総説

Whole genome sequencing による耐性因子検出

松村　康史[1, 2)]

[1)] 京都大学医学部附属病院検査部・感染制御部*

[2)] 京都大学大学院医学研究科臨床病態検査学

受付日：2019 年 7 月 2 日　　受理日：2019 年 8 月 29 日

Whole genome sequencing（WGS，全ゲノムシーケンス）とは，次世代シーケンサーを用いて微生物のゲノムすべてを短時間で解読する技術であり，近年研究目的での利用が急速に進んでいる。WGS では，プラスミド・染色体を含むすべての遺伝子配列が得られるため，従来の PCR やマイクロアレイと異なり，検出対象とする遺伝子を検査前にあらかじめ決める必要はない。また，耐性因子だけでなく，菌種，multilocus sequence type，プラスミドタイプ，病原遺伝子等についての情報も同時に得られる。WGS 解析の実際の手順としては，次世代シーケンス実施とデータ解析の 2 つに分けられる。データ解析では，まず得られた生データから *de novo* アセンブリを行い，draft genome を作成する。この中から耐性遺伝子の網羅的データベース（ResFinder など）を利用して耐性遺伝子を検索する。視覚的操作が可能な Web ベースの解析システムも利用できる。耐性遺伝子検出から，実際の薬剤感受性を予測する研究が行われており，腸内細菌科細菌では 95% の精度で可能という報告もなされている。WGS による耐性因子検出の課題として，コスト，迅速性，専門知識を有する人材不足に加え，解析技術の限界が挙げられる。耐性遺伝子の周辺構造解析，ゲノム上の位置決定，染色体性耐性機構の網羅的検出は難しく，解析の妥当性を保証する基準や標準化もなされていない。これらの課題を解決するため，多くの資金と人的リソースが投入されており，今後の技術革新や標準化により，WGS による臨床微生物検査が導入される可能性が期待される。

Key words: whole genome sequencing, next-generation sequencer, resistance gene, drug-resistant bacteria

I.　Whole genome sequencing

Whole genome sequencing（WGS，全ゲノムシーケンス）とは，2005 年に発売された次世代シーケンサーにより可能となった，微生物のゲノムすべてを短時間で解読する技術である。近年の性能向上とコストダウンにより，研究レベルにおいてその利用が急速に進んでいる。次世代シーケンサーでは網羅的タンパク発現解析など，耐性因子に関連する種々の解析が可能だが，本稿では基本手法となる WGS による耐性遺伝子検出について述べる。

次世代シーケンサーによる WGS の基本原理は，まず細菌のゲノムをランダムな短い断片にし，この断片の塩基配列（「リード」と呼ばれる）を大量に取得することにより，ゲノムすべてをカバーする断片化された塩基配列を得ることにある。これらのリードを，コンピューターを用いて結合することで，contig あるいは scaffold と呼ばれる断片化されたゲノムが得られる。これを draft genome と呼ぶ。一方，すべての染色体・プラスミドなどのゲノムが完全に再構成された場合，complete genome と呼ぶ。現在，最も普及しているイルミナ社のシーケンサーは，リードが 150～300 bp と短いものの，エラー率が 0.1% 程度と低く，1 回のシーケンスで 15 GB から 2 TB もの大量のデータが得られ，バーコードを

*京都府京都市左京区聖護院川原町 54

Table 1. Advantages and disadvantages of resistance gene detection using whole genome sequencing analysis

Advantages	Disadvantages
Comprehensive detection of resistance genes No need to predefine targets	Detectable genes depend on the database utilized
Data other than resistance genes, including species identification, sequence type (strain type), plasmid types, and virulence genes can be obtained	Location of resistance genes (chromosome or plasmids) can rarely be determined
	Expensive
	Need knowledge on bioinformatics

