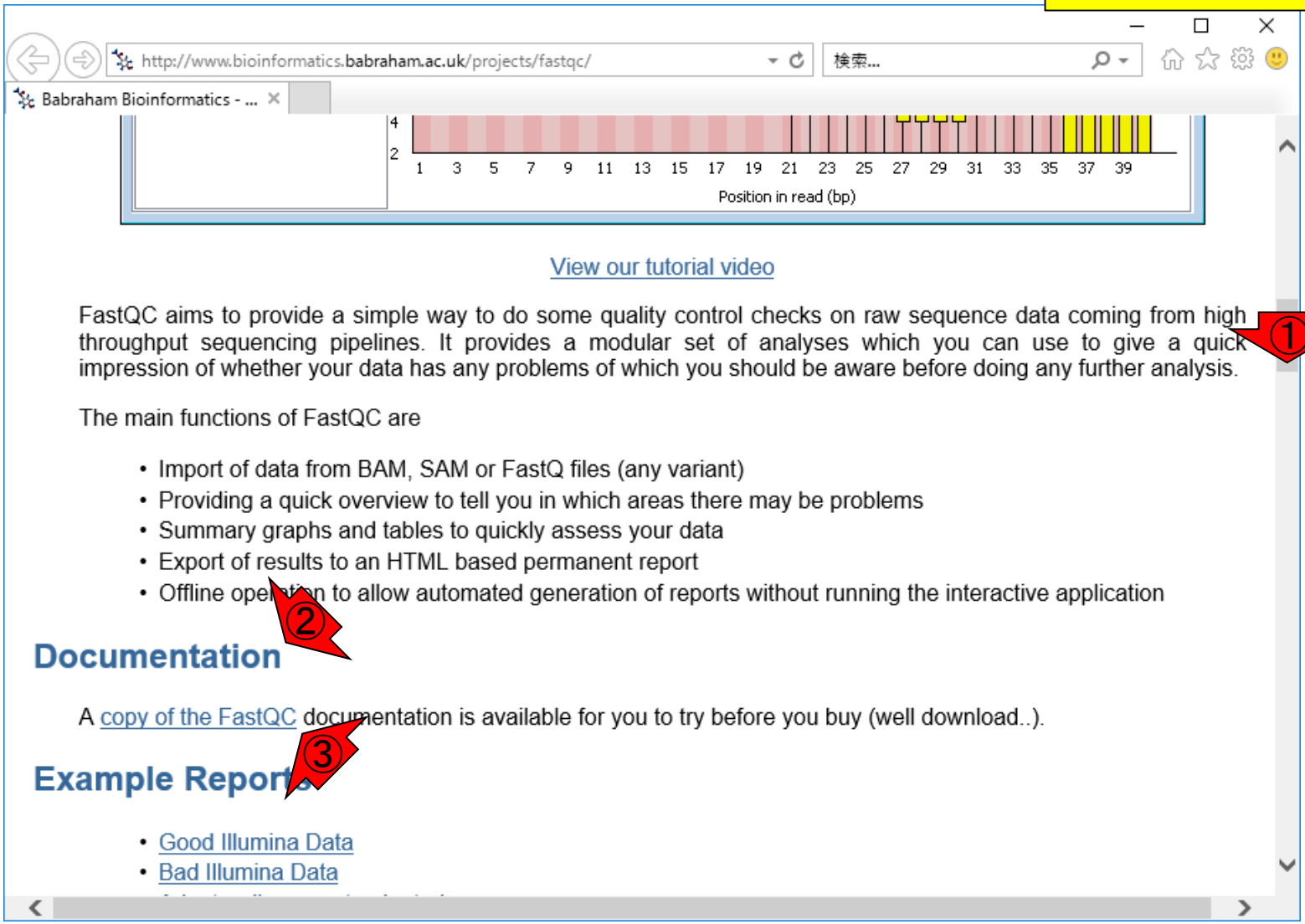
FastQC実行結果に関する説明は、①ページ下部の…

参考情報



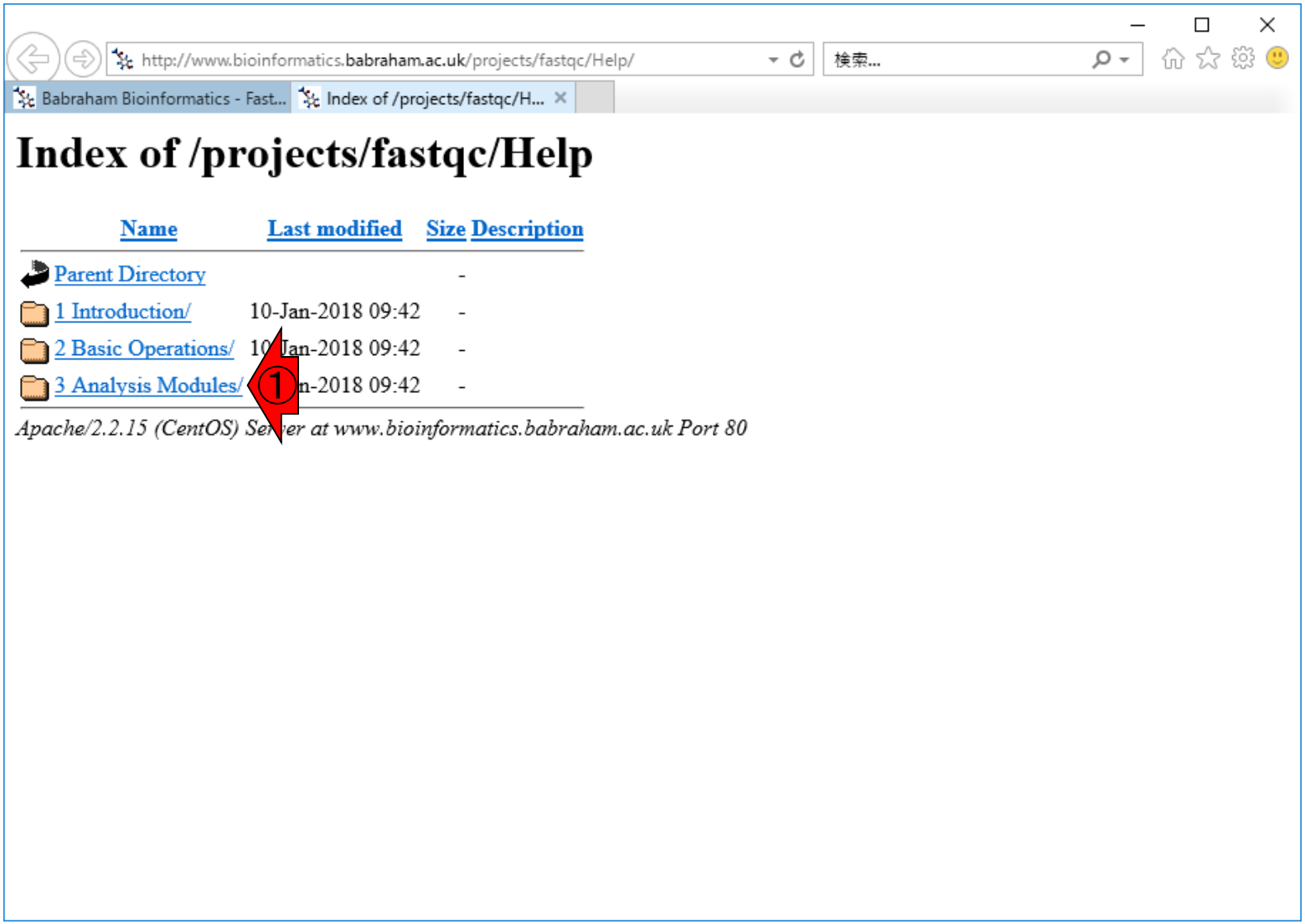
参考情報

①このあたりまで下がったところにある、②Documentationのすぐ下にある、③copy of the FastQCのリンク先の…



