

## 第3部:NGS解析(中～上級)

～Linux環境でのデータ解析:JavaやRの利用法～

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

門田幸二(かどた こうじ)

[kadota@iu.a.u-tokyo.ac.jp](mailto:kadota@iu.a.u-tokyo.ac.jp)

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

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# まずはFaQCs実行

## (Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランскриプトーム、正規化、発現変動、統計、モ

(last modified 2016/05/12) NEW

- ・書籍 | 日本乳酸菌学会誌 | について (last modified 2016/05/12) NEW
- ・書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2015/09/11)
- ・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 2015/09/11)
- ・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 2015/09/11)
- ・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 2015/09/11)
- ・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 2016/05/12) ①
- ・書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ (last modified 2016/05/12) ②
- ・書籍 | 日本乳酸菌学会誌 | 第7回統計的解析 (last modified 2016/05/12) ③
- ・インストロ

### 書籍 | 日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC

日本乳酸菌学会誌の第5回分です。Linuxコマンドのリリース先は主に日経BP社様です。

#### FaQCs(ver. 1.34)によるQC ②

- ・連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。  
連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v  
fastqc2 -q SRR616268sub_1.fastq.gz --outdir=/home/iu/Desktop/mac_share  
fastqc2 -q SRR616268sub_2.fastq.gz --outdir=/home/iu/Desktop/mac_share
```

連載第4回[W9-5]の手順通りに行ったヒトは、fastqc2がFastQC ver. 0.11.3の実行プログラムになっているはずです。実行結果ファイル [SRR616268sub\\_1\\_fastqc.html](#)(forward側)では、(古いバージョンFastQC ver. 0.10.1の結果ですが)連載第4回[W8-6]に示すように、Overrepresented sequencesという項目中に「TruSeq Adapter, Index2と3」がリード中に含まれていることがわかります。また、[SRR616268sub\\_2\\_fastqc.html](#)(reverse側)では、「 Illumina Single End PCR Primer 1」がリード中に含まれていることがわかります。

- ・Skewer: [Jiang et al., BMC Bioinformatics, 2014](#)
- ・FaQCs: [Lo and Chain, BMC Bioinformatics, 2014](#)
- ・図1。第4回の[W17-3]と基本的に同じ。[W1-1]

cd ~/Documents/cnp017156

①乳酸菌NGS連載第5回のサイト。②のあたりまでページ下部(概ね4ページ分)に移動。  
③のあたりにコピペ用コマンドがあります

# まずはFaQCs実行

FaQCs(ver. 1.34)によるQC

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。  
連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v  
fastqc2 -q SRR616268sub_1.fastq.gz --outdir=/home/iu/Desktop/mac_share  
fastqc2 -q SRR616268sub_2.fastq.gz --outdir=/home/iu/Desktop/mac_share
```

連載第4回[W9-5]の手順通りに行ったヒトは、fastqc2が FastQC ver. 0.11.3の実行プログラムになっているはずです。実行結果ファイル [SRR616268sub\\_1\\_fastqc.html](#)(forward側)では、(古いバージョンFastQC ver. 0.10.1の結果ですが)連載第4回W8-6に示すように、Overrepresented sequencesという項目中に「TruSeq Adapter, Index2と3」がリード中に含まれていることがわかります。また、[SRR616268sub\\_2\\_fastqc.html](#) (reverse側)では、「Illumina Single End PCR Primer 1」がリード中に含まれていることがわかります。

- Skewer: [Jiang et al., BMC Bioinformatics, 2014](#)
- FaQCs: [Lo and Chain, BMC Bioinformatics, 2014](#)

- 図1。第4回の[W17-3]と基本的に同じ。[W1-1]

```
cd ~/Documents/srp017156  
rm -f hoge*  
rm -f JS*  
rm -rf result*  
rm -f *.bz2
```

```
pwd  
ls -lh  
fastqc2 -v  
FaQCs.pl -v  
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2  
ls result2
```

①ここです。②作業ディレクトリを変更。③は無視でよい(第4回最後の状態から不必要的ものを削除しているだけ)。②をコピペ後に、④をまずはコピペ実行。約20分

スライド18まで説明

①

②

③

④

# W1-1 : FaQCs実行

```
File Edit View Terminal Help
iu@bielinux[srp017156] pwd
/home/iu/Documents/srp017156
iu@bielinux[srp017156] ls -lh
total 139M
-rw-rw-r-- 1 iu iu 74M 12月  9 15:24 SRR616268sub_1.fastq.gz
-rw-rw-r-- 1 iu iu 66M 12月  9 15:24 SRR616268sub_2.fastq.gz
iu@bielinux[srp017156] fastqc2 -v
FastQC v0.11.4
iu@bielinux[srp017156] FaQCs.pl -v
Version: 1.34
iu@bielinux[srp017156] time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR6
16268sub_2.fastq.gz -d result2
Bwa extension trimming algorithm is used.
Processing SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz file
Processed 2000000/2000000
Post Trimming Length(Mean, Std, Median, Max, Min) of 1972635 reads with Over
all quality 36.37
(99.33, 8.62, 107.0, 107, 50)
FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d 113
4.61s user 11.10s system 96% cpu 19:41.18 total
iu@bielinux[srp017156] [11:57午前]
[11:57午前]
[12:17午後]
```

第4回W17-3と基本的に同じ。①ls実行結果で入力として利用する2つのgzip圧縮ファイル(\*.fastq.gz)が見られる。ファイルサイズが多少違っていても気にしない。②第4回W9-1で示すように、2015年10月9日以降にFastQCをインストールしたヒトは、ここがver. 0.11.4以上。③がFaQCs実行部分。④result2ディレクトリに結果を保存するように指定している

# Contents

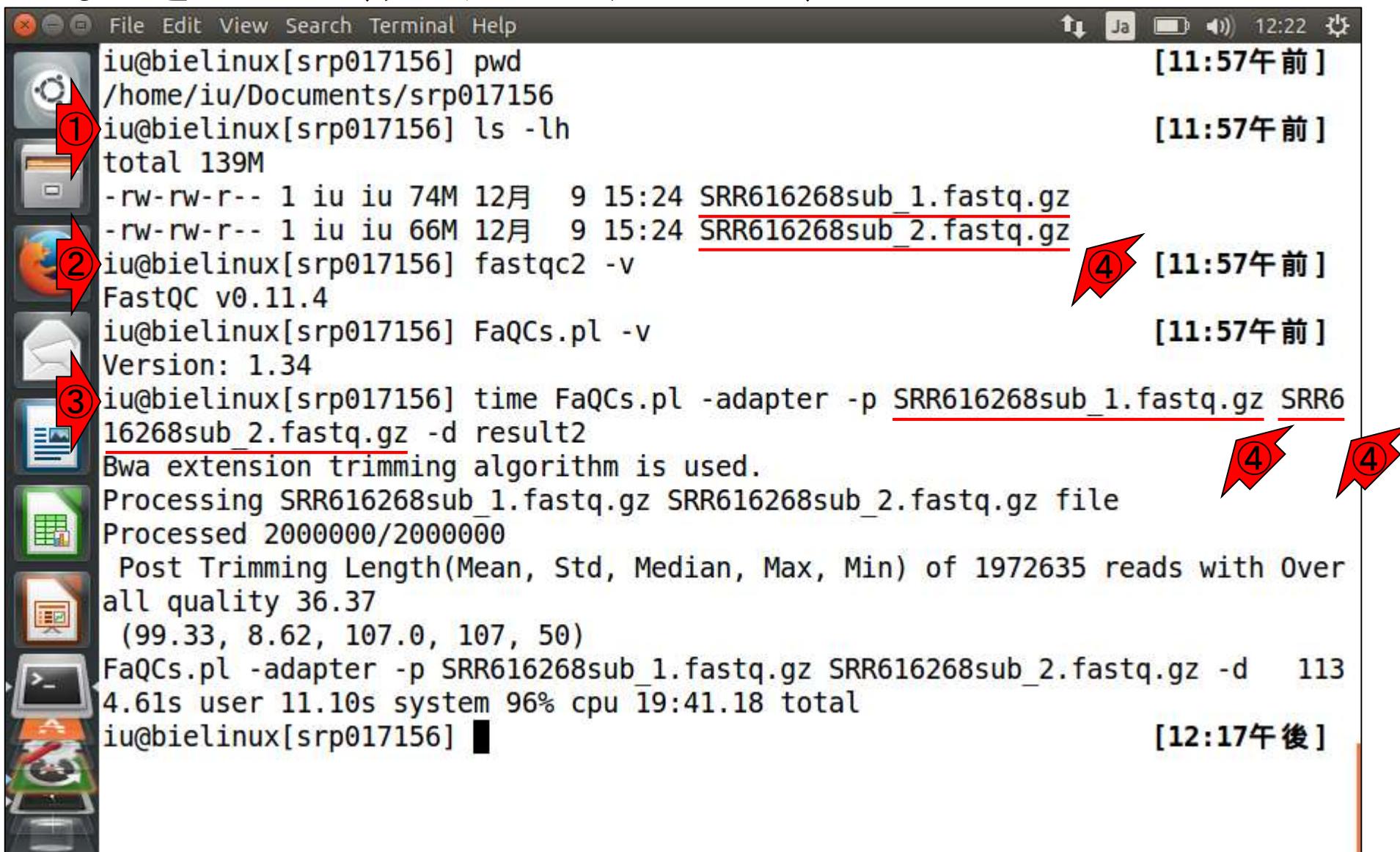
- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



④まずは赤下線部分のFaQCs  
の入力ファイルについておさらい

# おさらい(入力ファイル)

```
File Edit View Search Terminal Help [12:22]
①
iu@bielinux[srp017156] pwd [11:57午前]
/home/iu/Documents/srp017156
iu@bielinux[srp017156] ls -lh [11:57午前]
total 139M
-rw-rw-r-- 1 iu iu 74M 12月 9 15:24 SRR616268sub_1.fastq.gz
-rw-rw-r-- 1 iu iu 66M 12月 9 15:24 SRR616268sub_2.fastq.gz
iu@bielinux[srp017156] fastqc2 -v [11:57午前]
FastQC v0.11.4
iu@bielinux[srp017156] FaQCs.pl -v [11:57午前]
Version: 1.34
iu@bielinux[srp017156] time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2 [11:57午前]
Bwa extension trimming algorithm is used.
Processing SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz file
Processed 2000000/2000000
Post Trimming Length(Mean, Std, Median, Max, Min) of 1972635 reads with Overall quality 36.37
(99.33, 8.62, 107.0, 107, 50)
FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d 113
4.61s user 11.10s system 96% cpu 19:41.18 total
iu@bielinux[srp017156] [12:17午後]
```



# おさらい(入力ファイル)

## (Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランскриプトーム、正規化、発現変動、統計、モ

(last modified 2016/05/12) NEW

- ・書籍 | 日本乳酸菌学会誌 | について (last modified 2016/05/12) NEW
- ・書籍 | 日本乳酸菌学会誌 | 第1回インストラクション (last modified 2015/09/11)
- ・書籍 | 日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ (last modified 2016/05/12)
- ・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 2016/05/12)
- ・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 2016/05/12)
- ・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 2016/05/12)
- ・書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ (last modified 2016/05/23)
- ・書籍 | 日本乳酸菌学会誌 | 第7回統計的解析 (last modified 2016/05/23)
- ・書籍 | 日本乳酸菌学会誌 | 第8回機械学習による発現変動解析 (last modified 2016/05/23)
- ・インストラクション (last modified 2016/05/12)

### 書籍 | 日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC

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

- ・第5回分PDF
- ・ウェブ資料PDF(2015)

#### Linuxコマンド

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

#### 乳酸菌(*Lactobacillus casei* 12A) paired-end RNA-seqデータのおさらい

Illumina HiSeq 2000で得られたデータです。

- ・公共DB
  - [SRR616268](#) (DDBJ SRA; DRA)
  - [SRR616268](#) (EMBL-EBI ENA)
  - [SRR616268](#) (NCBI SRA)
- ・DRAから取得したbzip2圧縮ファイル。2つとも行数は539,023,984行、リード数は134,755,996リード(約1.35億)。
  - forward側: [SRR616268\\_1.fastq.bz2](#), 7,662,128,101 bytes (約7.2GB)、全リード107 bp。
  - reverse側: [SRR616268\\_2.fastq.bz2](#), 7,017,031,734 bytes (約6.6GB)、全リード93 bp。

```
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042
ls -l *.fastq.bz2
bzip2 -dc SRR616268_1.fastq.bz2 | wc
bzip2 -dc SRR616268_2.fastq.bz2 | wc
```

①乳酸菌NGS連載第5回のサイト  
②のおさらい部分を軽く説明。  
場所は、さきほどのFaQCs実行コマンドの上部。スライドを見るだけ

# おさらい(入力ファイル)

乳酸菌(*Lactobacillus casei* 12A) paired-end RNA-seqデータのおさらい

Illumina HiSeq 2000で得られたデータです。

- 公共DB
  - [SRR616268](#) (DDBJ SRA; DRA)
  - [SRR616268](#) (EMBL-EBI ENA)
  - [SRR616268](#) (NCBI SRA)
- DRAから取得したbzip2圧縮ファイル。2つとも行数は539,023,984行、リード数は134,755,996リード(約1.35億)。
  - forward側: [SRR616268\\_1.fastq.bz2](#)、7,662,128,101 bytes (約7.2GB)、全リード107 bp.
  - reverse側: [SRR616268\\_2.fastq.bz2](#)、7,017,031,734 bytes (約6.6GB)、全リード93 bp.

```
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042  
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042  
ls -l *.fastq.bz2  
bzip2 -dc SRR616268_1.fastq.bz2 | wc  
bzip2 -dc SRR616268_2.fastq.bz2 | wc
```

①乳酸菌のpaired-end RNA-seqデータ。  
②Illumina HiSeq 2000で得られ、③オリジナルはそれぞれ約1.35億リード。④リードの長さは、forward側が107 bp、reverse側が93 bp。bzip2圧縮ファイル状態でも、計約14GBに達する!見るだけ



# おさらい(入力ファイル)

乳酸菌(*Lactobacillus casei* 12A) paired-end RNA-seqデータのおさらい

Illumina HiSeq 2000で得られたデータです。

- ・ 公生DB

- ① [SRR616268](#) (DDBJ SRA; DRA)
  - [SRR616268](#) (EMBL-EBI ENA)
  - [SRR616268](#) (NCBI SRA)

- ・ DRAから取得したbzip2圧縮ファイル。2つとも行数は539,023,984行、リード数は134,755,996リード(約1.35億)。
  - forward側: [SRR616268\\_1.fastq.bz2](#)、7,662,128,101 bytes (約7.2GB)、全リード 107 bp。
  - reverse側: [SRR616268\\_2.fastq.bz2](#)、7,017,031,734 bytes (約6.6GB)、全リード 93 bp。

```
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042  
#wget -c ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA061/SRA061483/SRX2042  
ls -l *.fastq.bz2  
bzip2 -dc SRR616268_1.fastq.bz2 | wc } ③  
bzip2 -dc SRR616268_2.fastq.bz2 | wc }
```

①DDBJ SRA (DRA)から、②(#でコメントアウトしているが)FASTQファイルをダウンロードすると、bzip2圧縮ファイル(.bz2)として得られる。③は、元の.bz2ファイルを残しつつ、ファイルの行数をwcでカウントしている。④がその結果。⑤4で割って、リード数情報を得ている



# おさらい(入力ファイル)

①オリジナルはデカすぎるので、最初の  
②4000000行分のみからなるサブセットを  
③gzipした、④の.gzファイルが…

乳酸菌(*Lactobacillus casei* 12A) paired-end RNA-seqデータのおさらい

Illumina HiSeq 2000で得られたデータです。

- ・ 公共DB
  - [SRR616268](#) (DDBJ SRA; DRA)
  - [SRR616268](#) (EMBL-EBI ENA)
  - [SRR616268](#) (NCBI SRA)
- ・ DRAから取得したbzip2圧縮ファイル。2つとも行数は539,023,984行、リード数は134,755,996リード(約1.35億)。
  - forward側: [SRR616268\\_1.fastq.bz2](#), 7,662,128,101 bytes (約7.2GB), 全リード 107 bp.
  - reverse側: [SRR616268\\_2.fastq.bz2](#), 7,017,031,734 bytes (約6.6GB), 全リード 93 bp.

```
#wget -c
#wget -c
ls -l *
bzip2 -d
bzip2 -d
  • サブセットの取得。2つとも行数は4,000,000行、リード数は1,000,000リード(100万)。
  連載第3回の[W25-2]あたり。

- forward側: SRR616268sub\_1.fastq.gz, 74,906,576 bytes (約75MB), 全リード 107 bp.
- reverse側: SRR616268sub\_2.fastq.gz, 67,158,462 bytes (約67MB), 全リード 93 bp.


① bzip2 -dc SRR616268_1.fastq.bz2 | head -n 4000000 | gzip > SRR616268sub_1.fastq.gz
② bzip2 -dc SRR616268_2.fastq.bz2 | head -n 4000000 | gzip > SRR616268sub_2.fastq.gz
③
④
```

- ・ サブセットのgzip圧縮ファイルは以下のコピペでも取得可能。

```
wget -c http://www.iu.a.u-tokyo.ac.jp/~kadota/R_seq/SRR616268sub_1.fastq.gz
wget -c http://www.iu.a.u-tokyo.ac.jp/~kadota/R_seq/SRR616268sub_2.fastq.gz
ls -l *.fastq.gz
gzip -dc SRR616268sub_1.fastq.gz | wc
gzip -dc SRR616268sub_2.fastq.gz | wc
```

# おさらい(入力ファイル)

```
File Edit View Terminal Help
iu@bielinux[srp017156] pwd [11:57午前]
/home/iu/Documents/srp017156
iu@bielinux[srp017156] ls -lh [11:57午前]
total 139M
-rw-rw-r-- 1 iu iu 74M 12月  9 15:24 SRR616268sub_1.fastq.gz
-rw-rw-r-- 1 iu iu 66M 12月  9 15:24 SRR616268sub_2.fastq.gz
iu@bielinux[srp017156] fastqc2 -v [11:57午前]
FastQC v0.11.4
iu@bielinux[srp017156] FaQCs.pl -v [11:57午前]
Version: 1.34
iu@bielinux[srp017156] time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2 [11:57午前]
Bwa extension trimming algorithm is used.
Processing SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz file
Processed 2000000/2000000
Post Trimming Length(Mean, Std, Median, Max, Min) of 1972635 reads with Overall quality 36.37
(99.33, 8.62, 107.0, 107, 50)
FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d 113
4.61s user 11.10s system 96% cpu 19:41.18 total
iu@bielinux[srp017156] [12:17午後]
```

# おさらい(FaQCs)

FaQCs(ver. 1.34)によるQC

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータで[FastQC](#)連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v  
fastqc2 -q SRR616268sub_1.fastq.gz --outdir=/home/iu/Desktop/mac_share  
fastqc2 -q SRR616268sub_2.fastq.gz --outdir=/home/iu/Desktop/mac_share
```

①FaQCs原著論文。是非使ってもらいたいというわけではなく、分散型バージョン管理システムGitHub上で公開されているプログラムの一例(連載第4回W15)として、またプログラム内部で用いるPerlモジュール(第4回W15)のインストール法など、これまで紹介してこなかったテクニックの伝授用としての意味合いが大きい

連載第4回[W9-5]の手順通りに行ったヒトは、fastqc2が FastQC ver. 0.11.3の実行プログラムになっているはずです。実行結果ファイル[SRR616268sub\\_1\\_fastqc.html](#)(forward側)では、(古いバージョンFastQC ver. 0.10.1の結果ですが)連載第4回W8-6に示すように、Overrepresented sequencesという項目中に「TruSeq Adapter, Index2と3」がリード中に含まれていることがわかります。また、[SRR616268sub\\_2\\_fastqc.html](#) (reverse側)では、「Illumina Single End PCR Primer 1」がリード中に含まれていることがわかります。

- Skewer: [Jiang et al., BMC Bioinformatics, 2014](#)
- FaQCs: [Lo and Chain, BMC Bioinformatics, 2014](#)
- 図1。第4回の[W17-3]と基本的に同じ。[W1-1]

```
cd ~/Documents/srp017156  
rm -f hoge*  
rm -f JS*  
rm -rf result*  
rm -f *.bz2  
  
pwd  
ls -lh  
fastqc2 -v  
FaQCs.pl -v  
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2  
ls result2
```



# おさらい(FaQCs)

BMC Bioinformatics. 2014 Nov 19;15:366. doi: 10.1186/s12859-014-0366-2.

## Rapid evaluation and quality control of next genera

Lo CC<sup>1</sup>, Chain PS<sup>2,3</sup>.

### ⊕ Author information

#### Abstract

**BACKGROUND:** Next generation sequencing (NGS) technologies that parallelize the sequencing process and produce thousands to millions, or even hundreds of millions of sequences in a single sequencing run, have revolutionized genomic and genetic research. Because of the vagaries of any platform's sequencing chemistry, the experimental processing, machine failure, and so on, the quality of sequencing reads is never perfect, and often declines as the read is extended. These errors invariably affect downstream analysis/application and should therefore be identified early on to mitigate any unforeseen effects.

**RESULTS:** Here we present a novel FastQ Quality Control Software (FaQCs) that can rapidly process large volumes of data, and which improves upon previous solutions to monitor the quality and remove poor quality data from sequencing runs. Both the speed of processing and the memory footprint of storing all required information have been optimized via algorithmic and parallel processing solutions. The trimmed output compared side-by-side with the original data is part of the automated PDF output. We show how this tool can help data analysis by providing a few examples, including an increased percentage of reads recruited to references, improved single nucleotide polymorphism identification as well as de novo sequence assembly metrics.

**CONCLUSION:** FaQCs combines several features of currently available applications into a single, user-friendly process, and includes additional unique capabilities such as filtering the PhiX control sequences, conversion of FASTQ formats, and multi-threading. The original data and trimmed summaries are reported within a variety of graphics and reports, providing a simple way to do data quality control and assurance.

FaQCs自体はQuality Control用プログラム。(平成27年度受講生の要望として挙がっていた)paired-endデータの①クオリティフィルタリングや、(特にIlluminaの)アダプター除去がadapterオプションのみで簡単にできるので、精度に期待はしていなかったが使用感は思ったよりもよかったです(個人の感想です)という代物です

①

# おさらい(FaQCs)

## FaQCs(ver. 1.34)によるQC

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。  
連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v  
fastqc2 -q SRR616268sub_1.fastq.gz --outdir=/home/iu/Desktop/mac_share  
fastqc2 -q SRR616268sub_2.fastq.gz --outdir=/home/iu/Desktop/mac_share
```

連載第4回[W9-5]の手順通りに行ったヒトは、fastqc2がFastQC ver. 0.11.3の実行プログラムになっているはずです。実行結果ファイル[SRR616268sub\\_1\\_fastqc.html](#)(forward側)では、(古いバージョンFastQC ver. 0.10.1の結果ですが)連載第4回W8-6に示すように、Overrepresented sequencesという項目中に「TruSeq Adapter, Index2と3」がリード中に含まれていることがわかります。また、[SRR616268sub\\_2\\_fastqc.html](#)(reverse側)では、「Illumina Single End PCR Primer 1」がリード中に含まれていることがわかります。

- Skewer: [Jiang et al., BMC Bioinformatics, 2014](#)
- FaQCs: [Lo and Chain, BMC Bioinformatics, 2014](#)

- ① 第4回の[W17-3]と基本的に同じ。[W1-1]

```
cd ~/Documents/srp017156  
rm -f hoge*  
rm -f JS*  
rm -rf result*  
rm -f *.bz2  
  
pwd  
ls -lh  
fastqc2 -v  
FAQCs.pl -v  
time FAQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2  
ls result2
```

# おさらい(FaQCs)

GitHub This repository Search Explore Features Enterprise Blog Sign up Sign in

LANL-Bioinformatics / FaQCs forked from chianchi/FaQCs

45 commits 1 branch 1 release 1 contributor

branch master + FaQCs / +

This branch is 23 commits ahead, 16 commits behind chianchi/master

fix type

chianchi authored on Mar 21

example adjust margin of pdf plots 3 months ago

galaxy\_module Update FaQCs and 6 months ago

lib Add option "-5end off <bool>" Turn off trimming from 5' end. 4 months ago

COPYRIGHT add LICENSE a year ago

FaQCs.pl patch for jellyfish v2 3 months ago

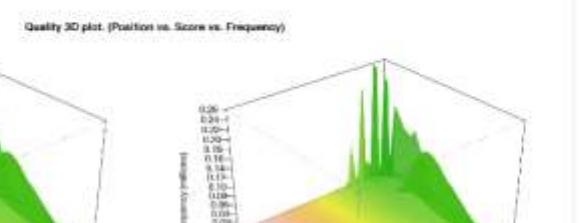
LICENSE add LICENSE a year ago

README.md fix typo 2 months ago

README.md

## FaQCs: Quality Control of Next Generation Sequencing Data

Quality 3D plot. (Position vs. Score vs. Frequency)



第4回W17-3。GitHubのページ中央あたりに、利用可能なオプションの説明あり。例えば①を眺めることで、-adapterオプションをつけられることを学ぶ。漫然と眺めるのではなく、FaQCsを通じてGitHubのページ構成や見方の経験を積むべし

## BASIC

- Trimming by quality 5 and filtering reads with any ambiguous base or low complexity.

```
$ perl FaQCs.pl -p reads1.fastq reads2.fastq -d out_directory
```

- Quality check only on subsamples of input, no trimming and filtering.

```
$ perl FaQCs.pl -p reads1.fastq reads2.fastq -d out_directory -qc_only
```

## Full USAGE

Usage: perl FaQCs.pl [options] [-u unpaired.fastq] -p reads1.fastq reads2.fastq -d out\_directory  
Version 1.34

Input File: (can use more than once)

-u <Files> Unpaired reads

-p <Files> Paired reads in two files and separate by space

Trim:

-mode "HARD" or "BWA" or "BWA\_plus" (default BWA\_plus)  
BWA trim is NOT A HARD cutoff! (see bwa's bwa\_trim\_read() function in bwa.c)

-q <INT> Targets # as quality level (default 5) for trimming

-5end <INT> Cut # bp from 5' end before quality trimming/filtering

-3end <INT> Cut # bp from 3' end before quality trimming/filtering

-adapter <bool> Filter reads with illumina adapter/primers (default: no)  
-rate <FLOAT> Mismatch ratio of adapters' length (default: 0.2, allow 20%)

-artifactFile <File> additional artifact (adapters/primers/contaminations) reference file

Filters:

-min\_L <INT> Trimmed read should have to be at least this minimum length (default: 0)

-avg\_q <NUM> Average quality cutoff (default: 0, no filtering)

①

# おさらい(FaQCs)

FaQCs(ver. 1.34)によるQC

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。  
連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v  
fastqc2 -q SRR616268sub_1.fastq.gz --outdir=/home/iu/Desktop/mac_share  
fastqc2 -q SRR616268sub_2.fastq.gz --outdir=/home/iu/Desktop/mac_share
```

連載第4回[W9-5]の手順通りに行ったヒトは、fastqc2がFastQC ver. 0.11.3の実行プログラムになっているはずです。実行結果ファイル[SRR616268sub\\_1\\_fastqc.html](#)(forward側)では、(古いバージョンFastQC ver. 0.10.1の結果ですが)連載第4回W8-6に示すように、Overrepresented sequencesという項目中に「TruSeq Adapter, Index2と3」がリード中に含まれていることがわかります。また、[SRR616268sub\\_2\\_fastqc.html](#)(reverse側)では、「Illumina Single End PCR Primer 1」がリード中に含まれていることがわかります。

- Skewer: [Jiang et al., BMC Bioinformatics, 2014](#)
- FaQCs: [Lo and Chain, BMC Bioinformatics, 2014](#)
- 図1。第4回の[W17-3]と基本的に同じ。[W1-1]

```
cd ~/Documents/srp017156  
rm -f hoge*  
rm -f JS*  
rm -rf result*  
rm -f *.bz2  
  
pwd  
ls -lh  
fastqc2 -v  
FaQCs.pl -v  
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2  
ls result2
```



①今実行中のコマンドには-adapterオプションがついている。この経緯について説明。まず、FaQCsを最初に実行したとき(第4回W17-1)は、-adapterをつけていなかった

# おさらい(FaQCs)

## FastQC Report

### Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)
- [Kmer Content](#)

Sequence	Count	Mean Quality	Notes
CTTTCGACAAATGGACCTTATCGCTCACTGTC	1349	0.13802869646272048	No Hit
ATTTAGCCTTGGGAGATGGCCTCCCGGATTG	1341	0.13721014229541006	No Hit
CGCTACTCATGCCGCATTCTCACTTCAAGG	1336	0.13669854594084105	No Hit
CCGGGGTGCTTTCACCTTCCCTACGGTACTGGTCACTATCGGTAC	1319	0.13495911833530638	No Hit
GCGGCCGCCAGCTATGTATTCACTGACAAGCAATACTGATGTGACT	1305	0.13352664854251314	No Hit
AGATCGGAAGAGCACACGCTGAACTCCAGTCACCGATGTATCTGTATG	1296	0.13260577510428892	TruSeq Adapter, Index 2 (100% over 49bp)
GTTTGGCTCTTCCCACCTCGCTCGCCGCTACTATGGGAATCGAGTTTC	1262	0.12912691989321962	No Hit
CGTCCCTCCATCGCTAACAAAAAAACTAGTGCAGGAATCTAACCTG	1258	0.12871764280956438	No Hit
CACACGGTTTCAGGAACTGTTCACTCCCCCTCGGGGTGCTTTCACCT	1257	0.1286153235386506	No Hit
CCCTAGTTCAAACAGTCGCTCACCTCCATGATCCATCCTCCGAGGCTAAC	1249	0.12779676937134016	No Hit
GTCATTTGCCAGTTCTAACGAGAGTTCGCTCGTCACCTGAGGATA	1249	0.12779676937134016	No Hit
ACCTTAACGGTACGGCTGTTCCCCTTCGACAATGGACCTTATCG	1223	0.1251364683275813	No Hit
CACAGTTCCGGTATTATGCTTAGCCCCGGTATATTTCGCGCAGTGCCA	1220	0.12482951051483988	No Hit
GGTCATTTGCCAGTTCTAACGAGAGTTCGCTCGTCACCTGAGGAT	1208	0.12360167926387423	No Hit
CCACTTAGCATAAATTAGGGACCTTAACGGTACGGTACGGCTGTTCCCC	1207	0.12349935999296043	No Hit
CCGGTTCAATTCTACAAAAGGCACGCCATTACCCGTTAACGGGTTTGA	1191	0.12186225165833958	No Hit
CTGCCGCCGGCCAGCTATGTATTCACTGACAAGCAATACTGATGTGA	1178	0.12053210113646014	No Hit

-adapterをつけずにFaQCsを実行したFASTQファイルを入力として、FastQCを実行した結果(第4回W17-2)。①Overrepresented sequencesを眺めると、②(Illuminaの)TruSeq Adapter, Index 2が見つかった。このことからデフォルトではアダプター配列除去までは行えないことを学び、第4回W17-3で-adapterをつけてみるといいんじゃないかと学び、それをまずは最初に実行した(という次第)

①

②

# W1-1 : FaQCs実行

```
File Edit View Search Terminal Help
iu@bielinux[srp017156] pwd
/home/iu/Documents/srp017156
iu@bielinux[srp017156] ls -lh
total 139M
-rw-rw-r-- 1 iu iu 74M 12月  9 15:24 SRR616268sub_1.fastq.gz
-rw-rw-r-- 1 iu iu 66M 12月  9 15:24 SRR616268sub_2.fastq.gz
iu@bielinux[srp017156] fastqc2 -v [11:57午前]
FastQC v0.11.4
iu@bielinux[srp017156] FaQCs.pl -v [11:57午前]
Version: 1.34
iu@bielinux[srp017156] time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2
Bwa extension trimming algorithm is used.
Processing SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz file
Processed 2000000/2000000
Post Trimming Length(Mean, Std, Median, Max, Min) of 1972635 reads with Overall quality 36.37
(99.33, 8.62, 107.0, 107, 50)
FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d 113
4.61s user 11.10s system 96% cpu 19:41.18 total
iu@bielinux[srp017156] ls result2 [12:17午後]
fastqCount.txt      QC.2.trimmed.fastq  QC.stats.txt
QC.1.trimmed.fastq  QC_qc_report.pdf   QC.unpaired.trimmed.fastq
iu@bielinux[srp017156] [12:25午後]
```

③最初にコピペ実行したFaQCsの実行時間は、約20分。画面上で見えているのは、赤枠内の一一番下の19:41.18という数値。これは19分41秒かかったことを示しており、ヒトによって異なる。④ result2ディレクトリをls。QC.1.trimmed.fastqとQC.2.trimmed.fastqがFaQCsの主な実行結果ファイル

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W1-2: FastQC実行

## FaQCs(ver. 1.34)によるQC

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v
fastqc2 -q SRR616268sub_1.fa
fastqc2 -q SRR616268sub_2.fa
```

