國立臺灣大學生命科學院漁業科學研究所

碩士論文

Institude of Fisheries Science

College of Life Science

National Taiwan University

Master Thesis

以表現標誌序列重組南美白蝦之轉錄基因體並比較不同組織間的基因表現

Decipher the transcriptome of *Litopenaeus vannamei* by assembling ESTs and compare gene expression in different libraries

林宜靜

Yi-Ching Lin

指導教授：林仲彥 博士

Advisor: Chung-Yen Lin, Ph. D.

中華民國 101 年 7 月

July, 2012

謝辭

碩士的求學期間受到許多人的幫助，隨著論文的完成，我的學業也將告一個段落，在此感謝一路上給予我幫助關心的所有人。

感謝我的指導老師 林仲彥博士，在論文上的悉心指導及生活上的關心，謝謝老師使我有機會進行跨領域的研究，並且學習從不同角度進行思考及探究問題。同時也要特別感謝陳淑華博士，在論文撰寫過程中給予我許多的意見，並且提醒我研究上需注意的眾多事項，使得我的論文能夠順利完成。此外也要感謝口試委員韓玉山博士及呂健宏博士，提供寶貴的意見及悉心協助論文的修改，使我的論文可以更加充實與嚴謹。

感謝實驗室的育彬學長在研究上的多方幫助，以及身為人生前輩的建議，不論是學術上或是生活上都使我受益良多。感謝聖堯學長、信宏學長、智偉、怡萱及昆霖學弟在碩士生涯中，給予我的所有協助及鼓勵，因為你們的幫助，讓我能夠順利完成我的學業。

最後感謝我最摯愛的父母及哥哥，總是扮演著支柱的角色，謝謝你們的支持及陪伴，讓我能夠專心致力於研究上。感謝我眾多的親朋好友以及長輩們，謝謝你們一直以來的關心及鼓勵，總讓我感到溫暖並且恢復元氣。僅將此論文獻給所有關心我、幫助我、愛我的人，謝謝你們。

中文摘要

南美白蝦（*Litopenaeus vannamei*）在蝦類養殖的地位日趨重要，進行其基因轉錄體的研究可了解該物種的基因組成，有助於對蝦生理、遺傳、育種、與病理研究等基礎領域，以因應對蝦繁養殖上所面臨的問題。EST (Expressed Sequence Tag) 是 cDNA 的部分定序序列，可提供轉錄基因的序列資訊及相對表現數量，對於尚未完成完整基因體的物種，藉由 EST 定序能得到大量蛋白質基因的資訊。本研究蒐集公開資料庫上的白蝦 ESTs 序列，共 161,241 筆資料，利用與白蝦在分類上相近且完成完整基因體定序解析資訊的水蚤 (*Daphnia pulex*) 及果蠅 (*Drosophila melanogaster*) 爲參考，利用序列的同源關係，設計序列分析流程，組裝得到 16,886 筆 contigs 及 20,515 筆 singletons，共 37,401 筆重組序列；經過資料庫比對的序列註解程序後，超過 40%的重組序列可在 non-redundant protein database 及 Pfam database 找到相似的序列或模組特徵。

在原始 EST 序列資料中，92%的是來自眼柄、鰓、血細胞、肝胰臟、淋巴器官及神經索，此六個組織的 EST 數量皆超過 20,000 筆。上述重組序列，以其 EST 來源組織的 EST 數爲基因表現量，代表六個組織的基因表現概況。分析這六個組織以重組序列的 Gene Ontology 註解資訊，發現眼柄、肝胰臟及淋巴器官的表現基因功能分類有較明顯集中（enrichment）的傾向，挑選其中差異較大的眼柄及肝胰臟，進一步分析兩組織的重組序列組成在 KEGG pathway 的配置方式，是否有集中程度的差異。綜合 KEGG pathway 與 GO term 的分析結果顯示，兩組織間都有共同表現的重組序列，大多是參與轉譯過程的基因，眼柄組織特有序列較集中表現 actin 及 myosin 兩個基因，肝胰臟組織的特有序列，則是表現血藍素 (hemocyanin)。本研究提供不同的序列組裝策略，運用重組序列之註解了解組織特異性，其結果能幫助我們對白蝦的基因更加了解，過程得到的資料也可提供後續