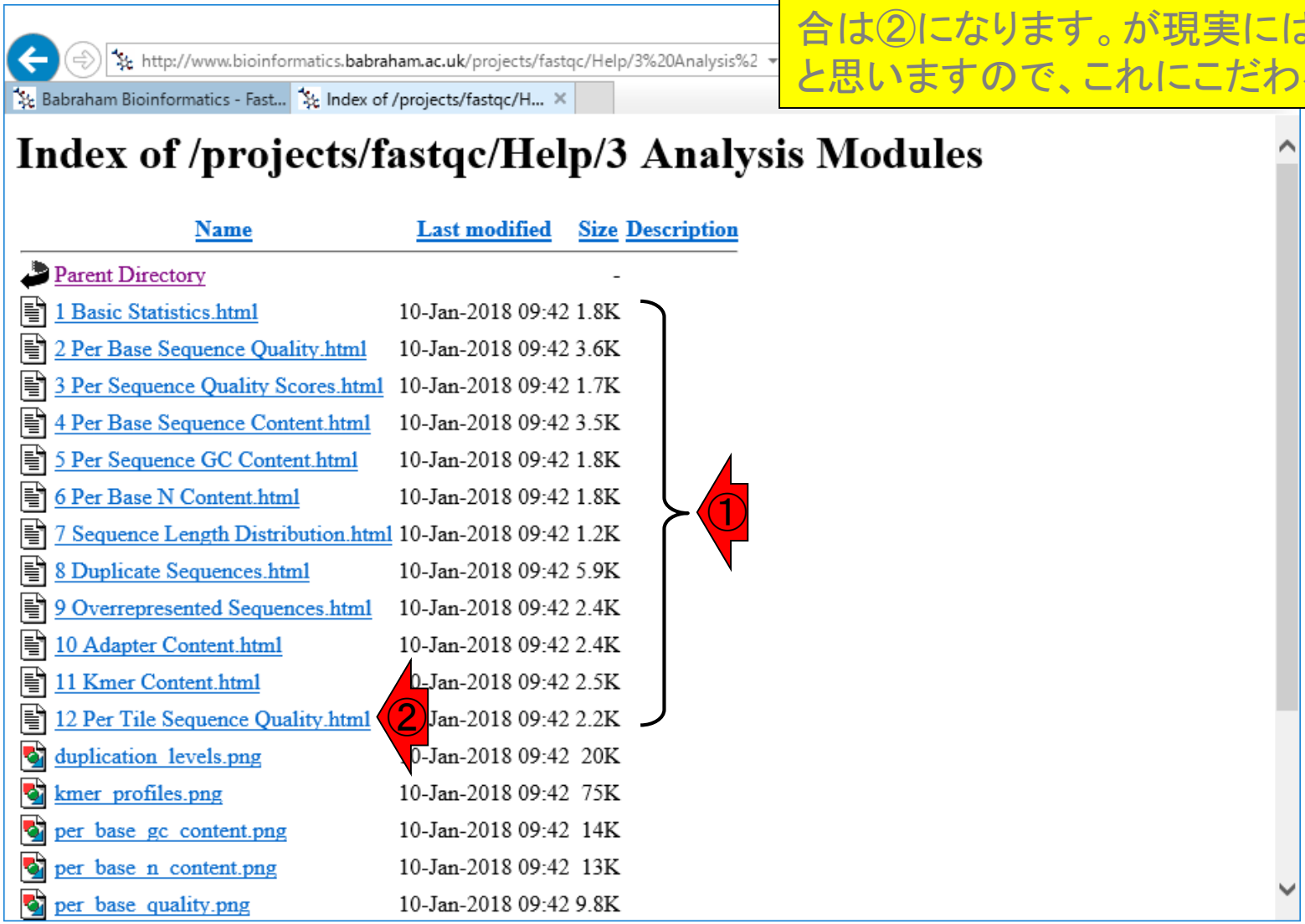
①3 Analysis Modulesのリンク先の…

参考情報



参考情報

①このあたりです。例えばDRR000031.fastq.bz2とDRR000031sub.fastqのFastQC実行結果で違いのあったPer tile sequence qualityの見方について知りたい場合は②になります。が現実には読んでも理解しづらいと思いますので、これにこだわる必要は全くありません



FastQC

FastQC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。例えば、RNA-QC-chainなどのプログラムを実行した結果のFASTQファイルをさらにFastQC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

