計畫編號：MOHW106-CDC-C-112-112703

衛生福利部疾病管制署 106 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

106 年 度/全 程 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：傅筱君

執行期間：106 年 1 月 1 日至 106 年 12 月 31 日

目錄

摘要

隨著次世代定序技術的成熟與成本的降低，細菌全基因體定序很快將成為微生物實驗室分析菌株之工具；而如何從大量的序列資料中挖掘出有意義的資訊，吸引許多生物資訊學者的投入研究。在分子流行病學領域，開發簡單準確且符合成本的基因分子分型技術是許多研究的重點；另外，基因分型結果是否能跨實驗室進行比對，是進行跨實驗室疾病監測的重點。全基因體多位址序列分型(whole genome multilocus sequence typing, wgMLST)是目前全球微生物公衛實驗室積極投入發展與評估的細菌基因分子分型方法，擁有傳統多位址序列分型(MLST)可跨實驗室比對的優點，且因加入了更多比對的基因點(loci)而比傳統 MLST 具有更高的區別效力(discriminatory power)。

全基因體序列基因分型(WGS-based genotyping)結果，若要供跨實驗室比對，必須使用共同的泛基因體對偶基因資料庫(pan-genome allele database, PGAdb)產生基因圖譜(genetic profiles)，在應用 wgMLST 分型方法前，必須先建立目標菌株的 PGAdb，並且評估適合於目標菌株分型的 loci 組合(scheme)。106年度計畫已利用自行開發的「PGAdb 建置工具」下載由 NCBI Genome 下載的沙門氏菌全基因體菌株建立沙門氏菌的 PGAdb，並透過自行開發的「genetic profiling 工具」與「親源關係樹建立工具」評估適合分型的 scheme，並由疾病管制署的實驗室進行 PGAdb 的應用評估，評估結果顯示核心基因體多位址序列分型(core genome MLST, cgMLST)具有高的區別效力，是群聚感染感染事件調查與疾病監測之有效工具。

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

With the advance of the next-generation sequencing (NGS), whole genome sequencing of bacterial isolates has become inexpensive. Therefore, how to fetch meaningful information from a huge amount of sequence data is a big challenge. Topics associated with NGS analysis have attracted many bioinformaticists to engage in. In the field of molecular epidemiology, to develop simple, accurate and cost-effective molecular typing techniques are many research focuses. In addition, cross-laboratory comparison is also an important part in developing a new molecular typing technique. Whole-genome multilocus sequence typing (wgMLST) is the up-to-date molecular typing method that CDC is engaged. The wgMLST is inherited from the conventional MLST that can be compared across laboratories and possesses a higher discriminatory power than the MLST because thousands of loci are included for the gene by gene comparison.

For the application of wgMLST approach for comparing across laboratory, a pan-genome allele database (PGAdb) must first be created to generate genetic profiles, and a panel of loci combination (scheme) for a specific species has to be evaluated. In the 106 project, we have built the *Salmonella Enterica* PGAdb based on Salmonella WGS data downloaded from NCBI Genome by using self-developted PGAdb builder. The PGAdb evaluation task was performed by using self-developed “genetic profiling tool” and “phylogenetic tree builder” in cooperation with Taiwan CDC for the typing scheme evaluation. The evaluation results demonstrate that core-genome MLST (cgMLST) possess high discriminatory power and might be an efficient tool for outbreak detaction and disease surveillance.

keywords：Next generation sequencing (NGS), Whole-genom multi-locous sequence typing (wgMLST), Pan-genome allele database (PGAdb)

