

地衛研 ゲノム解析精度管理マニュアル

細菌 SNP 解析編

Version 0.5

*本資料に記載されたプロトコルは研究および調査の目的に限り使用できるものであり、体外診断等の目的に使用されるものではありません。本資料の内容およびプロトコルの正当性はいかなる法令等により担保されるものではありません。使用される方ご自身の責任に基づき正しく実験を行いその結果を解釈してください。

変更履歴

- ・初版作成（）

集団感染事例における細菌 SNP 解析（情報解析）精度管理の概要

本マニュアルは、単一塩基多型（Single Nucleotide Polymorphism、以下 SNP）解析ツールとして SNPcaster（<https://github.com/leech-rr/SNPcaster>）を使用することを念頭に置いて作成されている。リードデータの品質評価（Quality Check、以下 QC）および SNP 解析両方で使用できる。

リードデータの QC

リードデータのトリミング

（アダプター配列や低クオリティデータの除去）



データカバレッジの計算

（解析対象微生物ゲノムに対して何倍量のデータが取得できたかの確認）



ドラフトゲノム配列のアセンブリ

（生物種、Total ゲノムサイズ、コンティグ数などの確認）



CheckM による最終評価

（生物種、コンタミネーションの確認）

集団事例における SNP 解析

対象となる微生物の参照株を選定（通常は標準株で構わない）



集団事例関連株のトリミング済みリードを参照株のゲノム配列にマッピング

（SNP の検出）



コアゲノムサイズの確認、系統解析を実施

（コアゲノムサイズは十分確保できているか、系統的に離れた株が解析に混ざっていないかを確認）



系統的に離れた株があった場合、それらを除いて再度 SNP 解析を実施



コアゲノムサイズが小さい場合など必要に応じて参照株を変更

（集団事例の 1 株を選択してアセンブルデータを用意し、新たな参照株として用いる）

1. はじめに

COVID-19 のパンデミック以降、次世代シーケンサー (NGS) を導入する地方衛生研究所が急増した。また、民間企業の NGS の受託解析についても以前より安価なものが増加しており、病原体のリードデータ取得のハードルは下がっている。一方で、NGS から得られるデータ量は膨大であり、データ解析のためにはバイオインフォマティクスの知識が必要となる側面があり、このことが病原体ゲノム解析におけるソフト面での問題となっている。病原細菌の分野では、集団事例における菌株解析技術として、主にパルスフィールドゲル電気泳動 (PFGE) が用いられてきたが、当該機器のサポート終了に伴う代替法の導入が急務となっている。

本マニュアルでは、これまで PFGE が担ってきた菌株の同一性に関する解析手法の代替法として、全ゲノム SNP 解析について解説する。これから全ゲノム SNP 解析を導入する施設などで、自身が採用する手法やデータが妥当であるかどうかの参考になれば幸いである。対象はショートリード (Illumina 等) を用いた単離株解析を想定している。ただし、「SNP が何個以内であれば同一の株とみなすことができるのか」という基準については、病原体ごとに状況が異なるため、汎用的な答えを用意するのは困難である。そこで本マニュアルでは、SNP 解析に至るまでのリードデータの精度や、解析に用いられた手法そのものの妥当性を評価する方法について紹介する。

なお、本マニュアルでは JIHS 国立感染症研究所細菌第一部で開発された SNPcaster (<https://github.com/leech-rr/SNPcaster>) の使用を前提とした記載となっているが、SNPcaster のインストールや使用方法については詳細には記載しない。インストールや使用方法については <https://github.com/leech-rr/SNPcaster/tree/main/doc/manual> より日本語マニュアルが入手可能である他、マニュアルを学習させた ChatGPT (URL: <https://chatgpt.com/g/g-69118b0d42f48191906ba605a540cba8-snpcaster-gpt>) において対話形式での情報入手が可能 (ただし無料ログインが必要) なので、それらを確認してほしい。

