細菌ゲノム配列情報から施設内遺伝子情報基盤データベースの構築とその活用に関する調査研究

綿引　正則　　内田　　薫　　磯部　順子　　木全　恵子
金谷　潤一　　加藤　智子　　大石　和徳

Construction and Utilization of In-house Genetic Infrastructure Databases from Bacterial Genome Sequence Information

Masanori WATAHIKI, Kaoru UCHIDA, Junko ISOBE, Keiko KIMATA,
Jun-ichi KANATANI, Tomoko KATO, and Kazunori OISHI

要　旨　次世代シーケンサー（NGS）による塩基配列決定技術の世界的な普及に伴い，公的 DNA データベースに登録されるゲノム配列情報も増大している．NGS から生産される膨大な配列情報を処理するゲノム解析は，主にLinux コンピュータを用いて世界的に進展しており，バイオインフォマティクスの知識が必要である．地方衛生研究所 (地研) の細菌検査部門においても，病原細菌のゲノム情報の活用は，感染源調査や病原性に関する解析等，公衆衛生上有用な情報が得られると期待されており，近年，導入が進んでいる．一方で，ヒューマンインターフェイスに優れたWindows 環境は広く普及しているが，バイオインフォマティクスの知識とLinux コンピュータの使用がメインになるため，地研におけるゲノム解析の利用は限定的であった．しかし，Windows10 の登場によりWSL(Windows Subsystem for Linux) 機能が追加され，Linux 用のアプリケーションがWindows 環境でシームレスに使用出来るようになり，地研におけるゲノム解析のためのバイオインフォマティクス技術が利用しやすくなった．Windows 環境で，インターネット上に存在する膨大なゲノム配列情報を地研の細菌検査部門で利用するためには，バイオインフォマティクスの知識が必要であり，普及の壁になっていると思われる．そこで，本研究では，膨大なゲノム情報の中から，地研の細菌検査部門で有効利用できる情報を抽出し，バイオインフォマティクスの知識に依存しなくても利用を可能とするin-house データベースの構築のためのワークフローの作成とその活用法について検討した．カルバペネム耐性腸内細菌科細菌検査，腸炎ビブリオ分離株のSNP 情報のデータベース化，およびこれまで蓄積してきた病原体検出情報も同じくデータベース化し，その活用法を検討したところ，検査結果の迅速化や的確な情報発信などに寄与することが明らかになった．まだまだ発展中の分野であり，今回示したワークフローの検証を含めて，日々の業務を支援するデータベースとしてどのように維持するかが今後重要である．

近年の塩基配列決定技術の進歩は著しく，DNA シーケンサと呼ばれる塩基配列を決定する装置が普及し，地方衛生研究所（地研）においても，検査機器の一つとして普及している．塩基配列情報は，アデニン（A），グアニン（G），シトシン（C）およびチミン（T）を示すAGCT の4つの塩基文字列によるテキストデータとして表現されている．そのためデータベース化しやすく，日本をはじめ，米国や欧州の３極協調による塩基配列データベース（「公的 DNA データベース」と称される；日本の場合は，DDBJ[1] が担う）が構築されており，インターネット上で誰でも自由に利用することができるのが大きな特徴である．公衆衛生領域に限れば，感染症の原因となる微生物の塩基配列情報も日々増加しており，その活用も広がっている．例えば，1）細菌，真菌及び原虫のリボゾーム RNA 配列解析による分子同定，2）ウイルスや細菌遺伝子の配列多型解析による病原体の遺伝系統の解析や分子疫学情報として利用した感染源や感染ルート調査への応用，3）公開された遺伝子配列から病原微生物を解析するPCR プライマーの設計等，既に日常的な検査に利用されている．

公的データベース上の配列情報は誰でも自由に利用することが出来る、有用性の高い情報である．それをどのように利用するか，特に近年増加する

病原微生物ゲノム情報は、細菌検査の現場で有用と思われるが，食中毒細菌（例えば，病原性大腸菌やサルモネラ属菌）のゲノムサイズは約４～6Mbp（メガ塩基対）と大きく，塩基配列を解析するコンピュータとソフトウェアが必須である．さらに近年，地研においても次世代シーケンサー（NGS）の導入が進み，バイオインフォマティクス技術の必要性が増大している．NGS から生産された大量の塩基配列データは，コンピュータの OS として，Windows ではなく Unix 系の Linux 上で動く様々なアプリケーションが開発されゲノム解析に広く利用されているが，Windows ユーザーにとって Linux の利用はこれまで難しかった．しかし，Linux 用のソフトをそのまま使えるようにする WSL（Windows Subsystem for Linux）機能が利用できるようになった Windows10 の普及により，NGS による細菌ゲノム解析は Windows 環境で利用できるようになってきている．

地研におけるゲノム解析は，地域で分離された病原ウイルスや病原細菌が対象となる．ウイルスと細菌のゲノムサイズは，大きく異なり，そのため，ゲノム解析の手法が異なる．本調査研究においては，地研における細菌検査部門に限定する．地研の細菌検査の現場で生産される日々の遺伝子検査結果と公的 DNA データベースから得られる膨大な塩基配列情報を利用するためには，その情報を如何に利用するかが重要である．適切なソフトウェアを用いても，かなりの時間を有する．従って，日々，検査の現場で利用できる有用な遺伝子情報を抽出し，施設内で簡便に利用できるデータベースの構築は有用と思われる．あらかじめ解析に利用できる in-house（施設内）データベースとして構築しておけば，必要な情報が短時間に得られ，結果的に検査時間の短縮につながると考えられる．

インターネット上で得られる情報は膨大であり，有用な情報が存在するものの，様々なファイル形式で存在しており，有効に利用するためには必要な情報のみを抽出することが必要となる．特に遺伝子に関する情報については，それぞれ目的に応じた拡張子や様々なファイル形式が定義されている．これらのファイル形式は基本的には文字列のみのテキストデータ（plain text とも呼ばれる）から成る．従って，このような文字列から，検査や調査研究に利用できる情報を効率的に抽出することが必要である．

