

結核菌検査におけるゲノム解析法に関する検討

安中 めぐみ^a, 長谷川 乃映瑠^a, 久保田 寛顕^a, 吉田 勲^a, 鈴木 淳^a, 貞升 健志^b

近年、ベンチトップ型次世代シーケンサー（NGS）が普及し、結核菌の分子疫学調査においても解像度の高いゲノム解析を利用することが期待されている。しかし、標準的なマニュアルに沿えばNGSでは均一なシーケンスデータが得られる一方で、そのデータを用いてゲノム解析を進め、分子疫学調査に利用していく点は検討の段階にある。本研究では、東京都で分離された結核菌株596株についてNGSによって得たシーケンスデータをもとに二種類の解析方法、すなわち、Core Genome Multilocus Sequence Typing（cgMLST）とSingle Nucleotide Polymorphism（SNP）解析について、異なったlocusの数とSNP数を比較した。その結果、cgMLSTにおいて疫学的関連性が強いとされる5 locusの違いの株同士のうち98.3%が、同じく疫学的関連性を示すとされる5 SNP以内の関係性となり、同一クラスターの可能性がある最大数である12 locus以内の関係の99.7%はSNPs解析において12 SNPs以内の差となった。以上のことから、一定の疫学的関連性が考えられる株同士においては二法の差はほとんどないと考えられた。

ゲノム解析は、一塩基の解像度で菌株の類似性を見ることのできる強力な分子疫学のツールであるが、感染経路の特定には疫学情報の収集が第一であり、これを支援する検査方法としての有用性を継続的に評価していくことが重要である。

キーワード：結核，分子疫学調査，感染経路，次世代シーケンサー，ゲノム解析

はじめに

結核は二類感染症・全数把握疾患として指定されている感染症である。東京都健康安全研究センターでは東京都積極的疫学調査の一環として、都内医療機関から搬入された検体に対し結核菌の検査を行っている。また、結核菌の感染拡大を防ぐには感染経路を特定することが重要であることから、保健所や関連部署との連携のもと、疫学的関連が考えられる感染者から分離された株同士の類似性を分子疫学的に解析し、行政対応へとフィードバックしている。

2024 年現在、菌株解析の中心となるのは Variable Number of Tandem Repeats（VNTR）法である。本法は結核菌ゲノムの中で、特定の配列が繰り返し見られ、その繰り返し回数に株によって異なる 24 領域を対象に PCR による遺伝子増幅とフラグメント解析を行う方法である。フラグメント長から換算された繰り返し回数の組み合わせパターンを指紋のように菌を特徴づけるプロファイルとし、パターンが同一であることが認められた株同士については極めて近縁な株であるとみなす。特に、疫学的関連性の強い感染者から分離された菌株がこうした同一パターンである場合、感染伝播が起こった可能性が高いと考えられる。都の事業においても VNTR を分子疫学調査に活用しており¹⁻⁴⁾、日本各地においてほぼ同一の VNTR パターンを示す「M」株が広がっていることが示されてきた^{5,6)}。

その一方で、VNTR 法よりも解像度の高い分子疫学解析法として期待されるのが全ゲノム解析である。近年、ベンチトップ型の次世代シーケンサー（NGS）が普及し、

細菌サイズのゲノムシーケンス（数十万から数百万塩基対）であれば数日で取得することが可能となってきた。これにより、数百塩基対程度の解析を目的とした従来の DNA シーケンサーでは不可能であった、結核菌全ゲノムに基づく解析が行えるようになり、菌株の相同性をより正確に調べることが実現できると考えられる。