用いることで 384 株まで同時解析することができる。一般細菌の WGS を行うと，50～250 個程度の contig が得られることが多い。耐性遺伝子の多くは可動性遺伝子と関連しているため繰り返し配列が多く，周辺構造の解析は一般的に困難であるが，耐性遺伝子の有無やタイピングには十分なデータが得られる。したがって，1 株当たり 1～3 万円程度と比較的にコストにも優れており疫学解析に適している。一方，Pacific Biosciences 社（PacBio RS II 等）や Oxford Nanopore Technologies 社（MinION）のシーケンサーは，ロングリードシーケンサーと呼ばれ，数 Kb 以上のリードが出力できるため complete genome を得るのに有用である。しかし，短いプラスミドが読みにくい，エラー率が 10～15% と高い，解析に必要なコンピューターのスペックが高い，1 株あたり 10～20 万円とコストが高いなどの欠点がある。

II. WGS による耐性遺伝子検出の特徴

WGS では，基本的にプラスミド・染色体を含むすべての遺伝子配列を決定することができる。したがって，PCR やマイクロアレイのように検出対象とする遺伝子を検査前にあらかじめ決める必要はなく，一度データが得られれば，コンピューター上で何度でも再解析が可能である。ただし，既知の耐性遺伝子以外は検出が難しく，検出可能な遺伝子は，用いる耐性遺伝子のデータベースに依存している。耐性遺伝子以外の有用な遺伝子情報，例えば菌種・multilocus sequence type・プラスミドタイプ・病原遺伝子等についての情報も同時に得られる。Pulse-field gel-electrophoresis（PFGE）等の既存の方法に比べ，高い分解能と客観性をもって菌株間の相同性に関する情報が得られるため，アウトブレイク株か否かの判定も正確に行うことができる[1]。利点と欠点のまとめを Table 1 に示す。

III. WGS による耐性遺伝子検出の実際

耐性遺伝子検出までのフローは，菌株やシーケンサーを操作し実験操作により塩基配列を得るステップ（A1～A4）と，コンピューターで得られた大量のデータを解析する（bioinformatics とも呼ばれる）ステップ（B1～B3）に分けられる（Fig. 1）。ここでは，例としてイルミナ社のシーケンサー，NextSeq 500 を用い大腸菌 96 株の WGS を行う場合について概説する。

WGS を行い，データを得るまでの実験作業の難易度は高くない。PCR 等の基本的知識に習熟していれば手技としては問題なく，プロトコルを遵守している限りまず失敗することはない。一方，データ解析の質は解析者の技能に大きく左右される。簡便に処理を行うソフトウェアも開発されつつあるが，手作業を要することが多い。データ処理の品質を保ちつつ，適切な解析を行うには，UNIX システムやプログラミングの知識が必要といえる。

A1. シーケンス計画の立案

使用するシーケンス試薬（出力データ量）に対して，シーケンスを行う株数が適切になるよう調整・確認を行う。大腸菌のゲノムサイズは約 5 Mb（500 万塩基対），NextSeq500 Mid Output 試薬キットは 40 Gb 程度の出力が見込める。この場合，同じ遺伝子領域を何個のリードでカバーできるかを表す depth of coverage（通常 50～100x が推奨される）は 80x となり，シーケンス試薬と株数の選択が妥当であることが確認できる。

A2. ライブラリー用 DNA の準備

ゲノム DNA の抽出は，QIAamp DNA Mini Kit（QIAGEN 社）等を用いて行う。次に，Qubit Fluorometer（Invitrogen 社）を用いて DNA 濃度を正確に測定し，規定濃度まで希釈する。菌株の培

