

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、Linux
 - 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qrrcでクオリティチェック

公共DB

NGSデータの公共DBは、①DDBJ SRA、②NCBI SRA、③EMBL-EBI 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](#) | [一般](#) | [配列取得](#) | [トランスクリプトーム配列](#) | [GenomicFeatures\(Lawrence 2009\)](#)
- ・ [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 2015/11/15)
- ・ [イントロ](#) | [一般](#) | [読み込み](#) | [xlsx形式](#) | [openxlsx](#) (last modified 2015/11/15)
- ・ [イントロ](#) | [NGS](#) | [様々なプラットフォーム](#) (last modified 2016/03/24)
- ・ [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 2015/02/23)
- ・ [イントロ](#) | [NGS](#) | [配列取得](#) | [FASTQ or SRA](#) | [SRADB\(Zhu 2013\)](#) (last modified 2015/02/24)
- ・ [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 2014/03/26)

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

- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)

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ー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見玉さんありがとうございますm(_ _)m)。

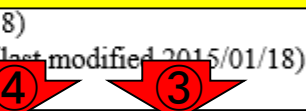
データの形式は基本的に [Sanger type](#) のFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです([Cock et al., Nucleic Acids Res., 2010](#))。しかし少なくとも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](#))。

- ① [DDBJ Sequence Read Archive \(DRA\): Kodama et al., Nucleic Acids Res., 2012](#)
 - ② [EMBL-EBI European Nucleotide Archive \(ENA\): Eddy et al., Nucleic Acids Res., 2015](#)
 - ③ [NCBI Sequence Read Archive \(SRA\): NCBI Resource Coordinators., Nucleic Acids Res., 2014](#)
- ・ [ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015](#)
 - ・ [GEO: Barrett et al., Nucleic Acids Res., 2013](#)
 - ・ [DBCLS SRA: Nakazato et al., PLoS One, 2013](#)

.sraから.fastqを作成

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

- ・ イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | [GenomicFeatures\(Lawrence 2009\)](#)
- ・ 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 2015/11/15)
- ・ イントロ | 一般 | 読み込み | xlsx形式 | [openxlsx](#) (last modified 2015/11/15)
- ・ イントロ | NGS | [様々なプラットフォーム](#) (last modified 2016/03/24)
- ・ 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 2015/02/23)
- ・ イントロ | NGS | 配列取得 | FASTQ or SRA | [SRADB\(Zhu 2013\)](#) (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 | [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 2014/03/26)



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

- ・ イントロ | NGS | アノテーション情報取得
- ・ イントロ | NGS | アノテーション情報取得
- ・ イントロ | NGS | アノテーション情報取得
- ・ [イントロ | NGS | アノテーション情報取得](#)
- ・ イントロ | NGS | アノテーション情報取得
- ・ イントロ | NGS | アノテーション情報取得

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ー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見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)。

- ① [DDBJ Sequence Read Archive \(DRA\): Kodama et al., Nucleic Acids Res., 2012](#)
- ② [EMBL-EBI European Nucleotide Archive \(ENA\): Ewings et al., Nucleic Acids Res., 2015](#)
- ・ [NCBI Sequence Read Archive \(SRA\): NCBI Resource Coordinators., Nucleic Acids Res., 2014](#)
- ・ [ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015](#)
- ・ [GEO: Barrett et al., Nucleic Acids Res., 2013](#)
- ・ [DBCLS SRA: Nakazato et al., PLoS One, 2013](#)

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

DRA000011

Browser address: <http://ddbj.nig.ac.jp/DRASearch/>

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show records Sort by

Data Last Update 2018-04-25

Statistics

Released Entries

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

Organism			Study Type			Center Name		
#	Organism Name	Study	#	Study Type	Study	#	Center Name	Study
1	Homo sapiens	13078	1	Other	50118	1	BioProject	80643
2	Mus musculus	10555	2	Whole Genome Sequencing	49553	2	GEO	23275
3	soil metagenome	3961	3	Metagenomics	19706	3	DOE - JOINT GENOME INSTITUTE	2590
4	marine metagenome	1690	4	Transcriptome Analysis	19161	4	UMIGS	2557
5	Arabidopsis thaliana	1688	5	Population Genomics	791	5	JGI	2364
6	Panicum virgatum	1557	6	Epigenetics	705	6	WUGSC	1398
7	Drosophila melanogaster	1547	7	Exome Sequencing	248	7	JCVI	1148
8	Oryza sativa	1505	8	Transcriptome Sequencing	170	8	BI	962
9	Populus trichocarpa	1187	9	Cancer Genomics	133	9	SC	903
10	Saccharomyces cerevisiae	1185	10	Pooled Clone Sequencing	35	10	The Wellcome Trust Sanger Institute	759

<http://ddbj.nig.ac.jp/DRASearch/query> of Japan

Last modified: Sep. 06, 2017 (V3.2)

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

DRASearch Search Home DRA Home

DRA000011 FTP

Submission Detail	
Alias	DRA000011
Submission ID	
Submission Date	2009-08-17
Center Name	UT-MGS
Lab Name	Laboratory of Functional Genomics, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo

Navigation	
Study	DRP000011
Experiment	DRX000011 FASTQ SRA
Sample	DRS000011
Run	DRR000031 FASTQ SRA

Website policy | © DNA Data Bank of Japan

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いでいいという話です。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

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

DRA000011 - DRA Search DRR000031 - DRA Search

DRASearch

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	
Experiment	DRX000011	FASTQ SR

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

```
>DRR000031.1
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
>DRR000031.2
TGTAATGATTATGATTCTCAGGGATTGGGAAAGGT
>DRR000031.3
TCAAAAAATACGAAAGTTAGGGTGACAAAGTTTGACA
>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
>DRR000031.5
TTCAGGAAGCTGGTGATGGAGCACCAGAGGGTGGT
>DRR000031.6
TTTTTTTAAAATAGCACTTTAAATATTTATTTTTTT
>DRR000031.7
GGAGGGTTAATTCAGGCGAGTATACTAACTTAAGG
>DRR000031.8
TTATCATCTTCACAATTCTAATNNNACTGACTATCC
>DRR000031.9
TTTTAAATGTAAATTTTTTATTTGGAAAACAAATAT
>DRR000031.10
TGGTAACAGCCTGATGGGTTATTTGACTGCACTAAG
```

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

FTP ディレクトリ / ddbj_database/dra/fastq/DRA000/DRA000011/DRX000011 / ftp.c

エクスプローラーでこの FTP サイトを表示するには、Alt キーを押して、[表示]をクリックし、[エクスプローラーで FTP サイトを開く]

-Welcome to DDBJ FTP Archive, running on ftp.ddbj.nig.ac.jp!
-Please contact ddbj@ddbj.nig.ac.jp when you have any problem for getting access to this archive, downloading the data, and etc.
-For details on the directory structure and file contents, please refer to the README.TXT placed in the "ddbj_database".

[1階層上のディレクトリへ](#)

06/06/2014 12:00午前	122,495,839	DRR000031.fastq.bz2
--------------------	-------------	-------------------------------------

DRR000031.fastq.bz2のプロパティ

全般 セキュリティ 詳細 以前のバージョン

DRR000031.fastq.bz2

ファイルの種類: BZ2 ファイル (.bz2)

プログラム: アプリの選択 変更(C)...

場所: C:\Users\kojik\Documents\2018\Lecture\09.機能ゲム

サイズ: 116 MB (122,495,839 バイト)

ディスク上のサイズ: 116 MB (122,499,072 バイト)

作成日時: 2018年4月19日、15:14:22

更新日時: 2018年4月19日、15:14:28

アクセス日時: 2018年4月19日、15:14:22

属性: 読み取り専用(R) 隠しファイル(H) 詳細設定(D)...

セキュリティ: このファイルは他のコンピューターから取得したものです。このコンピューターを保護するため、このファイルへのアクセスはブロックされる可能性があります。 ブロックの解除(K)

OK キャンセル 適用(A)

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

FTP ディレクトリ / ddbj_database/dra/fastq/DRA000/DRA000011/DRX000011

-Welcome to DDBJ FTP Archive, running on ftp.ddbj.nig.ac.jp!
-Please contact ddbj@ddbj.nig.ac.jp when you have any problem for getting access to this archive, downloading the data, and etc.
-For details on the directory structure and file contents, please refer to the README.TXT placed in the "ddbj_database".

[1階層上のディレクトリへ](#)

06/06/2014 12:00午前 122,495,839 [DRR000031.fastq.bz2](#) ①

DRR000031.fastq

ファイルの種類: FASTQ ファイル (.fastq) ②

プログラム: アプリの選択 変更(C)...

場所: C:\Users\%kojik%\Documents\2018\Lecture\09.機能ゲム

サイズ: 781 MB (819,218,014 バイト) ③

ディスク上のサイズ: 781 MB (819,220,480 バイト)

作成日時: 2018年4月19日、15:15:54

更新日時: 2018年4月19日、15:16:27

アクセス日時: 2018年4月19日、15:15:54

属性: 読み取り専用(R) 隠しファイル(H) 詳細設定(D)...

OK キャンセル 適用(A)

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrcでクオリティチェック

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

- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌](#) | [第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | [日本乳酸菌学会誌](#) | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 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 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/25)
- 書籍 | [日本乳酸菌学会誌](#) | [第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | [日本乳酸菌学会誌](#) | [第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第11回統合データベース](#)
- 書籍 | [日本乳酸菌学会誌](#) | [第12回Galaxy:ビ](#)

- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2015/11/26)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2015/11/26)
- イントロ | 一般 | [任意のキーワードを含む行を抽出](#) (last modified 2015/11/26)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2015/11/26)

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

Linux

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls
DRR000031.fastq.bz2
iu@bielinux[mac_share] █
```

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

[12:01午後]

Linux

①ls -(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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls [12:01午後]
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l [12:01午後]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] █ [12:12午後]
```



Linux

①ls -(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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l
total 119625
-rwxrwxrwx 1 iu iu 122495839  4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] █
```

12:01午後

DRR000031.fastq.bz2のプロパティ

全般 セキュリティ 詳細 以前のバージョン

DRR000031.fastq.bz2

ファイルの種類: BZ2 ファイル (.bz2)

プログラム: アプリの選択 変更(C)...

場所: C:\Users\kajok\Documents\2018\Lecture\09.機能ゲム

サイズ: 116 MB (122,495,839 バイト) **③**

ディスク上のサイズ: 116 MB (122,499,072 バイト)

作成日時: 2018年4月19日、15:14:22

更新日時: 2018年4月19日、15:14:28

アクセス日時: 2018年4月19日、15:14:22

属性: 読み取り専用(R) 隠しファイル(H) 詳細設定(D)...

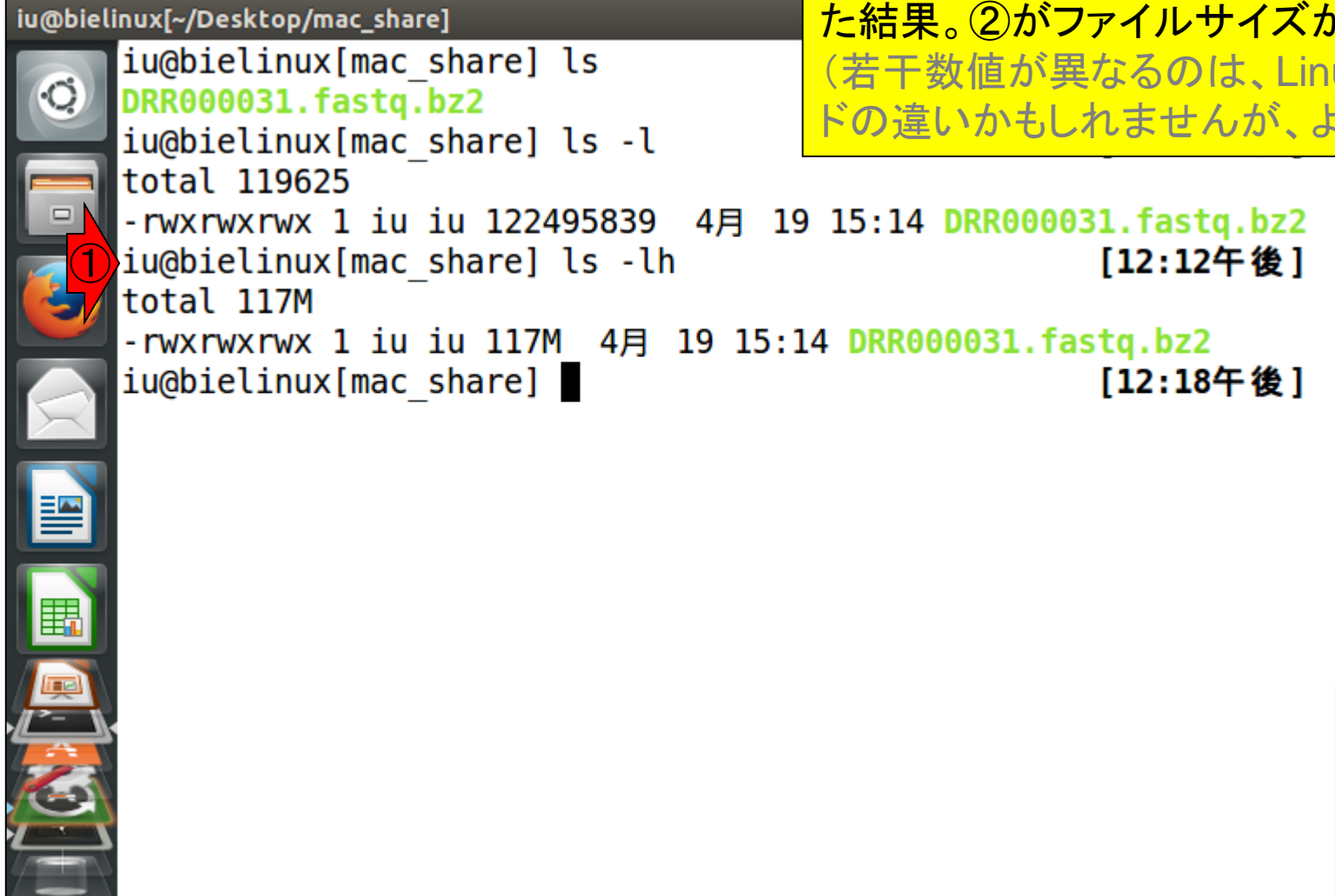
セキュリティ: このファイルは他のコンピューターから取得したものです。このコンピューターを保護するため、このファイルへのアクセスはブロックされる可能性があります。 ブロックの解除(K)

OK キャンセル 適用(A)

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)

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l
total 119625
-rwxrwxrwx 1 iu iu 122495839  4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh
total 117M
-rwxrwxrwx 1 iu iu 117M  4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] █
```



Linux

① bunzip2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

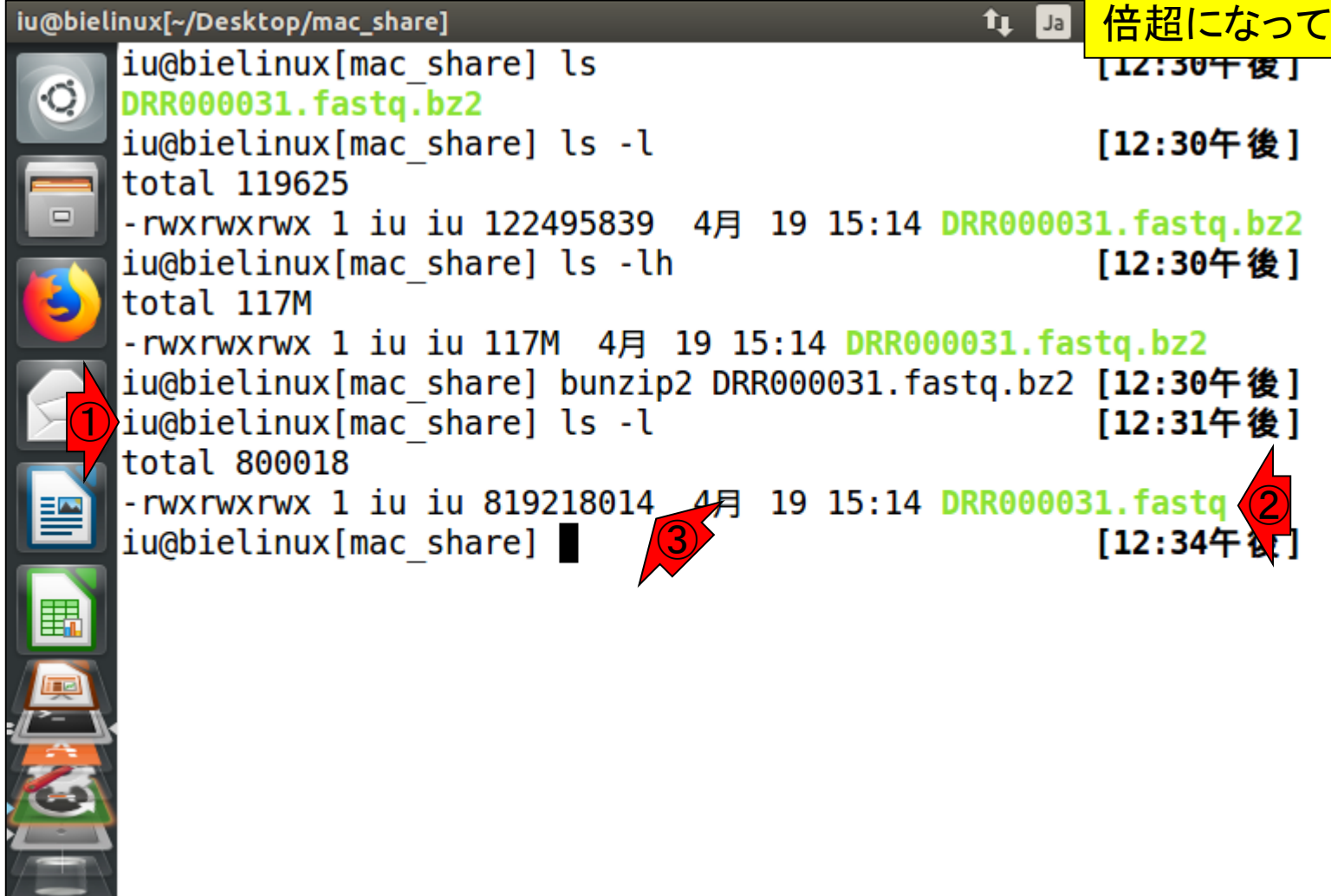
```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls [12:30午後]
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l [12:30午後]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh [12:30午後]
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
① iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] [12:31午後]
```


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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
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.fastq
iu@bielinux[mac_share] █
```

[12:30午後]
[12:30午後]
[12:30午後]
[12:30午後]
[12:31午後]
[12:31午後]
[12:34午後]



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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l [12:30午後]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh [12:30午後]
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] ls -l [12:31午後]
total 800018
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
① iu@bielinux[mac_share] wc DRR000031.fastq [12:34午後]
18359096 36718192 819218014 DRR000031.fastq
② ③ ④ iu@bielinux[mac_share] [ 1:20午後]
```