連載第4回[W9-5]の手順通りに行なったです。実行結果ファイル[SRR616268sub\\_1.fastq.gz](#)と[SRR616268sub\\_2.fastq.gz](#)が生成されましたが、連載第4回[W8-6]に示すよと3がリード中に含まれていることが、「Illumina Single End PCR Primer 1」が

- [Skewer: Jiang et al., BMC Bioinformatics, 2012](#)
- [FaQCs: Lo and Chain, BMC Bioinformatics, 2013](#)
- 図1。第4回の[W17-3]と基本的に同じ。

```
cd ~/Documents/srp017156
rm -f hoge*
rm -f JS*
rm -rf result*
rm -f *.bz2
```

```
pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1
ls result2
```



```
pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2
ls result2
```

①左上と右下のスクリーンショットの重複(のりしろ)部分。②FastQCコピペ実行用コード。全部で6行分あるが、必要最小限は赤枠真ん中のfastqc2から始まる2行分だけ。それゆえpwd, ls, dateなどのコマンド実行結果はイチイチ説明しないが、「今どこで作業をしていて、入力ファイルがあるかどうか」の確認は自分でやっておこう。

FaQCs ver. 1.34を実行し、結果をresult2ディレクトリに保存(約25分)。result2ディレクトリ上には、計6ファイルが生成されている。そのうちのQC結果のサマリーレポートは次の2つ:[QC.stats.txt](#)と[QC qc\\_report.pdf](#)

- トリム後(FaQCs実行後)のデータを入力としてFastQC (ver. 0.11.3)をデフォルトで実行。[W1-2]



```
pwd
ls result2/*.fastq
fastqc2 -q result2/QC.1.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
fastqc2 -q result2/QC.2.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
date
ls -lh /home/iu/Desktop/mac_share/QC.*
```

出力結果を共有フォルダ(/home/iu/Desktop/mac\_share)に直接保存。実行結果ファイル[QC.1.trimmed\\_fastqc.html](#)(forward側)と[QC.2.trimmed\\_fastqc.html](#)(reverse側)とともに、Overrepresented sequences項目に見えていたアダプターやプライマー配列情報がなくなっているのがわかります。[W1-3]

- [Rockhopper 2: Tjaden B, Genome Biol., 2015](#)
- [QuasR: Gaidatzis et al., Bioinformatics, 2015](#)

# W1-2: FastQC実行

FaQCs(ver. 1.34)によるQ③

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11.3)を実行。連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v
fastqc2 -q SRR616268sub_1.fa
fastqc2 -q SRR616268sub_2.fa
```

連載第4回[W9-5]の手順通りに行なったです。実行結果ファイル [SRR616268sub\\_1.fastq.gz](#) と [SRR616268sub\\_2.fastq.gz](#) が生成されましたが、連載第4回[W8-6]に示すように「3'」がリード中に含まれていることが、「Illumina Single End PCR Primer 1」が

- [Skewer: Jiang et al., BMC Bioinformatics, 2012](#)
- [FaQCs: Lo and Chain, BMC Bioinformatics, 2013](#)
- 図1。第4回の[W17-3]と基本的に同じ。

```
cd ~/Documents/srp017156
rm -f hoge*
rm -f JS*
rm -rf result*
rm -f *.bz2

pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1
ls result2
```

```
pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2
ls result2
```

以後は徐々に冗長な説明を省略していくが、③のW1-2のようなウェブ資料番号情報を頼りにして、自分でコピペ用コードを探して有効利用してください。Bio-Linux上でコピペできないなどのバグに遭遇して再起動に時間がかかることがあるでしょうが、コピペを有効利用すれば追いつけます。④の部分をコピペしたのが次のスライド

FaQCs ver. 1.34を実行し、結果をresult2ディレクトリに保存(約25分)。result2ディレクトリ上には、計6ファイルが生成されている。そのうちのQC結果のサマリーレポートは次の2つ:[QC.stats.txt](#)と[QC qc\\_report.pdf](#)

- トリム後(FaQCs実行後)のデータを入力としてFastQC (ver. 0.11.3)をデフォルトで実行。[W1-2]

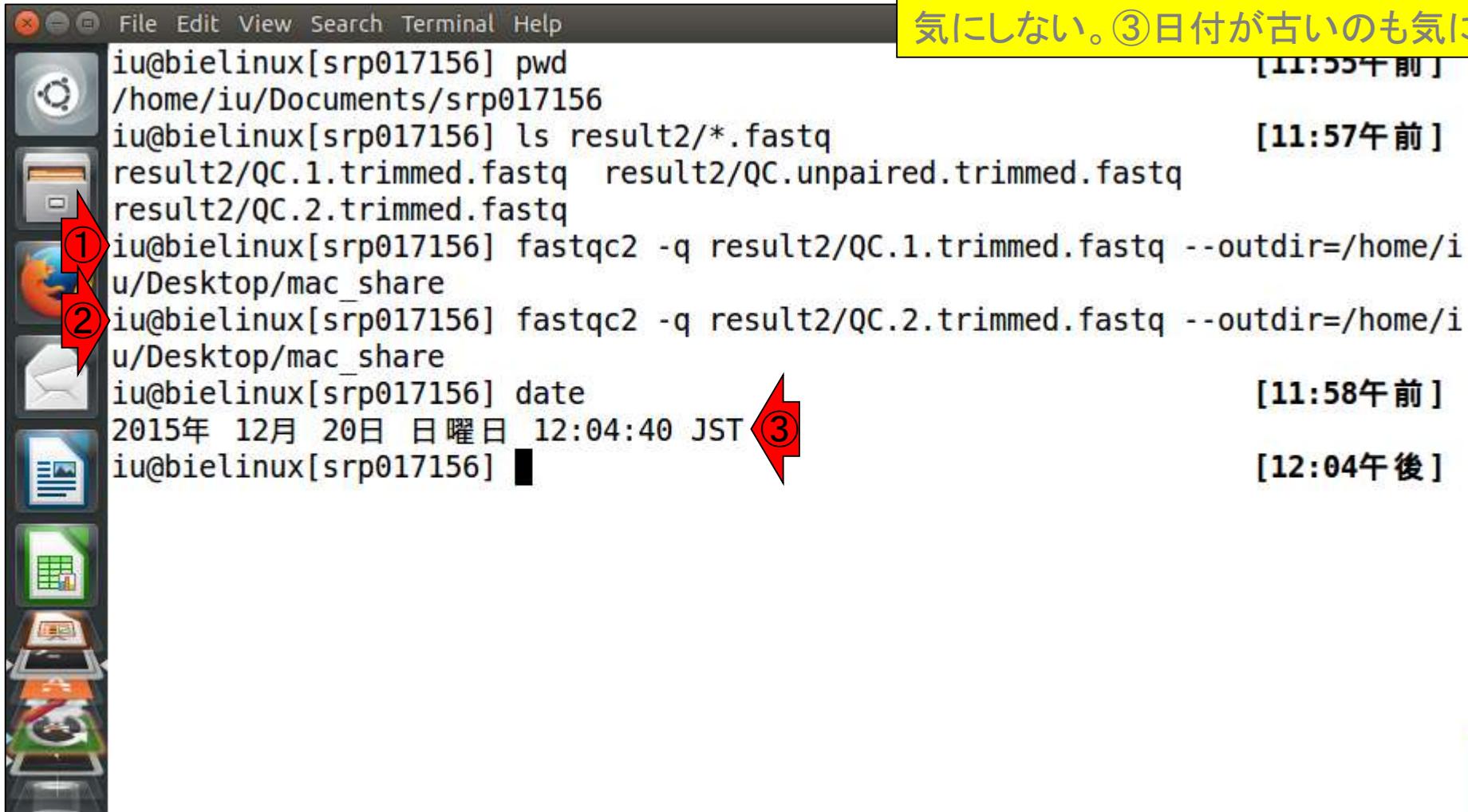
```
pwd
ls result2/*.fastq
fastqc2 -q result2/QC.1.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
fastqc2 -q result2/QC.2.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
date
ls -lh /home/iu/Desktop/mac_share/QC.*
```

出力結果を共有フォルダ(/home/iu/Desktop/mac\_share)に直接保存。実行結果ファイル [QC.1.trimmed\\_fastqc.html](#)(forward側)と [QC.2.trimmed\\_fastqc.html](#)(reverse側)とともに、Overrepresented sequences項目に見えていたアダプターやプライマー配列情報がなくなっているのがわかります。[W1-3]

- [Rockhopper 2: Tjaden B, Genome Biol., 2015](#)
- [QuasR: Gaidatzis et al., Bioinformatics, 2015](#)

# W1-2: FastQC実行

```
File Edit View Search Terminal Help  
iu@bielinux[srp017156] pwd [11:55午前]  
/home/iu/Documents/srp017156  
iu@bielinux[srp017156] ls result2/*.fastq [11:57午前]  
result2/QC.1.trimmed.fastq result2/QC.unpaired.trimmed.fastq  
result2/QC.2.trimmed.fastq  
iu@bielinux[srp017156] fastqc2 -q result2/QC.1.trimmed.fastq --outdir=/home/iu/Desktop/mac_share  
iu@bielinux[srp017156] fastqc2 -q result2/QC.2.trimmed.fastq --outdir=/home/iu/Desktop/mac_share  
iu@bielinux[srp017156] date [11:58午前]  
2015年 12月 20日 日曜日 12:04:40 JST ③ [12:04午後]  
iu@bielinux[srp017156]
```



# W1-2: FastQC実行

①dateで日付を表示。②保存先として指定した  
共有フォルダ(/home/iu/Desktop/mac\_share)  
中に、確かにFastQC実行結果ファイルが存在  
することがわかる

```
File Edit View Search Terminal Help
iu@bielinux[srp017156] pwd [11:55午前]
/home/iu/Documents/srp017156
iu@bielinux[srp017156] ls result2/*.fastq [11:57午前]
result2/QC.1.trimmed.fastq  result2/QC.unpaired.trimmed.fastq
result2/QC.2.trimmed.fastq
iu@bielinux[srp017156] fastqc2 -q result2/QC.1.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
iu@bielinux[srp017156] fastqc2 -q result2/QC.2.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
iu@bielinux[srp017156] date [11:58午前]
2015年 12月 20日 日曜日 12:04:40 JST
iu@bielinux[srp017156] ls -lh /home/iu/Desktop/mac_share/QC.* [12:04午後]
-rwxrwxrwx 1 iu iu 356K 12月 20 11:58 /home/iu/Desktop/mac_share/QC.1.trimmed_fastqc.html
-rwxrwxrwx 1 iu iu 404K 12月 20 11:58 /home/iu/Desktop/mac_share/QC.1.trimmed_fastqc.zip
-rwxrwxrwx 1 iu iu 333K 12月 20 11:58 /home/iu/Desktop/mac_share/QC.2.trimmed_fastqc.html
-rwxrwxrwx 1 iu iu 373K 12月 20 11:58 /home/iu/Desktop/mac_share/QC.2.trimmed_fastqc.zip
iu@bielinux[srp017156] [12:07午後]
```

# W1-3: FastQCで確認

①ホストOS上の共有フォルダshare上で  
、②forward側のFastQC実行結果ファイル(QC.1.trimmed\_fastqc.html)を眺める。

The screenshot illustrates the workflow for viewing a FastQC report:

- Left Panel (File Explorer):** Shows the contents of the 'share' folder on the host OS. A red arrow labeled ① points to the 'share' icon in the sidebar. Another red arrow labeled ② points to the 'QC.1.trimmed\_fastqc.html' file listed in the folder.
- Right Panel (FastQC Report):** The 'QC.1.trimmed\_fastqc.html' file is opened in a browser window. A red arrow labeled ② points to the browser's address bar, which displays the full URL: C:\Users\kadota\Desktop\share\QC.1.trimmed\_fastqc.html.
- Report Content:** The report shows a 'Summary' section with various quality metrics, many of which are marked as 'Basic Statistics' (green checkmark). It also includes a detailed 'Basic Statistics' table and a 'Per base sequence quality' chart.

Produced by [FastQC](#) (version 0.11.4)

# W1-3: FastQCで確認

FaQCs(ver. 1.34)によるQ①

- 連載第5回[W1-1]結果との比較用に、FaQCs実行前のデータでFastQC (ver. 0.11)連載第4回の[W7, W8, W9-7]あたり。

```
fastqc2 -v
fastqc2 -q SRR616268sub_1.fa
fastqc2 -q SRR616268sub_2.fa
```

連載第4回[W9-5]の手順通りに行なったです。実行結果ファイル[SRR616268su](#)結果ですが、連載第4回W8-6に示すよと3」がリード中に含まれていることが、「Illumina Single End PCR Primer 1」が

- [Skewer: Jiang et al., BMC Bioinformat](#)
- [FaQCs: Lo and Chain, BMC Bioinform](#)
- 図1。第4回の[W17-3]と基本的に同じ。

```
cd ~/Documents/srp017156
rm -f hoge*
rm -f JS*
rm -rf result*
rm -f *.bz2

pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1
ls result2
```

```
pwd
ls -lh
fastqc2 -v
FaQCs.pl -v
time FaQCs.pl -adapter -p SRR616268sub_1.fastq.gz SRR616268sub_2.fastq.gz -d result2
ls result2
```

FaQCs ver. 1.34を実行し、結果をresult2ディレクトリに保存(約25分)。result2ディレクトリ上には、計6ファイルが生成されている。そのうちのQC結果のサマリーレポートは次の2つ:[QC.stats.txt](#)と[QC qc\\_report.pdf](#)

- トリム後(FaQCs実行後)のデータを入力としてFastQC (ver. 0.11.3)をデフォルトで実行。[W1-2]

```
pwd
ls result2/*.fastq
fastqc2 -q result2/QC.1.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
fastqc2 -q result2/QC.2.trimmed.fastq --outdir=/home/iu/Desktop/mac_share
date
ls -lh /home/iu/Desktop/mac_share/QC.*
```

出力結果を共有フォルダ(/home/iu/Desktop/mac\_share)に直接保存。実行結果ファイル[QC.1.trimmed\\_fastqc.html](#)(forward側)と[QC.2.trimmed\\_fastqc.html](#)(reverse側)とともに、Overrepresented sequences項目に見えていたアダプターやプライマー配列情報がなくなっているのがわかります。[W1-3]

- [Rockhopper 2: Tjaden B, Genome Biol., 2015](#)
- [QuasR: Gaidatzis et al., Bioinformatics, 2015](#)

トランскryptームアセンブリ

# W1-3: FastQCで確認

①Overrepresented sequencesを眺めて  
Possible Sourceのところが全てNo Hitになっているかどうかをチェック

Sun 20 Dec 2015  
QC.1.trimmed.fastq

## FastQC Report

### Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)
- [Kmer Content](#)

### Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCCGGTATTTGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCA	14289	1.4622360576421252	No Hit
GGCCTATTCACTGCGGCTGACCTTGCAGTCAGCACCCCTTCTTCCGAAGT	10984	1.124025534127028	No Hit
GTCCTTTCACCTTCCCTCACGGTACTGGTCACTATCGTCACTAGGG	8746	0.8950043082187716	No Hit
CCCGGTATTTGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCAC	8446	0.8643044119844209	No Hit
GCCGGCATTCTCACTTCTAACCGCTCCAGCCGCTCACGATCGACCTTC	8081	0.8269528715659608	No Hit
GTCACTAGGGAGTATTAGCCTTGGGAGATGGCCTCCGGATTCCGACG	7943	0.8128309192981594	No Hit
GTCAGTCCTACAACCCGAGAAGCAAGCTCTCGGTTGGCTTCCC	6618	0.6772397109297771	No Hit
GTCGGTTGCGGTACGGTAGTTATTCTCACTAGAAGCTTCTTGGC	6373	0.6521681290050573	No Hit
GGTCACTTGGTTGGGTCTACATCTGTTACTCATCGCCCTGTTAGA	5461	0.558840444452631	No Hit
GCCGGCATTCTCACTTCTAACCGCTCCAGCCGCTCACGATCAACCTTC	4804	0.491607671699403	No Hit
CCCGGTATTTGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCACT	4378	0.44801381904662496	No Hit
CCCTCCATCGCTAAACAAAATAAACTAGTGAGGAATCTAACCTGCTT	4342	0.44432983149850286	No Hit
CCCGCGTCTGCCGCGGCCAGCTATGTATTCACTGACAAGCAATACACTG	4335	0.443613500586368	No Hit
CCACAGTTGGTATTATGCTTAGCCCCGGTATTTGGCGCAGTGCC	4292	0.4392131821261111	No Hit
CTGGGCTGTTCCCTTGCACAATGGACCTTATCGCTCACTGTCTGACTC	4087	0.41823491969930476	No Hit
CCGGCATTCTCACTTCTAACCGCTCCAGCCGCTCACGATCGACCTTC	3839	0.39285633881224147	No Hit

# W1-3: FastQCで確認

## FastQC Report

### Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)
- [Kmer Content](#)

Sequence	Length	Quality Score	Status
CGCTACTCATGCCGGCATTCTCACTTCAA			
CCGGGGTGCTTTCACCTTCCTCACGGTACTGGTTCACTATCGGTAC	1349	0.13804720006713045	No Hit
GGGCCTATTCACTGCGGCTGACCTTGCGGTAGCACCCCTTCTCCGAAG	1341	0.13722853616754777	No Hit
CTGGTGATCTGGGCTGTTCCCCTTCGACAATGGACCTTATCGCTCACTG	1336	0.13671687123030857	No Hit
GCCGCCGGCCAGCTATGTATTCACTGACAAGCAATAACACTGATGTGACT	1319	0.13497721044369537	No Hit
<b>1) GCCGAGTTCTTAACGAGAGTTCGCTCGCTACCTGAGGATACTCTCCTC</b>	1305	0.13354454861942566	No Hit
<b>GTTTGGGCTTCCCACCTCGCTCGCGCTACTATGGGAATCGAGTTTC</b>	1262	0.12914423015916873	No Hit
CGTCCCTCCATCGTTAACAAAAATAAAACTAGTGCAGGAATCTCAACCTG	1258	0.1287348982093774	No Hit
CACACGGTTTCAGGAACTGTTCACTCCCTTGGGGTGCTTTCACCT	1257	0.12863256522192956	No Hit
CCCTAGTTCAAACAGTGCTCTACCTCCATGATCCATCCTCCGAGGCTAAC	1249	0.12781390132234688	No Hit
GTCATTTGCCAGTTCTTAACGAGAGTTCGCTCGCTACCTGAGGATA	1249	0.12781390132234688	No Hit
ACCTTAACTGGTACTGGGCTGTTCCCCTTCGACAATGGACCTTATCG	1223	0.12515324364870312	No Hit
CACAGTTCCGTATTATGCTTAGCCCCGGTATATTCGGCGCAGTGCCA	1220	0.12484624468635963	No Hit
GGTCATTTGCCAGTTCTTAACGAGAGTTCGCTCGCTACCTGAGGAT	1208	0.1236182488369856	No Hit
CCACTTAGCATAAATTAGGGACCTTAACGGTACTGGTACTGGCTGTTCCC	1207	0.12351591584953776	No Hit
CCGGTTCATTCTACAAAAGGCACGCCATTACCGTTAACGGGCTTGACT	1191	0.1218785880503724	No Hit
CTGCCGCCGCCAGCTATGTATTCACTGACAAGCAATAACACTGATGTGA	1178	0.12054825921355053	No Hit
CGCTACTCACCACCTCCACCCACCCCTCCACCTTCCACACCCCT	1174	0.12012002720275017	No Hit

①連載第4回W17-2で見られたアダプター配列がなくなっていることがわかる。これを含め、全てNo Hitになっていたことから、-adapterオプションがうまく機能していることがわかる。同じ手順で、reverse側のFastQC実行結果ファイル(QC.2.trimmed\_fastqc.html)も眺めておこう

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# 第5回原稿PDF

として、FastQCによるクオリティチェックを行えばよい [W1-2]。著者らは、FastQC実行結果ファイルの項目(Overrepresented sequences)を眺めて、トリム前に見えていた既知のアダプターやプライマー配列が、トリム後に正しく見えなくなっていることを確認して安心している[W1-3]。

このデータに関して結論からいえば、forward側の107 bpのリードファイル(SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq)のうち、100-107塩基付近に乳酸菌由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけQC段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper2<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そしてQC再実行である。

## トランスクリプトームアセンブリ

ゲノムのアセンブリは、断片化されたゲノム配列由来リードをつなぎ合わせて、元のゲノム配列を再構築する作業である。この再構築に相当する英語がアセンブリ(assembly)であり、再構築を行うプログラムをアセンブラー(assembler)という。デノボ(*de novo*)という言葉が同時に用いられることが多いが、これは「最初から」と

ここまでで、第5回原稿PDFの①p195の②左上あたりまでを述べた。FaQCs実行によって、FastQCのOverrepresented sequences項目で見えていたIllumina adapterが消えたので第一段階クリアということで安心している状態

か「一から」という意味  
入力として(つまり他の  
する際には、*de novo*)

トランスクリプトームアセンブリとは、アセンブル対象がゲノムではなく解析サンプル中で発現している全転写物(トランスクリプトーム)の場合を指す。RNA-seqデータのみを入力として一からアセンブルする場合は、*de novo transcriptome assembly*などと呼ばれる。

Multiple-k<sup>20)</sup>やTrans-ABySS<sup>21)</sup>などの初期のトランスクリプトーム用アセンブラーは、ゲノム用を内部的に用いていた。詳細は省くが、上述のk-merのkの値(正の整数)を大きくすればするほど、得られるコンティグは長くなり、高発現のものに偏る傾向にある<sup>22)</sup>。kの値は、アセンブル時の「のりしろ」に相当するものである。パリンドロームを避けるべく、通常は奇数が採用される<sup>23)</sup>。kの値を小さくすればするほど、低発現転写物を拾いあげることが原理的には可能であるが、得られるコンティグは短くなり(断片化)、似た配列からなるコンティグが多く得られる傾向(重複)にある。このためこれらのプログラムは、複数のkの値を用いて独立にゲノム用アセンブラーを適用し、できるだけ多くの転写物配列をコンティグとして得ることに主眼を置いていた。それゆえ、コンティグ集合からいかに重複を取り除くかが課題であった。

おそらく現在もっとも頻用されているトランスクリプトーム用アセンブラーは、Trinity<sup>24)</sup>である。Trinityは、トランスクリプトーム専用としてデザインされた最初のブ

# 第5回原稿PDF

として、FastQCによるクオリティチェックを行えばよい [W1-2]。著者らは、FastQC実行結果ファイルの項目(Overrepresented sequences)を眺めて、トリム前に見えていた既知のアダプターやプライマー配列が、トリム後に正しく見えなくなっていることを確認して安心している[W1-3]。

このデータに関して結論からいえば、forward側の107 bpのリードファイル(SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq)のうち、100-107塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけQC段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そしてQC再実行である。

## トランスクリプトームアセンブリ

ゲノムのアセンブリは、断片化されたゲノム配列由来リードをつなぎ合わせて、元のゲノム配列を再構築する作業である。この再構築に相当する英語がアセンブリ(assembly)であり、再構築を行うプログラムをアセンブラー(assembler)という。デノボ(de novo)という言葉が同時に用いられることが多いが、これは「最初から」と

か「一から」とい入力として(つま  
ルする際には、de  
トランスクリプト  
ゲノムではなく解析サンプル中で発現している全転写物  
(トランスクリプトーム)の場合を指す。RNA-seqデータ  
のみを入力として一からアセンブルする場合は、*de novo*  
transcriptome assembly. ハビレ呼ばれフ

赤枠の最初のほうでオチを先に述べているが、当時は、この乳酸菌RNA-seqデータの次の解析手段として、バクテリア用*de novo*トランスクリプトームアセンブラーである①Rockhopper2の論文を発見した。バクテリア用だから、有名なTrinityなどよりもさぞかしよりよい結果が得られるだろうと思っていた

このデータに関して結論からいえば、forward側の107 bpのリードファイル(SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq)のうち、100-107塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけQC段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そしてQC再実行である。



# W2-1 : Rockhopper

トランскриプトームアセンブリ

①

- Multiple-k: Surget-Groba and Montoya-Burgos, Genome Res., 2009
- Trans-ABYSS: Robertson et al., Nat Methods, 2010
- Gibbons et al., Mol Biol Evol., 2009
- Miller et al., Genomics, 2010
- Trinity: Grabherr et al., Nat Biotechnol, 2011
- Bridger: Chang et al., Genome Biol., 2015
- CD-HIT: Fu et al., Bioinformatics, 2012
- IFRAT: Mbandi et al., BMC Bioinformatics, 2015
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- Rockhopper(バクテリア用): McClure et al., Nucleic Acids Res., 2013
- TruHMM(バクテリア用): Li et al., BMC Genomics, 2013
- Rockhopper 2(バクテリア用): Tjaden B, Genome Biol., 2015

②

③

De novo transcriptome assembly (Rockhopper2 ver. 2.0.3)

③

- Rockhopper2(バクテリア用): Tjaden B, Genome Biol., 2015
- Java確認など[W2-2]

①このあたり。トランскриプトーム用 *de novo* アセンブラーの②Trinityと③Rockhopper2。ここでは、トランскриプトームアセンブリというよりも、Javaプログラムの1つとしてのRockhopper2をどうやって実行させるかを紹介するという位置づけ

# W2-1 : Rockhopper



# ROCKHOPPER

## A system for analyzing bacterial RNA-seq data

HOME

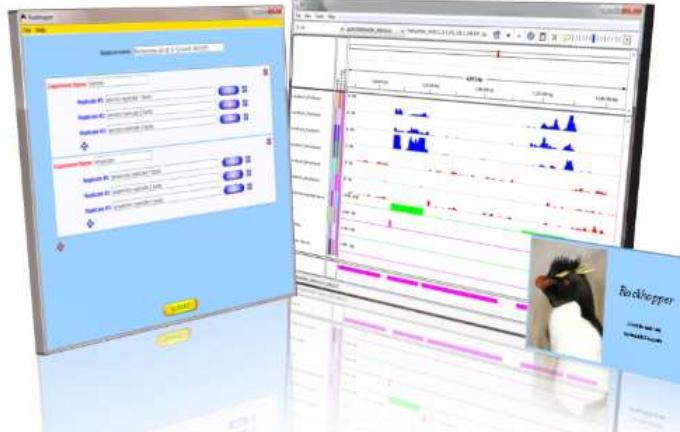
Download

User Guide

FAQ

①

- Rockhopper is a comprehensive and user-friendly system for computational analysis. As input, Rockhopper takes RNA sequencing reads output by high-throughput seqQSEQ, FASTA, SAM, or BAM files). Rockhopper supports the following tasks:
- Reference based transcript assembly (when one or more reference genomes
    - Aligning reads to genomes
    - Assembling transcripts
    - Identifying transcript boundaries and novel transcripts such as small RNAs
  - *De novo* transcript assembly (when reference genomes are unavailable)
  - Normalizing data from different experiments
  - Quantifying transcript abundance
  - Testing for differential gene expression
  - Characterizing operon structures
  - Visualizing results in a genome browser



## System Requirements

Rockhopper is implemented in Java, so you must have Java installed on your computer to use Rockhopper. To check if Java is installed on your computer, type "java -version" at any shell or terminal window or command prompt:

```
C:\>java -version
java version "1.7.0_11"
Java(TM) SE Runtime Environment (build 1.7.0_11-b21)
Java HotSpot(TM) 64-Bit Server VM (build 23.6-b04, mixed mode)
```

It is recommended that your computer has Java version 1.6 or later and your computer has at least 2 gigabytes of RAM. If you do not have Java or you need to update to a more recent version, you can do so by clicking the Java icon on the right:



## Download Latest Release (Rockhopper version 2.0.3)

Rockhopper  
for Windows



Rockhopper  
for Mac



Rockhopper for any  
platform



Rockhopper source code



Error opening  
Rockhopper on  
Mac?

To execute the GUI  
version of Rockhopper,  
use the following  
command:  
`java -Xmx1200m -jar  
Rockhopper.jar`

To execute the command  
line version of Rockhopper,  
use the following  
command:  
`java -Xmx1200m -cp  
Rockhopper.jar  
Rockhopper`

To extract the source code from the  
JAR file, use the following command:  
`jar xf Rockhopper.jar`

To extract the source code from the  
compressed TAR archive, use the  
following command:  
`tar xjf  
Rockhopper-2.0.3.tar.bz2`

①Download。Bio-Linux8の場合は、②「Rockhopper for any platform」でよい。  
。実際にはダウンロード済みなのでエーアハズオン(やったつもり)

# W2-1 : Rockhopper



# ROCKHOPPER

## A system for analyzing bacterial RNA-seq data

HOME

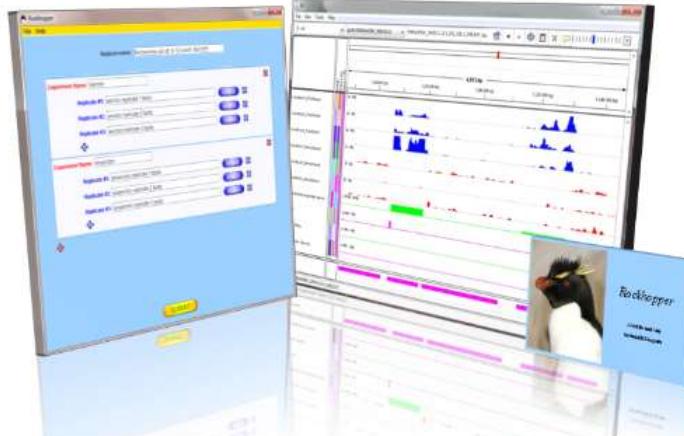
Download

User Guide

FAQ

①

- Rockhopper is a comprehensive and user-friendly system for computational analysis. As input, Rockhopper takes RNA sequencing reads output by high-throughput seqQSEQ, FASTA, SAM, or BAM files. Rockhopper supports the following tasks:
- Reference based transcript assembly (when one or more reference genomes)
    - Aligning reads to genomes
    - Assembling transcripts
    - Identifying transcript boundaries and novel transcripts such as small RNAs
  - *De novo* transcript assembly (when reference genomes are unavailable)
  - Normalizing data from different experiments
  - Quantifying transcript abundance
  - Testing for differential gene expression
  - Characterizing operon structures
  - Visualizing results in a genome browser



## System Requirements

Rockhopper is implemented in Java, so you must have Java installed on your computer to use Rockhopper. To check if Java is installed on your computer, type "java -version" at any shell or terminal window or command prompt:

```
C:\>java -version
java version "1.7.0_11"
Java(TM) SE Runtime Environment (build 1.7.0_11-b21)
Java HotSpot(TM) 64-Bit Server VM (build 23.6-b04, mixed mode)
```

It is recommended that your computer has Java version 1.6 or later and your computer has at least 2 gigabytes of RAM. If you do not have Java or you need to update to a more recent version, you can do so by clicking the Java icon on the right:



## Download Latest Release (Rockhopper version 2.0.3)

Rockhopper for Windows



Rockhopper for Mac



Rockhopper for any platform



To execute the line version of Rockhopper, use the following command:  
java -Xmx12G Rockhopper

Error opening Rockhopper on Mac?

Rockhopper source code



開く(O)  
新しいタブで開く(W)  
新しいウィンドウで開く(N)  
対象をファイルに保存(A)...  
対象を印刷(P)

②

切り取り  
コピー(C)  
ショートカットのコピー(T)

③

from the command:  
java -jar

from the use the  
command:  
java -jar

貼り付け(P)

Bing で翻訳  
電子メール (Windows Live Hotmail)  
すべてのアクセラレータ

基本的には右クリックで②「対象をファイルに保存」でよいが、wgetコマンドを利用したい場合は③「ショートカットのコピー(Windowsの場合)」でURL情報を取得する

# W2-1 : Rockhoppe

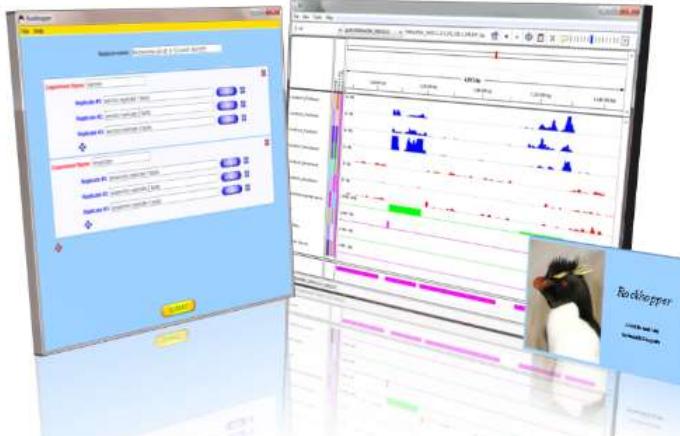


# ROCKHOPPER

## A system for analyzing bacterial RNA-seq data

Rockhopper is a comprehensive and user-friendly system for computational analysis. As input, Rockhopper takes RNA sequencing reads output by high-throughput seqQSEQ, FASTA, SAM, or BAM files). Rockhopper supports the following tasks:

- Reference based transcript assembly (when one or more reference genomes
  - Aligning reads to genomes
  - Assembling transcripts
  - Identifying transcript boundaries and novel transcripts such as small RNAs
- *De novo* transcript assembly (when reference genomes are unavailable)
- Normalizing data from different experiments
- Quantifying transcript abundance
- Testing for differential gene expression
- Characterizing operon structures
- Visualizing results in a genome browser



①System Requirements。これはFaQCsのときのPrerequisite(前もって必要な事柄)と似たようなもの。Rockhopper2はJavaプログラムなので、Javaがver. 1.6以上かどうかを②「java -version」で確認(次スライド)

### System Requirements

Rockhopper is implemented in Java, so you must have Java installed on your computer to use Rockhopper. To check if Java is installed on your computer, type "java -version" at any shell or terminal window or command prompt:

```
C:\>java -version  
java version "1.7.0_11"  
Java(TM) SE Runtime Environment (build 1.7.0_11-b21)  
Java HotSpot(TM) 64-Bit Server VM (build 23.6-b04, mixed mode)
```

It is recommended that your computer has Java version 1.6 or later and your computer has at least 2 gigabytes of RAM. If you do not have Java or you need to update to a more recent version, you can do so by clicking the Java icon on the right:



### Download Latest Release (Rockhopper version 2.0.3)

Rockhopper for Windows



Rockhopper for Mac



Rockhopper for any platform



To execute the GUI version of Rockhopper, use the following command:  
`java -Xmx1200m -jar Rockhopper.jar`

Error opening Rockhopper on Mac?

To execute the command line version of Rockhopper, use the following command:  
`java -Xmx1200m -cp Rockhopper.jar Rockhopper`

Rockhopper source code



To extract the source code from the JAR file, use the following command:  
`jar xf Rockhopper.jar`



To extract the source code from the compressed TAR archive, use the following command:  
`tar xjf Rockhopper-2.0.3.tar.bz2`

## W2-2: Java確認など

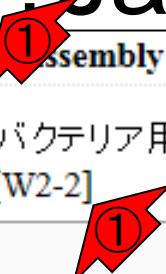
①Javaのバージョンを確認。  
②赤枠部分が次のスライド

### De novo transcriptome assembly (Rockhopper2 ver. 2.0.3)

- Rockhopper2(バクテリア用): [Tjaden B, Genome Biol., 2015](#)
- Java確認など[W2-2]

```
cd
```

```
pwd  
java -version  
cd ~/Downloads  
pwd  
ls
```



①



②

- wgetで取得[W2-3]

```
wget -c http://cs.wellesley.edu/~btjaden/Rockhopper/download/current/Rockhopper.jar  
ls  
ls -l Rockhopper.jar
```

- GUI版を実行[W2-4]

```
ls -l Rockhopper.jar  
java -Xmx1200m -jar Rockhopper.jar
```

## W2-2: Java確認など

```
iu@bielinux[~/Downloads]
iu@bielinux[iu] pwd
/home/iu
iu@bielinux[iu] java -version
java version "1.7.0_79"
OpenJDK Runtime Environment (IcedTea 2.5.5) (7u79-2.5.5-0ubuntu0.14.04.2)
OpenJDK 64-Bit Server VM (build 24.79-b02, mixed mode)
iu@bielinux[iu] cd ~/Downloads
iu@bielinux[Downloads] pwd
/home/iu/Downloads
iu@bielinux[Downloads] ls
boost_1_61_0.tar.bz2      kmergenie-1.6982.tar.gz
Bridger_r2014-12-01.tar.gz master.zip
FAQCs
FastQC
fastqc_v0.11.4.zip
IGV_2.3.67
IGV_2.3.67.zip
iu@bielinux[Downloads]
```

[ 2:07 午後 ]  
[ 2:07 午後 ]  
[ 2:07 午後 ]  
① 「java -version」実行結果(作業ディレクトリはどこでもよい)。このPCには、ver. 1.7.0\_79がインストールされていることがわかる。  
② ~/Downloadsに移動。  
③ ls 実行結果で見えるものはヒトによって異なるが、基本気にしなくてもよい。  
④ 講習会では、ダウンロード済みなのでRockhopper.jarが見えている

②  
③  
④

# W2-3: wgetで取得

やらずに見るだけ! ①wget実行したときのイメージ。  
赤下線部のURL情報の最後がダウンロードしたいファイル名に相当する。②約13MB (14,039,789 bytes)

The screenshot shows a terminal window with a red arrow pointing to the number 1. The terminal output is as follows:

```
iu@bielinux[Downloads] wget -c http://cs.wellesley.edu/~btjaden/Rockhopper/download/current/Rockhopper.jar
--2015-12-20 14:23:49-- http://cs.wellesley.edu/~btjaden/Rockhopper/download/current/Rockhopper.jar
Resolving cs.wellesley.edu (cs.wellesley.edu)... 149.130.136.40
Connecting to cs.wellesley.edu (cs.wellesley.edu)|149.130.136.40|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 14039789 (13M) [application/x-java-archive]
Saving to: 'Rockhopper.jar'

100%[=====] 14,039,789 1.72MB/s in 9.6s

2015-12-20 14:23:59 (1.39 MB/s) - 'Rockhopper.jar' saved [14039789/14039789]
```

After the download, the user runs an ls command to list the contents of the Downloads directory. A red arrow points to the number 2, highlighting the file "Rockhopper.jar". The terminal output is:

```
iu@bielinux[Downloads] ls [ 2:23午後]
FaQCs  fastqc_v0.11.4.zip  IGV_2.3.67.zip
FastQC  IGV_2.3.67          Rockhopper.jar
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 2:23午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 2015 Rockhopper.jar
iu@bielinux[Downloads] [ 2:24午後]
```

# W2-4: GUI版を実行

- wgetで取得[W2-3]

```
wget -c http://cs.wellesley.edu/~btjaden/Rockhopper/download/current/Rockhopper.jar  
ls  
ls -l Rockhopper.jar
```

- GUI版を実行[W2-4]

```
ls -l Rockhopper.jar  
java -Xmx1200m -jar Rockhopper.jar
```

- Tips: バックグラウンドジョブ(background job)[W3-1]

```
java -Xmx1200m -jar Rockhopper.jar&
```

- Tips: psコマンド[W3-3からW3-5]

```
ps -f
```

```
ps -f
```

①W2-4。Rockhopper2のGUI版を実行。②赤枠部分が次のスライド。以後はこのようなウェブページのスクリーンショットを基本的に示さないので、自分で辿っていってください

# W2-4: GUI版を実行

iu@bielinux[~/Downloads]

```
iu@bielinux[Downloads] ls -l Rockhopper.jar
```

```
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
```

```
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
```

①RockhopperのGUI版を実行したい場合のやり方が書いてあるので、②その通りに実行。使ったことのないプログラムは、まずは例題の実行が基本。③リターンキーを押すと…

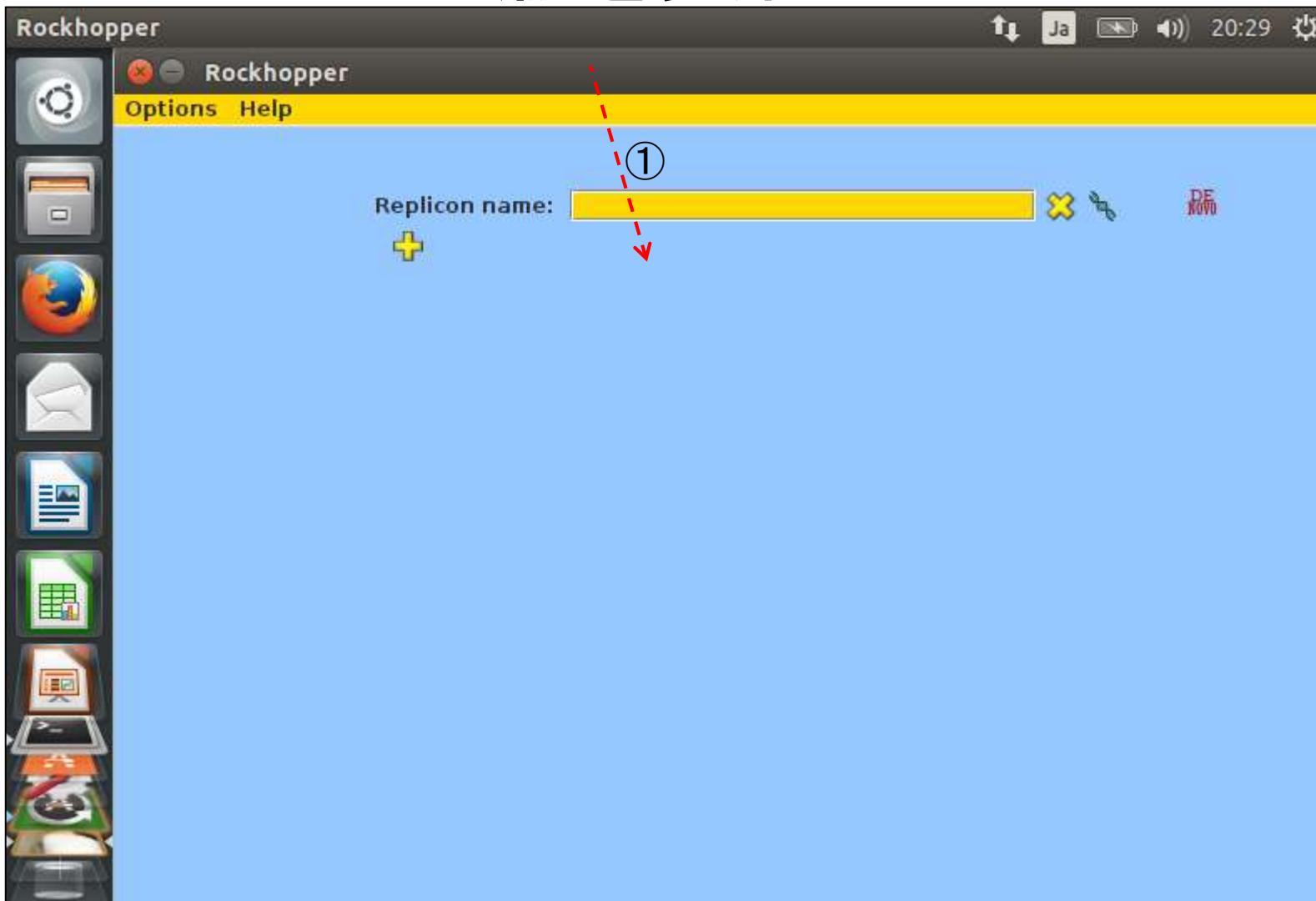


Download Latest Release (Rockhopper version 2.0.3)

Rockhopper for Windows	Rockhopper for Mac	Rockhopper for any platform	Rockhopper source code
		To execute the GUI version of Rockhopper, use the following command: <code>java -Xmx1200m -jar Rockhopper.jar</code>	To extract the source code from the JAR file, use the following command: <code>jar xf Rockhopper.jar</code>
Error opening Rockhopper on Mac?		To execute the command line version of Rockhopper, use the following command: <code>java -Xmx1200m -cp Rockhopper.jar Rockhopper</code>	To extract the source code from the compressed TAR archive, use the following command: <code>tar xjf Rockhopper-2.0.3.tar.bz2</code>

RockhopperのGUI版が起動する。①GUIを赤矢印の始点から終点に移動させると…

# W2-4 : GUI版を実行

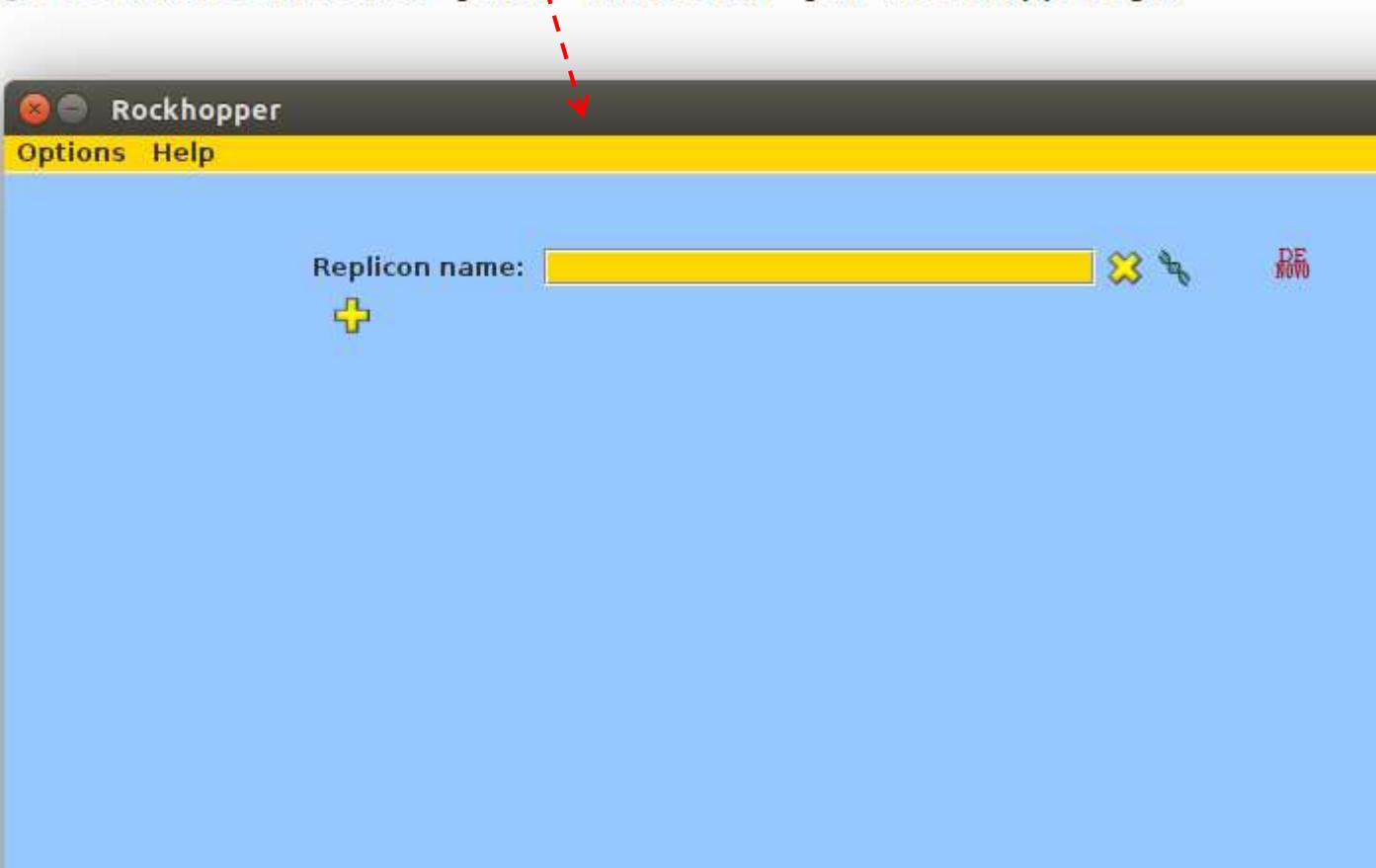


# W2-4 : GUI版を実行



```
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
```

RockhopperのGUI版を起動したコマンドが見える。RockhopperのGUI起動中は、①のようにコマンドプロンプトが出ないため、このターミナル上では新たに何もコマンドを打つことができない



# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W3-1 : background job



```
iu@bielinux[Downloads] ls -l Rockhopper.jar  
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Roc  
iu@bielinux[Downloads] java -Xmx1200m -jar Ro
```

何もコマンドを打つことができない状態(W2-4)を回避する1つのやり方が「バックグラウンドジョブ」。この場合は、RockhopperのGUIを裏(background)で実行させる(jobを流す)ことで、コマンドを実行したターミナル上で、次のコマンドを自在に打てるようになります。① RockhopperのGUIを一旦終了させると…

# W3-1 : background job

```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar [ 8:57 午後]
iu@bielinux[Downloads]
```

The screenshot shows a terminal window with the command history and output. A red arrow points to the icon for the application menu in the dock, which is circled with a red circle labeled '①'. The application menu is open, showing various icons for applications like a terminal, file manager, browser, and others.

①通常のコマンド打ち込み可能状態となる。バックグラウンドジョブとは、RockhopperのGUIを起動しつつも、このような状態にするテクニックです

基本的には、①通常のコマンドの最後に「&」をつけるだけ

# W3-1 : background job

The screenshot shows a Linux desktop environment. On the left, there is a vertical docked application menu with various icons. One icon, which appears to be a terminal or command-line interface, has a red circle with the number '1' and a red arrow pointing to it, indicating it is the active application. To the right of the dock is a terminal window titled 'iu@bielinux[~/Downloads]'. The terminal displays the following commands:

```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24 午後 ]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar&
```

# W3-2: background job

Rockhopper

```
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276 [ 9:04 午後 ]
iu@bielinux[Downloads]
```

①リターンキーを押した結果。確かにRockhopper GUIが起動しつつ、②コマンド打ち込み可能状態になっていることがわかる。赤枠内の数値は、ヒトによって異なる

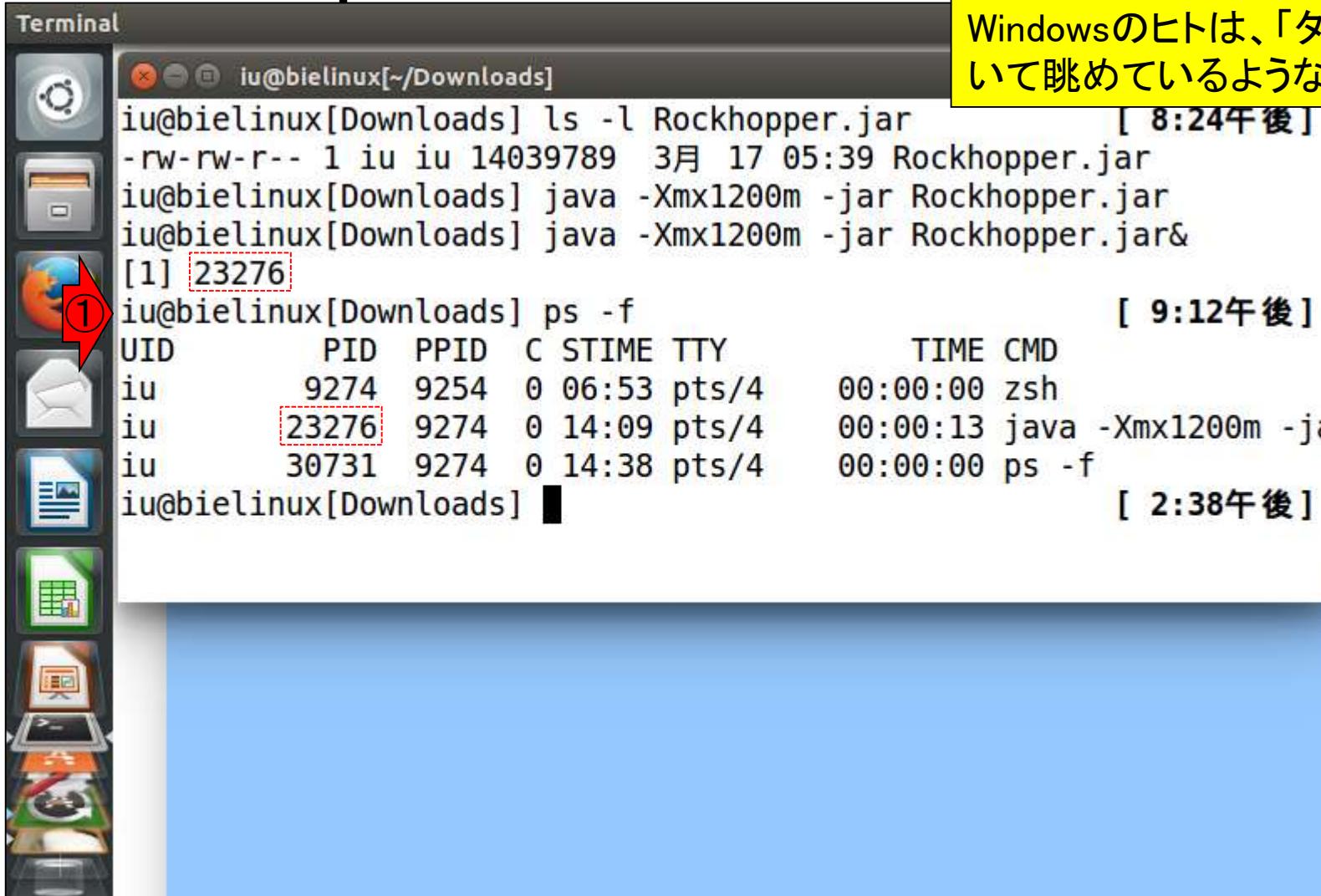
# W3-2:background job

The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "Terminal" and the command line shows the user "iu" at "bielinux" in the directory "~/Downloads". The user runs "ls -l Rockhopper.jar" which lists a file "Rockhopper.jar" with permissions "-rw-rw-r--" and size 14039789. The user then runs "java -Xmx1200m -jar Rockhopper.jar" and immediately follows it with "java -Xmx1200m -jar Rockhopper.jar&". The terminal window has a dark theme and includes a vertical scroll bar. The desktop background is blue. A vertical dock on the left contains icons for various applications like the Dash, Home, and several system monitors.

```
iu@bielinux[~/Downloads]
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 8:24 午後 ]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] [ 9:12 午後 ]
```

# W3-3:psコマンド

Terminal



iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24 午後]  
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar  
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar  
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&  
[1] 23276  
iu@bielinux[Downloads] ps -f [ 9:12 午後]  
UID PID PPID C STIME TTY TIME CMD  
iu 9274 9254 0 06:53 pts/4 00:00:00 zsh  
iu 23276 9274 0 14:09 pts/4 00:00:13 java -Xmx1200m -ja  
iu 30731 9274 0 14:38 pts/4 00:00:00 ps -f  
iu@bielinux[Downloads]

①psコマンドで実行中のプロセスを表示。プロセスと表現する場合が多いのでそう書いているが、jobやタスクという理解でもよい。Windowsのヒトは、「タスクマネージャー」を開いて眺めているようなものだと思えばよい

# W3-3:psコマンド

①CMD列が、現在実行中のコマンド  
。②zshのみ打ち込んだ記憶がないだ  
ろうが、これは③のターミナルボタン  
を押して起動中のターミナルそのもの

Terminal

```
iu@bielinux[~/Downloads]
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f
UID      PID  PPID  C STIME TTY          TIME CMD
iu        9274  9254  0 06:53 pts/4    00:00:00 zsh
iu       23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu       30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads]
```

① [ 9:12 午後 ]  
② [ 2:38 午後 ]

③

# W3-4: psコマンド

①ターミナルボタンを押して起動したターミナルのプロセスID (PID)は9274

Terminal

```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[~/Downloads] ps -f [ 9:12 午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[~/Downloads] [ 2:38 午後]
```

# W3-4: psコマンド

①ターミナルボタンを押して起動したターミナルのプロセスID (PID)は9274。②で打ち込んだコマンドのPIDは23276

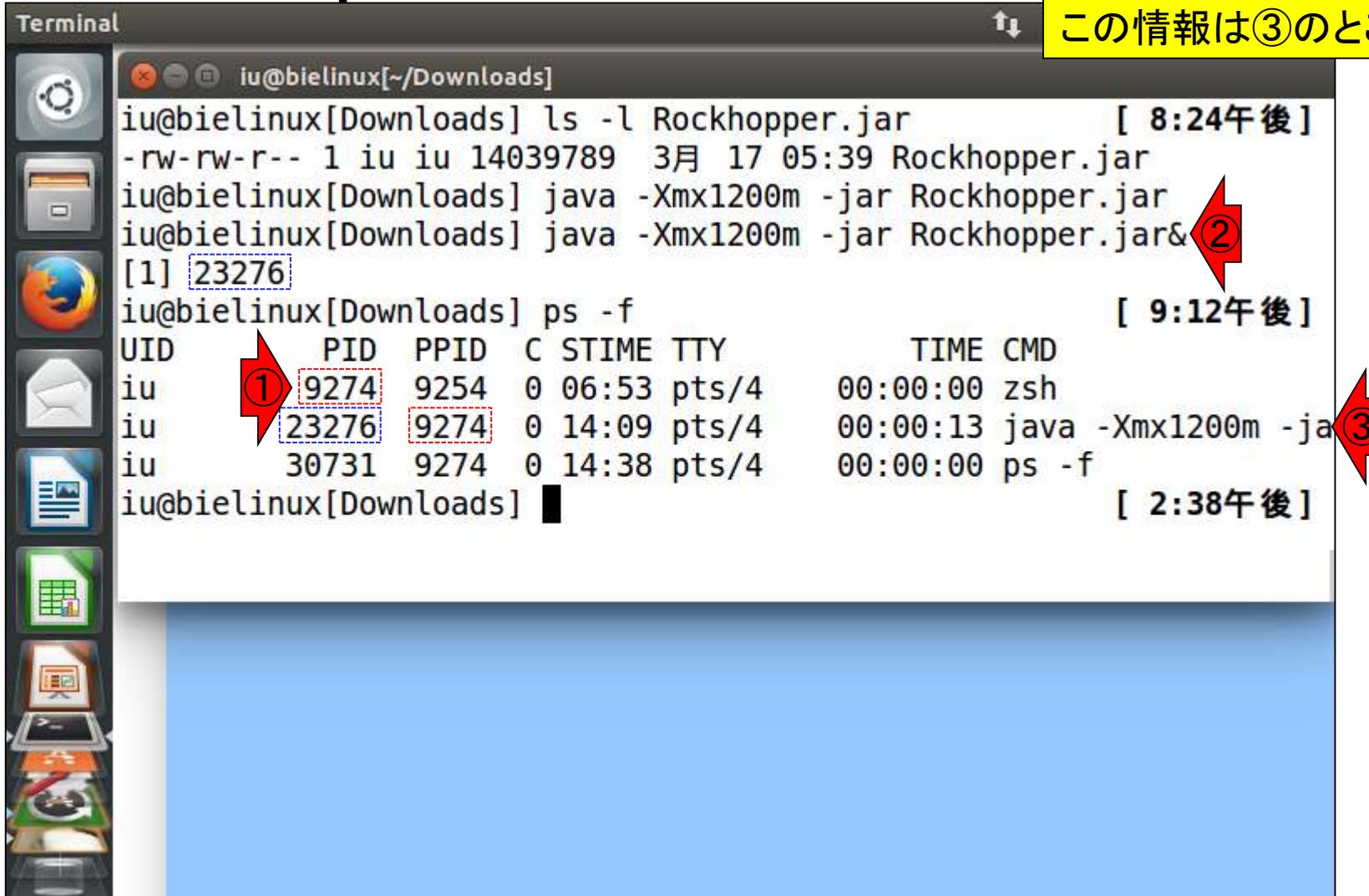
```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f [ 9:12午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads]
```

# W3-4 : psコマンド

①ターミナルボタンを押して起動したターミナルのプロセスID (PID)は9274。②で打ち込んだコマンドのPIDは23276。この情報は③のところに相当

Terminal

```
iu@bielinux[~/Downloads]
iu@bielinux[Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f [ 9:12 午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads]
```



# W3-4: psコマンド



