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「EHEC 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）O157、O26、および O111 においてそれぞれ計 1,636 株、585 株、および 285 株の全ゲノム配列から単一塩基多型（single nucleotide polymorphism : SNP）を抽出した。さらに、菌株間の SNP が 5 または 10 か所以内の株をクラスター化し、重症化率等を計算するプログラムを Perl にて作製した。また、分担研究者が構築したモデルの評価を行い、O157、O26、および O111 のいずれにおいても敏感度（SNP で 10 以内のペアを「近縁株」として検出している割合）の顕著な増加が認められた。以上の結果から、本研究の機械学習モデルは、「MLVA での差異がある程度あるがゲノムレベルでは近縁な株」を効率よく抽出可能であることが示された。

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、O26、および O111 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。また、従来のように MLVA のみで近縁株を抽出した場合と比べた際と、機械学習モデルを用いた際との結果を比較することで、同モデルの評価を行った。

B. 研究方法

2014 年から 2021 年に分離された EHEC O157 384 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データと合わせ、計 1,636 株で解析を行った。O26 および O111 では、それぞれ 585 株および 285 株についての全ゲノム配列解析を行った。一部の菌株については上記の方法で新たに WGS を解読した。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。

また、モデル構築に用いた株のデータを用いて、菌株間の SNP が 5 または 10 か所以内の株をクラスター化し、重症化率等を計算するプログラムを Perl にて作製した。

機械学習モデルの評価として、近縁株を検出する能力を敏感度、特異度、陽性的中率、および陰性的中率の 4 種の指標を用いた。近縁株の定義としては、O157 の MLVA では 1 アリール以内の差異、O26 および O111 では同一の MLVA 型、を用いた。機械学習モデルでは、最も成績の良かった 10 か所以内・11 か所以上のカテゴリー分けデータを用いた。

C. 研究結果

計 1636 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。さらに、*in silico* で clade を決定した（表 1）。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。

また、機械学習にて近縁株を抽出した後に、病原性や国内での分布を予測するための Perl プログラムを作製した。本プログラムでは、まず SNP 情報に基づいて 5 か所または 10 か所以内の株同士をクラスター化する。クラスター化された株について、菌株情報をもとに重症化率（溶血性尿毒症症候群および血便の割合）、無症状保菌の割合、分離地の中央値、最小値、および最大値を算出した。結果例を表 2 に示す。本プログラムによって、機械学習モデルによって近縁株を抽出した後に、関連株の病原性等を予測することが可能となった。

機械学習の評価では、O157 では MLVA のみで近縁株を抽出した場合、敏感度以外の指標は 0.95 以上と非常に高い値を示した（表 3）。一方、敏感度（SNP で 10 以内のペアを「近縁株」として検出している割合）は比較的低い値（0.61）であった。機械学習の結果を用いると、clade 7 以外の敏感度は 0.88 以上となり、より多くの

近縁株の検出が可能であった。また、実際のサーベイランスにおける運用の際には、MLVA で 1 段階目の近縁株の抽出を行い、次に機械学習によって 2 段階目の抽出を行うと考えられる。そこで、MLVA 結果と機械学習結果を組み合わせた際の、敏感度などの指標を計算した。その結果、いずれの clade においても敏感度の値が向上した。

次に、O26 および O111 において O157 と同様の評価を行った。これらの血清型では、MLVA の結果のみを用いる場合には敏感度の値が低く、O26 では 0.51、O111 では 0.16 であった（表 4）。しかし、機械学習モデルを適用することによって、それぞれ 0.90 および 0.78 に向上した。その他の指標は、MLVA および機械学習に関わらず 0.95 以上と高値であった。

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。O157 は、遺伝的に多様であり、複数の亜系統（clade）に分けられる。本研究では、国内分離株の 9 割以上を占める clade 2, 3, 7 および 8 を対象にして機械学習モデルを構築することで、解析の精度を高めることが可能であった。

さらに、過去 2 年間に構築した機械学習モデルを用いて近縁株の抽出を行った際の敏感度などの指標を評価した。その結果、機械学習モデルを用いることで、敏感度の顕著な増加が認められた。この結果は、「MLVA での差異がある程度あるがゲノムレベルでは近縁な株」を効率よく抽出出来ていることを示している。詳細な機序は不明であるが、差異のある MLVA アリールやリピート数の違いによって、ゲノムレベルでの遺伝的距離を推測していると考えられる。Clade 7 では敏感度の上昇は認められなかったが、MLVA 結果と組み合わせることによって、成績の向上が可能であった。実際のサーベイランスでは、そのような従来法（O157 では MLVA で 1 か所以内の差異）と組み合わせた運用がなされると考えられるため、実用的にも高い型別能を有すると考えら