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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls [12:30午後]
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l [12:30午後]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh [12:30午後]
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] ls -l [12:31午後]
total 800018
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
iu@bielinux[mac_share] wc DRR000031.fastq [12:34午後]
18359096 36718192 819218014 DRR000031.fastq
① iu@bielinux[mac_share] head -n 2 DRR000031.fastq [ 1:20午後]
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
iu@bielinux[mac_share] [ 1:46午後]
```

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

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] ls [12:30午後]
DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -l [12:30午後]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh [12:30午後]
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] ls -l [12:31午後]
total 800018
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
iu@bielinux[mac_share] wc DRR000031.fastq [12:34午後]
18359096 36718192 819218014 DRR000031.fastq
iu@bielinux[mac_share] head -n 2 DRR000031.fastq [ 1:20午後]
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATA...
iu@bielinux[mac_share] head -n 24000 DRR000031.fastq > DRR000031sub.fastq
iu@bielinux[mac_share] [ 2:28午後]
```

①

②

③

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

```
iu@bielinux[~/Desktop/mac_share]
total 119625
-rwxrwxrwx 1 iu iu 122495839  4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] ls -lh [12:30午後]
total 117M
-rwxrwxrwx 1 iu iu 117M  4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] ls -l [12:31午後]
total 800018
-rwxrwxrwx 1 iu iu 819218014  4月 19 15:14 DRR000031.fastq
iu@bielinux[mac_share] wc DRR000031.fastq [12:34午後]
18359096  36718192 819218014 DRR000031.fastq
iu@bielinux[mac_share] head -n 2 DRR000031.fastq [ 1:20午後]
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATA...
iu@bielinux[mac_share] head -n 24000 DRR000031.fastq > DRR000031sub.fastq
iu@bielinux[mac_share] ls -l [ 2:28午後]
total 801012
-rwxrwxrwx 1 iu iu 819218014  4月 19 15:14 DRR000031.fastq
-rwxrwxrwx 1 iu iu  1016982  5月  2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] █
```



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、 $6,000 \times 4 =$ 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

```
iu@bielinux[~/Desktop/mac_share]
total 119625
-rwxrwxrwx 1 iu iu 122495839 4月
iu@bielinux[mac_share] ls -lh
total 117M
-rwxrwxrwx 1 iu iu 117M 4月 19 15:14 DRR000031.fastq.bz2
iu@bielinux[mac_share] bunzip2 DRR000031.fastq.bz2 [12:30午後]
iu@bielinux[mac_share] ls -l [12:31午後]
total 800018
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq
iu@bielinux[mac_share] wc DRR000031.fastq [12:34午後]
18359096 36718192 819218014 DRR000031.fastq ①
iu@bielinux[mac_share] head -n 2 DRR000031.fastq [1:20午後]
②
@DRR000031.1 20E06AAXX:5:1:121:512 length=36
TTTAAAAGATAATGTCATCACAACGCAACATA ④
iu@bielinux[mac_share] head -n 24000 DRR000031.fastq > DRR00003
lsub.fastq
iu@bielinux[mac_share] ls -l [2:28午後]
total 801012 ③
-rwxrwxrwx 1 iu iu 819218014 4月 19 15:14 DRR000031.fastq ①
-rwxrwxrwx 1 iu iu 1016982 5月 2 14:28 DRR000031sub.fastq
iu@bielinux[mac_share] █ [2:32午後]
```

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.fastq
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
lsub.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] █
```

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
[.gff3ファイル](#) (約1.3MB)
[.faファイル](#) (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
[plasmid1.gff3](#)(課題用)
[plasmid2.gff3](#)(課題用)
- 平成30年05月15日
講義資料PDF
[DRR000031sub.fastq](#) 
[RNA-QC-chain : Zhou et al., BMC Genomics, 2018](#)
[Biostar : Parnell et al., PLoS Comput Biol., 2011](#)
[FastQC](#)
[DRR000031sub_fastqc.html](#)
[DRR000031_fastqc.html](#)(課題用)
[report.html](#)(qrrqcを用いたQC結果)
- 平成30年05月22日



Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrcでクオリティチェック

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrcでクオリティチェック

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でしょう。それをこの後確認していきます

DRASearch

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
TGTAATGATTATGATTCTCAGGGATTGGGAAAGGT
>DRR000031.3
TCAAAAAATCGAAGTTAGGGTGACAAAGTTTGACA
>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGGGTGT
>DRR000031.5
TTCAGGAAGCTGGTGATGGAGCACCAGAGGGTGT
>DRR000031.6
TTTTTTTAAATAGCACTTTAAATATTTATTTTTT
>DRR000031.7
GGAGGTTAATCTGAGGCAGTATAGTAACCTTAAGG
>DRR000031.8
TTATCATCTTCACAATTCTAATNNNACTGACTATCC
>DRR000031.9
TTTTAAATGTAATTTTTTATTGGAAAACAAATAT
>DRR000031.10
TGGTAACAGCCTGATGGGTTATTTGACTGCACTAAG
```

C:\Users\kajok\Documents\2018\Lecture\09.機能ゲム学\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 x

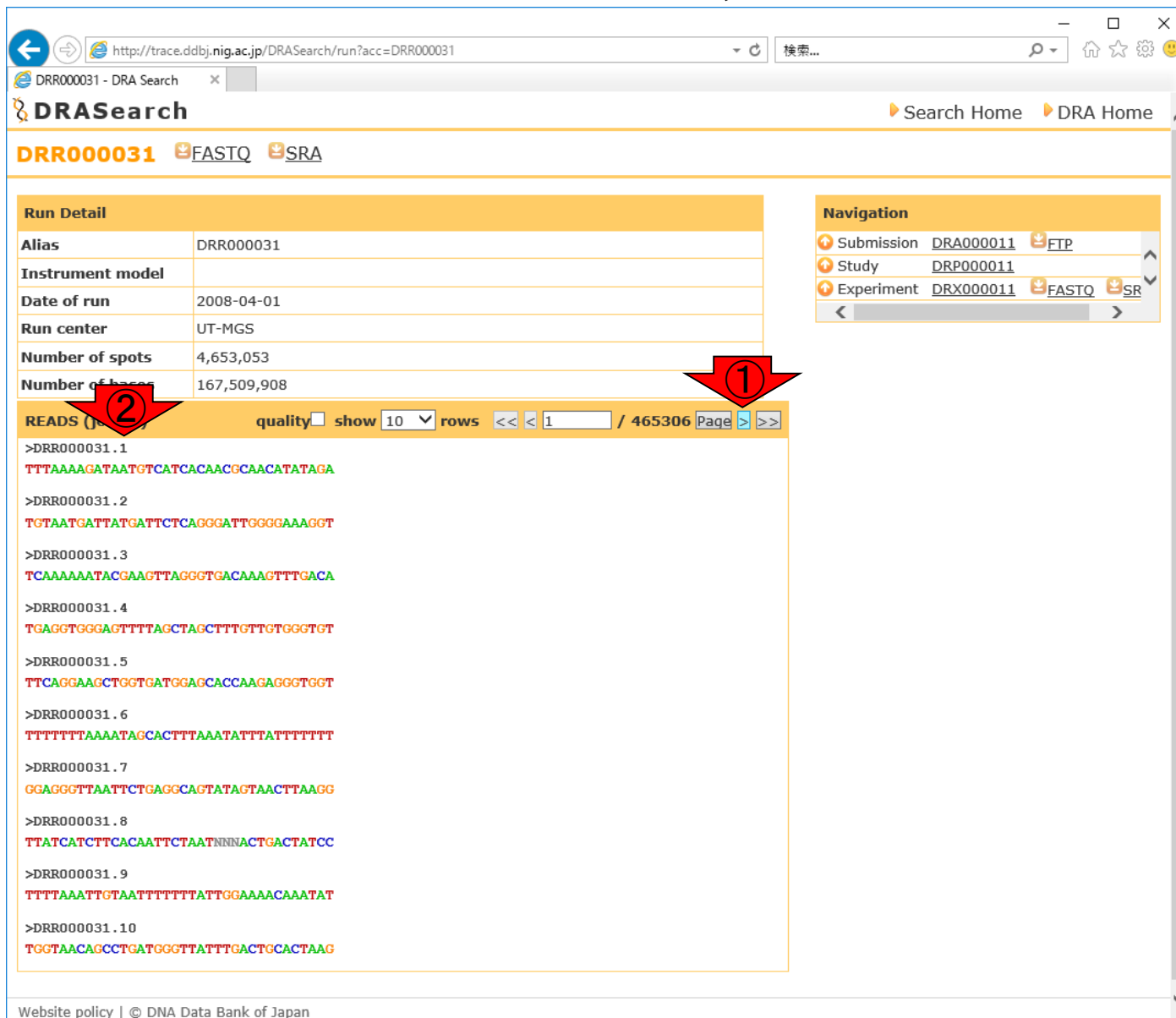
```
@DRR000031.77 20E06AAXX:5:1:540:232 length=36
GGATACCATTAGTGTGTATAACAGATTATTGTTTCAT
+DRR000031.77 20E06AAXX:5:1:540:232 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@DRR000031.79 20E06AAXX:5:1:796:395 length=36
TGAACCAGATCAATAGTGATAACATTATTCTCATAC
+DRR000031.79 20E06AAXX:5:1:796:395 length=36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

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



DRASearch

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
- Experiment DRX000011 FASTQ SR

READS (0) quality show 10 rows << < 1 / 465306 Page >>