本調査研究は，地研の検査の現場でコンピュータの OS として広く利用されている Windows 環境で，細菌検査や調査研究に有用となる病原体細菌の遺伝子情報に関する in-house データベースを構築する方法を検討した．さらに構築したデータベースの活用法についても検討を加えたので報告する．

材料と方法

１．使用コンピュータの環境，Web ツールおよびアプリケーション

１）使用したコンピュータは，64-bit 仕様の Windows 10 で，Ubuntu on Windows Subsystem for Linux（WSL）の使用及びインターネット接続環境とした．

２）Web ツール

・DDBJ [1], NCBI [2]：公的 DNA データベースで，あらゆる生物の解析された DNA 配列の登録，検索（blast 検索サービス）等の利用がインターネット経由で可能．登録 DNA 塩基配列は genebank ファイル（拡張子 gbk あるいは gb）として，遺伝子の情報と配列が記載された書式（gbk ファイル）としてダウンロード可能．

・RAST [3]：ゲノム解析された塩基配列に遺伝子と機能を自動的に割当て（annotation），gbk ファイルや Excel ファイル形式でダウンロード可能．

３）　使用ソフト

①　Windows 環境で使用したソフトウェアおよび使用目的は以下の通りである．

・CLC Genomic Workbench（フィルジェン株式会社）:de novo アッセンブル

・秀丸エディタ [4]：テキストデータの加工（正規表現，検索，置換，grep 検索），csv ファイルから fasta ファイル * 変換，文字コード・改行コードの変換（Windows から Linux）

*fasta ファイルとは、拡張子 fasta, fas あるいは fa：配列の識別名称と DNA 配列を含む文字列より成る plain text のこと．

・エクセル（Microsoft Corp）：テキストデータの加工 , 抽出（関数使用）

・FileMaker（FileMaker Pro 16 Advanced; FileMaker, Inc）: 検索，関数を使用したテキストデータの加工とゲノム情報データベースを構築．

・Artemis [5]:gbk ファイルを読み込み，CDS（coding sequence，蛋白質をコードする遺伝

子）情報を視覚化およびCDS配列（塩基配列，アミノ酸配列）の抽出，gbkファイルの編集.
・Sequence Assistant ver2.7 [6]：塩基配列の連結，DNA配列の相補鎖配列の変換.
・Ruby script[7]: 複数の塩基配列情報を含むmulti-fastaファイルのcsv形式への一括変換（Windows付属のコマンドプロンプト使用）
・FastQC [8]: NGSデータのクオリティチェック
・trimmomatic [9]: Miseqで得られた配列情報からのアダプター除去と及び配列のトリミング.
② Linux（ubuntu on WSL）で使用したソフトおよび使用目的は以下の通りである.
・nucmer [10]：相同（共通）CDSの抽出，および一塩基多型（Single Nucleotide Polymorphisum; SNP）の抽出
・raxml-ng [11]：系統樹作成
尚，使用アプリケーションについては，Windows10，WSL及びWeb上のサービスを利用した場合には，アプリケーション名やWeb上のサービス名の後に，必要に応じて，それぞれ「/Win」,「/Linもしくは/Linux」および「/Web」と付加して表記した.

２．供試菌

SNP解析，SNPデータベースの構築とその評価のため，腸炎ビブリオの保存株を使用した．内訳は，1997-8年分離4株（臨床株：血清型O4:K68 1株，O3:K6 3株），2010年分離1株（海水からの分離株：血清型O3:K6）および2016-7年分離3株（臨床株：血清型O3:K6），合計8株である.

３．ゲノム解析 , SNP 抽出

細菌分離株のゲノム配列はNGSであるMiSeq（illumina, Inc）を使用し取得した．DNAライブラリーはNextera XT DNA Library Preparation Kit（illumina, Inc）を用いて作製し，試薬キットはMiSeq Reagent Kit v3（PE，600サイクル）を使用した．MiSeqから得られたfastqファイルからゲノム配列の作成は，品質トリミング後（FastQC/Win，trimmomatic/Win），CLC Genomics Workbench/WinのDe Novoアセンブルツールを用いて，contig配列より成るfastaファイルとして得た．得られた配列から，比較したい株すべてに共通に存在するCDSの抽出，および参照配列と比較したSNPの抽出にはnucmer/Linを用いた．参照配列とゲノム配列）に共通するCDSを選択し，共通のCDSのみからなるSNP抽出用の参照配列を作成（multi-fasta形式）した．これを参照配列として，nucmer/Linを利用して，各ゲノム配列と比較し塩基の変異データを抽出し，塩基置換部位を記述するファイル形式であるvcf（variant call format）ファイル（参照配列に対する位置と塩基情報と比較配列の代替塩基情報等を含む）として出力した．このvcfファイルから，秀丸エディタを使って，CDS名，参照配列に対する位置，参照配列塩基，解析株の置換塩基を抽出し，csvファイルに変換し，FileMakerにインポートして，SNPデータベースとした．今回は，塩基置換だけでなく，挿入と欠失（Indel）情報も得た.

本研究では、腸炎ビブリオの保存株を用いたSNPデータベースの構築と評価を行うため、参照配列は*Vibrio parahaemolyticus* RIMD 2210633のゲノム配列（Accession No. BA000031, BA000032）を用いた.

４．系統樹作成

FileMakerによるSNPデータベース上で，indelを排除し，すべてのSNP位置の塩基をExcelファイルとしてエキスポートした．参照配列と各株のSNP塩基はそれぞれExcelの連結関数を用いて連結し，秀丸エディタにコピー，貼り付けし，multi-fasta形式のファイルを作成した．このfastaファイルを用いて，raxml-ng/Linを用いて系統樹を作成した．系統樹の表示は，FigTree/Win[12]を用いた.