```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f
[ 9:12 午後 ]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads]
```

①ターミナルボタンを押して起動したターミナルのプロセスID (PID)は9274。②で打ち込んだコマンドのPIDは23276。この情報は③のところに相当。全体像から④のPPIDがPIDの親プロセスIDであることがわかる。Parent PIDと解釈

[ 9:12 午後 ]

[ 2:38 午後 ]

①

④

③

②

# W3-5: psコマンド

①「ps -f」のプロセスID (PID)は30731。この親プロセスID (PPID)が9274なのは妥当。理由は、このPID9274のターミナル上で実行したコマンドだから

Terminal

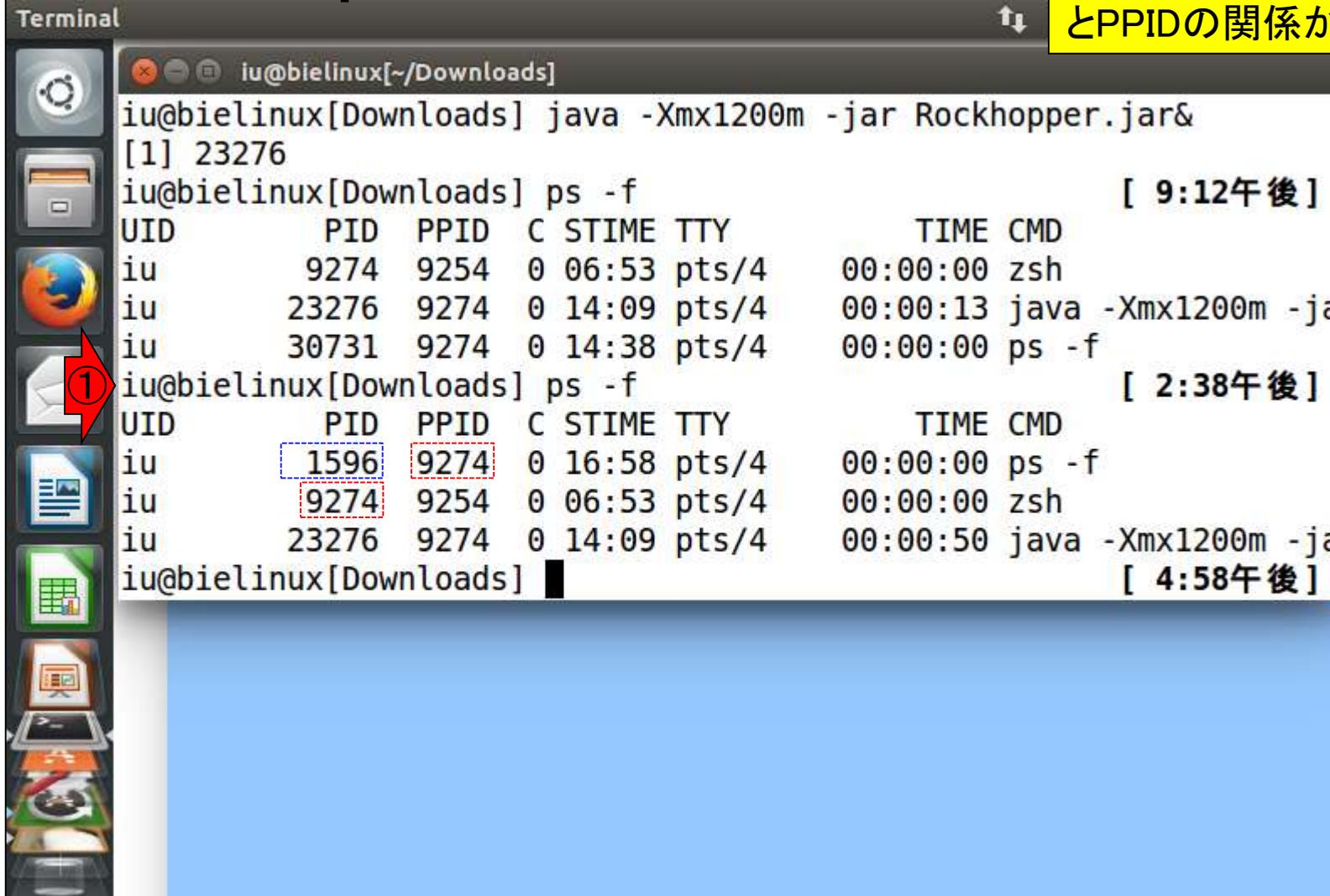
```
iu@bielinux[~/Downloads] ls -l Rockhopper.jar [ 8:24 午後]
-rw-rw-r-- 1 iu iu 14039789 3月 17 05:39 Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f [ 9:12 午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f ① 2:38 午後
iu@bielinux[Downloads]
```

# W3-5: psコマンド

①もう一度「ps -f」を実行。このプロセスID (PID)は1596。このように数値はコロコロ変わるものなので、基本的にPIDとPPIDの関係がわかつていればよい

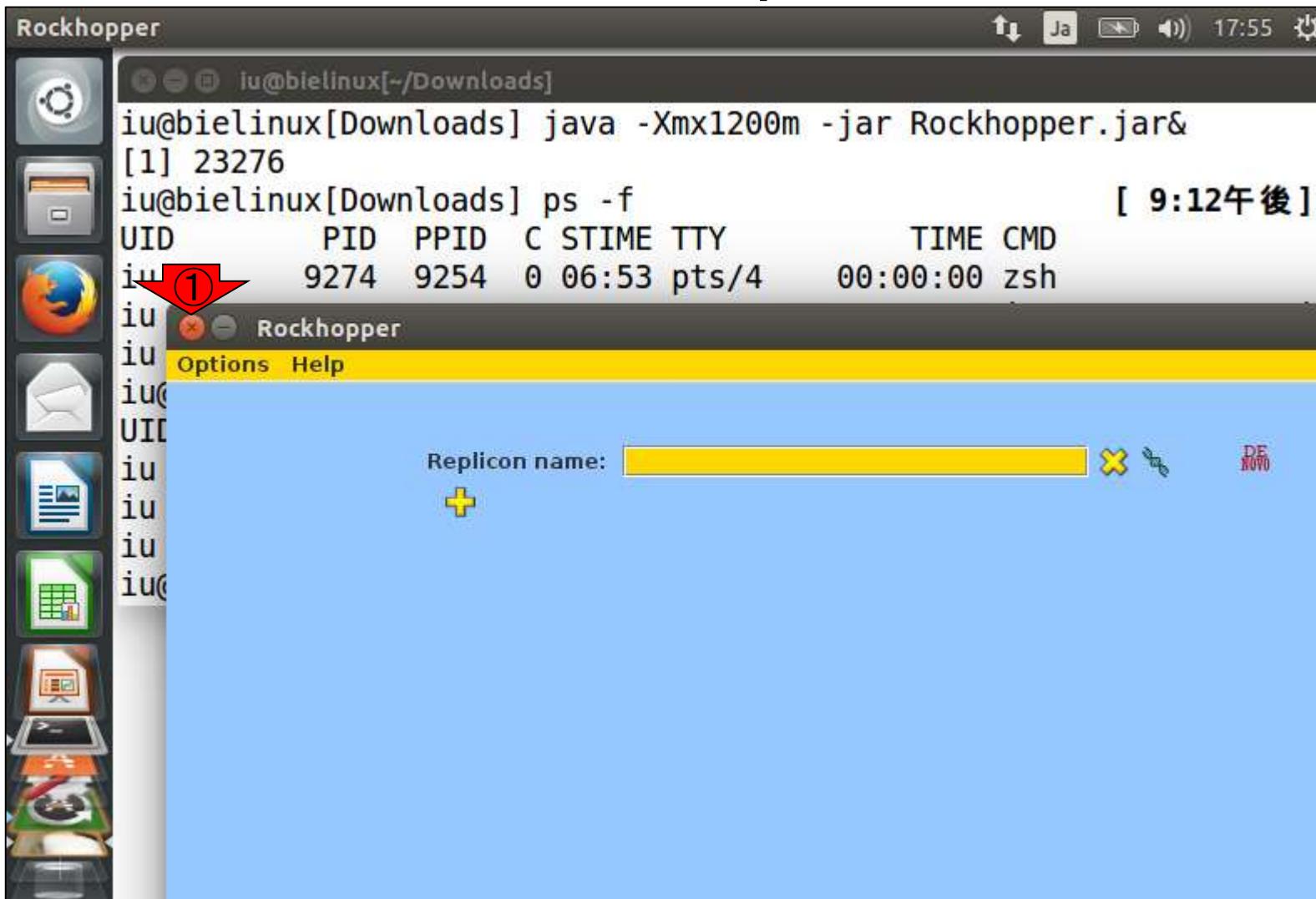
Terminal

```
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f [ 9:12午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] ps -f [ 2:38午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      1596  9274  0 16:58 pts/4    00:00:00 ps -f
iu     9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:50 java -Xmx1200m -ja
iu@bielinux[Downloads] [ 4:58午後]
```



①GUIベースでやる場合は、×ボタンだが、ここでは押さないで!

# W3-6: プロセスの終了



# W3-6: killコマンド

①「kill プロセスID」で終了させることができる。やたらとメモリを消費している意味不明なプロセスが実行されている場合に、このような処理を行って終了させる。各自のターミナル上で見えているPIDを与えてkillしよう

Terminal

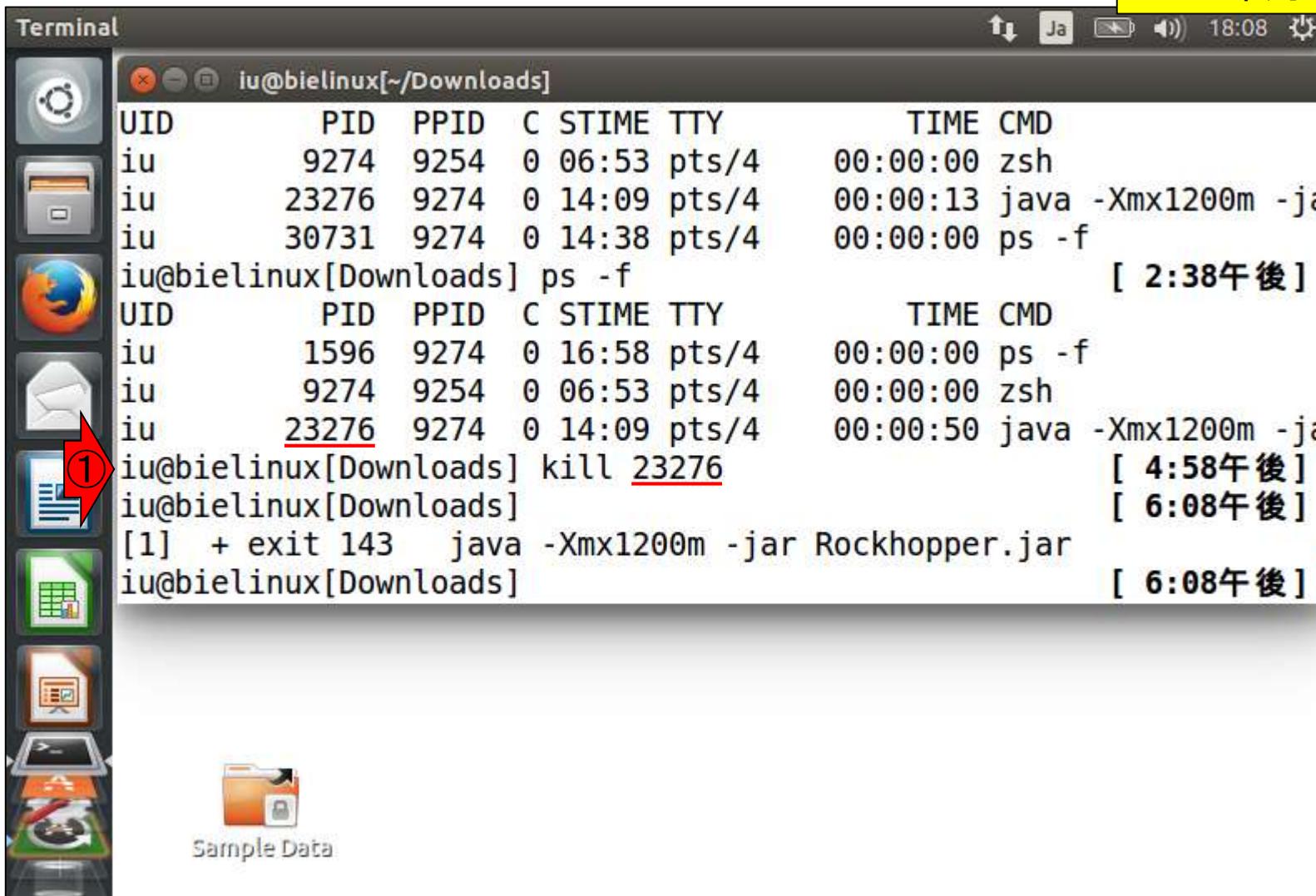
```
iu@bielinux[~/Downloads] java -Xmx1200m -jar Rockhopper.jar&
[1] 23276
iu@bielinux[Downloads] ps -f [ 9:12午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274  0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] ps -f [ 2:38午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu      1596  9274  0 16:58 pts/4    00:00:00 ps -f
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     23276  9274  0 14:09 pts/4    00:00:50 java -Xmx1200m -ja
iu@bielinux[Downloads] kill 23276 [ 4:58午後]
```



## W3-6: killコマンド

①「kill プロセスID」でリターンキーを押した後の状態。RockhopperのGUIが終了していることがわかる

Terminal



```
iu@bielinux[~/Downloads] ps -f
UID      PID  PPID   C STIME TTY          TIME CMD
iu      9274  9254   0 06:53 pts/4    00:00:00 zsh
iu     23276  9274   0 14:09 pts/4    00:00:13 java -Xmx1200m -ja
iu     30731  9274   0 14:38 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] ps -f                                [ 2:38午後]
UID      PID  PPID   C STIME TTY          TIME CMD
iu      1596  9274   0 16:58 pts/4    00:00:00 ps -f
iu      9274  9254   0 06:53 pts/4    00:00:00 zsh
iu     23276  9274   0 14:09 pts/4    00:00:50 java -Xmx1200m -ja
iu@bielinux[Downloads] kill 23276                            [ 4:58午後]
iu@bielinux[Downloads]                                         [ 6:08午後]
[1] + exit 143  java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads]                                         [ 6:08午後]
```

Sample Data

## W3-6: killコマンド

Terminal

```
iu@bielinux[~/Downloads] ps -f [ 2:38午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu        1596  9274  0 16:58 pts/4    00:00:00 ps -f
iu        9274  9254  0 06:53 pts/4    00:00:00 zsh
iu      23276  9274  0 14:09 pts/4    00:00:50 java -Xmx1200m -ja
iu@bielinux[Downloads] kill 23276 [ 4:58午後]
iu@bielinux[Downloads] [ 6:08午後]
[1] + exit 143  java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] ps -f [ 6:08午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu        9274  9254  0 06:53 pts/4    00:00:00 zsh
iu      21567  9274  0 18:16 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] [ 6:16午後]
```

①

②

The screenshot shows a Linux desktop environment with a terminal window open in the foreground. The terminal window title is 'Terminal' and the command entered is 'iu@bielinux[~/Downloads] ps -f'. The output shows several processes running, including a Java application with PID 23276. A red arrow labeled '①' points to the terminal window. Another red arrow labeled '②' points to the application dock on the left side of the screen, which contains icons for various desktop applications like the Dash, Home, and Dash to Dock.

# W3-7: nohupコマンド

Terminal

```
iu@bielinux[~/Downloads] ps -f
UID      PID  PPID  C STIME TTY          TIME CMD
iu        1596  9274  0 16:58 pts/4    00:00:00 zsh
iu        9274  9254  0 06:53 pts/4    00:00:00 ps -f
iu       23276  9274  0 14:09 pts/4   00:00:50 java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] kill 23276
iu@bielinux[Downloads]
[1] + exit 143  java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] ps -f
UID      PID  PPID  C STIME TTY          TIME CMD
iu        9274  9254  0 06:53 pts/4    00:00:00 zsh
iu       21567  9274  0 18:16 pts/4   00:00:00 ps -f
iu@bielinux[Downloads] nohup java -Xmx1200m -jar Rockhopper.jar&

```

①

Sample Data

バックグラウンドジョブ時には、通常コマンドの最後に&をつけるだけでなく、①コマンドの最初にnohupをつける。一般的なNGS解析の利用法は、SSH経由で大型計算機にアクセスし、そこで長時間の計算を実行する。このとき、&だけだとログアウト時に計算が終了してしまうが、nohupをつけることで、ログアウト後も計算を継続させることができる

# W3-7: nohupコマンド

Rockhopper

The screenshot shows a desktop environment with a terminal window open in the background and the Rockhopper application in the foreground. The terminal window displays a command-line session where the user runs 'nohup' to start the Rockhopper application in the background. The Rockhopper application window is titled 'Rockhopper' and has a yellow header bar with 'Options' and 'Help' menu items. It features a text input field labeled 'Replicon name:' with a placeholder '██████████' and a yellow 'X' button. Below the input field is a yellow plus sign button.

```
iu@bielinux[~/Downloads] iu      23276  9274  0 14:09 pts/4    00:00
iu@bielinux[Downloads] kill 23276
iu@bielinux[Downloads]
[1] + exit 143  java -Xmx1200m -jar Rockhopper.jar
iu@bielinux[Downloads] ps -f
UID      PID  PPID  C STIME TTY          TIME CMD
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu     21567  9274  0 18:16 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] nohup java -Xmx1200m -jar Rockhopper.jar&
[1] 8785
iu@bielinux[Downloads] nohup: ignoring input and appending output
to 'nohup.out'
```

①でリターンキーを1回押した直後の状態。ターミナル上では、一見コマンド打ち込み不可能なように見えるが、ターミナル画面をアクティブにしてもう一度リターンキーを押すと、ちゃんとバックグラウンドジョブとしてRockhopper GUIが起動していることが確認できる。

# W3-7: nohupコマンド

①SSHの話までするとややこしいので  
これ以上深入りしないが、遺伝研ス  
パコンなどを利用するようになったと  
きに、この意味が理解できるでしょう

The screenshot shows a Linux desktop environment with several icons in the dock on the left. A terminal window is open in the center, showing the following command history:

```
iu@bielinux[~/Downloads] kill 23276 [ 4:58 午後]
iu@bielinux[~/Downloads]
[1] + exit 143  java -Xmx1200m -jar Rockhopper.jar [ 6:08 午後]
iu@bielinux[~/Downloads] ps -f [ 6:08 午後]
UID      PID  PPID  C STIME TTY          TIME CMD
iu        9274  9254  0 06:53 pts/4    00:00:00 zsh
iu       21567  9274  0 18:16 pts/4    00:00:00 ps -f
iu@bielinux[~/Downloads] nohup java -Xmx1200m -jar Rockhopper.jar& [ 7:34 午後]
[1] 8785
iu@bielinux[~/Downloads] nohup: ignoring input and appending output
to 'nohup.out'
```

A red arrow points to the icon in the dock, which is circled with a red circle labeled '①'.

①「ps -f」で確認。確かにPID8785が存在することがわかる

# W3-7: nohupコマンド

The screenshot shows a Linux desktop environment with several icons in the dock on the left. A terminal window titled "iu@bielinux[~/Downloads]" is open, displaying the following command history:

```
iu      9274  9254  0 06:53 pts/4    00:00:00 zsh
iu      21567  9274  0 18:16 pts/4    00:00:00 ps -f
iu@bielinux[Downloads] nohup java -Xmx1200m -jar Rockhopper.jar&
[1] 8785
iu@bielinux[Downloads] nohup: ignoring input and appending output
to 'nohup.out'
```

A red arrow points to the first "nohup" command, which is highlighted with a red circle labeled "①". Below this, another "ps -f" command is run, showing the process details:

UID	PID	PPID	C	STIME	TTY	TIME	CMD
iu	<u>8785</u>	9274	0	08:59	pts/4	00:00:06	java -Xmx1200m -ja
iu	9274	9254	0	9月 05	pts/4	00:00:00	zsh
iu	12373	9274	0	09:12	pts/4	00:00:00	ps -f

The terminal window has a yellow bar at the bottom with "Options Help". A file manager window is visible in the background, showing a "Replicon name:" input field and a yellow "DE" button.

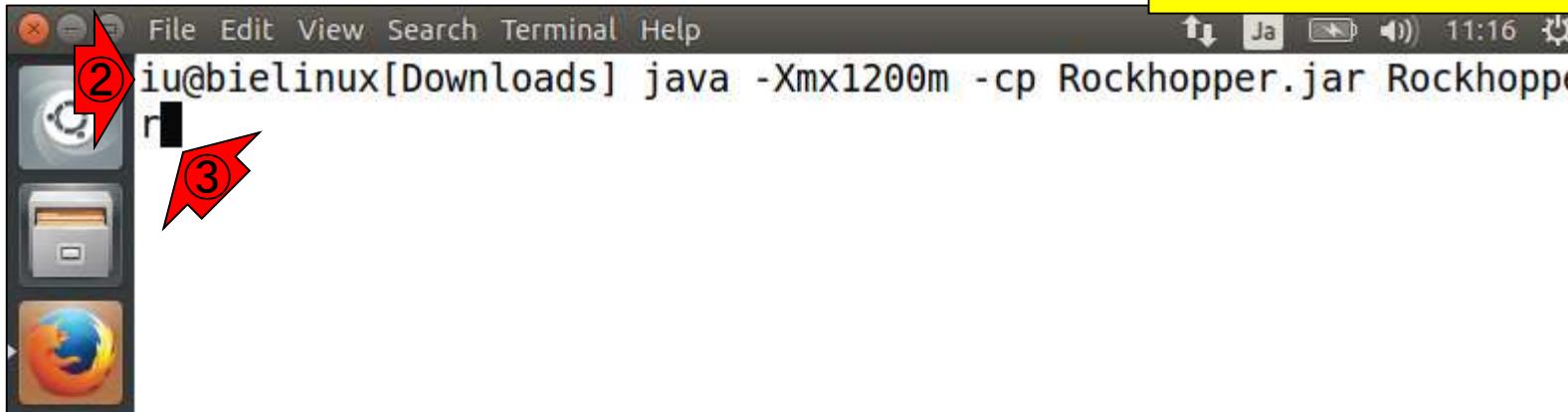
# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用

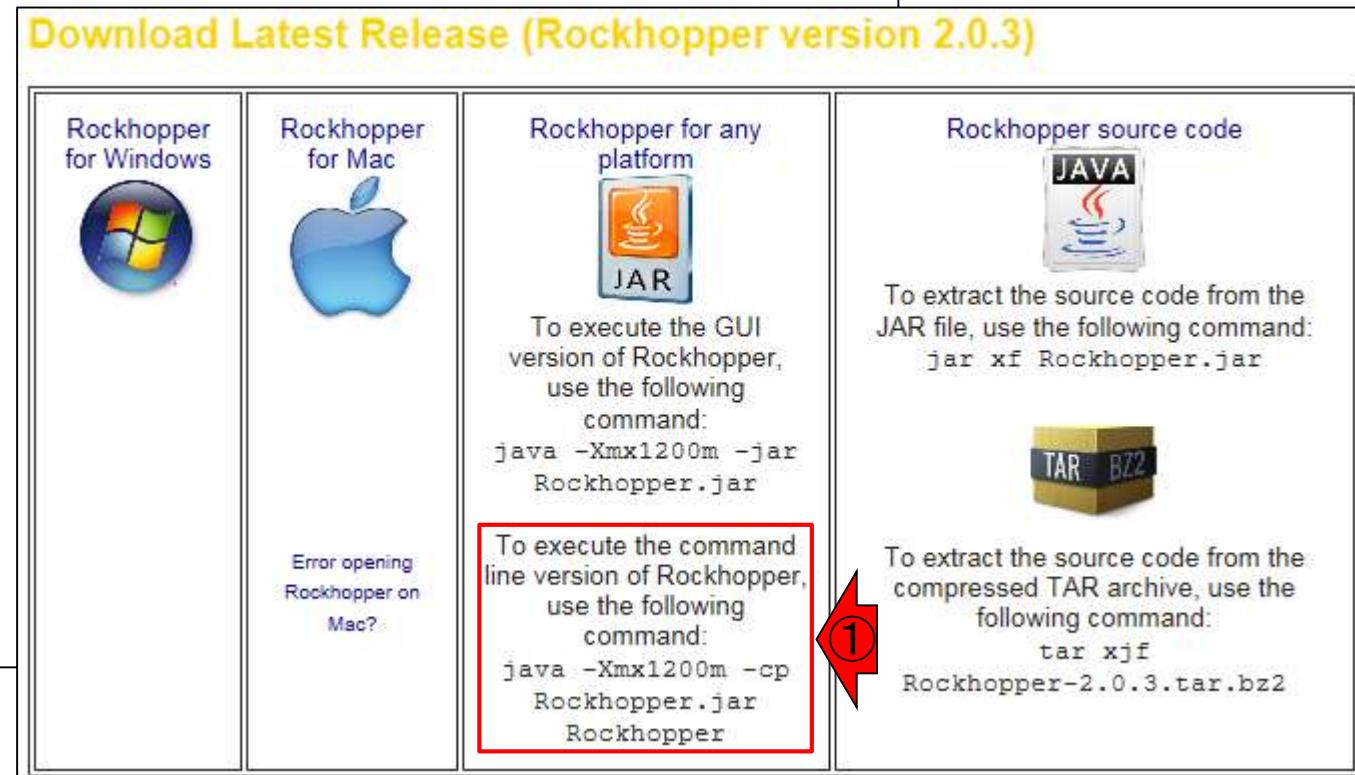


# W4-1: コマンドライン版

①Rockhopperのコマンドライン版を実行したい場合のやり方が書いてあるので、②その通りに実行。③リターンキーを押すと…



```
iu@bielinus[Downloads] java -Xmx1200m -cp Rockhopper.jar Rockhopper
```



# W4-1: コマンドライン版

File Edit View Search Terminal Help

```
java Rockhopper <options> -g genome_DIR1,genome_DIR2  
cate1_pairedend1.fastq%aerobic_replicate1_pairedend2  
_replicate2_pairedend1.fastq%aerobic_replicate2_paire  
naerobic_replicate1_pairedend1.fastq%anaerobic_replicat  
e2.fastq,anaerobic_replicate2_pairedend1.fastq%anaerobic_replicat  
e2_pairedend2.fastq
```

## EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH SINGLE-END READS

```
java Rockhopper <options> aerobic_replicate1.fastq,aerobic_replica  
te2.fastq anaerobic_replicate1.fastq,anaerobic_replicate2.fastq
```

## EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH PAIRED-END READS

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq%aero  
bic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fast  
q%aerobic_replicate2_pairedend2.fastq anaerobic_replicate1_paire  
nd1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicat  
e2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

```
iu@bielinux[Downloads] java -Xmx1200m -cp Rockhopper.jar Rockhopper  
r | more
```

マニュアルが一気に流れる。赤下線で示すように、最後のほうに *de novo* アセンブリのコマンド実行例があるのでなんとなくわかる。  
①マニュアルを最初から眺めるべく、「| more」をつけて、直前のコマンドを再実行

①

# W4-1:コマンドライン版

moreコマンドは、「Returnキー」で1行分づつ、「Spaceキー」で1画面分づつスクロールできる。ただの復習

File Edit View Search Terminal Help

↑ ↓ Ja 11:34

```
*****
***** Rockhopper version 2.03 *****
*****
```

The Rockhopper application has the following required command line arguments.

## REQUIRED ARGUMENTS

tq        exp1A.fastq,exp1B.fastq,exp1C.fastq    exp2A.fastq,exp2B.fas  
tq        a comma separated list of sequencing files (in FASTQ, QSEQ,  
, FASTA, SAM, or BAM format) for replicate experiments, one list per experimental condition (mate-pair files should be delimited by '%')

## REFERENCE BASED ASSEMBLY VS. DE NOVO ASSEMBLY:

IF THE -g OPTION IS USED THEN ROCKHOPPER ALIGNS READS TO ONE OR MORE REFERENCE GENOMES,  
OTHERWISE, ROCKHOPPER PERFORMS DE NOVO TRANSCRIPT ASSEMBLY.

--More--

# W4-1: オプション

この画面あたりが *de novo* アセンブリで使うオプションの説明。①kの値は、Trinityと同じく25がデフォルトのようだ。②アセンブリ後のコンティグの最低配列長は $2*k = 2*25 = 50$ だと解釈

File Edit View Search Terminal Help

## OPTIONAL ARGUMENTS FOR DE NOVO ASSEMBLY ONLY

-k <integer> size of k-mer, range of values is 15 to 31  
(default is 25) 

-j <integer> minimum length required to use a sequencing read after trimming/processing (default is 35)

-n <integer> size of k-mer hashtable is  $\sim 2^n$  (default is 25). HINT: should normally be 25 or, if more memory is available, 26. WARNING: if increased above 25 then more than 1.2M of memory must be allocated

-b <integer> minimum number of full length reads required to map to a de novo assembled transcript (default is 20) 

-u <integer> minimum length of de novo assembled transcripts (default is  $2*k$ )

-w <integer> minimum count of k-mer to use it to seed a new de novo assembled transcript (default is 50)

-x <integer> minimum count of k-mer to use it to extend an existing de novo assembled transcript (default is 5)

## EXAMPLE EXECUTION: REFERENCE BASED ASSEMBLY WITH SINGLE-END READS

--More--

## W4-2: 実行コマンド例

File Edit View Search Terminal Help

```
java Rockhopper <options> -g genome_DIR1,genome  
cate1_pairedend1.fastq%aerobic_replicate1_paired  
_replicate2_pairedend1.fastq%aerobic_replicate2  
naerobic_replicate1_pairedend1.fastq%anaerobic  
d2.fastq,anaerobic_replicate2_pairedend1.fastq%  
2_pairedend2.fastq
```

オプションは特に指定する必要はないなさそうだと解釈し、とりあえず実行例を再確認。これはリターンキーを連打してマニュアルの最後のほうを示している。赤下線部分あたりをじっくり眺めて、①paired-endの2つのファイルは「%」で連結する、②反復実験データがあれば「,」でつなげていけばよいだろうということを学習する。スペースが入っていないことも記憶にとどめておく

EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH SINGLE-END READS

```
java Rockhopper <options> aerobic_replicate1.fastq,aerobic_replica  
te2.fastq anaerobic_replicate1.fastq,anaerobic_replicate2.fastq
```

EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH PAIRED-END READS

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq%aero  
bic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fast  
q%aerobic_replicate2_pairedend2.fastq anaerobic_replicate1_paire  
dend1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicat  
e2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

iu@bielinux[Downloads]  
iu@bielinux[Downloads]

[ 2:47 午後 ]  
[ 2:47 午後 ]

## W4-3: Tips

File Edit View Search Terminal Help

```
java Rockhopper <options> aerobic_replicate1.f  
te2.fastq anaerobic_replicate1.fastq,anaerobic
```

EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH PAIRE

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq%aero  
bic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fast  
q%aerobic_replicate2_pairedend2.fastq anaerobic_replicate1_pairede  
nd1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicat  
e2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

```
iu@bielinux[Downloads]  
iu@bielinux[Downloads] ls  
boost_1_61_0.tar.bz2  
Bridger_r2014-12-01.tar.gz  
FAQCs  
FastQC  
fastqc_v0.11.4.zip  
IGV_2.3.67  
IGV_2.3.67.zip  
kmergenie-1.6982.tar.gz  
iu@bielinux[Downloads]
```

[ 2:47 午後 ]  
[ 2:47 午後 ]

```
master.zip  
nohup.out  
Rockhopper.jar  
Rockhopper_Results  
sratoolkit.2.6.3-ubuntu64.tar.gz  
v2.2.0.tar.gz  
velvet_1.2.10.tgz
```

[ 2:48 午後 ]

①何気なしにls。②nohup.outファイルは、nohupコマンドを実行すると自動的に生成される。③Rockhopper\_Resultsディレクトリは、Rockhopper実行時に自動生成される。これらの何気なしに実行したlsコマンドの結果と、nohupやRockhopperのマニュアルを見比べると、より理解が深まっていくものです

# W4-4: クラスパスの設定

File Edit View Search Terminal Help



①EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH PAIRED-END READS

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq%aerobic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fastq%aerobic_replicate2_pairedend2.fastq anaerobic_replicate1_pairedend1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicate2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

iu@bielinux[Downloads]

[ 2:47 午後 ]

iu@bielinux[Downloads] ls

[ 2:47 午後 ]

boost\_1\_61\_0.tar.bz2

master.zip

Bridger\_r2014-12-01.tar.gz

nohup.out

FaQCs

Rockhopper.jar

FastQC

Rockhopper\_Results

fastqc\_v0.11.4.zip

sratoolkit.2.6.3-ubuntu64.tar.gz

IGV\_2.3.67

v2.2.0.tar.gz

IGV\_2.3.67.zip

velvet\_1.2.10.tgz

kmergenie-1.6982.tar.gz

②iu@bielinux[Downloads] java Rockhopper

[ 2:48 午後 ]

Error: Could not find or load main class Rockhopper

[ 2:48 午後 ]

①EXAMPLE EXECUTIONのところを眺めると、実行は赤下線のようなコマンドを打つと書いてある。しかし、②それを実際に打ってみてもエラーが出る。理由はクラスパスの設定ができていないから



# W4-4: クラスパスの設定

File Edit View Search Terminal Help

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq,aerobic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq anaerobic_replicate1_pairedend1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicate2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

iu@bielinux[Downloads] [ 2:47 午後]  
iu@bielinux[Downloads] ls [ 2:47 午後]

boost_1_61_0.tar.bz2	master.zip
Bridger_r2014-12-01.tar.gz	nohup.out
FaQCS	Rockhopper.jar
FastQC	Rockhopper_Results
fastqc_v0.11.4.zip	sratoolkit.2.6.3-ubuntu64.tar.gz
IGV_2.3.67	v2.2.0.tar.gz
IGV_2.3.67.zip	velvet_1.2.10.tgz

iu@bielinux[Downloads] java Rockhopper [ 2:48 午後]  
Error: Could not find or load main class Rockhopper

iu@bielinux[Downloads] export CLASSPATH=/home/iu/Downloads/Rockhopper.jar  
iu@bielinux[Downloads] java Rockhopper [ 2:50 午後]

①クラスパスの設定は「export CLASSPATH=設定したいjarファイルの絶対パス」。講習会の環境では、Rockhopper.jarの絶対パスは赤下線のように書く。②設定後にもう一度「java Rockhopper」と打つ



# W4-4: クラスパスの設定

リターンキーを押した後の状態。  
エラーメッセージではなく、正しくマニュアルが表示される



```
java Rockhopper <options> -g genome_DIR1,genome_DIR2 aerobic_repli  
cate1_pairedend1.fastq%aerobic_replicate1_pairedend2.fastq,aerobic  
_replicate2_pairedend1.fastq%aerobic_replicate2_pairedend2.fastq a  
naerobic_replicate1_pairedend1.fastq%anaerobic_replicate1_paireden  
d2.fastq,anaerobic_replicate2_pairedend1.fastq%anaerobic_replicate  
2_pairedend2.fastq
```

EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH SINGLE-END READS

```
java Rockhopper <options> aerobic_replicate1.fastq,aerobic_replica  
te2.fastq anaerobic_replicate1.fastq,anaerobic_replicate2.fastq
```

EXAMPLE EXECUTION: DE NOVO ASSEMBLY WITH PAIRED-END READS