```
>DRR000031.1
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
>DRR000031.2
TGTAATGATTATGATTCTCAGGGATTGGGAAAGGT
>DRR000031.3
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
>DRR000031.5
TTCAGGAAGCTGGTGATGGAGCACCAGAGGGTGGT
>DRR000031.6
TTTTTTTAAAATAGCACTTTAAATATTTATTTTTTT
>DRR000031.7
GGAGGGTTAATTCAGGCCAGTATACTAACTTAAGG
>DRR000031.8
TTATCATCTTCACAATCTAATNNNACTGACTATCC
>DRR000031.9
TTTTAAATGTAATTTTTTATTTGGAAAACAAATAT
>DRR000031.10
TGGTAACAGCCTGATGGGTTATTTGACTGCACTAAG
```


ロングリードデータの場合

(Rで)塩基配列解析

(last modified 2018/04/17, since 2010)

このウェブページのR関連部分は、[インストール](#) | [|](#) についての推奨手順 ([Windows2018.03.12版](#)と[Macin](#)
[版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は
[基本的な利用法](#)([Windows2015.04.03版](#)と[Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にま
とめた書籍もあります。(2015/04/03)

①PacBio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①をご覧ください

What's new?

- Silhouetteスコアの新たな使い道提唱 [Silhouette scores](#)(シルエットスコア) | [|](#)
- Silhouetteスコアの新たな使い道提唱
- 「平成29年度NGSハンズオン講習会」
- [門田からメール返信をもらえない場合](#)
- [はじめに](#) (last modified 2015/03/31)
- 参考資料 | [書籍、学会誌](#) (last modified 2015/04/03)
- 参考資料 | [講習会 講義 講演資料](#)

- 書籍 | [トランスクリプトーム解析 | 4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | [トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | について](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 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 2017/06/27)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/25)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last modified 2017/03/13)
- 書籍 | [日本乳酸菌学会誌 | 第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第11回統合データ解析環境Galaxy](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌 | 第12回Galaxy:ヒストリーとワークフロー](#) (last modified 2018/03/23)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)
- イントロ | 一般 | [任意の長さの可能な全ての塩基配列を作成](#) (last modified 2015/02/19)
- イントロ | 一般 | [任意の位置の塩基を置換](#) (last modified 2013/09/12)

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrcでクオリティチェック

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があります。①NGS連載第6回後半と第7回で紹介しているDDBJ Pipelineです

(Rで)塩基配列解析

(last modified 2018/04/17, since 2010)

このウェブページのR関連部分は、[インストール](#) | についての推奨手順 ([Windows2018.03.12版](#)と[Macintosh2015.04.03版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は[基本的な利用法](#)([Windows2015.04.03版](#)と[Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にまとめた書籍もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱 [Silhouette scores](#)(シルエットスコア) | [こちら](#)
- Silhouetteスコアの新たな使い道提唱 [「平成29年度NGSハンズオン講習会」](#) | [こちら](#)
- [門田からメール返信をもらえない場合](#) | [こちら](#)
- [はじめに](#) (last modified 2015/03/31)
- 参考資料 | [書籍、学会誌](#) (last modified 2017/06/28)
- 参考資料 | [講習会 講義 講演資料](#) (last modified 2017/06/28)

- 書籍 | [トランスクリプトーム解析 | 4.3.2 群間比較](#) (last modified 2014/04/28)
- 書籍 | [トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | について](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 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 2017/06/27)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/25)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last modified 2017/03/13)
- 書籍 | [日本乳酸菌学会誌 | 第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌 | 第11回統合データ解析環境Galaxy](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌 | 第12回Galaxy:ヒストリーとワークフロー](#) (last modified 2018/03/23)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)
- イントロ | 一般 | [任意の長さの可能な全ての塩基配列を作成](#) (last modified 2015/02/19)

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

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があります。もう1つは①NGS連載第11回と12回で紹介しているGalaxyです

(Rで)塩基配列解析

(last modified 2018/04/17, since 2010)

このウェブページのR関連部分は、[インストール](#) | についての推奨手順 ([Windows2018.03.12版](#)と[Macintosh2015.04.03版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は[基本的な利用法](#)([Windows2015.04.03版](#)と[Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にまとめた書籍もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱 [Silhouette scores](#)(シルエットスコア) | [こちら](#)
- Silhouetteスコアの新たな使い道提唱
- 「平成29年度NGSハンズオン講習会」
- [門田からメール返信をもらえない場合](#)
- [はじめに](#) (last modified 2015/03/31)
- 参考資料 | [書籍、学会誌](#) (last modified 2015/03/31)
- 参考資料 | [講習会 講義 講演資料](#)

- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌](#) | [第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | [日本乳酸菌学会誌](#) | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 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 2017/06/27)
- 書籍 | [日本乳酸菌学会誌](#) | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/25)
- 書籍 | [日本乳酸菌学会誌](#) | [第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | [日本乳酸菌学会誌](#) | [第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last modified 2017/03/13)
- 書籍 | [日本乳酸菌学会誌](#) | [第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第11回統合データ解析環境Galaxy](#) (last modified 2017/11/13)
- 書籍 | [日本乳酸菌学会誌](#) | [第12回Galaxy:ヒストリーとワークフロー](#) (last modified 2018/03/23)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)
- イントロ | 一般 | [任意の長さの可能な全ての塩基配列を作成](#) (last modified 2015/02/19)

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

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrcでクオリティチェック

公共DB

①EMBL-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

- ・ [イントロ](#) | [一般](#) | [配列取得](#) | [トランスクリプトーム配列](#) | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2016/02/10)
- ・ [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](#) | [読み込み](#) | [xlsx形式](#) | [openxlsx](#) (last modified 2015/11/15)
- ・ [イントロ](#) | [NGS](#) | [様々なプラットフォーム](#) (last modified 2016/03/24)
- ・ [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 2015/02/23)
- ・ [イントロ](#) | [NGS](#) | [配列取得](#) | [FASTQ or SRA](#) | [SRADB\(Zhu 2013\)](#) (last modified 2015/02/24)
- ・ [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 2014/03/26)

- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)
- ・ [イントロ](#) | [NGS](#) | [アノテーション情報取得](#)

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

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ー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(DBJ見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., 2012](#)
- ・ [EMBL-EBI European Nucleotide Archive \(ENA\): Ester et al., Nucleic Acids Res., 2015](#)
- ・ [NCBI Sequence Read Archive \(SRA\): NCBI Resource Coordinators., Nucleic Acids Res., 2014](#)
- ・ [ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015](#)
- ・ [GEO: Barrett et al., Nucleic Acids Res., 2013](#)
- ・ [DBCLS SRA: Nakazato et al., PLoS One, 2013](#)

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

EMBL-EBI European Nucleotide Archive

Services Research Training About us

DRR000031

Examples: [BN000065](#), [histone](#)

Search

Advanced Search

Home Search & Browse Submit & Update Software About ENA Support

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)

Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.

Text Search

Examples: [BN000065](#), [histone](#)

Search

[Advanced search](#)

Popular

- [Submit and update](#)
- [Sequence submissions](#)
- [Genome assembly submissions](#)
- [Submitting environmental sequences](#)
- [Citing ENA data](#)
- [Rest URLs for data retrieval](#)
- [Rest URLs to search ENA](#)

Latest ENA news

19 Mar 2018: [ENA Release 135](#)

Release 135 of ENA's assembled/annotated sequences is now available

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をブラウザでみる

The screenshot shows a web browser window displaying the ENA search results for the query *DRR000031*. The browser address bar shows <https://www.ebi.ac.uk/ena/data/search?query=DRR000031>. The ENA logo and navigation menu are visible at the top. The search results are displayed under the heading "Search results for *DRR000031*".

On the left side, there are links for "Read", "Experiment (1)", and "Run (1)".

The search results are organized into two sections:

- Experiment (1 results found)**
 - ① DRX000011 Illumina Genome Analyzer sequencing; HT29_Cytoplasm_Control
[View all 1 results](#)
- Run (1 results found)**
 - ② DRR000031 Illumina Genome Analyzer sequencing; HT29_Cytoplasm_Control
[View all 1 results](#)

Red arrows with circled numbers 1 and 2 point to the first and second results respectively.

At the bottom left, it says "Powered by [EBI Search](#)".

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ません**、です

ebi.ac.uk

EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

Search

Examples: [BN000065](#), [histone](#)

Advanced
Sequence

Home Search & Browse Submit & Update Software About ENA Support

EMBL-EBI

News
Our impact
Contact us
Intranet

Services
By topic
By name (A-Z)
Help & Support

Research
Overview
Publications
Research groups
Postdocs & PhDs

Training
Overview
Train at EBI
Train outside EBI
Train online
Contact organisers

Industry
Overview
Members Area
Workshops
SME Forum
Contact Industry programme

About us
Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News
Events
Visit us
Contact us

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44
Copyright © EMBL-EBI 2018 | EBI is an outstation of the European Molecular Biology Laboratory | [Terms of use](#)

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...

European Nucleotide Archive

EMBL-EBI

Services Research Training About us

ENA
European Nucleotide Archive

DRR000031

Examples: BN000065, histone

Search

Advanced
Sequence

Home Search & Browse Submit & Update Software About ENA Support

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)

Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.

Text Search

Examples: [BN000065](#), [histone](#)

Search

[Advanced search](#)

Popular

- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- Rest URLs to search ENA

Latest ENA news

19 Mar 2018: [ENA Release 135](#)

Release 135 of ENA's assembled/annotated sequences is now available

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

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

ENA
European Nucleotide Archive

Search: [Search](#)
Examples: BN000065, his...
[Advanced Sequence](#)

Home **Search & Browse** [Submit & Update](#) [Software](#) [About ENA](#) [Support](#)

Search results for *DRR000031* [Show more data from EMBL-EBI](#)

Read
Experiment (1)
Run (1)

Experiment (1 results found)
DRX000011 Illumina Genome Analyzer sequencing; HT29_Cytoplasm_Control
[View all 1 results](#)

Run (1 results found)
DRR000031 Illumina Genome Analyzer sequencing; HT29_Cytoplasm_Control
[View all 1 results](#)

Powered by [EBI Search](#)

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でみる

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

ENA

European Nucleotide Archive

Search

Examples: [BN000065](#), [histone](#)

[Advanced](#)
[Sequence](#)

[Home](#) [Search & Browse](#) [Submit & Update](#) [Software](#) [About ENA](#) [Support](#)

[Contact Helpdesk](#)

Run: DRR000031

Illumina Genome Analyzer sequencing; HT29_Cytoplasm_Control

View: [XML](#) Download: [XML](#)

Submitting Centre	Platform	Model	Read Count	Base Count
UT-MGS	ILLUMINA	Illumina Genome Analyzer	4,653,053	167,509,908
Library Layout	Library Strategy	Library Source	Library Selection	Library Name
SINGLE	FL-cDNA	TRANSCRIPTOMIC	cDNA	HT29_Cytoplasm_Control
Broker Name	DDBJ			

Navigation Read Files

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

Navigation Read Files

This table contains the files for run DRR000031

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

Download: 1 - 1 of 1 results in [TEXT](#)

[Select columns](#)

Showing results 1 - 1 of 1 results

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 (Galaxy)
PRJDA34559	SAMD00009330	DRS000011	DRX000011	DRR000031	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

Prev Next Go to page: 1 Go

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ページ下部に移動すると...

EMBL-EBI Services Research Training About us

ENA

European Nucleotide Archive

Home **Search & Browse** Submit & Update Software About ENA Support

Study: PRJDA34559 Contact Helpdesk

Homo sapiens transcriptome project

View: [Project XML](#) Download: [Project XML](#)

Name	Submitting Centre	Organism
Homo sapiens	Univeristy of Tokyo Department of Medical Genome Sciences	Homo sapiens

Secondary accession(s)

[DRP000003](#), [DRP000004](#), [DRP000005](#), [DRP000006](#), [DRP000007](#), [DRP000008](#), [DRP000011](#), [DRP000012](#), [DRP000013](#), [DRP000014](#), [DRP000015](#), [DRP000016](#), [DRP000017](#), [DRP000018](#), [DRP000019](#), [DRP000020](#), [DRP000021](#), [DRP000022](#), [DRP000023](#), [DRP000024](#), [DRP000025](#), [DRP000026](#), [DRP000027](#), [DRP000028](#), [DRP000032](#), [DRP000033](#), [DRP000034](#), [DRP000035](#), [DRP000036](#), [DRP000037](#), [DRP000038](#), [DRP000040](#), [DRP000041](#), [DRP000042](#), [DRP000043](#), [DRP000044](#), [DRP000045](#), [DRP000046](#), [DRP000047](#), [DRP000048](#), [DRP000049](#), [DRP000050](#), [DRP000051](#), [DRP000052](#), [DRP000053](#), [DRP000054](#), [DRP000055](#), [DRP000056](#), [DRP000057](#), [DRP000058](#), [DRP000059](#), [DRP000060](#), [DRP000061](#), [DRP000062](#), [DRP000063](#), [DRP000064](#), [DRP000065](#), [DRP000066](#), [DRP000073](#), [DRP000074](#), [DRP000075](#), [DRP000076](#), [DRP000077](#), [DRP000078](#), [DRP000079](#), [DRP000080](#), [DRP000081](#), [DRP000082](#), [DRP000089](#), [DRP000090](#), [DRP000091](#), [DRP000092](#), [DRP000093](#), [DRP000094](#), [DRP000095](#), [DRP000096](#), [DRP000097](#), [DRP000098](#), [DRP000099](#), [DRP000100](#), [DRP000101](#), [DRP000102](#), [DRP000103](#), [DRP000104](#), [DRP000105](#), [DRP000106](#), [DRP000107](#), [DRP000108](#), [DRP000109](#), [DRP000110](#), [DRP000111](#), [DRP000112](#), [DRP000113](#), [DRP000114](#), [DRP000115](#), [DRP000116](#), [DRP000117](#), [DRP000118](#), [DRP000119](#), [DRP000120](#), [DRP000121](#), [DRP000122](#), [DRP000123](#), [DRP000124](#), [DRP000125](#), [DRP000126](#), [DRP000127](#), [DRP000128](#), [DRP000129](#), [DRP000130](#), [DRP000131](#), [DRP000132](#), [DRP000133](#), [DRP000134](#), [DRP000135](#), [DRP000136](#), [DRP000137](#), [DRP000138](#), [DRP000139](#), [DRP000140](#), [DRP000141](#), [DRP000142](#), [DRP000143](#), [DRP000144](#), [DRP000145](#), [DRP000146](#),

ENAをChromeでみる

①このあたりまで移動。今回の対象であるDRR000031は、②で見られる

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

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 (Galaxy)
PRJDA34559	SAMD00013986	DRS0000003	DRX0000003	DRR0000003		Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS0000003	DRX0000003	DRR0000004	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS0000003	DRX0000003	DRR0000005	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS0000003	DRX0000003	DRR0000006	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD00013986	DRS0000003	DRX0000003	DRR0000007	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		
PRJDA34559	SAMD0002711	DRS0000004	DRX0000004	DRR0000008	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

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

ENAをChromeでみ

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

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 463 results

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 (Galaxy)
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		
PRJDA34559	SAMD0002711	DRS000004	DRX000004	DRR000008	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

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

ENAをChromeでみ

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

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 (Galaxy)
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		
PRJDA34559	SAMD0002711	DRS000004	DRX000004	DRR000008	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

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

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

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 (Galaxy)
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		
PRJDA34559	SAMD0002711	DRS000004	DRX000004	DRR000008	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		



ENAをChromeでみる

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

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 (Galaxy)
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		
PRJDA34559	SAMD0002711	DRS000004	DRX000004	DRR000008	9606	Homo sapiens	Illumina Genome Analyzer	SINGLE	File 1	File 1			File 1	File 1		

ENAをChromeでみる

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

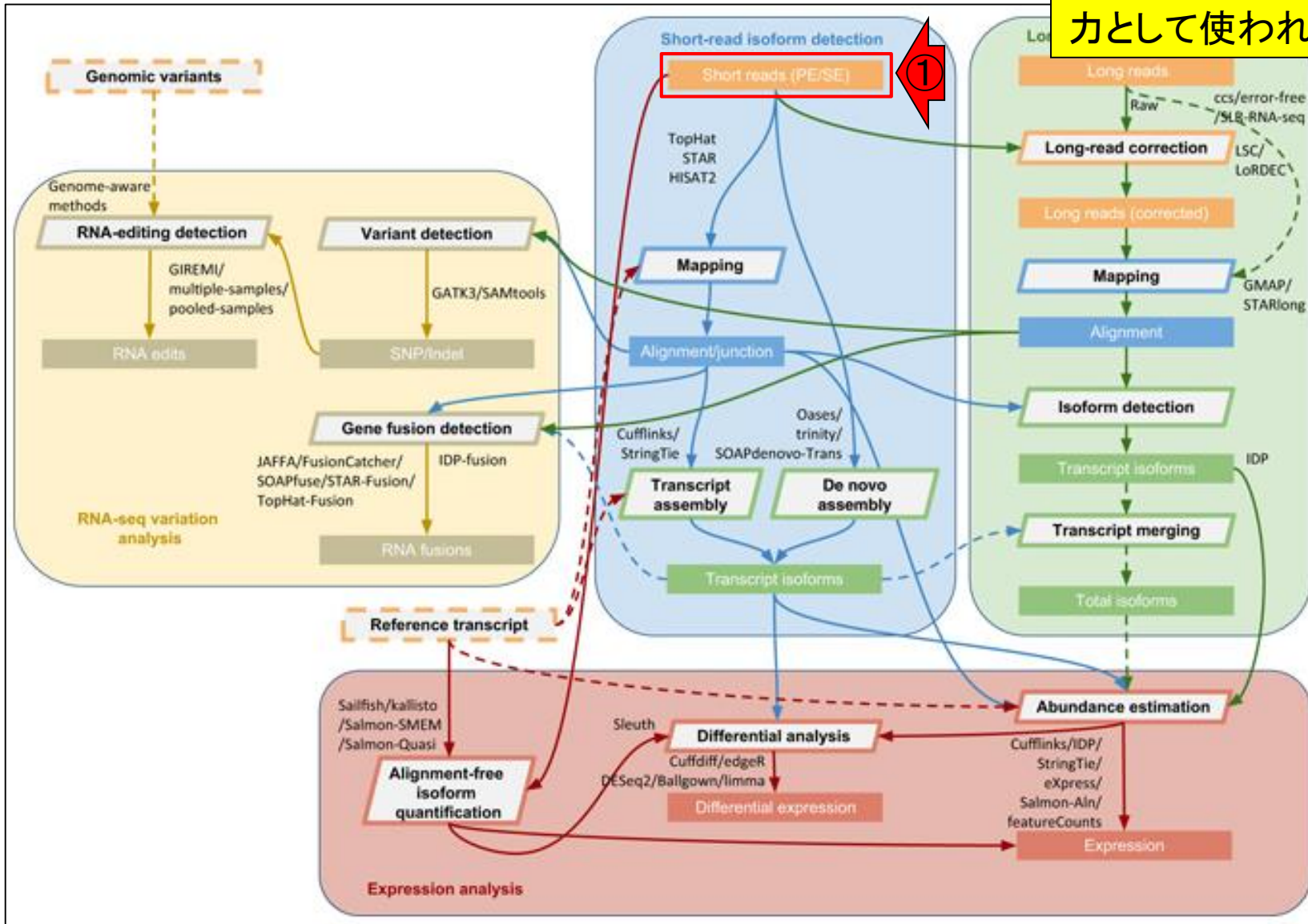
Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
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		
PRJDA34559	SAMD0002711	DRS000004	DRX000004	DRR000008	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、Linux
 - 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qrrc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



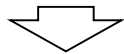
Sahraeian et al., Nat Commun., 8(1): 59, 2017

前処理 (preprocessing)

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)

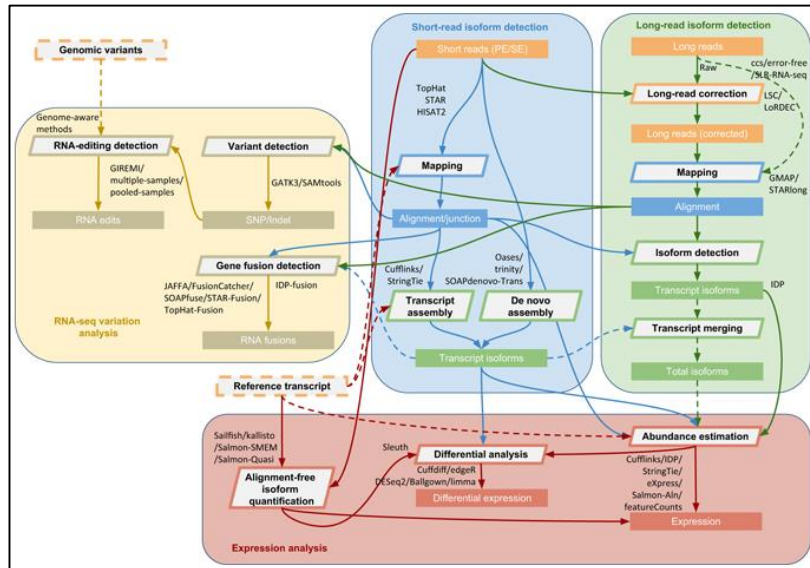
①

③

④

②

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



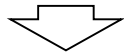
前処理 (preproce

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。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

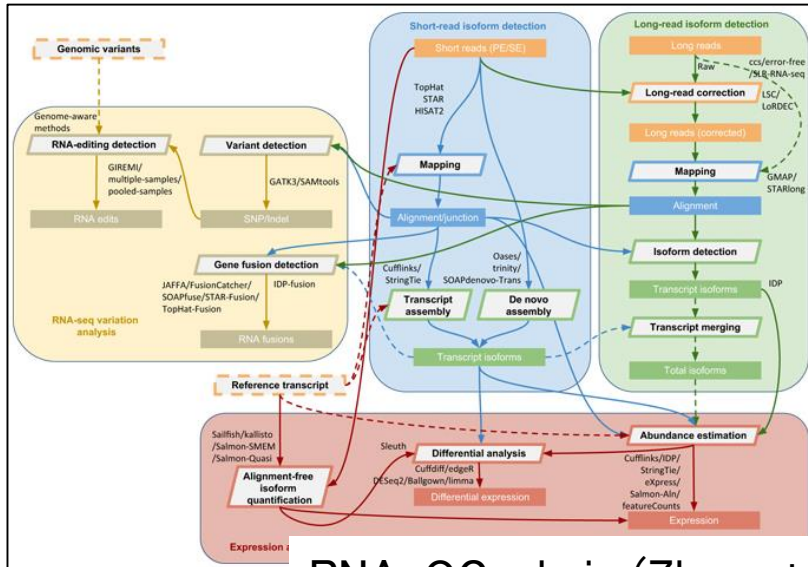
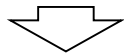
NGSリードデータ (SRAファイル)



NGSリードデータ (FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)



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

～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 Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

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](#)

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

May 15, 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 Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

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 Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

Author information

Abstract

BACKGROUND: RNA-Seq has become one of the most widely used applications of high-throughput sequencing technology. However, raw RNA-Seq data may have quality issues, which can significantly affect downstream analysis and lead to erroneous conclusions. Therefore, the raw data must be subjected to vigorous quality control before downstream analysis. Currently, an accurate and complete QC of RNA-Seq data is difficult to achieve. Existing tools used consecutively, which is inefficient in terms of usability, running time, file size, and storage space.

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](#)

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

~NGSデータ共通の事柄~

- ②・クオリティの低いリード
- ②・アダプター配列
- ④・解析対象外の生物種に由来するリード

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 Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

Author information

Abstract

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

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 sequencing depth and pair-end read mapping information). This package was developed as a command-line tool for general QC of next-generation sequencing (NGS) data called QC-Chain, with support for both paired-end and single-end RNA-Seq data. It has several features that are not available yet in other QC tools for RNA-Seq data. It includes sequence trimming, automatic rRNA detection and automatic contaminating species identification. QC-Chain can be run either sequentially or independently, enabling RNA-QC-Chain as a comprehensive and efficient QC tool with high usability. Moreover, parallel computing and optimizations are embedded in most of the steps to ensure 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

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)

～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
- ・RNA degradation (RNA分解)
- ・多様なリードカバレッジ (varied read coverage)

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

Zhou Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

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: [PMC58131](#)

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

May 15, 2018

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

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 Q¹, Su X^{2,3}, Jing G², Chen S⁴, Ning K⁵.

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](#)

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

May 15, 2018



Read free
full text at BMC

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 qualiti \[Genomics Proteomics Bioinforma...\]](#)

[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

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

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> | © The Author(s). 2018

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

Abstract

Keywords

Background

Results and discussion

Conclusions

Availability and requirements



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

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>

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。

インストール失敗例

```
iu@bielinux[~/Downloads/RNA-QC-Chain]
make[3]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/sam/bamtools/build'
make[2]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/sam/bamtools/build'
g++ -o bin/RQC-SAM-stats sam/sam.cpp -I ./common -Wno-deprecated -I /sam/bamtools/include/ -L /sam/bamtools/lib/ -lbamtools -Wl,-rpath,/sam/bamtools/lib/
sam/sam.cpp:4:27: fatal error: api/BamReader.h: No such file or directory
#include <api/BamReader.h>
^
compilation terminated.
make[1]: *** [sam_parser] Error 1
make[1]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain'
make: *** [all] Error 2
iu@bielinux[RNA-QC-Chain] ls [ 6:11午後 ]
bin                Makefile           qc                 sam
common             models             rrna              User's manua.pdf
Default_tag_sequence parallel-meta      Rscript
iu@bielinux[RNA-QC-Chain] [ 6:41午後 ]
```

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、そのLinux版のようなものという理解でよい

```
iu@bielinux[~/Downloads/RNA-QC-Chain]
make[3]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/sam

C:\Users\kojik\Documents\html\lectures\AG09\180515\RNA-QC-chain_install.txt - EmEditor
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)
RNA-QC-chain_install.txt x
wget -c http://bioinfo.single-cell.cn/Released_Software/rna-qc-chain/RNA-QC-Chain-1.0.tar.gz
tar -zxvf RNA-QC-Chain-1.0.tar.gz
cd RNA-QC-Chain
make
iu@bielinux[RNA-QC-Chain] ls
bin          Makefile    qc          sam
common      models     rrna       User's manua.pdf
Default_tag_sequence parallel-meta Rscript
iu@bielinux[RNA-QC-Chain]
```

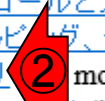
make

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)

```
C:\Users\kojik\Documents\html\lectures\AG09\180515\RNA-QC-chain_install.txt - EmEditor
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)
RNA-QC-chain_install.txt x
wget -c http://bioinfo.single-cell.cn/Released_Software/rna-qc-chain/RNA-QC-Chain-1.0.tar.gz
tar -zxvf RNA-QC-Chain-1.0.tar.gz
cd RNA-QC-Chain
make
```



- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2017/11/13)
- 書籍 | 日本乳酸菌学会誌 | [第1回イントロダクション](#) (last modified 2016/12/22)
- 書籍 | 日本乳酸菌学会誌 | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 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 2017/07/07)
- 書籍 | 日本乳酸菌学会誌 | [第4回クオリティコントロールとプログラムのインストール](#) (last modified 2017/06/25)
- 書籍 | 日本乳酸菌学会誌 | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2017/06/25)
- 書籍 | 日本乳酸菌学会誌 | [第6回ゲノムアセンブリ](#) (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | [第7回ロングリードアセンブリ](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第8回アセンブリ後の解析](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第9回ゲノムアノテーションとその可視化、DDBJへの登録](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第10回DDBJへの塩基配列の登録\(後編\)](#) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | [第11回統合データ解析環境Galaxy](#) (last modified 2017/11/13)
- 書籍 | 日本乳酸菌学会誌 | [第12回Galaxy:ヒストリーとワークフロー](#) (last modified 2018/03/23)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)



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。原因はどうか、③BamReader.h、④というファイルがない、ことに起因するようだ

```
iu@bielinux[~/Downloads/RNA-QC-Chain]
make[3]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/sam/bamtools/build'
make[2]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/sam/bamtools/build'
g++ -o bin/RQC-SAM-stats sam/sam.cpp -I ./common -Wno-deprecated -I /sam/bamtools/include/ -I /sam/bamtools/lib/ -lbamtools -Wl,-rpath,/sam/bamtools/lib/
sam/sam.cpp:4:27: fatal error: api/BamReader.h: No such file or directory
#include <api/BamReader.h>
^
compilation terminated.
make[1]: *** [sam_parser] Error 1
make[1]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain'
make: *** [all] Error 2

iu@bielinux[RNA-QC-Chain] ls
bin          Makefile    qc          sam
common      models     rrna       User's manua.pdf
Default_tag_sequence parallel-meta Rscript
iu@bielinux[RNA-QC-Chain]
```

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)。それが無い段階で「不親切だな」という感想を(少なくとも私は)もつ