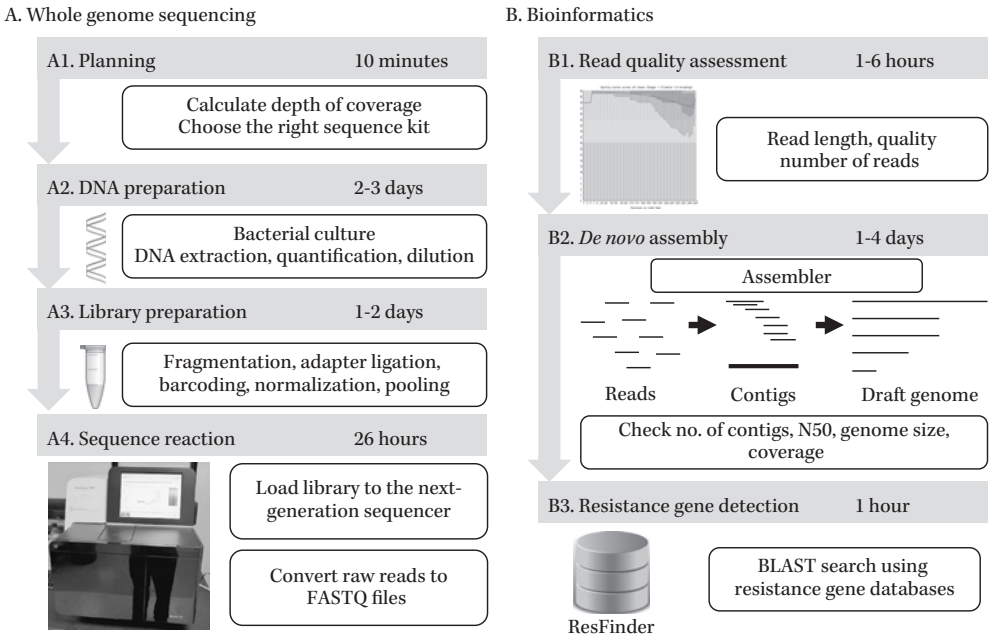

Fig. 1. Procedures for whole genome sequencing analysis-based resistance gene detection.

養に 1 日，96 株分の DNA 抽出に約 1 日，DNA 希釈は数時間～半日を要する。

A3. ライブラリー作成

抽出したゲノム DNA を，シーケンスに適した状態とする作業がライブラリー作成である。イルミナ社の Nextera XT キットは，多数株の WGS を行うのに適したライブラリー作成キットである。わずか 1 ng のゲノム DNA しか必要とせず，また複数検体を混合する際の濃度調整を簡単に行うことができる。ゲノム DNA の断片化とアダプター配列付加，マルチプレックス化用のバーコード配列付加と増幅，DNA 精製，濃度均質化と DNA の一本鎖化，ライブラリー混合（プール）の順に処理を行う。96 株あたり 1～2 日を必要とする。

A4. シーケンス反応

前項で得られたプール済み DNA ライブラリー，シーケンス試薬，フローセルを次世代シーケンサーにセットし反応を開始する。26 時間後の反応終了時には，生データが得られる。生データをシーケンサーからダウンロードし，バーコード配列の除去と FASTQ 形式への変換を行う（数十分）。得られたデータが解析の基本となるリードである。システム構成によっては反応終了時点で FASTQ ファイルが生成できる場合もある。最初のステップである菌株の培養から考えると，トータルで約 5 日，試薬等の消耗品に約 100 万円が必要となる。

B1. リードの品質確認とトリミング

各ゲノムと全体のシーケンス反応の品質を確認するため，得られたリードの品質を確認する。FastQC などのプログラムを用いて，リード数，リードの長さ，クオリティ（エラー率）を算出する。定まった評価基準はないが，エラー率 0.1% 未満（＞Q30）の総データ量が，予測値（今回は，400 Mb 程度）と大きく離れていなければ，シーケンス反応が良好であったと考えてよいだろう。リードの品質が確認できれば，次にクオリティの低いリードや極端に短いリード等を除去するトリミングを行う。一定のクオリティを有するリードのみを用いることで解析の質が向上するため，以後はトリミングされたリードを用いる。

