

厚生労働行政推進調査事業費補助金（健康安全・危機管理対策総合推進事業）
分担研究報告書

次世代シーケンサーに関わる各種検査マニュアルの作成：
疫学調査における NGS 解析の利用に関する研究

研究分担者 久保田 寛顕 東京都健康安全研究センター 統括課長代理
野本 竜平 神戸市健康科学研究所 副部長

研究要旨

集団食中毒をはじめ病原性細菌に起因する集団感染が発生した場合、地方衛生研究所ではパルスフィールド電気泳動（PFGE）による系統解析を行い、その結果をもって保健所等と連携し、事例究明と収束に向けた対応に当たってきた。今後、PFGE に代わり、一塩基単位の解像度で全ゲノム比較を行うことが可能な次世代シーケンサー（NGS）を用いた菌の比較が求められる。

本研究課題では NGS を用いた数値化の方法として一塩基多型（SNP）解析を行うことを前提とした。その際、検出される一塩基変異（SNV）は解析方法によって変化しうることから、特に影響を与える要因を特定したところ、リファレンスの選定や、遠縁株を含むか否かという条件によって検出される SNV の数が影響されることを確認した。

検討結果を踏まえ、最終的に SNP 解析精度管理用マニュアルの草案を作成した。本マニュアルでは最初に粗く SNP 解析をすることで遠縁株を判断し、近縁株のみで再度 SNP 解析を行う二段階のワークフローを採用した。各地衛研が同等の解析を行うための精度保証として、リードデータ品質の基準の統一ならびに SNP 解析の方法の指定を盛り込んだ。今後は本マニュアルの完成版をウェブ上で公開するとともに、地衛研ネットワークの中で研修を実施するなどし、全国地衛研への浸透を図っていく。

A. 研究目的

集団食中毒をはじめとする病原体による集団感染が発生した際、全国の地方衛生研究所（以下、地衛研）では本庁、保健所等との連携のもと、事例の究明と収束に向けた対応にあたる。この際、地衛研では主としてヒトや環境サンプルなどの分析を行い、科学的データを与えることをもって疫学調査をより強固なものとする役割を担う。特に腸管出血性大腸菌など、細菌が原因となる集団感染が起きた際は、分離株の一致、不一致を見る方法として、これまでパルスフィールド電気泳動（PFGE）法がゴールドスタンダードとして用いられてきた。しかし、PFGE 法に用いる標準的な泳動装置が今後

製造終了になり、使用不能となることが知られている。このため、集団感染を追跡する上で、PFGE 法に代わる方法の策定が喫緊の課題である。

そこで、本研究課題では PFGE の代替法として次世代シーケンサー（NGS）の活用を検討する。従来、菌型を比較する方法として PFGE は最も解像度の良い方法であったが、NGS の登場により一塩基単位での比較が可能となり、PFGE と同等か、それ以上の解像度で集団感染を追跡できるようになってきた。その一方で、NGS は機器から出力される膨大なデジタルデータをバイオインフォマティクスにより処理して結果としていく必要がある。また、疫学調査との比

較として一定の方法論（表 1）がコンセンサスを得ている PFGE に比べると、NGS では一塩基多型（SNP）解析によって得られる一塩基変異（SNV）を指標とした方法論がいまだ統一されていない状況にある。

表 1 PFGE における解釈
(Tenover et al., 1995. JCM)

相違バンド数	近縁関係	疫学的解釈
0 本	区別不能	集団感染
2 から 3 本	近縁	集団感染の可能性が高い
4 から 6 本	近縁の可能性	集団感染かもしれない
7 本以上	遠縁	非集団感染

そこで、NGS を実務的な集団感染対応へ活用するためにも解析手法の統一化やその精度を保証する仕方を整理するための検討を行っていく。なお、次世代シーケンサーにより生データを取得する工程、すなわちウェットの作業については地衛研ゲノム解析ワーキンググループにおける統一的マニュアルの作成と重複する。さらに、実際に行政依頼に対して保健所等に結果を発行する形式に関しては各自治体に適した形があると考えられる。したがって、本研究課題ではその間をつなぐデータ解析部分を対象とする（図 1）。