研究者作爲研究之參考。

參考網站：http://ips.sinica.edu.tw/lv

關鍵詞：南美白蝦、表現標誌序列、參考基因體、重組、註解

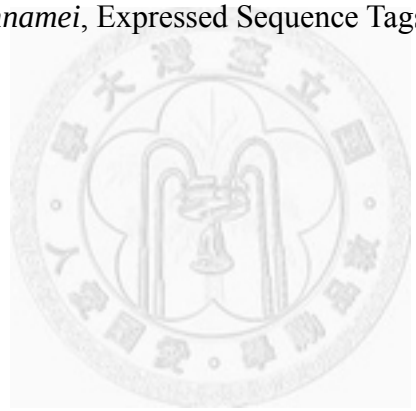
Abstract

The Pacific white shrimp, *Litopenaeus vannamei,* has became a more important animal in shrimp aquaculture. Although its genome has not been fully sequenced, this economically important species has attracted researchers for building a rich source of ESTs for benefiting both the basic biology and the applied science. Expressed Sequence Tags (ESTs), are short sequences derived from a batch-wise partial sequencing result of cDNA library, which may represents the relative expression level of transcripts in the library. It is an effective approach to get numerous sequences of protein coding genes of a species without previous knowing about the genome context. In this study, we collected 161,241 ESTs of white shrimp from public database. Reference sequences from two taxonomically closed species *Daphnia pulex* and *Drosophila melanogaster* were used for selecting subset for *de novo* assembling. Totally, 37,401 assembled sequences, including 16,886 contigs and 20,515 singletons, were obtained. Over 40% of the assemblies could be matched to homolog sequences in non-redundant protein database and/or protein domain feature in Pfam database. Six tissues including eyestalk, gills, hemocyte, hepatopancreas, lymphoid organ and nerve cord are the major tissue source of this collection. Assemblies were annotated and the expression level of each assembled contig/ singleton in each tissue was calculated by the number of ESTs contributed to this contig/singleton. Enrichment analyses were further applied to detect

Gene Ontology terms, and KEGG pathway for describing the unique function of each tissue. For example, the unique expressing subset of hepatopancreas is related to hemocyanin in KEGG pathway enrichment analysis result, while the unique subset in eyestalk represent actin and myosin genes. In summary, this study provides an alternative strategy of assembling EST and the data produced from this research can assist any studies about white shrimp in the future.

Reference URL: http://ips.iis.sinica.edu.tw/lv

Keywords: *Litopenaeus vannamei*, Expressed Sequence Tags, reference genome, assembly, annotation



目 錄

圖目錄

表目錄

第一章 簡介

1.1 研究背景

台灣是個四面環海的國家，具有發展漁業的優良條件。除了沿岸近海與遠洋的捕撈漁獲之外，養殖漁業是一個重要且富潛力的產業，水產養殖技術的發展與經營開發，將有助於提供人類高品質的蛋白質來源，避免過度漁撈導致海洋生物資源枯竭。

台灣最早的養殖記錄始於十七世紀，在 20 世紀後半數十年間，台灣的養殖產業迅速發展，其中蝦類養殖爲重要的一環。1968 年廖一久博士建立蝦苗人工養殖技術，促使台灣養蝦業有了突破性的成長(Liao, 1969)，根據 FAO 的統計，台灣的蝦類養殖量從 1970 年的 500 噸增加到 1987 年的 88,264 噸(FAO, 2010-2012)，其中草蝦(*Penaeus monodon*)的養殖年產量則從 73 噸增加到 78,548 噸，由統計數字可看出，在 1970 年到 1987 年這段時間內，台灣主要的蝦類養殖對象爲草蝦，台灣在當時也獲得養蝦王國的美譽。然而 1988 年的蝦類產量卻忽然下降到了 39,507 噸，之後產量便逐年下滑，到了 1998 年只剩下 5,549 噸(圖 1-1)。

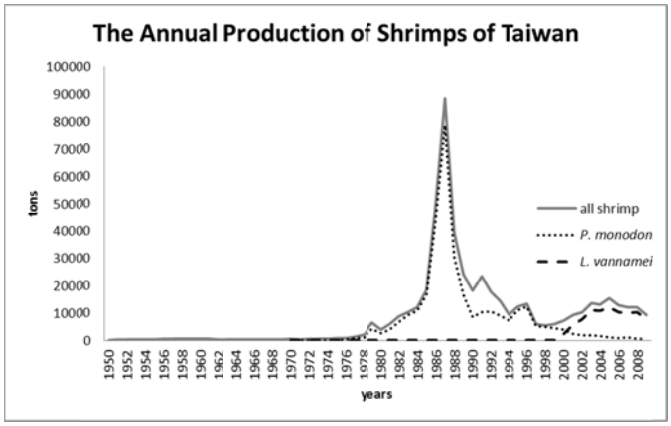


圖 1 -1 台灣蝦類養殖產值，實線是所有蝦類的總和值，點線是草蝦單計值，斷線是白蝦單計值，數據來源：(FAO. Fisheries and Aquaculture Department.)

