計畫編號：MOHW107-CDC-C-114-122508

衛生福利部疾病管制署 107 年委託科技研究計畫

計畫名稱：發展病原菌株全基因體 DNA 序列分析技術與應用平台

107 年 度 研 究 報 告

執行機構：財團法人國家衛生研究院

計畫主持人：熊昭

協同主持人：廖玉潔

研究人員：劉儼毅、余梓芸

執行期間：107 年 1 月 1 日至 107 年 12 月 31 日

＊本研究報告僅供參考，不代表本署意見，如對媒體發布研究成果應事先徵求本署同意*

目錄

	頁　碼
目錄	
計畫中文摘要	
計畫英文摘要	
計畫內容	
一、前言	（4~5　）
二、材料與方法	（6~8　）
三、結果	（9~17　）
四、討論	（18~19）
五、結論與建議	（20　　）
六、計畫重要研究成果及具體建議	（21　　）
七、參考文獻	（22~23）
	計 23 頁

摘要

我們於107年度持續開發「病原菌株全基因體DNA序列分析技術與應用平台」相關軟體與資料庫。包含「開發利用菌株 core genome 產生 SNP profiles與phylogenetic tree之軟體工具」、「開發細菌全基因體抗藥性基因鑑定工具」、「建立鮑氏不動桿菌」PGAdb、「建立克雷氏肺炎桿菌」PGAdb、利用鮑氏不動桿菌、克雷氏肺炎桿菌PGAdb產生相應之wgMLST等。開發項目中利用菌株core genome產生SNP profile與phylogenetic tree之工具，我們引入了 MDS 的視覺化呈現方式，可以協助使用者以很簡單的方式進行群聚偵測。而全基因體抗藥性基因鑑定工具的開發則整合了106年度所開發之「genetic profiling工具」除了可以快速檢測具有相似抗藥性pattern之菌株外，也為之後 metagenome 的抗藥性檢測工具開發提供基礎。另外，本年度所建立之 PGAdb 以及相關軟體工具也將交付給疾管署進行後續之評估與應用。

關鍵詞： 次世代定序、全基因體多位址序列分型、泛基因體對偶基因資料庫

Abstract

This year, we continuely developed programs and databases of pathogenic whole-genome DNA sequencing analysis technology and application plateform. These items include “development of tools for generating SNP profiles and phylogenetic tree based on bacterial core genomes (PathoBacTyper)”, “development tools of bacterial whole-genome AMR genes detection (AMR-dt)”, “*A. baumannii* PGAdb construction”, “*k. pneumonia* PGAdb construction” and “*A. baumannii* and *k. pneumonia* wgMLST profiling”. In PathoBacTyper, we introduced the multi-dimisional scaling method (MDS) for visulising clustering results. The MDS method can assist user doing outbreak detection in a very trivel way. The AMR-dt tool integrated the “genetic profiling tool”, which is developed in 2017, that can fast detect strains with similar AMR patterns and provide a basis for AMR detection of metagenomics. On the other hand, the PGAdbs and tools developed this year will deliver to Taiwan CDC for furture evaluation and field application.

keywords： Next Generation Sequencing (NGS), Whole-genome Multiple Locus Sequence Typing (wgMLST), Pangenome allele database (PGAdb)

一、前言：

由於 WGS 技術越來越普及，定序成本也越來越低廉，應用 WGS 技術來實踐流行病學的監測逐漸成為趨勢。傳統的分子檢驗方法通常需要投入較多的實驗室人力，而且如果需要得到更多的資訊例如抗藥性、致病性、質體種類等又需要再啟動其他的實驗，對於時效性以及投入成本上的掌握不易。近年來，次世代定序技術的成熟，加上定序成本的降低，有越來越多成功案例顯示 WGS 應用於院感監測與調查的優勢。另外還有一點是傳統分子檢驗技術所做不到的，那就是 metagenomics 的技術，這是能直接從檢體，不經過細菌培養與分離單一菌株的步驟，直接定序分析菌種的組成，這項技術雖然不適合拿來做院感事件調查的分子檢驗工具，但是在檢測未知感染病源佔有很大優勢，多個成功的例子也被發表在高影響力的期刊(Koch, 2016; Lipkin, 2013; Miller, Montoya, Gardy, Patrick, & Tang, 2013; Pallen, 2014; Rose et al., 2015)。

2018 年開始，李斯特菌將加入法定傳染病通報名單，從 2013 開始美國 CDC 就已經開始評估 WGS 於李斯特菌群聚的偵測效果(https://goo.gl/j6UfiQ)，並證實確實可以得到較過去 PFGE 更精確的結果。另外，法國巴斯德研究院也在啟動大規模的評估之後將 WGS 的技術正式應用於李斯特菌的監測(Moura et al., 2016)。而在院內感染控制方面，院內感染菌對於醫院來說是必需要面對的重要議題，控制的方法主要就是要能找到其傳播的路徑，由英國惠康基金會(Welcome Trust)所支持的針對四種院內感染菌的全基因體定序應用在院內感染途徑的研究上，指出了 WGS 應用於院感傳播路徑調查的潛力。這個研究也於 2013 發表在國際知名的醫學期刊新英格蘭醫學期刊(New England Journal of Medicine)(Eyre et al., 2013). 除了群聚偵測外，抗藥性的檢測也是疾病監測上很重要的工作，尤其目前多重抗藥性的議題越來越受到國際的重視，藉由比對預先建立的抗藥性基因資料庫，即可利用 WGS 來預測菌株抗藥性，多篇國際研究也成功運用 WGS 來預測細菌抗藥性(Gordon et al., 2014; Koser, Ellington, & Peacock, 2014; Nair et al., 2016; Punina, Makridakis, Remnev, & Topunov, 2015; Tyson et al., 2015)。

以全基因體為基底的 whole-genome MLST (wgMLST)分型技術是繼 PFGE 分型技術之後在分子流行病學上最具發展潛力的檢測技術(Nadon et al., 2017)。而如何建置與評估