結　果

１．遺伝子情報基盤データベースの構築：β - ラクタマーゼ遺伝子

FileMaker/Winを用いて，in-houseデータベースを構築した．本データベースで収集する情報は，主に蛋白質をコードする遺伝子（CDS）であり，付随する情報として，菌種名，株情報，accession番号（公的データベースで付与された受入番号），遺伝子名，塩基配列とアミノ酸配列及びそのサイズ（塩基数とアミノ酸数）を含み，その他，データベースの目的に応じて収集すべき情報を追加した．ここで主な情報元は，3つで，1つ目は，公的データベース（DDBJとNCBI）に登録されているgbkファイル，2つ目は特定の遺伝子あるいは菌種に特化した遺伝子データベースで通常fastaファイルとして入手できる．3つ目

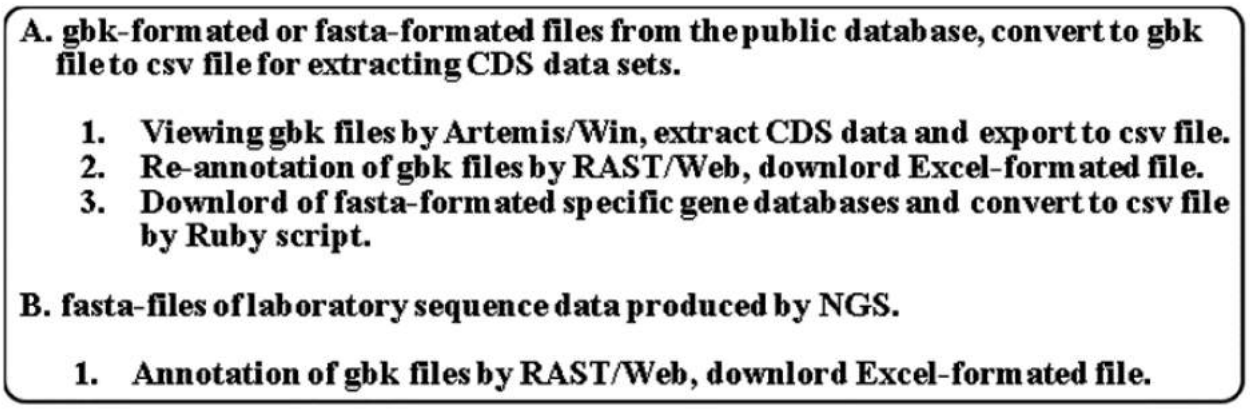


Fig.1. Workflow of Construction of In-house Database using FileMaker/Win.

は，自施設で解析した配列データで，ゲノム解析の結果として，de novo アセンブル後，繋がったcontig 配列のfasta ファイルとして得られる．登録されている細菌の完全ゲノム配列のgbk ファイルは，ファイル容量は通常約 10M バイトで、テキストデータとしては大きなファイルである。ここでこのようなファイルから抽出した情報をFileMaker にインポートすることは直接には不可能であり，FileMaker にデータインポートを可能とするファイル形式である，csv あるいはExcel ファイル形式に先ず変換が必要である．そこで本研究では，3つの情報元から得た配列情報をFileMaker にインポートするためのワークフローをFig.1 に示した．このワークフローを利用して，薬剤耐性菌の遺伝子検査に有用と思われるβ - ラクタマーゼ遺伝子情報データベースを構築した．β - ラクタマーゼ CDS の塩基配列及びアミノ酸配列は，NCBI のWeb サイト [13] からクラス AからD に分類されているβ - ラクタマーゼの遺伝子配列およびアミノ酸配列のfasta 形式のファイルをそれぞれダウンロードし，ワークフローの工程に従いFileMaker にインポートし，β - ラクタマーゼ遺伝子情報のin-house データベースとした（Fig.1A3）．β - ラクタマーゼ遺伝子情報については，日々，新しい遺伝子情報が公的データベースに登録されており，データベースの更新や訂正を以下のように行った．DDBJ の検索 Web ツールであるARSA[1] を利用して，キーワード検索して新しい遺伝子情報を調査し、必要であればgbk ファイルとしてダウンロードし、塩基配列とアミノ酸配列情報を抽出し，データベースに追記した（Fig.1 A1）．その結果，2,338 配列よりなるβ - ラクタマーゼ遺伝子情報データベースを作成した．その内訳は、β - ラクタマーゼの機能別の分類では，carbapenemase 632 遺伝子，Extended Spectrum β -lactamase（ESBL）488 遺伝子，AmpC β -lactamase 606 遺伝子およびその他のβ - ラクタマーゼとして662 遺伝子（公的データベースから得たgbk フィイル中の情報から特定できなかったβ - ラクタマーゼ）を含み，クラス別収録遺伝子の分類では，β - ラクタマーゼとして，classA 型 845 遺伝子，classB 型 252 遺伝子，classC 型 612 遺伝子，classD 型 608 遺伝子 およびその他遺伝子（公的データベースから得たgbk フィイル中の情報から特定できなかったβ - ラクタマーゼ）から構成されているin-house データベースとなった．

２．β - ラクタマーゼ遺伝子配列データベースの活用

このデータベースを用いて，PCR に使用するプライマー配列がCDS 中に存在するかどうか確認が可能であり，日常的に実施される薬剤耐性菌のβ - ラクタマーゼ遺伝子のPCR 検査に活用が期待できる．また，このデータベースからエキスポートして得られるβ - ラクタマーゼ遺伝子配列のfasta ファイルは，ローカル blast 検索で使用可能なデータベースファイルに変換すると，検査で得られたβ - ラクタマーゼ遺伝子の塩基配列から，β - ラクタマーゼ遺伝子の遺伝子型別を簡単に行うことが可能であった．

