バクテリアゲノム解析
実習編

参考書籍

次世代シークエンサー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

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

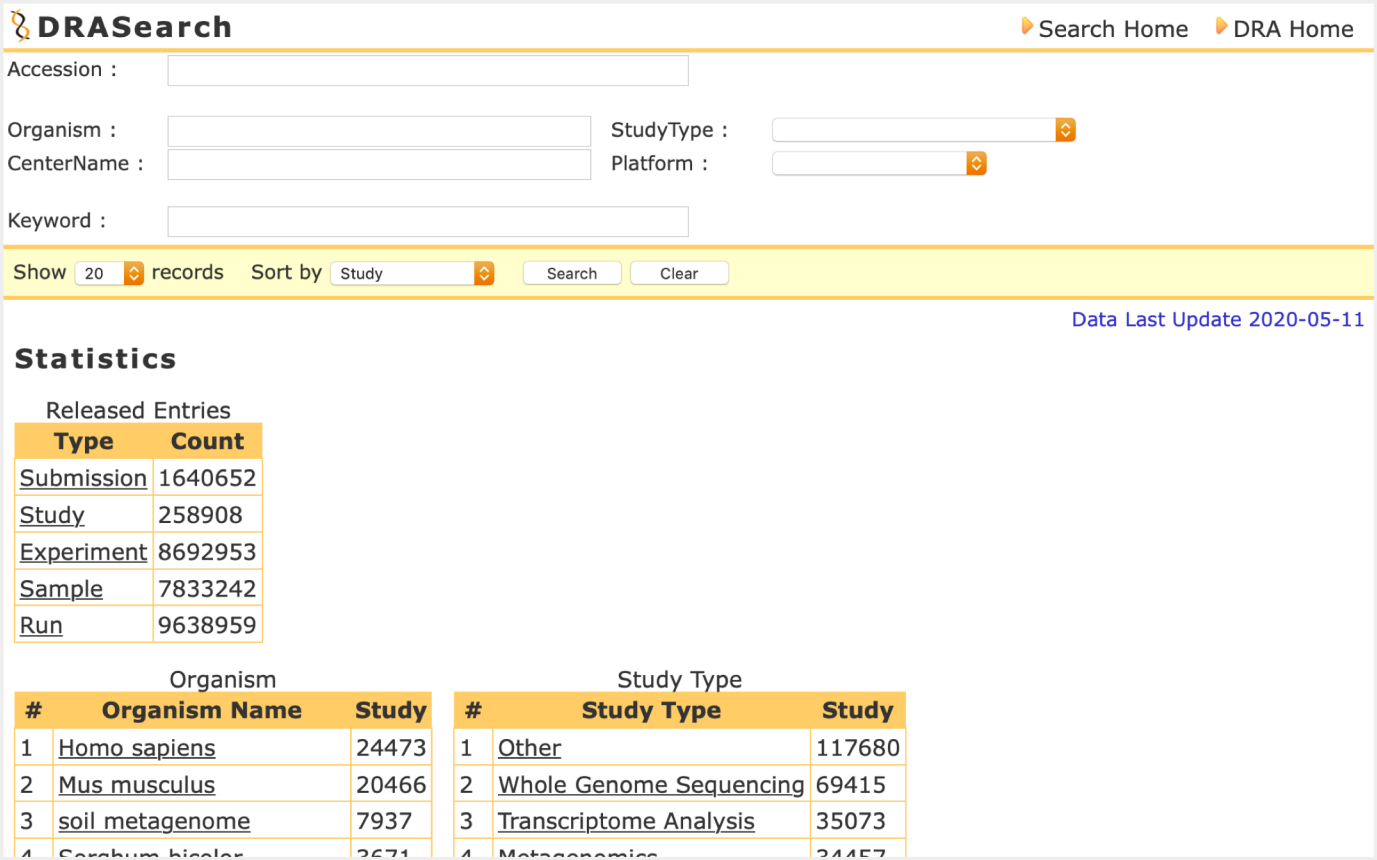
spades: アセンブルのためのソフト

解析に用いるデータの取得１

DRA: DDBJ Sequence Read Arcive から取得

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