PGAdb 技術是上述公衛應用的基礎。本計畫在 106 年度已經完成「PGAdb 建置工具」、-「運用 PGAdb 之 wgMLST profiling 工具」、「利用 wgMLST profile 之 phylogenetic tree 建置工具」以及「Salmonella-PGAdb」。本年度計畫，應用上述已開發之工具程式分別建置李斯特菌(*Listeria monocytogenes*)、鮑式不動桿菌(*Acinetobacter baumannii*)、克雷式肺炎桿菌(*Klebsiella pneumoniae*)之 PGAdb，並且與疾管署共同評估這三種菌的 PGAdb 在流病分型上的效力。評估好的 PGAdb 被用來產生 wgMLST profile，我們建置了李斯特菌、鮑式不動桿菌、雷式肺炎桿菌 wgMLST profile 資料庫可供後續定義 wgMLST type 甚至是利用來建立預測模型的基礎。另外，由於單純透過 wgMLST profile 的比較僅能呈現菌株之間的基因組相似性，無法用來計算演化關係，因此，我們在本年度計畫中開發以 PGAdb 為基底的 core-genome SNP calling 以及演化樹建立工具(PathoBacTyper)，可提供病原菌溯源之應用。以上應用為診斷部分的發展。而在治療決策輔助方面，本年度計畫中我們也開發了以 PGAdb 為基底的抗藥性基因檢測用來預測菌株抗藥性以協助公衛防治與疾病治療。

二、材料與方法

計畫期程共計 2 年，107 年度為最後一年，預定「開發利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具」、「開發細菌全基因體抗藥性基因鑑定工具」、建立「院內感染病原菌鮑氏不動桿菌(*A. baumannii*)與克雷氏肺炎桿菌(*K. pneumonia*)及食媒性病原菌李斯特菌(*L. monocytogenes*)泛基因體等位基因資料庫(PGAdb)」，建立「鮑氏不動桿菌(*A. baumannii*)、克雷氏肺炎桿菌(*K. pneumonia*)及李斯特菌(*L. monocytogenes*) 全基因體多位址序列分型圖譜(wgMLST)」。開發工具方法詳述如下:

1. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌全基因體等位基因資料庫:

我們從公共資料庫 NCBI 下載鮑氏不動桿菌 2625 條、克雷氏肺炎桿菌 3659 條、李斯特菌 2231 條之全基因體序列資料，並利用第一年開發的 PGAdb 建置工具來建置 PGAdb。之後與疾管署合作進行評估與測試。評估方式以能否準確預估群聚感染為原則，我們將從泛基因組資料庫挑選不同的 panel 來做測試，利用已知的群聚感染菌株全基因體序列資料當測試資料集，經由第一年開發的 wgMLST 產生工具轉換成 wgMLST profile，進一步應用第一年開發的親緣關系樹建立工具計算出測試菌株彼此間的演化關聯性，並與已知之流行病學資料進行對照，藉此評估資料庫應用於群聚預測的準確性。

2. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST profile:

利用第一年已建立之 wgMLST profiling 工具建立鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST。目前先建置由 NCBI Genome 下載的鮑氏不動桿菌、克雷氏肺炎桿菌全基因體多位址分型波譜資料庫。

3. 細菌全基因體抗藥性基因找尋工具:

我們透過收集已經過發表驗證之細菌抗藥基因，這些收集到的基因序列被製作成等位基因資料庫，抗藥同源基因群集組將被定義為一個 panel，透過我們自行開發的 wgMLST 產生工具搭配我們製作的等位基因資料庫，即可從使用者輸入的菌株全基因體資料中截取出抗藥基因 panel 的等位基因波譜(allelic profile)與序列(allelic sequence)。

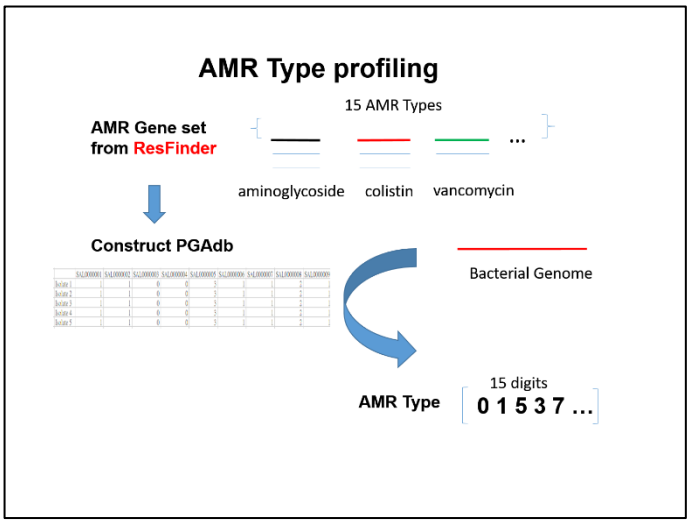
本研究中我們選用 ResFinder 抗藥基因資料庫(Zankari et al., 2012)，裡面含有 15 類抗生素共包含有 3037 條抗藥基因序列(圖 1)，我們將此抗藥基因資料庫建置成具有 15 個 loci 的 PGAdb，透過此 PGAdb 再加上 wgMLST profiling 工具即可以將菌株基因體

序列轉換成抗藥波譜，達到抗藥性偵測的目的，此流程如圖 2 所示。

圖 1

抗生素	抗藥基因序列
aminoglycoside	247
beta-lactam	1963
colistin	52
fosfomycin	40
fusidicacid	2
glycopeptide	44
macrolide	174
nitroimidazole	14
oxazolidinone	24
phenicol	45
quinolone	121
rifampicin	9
sulphonamide	54
tetracycline	143
trimethoprim	105

圖 2



4. 利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具:

我們撰寫程式自動化執行 SNP caller，讓輸入的菌株序列(raw reads and assemblies）產生 SNP profile，進一步利用這些 SNP profile 計算演化距離與建構 phylogenetic tree (Maximum likelihood)，ETE3 python module 最後被利用來做演化樹視覺的呈現。完整的分析流程以網站形式展示。所設計的 pipeline 包含 2 個主要功能:「**物種辨識**」及「**細菌分型**」。在物種辨識的模組裡，我們預先建置一個包含有 478 個物種(Scholz et al., 2016)的參考序列資料庫，透過與參考序列資料庫做比對(mapping)我們可以辨識比對的物種。在這裡，我們設計了一個 hash-based 的比對方法，分別將參考序列與比對序列切成以 25-mer 長度為單位的組合，這些組合存在 hash table 當中，mapping 的時候則 randomly 取 10%比對序列的 25-mer 組合與參考序列的 hash table 做比較，並且計算與參考序列的覆蓋率，依照覆蓋率高到低排列，最高者為建議之預測物種，網頁會列出覆蓋率 top10 的物種，假使覆蓋率很低，也許是比對物種不在參考序列資料庫中，則需要自行輸入參考物種序列。而比對到的物種序列將當成 SNP calling 的參考序列，

這邊串接的 SNP caller 是 Lyve-SET(Katz et al., 2017)（如果 input 是 assembly）以及 Parsnp(Treangen, Ondov, Koren, & Phillippy, 2014)（如果 input 是 raw reads）。SNP calling 所產生的 profile 將進一步被拿來計算演化距離，這邊使用 FastTree(Price, Dehal, & Arkin, 2009)來計算演化距離與建構 SNP tree。演化樹則利用 python ETE3(Huerta-Cepas, Serra, & Bork, 2016)模組建置。比較特別的是，我們另外利用菌株之間的 SNP 距離(Euclidean distance)來計算 MDS 的分布圖，可以快速檢視比對的菌株之間基因體相似程度，提供群聚偵測的視覺化工具。圖 3 顯示我們 pipeline 的流程。

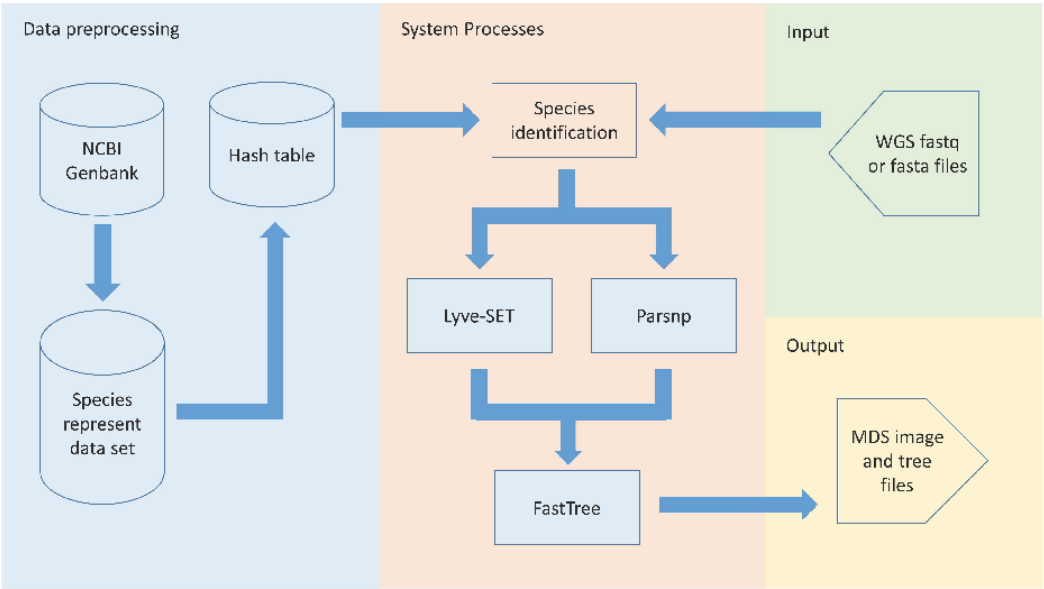


圖 3. PathoBacTyper 程式流程圖

三、 結果

1. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌全基因體等位基因資料庫:

為了清楚描述 PGAdb 的內容，我們藉由展示「locus 在不同菌株基因體出現率」以及「locus 中 allele 長度分佈」視覺化呈現資料庫中的資料分佈型態。各病原菌皆以 4 張圖表展現，以 *A. baumannii* PGAdb 為例解釋各圖如下：

(圖 4.2) PGAdb 中 locus 在不同菌株基因體出現率分佈，

(圖 4.3) PGAdb 中 locus 在不同菌株基因體出現率分佈(去除出現率小於 2%的 locus)，

(圖 4.4) PGAdb 中 locus 在不同菌株基因體出現率累進分佈，

(圖 4.5) PGAdb 中 locus 中 allele 長度分佈(僅示範性呈現前 103 個 locus)。

(1) *A. baumannii* PGAdb

我們由 NCBI Genome 資料庫下載鮑氏不動桿菌 fasta 序列共計 2,625 筆，並利用「PGAdb 建置工具」建構 PGAdb。下面分佈圖呈現資料庫中 locus 內含 allele 數量的統計:

圖 4.1

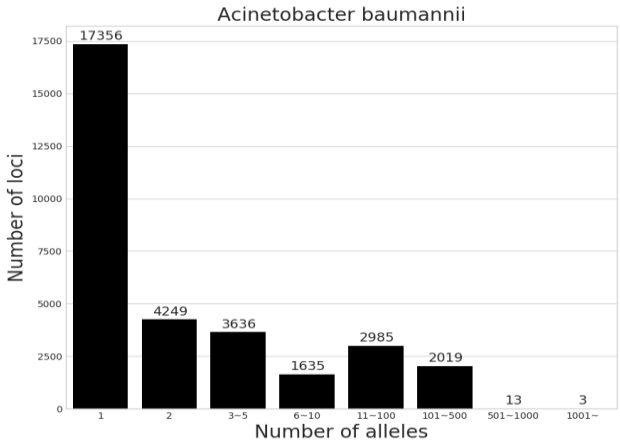


圖 4.1 中顯示絕大多數的 locus 僅含有 1 個 allele，表示 unique gene/locus 在 PGAdb 中佔大多數。

圖 4.2

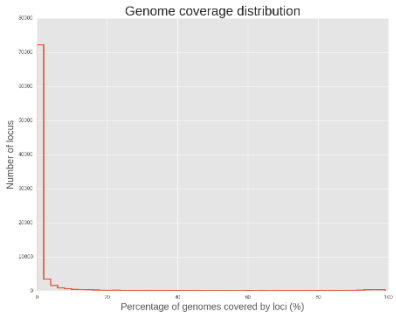


圖 4.3

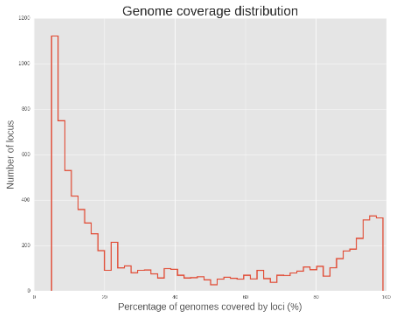


圖 4.4

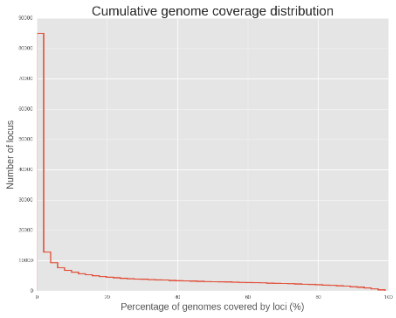
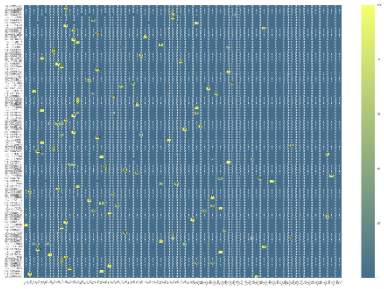


圖 4.5



「圖 4.2」顯示在所有的 locus 中，在不同菌株基因體出現率極低(< 2%)的 locus 佔大多數，unique locus (<1%)數量佔了將近 90%，去掉出現率< 2%的 locus，得到了如「圖 4.3」的分佈，這是一個雙峰分佈，在出現率大於 80% (core-gene region)以及出現率小於 20% (unique-gene region)處皆有一個峰值，代表用來建庫的菌株基因體之間存在差異，而雙峰之間(20%~80%, dispensable-gene region)則是平緩曲線，代表建庫菌株無存在多筆序列高度相似的基因體資料。「圖 4.4」是累進分佈圖，可以用來輔助「圖 4.2」的解讀。「圖 4.2」至「圖 4.4」可以提供我們了解用來建庫的菌株基因體相似性分佈情況，以及評估可利用來做後續基因型鑑定(genotyping)的 locus 數量。「圖 4.5」示範性展示前 103 個 locus allele 長度的分佈情形，顏色愈偏黃代表數量愈多，分析結果顯示 locus 中長度最長的 allele 通常數量也最多(眾數)，這可以提供我們挑選每個 locus 代表序列的參考。

(2) *K. pneumonia* PGAdb

我們由 NCBI Genome 資料庫下載鮑氏不動桿菌 fasta 序列共計 3,659 筆，並利用「PGAdb 建置工具」建構 PGAdb。圖 5.1 呈現資料庫中 locus 內含 allele 數量的統計:

圖 5.1

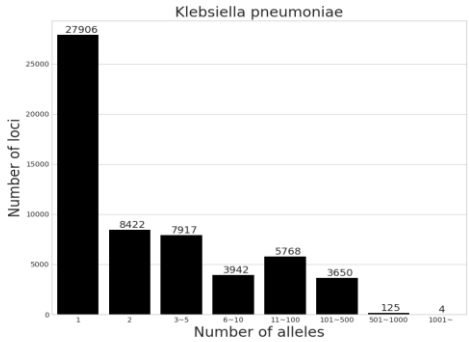


圖 5.1 中顯示絕大多數的 locus 僅含有 1 個 allele，表示 unique gene/locus 在 PGAdb 中佔大多數。

圖 5.2

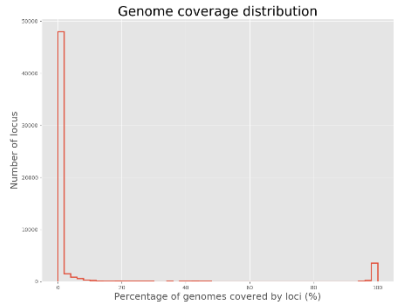


圖 5.3

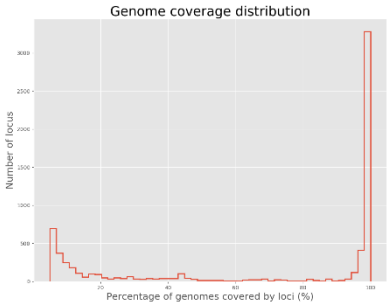


圖 5.4

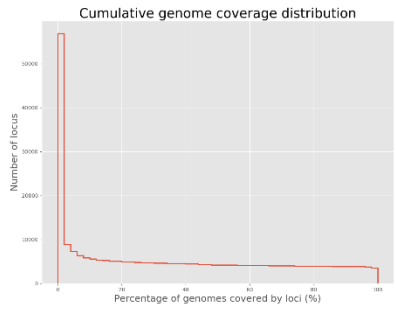
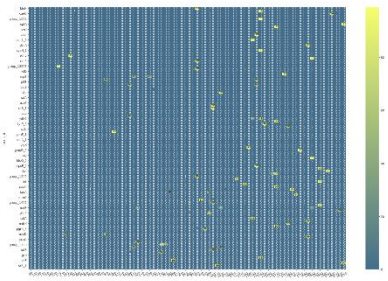


圖 5.5



.