造成草蝦的產量急劇下滑的可能原因，包括台灣採取集約式的養殖，過

度密集的養殖造成養殖環境的品質管控不易、水質惡化(Primavera, 1998)，在1990 年時疫情爆發，養殖蝦大量死亡，產量大幅的下降，除了少數地區以外，全國草蝦養殖池連續數年無法收成，重重打擊了養殖業。台灣在 1998 年引進無特定病原 (Specific Pathogen Free，SPF) 白蝦(*Litopenaeus vannamei*)種蝦，並改變養殖規模、加強疫病監測與管理方式之後，蝦類養殖的產量才又逐漸開始增加(鄭金華, 2007)。從圖 1-1 可看出，自 2000 年以後，白蝦是現今台灣蝦類養殖的主要品種，蝦類養殖總產量幾乎都是由白蝦的產量所決定。

在全世界蝦類產量的部分(圖 1-2)，也可看到白蝦逐漸佔有一席之地，2000 年以前，白蝦的產量皆沒有超過蝦類總產量的 20%，但是 2000 年以後，白蝦的產量開始呈現爆炸性的增長，從 2000 年的 146,362 噸增加爲 2009 年的 2,327,534 噸，總產量的佔有率快速地上升到 66%左右，顯示現今世界蝦類的總量，約有 2/3 都是來自於白蝦這一個蝦種(FAO, 2010-2012)。由此可看出白蝦不僅在台灣的養殖業佔有重要地位，對全世界來說也是一重要物種，因此，白蝦的研究變得越來越迫切。

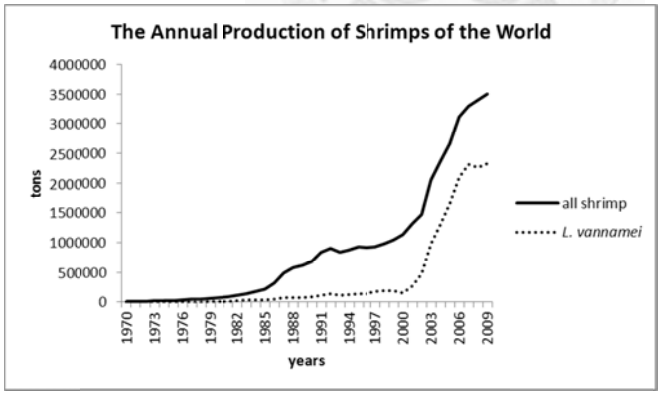


圖 1 -2 世界蝦類養殖產值，實線是所有蝦類的總和值，點線是單計白蝦產值，數據來源：(FAO. Fisheries and Aquaculture Department)

影響蝦類養殖的因素有很多，環境因子包括水質、溶氧量、鹽度及溫度等(Preston, 1985)，而生物性的因子則包括蝦苗體的狀況以及病原體的感染。一般來說，一隻健康蝦子的體內與其所處環境中有許多常在菌，也可能會包括一些致病病原，當蝦體感受到環境緊迫時，便可能被這些病原所感染；在

集約或是半集約的養殖方式下，養殖環境會因爲族群太過密集而惡化，如溶氧量下降或 pH 值改變，這些物理因子的變動會使得蝦體變得脆弱，容易受到病原侵襲，造成疾病爆發，如白點病 (White Spot virus disease)、哈威弧菌的感染等(Cheng *et al.*, 2003,Cheng *et al.*, 2003,Selvam *et al.*, 2012)。從 1970 年代蝦類水產養殖開始興起，許多國家都曾先後居於世界最高產量國的地位，但是皆因疾病的爆發導致產量迅速銳減，如 1988 年的台灣、1993 年的中國及 1996 年的泰國。爲了維持這個產業且避免再次發生大規模的疾病爆發，除了重新審視養殖方式之外，對於蝦類的物種研究也成爲另一個急需面對的課題。

1.2 研究動機及目的

Expressed sequenced tag (EST，表現標誌序列) ，是由特定生物檢體製備的 cDNA 基因庫中，隨機抽取載體殖株，經單次單向（或雙向）定序所得的基因表現序列集合。EST 的序列長度受限於定序分析方法，不能確保涵蓋基因表現序列的全長；另外，序列內容是單次定序判讀結果，也使序列本身內容可靠性稍顯不足。但是，這些表現基因的「部分序列」帶有轉錄基因的資訊可供利用，特別是在研究尚未有完整基因體資訊的物種時，ESTs 可以讓我們得到許多基因體轉錄產物的資訊，包括序列本身以及其相對表現數量(Nagaraj *et al.*, 2007)。1991 年 ESTs 開始被使用於人類基因體研究；採樣頻度足夠的 ESTs 可以提供特定組織的基因表現概況，整合基因體與 EST 資料，可作爲基因體註解的補充，發現新的基因，定義基因體的結構以及促進蛋白質體的分析(Adams *et al.*, 1991,Jongeneel, 2000,Rudd, 2003)。