３．Windows によるSNP データベースの構築：腸炎ビブリオ分離株のゲノム解析とSNP データベース

分離株の高精度分子疫学的マーカーとして利用可能なゲノム SNP データベースをWindows 環境のFileMaker で構築した（Fig.2）．構築法とデー

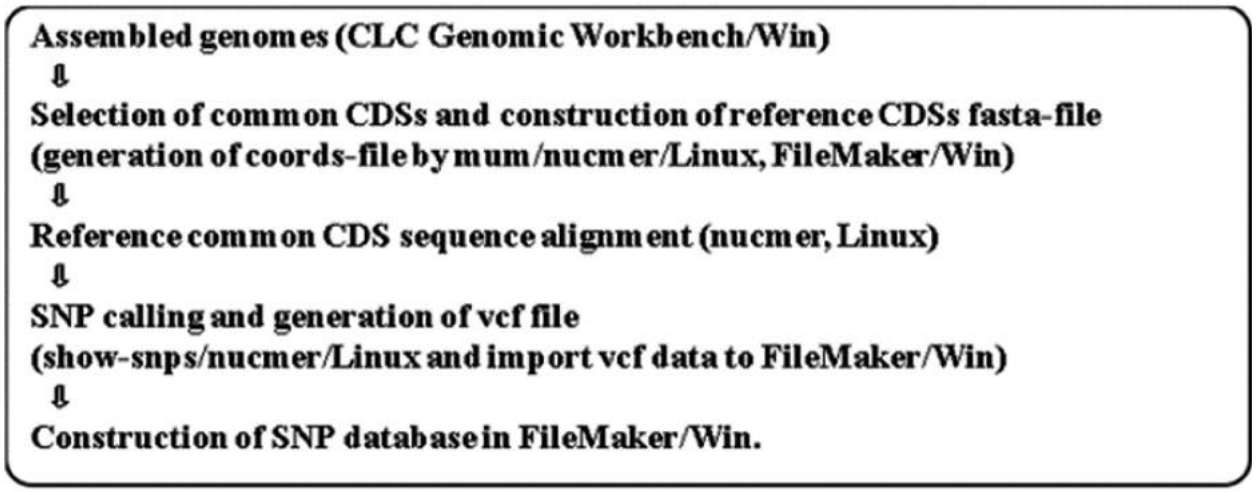
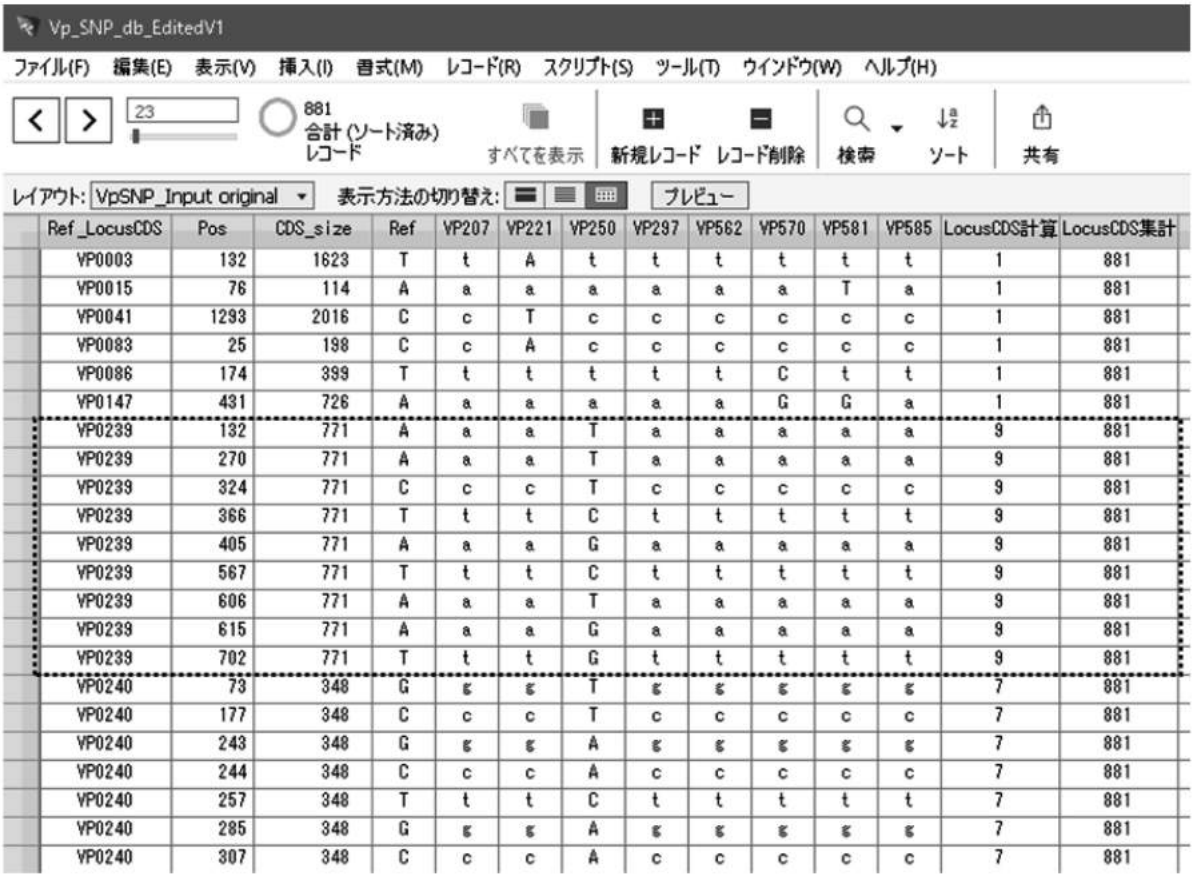


Fig.2. Workflow of Construction of SNP Databases using FileMaker/Win.