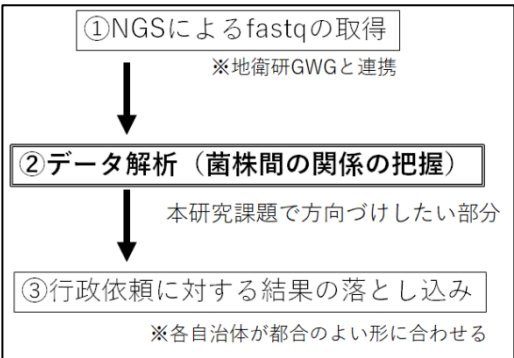


図 1 地衛研における NGS 対応

B. 研究方法

1. SNP 解析に影響を与える解析条件の特

定

集団感染時には PFGE のバンド数の違いのように、SNV 数という数値化された菌株間の差異を判断材料とすることが想定される。そこで、SNP 解析のマニュアル化を行う上で重要と考えられる、解析結果に対して影響を与えうる要因を特定するために同一データに対して解析条件を変化させ、その結果を評価することとした。

まず、テストサンプルとして、過去に分離された *Acinetobacter baumannii* 14 株（うち、13 株は ST208 であり、残り 1 株は ST1316）について、SNPcaster（<https://github.com/leechrr/SNPcaster/>）により SNV を検出し、リファレンス配列の違い等に起因する解析結果の変動について評価を行った。

2. 集団感染対応を目的とした精度管理マニュアルの作成

1. で特定された SNP 解析に与える解析条件を念頭に置き、地衛研が実際に集団感染に対応するにはどのような手順を踏むべきかについて、ワークフローを整えるとともに、SNPcaster の活用を前提とした検査マニュアルの作成を行った。SNPcaster のテストデータを用いて実際に SNP 解析を実施し、その結果を参考情報としてマニュアル内に記載した。また、これらのことについて、ゲノムデータ解析研修において一部の地衛研に対して情報共有を図った。

（倫理面への配慮）

本研究で用いた臨床分離株は、分離元とは連結不能な匿名化がされていることに加えて、本研究課題では一切の個人情報を用いず、個人が特定されることはない。また、東京都健康安全研究センター倫理審査委員会により、承認を受けているものである（審査番号：6 健研健第 906 号）

C. 研究結果

1. SNP 解析に影響を与える解析条件

NGS で標準株をリファレンスとした場合、リファレンスを標準株から ST208 株に変更した場合、ならびに解析対象から

ST1316 の一株を除いた場合に算出された SNV 数の変化を図 2 に示した。比較しやすいように、標準株と同じ SNV 数だった場合の理論直線を同時に示している。

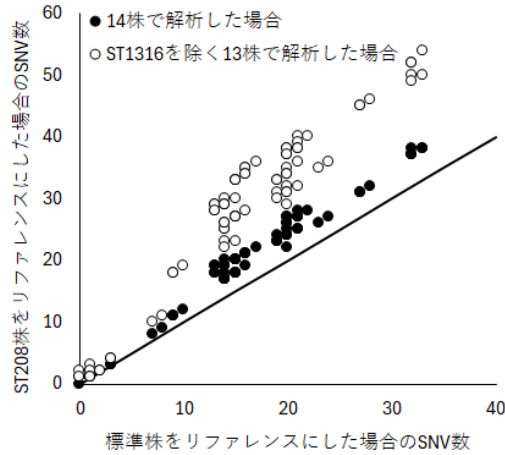


図 2 リファレンス、解析株数による変化

今回の解析でリファレンスやサンプルセットを変更した際のコアゲノム（全株に共通の領域）の変化を図 3 に、またそれぞれの解析での総 SNP 数と組換え領域内として除去された SNP 数を表 2 に示した。

