致謝辭

隨著論文付梓，在東海大學生命科學系大學四年、研究所兩年的學習過程中到如今研究告一段落，一路上的經歷只有感謝。

本論文能順利完成，首先要感謝我的指導教授─邱偉欣老師的指導與教誨，從最初研究題目的擬定、程式語言的指導、到論文架構的匡正、資料的提供與求學態度等均不厭其煩地教導並予以建議，當我研究過程中受挫時，老師給予關懷與提攜，使我受益良多，在此獻上最深的敬意與謝意。此外，感謝口試委員謝明麗老師和陳荷明老師對於我論文及實驗上給予的諸多寶貴意見與疏漏之處予以指正，使得我在實驗及論文上更趨完善，在此謹致謝忱。

感謝系上教授們悉心安排紮實的課程與諄諄教誨，除了帶領我們邁向學術殿堂也成為我們人生路上的表率，讓我在這兩年內收穫良多。感謝學長姐及雅玄、閔暄、柏凱在研究所兩年的時間修課期間的切磋討論、課外時間給予實驗上的建議，感謝怡靜這六年來同學兼好友給予的支持與鼓勵，同時也感謝助學勞作組衛生助理團隊的學長姐與夥伴們提供機會讓我經濟無虞並在人生道路上不斷成長。

最後，特別感謝我最敬愛的父母，感謝雙親的關懷照顧與殷殷叮嚀及其他家庭成員對我的關懷與支持。

碩士生涯即將畫上句點，但這只是人生的一段，學習的旅程仍將繼續。今後，我將把研究生涯中學習到的態度，及論文寫作上閱讀與論述批判所給予的啟發，實際應用在日常生活及工作中，並期勉自己的未來能成為有批判素養的人。

蔡淑閔　謹誌於

中華民國 101 年七月

目錄

頁次

表目錄

頁次

圖目錄

頁次

中文摘要

多源性基因體學(metagenomics)是近年來快速發展的一門研究領域，其目的是研究環境中微生物。而由於高通量定序技術的發展，更促使多源性基因體學的快速進展，使得以定序為主的多源性基因體學(sequence-based metagenomics)的方法廣泛用於多源性基因體學的研究上。以定序為主的多源性基因體學：直接從環境樣本中萃出的多源性基因體會被定序，經由定序所得到的 DNA 片段(稱為 reads)會更進一步地被組合成 contigs 等較長的片段再更進一步作序列分析，或是藉由以基因為中心的方法直接進行序列分析。blastx 廣泛被應用在以基因為中心的方法來分析多源性基因體。然而，blastx 的敏感度和特異度偏低，而大多數的研究人員無法完全意識到這個缺點，因此，我們提出了數個問題來檢測主要影響 blastx 準確度的參數。為了模擬以序列分析為主的多源性基因體研究，我們取得 *Escherichia coli* BL21-Gold(DE3)pLysS AG 基因體並在電腦上模擬高通量定序法，並進行 blastx。我們測試了序列長度、不同期望值、不同原核生物基因密度對 blastx 表現的影響，目前的結果顯示序列長度、期望值、原核生物基因密度都是影響 blastx 敏感度和特異度的重要因素。

英文摘要

Metagenomics is a discipline that studies environmental microbes. Sequence-based methods are widely adopted in metagenomics. In sequence-based metagenomics, genomes are extracted from environmental samples and DNA fragments are then sequenced. The sequenced DNA fragments (reads) are subjected to be further assembled to generate contigs for further analyzing or they can be directly analyzed by gene-centric methods, in which functional genes are annotated of reads and species diversity postulated. Blastx is a tool widely adopted in gene-centric analyzed. However, its sensitivity and specificity are usually low and most of researchers are not fully aware of the drawbacks. Therefore, we address several questions to check potential parameters that affect the performance of blastx. To mimic sequence-based metagenomic studies, we used *Escherichia coli* BL21-Gold(DE3)pLysS AG genome and simulated high-throughput sequencing in silico followed by blastx analysis. We first tested whether sequence lengths affect blastx performance. Second, we tested whether e-value affect blastx performance. Third, we checked whether the gene density is a critical factor to affect blastx performance. Recent results showed that sequence lengths, e-values and gene densities are important factors to affect the sensitivity and the specificity of blastx.

前言

多源性基因體學 (metagenomics)是近年來快速發展的一門學問，主要是研究無法於實驗室條件下被培養的微生物在環境中的分布、多樣性、功能、交互作用關係、甚至是演化關係；多源性基因體學研究的環境相當廣泛，例如：土壤、水、動物的古老遺骸，或是動物和人類的消化系統(Huang Y *et al.*, 2009、Hugenholtz P *et al.*, 2008) 。近年來由於定序技術發展快速，使得以定序為基礎的多源性基因體學 (sequence-based metagenomics) 也快速發展(Adams IP *et al.*, 2009、Tringe SG *et al.*, 2005) ，其方法為：研究人員從環境中取得了樣本，萃取出其中的基因體 (genome) ，利用高通量定序技術定序基因體；定序出來的片段稱之為 reads，短片段 reads 可以透過被組裝成較長的片段 (像是 contigs、scaffolds) 再進行序列分析 (Markowitz VM *et al.*, 2006、Wooley J C *et al.*, 2010) 。然而，將 reads 進行組裝會有以下的缺點：1. 高通量定序出來的短片段難以進行組裝；2. 在複雜的微生物族群中組裝多源性基因體，嵌合體 (chimeras，組裝短序列的過程中將無同源關係的序列組裝在一起)的產生是無法避免的；3. 組裝短序列難以涵蓋樣本中所有的微生物(Weng FC *et al.*, 2010) 。以基因為中心的分析方法(gene-centric analysis)，則是跳過組裝的過程，將短片段直接進行序列分析(Huson DH *et al.*, 2009) ；相較於其他多源性基

因體學的分析是研究環境中微生物的組成，以基因為中心的分析方法，著重在環境族群中基因的組成，透過在某族群中所佔有的多數基因，進一步去分析物種的組成，或是預測代謝途徑(Hugenholtz P *et al.*, 2008) 。

多源性基因體學研究中常用的序列分析工具—Basic Local Alignment Search Tool(以下簡稱 blast) ，是美國全國生物技術信息中心(National Center for Biotechnology Information，以下簡稱 NCBI)進行序列相似性比對分析的工具；透過輸入一段未知序列，和已知的資料庫序列進行序列比對，得到的排比結果要小於期望值才會是有比對到同源序列的結果(Altschul SF *et al.*,1990)。 blast 中的一個工具—blastx，是利用輸入未知的 DNA 序列透過 blastx 工具的六框轉譯(6-frame translation)後所得到的胺基酸序列和蛋白質的資料庫進行序列排比，blastx 被廣泛應用在以基因為中心的多源性基因體分析方法，進行基因預測和蛋白質功能註解(Dalevi D *et al.*, 2008)。

然而，blastx 的效能卻相當不穩定。以 2009 年 Rosario 等人的研究報告為例，他們研究再生水中的多源性基因體，並將自來水中的多源性基因體 DNA 組成和再生水中的多源性基因體 DNA、RNA 組成比較，以 blastx 和 Genbank 比對的結果發現有絕大多數序列無法在資料庫中比對出同源序列(Rosario K *et al.*, 2009) 。

過去的文獻曾經提到影響 blast 效能可能因素：1. 定序出來的短片段 (Markowitz VM *et al*., 2006、Essinger SD *et al.*, 2010)；2. 資料庫中的同源基因體多寡 (Essinger SD *et al.*, 2010)。目前研究人員對於 blastx 效能的準確程度、確切影響 blastx 的因素等相關研究仍相當不足。因此本研究想要深入了解哪些因素會影響 blastx 效能；我們提出了以下的問題，第一、序列長度是否確實會影響 blastx 的效能，第二、期望值對 blastx 效能的影響，第三、原核生物的基因密度是否會影響 blastx 的效能；透過已知物種來檢測並驗證影響 blastx 效能的因子，使得我們瞭解使用 blast 時應該注意哪些因子及如何達到最佳效能。我們利用敏感度和特異度(Altman DG *et al.*, 1994、Altschul SF *et al.*, 1990)來檢測 blastx 的效能：敏感度就是工具偵測正確的能力，也就是原本正確的結果中，blastx 比對正確所佔的百分比；特異度就是工具預測錯誤的能力，或是真正偵測正確的可能性，也就是 blastx 比對到的結果中，真正比對正確所佔的百分比。計算敏感度和特異度的公式如算式(1)、(2)。

$$\text{敏感度 Sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (1)$$

$$\text{特異度 Specificity} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (2)$$

我們模擬定序出來的片段產生了模擬片段(pseudo-reads)、註解模

擬片段、將模擬片段進行 blastx，並根據結果來計算 blastx 在序列長度、期望值或基因密度等因子改變時 blastx 的敏感度和特異度。從結果中，我們發現輸入 blastx 的序列長度越長，blastx 效能越佳；而 blastx 設定的期望值(以下簡稱 E 值)越大，blastx 效能越佳；原核生物不同的基因密度影響 blastx 效能，基因密度越大，blastx 效能越佳。

材料與方法

壹、硬體

在本研究中所使用的工作站電腦是 LANCOOL，內含八核心的中央處理器(8X Intel(R) Core(TM) i7 CPU950 @3.076Hz)，記憶體為 12328MB，硬碟容量是 472.3GB。

貳、軟體

我們使用的作業系統是 Ubuntu Linux 2.6.32-41-Server(X86-64) ；利用工作站的電腦終端機安裝 python 軟體在終端機中進行分析。python 是物件導向、直譯式的程式語言，而且可以在許多作業系統上使用(http://www.python.org/) ，所以我們選用 python 作為我們撰寫的程式語言；我們使用的 python 的版本是 python2.6.5 及 python3.1.2，同時我們也使用 python 的模組—biopython(http://biopython.org/wiki/Biopython)處理結果，利用 Genay(http://www.geany.org/)這個軟體撰寫 python 程式。

blast 是比對序列間相似性的工具，透過和核苷酸或蛋白質資料庫的序列進行比對、具統計學意義的排比結果，可以用來預測序列間的功能和演化關係，並幫助辨識同源基因(http://blast.ncbi.nlm.nih.gov/) 。NCBI 的 blast 提供了網路版的序列排比工具，可以在網站上直接使用

不同的 blast 工具；此外，也提供了可以從 NCBI 網站 FTP 下載至工作站電腦使用的本地 blast(blast+) ；blast+使用上除了不受網路連線的穩定性與否的限制，同時也可以根據使用者需求建構資料庫。本研究是檢測本地的 blast(blast+) 中的 blastx 效能。本研究中結果的圖表是使用 Microsoft Office Excel 2010 年版本、Microsoft Office PowerPoint 2010 年版本繪製而成。

參、序列收集

物種的選用上，我們首先使用了 *Escherichia coli* BL21-Gold(DE3)pLysS AG(以下簡稱 *E.coli* BL21) 來分析不同序列長度是否影響 blastx 效能、blastx 設定的不同 E 值對 blastx 效能的影響。我們從 NCBI 網站下載了 *E.coli* BL21 基因體序列、基因註解(截圖見圖一)，並藉由 Parse-genome-txt2.py(附錄一) 這個程式來整理下載的基因註解成我們方便分析的檔案格式。

為了探究不同原核生物的基因密度是否會影響 blastx 的效能，我們參考了 2010 年 Lagesen 等人的研究(Lagesen K *et al.*, 2010) ，從 NCBI 的 Genome 網站中(http://www.ncbi.nlm.nih.gov/genome/browse/)，找出所有完整基因體的原核生物，從中選了大於及小於 *E.coli* BL21 基因密度的已知物種(包括 *E.coli* BL21 共五種物種，見表一) 。

肆、產生模擬片段

為了產生模擬片段，我們考慮了以高通量定序技術進行多源性基因體分析時，影響精確程度的兩個主要因子：序列長度和覆蓋率；廣被使用的高通量定序技術(454 焦磷酸定序技術)所定序出來的長度約為 400~500 bp；而定序所考慮的覆蓋率(coverage rate)是指序列上單一鹼基被定序的次數，而在以定序為主的多源性基因體學研究上，常用的覆蓋率最低限度是 6 (Shan GA O *et al.*, 2009； Jiang J *et al.*, 2012.) 。

我們將利用工作站模擬定序出來的片段，稱為模擬片段(pseudo-reads)。在研究「序列長度對 blastx 效能的影響」、「期望值對 blastx 效能的影響」時，我們利用 randomly-chomp-genome.py(附錄二)這個程式，在工作站的電腦隨機切割 *E.coli* BL21 的基因體以產生不同長度的模擬片段：500、1000、1500 及 2000 bp；在研究「原核生物不同的基因密度對 blastx 效能的影響」時，變因是原核生物不同的基因密度，我們產生的模擬片段均為 500 bp。我們產生的模擬片段覆蓋率均固定為 6。

