

# バクテリアゲノム解析 実習編

# 参考書籍



次世代シーケンサーDRY解析教本  
改定第2版（学研メディカル秀潤社）  
2019年12月発行

本講義では 0 から始めるバクテリアゲノム解析の内容を一部改変して紹介します。

- **実行環境**
- **Miniconda のインストール**
- **必要な解析ソフトのインストール**
- **解析に用いるデータの取得**
- **解析**
  - **データ前処理**
  - **ゲノムアセンブリ**
  - **アノテーション**

# 実行環境

## Mac • Linux

テキストエディタ

ターミナルを使用してコマンド実行

## Windows

Windows Subsystem for Linux (WSL) または、  
VirtualBox等の仮想マシンを使って Linux のコマンドが  
実行できる環境が必要

## 遺伝研スパコン

無料で利用可能

# Miniconda のインストール

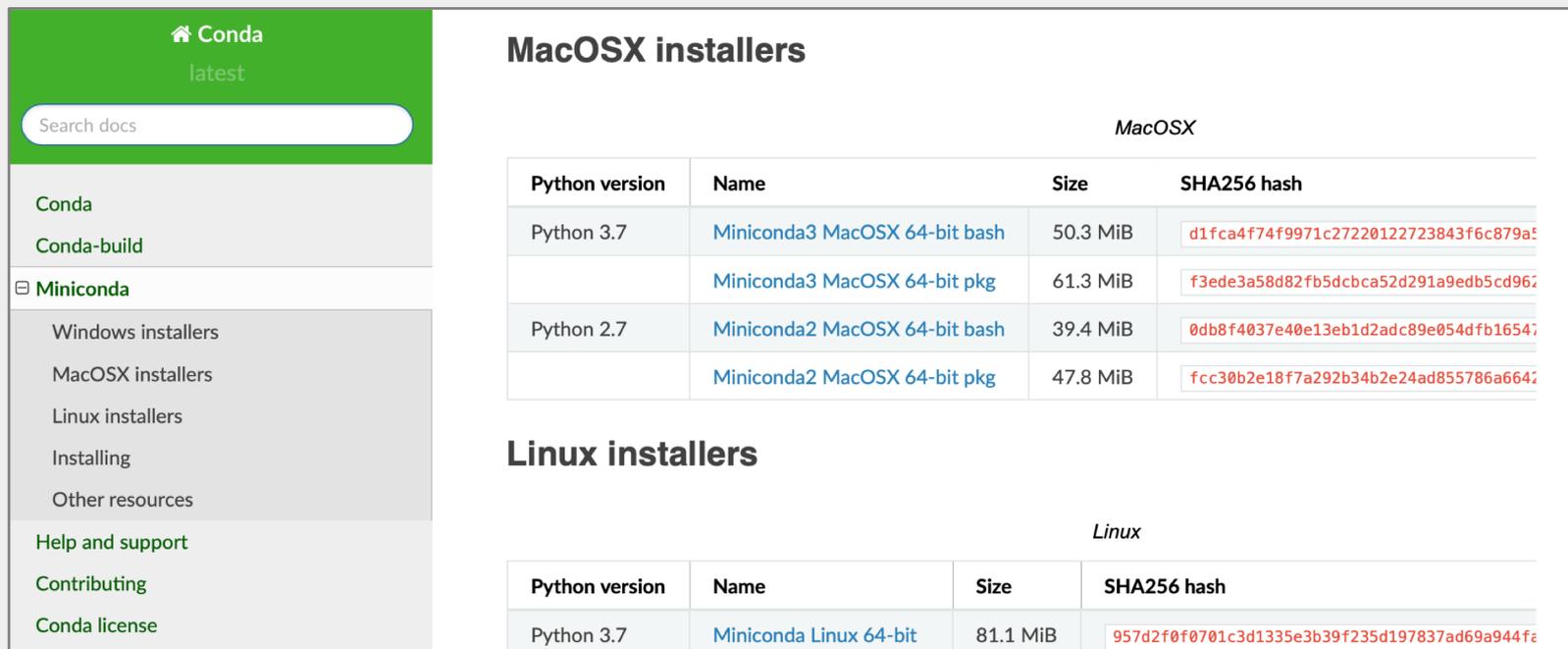
## Miniconda: データ解析プラットフォーム Anaconda の最小構成