タベースの評価を行うため，1997-8 年と 2016-7 年に分離された腸炎ビブリオ食中毒患者及び海水から分離された 8 株の腸炎ビブリオのゲノム配列を MiSeq より得た．得られた fastq ファイルは，アダプター配列除去，品質評価後のトリミングを実施し，CLCgenomic Workbench/Win で de novo アセンブルを行い，40 ～ 60 contig 配列を得た．得られた配列は，参照配列として，既に報告されているゲノム配列（BA000031 及び BA000032）と比較して，塩基置換部位（single nucleotide variation site）を検索し，これを SNP（single nucleotide polymorphism）として利用した．SNP の抽出は，蛋白質をコードする領域（CDS）を対象とし，2,254 個の CDS からなる参照配列と 8 株のゲノム配列に共通に存在する CDS を対象として選択した．この共通する CDS の選択には，nucmer/Lin（-mum，オプションスイッチ使用）を使って，当所の供試 8 株の contig 配列から，今回は 100% 配列が一致した CDS（一塩基多型が存在しないため）を除き，また，2 コピーあるいは部分的に 2 コピー以上存在する CDS も除いたところ，最終的に 171 個の CDS を得た．そこで参照配列からこの 171 個の CDS 塩基配列を抽出し，multi-fasta ファイルとして，これを今回，NGS 解析した 8 株の腸炎ビブリオゲノム配列から SNP 抽出に使用した．SNP 抽出には，nucmer を用いてそれぞれの株の配列毎に，参照配列に対する塩基置換位置と置換塩基を検出し，vcf ファイル形式で結果を得た．この結果を秀丸エディタとエクセルを用いて，必要な情報（CDS の識別名称，CDS 内の変異部位の位置情報，参照配列及び解析株配列の塩基情報）を抽出し，FileMaker に逐次インポートし，SNP データベースを構築した．複数株の塩基置換部位について，FileMaker へインポート時に参照配列と照合し，同一部位で既に検出されているか，あるいは新しい部位かを判定し，新しい場合にはデータを逐次追加するインポート法を採用した．今回の nucmer による塩基変異部位の検出は，indel も検出しているため，FileMaker にインポートした後，この indel についてはすべて除去した．その結果，今回，171 個の CDS（合計塩基数，180,648 塩基）から，881 か所で SNP 座を検出した (Fig.3)．参照配列の塩基と，参照配列と較べて SNP として検出された塩基は，大文字で，参照配列と同一塩基は小文字で標記した．また，ここで示した参照配列の「Ref_LocusCDS」フィールドで，VP239 は 771 塩基の CDS であることを示しており，ここでは 9 か所の SNP が存在することを示されている．ここで例示した VP0239 は *Vibrio parahaemolyticus* RIMD 2210633 の第一染色体にコードされている triosephosphate isomerase であった．

４．Windows による SNP データベースの構築：腸炎ビブリオ分離株の SNP 配列による系統解析

SNP データベースのデータを用いた系統解析のワークフローは Fig.4 に示した．参照株を含めたすべての株の 881 か所の SNP を含む塩基は，複数塩基配列をもつ fasta 形式のファイル（multi fasta）に変換した．作成した fasta ファイルは，文字コード及び改行コードを Linux 形式に変換し，raxml-ng/Lin を用いて，系統樹を作成した


Vp_SNP_db_EditedV1

ファイル(F)　編集(E)　表示(V)　挿入(I)　書式(M)　レコード(R)　スクリプト(S)　ツール(T)　ウインドウ(W)　ヘルプ(H)

23　881 合計 (ソート済み) レコード　すべてを表示　新規レコード　レコード削除　検索　ソート　共有

レイアウト: VpSNP_Input original　表示方法の切り替え:　プレビュー

Ref_LocusCDS	Pos	CDS_size	Ref	VP207	VP221	VP250	VP297	VP562	VP570	VP581	VP585	LocusCDS計算	LocusCDS集計
VP0003	132	1623	T	t	A	t	t	t	t	t	t	1	881
VP0015	76	114	A	a	a	a	a	a	a	T	a	1	881
VP0041	1293	2016	C	c	T	c	c	c	c	c	c	1	881
VP0083	25	198	C	c	A	c	c	c	c	c	c	1	881
VP0086	174	399	T	t	t	t	t	t	C	t	t	1	881
VP0147	431	726	A	a	a	a	a	a	G	G	a	1	881
VP0239	132	771	A	a	a	T	a	a	a	a	a	9	881
VP0239	270	771	A	a	a	T	a	a	a	a	a	9	881
VP0239	324	771	C	c	c	T	c	c	c	c	c	9	881
VP0239	366	771	T	t	t	C	t	t	t	t	t	9	881
VP0239	405	771	A	a	a	G	a	a	a	a	a	9	881
VP0239	567	771	T	t	t	C	t	t	t	t	t	9	881
VP0239	606	771	A	a	a	T	a	a	a	a	a	9	881
VP0239	615	771	A	a	a	G	a	a	a	a	a	9	881
VP0239	702	771	T	t	t	G	t	t	t	t	t	9	881
VP0240	73	348	G	g	g	T	g	g	g	g	g	7	881
VP0240	177	348	C	c	c	T	c	c	c	c	c	7	881
VP0240	243	348	G	g	g	A	g	g	g	g	g	7	881
VP0240	244	348	C	c	c	A	c	c	c	c	c	7	881
VP0240	257	348	T	t	t	C	t	t	t	t	t	7	881
VP0240	285	348	G	g	g	A	g	g	g	g	g	7	881
VP0240	307	348	C	c	c	A	c	c	c	c	c	7	881

Fig.3. Partial Depiction of SNP Database of *Vibrio parahaemolyticus* Isolates in Toyama
Dotted box shows the details of the SNP locus present in the referrence CDS, VP0239.
See text for details.

Export SNP data to Excel format after searching on FileMaker/Win.
⇩
Convert the concatenated SNP data sets to multi-fasta format file (Excel, Text editor/Win).
⇩
Construction of phylogenetic trees (raxml-ng/Linux)
⇩
Veiwing of trees (FigTree/Win)

Fig.4. Construction of SNP phylogenetic tree

(Fig.5).