伍、註解模擬片段

我們將模擬片段進行註解，當模擬片段和基因有重疊或涵蓋的情

況，我們定義該模擬片段是陽性(positive)；而當模擬片段和基因沒有任何重疊或涵蓋，這種情況我們定義模擬片段為陰性(negative)。模擬片段和基因之間有四種可能的關係(見圖二)；1. 模擬片段和一個基因(或多個基因)有重疊；2. 模擬片段位在基因內；3. 模擬片段內涵蓋一個或多個基因； 4. 模擬片段位在非基因的區域。符合 1、2 或 3 的情況，模擬片段都將被註解為陽性，而第 4 種情況的模擬片段則為陰性。

考慮了上述四種情況，我們在註解模擬片段時，先考慮其起始位置或結束位置是否被涵蓋在某個基因內，然後再進一步比對在模擬片段範圍內是否有涵蓋其他基因。1st-compare.py(附錄三)這個程式將模擬片段起始位置和結束位置和已知物種(*E.coli* BL21)基因註解中的基因位置比對，reconfirmtxt.py(附錄四)這個程式，則是比對模擬片段範圍內是否有涵蓋基因。最後的結果，則是利用 reconfirmsplit.py(附錄五)這個程式，將註解紀錄在 positive.txt、negative.txt 這兩個檔案。

陸、蛋白質資料庫的建立

我們使用本地端的 blast(blast+)，並參考 NCBI 網站上 BLAST® Help 中的 BLAST Command Line Applications User Manual (http://www.ncbi.nlm.nih.gov/books/NBK1763/)，建構我們使用的本地

blastx 所需的蛋白質資料庫。

從 NCBI 的 Genbank 下載特定物種蛋白質胺基酸序列(fasta 格式)，以如下指令讓 blast 自動建立蛋白質資料庫：「makeblastdb -in foo.faa –title foo –out foo.db」；foo.faa 是下載的胺基酸序列；-title foo 是蛋白質資料庫的標題；-out foo.db 是蛋白質資料庫的檔名，在「序列長度對 blastx 效能的影響」、「期望值對 blastx 效能的影響」研究中，我們建構的蛋白質資料庫是 *E.coli* BL21(ecoli_bl21_gold_de3.db，附錄六)；而在「原核生物不同的基因密度對 blastx 效能的影響」研究中，我們依據不同原核生物建立個別的蛋白質資料庫，包括 *Ehrlichia ruminantium* str. Gardel(Er.db，附錄七)、*Mycoplasma hyopneumoniae* 168(16.db，附錄八)、*Mycoplasma hyorhinis* HUB-1(13.db，附錄九)、*Mycoplasma haemofelis* str. Langford 1(Mh.db，附錄十)蛋白質資料庫及先前研究使用的 *E.coli* BL21 蛋白質資料庫(附錄六)。

柒、對模擬片段進行 blastx

在終端機中輸入「blastx –query foo.fasta -db foo.db -out foo.xml -evalue 1e-40 -outfmt 5」；其中，foo.fasta 是要進行 blastx 的檔案名稱；foo.db 是進行 blastx 比對的蛋白質資料庫；foo.xml 是 blastx 輸出的檔案。資料庫是已知物種胺基酸序列所建構成的蛋白質資料庫；evalue

是設定 blastx 的 E 值，E 值越小，blastx 對結果的篩選越嚴苛，得到的結果數目較少，理論上準確度會較高；outfmt：輸出的檔案格式，我們設定 outfmt 為 5，表示輸出檔案格式是 xml 格式。

在「序列長度是否會影響 blastx 的效能」研究中，我們設定的 blastx 的 E 值是 1E-40。在「期望值對 blastx 效能的影響」的研究中，我們參考了西元 2000 年到 2012 年多源性基因體學研究 (blastx-evalue-reference.pdf，附錄十一)，將文獻使用的 E 值作圖(見圖三)，使用 blastx 作為分析工具時常用設定的 E 值作為我們設定 E 值的依據，分別為：1E-2、1E-3、1E-5、1E-10。另外在「原核生物的基因密度是否會影響 blastx 的效能」的研究中，我們設定 blastx E 值是 1E-2。將模擬片段進行 blastx 的結果有兩種情況：模擬片段和資料庫有比對到同源序列者為 hits，模擬片段和資料庫無法比對到同源序列者為 no hits。我們利用 list-blast+.py(附錄十二)這個程式，將 blastx 結果從 xml 檔案格式(foo.xml)儲存成 txt 檔案格式(foo-blastx_hits.txt、foo-blastx_nohits.txt)。

捌、計算 blastx 效能

我們定義 blastx 的效能是根據敏感度和特異度(見算式(1)、(2))。我們首先定義「真陽性」、「真陰性」、「假陽性」、「假陰性」，當模擬

片段被註解為陽性，並且 blastx 比對得到相同的基因，我們定義為「真陽性」(True Positive，TP)；當模擬片段被註解為陰性，而 blastx 的結果為 no hits，這種情況極為「真陰性」(True Negative，TN)；當模擬片段被註解為陽性，blastx 的結果為 no hits，我們定義此種結果為「假陰性」(False negative，FN)，反之，當模擬片段被註解為陽性，blastx 卻比對到不同的基因，我們定義這種結果為「第一型假陽性」(False positive，FP1)；而當模擬片段被註解為陰性，blastx 結果卻有得到 hits，我們將這種情況定義為「第二型假陽性」(False positive，FP2)。FP1 和 FP2 的加總為全部假陽性的結果。

首先我們利用 p-negativeblast.py(附錄十三)這個程式，將模擬片段註解的結果和 blastx 的結果進行比對並得到 TN、FN、FP2 的結果。比對出 TP 的方法除了比對被註解為陽性的模擬片段是否位於 hits 的結果中，也比對模擬片段註解結果的基因和 hits 結果的基因是否相同，我們選用了基因註解(見圖四)中的基因敘述(Description)和 blastx hits 結果的基因進行比對，當兩者相符會被判定成 TP，兩者不相符時便會被判定為 FP1。

由於蛋白質資料庫中的基因敘述與基因體資料庫中的基因敘述不一致，因此我們必須將兩者統一。以 dnaA 這個基因為例，在基因註解中的基因敘述是 chromosomal replication，但是在 *E.coli BL21*

胺基酸序列中的基因敘述卻是 chromosomal replication initiator protein DnaA，雖然同一個基因會有不同的基因敘述，但是基因的起始和終止位置則是一致的。因此，我們先利用 faa-record.py(附錄十四)將基因起始位置、基因結束位置、基因敘述整理後儲存成 csv 格式的檔案(faa-record.csv 附錄十五)，截圖如圖五所示)，再利用 recordC.py(見附錄十六)這個程式將 faa-record.csv 的結果和基因註解進行比對，當基因註解和 faa-record.csv(附錄十五)的基因位置兩者相同時，將 faa-record.csv 的基因敘述和基因註解的基因名稱儲存成 csv 檔案格式的結果(recordC.csv(附錄十七)，截圖如圖六所示)。此外，模擬片段註解的基因敘述為「hypothetical protein」時，有比對到相同的基因時，在 blastx 的結果會以「hypothetical protein ECBD_XXXX」來表示，便可能名稱不同而被判定成是 FP1，因此我們在比對的 python 程式(0718p.py，附錄十八)中，加了一個條件─當模擬片段的基因敘述為「hypothetical protein」，比對上會將基因敘述的名稱加上 Gene name(ECBD_XXXX)。我們利用 0718p.py 這個程式將結果比對並計算 blastx 敏感度和特異度的結果。

結果

壹、序列長度對 blastx 效能的影響

我們提出的第一個問題是「序列長度是否影響 blastx 的效能」。根據我們的問題，我們需要已知物種基因體序列、基因註解、產生不同長度的模擬片段、註解不同長度的模擬片段，最後比對模擬片段 blastx 的結果並計算敏感度和特異度。

我們從 NCBI 的 Genbank 中下載 *E.coli* BL21 基因體序列和基因註解，利用 Parse-genome-txt2.py(附錄一)，根據 *E.coli* BL21 基因註解中的「基因名稱(Gene name)」、「敘述(Description)」、「基因編碼(Gene ID)」、「基因起始的位置(Start position)」、「基因結束的位置(End position)」，寫成利於我們之後分析的表格(csv)(附錄十九、結果截圖見圖四)。為了產生不同序列長度的模擬片段，利用 randomly-chomp-genome.py(附錄二)，我們將 *E.coli* BL21 基因體序列隨機產生不同長度(500、1000、1500、2000 bp)、相同覆蓋率(6)的模擬片段。我們利用 1st-compare.py(附錄三)、reconfirmtxt.py(附錄四)、reconfirmsplit.py(附錄五)將模擬片段和 *E.coli* BL21 基因註解比對並記錄模擬片段的註解結果，從註解結果(見表二)中我們發現不同長度的模擬片段大多被註解為陽性，且序列長度越長模擬片段越少被註解為陰性；因此理論上將模擬片段進行 blastx 有較高的機會和蛋白質資

料庫比對到同源序列。我們建構了 *E.coli* BL21 蛋白質資料庫(ecoli_bl21_gold_de3.db，見附錄六)並將模擬片段進行 blastx，設定的 E 值為 1E-40，我們將 blastx 的結果作成表(見表三)；從 blastx 結果中 hits 和 no hits 的數量及百分比顯示，絕大多數的模擬片段和資料庫有比對到同源序列。為了計算 blastx 的敏感度和特異度，我們將模擬片段註解的結果和 blastx 結果、recordC.csv(附錄十七)進行比對，比對出 TP、TN、FP、FN 並計算出敏感度和特異度(結果見表四)，我們將結果(表四)作圖並畫上趨勢線(見圖七、八)，從敏感度結果(圖七)、特異度結果(圖八)中我們發現序列長度的確會影響 blastx 敏感度和特異度，而且序列長度越長，敏感度和特異度的值越高。

在敏感度的結果中，500bp 模擬片段 blastx 敏感度為(84.78%)，相較於 1000bp 模擬片段 blastx 敏感度(96.62%)，敏感度略低；500bp 敏感度偏低，是否和序列涵蓋的基因資訊多寡有關，我們查到 *E.coli BL21* 基因密度是 0.949(gene/kbp)(Lukjancenko O *et al., 2010*)，換算不同長度的模擬片段平均所涵蓋的基因數(見表五)，500bp 模擬片段所涵蓋的基因數為 0.4745，1000bp 模擬片段所涵蓋的基因數為 0.949，由於 1000bp 模擬片段所涵蓋的基因數較 500bp 模擬片段多，所涵蓋的基因資訊較多，故 blastx 敏感度較高。

貳、期望值對 blastx 效能的影響

我們搜尋了西元 2000 到 2012 年的文獻(blastx-evalue-reference.pdf，附錄十一)，多源性基因體學的研究中，以 blastx 進行同源序列搜尋的時候多數研究團隊所使用的 E 值，發現 E 值的使用上沒有定值與規則，以 1E-2、1E-3、1E-5、1E-10 為最常用的設定值(圖三)，因此我們提出了第二個問題是：期望值是否影響 blastx 效能？

我們想要了解 blastx 設定的 E 值對 blastx 效能的影響，我們將不同序列長度(500、1000、1500、2000bp)的模擬片段，在不同的 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)下進行 blastx，並將模擬片段註解的結果和 blastx 結果、recordC.csv(附錄十七)進行比對並計算 blastx 敏感度和特異度(見表六)。我們將不同長度的模擬片段在不同 E 值下 blastx 的敏感度和特異度結果作圖並畫上趨勢線(圖九、十)。不同長度的模擬片段當 blastx 設定的 E 值越大，敏感度和特異度均有增加的趨勢，使用 500bp 的模擬片段進行 blastx 時，E 值的改變對 blastx 敏感度的影響最為明顯，當 E 值設定為 1E-40 時，blastx 敏感度為 84.78%；而 E 值設定從 1E-40 到 1E-10 時，blastx 敏感度提升至 96.80%(表六、圖九)。為了解釋敏感度增加的原因，我們將計算敏感度的 TP 和 FN 在不同 E 值下的數量作圖並畫上趨勢線(見圖十一、十二)，當 blastx 設

定的 E 值變大，blastx 條件限制放寬後，模擬片段被判定成 FN 的結果減少，尤其使用 500bp 的模擬片段進行 blastx 時，E 值的改變使得被判定為 FN 的 500bp 模擬片段數目減少最為明顯；當 blastx 設定的 E 值改變時，不同長度的模擬片段進行 blastx 時，模擬片段被判定成 TP 的數目也隨著 E 值變大而增加。從以上結果，敏感度隨著 E 值變大的增加是因為模擬片段被判定為 FN 的數目減少與被判定為 TP 的模擬片段數目增加。