```
java Rockhopper <options> aerobic_replicate1_pairedend1.fastq%aero  
bic_replicate1_pairedend2.fastq,aerobic_replicate2_pairedend1.fast  
q%aerobic_replicate2_pairedend2.fastq anaerobic_replicate1_paire  
dend1.fastq%anaerobic_replicate1_pairedend2.fastq,anaerobic_replicat  
e2_pairedend1.fastq%anaerobic_replicate2_pairedend2.fastq
```

iu@bielinux[Downloads]

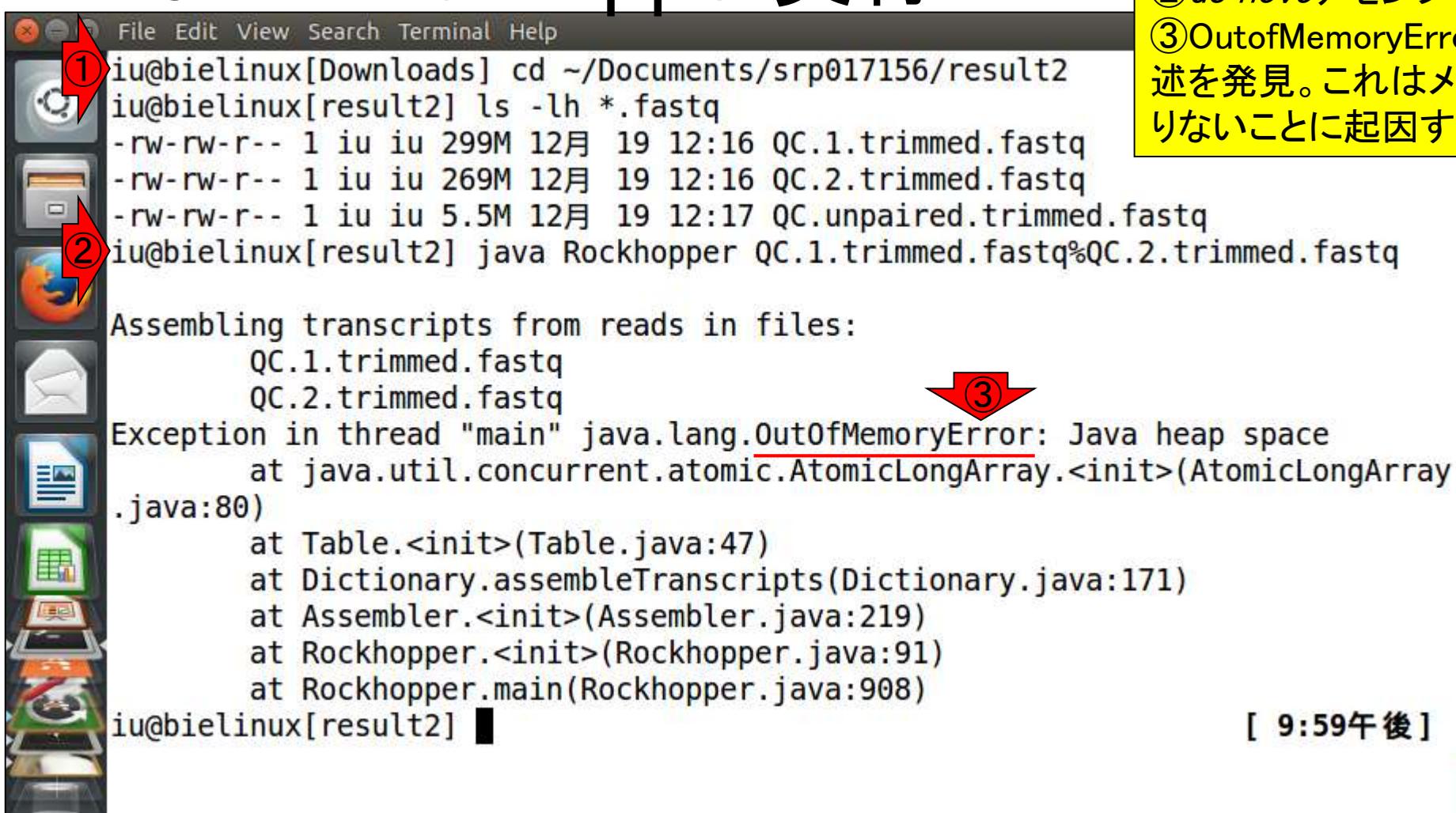
[ 2:52 午後 ]

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W5-1 : Rockhopper実行



```
File Edit View Search Terminal Help
iu@bielinux[Downloads] cd ~/Documents/srp017156/result2
iu@bielinux[result2] ls -lh *.fastq
-rw-rw-r-- 1 iu iu 299M 12月 19 12:16 QC.1.trimmed.fastq
-rw-rw-r-- 1 iu iu 269M 12月 19 12:16 QC.2.trimmed.fastq
-rw-rw-r-- 1 iu iu 5.5M 12月 19 12:17 QC.unpaired.trimmed.fastq
iu@bielinux[result2] java Rockhopper QC.1.trimmed.fastq%QC.2.trimmed.fastq

Assembling transcripts from reads in files:
    QC.1.trimmed.fastq
    QC.2.trimmed.fastq
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
    at java.util.concurrent.atomic.AtomicLongArray.<init>(AtomicLongArray.java:80)
        at Table.<init>(Table.java:47)
        at Dictionary.assembleTranscripts(Dictionary.java:171)
        at Assembler.<init>(Assembler.java:219)
        at Rockhopper.<init>(Rockhopper.java:91)
        at Rockhopper.main(Rockhopper.java:908)
iu@bielinux[result2] [ 9:59午後 ]
```

①FaQCs実行結果ファイルを含むディレクトリに移動して、  
②*de novo*アセンブリを実行。  
③OutOfMemoryErrorという記述を発見。これはメモリが足りないことに起因するエラー

# W5-1 : Rockhopper実行

```
File Edit View Terminal Help
Assembling transcripts from reads in files:
    QC.1.trimmed.fastq
    QC.2.trimmed.fastq
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
    at java.util.concurrent.atomic.AtomicLongArray.<init>(AtomicLongArray.java:80)
    at Table.<init>(Table.java:47)
    at Dictionary.assembleTranscripts(Dictionary.java:171)
    at Assembler.<init>(Assembler.java:219)
    at Rockhopper.<init>(Rockhopper.java:91)
    at Rockhopper.main(Rockhopper.java:908)
iu@bielinux[result2] ls [ 9:59 午後 ]
fastqCount.txt      QC_qc_report.pdf          Rockhopper_Results
QC.1.trimmed.fastq QC.stats.txt
QC.2.trimmed.fastq QC.unpaired.trimmed.fastq
iu@bielinux[result2] ls -l Rockhopper_Results [ 10:02 午後 ]
total 4
drwxrwxr-x 2 iu iu 4096 12月 20 21:58 genomeBrowserFiles
-rw-rw-r-- 1 iu iu     0 12月 20 21:58 summary.txt
iu@bielinux[result2] more Rockhopper_Results/summary.txt [ 10:02 午後 ]
iu@bielinux[result2]
```

① lsすると、一応Rockhopper\_Resultsディレクトリはできている。②その中身を眺めている。summary.txtのファイルサイズも0なうえ、コンティグファイルもできていないことがわかる。③ファイルサイズが0ということは中身がないということ

# W5-2: Rockhopper再実行

```
File Edit View Search Terminal Help
QC.1.trimmed.fastq
QC.2.trimmed.fastq
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
    at java.util.concurrent.atomic.AtomicLongArray.<init>(AtomicLongArray.java:80)
    at Table.<init>(Table.java:47)
    at Dictionary.assembleTranscripts(Dictionary.java:171)
    at Assembler.<init>(Assembler.java:219)
    at Rockhopper.<init>(Rockhopper.java:91)
    at Rockhopper.main(Rockhopper.java:908)
iu@bielinux[result2] ls [ 9:59 午後 ]
fastqCount.txt      QC_qc_report.pdf          Rockhopper_Results
QC.1.trimmed.fastq  QC.stats.txt
QC.2.trimmed.fastq  QC.unpaired.trimmed.fastq
iu@bielinux[result2] ls -l Rockhopper_Results [10:02 午後 ]
total 4
drwxrwxr-x 2 iu iu 4096 12月 20 21:58 genomeBrowserFiles
-rw-rw-r-- 1 iu iu     0 12月 20 21:58 summary.txt
iu@bielinux[result2] more Rockhopper_Results/summary.txt [10:02 午後 ]
iu@bielinux[result2] java -Xmx2000m Rockhopper QC.1.trimmed.fastq%QC.2.trimmed.fastq
```

①

# W5-2: Rockhopper再実行

```
File Edit View Search Terminal Help
at Rockhopper.main(Rockhopper.java:908)
iu@bielinux[result2] ls
fastqCount.txt      QC_qc_report.pdf          Rockhopper_Results
QC.1.trimmed.fastq QC.stats.txt
QC.2.trimmed.fastq QC.unpaired.trimmed.fastq
iu@bielinux[result2] ls -l Rockhopper_Results
total 4
drwxrwxr-x 2 iu iu 4096 12月 20 21:58 genomeBrowserFiles
-rw-rw-r-- 1 iu iu    0 12月 20 21:58 summary.txt
iu@bielinux[result2] more Rockhopper_Results/summary.txt      [10:02午後]
iu@bielinux[result2] java -Xmx2000m Rockhopper QC.1.trimmed.fastq%QC.2.trimmed.fastq

Assembling transcripts from reads in files:
QC.1.trimmed.fastq
QC.2.trimmed.fastq

Aligning reads to assembled transcripts using files:
QC.1.trimmed.fastq
QC.2.trimmed.fastq
```

①「-Xmx2000m」オプションをつける前は、赤枠の途中経過が出る前にOutOfMemoryErrorとなっていたので、②このようなメッセージが出るのを見られただけでもうれしいものです

# W5-2: Rockhopper再実行

```
File Edit View Search Terminal Help

Aligning reads to assembled transcripts using files:
  QC.1.trimmed.fastq
  QC.2.trimmed.fastq

  Total reads in files: 976468
  Perfectly aligned reads: 35  0%
                                         ③

  Total number of assembled transcripts: 0  ②
  Average transcript length: 0
  Median transcript length: 0
  Total number of assembled bases: 0

  Summary of results written to file:
  summary.txt
  Details of assembled transcripts written to file:
  transcripts.txt

  FINISHED.

iu@bielinux[result2] [10:10午後]
```

①無事 *de novo*アセンブリが終了し、コマンド入力待ち状態になっている。但し、②アセンブルされた転写物(transcripts)は1つもないことがわかる。③おそらくこれはバグ。アセンブルされたコンティグ(転写物)が1つもないのに、35リードがマップされたというのは論理的におかしい

# W5-3: 実行結果概観

iu@bielinux[~/Documents/srp017156/result2]

Median transcript length:

Total number of assembled bases:

Summary of results written to file:  
summary.txt

Rockhopper\_Results/su

Details of assembled transcripts written to file:  
transcripts.txt

Rockhopper\_Results/tr

FINISHED.

iu@bielinux[result2] ls

[10:10午後]

fastqCount.txt QC\_qc\_report.pdf

Rockhopper\_Results

QC.1.trimmed.fastq QC.stats.txt

QC.2.trimmed.fastq QC.unpaired.trimmed.fastq

iu@bielinux[result2] ls -l Rockhopper\_Results

[10:17午後]

total 16

drwxrwxr-x 2 iu iu 4096 12月 20 21:58 genomeBrowserFiles

drwxrwxr-x 2 iu iu 4096 12月 20 22:10 intermediary

-rw-rw-r-- 1 iu iu 608 12月 20 22:10 summary.txt

-rw-rw-r-- 1 iu iu 29 12月 20 22:10 transcripts.txt

iu@bielinux[result2]

[10:17午後]

① Rockhopper\_Resultsディレクトリの中身は、エラーを吐いたとき(W5-1)とは異なることがわかる。②summary.txtの中身は、赤枠でも示されているように、基本的に画面に表示されていったアセンブル結果の要約情報が含まれている。

# W5-3: 実行結果概観

File Edit View Search Terminal Help

Total number of assembled bases:

Summary of results written to file:

Results/summary.txt

Details of assembled transcripts written to file:

Results/transcripts.txt

FINISHED.

```
iu@bielinux[result2] ls  
fastqCount.txt      QC_qc_report.pdf  
QC.1.trimmed.fastq QC.stats.txt  
QC.2.trimmed.fastq QC.unpaired.trimmed.fastq  
iu@bielinux[result2] ls -l Rockhopper_Results  
total 16
```

```
drwxrwxr-x 2 iu iu 4096 9月  8 11:01 genomeBrowserFiles  
drwxrwxr-x 2 iu iu 4096 9月  8 11:02 intermediary  
-rw-rw-r-- 1 iu iu  608 9月  8 11:02 summary.txt  
-rw-rw-r-- 1 iu iu   29 9月  8 11:02 transcripts.txt
```

```
iu@bielinux[result2] more Rockhopper_Results/transcripts.txt  
Sequence      Length  Expression 1  
iu@bielinux[result2] ■
```

①アセンブルされた転写物配列情報は transcripts.txt ファイルに格納される。ただし、この場合は1つもコンティグがないので、②more でファイルの中身を表示させても「Sequence Length Expression 1」というヘッダー一行しかないことがわかる

①

[11:02午前]

Rockhopper\_Results

[11:55午前]

②

[11:58午前]

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W6-1 : single-endで実行

```
File Edit View Terminal Help
summary.txt
Details of assembled transcripts written to file:
transcripts.txt

FINISHED.

iu@bielinux[result2] ls -l Rockhopper_Results
total 16
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 22:43 intermediary
-rw-rw-r-- 1 iu iu 608 12月 20 22:43 summary.txt
-rw-rw-r-- 1 iu iu 29 12月 20 22:43 transcripts.txt
iu@bielinux[result2] more Rockhopper_Results/transcripts.txt      [10:45午後]
Sequence          Length Expression 1
iu@bielinux[result2] nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq&
[2] 25122
iu@bielinux[result2] nohup: ignoring input and appending output to 'nohup.out'
,
[2] + done      nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq
iu@bielinux[result2] [10:49午後]
```

①single-endとしてforward側のみのファイル(QC.1.trimmed.fastq)を入力として実行。nohupをつけてバックグラウンドで実行したので、途中経過はターミナル画面上には表示されない。画面出力される内容は、赤下線で示すようにnohup.outというファイルに追加で書き込まれる。リターンキーなどを押さずとも、30秒ほどで②計算終了(done)となる

# W6-1 : single-endで実行

②Rockhopper\_Results中のtranscripts.txtが145 bytesとなっていることから、何かしらアセンブリ結果があるのだろうと解釈する。③summary.txtをlessで眺める

```
File Edit View Terminal Help
iu@bielinux[result2] ls -l Rockhopper_Results [10:43午後]
total 16
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 22:43 intermediary
-rw-rw-r-- 1 iu iu 608 12月 20 22:43 summary.txt
-rw-rw-r-- 1 iu iu 29 12月 20 22:43 transcripts.txt
iu@bielinux[result2] more Rockhopper_Results/transcripts.txt [10:45午後]
Sequence Length Expression 1
iu@bielinux[result2] nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq&
[2] 25122
iu@bielinux[result2] nohup: ignoring input and appending output to 'nohup.out'
'

[2] + done      nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq
iu@bielinux[result2] ls -l Rockhopper_Results [10:49午後]
total 16
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 22:49 intermediary
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary.txt
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts.txt
iu@bielinux[result2] less Rockhopper_Results/summary.txt [10:54午後]
```

# W6-2:single-endで実行

File Edit View Search Terminal Help

Assembling transcripts from reads in file:

Aligning reads to assembled transcripts using file: QC.1.trimmed.fastq

Total reads in file: 976519

Perfectly aligned reads: 119 0%

Total number of assembled transcripts: 1

Average transcript length: 107

Median transcript length: 107

Total number of assembled bases: 107

Summary of results written to file:

summary.txt

Details of assembled transcripts written to file:

transcripts.txt

FINISHED.

Rockhopper\_Results/summary.txt (END)

summary.txtのless実行結果。①アセンブル結果として転写物が1つだけ得られたと解釈する。しかし、その長さは107 bp。入力も107 bpなので、どれか1つのリードを出力したのと同じじゃないかと苦笑。②qで抜ける。quitのqです

①

②

Rockhopper\_Results/su

Rockhopper\_Results/tr

# W6-2:single-endで実行

```
File Edit View Terminal Help ↑ Ja 22:58 ⌂
total 16
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 22:43 intermediary
-rw-rw-r-- 1 iu iu 608 12月 20 22:43 summary.txt
-rw-rw-r-- 1 iu iu 29 12月 20 22:43 transcripts.txt
iu@bielinux[result2] more Rockhopper_Results/transcripts.txt      [10:45午後]
Sequence          Length Expression 1
iu@bielinux[result2] nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq&
[2] 25122
iu@bielinux[result2] nohup: ignoring input and appending output to 'nohup.out'
'

[2] + done      nohup java -Xmx2000m Rockhopper QC.1.trimmed.fastq
iu@bielinux[result2] ls -l Rockhopper_Results                      [10:49午後]
total 16
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 22:49 intermediary
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary.txt
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts.txt
iu@bielinux[result2] less Rockhopper_Results/summary.txt           [10:54午後]
iu@bielinux[result2] █
```

## W6-3:mvでrename

```
File Edit View Search Terminal Help  
iu@bielinux[result2] ls -l Rockhopper_Results  
total 16  
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles  
drwxrwxr-x 2 iu iu 4096 12月 20 22:49 intermediary  
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary.txt  
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts.txt  
iu@bielinux[result2] less Rockhopper_Results/summary.txt  
iu@bielinux[result2] mv Rockhopper_Results/summary.txt Rockhopper_Results/summary_1.txt  
iu@bielinux[result2] mv Rockhopper_Results/transcripts.txt Rockhopper_Results/transcripts_1.txt  
iu@bielinux[result2] ls -l Rockhopper_Results  
total 16  
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles  
drwxrwxr-x 2 iu iu 4096 12月 20 22:49 intermediary  
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary_1.txt  
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts_1.txt
```

ここまででの作業で、Rockhopperはアセンブリ結果ファイルを上書き保存していることがわかる。この後に行うreverse側のsingle-endのアセンブリで結果が消えてしまわぬように、forward側の実行結果ファイルの名前を変更しておく

# W6-4 : reverse側を実行

```
File Edit View Terminal Help
iu@bielinus[result2] java -Xmx2000m Rockhopper QC.2.trimmed.fastq
Assembling transcripts from reads in file: QC.2.trimmed.fastq
Aligning reads to assembled transcripts using file: QC.2.trimmed.fasta
Total reads in file: 977151
Perfectly aligned reads: 706568 72%
Total number of assembled transcripts: 423
Average transcript length: 437
Median transcript length: 228
Total number of assembled bases: 184929
Summary of results written to file: Rockhopper_Results/summary.txt
Details of assembled transcripts written to file: Rockhopper_Results/transcripts.txt
FINISHED.
iu@bielinus[result2] [11:10午後]
```

①reverse側ファイル(QC.2.trimmed.fastq)を入力としてsingle-endのアセンブリを実行。nohupと&をつけてないので、途中経過(summary.txtと同じもの)がターミナル画面上に出力される

# W6-4 : reverse側を実行

①アセンブルされた転写物数は423個!  
②総塩基数は184,929。③入力リード数  
977,151個のうち、72% (706,568個)がマ  
ップされていることがわかる

iu@bielinux[result2] java -Xmx2000m Rockhopper QC.2.trimmed.fastq

Assembling transcripts from reads in file: QC.2.trimmed.fastq

Aligning reads to assembled transcripts using file: QC.2.trimmed.fastq

Total reads in file: 977151

Perfectly aligned reads: 706568 72%

③

Total number of assembled transcripts: 423

①

Average transcript length: 437

Median transcript length: 228

Total number of assembled bases: 184929

②

Summary of results written to file:

summary.txt

Rockhopper\_Results/su

Details of assembled transcripts written to file:  
assembled\_transcripts.txt

Rockhopper\_Results/tr

FINISHED.

iu@bielinux[result2] ■

[11:10午後]

# W6-4: mvでrename

```
File Edit View Search Terminal Help [11:10午後]
iu@bielinux[result2] ls -l Rockhopper_Results
total 208
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 23:09 intermediary
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary_1.txt
-rw-rw-r-- 1 iu iu 591 12月 20 23:09 summary.txt
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts_1.txt
-rw-rw-r-- 1 iu iu 188726 12月 20 23:09 transcripts.txt
①iu@bielinux[result2] mv Rockhopper_Results/summary.txt Rockhopper_Results/summary_2.txt
②iu@bielinux[result2] mv Rockhopper_Results/transcripts.txt Rockhopper_Results/transcripts_2.txt
iu@bielinux[result2] ls -l Rockhopper_Results [11:15午後]
total 208
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 23:09 intermediary
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary_1.txt
-rw-rw-r-- 1 iu iu 591 12月 20 23:09 summary_2.txt
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts_1.txt
-rw-rw-r-- 1 iu iu 188726 12月 20 23:09 transcripts_2.txt
iu@bielinux[result2] [11:15午後]
```

# W6-5: 結果を眺める

①reverse側の実行結果ファイルtranscripts\_2.txtの行数をwcで調べる。424行だったが、最初の1行目はヘッダー行なので、423 transcriptsの結果と矛盾はない。②lessで眺める。

```
iu@bielinux[result2] pwd [11:18午後]
/home/iu/Documents/srp017156/result2
iu@bielinux[result2] ls -l Rockhopper_Results [11:19午後]
total 208
drwxrwxr-x 2 iu iu 4096 12月 20 22:42 genomeBrowserFiles
drwxrwxr-x 2 iu iu 4096 12月 20 23:09 intermediary
-rw-rw-r-- 1 iu iu 582 12月 20 22:49 summary_1.txt
-rw-rw-r-- 1 iu iu 591 12月 20 23:09 summary_2.txt
-rw-rw-r-- 1 iu iu 145 12月 20 22:49 transcripts_1.txt
-rw-rw-r-- 1 iu iu 188726 12月 20 23:09 transcripts_2.txt
iu@bielinux[result2] wc Rockhopper_Results/transcripts_2.txt [11:19午後]
424 1273 188726 Rockhopper_Results/transcripts_2.txt
iu@bielinux[result2] less Rockhopper_Results/transcripts_2.txt [11:19午後]
```



# W6-5: 結果を眺める

①lessで開いた直後の状態。赤枠部分がヘッダー行。ファイル末尾に移動したい場合は「G」、先頭に移動したい場合は「g」

```
File Edit View Search Terminal Help
Sequence Length Expression 1
CAAGCATTGATTGCTTGTGATAATATATCAAAGTTGTCTTTGACAAGGTCGTCTTGAAACTG
AACAAAGTTCGTTAAATGTGCAGGGCCTTGGTACCTCGGTACCGAGGCAAAAGTAACATTG
CGAAGTCAATTGCTAGAACAAACAAATCGAGCTATTGAACAGCTCATATTATGAGAGTTGA
TCCTGGCTCAGGATGAACGCTGGCGCGTGCCTAATACATGCAAGTCGAACGAGTTCTGTTGATG
ATCGGTGCTTGCAGTGGAGATTCAACATGGAACAGTGGCGGACGGGTGAGTAACACGTGGGTAACC
TGCCCTTAAGTGGGGATAAACATTGGAAACAGATGCTAATACCGCATAGATCCAAGAACCGCATG
GTTCTTGGCTGAAAGATGGCGTAAGCTATCGCTTGGATGGACCCGCGCGTATTAGCTAGTTGG
TGAGGTAATGGCTACCAAGGCGATGATACGTAGCCGAAGTGGAGGTTGATCGGCCACATTGGGA
CTGAGACACGGCCCAAACCTACGGGAGGCAGCAGTAGGAAATCTTCCACAATGGACGCAAGTCT
GATGGAGCAACGCCCGTGAAGTAAGAAGGCTTGGGTGTAACACTCTGTTGGAGAAGAAT
GGTCGGCAGAGTAACGTGTCGGCGTGACGGTATCCAACCAGAAAGCCACGGCTAACTACGTGCC
AGCAGCCCGCGTAATACGTAGGTGGCAAGCGTTATCCGGATTTATTGGCGTAAAGCGAGCGCAGG
CGGTTTTTAAGTCTGATGTGAAAGCCCTGGCTTAACCGAGGAAGCGCATCGGAAACTGGGAAAC
TTGAGTGCAGAAGAGGACAGTGGAACTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAAGAAC
ACCACTGGCGAAGGCCGCTGTCTGGTCTGTAACGTACGCTGAGGCTGAAAGCATGGTAGCGAAC
AGGATTAGATAACCCTGGTAGTCCATGCCGTAAACGATGAATGCTAGGTGTTGGAGGGTTCCGCC
TTCAGTGCCCGAGCTAACGCATTAAGCATTCCGCCTGGGAGTACGACCGCAAGGTTGAAACTCAA
AGGAATTGACGGGGCCCGACAAGCGGTGGAGCATGTGGTTAATTGAAAGCAACCGCAAGAAC
TTACCAAGGTCTTGACATCTTGTACCTGAGAGATCAGGTTCCCTCGGGGGCAAAATGACA
GGTGGTGCATGGTTGTCGTCAGCTCGTGTGAGATGTTGGGTTAAGTCCCACGAGCGAAC
CCTTATGACTAGTTGCCAGCATTAGTTGGGCACTCTAGTAAGACTGCCGGTGACAAACCGGAGGA
Rockhopper_Results/transcripts_2.txt
```

# W6-5: 結果を眺める

ファイル末尾に移動。赤枠内に見えているのは、計4 transcripts分の情報。2列目がLengthなので、赤下線部分が転写物の長さ情報に相当。qで抜ける

```
File Edit View Search Terminal Help 18:06  
128 113  
TATGCTGATCGGGAAATTCTGAATTATCCGGTGGTCAACAAACAGCGAGTTGCCATTGCTCGAGCG  
ATTGTACTCGAGCCGCAAGTATTGCTGCTAGATGAACCGTTATCAGCACTTGACGCCAATTGCGT  
AAGGATATGCAATATGAATTGCGCGAATTGCAGGAACGGTTGGGGATCACTTCTATTGTGACA  
CACGACCAAGAAGAACGTTGGCTTGTGGACGAAATTTCATGAACGATGGTGAAGTGCAA  
CAAAGTGGTACGCCAGTTGATATTATGATGAGCCGGTCAATCATTGTGGCGGATTTCATTGGT  
GAAAGTAACATCATTCAAGGGCACATGATTAAGGACTTTAGTTGAGTTCAATGGCAAACGGTTT  
GAATGTGCCGATGCCGGAATGCG 419 138  
TCACTCAACCAACCCGCCAAAGACGGCGAAACGGTTGGCTTGACGTTGATCCTGAGGACATCCA  
TGTATGCGGCTAACGAAATCTGAAGAAGATTCGACGCTCGGCTGGAAACCTACGAAGGGGAATA  
ACGTCAGTAAGGATCAGTGGGCAGTGTATCGAAGCGTTTGACGATTAGATAGACACTGGTCG  
GCGCTGATCGATTTCAACTGAGGAGGGTCATCGTAAAAAAATCCACCACAAACGCCGCATTCTA  
CACACCTTATGTGATGTGGCTGGCATTATTGTGA 299 162  
CTTGCCAATTATCAGACCTATTTCAATCAGGCACTTATATTGATGACGATCAATTCCGTCTGG  
TATGCGTTCTGATCACGTTAGCAACTTGTGAAATCAGCTATCCAACAGCATATTGCTGCATTAT  
GCCAAGCATAAACAGTTATGGCTGTTGCTGATTATTCTACCGACTTGGATTAAC 186  
139  
AATAGTTCTGGGAATGTTGGCATCGCGCCCCAGCAATTCTGTTACGGATTTAGTTCTATT  
TTTGTGCTGCTTATATTGAAATCCCCTCATGATCTTCCAATTTCATGCAATCGAAGAATT  
CCGGAAAACCTTGTCAATGCTGCCAAGATTAGGGGCCAAAGCTGGCAGACCTTCACCAAGGTG  
ATCTGGCCGTTGACAATTCTGGCGTGAAATCCGGGTTCAAGCTGTTTCATTCCAAGTTGAGT  
TTATTGATGT 274 121  
(END)
```

# 第5回原稿PDFのp195

ここまでが、バクテリア用*de novo*トランスクリプトームアセンブラーである①Rockhopper2実行部分

として、FastQCによるクオリティチェックを行えばよい [W1-2]。著者らは、FastQC実行結果ファイルの項目(Overrepresented sequences)を眺めて、トリム前に見えていた既知のアダプターやプライマー配列が、トリム後に正しく見えなくなっていることを確認して安心している [W1-3]。

このデータに関して結論からいえば、forward側の107 bp のリードファイル (SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq) のうち、100-107 塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけ QC 段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper2<sup>18)</sup>によるトランスクリプトームアセンブリー、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そして QC 再実行である。

## トランスクリプトームアセンブリー

ゲノムのアセンブリは、断片化されたゲノム配列由来リードをつなぎ合わせて、元のゲノム配列を再構築する作業である。この再構築に相当する英語がアセンブリ(assembly)であり、再構築を行うプログラムをアセンブラー(assembler)という。デノボ(*de novo*)という言葉が同時に用いられることが多いが、これは「最初から」と

か「一から」という意味である。このため、リードのみを入力として(つまり他の情報を一切利用せずに)アセンブルする際には、*de novo assembly*という表現がなされる。トランスクリプトームアセンブリとは、アセンブル対象がゲノムではなく解析サンプル中で発現している全転写物(トランスクリプトーム)の場合を指す。RNA-seq データ

のみを入力として一からアセンブルする場合は、*de novo transscriptome assembly* と呼ばれる。

このデータに関して結論からいえば、forward側の107 bp のリードファイル (SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq) のうち、100-107 塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけ QC 段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper2<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そして QC 再実行である。



# ここまでまとめ

## ■ オリジナル(SRR616268)

- 乳酸菌paired-end RNA-seqデータで、最初の100万リードのみ抽出
- forward側(SRR616268sub\_1.fastq.gz)のリード長は107 bp
- reverse側(SRR616268sub\_2.fastq.gz)のリード長は93 bp



## ■ FaQCs実行結果(W1-1)

- 1,000,000リード → 977,202リード (W1-3)
- forward側(QC.1.trimmed.fastq)
- reverse側(QC.2.trimmed.fastq)
- リード長はバラバラ。FastQC上で見られるIllumina adapterは消滅状態

## ■ *de novo*トランスクリプトームアセンブリ(Rockhopper 2)実行結果

- paired-end (QC.1.trimmed.fastqとQC.2.trimmed.fastq) : 0 transcript or contig (W5-2)
- single-end (forward側のみ; QC.1.trimmed.fastq) : 1 transcript (W6-2)
- single-end (reverse側のみ; QC.2.trimmed.fastq) : 423 transcripts (W6-4)

①

# 第5回原稿PDFのp197

イルが得られ、これが実行ファイルになる。つまり、基本的にJavaファイルのダウンロード完了がインストール完了を意味する [W2-3]。これはWindows版 (Rockhopper.exe) やMacintosh版 (Rockhopper.dmg) についても同じである。Bio-Linux 8では、GUI版とコマンドライン版の両方が利用可能であり、基本的に指示された通りのコマンドを打てばよい [W2-4]。バックグラウンドジョブ (nohupと&の付加) やプロセス管理 (psとkill) は、特に遺伝研スパコンなどの大型計算機にセキュアシェル (secure shell; SSH) 経由でログインして解析する際に利用すると思われる。このため、GUI版の起動説明 (java -Xmx1200m -jar Rockhopper.jar) と絡めて、これらの基本的な利用法を示した [W3]。

コマンドライン版の実行コマンド (java -Xmx1200m -cp Rockhopper.jar Rockhopper) も、GUI版と似ている [W4-1]。「-Xmx1200m」は、最大メモリを1200MB分確保するという意味である。「-cp」は、クラスパス(classpath)を意味し、「-classpath」と書いててもよい。これは、「パスを通す」ことと本質的に同じ概念である。しかしながら、第4回 (W9-5; W15-5; W18-3) で示したような「sudo ln -s /home/iu/Downloads/Rockhopper.jar /usr/local/bin」を実行してもRockhopper.jarのタブ補完がうまくいくようになるだけである。この作業では、コマンドライン版をうまく実行できない。RockhopperのEXAMPLE EXECUTIONは「java Rockhopper <options> ...」となっているが、「java Rockhopper」でエラーが出ないようにするには、クラスパスを正しい手順で設定する必要がある [W4-4]。Java特有の概念であること、Rockhopper

うに、アセンブルされた転写物は1つもなかったことがわかる。この原因は、前述のようにforward側の107 bpのリードファイル (QC.1.trimmed.fastq) にある。特に、100-107塩基付近に乳酸菌に由来しないもの（以下、 $f_{100-107}$ ）がトリムしきれずに多く残っているためである。ただし、これは乳酸菌ゲノム配列にQuasR<sup>19)</sup>を用いてリードのマッピングを行った結果（後述）を眺めることで後に判明したことである。

アセンブル結果のみを眺めていた当時は、single-endのみで実行した結果よりも paired-endの結果のほうが悪いという、理解に苦しむ現象に苦惱していた [W6]。具体的には、①forward側ファイル (QC.1.trimmed.fastq) のsingle-endアセンブル結果が1 transcripts (107 bp)、②reverse側ファイル (QC.2.trimmed.fastq) のsingle-endアセンブル結果が423 transcripts (平均437 bp)、そして③paired-endのアセンブル結果が0 transcriptsであった (Rockhopper2 ver. 2.0.3)。



## Rの基本的な利用法とパッケージのインストール

Bio-Linux 8にはR<sup>31)</sup>がプレインストールされている。著者らの環境では、2015年4月にリリースされたR(ver. 3.2.0)が利用可能である。Biostringsなどいくつかの代表的なパッケージもプレインストールされているものの、マッピングからカウントデータ取得まで行えるQuasRを含む比較的最近のパッケージは、インストールから行う必要がある。ここでは、ゲストOS (Bio-Linux8) 上でのRの基本的な利用法とQuasRパッケージのインストール法を示す。ホストOS (WindowsやMacintosh) 上でのR

# 第5回原稿PDFのp195

として、FastQCによるクオリティチェックを行えばよい [W1-2]。著者らは、FastQC実行結果ファイルの項目(Overrepresented sequences)を眺めて、トリム前に見えていた既知のアダプターやプライマー配列が、トリム後に正しく見えなくなっていることを確認して安心している [W1-3]。

このデータに関して結論からいえば、forward側の107 bp のリードファイル (SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq) のうち、100-107 塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけ QC 段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper2<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そして QC 再実行である。

## トランスクリプトームアセンブリ

ゲノムのアセンブリは、断片化されたゲノム配列由来リードをつなぎ合わせて、元のゲノム配列を再構築する作業である。この再構築に相当する英語がアセンブリ(assembly)であり、再構築を行うプログラムをアセンブラー(assembler)という。デノボ(*de novo*)という言葉が同時に用いられることが多いが、これは「最初から」と

か「一から」という意味である。このため、入力として(つまり他の情報を一切利用せずに)アセンブルする際には、*de novo assembly*という表現がなされる。トランスクリプトームアセンブリとは、アセンブル対象がゲノムではなく解析サンプル中で発現している全転写物(トランスクリプトーム)の場合を指す。RNA-seq データのみを入力として一からアセンブルする場合は、*de novo transcriptome assembly*と呼ばれる。

この後は、①RパッケージQuasRによる乳酸菌ゲノムへのマッピングの話。そしてそれに関連したLinux環境でのRの利用の話に移行

このデータに関して結論からいえば、forward側の107 bp のリードファイル (SRR616268sub\_1.fastq.gz → QC1.trimmed.fastq) のうち、100-107 塩基付近に乳酸菌に由来しないものがトリムしきれずに多く残っている。これは、アセンブルやマッピングがうまくできない、という実害を被ることでわかる。計算時間がかかるため、できるだけ QC 段階で問題解決するという方針もある。しかし、やってみてはじめてわかることがある。以降の内容は、著者らが実際に行ったことを問題解決に至る思考回路とともに述べる。大まかに述べると、Rockhopper2<sup>18)</sup>によるトランスクリプトームアセンブリ、QuasR<sup>19)</sup>による乳酸菌ゲノムへのマッピング、そして QC 再実行である。

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W7-1 : Rの起動

①Rの起動は、「R」と打ってリターンキーを押すだけ。②Rのバージョンは3.2.0であることがわかる。③「>」となっていれば、コマンド入力待ち状態

iu@bielinux[~/Documents/srp017156/result2]

[ 4:51午後 ]