1997-8 年 の O3:K6 流 行 株（VP207, VP221, VP297）と 2016-7 年の O3:K6 株（VP570, VP581, VP585）に加えて，2010 年に海水から分離 O3:K3 株（P562），血清型が異なる VP250（O4:K68）の 8 株では，1997-8 年の分離株間の SNP は 6 ～ 13 違いで系統関係が極めて近いことが示されたが，年代が上がるとその SNP の違いは 100 前後であることが示された．血清型が異なる分離株は，遺伝的に離れており，今回解析した血清型の異なる VP250 株と参照株（Ref）の SNP 数は 658 であった．今回，解析した株のなかで最も参照配列

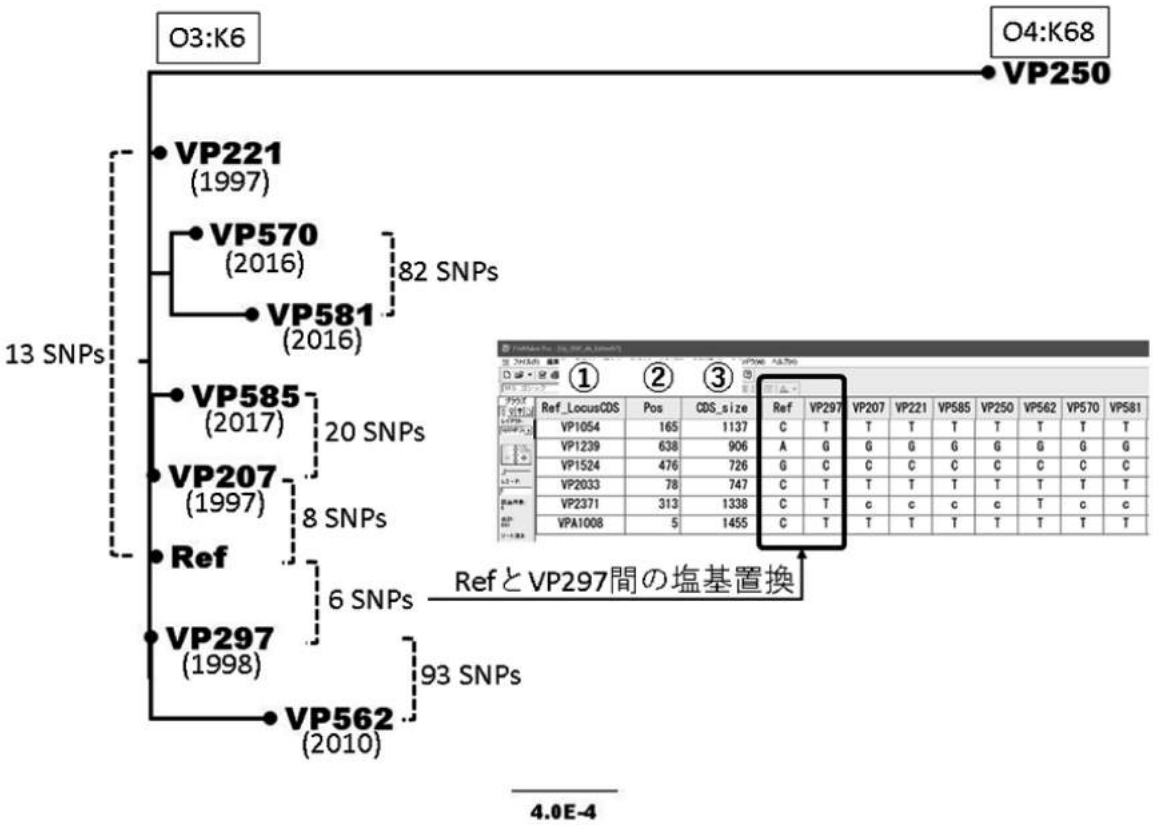


Fig.5. SNP-based Phylogenetic Tree of *Vibrio parahaemolyticus* Isolates in Toyama
Internal figure is described 6 SNPs after searching loci of base substitutions by FileMaker.

（Ref）に近縁であったのはVP297株で，6 SNPの違いであり，PFGEパターンが同一になる可能性もあると推測される。本データベースから，参照配列のCDSにおける6つのSNP位置の特定が可能であった（Fig.5）.

５．病原微生物検出情報データベースの構築

検体別に病原細菌の検出数の集計が記録されている2000年からの富山県病原微生物検出情報から，菌種名，分離年，分離月，分離材料に関する情報をFileMakerにインポートして，データベースを構築した．令和2年3月31日時点で，報告された分離菌株総数は，350,711株となった．検査材料別分離株数の割合は，尿及び呼吸器材料からの検出が多かった（Fig.6）．このデータベースを用いることにより，富山県で検出されている病原細菌の動態を迅速に把握することが可能となった.

考　察

細菌検査は，感染症や食中毒の原因となる病原細菌を分離し，分離した細菌の特徴を明らかにし，疾病診断や治療に資することが目的である．その中で，地研における細菌検査は公衆衛生上，感染症や食中毒が発生した場合，それが集団感染なのか，集団発生の場合，その感染源や感染ルートを明らかにし，蔓延防止に資するため，病原細菌の分離とその病原細菌の特徴を明らかにすることが求められる．また，平時においては，食品や環境中に存在する病原微生物のサーベイランスなどを行い，環境中のリスク評価を行なっている．その中で遺伝子検査は，その迅速性と再現性に優れ，得られる情報量の多さから近年，急速に普及が進み，塩基配列決定技術の需要は増大しつつある．また，ゲノム塩基配列情報が増大しているなかで，我々の検査や研究活動から生産される塩基配列情報の有効利用はいまだ限定的である．特にバイオインフォマティクスの知識が必要で，主にLinux