れた。

また、クラスター化された株について病原性等の情報を自動的に得られるプログラムによって、集団感染等が起こった際の危険度を予測することが可能になると考えられる。

E. 結論

本研究では、国内 EHEC O157、O26、および O111 において、MLVA 結果からゲノムレベルで近縁な株を高精度で抽出するプログラムを構築した。本モデルは、散発事例株や地理的に離れた株の関連性を推定するうえで有用となると考えられる。

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

1. 伊澤和輝, 李 謙一, 泉谷秀昌, 伊豫田 淳, 大西 真, 明田幸宏. MLVA結果と機械学習モデルを用いた腸管出血性大腸菌の遺伝的距離の予測, 第42回日本食品微生物学会学術総会
2. 李 謙一. 腸管出血性大腸菌の全ゲノム解析法について. 第34回 地方衛生研究所全国協議会 関東甲信静支部細菌研究部会. 横浜, 2023.
3. 泉谷秀昌, 李 謙一, 伊豫田 淳, 明田幸宏. 腸管出血性大腸菌のMLVAによる分子疫学解析. 第43回日本食品微生物学会学術総会. 東京, 2022.
4. 李 謙一. 全ゲノム配列解析を用いた腸管出血性大腸菌サーベイランスとクラスター検出事例 衛生微生物技術協議会42回研究会. Web, 2022.
5. 泉谷秀昌, 李 謙一, 伊豫田 淳, 大西 真. 2021年に分離された腸管出血性大腸菌のMLVA法による解析. 2022. Infectious Agents Surveillance Report 43:108-109.

