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クエ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、ナノポアシークエンサーにて 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クエ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、ナノポアシークエンサーにて 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クエ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　ナノポアシークエ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

■ファイル

冨田研ファイルサーバの

```
/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)、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。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。

■ 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、

```
awk 'BEGIN {OFS = "\n"} {header = $0 ; getline seq ; getline qheader ; getline qseq ; if (length(seq) >= 10000 ) {print header, seq, qheader, qseq}}' < input.fq > filtered-10000.fastq
```

のように（ここの input.fq を入力ファイル , あとは 10000 を任意の数字に）すると、任意の長さ以上の配列だけを取得できます。あるいは、BBMap の reformat.sh を使って、

```
/home/gaou/kumamushi/software_smith/bbmap/reformat.sh in=BC01 .fq out=BC01 -filter10 k.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。

```
/home/gaou/kumamushi/software_smith/bbmap/stats.sh filtered-10000.fastq
```

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。

```
[gaou@smith7 NanoPlot]$ NanoPlot -t 24 --fastq BC01.fq -p BC01
```

■ Canu でのアセンブリー

次に canu (https://github.com/marbl/canu/releases) をインストールします。Canu は最新版だと場合によってはまだバイナリが配布されていないので、自分でコンパイルが必要です。その場合、一つ前のバージョンだと Linux 版のバイナリが配布されています。指示にしたがってコンパイルするか、バイナリの場合は解凍してください。コンパイルの場合、Java のバージョンに依存性があるので、あまり古いOSだとコンパイルできない可能性がある点ご注意ください。kingではワーニングたくさんでますが問題なくコンパイルできました。

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

で king のノードにログインし、

```
/home/gaou/kumamushi/software_smith/canu-2.2/bin/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ね）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。

```
/home/gaou/kumamushi/software_smith/bbmap/stats.sh BC01/BC01.contigs.fasta
```

と打って、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。長さが 10kbp に満たない suggestCircular が no の contig はゴミの可能性が高いですので、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 CEGMA や BUSCO といった手法 (http://kazumaxneo.hatenablog.com/entry/2017/07/19/145640) を用います。これらのソフトウェアはインストールが面倒なのですが、理研が開発している gVolante というウェブサーバが非常に簡単に使えるようにしてくれています。https://gvolante.riken.jpBUSCO は v.1 がバクテリアに対応しています。

バクテリア用には CheckM というソフトウェアの方がより詳細に completeness を計算できます。CheckM は DFAST Quality Control ツール (https://dfast.nig.ac.jp/dqc/submit/) で簡単に検証できるので、Genome Report には BUSCO よりも CheckM の値を載せた方が良いです。

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)。

ナノポアの配列のみでアセンブルした場合、大抵の場合 BUSCO スコアが目安となる 90% 程度を著しく下回ります。これは、多くの場合 indel 系のエラーが解決されないことに起因します。

■ 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。ただし、Illumina リードがを用いた方が確実に良い結果が得られます。今回この授業では Illumina シーケンスを用意しましたので、これを使ってエラーコレクションを行いましょう。新規で読んでいて Illumina リードがない場合には nanopore オンリーのエラーコレクションを行います（非推奨)。

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。

```
 /home/gaou/kumamushi/software_smith/bwa-0.7.11/bwa index BC01.contigs.fasta
 /home/gaou/kumamushi/software_smith/bwa-0.7.11/bwa mem -t 8 BC01.contigs.fasta BC01_S1_merged_R1.fq
| /home/gaou/kumamushi/software_smith/samtools-1.9/samtools view -@ 4 -b -o aln.bam -
 /home/gaou/kumamushi/software_smith/samtools-1.9/samtools sort -T sort.tmp -o aln.sorted.bam -@ 4
aln.bam
 /home/gaou/kumamushi/software_smith/samtools-1.9/samtools index aln.sorted.bam
```

次に pilon。

```
java -Xms8g -jar /home/gaou/kumamushi/software_smith/pilon-1.23.jar --genome BC01.contigs.fasta
--bam aln.sorted.bam --threads 4 --output pilon1
```

pilon 後に再度 gVolante で BUSCO スコアを算出すると、大幅に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。

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_smith/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。

2. Nanopore での補正 nanopolish を使います。https://github.com/jts/nanopolishnanopolish index で fastq と fast5 を対応付け、fastq を bwa でリファレンスにマッピングし、実行します。マニュアルだと parallel を用いた方法が書いてありますが、以下のように実行するとシングルで実行できます。が、非常に時間がかかるので、parallel をインストールできる場合 parallel を使って 16 並列 x 4 スレッド、あるいは 8 並列 x 8 スレッドくらいでやったほうがいいかもしれません。

```
nanopolish variants --consensus -r BC01.dedup.fastq -b BC01/reads.sorted.bam -g BC01/BC01
.contigs.fasta -o BC01.nanopolished.fasta -t 64 -q dcm,dam -w tig00000001:1-3640229
```

マニュアルに書いていませんが、この時、-q dcm,dam オプションをつけることは重要です。これは DNA のメチル化を考慮に入れたベースコールをするオプションで、これが入るとエラーコレクションの精度が大分向上します。（バクテリアゲノムは特にメチレーションが多いため）

- king だと例によって warning は出ますが問題なくコンパイルできます。
- fastq エントリが duplicate だと言われる場合： /home/gaou/bin/fastq-dedup.pl BC01.fastq > BC01.dedup.fastq
- nanopolish index 時は -f オプションで sequencing_summary の場所を必ず指定すること（でないと異様に遅い）
- nanopolish の最適カバレッジは x100~200 です。これ以上ある場合には x200 以下までダウンサンプリングしてください。
- nanopolish index と bwa mem はいずれもかなり時間がかかるので、同時にかけると良いでしょう。
- nanopolish は異様に時間かかるので、先に Racon で様子を見てみるのもアリです。http://kazumaxneo.hatenablog.com/entry/2018/03/22/013006

■ 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 Annoucementshttps://mra.asm.org/content/getting-started

■ フォーマット

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

Abstract: 50 wordsTotal 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

■ 2021 年

- 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)