厚生労働省科学研究費補助金　食品の安全確保推進研究事業
「ゲノム情報および機械学習を用いた腸管出血性大腸菌の高精度型別モデルの構築」
（20KA3002）
研究分担報告書

分担研究課題「O157 菌株の全ゲノム解析および MLVA との比較」
研究代表者　李　謙一（国立感染症研究所　細菌第一部）、

研究要旨

機械学習の基礎となるデータを得るために、腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*：EHEC）計 882 株の全ゲノム配列から単一塩基多型（single nucleotide polymorphism：SNP）を抽出し、MLVA との関連性を解析した。この結果、大部分の株間では SNP と MLVA の結果は正の相関関係を示したが、大きなばらつきが認められた。さらに、MLVA に加えて分離日間隔を含めても、SNP を線形的に予測することは困難であり、機械学習等のより複雑なモデル化が必要であることが明らかとなった。

## A. 研究目的

腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*: EHEC）の全国サーベイランスでは、現在反復配列多型解析（multi locus variable tandem repeat analysis：MLVA）法が用いられている。これまでに EHEC O157 を対象にした、MLVA 法と全ゲノム配列（whole-genome sequence : WGS）解析法との比較では、MLVA 法は短期間の集団感染調査には十分高い孝別能を有することが示されている。しかしながら、MLVA 型が 2 座位以上異なる株間では、近縁な株と遠縁な株が混在していることが明らかとなっている。そこで本分担研究では、機械学習に供するための EHEC O157 の WGS 解読を行い、各菌株間の遺伝的距離を単一塩基多型（single nucleotide polymorphism）にて算出した。得られた結果と MLVA 結果を比較し、機械学習の基礎となるデータを得た。

## B. 研究方法

2014 年から 2019 年に分離された EHEC O157 192 株について、ゲノム DNA 抽出を行い、Nextera XT DNA Library Prep Kit（illumina）または QIAseq FX DNA Library Kit（QIAGEN）を用いてライブラリー調製を行った。作製したライブラリーを使用して、HiSeqX（illumina）によってペアエンドシークエンシング（150-mer×2）を行った。得られたショートリードは、これまでに感染研・細菌第一部で既に解読した 690 株の WGS と合わせ、計 882 株で解析を行った。SNP 抽出は、BactSNP および snippy などを用いた解析パイプラインを用いて行い、Gubbins によって組換え領域

の検出・削除を行った。

**C. 研究結果**

計 882 株の WGS 解析を行い、全株総当たりのペアを作製し、各ペアでの SNP 数および MLVA で異なる座位数を算出した。過去の同様の解析では、MLVA での差異が 1 か所以内の株間では少数の SNP のみ存在することが示されている。今回の解析は、散発事例株が含まれるため、SNP のばらつきはより大きく表れた。MLVA での差異が 2 座位以内の場合には、cgSNP の中央値は 10 以内に収まり、近縁な株が大部分であった（図 1）。しかし、MLVA が同一でも SNP が 400 か所以上存在する株や、MLVA の差異が 11 か所存在する場合にも、SNP が 8 か所である株が存在した。経時的な SNP の蓄積速度を調べるために、MLVA の差異ごとに SNP と分離日の間隔を用いて、回帰分析を行った。その結果、異なる MLVA 座位数が大きくなるにつれ、回帰式の傾きが小さくなる傾向が認められ、5 座位が異なる株間では相関は認められなくなった。

**D. 考察**

国内株の O157 の SNP 解析データをさらに蓄積し、機械学習の基礎となるデータを得た。これまでのデータでは、集団感染株や関連する MLVA 型の株の割合が高かったが、本研究では散発事例株も含む株の解析を行った。この結果、MLVA と SNP の相関関係は先行研究と同様に認められたが、例外的な株（MLVA で類似しているが多数の SNP が存在する、または MLVA での差異が大きいが少数の SNP のみ存在する）が多数認められた。これらの株については、差異が存在する MLVA の座位やリピート数についてより詳細に検討する必要がある。また、異なる MLVA 座位数別に経時的な SNP の蓄積を回帰分析で解析したところ、強い相関は認められなかった。さらに、異なる MLVA 座位数が大きくなるにつれて、SNP と分離日間隔の相関性が弱くなる傾向が認められた。これは、経時的に MLVA 型も変化しているためと考えられる。つまり、分離日が数年離れている同一型のケースは、特殊な事例（冷凍保存食品など）の影響が強く出ている可能性がある。このことから、SNP 数は分離日と MLVA 型の差異数から単純に予測することはできず、多変量解析や機械学習等のより複雑なモデルによる予測が必要と考えられた。