また、本マニュアルにはデータの準備や検証を行う際に一部 Linux を使用する箇所が存在するが、コマンドについての詳細な解説は省略する。Windows パソコン (PC) を用いた Linux コマンドでのゲノムデータ取り扱いについては、厚生労働行政推進調査事業費補助金健康安全・危機管理対策総合研究事業「地方衛生研究所におけるゲノム検査などに係る人員体制及び人材育成法を確立するための研究」(研究代表者 貞升健志) により実施された、ゲノムデータ解析研修会の研修資料が地方衛生研究所全国協議会のホームページ (<https://www.chieiken.gr.jp/>) に掲載されているので、そのマニュアルにより操作法を身に付けていただきたい。

2. リードデータの QC (品質評価)

適用範囲

- 対象：単離された病原細菌株
- 入力データ：ペアエンド FASTQ ファイル
- 推奨 PC スペック：メモリ 16-128 GB (サンプル数・アセンブル方法に依存)、CPU 8 コア以上推奨

解析ツール

本マニュアルは SNPcaster(<https://github.com/leech-rr/SNPcaster>)での操作を前提としている。ただし、市販のソフトウェアであっても、同等の内容が担保されていれば本マニュアルの流れに沿った解析を実行可能である。あるいは、独自にオープンソースのプログラムを組み合わせて解析することも可能である。各地衛研におけるバイオインフォマティクスの習熟度や予算を考慮して選択されたい。

QC フロー

ここではショートリードデータの品質確認方法について、以下の 4 ステップを紹介する。SNPcaster では 1 度の操作で下記の一連の解析が実行されレポートファイルが出力される。

1. トリミング (fastp など)
2. カバレッジ (深度) 確認 (SeqKit など)
3. アセンブルによるゲノムサイズ確認 (SPAdes, SKESA など)
4. CheckM (次ページ参照) による Completeness, Contamination 確認

カバレッジとはリードデータの厚みのことであり、下記の計算式により計算される。

$$\text{カバレッジ} = (\text{R1} + \text{R2 の総塩基長}) / \text{解析対象微生物のゲノムサイズ(bp)}$$



図 1. データカバレッジの概念図

信頼性のある SNP 解析を実施するうえで必要なカバレッジの目安は以下の通りである。

- ✓ 必要最低限:30x
- ✓ 推奨:50x 以上

カバレッジが 30x を下回るデータ量の検体を用いた解析では SNP の誤判定を生む可能性が高くなるため、このような検体は解析からは除外することを強く推奨する。
その他、1〜4 まで考慮した QC のチェックリスト（目安）を以下に記載する。

- ☐ カバレッジ 30x 以上
- ☐ アセンブルゲノムサイズ(標準株の±10%)
- ☐ CheckM completeness(97%以上)
- ☐ CheckM contamination(3%以下)

メモ

CheckM の completeness（完全性）と contamination（汚染度）は、ゲノムアセンブリの品質を評価する上で重要な指標となる。CheckM は分類系統（門・綱・目…）ごとにそのグループが必ず持っている必須遺伝子セットが定義されたデータベースがあり、completeness はその必須遺伝子がどれだけ揃っているかを示す指標で、100%に近いほど欠落が少なく完全性が高いと判断される。contamination は本来 1 コピーのはずの必須遺伝子が複数コピー検出された場合に、多生物由来の配列が混ざっている可能性があるとして数値化される。

QC の実行 (SNPcaster)

SNPcaster (Jupyter Lab 版) を使用する場合は、プロジェクト作成後にプロジェクトフォルダ内に「grape_qc_assembly.ipynb」という QC 解析用のノートブックが作成されているので、それを利用する。ノートブック内に詳細な解析手順についての記載があるので、それを確認しながら実行する。

なお「grape_qc_assembly.ipynb」で選択可能なアセンブラ（コンティグを作成するソフト）は SKESA と SPAdes との 2 種類があり、前者のほう解析に要する時間が短い。QC チェックリストさえクリアされれば、いずれのアセンブラを用いても問題ないが、集団事例中の株を参照配列に用いたい場合（3-4 で後述する）は SKESA の利用を推奨する。

QC を行う場合は、configuration.txt の作成を行う。まず、<https://github.com/leech-rr/SNPcaster> よりダウンロードした SNPcaster フォルダ内の app/notebook/sample_data/ehc 内に腸管出血性大腸菌用の「configuration.txt」ファイルがあるので、これを自身のプロジェクトフォルダにコピーする。初期値は右図の通りとなっているので、解析する菌種に応じて適宜編集すれば良い。
解析が終了すると、プロジェクトフォルダ内に qc_assembly_[日付]_[時間]_[リスト名] のフォルダが作製される。解析フォルダ内に「qc_results.xlsx」というファイルが出力されているので、