一、前言：

隨著次世代定序(next generation sequencing, NGS)技術的成熟與成本的降低，細菌全基因體定序(whole genome sequencing, WGS)很快將成為微生物實驗室分析菌株之主要工具。在食媒疾病監控方面，根據美國 CDC 分子分型食媒性疾病監控網絡(PulseNet)對於全基因體序列基因分型(WGS-based genotyping)應用於群聚感染感染偵測的初步評估結果，指出 WGS 基因分型法比目前普及的標準分子分型方法— pulsed-field gel electrophoresis (PFGE)具更高的菌株區別效力(discriminatory power)，而且可在單一次的 WGS 資料擷取過去需分成多項實驗分析才能得到的資訊(Salipante et al., 2015)，能節省更多的人力以及分析成本。美國疾病管制預防中心(US centers for disease control and prevention, USCDC) 2013 年就開始評估應用 WGS 於李斯特菌群聚感染感染的偵測效果，證實確實可以得到較 PFGE 更靈敏精確的偵測效力。也因此，積極開發 WGS 的方法應用於其他的食因性病原菌上例如沙門氏菌(Salmonella)的疾病監控與調查。

多重抗藥院內感染菌是病人健康的嚴重威脅，加重疾病負擔，是醫院感染控制面對的重大難題。英國惠康基金會(Welcome Trust)所支持的針對四種院內感染菌的 WGS 應用在院內感染途徑的研究上，指出了 WGS 是院感傳播路徑調查的強力工具，此研究於 2013 年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)。

除了群聚感染事件的偵測外，抗藥性的檢測也是醫療上很重要的工作，尤其目前多重抗藥性的問題越來越嚴重，藉由比對預先建立的抗藥性基因資料庫，即可利用分析 WGS 序列來鑑定抗藥基因與預測菌株抗藥性，目前已有相當多運用 WGS 來預測細菌抗藥性的論文被發表(Gordon et al., 2014; Koser, Ellington, & Peacock, 2014; Nair et al., 2016; Punina, Makridakis, Remnev, & Topunov, 2015; Tyson et al., 2015)。

目前國際食媒疾病監測網(PulseNet International)實驗室使用標準化的

PFGE 做為細菌的基因分型工具，這種方法的優點是其結果易於判讀，且具有高度可重複性，且對大部份監測的病原菌且有高的分型效力。但是，PFGE 耗時耗力，而且必須由技術熟練的人員來進行操作，才能確保結果的品質與一致性。另外，對於一些變異度小的菌種如宋內志賀氏菌(*Shigella sonnei*)以及腸炎沙門氏桿菌(*Salmonella enterica* serovar Enteritidis)則無法有效的分型(Boxrud et al., 2007; Liang et al., 2007)。另一種曾經被考慮取代 PFGE 的分子分型方法－多位址變異分析(multilocus variable-number tandem repeat analysis, MLVA)，則是因為具高度物種專一性(organism-specificity)而難以做為通用的細菌基因分型方法(Chiou, 2010; Chiou et al., 2013)。因此開發具有高度分型效力且通用性(universality)的細菌基因分型技術，是公衛微生物實驗室努力的目標。

細菌株全基因體定序產生上百萬條的短序列(raw reads)，如何從裡面擷取出所需要的資訊是許多公衛領域與生物資訊研究人員積極努力的目標。在 WGS 基因分型分析，目前主流的分析方法偏重由 raw reads 中經過與參考基因體序列的比較，萃取(calling)出單核苷酸多型性(single nucleotide polymorphism, SNP)序列(Kohl et al., 2014; Kong et al., 2016; McGann et al., 2016; Stasiewicz, Oliver, Wiedmann, & den Bakker, 2015; Taylor et al., 2015)，而這個方法主要的限制在於參考基因體序列的選擇，但一些含有多種血清型的菌種(例如沙門氏桿菌具有 2,600 多種血清型)，不容易挑選出一個代表參考序列；另外，由於缺乏統一的比較標準，所得到的結果無法跨實驗室來做比對。另一種由多位址序列分型(MLST)(Maiden et al., 1998)方法所擴大衍生出來的方法，即全基因體 MLST (wgMLST) (Maiden et al., 2013)，因為次世代定序技術的成熟與成本下降正漸成為 WGS 基因分型的主流分析方法。

過去的 MLST 是選定數個管家基因(house-keeping gene)，透過預先建立之 MLST allele database 來做為菌株之間比較的標準。菌株可透過此資料庫產生標準化的 MLST allelic profile，因此可跨實驗室相互比較。但是由於管家基因演化速