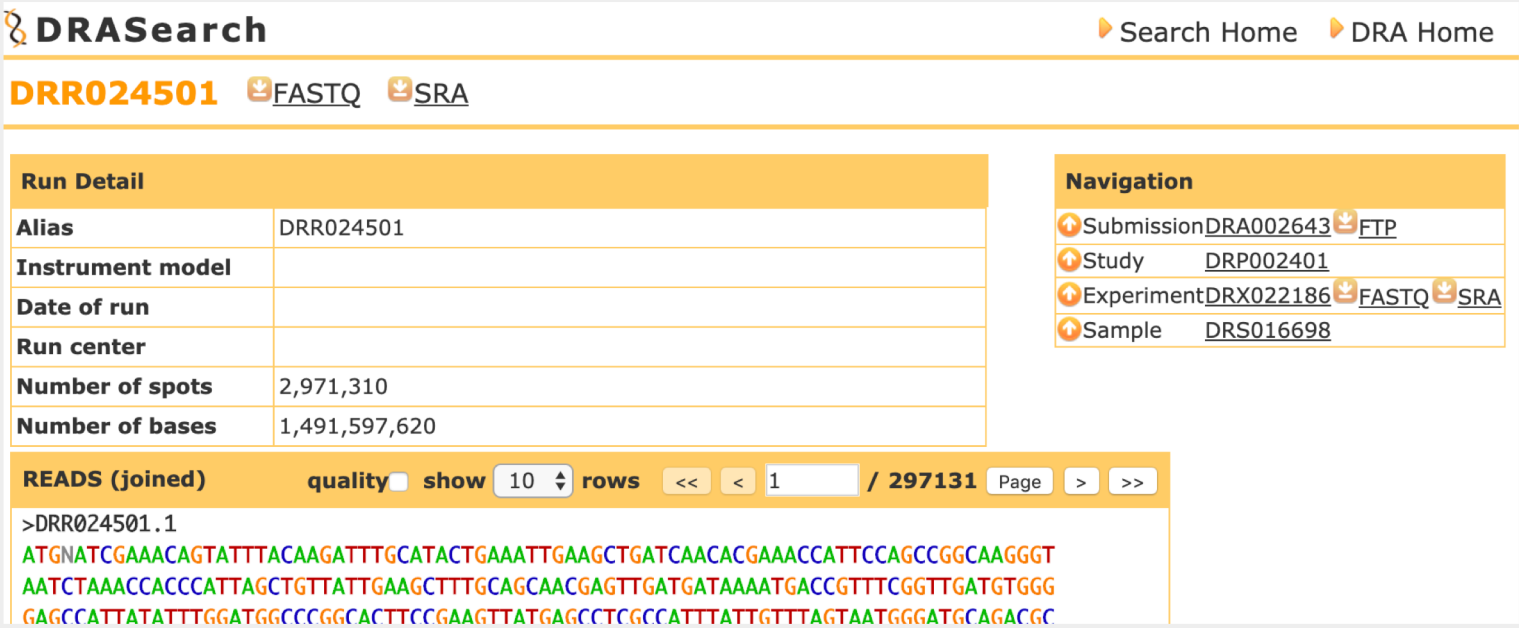
生物種名や論文記載のアクセッション番号で検索



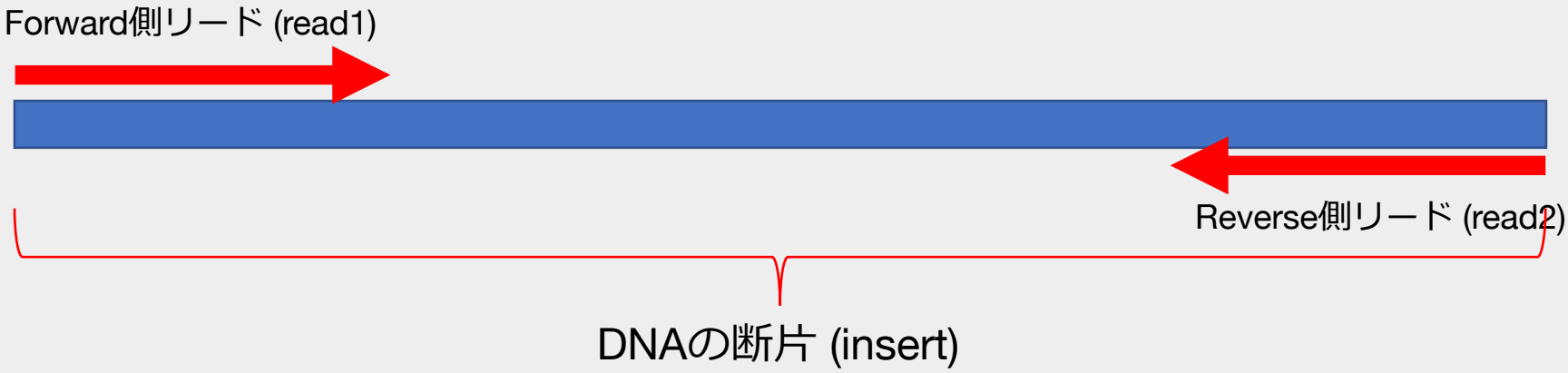
解析に用いるデータの取得２

論文* で使われているシークエンスデータ DRR024501 を検索

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



解析に用いるデータの取得３（DRAのリード情報の意味）



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)