Genome size [Mbp]:	5.5
Coverage threshold:	40
Maximum number of contigs:	1000
Minimum genome size (Mbp):	4
Maximum genome size (Mbp):	6.5
Minimum completeness (%):	98
Maximum contamination (%):	2

qc のシートのデータを確認する。特に、coverage、total_length、completeness、contamination について確認し、その後の解析（SNP 解析等）に用いて良いかどうか判断する。

3. SNP 解析

本マニュアルでは、https://github.com/leech-rr/WGS_benchmark_datasetに記載されている、腸管出血性大腸菌の集団事例で解析されたリードデータ、を使用してSNPcasterで解析した例を示す。自施設にSNPcaster以外のツールを導入してSNP解析を試みる際も、このテストデータを使用して解析を実施し、SNPcasterの結果と比較して大きな齟齬が無いことを確認することを推奨する。

上記ページに記載のリードデータのアクセッション番号をもとに、データベースからダウンロードしても良いが作業が煩雑であるため、ここではLinux（Ubuntu）環境（WindowsのWSL2を含む）でコマンドを利用したデータの準備方法を紹介する。

なお、以下 の網掛けでハイライトした部分はUbuntuターミナルで使用するコマンドを表す。

Ubuntuのターミナルを起動し、デスクトップなどの任意の場所に `cd` コマンドで移動して下記コマンドをタイプする

```
git clone https://github.com/leech-rr/WGS_benchmark_dataset.git
```

「WGS_benchmark_dataset」というフォルダがダウンロードされるので、そこに `cd` コマンドで移動する。次に、https://github.com/leech-rr/WGS_benchmark_datasetの「How to download short-read files」を参考にダウンロードする。「fasterq-dump_2023-05-30.sh」の利用には、SRA Toolkitのインストールが必要である。自身のPC環境に応じて、必要であればconda環境を作成してインストールする。

*SRA-Toolkitインストール参考コマンド

```
conda create -n sratools sra-tools
```

その後は、「fasterq-dump_2023-05-30.sh」と「list」ファイルがフォルダ内にあることを確認して下記コマンドでリードデータがダウンロードできる。`THREADS`は計算に使用するCPUのスレッド数なので、自身のPC環境に応じて任意の数字に入れ替える。

```
fasterq-dump_2023-05-30.sh list THREADS
```

各ファイルサイズを巻末の表に示すので、問題なくダウンロードできたかどうかを確認する際に参考にされたい。

3-1. SNP解析の実行

ダウンロードしたデータを用いてSNP解析を実行する。マニュアルとノートブック内に詳

細な解析手順についての記載があるので、それを確認しながら実行する。参照配列には SNPcaster に付属している Sakai_BA000007.fasta を用いた。なお、SNPcaster は図2に示すフローで解析が進むパイプラインであり、BactSNP（SNP検出ソフト）にトリミングツールが付属しているため、解析にあたってはリードデータのトリミングの必要はない。