図 3. 各解析によるコア領域の変化

表 2 各解析における SNP 数

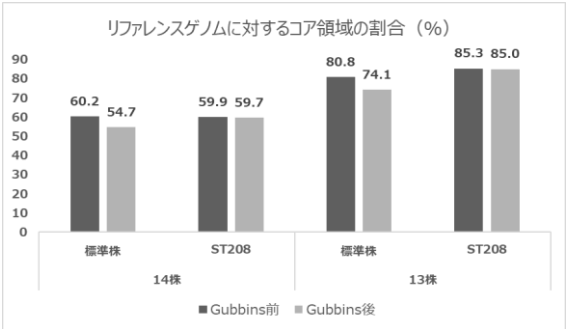
解析株数	14 株		13 株	
	標準株	ST208	標準株	ST208
総 SNP 数 (Gubbins 前)	242,951	228,239	55,036	648
除去された SNP 数	23,950 (9.9%)	924 (0.4%)	17,392 (31.6%)	478 (73.8%)

リファレンスゲノムに対するコアゲノム領域の割合は 14 株で解析した場合は 60% 前後であったのに対し、異なる ST の 1 株を除いた 13 株の解析では 80~85% となった。一方でリファレンスを変更した場合はコアゲノム領域の割合に与える影響は軽微であった。

また、各解析における総 SNP 数は 14 株で解析した場合はリファレンスに関わらず

20 万個を超える SNP が検出された一方で、13 株での解析では標準株をリファレンスにした場合は約 55,000 個、同一 ST をリファレンスにした場合は約 650 個の SNP となった。Gubbins による組換え領域除去を実施すると、14 株、13 株どちらの解析においても、リファレンスが標準株の場合は 17,000~23,000 個程度の SNP が除去され、リファレンスを ST208 にした場合は数百個の SNP のみが除去された。

算出される SNV 数がリファレンスとする配列や、解析に含む株によっても変動することが確認された一方で、コアゲノム領域については、リファレンス配列も含めた解析配列全体で判定されるため、今回の解析において異なる ST が混入していた 14 株の解析では大きくコアゲノム領域が減少することとなった。コアゲノム領域外にある SNP は解析には含まれないため、解析の分解能を向上させるためには遺伝系統的に離れた株は解析から排除することが望ましいと考えられる。また、組換え領域の除去に関してはリファレンスの影響が大きかった。これは Gubbins というソフトの仕様上、一株のみに検出される SNV では組換え領域



として判定されないという事が寄与している。今回の解析では遺伝系統的に離れた株は 1 株だったため、その株を除去するかどうかは組換え領域の解析ではあまり影響せず、リファレンスを変更することにより近接 SNV がどれだけ検出されるかどうかの方が影響が大きかったと考えられる。

2. SNP 解析精度管理用マニュアル

前項において特定された要因を考慮することで、マニュアル化におけるワークフローには二段階の SNP 解析を採用すること

が望ましいと考えられた（図 3）。

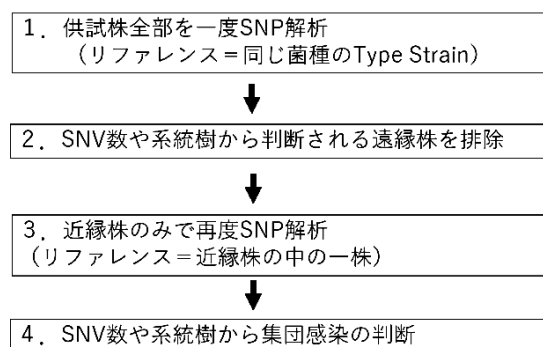


図 3 集団感染対応時のワークフロー

集団感染はいつ何時、どのような菌種、系統の病原細菌によって引き起こされるかはわからず、データベース上から取得することを前提とした場合に常に理想的なリファレンスが準備できるとは限らない。そこで、あらゆる菌種に対応可能となるように、集団感染を構成するうちのひとつの分離株を SNP 解析のリファレンスとして用いる形を考えた。その一方、全て菌種に対して菌種的一致、不一致を決める統一的な SNV の線引きを設定することは避けることとした。また、これらの内容についてゲノムデータ解析研修・中国四国地区研修会（2025 年 12 月実施）において説明し、同地区の地衛研との情報共有を行った。

