

ゲノム工学実習 大学院科目

・【開講日程】 2018 年度 秋学期 特定期間集中 【担当教員】 荒川 和晴

【前提科目（推奨）】 B6161: 基礎分子生物学 4
【前提科目（推奨）】 B6160: 基礎分子生物学 3
【前提科目（推奨）】 B6159: 基礎分子生物学 2
【前提科目（推奨）】 B6158: 基礎分子生物学 1
【前提科目（推奨）】 34190: 基礎分子生物学 4
【前提科目（推奨）】 34180: 基礎分子生物学 3
【前提科目（推奨）】 34170: 基礎分子生物学 2
【前提科目（推奨）】 34160: 基礎分子生物学 1
【前提科目（推奨）】 B3215: 生命科学実験の基礎
【前提科目（推奨）】 C2038: 遺伝子解析実習
【前提科目（推奨）】 34130: 遺伝子解析実習

・【開講場所】 TTCK 【授業形態】 講義、実習
・【履修条件】 TTCK 生のみ履修可
・【連絡先】 gaou@sfc.keio.ac.jp

注意

配布資料などは [SFC-SFS](#) の授業ページで公開します。

科目概要

DNA 解析技術の飛躍的な向上により、微生物程度のゲノム解析はもはや「誰でも」「どこでも」可能なレベルにまで簡単になってきている。特に、携帯型ナノポアシーケンサーの登場は初期投資をほぼ必要とせずに長鎖 DNA の解析を安価に可能とした。このような現状を踏まえれば、微生物程度のゲノムであれば遺伝子単位ではなくもはやゲノム単位で DNA を解析することが第一選択肢となる時代が到来していることを意味する。そこで、本実習では任意の微生物から長鎖 DNA を抽出・精製し、ナノポアシーケンサーにて DNA を読み取り、それをバイオインフォマティクスによりアセンブル・アノテーションし、解析可能なゲノム情報にして、さらにそれを Genome Reports の形にして国際誌に投稿するまでの全過程を学ぶ。

授業シラバス

主題と目標 / 授業の手法など

DNA 解析技術の飛躍的な向上により、微生物程度のゲノム解析はもはや「誰でも」「どこでも」可能なレベルにまで簡単になってきている。特に、携帯型ナノポアシーケンサーの登場は初期投資をほぼ必要とせずに長鎖 DNA の解析を安価に可能とした。このような現状を踏まえれば、微生物程度のゲノムであれば遺伝子単位ではなくもはやゲノム単位で DNA を解析することが第一選択肢となる時代が到来していることを意味する。そこで、本実習では任意の微生物から長鎖 DNA を抽出・精製し、ナノポアシーケンサーにて DNA を読み取り、それをバイオインフォマティクスによりアセンブル・アノテーションし、解析可能なゲノム情報にして、さらにそれを Genome Reports の形にして国際誌に投稿するまでの全過程を学ぶ。

前半の過程では実際にナノポアシーケンサーに適した長鎖 DNA をシーケンスする実験を実習として行い、後半ではシーケンスされた DNA をコンピュータを用いて解析する。よって、実験・バイオインフォマティクス双方の過程を学ぶが、知識としては少なくとも実験の経験があれば構わない。