一般來說，EST 定序研究規模大約是數百條至數萬條，這種大量定序工作後所產生的資料，以人工的方式來分析不僅費日耗時且容易產生品質不一的序列分析結果與註解，因此必須借助資訊科學的方法，如字串比對、資料

庫管理、對整批次的 EST 序列做處理等，從序列的定序品質檢查、載體序列清除（vector masking）、分群（clustering）、序列組裝（assembly）、序列定址對映（mapping）到功能註解（annotation）與基因表現概況分析（gene expression profiling），EST 定序研究可增加各種未知序列的資訊 (Sathiyamoorthy *et al.*, 2010)。

過去由於基因體定序研究的耗費龐大，完成完整基因體序列解析的物種侷限在少數生物模式物種。相對地，EST 定序研究是能夠以比較小的花費，定序到大量能夠製造蛋白質產物的基因，而且序列組裝的問題較小，因此受到研究者的重視。就對蝦類研究來說，較具規模的 EST 定序物種有四：草蝦（*Penaeus monodon*）、白蝦（*Litopenaeus vannamei*）、中國明對蝦（*Fenneropenaeus chinensis*）與斑節蝦（*Marsupenaeus japonicus*），約佔所有十足目 EST 序列數量的一半(表 1-1)，其中以白蝦 ESTs 數量最大，達十六萬餘筆。

表 1-1 ESTs 定序數量超過 3000 筆的十足目（Decapoda）物種，資料來源：NCBI Taxonomy (Date: March, 2012)

taxonomy	sequence number
Decapoda	429,499
Dendrobranchiata	216,436
Penaeidae (penaeid shrimps)	216,436
Fenneropenaeus chinensis (fleshy prawn)	10,446
Litopenaeus vannamei (Pacific white shrimp)	161,241
Penaeus monodon (black tiger shrimp)	39,397
Marsupenaeus japonicus	3,156
Pleocyemata	213,063
Anomura (hermit crabs)	97,806
Petrolisthes cinctipes (flat porcelain crab)	97,806
Astacidea (true lobsters and crayfishes)	30,921

Homarus americanus (American lobster)	29,957
Brachyura (short-tailed crabs)	68,583
Callinectes sapidus (blue crab)	10,563
Carcinus maenas (green crab)	15,558
Portunus trituberculatus (swimming crab)	13,985
Scylla paramamosain (green mud crab)	3,841
Eriocheir sinensis (Chinese mitten crab)	16,987
Celuca pugilator (Atlantic sand fiddler crab)	3,646
Caridea	13,064
Macrobrachium nipponense (oriental river prawn)	8,458
Macrobrachium rosenbergii (giant freshwater prawn)	4,427

白蝦迄今仍沒有完整的基因體資料，因此我們希望運用大型的公開資料庫 NCBI 上所儲存的 EST 資料，將 ESTs 經過 assembly，得到白蝦的轉錄基因體（transcriptome），運用生物資訊的工具，在公開的 non-redundant database 中找尋其同源性基因，藉此預測分子功能的註解及可能參與的生理代謝途徑，進而比較不同組織間其表現基因的差異，希望藉由不同的組裝方式得到較爲可信的轉錄基因體，預測轉錄基因的功能，可讓我們對白蝦的生理狀態有進一步的了解，或許對於疾病的感染防治能有所助益，此外，也能作爲白蝦基因體研究，以及其他對蝦物種的轉錄體研究的參考模式。

1.3 文獻探討

1.3.1 單一組織的基因表現

2002 年，Supungul *et al*. 以 ESTs 爲研究方法分析草蝦 (*P. monodon*) 血細胞 (hemocyte) 的基因表現(Supungul *et al.*, 2002)；共有 615 條 ESTs，其中 315 條在 GenBank 的資料中有找到同源基因，剩下的 300 條則沒有找到適當對應。其中，表現量高的 ESTs 其所轉譯的蛋白質皆參與基因表現或蛋白質生合成

的過程，這部分的表達序列標籤爲全部的 17.7%，在這 615 條表達序列標籤中也發現擁有防禦功能及維持生理平衡的蛋白質有 55 條，同時，在這篇研究的結果中第一次發現對蝦類的 antilipopolysaccharide factor (ALF) 及熱休克蛋白 – cpn 10。