「圖 5.2」顯示在所有的 locus 中，在不同菌株基因體出現率極低(< 2%)的 locus 佔大多數，unique locus (<1%)數量佔了將近 90%，去掉出現率< 2%的 locus，得到了如「圖 5.3」的分佈，這是一個雙峰分佈，在出現率大於 80% (core-gene region)以及出現率小於 20% (unique-gene region)處皆有一個峰值，但以出現在 core-genome region 的 peak 為主，意味著 *K. pneumonia* 基因體高度相似，因此 core-genome 很 conserve，更進一步推，建庫菌株中出現率<2%的比例卻一樣很高，意味著 *K. pneumonia* 族群中可能涵蓋有多樣性的質體序列。雙峰之間(20%~80%, dispensable-gene region)則是平緩曲線，代表建庫菌株無存在多筆序列高度相似的基因體資料。

(3) *L. monocytogenes* PGAdb

我們由 NCBI Genome 資料庫下載李斯特菌 fasta 序列共計 2,231 筆，並利用「PGAdb 建置工具」建構 PGAdb。圖 6.1 呈現資料庫中 locus 內含 allele 數量的統計:

圖 6.1

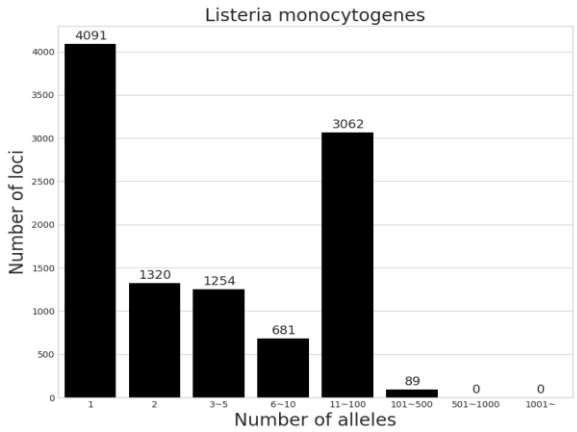


圖 6.1 中顯示絕大多數的 locus 僅含有 1 個 allele，表示 unique gene/locus 在 PGAdb 中佔大多數，但是也有極高比例的 loci 中 allele 的數量在 11~100 之間，推測可能是 *L. monocytogenes* 基因體不同個體間基因的保守性較高所致。

圖 6.2

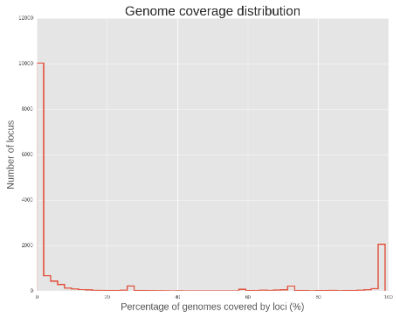


圖 6.3

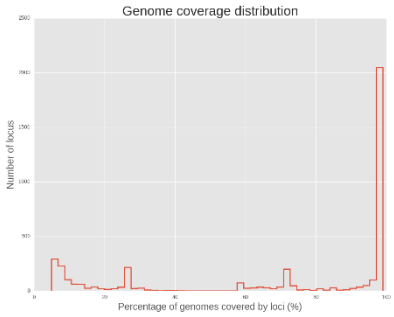


圖 6.4

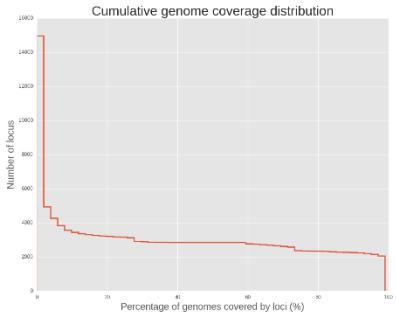
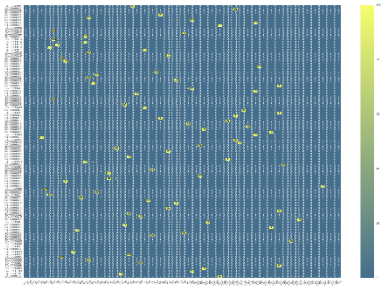


圖 6.5



「圖 6.2」顯示在所有的 locus 中，在不同菌株基因體出現率極低(< 2%)的 locus 佔大多數，unique locus (<1%)數量佔了將近 90%，去掉出現率< 2%的 locus，得到了如「圖 6.3」的分佈，這是一個雙峰分佈，在出現率大於 80% (core-gene region)以及出現率小於 20% (unique-gene region)處皆有一個峰值，但出現在 core-genome region 的 peak 比 20% region 高很多，意味著 *L. monocytogenes* 基因體高度相似，因此 core-genome 很 conserve，更進一步推，建庫菌株中出現率<2%的比例卻一樣很高，意味著 *L. monocytogenes* 族群中可能涵蓋有多樣性的質體序列。雙峰之間(20%~80%, dispensable-gene region)則是平緩曲線，代表建庫菌株無存在多筆序列高度相似的基因體資料。

2. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST profile:

我們利用 106 年度已開發的 wgMLST profiling 工具透過本年度建置的 *A. baumannii*, *K. pneumonia*, *L. monocytogenes* PGAdb 將每個菌株的基因體序列轉換成 wgMLST

(cgMLST) profiles。每個菌株基因體序列先進行基因預測，之後將預測的基因 (Open Reading Frame, ORF) 序列在已建置好的 PGAdb 中查找相對應(序列相同)的 allele 序列，如果有找到就給予 PGAdb 中已經定義好的 allele number。下表 1 為輸出範例:

表 1. wgMLST profile 範例

	SAL0000001	SAL0000002	SAL0000003	SAL0000004	SAL0000005	SAL0000006	SAL0000007	SAL0000008	SAL0000009	SAL0000010	SAL0000011	SAL0000012	SAL0000013	SAL0000014	SAL0000015	SAL0000016
Isolate 1	1	1	0	0	3	1	1	2	1	1	1	2	1	1	1	1
Isolate 2	1	1	0	0	3	1	1	2	1	1	1	2	1	1	1	1
Isolate 3	1	1	0	0	3	1	1	2	1	1	1	2	1	1	1	1
Isolate 4	1	1	0	0	3	1	1	2	1	1	1	2	1	1	1	1
Isolate 5	1	1	0	0	3	1	1	2	1	1	1	2	1	1	1	1