**E. 結論**

本研究では、国内 EHEC O157 のゲノム解析によって、より網羅的に MLVA と SNP 解析の関係性を明らかにした。また、MLVA と分離日間隔を用いた線形予測によって SNP を予測することは困難であることが明らかとなった。

**F. 健康危険情報**

なし

**G. 研究発表**

1) 誌上発表

なし

2) 学会発表

なし

**H. 知的財産権の出願・登録状況**

1. 特許取得

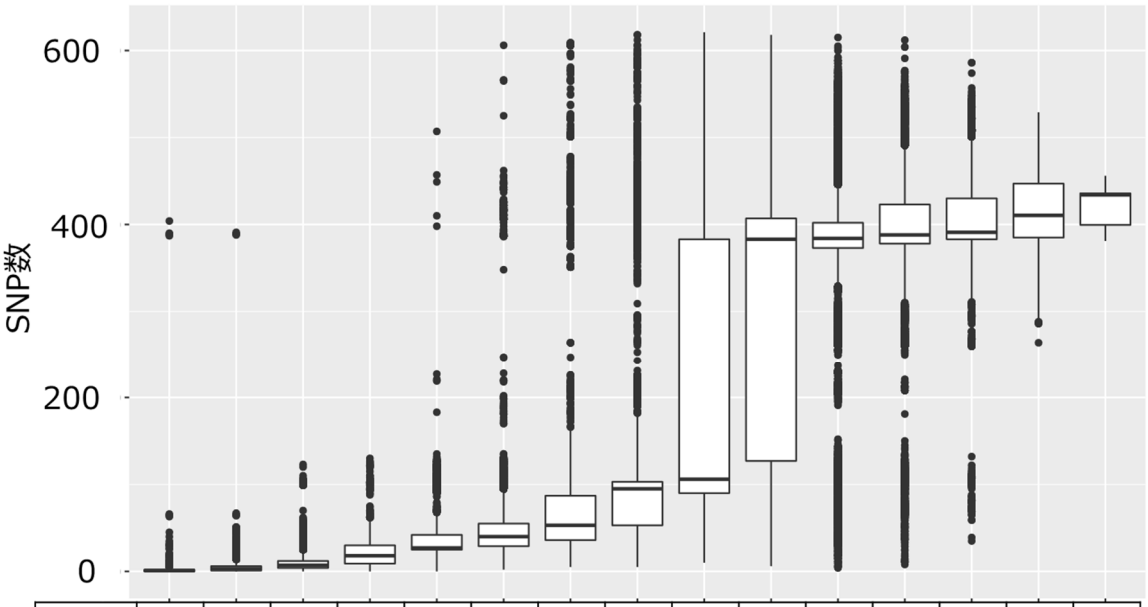
なし

２. 実用新案登録

なし

**図 1. MLVA と SNP の関連性**

MLVA の異なる座位数別に見た SNP の分布を箱ひげ図で示す。




| MLVA差異 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Median | 1 | 3 | 7 | 18 | 27 | 40 | 53 | 95 | 106 | 383 | 384 | 388 | 391 | 410.5 | 434 |
| Min | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 5 | 10 | 6 | 4 | 8 | 35 | 264 | 381 |
| Max | 404 | 391 | 123 | 130 | 507 | 606 | 609 | 618 | 621 | 618 | 615 | 612 | 586 | 529 | 456 |

**図 2. 異なる MLVA の座位数別にみた SNP と分離日間隔の関連性**

各カラムの右上に回帰式および決定係数（$R^2$）を示す。