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クエンスされているか？
(Depth of coverage とも呼ぶ)

この場合はカバレッジは 5x

$$\text{カバレッジ} = \frac{\text{リード数} \times \text{1リードあたり長さ}}{\text{ゲノムサイズ}}$$

$$= \frac{1\text{M} \times 250\ \text{bp}}{2.5\text{Mbp}}$$

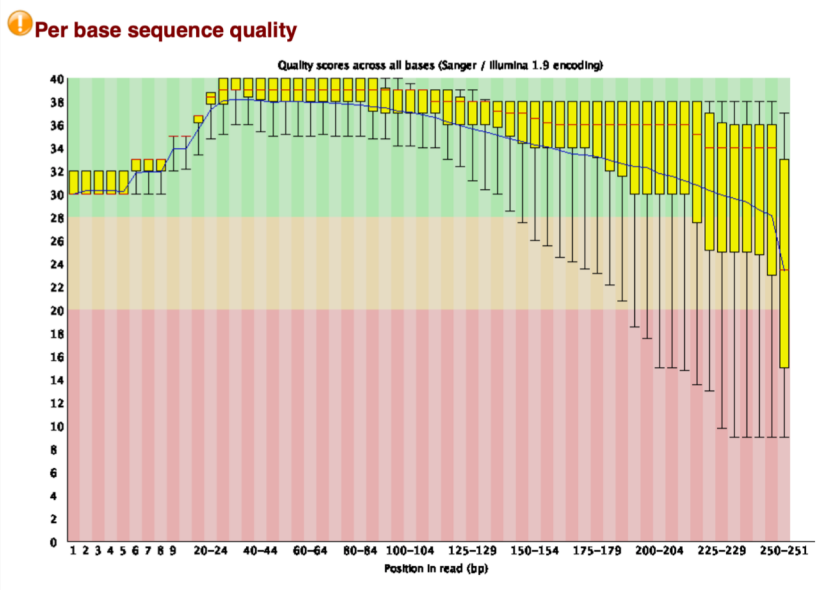
$$= 100\text{x}$$

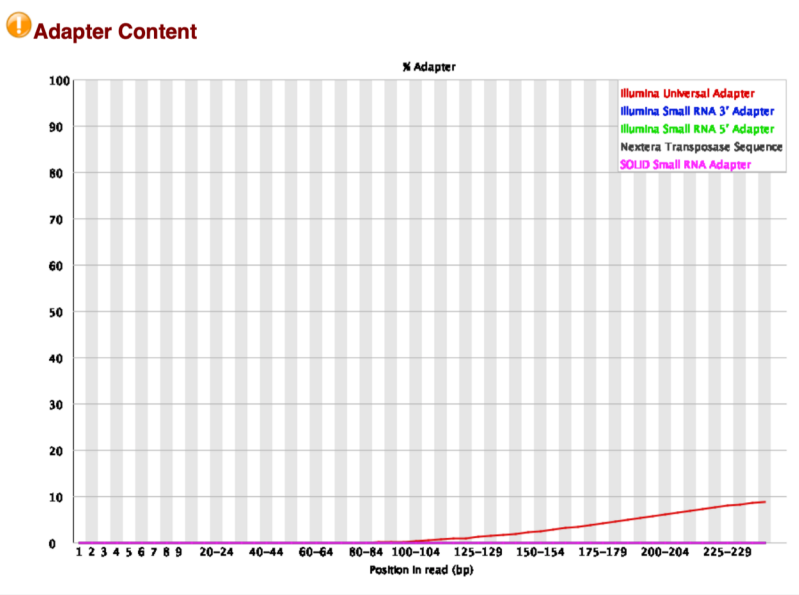
M = 百万 (million)