最終的に、別添のとおり 2026 年 2 月末日の時点でのマニュアル草案である「地衛研 ゲノム解析精度管理マニュアル SNP 解析編（仮）」を作成した。本マニュアルは、SNPcaster の使用を前提とするほか、例えば近縁であっても低品質のリードデータが含まれると、その株のみならず全体の解析結果に悪影響を及ぼす可能性があることから、最低限必要なデータ深度の確認やコンタミネーションの有無など、リードデータの品質評価を行ったうえでの SNP 解析の具体的な手順について定めた。

また、本マニュアルは各地衛研の事情に応じて柔軟に解析手法を選択できるように、SNPcaster の使用を必須とするものではなく、オープンソース、あるいは市販のソフト

ウェアを用いて解析環境を整える場合に、SNPcaster のテストデータを使用して解析環境の検証を実施することを推奨した。そのため、データベースからのテストデータのダウンロード方法についても詳細に記述した。

テストデータを用いた解析では、1. で特定した解析条件を反映し、最初に異なる血清型も含む腸管出血性大腸菌の SNP 解析を実施し、続いて遠縁株を排除した解析、最後に集団事例中の 1 株をリファレンス株として用いる解析、という流れを結果を付けて記載した。

D. 考察

本研究課題において解析結果に強い影響を及ぼすと考えられた要素はリファレンスの選定と、遠縁株の含有であった。このことから、マニュアルでは二段階の SNP 解析を取り入れ、さらに一定の精度管理基準を設けた。統一的なマニュアルを提案することにより、複数の自治体間にまたがる大規模な集団感染事例が発生した場合にも地衛研間の解析結果をすり合わせ、広域的な連携をとることが可能となる。

この点において、NGS による集団感染の解析は単純に PFGE の代わりとなるだけでなく、PFGE を超える利点も生まれることが期待される。というのは、複数の自治体間にまたがる大規模な集団感染事例が発生した場合、PFGE ではひとつの地衛研に菌株を集め、同一手順によるプラグ作成から電気泳動を行う必要があった。NGS では分析済みの電子データの送付を行うだけで迅速に集約した解析が可能となる。

自治体間でのデータ共有は行政検査の秘匿性から容易ではないことも考えられるが、SNPcaster では生データを用いることなしに、途中まで解析したデータを共有することでその後の SNV 算出まで実施することができ、一定の秘匿性が担保される。こうした点においても菌株の共有に比べると、自治体間の連携に向けたハードルが低くなることが期待される。

E. 結論

集団感染を追跡するツールとして PFGE から NGS へと移行するにあたり、まず SNV 数による線引きを決めていくことに目が行きがちであるが、そうした基準ばかりが先行し、解析方法由来で SNV の検出数が変化してしまっは元も子もない。全国地衛研において同等の検査精度を確保するためには解析方法についても一定の道筋を指し示すことが重要である。本研究課題で作成した解析マニュアルの草案を元に、地方衛生研究所全国協議会と連携するなどし、迅速に全国地衛研との意識共有を進める。完成したマニュアルは、同協議会のホームページを活用して公開することを想定している。

さらに、地衛研間のネットワークを通じ、全国、あるいは地方ごとの集合研修または Web 研修によりマニュアルの浸透ならびに継続的なフォローアップを図っていくことが求められる。