2006 年，Yamano 和 Unuma 爲了瞭解參與雌性生殖過程中的基因，因此製作了斑節蝦 (*M. japonicus*) 眼柄的 cDNA 基因庫，通過篩選條件後，共得到 1,988 條 ESTs 以進行後續分析(Yamano and Unuma, 2006)。透過同源性搜尋的步驟後發現，1,988 條的 ESTs 大多是屬於核醣體 RNA 以及粒線體呼吸作用的酵素，這之中只有 4 條 ESTs 是已知與生殖功能相關的序列；其中三條與已知的斑節蝦眼柄賀爾蒙不同，但經過序列相似性比對後，作者認爲它們是新的 pigment-dispersing hormone (PDH，色素擴張激素)，molt-inhibiting hormone (MIH，褪殼抑制激素) 以及 crustacean hyperglycemic hormone (CHH，高醣激素)，剩下的一條表達序列標籤則是 farnesoic acid *O*-methyltransferase 的同源基因，可產生類青春激素(methyl farnesoate，MF)。

2007 年，Preechaphol *et al.* 使用 ESTs 定序分析，來研究性別相關基因及其表現時期。卵巢的 cDNA 基因庫來自野外採集的雌草蝦種蝦 (*P. monodon*)。共取 1,051 條 ESTs 進行分析與同源性比對註解，發現在此 cDNA 基因庫表現量最高的基因是 *peritrophin* 及 *thrombospondin* (TSP)(Preechaphol *et al.*, 2007)；peritrophin 是 cortical rods 的主要成分，同時也是蝦卵膠質層的前驅物。另外，在 25 個與性別相關的同源性基因中，*female sterile* 及 *ovarian lipoprotein receptor* 只表現於成年的蝦體卵巢組織內，其餘如 *chromobox protein*，*phosphatidylinositol 4 kinase*，*thioredoxin peroxidase* 及 *ubiquitin specific protease 9* 等基因在卵巢中的表現量高於在精巢中的表現量，此篇研究不僅發現在卵巢中有許多性別相關的基因，同時也有許多基因參與並調控卵巢的發育。

2007 年，Dong and Xiang 定序中國對蝦(*F. chinensis*)血細胞的 ESTs 共

2,371 條，經過序列相似性比對後，發現 177 條 ESTs 與免疫防禦功能相關(Dong and Xiang, 2007)，這些與免疫功能相關的基因大致可分爲五類：第一類，anti-microbial peptides，包括 *penaeidins*，*thymosin* 及 *ALFs*，歸屬於此類的 ESTs 共有 71 條，這個類別中包含表現量最高的免疫相關基因；第二類，prophenoloxidase activating system，包括 *prophenoloxidase*，*serine proteinase*，*serine proteinase inhibitor* 及 *SOD-protein*，歸屬於此類的 ESTs 有 44 條；第三類，clotting proteins，包括 *transglutaminase*，*thrombospondin* 及 *lectin*，此類包含 39 條 ESTs；第四類，intercellular signal transduction，包括 *peroxinectin* 及 *integrin*，共有 5 條 ESTs；第五類，chaperone protein，此類中只有兩個蛋白質，HSP70 及 thioredoxin peroxidase，共 19 條 ESTs 。

2007 年，Clavero-Sales *et al.* 利用 EST 的分析方法研究白蝦(*L. vannamei*)的鰓部感染白點病毒 (White Spot Syndrome Virus，WSSV) 後的基因表現，共取 802 個 clones 進行 5 端定序，去除載體序列後得到 601 條高品質的 ESTs，進而組裝爲 79 個 contigs 及 197 個 singletons，共 276 個重組序列（assemblies）(Clavero-Salas *et al.*, 2007)。以 BLASTN 及 BLASTX 在 E value < 10^{-2} 的條件下對 NCBI nr database 進行同源性搜尋，有 87% (522/601) 的 ESTs 與 GenBank 中的已知序列有高度相似性，在這些序列中對應到 276 種不同的蛋白質，其中對應量最高的是 40S ribosomal protein S13 基因；同時也發現 148 個 ESTs 與 WSSV 的序列有同源性，此外定序出多條 cDNA 全長序列，如 *keratinacyte associated protein 2*, *seleno-protein M*, *profilin prohibition* 及 *oncoprotein nm23*。

2008 年，Xiang *et al.* 由雌性中國對蝦 (*F. chinensis*) 的單一成蝦頭胸部製備 cDNA 基因庫，共定序 10,446 條 ESTs(Xiang *et al.*, 2008)，重組成 1,399 條 contigs 與 1,721 條 singletons，以 BLASTN 及 BLASTX 分別在 E value < 10^{-7} 及 E value < 10^{-3} 的條件下，對 NCBI 的 nt 及 nr database 進行同源性搜尋，有 44%的重組序列(contigs + singletons)與已知序列有高度相似性，表現量高的重