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
```

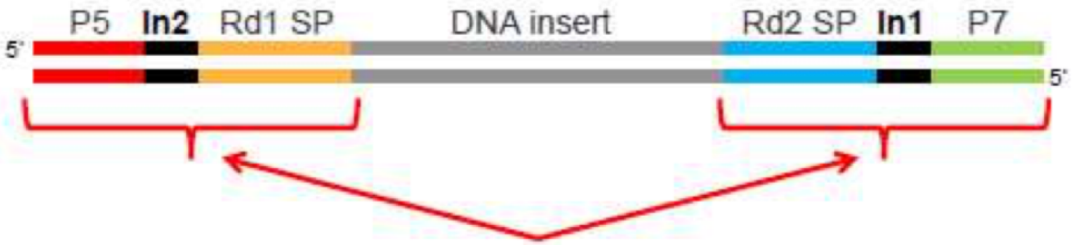




データ前処理４

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 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
```

ゲノムアセンブリ２　N50

配列を長いものから順に足していったときに、全長の50%に達した時の配列の長さ
（配列長を考慮した加重平均）

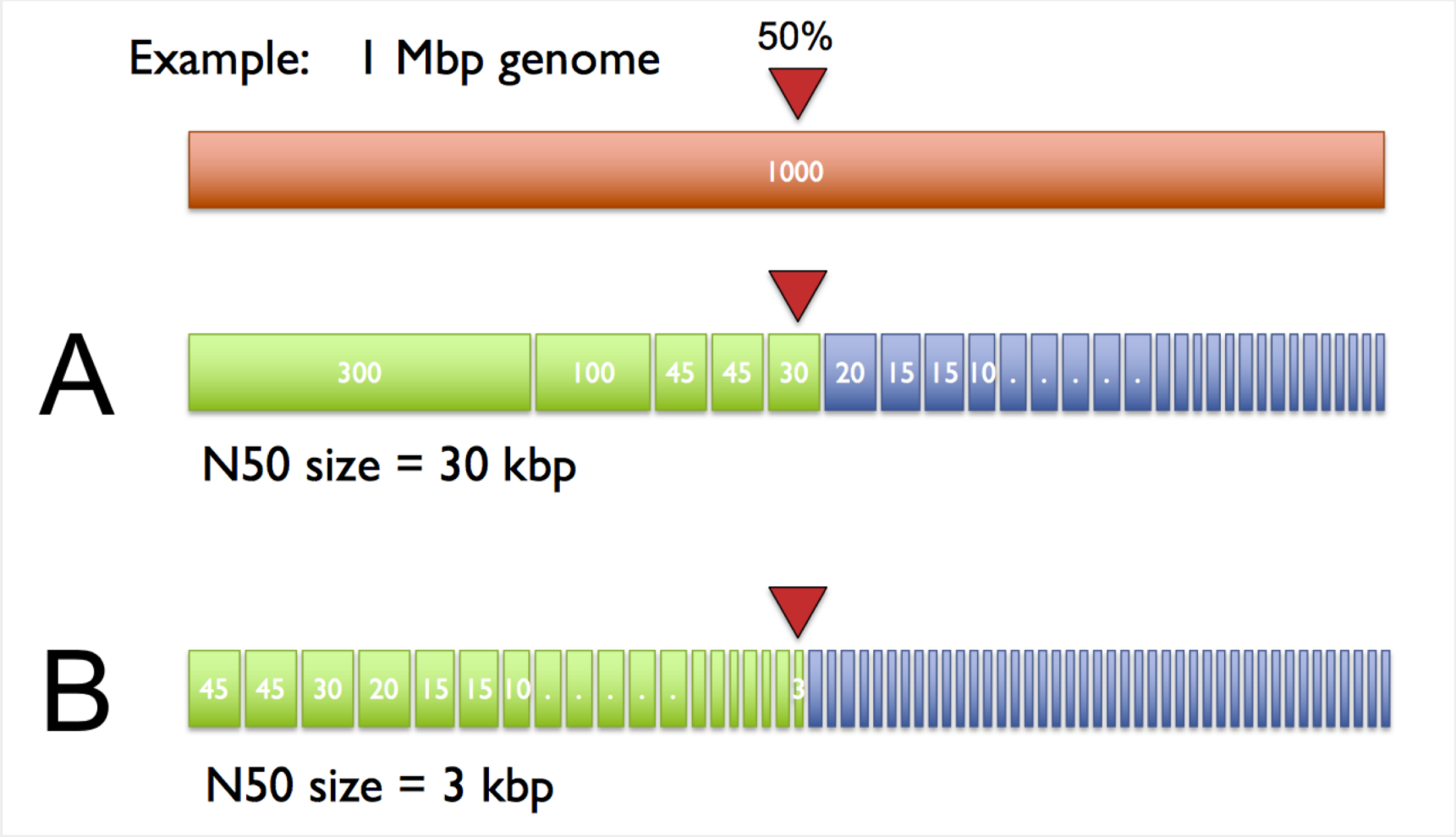