```
iu@bielinux[result2] pwd  
/home/iu/Documents/srp017156/result2  
iu@bielinux[result2] R
```

[ 4:51午後 ]

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

> [ 3 ]

# W7-1 : Rの終了

```
File Edit View Search Terminal Help [ 16:40 ]  
iu@bielinux[result2] pwd [ 3:59 午後 ]  
/home/iu/Documents/srp017156/result2  
iu@bielinux[result2] R [ 4:35 午後 ]  
  
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
①> q()
```

①「Save workspace image?」と聞かれる。この意味  
がよくわからないうちは、Noに相当する「n」を打つ

# W7-1: Rの終了

```
File Edit View Search Terminal Help
/home/iu/Documents/srp017156/result2
iu@bielinix[result2] R [ 5:04 午後]

R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

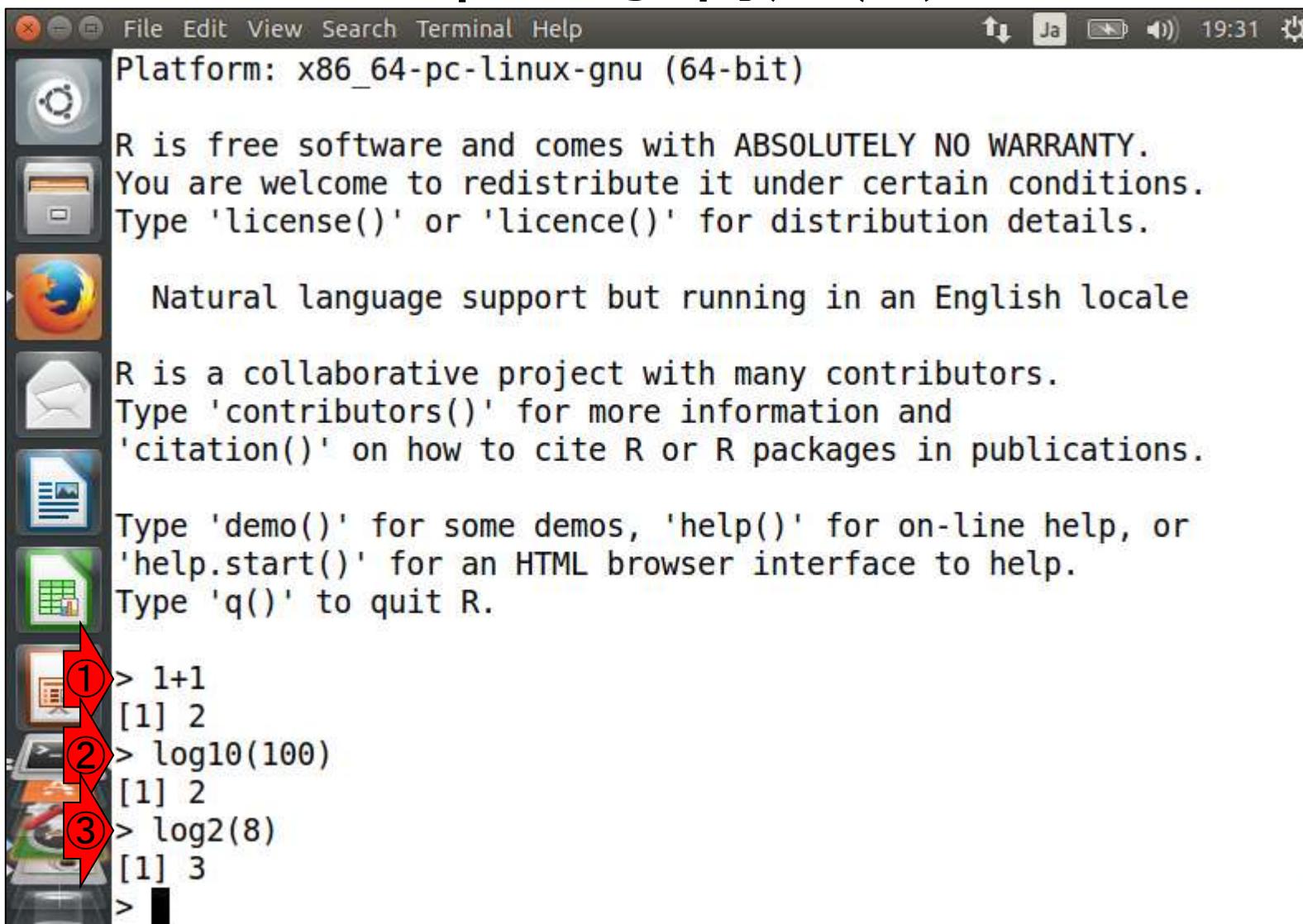
> q()
① Save workspace image? [y/n/c]:
```

①「n」と打ってリターンした直後の状態。②通常のBio-Linuxのコマンド入力待ち状態に戻ったことがわかる

# W7-1: R の終了

```
File Edit View Search Terminal Help  
iu@bielinux[result2] R [ 5:04 午後 ]  
  
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> q()  
Save workspace image? [y/n/c]: n ①  
iu@bielinux[result2] [ 5:08 午後 ] ②
```

# W7-2: 基本的な利用法



The screenshot shows a Linux desktop environment with the Unity interface. A terminal window titled 'Platform: x86\_64-pc-linux-gnu (64-bit)' is open, displaying the R startup message. On the left, there is a vertical dock of icons for various applications like the Dash, Home, and specific productivity tools. Three red numbered arrows (1, 2, 3) point from the bottom left towards the terminal window, indicating the sequence of steps being demonstrated.

```
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

① > 1+1
[1] 2
② > log10(100)
[1] 2
③ > log2(8)
[1] 3
> █
```

## W7-3: パッケージのロード

File Edit View Search Terminal Help

Type 'license()' or 'licence()' for distribution de-

Natural language support but running in an English

R is a collaborative project with many contributors

Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publ

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

> 1+1

[1] 2

> log10(100)

[1] 2

> log2(8)

[1] 3

> library(QuasR)

Error in library(QuasR) : there is no package called 'QuasR'

②

> q()

Save workspace image? [y/n/c]: n

iu@bielinux[result2]

①QuasRパッケージを利用したい場合はlibraryという関数を用いてロードする。Bio-Linux8にはまだQuasRがインストールされていないので、「そのようなパッケージはない」と文句を言っていることがわかる。②一旦終了。講習会環境ではインストール済みなので、エラーは出ない。このあたりは画面を見るだけ(エアーハンズオン)

[ 7:23午後 ]

## W7-4: パッケージインストール

The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "iu@bielinux[~/Documents/srp017156/result2]". Inside the terminal, the command "sudo R" is being typed, followed by a password prompt "[sudo] password for iu:". A red arrow labeled ① points to the "sudo" command, and another red arrow labeled ② points to the password prompt.

```
iu@bielinux[~/Documents/srp017156/result2]
iu@bielinux[result2] sudo R
[sudo] password for iu:
```

①作業ディレクトリはどこでもいいので「sudo R」。これはroot(管理者)権限でRを実行するという指令。パッケージのインストール時に書き込み権限がないことに起因するエラー回避が目的。②パスワードを聞かれたらログインパスワード(推奨手順通りだとpass1409)を打ち込む。**見るだけ**

## W7-4: パッケージインストール

```
File Edit View Search Terminal Help  
iu@bielinux[result2] sudo R  
[sudo] password for iu:  
  
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computation  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
① > source("http://bioconductor.org/biocLite.R")
```

①「source("…")」を打ち込む。これは赤下線部分で示す biocLiteというインストール用の関数を利用できるようにするためのおまじないのようなもの。ネットワーク経由でのインストール作業になる。有線LAN環境が望ましい。[見るだけ](#)

# W7-4: QuasRインストール

```
File Edit View Search Terminal Help  
Ubuntu icon  
Terminal icon  
Ja icon  
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> source("http://bioconductor.org/biocLite.R")  
Bioconductor version 3.1 (BiocInstaller 1.18.3), ?biocLite for help  
> biocLite("QuasR")
```

①biocLite関数を用いてQuasRをインストール。赤下線部分を変えることで、同じノリで他の任意のパッケージをインストール可能。東大有線LAN環境でインストール完了まで約20分。講習会当日できなかつたら後々困るので危ない橋は渡らないのです。見るだけ

# W7-5:途中経過1

```
File Edit View Search Terminal Help 20:08  
trying URL 'http://cran.rstudio.com/src/contrib/latticeExtra_0.6-2  
6.tar.gz'  
Content type 'application/x-gzip' length 1295154 bytes (1.2 MB)  
=====  
downloaded 1.2 MB  
  
trying URL 'http://cran.rstudio.com/src/contrib/futile.logger_1.4.  
1.tar.gz'  
Content type 'application/x-gzip' length 17806 bytes (17 KB)  
=====  
downloaded 17 KB  
  
trying URL 'http://bioconductor.org/packages/3.1/bioc/src/contrib/  
Rbowtie_1.8.0.tar.gz'  
Content type 'application/x-gzip' length 885959 bytes (865 KB)  
=====  
downloaded 865 KB  
  
trying URL 'http://bioconductor.org/packages/3.1/bioc/src/contrib/  
BSgenome_1.36.3.tar.gz'  
Content type 'application/x-gzip' length 6061522 bytes (5.8 MB)  
=====
```

# W7-5:途中経過2

```
File Edit View Search Terminal Help  
** building package indices  
** installing vignettes  
** testing if installed package can be loaded  
* DONE (QuasR)  
  
The downloaded source packages are in  
      '/tmp/RtmpS1T00B/downloaded_packages'  
Old packages: 'annotate', 'biomaRt', 'Biostrings', 'BitSeq', 'DESeq',  
  'edgeR',  
  'evaluate', 'gdata', 'gee', 'GenomeInfoDb', 'GenomicRanges', 'Hi  
lbertVis',  
  'IRanges', 'limma', 'lme4', 'matrixStats', 'pcaMethods', 'plotri  
x',  
  'prettyR', 'qvalue', 'Rcpp', 'RcppEigen', 'RCurl', 'S4Vectors',  
'scales',  
  'scatterplot3d', 'sp', 'stringi', 'tcltk2', 'testthat', 'XML', 'xtable',  
  'zlibbioc', 'boot', 'class', 'cluster', 'codetools', 'foreign',  
'KernSmooth',  
  'lattice', 'MASS', 'Matrix', 'mgcv', 'nlme', 'nnet', 'rpart', 'spatial',  
  'survival'  
Update all/some/none? [a/s/n]: a
```

①リターンキーを押してから約5分後にこのような状態になる。古いパッケージのアップデートをするかどうかを聞かれている。基本はすべてアップデートの「a」か、アップデートしないの「n」。②ここでは「a」と打ってリターン

# W7-5:途中経過3

```
File Edit View Search Terminal Help  
* DONE (prettyR)  
* installing *source* package 'qvalue' ...  
** R  
** data  
** inst  
** preparing package for lazy loading  
** help  
*** installing help indices  
** building package indices  
** installing vignettes  
** testing if installed package can be loaded  
* DONE (qvalue)  
* installing *source* package 'Rcpp' ...  
** package 'Rcpp' successfully unpacked and MD5 sums checked  
** libs  
g++ -I/usr/share/R/include -DNDEBUG -I../inst/include/ -fpic  
-g -O2 -fstack-protector --param=ssp-buffer-size=4 -Wformat -Werro  
r=format-security -D_FORTIFY_SOURCE=2 -g -c Date.cpp -o Date.o  
g++ -I/usr/share/R/include -DNDEBUG -I../inst/include/ -fpic  
-g -O2 -fstack-protector --param=ssp-buffer-size=4 -Wformat -Werro  
r=format-security -D_FORTIFY_SOURCE=2 -g -c Module.cpp -o Module.  
o
```

特にエラーを吐くことなく順調にインストールが進んでいるようだ。この間は、基本的に画面が流れているかどうかに注意を払っていればよい。もし止まっているようだったら、「何か聞かれているかエラーかも」という視点でメッセージを見る

# W7-5: 終了後の状態

①コマンド入力待ち状態になれば基本的にOK。ぱッと見、エラーメッセージが出ていないようだ

```
File Edit View Search Terminal Help ↑ ↓ Ja 20:28
installing to /usr/lib/R/library/mgcv/libs
** R
** data
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (mgcv)

The downloaded source packages are in
  '/tmp/RtmpS1T00B/downloaded_packages'
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
Warning messages:
1: In install.packages(update[instlib == l, "Package"], l, contrib
url = contriburl, :
  installation of package 'RCurl' had non-zero exit status
2: In install.packages(update[instlib == l, "Package"], l, contrib
url = contriburl, :
  installation of package 'XML' had non-zero exit status
> █
```

①

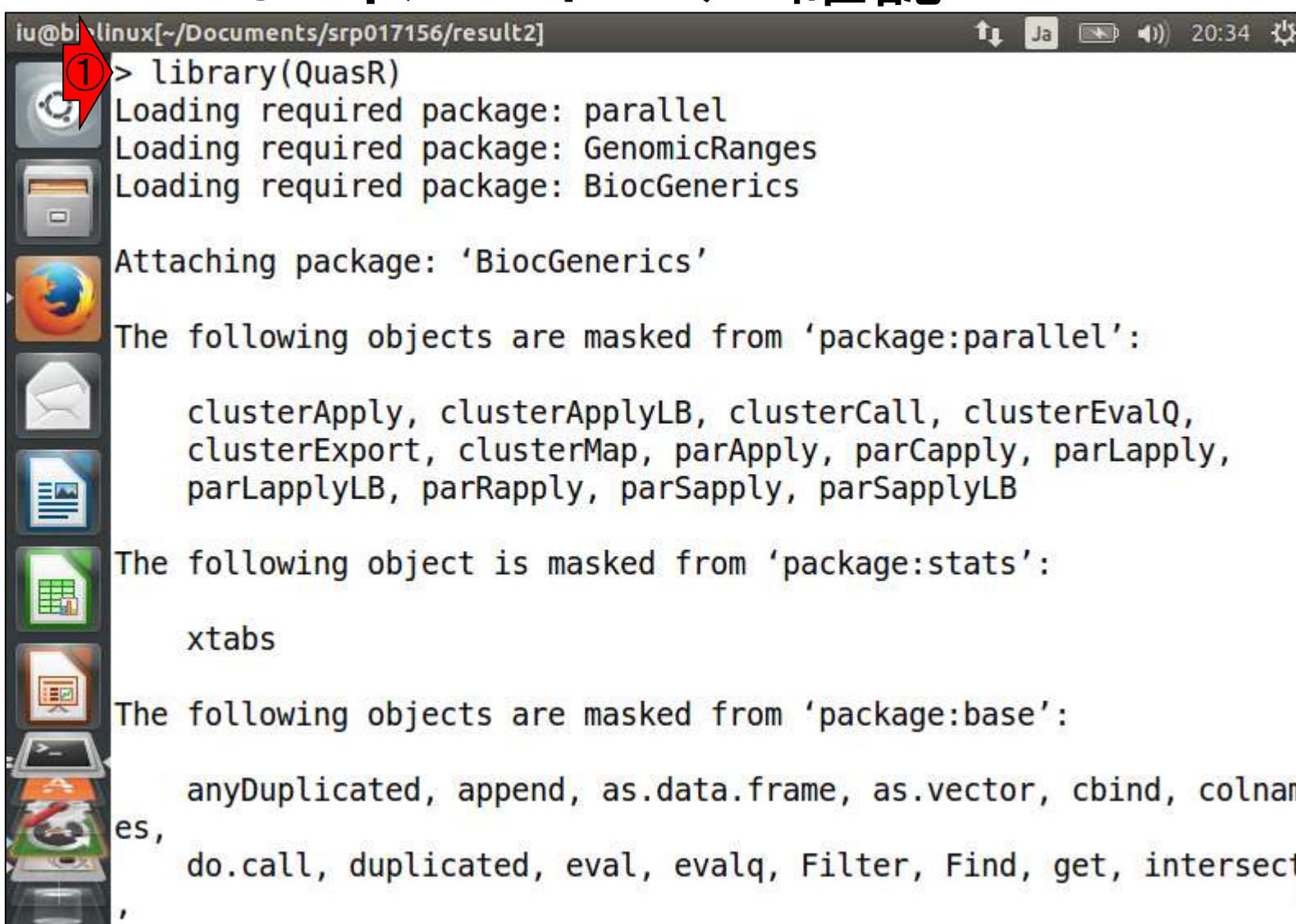
# W7-6:インストール確認

```
File Edit View Terminal Help 20:32
installing to /usr/lib/R/library/mgcv/libs
** R
** data
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (mgcv)

The downloaded source packages are in
  '/tmp/RtmpS1T00B/downloaded_packages'
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
Warning messages:
1: In install.packages(update[instlib == l, "Package"], l, contrib
url = contriburl, :
  installation of package 'RCurl' had non-zero exit status
2: In install.packages(update[instlib == l, "Package"], l, contrib
url = contriburl, :
  installation of package 'XML' had non-zero exit status
> library(QuasR)
```

①

# W7-6:インストール確認



```
iu@bi...:~/Documents/srp017156/result2] 20:34
① > library(QuasR)
Loading required package: parallel
Loading required package: GenomicRanges
Loading required package: BiocGenerics

Attaching package: 'BiocGenerics'

The following objects are masked from 'package:parallel':

clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
clusterExport, clusterMap, parApply, parCapply, parLapply,
parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats':

xtabs

The following objects are masked from 'package:base':

anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
```

リターンキーを押した最後のほうの画面。  
特にエラーメッセージは出ていないようだ

# W7-6:インストール確認

```
File Edit View Search Terminal Help
xtabs
The following objects are masked from 'package:base':
anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
is.unsorted, lapply, Map, mapply, match, mget, order, paste, pmax,
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,
rownames, sapply, setdiff, sort, table, tapply, union, unique,
unlist, unsplit
Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in package 'S4Vectors'
Loading required package: IRanges
Loading required package: GenomeInfoDb
Loading required package: Rbowtie
> ■
```

①

# W7-6: インストール確認

File Edit View Search Terminal Help

xtabs

The following objects are masked from 'package:base':

anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,  
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,  
,  
is.unsorted, lapply, Map, mapply, match, mget, order, paste, p  
max,  
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.i  
nt,  
rownames, sapply, setdiff, sort, table, tapply, union, unique,  
unlist, unsplit

Loading required package: S4Vectors

Loading required package: stats4

Creating a generic function for 'nchar' from package 'base' in pac  
kage 'S4Vectors'

Loading required package: IRanges

Loading required package: GenomeInfoDb

Loading required package: Rbowtie

> library(QuasR)

画面がばーっと流れてエラーの確認が  
しづらいときは、もう一度同じコマンドを  
実行するとよい。このとき、一般的な  
LinuxのTipsと同様に、キーボードの上  
矢印キーを押すと直前に打ったコマン  
ドが表示される。有効利用すべし。この  
あたりはホストOS上のRと同じ使用感

①

# W7-6: インストール確認

2回目は、特に何も表示されない。このような場合は、QuasRパッケージのロードに成功していることを意味する。何らかのエラーに遭遇していれば、その旨表示がなされる。**②**一旦Rを終了

```
anyDuplicated, append, as.data.frame, as.vector, cbind, column  
es,  
do.call, duplicated, eval, evalq, Filter, Find, get, intersect  
,  
is.unsorted, lapply, Map, mapply, match, mget, order, paste, p  
max,  
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.i  
nt,  
rownames, sapply, setdiff, sort, table, tapply, union, unique,  
unlist, unsplit
```

```
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in pac  
kage 'S4Vectors'  
Loading required package: IRanges  
Loading required package: GenomeInfoDb  
Loading required package: Rbowtie  
> library(QuasR)  
> q()  
Save workspace image? [y/n/c]: n  
iu@bielinux[result2]
```

[ 8:48 午後 ]



# W7-7: QuasRウェブページ



Home > Bioconductor 3.1 > Software Packages > QuasR

## QuasR

platforms: 3/3 | downloads: 109 5% | posts: 3/3 3/8 | in BioC: 2.5 years  
build: ok | commits: 4.00 | test coverage: 21%

### Quantify and Annotate Short Reads in R

Bioconductor version: Release (3.1)

This package provides a framework for the quantification and analysis of Short Reads. It covers a complete workflow starting from raw sequence reads, over creation of alignments and quality control plots, to the quantification of genomic regions of interest.

Author: Anita Lerch, Dimos Gaidatzis and Michael Stadler

Maintainer: Michael Stadler <michael.stadler@fmiech.de>

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

Gaidatzis D, Lerch A, Hahne F and Stadler MB (2015). "QuasR: Quantification and annotation of short reads in R." *Bioinformatics*, 31(7), pp. 1130-1132. <http://dx.doi.org/10.1093/bioinformatics/btu781>. PMID:25417205.

Langmead B, Trapnell C, Pop M and Salzberg SL (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biology*, 10(2), pp. R25. <http://dx.doi.org/10.1186/gb-2009-10-2-r25>. PMID:19261174.

Au KF, Jiang H, Lin L, Xing Y and Wong WH (2010). "Detection of splice junctions from paired-end RNA-seq data by SpliceMap." *Nucleic Acid Research*, 38(14), pp. 4570-4578. <http://dx.doi.org/10.1093/nar/gkq211>. PMID:20371516.

### Installation

To install this package, start R and enter:

```
#>.try http if https is not available
source("https://bioconductor.org/biocLite.R")
biocLite("QuasR")
```

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("quask")
```

### Documentation >

#### Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

### Support >

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

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

To install this package, start R and enter:

```
## try http if https is not available
source("https://bioconductor.org/biocLite.R")
biocLite("QuasR")
```

BioconductorのQuasRウェブページ。  
QuasRパッケージのインストール手順  
と実際に行ったこととの対応関係がよ  
くわかるでしょう。スライドを見るだけ

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W8-1 : pwd ⇄ getwd()

The screenshot shows a terminal window on a Linux desktop. The desktop icons on the left include a terminal, file manager, browser, email, and others. The terminal window title bar says "iu@bielinux[~/Documents/srp017156/result2]". The terminal content shows:

- Line 1: "iu@bielinux[result2] pwd" followed by the output "/home/iu/Documents/srp017156/result2" with a red box around it.
- Line 2: "iu@bielinux[result2] R"
- Line 3: R startup message:

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```
- Line 4: License information:

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```
- Line 5: Natural language support message:

```
Natural language support but running in an English locale
```
- Line 6: Collaborative project information:

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```
- Line 7: Help information:

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```
- Line 8: User input: "> getwd()"

Red numbered arrows point to specific parts: 1 points to the terminal title bar, 2 points to the pwd command output, and 3 points to the user input at the bottom.

①現在の作業ディレクトリは赤下線で示したところ。②Rを起動。③Linuxのpwdコマンドに対応するR上での作業ディレクトリ表示コマンドは「getwd()」

ホストOS (WinやMac)上のR GUI版との違いは、起動時の作業ディレクトリが、Rを起動した場所になる点

# W8-1 : pwd ⇔ getwd()

The screenshot shows a terminal window on a Linux desktop. The desktop has a vertical dock on the left containing icons for various applications like a file manager, terminal, browser, and email. A red arrow labeled ② points to the terminal icon in the dock. The terminal window title bar says "iu@bielinus[result2] R" and the status bar says "[12:54午後]". The terminal content shows the standard R startup message, followed by the user's command and its output:

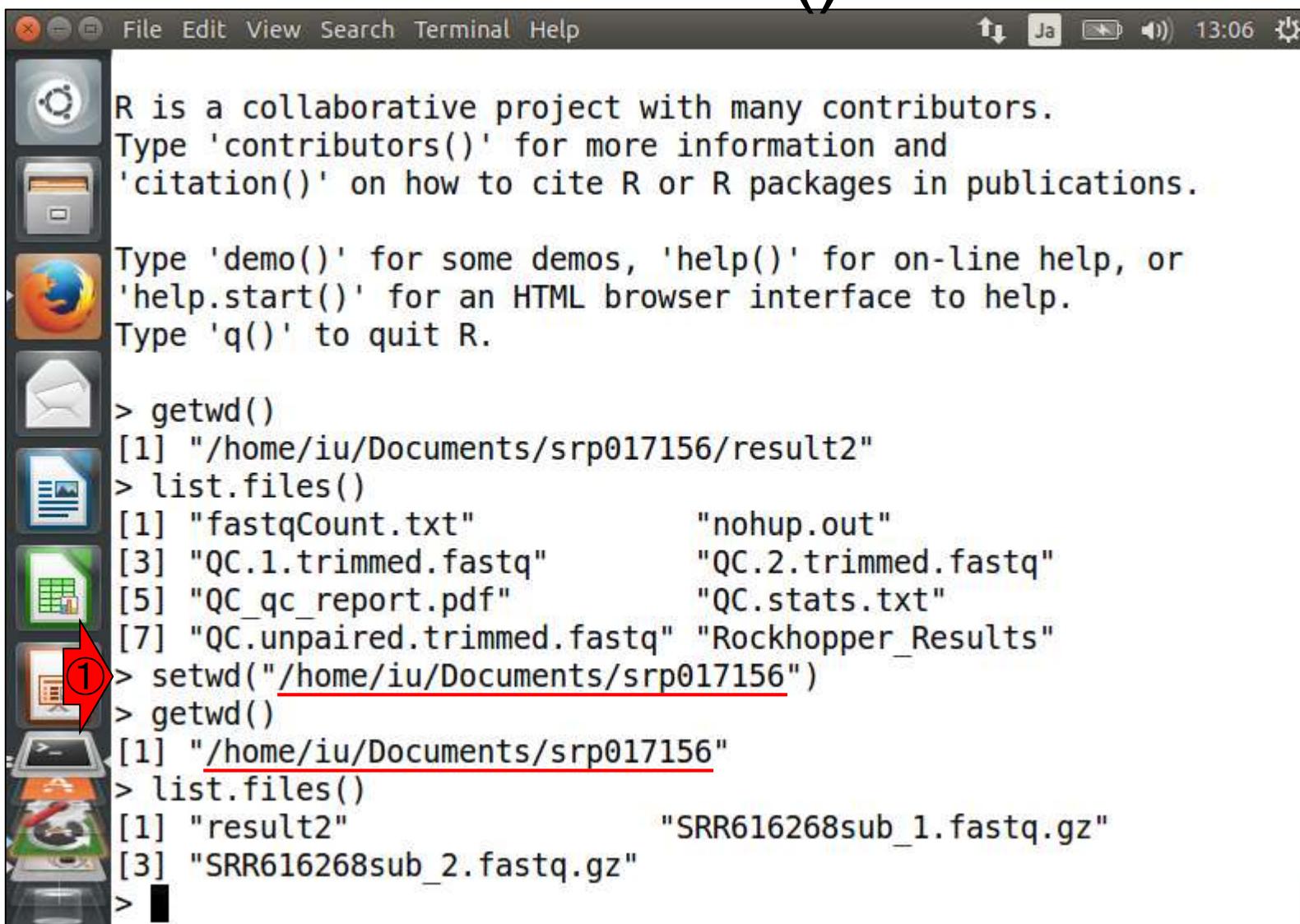
```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> getwd()  
[1] "/home/iu/Documents/srp017156/result2"  
> █
```

A red arrow labeled ③ points to the output of the "getwd()" command, which is highlighted in red.