```
iu@bielinux[~/Downloads/RNA-QC-Chain]
make[3]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/bamtools/build'
make[2]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain/bamtools/build'
g++ -o bin/RQC-SAM-stats sam/sam.cpp -I /sam/bamtools/include/ -L /sam/bamtools/lib/ -l, -rpath,/sam/bamtools/lib/
sam/sam.cpp:4:27: fatal error: api/BamReader.h: No such file or directory
#include <api/BamReader.h>
^
compilation terminated.
make[1]: *** [sam_parser] Error 1
make[1]: Leaving directory `/home/iu/Downloads/RNA-QC-Chain'
make: *** [all] Error 2
iu@bielinux[RNA-QC-Chain] ls [ 6:11午後 ]
bin          Makefile    qc          sam
common       models     rrna       User's manua.pdf
Default_tag_sequence parallel-meta Rscript
iu@bielinux[RNA-QC-Chain] [ 6:41午後 ]
```

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

rna-qc-chain bamreader.h

①

すべて 画像 動画 地図 ニュース | 保存した項目

40,600 件の検索結果 時間指定なし 言語で絞り込む

実践 データ分析講座 1 : 基礎編 | 日本規格協会 / 公式

広告 · webdesk.jsa.or.jp/

分析に利用する基本的な統計手法をしっかり学べるセミナーです！

データ解析 · 統計解析ソフト · JMP

3 品質工学 · 2 統計的品質管理 · 1 データ分析 : 基礎編 · 4 実験計画法 : 応用編

5Sから改善が速やかに実践可能 - jiet.co.jp

広告 · www.jiet.co.jp/

《タイムプリズム》作業の見える化・カイゼンが効率よく実践できます

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

<https://bmcbgenomics.biomedcentral.com/articles/10.1186/s...> このページを翻訳

2017/09/20 · 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 ...

KAPA hgDNA Quantification and QC Kit - Roche ...

sequencing.roche.com/en/products-solutions/by-category/... このページを翻訳

KAPA hgDNA Quantification and QC Kits contain all the consumables needed for the qPCR-based quantification and quality assessment of human genomic DNA prior to NGS library construction. Each kit contains KAP...

②

Hey any one used QC-Chain tool? - Latest Posts

<https://www.biostars.org/p/142403> このページを翻訳

any one used QC-Chain tool for Quality check of NGS data. how accurate is the tool to check contamination of seq data? Need help. Thanks ... QC-Chain is kind of messy to install, it is a pipeline composed of several tools ...

③

Biostar

①Biostarのリンク先ページ。赤枠の記載内容を見て、(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検索しても引っかけられないので注意

LATEST OPEN RNA-SEQ CHIP-SEQ SNP ASSEMBLY



Community

Log In

Sign Up

Welcome to Biostar!

Question: Hey any one used QC-Chain tool?

any one used QC-Chain tool for Quality check of NGS data.
how accurate is the tool to check contamination of seq data?

0 Need help.

Thanks
qc-chain • 902 views

ADD COMMENT • link • Not following

modified 3.0 years ago by [h.mon](#) • 14k • written 3.0 years ago by [gskbioinfo143](#) • 50



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.

2 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.

In the end, I decided to use either [FastQ Screen](#) (really easy to install, not so pretty graphics) or [MGA](#) (better output but not so simple to install and particularly run).

There is also [BBduk](#) and [SeqyClean](#), both of them may check for contamination using kmers, it may be more sensitive than FastQ Screen and MGA, which use Bowtie.

ADD COMMENT • link



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

Biostar

①Biostarについては、②アグリバイオ講義科目「ゲノム情報解析基礎」の③第1回で紹介しています。④ここ

2. ゲノム情報解析基礎

授業の目標・概要

次世代シーケンサーの普及により、ゲノム情報を基盤とした膨大な塩基配列情報を自在に解析するスキルが要求される時代になっています。フリーソフトRを用いて、配列決定後の基礎情報取得など各種配列解析の基本スキル向上を目指した実習を含む講義を行います。また、ウェブツールなどを用いて遺伝子領域の予測やアノテーションなどゲノム情報を比較または解析するための手法について解説します。

担当教員

嶋田 透 (東大・農・生産・環境生物学専攻 / 教授)
勝間 進 (東大・農・生産・環境生物学専攻 / 准教授)
門田幸二 (東大・農・アグリバイオ / 准教授)

参考図書

坊農秀雄 著、生命科学データ解析、MEDSI、2017

お知らせ

講義では、Rの様々なパッケージを利用します。
持ち込み用PC利用希望者は [インストール | について](#) を参考にしてR本体および必要なパッケージ群を必ずインストールしておいてください。

講義日程 (平成30年度)

1. 平成30年04月16日 (PC使用)
講師：嶋田 透
講師：門田幸二
[バイオインフォマティクス基礎知識](#)
[講義資料PDF\(Win版 ; 完全版\)](#)
[講義資料PDF\(Mac版 ; Rの説明](#)

[ゲノム情報解析基礎](#) バイオインフォマティクス基礎知識 (2018年04月16日) 門田幸二

バイオインフォマティクス系学会など

- ・ [ISCB\(International Society for Computational Biology\)](#)
- ・ [JSBI\(日本バイオインフォマティクス学会\)](#)
- ・ ...

バイオインフォマティクス系?!よろず相談所

- ・ [SEQanswers: Li et al., Bioinformatics, 2012](#)
◦ NGSに特化した相談所。
- ・ [Biostar: Parnell et al., PLoS Comput Biol., 2011](#)
◦ バイオインフォ全般。
- ・ [ライフサイエンスQA](#)
◦ バイオインフォ全般。過疎ってる印象。
- ・ [Bio Technical フォーラム](#)
◦ 主に実験系だがときどきバイオインフォ系トピックも。。。活発な印象。

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

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrqcでクオリティチェック

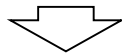
前処理 (preprocess)

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)

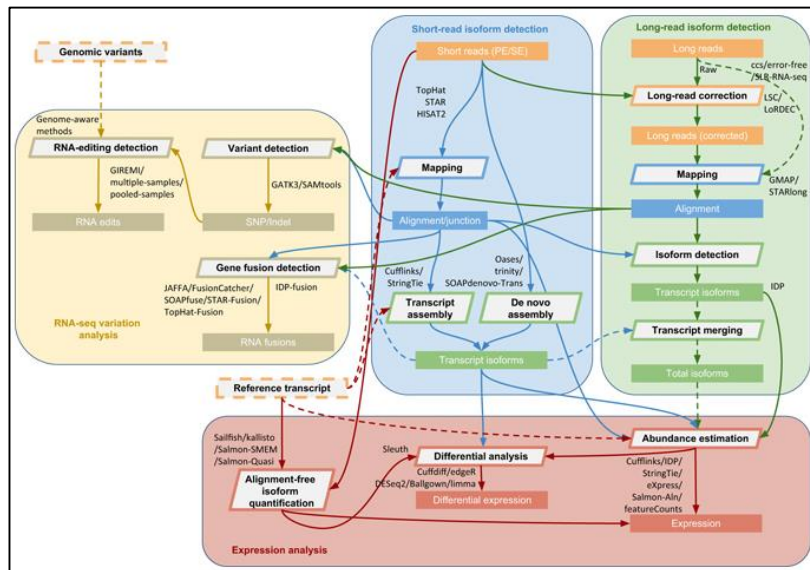
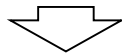


NGSリードデータ (FASTQファイル)



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

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

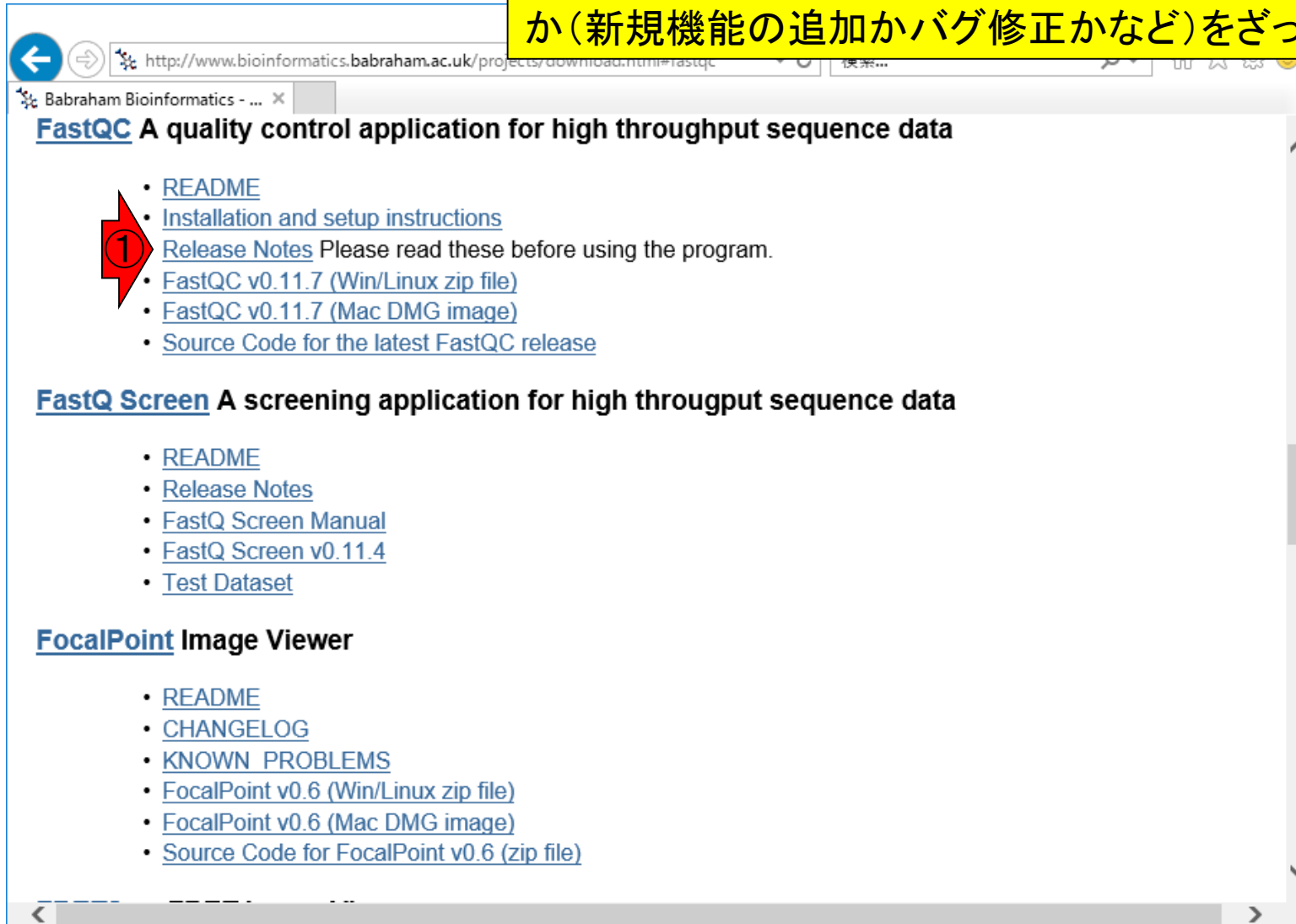
FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)

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



http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

Babraham Bioinformatics - ... x

FastQC A quality control application for high throughput sequence data

- [README](#)
- [Installation and setup instructions](#)
- [Release Notes](#) Please read these before using the program.
- [FastQC v0.11.7 \(Win/Linux zip file\)](#)
- [FastQC v0.11.7 \(Mac DMG image\)](#)
- [Source Code for the latest FastQC release](#)

FastQ Screen A screening application for high throughput sequence data

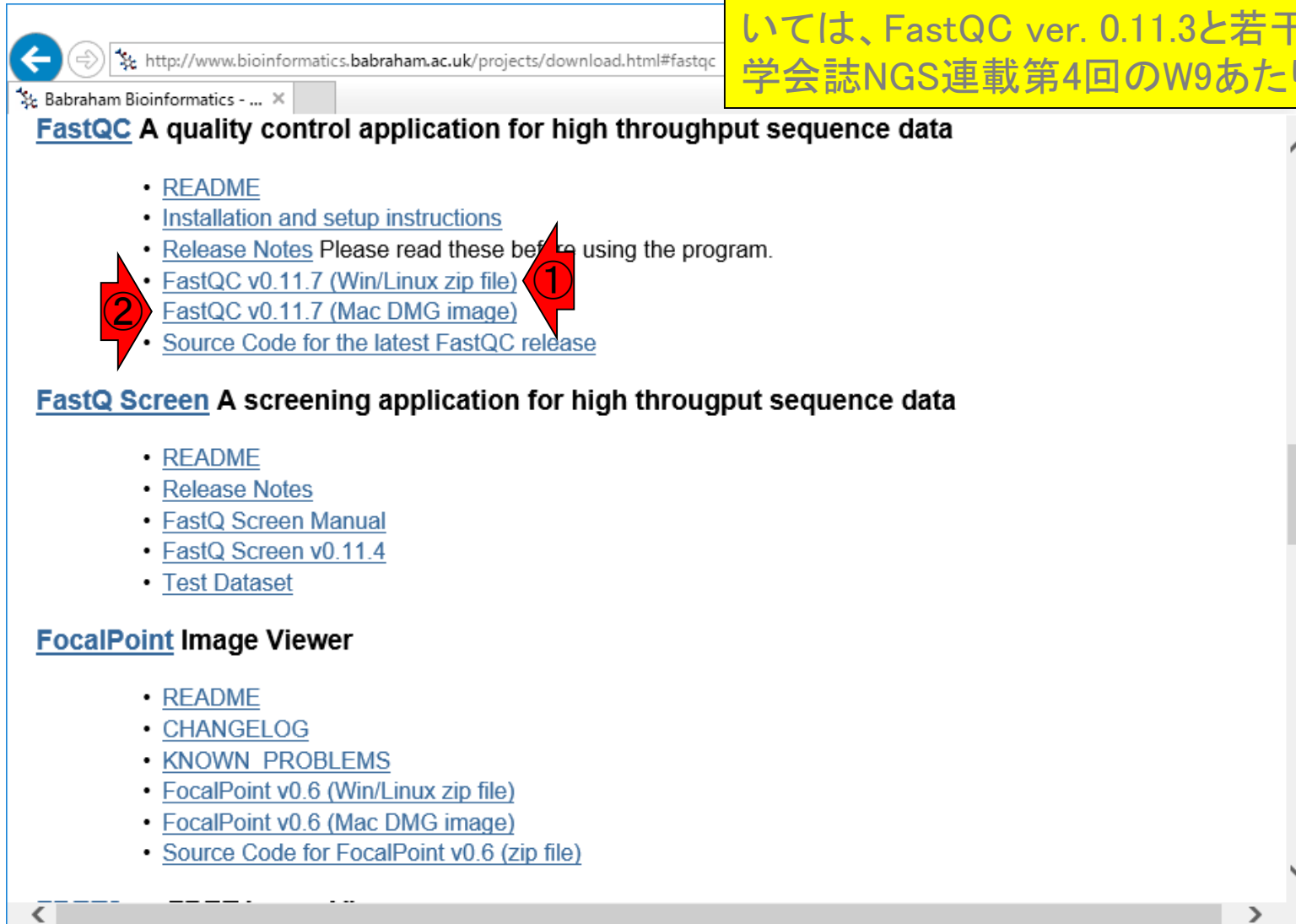
- [README](#)
- [Release Notes](#)
- [FastQ Screen Manual](#)
- [FastQ Screen v0.11.4](#)
- [Test Dataset](#)

FocalPoint Image Viewer

- [README](#)
- [CHANGELOG](#)
- [KNOWN PROBLEMS](#)
- [FocalPoint v0.6 \(Win/Linux zip file\)](#)
- [FocalPoint v0.6 \(Mac DMG image\)](#)
- [Source Code for FocalPoint v0.6 \(zip file\)](#)

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



<http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc>

Babraham Bioinformatics - ... x

FastQC A quality control application for high throughput sequence data

- [README](#)
- [Installation and setup instructions](#)
- [Release Notes](#) Please read these before using the program.
- [FastQC v0.11.7 \(Win/Linux zip file\)](#) ①