NGSリードデータ（SRAファイル）

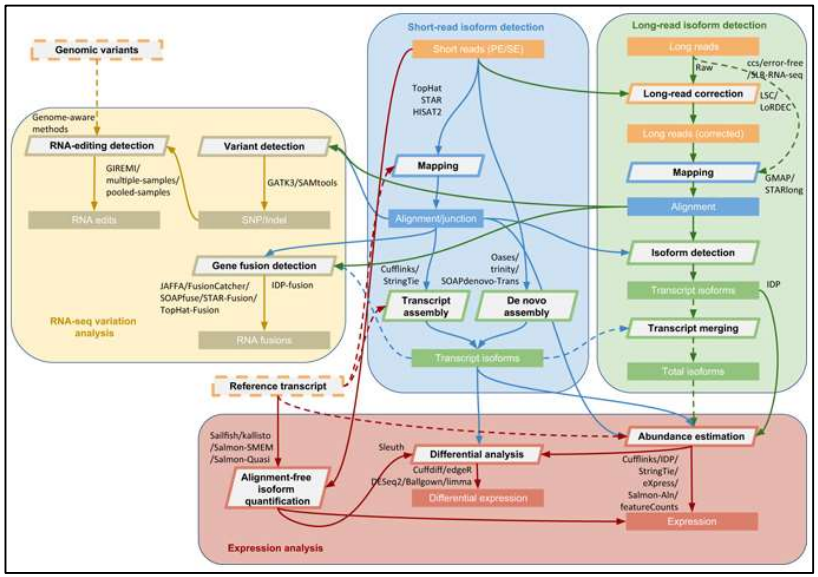
⇩

NGSリードデータ（FASTQファイル）

⇩

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

⇩

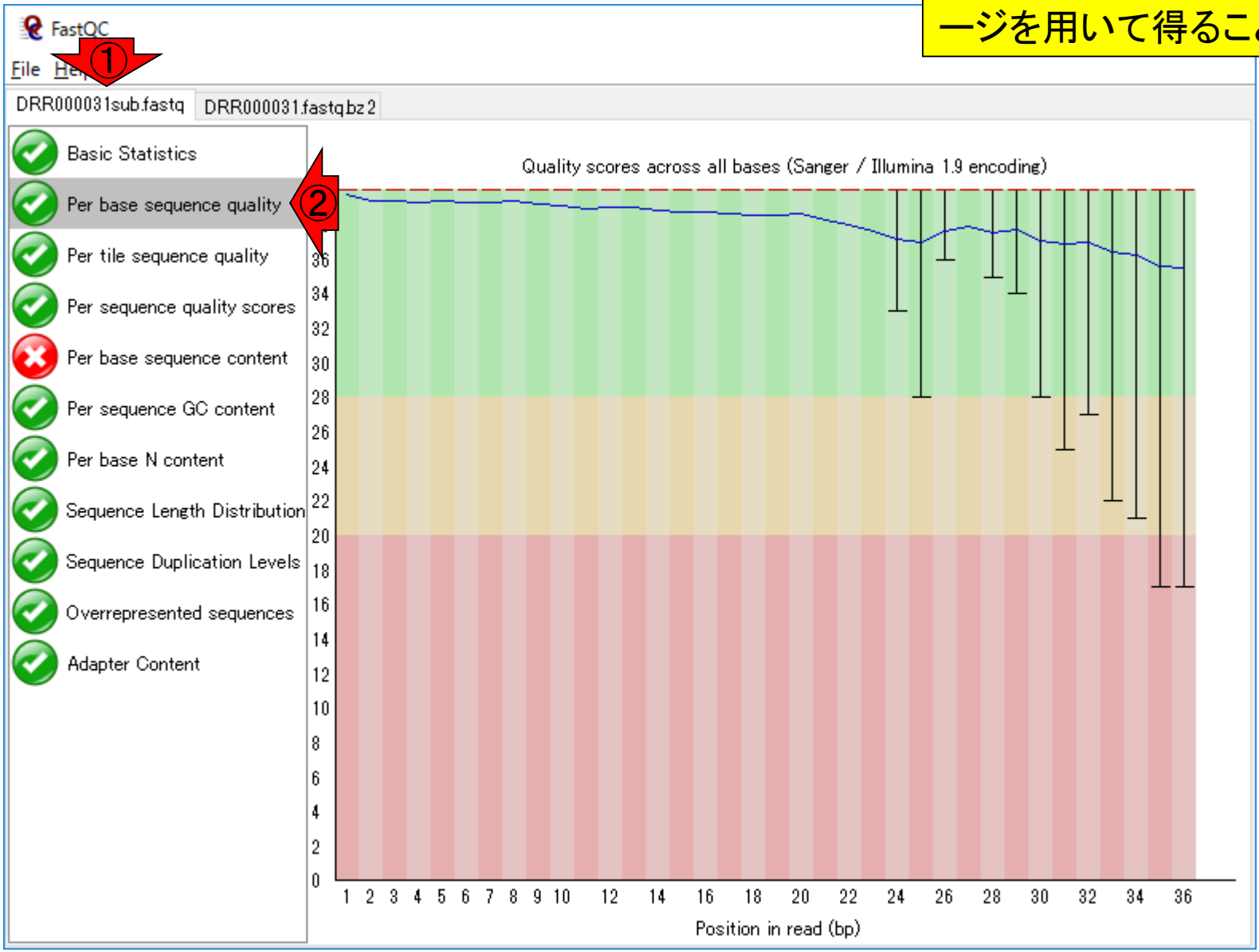


Contents

- 公共DB関連のTips
  - 公共DB、Bio-Linux、WSLのインストール、Linux上でbzip2ファイルの解凍
  - FASTQファイルの説明、リード数の違い
  - ウェブツール、ウェブブラウザに注意
- 前処理（Preprocessing） or Quality Control （QC）
  - RNA-QC-chain
  - FastQCのインストールと実行
  - FastQC実行結果の解説
  - 圧縮ファイルでFastQC、課題
  - Rパッケージqrqcでクオリティチェック

Rでクオリティチェック

①6,000リードからなるDRR000031sub.fastqのFastQC実行結果と似たような結果（例えば②Per base sequence quality）をRパッケージを用いて得ることが一応できます



Rでクオリティチェック

①qrqcというRパッケージ。圧縮ファイルを取り扱えないなどの制約があるため、実用的とは言い難いですが、6,000リードからなるDRR000031sub.fastqの結果程度であれば問題なく動きます。②例題1の入力ファイル名部分を変更して実行

(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 | 指定した末端塩基
- 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:

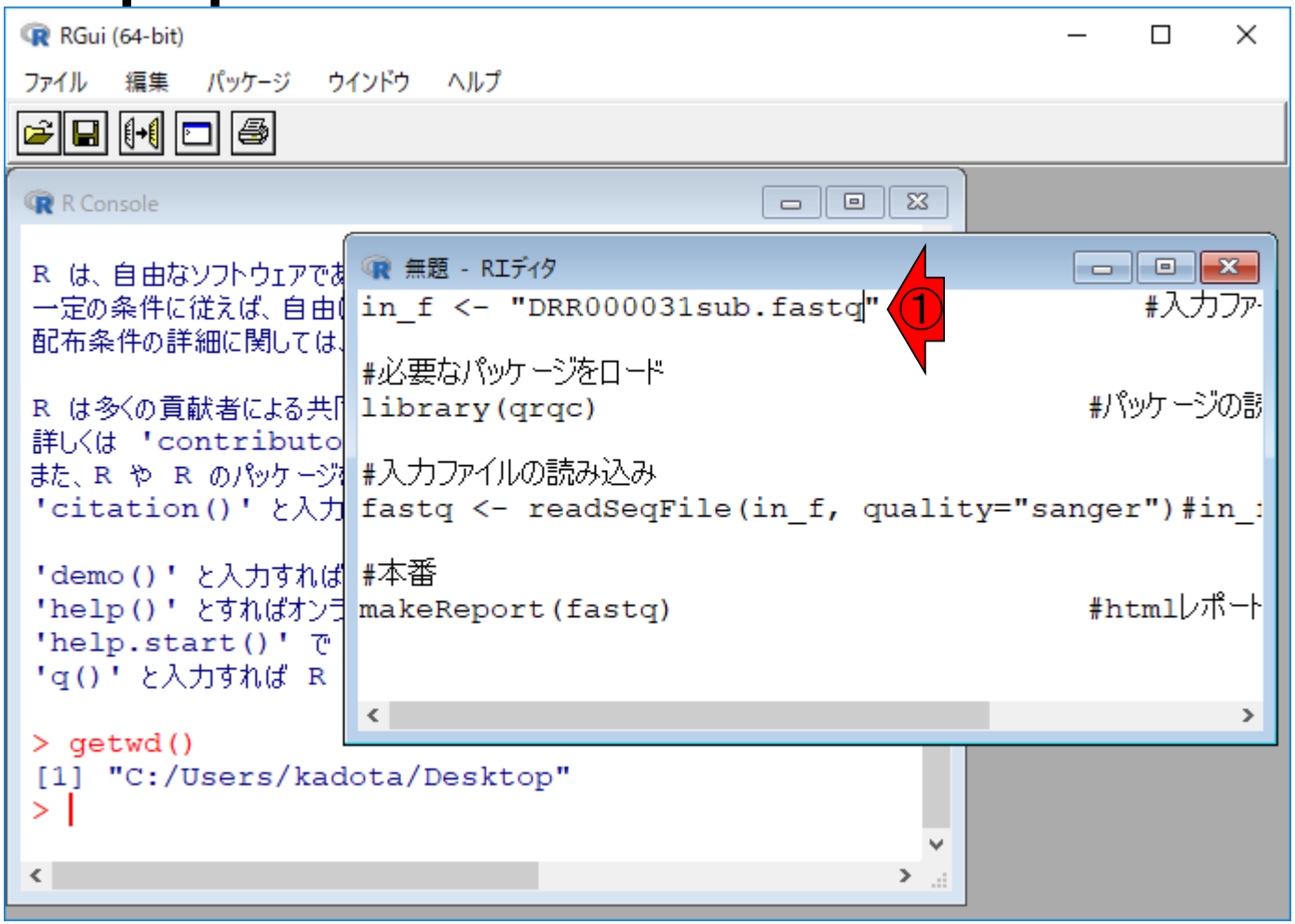
SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです (Bullard et al., 2010)。下記を実行すると「SRR037439-report」という名前のフォルダが作成されます。中にあるreport.htmlをダブルクリックするとhtmlレポートを見ることができます。