F. 健康危険情報 なし

G. 研究発表

1. 論文発表

- 1) Mitobe M, Kubota H, Kobayashi K, Miyake H, Takano M, Mizushima D, Gatanaga H, Oka S, Suzuki J, Sadamasu K. Clustering of Polymorphic Membrane Protein E Clade in *Chlamydia trachomatis* Lineages from Men Who Have Sex with Men. *Emerg Infect Dis.* 30(10):2183-2187. 2024.
- 2) Uchitani Y, Okuno R, Ariyoshi T, Kubota H, Suzuki J, Sadamasu K. Genetic characteristics of invasive pneumococcal disease-derived *Streptococcus pneumoniae* of serogroup 24 isolated in Tokyo, Japan. *J Infect Chemother.* 202. 31(1):102484. 2025.
- 3) Ando N, Mizushima D, Takano M, Mitobe M, Kobayashi K, Kubota H, Miyake H, Suzuki J, Sadamasu K, Aoki T, Watanabe K, Oka S, Gatanaga H. Prolonged sitafloxacin and doxycycline combination regimen for treating infections by highly resistant *Mycoplasma genitalium*. *J Antimicrob Chemother.* 2025. 80(1):247-253. 2025.
- 4) Kubota H, Okuno R, Kenri T, Uchitani Y, Ariyoshi T, Yoshida I, Kobayashi K, Mitobe M, Suzuki J, Sadamasu K. Multiplex Amplicon Sequencing for the Comprehensive Genotyping of *Mycoplasma pneumoniae*. *Microbiol. Spectr.* 13:e0271924. 2025.
- 5) Ishitsuka T, Sugiyama M, Kubota H, Takai S, Kakuda T, Suzuki Y. Isolation of *Rhodococcus equi* bacteriophage ReT1 and investigation of its bactericidal effect. *Microb Pathog.* 207:107868. 2025.
- 6) Aoki W, Uwamino Y, Niida N, Kubota H, Kamoshita Y, Inose R, Nagata M, Ishihara O, Uno S, Yoshifuji A, Namkoong H, Hasegawa N, Matsushita H. Cefiderocol susceptibility of *Stenotrophomonas maltophilia* species complex and carbapenem-resistant *Pseudomonas aeruginosa* isolates from blood cultures at a university hospital in Tokyo, Japan. *J Glob Antimicrob Resist.* 44:251-255. 2025.
- 7) Aoki W, Uwamino Y, Kubota H, Kamoshita Y, Inose R, Nagata M, Namkoong H, Matsushita H. Evaluation of in vitro efficacy of aztreonam-nacubactam and cefepime-nacubactam against clinical isolates of *Stenotrophomonas maltophilia*. *Antimicrob Agents Chemother.* 69:e0075525. 2025.
- 8) Kobayashi K, Kubota H, Ariyoshi T, Suzuki Y, Mori K, Suzuki J, Ochiai Y, Sadamasu K, Chiba T. Differential Copy Number of Chromosomal blaCTX-M-14 in *Escherichia coli* Sequence Type ST38 from Companion Dogs and Cats. *Curr Microbiol.* 83:189. 2026.
- 9) Nakamura K, Taniguchi I, Gotoh Y, Isobe J, Kimata K, Igawa Y, Kitahashi T, Takahashi Y, Nomoto R, Iwabuchi K,