- [FastQC v0.11.7 \(Mac DMG image\)](#) ②
- [Source Code for the latest FastQC release](#)

FastQ Screen A screening application for high throughput sequence data

- [README](#)
- [Release Notes](#)
- [FastQ Screen Manual](#)
- [FastQ Screen v0.11.4](#)
- [Test Dataset](#)

FocalPoint Image Viewer

- [README](#)
- [CHANGELOG](#)
- [KNOWN PROBLEMS](#)
- [FocalPoint v0.6 \(Win/Linux zip file\)](#)
- [FocalPoint v0.6 \(Mac DMG image\)](#)
- [Source Code for FocalPoint v0.6 \(zip file\)](#)

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

The screenshot shows a web browser window with the URL <http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc>. The page title is "FastQC A quality control application for high throughput sequence data".

Under the "FastQC" section, there is a list of links:

- [README](#)
- [Installation and setup instructions](#)
- [Release Notes](#) Please read these before using the program.
- [FastQC v0.11.7 \(Win/Linux zip file\)](#) ①
- [FastQC v0.11.7 \(Mac DMG image\)](#)
- [Source Code for the latest FastQC release](#)

Under the "FastQ Screen" section, there is a list of links:

- [README](#)
- [Release Notes](#)
- [FastQ Screen Manual](#)
- [FastQ Screen v0.11.4](#)
- [Test Dataset](#)

Under the "FocalPoint Image Viewer" section, there is a list of links:

- [README](#)
- [CHANGELOG](#)
- [KNOWN PROBLEMS](#)
- [FocalPoint v0.6 \(Win/Linux zip file\)](#)

A download dialog box is open at the bottom of the browser window, showing the text: "bioinformatics.babraham.ac.uk から fastqc_v0.11.7.zip (9.77 MB) を開くか、または保存しますか?". The dialog has three buttons: "ファイルを開く(O)", "保存(S)", and "キャンセル(O)". Callout ② points to the dialog box, and callout ③ points to the "ファイルを開く(O)" button.

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

http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

Babraham Bioinformatics - ... x

FastQC A quality control application for high throughput sequence data

- [README](#)
- [Installation and setup instructions](#)
- [Release Notes](#) Please read these before using the program.
- [FastQC v0.11.7 \(Win/Linux zip file\)](#)
- [FastQC v0.11.7 \(Mac DMG image\)](#)
- [Source Code for the latest FastQC release](#)

FastQ Screen A screening application for high throughput sequence data

- [README](#)
- [Release Notes](#)
- [FastQ Screen Manual](#)
- [FastQ Screen v0.11.4](#)
- [Test Dataset](#)

FocalPoint Image Viewer

- [README](#)
- [CHANGELOG](#)
- [KNOWN PROBLEMS](#)
- [FocalPoint v0.6 \(Win/Linux zip file\)](#)
- [FocalPoint v0.6 \(Mac DMG image\)](#)
- [Source Code for FocalPoint v0.6 \(zip file\)](#)

セキュリティスキャンを実行中..

ダウンロードの表示(Y) x

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

ファイル ホーム 共有 表示

PC > デスクトップ > FastQC

名前	更新日時	種類	サイズ
Configuration	2018/05/04 21:13	ファイル フォルダ	
Help	2018/05/04 21:13	ファイル フォルダ	
net	2018/05/04 21:13	ファイル フォルダ	
org	2018/05/04 21:13	ファイル フォルダ	
Templates	2018/05/04 21:13	ファイル フォルダ	
uk	2018/05/04 21:13	ファイル フォルダ	
cisd-jhdf5.jar	2017/12/21 11:50	Executable Jar File	9,051 KB
fastqc	2017/12/21 11:50	ファイル	15 KB
fastqc_icon.ico	2017/12/21 11:50	アイコン	3 KB
INSTALL.txt	2017/12/21 11:50	テキスト ドキュメント	7 KB
jbzip2-0.9.jar	2017/12/21 11:50	Executable Jar File	49 KB
LICENSE	2017/12/21 11:50	ファイル	18 KB
LICENSE.txt	2017/12/21 11:50	テキスト ドキュメント	35 KB
LICENSE_JHDF5.txt	2017/12/21 11:50	テキスト ドキュメント	12 KB
README.md	2017/12/21 13:21	MD ファイル	2 KB
README.txt	2017/12/21 13:21	テキスト ドキュメント	3 KB
RELEASE_NOTES.txt	2018/01/10 9:29	テキスト ドキュメント	37 KB
run_fastqc.bat	2017/12/21 11:51	Windows バッチ ファ...	1 KB
sam-1.103.jar	2017/12/21 11:51	Executable Jar File	630 KB

19 個の項目

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

C:\Windows\system32\cmd.exe

C:\Users\koyik\Desktop\FastQC\fastqc application

FastQC

File Help

FastQC High Throughput Sequence QC Report
Version: 0.11.7
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011-17,
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression ©Matthew J. Francis, 2011
Base64 encoding ©Robert Harder, 2012
Java HDF5 reader ©ETH, CISD and SIS, 2007-14

Use File > Open to select the sequence file you want to check

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



FastQC High Throughput Sequence QC
Version: 0.11.7
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewings
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression ©Matthew J. Francis, 2011
Base64 encoding ©Robert Harder, 2012
Java HDF5 reader ©ETH, CISD and SIS, 2007-14

Use File > Open to select the sequence file you want to check

講義日程 (平成30年度)

- 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
- 平成30年05月15日
講義資料PDF
① DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
- 平成30年05月22日

入力ファイルの指定

The screenshot shows the FastQC application window with an 'Open' dialog box overlaid. The dialog box is titled '開く' and shows the 'ドキュメント' (Documents) folder selected. A red arrow points to the 'デスクトップ' (Desktop) folder in the left sidebar. The main pane of the dialog shows a list of folders and files. The 'ファイル名(N):' field is empty, and the 'ファイルのタイプ(T):' dropdown is set to 'Sequence Files'. The '開く' (Open) button is highlighted.

FastQC High T

www.bioinf

© Simon Andrews, Pierre

Picard B

BZip d

Bas

Java HI

Use File > Open to select th

開く

ファイルの場所(D): ドキュメント

最近使った項...

デスクトップ

ドキュメント

PC

ネットワーク

- 2015
- 2016
- 2017
- 2018
- CyberLink
- Fax
- Fuji Xerox
- html
- Office のカスタム テンプレート
- Outlook ファイル
- paper
- public_html
- R
- Scanned Documents
- サウンドレコーディング
- その他

- data_rma_2_nr.txt
- hoge2_hoge.txt
- hoge3_GPL1355.txt
- hoge5.txt
- hoge6.txt
- result_Add1a.txt
- result_Add1b.txt
- result_Add1c.txt
- result_Fig1b.txt
- table.txt
- table2.txt

ファイル名(N):

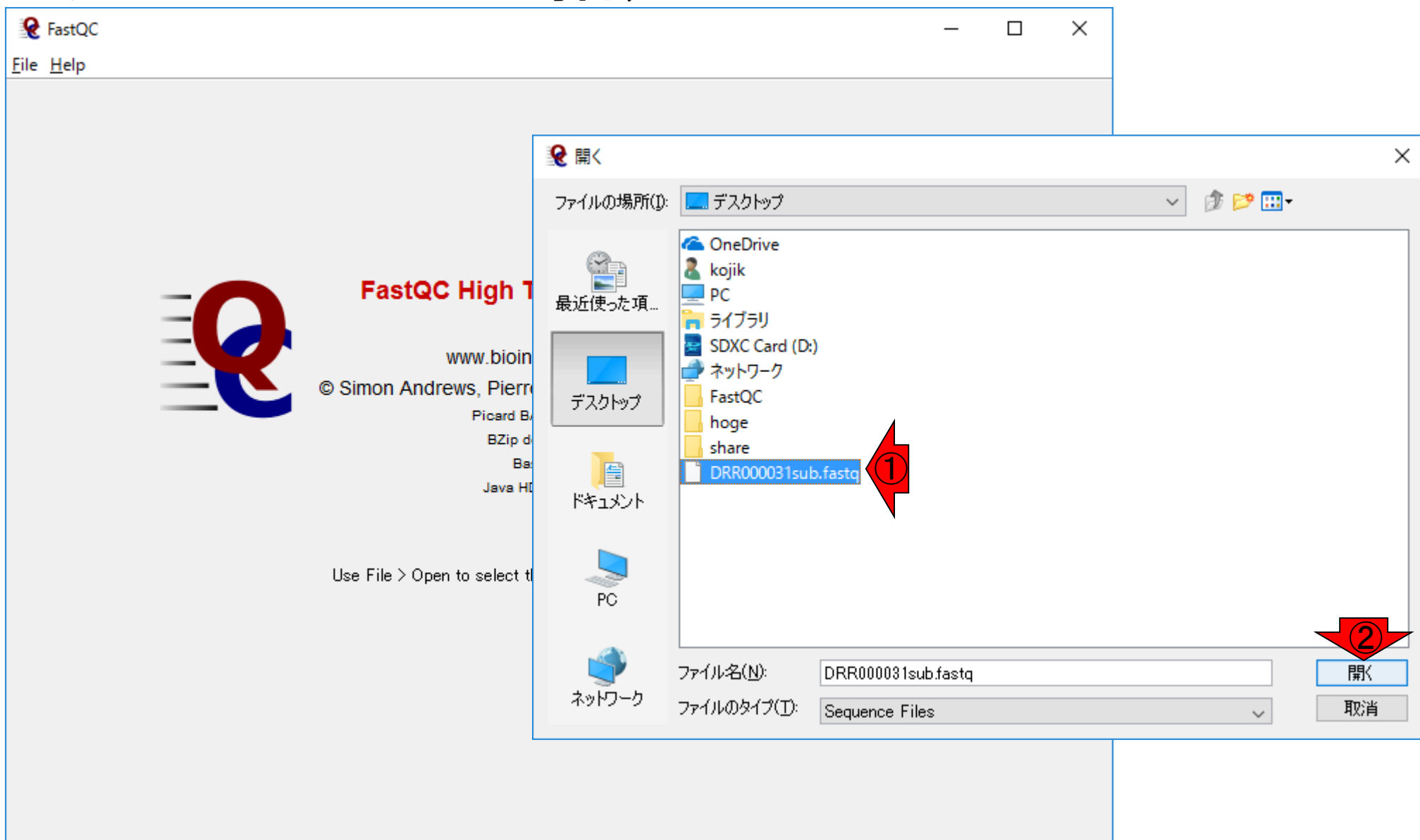
ファイルのタイプ(T): Sequence Files

開く

取消

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



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

FastQC実行結果