率慢，較適合作為研究長期物種演化的生物標誌(biomarker)(Urwin & Maiden, 2003)，而群聚感染調查與疾病監測需要能偵測在短時間內的物種變化，因此，選取演化速率較快的 biomarker 應該較為適當。過去從事 MLST 分型方法，需要設計多組引子組(primers)以便能利用 PCR-Sanger sequencing 的方法決定 MLST 的序列型別(sequence type, ST)，花費成本頗高。次世代定序技術成本下降，只要設計不同的 scheme 就可以由全基因體序列中彈性選擇不同 loci 的組合，產出可跨實驗室比對的基因圖譜(genetic profiles)，然而要跨實驗室比對的基因型別，需使用共同的泛基因體對偶基因資料庫(pan-genome allele database，PGAdb)產出基因圖譜，因此建立各別菌種的 PGAdb 是發展 WGS 基因分型技術的基礎。

本研究之整體目標在發展「病原菌株全基因體 DNA 序列分析技術與應用平台」以協助公衛實驗室與醫院實驗室運用次世代定序技術進行菌株基因分型，以進行群聚感染調查(院內群聚感染)與疾病監測，以早期偵測到群聚感染流行。本年度研究，開發優化用於建立 PGAdb 與產生基因圖譜(wgMLST/cgMLST)之程式，並用於建立沙門氏菌之 PGAdb，並進行資料庫內容之分析。此 *Salmonella* PGAdb 將由疾病管制署實驗室使用，進行菌株基因分型，以評估其區別流病相關(epidemiologically-related)與流病不相關(epidemiologically-unrelated)菌株之能力。

二、材料與方法

計畫期程共計 2 年，106 年度優先建置食媒病原菌沙門氏桿菌泛基因體等位基因資料庫(*Salmonella* PGAdb)，先後開發與優化「泛基因體等位基因資料庫建置工具」、「全基因體多位址序列分型圖譜 (wgMLST profile)產生工具」、「親源關係樹建立工具(利用 wgMLST profiles)」。106 年度已完成之開發工具方法詳述如下:

1. 泛基因體等位基因資料庫建置工具：

首先，運用基因預測(註釋)軟體 Prokka (Seemann, 2014)，對細菌全基因體 DNA 序列片段 (contigs)進行註釋以定義出每個基因在 DNA 序列上的區域，接下來利用序列比對程式 BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990)與 CD-HIT(Li & Godzik, 2006)建構同源基因 (othologous gene)群集。需要注意的是旁系同源基因 (paralogous gene)將在建構群集時將被剔除。建構完成後將得到多個同源基因群集，每個群集內包含彼此間序列有若干差異的同源基因(或稱等位基因)，我們可以為每個群集設定一個不重複的名稱，並且為群集內的等位基因編流水號。在依據群集內等位基因在基因體上之分布個數多寡將群集進行排序後定義出核心基因群 (core genes)、泛基因群 (pan genes)、附屬基因群 (accessory genes)、單獨基因群 (unique genes)等不同的基因組合 (scheme)。上述步驟將以腳本語言 (script language)撰寫串接成一自動化流程。

2. 全基因體多位址序列分型波譜產生(wgMLST profiling)工具

當特定病原菌泛基因體等位基因資料庫已建構完成並評估之後，即可做為 wgMLST profiling 工具之序列比對資料庫。透過 BLAST 將病原菌全基因體片斷序列 (contigs)與泛基因體等位基因資料庫比對，由於每一個在資料庫中的等位基因皆已編號(同源基因群集名稱與等位基因流水編號，如：A00001:1,

A00002:1 …)，經過比對即可得到一組由不同之同源基因群集所對應的等位基因流水編號之數字序列 (如：A00001:1, A00002:3 …)，此數字序列我們稱為”wgMLST profile”。上述之步驟將以腳本語言撰寫，由於此工具內含大量序列比對運算，將妥善評估執行效能以期能符合實際應用。

3. 親源關係樹建立工具(利用 wgMLST profile)