教材・参考文献

参考文献：

1. 荒川和晴(企画), "どこでも 誰でも より長く ナノポアシーケンサーが研究の常識を変える！", 実験医学 2018 年 1 月号 Vol.36 No.1

提出課題・試験・成績評価の方法など

実験ノート及び最終レポートをもって評価する

履修上の注意

実験経験のある TTCK 生のみ履修可。

授業計画

■第1回 イントロダクション

ナノポアシーケンスと、ゲノム解析の流れについて講義します。

■第2回 長鎖 DNA 抽出 1

ナノポアシーケンス用長鎖 DNA を抽出します。

■第3回 長鎖 DNA 抽出 2

ナノポアシーケンス用長鎖 DNA を抽出します。

■第4回 長鎖 DNA 抽出 3

ナノポアシーケンス用長鎖 DNA を抽出します。

■第5回 長鎖 DNA 抽出 4

ナノポアシーケンス用長鎖 DNA を抽出します。

■第6回 長鎖 DNA QC 1

長鎖 DNA の品質をパルスフィールド電気泳動を用いて検証します。

■第7回 長鎖 DNA QC 2

長鎖 DNA の品質をパルスフィールド電気泳動を用いて検証します。

■第8回 長鎖 DNA QC 3

長鎖 DNA の品質をパルスフィールド電気泳動を用いて検証します。

■第9回 ナノポアライブラリ作製

ナノポアシーケンス用ライブラリを作成します。

■第10回 ナノポアシーケンシング

ライブラリをシーケンスにかけます。

■第11回 ゲノムアセンブリー

得られたゲノムをアセンブルします。

■第12回 エラー補正

Nanopolish を用いてエラー補正します。

■第13回 ゲノムアノテーション

D-FAST を用いてゲノムをアノテーションします。

■第14回 Genome Report 執筆

これまでに得られたデータを Genome Report の形にまとめます。

■その他

毎回実験ノートをまとめ、次回の準備をする

ゲノムのアセンブリー

■環境設定

ソフトウェアのインストールには miniconda が便利です。
<https://docs.conda.io/projects/miniconda/en/latest/>

```
$ mkdir -p ~/miniconda3
$ wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/miniconda.sh
$ bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
```

After installing, initialize your newly-installed Miniconda. The following commands initialize for bash and zsh shells:

```
~/miniconda3/bin/conda init bash
~/miniconda3/bin/conda init zsh
```

conda で作業するには base 環境じゃない方が良いので、例えば新規に gew という環境を作り

```
$ conda create -n gew
```

その環境内で作業するようにすると良いです。

```
$ conda activate gew
```

ソフトウェアのインストールは以下のようにします。

```
$ conda install -c bioconda bbmap
```

bbmap のところを porechop, canu, nanoplot, などとソフトウェアの名前に変更してインストール。

サーバは king を使えると良いですが、Torque の使い方がわからない場合には smith6/7 を使っても良いです（授業以外での利用はご遠慮ください）。

■ファイル

mt-rg ファイルサーバの

```
/home/gaou/gew/
```

の該当する年度のフォルダの下にそれぞれのバーコードに相当するナノポア配列 (fastq)、~~<strike>illumina~~ フォルダ以下にバーコードに対応する Illumina 配列ファイル (fastq) があります~~</strike>~~。guppy ソフトウェアでベースコール後、バーコードの demultiplex を行なっています。各自自分のホームにファイルをコピーして以降の解析を実施してください。

■ アダプターのフィルタリング

微生物ゲノムの解析だとやらなくてもなんとかありますが、アダプタートリミングをしておいた方が良いでしょう。アダプタートリミングを行えるソフトウェアとしては NanoFilt, porechop, filtlong などがあります。ただし、配列ベースでアダプター除去ができる porechop のサポートは終了しています。

```
$ porechop -i BC01.fastq -o BC01.porechop.fastq --discard_middle
```

NanoFilt の場合 (Quality 10 以上の配列のみを残す、先頭の 100 文字を除去 バーコードを使わない場合は 50 文字で ok)

```
$ NanoFilt -q 10 --headcrop 100 BC01.fastq > BC01.nanofilt.fastq
```

■ リードのフィルタリング

まず、現状ではリードが多すぎるので、だいたい x50~x100 になるようにリードを調整します。この時、長いリードは残したいので、BBMap の reformat.sh を使って、

```
# reformat.sh in=BC01.fq out=BC01-filter10k.fq minlength=10000 qin=33
```

のようにします。

以下のコマンド (BBMap: <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/>) で統計値を算出して、カバレッジを合わせます。

```
$ stats.sh BC01-filter10k.fq
```

基本的な配列の統計を可視化するには、NanoPlot (<https://github.com/wdecoster/NanoPlot>) が便利です。

```
$ NanoPlot -t 24 --fastq BC01.fq -p BC01
```

■ Canu のアセンブリー

その後 canu でアセンブリーを行います。メモリや CPU の関係上 king でやるのが良いかと思いますが、(king 利用についてはこちらを参照。

https://www.bioinfo.ttck.keio.ac.jp/wordpress/?page_id=2383)

```
qsub -l -l nodes=1:ppn=32
```

で king のノードにログインし、

```
canu -nanopore BC01.fastq -d BC01 -p BC01 -fast useGrid=false genomeSize=4m maxThreads=8
```