不同序列長度的模擬片段在不同 E 值下的特異度結果(圖十)卻有所不同，以 500bp 的模擬片段進行 blastx 時，隨著 E 值越大 blastx 特異度也越高。我們將不同長度模擬片段在不同 E 值下被判定為 FP 的結果作圖並畫上趨勢線(圖十三)。由於 E 值越大 blastx 比對的準確度會相對下降，因此使得 500bp 模擬片段被判定為 FP 數目隨著 E 值增加而增加；隨著 E 值改變 blastx 特異度在 E 值設定為 1E-40 增加至 1E-10 時，以 1000、1500、2000bp 的模擬片段進行 blastx 時 blastx 特異度有略為下降，從 1000、1500、2000bp 的模擬片段在不同 E 值下進行 blastx 被判定為 FP 的數目(圖十三)，當 E 值設定為 1E-40 增加至 1E-10 時，FP 的數目也略為減少，我們根據 python 程式(fptp.py，附錄二十)計算出當 E 值設定的改變，使得模擬片段從 FP1 轉而被判定為 TP 的結果：500bp 的模擬片段中有 215 條模擬片段、1000bp 的

模擬片段中有 142 條模擬片段、1500bp 的模擬片段中有 79 條模擬片段、在 2000bp 的模擬片段中有 79 條模擬片段。

從特異度的結果(表六、圖十)中，當 E 值設定從 1E-10 到 1E-2，不同長度模擬片段被判定為 FP2 的結果增加了,隨著 E 值設定越大，由於 blastx 準確度越低，因此在 E 值較小時可以被判定成是 TN 的模擬片段，在 E 值變大時，會被判定成 FP2 的結果。從本研究的結果中也符合西元 2000 到 2012 年，多源性基因體學的研究中(圖三)，以 blastx 進行同源序列搜尋的時候多數研究團隊所使用的 E 值，相較於我們結果中敏感度和特異度最佳的 E 值是 1E-2；在過去研究中最常用的 E 值是 1E-3、1E-5(見圖三)，從不同 E 值設定時 blastx 的敏感度和特異度結果中(表六)，設定 E 值為 1E-3、1E-5 時，模擬片段被判定為 FP2 的數目較少，而在設定 E 值為 1E-2 時，模擬片段被判定為 FP2 的數目較多(見表五)；因此將多源性基因體以 blastx 進行分析時，相較於設定 E 值為 1E-2，使用 E 值為 1E-3 或 1E-5 便能避免序列被判定為 FP2。

參、原核生物不同的基因密度對 blastx 效能的影響

承接以上的結果：「序列長度對 blastx 效能的影響」結果中，模擬片段從 500 到 1000bp，由於 1000bp 所涵蓋的基因數(0.949)較 500bp

模擬片段涵蓋的基因數(0.4745)多，因此使得敏感度提高近十倍，因此我們推論 blastx 效能可能和基因密度有關。再者，環境的多源性基因體是由許多生物組成；不同原核生物具有不同的基因密度。因此我們提出的第三個問題是；原核生物不同的基因密度對 blastx 效能是否有影響。

首先我們選用大於和小於 *E.coli* BL21 基因密度(949 genes/Mbp)(Lukjancenko O *et al.*, 2010)的物種 *Mycoplasma haemofelis* str. Langford 1(1374 genes/Mbp，以下簡稱 *M.h.*-L1)、*Ehrlichia ruminantium* str. Gardel(659 genes/Mbp，以下簡稱 *E.r.*-G)、*Mycoplasma hyopneumoniae* 168(795 genes/Mbp，以下簡稱 *M. h.*-168)、*Mycoplasma hyorhinis* HUB-1(846 genes/Mbp，以下簡稱 *M.h.*-HUB-1)，選用物種的依據除了根據基因密度之外，除了 *E.coli* BL21(4.57Mbp)之外，物種基因體大小均差異不大：*M.h.*-L1(1.15Mbp)、*E.r.*-G(1.5Mbp)、*M. h.*-168(0.93bp)、*M.h.*-HUB-1(0.83Mbp)。

由於變數是基因密度，因此模擬片段的長度和覆蓋率均為定值；因此我們產生的模擬片段覆蓋率為 6、500bp 的模擬片段，分別根據 *M.h.*-L1、*E.r.*-G、 *M. h.*-168、*M.h.*-HUB-1 的基因註解，註解模擬片段並分別將模擬片段進行 blastx，blastx 設定的 E 值是根據「期望值對 blastx 效能的影響」結果中 blastx 敏感度和特異度最佳的 E 值(1E-2)。

我們將不同基因密度物種，blastx 的敏感度和特異度結果(見表七)作圖並畫上趨勢線(如圖十四、圖十五)，基因密度越大，blastx 敏感度和特異度有上升的趨勢，而且基因密度對 blastx 效能的影響並不會因為 *E.coli* BL21 基因體(4.57Mbp)相較於其他原核生物的基因體較大而影響 blastx 效能。

討論

過去研究(Essinger SD, Rosen GL, 2010)僅驗證 blast 準確度，然而 blastx 在多源性基因體學的研究上應用相當廣泛，因此本研究探討影響 blastx 的因素，透過利用已知物種來探討不同因素(序列長度、E 值、基因密度)下 blastx 的效能；以下針對本研究探討不同因素對 blastx 效能的影響作討論。

壹、序列長度對 blastx 效能的影響

過去文獻(Markowitz V. M. *et al*., 2006、Essinger S. D. *et al*., 2010)均提到序列長度可能會影響 blastx 效能，本研究針對過去文獻再次驗證序列長度對 blastx 效能的影響，從結果中我們得知當模擬片段的序列長度越長 blastx 敏感度和特異度越高，本研究驗證了序列長度對於 blastx 效能的影響。當定序技術發展日趨先進，從次世代定序的 400~500 bp，到第三代定序技術可以定序出 1000bp 或是更長的片段 (Schadt E. E. *et al*., 2010) ，定序出來的片段所涵蓋的基因資訊越多，blastx 應用在多源性基因體學的研究時，序列長度影響 blastx 敏感度和特異度的問題也可以隨之解決。本研究產生模擬序列是以次世代定序法定序出來的片段為 400~500bp，因此我們由 *E.coli* BL21 基因體序列產生的模擬片段最短是 500bp；由於定序技術至今不斷進步中，

因此我們也產生 1000、1500、2000bp 長的模擬片段，因應未來第三代定序技術的發展或是定序技術發展更為成熟時，我們的研究結果仍具有可信度。

貳、期望值對 blastx 效能的影響

過去沒有任何研究指出 E 值是否影響 blastx 效能，因此本研究證實不同 E 值確實影響 blastx 敏感度和特異度。相較於我們過去將單一物種進行 blastx 時，設定的 E 值越小才得以使得結果較具統計意義 (Altschul SF *et al.*,1990)，而將多源性基因體進行 blastx 時，從本研究結果顯示使用 blastx 作多源性基因體學的分析時，可依據不同長度來設定不同 E 值以求最佳結果，例如當序列較短(400~500bp)的時候，我們可以選擇設定較大的 E 值 (如 1E-2)，而當序列大於 1000bp 時，可以考慮使用較小的 E 值，來增加 blastx 偵測的效能。此外，本研究在這部分只探討在 blastx 資料庫搜尋的條件嚴謹與否，但若能將影響 blastx 的參數(例如 blastx 得分計算的 Matrix、Gap Costs 等)分別進行分析，便可篩選出在不同條件下，最佳的 blastx 參數，進而在多源性基因體學研究上，blastx 效能達到最佳的狀態。

參、原核生物不同的基因密度對 blastx 效能的影響

過去對於多源性基因體中原核生物的基因密度和 blastx 效能的影

響與否未知，因此本研究將已知不同原核生物基因密度在相同序列長度(500bp)、設定相同的 E 值(1E-2)進行 blastx，所得到的結果顯示基因密度越大，blastx 敏感度和特異度越佳。過去研究提出多源性基因體序列和資料庫是否有同源序列 (Essinger S. D. *et al.*, 2010)，在本研究中發現了基因密度也會影響 blastx 的效能，因此我們可以推論在多源性基因體學的研究中，不同原核生物的基因密度大小差異也會影響 blastx 在偵測不同物種的效能。

肆、後續研究建議

本研究是利用已知物種基因體序列所產生的模擬片段分別和已知物種建構的蛋白質資料庫進行 blastx，相較於將多源性基因體進行 blastx 是利用”non-redundant” database 比對，本研究所得到的 blastx 效能較佳，在多源性基因體的研究中，blastx 的效能可能不如本研究的結果，但是研究人員使用 blastx 來分析多源性基因體時可透過本研究的結果，進而調整影響 blastx 效能的因素，使 blastx 效能使用上有所提升。

參考文獻

Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N. (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Molecular plant pathology, 10, 537–45

Altman DG, Bland JM. (1994) Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ British Medical Journal, 308, 6943, p. 1552.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. Journal of molecular biology, 215, 403-10.
Bassi S. (2007) A primer on python for life science researchers. PLoS computational biology, 3, e199

Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N, Sederoff R, Carlson JE. (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. BMC Plant Biol, 9, 51.

Berg Miller ME, Yeoman CJ, Chia N, Tringe SG, Angly FE, Edwards RA, Flint HJ, Lamed R, Bayer EA, White BA. (2012) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. Environ Microbiol, 14, 207-27.

Bibby K, Viau E, Peccia J. (2011) Viral metagenome analysis to guide human pathogen monitoring in environmental samples. Lett Appl Microbiol, 52, 386-92.

Biddle JF, White JR, Teske AP, House CH. (2011) Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. ISME J., 5, 1038-47.

Boyer M, Gimenez G, Suzan-Monti M, Raoult D. (2012) Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. Intervirology, 53, 310-20.

Carpi G, Cagnacci F, Wittekindt NE, Zhao F, Qi J, Tomsho LP, Drautz DI, Rizzoli A, Schuster SC. (2011) Metagenomic profile of the bacterial communities associated with Ixodes ricinus ticks. PLoS One, 6, e25604

Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, Hugenholtz P, Kyrpides NC, Markowitz VM. (2008) Annotation of metagenome short reads using proxygenes. Bioinformatics (Oxford, England), 24, 16, pp. i7-13.

Debroas D, Humbert JF, Enault F, Bronner G, Faubladier M, Cornillot E. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget--France). Environ Microbiol, 11, 2412-24.

Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG. (2011) De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum. BMC Genomics, 12, 99.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. (2008) Functional metagenomic profiling of nine biomes. Nature, 452, 629-32.

Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. (2009) Metagenomic analysis of RNA viruses in a fresh water lake. PLoS One, 4, e7264.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature, 452, 340-3

Edwards JL, Smith DL,Connolly J, McDonald JE,Cox MJ, Joint I, Edwards C, McCarthy AJ. (2010) Identification of Carbohydrate Metabolism Genes in the Metagenome of a Marine Biofilm Community Shown to Be Dominated by Gammaproteobacteria and Bacteroidetes. Genes, 1, 371-384

Essinger SD, Rosen GL (2010) The Effect of Sequence Error and Partial Training Data on BLAST Accuracy. 2010 IEEE International Conference on BioInformatics and BioEngineering, 257-62.

Frank JA, Sørensen SJ. (2011) Quantitative metagenomic analyses based on average genome size normalization. Appl Environ Microbiol, 77, 2513-21.

Folino G, Gori F, Jetten MSM, Marchiori E. (2009) Evidence-Based Clustering of Reads and Taxonomic Analysis of Metagenomic Data. Lecture Notes in Computer Science, 5780, 102-112.

Ge X, Li Y, Yang X, Zhang H, Zhou P, Zhang Y, Shi Z. (2012) Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. J Virol, 86, 4620-30

Gerlach W, Stoye J. (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. , 39, e91.

Haiying Liang, Saravanaraj Ayyampalayam, Norman Wickett, Abdelali Barakat, Yi Xu, Lena Landherr, Paula E. Ralph, Yuannian Jiao, Tao Xu, Scott E. Schlarbaum, Hong M, James H, Leebens-M, Claude WD (2011) Generation of a large-scale genomic resource for functional and comparative genomics in Liriodendron tulipifera L. TREE GENETICS & GENOMES, 7, 941-54

Hewson I, Paerl RW, Tripp H J, Zehr JP, Karl DM. (2009) Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnol. Oceanogr, 1981–1994.

Huang Y, Gilna P, Li W. (2009) Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics (Oxford, England), 25, 1338-40.
Hugenholtz P, Tyson GW. (2008) Microbiology: Metagenomics. Nature, Nature. 455, 7212, P. 481-3.

Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC. (2009) Methods for comparative metagenomics. BMC bioinformatics, 10, S12.

Ibarra-Laclette E, Albert VA, Pérez-Torres CA, Zamudio-Hernández F, Ortega-Estrada Mde J, Herrera-Estrella A, Herrera-Estrella L. (2011) Transcriptomics and molecular evolutionary rate analysis of the bladderwort (Utricularia), a carnivorous plant with a minimal genome. BMC Plant Biol, 11, 101.

Jaenicke S, Ander C, Bekel T, Bisdorf R, Dröge M, Gartemann KH, Jünemann S, Kaiser O, Krause L, Tille F, Zakrzewski M, Pühler A, Schlüter A, Goesmann A. (2011) Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. PLoS One, 6, e14519.