建立親源關係樹，需先建立距離矩陣。由於計算 wgMLST profile 之間的差異(distance)，profile 僅包含一串數字，不同於計算 DNA 序列之間的差異有算分矩陣 (scoring matrix)可以計算差異分數，因此我們參考 Lars Snipen (Snipen & Ussery, 2010)在所發表建立泛基因樹的方法來計算 profile 之間的差異。我們計算 profile 之間的曼哈頓距離(Manhattan distance)，換句話說，我們僅計算兩個 profile 彼此相對應 locus (同源基因群集)的等位基因差異數量做為差異分數。我們採用可以直接從成對距離 (pairwise distance)來建構演化樹的演算法－Unweighted Pair Group Method with Arithmetic Mean (UPGMA)。Boostrap 統計方法也應用來計算演化樹分支之可信度。在本研究中，我們將利用 Phylip (Felsenstein, 1981) 程式組來建構 UPGMA 演化樹，Boostrap 統計值計算以及演化樹檔案輸出則會利用 ETE toolkit (Huerta-Cepas, Dopazo, & Gabaldon, 2010)。腳本語言將被運用來撰寫上述之親源關係樹建立工具。

4. 建立沙門氏菌泛基因體對偶基因資料庫(Salmonella PGAdb)

我們由 NCBI Genome 資料庫下載沙門氏菌全基因體資料共 7,449 筆，下載下來的菌株全基因體序列(包含 complet 及 contigs)先使用 Prokka(Seemann, 2014)進行基因註解並輸出 gff 檔提供給自行開發的「泛基因體等位基因資料庫建置工具」作為輸入檔進行「沙門氏菌泛基因體對偶基因資料庫」的建置工作。

三、 結果

1. 泛基因體等位基因資料庫(PGAdb)建置工具

在使用我們所開發的 PGAdb 建置工具前，需要利用 Prokka(Seemann, 2014)將全基因體序列進行基因註解並產生 gff 檔，接著利用 PGAdb 建置工具建置 PGAdb。此建庫工具能大規模地將 7449 株沙門氏菌所含有約三千萬的總基因序列進行 othologous 分群(gene locus)，並給予每個 locus 中的 allele 編號。目前測試沙門氏菌 PGAdb 建置，以 24 core 的 128GB RAM 約 200hr 可完成。運算時間還可經由串聯伺服器的方式大幅減少。

2. 全基因體多位址序列分型波譜產生(wgMLST profiling)工具

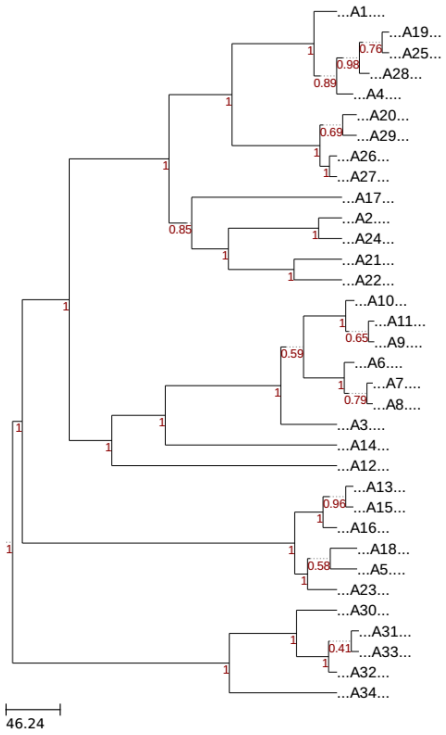
我們所開發的 wgMLST profiling 工具可以利用所建置的 PGAdb 將每個菌株的基因體序列轉換成波譜。下圖為輸出範例，每個 isolate 的基因體序列根據 PGAdb 中所定義的 locus，逐一尋找是否在每個 locus 中有找到相對應的 allele 序列，如果有找到就給予 PGAdb 中已經定義好的 allele number，圖例中「SALxxxxxxx」代表 locus name，表格內的數字代表 isolate 在某個 locus 中所找到相對應(序列相同)allele 的編號，每一列代表一個 isolate 的資料，此範例顯示 5 個 isolate 在前 16 個 locus 的 profiling 資料。