H. 知的財産権の出願・登録状況

1. 特許取得

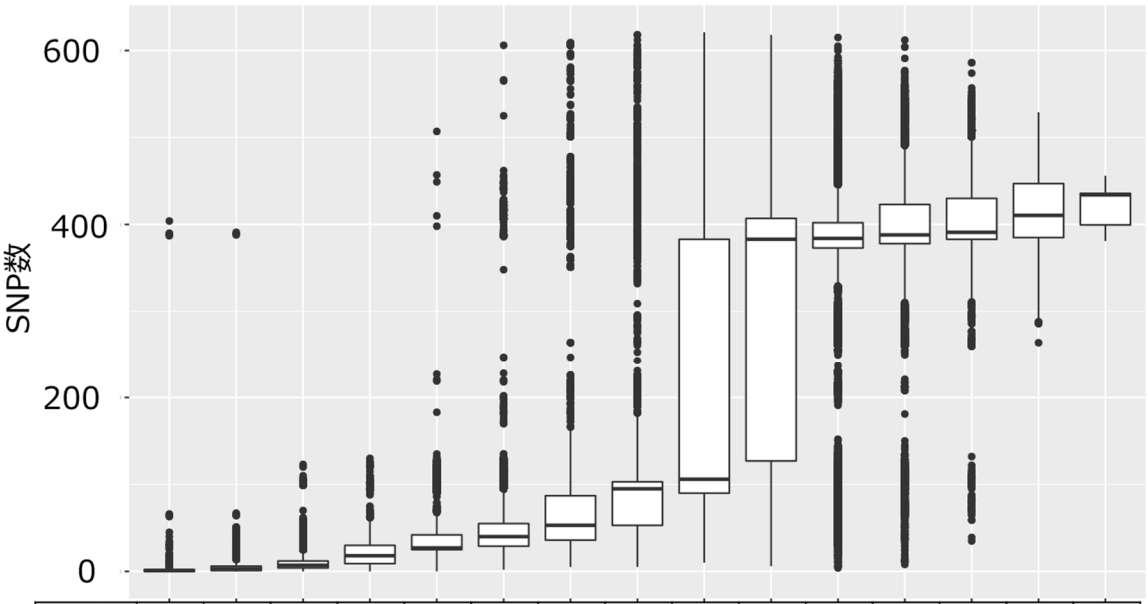
なし

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。

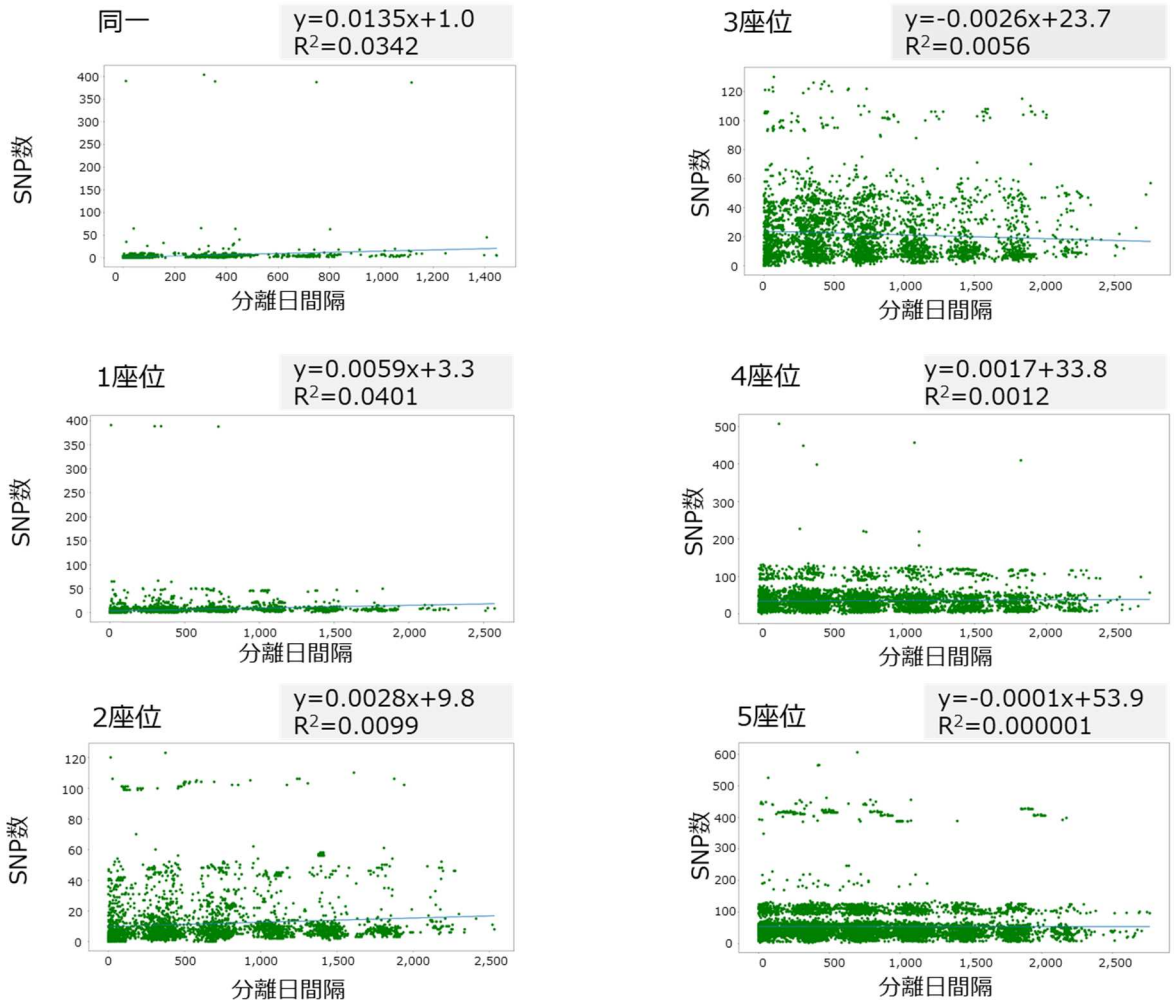


図 1. 解析菌株の clade 分布

Clade	株数
1	1
2	396
3	471
4/5	8
6	7
7	346
8	395
9	2
同定不可	10
計	1636

表 2. クラスター検出プログラムの出力例

株名	SNP5_cluster							SNP10_cluster						
	クラスター株	株数	重症化率(%)	無症状保菌者率(%)	距離中央値(km)	距離最小値(km)	距離最大値(km)	クラスター株	株数	重症化率(%)	無症状保菌者率(%)	距離中央値(km)	距離最小値(km)	距離最大値(km)
JNE130772	NA							NA						
JNE130856	JNE131493,JNE131896,JN	118	61.9	11.0	351.4	10.9	891.9	JNE131493,JNE131896,JN	229	61.1	11.8	304.9	10.9	891.9
JNE131070	NA							JNE160912,JNE171012,JN	6	50.0	33.3	318.5	37.4	872.1
JNE131158	NA							NA						
JNE131281	JNE131486,JNE131487,JN	25	52.0	32.0	53.0	14.3	913.7	JNE131486,JNE131487,JN	42	57.1	21.4	77.8	14.3	1032.7

表 3. O157 における MLVA と機械学習（ML）結果との比較

Stats	MLVA 1 アリール以内	Clade							
		2		3		7		8	
		ML のみ	MLVA＋ML	ML のみ	MLVA＋ML	ML のみ	MLVA＋ML	ML のみ	MLVA＋ML
敏感度	0.61	0.88	0.90	0.95	0.97	0.62	0.89	0.99	1.00
特異度	1.00	0.93	0.93	1.00	0.99	1.00	1.00	1.00	0.99
陽性的中率	0.97	0.82	0.82	0.98	0.96	0.90	0.87	1.00	0.99
陰性的中率	0.98	0.95	0.96	0.99	0.99	1.00	1.00	1.00	1.00

表 4. O26 および O111 における MLVA と機械学習（ML）結果との比較

Stats	O26		O111	
	MLVA	ML	MLVA	ML
敏感度	0.51	0.90	0.16	0.78
特異度	1.00	1.00	1.00	1.00
陽性的中率	0.95	0.99	1.00	0.95
陰性的中率	0.99	1.00	1.00	1.00