Jeffries TC, Seymour JR, Gilbert JA, Dinsdale EA, Newton K, Leterme SS, Roudnew B, Smith RJ, Seuront L, Mitchell JG. (2011) Substrate type determines metagenomic profiles from diverse chemical habitats. PLoS One, 6, e25173.

Jiang J, Li J, Kwan HS, Au CH, Wan Law PT, Li L, Kam KM, Lun Ling JM, Leung FC. (2012) A cost-effective and universal strategy for complete prokaryotic genomic sequencing proposed by computer simulation. BMC Research Notes, 5, 80-90.

Kanokratana P, Uengwetwanit T, Rattanachomsri U, Bunterngsook B, Nimchua T, Tangphatsornruang S, Plengvidhya V, Champreda V, Eurwilaichitr L. (2011) Insights into the phylogeny and metabolic potential of a primary tropical peat swamp forest microbial community by metagenomic analysis. Microb Ecol., 61, 518-28.

Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suciu D, Levine SR, Markowitz VM, Rigoutsos I, Tringe SG, Bruce DC, Richardson PM, Lidstrom ME, Chistoserdova L. (2009) High-resolution metagenomics targets major functional types in complex microbial communities. Data Management.

Kelly LW, Barott KL, Dinsdale E, Friedlander AM, Nosrat B, Obura D, Sala E, Sandin SA, Smith JE, Vermeij MJ, Williams GJ, Willner D, Rohwer F. (2011) Black reefs: iron-induced phase shifts on coral reefs. ISME J. , 6, 638-49.

Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. (2010) New dimensions of the virus world discovered through metagenomics. Trends Microbiol., 18, 11-9.

Kvist T, Ahring BK, Lasken RS, Westermann P. (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. Appl Microbiol Biotechnol, 74, 926-35.

Ladner JT , Barshis DJ, Palumbi SR. (2012) Evolutionary comparison of transcriptome sequences from four clades of coral endosymbionts in the genus Symbiodinium: a search for genes involved in the thermotolerance of clade D. stacks.stanford.edu, 145-171

Lagesen K, Ussery DW, Wassenaar TM. (2010) Genome update: the 1000th genome--a cautionary tale. Microbiology, 156, 603-608.

Lalkhen AG, McCluskey A.(2008) Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain, 8, 6, pp. 221-223.

Li N, Zhang L, Li F, Wang Y, Zhu Y, Kang H, Wang S, Qin S. (2011) Metagenome of microorganisms associated with the toxic Cyanobacteria Microcystis aeruginosa analyzed using the 454 sequencing platform. CHINESE JOURNAL OF OCEANOLOGY AND LIMNOLOGY, 29, 505-513.

Li L, Victoria JG, Wang C, Jones M, Fellers GM, Kunz TH, Delwart E. (2010) Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J Virol, 84, 6955-65.

Li LL, McCorkle SR, Monchy S, Taghavi S, van der Lelie D. (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnology for biofuels, 2, 10.

Littman R, Willis BL, Bourne DG. (2011) Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. Microbial Ecology, 3, 651-60.

Liu B, Pop M. (2010) Identifying Differentially Abundant Metabolic Pathways in Metagenomic Datasets. Bioinformatics Research and Applications Lecture Notes in Computer Science, 6053, 101-12

Liu B, Pop M. (2011) MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc, 5, S9.

Lu J, Domingo JS. (2008) Turkey fecal microbial community structure and functional gene diversity revealed by 16S rRNA gene and metagenomic sequences. J Microbiol., 46, 469-77.

Lu J, Santo Domingo J, Shanks OC. (2007) Identification of chicken-specific fecal microbial sequences using a metagenomic

approach. Water Res., 41, 3561-74.

Lukjancenko O, Wassenaar TM, Ussery DW. (2010) Comparison of 61 sequenced Escherichia coli genomes. Microbial ecology, 60, 708-20.

McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F, Paul JH. (2008) Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. PLoS One, 3, e3263.

Macdonald NJ, Parks DH, Beiko RG. (2012) Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res, 40, e111.

Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature, 480, 368-71.

Marhaver KL, Edwards RA, Rohwer F. (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol, 10, 2277-86.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P, Kyrpides NC. (2006) An experimental metagenome data management and analysis system. Bioinformatics

(Oxford, England), 22, e359-67.

Martiny AC, Huang Y, Li W (2011) Adaptation to Nutrient Availability in Marine Microorganisms by Gene Gain and Loss, in Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats (ed F. J. de Bruijn), John Wiley & Sons, Inc., Hoboken NJ, USA.

Martiny AC, Kathuria S, Berube PM. (2009) Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. Proc Natl Acad Sci U S A., 106, 10787-92.

McHardy AC, Rigoutsos I. (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. Current opinion in microbiology, 10, 499-503

Menor M, Baek K, Belcaid M, Gingras Y, Poisson G. (2010) Virus DNA-fragment classification using taxonomic hidden Markov model profiles. Proceedings of the 2010 ACM Symposium on Applied Computing, 1567-171

Montaña JS, Jiménez DJ, Hernández M, Angel T, Baena S. (2011) Taxonomic and functional assignment of cloned sequences from high Andean forest soil metagenome. Antonie Van Leeuwenhoek, 101, 205-15.

Muhammad AG, William DG, Shaun H. (2010) Metagenomic Accessing

of Genes From Environmental Clone Library From an Oxidation Tank of a Waste Treatment Plant Using 59-be Conserved Sequence Specific Primers. University of Leicester,

Nakamura S, Maeda N, Miron IM, Yoh M, Izutsu K, Kataoka C, Honda T, Yasunaga T, Nakaya T, Kawai J, Hayashizaki Y, Horii T, Iida T. (2008) Metagenomic Diagnosis of Bacterial Infections. Emerging Infectious Diseases, 14, 1784-86.

Nguyen VH. (2009) Traitement parallèle des comparaisons intensives de séquences génomiques. Université Rennes 1, Dominique Lavenier (Dir.)

Park EJ, Kim KH, Abell GC, Kim MS, Roh SW, Bae JW. (2011) Metagenomic analysis of the viral communities in fermented foods. Appl Environ Microbiol, 77, 1284-91.

Park EJ, Kim KH, Abell GC, Kim MS, Roh SW, Bae JW. (2011) Metagenomic analysis of the viral communities in fermented foods. Appl Environ Microbiol, 77, 1284-91.

Pasić L, Rodriguez-Mueller B, Martin-Cuadrado AB, Mira A, Rohwer F, Rodriguez-Valera F. (2009) Metagenomic islands of hyperhalophiles: the case of Salinibacter ruber. BMC Genomics, 10, 570.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT. (2005) Analysis of microbial gene

transcripts in environmental samples. Appl Environ Microbiol, 71, 4121-6.

Ray J, Dondrup M, Modha S, Steen IH, Sandaa R-A, Clokie M. (2012) Finding a Needle in the Virus Metagenome Haystack - Micro-MetagenomeAnalysis Captures a Snapshot of the Diversity of a Bacteriophage Armoire. PLoS ONE, 7, e34238

Rho M, Tang H, Ye Y.Rho M, Tang H, Ye Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res., 38, e191.

Rosario K, Duffy S, Breitbart M. (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol, 90, 2418-24.

Rosario K., Nilsson C., Lim YW, Ruan Y., and Breitbart M. (2009) Metagenomic analysis of viruses in reclaimed water. Environmental microbiology, 11, 2806-20.

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One, 7, e33641.

Schadt EE, Turner S., Kasarskis A. (2010) A window into

third-generation sequencing. Human molecular genetics, 19, R2, pp. R227-40.

Schoenfeld T, Liles M, Wommack KE, Polson SW, Godiska R, Mead D. (2010) Functional viral metagenomics and the next generation of molecular tools. Trends Microbiol, 20-9.

Schlüter A, Krause L, Szczepanowski R, Goesmann A, Pühler A. (2008) Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. J Biotechnol, 136, 65-76

Schwartz TS, Tae H, Yang Y, Mockaitis K, Van Hemert JL, Proulx SR, Choi JH, Bronikowski AM. (2010) A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. BMC Genomics, 11, 694.

Shan GAO, Zheng WZ. (2009) An ℓ-mer component distribution for genome size esimation, pp. 1-7.

Shan T, Li L, Simmonds P, Wang C, Moeser A, Delwart E. (2011) The fecal virome of pigs on a high-density farm. J Virol, 85, 11697-708.

Shrestha PM, Kube M, Reinhardt R, Liesack W. (2009) Transcriptional activity of paddy soil bacterial communities. Environ Microbiol, 11, 960-70.

Singh KM, Ahir VB, Tripathi AK, Ramani UV, Sajnani M, Koringa PG, Jakhesara S, Pandya PR, Rank DN, Murty DS, Kothari RK, Joshi CG. (2011) Metagenomic analysis of Surti buffalo (Bubalus bubalis) rumen: a preliminary study. Mol Biol Rep., 39, 4841-8.

Smith RJ, Jeffries TC, Roudnew B, Fitch AJ, Seymour JR, Delpin MW, Newton K, Brown MH, Mitchell JG. (2012) Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems.Environ Microbiol, 14, 240-53.

Sorokin VA, Gelfand MS, Artamonova II. (2010) Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. Appl Environ Microbiol., 76, 2136-44.

Steward GF, Preston CM. (2011) Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. Virol J, 8, 287.

Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol, 4, 168-83.

Tamaki H, Zhang R, Angly FE, Nakamura S, Hong PY, Yasunaga T, Kamagata Y, Liu WT. (2012) Metagenomic analysis of DNA viruses in a

wastewater treatment plant in tropical climate. Environ Microbiol, 14, 441-52.

Thomas CJ, Hons BSc. (2011) MICROBIAL COMMUNITY COMPOSITION OF A NATURAL SEDIMENT SALINITY GRADIENT: TAXONOMIC AND METABOLIC PATTERNS AND CONTROLLING FACTORS. A THESIS SUBMITTED FOR THE DEGREE DOCTOR OF PHILOSOPHYSchool of Biological Sciences, Flinders University, Adelaide, Australia

Tringe SG, Rubin EM. (2005) Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics, 11, 805-14.

Tully BJ, Nelson WC, Heidelberg JF. (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. Environ Microbiol, 14, 254-67.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444, 1027-31.

Varaljay VA, Howard EC, Sun S, Moran MA. (2010) Deep sequencing of a dimethylsulfoniopropionate-degrading gene (dmdA) by using PCR primer pairs designed on the basis of marine metagenomic data. Appl Environ Microbiol., ;76, 609-17.

Vega Thurber R, Willner-Hall D, Rodriguez-Mueller B, Desnues C, Edwards RA, Angly F, Dinsdale E, Kelly L, Rohwer F. (2009) Metagenomic analysis of stressed coral holobionts. Environ Microbiol, 11, 2148-63

Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E. (2009) Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol, 83, 4642-51.

Weng FC, Su CH, Hsu MT, Wang TY, Tsai HK, Wang D. (2010) Reanalyze unassigned reads in Sanger based metagenomic data using conserved gene adjacency. BMC bioinformatics, 11, 1, p. 565-75..

Willner D, Furlan M. (2009) Deciphering the role of phage in the cystic fibrosis airway. Virulence, 1, 309-13.

Wooley JC, Godzik A., and Friedberg I.(2010) A primer on metagenomics. PLoS computational biology, 6, e1000667.

Woyke T. (2010) Metagenomic Analysis of Microbial Symbionts in a Gutless Worm. Lawrence Berkeley National Laboratory: Lawrence Berkeley National Laboratory. LBNL Paper LBNL-2723E.

Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y, Wang P, Tang X, Wang H, Kang K, Zhang S, Zhao G, Wu W, Yang Y, Chen H, Mu F, Chen W. (2011) Metagenomic analysis of fever, thrombocytopenia and leukopenia

syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog, 7, e1002369.

Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. (2012) Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis., 6, e1485.