The screenshot shows the FastQC application window. The title bar reads 'FastQC'. The menu bar contains 'File' and 'Help'. The file name 'DRR000031sub.fastq' is displayed in the top-left corner. On the left side, there is a vertical list of analysis modules, each with a status icon: a green checkmark for 'Basic Statistics', 'Per base sequence quality', 'Per tile sequence quality', 'Per sequence quality scores', 'Per sequence GC content', 'Per base N content', 'Sequence Length Distribution', 'Sequence Duplication Levels', 'Overrepresented sequences', and 'Adapter Content'. A red 'X' icon is next to 'Per base sequence content'. The main area displays a table titled 'Basic sequence stats' with two columns: 'Measure' and 'Value'. The table contains the following data:

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrc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

File Help

DRR000031sub.fastq

Basic Statistics

Per base sequence quality

Per tile sequence quality

Per sequence quality scores

Per base sequence content

Per sequence GC content

Per base N content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

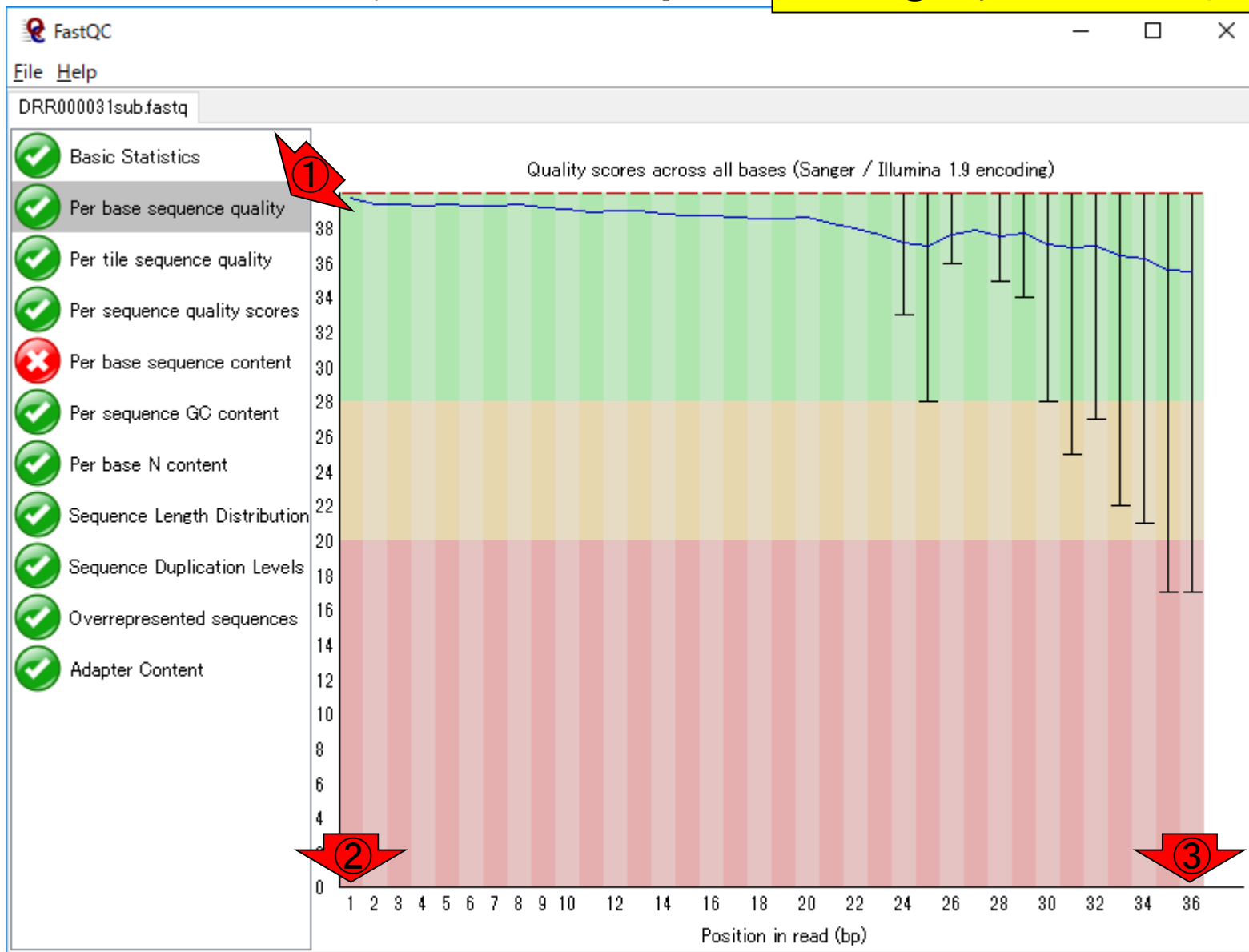
Adapter Content

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

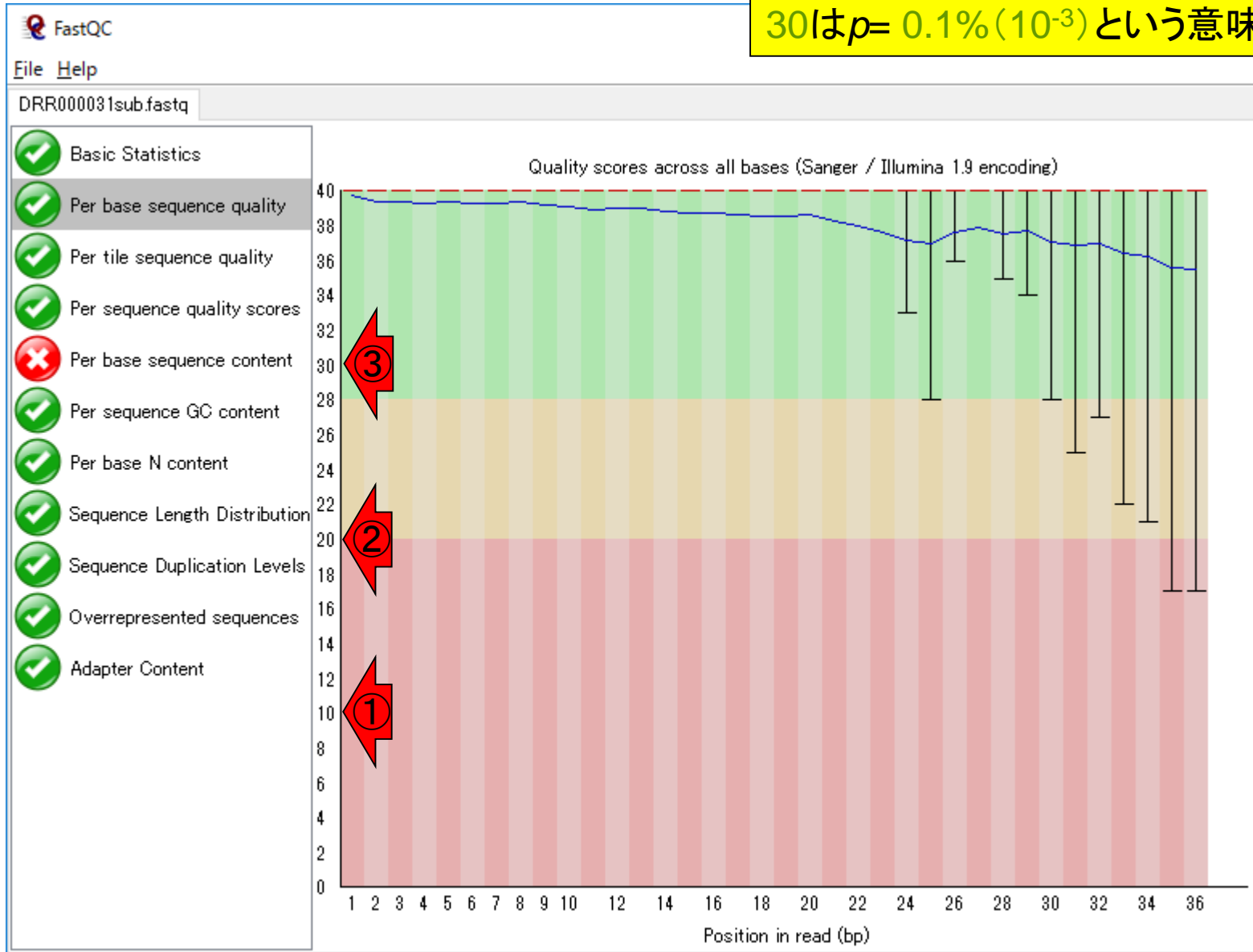
FastQC実行結果

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



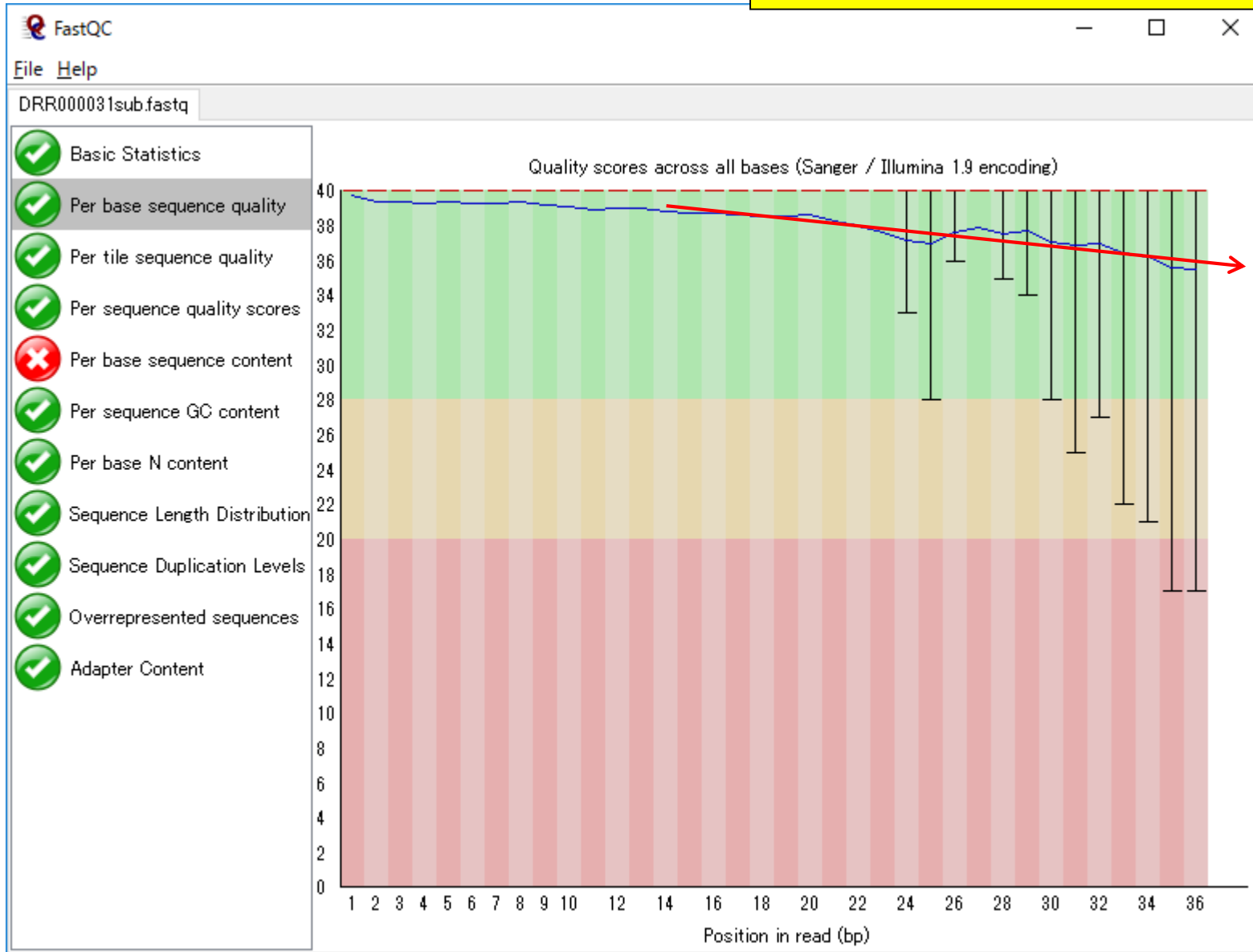
FastQC実行結果

縦軸がクオリティスコア q 。高ければ高いほどよい。① $q = 10$ 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、② $q = 20$ は $p = 1\%$ (10^{-2})、③ $q = 30$ は $p = 0.1\%$ (10^{-3})という意味でした(前回の講義)



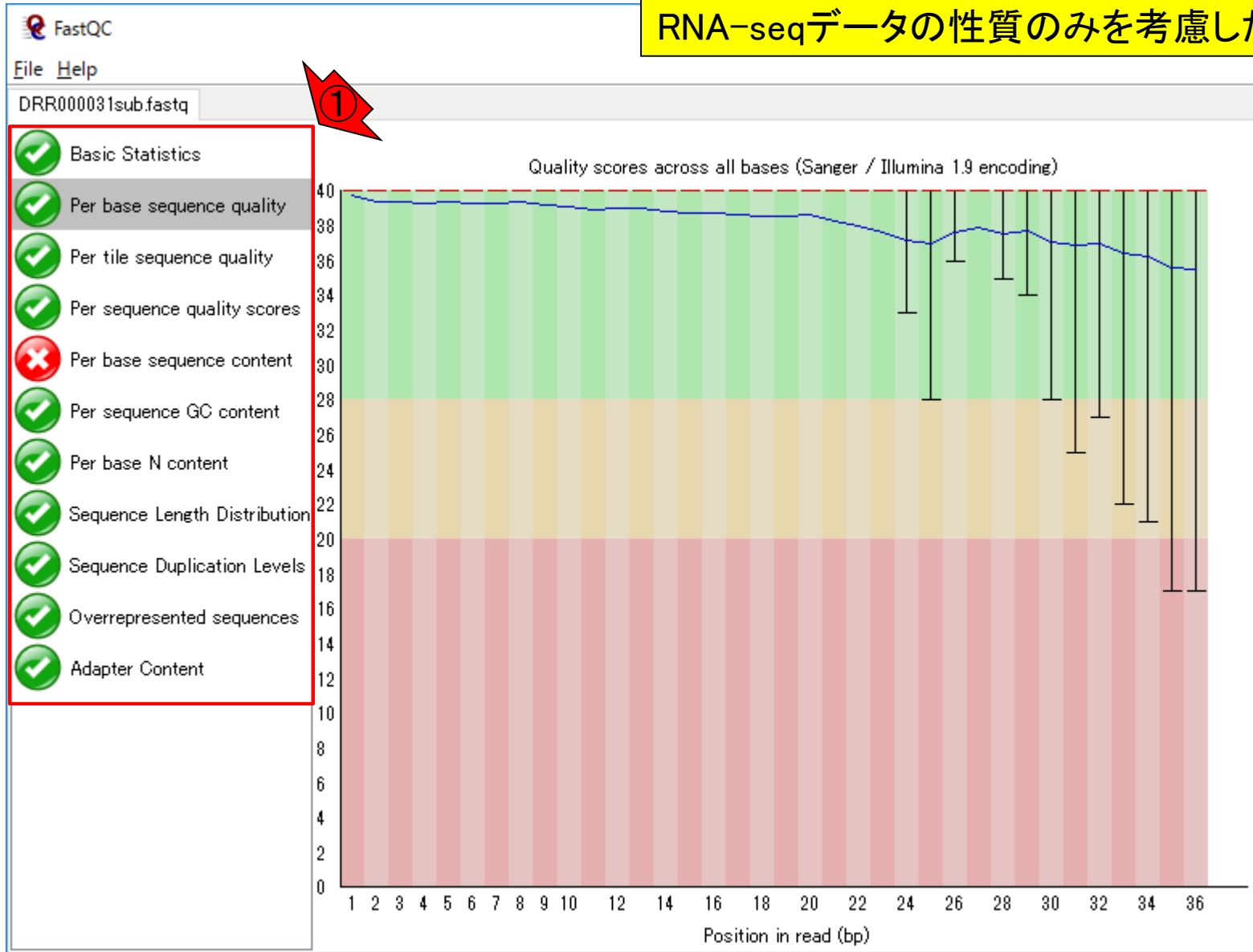
解釈

塩基配列決定精度は、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。



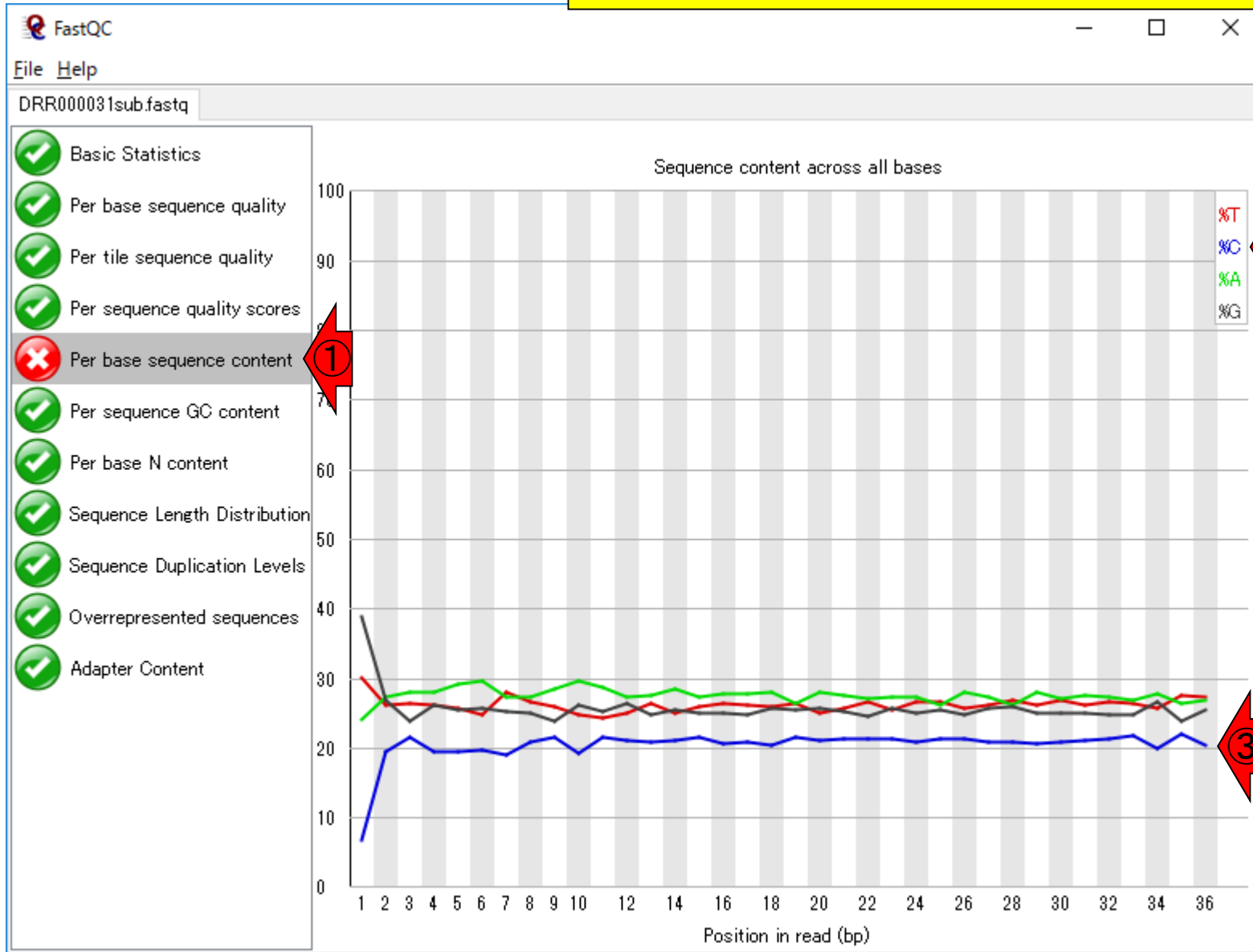
信号機と同じ

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

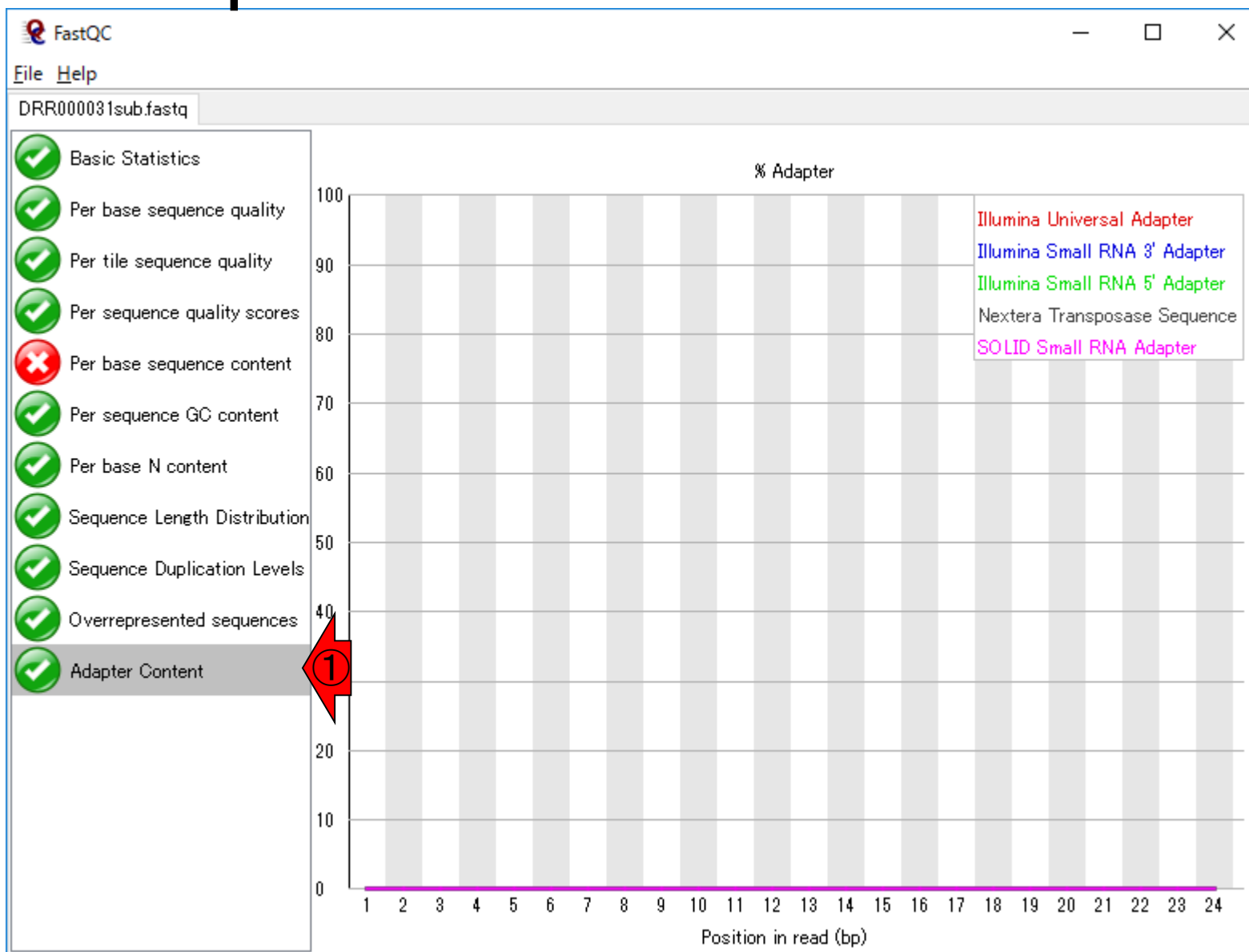


Overrepresented

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)

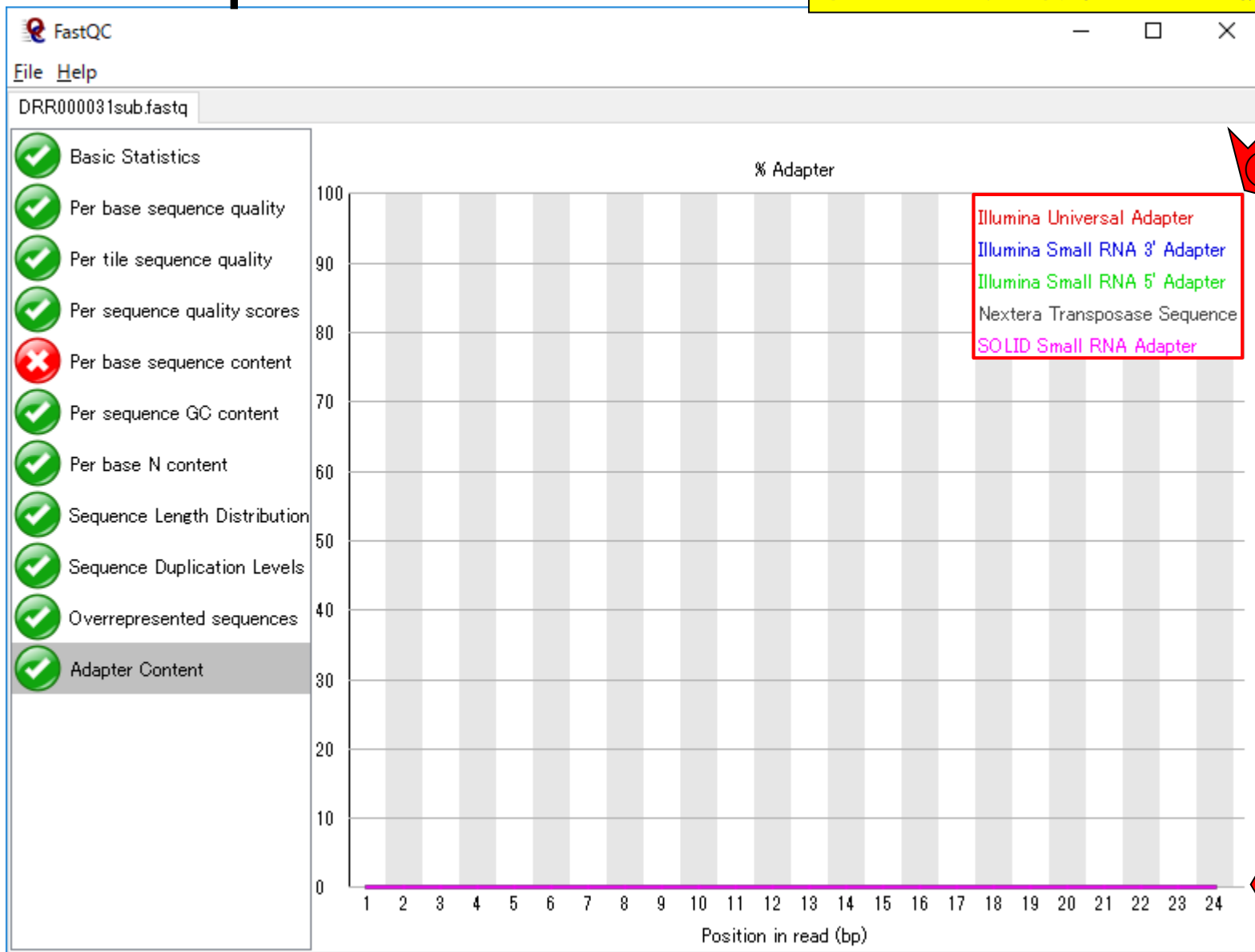
The screenshot shows the FastQC web interface for a file named 'DRR000031sub.fastq'. On the left, a sidebar contains a list of analysis modules, each with a green checkmark icon, except for 'Per base sequence content' which has a red 'X' icon. The 'Overrepresented sequences' module is highlighted with a grey background and a red arrow pointing to it, with a circled '1' next to the arrow. The main content area on the right is titled 'Overrepresented sequences' and displays the text 'There are no overrepresented sequences'.

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



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...

The screenshot shows the FastQC application window. The 'File' menu is open, and 'Save report...' is highlighted with a red arrow labeled ①. A '保存' (Save) dialog box is open, showing the 'デスクトップ' (Desktop) location selected with a red arrow labeled ②. The file name is 'DRR000031sub_fastqc.html' and the file type is 'HTML files'. A red arrow labeled ③ points to the '保存' (Save) button. In the background, a bar chart displays sequence quality metrics across 24 positions in a read. The y-axis represents a percentage from 0 to 100. The x-axis is labeled 'Position in read (bp)'. A legend on the left lists the metrics with their status:

- Per base sequence content: ❌
- Per sequence GC content: ✅
- Per base N content: ✅
- Sequence Length Distribution: ✅
- Sequence Duplication Levels: ✅
- Overrepresented sequences: ✅
- Adapter Content: ✅

DRR000031sub_fastqc.html

講義日程 (平成30年度)

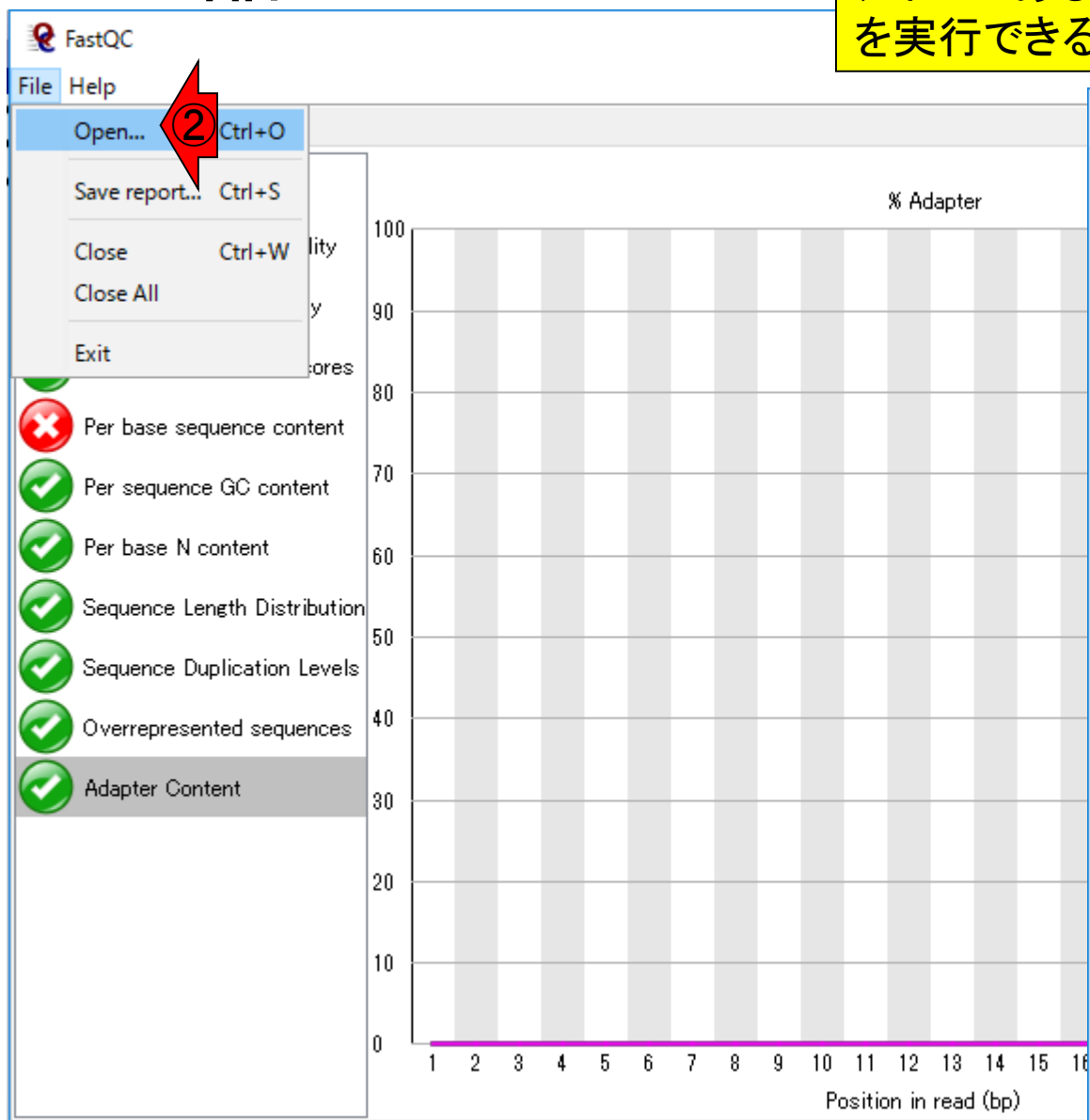
1. 平成30年05月08日
講義資料PDF
.gff3ファイル (約1.3MB)
.faファイル (約2.2MB)
(Rで)塩基配列解析
(Rで)マイクロアレイデータ解析
plasmid1.gff3(課題用)
plasmid2.gff3(課題用)
2. 平成30年05月15日
講義資料PDF
DRR000031sub.fastq
RNA-QC-chain : Zhou et al., BMC Genomics, 2018
Biostar : Parnell et al., PLoS Comput Biol., 2011
FastQC
DRR000031sub_fastqc.html 
DRR000031_fastqc.html(課題用)
report.html(qrqcを用いたQC結果)
3. 平成30年05月22日

Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrc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!



DRR000031.fastq.bz2のプロパティ

全般 セキュリティ 詳細 以前のバージョン

DRR000031.fastq.bz2

ファイルの種類: BZ2 ファイル (.bz2)

プログラム: アプリの選択 変更(C)...

場所: C:\Users\kajok\Documents\2018\Lecture\09.機能ゲノム

サイズ: 116 MB (122,495,839 バイト)

ディスク上のサイズ: 116 MB (122,499,072 バイト)

作成日時: 2018年4月19日、15:14:22

更新日時: 2018年4月19日、15:14:28

アクセス日時: 2018年4月19日、15:14:22

属性: 読み取り専用(R) 隠しファイル(H) 詳細設定(D)...

セキュリティ: このファイルは他のコンピューターから取得したものです。このコンピューターを保護するため、このファイルへのアクセスはブロックされる可能性があります。 ブロックの解除(K)

OK キャンセル 適用(A)

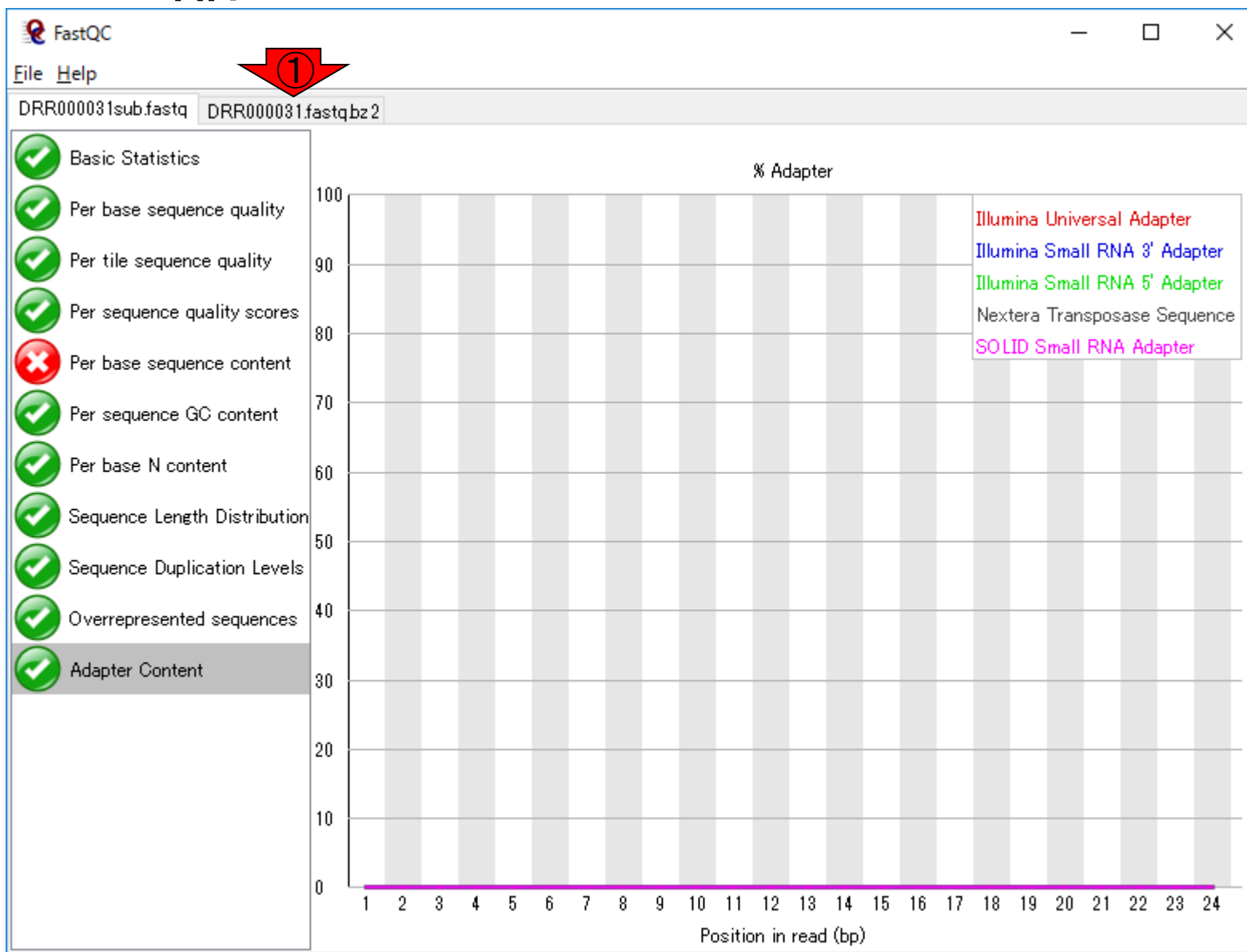
圧縮ファイルも可

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!

The screenshot displays the FastQC application interface. The 'File' menu is open, showing 'Open...' (Ctrl+O) as the selected option. An '開く' (Open) dialog box is overlaid, showing the file 'DRR000031.fastq.bz2' selected in a folder named '09機能ゲノム学'. A red arrow with the number '1' points to the selected file. Below the dialog, the file name 'DRR000031.fastq.bz2' and type 'Sequence Files' are visible. A second red arrow with the number '2' points to the '開く' (Open) button in the dialog. In the background, a 'DRR000031.fastq.bz2のプロパティ' (Properties) window is open, showing the file name and a '変更(C)...' (Change...) button. The main FastQC window shows a bar chart for '% Adapter' content, with a y-axis labeled 'Quality' and a value of 100. The left sidebar contains a list of analysis options, with 'Adapter Content' checked and highlighted.

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



圧縮ファイルも可

一瞬なんじゃこりゃ!となりますが、よく見ると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らうと納得

FastQC

File Help

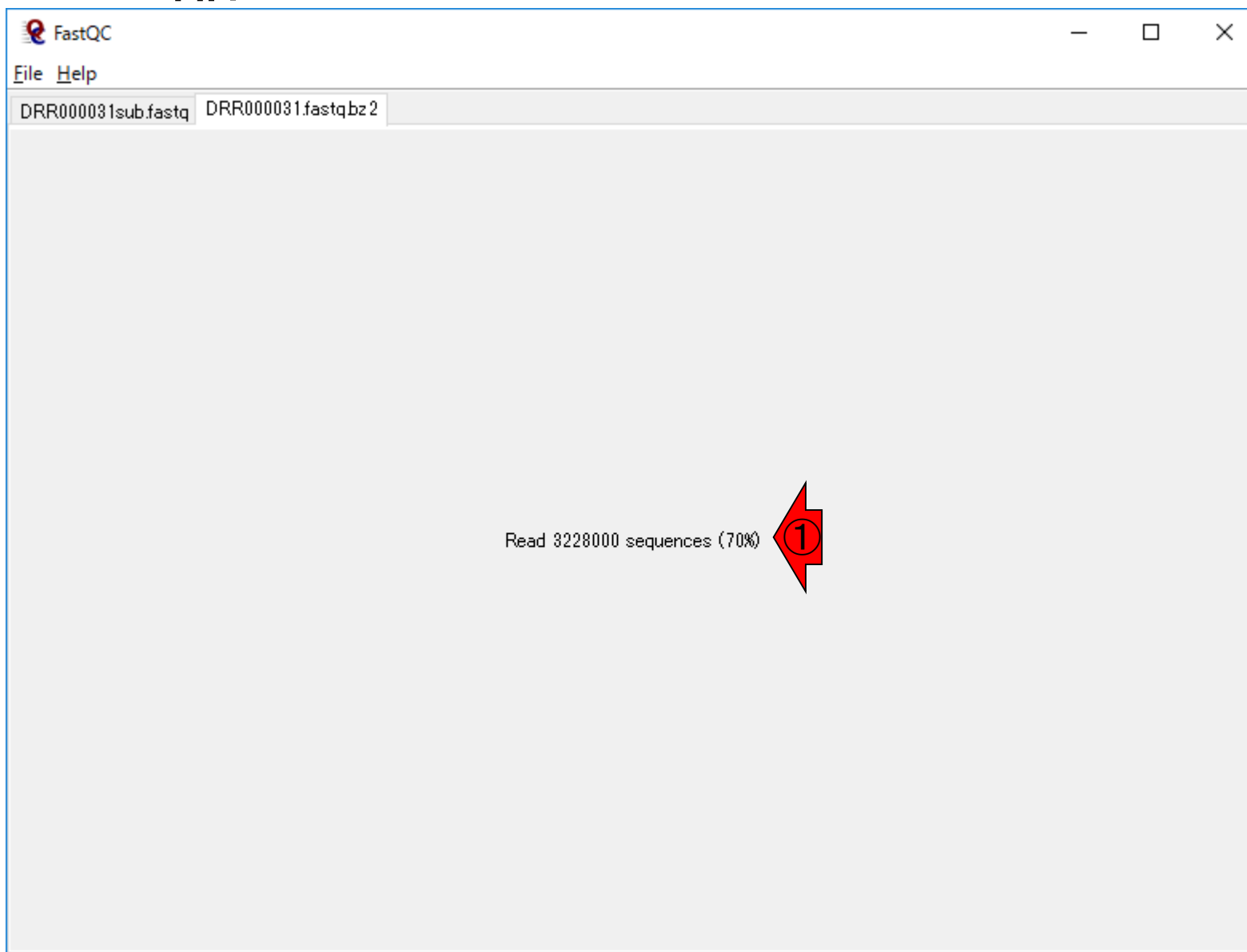
DRR000031sub.fastq DRR000031.fastq.bz2

Read 2563000 sequences (55%)

The screenshot shows the FastQC application window. At the top, there are two tabs: 'DRR000031sub.fastq' and 'DRR000031.fastq.bz2'. The main area is a progress bar. A red arrow with the number '1' inside points to the text 'Read 2563000 sequences (55%)' on the progress bar.

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