のように打って (BC01.fastq は上でフィルタリングした fastq のファイル名に、BC01 のところは自分のバーコード、あるいは任意の名前に変えてください) アセンブリーを実行してください。-fast オプションは精度を犠牲にして実行速度を上げるオプションで、今回くらいのカバレッジ (x100) があればつけても問題なくアセンブリーできると思いますが、時間がかかっても構わないなら外してください。genomeSize オプション (頭にハイフンをつけない点に注意) は予想ゲノムサイズより少し大きめを設定してください。バクテリアの場合大抵 4m で良いと思います。maxThreads は、使用するサーバに合わせて設定してください。king の場合、32 に設定してください。

大きな問題がなければこれで数時間でアセンブリーが終了します。

```
$ stats.sh BC01/BC01.contigs.fasta
```

と打って、ゲノムサイズのちゃんとアセンブリーが終了したか確認します。

■ 末端処理

まず環状化ができているかを確認します。

```
$ grep ">" BC01.contigs.fasta
```

で各 contig の FASTA ヘッダを見て、右端の suggestCircular が yes になっているか確認します。(環状染色体の場合) 長さが 10kbp に満たない suggestCircular が no の contig はゴミの可能性が高いので、多くの場合破棄して構いません。

suggestCircular が yes のものは、染色体かプラスミドの可能性が高いです。長さが 1Mbp に満たないものはプラスミドの可能性が高いので、適当に数 kbp 分をコピーして NCBI BLAST に投げてみましょう。プラスミド配列にヒットするようでしたら高い確率でプラスミドと言えます。

ここで suggestCircular が yes になっていても、環状であることを確認しているだけで環状化されているわけではありません。最初の 50 文字程度で自身の配列内を検索して、末端部分に該当する場所を見つけてください。基本的に、その後に続く配列が先と部分と末端の該当部分で一致するはずです。一致を確認したら、末端の一致部分を削除します。

■ アセンブリークオリティの検証

アセンブリーのクオリティ確認は N50 だけではだめで、ちゃんとゲノムとして全遺伝情報がカバーできているか、を確認する必要があります。このためには微生物だと CheckM, 真核生物だと BUSCO といった手法 (<http://kazumaxneo.hatenablog.com/entry/2017/07/19/145640>) を用います。

CheckM は DFAST Quality Control ツール (<https://dfast.nig.ac.jp/dqc/submit/>) で簡単に検証できます。

ローカルで CheckM を利用する際にはこちらを参考にしてください (<https://kazumaxneo.hatenablog.com/entry/2017/09/22/012544>)。

次のエラー補正で、この completeness が 100%(かそれに限りなく近くなるよう)にしていきます。ナノポアの場合、アセンブル直後だと多くの場合 indel 系のエラーが解決されないで残ります。

■ エラーコレクション

エラー補正には、ナノポアリードを使う場合と、Illumina リードを使う場合の二通りがあります。今回この授業ではナノポアのみでエラーコレクションを行います。

nanopore での補正には medaka を使います。

```
$ medaka_consensus -i BC01.fq -d BC01.contigs.fasta -o medaka_output -m r1041_e82_400bps_sup_v4.3.0
```

medaka 後に再度 checkM で確認すると、精度が向上していることが確認できます。ただし、一回では不十分なことが多いので、ここで作成したエラーコレクション後のファイル (pilon1.fasta) に再度 Illumina リードをマッピングし、pilon をかけ直し、BUSCO スコアが向上する限りこれを繰り返します。

番外編：

0. Illumina データのフィルタリング Illumina のデータは正確ですが、とはいえ末端の配列などはクオリティが高くない場合があります。クオリティが悪い配列でエラーコレクションしようとすると余計にエラーを導入することになりかねないので、事前にクオリティでフィルタリングをした方が良いでしょう。この時には fastp というソフトウェアを使うとほぼ全自動でやってくれます。 (<https://kazumaxneo.hatenablog.com/entry/2018/05/21/111947>)

```
$ fastp -i BC01_S1_L001_R1_001.fastq -I BC01_S1_L001_R2_001.fastq -o fastpR1.fq -O fastpR2.fq -g -c -w 8
```