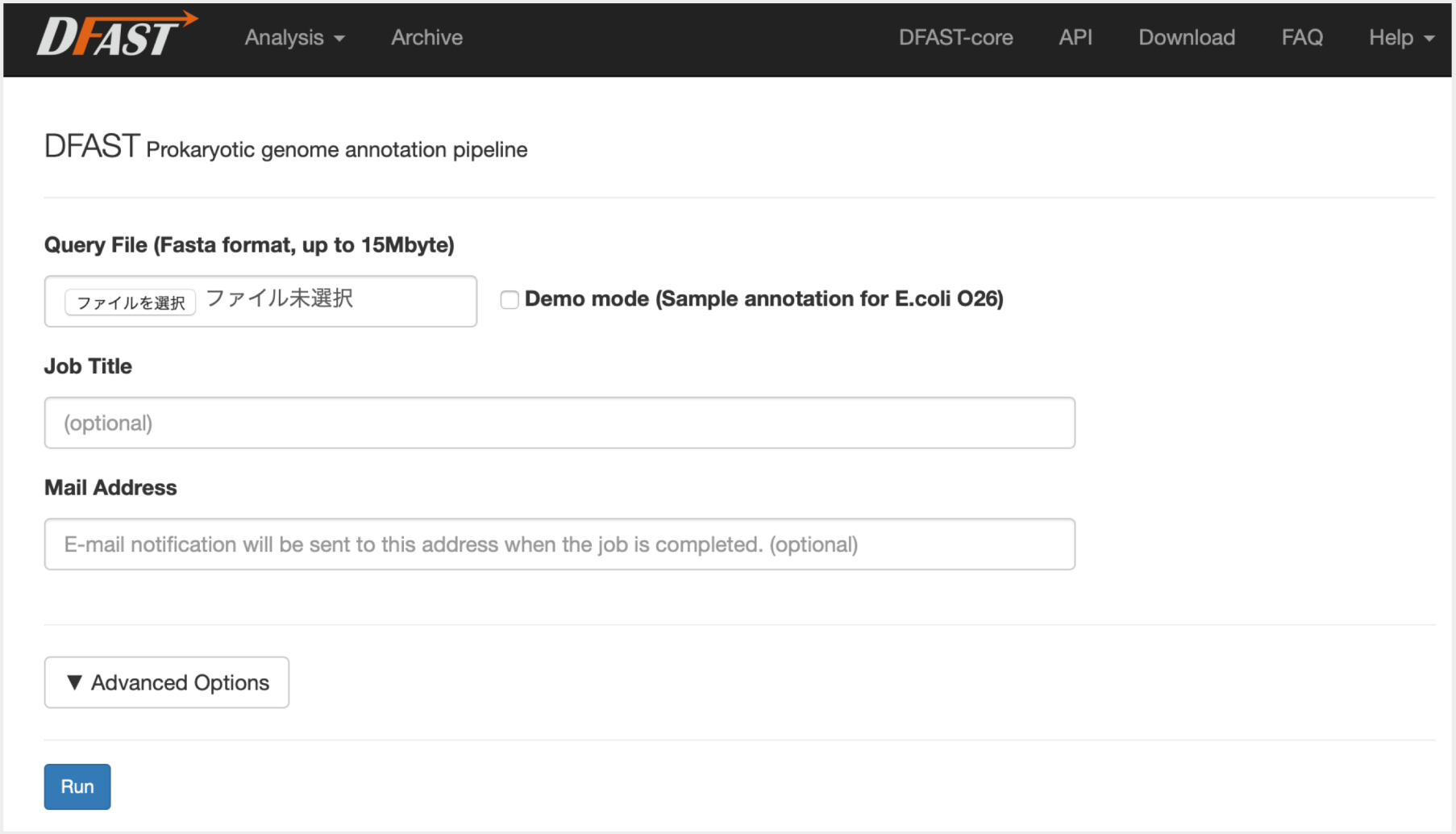


図. M. Schatz　http://schatz-lab.org/appliedgenomics2018/lectures/03.GenomeAssembly.pdf

アノテーション

DFAST の実行（ファイルをアップロード）



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^{3}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

https://www.biorxiv.org/content/biorxiv/early/2020/03/02/2020.02.28.969394.full.pdf より引用