## W8-2: ls ⇔ list.files()

```
File Edit View Search Terminal Help ↑ Ja 13:01 ⚡  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> getwd()  
[1] "/home/iu/Documents/srp017156/result2"  
> list.files()  
[1] "fastqCount.txt"           "nohup.out"  
[3] "QC.1.trimmed.fastq"      "QC.2.trimmed.fastq"  
[5] "QC_qc_report.pdf"        "QC.stats.txt"  
[7] "QC.unpaired.trimmed.fastq" "Rockhopper_Results"  
> █
```

## W8-3: cd ⇔ setwd()



R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

```
> getwd()
[1] "/home/iu/Documents/srp017156/result2"
> list.files()
[1] "fastqCount.txt"                  "nohup.out"
[3] "QC.1.trimmed.fastq"             "QC.2.trimmed.fastq"
[5] "QC_qc_report.pdf"               "QC.stats.txt"
[7] "QC.unpaired.trimmed.fastq"     "Rockhopper_Results"
> setwd("/home/iu/Documents/srp017156")
> getwd()
[1] "/home/iu/Documents/srp017156"
> list.files()
[1] "result2"                      "SRR616268sub_1.fastq.gz"
[3] "SRR616268sub_2.fastq.gz"
>
```

# W8-4: Rは閉じた世界

File Edit View Search Terminal Help

①作業ディレクトリ変更後に②Rを終了。  
Linuxの世界に戻ったのち③pwd。R起動  
前のディレクトリと同じ。このことから、R  
の中には閉じた世界であることがわかる

```
> getwd()
[1] "/home/iu/Documents/srp017156/result2"
> list.files()
[1] "fastqCount.txt"           "nohup.out"
[3] "QC.1.trimmed.fastq"      "QC.2.trimmed.fastq"
[5] "QC_qc_report.pdf"        "QC.stats.txt"
[7] "QC.unpaired.trimmed.fastq" "Rockhopper_Results"
> setwd("/home/iu/Documents/srp017156")
> getwd()
[1] "/home/iu/Documents/srp017156"
> list.files()
[1] "result2"                  "SRR616268sub_1.fastq.gz"
[3] "SRR616268sub_2.fastq.gz"
> q()
Save workspace image? [y/n/c]: n
iu@bielinux[result2] pwd
/home/iu/Documents/srp017156/result2
iu@bielinux[result2] ls
fastqCount.txt      QC.2.trimmed.fastq  QC.unpaired.trimmed.fastq
nohup.out          QC_qc_report.pdf    Rockhopper_Results
QC.1.trimmed.fastq QC.stats.txt
iu@bielinux[result2]
```

[ 1:07 午後 ]

[ 1:07 午後 ]

[ 1:08 午後 ]

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W9-1：ゲノム配列取得

```
File Edit View Search Terminal Help
iu@bielinux[result2] cd ~/Documents
iu@bielinux[Documents] mkdir genomes
iu@bielinux[Documents] cd genomes
iu@bielinux[genomes] pwd
/home/iu/Documents/genomes
iu@bielinux[genomes] wget -cq ftp://ftp.ensemblgenomes.org/pub/bacteria/release-22/fasta/bacteria_15_collection/lactobacillus_casei_12a/dna/Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz
iu@bielinux[genomes] ls -l
total 884
-rw-rw-r-- 1 iu iu 902847 12月 21 16:28 Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz
iu@bielinux[genomes] [ 4:28午後]
iu@bielinux[genomes] [ 4:28午後]
iu@bielinux[genomes] [ 4:28午後]
iu@bielinux[genomes] [ 4:28午後]
```

①作業ディレクトリは「~/Documents/genomes」。  
②wget実行時にqオプションをつけてるので途中経過が表示されなくてスッキリ。  
③ls  
。②のwget部分は、実際には行わないで!

# W9-1：ゲノム配列取得

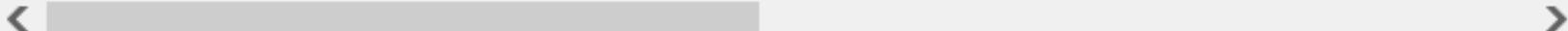
## Rでゲノム解析(Linux版)

- [Lactobacillus casei 12A \(Taxonomy ID: 1051650\): Broadbent et al.](#)
- [Ensembl Genomes \(Cunningham et al., Nucleic Acids Res., 2015\)](#)の
- ゲノム配列取得[W9-1]

①普通にコピペしても行頭に#をつけてコメントアウトしているので、「#wgetというコマンドはない!」と言われて実行されません。講習会では「~/Desktop/backup」フォルダ上にダウンロード済みのファイルを置いてあります。それゆえ、実際には①wgetの代わりに②をコピペ実行

Ensembl release 22の乳酸菌(Lactobacillus casei 12A)ゲノム配列を取得。

```
cd ~/Documents  
mkdir genomes  
cd genomes  
pwd  
① #wget -cq ftp://ftp.ensemblgenomes.org/pub/bacteria/release-22/fasta/bacteria_15_collec  
cp ~/Desktop/backup/Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz .  
ls -l  
  
gunzip Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz  
ls -l  
ls -lh
```



つまり、①の部分が「cp ~/Desktop/backup/Lac\*.」  
になるということです。②gunzipでgzファイルを解凍。  
解凍後のファイルサイズは2,935,945 bytes (約2.8MB)

# W9-1:ゲノム配列取

```
File Edit View Search Terminal Help  ↑ ↓ Ja 16:30
iu@bielinux[result2] cd ~/Documents [ 4:28午後]
iu@bielinux[Documents] mkdir genomes [ 4:28午後]
iu@bielinux[Documents] cd genomes [ 4:28午後]
iu@bielinux[genomes] pwd [ 4:28午後]
/home/iu/Documents/genomes
iu@bielinux[genomes] wget -cq ftp://ftp.ensemblgenomes.org/pub/bacteria/release-22/fasta/bacteria_15_collection/lactobacillus_casei_12a/dna/Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz
iu@bielinux[genomes] ls -l [ 4:28午後]
total 884
-rw-rw-r-- 1 iu iu 902847 12月 21 16:28 Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz
iu@bielinux[genomes] gunzip Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz
iu@bielinux[genomes] ls -l [ 4:30午後]
total 2868
-rw-rw-r-- 1 iu iu 2935947 12月 21 16:28 Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] ls -lh [ 4:30午後]
total 2.9M
-rw-rw-r-- 1 iu iu 2.8M 12月 21 16:28 Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
```

## W9-2: Rを起動

The screenshot shows a terminal window titled "iu@bielinux[genomes] R" with the following content:

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

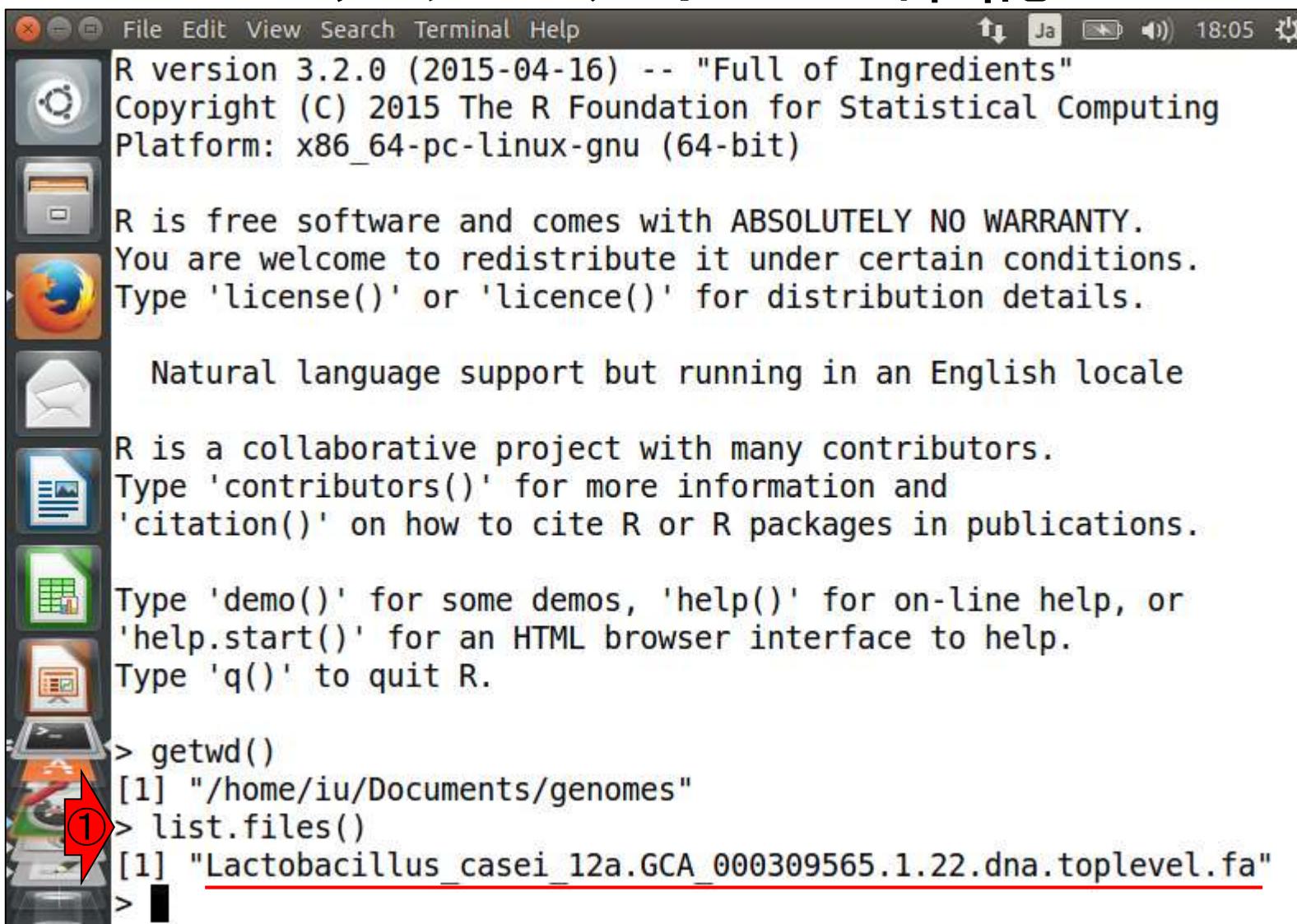
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "/home/iu/Documents/genomes"
> █
```

A red arrow labeled ① points to the terminal icon in the dock on the left. A red arrow labeled ② points to the command "getwd()" in the terminal history.

# W9-3: 入力ファイルの確認



```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "/home/iu/Documents/genomes"
> list.files()
[1] "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa"
```

## W9-4: コピペ用コード

- 連載第1回の「Rでゲノム解析」と同じコード。

このウェブページ中の「[intro](#) | [NGS](#) | [読み込み](#) | [FASTA形式](#) | [基本情報を取得](#)」の5と基本的に同じ。

以下の内容は[JSLAB5\\_1.R](#)の中身と同じです。

```
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa" #①入力ファイル名を  
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してout_fに格納  
  
# 必要なパッケージをロード  
library(Biostrings) #パッケージの読み込み  
  
# 入力ファイルの読み込み  
fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み  
  
# 本番(基本情報取得)  
Total_len <- sum(width(fasta)) #配列の「トータルの長さ」を取得  
Number_of_contigs <- length(fasta) #「配列数」を取得  
Average_len <- mean(width(fasta)) #配列の「平均長」を取得  
Median_len <- median(width(fasta)) #配列の「中央値」を取得  
Max_len <- max(width(fasta)) #配列の長さの「最大値」を取得  
Min_len <- min(width(fasta)) #配列の長さの「最小値」を取得  
  
# 本番(N50情報取得)  
sorted <- rev(sort(width(fasta))) #長さ情報を降順にソートした結果をsortedに格納  
obj <- (cumsum(sorted) >= Total_len*0.5) #条件を満たすかどうかを判定した結果をobjに格納  
N50 <- sorted[obj][1] #objがTRUEとなる1番最初の要素のみ抽出した結果をN50に格納
```



#パッケージの読み込み

# W9-5: コピペ

- 連載第1回の「Rでゲノム解析」と同じコード。  
このウェブページ中の「[intro](#) | [NGS](#) | [読み込み](#) | [FASTA形式](#) | [基本情報を取得](#)」の5.と基本的に同じ。  
以下の内容は[JSLAB5\\_1.R](#)の中身と同じです。

```
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa" #入力ファイル名を指定  
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してout_fに格納  
  
#必要なパッケージをロード  
library(Biostrings)  
  
#入力ファイルの読み込み  
fasta <- readDNAStringSet(in_f, ...)  
  
#本番(基本情報取得)  
Total_len <- sum(width(fasta))  
Number_of_contigs <- length(fasta)  
Average_len <- mean(width(fasta))  
Median_len <- median(width(fasta))  
Max_len <- max(width(fasta))  
Min_len <- min(width(fasta))  
  
#本番(N50情報取得)  
sorted <- rev(sort(width(fasta)))  
obj <- (cumsum(sorted) >= Total_len)  
N50 <- sorted[obj][1]
```

The screenshot shows an RStudio environment with a context menu open over the code. The menu items are:

- 切り取り(T)
- コピー(C)** (highlighted with a red arrow and circled with a red number 1)
- 貼り付け
- すべて選択(A)
- 印刷(I)...
- 印刷プレビュー(N)...
- Bing でマップ
- Bing で翻訳
- Google で検索
- 電子メール (Windows Live Hotmail)
- すべてのアクセラレータ
- Send to OneNote

The code itself is a script for basic genome statistics and N50 calculation.

## W9-5: コピペ

Rの画面上で①ペースト。ホスト - ゲスト間でコピペがうまくできないときは、②のfirefoxを起動してコードのコピーをすればよい。あるいは一旦Bio-Linuxを保存せずに終了して再起動とか

iu@bielinux[~/Documents/genomes]

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

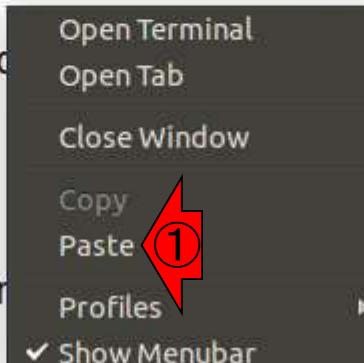
```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos  
'help.start()' for an HTML browser  
Type 'q()' to quit R.
```

```
> getwd()  
[1] "/home/iu/Documents/genomes"  
> list.files()  
[1] "Lactobacillus_casei_12a.GCA_000509565.1.22.dna.toplevel.fa"  
> 
```



# W9-6: コピペ後

エラーなく実行できたときの全貌(っていうか最後のほう)。①最後の行は、`write.table`という関数を用いてtmpの中身をout\_fで指定したファイルに保存するコマンド

```
File Edit View Search Terminal Help  
> #本番 (GC含量情報取得)  
> hoge <- alphabetFrequency(fasta)          #A,C,G,T,...の数を配列ごと  
にカウントした結果を hogeに格納  
> CG <- rowSums(hoge[,2:3])                  #C,Gの総数を計算してCGに  
格納  
> ACGT <- rowSums(hoge[,1:4])                #A,C,G,Tの総数を計算してA  
CGTに格納  
> GC_content <- sum(CG)/sum(ACGT)           #トータルのGC含量の情報を  
取得  
>  
> #ファイルに保存  
> tmp <- NULL  
> tmp <- rbind(tmp, c("Total length (bp)", Total_len))  
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))  
> tmp <- rbind(tmp, c("Average length", Average_len))  
> tmp <- rbind(tmp, c("Median length", Median_len))  
> tmp <- rbind(tmp, c("Max length", Max_len))  
> tmp <- rbind(tmp, c("Min length", Min_len))  
> tmp <- rbind(tmp, c("N50", N50))  
> tmp <- rbind(tmp, c("GC content", GC_content))  
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F  
, col.names=F) #tmpの中身を指定したファイル名で保存  
>
```

①

# W9-6: コピペ後

それゆえ、out\_fで指定した出力ファイル(result\_JALAB1.txt)をイチイチ開いて確認しなくても、①tmpと打って、tmpの中身をR画面上で確認するのもよい。ただの復習

```
File Edit View Search Terminal Help  
> #本番 (GC含量情報取得)  
> hoge <- alphabetFrequency(fasta)          #A,C,G,T,...の数を配列ごと  
にカウントした結果を hogeに格納  
> CG <- rowSums(hoge[,2:3])                  #C,Gの総数を計算してCGに  
格納  
> ACGT <- rowSums(hoge[,1:4])                #A,C,G,Tの総数を計算してACGTに格納  
> GC_content <- sum(CG)/sum(ACGT)           #トータルのGC含量の情報を  
取得  
>  
> #ファイルに保存  
> tmp <- NULL  
> tmp <- rbind(tmp, c("Total length (bp)", Total_len))  
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))  
> tmp <- rbind(tmp, c("Average length", Average_len))  
> tmp <- rbind(tmp, c("Median length", Median_len))  
> tmp <- rbind(tmp, c("Max length", Max_len))  
> tmp <- rbind(tmp, c("Min length", Min_len))  
> tmp <- rbind(tmp, c("N50", N50))  
> tmp <- rbind(tmp, c("GC content", GC_content))  
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F  
, col.names=F) #tmpの中身を指定したファイル名で保存  
> tmp
```

①

①tmpの中身を表示させた結果。これが「連載第1回の図2」と基本的に同じもの

# W9-6: コピペ後

```
File Edit View Search Terminal Help ↑ ↓ Ja 19:40 ⚡  
x - +    19:40 ⚡  
  
> #ファイルに保存  
> tmp <- NULL  
> tmp <- rbind(tmp, c("Total length (bp)", Total_len))  
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))  
> tmp <- rbind(tmp, c("Average length", Average_len))  
> tmp <- rbind(tmp, c("Median length", Median_len))  
> tmp <- rbind(tmp, c("Max length", Max_len))  
> tmp <- rbind(tmp, c("Min length", Min_len))  
> tmp <- rbind(tmp, c("N50", N50))  
> tmp <- rbind(tmp, c("GC content", GC_content))  
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F  
, col.names=F) #tmpの中身を指定したファイル名で保存  
> tmp  
[1] [2]  
[1,] "Total length (bp)" "2885619"  
[2,] "Number of contigs" "28"  
[3,] "Average length" "103057.821428571"  
[4,] "Median length" "58047"  
[5,] "Max length" "472701"  
[6,] "Min length" "899"  
[7,] "N50" "222389"  
[8,] "GC content" "0.464045193416998"  
>
```



# ほらね

File Edit View Search Terminal Help

R Console

```

> #ファイルに保存
> tmp <- NULL
> tmp <- rbind(tmp, c("Total length", "Total_len"))
> tmp <- rbind(tmp, c("Number of contigs", "Number_of_contigs"))
> tmp <- rbind(tmp, c("Average length", "Average_len"))
> tmp <- rbind(tmp, c("Median length", "Median_len"))
> tmp <- rbind(tmp, c("Max length", "Max_len"))
> tmp <- rbind(tmp, c("Min length", "Min_len"))
> tmp <- rbind(tmp, c("N50", "N50"))
> tmp <- rbind(tmp, c("GC content", "GC_content"))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F)
, col.names=F) #tmpの中身を指定します
> tmp
     [,1] [,2]
[1,] "Total length (bp)" "2885619"
[2,] "Number of contigs" "28"
[3,] "Average length"    "103057.821428571"
[4,] "Median length"     "58047"
[5,] "Max length"        "472701"
[6,] "Min length"        "899"
[7,] "N50"                "222389"
[8,] "GC content"         "0.464045193416998"
>

```



① 図2. Rでゲノム配列解析。Rコード実行結果のスクリーンショットを示している。

解析するという手間が軽減されることが期待される。Expression Atlasと似たウェブツールであるDBCLSの今後の活動に期待したい。

原著論文<sup>36)</sup>のTable 1の記載内容(コンタル塩基数: 2,885,619 bp、%GC: 46.4)と一致していることがわかる。また、最長コ

①確かにW9-3のkopペ前には存在しなかつたresult\_JSLAB1.txtがある

# W9-7: 存在確認

```
File Edit View Search Terminal Help Ja 19:50
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))
> tmp <- rbind(tmp, c("Average length", Average_len))
> tmp <- rbind(tmp, c("Median length", Median_len))
> tmp <- rbind(tmp, c("Max length", Max_len))
> tmp <- rbind(tmp, c("Min length", Min_len))
> tmp <- rbind(tmp, c("N50", N50))
> tmp <- rbind(tmp, c("GC content", GC_content))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F
, col.names=F)#tmpの中身を指定したファイル名で保存
> tmp
 [,1] [,2]
[1,] "Total length (bp)" "2885619"
[2,] "Number of contigs" "28"
[3,] "Average length" "103057.821428571"
[4,] "Median length" "58047"
[5,] "Max length" "472701"
[6,] "Min length" "899"
[7,] "N50" "222389"
[8,] "GC content" "0.464045193416998"
> list.files()
[1] "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa"
[2] "result_JSLAB1.txt"
>
```



①Rを終了させて、②lsで存在確認しているだけです

# W9-7: 存在確認

```
File Edit View Search Terminal Help Ja 19:53
> tmp <- rbind(tmp, c("N50", N50))
> tmp <- rbind(tmp, c("GC content", GC_content))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F
, col.names=F) #tmpの中身を指定したファイル名で保存
> tmp
     [,1]          [,2]
[1,] "Total length (bp)" "2885619"
[2,] "Number of contigs" "28"
[3,] "Average length"    "103057.821428571"
[4,] "Median length"     "58047"
[5,] "Max length"        "472701"
[6,] "Min length"        "899"
[7,] "N50"                "222389"
[8,] "GC content"         "0.464045193416998"
> list.files()
[1] "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa"
[2] "result_JSLAB1.txt"
> q()
Save workspace image? [y/n/c]: n
iu@bielinux[genomes] ls [ 7:53午後 ]
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] [ 7:53午後 ]
```

# W9-8: R起動時のTips

①R起動時に-qオプションをつけることで、スタートアップメッセージを非表示にすることができる。



# W9-8: R終了時のTips

①R終了時に「save="no"」オプションをつけることで、「Save workspace image? [y/n/c]」と毎回聞かれる苦行から解放される

The screenshot shows a terminal window with the following content:

```
iu@bielinux[genomes] R -q
> q(save="no")
iu@bielinux[genomes]
```

A red arrow labeled '①' points to the 'q(save="no")' command in the terminal.

The desktop environment includes a docked application bar at the top with icons for File, Edit, View, Search, Terminal, Help, and system status indicators like battery level and network. The system tray shows the date and time: [11:29午前] and [11:33午前].

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



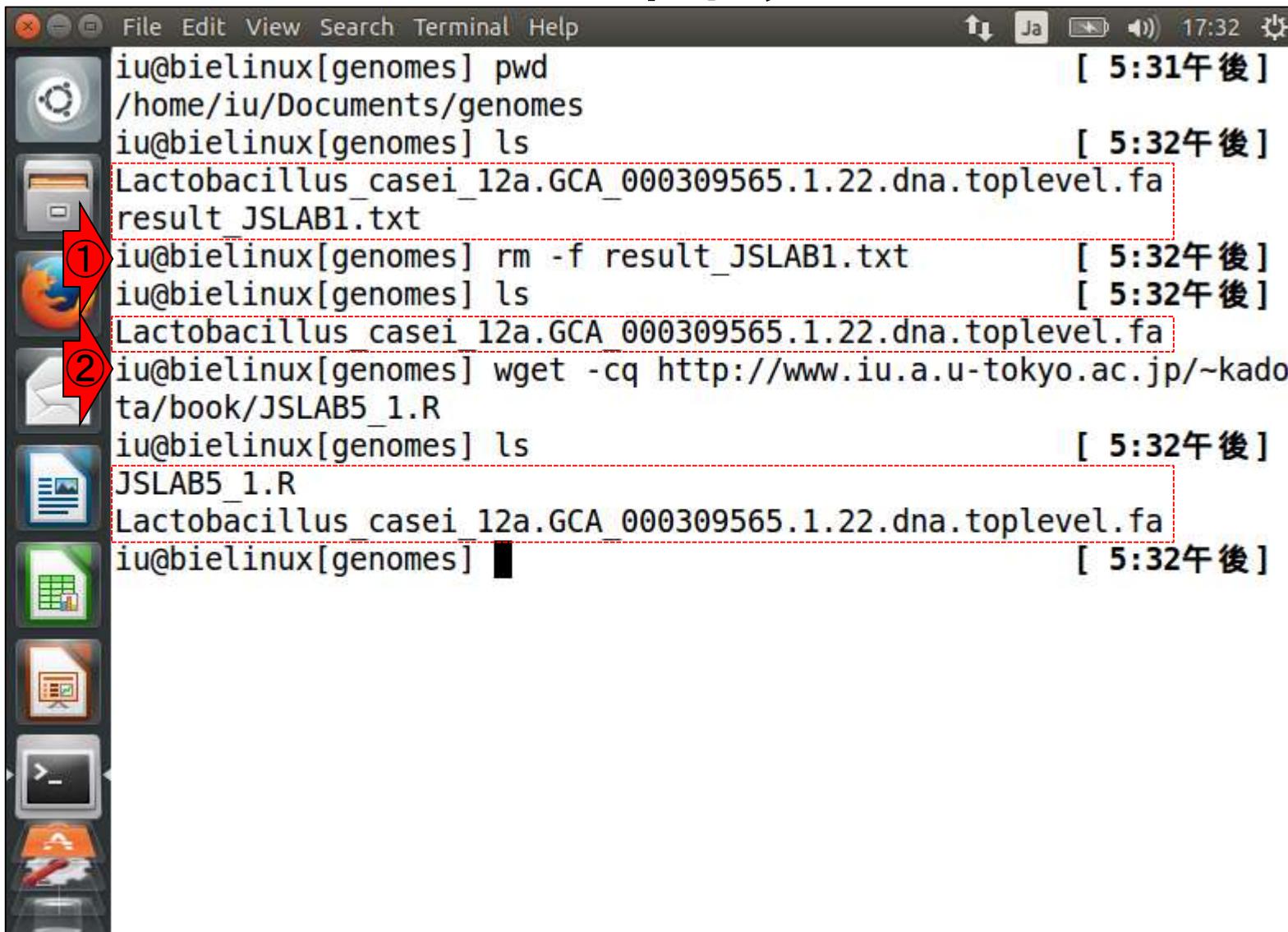
# wgetは基本やらない

```
File Edit View Search Terminal Help  
① iu@bielinux[backup] pwd  
/home/iu/Desktop/backup  
iu@bielinux[backup] ls  
DRR054113_163380_fastqc.html  
DRR054113.fastq.bz2  
fastaLengthFilter.py  
JSLAB5_1.R  
JSLAB5_2.R  
JSLAB5_3.R  
JSLAB5_4.txt  
JSLAB5_5.R  
JSLAB5_6.R  
JSLAB5_7.txt  
JSLAB5_8.R  
JSLAB6_1.R  
kmergenie-1.6982.tar.gz  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa.gz  
Lactobacillus_casei_12a.GCA_000309565.2.30.dna.toplevel.fa.gz  
platanusResult.zip  
QC.1.trimmed.fastq.gz  
QC.2.trimmed.fastq.gz  
result.zip
```

この後何回かwgetを利用しますが、基本的にはwgetしたつもりにしてください。実際に行うのは、①「~/Desktop/backup」上にある、対応するファイルのコピーを行ってください。大量同時アクセスにならないタイミングでひっそり自己責任でwgetするのは黙認します。<http://www.iu.a.u-tokyo.ac.jp>以下のアグリバイオサーバ上にあるものはwgetを試みてもOK

Rockhopper.jar  
test\_out.txt  
Trinity1.fasta  
Trinity2.fasta  
velvet\_1.2.10.tgz  
velvet.zip

# W10-1 : source関数



The screenshot shows a terminal window with a dark theme. On the left, there is a vertical dock with various icons, some of which have red arrows pointing to them. The terminal window has a title bar with the text "File Edit View Search Terminal Help". The status bar at the top right shows the date and time as "17:32". The terminal history is as follows:

```
iu@bielinux[genomes] pwd [ 5:31午後]
/home/iu/Documents/genomes
iu@bielinux[genomes] ls [ 5:32午後]
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
①
iu@bielinux[genomes] rm -f result_JSLAB1.txt [ 5:32午後]
iu@bielinux[genomes] ls [ 5:32午後]
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
②
iu@bielinux[genomes] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kado
ta/book/JSLAB5_1.R
iu@bielinux[genomes] ls [ 5:32午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] [ 5:32午後]
```

Red boxes highlight the file "result\_JSLAB1.txt" and the command "rm -f result\_JSLAB1.txt". Red numbers ① and ② point to the "result\_JSLAB1.txt" file and the "wget" command respectively.

# W10-1 : source関数

①headで最初の5行分を表示。文字化けしているが、結果に影響しないコメント部分なので、ここでは気にしない

```
iu@bielinux[genomes] pwd [ 5:31午後]
/home/iu/Documents/genomes
iu@bielinux[genomes] ls [ 5:32午後]
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] rm -f result_JSLAB1.txt [ 5:32午後]
iu@bielinux[genomes] ls [ 5:32午後]
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kada/book/JSLAB5_1.R
iu@bielinux[genomes] ls [ 5:32午後]
JSLAB5_1.R
iu@bielinux[genomes] head -n 5 JSLAB5_1.R [ 5:32午後]
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.
fa" #000t@0C00000000軒 00in_f0Qi0[
out_f <- "result_JSLAB1.txt" #0t0@0C00000000軒 00out_f0Qi
#[

#KôvôâpôbôPô[ôWôôôôôô[ôh
library(Biostrings) #ôpôbôPô[ôWôôôôôô
iu@bielinux[genomes] [ 5:33午後]
```

# W10-1 : source関数

```
File Edit View Search Terminal Help [ 17:34 ]  
iu@bielinux[genomes] pwd [ 5:31 午後 ]  
/home/iu/Documents/genomes  
iu@bielinux[genomes] ls [ 5:32 午後 ]  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
result_JSLAB1.txt  
iu@bielinux[genomes] rm -f result_JSLAB1.txt [ 5:32 午後 ]  
iu@bielinux[genomes] ls [ 5:32 午後 ]  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
iu@bielinux[genomes] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kado  
ta/book/JSLAB5_1.R  
iu@bielinux[genomes] ls [ 5:32 午後 ]  
JSLAB5_1.R  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
iu@bielinux[genomes] head -n 5 JSLAB5_1.R [ 5:32 午後 ]  
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.  
fa" # 亂数  
out_f <- "result_JSLAB1.txt" # 亂数  
#[  
  
# 亂数  
library(Biostrings)  
iu@bielinux[genomes] R -q [ 5:33 午後 ]  
>
```



# W10-1 : source関数

目的は「source("JSLAB5\_1.R")」と打ち込むこと。Rの対話モードでもタブ補完が有効なので、①のような状態でTabキーを押すと…



```
iu@bielinux[genomes] R -q
> source("JS")
```

# W10-1 :source関数

The screenshot shows a Linux desktop environment. On the left, there is a vertical docked application bar with various icons. The top icon, which is a terminal window, has a red arrow pointing to it with the number '1' in a red circle. The terminal window title bar says 'iu@bielinux[genomes] R -q'. The main workspace contains a terminal window with the command 'source("JSLAB5\_1.R")' entered. The status bar at the bottom of the terminal window shows '[ 5:35午後 ]'.