	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

目前產生一個含有 3000 個(core-genome genes)locus 的 genetic profile 使用 24 core CPU 128GB RAM 大約需要 90~120 秒。

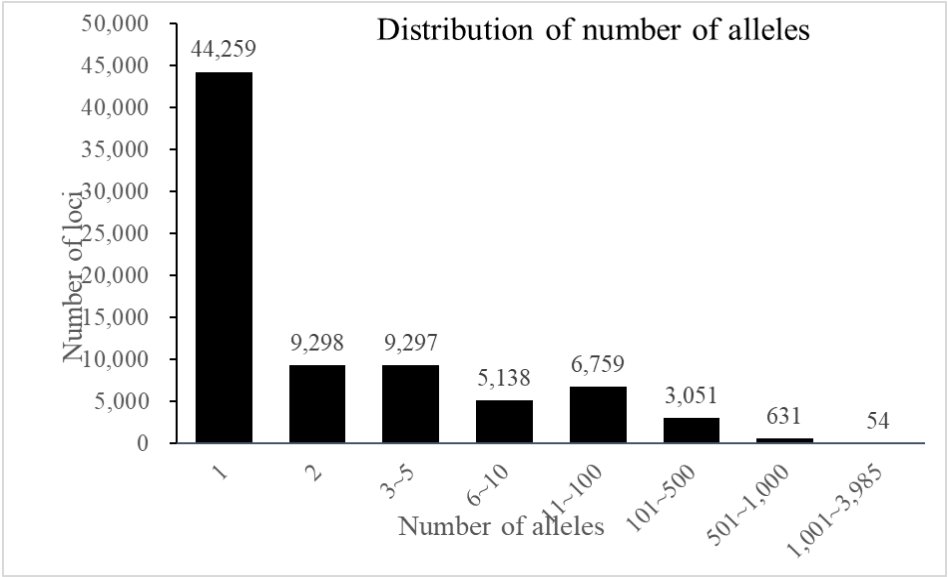
3. 親源關係樹建立工具(利用 wgMLST profile)

我們開發了一套利用上述 wgMLST profiling 工具所產出的 profile 來建立親源關係樹(更精確應稱為「基因關聯相似樹」)。我們採用 UPGMA 的方法來建立基因關聯相似樹，每一個分支(split)的 boosttrap 值也有被計算並標註。右圖範例是利用 34 株 *Salmonella Typhimurium* wgMLST profile 所建置的基因關聯相似樹，所用菌株可參考(Leekitcharoenphon, Nielsen, Kaas, Lund, & Aarestrup, 2014)。



4. Salmonella PGAdb

我們由 NCBI Genome 資料庫下載沙門氏菌 fasta 序列共計 7449 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:



為了更清楚描述 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。以 3 張圖表展現，各圖解釋如下：

(圖一) PGAdb 中 locus 在不同菌株基因體出現率分佈，

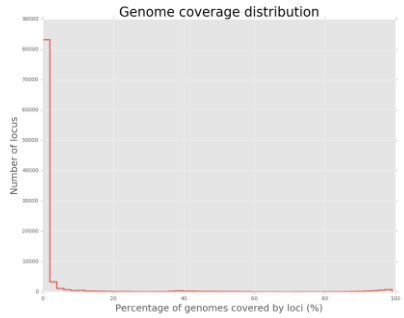
(圖二) PGAdb 中 locus 在不同菌株基因體出現率分佈(去除出現率小於 5%的 locus)，

(圖三) PGAdb 中 locus 在不同菌株基因體出現率累進分佈，

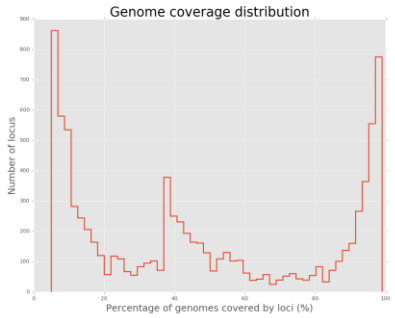
(圖四) PGAdb 中 locus 中 allele 長度分佈(僅示範性呈現前 103 個 locus)。