範例中「SALxxxxxxx」代表 locus name，表格內的數字代表 isolate 在某個 locus 中所找到相對應 allele 的編號，每一列代表一個菌株基因體的資料，此範例顯示 5 個菌株在前 16 個 locus profiling 的結果。我們利用此工具分別建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST profile 並將這些 profile 以資料庫形式儲存這些資料。

3. 細菌全基因體抗藥性基因找尋工具:

我們利用此工具測試已知抗藥性之菌株以評估工具之可靠性。利用已知抗藥性資料之沙門氏菌菌株共 19 株進行測試(自行進行 NGS 定序及藥敏試驗之菌株)，並計算預測之準確率(ACC)，結果如下表 2:

表 2. 抗藥性預測效果評估

Isolates	ACC	Isolates	ACC
CD14.202	0.93	R16.0309	0.86
R15.0334	1.00	R16.0460	0.79
R15.0475	0.93	R16.0676	0.93
R15.0609	0.86	R16.2821	0.93
R15.0612	1.00	R16.3355	0.93
R15.0614	0.93	R17.0464	0.93
R15.0626	1.00	R17.1311	0.86
R15.1006	1.00	R17.3022	0.86
R16.0193	0.93	R17.3161	0.93
R16.0267	0.86		

結果顯示我們所開發的工具能準確偵測相對應的抗藥基因。各菌株之抗藥性預測與藥敏試驗結果之比較如下表 3:

表 3. 各菌株之抗藥性預測與藥敏試驗結果之比較。標註”1”代表有此類抗生素之抗藥性，黑色為預測之抗藥性，紅色為藥敏試驗之結果。欄位 A~N 分別代表 A: aminoglycoside, B: beta-lactam, C: colistin, D: fosfomycin, E: fusidicacid, F: glycopeptide, G: nitroimidazole, H: oxazolidinone, I: phenicol, J: quinolone, K: rifampicin, L: sulphonamide, M: tetracycline, N: trimethoprim。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
CD14.202	1	1	1						1	1	1	1	1	1
	1	1	1						1	1		1	1	1
R15.0334	1	1							1			1	1	1
	1	1							1			1	1	1
R15.0475	1	1							1	1		1	1	1
	1	1							1			1	1	1
R15.0609	1	1										1	1	
	1	1										1		
R15.0612	1	1										1	1	
	1	1										1	1	
R15.0614	1	1							1			1	1	1
	1	1							1	1		1	1	1
R15.0626	1	1	1						1			1	1	
	1	1	1						1			1	1	
R15.1006	1	1							1			1	1	
	1	1							1			1	1	
R16.0193	1	1							1	1		1	1	1
	1	1							1			1	1	1
R16.0267	1	1							1			1		1
	1	1							1	1		1	1	1
R16.0309	1	1							1			1	1	1
	1	1							1				1	
R16.0460	1								1					
		1							1	1				
R16.0676	1	1							1	1		1	1	1
	1	1							1			1	1	1
R16.2821	1													
R16.3355	1	1							1	1		1	1	1
	1	1							1			1	1	1
R17.0464	1	1										1		
	1	1								1		1		
R17.1311	1	1							1	1			1	1
	1	1							1				1	
R17.3022	1			1										
R17.3161	1	1							1	1		1	1	1
	1	1							1			1	1	1

4. 利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具:

為了測試我們所設計的 pipeline，我們利用 Bekal 等人所發表的 59 株 *S. Heidelberg* 菌

株(Bekal et al., 2016) 進行測試。我們從 NCBI SRA 資料庫將這 59 株菌株的 raw reads 下載回來，利用 CLC v9.5.2 軟體進行 de novo assembly。我們同時對 pipeline 進行菌株 assembly 以及 raw reads 輸入測試。59 株 *S. Heidelberg* Assembly 以及 raw reads 分別花了 587 以及 101 分鐘完成。我們使用的伺服器規格為 two 2.40-GHz Intel Xeon processors comprising 16 cores。我們的 pipeline 可從 59 株菌株中成功區分出 3 群不同的 outbreak 菌株，由分析結果得知，被歸在同一個 cluster 的菌株，彼此間差異在 10 個 SNP 以內，而不同 cluster 的菌株彼此間的差異皆大於 50 個 SNP。而利用 maximum likelihood 方法所建構的 SNP tree 其菌株親緣關係也與 Bekal 等人所發表的結果一致。另外，從 SNP 個數差異矩陣所描繪的 MDS 也可清楚呈現三個 cluster。利用 59 株 *S. Heidelberg* 所建構的 MDS (圖 7.1)與親緣關係樹(圖 7.2)。

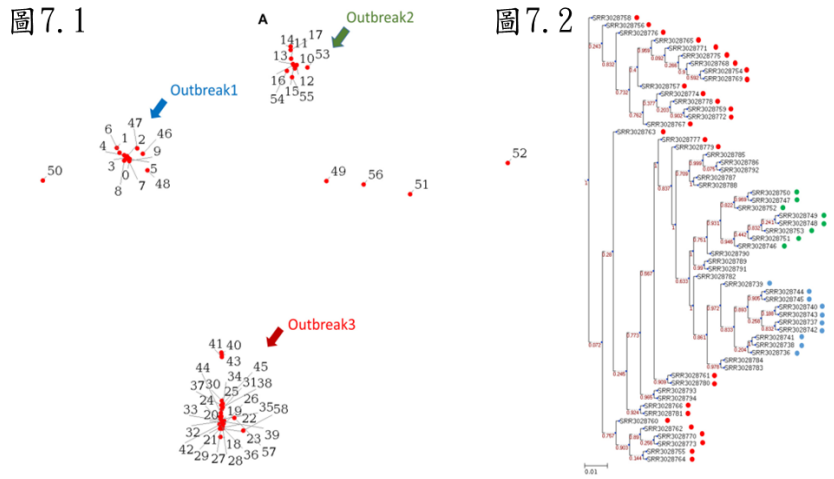


圖 7. 利用 59 株 *S. Heidelberg* 所建構的 MDS (圖 7.1)與親緣關係樹(圖 7.2)