conda コマンドを使用して様々なソフトウェアを簡単にインストールすることができる

生命科学用のソフトウェアの多くは Bioconda から利用可能

<https://bioconda.github.io>

## Miniconda インストーラーの取得



The screenshot shows the Miniconda documentation page. On the left is a navigation sidebar with a search bar and links to 'Conda', 'Conda-build', 'Miniconda', 'Windows installers', 'MacOSX installers', 'Linux installers', 'Installing', 'Other resources', 'Help and support', 'Contributing', and 'Conda license'. The main content area is titled 'MacOSX installers' and includes a sub-header 'MacOSX'. Below this is a table with columns for 'Python version', 'Name', 'Size', and 'SHA256 hash'. The table lists four installers: two for Python 3.7 (bash and pkg) and two for Python 2.7 (bash and pkg). Below the MacOSX section is the 'Linux installers' section with a sub-header 'Linux' and a table listing one installer for Python 3.7 (64-bit).

MacOSX			
Python version	Name	Size	SHA256 hash
Python 3.7	Miniconda3 MacOSX 64-bit bash	50.3 MiB	d1fca4f74f9971c27220122723843f6c879a5
	Miniconda3 MacOSX 64-bit pkg	61.3 MiB	f3ede3a58d82fb5dcbca52d291a9edb5cd96
Python 2.7	Miniconda2 MacOSX 64-bit bash	39.4 MiB	0db8f4037e40e13eb1d2adc89e054dfb1654
	Miniconda2 MacOSX 64-bit pkg	47.8 MiB	fcc30b2e18f7a292b34b2e24ad855786a664

Linux			
Python version	Name	Size	SHA256 hash
Python 3.7	Miniconda Linux 64-bit	81.1 MiB	957d2f0f0701c3d1335e3b39f235d197837ad69a944f

<https://docs.conda.io/en/latest/miniconda.html>

# Miniconda のインストール

作業ディレクトリの準備

```
mkdir agribio  
cd agribio
```

インストーラーのダウンロード

```
curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh
```

インストーラーの実行

```
sh Miniconda3-latest-MacOSX-x86_64.sh
```

実行後、指示にしたがって進める

不要になったインストーラーの削除

```
rm Miniconda3-latest-MacOSX-x86_64.sh
```

# 必要な解析ソフトウェアの用意

FastQC: リードのクオリティチェックのためのソフト

```
conda install -c bioconda fastqc
```

Bioconda を conda の追加チャンネルとして指定するため、'-c' オプションを指定している。

fastp, seqkit, spades のインストール

```
conda install -c bioconda fastp seqkit spades
```

fastp: アダプター配列の除去や、低クオリティのリードを除くソフト

seqkit: 配列データに対する様々な操作を行うユーティリティソフト

spades: アセンブルのためのソフト

# 解析に用いるデータの取得 1

DRA: DDBJ Sequence Read Archive から取得

<https://www.ddbj.nig.ac.jp/dra/index.html>

生物種名や論文記載のアクセッション番号で検索

**DRASearch** [Search Home](#) [DRA Home](#)

Accession :

Organism :  StudyType :

CenterName :  Platform :

Keyword :

Show  records Sort by

Data Last Update 2020-05-11

### Statistics

Released Entries

Type	Count
<a href="#">Submission</a>	1640652
<a href="#">Study</a>	258908
<a href="#">Experiment</a>	8692953
<a href="#">Sample</a>	7833242
<a href="#">Run</a>	9638959

Organism			Study Type		
#	Organism Name	Study	#	Study Type	Study
1	<a href="#">Homo sapiens</a>	24473	1	<a href="#">Other</a>	117680
2	<a href="#">Mus musculus</a>	20466	2	<a href="#">Whole Genome Sequencing</a>	69415
3	<a href="#">soil metagenome</a>	7937	3	<a href="#">Transcriptome Analysis</a>	35073
4	<a href="#">Saccharum bicolor</a>	2671	4	<a href="#">Metagenomics</a>	24157