B2. Draft genome の作成と品質確認

トリミングされたリードをもとに *de novo* アセンブリを行い，draft genome を作成し，そのクオリティを確認する。通常 1 ゲノムあたり数十分～数時間程度の時間を要するが，リード量が多い，あるいはゲノムサイズ大きい場合，CPU パワーと大量のメモリーが必要となるため，いわゆるパーソナルコンピューターでは処理が難しく，高スペックの UNIX サーバーやグリッドシステムが用いられることが多い。Contig 数，N50，ゲノムサイズ，depth

Table 2. Resistance gene databases

Database name	Number of genes[a]	Description	Reference
ResFinder	3,078	A web-accessible tool is available for both draft assemblies and raw reads.	4)
ARG-ANNOT	2,038	Only the database itself is available and distributed at the official site.	13)
CARD	2,570	A web-accessible tool is available. With the RGI (Resistance Gene Identifier) tool, resistance genes can be sought based on homology and SNP models which enables detection of mutations in intrinsic genes.	5)

[a] Data were accessed on 23 June, 2019.

of coverage を確認し，解析に適しているかを判定する[2]。

B3. データベースを用いた耐性遺伝子検出

耐性遺伝子の塩基配列を収載した網羅的データベースが公開されている（Table 2）。これを用いて，作成した draft genome 中に，耐性遺伝子と相同性の高い DNA 配列があるかを BLAST[3]と呼ばれるアルゴリズムを用い検索する方法が一般的である。塩基配列ベースの遺伝子検索は数秒～数分単位で終了するが，タンパク翻訳を行うアルゴリズムの場合は数倍以上の時間を要する。結果は，耐性遺伝子に対する相同性（% identity）やカバー長（% coverage）として得られ，それぞれカットオフを設定しフィルタリングを行う。データベースをダウンロードし，それぞれのコンピューターで BLAST 検索を行うのが基本であるが，視覚的操作が可能なソフトウェアや Web ベースの簡便な解析システムもある。

ResFinder データベース[4]は，Web サイト上でゲノムをアップロードすれば匿名で利用できるため，現在広く利用されている（https://cge.cbs.dtu.dk/services/ResFinder/）。データベース利用時の注意点として，（1）塩基レベルで 100% 一致しない場合であっても，通常同じタンパクに翻訳されるなら同じ耐性遺伝子とみなされるため，場合によってはアミノ酸翻訳を追加する必要があること，（2）遺伝子の一部しか存在しない場合（% coverage が 100% 未満の場合）には，基本的には遺伝子は機能していないと考えられるが，変異型やアセンブリの問題である場合があること，（3）データベース上に間違った遺伝子名での登録がなされていること，が挙げられる。

CARD データベースは，専用のツールと組み合わせることで，タンパクレベルの比較，SNP の検出，耐性の表現型予測も可能である[5]。獲得性耐性遺伝子のみならず，染色体性の変異による耐性検出にも対応している。

Resfams[6]は，既知の耐性遺伝子から得られたタンパクモデルを利用して，遺伝子の機能を推定することで耐性遺伝子の検出を行うシステムである。耐性遺伝子の大まかなタイプや機能しか同定できない反面，未知の遺伝子も同定することが可能であり，網羅的データベースにない耐性を予測する際に有用である可能性が示唆されている。

IV. WGS による薬剤感受性予測の精度

臨床分離株を用いた研究により，WGS データを用いた薬剤感受性予測の臨床応用が期待されている。*Escherichia coli* や *Klebsiella pneumoniae* においては，耐性遺伝子検出から β ラクタム・アミノグリコシド・フルオロキノロンの薬剤感受性予測をすることが，95% の精度で可能という報告がある[7]。黄色ブドウ球菌については，専用の予測システムを用いることで，感度 97%，特異度 99% で薬剤感受性が検出可能であった[8]。抗酸菌は培養に時間を要するため，菌種・感受性の迅速診断が臨床的に大きな意味をもつ。結核菌における検討では，93% の精度で菌種と薬剤耐性の予測が可能で，さらにタイピングによりアウトブレイク株か否かの判定も可能であったとの報告がある[9]。分離培養された菌体ではなく，臨床検体からの直接検査も検討されており，MinION シーケンサーにより結核菌の即日同定は全例で可能で，62% で感受性予測も可能であった[10]。