- Morimoto Y, Iyoda S, Hayashi T. Diversity of Shiga toxin transducing phages in *Escherichia coli* O145:H28 and the different Shiga toxin 2 production levels associated with short- or long-tailed phages. *Front Microbiol.* 15:1453887. 2024.
- 10) Nomoto R, Nakanishi N, Komatsu S, Matsui M, Suzuki S. Complete genome sequences of *Escherichia coli* KA0011 clinical isolate used as a quality control strain of carbapenem susceptibility testing in Japan. *Microbiol Resour Announc.* 13:e0015524. 2024.
- 11) Nomoto R, Osawa K, Kinoshita S, Kitagawa K, Nakanishi N, Sarassari R, Raharjo D, Fujisawa M, Kuntaman K, Shirakawa T. Metagenome and Resistome Analysis of Beta-Lactam-Resistant Bacteria Isolated from River Waters in Surabaya, Indonesia. *Microorganisms.* 12:199. 2024.
- 12) Tanimoto Y, Ohyama M, Ito E, Akiyoshi K, Onishi Y, Mori A, Nomoto R. Whole genome-based surveillance for human adenovirus-related diseases in Kobe City, Japan, 2018-2022. *BMC Res Notes.* 2025;18:170.
- 13) Komatsu S, Nomoto R, Nakanishi N. Stability of a blaGES-5-carrying plasmid in a *Serratia marcescens* nosocomial outbreak. *J Infect Chemother.* 2025;31:102742.
- 14) Hiratsuka T, Nakanishi N, Akita H, Masuda K, Komatsu S, Nomoto R, Amemura-Maekawa J. A large outbreak investigation of Legionnaires' disease associated with a public bath facility in Hiroshima, Japan, using PFGE, SBT, MLVA, and whole-genome sequencing. *Int J Med Microbiol.* 2025 ;321:151680.
- 15) Izuta K, Nishikubo M, Nomoto R, Sumitomo S, Matsuoka Y, Seo R, Nishioka H, Ariyoshi K, Sando E. Systemic lupus erythematosus mimicking Japanese spotted fever due to *Rickettsia japonica* detection in an eschar: a case report. *BMC Infect Dis.* 2026. Epub ahead of print.
- 16) 長島真美, 藤原卓士, 村田ゆかり, 長谷川乃映瑠, 有吉司, 上原さとみ, 小西典子, 門間千枝, 浅倉弘幸, 林真輝, 熊谷遼太, 原田幸子, 永野美由紀, 山崎貴子, 河上麻美代, 糟谷文, 矢尾板優, 黒木絢士郎, 天野有紗, 北村有里恵, 磯貝まや, 小杉知宏, 加來英美子, 鈴木薫, 南須原亮, 田部井由紀子, 小川麻萌, 中村三琴, 後藤千恵, 村内このみ, 古谷実, 原田順子, 川合由華, 小池浩二, 灘岡陽子, 中村貴恵子, 久保田寛顕, 安中めぐみ, 吉田勲, 小林甲斐, 水戸部森歌, 奥野ルミ, 内谷友美, 田淵優里, 森功次, 高橋由美, 村上昂, 和田紀乃, 河村真保, 赤瀬悟, 三関詞久, 中里彩乃, 小野明日香, 浅山睦子, 新井麻里, 鈴木愛, 岡田若葉, 古田菜摘, 齊木大, 前田雅子, 尾畑浩魅, 西野由香里, 井田美樹, 宗村佳子, 千葉隆司, 新開敬行, 横山敬子, 三宅啓文, 鈴木淳, 貞升健志. 健康危機管理に関連する微生物の分子疫学解析と検査法の開発に関する研究. 東京健安研セ年報. 75. 27-40. 2024.
- 17) 神田真軌, 小林甲斐, 西野由香里, 大場由実, 中島崇行, 栗田翔, 久保田寛顕, 有吉司, 神門幸大, 吉田勲, 安中めぐみ, 水戸部森歌, 長谷川乃映瑠, 尾畑浩魅, 畠山薫, 福井理恵, 市川めぐみ, 三橋華子, 黒田寿美代, 山崎華恵子, 井田美樹, 小林真紀子, 上原さとみ, 高橋由美, 高林晃子, 浅山睦子, 吉川聡一, 林洋, 松島陽子, 小池裕, 林もも香, 永野智恵子, 富澤早苗, 八巻ゆみこ, 上條恭子, 山本和興, 高田朋美, 小鍛冶好恵, 渡邊趣衣, 志良堂裕子, 大澤佳浩, 小山彩音, 野口舞子, 大貫文, 岡優香, 山崎貴子, 相原三菜子, 渡邊喜美代, 鈴木淳也, 小田智子, 鈴木淳, 横山敬子, 大塚健治, 横山知子, 新藤哲也, 中嶋順一, 木下輝昭, 貞升健志, 笹本剛生, 猪又明子. ワンヘルスアプローチに基づく抗微生物薬と薬剤耐性微生物の実態把握に関する研究. 東京健安研セ年報. 75. 41-57. 2024.
- 18) 安中めぐみ, 長谷川乃映瑠, 久保田寛顕, 吉田勲, 鈴木淳, 貞升健志. 結核菌検査におけるゲノム解析法に関する検討. 東