# 解析に用いるデータの取得 2

論文\* で使われているシーケンスデータ DRR024501 を検索

\* <https://dx.doi.org/10.1186%2Fs12864-015-1435-2>

**DRASearch** [Search Home](#) [DRA Home](#)

**DRR024501** [FASTQ](#) [SRA](#)

Run Detail	
Alias	DRR024501
Instrument model	
Date of run	
Run center	
Number of spots	2,971,310
Number of bases	1,491,597,620

Navigation	
Submission	<a href="#">DRA002643</a> <a href="#">FTP</a>
Study	<a href="#">DRP002401</a>
Experiment	<a href="#">DRX022186</a> <a href="#">FASTQ</a> <a href="#">SRA</a>
Sample	<a href="#">DRS016698</a>

**READS (joined)**    quality  show 10 rows << < 1 / 297131 Page > >>

```
>DRR024501.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGATCAACACGAAACCATTCCAGCCGGCAAAGGT
AATCTAAACCACCCATTAGCTGTTATTGAAGCTTTGCAGCAACGAGTTGATGATAAAATGACCGTTTCGGTTGATGTGGG
GAGCCATTATTTGGATGGCCCGGCACTTCCGAAATTATGAGCCTCGCCATTTATTGTTAGTAATGGGATGCAGACGC
```

# 解析に用いるデータの取得3 (DRAのリード情報の意味)

Forward側リード (read1)



Reverse側リード (read2)

DNAの断片 (insert)

Library Description	
Name	LH_LOOC260_lib2
Strategy	WGS
Source	GENOMIC
Selection	RANDOM
Layout	PAIRED
Orientation	
Nominal Length	520 → insert の長さ
Nominal Sdev	50.0
Construction Protocol	

Spot Information	
Number of Reads per Spots	0
Spot Length	502 → read1 と read2 の の長さの合計

# 解析に用いるデータの取得 4 (データのダウンロード)

/ddbj\_database/dra/fastq/DRA002/DRA002643/DR

 [親ディレクトリ]

名前	サイズ	更新日
 <a href="#">DRR024501_1.fastq.bz2</a>	449 MB	2014/11/12 9:00:00
 <a href="#">DRR024501_2.fastq.bz2</a>	504 MB	2014/11/12 9:00:00

## ダウンロード

```
curl -O  
ftp://ftp.ddbj.nig.ac.jp//ddbj_database/dra/fastq/DRA002/DRA002643/DRX02  
2186/DRR024501_1.fastq.bz2
```

```
curl -O  
ftp://ftp.ddbj.nig.ac.jp//ddbj_database/dra/fastq/DRA002/DRA002643/DRX02  
2186/DRR024501_2.fastq.bz2
```

## 展開

```
bunzip2 *.bz2
```

アスタリスク \* はワイルドカードとして使える。

# データ前処理 1

データのサブサンプリング (時間短縮のため)

合計100万件のリードを抽出

```
seqkit sample -n 500000 DRR024501_1.fastq > DRR024501_1.1M.fastq
```

```
seqkit sample -n 500000 DRR024501_2.fastq > DRR024502_1.1M.fastq
```