表

表一、「基因密度是否會影響 blastx 成效」所採用的已知物種。

Organism/Name	Genes	Genome size (Mbp)	Gene density (genes/ Mbp)
Ehrlichia ruminantium str. Gardel	989	1.5	659
Mycoplasma hyopneumoniae 168	739	0.93	795
Mycoplasma hyorhinis HUB-1	711	0.84	846
Escherichia coli BL21-Gold(DE3)pLysS AG	4439	4.57	949
Mycoplasma haemofelis str. Langford 1	1580	1.15	1374

我們選用了目前已定序完成的原核生物：*Ehrlichia ruminantium* str. Gardel(659(genes/ Mbp))、*Mycoplasma hyopneumoniae* 168(795(genes/ Mbp))、*Mycoplasma hyorhinis* HUB-1(846(genes/ Mbp))、*Escherichia coli* BL21-Gold(DE3)pLysS AG(949(genes/ Mbp))、*Mycoplasma haemofelis* str. Langford 1 (1374(genes/ Mbp))，其基因密度的計算是從基因數除以基因體大小(Mbp)得來，我們利用不同原核生物基因密度來了解基因密對 blastx 效能的影響。

表二、不同長度的模擬片段註解的結果。

模擬片段長度(bp)	模擬片段總量	註解模擬片段結果	
		陽性	陰性
500	54852	54592	260
1000	27426	27393	33
1500	18284	18279	5
2000	13713	13711	2

我們利用程式 randomly-chomp-genome.py(附錄二)模擬定序產生不同長度(bp)的模擬片段，再將模擬片段和 *E.coli* BL21 基因註解比對後，和基因有重疊的模擬片段會被註解為陽性，位於非基因區域的模擬片段則會被註解成陰性。

表三、將模擬片段進行 blastx 的結果。

模擬片段長度(bp)	模擬片段總量	將模擬片段進行blastx的結果			
		Hits	百分比%	No hits	百分比%
500	54852	46368	84.53	8484	15.47
1000	27426	26479	96.55	947	3.45
1500	18284	17960	98.23	324	1.77
2000	13713	13478	99.01	135	0.99

我們將不同長度的模擬片段和 *E.coli* BL21 胺基酸序列建構成的蛋白質資料庫進行 blastx 後，和資料庫有比對到同源序列者為 hits，和資料庫沒有比對到同源序列者為 no hits。Hits 和 No hits 的百分比是計算分別是 Hits(No hits)的數量除以模擬片段的總量，從模擬片段進行 blastx 的結果中，模擬片段大多和資料庫有比對到同源序列(Hits)。

表四、不同長度的模擬片段 blastx 的敏感度與特異度。

(bp)	TP	FN	TN	FP= FP1+FP2	Sensitivity= TP/(TP+FN)	Specificity= TP/(TP+FP)
500	45825	8224	260	1059	84.78	97.74
1000	26134	914	33	483	96.62	98.19
1500	17808	319	5	275	98.24	98.48
2000	13454	133	2	138	99.02	98.98

不同長度的模擬片段(pseudo-reads)，blastx 設定的 E 值為 1E-40，將模擬片段註解的結果和 blastx 的結果比對後，比對出 TP、FN、TN、FP，並計算出不同長度模擬片段 blastx 的敏感度和特異度。

表五、不同長度的模擬片段所涵蓋的基因數。

模擬片段長度(bp)	基因數
500	0.4745
1000	0.949
1500	1.4235
2000	1.898

從文獻中 *E.coli* BL21 基因密度(Lukjancenko O. et al., 2010)，換算成不同序列長度下，模擬片段所涵蓋的基因數。

表六、不同 E 值下 blastx 的敏感度和特異度結果。

(bp)	E-value	TP	FN	TN	FP1	FP2	Sensitivity= TP/(TP+FN)	Specificity= TP/(TP+FP1+FP2)
500	1.0E-02	52719	1197	209	1268	51	97.78	97.56
	1.0E-03	52682	1255	218	1247	42	97.67	97.61
	1.0E-05	52565	1369	223	1250	37	97.46	97.61
	1.0E-10	52244	1729	223	1205	13	96.80	97.72
	1.0E-40	45825	8224	260	1059	0	84.78	97.74
1000	1.0E-02	26794	334	19	403	14	98.77	98.47
	1.0E-03	26789	347	19	395	14	98.72	98.50
	1.0E-05	26766	366	19	399	14	98.65	98.48
	1.0E-10	26726	396	29	409	4	98.54	98.48
	1.0E-40	26134	914	33	483	0	96.62	98.19
1500	1.0E-02	18023	164	0	166	55	99.10	98.79
	1.0E-03	18021	168	0	164	55	99.08	98.80
	1.0E-05	18014	173	0	166	55	99.05	98.79
	1.0E-10	17999	187	0	167	55	98.97	98.78
	1.0E-40	17808	319	5	225	50	98.24	98.48
2000	1.0E-02	13564	83	0	78	2	99.39	99.41
	1.0E-03	13564	83	0	78	2	99.39	99.41
	1.0E-05	13560	90	0	75	2	99.34	99.44
	1.0E-10	13549	96	0	80	2	99.30	99.40
	1.0E-40	13454	133	2	138	0	99.02	98.98

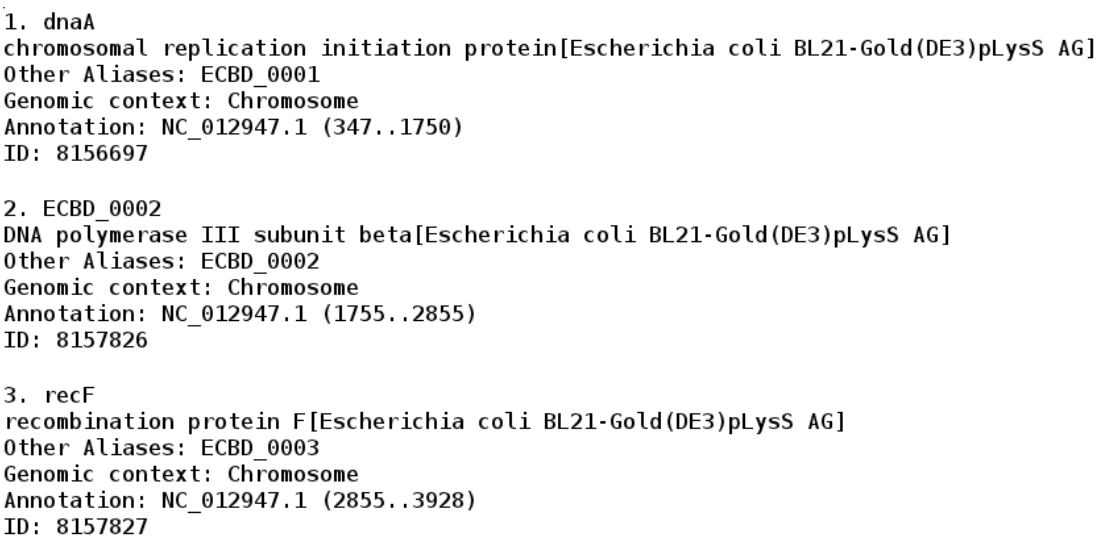
不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，blastx 敏感度和特異度的結果(%)。

表七、blastx 在原核生物不同基因密度下的敏感度和特異度。

Organism/Name	Gene density (genes/ Mbp)	Sensitivity (%)	Specificity (%)
Ehrlichia ruminantium str. Gardel	659	96.43	96.33
Mycoplasma hyopneumoniae 168	795	97.49	96.58
Mycoplasma hyorhinis HUB-1	846	96.77	97.03
Escherichia coli BL21-Gold(DE3)pLysS AG	949	97.78	97.56
Mycoplasma haemofelis str. Langford 1	1374	99.60	99.37

將不同基因密度的原核生物基因體序列分別產生 500(bp)的模擬片段，blastx 設定的 E 值為 1E-2 作 blastx，將模擬片段的註解結果和 blastx 結果進行比對並計算出 blastx 敏感度和特異度。

圖



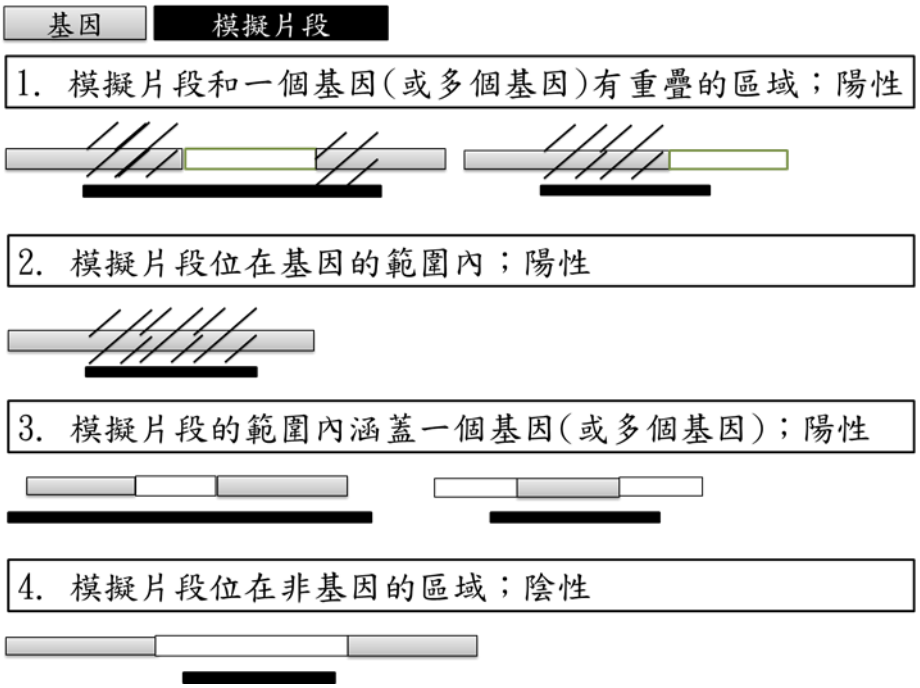
```
1. dnaA
chromosomal replication initiation protein[Escherichia coli BL21-Gold(DE3)pLysS AG]
Other Aliases: ECBD_0001
Genomic context: Chromosome
Annotation: NC_012947.1 (347..1750)
ID: 8156697

2. ECBD_0002
DNA polymerase III subunit beta[Escherichia coli BL21-Gold(DE3)pLysS AG]
Other Aliases: ECBD_0002
Genomic context: Chromosome
Annotation: NC_012947.1 (1755..2855)
ID: 8157826

3. recF
recombination protein F[Escherichia coli BL21-Gold(DE3)pLysS AG]
Other Aliases: ECBD_0003
Genomic context: Chromosome
Annotation: NC_012947.1 (2855..3928)
ID: 8157827
```

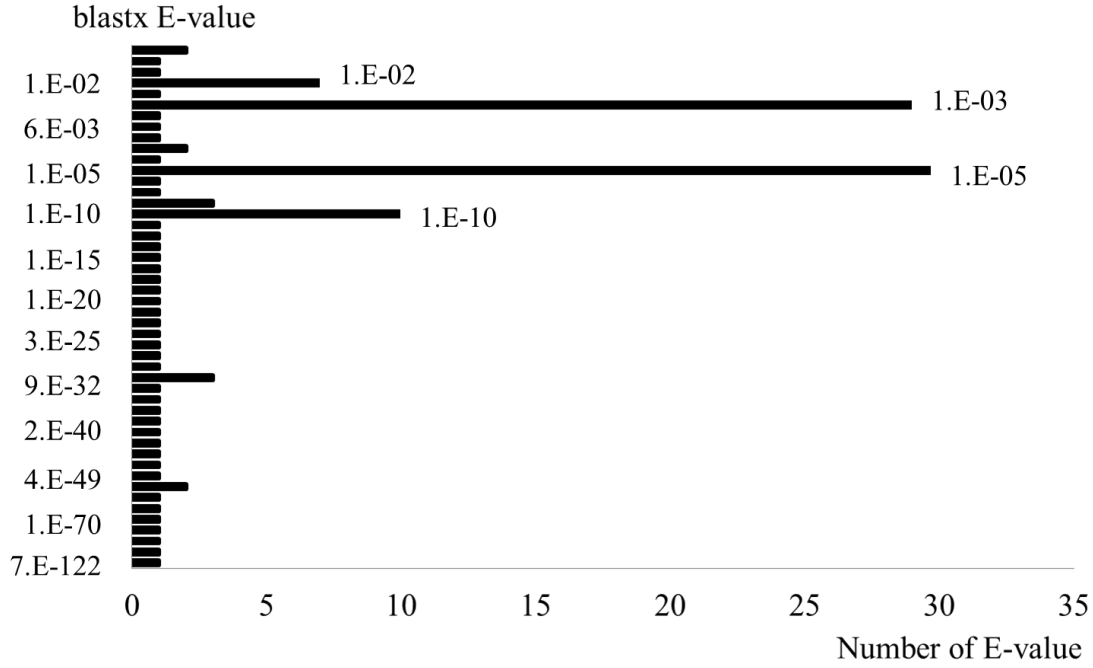
圖一、*E.coli* BL21 基因註解 txt 檔案格式截圖。

我們從 NCBI Gene 下載了 *E.coli* BL21 基因註解，下載的檔案格式我們選用 txt 檔案格式。



圖二、模擬片段和基因可能關係示意圖。

如圖所示，灰色代表基因區域，白色代表非基因區域，黑色代表模擬片段:註解模擬片段時,我們考慮到模擬片段和基因可能的關係。和基因有重疊的模擬片段會被註解為陽性,模擬片段位於非基因區域則會被註解為陰性。



圖三、多源性基因體學研究以 blastx 作分析時設定的 E 值。

我們參考了西元 2000~2012 年，多源性基因體學研究上採用 blastx 作分析工具時設定的 E 值，將 blastx-evalue-reference.pdf (附錄十一) 結果作圖，橫軸是使用 E 值的次數，縱軸是 blastx 的 E 值。