組序列其同源基因爲 *peritrophin-like protein I*，爲跨膜蛋白，擁有防止微生物及寄生蟲侵入的功用，*transmembrane-4-superfamily-8*，在細胞黏著、移動、活化及增生上扮演重要角色，*thrombospondin*，參與細胞及細胞間或是細胞與基質間的交流，並在組織生成及修復過程中調控細胞的表型，其他表現量高的基因爲 *elongation factor 1-α*，*tubulin β* 及 *hemocyanin* 等。約有 56%的 assemblies 沒有找到適當對應，作者藉由搜尋 InterPRO database 找尋與其相似的序列，發現表現量最高的重組序列具有 Type I antifreeze protein 的特徵，作者推測這也許是中國對蝦與其他對蝦相比，能夠存活在較低溫環境的原因。同時透過分析 ESTs，也首次在蝦類中發現 *trehalosephosphate synthase gene* 的存在，前人研究的結果顯示 trehalosephosphate 可能在蝦類的免疫系統中扮演重要腳色。

2009 年，Leelatanawit *et al.* 製作草蝦(*P. monodon*)的精巢 cDNA 基因庫，共定序 896 條 ESTs，其中並未發現表現量特別高的 ESTs(Leelatanawit *et al.*, 2009)，但定義出幾條全長的 cDNA，其蛋白質的功能與精巢發育有關，如 *cyclophilin A* (*PMCYA*)，*small ubiquitin-like modifier 1* (*PMSUMO - 1*)，*ubiquitin conjugating enzyme E2*，*dynactin subunit 5*，*cell division cycle 2* (*cdc2*) 以及 *mitotic checkpoint BUB3*。此外也首次在甲殼類中發現參與性別決定過程的基因 *Tra-2* (*PMTra - 2*) 的同源序列；在基因表現的部分，發現 *testis-specific transcript 1* (*PMTST1*) 只會表現在精巢而不會出現在卵巢，*multiple inositol polyphosphate 2* (*MIPP2*) 及 *heat shock-related 70 kDa protein 2* (*HSP70-2*) 在精巢中的表現量高於在卵巢中的表現量，作者推測這些基因可能參與精巢的發育過程。

1.3.2 健康蝦體與被病菌感染蝦體的基因表現之比較

2002 年，Rojtinnakorn *et al.* 使用 ESTs 作爲實驗工具，探討被白點症病

毒感染的斑節蝦 (*M. japonicas*) 的血細胞之基因表現(Rojtinnakorn *et al.*, 2002)。此研究從健康蝦體的血細胞 cDNA 基因庫中得到 635 條 ESTs，從被 WSSV 感染蝦體的血細胞 cDNA 基因庫中得到 370 條 ESTs。在健康蝦體的 635 條 ESTs 中有 284 條與已知的序列有高度相似性，而被感染的蝦體中的 370 條則有 174 條與已知的序列有高度相似性，這些具有已知相近序列的 ESTs 共對應到 152 個蛋白質基因，表現量最高的是粒線體呼吸酵素的蛋白質基因，其次是核醣體蛋白質基因。這些已知蛋白質基因在兩個基因庫中大多都有表現，僅表現量不同，但仍然有一些蛋白質只出現於某一個基因庫中。在這 152 個蛋白質基因中，有 28 個的功能是與防禦相關，其中有 15 個是首次在對蝦類中發現；上述的 28 個蛋白質皆參與 prophenoloxidase (proPO) system 及 clotting process。此外還發現 3 個抗菌胜肽的分子 (bacteinecin-11, penaeidin-2, lysozyme C type) 及 6 個細胞凋亡蛋白及腫瘤相關蛋白，如 β-integrin, cell adhesion molecule 及三型的 collagen (是三個 collagen protein genes)，作者認為這些蛋白質也可能具有防禦病毒感染的功能。

2004 年，Supungul *et al.* 分析來自被哈威弧菌 (*Vibrio harveyi*) 感染的草蝦(*P. monodon*)血細胞 cDNA 基因庫的 447 條 ESTs，且與其在 2002 年自正常血細胞中選出的 615 條 ESTs 做比較(Supungul *et al.*, 2004)。被感染的血細胞基因庫與正常的血細胞基因庫中，表現量最多的都是抗菌肽 (*AMP*，*antimicrobial peptides*)，在正常的血細胞基因庫中量最多的是對蝦素 (*penaeidin*)，其次為 *crustins* 及 *antilipopolysaccharide factor* (*ALF*)，然而，在弧菌感染個體的血細胞基因庫中表現量最多的則是 *antilipopolysaccharide factor* (*ALF*)，其次是 *crustins*，最後是對蝦素，而且其 *ALF* 的表現量高達 50%。根據此結果，作者推測弧菌的感染會導致 *ALF* 被活化導致表現量增加，另外，在被哈威弧菌感染 3 小時後的蝦體，其 *crustins* 及對蝦素的表現量卻呈現下降趨勢，原因可能為產生對蝦素的血細胞經過血流或是脫顆粒作用，將對蝦素釋放到循環系統，