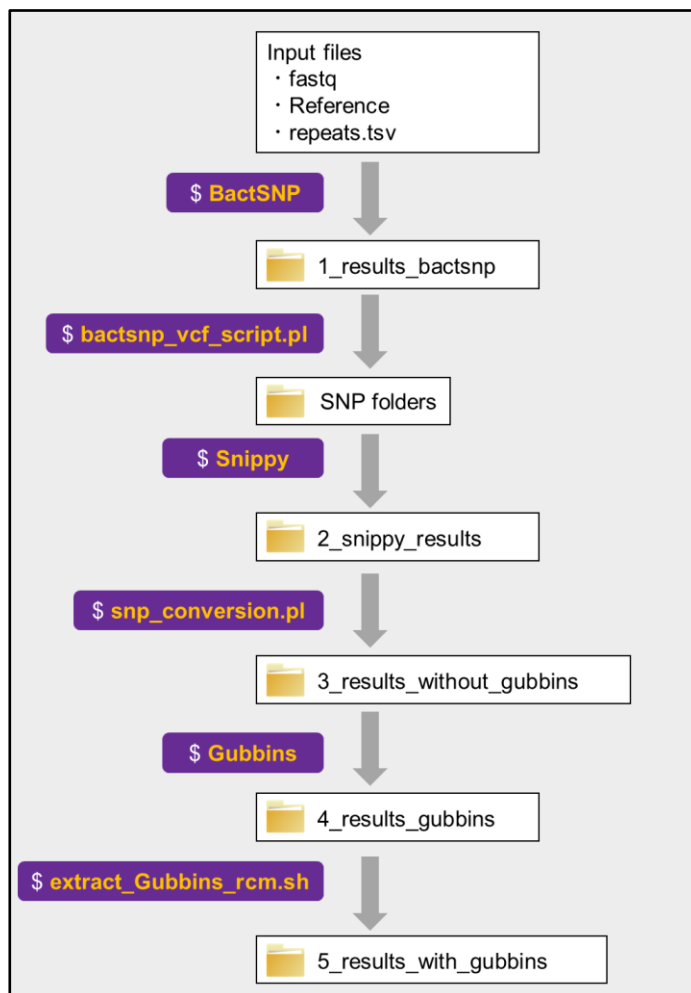


図2. SNPcasterの比較ゲノム解析フロー図

3-2. 全検体を用いたスクリーニング解析

本マニュアルでは、まず、ダウンロードした15検体（O157：10株、O121：5株）全てを用いて解析を実施した例を示す。これは、集団事例における分離株すべてを用いた解析に即した流れであると考えて良い。解析パラメータはほぼ基本値（変更なし）のままで、Gubbinsについては「実行する（1）」を選択した。解析が終了すると、プロジェクトフォルダ内に「snpcaster_[日付]_[時間]_[リスト名]」という名称のフォルダが作製され、出力ファイルが格納される。解析結果は「SNPcaster_report.txt」に集計されており、本解析では右記のようになった。

Core genome size (bp)
Before Gubbins : 4291372
After Gubbins : 4204425
Number of informative SNP sites (with reference)
No masking: 74374
After Gubbins : 65665
Number of informative SNP sites (without reference)
No masking: 74327
After Gubbins : 65627
Removed SNPs
Gubbins: 8709

コアゲノムサイズ（Core genome size）は参照配列の5.5Mに対して約4.3M（78%）であった。検出されたSNP数は組換え領域除去前で74,327個であり、多数のSNPsが検出されたことがわかる。それぞれの株間のSNP数をマトリックスで表示したファイルは、SNPcasterの出力フォルダ内の

「3_results_without_gubbins」（組換え領域除去前）もしくは「5_results_with_gubbins」（組換え領域除去後）に格納されている。図3は、そのうち組換え領域除去前のファイルを示したものである。このうち7万以上のSNPsが検出されている株を黄色でハイライトしているが、これらは血清型O121のグループである。

	REF	DRR14706	DRR14706	DRR14707	DRR14707	DRR14707	DRR14958	DRR14958	DRR14958	DRR14964	DRR14964	DRR14974	DRR14974	DRR14976	DRR14976	DRR14977
REF	0	73759	73758	73759	73759	73760	138	137	136	137	137	981	980	982	982	980
DRR147067	73759	0	1	2	2	3	73784	73783	73782	73783	73783	73845	73844	73846	73846	73844
DRR147069	73758	1	0	1	1	2	73783	73782	73781	73782	73782	73844	73843	73845	73845	73843
DRR147072	73759	2	1	0	2	3	73784	73783	73782	73783	73783	73845	73844	73846	73846	73844
DRR147075	73759	2	1	2	0	3	73784	73783	73782	73783	73783	73845	73844	73846	73846	73844
DRR147079	73760	3	2	3	3	0	73785	73784	73783	73784	73784	73846	73845	73847	73847	73845
DRR149584	138	73784	73783	73784	73784	73785	0	3	2	3	3	1025	1024	1026	1026	1024
DRR149590	137	73783	73782	73783	73783	73784	3	0	1	2	2	1024	1023	1025	1025	1023
DRR149598	136	73782	73781	73782	73782	73783	2	1	0	1	1	1023	1022	1024	1024	1022
DRR149646	137	73783	73782	73783	73783	73784	3	2	1	0	2	1024	1023	1025	1025	1023
DRR149690	137	73783	73782	73783	73783	73784	3	2	1	2	0	1024	1023	1025	1025	1023
DRR149748	981	73845	73844	73845	73845	73846	1025	1024	1023	1024	1024	0	1	3	3	1
DRR149750	980	73844	73843	73844	73844	73845	1024	1023	1022	1023	1023	1	0	2	2	0
DRR149762	982	73846	73845	73846	73846	73847	1026	1025	1024	1025	1025	3	2	0	4	2
DRR149766	982	73846	73845	73846	73846	73847	1026	1025	1024	1025	1025	3	2	4	0	2
DRR149771	980	73844	73843	73844	73844	73845	1024	1023	1022	1023	1023	1	0	2	2	0