京健安研セ年報. 75. 115-118. 2024.

なし

19) 安中めぐみ, 久保田寛顕, 長谷川乃映瑠, 伊藤仁, 小林甲斐, 森功次, 貞升健志, 千葉隆司. 結核菌における薬剤耐性遺伝子変異検出の有用性の検討. 東京健安研セ年報. 76. 57-61. 2025.

2. 実用新案登録

なし

3. その他

なし

20) 久保萌加, 濱畑啓悟, 森愛, 野本竜平, 磯目賢一, 川崎悠, 岩田あや, 松原康策. エコーウイルス 11 髄膜炎の小児例:神戸市 2 施設の後方視的観察研究 2023~2024. 小児感染免疫. 37. 165-175. 2025

21) 平良由貴, 楫理恵子, 須賀知子, 伏屋智明, 谷本佳彦, 谷本佳彦, 森愛, 野本竜平. 2024 年 8 月に神戸市で検出された麻疹ウイルス遺伝子の分子疫学解析. IASR. 46. 63-64. 2025

22) 中西典子, 藤永千波, 小松頌子, 田中忍, 向井健悟, 野本竜平, 塩谷紀代, 森本るみ, 潮見千賀, 菅澄子, 宗田由紀, 楠信也. 神戸市における *Streptococcus pyogenes* M1UK 系統株の発生動向と分子疫学解析 感染症学雑誌. 99. 301-309. 2025

2. 学会発表

1) 小林甲斐, 久保田寛顕, 水戸部森歌, 有吉司, 鈴木淳, 貞升健志. 染色体上の *blactX-M-14* の増幅が認められたイヌおよびネコ由来基質特異性拡張型 β -ラクタマーゼ産生大腸菌 sequence type 38 の分子疫学的性状. 第 36 回日本臨床微生物学会総会・学術総会. 2025.

2) 内谷友美, 奥野ルミ, 有吉司, 吉田勲, 久保田寛顕, 鈴木淳, 貞升健志. 侵襲性肺炎球菌感染症由来株の血清型と血清群 24 の分子疫学解析. 第 36 回日本臨床微生物学会総会・学術総会. 2025.

3) 有吉司, 内谷友美, 奥野ルミ, 吉田勲, 久保田寛顕, 鈴木淳, 貞升健志. 都内医療機関から分離された *bla*NDM 搭載 IncX3 プラスミド保有腸内細菌目細菌の分子疫学解析. 第 36 回日本臨床微生物学会総会・学術総会. 2025.