因此導致對蝦素有部分減少的情況。

2007 年，Leu *et al.* 製作健康的草蝦(*P. monodon*)後幼蟲期（postlarval stage PL20）及被白點症病毒感染的草蝦後幼蟲期 cDNA 基因庫，以 ESTs 定序分析觀察後幼蟲期受到白點症病毒感染後的基因表現變化，並發現新的蝦類免疫相關基因(Leu *et al.*, 2007)。由健康的草蝦後幼蟲期的 cDNA 基因庫共得到 7,200 條 ESTs，由被白點症病毒感染的草蝦後幼蟲期基因庫得到 8,064 條 ESTs。將載體序列及品質較差的序列移除後，分別得到 6,658 條健康的草蝦後幼蟲期 ESTs 及被白點症病毒感染的草蝦後幼蟲期 7,276 條 ESTs。爲了進一步增加被註解的 ESTs 的比例，因此隨機選取無註解的 3' ESTs 載體選殖株，進行 5 端定序，得到 978 條來自健康後幼蟲期的 5' ESTs 及 1,069 條來自被白點症病毒感染的後幼蟲期的 5' ESTs，將之與前述的 3' ESTs 合併，最終得到 15,981 條 ESTs，經過序列組裝後得到 1,364 條 contigs 及 8,258 條 singletons。針對各大資料庫以 BLASTX 進行同源性搜尋，在 E value < 10^{-10} 的條件下，有 2,027 條重組序列 (21.07%) 可在 NCBI nr database 中找到有高度相似性的蛋白質基因序列，有 2,026 (21.06%) 重組序列可經由 UniProt database 的同源序列得到 GO 註解。作者藉由比較 ESTs 的分析方法得到以下的結論：（1）在後期幼體時期，白點症病毒的感染會影響許多組織的基因表現，包括肝胰腺、肌肉、蝦體表皮及眼柄等組織；（2）感染的過程中也影響許多基礎細胞代謝的過程，如 oxidative phosphorylation, protein synthesis, the glycolytic pathway 及 calcium ion balance；（3）在白點症病毒感染後，immune-related chitin-binding protein 的基因表現量會顯著上升。

2008 年，Pongsomboon *et al.* 比較健康之草蝦(*P. monodon*)蝦體的淋巴器官及被哈威弧菌 (*V. harveyi*) 感染的蝦體的淋巴器官的基因表現差異性 (Pongsomboon *et al.*, 2008)；淋巴器官在對蝦類中被認爲是參與免疫防禦的重要組織。從健康蝦體淋巴器官的 cDNA 基因庫中得到 408 條 ESTs，自被哈

威弧菌感染的蝦體淋巴器官的 cDNA 基因庫中得到 625 條 ESTs，在兩個基因庫中與免疫相關 ESTs 的數量皆擁有的 15%左右比例，但是所對應到的蛋白質基因組成及數量則不相同；在健康蝦體的淋巴器官中表現量最高的是 *cathepsin L* 及 *cathepsin B*，然而在被哈威弧菌感染的蝦體的淋巴器官中表現量最高的則是 *peritrophin* 及 *thrombospondin*；*cathepsin L* 及 *cathepsin B* 在健康蝦體的淋巴器官及感染蝦體的淋巴器官中皆有表現，但 *peritrophin* 及 *thrombospondin* 則只出現於被感染蝦體的淋巴器官中。作者藉由 real time PCR 定量分析，發現當蝦體被哈威弧菌或白點症病毒感染時，*cathepsin L* 及 *cathepsin B* 的表現量只有些許的改變，但 *peritrophin* 及 *thrombospondin* 的表現則有明顯的上升趨勢，因此作者認爲在草蝦的淋巴器官中，*cathepsin L* 及 *cathepsin B* 兩個基因在蝦體內爲常態性高量表現的基因，也許與淋巴器官的正常生理功能有關，*peritrophin* 及 *thrombospondin* 兩基因的表現量則會受到病原菌感染的刺激而上升，推測這兩個基因在淋巴器官的免疫防禦功能方面扮演重要的角色。