以下就詳述 *Salmonella* PGAdb 的分析內容:

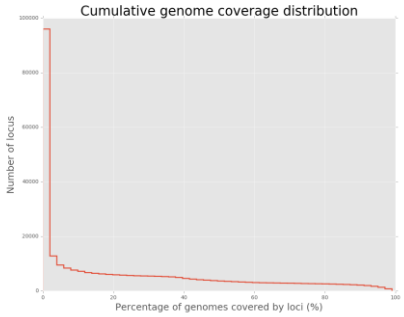
圖一



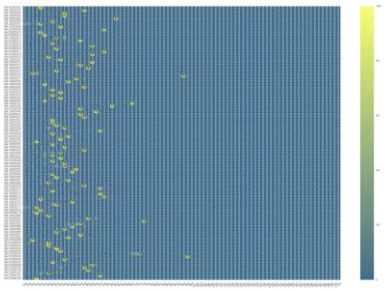
圖二



圖三



圖四



「圖一」顯示在所有的 locus 中，在不同菌株基因體出現率極低(< 5%)的 locus 佔大多數，unique locus (<1%)數量佔了將近 90%，去掉出現率< 5%的

locus，得到了如「圖二」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，雙峰之間(20%~80%, dispensable-gene region)則是平緩曲線，代表用來建庫的菌株基因體之間存在差異，但 *Sal*PGAdb 卻在 40%~50%的區域出現一明顯峰值暗示可能有一些建庫菌株有高度相似性。「圖三」是累進分佈圖，可以用來輔助「圖一」的解讀。「圖一」至「圖三」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。「圖四」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。

四、 討論

泛基因體等位基因資料庫建置工具

我們利用 Roary(Page et al., 2015)為核心程式撰寫泛基因體等位基因資料庫建置工具，優點是可以在可預期的時間內完成大尺度(>1,000 genome)的泛基因體建置，目前 beta 版工具版本以 PHP 程式語言撰寫，現在已著手進行程式效能的優化，目前僅提供 Linux 單機版本，但為了日後的持續維護與提供穩定的雲端服務，已經開始進行 Python 程式語言版本的改寫，並規劃是否能提供雲端服務版本。此建置工具預期將可提供疾管署大規模建置各種病原菌的泛基因體等位基因資料庫。至於資料庫中 scheme 要如何挑選才符合實際 typing 的需求，則需與疾管署相關實驗室共同評估。

全基因體多位址序列分型波譜產生(wgMLST profiling)工具

我們所開發的利用「泛基因體等位基因資料庫」產生「全基因體多位址分型波譜工具」可以大幅減少全基因體序列資料所需的儲存空間(~500MB → 5MB)，並且在經過資料庫產生波譜的過程相當於做了標準化，因此，只要不同實驗室皆使用同樣的泛基因體等位基因資料庫，所產出的 wgMLST profile 皆可跨實驗室互相比對。這樣的特性將為全基因體定序在流病防疫上的應用建構堅實的基礎。但是後續使用上如: wgMLST profile 的交換機制、產出波譜的資料庫如何統一、以及推廣則須進一步的評估。

親源關係樹建立工具

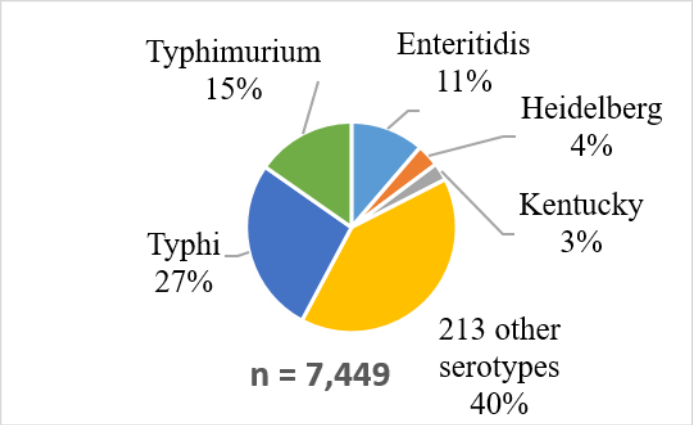
細菌基因體間的差異可以利用比較 wgMLST profile 間的差異來得到，透過 UPGMA 的方法可以將 profile 間的差異建置成親源關係樹(更精確地說，應

