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クラスター検出
プログラムの開発」
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

機械学習の精度を向上させるデータを得るために、腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*：EHEC）計 1,636 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。菌株の clade は主に 2、3、7、および 8 であり、特にこれまでに不足していた散発事例由来 clade 7 の菌株を追加した。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。

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型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。

そこで本研究では、機械学習によって MLVA および菌株情報から菌株間の距離を推定するモデルの作製を目的とした。本課題の 1 年度目では基本となるモデルの作製を行ったが、散発事例由来株が少ないなどの課題があった。本年度は、機械学習のモデル構築および評価に必要な WGS データを取得するとともに、近縁株抽出に用いる一部プログラムの作成を行った。

B. 研究方法

2020 年から 2021 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

データと合わせ、計 1,636 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。

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。

C. 研究結果

計 1,636 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。*in silico* 解析による clade の分布を表 1 に示す。全体の 98%以上が clade 2, 3, 7, および 8 であった。

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。

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（特に clade 7）も含む株の解析を行った。この結果、より正確に遺伝的距離を推定することが可能になったと考えられる。

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 のゲノム情報を大幅に追加することによって、機械学習モデルの精度を向上させることが可能となった。

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

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

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. 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