V. WGS による耐性遺伝子検出の限界・課題

WGS による耐性遺伝子検出は未だ研究レベルにある。まず，正確性・迅速性・コストといったシーケンス技術に多くの改善の余地がある。解析についても，耐性遺伝子データベースや表現型予測手法，

WGS 技術の標準化には課題が山積しており，さらに WGS を扱うことができる十分なスキルをもった人的資源も不足しているといえる。

1．シーケンス技術

現時点で，WGS に必要なコスト（数万円/1 株），時間（数日）は汎用に耐えうるレベルに達しておらず，また WGS で得られた情報も完全とはいえない。現在の次世代シーケンサーからのデータにはエラーが存在し，さらにデータ解析過程でのエラー（ミスアセンブリ・ミスマッピングなど）も常に存在していると考えられる。ゲノムには，繰り返し領域やマルチコピー遺伝子が存在するため，リードから完全に再構成することは容易でない。Draft genome はあくまでスナップショットであり，誤った配列の生成，特定領域の欠損などもあり得る。既知の遺伝子配列へのリードマッピングを用いて遺伝子検出を行った場合は，偽陽性が生じやすい。加えて，mobile elements 等の近傍に存在する獲得性遺伝子，類似性の高い遺伝子（例：$bla_{SHV\text{-}11}$ と $bla_{SHV\text{-}12}$），複数のコピーがある遺伝子において点突然変異の検出は難しいことがあり，偽陰性となる場合がある。現在主流のショートリードシーケンサーでは，耐性遺伝子の周辺構造は解析できないことが多く，局在診断（遺伝子がどこに存在するのか）は難しいことが多い。筆者の検討では，薬剤耐性腸内細菌科細菌においてESBL 遺伝子の局在が明らかとなったのはわずか 5%[11]，IMP 遺伝子では 13% であった[12]。

2．データベース

耐性遺伝子のデータベースには，一部染色体性変異も収載されているが，基本的に獲得性遺伝子のみが登録されている。また，当然ではあるが，データベースに登録されていない遺伝子を検出することはできない。十分に検討されずに登録されている場合もあるため（誤登録など），こういった限界を認識したうえでスクリーニングに用いるのが適当と考えられる。目的の耐性遺伝子がある場合は，信頼できるデータソースを用いて確認する必要がある。

WGS で検出可能なのはあくまで遺伝子であるので，感受性（表現型）を予測するためには，この対応を明らかにする必要があるが，明らかにデータが不足している。腸内細菌科細菌における β ラクタム耐性のように，限られた獲得性耐性遺伝子・耐性関連変異が耐性機序において優勢であれば，単純な一対一対応による予測でも高精度となる。しかし，多くの薬剤耐性は，獲得性遺伝子に加えポーリン・エフラックスポンプ・LPS 構造変異など内因性遺伝子における変異の複合的な作用により表現型が決定されるため，予測が困難である。また耐性の責任遺伝子すら同定されていない薬剤も存在する。

3．解析法の品質評価・標準化

EUCAST の報告書によれば，WGS データの妥当性を保証する基準，国際的な耐性因子の標準データベース整備，耐性遺伝子検出ソフトウェアの性能評価基準などは現在存在しておらず，臨床応用は推奨されていない[2]。そもそも耐性遺伝子が存在することは，表現型において epidemiological cutoff と対応するものであり，臨床効果を予測するブレイクポイントを予測するものではない。解析手法・判定基準の標準化を進めるために，(1) 適切な処理法・品質管理基準に関する国際的合意が必須であり，品質管理基準をクリアしたデータのみを使用する，(2) 異なった解析ツールを使っても，一定の基準を満たす結果が得られる，(3) すべての耐性遺伝子（変異）が登録され，管理・更新されたデータベースを確立する，こと等が必要と述べられている。