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
```

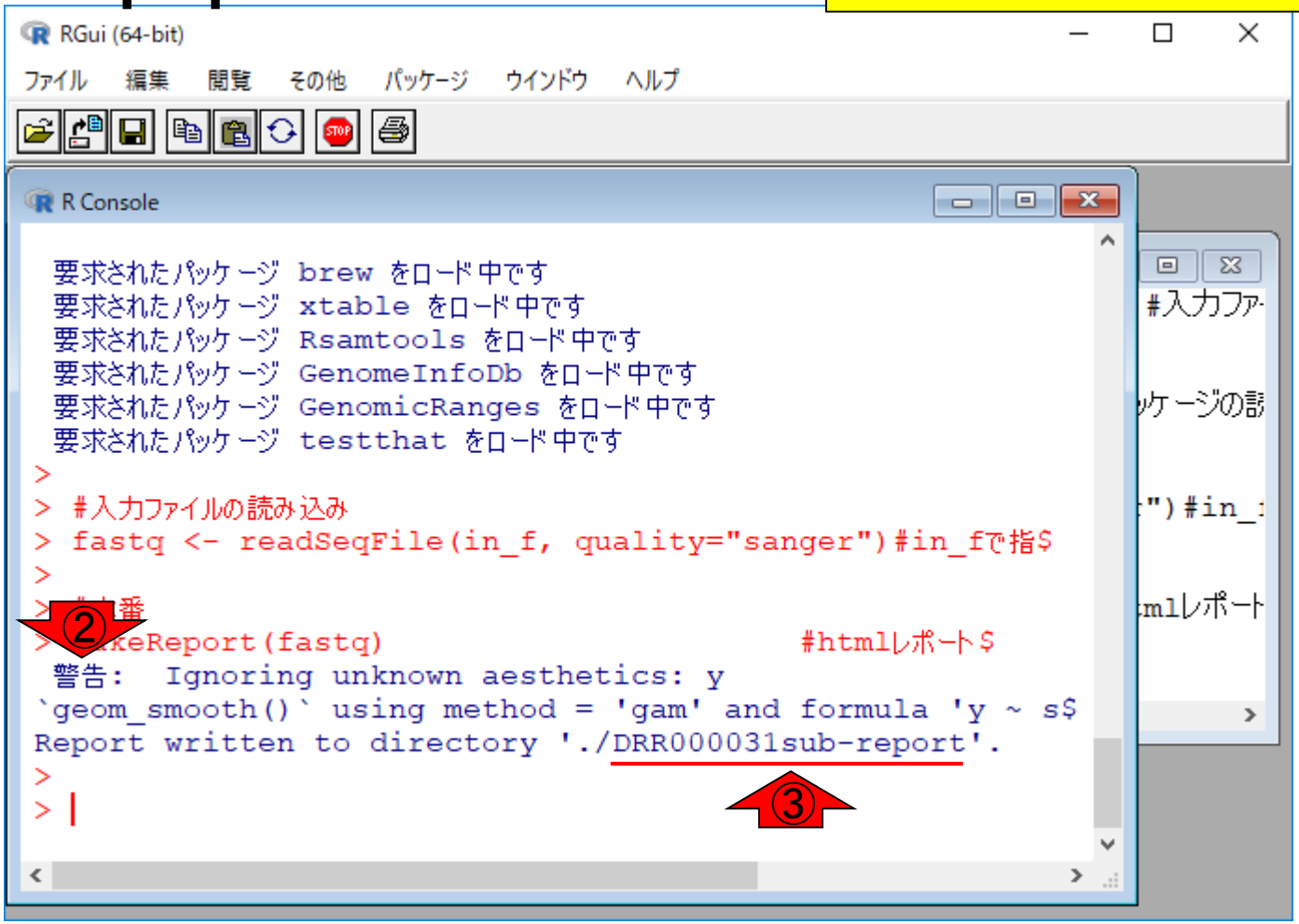
①入力ファイル名部分を変更してコピペ実行。

qrqc実行



qrqc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



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

講義日程（2019年度）

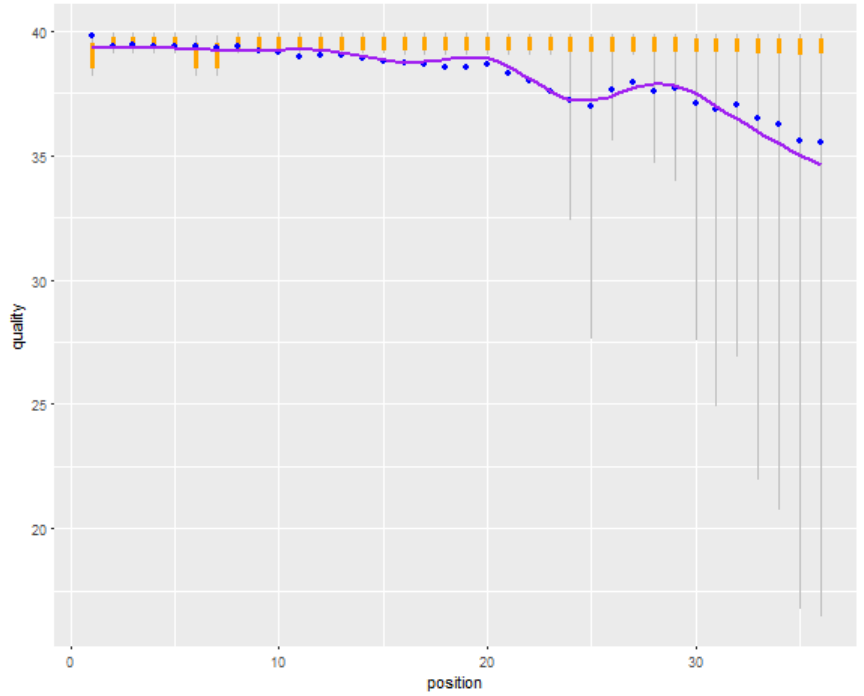
1. 2019年05月27日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)
   - Maser : Kinjo et al., Database (Oxford), 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
   - 講義資料PDF
   - (R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)

General Information

File: DRR000031sub.fastq
Type: FASTQ
Sequence Length Range: 36 to 36
Total Sequences: 6000
Unique Sequences: 5368

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