The screenshot shows the FastQC application window. The left sidebar contains a list of analysis modules, most of which are checked with green icons. The main panel displays the 'Basic sequence stats' table for the file 'DRR000031.fastq.bz2'. A red box highlights the 'Total Sequences' row, which has a value of 4,589,774. A red arrow with the number '1' points to this value. A larger red box below it shows a zoomed-in view of the same table.

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

FastQC interface showing two files: DRR000031sub.fastq and DRR000031.fastqbz2. The 'Basic Statistics' tab is selected. The table below shows the 'Basic sequence stats' for the selected file.

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

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...

htmlファイルで保存

The screenshot shows the FastQC application interface. The 'File' menu is open, and 'Save report...' is selected. A file save dialog box is overlaid, showing the 'Desktop' location selected. The file name is 'DRR000031_fastqc.html' and the type is 'HTML files'. The 'Save' button is highlighted.

① 0001.fastq.bz2
② Save report...
③ デスクトップ
④ 保存

DRR000031_fastqc.html

講義日程 (平成30年度)

1. 平成30年05月08日

[講義資料PDF](#)

[.gff3ファイル](#) (約1.3MB)

[.faファイル](#) (約2.2MB)

[\(Rで\)塩基配列解析](#)

[\(Rで\)マイクロアレイデータ解析](#)

[plasmid1.gff3](#)(課題用)

[plasmid2.gff3](#)(課題用)

2. 平成30年05月15日

[講義資料PDF](#)

[DRR000031sub.fastq](#)

[RNA-QC-chain : Zhou et al., BMC Genomics, 2018](#)

[Biostar : Parnell et al., PLoS Comput Biol., 2011](#)

[FastQC](#)

[DRR000031sub_fastqc.html](#)

[DRR000031_fastqc.html](#)(課題用)

[report.html](#)(qrrqcを用いたQC結果)



3. 平成30年05月22日

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

FastQC interface for file DRR000031sub.fastq. The 'Basic sequence stats' table is as follows:

Measure	Value
Filename	DRR000031sub.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6000
Sequences flagged as poor quality	0
Sequence length	36
%GC	46

FastQC interface for file DRR000031.fastq.bz2. The 'Basic sequence stats' table is as follows:

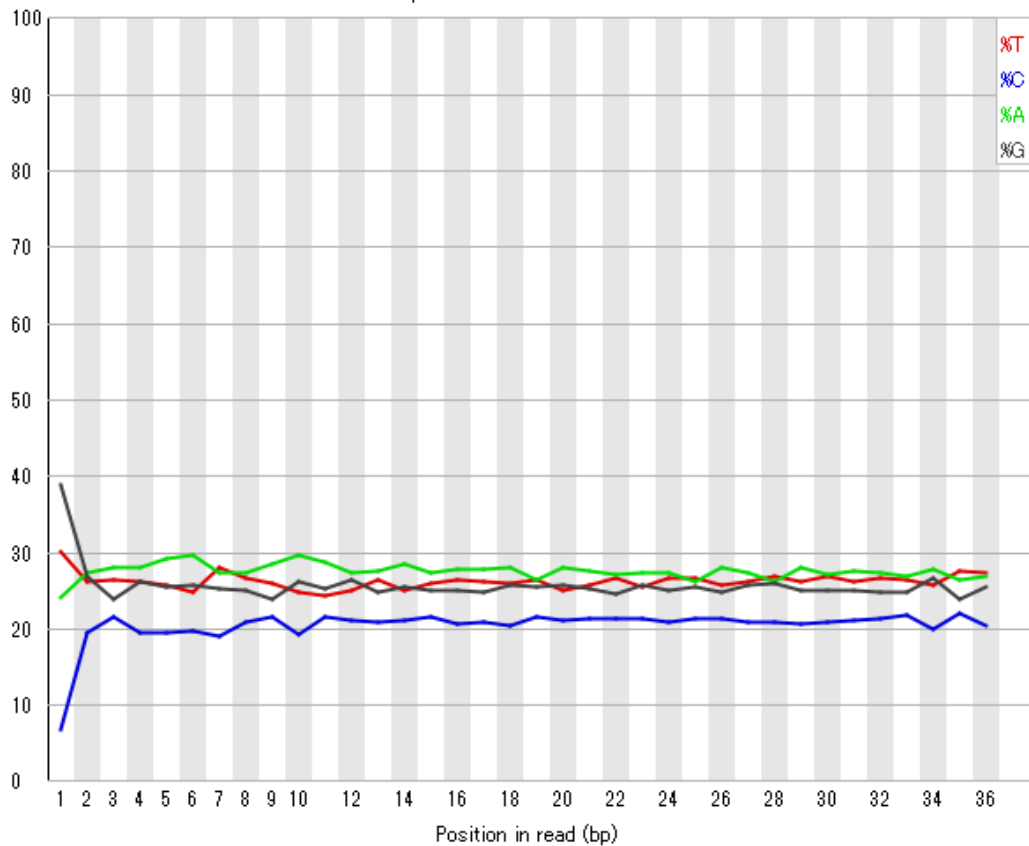
Measure	Value
Filename	DRR000031.fastq.bz2
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4589774
Sequences flagged as poor quality	0
Sequence length	36
%GC	47

課題：論点1

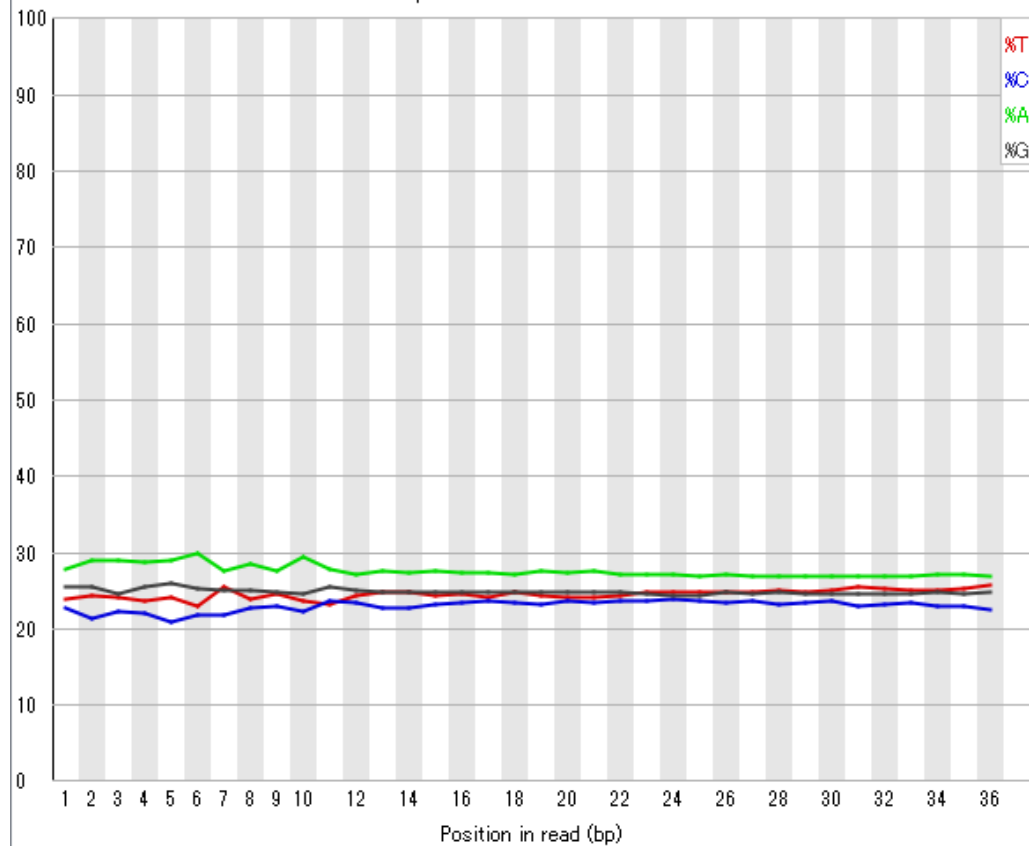
FastQC実行結果のPer base sequence content
。入力ファイルが①4,589,774リードからなる
DRR000031.fastq.bz2と、②6,000リードからなる
DRR000031sub.fastqの違いについて簡単に考察



Sequence content across all bases



Sequence content across all bases



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



Overrepresented sequences

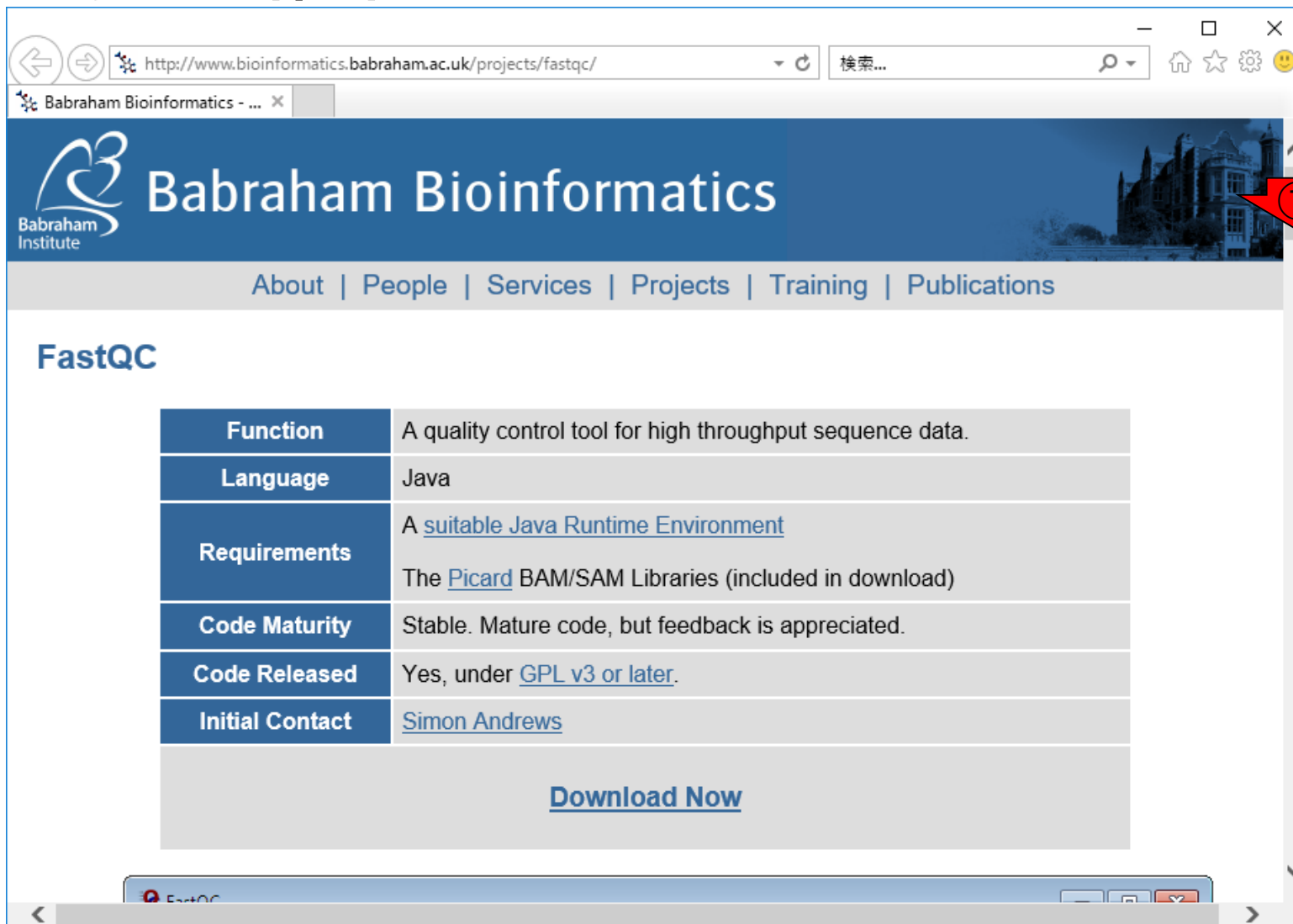
There are no overrepresented sequences



Overrepresented sequences

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

参考情報



The screenshot shows a web browser window displaying the Babraham Bioinformatics website. The URL in the address bar is <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. The page features the Babraham Institute logo and a navigation menu with links for About, People, Services, Projects, Training, and Publications. The main content area is titled "FastQC" and contains a table with the following information:

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

Below the table, there is a prominent [Download Now](#) button. A red arrow with the number 1 points to the bottom of the page, indicating the location of the explanation mentioned in the header.

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…

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

検索...

Babraham Bioinformatics - ...

4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

[View our tutorial video](#)

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

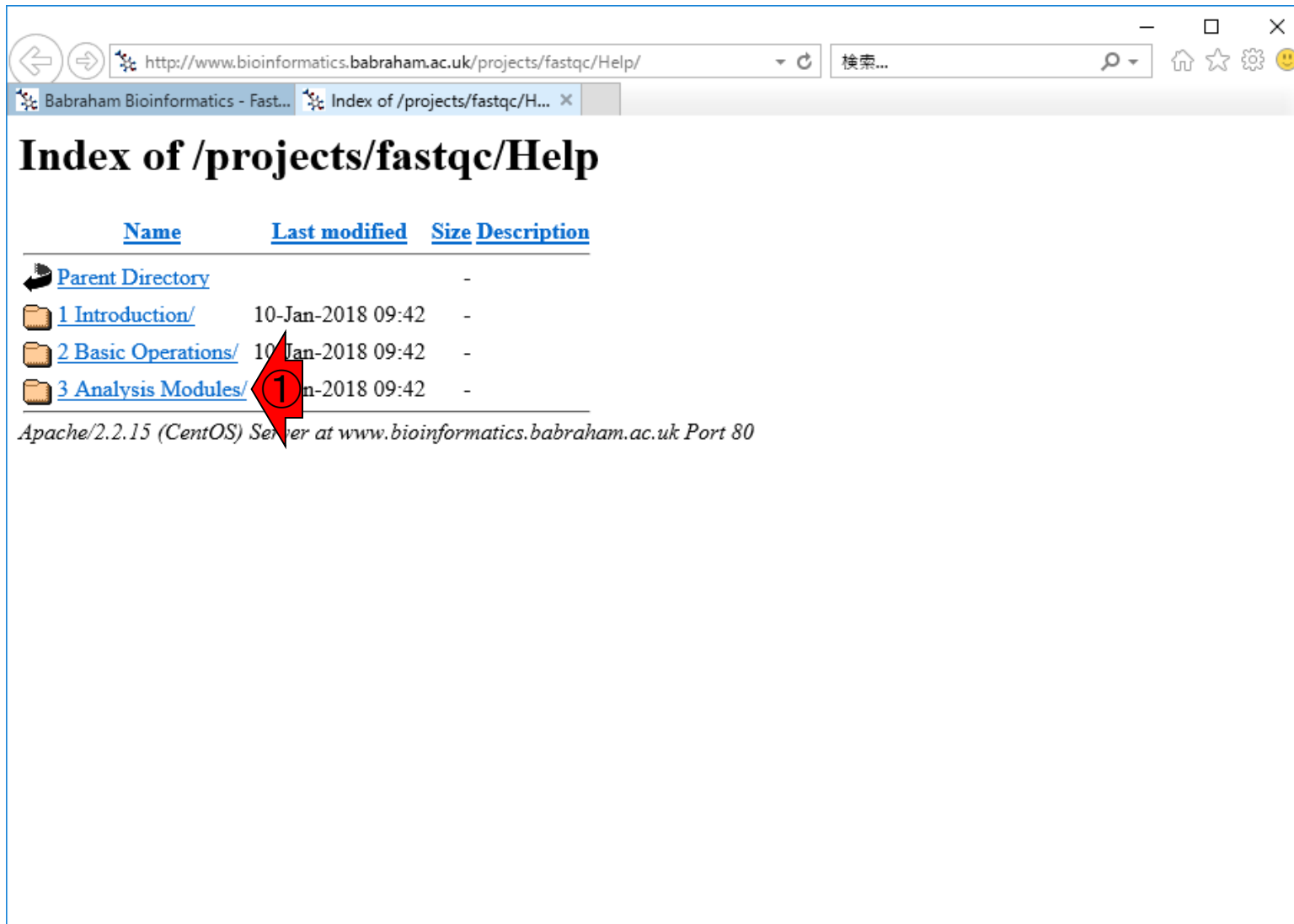
Documentation

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

Example Reports

- [Good Illumina Data](#)
- [Bad Illumina Data](#)

参考情報



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Index of /projects/fastqc/Help

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
Parent Directory	-	-	-
1 Introduction/	10-Jan-2018 09:42	-	-
2 Basic Operations/	10-Jan-2018 09:42	-	-
3 Analysis Modules/	10-Jan-2018 09:42	-	-

Apache/2.2.15 (CentOS) Server at www.bioinformatics.babraham.ac.uk Port 80

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

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
Parent Directory	-	-	-
1 Basic Statistics.html	10-Jan-2018 09:42	1.8K	
2 Per Base Sequence Quality.html	10-Jan-2018 09:42	3.6K	
3 Per Sequence Quality Scores.html	10-Jan-2018 09:42	1.7K	
4 Per Base Sequence Content.html	10-Jan-2018 09:42	3.5K	
5 Per Sequence GC Content.html	10-Jan-2018 09:42	1.8K	
6 Per Base N Content.html	10-Jan-2018 09:42	1.8K	
7 Sequence Length Distribution.html	10-Jan-2018 09:42	1.2K	
8 Duplicate Sequences.html	10-Jan-2018 09:42	5.9K	
9 Overrepresented Sequences.html	10-Jan-2018 09:42	2.4K	
10 Adapter Content.html	10-Jan-2018 09:42	2.4K	
11 Kmer Content.html	10-Jan-2018 09:42	2.5K	
12 Per Tile Sequence Quality.html	10-Jan-2018 09:42	2.2K	
duplication_levels.png	10-Jan-2018 09:42	20K	
kmer_profiles.png	10-Jan-2018 09:42	75K	
per_base_gc_content.png	10-Jan-2018 09:42	14K	
per_base_n_content.png	10-Jan-2018 09:42	13K	
per_base_quality.png	10-Jan-2018 09:42	9.8K	

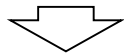
FastQC

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)

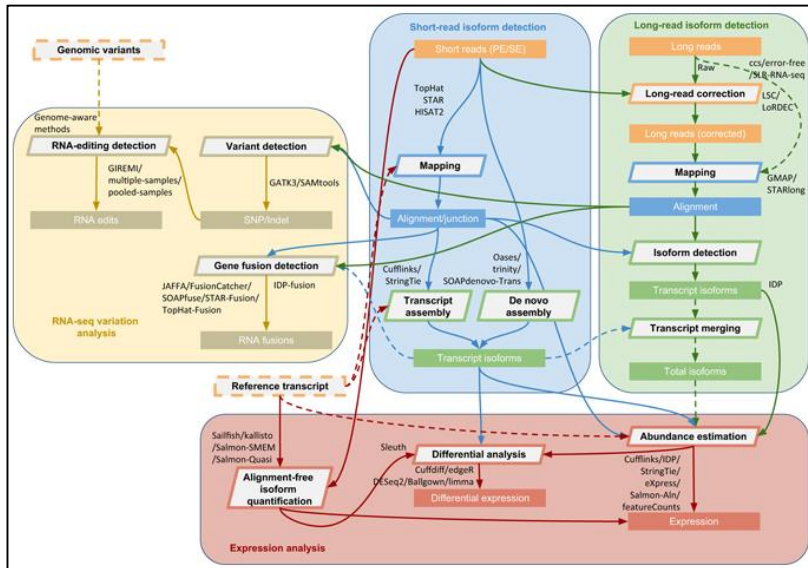
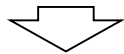


NGSリードデータ(FASTQファイル)



前処理 (preprocessing) or Quality Control (QC)

1

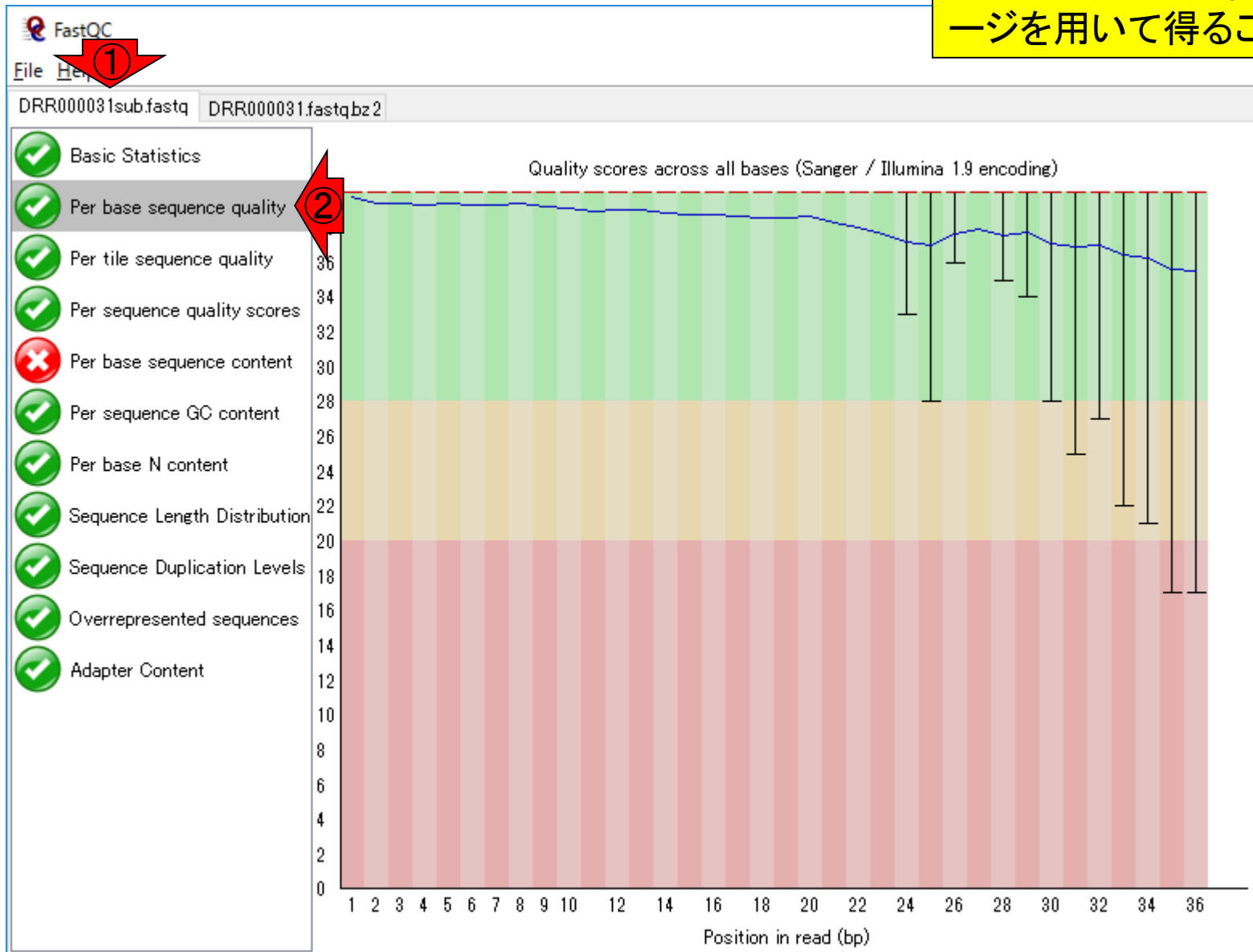


Contents

- 公共DB関連のTips
 - 公共DB、Linux
 - 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qrrc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)

このウェブページは、ツール済みで的にまとめた

What's new

- Silhouette
- Silhouette
- 「平成29年

- 門田から
- はじめに
- 参考資料
- 参考資料

- 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](#)
- 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 | [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](#)

前処理 | クオリティチェック | qrcq

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

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

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

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