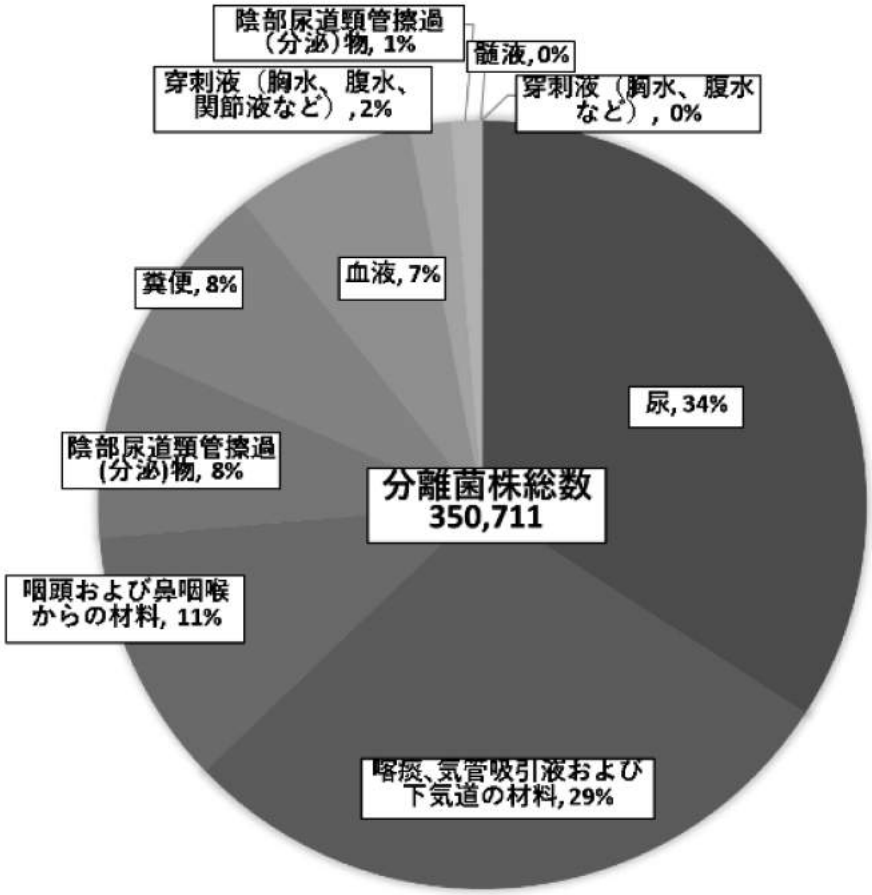


Fig.6. Isolation Rates of Pathogenic Bacteria by Clinical Specimen Detected in Toyama from 2000 to 2019.

環境で発展しており，Windows を主に使っている現場ではその普及も限定的であった．しかし，最近発表されたWindows10 の機能として追加されたWSL を利用すると，これまでLinux 上で利用されてきたバイオインフォマティクス用のアプリケーションがWindows とLinux をシームレスで利用できるようになった。そこで本研究では、我々が必要とする情報をWindows 環境で広く利用されているテキストエディタやExcel と言ったソフトの機能を使用し，様々なファイル形式で記述されている大量の情報から必要な情報を抽出し，我々が日常業務で簡単に利用できる情報に特化したデータベースの構築するためのワークフローの構築を試みた。また、その活用法を評価することも本研究の目的とした。

今回，作成したデータベースは，CRE 検査に有用と思われるβ-ラクタマーゼ遺伝子に関する情報，分離株の高精度分子疫学マーカーとして利用価値が高いといわれるSNP 情報，さらに本県で長期にわたり蓄積されている病原微生物検出情報に関する．何れも，FileMaker の検索機能が利用可能で，検索して得られた情報をエキスポートすることにより，blast 検索や遺伝子配列に基づく遺伝子系統解析などに，すぐに利用することが可能になるというのが大きな利点である．

β-ラクタマーゼのデータベースは，カルバペネム耐性腸内細菌科細菌（CRE）感染症の起因菌の検査に有用である．地研のCRE の検査は，カルバペネマーゼ産生腸内細菌科細菌（CPE）を鑑別するため，スクリーニングとしてPCR[14] を実施し，カルバペネマーゼ遺伝子の有無を確認している．CPE と決定され，医療関連感染が疑われる場合には，カルバペネマーゼ遺伝子の亜型の決定が必要になる場合がある．今回，構築したβ-ラクタマーゼのデータベースから，カルバペネマーゼ遺伝子配列と亜型情報を含めた遺伝子名を出力し，予め，ローカルブラスト検索のデータベースとして準備しておくと，瞬時にカルバペネマーゼ遺伝子の鑑別が可能であった．このβ-ラクタマーゼのデータベースは，地研の検査の現場

では有用であり，既に複数の地研でこのデータベースの情報が利用されている．また，このデータベースを利用して，新しいPCR検査の開発，評価に利用された[15].

地研で有用と思われる細菌のゲノム解析の一つにSNP解析が挙げられる．地研の役割は，分子疫学情報として利用できる遺伝子型を決定することであり，アウトブレイクの早期探知，感染源や感染ルートの特定に欠かせない検査である．この目的に実施されている遺伝子検査法として，パルスフィールドゲル電気泳動（PFGE）[16]，IS-Printing System [17], Mutiple-Locus Variable number tandem repeat Analysis（MLVA）[18], Multi Locus Sequencing Typing（MLST）[19]などが知られている．しかし，菌種やその検査原理により，迅速性，使い勝手，解像度やコストなどが異なり，一長一短があることも事実である．特に，PCR法をベースにした方法は，迅速性は優れているものの解像度が低い傾向にあり，ターゲットとする遺伝子により，分子疫学マーカーとして有用性が異なり，個別に評価が必要である．しかし，一方でコストやバイオインフォマティクスの知識が必要となるが，ゲノム配列上の一塩基多型を利用した解析として，高精度疫学マーカーとして利用できるSNP解析が注目されている．これまで既存の解析法では解像度の点で問題のあったサルモネラ属菌や結核菌は，その有用性が報告されている[20, 21]．ここでSNP抽出は，基準となるゲノム配列（参照配列）に対して，NGSから得られた配列を張り付けて変異塩基を抽出する．また，

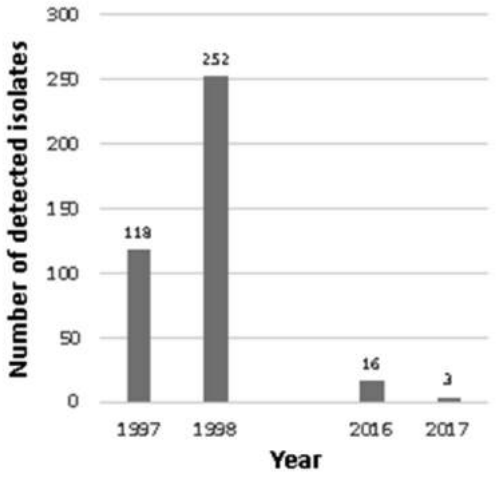


Fig.7. *Vibrio parahaemolyticus* Detected in Toyama at Specific Times.