Gene name	Description	Gene ID	Start position	End Position
dnaA	chromosomal replication initiation protein	8156697	347	1750
ECBD_0002	DNA polymerase III subunit beta	8157826	1755	2855
recF	recombination protein F	8157827	2855	3928
gyrB	DNA gyrase subunit B	8157828	3957	6371
ECBD_0005	hypothetical protein	8157019	6611	7009
ECBD_0006	sugar phosphatase	8160210	7124	7936
ECBD_0007	hypothetical protein	8160211	7982	8638
ECBD_0008	GntR domain protein	8160212	8916	9605
ECBD_0009	2-dehydro-3-deoxygalactonokinase	8160213	9602	10480
ECBD_0010	2-dehydro-3-deoxy-6-phosphogalactonate aldolase	8160214	10464	11081
ECBD_0011	galactonate dehydratase	8160215	11078	12226
ECBD_0012	d-galactonate transporter	8160216	12346	13638
ECBD_0013	putative oxidoreductase	8160217	13635	14699
ECBD_0014	hypothetical protein	8160218	14800	16014
ECBD_0015	hypothetical protein	8160219	16016	16423
ECBD_0016	heat shock protein IbpA	8160220	16654	17067
ECBD_0017	heat shock chaperone IbpB	8160221	17179	17607
ECBD_0018	hypothetical protein	8160222	17803	19464
ECBD_0019	pseudo	8160223	19461	19946
…	…	…	…	…

圖四、*E.coli* BL21 基因註解轉換格式成 csv 檔案格式截圖。

我們從 NCBI Gene 下載的 *E.coli* BL21 基因註解檔案格式是 txt 檔案格式，為了之後便於分析，我們經由 Parse-genome-txt2.py 程式(附錄一)整理成 csv 檔案格式(record.csv)的結果截圖。將基因註解轉換格式成表格(csv)的形式(record.csv，附錄十九)。

Start position	End Position	Description
347	1750	chromosomal replication initiator protein DnaA
1755	2855	DNA polymerase III
2855	3928	DNA replication and repair protein RecF
3957	6371	DNA gyrase
6611	7009	protein of unknown function DUF937
7124	7936	Cof-like hydrolase
8638	7982	conserved hypothetical protein
8916	9605	GntR domain protein
9602	10480	2-dehydro-3-deoxygalactonokinase
10464	11081	KDPG and KHG aldolase
11078	12226	Mandelate racemase/muconate lactonizing protein
12346	13638	d-galactonate transporter
14699	13635	putative oxidoreductase
14800	16014	conserved hypothetical protein
16423	16016	protein of unknown function DUF1375
16654	17067	heat shock protein Hsp20
17179	17607	heat shock protein Hsp20
17803	19464	YidE/YbjL duplication
20394	22010	PTS system
22010	23332	glycoside hydrolase family 4
...	…	…

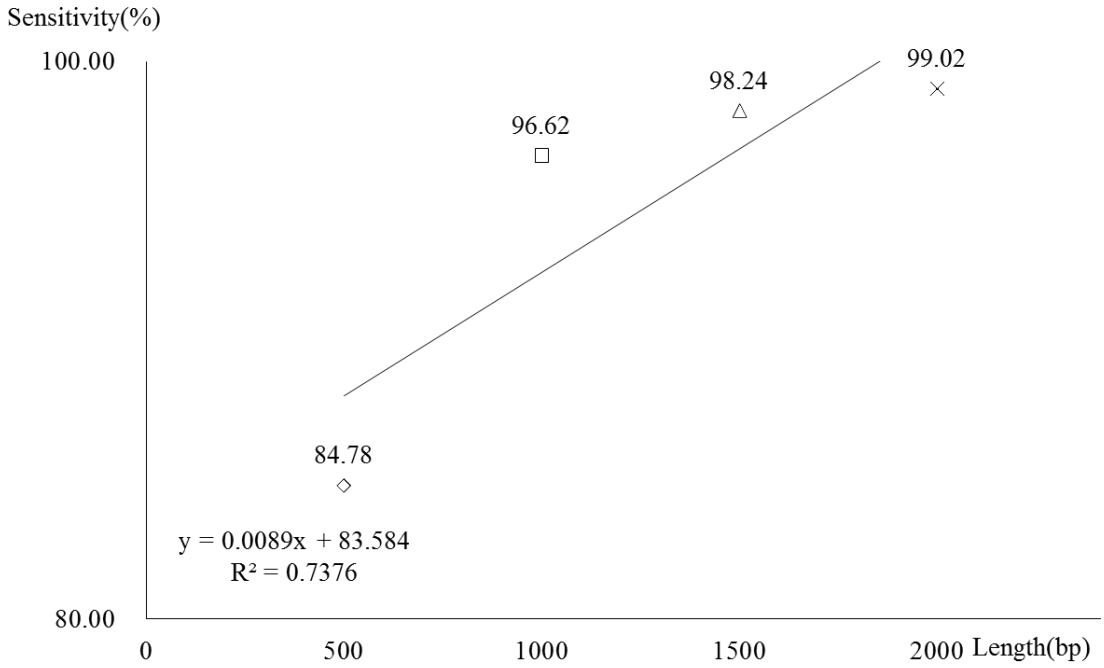
圖五、將 *E.coli* BL21 胺基酸、DNA 序列整理後的結果截圖。

將 *E.coli* BL21 胺基酸序列(faa)的基因敘述和 DNA 序列(fnn)的基因敘述，經由 python 程式(faa-record.py，附錄十四)整理成 csv 檔案格式(faa-record.csv，附錄十五)。

Gene name	Description
dnaA	chromosomal replication initiator protein DnaA
ECBD_0002	DNA polymerase III
recF	DNA replication and repair protein RecF
gyrB	DNA gyrase
ECBD_0005	protein of unknown function DUF937
ECBD_0006	Cof-like hydrolase
ECBD_0007	conserved hypothetical protein
ECBD_0008	GntR domain protein
…	…

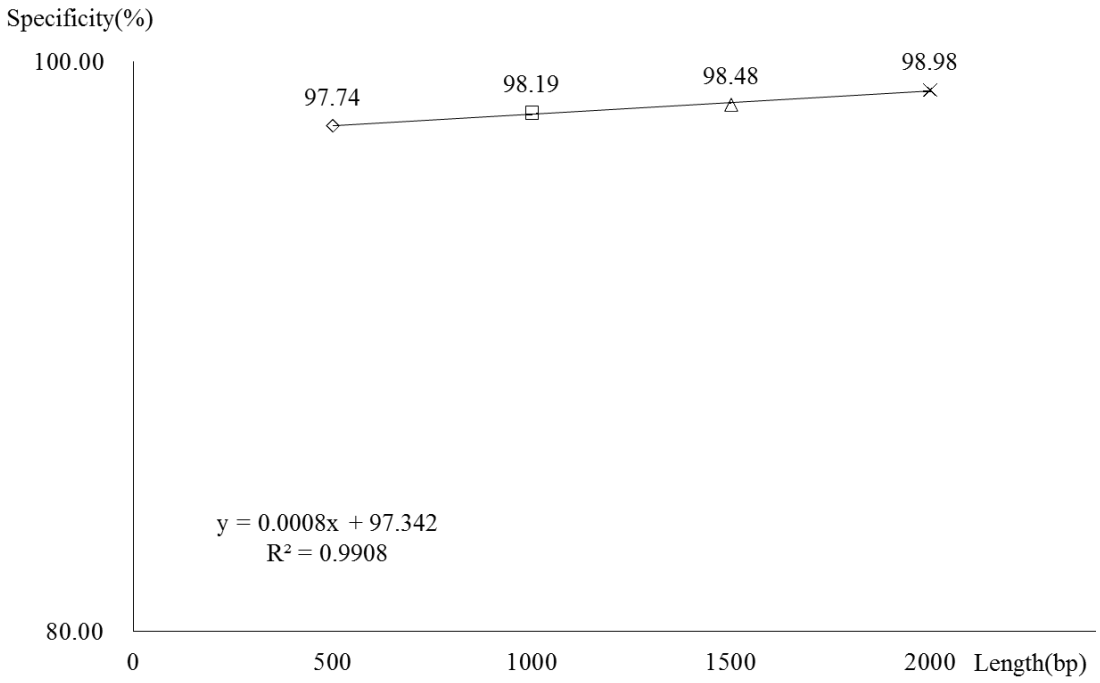
圖六、將 *E.coli* BL21 基因註解重新整理後的結果截圖。

將 *E.coli* BL21 基因註解(record.csv，附錄十九)和 faa-record.csv(附錄十五)經由 python 程式(recordC.py，附錄十六)比對後，重新整理成 csv 檔案格式(recordC.csv，附錄十七)的結果截圖。



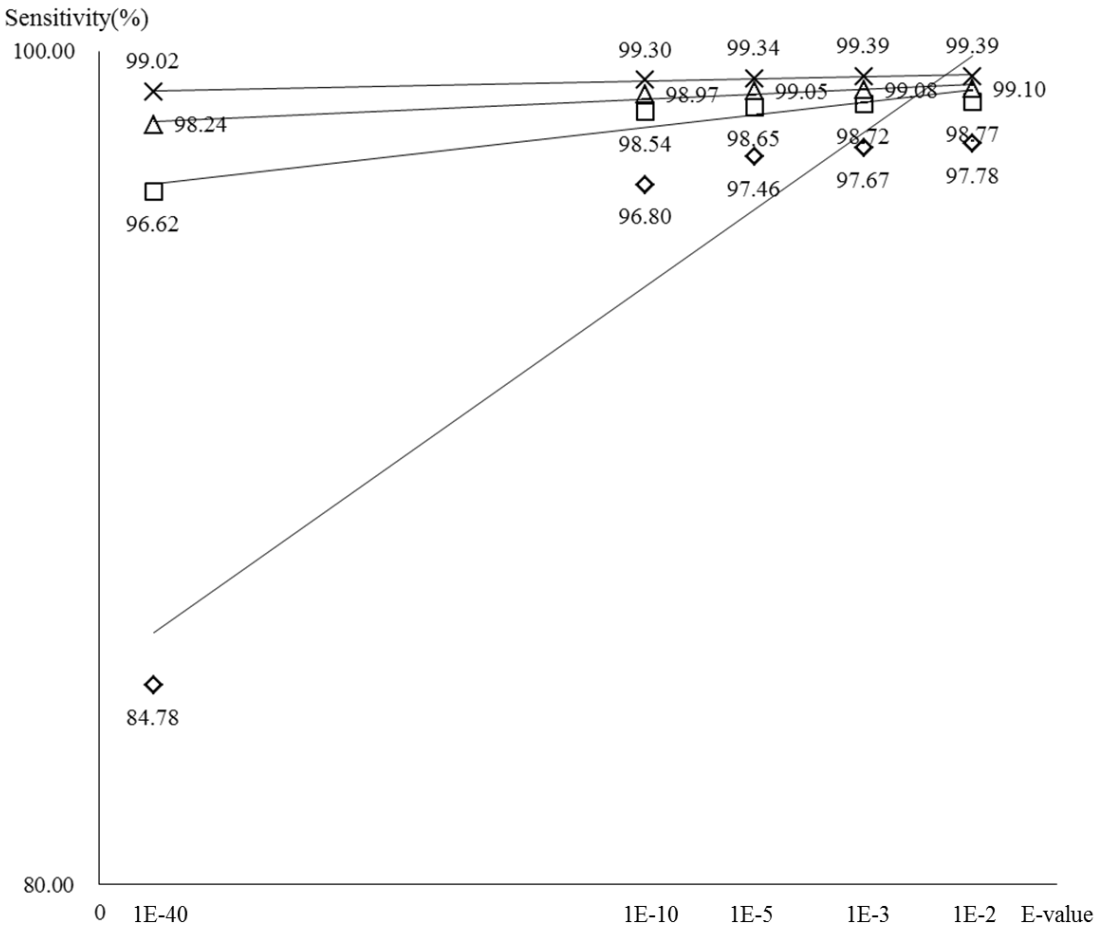
圖七、序列長度對 blastx 敏感度的影響結果。

不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的 E 值為 1E-40，blastx 敏感度的結果(表四)作圖，其中◇為 500bp 模擬片段，❐為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同長度(500、1000、1500、2000bp)，縱軸是敏感度，數值以百分比表示。