データの確認

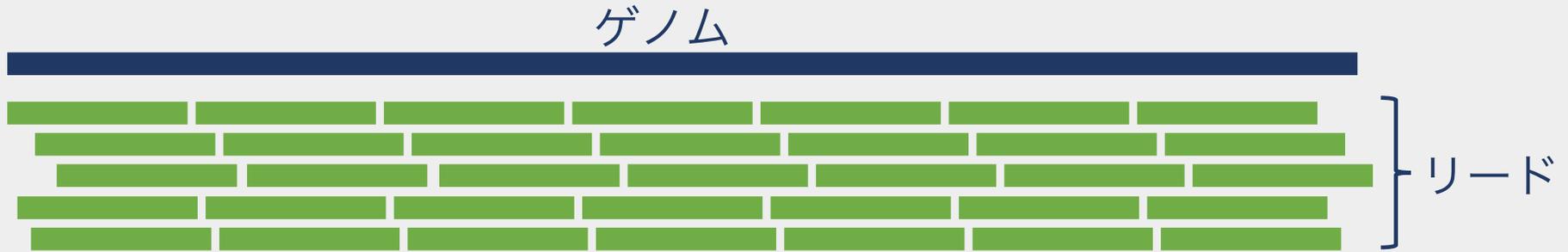
```
seqkit stat *fastq
```

結果

file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
DRR024501_1.1M.fastq	FASTQ	DNA	499,916	125,478,916	251	251	251
DRR024501_1.fastq	FASTQ	DNA	2,971,310	745,798,810	251	251	251
DRR024501_2.1M.fastq	FASTQ	DNA	499,916	125,478,916	251	251	251
DRR024501_2.fastq	FASTQ	DNA	2,971,310	745,798,810	251	251	251

## データ前処理 2 (データ量について)

カバレッジ：1塩基あたり何回シーケンスされているか？  
(Depth of coverage と呼ぶ)



この場合はカバレッジは 5x

$$\begin{aligned} \text{カバレッジ} &= \frac{\text{リード数} \times \text{1リードあたり長さ}}{\text{ゲノムサイズ}} \\ &= \frac{1\text{M} \times 250 \text{ bp}}{2.5\text{Mbp}} \\ &= 100\text{x} \end{aligned}$$

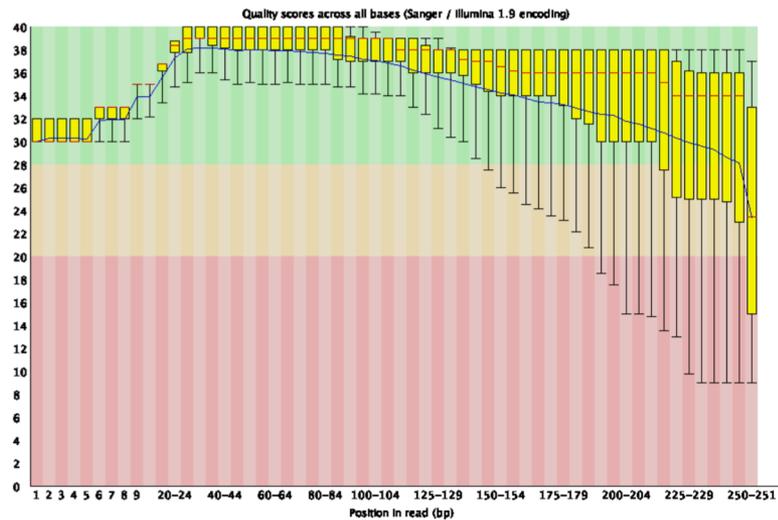
M = 百万 (million)

# データ前処理 3

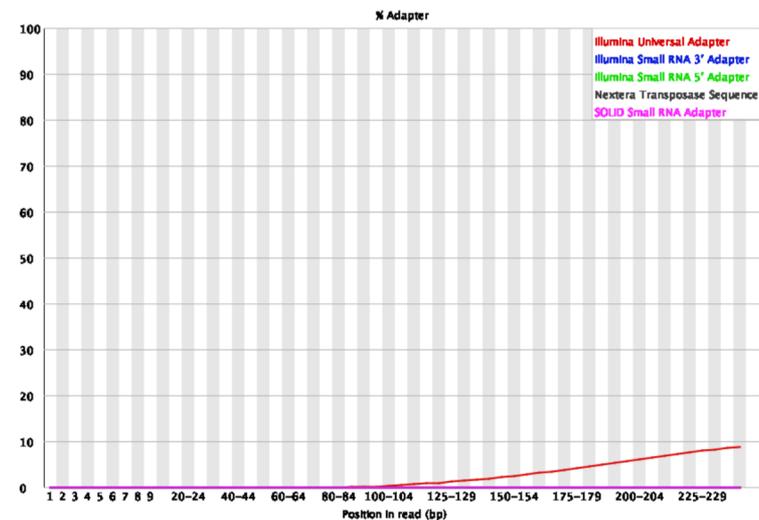
## FastQCによるクオリティチェック

```
fastqc DRR024501_1.1M.fastq
```