我們設計一個讓使用者可以簡單透過網頁上傳菌株全基因體序列資料後，自動進行參考菌株的挑選，以及利用此參考菌株進行親緣關係的建構，並提供 MDS 的視覺化群集分析方法，所顯示的計算結果皆可提供下載。使用者可藉由提供的親緣關係樹以及 MDS 偵測群聚感染菌株。由於許多公衛實驗室業務同仁較不具備生物資訊分析的技術以及資訊設備，在取得菌株全基因體序列資料之後常常不知如何處理。此網頁工具平台將可針對手上已定序完的菌株序列資料提供直覺化的分析功能。下圖 8 展示本工具網頁介面：

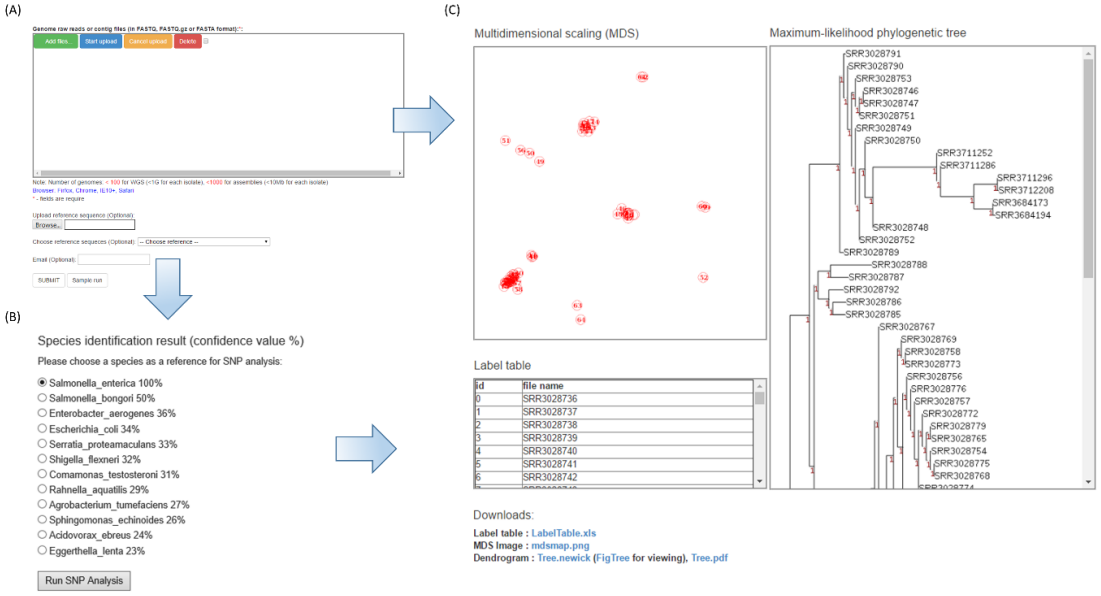


圖 8. 網頁操作介面與結果頁示意圖

四、 討論

1. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌全基因體等位基因資料庫

雖然目前已分別利用上千株菌株全基因體序列建置全基因體等位基因資料庫，但是以定序技術的發展與價格的調降，公共資料庫上的菌株全基因體序列增加速度很快，理想上應該需要撰寫自動化更新程式，並且規劃將資料庫放置於穩定的雲端服務環境中以利後續的使用與維護，另外，如果進行資料庫更新，前後版本的相容性問題也還需要解決。這部分需要與疾管署相關實驗室進行評估與討論。

2. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST profile

由於 wgMLST profile 的產生是以全基因體等位基因資料庫為基底，因此，如果資料庫進行更新產生前後版本相容性問題的話，所產生的 wgMLST profile 也會因版本而不同，造成資料在跨實驗室交換與比對的時候會有問題。這部分還是需要與疾管署相關實驗室討論解決。

3. 細菌全基因體抗藥性基因找尋工具

利用本計畫所開發的 PGAdb 建置工具與 wgMLST profiling 工具，可以將抗藥基因序列資料庫轉換成 PGAdb，而後基於此 PGAdb 利用 wgMLST profiling 工具可進行抗藥性圖譜的 profiling，達到抗藥基因偵測的目的。但是利用檢測抗藥基因預測抗藥性的方法有二個根本性的問題，其一，此方法僅能適用於資料庫有收集到的抗藥基因的檢測，如果菌株基因體含有新的抗藥基因就無法檢測到，因此，還是需要持續收集更多的抗藥基因;其二，抗藥基因與抗藥性的關聯性還需要有更多的研究佐證。基於以上，此工具開發的目的是在快速提供公衛或臨床人員菌株抗藥性的參考，達到「定序一次，多方面使用」的目的。

4. 利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具

我們設計一個讓使用者可以簡單透過網頁上傳菌株全基因體序列資料後，自動進行參考菌株的挑選，以及利用此參考菌株進行親緣關係的建構，並提供 MDS 的視覺化群集分析方法，所顯示的計算結果皆可提供下載。使用者可藉由提供的親緣關係樹以及 MDS 偵測群聚感染菌株。由於許多公衛實驗室業務同仁較不具備生物資訊分析的技術以及資訊設備，在取得菌株全基因體序列資料之後常常不知如何處理。此網頁工具平台將可針對手上已定序完的菌株序列資料提供直覺化的分析功能。

五、結論與建議

結論部分

本年度(107)是計畫的最後一年，已完成 1. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌全基因體等位基因資料庫、2. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌 wgMLST profile、3. 細菌全基因體抗藥性基因找尋工具、4. 利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具。所建置之全基因體等位基因資料庫、wgMLST profile 資料庫與各項工具程式將交由疾管署相關實驗室進行後續之更新與維護。