図3. dist_final_snp_matrix.tsv（全検体）の中身

3-3. 遠縁菌株の除外

集団事例におけるSNP解析では、異なる血清型やSequence Type (ST) の菌株は、解析群からは除外することが望ましい。3-2.のような例では、2つの血清型 (O157とO121) が含まれており、ここでは血清型O121を除いた10検体を解析することとした。その結果の集約 (SNPcaster_report.txt) は右記の通りである。コアゲノムサイズは4.9M (89%) まで上昇している一方で、総SNP数は組換え領域除去前で1,578個まで減少した。

Core genome size (bp)
Before Gubbins : 4959266
After Gubbins : 4916717
Number of informative SNP sites (with reference)
No masking: 1578
After Gubbins : 750
Number of informative SNP sites (without reference)
No masking: 1521
After Gubbins : 701
Removed SNPs
Gubbins: 828

それぞれの株間のSNP数マトリックス表を図4に示す。血清型O157株間のSNP数は、図3および図4において緑色でハイライトされた部分に相当する。血清型O121を除外する前後のO157株間のSNP数を比較すると、SNP数が4桁を超える株間では、O121を除外した場合に検出SNP数が500個近く増加しているのがわかる。これは、遠縁菌株であるO121を除去したことによってコアゲノム領域が拡張され、検出対象となったSNPsが増加したことに起因する。一方で、このような場合でもSNP数が1桁の株間では大きな変化は認められなかったのは、ほとんどの検出SNPsが最初の解析の際に定義されたコア領域に含まれていたからである。

	REF	DRR14958	DRR14959	DRR14959	DRR14964	DRR14969	DRR14974	DRR14975	DRR14976	DRR14976	DRR14977
REF	0	256	255	254	255	255	1372	1371	1373	1373	1371
DRR149584	256	0	3	2	3	3	1514	1513	1515	1515	1513
DRR149590	255	3	0	1	2	2	1513	1512	1514	1514	1512
DRR149598	254	2	1	0	1	1	1512	1511	1513	1513	1511
DRR149646	255	3	2	1	0	2	1513	1512	1514	1514	1512
DRR149690	255	3	2	1	2	0	1513	1512	1514	1514	1512
DRR149748	1372	1514	1513	1512	1513	1513	0	1	3	3	1
DRR149750	1371	1513	1512	1511	1512	1512	1	0	2	2	0
DRR149762	1373	1515	1514	1513	1514	1514	3	2	0	4	2
DRR149766	1373	1515	1514	1513	1514	1514	3	2	4	0	2
DRR149771	1371	1513	1512	1511	1512	1512	1	0	2	2	0

図4. dist_final_snp_matrix.tsv (O121除外) の中身

3-4. 集団事例中の1株を参照株とした解析

3-3. で示したO157 Sakai株を参照配列とした解析では、解析対象株を血清型O157のみに限定することで十分なコア領域が確保され、近接な菌株は検出SNP数が1桁となった。このような明解なケースでは通常必要ないと考えられるが、本マニュアルでは更に集団事例のうちの1株を参照配列として解析する例を示す。