### Per base sequence quality



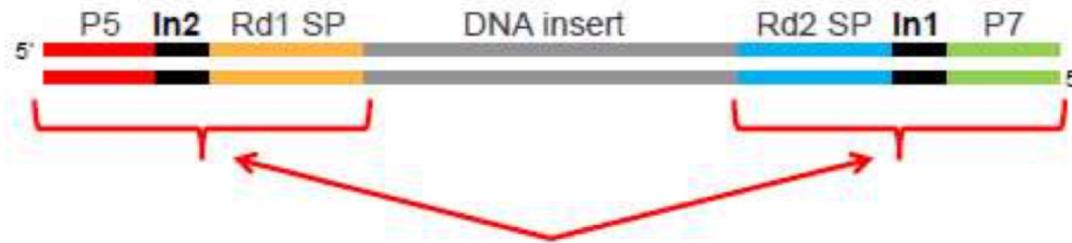
### Adapter Content



# データ前処理 4

## アダプターとは

### イルミナ ライブラリの構造



### イルミナシーケンサー専用オリゴヌクレオチドアダプター

- DNA インサート : 数百bpに断片化したDNA. 読みたい目的サンプル配列.
- P5, P7 : フローセルへの結合部位
- SP : シーケンシングプライマー結合部位
- In (Index) : 複数サンプル同時解析用のバーコード (目印配列)

**ライブラリ = DNA インサート + 両端にそれぞれ別のアダプター**

イルミナシーケンサーでシーケンスするため、この構造をとるようにサンプル調整する

[https://jp.illumina.com/content/dam/illumina-marketing/apac/japan/documents/pdf/2015\\_techsupport\\_session10.pdf](https://jp.illumina.com/content/dam/illumina-marketing/apac/japan/documents/pdf/2015_techsupport_session10.pdf) より引用

DNA insertの長さが短いと、リード後半にアダプター配列が含まれることがある

# データ前処理 5

## アダプター配列の除去

```
fastp -i DRR024501_1.1M.fastq -o DRR024501_1.1M.fastp.fastq -I  
DRR024501_2.1M.fastq -O DRR024501_2.1M.fastp.fastq
```

## データ再確認

```
seqkit stat *.fastp.fastq
```

```
fastqc DRR024501_1.1M.fastp.fastq
```

# ゲノムアセンブリ 1

## Spadesの実行

```
spades.py -o assemble -1 DRR024501_1.1M.fastp.fastq -2  
DRR024501_2.1M.fastp.fastq
```

## 結果の確認

```
seqkit stat -G N -a assemble/contigs.fasta
```

```
seqkit stat -G N -a assemble/scaffolds.fasta
```



# アノテーション

DFAST の実行 (ファイルをアップロード)

 Analysis ▾ Archive DFAST-core API Download FAQ Help ▾

DFAST Prokaryotic genome annotation pipeline

---

**Query File (Fasta format, up to 15Mbyte)**

**Demo mode (Sample annotation for E.coli O26)**

**Job Title**

**Mail Address**

---

---

# 大量処理・自動処理のニーズが高まり、 All-in-one の解析ツールが増えている。

Feature	Bactopia	ASA <sup>3</sup> P	Nullarbor	TORMES
Version	1.3.0	1.2.2	2.0.20191013	1.0
Release Date	February 19th, 2020	November 20th, 2019	October 13th, 2019	February 4th, 2019
Latest Commit	February 23th, 2020	February 19th, 2020	February 23th, 2020	January 30th, 2020
Sequence Technology	Illumina	Illumina, Nanopore, PacBio	Illumina	Illumina
Single End Reads	Yes	Yes	No	No
Workflow	Nextflow	Groovy	Perl + Make	Bash
Resume If Stopped	Yes	No	Yes	No