H. 知的財産権の出願・登録状況

1. 特許取得

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。

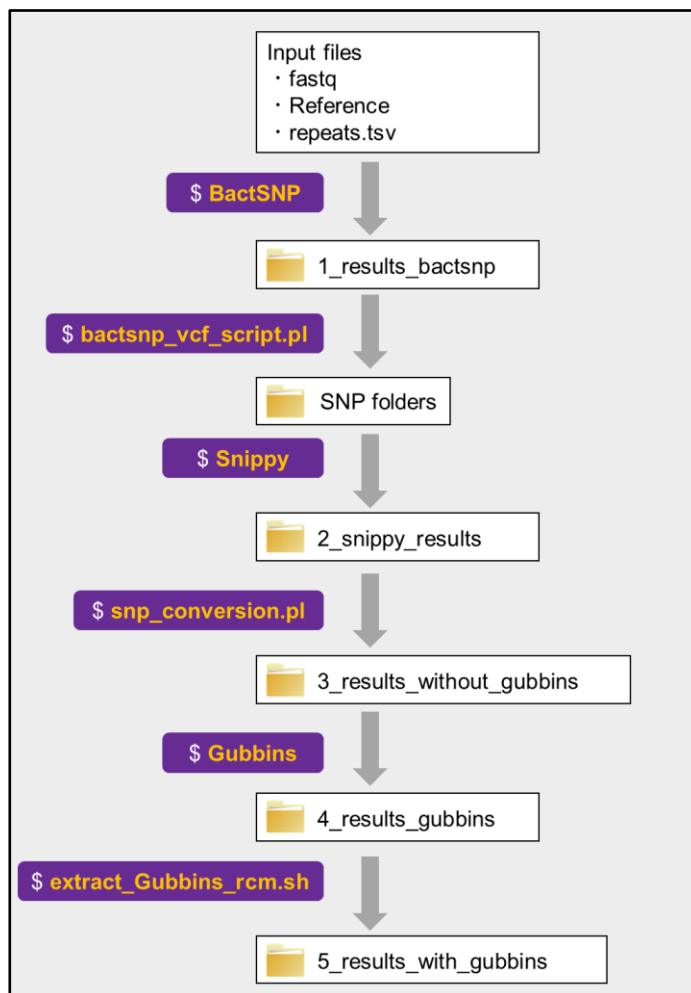


図2. SNPcasterの比較ゲノム解析フロー図

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。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。

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

	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

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。

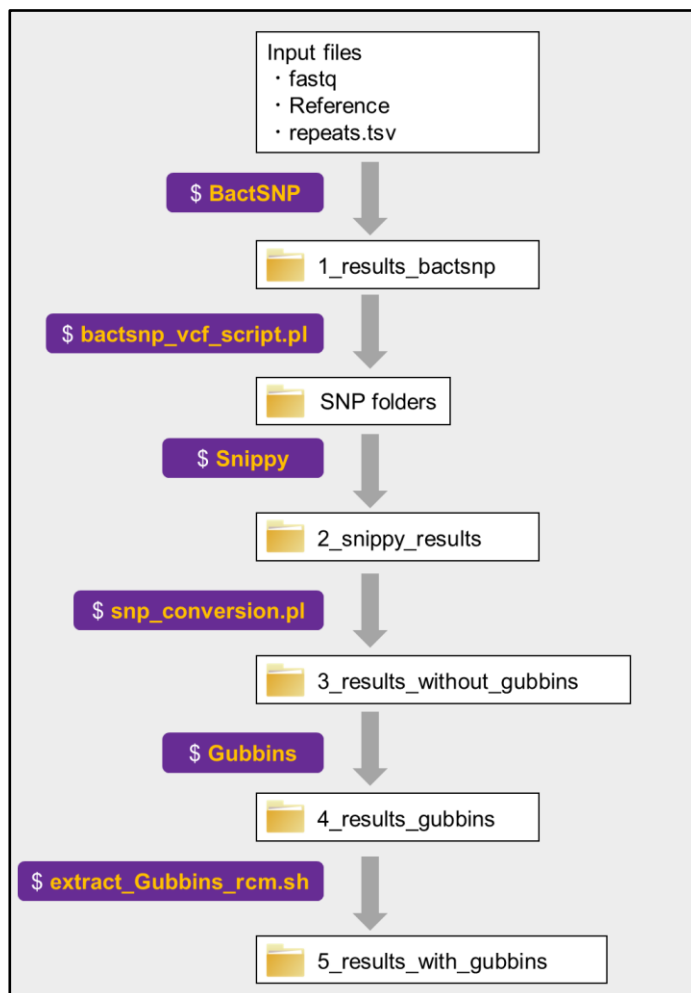


図2. SNPcasterの比較ゲノム解析フロー図

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。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。

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

	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