建議部分

目前，委外計畫僅通過 2 年，與原先規劃的 4 年期程縮短不少，因此，工具開發後的效能評估、資料庫更新版本相容性問題、資料庫與工具程式後續維運問題、甚至是後續的雲端上線等，在與疾管署交接後，相關的問題必須要妥善評估，避免造成所開發的工具後續無法修改與維護的問題。

六、計畫重要研究成果及具體建議

本年度(107)主要完成之具體成果如下:

1. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌全基因體等位基因資料庫
2. 建置鮑氏不動桿菌、克雷氏肺炎桿菌、李斯特菌wgMLST profile
3. 細菌全基因體抗藥性基因找尋工具
4. 利用菌株core genome產生SNP profiles與phylogenetic tree之軟體工具

由於計畫期程由4年縮短至2年，僅能完成所羅列之開發項目。工具開發後的優化與雲端服務上線等工作勢必要交付疾管署完成後續工作。相關交接事項已與疾管署會商，以利後續系統與資料庫的修改與維護。

七、參考文獻：

Bekal, S., Berry, C., Reimer, A. R., Van Domselaar, G., Beaudry, G., Fournier, E., . . . Tremblay, C. (2016). Usefulness of High-Quality Core Genome Single-Nucleotide Variant Analysis for Subtyping the Highly Clonal and the Most Prevalent Salmonella enterica Serovar Heidelberg Clone in the Context of Outbreak Investigations. *J Clin Microbiol, 54*(2), 289-295. doi:10.1128/JCM.02200-15

Eyre, D. W., Cule, M. L., Wilson, D. J., Griffiths, D., Vaughan, A., O'Connor, L., . . . Walker, A. S. (2013). Diverse sources of C. difficile infection identified on whole-genome sequencing. *N Engl J Med, 369*(13), 1195-1205. doi:10.1056/NEJMoa1216064

Gordon, N. C., Price, J. R., Cole, K., Everitt, R., Morgan, M., Finney, J., . . . Golubchik, T. (2014). Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol, 52*(4), 1182-1191. doi:10.1128/JCM.03117-13

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol, 33*(6), 1635-1638. doi:10.1093/molbev/msw046

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., . . . Carleton, H. A. (2017). A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front Microbiol, 8*, 375. doi:10.3389/fmicb.2017.00375

Koch, L. (2016). Metagenomics: Pathogen diagnostics for the masses. *Nat Rev Genet, 17*(7), 378. doi:10.1038/nrg.2016.76

Koser, C. U., Ellington, M. J., & Peacock, S. J. (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet, 30*(9), 401-407. doi:10.1016/j.tig.2014.07.003

Lipkin, W. I. (2013). The changing face of pathogen discovery and surveillance. *Nat Rev Microbiol, 11*(2), 133-141. doi:10.1038/nrmicro2949

Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., & Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med, 5*(9), 81. doi:10.1186/gm485

Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., . . . Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat Microbiol, 2*, 16185. doi:10.1038/nmicrobiol.2016.185

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., . . . Panel, F.-N. E. (2017). PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill, 22*(23). doi:10.2807/1560-7917.ES.2017.22.23.30544

Nair, S., Ashton, P., Doumith, M., Connell, S., Painset, A., Mwaigwisya, S., . . . Day, M. (2016). WGS for surveillance of antimicrobial resistance: a pilot study to detect the prevalence and mechanism of resistance to azithromycin in a UK population of non-typhoidal Salmonella. *J Antimicrob Chemother*. doi:10.1093/jac/dkw318

Pallen, M. J. (2014). Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology, 141*(14), 1856-1862. doi:10.1017/S0031182014000134

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol, 26*(7), 1641-1650. doi:10.1093/molbev/msp077

Punina, N. V., Makridakis, N. M., Remnev, M. A., & Topunov, A. F. (2015). Whole-genome sequencing targets drug-resistant bacterial infections. *Hum Genomics, 9*, 19.

doi:10.1186/s40246-015-0037-z
Rose, G., Wooldridge, D. J., Anscombe, C., Mee, E. T., Misra, R. V., & Gharbia, S. (2015). Challenges of the Unknown: Clinical Application of Microbial Metagenomics. *Int J Genomics, 2015*, 292950. doi:10.1155/2015/292950
Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., . . . Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods, 13*(5), 435-438. doi:10.1038/nmeth.3802
Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol, 15*(11), 524. doi:10.1186/PREACCEPT-2573980311437212
Tyson, G. H., McDermott, P. F., Li, C., Chen, Y., Tadesse, D. A., Mukherjee, S., . . . Zhao, S. (2015). WGS accurately predicts antimicrobial resistance in Escherichia coli. *J Antimicrob Chemother, 70*(10), 2763-2769. doi:10.1093/jac/dkv186
Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., . . . Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother, 67*(11), 2640-2644. doi:10.1093/jac/dks261

衛生福利部疾病管制署委託科技研究計畫
107 年度計畫重要研究成果及具體建議

計畫名稱：發展病原菌株全基因體 DNA 序列分析技術與應用平台

主 持 人：熊昭

計畫編號：MOHW107-CDC-C-114-122508

1.計畫之新發現或新發明

(1) 開發利用菌株 core genome 產生 SNP profiles 與 phylogenetic tree 之軟體工具 (PathoBacTyper)

(2) 利用細菌全基因體序列找尋抗藥性基因之平台

2.計畫對民眾具教育宣導之成果

此計畫所開發的分子分型工具導入 MDS 視覺化呈現方式可以輔助群聚感染判斷，可以協助流行病學的調查工作，並進行相關的防疫作為。

3.計畫對醫藥衛生政策之具體建議

在台灣食媒性細菌造成食物中毒案例不勝枚舉，且所分離的細菌中具多重抗藥性的菌株比例也逐漸升高，利用此計畫所開發的工具可以達到快速偵測群聚感染與菌株抗藥性，幫助衛生單位提供民眾食安方面問題的預警。