NGS によるデータ取得が試薬類や分析機器のマニュアルに従えば作業者に依らずほぼ安定的に実施できる一方で、全ゲノム解析を行政対応に活用するには、膨大なシーケンスデータの情報処理と、菌株を区別する明確な基準を設定することが課題である。例えば、食中毒菌などによる集団感染に対してはこれまでパルスフィールド電気泳動法が用いられてきたが、同法では電気泳動によって検出された制限酵素切断パターンについて、少なくとも 10 本以上のバンドが見られるとともに、異なる位置のバンドが 2 本ないし 3 本以内であれば「Closely related（密接な関係性）」とされている⁷⁾。結核菌全ゲノム解析においても判断基準を設けることが望ましく、その一方、一塩基単位でシーケンスデータが取得される全ゲノム解析では複数の比較方法が存在する。本研究では一定の判断基準が論文等において報告されている core genome MultiLocus Sequence Typing（cgMLST）⁸⁾と Single Nucleotide Polymorphism（SNP）解析⁹⁾を選出し、これら二法を比較したのでその結果を報告する。

^a 東京都健康安全研究センター微生物部病原細菌研究科

^b 東京都健康安全研究センター微生物部

実験方法

1. 結核菌株

2005年度から2022年度にかけ、東京都健康安全研究センターに搬入された結核菌株のうち、596株を対象とした。

2. NGSによるデータ取得

界面活性剤存在下でジルコニアビーズ破碎を行った後、熱処理した結核菌株に対してRNase処理を施し、フェノールクロロホルム沈殿法または市販のDNA精製カラム（QIAquick column, キアゲン）を用いてDNAを精製した。次に、精製したDNAをもとにNextera XT Library Preparation Kit（イルミナ）を用いてライブラリ調製を行った後、MiSeq Reagent Kit v3（イルミナ）によって2×300塩基のfastqデータを所得した。

3. cgMLST

NGSにより取得したfastqデータをRidom SeqSphere+（フイルジェン）にインポートし、結核菌cgMLSTスキームにより各locusのalleleを決定した。全株について本作業を行った後、同ソフトウェアを用いて総当たり（177,310組）で比較解析を行い、株間で異なるlocusの数を調べた。総当たりのデータを596行596列の距離行列として出力した後はMicrosoft Excelを用いて一列に整列し、以降の解析を行った。

4. SNPs解析

cgMLSTを実施したデータについてRidom SeqSphere+上の「Find SNV」モードを用い、全alleleに存在するH37Rv株の全ゲノムシーケンス(GenBank NC000962.2)をリファレンスとした場合のSNPsを検出した後、同ソフトウェアを用いて総当たり（177,310組）で比較解析を行い、株間のSNPsの数を調べた。SNPを検出する際は、過去の文献⁹⁾に従い、12塩基内に隣接したものについては検出せず、また、InDelは解析の対象外とした。総当たりのデータを596行596列の距離行列として出力した後は、Microsoft Excelを用いて一列に整列し、以降の解析を行った。

5. cgMLSTとSNPs解析の比較

一列に整列した、cgMLSTにおいて異なるlocus数およびSNPs数を横に並べて177,310行2列のデータとし、Microsoft Excelを用いて散布図とヒストグラムの作成を行った。

結果及び考察

1. 異なるlocus数とSNPs数の関係性

cgMLSTにおけるlocus数の違いとSNPsの関係を散布図に示した（図1）。なお、図1の直線は、locus数の違いとSNPs数が等しい場合の理論直線として示した。概ねlocus数が大きくなるに従ってSNPsは下に凸の曲線を描き、locus数が400を超えてくるあたりからSNPs数との差は大きくなることが確認された。このことは、cgMLSTでは同一

locus内に複数のSNPsが存在したとしても、ひとつのlocusの差としてカウントされることに由来すると考えられる。特に、菌株間の類似性が低い、異なるlocusの数が多い株同士においては、ひとつのlocus内のSNPs数も大きくなることが考えられるため、locus数が多くなるにつれ、SNPs数との乖離も大きくなったことが示唆される。