```
in_f <- "SRR037439.fastq" #入力ファイル名を指定してin_fに格納
library(qrcq) #パッケージの読み込み
#入力ファイルの読み込み
fastq <- readSeqFile(in_f, quality="sanger") #in_fで指定したファイルの読み込み
#本番
makeReport(fastq) #htmlレポートの作成
```

qrqc実行

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

```
in_f <- "DRR000031sub.fastq" #入力ファイル名を指定してin_fに格納  
  
#必要なパッケージをロード  
library(qrhc) #パッケージの読み込み  
  
#入力ファイルの読み込み  
fastq <- readSeqFile(in_f, quality="sanger") #in_fで指定したファイルの読み込み  
  
#本番  
makeReport(fastq)
```

```
R Console  
  
要求されたパッケージ testthat をロード中です  
警告メッセージ:  
パッケージ 'GenomicRanges' はバージョン 3.4.4 の R の$  
>  
> #入力ファイルの読み込み  
> fastq <- readSeqFile(in_f, quality="sanger") #in_fで指$  
>  
> #本番  
> makeReport(fastq) #htmlレポート$  
警告: Ignoring unknown aesthetics: y  
'geom_smooth()' using method = 'gam'  
Report written to directory './DRR000031sub-report'.  
> |
```

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 (平成30年度)

1. 平成30年05月08日

[講義資料PDF](#)

[.gff3ファイル](#) (約1.3MB)

[.faファイル](#) (約2.2MB)

(Rで)塩基配列解析

(Rで)マイクロアレイデータ解析

[plasmid1.gff3](#)(課題用)

[plasmid2.gff3](#)(課題用)

2. 平成30年05月15日

[講義資料PDF](#)

[DRR000031sub.fastq](#)

[RNA-QC-chain : Zhou et al., BMC Genomics, 2018](#)

[Biostar : Parnell et al., PLoS Comput Biol., 2011](#)

[FastQC](#)

[DRR000031sub_fastqc.html](#)

[DRR000031_fastqc.html](#)(課題用)

[report.html](#)([qrqc](#)を用いたQC結果)

3. 平成30年05月22日

General Information

File: DRR000031sub.fastq

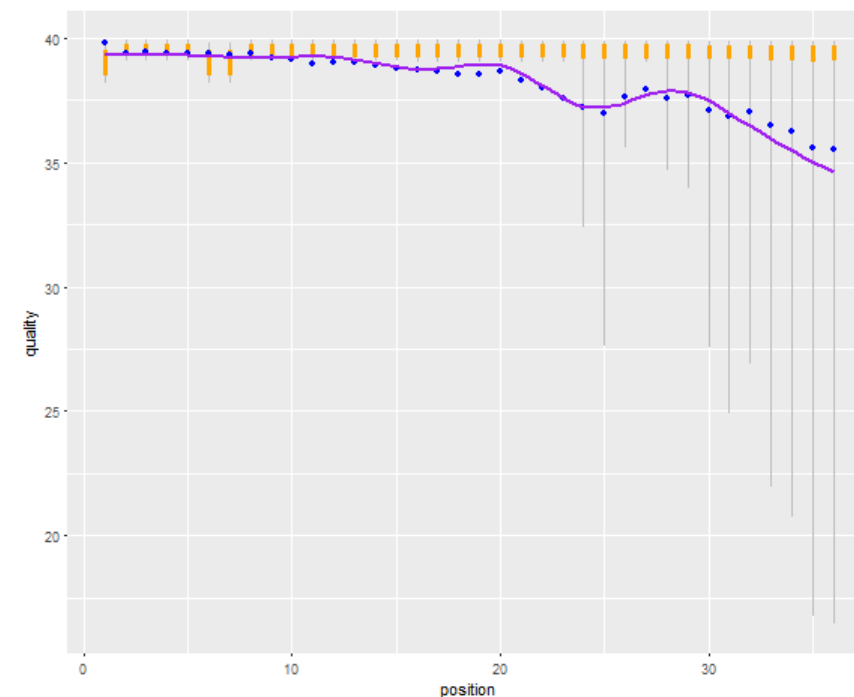
Type: FASTQ

Sequence Length Range: 36 to 36

Total Sequences: 6000

Unique Sequences: 5368

Quality by Position



Grey lines: 10% and 90% quartiles

Orange lines: 25% and 75% quartiles

Blue point: median

Green dash: mean

Purple line: lowess curve