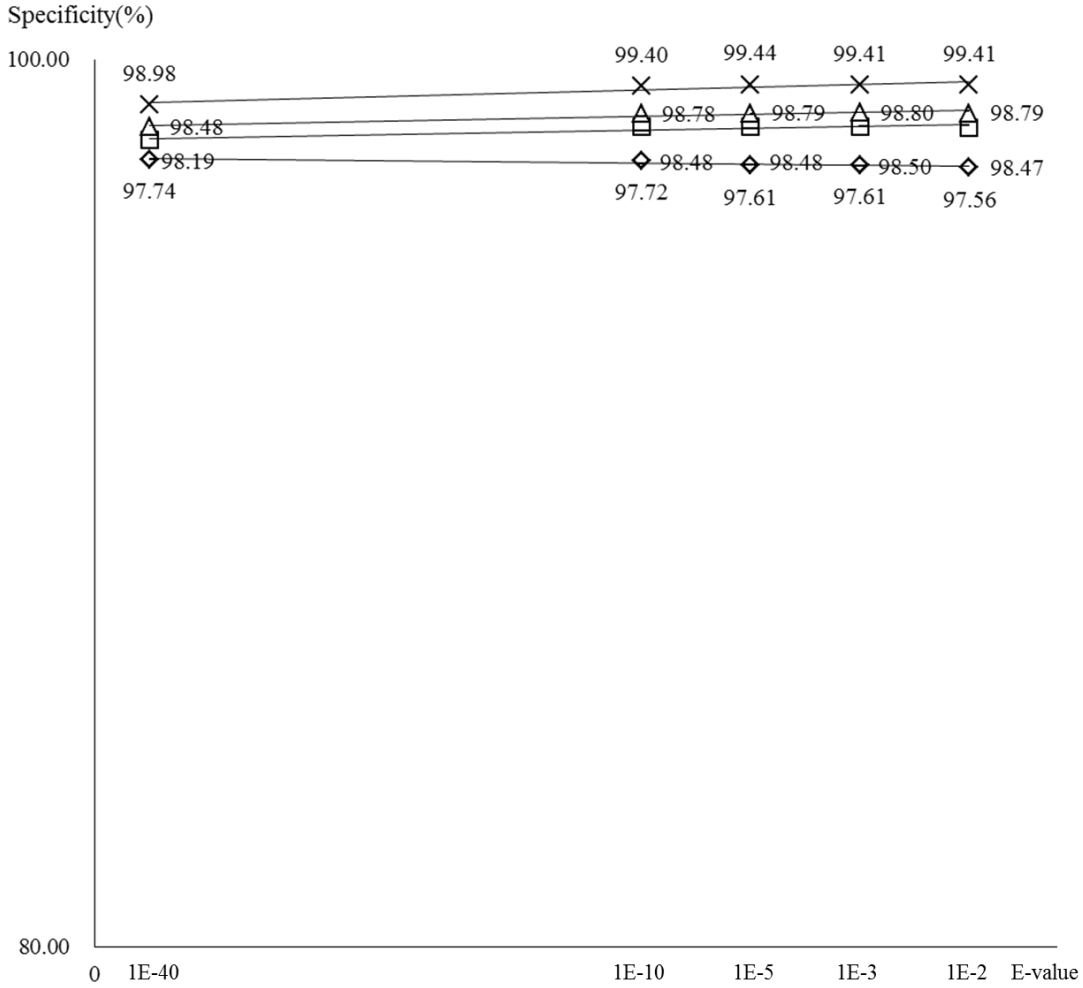
圖八、序列長度對 blastx 特異度的影響結果。

不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的 E 值為 1E-40，blastx 特異度的結果(表四)作圖，其中◇為 500bp 模擬片段，❐為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同長度(500、1000、1500、2000bp)，縱軸是特異度，數值以百分比表示。



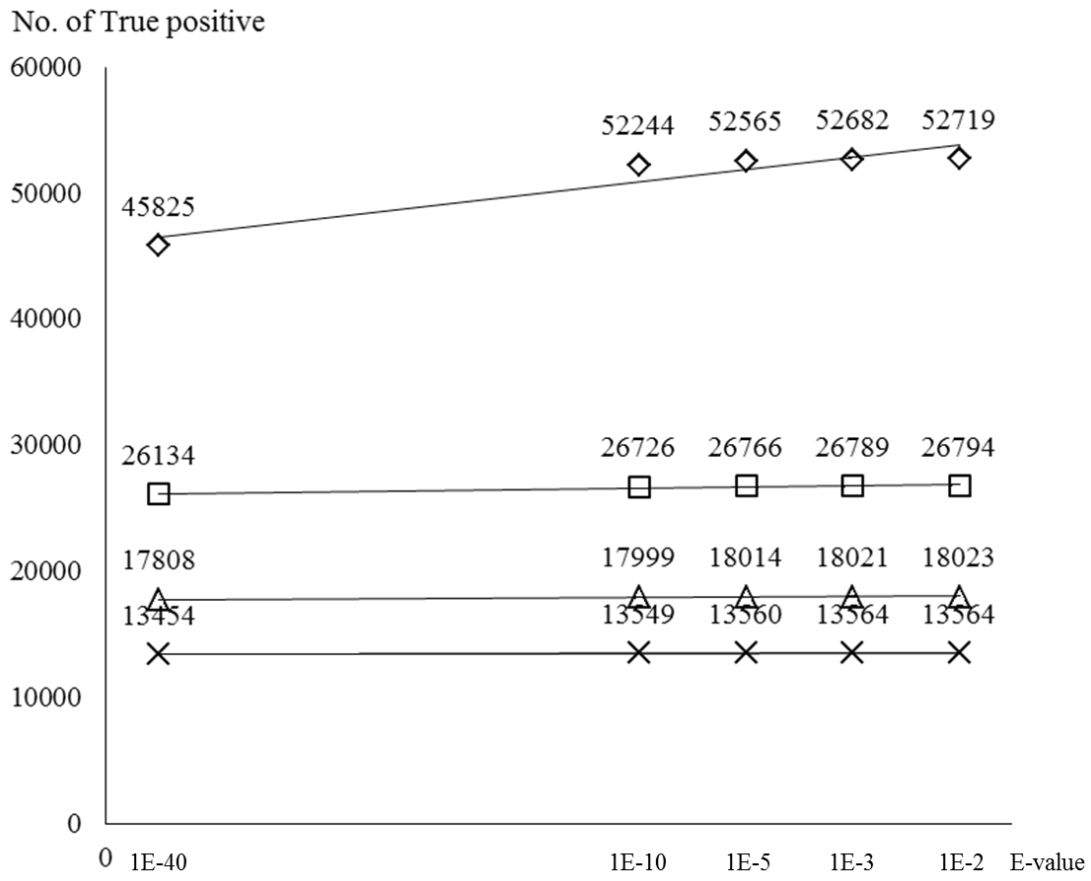
圖九、不同序列長度在不同 E 值下 blastx 的敏感度結果。

不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，blastx 敏感度的結果(表六)作圖其中◇為 500bp 模擬片段，□為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，縱軸是敏感度，數值以百分比表示。



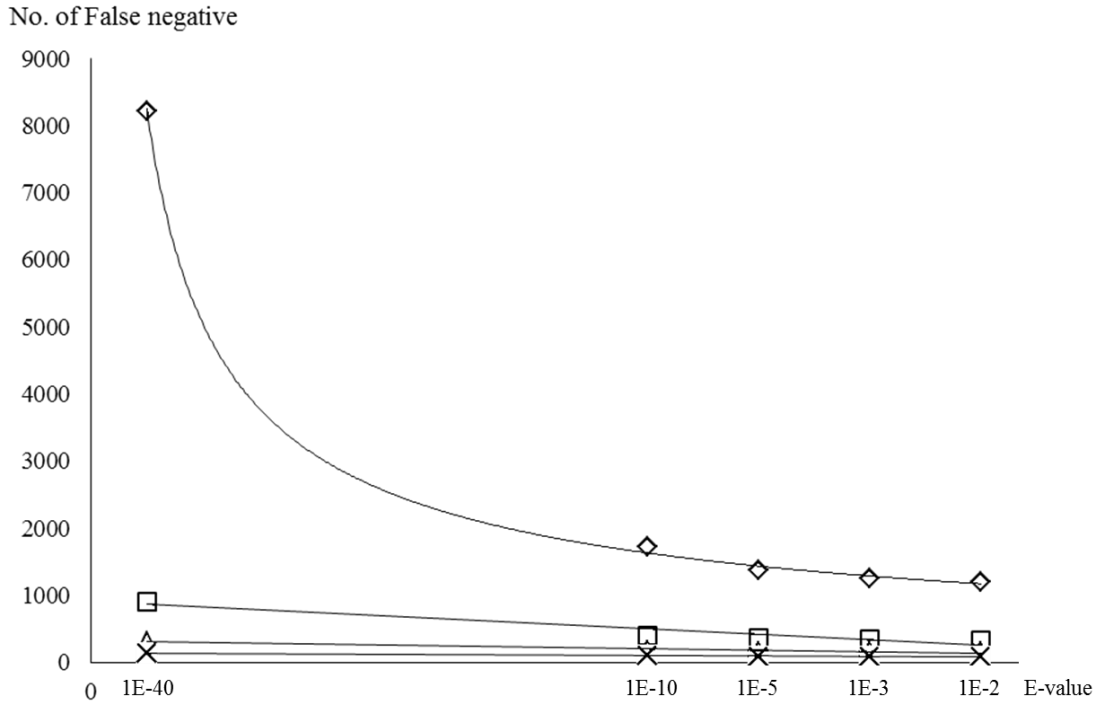
圖十、不同序列長度在不同 E 值下 blastx 的特異度結果。

不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，blastx 特異度的結果(表六)作圖，其中◇為 500bp 模擬片段，□為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，縱軸是特異度，數值以百分比表示。



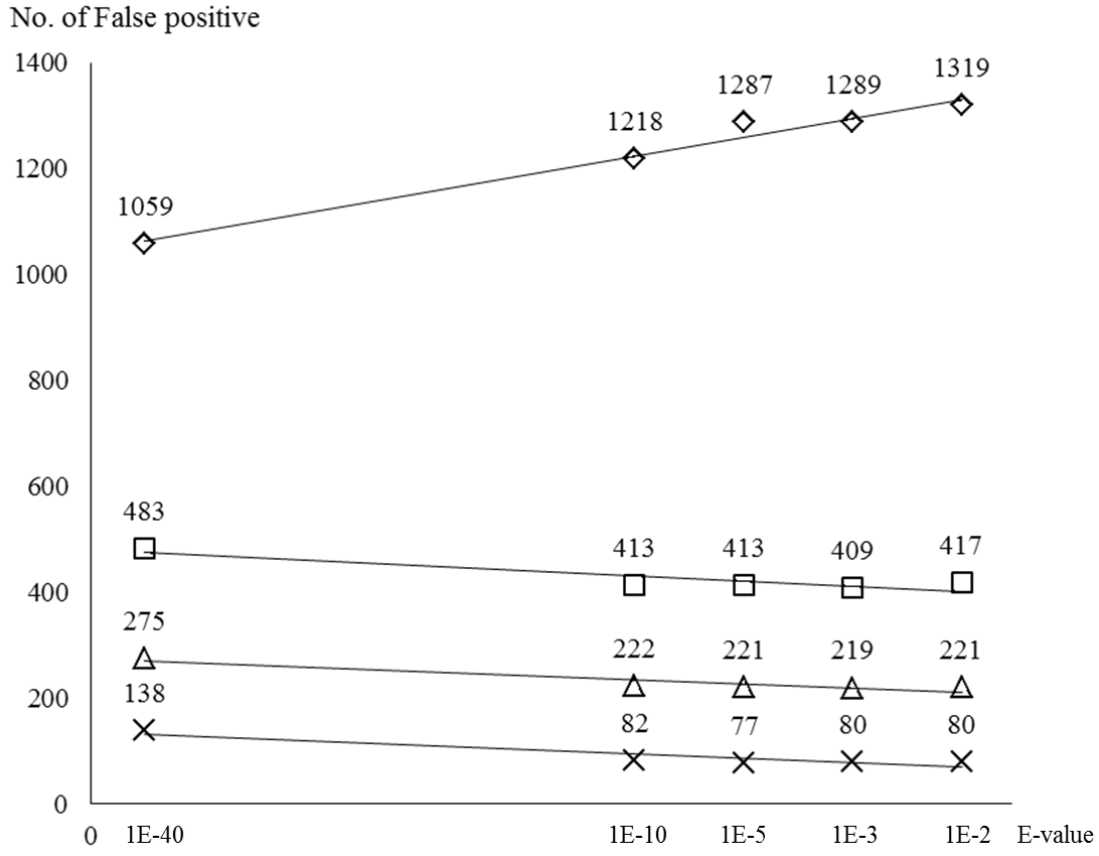
圖十一、不同長度模擬片段在不同 E 值下被判定為 TP 的數量。

我們將不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，模擬片段被判定為 TP 的數目(表六)作圖並畫上趨勢線，其中◇為 500bp 模擬片段，□為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，縱軸是 TP 的數量。