VI.　今後の展望

多くの研究者と資金が WGS による耐性遺伝子検出の発展に寄与しており，今後の技術革新や標準化により，さらに低コスト・簡便で信頼性が高い解析が可能になり，WGS による臨床微生物検査が導入される可能性もあるといえる。

利益相反自己申告：申告すべきものなし。

文献

1) Quainoo S, Coolen J P M, van Hijum S A F T, Huynen M A, Melchers W J G, van Schaik W, et al: Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev 2017; 30: 1015-63
2) Ellington M J, Ekelund O, Aarestrup F M, Canton R, Doumith M, Giske C, et al: The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect 2017; 23: 2-22
3) Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al: BLAST+: architecture and applications. BMC Bioinformatics 2009; 10: 421
4) Zankari E, Hasman H, Cosentino S, Vester-

gaard M, Rasmussen S, Lund O, et al: Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012; 67: 2640-4
5) Jia B, Raphenya A R, Alcock B, Waglechner N, Guo P, Tsang K K, et al: CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017; 45: D566-73
6) Gibson M K, Forsberg K J, Dantas G: Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 2015; 9: 207-16
7) Stoesser N, Batty E M, Eyre D W, Morgan M, Wyllie D H, Del Ojo Elias C, et al: Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother 2013; 68: 2234-44
8) Bradley P, Gordon N C, Walker T M, Dunn L, Heys S, Huang B, et al: Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun 2015; 6: 10063
9) Votintseva A A, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al: Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol 2017; 55: 1285-98
10) Lemon J K, Khil P P, Frank K M, Dekker J P: Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. J Clin Microbiol 2017; 55: 3530-43
11) Matsumura Y, Pitout J D, Gomi R, Matsuda T, Noguchi T, Yamamoto M, et al: Global *Escherichia coli* Sequence Type 131 Clade with $bla_{CTX\text{-}M\text{-}27}$ Gene. Emerg Infect Dis 2016; 22: 1900-7
12) Matsumura Y, Peirano G, Motyl M R, Adams M D, Chen L, Kreiswirth B, et al: Global Molecular Epidemiology of IMP-Producing *Enterobacteriaceae*. Antimicrob Agents Chemother 2017; 61: e02729-16
13) Gupta S K, Padmanabhan B R, Diene S M, Lopez-Rojas R, Kempf M, Landraud L, et al: ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014; 58: 212-20

Detection of antimicrobial resistance determinants using whole genome sequencing

Yasufumi Matsumura[1,2)]

[1)] Departments of Clinical Laboratory, Infection Control and Prevention, Kyoto University Hospital, 54 Shogoin-kawahara, Sakyo-ku, Kyoto, Japan

[2)] Department of Clinical Laboratory Medicine, Kyoto University Graduate School of Medicine

Whole genome sequencing (WGS) is a technology that can obtain genetic information on the entire microbial genome swiftly by the use of next-generation sequencers. The use of WGS has been rapidly increasing in research fields. WGS provides all nucleotide sequences including chromosomes and plasmids. We therefore do not need to predefine target genes before experiments in contrast to PCR or microarray methods. WGS can provide not only data on antimicrobial resistance (AMR) genes but also species identification, multilocus sequence types, plasmid types, and virulence genes. WGS analysis consists of a next-generation sequence reaction and a bioinformatics analysis. Bioinformatics analyses utilize public AMR genes databases after draft genome assembly. Studies have already reported a high accuracy for resistance phenotype prediction among clinical *Enterobacteriaceae* isolates. However, AMR prediction using WGS still has issues to be solved: high cost, long turn-around time, lack of skilled technicians, technical limitations in analysis methods, lack of international standards for analysis and quality control. Further technical advances and standardization will enhance the possibility of introduction of WGS in clinical microbiology laboratories.