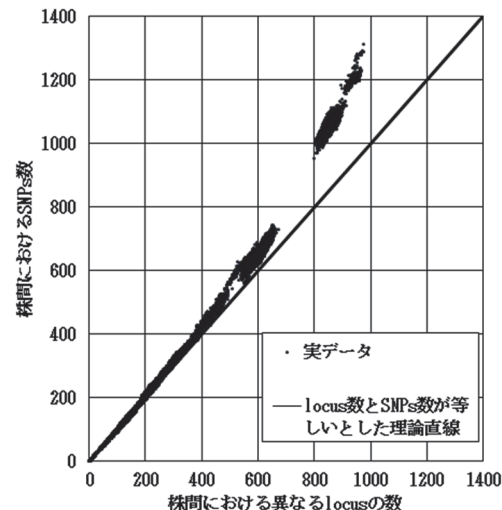


図1. 異なるlocus数とSNPs数の比較（n=177,310）

2. 疫学的関連性の判断基準付近における比較

cgMLSTでは異なるlocus数が5以内であれば疫学的関連性があるとされ、最大12 locusの差であれば同一クラスターの可能性があるとして報告されている⁸⁾。図2にcgMLSTにおける5 locus以内、図3に12 locus以内の差となった組み合わせについてのSNPs数を示した。

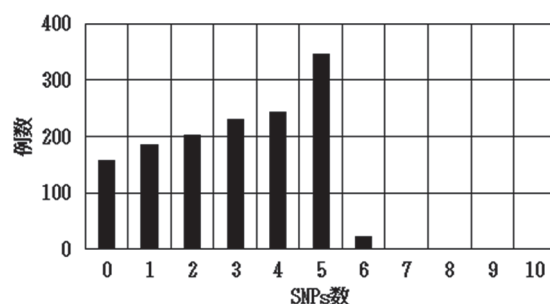


図2. 5 locus以内の株間におけるSNPs数（n=1,392）

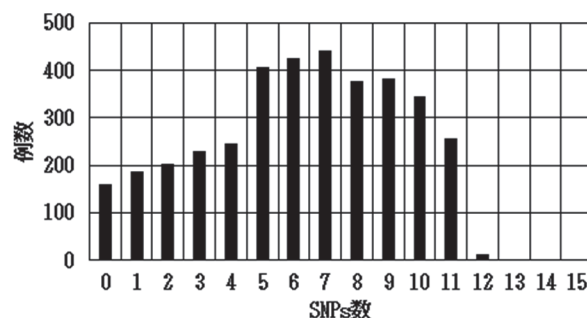


図3. 12 locus以内の株間におけるSNPs数（n=3,998）

5 locus以内の株間では98.3% (1,369/1,392) , 12 locus内の株間では99.8% (3,990/3,998) が, それぞれ5 SNPs以内, 12 SNPs以内の関係性となった. SNPs解析ではcgMLSTと同様に5 SNPsが疫学的関連性を支持し, 疫学的関連性を排除できない最大数として12 SNPsが判断基準とされている⁹⁾, ¹⁰⁾. cgMLST, SNPs解析のいずれの方法によっても同等の判断を行えることを示している.

ただし, cgMLSTと比べてSNPs解析にはまだ検討の余地がある. それは分析対象のlocusがあらかじめ定められているcgMLSTとは異なり, SNPs解析は5 SNPs等を判断基準として発表した文献³⁾では本研究でも採用した「12塩基以内の隣接したSNPsやInDelは除く」といった方法ととられるものの, 具体的にどの領域を解析対象とするかなどについては示しておらず, 統一的な解析方法が定められていない. 本研究ではcgMLSTの対象領域内のSNPsを検出するという形をとり, 二法間での比較検討を行ったが, こうした対象領域の選定はSNPs解析を行う上で必要不可欠である. また, あらゆる結核菌株の差を見るためのリファレンスとしてH37Rv株のゲノムシーケンスが適切であるかについても定かではない.

以上のことから, cgMLST, SNPs解析は同等の解像度であったが, 分析対象のlocusが定められているcgMLSTと異なりSNPs解析を行政検査において活用する際には, その方法と判断基準に対して関係者相互で一定のコンセンサスが得られていることが望ましいと考えられる. また, VNTR以上の分解能が期待されるゲノム解析ではあるものの, 感染経路を特定する上では疫学情報の収集が第一であることには変りがなく, 分子疫学解析はそれを裏付ける材料として利用することが重要である.