```
iu@bielinux[genomes] R -q
> source("JSLAB5_1.R")
```

# W10-2: 実行結果

```
File Edit View Search Terminal Help 17:39  
int,  
  rownames, sapply, setdiff, sort, table, tapply, union, unique  
,  
  unlist, unsplit  
  
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in pa  
ckage 'S4Vectors'  
Loading required package: IRanges  
Loading required package: XVector  
Warning messages:  
1: In grepl("\n", lines, fixed = TRUE) :  
  input string 1 is invalid in this locale  
2: In grepl("\n", lines, fixed = TRUE) :  
  input string 2 is invalid in this locale  
3: In grepl("\n", lines, fixed = TRUE) :  
  input string 4 is invalid in this locale  
4: In grepl("\n", lines, fixed = TRUE) :  
  input string 5 is invalid in this locale  
5: In grepl("\n", lines, fixed = TRUE) :  
  input string 7 is invalid in this locale  
> ■
```

①

# W10-3：確認

①list.files()で確認。確かに出力ファイルとして指定したresult\_JSLAB1.txtが生成されている。②Rを終了。③lsで念のため確認。④moreでファイルの中身を表示。妥当な結果である

```
File Edit View Search Terminal Help
1 4: In grepl("\n", lines, fixed = TRUE) :
  input string 5 is invalid in this locale
2 5: In grepl("\n", lines, fixed = TRUE) :
  input string 7 is invalid in this locale
> list.files()
[1] "JSLAB5_1.R"
[2] "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa"
[3] "result_JSLAB1.txt"
> q(save="no")
iu@bielinux[genomes] ls [ 5:44午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] more result_JSLAB1.txt [ 5:44午後]
Total length (bp)      2885619
Number of contigs       28
Average length          103057.821428571
Median length           58047
Max length              472701
Min length              899
N50                     222389
GC content              0.464045193416998
iu@bielinux[genomes] [ 5:45午後]
```

# W10-4: Rのバージョン確認

The screenshot shows a terminal window titled "iu@bielinux[genomes]". The window contains the following text:

```
File Edit View Search Terminal Help
iu@bielinux[genomes] R --version
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Comp
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY
You are welcome to redistribute it under the terms of the
GNU General Public License versions 2 or 3.
For more information about these matters see
http://www.gnu.org/licenses/.
```

On the left side of the terminal window, there is a vertical dock with various icons, including a terminal icon with a red arrow labeled "①" pointing to it.

Below the terminal window, there is another terminal window with the following text:

```
iu@bielinux[genomes] R --version | grep "version" [11:58午前]
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
GNU General Public License versions 2 or 3.
iu@bielinux[genomes] ■ [12:06午後]
```

On the left side of this second terminal window, there is a vertical dock with various icons, including a terminal icon with a red arrow labeled "②" pointing to it.

①「R --version」と打つことで、Rを起動することなくバージョン確認をすることができる。正確には一旦起動してすぐに終了しているようだが、エンドユーザからすれば起動していないのと同じ。②grepを併用することで、さらに表示結果から"version"を含む行のみに限定させることができる

# W10-5: バッチモード

①一旦 result\_JSLAB1.txt を削除。②lsで出力予定ファイル(result\_JSLAB1.txt)がないことを確認して、③バッチモードの基本形を実行

```
File Edit View Search Terminal Help
iu@bielinux[genomes] pwd [ 5:51午後]
/home/iu/Documents/genomes
iu@bielinux[genomes] ls [ 5:51午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] rm -f result_JSLAB1.txt ① [ 5:51午後]
iu@bielinux[genomes] ls [ 5:51午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] R --vanilla < JSLAB5_1.R [ 5:51午後]
```

The screenshot shows a terminal window with a command-line session. The terminal window has a dark theme with white text. The application bar at the top includes icons for file operations, search, terminal, and help, along with system status indicators like battery level and time (17:52). A vertical dock on the left contains icons for various applications, including a browser (Firefox), email, and file management tools. Red numbered arrows (2 and 3) point to the Firefox icon and the file management icons respectively, likely indicating steps in a workflow. The terminal session shows the user navigating to a directory, listing files, removing a specific file named 'result\_JSLAB1.txt', and then running an R script from a file named 'JSLAB5\_1.R'.

# W10-5: バッチモード

①リターンを押して実行した結果。確かにq()などでRを終了させる手間なく、通常のLinuxコマンド入力待ち状態になっていることがわかる

```
File Edit View Search Terminal Help
> #0{00(GC0@Y00000)
> hoge <- alphabetFrequency(fasta)      #A,C,G,T,...00000Z0EBJ0E
000g00000000$0hoge@Qi0[
> CG <- apply(as.matrix(hoge[,2:3]), 1, sum)#C,G000000v000CG0Qi
0[
> ACGT <- apply(as.matrix(hoge[,1:4]), 1, sum)#A,C,G,T0000000v000
0ACGT0Qi0[
> GC_content <- sum(CG)/sum(ACGT)        #0g0[0^0000GC0@S0000000
>
> #0t0@0C000c,0
> tmp <- NULL
> tmp <- rbind(tmp, c("Total length (bp)", Total_len))
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))
> tmp <- rbind(tmp, c("Average length", Average_len))
> tmp <- rbind(tmp, c("Median length", Median_len))
> tmp <- rbind(tmp, c("Max length", Max_len))
> tmp <- rbind(tmp, c("Min length", Min_len))
> tmp <- rbind(tmp, c("N50", N50))
> tmp <- rbind(tmp, c("GC content", GC_content))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=
F, col.names=F)#tmp000g000貯000t0@0C00000r,0
>
iu@bielinux[genomes] [ 5:52午後]
```

# W10-5：確認

①lsで確認。確かにresult\_JSLAB1.txtが出力結果として得られており、②その中身も妥当

```
File Edit View Search Terminal Help ↑ ↓ Ja ⌂ 17:54 ⌂
> tmp <- rbind(tmp, c("Average length", Average_len))
> tmp <- rbind(tmp, c("Median length", Median_len))
> tmp <- rbind(tmp, c("Max length", Max_len))
> tmp <- rbind(tmp, c("Min length", Min_len))
> tmp <- rbind(tmp, c("N50", N50))
> tmp <- rbind(tmp, c("GC content", GC_content))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=
F, col.names=F) #tmp@bielinux[genomes] ls [ 5:52 午後 ]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] more result_JSLAB1.txt [ 5:53 午後 ]
Total length (bp) 2885619
Number of contigs 28
Average length 103057.821428571
Median length 58047
Max length 472701
Min length 899
N50 222389
GC content 0.464045193416998
iu@bielinux[genomes] [ 5:53 午後 ]
```

# W11：文字化け対策

```
File Edit View Search Terminal Help
iu@bielinux[genomes] head -n 5 JSLAB5_1.R
in_f <- "Lactobacillus_casei_12a.GCA_0003095
fa">#??t@?C?????@?in_f@Qi@[
out_f <- "result_JSLAB1.txt" #@t@[
#[
library(Biostrings) #@p@b@P@[@W@?+v@]
iu@bielinux[genomes] file JSLAB5_1.R [ 5:57 午後]
JSLAB5_1.R: Non-ISO extended-ASCII text
iu@bielinux[genomes] nkf JSLAB5_1.R > hoge.R [ 5:57 午後]
iu@bielinux[genomes] █ [ 5:57 午後]
```

気になるヒト用に文字化け対策の復習。連載第4回W13-6にも記載あり。nkfのインストールができない人は第4回W13-5を参照して自己でインストール。  
①JSLAB5\_1.Rの最初の5行分を表示。  
②fileコマンドで文字コードを確認。  
赤下線のような見たこともない形式のようだが、  
③nkfが自動認識してくれるので気にもしない

# W11: 文字化け対策

```
File Edit View Search Terminal Help
iu@bielinux[genomes] head -n 5 JSLAB5_1.R
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.
fa"#入力ファイル名を指定してin_fに格納
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してout_fに格納
[

#必要なパッケージをロード
library(Biostrings) #パッケージの読み込み
iu@bielinux[genomes] file JSLAB5_1.R [ 5:57午後]
JSLAB5_1.R: Non-ISO extended-ASCII text
iu@bielinux[genomes] nkf JSLAB5_1.R > hoge.R [ 5:57午後]
iu@bielinux[genomes] head -n 5 hoge.R [ 5:57午後]
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.
fa"#入力ファイル名を指定してin_fに格納
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Biostrings) #パッケージの読み込み
iu@bielinux[genomes] file hoge.R [ 5:59午後]
hoge.R: UTF-8 Unicode text
iu@bielinux[genomes] [ 5:59午後]
```

④nkf実行後のファイルhoge.Rの最初の5行分を表示。確かに文字化けが解消されていることがわかる。⑤hoge.Rの文字コードは「UTF-8 Unicode text」というものらしいが、判読できていればよいので、特に気にしない

# Contents

- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W12-1：発展形1

①--slaveオプションをつけて実行させると、実行中に画面表示されるものが減るので、多少見づらさが緩和される

The screenshot shows a Linux desktop environment with a terminal window open in the foreground. The terminal window has a dark background and contains the following command-line session:

```
iu@bielinux[genomes] pwd  
/home/iu/Documents/genomes  
iu@bielinux[genomes] ls  
hoge.R  
JSLAB5_1.R  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
result_JSLAB1.txt  
iu@bielinux[genomes] rm -f hoge.R result_JSLAB1.txt [ 6:01午後]  
iu@bielinux[genomes] ls [ 6:01午後]  
JSLAB5_1.R  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
iu@bielinux[genomes] R --vanilla --slave < JSLAB5_1.R
```

A red arrow points to the icon for the application 'R' in the docked application bar on the left side of the screen. The number '1' is circled in red above this arrow.

# W12-1: 発展形1

```
File Edit View Search Terminal Help 18:03  
xtabs  
  
The following objects are masked from 'package:base':  
  
  anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,  
  do.call, duplicated, eval, evalq, Filter, Find, get, intersect,  
  is.unsorted, lapply, Map, mapply, match, mget, order, paste,  
  pmax,  
  pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,  
  rownames, sapply, setdiff, sort, table, tapply, union, unique  
,  
  unlist, unsplit  
  
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in package 'S4Vectors'  
Loading required package: IRanges  
Loading required package: XVector  
iu@bielinux[genomes] [ 6:03午後 ]
```

①

# W12-1: 発展形1

```
File Edit View Search Terminal Help 18:04
anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
is.unsorted, lapply, Map, mapply, match, mget, order, paste,
pmax,
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,
rownames, sapply, setdiff, sort, table, tapply, union, unique
,
unlist, unsplit

Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in package 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
iu@bielinux[genomes] ls [ 6:03午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
result_JSLAB1.txt
iu@bielinux[genomes] [ 6:04午後]
```

①

# W12-2: 発展形2

```
File Edit View Search Terminal Help  
iu@bielinux[genomes] rm -f result_JSLAB1.txt  
iu@bielinux[genomes] ls  
JSLAB5_1.R  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
iu@bielinux[genomes] nkf JSLAB5_1.R | head -n2 [ 4:09午後 ]  
in_f <- "Lactobacillus casei 12a.GCA 000309565.1.22.dna.toplevel.f  
a"#入力ファイル名を指定してin_fに格納  
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してou  
t_fに格納  
iu@bielinux[genomes] pwd [ 4:09午後 ]  
/home/iu/Documents/genomes  
iu@bielinux[genomes] [ 4:09午後 ]
```

①JSLAB5\_1.Rの最初の2行分を表示。  
nkf実行結果とパイプさせているのは、  
文字化け対策。JSLAB5\_1.Rが正常動作  
するのは、②作業ディレクトリ上に、in\_f  
で指定した入力ファイルが存在するから

# W12-2: 発展形2

- ①result2ディレクトリにJSLAB5\_1.Rをコピー。  
②移動先にはJSLAB5\_1.Rが入力として読み込む乳酸菌ゲノムファイルは存在しない。③result2上でJSLAB5\_1.Rを実行してみると…

```
iu@bielinux[genomes] rm -f result_JSLAB1.txt [ 4:09午後]
iu@bielinux[genomes] ls [ 4:09午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] nkf JSLAB5_1.R | head -n2 [ 4:09午後]
in_f <- "Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.f
a"#入力ファイル名を指定してin_fに格納
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してou
t_fに格納
iu@bielinux[genomes] pwd [ 4:09午後]
/home/iu/Documents/genomes
iu@bielinux[genomes] cp JSLAB5_1.R ~/Documents/srp017156/result2
iu@bielinux[genomes] cd ~/Documents/srp017156/result2 [ 4:21午後]
iu@bielinux[result2] pwd [ 4:21午後]
/home/iu/Documents/srp017156/result2
iu@bielinux[result2] ls [ 4:21午後]
fastqCount.txt QC.1.trimmed.fastq QC.stats.txt
JSLAB5_1.R QC.2.trimmed.fastq QC.unpaired.trimmed.fastq
nohup.out QC_qc_report.pdf Rockhopper_Results
iu@bielinux[result2] R --vanilla --slave < JSLAB5_1.R [ 4:21午後]
```

①cannot open file…や②Execution haltedというネガティブなメッセージからも、実行失敗の想像がつく

## W12-2: 発展形2

```
File Edit View Search Terminal Help Ja 16:39
do.call, duplicated, eval, evalq, Filter, Find, get, intersect
,
is.unsorted, lapply, Map, mapply, match, mget, order, paste, p
max,
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.i
nt,
rownames, sapply, setdiff, sort, table, tapply, union, unique,
unlist, unsplit

Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in pac
kage 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
Error in .Call2("new_input_ExternalFilePtr", fp, PACKAGE = "Biostr
ings") :
  cannot open file 'Lactobacillus_casei_12a.GCA_000309565.1.22.dna
.toplevel.fa'
Calls: readDNAStringSet ... .open_input_files -> lapply -> lapply
-> FUN -> .Call2 -> .Call
Execution halted
iu@bielinux[result2] [ 4:32午後 ]
```

①

②

# W12-2: 発展形2

File Edit View Search Terminal Help

unlist, unsplit

```
Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in pac
kage 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
Error in .Call2("new_input_ExternalFilePtr", fp, PACKAGE = "Biostr
ings") :
  cannot open file 'Lactobacillus_casei_12a.GCA_000309565.1.22.dna
.toplevel.fa'
Calls: readDNAStringSet ... .open_input_files -> lapply -> lapply
-> FUN -> .Call2 -> .Call
Execution halted
iu@bielinux[result2] ls [ 4:32 午後]
fastqCount.txt QC.1.trimmed.fastq QC.stats.txt
JSLAB5_1.R QC.2.trimmed.fastq QC.unpaired.trimmed.fastq
nohup.out QC_qc_report.pdf Rockhopper_Results
iu@bielinux[result2] ls ~/Documents/genomes [ 4:48 午後]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[result2] [ 4:52 午後]
```

③lsした結果。result\_JSLAB1.txtが生成され  
ていないことがわかる。④当然  
~/Documents/genomes上にもない。理由は  
シンプル。JSLAB5\_1.Rは、入力ファイルをカ  
レントディレクトリ上でのみ探索しているから



# W12-3: 発展形3

①wgetでJSLAB5\_2.Rをダウンロード。②最初の2行分を表示。赤下線で示すように、入力ファイルを絶対パスで指定している。  
③JSLAB5\_2.Rをバッチモードで実行

```
iu@bielinux[result2] pwd [ 7:04 午後]
/home/iu/Documents/srp017156/result2
iu@bielinux[result2] ls [ 7:05 午後]
fastqCount.txt QC.1.trimmed.fastq QC.stats.txt
JSLAB5_1.R QC.2.trimmed.fastq QC.unpaired.trimmed.fastq
nohup.out QC_qc_report.pdf Rockhopper_Results
iu@bielinux[result2] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kadota/book/JSLAB5_2.R
iu@bielinux[result2] ls [ 7:05 午後]
fastqCount.txt QC.1.trimmed.fastq QC.unpaired.trimmed.fastq
JSLAB5_1.R QC.2.trimmed.fastq Rockhopper_Results
JSLAB5_2.R QC_qc_report.pdf
nohup.out QC.stats.txt
iu@bielinux[result2] nkf JSLAB5_2.R | head -n 2 [ 7:05 午後]
in_f <- "/home/iu/Documents/genomes/Lactobacillus_casei_12a.GCA_00
0309565.1.22.dna.toplevel.fa" #入力ファイル名を指定してin_fに格納
out_f <- "result_JSLAB1.txt" #出力ファイル名を指定してout_fに格納
iu@bielinux[result2] R --vanilla --slave < JSLAB5_2.R [ 7:05 午後]
```



# W12-3: 発展形3

```
xtabs  
  
The following objects are masked from 'package:base':  
  
  anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,  
  do.call, duplicated, eval, evalq, Filter, Find, get, intersect  
,  
  is.unsorted, lapply, Map, mapply, match, mget, order, paste, p  
max,  
  pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.i  
nt,  
  rownames, sapply, setdiff, sort, table, tapply, union, unique,  
  unlist, unsplit  
  
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in pac  
kage 'S4Vectors'  
Loading required package: IRanges  
Loading required package: XVector  
iu@bielinux[result2] [ 7:19午後 ]
```

①lsで確認。確かに出力ファイルが存在し、②その中身も正しい

# W12-3: 発展形3

```
File Edit View Search Terminal Help  
unlist, unsplit  
  
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in pac  
kage 'S4Vectors'  
Loading required package: IRanges  
Loading required package: XVector  
iu@bielinux[result2] ls [ 7:19午後]  
fastqCount.txt QC.1.trimmed.fastq QC.unpaired.trimmed.fastq  
JSLAB5_1.R QC.2.trimmed.fastq result_JSLAB1.txt  
JSLAB5_2.R QC_qc_report.pdf Rockhopper_Results  
nohup.out QC.stats.txt  
iu@bielinux[result2] more result_JSLAB1.txt [ 7:21午後]  
Total length (bp) 2885619  
Number of contigs 28  
Average length 103057.821428571  
Median length 58047  
Max length 472701  
Min length 899  
N50 222389  
GC content 0.464045193416998  
iu@bielinux[result2] [ 7:21午後]
```

# Contents

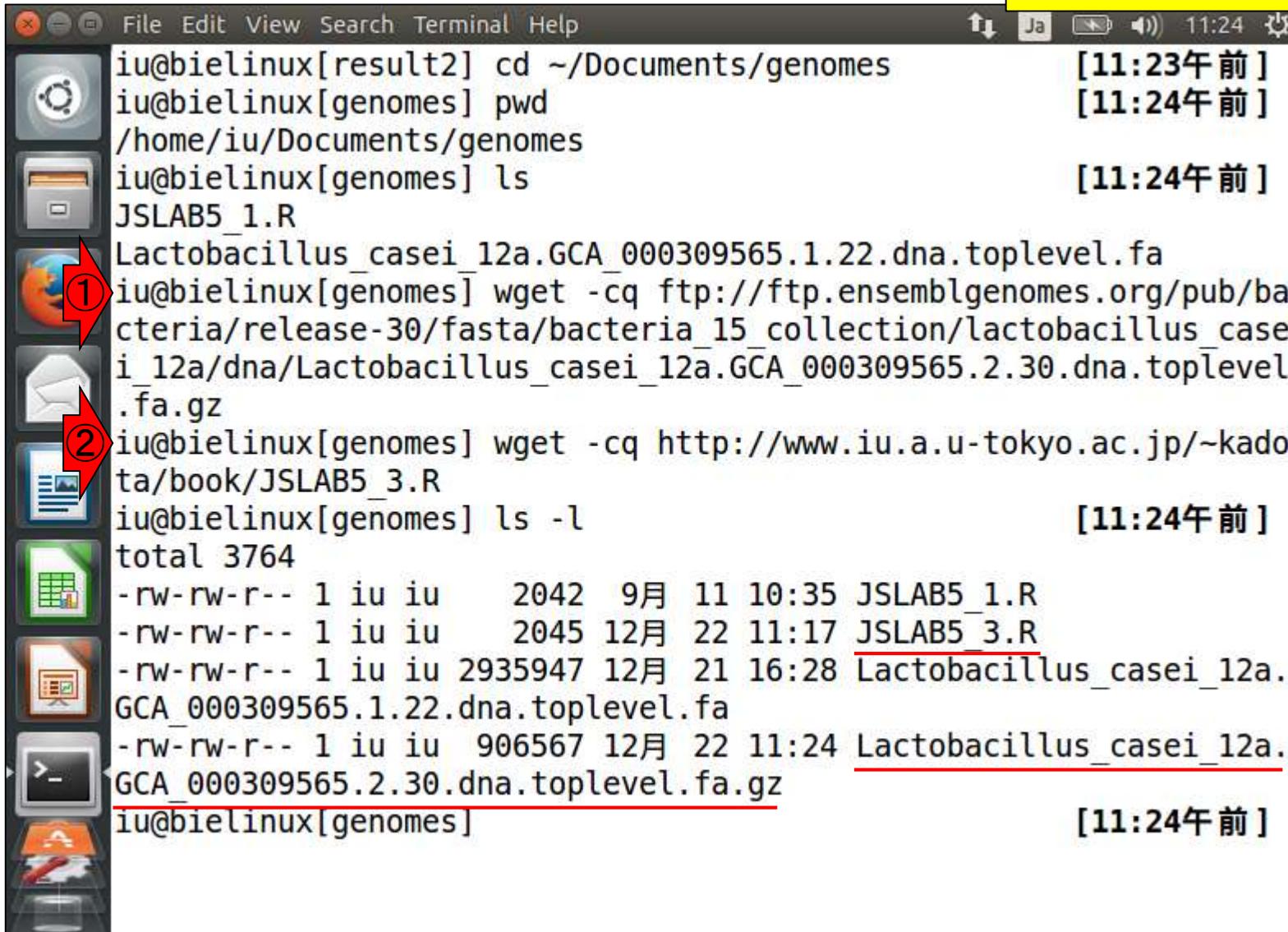
- 日本乳酸菌学会誌のNGS連載第4回までの復習(特にFastQCとFaQCs)
  - まずはFaQCs実行、おさらい、FastQCでIllumina adapterの消滅確認
- Javaプログラムの設定と実行(Rockhopper2)
  - W2: Javaの確認とダウンロード、GUI版の実行
  - W3: Linux Tips (&, ps, kill, and nohup)
  - W4とW5:コマンドライン版の実行(paired-end)、クラスパスの設定、再実行
  - W6:コマンドライン版の実行(single-end)
- Linux環境でのRの利用法
  - W7:起動と終了、QuasRパッケージのインストール(エアーハンズオン)
  - W8:R基本コマンド、W9:乳酸菌ゲノム配列取得と基本情報取得(連載第1回の図2)
  - W10:source関数、バッチモードでの利用
  - W12:バッチモードでの利用の発展形、入力ファイルの絶対パス指定
  - W13:gzip圧縮状態での利用



# W13-1：最新版で解析

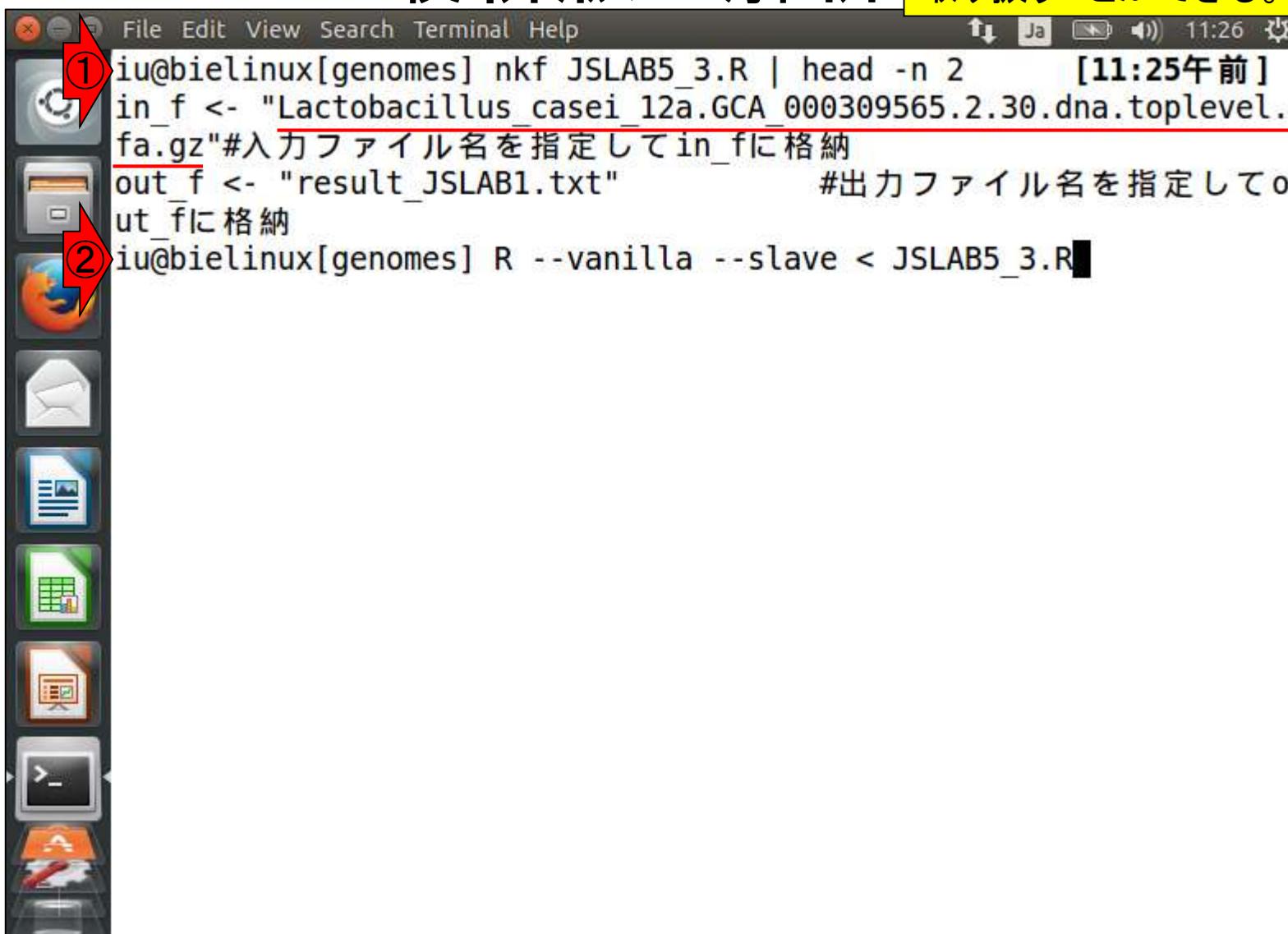
①ゲノムファイルと②Rスクリプトファイルのダウンロード。赤下線で示すように、正しく取得できてるっぽいことがわかる

```
File Edit View Search Terminal Help Ja 11:24
iu@bielinux[result2] cd ~/Documents/genomes [11:23午前]
iu@bielinux[genomes] pwd [11:24午前]
/home/iu/Documents/genomes
iu@bielinux[genomes] ls [11:24午前]
JSLAB5_1.R
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
iu@bielinux[genomes] wget -cq ftp://ftp.ensemblgenomes.org/pub/bacteria/release-30/fasta/bacteria_15_collection/lactobacillus_casei_12a/dna/Lactobacillus_casei_12a.GCA_000309565.2.30.dna.toplevel.fa.gz
iu@bielinux[genomes] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kodata/book/JSLAB5_3.R
iu@bielinux[genomes] ls -l [11:24午前]
total 3764
-rw-rw-r-- 1 iu iu 2042 9月 11 10:35 JSLAB5_1.R
-rw-rw-r-- 1 iu iu 2045 12月 22 11:17 JSLAB5_3.R
-rw-rw-r-- 1 iu iu 2935947 12月 21 16:28 Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa
-rw-rw-r-- 1 iu iu 906567 12月 22 11:24 Lactobacillus_casei_12a.GCA_000309565.2.30.dna.toplevel.fa.gz
iu@bielinux[genomes] [11:24午前]
```



# W13-1：最新版で解析

①Rスクリプトファイルの最初の2行分を表示。  
赤下線で示すようにgzip圧縮ファイルのままで  
取り扱うことができる。②JSLAB5\_3.Rを実行



```
File Edit View Search Terminal Help [11:25午前]
iu@bielinux[genomes] nkf JSLAB5_3.R | head -n 2
in_f <- "Lactobacillus_casei_12a.GCA_000309565.2.30.dna.toplevel.
fa.gz"#入力ファイル名を指定してin_fに格納
out_f <- "result_JSLAB1.txt"          #出力ファイル名を指定してo
ut_fに格納
iu@bielinux[genomes] R --vanilla --slave < JSLAB5_3.R
```

# W13-1 : 最新版で解析

```
File Edit View Search Terminal Help ↑ ↓ Ja 11:28 ⚙
xtabs

The following objects are masked from 'package:base':
anyDuplicated, append, as.data.frame, as.vector, cbind, colnames,
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
is.unsorted, lapply, Map, mapply, match, mget, order, paste,
pmax,
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,
rownames, sapply, setdiff, sort, table, tapply, union, unique,
unlist, unsplit

Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in package 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
iu@bielinux[genomes] [11:27午前]
```

# W13-1：最新版で解析

①lsで確認。出力ファイルresult\_JSLAB1.txtが確かにできている。②moreで中身を表示。1 contig、2,907,892 bpであることがわかる

```
File Edit View Search Terminal Help  
↑ ↓ Ja 11:28  
  
Loading required package: S4Vectors  
Loading required package: stats4  
Creating a generic function for 'nchar' from package 'base' in pa  
ckage 'S4Vectors'  
Loading required package: IRanges  
Loading required package: XVector  
iu@bielinux[genomes] ls [11:27午前]  
JSLAB5_1.R  
JSLAB5_3.R  
Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa  
Lactobacillus_casei_12a.GCA_000309565.2.30.dna.toplevel.fa.gz  
result_JSLAB1.txt  
iu@bielinux[genomes] more result_JSLAB1.txt [11:28午前]  
Total length (bp) 2907892  
Number of contigs 1  
Average length 2907892  
Median length 2907892  
Max length 2907892  
Min length 2907892  
N50 2907892  
GC content 0.464441595492542  
iu@bielinux[genomes] [11:28午前]
```

① *L. casei* 12Aの詳細情報はここからみられる。② wgetでgzip圧縮FASTA形式ファイル取得する際のURL情報はここからゲットできます。スライドを見るだけ

# W13-2: Ensembl

**EnsemblBacteria** ▾ Sequence Search | BLAST | Tools | More ▾ Login/Register

Lactobacillus casei 12A

## Lactobacillus casei 12A

Lactobacillus casei 12A

Provider European Nucleotide Archive | Taxonomy ID 1051650

Search Lactobacillus casei 12A...

e.g. spaB or Chromosome:502057-502915 or synthetase

### About Lactobacillus casei 12A

**Information and statistics**

Genome assembly: [GCA\\_000304565.2](#)  ①

More information and statistics  ②

Download DNA sequence (FASTA)

Display Ensembl Bacteria

**Gene annotation**

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

phnl metC hisA hpt lacZ accD Example gene

More about this genebuild

Download genes, cDNAs, ncRNA, proteins - FASTA - GFF3

Update your old Ensembl IDs

**Example transcript**

**Comparative genomics**

What can I find? Gene families based on HAMAP and PANTHER classification.

More about comparative analyses

**Variation**

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

Variant Effect Predictor 

About Ensembl Genomes | Contact Us | EMBL-EBI Terms of use | Privacy | Cookies | Help

Ensembl Bacteria release 28 - August 2015 © EBI

# W13-2: Ensembl

①をクリックした結果。連載第1回当時はコンティグレベルだったが、第5回執筆時には②染色体レベルになっていることがわかる。③トータル塩基数は2,907,892 bp

Lactobacillus casei 12A

**Lactobacillus casei 12A**

Lactobacillus casei 12A

Provider European Nucleotide Archive | Taxonomy ID 1051650

Search Lactobacillus casei 12A... Go

e.g. spaB or Chromosome:502057-502915 or synthetase

**About Lactobacillus casei 12A**

i Information and statistics

Genome assembly: GCA\_000304565.2 ①

i More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl Bacteria

View karyotyp

Example region

**Comparative genomics**

What can I find? Gene families based on HAMAP and PANTHER classification.

i More about comparative analyses

Ensembl Bacteria release 28 - August 2015 © EBI

**Lactobacillus casei 12A Assembly and Gene Annotation**

**Lactobacillus casei 12A**

**Organism**

Taxonomy ID 1051650

Name Lactobacillus casei 12A  
Wikipedia

Aliases Lactobacillus casei str. 12A  
Lactobacillus casei strain 12A

Classification > cellular organisms  
> Bacteria  
> Firmicutes  
> Bacilli  
> Lactobacillales  
> Lactobacillaceae  
> Lactobacillus  
> Lactobacillus casei group  
> Lactobacillus casei  
> Lactobacillus casei 12A

**European Nucleotide Archive Records**

CP006690.1

**References**

- Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. Broadbent J.R., Neeno-Eckwall E.C., Stahl B., Tandee K., Cai H., Morovic W., Horvath P., Heidenreich J., Perna N.T., Barrangou R., Steele J.L. - *BMC Genomics* 2012, 13:533 PubMed: 23035691

**Ensembl Genomes API Example**

This example Perl script shows how to create a database adaptor for this species. For more information see the Ensembl Bacteria documentation.

**Statistics**

Assembly: ASM30956v2, INSDC Assembly GCA\_000309565.2

Database version: 81.1

Base Pairs: 2,907,892 } ③

Golden Path Length: 2,907,892 }

Data source: European Nucleotide Archive

Genebuild method: Generated from ENA annotation

**Gene counts**

Coding genes: 2,681

Non coding genes: 72

Small non coding genes: 72

Pseudogenes: 46

Gene transcripts: 2,799

**Coordinate Systems**

chromosome 1 sequence

Filter

Sequence ▲ Length (bp) ▼

Chromosome 2907892