QCの実行 (SNPcaster) の項目でアセンブルされたfastaファイルのうち、解析したい集団株から1株を選択し、SNPcaster.ipynbの「6. 参照配列をつなげて一つのFASTAファイルにする」を用いて、ドラフトゲノムのコンティグを連結してSNPcasterで使用可能な参照配列を作成する。その後は、作成された配列を参照配列としてSNPcasterを実行すれば良い。なお、この方法で作成された配列を参照配列とする場合は、Gubbinsによる組換え領域の除去は実施できないので注意されたい。今回はDRR149598を参照株として、それとSNP数が1桁であった近縁の5株を解析した。その結果の集約 (SNPcaster_report.txt) は右記の通りである。コアゲノムサイズは5.2M (96%) となり、参照配列のほぼ全域がカバーされている一方で、検出SNP数は6個にまで減少した。それぞれの株間のSNP数マトリックス表を図5に示す。これまでの解析と比較して、わずかに検出SNP数の変化がみられる。これは、より近接な菌株ゲノムを参照配列としたことによってコア領域がさらに拡張され、SNP解析の解像度がさらに増加したことによるものである。

```
Core genome size (bp)
Before Gubbins : 5201544

Number of informative SNP sites (with reference)
No masking: 6

Number of informative SNP sites (without reference)
No masking: 6
```

	REF	DRR14958	DRR14959	DRR14959	DRR14964	DRR14969
REF	0	2	1	0	1	2
DRR149584	2	0	3	2	3	4
DRR149590	1	3	0	1	2	3
DRR149598	0	2	1	0	1	2
DRR149646	1	3	2	1	0	3
DRR149690	2	4	3	2	3	0

図5. dist_final_snp_matrix.tsv (近接株のみ) の中身

3-2から3-4までで述べたように、コアゲノムSNPsの検出は「解析するサンプル群」や「参照配列」によって大きな影響を受ける。そのため、集団事例におけるコアゲノムSNP解析においては「近接な株同士を比較すること」が大前提となる。集団事例にNGS解析を適用する際はこれを念頭に置き、解析しているサンプル群に遠縁の株が混ざっていないか、コアゲノム領域が十分に確保されているか (概ね75%以上を推奨)、について確認するよう努める必要がある。

冒頭でも述べたように、今後自施設において集団事例のSNP解析を導入する際には、本マニュアルに記載のデータを参考にして、自施設の解析環境が適切に機能するかどうかを確認してから解析を実行することを推奨する。

謝辞

本プロトコルの作成は、下記の支援を受け実施しました。

厚生労働行政推進調査事業費補助金 健康安全・危機管理対策総合研究事業「地方衛生研究所におけるゲノム検査などに係る人員体制及び人材育成法を確立するための研究」（研究代表者 貞升 健志）

別表. テストデータのファイルサイズ

DRR147067_R1.fastq.gz	105,783K
DRR147067_R2.fastq.gz	137,620K
DRR147069_R1.fastq.gz	112,819K
DRR147069_R2.fastq.gz	147,869K
DRR147072_R1.fastq.gz	110,675K
DRR147072_R2.fastq.gz	133,088K
DRR147075_R1.fastq.gz	95,208K
DRR147075_R2.fastq.gz	125,344K
DRR147079_R1.fastq.gz	104,785K
DRR147079_R2.fastq.gz	132,131K
DRR149584_R1.fastq.gz	144,066K
DRR149584_R2.fastq.gz	180,278K
DRR149590_R1.fastq.gz	116,620K
DRR149590_R2.fastq.gz	144,171K
DRR149598_R1.fastq.gz	154,462K
DRR149598_R2.fastq.gz	196,001K
DRR149646_R1.fastq.gz	103,696K
DRR149646_R2.fastq.gz	123,745K
DRR149690_R1.fastq.gz	146,901K
DRR149690_R2.fastq.gz	175,732K
DRR149748_R1.fastq.gz	183,890K
DRR149748_R2.fastq.gz	216,995K
DRR149750_R1.fastq.gz	163,170K
DRR149750_R2.fastq.gz	198,485K
DRR149762_R1.fastq.gz	187,530K
DRR149762_R2.fastq.gz	225,209K
DRR149766_R1.fastq.gz	160,643K
DRR149766_R2.fastq.gz	188,637K
DRR149771_R1.fastq.gz	89,887K
DRR149771_R2.fastq.gz	107,189K