ま と め

結核菌の分子疫学調査法としての全ゲノム解析について, cgMLSTとSNPs解析の比較検討を行った. 結果として, 疫学的関連性を判断する上では両法に差は見られず, 同等の行政対応につなげることが可能と考えられた. そして, 今後, 引き続き分析精度を高め, より信頼性のある判断基準を設定するためにもデータの集積と解析方法の改良を進めることが重要である.

文 献

- 1) 向川 純, 三宅啓文, 吉田 勲, 他: 東京健安研七 年報, **59**, 53–57, 2008.
- 2) 向川 純, 三宅啓文, 山本宣和, 他: 東京健安研七 年報, **60**, 45–48, 2009.
- 3) 向川 純, 山本宣和, 三宅啓文, 他: 東京健安研七 年報, **61**, 111–116, 2010.
- 4) 向川 純, 山本宣和, 三宅啓文, 他: 東京健安研七 年報, **62**, 79–84, 2011.
- 5) 東京都微生物検査情報, **33**, 1–2, 2012.
- 6) 和田崇之, 田丸亜貴, 岩本朋忠, 他: 結核, **88**, 393–398, 2013.
- 7) Tenover, FC., Arbeit, RD., Goering, RV., *et al.*: *J. Clin. Microbiol.* **33**, 2233–2239, 1995.
- 8) O'Toole RF.: *EBioMedicine*. **34**, 6–7, 2018.
- 9) Walker, TM., Ip, CL., Harrell, RH., *et al.*: *Lancet Infect. Dis.* **13**, 137–146, 2013.
- 10) Fiebing, L., Kohl, TA., Popovici, O., *et al.*: *Euro Surveill.* **22**, 30439, 2017.

Study on the whole-genome analysis methods of *Mycobacterium tuberculosis*

Megumi ANNAKA^a, Noeru HASEGAWA^a, Hiroaki KUBOTA^a, Isao YOSHIDA^a, Jun SUZUKI^a, and Kenji SADAMASU^a

Molecular epidemiological examination is an important tool for the identification of the infection route of *Mycobacterium tuberculosis*. Genomic analysis is recently expected as a powerful tool with higher resolution, which is comparable to the routine diagnosis according to the development of a benchtop-type next-generation sequencer (NGS). To date, the interpretation of output data from NGS should be carefully considered in the molecular epidemiological study and the administrative responses; data collection using NGS is not so difficult according to the manufacturing protocols. However, data analysis requires some specialized procedures. In this study, two genomic analysis procedures, namely, a core genome multilocus sequence typing (cgMLST) and single nucleotide polymorphism (SNP) analysis, were compared using the NGS data from 596 *M. tuberculosis* strains isolated in Tokyo, Japan. When comparing the numbers of different loci in cgMLST and SNPs, 98.3% of the epidemiological linkages defined as ≤ 5 different loci in cgMLST were also found to be epidemiologically linked in SNP analysis (defined as ≤ 5 SNPs); 99.7% of undeniable epidemiological linkage in cgMLST (≤ 12 different loci) were ≤ 12 SNPs, which is also considered the maximum number of the included identical clusters. Therefore, these two methods can be compatible when handling cases with close epidemiological relationships.

In summary, genomic analysis can determine the phylogenetic similarity of *M. tuberculosis* with single-nucleotide resolution, providing powerful molecular epidemiological evidence; however, the collection of epidemiological information is mostly important as basal materials in identifying the infection route, and genomic analysis should be placed as a supporting tool. The usefulness of genomic analysis should be continuously evaluated.

Keywords: tuberculosis , molecular epidemiological investigation , infection route , next-generation sequencer , genomic analysis

^a Tokyo Metropolitan Institute of Public Health,
3-24-1, Hyakunin-cho, Shinjuku-ku, Tokyo 169-0073, Japan