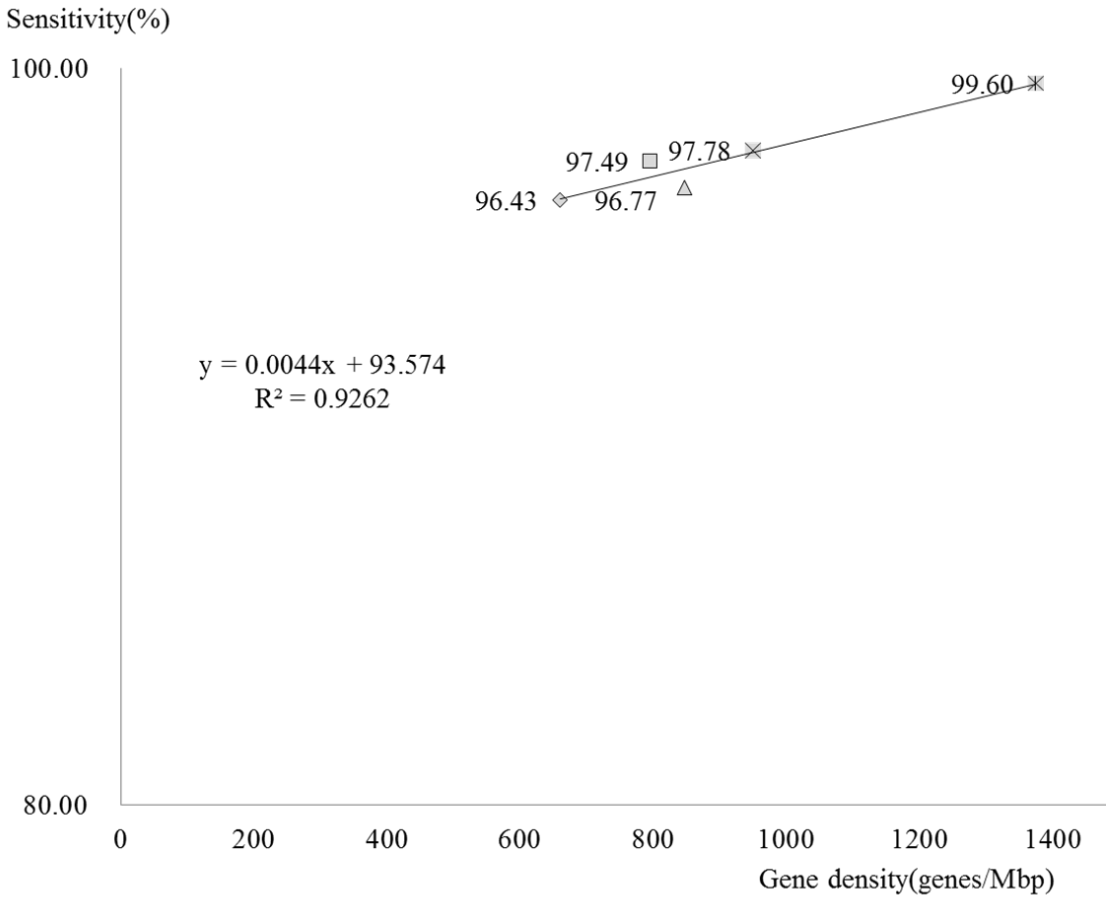
圖十二、不同長度模擬片段在不同 E 值下被判定為 FN 的數量。

我們將不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，模擬片段被判定為 FN 的數目(表六)作圖並畫上趨勢線，其中◇為 500bp 模擬片段，□為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，縱軸是 FN 的數量。



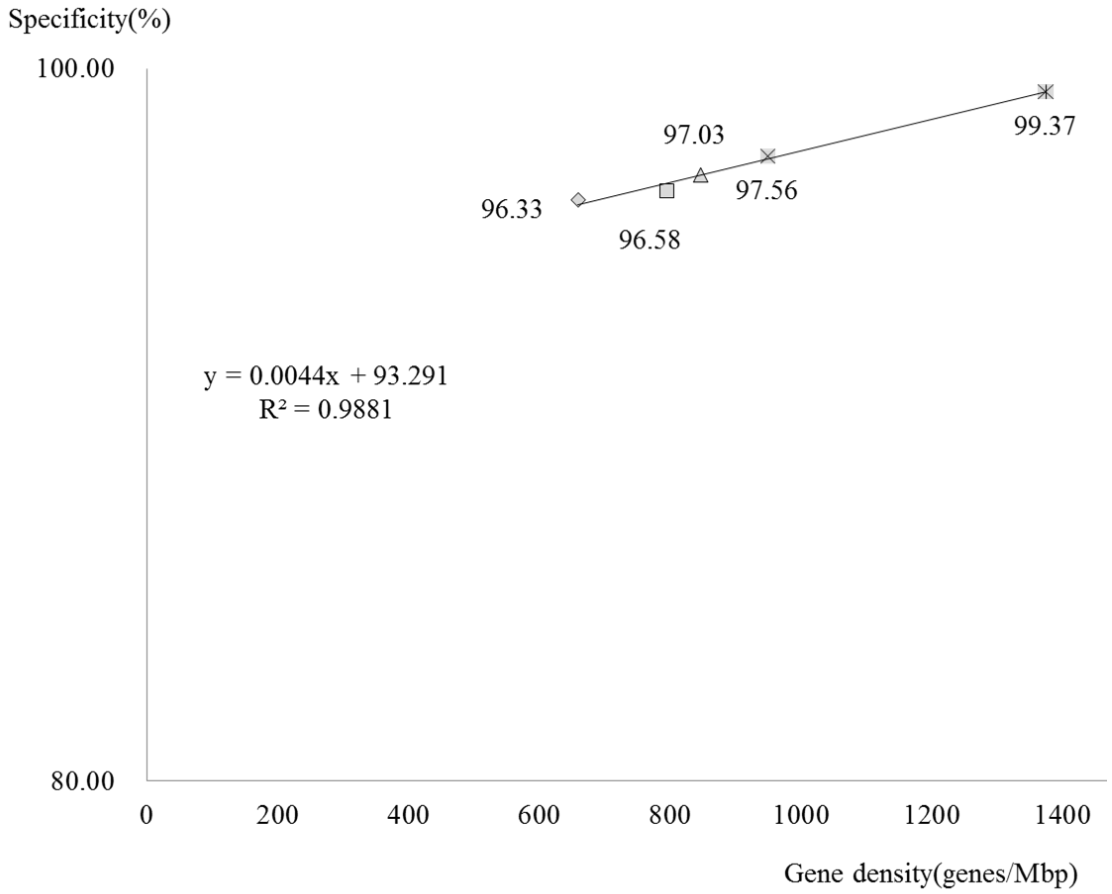
圖十三、不同長度模擬片段在不同 E 值下被判定為 FP 的數量。

我們將不同長度(500、1000、1500、2000bp)模擬片段，blastx 設定的不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，模擬片段被判定為 FP 的數目(表六)作圖並畫上趨勢線，其中◇為 500bp 模擬片段，□為 1000bp 模擬片段，△為 1500bp 模擬片段，×為 2000bp 模擬片段：橫軸是不同 E 值(1E-2、1E-3、1E-5、1E-10、1E-40)，縱軸是 FP 的數量 。



圖十四、不同密度的原核生物 blastx 敏感度的結果。

不同基因密度的原核生物分別產生 500bp 模擬片段並在 blastx 設定 E 值為 1E-2 將模擬片段進行 blastx，我們將 blastx 的敏感度結果(表七)作圖並畫上趨勢線，橫軸是基因密度，單位是在每 1Mbp 下所涵蓋的基因數量；其中◇代表的是 *E.r.*-G (659 genes/Mb)；□代表的是 *M. h.*-168 (795 genes/Mbp)；△代表的是 *M.h.*-HUB-1 (846 genes/Mbp)；✕代表的是 *E.coli* BL21；✳代表的是 *M.h.*-L1(1374 genes/Mbp)；縱軸是敏感度，數值以百分比表示。



圖十五、不同密度的原核生物 blastx 特異度的結果。

不不同基因密度的原核生物分別產生 500bp 模擬片段並在 blastx 設定 E 值為 1E-2 將模擬片段進行 blastx，我們將 blastx 的特異度結果(表七)作圖並畫上趨勢線，橫軸是基因密度，單位是在每 1Mbp 下所涵蓋的基因數量；其中◇代表的是 *E.r.*-G (659 genes/Mb)；□代表的是 *M. h.*-168 (795 genes/Mbp)；△代表的是 *M.h.*-HUB-1 (846 genes/Mbp)；×代表的是 *E.coli* BL21；✳代表的是 *M.h.*-L1(1374 genes/Mbp)；縱軸是特異度，數值以百分比表示。

附錄

以下為附錄檔案名稱，附錄內容說明及檔案另附在光碟中。

附錄一、Parse-genome-txt2.py (python2 版本)

附錄二、randomly-chomp-genome.py(python3 版本)

附錄三、1st-compare.py(python3 版本)

附錄四、reconfirmtxt.py(python3 版本)

附錄五、reconfirmsplit.py(python2 版本)

附錄六、ecoli_bl21_gold_de3.db

附錄七、Er.db

附錄八、16.db

附錄九、13.db

附錄十、Mh.db

附錄十一、blastx-evalue-reference.pdf

附錄十二、list-blast+.py(python2 版本，biopython 模組)

附錄十三、p-negativeblast.py(python2 版本)

附錄十四、faa-record.py (python2 版本)

附錄十五、faa-record.csv

附錄十六、recordC.py

附錄十七、recordC.csv

附錄十八、0718p.py

附錄十九、record.csv

附錄二十、fptp.py

附錄二十一、2000-6-negative.txt

附錄二十二、2000-6-positive.txt

附錄二十三、2000-6-40-blastx_hits.txt

附錄二十四、2000-6-40-blastx_nohits.txt

附錄二十五、2000-6-10-blastx_hits.txt

附錄二十六、2000-6-10-blastx_nohits.txt

附錄二十七、2000-6-5-blastx_hits.txt

附錄二十八、2000-6-5-blastx_nohits.txt

附錄二十九、2000-6-3-blastx_hits.txt

附錄三十、2000-6-3blastx_nohits.txt

附錄三十一、2000-6-2-blastx_hits.txt

附錄三十二、2000-6-2-blastx_nohits.txt

附錄三十三、BL21-2000-6.fasta

附錄三十四、BL21-1500-6.fasta

附錄三十五、1500-6-negative.txt

附錄三十六、1500-6-positive.txt

附錄三十七、1500-6-40-blastx_hits.txt

附錄三十八、1500-6-40-blastx_nohits.txt

附錄三十九、1500-6-10-blastx_hits.txt

附錄四十、1500-6-10-blastx_nohits.txt

附錄四十一、1500-6-5-blastx_hits.txt

附錄四十二、1500-6-5-blastx_nohits.txt

附錄四十三、1500-6-3-blastx_hits.txt

附錄四十四、1500-6-3-blastx_nohits.txt

附錄四十五、1500-6-2-blastx_hits.txt

附錄四十六、1500-6-2-blastx_nohits.txt

附錄四十七、BL21-1000-6.fasta

附錄四十八、1000-6-negative.txt

附錄四十九、1000-6-positive.txt

附錄五十、1000-6-40-blastx_hits.txt

附錄五十一、1000-6-40-blastx_nohits.txt

附錄五十二、1000-6-10-blastx_hits.txt

附錄五十三、1000-6-10-blastx_nohits.txt

附錄五十四、1000-6-5-blastx_hits.txt

附錄五十五、1000-6-5-blastx_nohits.txt

附錄五十六、1000-6-3-blastx_hits.txt

附錄五十七、1000-6-3-blastx_nohits.txt

附錄五十八、1000-6-2-blastx_hits.txt

附錄五十九、1000-6-2-blastx_nohits.txt

附錄六十、BL21-500-6.fasta

附錄六十一、500-6-negative.txt

附錄六十二、500-6-positive.txt

附錄六十三、500-6-40-blastx_hits.txt

附錄六十四、500-6-40-blastx_nohits.txt

附錄六十五、500-6-10-blastx_hits.txt

附錄六十六、500-6-10-blastx_nohits.txt

附錄六十七、500-6-5-blastx_hits.txt

附錄六十八、500-6-5-blastx_nohits.txt

附錄六十九、500-6-3-blastx_hits.txt

附錄七十、500-6-3-blastx_nohits.txt

附錄七十一、500-6-2-blastx_hits.txt

附錄七十二、500-6-2-blastx_nohits.txt

附錄七十三、CP001665.faa

附錄七十四、CP001665.ffn

附錄七十五、E-500-6.fasta

附錄七十六、E-Crecord2.csv

附錄七十七、E-negative.txt

附錄七十八、E-positive.txt

附錄七十九、E-500-6-2-blastx_hits.txt

附錄八十、E-500-6-2-blastx_nohits.txt

附錄八十一、M-500-6.fasta

附錄八十二、M-Crecord2.csv

附錄八十三、M-negative.txt

附錄八十四、M-positive.txt

附錄八十五、M-500-6-2-blastx_hits.txt

附錄八十六、M-500-6-2-blastx_nohits.txt

附錄八十七、Mh168.fasta

附錄八十八、Mh168-Crecord2.csv

附錄八十九、Mh168-negative.txt

附錄九十、Mh168-positive.txt

附錄九十一、Mh168-500-6-2-blastx_hits.txt

附錄九十二、Mh168-500-6-2-blastx_nohits.txt

附錄九十三、MhHUB-1.fasta

附錄九十四、MhHUB-1-Crecord2.csv

附錄九十五、MhHUB-1-negative.txt

附錄九十六、MhHUB-1-positive.txt

附錄九十七、MhHUB-1-500-6-2-blastx_hits.txt

附錄九十八、MhHUB-1-500-6-2-blastx_nohits.txt

附錄九十九、readme.pdf