配列が多すぎる場合にはだいたい 100x くらいの量にサブサンプリングした方がいい場合があります。300bp paired の配列で、ゲノムサイズが 4M の場合、100x だと 600000 ペアの配列があれば十分です。

```
/home/gaou/kumamushi/software/bbmap/reformat.sh in1=fastpR1.fq in2=fastpR2.fq out1=subsampledR1.fq out2=subsampledR2.fq sampleradtarget=600000 qin=auto qout=auto
```

1. Illumina での補正 Pilon を使います。 <https://github.com/broadinstitute/pilon/wiki> 基本的には BWA で mapping 後 pilon をかけるだけです。

まずは mapping。

```
bwa index BC01.contigs.fasta  
bwa mem -t 8 BC01.contigs.fasta BC01_S1_merged_R1.fq | samtools view -@ 4 -b -o aln.bam -
```

```
samtools sort -T sort.tmp -o aln.sorted.bam -@ 4 aln.bam
samtools index aln.sorted.bam
```

次に pilon。

```
pilon --genome BC01.contigs.fasta --bam aln.sorted.bam --threads 4 --output pilon1
```

Illumina データを SRA からダウンロードする場合、SRA 形式から FASTQ 形式に変換する必要があります。

```
/home/gaou/kumamushi/software/sratoolkit.2.8.2 -1 -centos_linux64 /bin/fastq-dump SRR390728
--split-files
```

ペアドエンドの場合には --split-files オプションをつけてください。

■ アノテーション

アノテーションには、遺伝子予測、機能予測、ゲノム開始位置の dnaA への調整、などなど非常に手間のかかる作業がたくさんあるのですが、今はいい時代なので DDBJ DFAST というオンラインツールで全自動でやってくれます。 <https://dfast.nig.ac.jp>

Genome Report

■ 投稿先

ASM Microbiology Resource Announcements <https://mra.asm.org/content/getting-started>

■ フォーマット

<https://mra.asm.org/content/organization-and-format>

Abstract: 50 words Total word: 500 words (abstract と acknowledgements は除く) タイトル: 54 文字以内

■ チェックリスト

https://mra.asm.org/sites/default/files/additional-assets/thumbs/MRA_Author_Checklist.pdf

過去の Genome Report

■ 2022 年

- Kato S, Yuzawa S, Takeda T, Arakawa K, “ Complete genome sequence of *Kitasatospora aureofaciens* Tü117 ”, Microbiology Resource Announcements, 2024, 13:e01014-23. ([Publisher](#))
- Naruki M, Watanabe A, Warashina T, Morita T, Arakawa K*, “ Complete genome sequence of *Limnobacter thiooxidans* CS-K2T, isolated from freshwater lake sediments in Bavaria, Germany ”, Microbiology Resource Announcements, 2023,13:e00992-23. ([Publisher](#))
- Oshibuchi K, Yang J, Obana N, Fukuda Shinji, Arakawa K*, “ Complete genome sequence of *Solobacterium moorei* JCM 10645T isolated from a human stool sample ”, Microbiology Resource Announcements, 2023,13:e00965-23. ([Publisher](#))
- Suzuki Y, Fukazawa A, Sugawara K, Galipon J, Arakawa K*, "Complete genome sequence of PETase type IIa-harboring *Marinobacter nanhaiticus* D15-8W, isolated from a South China Sea

sediment", Microbiol Resour Announc, 2023, 12:e00868-23. ([Publisher](#))

■ 2021 年

- Ishikawa S, Huang M, Tomita A, Kurihara Y, Watanabe R, Iwai H, Arakawa K*, "Complete Genome Sequences of Four Bacteria Isolated from the Gut of a Spiny Ant (*Polyrhachis lamellidens*)", Microbiol Resour Announc, 2022, 11:e00333-22. ([Publisher](#))
- Takeda T, Fukumitsu N, Yuzawa S, Arakawa K*, "Complete Genome Sequence of *Streptomyces albus* Strain G153", Microbiol Resour Announc, 2022, 11:e00332-22. ([Publisher](#))

■ 2020 年

- Takahashi H, Yang J, Yamamoto H, Fukuda S, Arakawa K*, "Complete Genome Sequence of *Adlercreutzia equolifaciens* subsp. *celatus* DSM18785", Microbiol Resour Announc, 2021, 10:e00354-21. ([Publisher](#))
- Warashina, T, Yamamura S, Suzuki H, Amachi S, Arakawa K, "Complete Genome Sequence of *Geobacter* sp. Strain SVR, an Antimonate-reducing Bacterium Isolated from Antimony-rich Mine Soil", Microbiol Resour Announc, 2021, 10:e00142-21. ([Publisher](#))