参照配列として，利用するゲノム配列により，core genome（cg）MLST，whole genome（wg）MLST [22]などSNP解析には，操作性や使い勝手により様々な方法が報告されるようになっている．そこで，今回，データベース化しやすいと思われる，共通の遺伝子（CDS）からSNPを抽出するcgMLSTの考え方をベースとした方法を開発した．その大きな特徴は，WSLのLinux上でSNP抽出を行い，得られたSNPデータを逐次保存可能で，どの遺伝子上のSNPとして検出されたかを知ることができる．今回使用した菌種は，腸炎ビブリオ食中毒で分離された研究室保存株である．本菌における，富山県の臨床分離株の検出株数は，1998年に252件とピークで，当時世界的に流行し，公衆衛生上問題となった代表的な食中毒起因菌である（Fig.7）．当時から，PFGE解析ではほとんど区別できない流行株として血清型O3:K6が注目されていた．この食中毒は，最近はほとんど報告されなくなってきたが，2016年に同じ血清型が16件検出された．約20年前に分離された同血清型の臨床分離株の遺伝的関係を調べるため，SNP解析を実施すると同時にSNPデータベースを構築し，その有用性を評価した．その結果1997-8年の分離株間のSNPは6～13違いで系統関係が極めて近いことが示されたが，年代が上がるとそのSNPの違いは100前後であることが示された．また，このデータベースは，参照配列のCDSと塩基置換部位を簡単に知ることができるため，塩基置換の蓄積しやすい遺伝子など推定が可能になると期待される（Fig.5）．現在，本研究のワークフローを用いて，レジオネラ菌のSNPデータベースを構築し，地域で分離された株間遺伝的関係を解析するツールとしての，本SNPデータベースの有用性の評価を進めている

FileMakerによって構築した細菌ゲノム配列情報は，地研で実施されている遺伝子検査を支援し，迅速で，的確な検査結果を提供することが期待され，今後，他の菌種についても作成中である．また，我々は，これまで感染症サーベイランスの一環として，以前から県内の定点医療機関で分離された病原微生物の検出情報を集計し，情報還元し，NESID（感染症サーベイランスシステム）のサブシステムである病原体検出情報システムにも報告している．これらの情報もFileMakerに移行し，現在，効率的な情報還元方法を検討している．今後，当所のホームページを利用した情報還元等に利用していきたい．

今回，細菌ゲノム配列情報に基づいたin-houseデータベースの構築と活用法について報告した．細菌ゲノム配列情報は，日々，増大しており，新しい情報が追加されており，感染症や食中毒の流行する菌種や感染様式も時代とともに変化しているのも事実である．従って，今回，構築したデータベースも時代の変化に合わせて，修正，追加だけでなく，新しいデータベースを構築する努力も必要になる．今後は，日々の業務を支援するデータベースをどう維持していくか，システム化することが重要である．

謝　辞

この調査研究は，平成29～31年度の富山県重点事業「病原体の遺伝子情報基盤整備による感染症対策強化事業」により遂行された．

参考文献

1．DDBJ（DNA Data Bank of Japan）, https://www.ddbj.nig.ac.jp/services.html（2020年7月1日アクセス可能）
2．NCBI（National Center for Biotechnology Information）, https://www.ncbi.nlm.nih.gov/nucleotide/（2020年7月1日アクセル可能）
3．Aziz RK, Bartels D, Best AA, et al.（2008）. BMC Genomics, 9:75
4．秀丸エディタ, https://hide.maruo.co.jp/software/hidemaru.html（2020年7月1日アクセス可能）
5．Carver T, Harris SR, Matthew Berriman M, et al.（2012）. Bioinformatics, 28（4）, 464–469
6．Sequence Assistant Ver2.7, http://www2s.biglobe.ne.jp/~haruta/（2020年7月1日 アクセス可能）
7．Ruby script, https://www.biostars.org/p/3722/（2020年7月1日アクセス可能）
8．FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/（2020年7月1日アクセル可能）
9．Bolger AM, Lohse M, Usadel B.（2014）. Bioinformatics, 30, 2114–2120
10. Marçais G, Delcher AL, Phillippy AM, et al.（2018）.PLoS Comput. Biol., 14, e1005944
11. Kozlov A, Darriba D, Flouri T, et al.（2019）. Bioinformatics, 35, 4453–4455
12. FigTree, http://tree.bio.ed.ac.uk/software/figtree/（2020年7月1日アクセス可能）
13. NCBI Beta-Lactamase Data Resources, https://www.ncbi.nlm.nih.gov/pathogens/beta-lactamase-data-resources/（2020年7月1日アクセス可能）
14. 地方衛生研究所全国協議会・国立感染症研究所.（2016）. 病原体検出マニュアル　薬剤耐性菌，2016年12月版
15. Watahiki M, Kawahara R, Suzuki M, et al.（2020）. Jpn J Infect Dis., 73, 166–172
16. Tenover FC, Arbeit RD, Goering RV, et al.（1995）. J Clin Microbiol., 33, 2233–2239
17. Ooka T, Terajima J, Kusumoto M, et al.（2009）. J Clin Microbiol., 47, 2888–2894
18. Izumiya H, Pei Y, Terajima J, et al.（2010）. Microbiol.Immunol., 54, 569–577
19. Maiden MC, Bygraves JA, Feil E, et al.（1998）.Proc Natl Acad Sci USA, 95, 3140–3145
20. Octavia S, Lan R. J Clin Microbiol.,（2007）. 45, 3795–3801
21. Walker TM, Clp CL, Harrell RH, et al.（2013）. Lancet Infect Dis., 13, 137–146
22. Maiden MC, Van Rensburg MJ, Bray JE, et al.（2013）. Nat Rev Microbiol., 11, 728–736