該是「基因體相似度關聯性」)，總的來說可以用樹的形式表達。我們利用 ETE3 toolkit(Huerta-Cepas, Serra, & Bork, 2016)為我們工具的核心，這是一套用來建置演化樹的 python 模組，並利用 FastTree 裡面計算 Booststrap value 的程式計算每個 split 的 Booststrap value 並標註於產出的樹上。親源關係樹建立工具預設會提供使用者 tree 的 PDF 檔，另外會提供 newick 檔讓使用者可以用其他的工具例如: FigureTree 去修改所產出的 tree。這個工具將可提供使用者方便易用的視覺化結果呈現。

沙門氏菌 PGAdb

我們目前利用由 NCBI Genome 下載的沙門氏菌株共 7449 株來建置 PGAdb，在 24 core CPU 128GB RAM 的 server 上運行時間為 8 天半。數量 Top 5 的血清型佔了全部 serotype 的 60% (圖五)，雖然我們目前取用的 scheme 是 core genome，因此利用來做基因分型應該不會有很大影響，serotypes 對於 core genes 的影響應該是在 allele 數量，但是如果需要做血清型的預測或許需要對不同 serotype 之間 core genome 差異做進一步的分析。

圖五



五、結論與建議

結論部分

本年度(106)已初步完成 1.泛基因體等位基因資料庫建置工具(PGAdb-builder)、2.全基因體多位址序列分型波譜產生(wgMLST profiling)工具、3.親源關係樹建立工具(利用 wgMLST profile)的開發工作，並利用PGAdb-builder 建立沙門氏菌泛基因體資料庫。由於期中報告委員建議委外計畫部分需與署內計畫專業分工，因此所建置之沙門氏菌 PGAdb 交由疾管署分析與評估所建立資料庫之流行病學適用性。107 年的計畫將持續進行已開發工具的除錯與優化，並開發其餘已規劃的工具程式與資料庫。

建議部分

PGAdb 建置與 wgMLST 方法學基礎是”Gene-by-Gene”的方法(Maiden et al., 2013)，即透過比較不同的細菌株基因體中彼此間基因在序列上的差異，之後進行分群的方法。我們開發的工具主要是利用腳本程式串接多個 open source 的軟體進行「基因預測」、「同源基因分群」、「同源基因群內等位基因的編號」等工作。 在工作流程(workflow)建置中有許多參數需要調整，例如:要挑選哪些同源基因群當成後面分子分型的標準等，尚需與疾管署委託實驗室共同評估。

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，甚至是後續的雲端上線等，勢必於 107 年度要與疾管署進行交接，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(106)主要完成之具體成果如下:

1. 泛基因體等位基因資料庫建置工具
2. 全基因體多位址序列分型波譜產生工具
3. 親源關係樹建立工具
4. 沙門氏菌泛基因體資料庫

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。相關交接事項需要於107年的計劃執行中與疾管署會商，以利後續系統與資料庫的修改與維護。

七、參考文獻：

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Boxrud, D., Pederson-Gulrud, K., Wotton, J., Medus, C., Lyszkowicz, E., Besser, J., & Bartkus, J. M. (2007). Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of Salmonella enterica serotype Enteritidis. *J Clin Microbiol, 45*(2), 536-543. doi:10.1128/JCM.01595-06

Chiou, C. S. (2010). Multilocus variable-number tandem repeat analysis as a molecular tool for subtyping and phylogenetic analysis of bacterial pathogens. *Expert Review of Molecular Diagnostics, 10*(1), 5-7.

Chiou, C. S., Izumiya, H., Thong, K. L., Larsson, J. T., Liang, S. Y., Kim, J., & Koh, X. P. (2013). A simple approach to obtain comparable Shigella sonnei MLVA results across laboratories. *Int J Med Microbiol, 303*(8), 678-684. doi:10.1016/j.ijmm.2013.09.008

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

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol, 17*(6), 368-376.