■ 2019 年

- Takeyama N, Huang M, Sato K, Galipon J, Arakawa K*, "Complete Genome Sequence of *Halomonas hydrothermalis* Strain Slthf2, a Halophilic Bacterium Isolated from a Deep-Sea Hydrothermal-Vent Environment", Microbiol Resour Announc, 2020, 9:e00294-20. ([Publisher](#))
- Takahashi Y, Takahashi H, Galipon J, Arakawa K*, "Complete Genome Sequence of *Halomonas meridiana* Strain Slthf1, Isolated from a Deep-Sea Thermal Vent", Microbiol Resour Announc, 2020, 9:e00292-20. ([Publisher](#))
- Seo K, Tanaka K, Fukuda S, Arakawa K*, "Complete Genome Sequences of Two *Cutibacterium acnes* Strains Isolated from an Orthopedic Surgical Site", Microbiol Resour Announc, 2020, 9:e00290-20. ([Publisher](#))
- Kurihara Y, Kawai S, Sakai A, Galipon J, Arakawa K*, "Complete Genome Sequence of *Halomonas meridiana* Strain Eplume2, Isolated from a Hydrothermal Plume in the Northeast Pacific Ocean", Microbiol Resour Announc, 2020, 9:e00330-20. ([Publisher](#))
- Inoue H, Shibata S, Ii K, Inoue J, Fukuda S, Arakawa K, "Complete Genome Sequence of *Bifidobacterium longum* Strain Jih1, Isolated from Human Feces", Microbiol Resour Announc, 2020, 9:e00319-20. ([Publisher](#))
- Nishimura K, Ikarashi M, Yasuda Y, Sato M, Cano Guerrero M, Galipon J, Arakawa K, "Complete Genome Sequence of *Sphingomonas paucimobilis* Strain Kira, Isolated from Human Neuroblastoma SH-SY5Y Cell Cultures Supplemented with Retinoic Acid.", Microbiol Resour Announc, 2021, 10(6):e01156-20. ([PubMed](#))

■ 2018 年

- Tsurumaki M, Deno S, Galipon J, Arakawa K*, "Complete Genome Sequence of Halophilic Deep-Sea Bacterium *Halomonas axialensis* Strain Althf1", Microbiol Resour Announc, 2019, 8:e00839-19. <https://mra.asm.org/content/8/31/e00839-19>
- Evans-Yamamoto D, Takeuchi N, Masuda T, Murai Y, Onuma Y, Mori H, Masuyama N, Ishiguro S, Yachie N, Arakawa K*, "Complete genome sequence of *Psychrobacter* sp. strain KH172YL61,

isolated from deep-sea sediments in the Nankai Trough, Japan", *Microbiol Resour Announc*, 2019, 8:e00326-19. <https://mra.asm.org/content/8/16/e00326-19>

- Nagata S, Ii KM, Tsukimi T, Miura MC, Galipon J, Arakawa K*, "Complete genome sequence of *Halomonas olivaria*, a moderately halophilic bacterium isolated from olive processing effluents, obtained by nanopore sequencing", *Microbiol Resour Announc*, 2019, 8:e00144-19. <https://mra.asm.org/content/8/18/e00144-19>
- Saito M, Nishigata A, Galipon J, Arakawa K*, "Complete Genome Sequence of *Halomonas sulfidaeris* Strain Esulfide1 Isolated from a Metal Sulfide Rock at a Depth of 2,200 Meters, Obtained Using Nanopore Sequencing", *Microbiol Resour Announc*, 2019, 8(23):e00327-19. <https://mra.asm.org/content/8/23/e00327-19>
- Murai Y, Masuda T, Onuma Y, Evans-Yamamoto D, Takeuchi N, Mori H, Masuyama N, Ishiguro S, Yachie N, Arakawa K*, "Complete Genome Sequence of *Bacillus* sp. Strain KH172YL63, Isolated from Deep-Sea Sediment", *Microbiol Resour Announc*, 2020, 9:e00291-20. ([Publisher](#))