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., Dopazo, J., & Gabaldon, T. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics, 11*, 24. doi:10.1186/1471-2105-11-24

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

Kohl, T. A., Diel, R., Harmsen, D., Rothganger, J., Walter, K. M., Merker, M., . . . Niemann, S. (2014). Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *Journal of Clinical Microbiology, 52*(7), 2479-2486. doi:10.1128/JCM.00567-14

Kong, Z., Zhao, P., Liu, H., Yu, X., Qin, Y., Su, Z., . . . Chen, J. (2016). Whole-Genome Sequencing for the Investigation of a Hospital Outbreak of MRSA in China. *PLoS One, 11*(3), e0149844. doi:10.1371/journal.pone.0149844

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

Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O., & Aarestrup, F. M. (2014). Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. *PLoS One, 9*(2), e87991. doi:10.1371/journal.pone.0087991

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659. doi:10.1093/bioinformatics/btl158

Liang, S. Y., Watanabe, H., Terajima, J., Li, C. C., Liao, J. C., Tung, S. K., & Chiou, C. S. (2007). Multilocus variable-number tandem-repeat analysis for molecular typing of Shigella sonnei. *J Clin Microbiol, 45*(11), 3574-3580. doi:10.1128/JCM.00675-07

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., . . . Spratt, B. G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America, 95*(6), 3140-3145.

Maiden, M. C., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews: Microbiology, 11*(10), 728-736. doi:10.1038/nrmicro3093

McGann, P., Bunin, J. L., Snesrud, E., Singh, S., Maybank, R., Ong, A. C., . . . Lesho, E. (2016). Real time application of whole genome sequencing for outbreak investigation - What is an achievable turnaround time? *Diagn Microbiol Infect Dis, 85*(3), 277-282. doi:10.1016/j.diagmicrobio.2016.04.020

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

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., . . . Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics, 31*(22), 3691-3693. doi:10.1093/bioinformatics/btv421

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

Salipante, S. J., SenGupta, D. J., Cummings, L. A., Land, T. A., Hoogestraat, D. R., & Cookson, B. T. (2015). Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *J Clin Microbiol, 53*(4), 1072-1079. doi:10.1128/JCM.03385-14

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069. doi:10.1093/bioinformatics/btu153

Snipen, L., & Ussery, D. W. (2010). Standard operating procedure for computing pangenome trees. *Stand Genomic Sci, 2*(1), 135-141. doi:10.4056/sigs.38923

Stasiewicz, M. J., Oliver, H. F., Wiedmann, M., & den Bakker, H. C. (2015). Whole-Genome Sequencing Allows for Improved Identification of Persistent Listeria monocytogenes in Food-Associated Environments. *Appl Environ Microbiol, 81*(17), 6024-6037. doi:10.1128/AEM.01049-15

Taylor, A. J., Lappi, V., Wolfgang, W. J., Lapierre, P., Palumbo, M. J., Medus, C., & Boxrud, D. (2015). Characterization of Foodborne Outbreaks of Salmonella enterica Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection. *J Clin Microbiol, 53*(10), 3334-3340. doi:10.1128/JCM.01280-15

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

Urwin, R., & Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol, 11*(10), 479-487.

衛生福利部疾病管制署委託科技研究計畫
106 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：MOHW106-CDC-C-112-114703

1.計畫之新發現或新發明

(1) 開發細菌泛基因體等位基因資料庫建置工具

(2) 開發全基因體多位址序列分型波譜(wgMLST profile)產生工具

2.計畫對民眾具教育宣導之成果

高分辨力的分子分型工具可以提供更即時性的可能群聚感染預警偵測，可以協助相關單位及早介入進行流行病學調查，並進行防疫作為與衛教宣導。

3.計畫對醫藥衛生政策之具體建議

在台灣，食媒性細菌感染所造成的食物中毒事件時有所聞，利用此計畫所開發的工具提高了菌株基因體差異分辨率，可偵測出舊的檢測方法無法得知的可能群聚感染，因此可以幫助衛生相關單位對於民眾在食品衛生方面的問題提供預警。