1.3.3 不同組織間的基因表現

1999 年 Lehnert *et al.* 挑選草蝦(*P. monodon*)的頭胸部、眼柄以及泳足分別製作三個 cDNA 基因庫，定序出 176 條 ESTs(Lehnert *et al.*, 1999)。比較三個基因庫，發現在泳足的 cDNA 基因庫相對於頭胸部及眼柄兩個 cDNA 基因庫有較高量的粒線體相關序列，作者認爲這應該是與泳足是運動的器官，需要較多的能量有關；而頭胸部是個擁有許多器官的地方，因此其 ESTs 的種類較多樣化，在這之中，以血藍蛋白 (hemocyanin) 基因相關序列最多；草蝦的眼柄是視覺器官，同時也是重要的神經內分泌組織所在部位，在這個基因庫中則有較多是參與視覺及神經內分泌系統的基因。

2001 年，Gross *et al.* 針對白蝦 (*L. vannamei*)與白濱對蝦 (*Litopenaeus*

setiferus) 兩物種做了研究(Gross *et al.*, 2001)，此篇研究是以兩物種各自的血細胞 (hemocyte) 以及肝胰臟(hepatopancreas) 製作 cDNA 基因庫，比較其與免疫相關的基因；此篇研究中四個基因庫共得到 2, 045 條 ESTs，結果顯示免疫相關基因在兩物種中的表現量皆爲血細胞高於肝胰臟，其他與免疫相關的物質如：抗菌肽(*AMP*，*antimicrobial peptides*)，其在血細胞中的表現量亦是最高的；然而凝集素(*Lectin*)則是只表現於肝胰臟中。

2006 年，O' Leary *et al.* 收集了白蝦(*L. vannamei*)的血細胞、肝胰臟、鰓、淋巴器官、眼柄以及腹神經索等六種組織的 ESTs，共定序了 13,656 條 ESTs，藉由 EST 彼此間重疊的序列，以 cap3 assembler 將 EST 組成更長的 contigs，共得到 7,466 條重組序列(O'Leary *et al.*, 2006)。此篇研究著重於免疫相關基因的發現及分佈，另外也藉由 Gene Ontology 的註解，得到基因功能的比率分佈及基因參與的生理途徑等資訊。2007 年時，同實驗團隊又發表了另一篇的研究(Robalino *et al.*, 2007)；作者選擇被白點症病毒感染的鰓部組織的 cDNA 基因庫進行 ESTs 定序分析，共有 872 個 clone 被定序，但最後只留下 601 條品質好的 ESTs，組裝成 276 條重組序列。作者除了得到基因的分佈之外，同時也藉由RT-PCR得知下列基因的表現量會受到WSSV感染的影響而有所改變：*keratinocyte associated protein 2* (*KCP2*)，其功能尚未被了解，但作者根據實驗結果推測 KCP2 可能參與蝦體被病毒感染後的反應過程、*selenoprotein M* (*SelM*)，白點症病毒的感染導致氧化壓力產生，SelM 可能參與減輕此氧化壓力的過程、*prohibitin* (*Phb*)，擁有抑制細胞增生的功能、*profilin* 及 *oncoprotein nm23* 等五個基因。

2006 年，Tassanakajon *et al.* 希望發現組織特異表現的基因，因此以健康的草蝦(*P. monodon*)爲實驗材料，選取六個組織：眼柄、肝胰腺、造血組織、血細胞、淋巴組織及卵巢，進行 ESTs 定序研究(Tassanakajon *et al.*, 2006)。將載體的序列及品質差的序列移除後共得到 10,100 條 ESTs。在六個組織的所有

ESTs 中，表現量最高的 ESTs 是粒線體相關的基因序列，其餘如 *thromobospondin*, *elongation factor I-*α, *ovarian peritrophin*, *ALF*。另外，依據 ESTs 對應結果，有一些目前註解爲 hypothetical proteins 且功能不明的基因，也具有相當高量的表現，值得進一步以序列比對或是蛋白模組分析來探討其可能的功能。若以有註解的序列爲主，根據功能將之分類，發現只有 3.9%的 ESTs 是與防禦功能或是生理平衡功能相關，6.2%與生理代謝有關，5.5%則與基因表現、調節及蛋白質生合成相關。將六個組織的 cDNA 基因庫相互比較，發現某些基因有組織特異性，如褪殼抑制激素（MIH）及色素擴張激素（PDH）只出現眼柄基因庫中，*hemocyanin* 只出現在肝胰腺 library 等，此外，也第一次在甲殼類中發現 α*- NAC protein* 及 *dystrobrevin-like protein* 基因，爲參與血球細胞分化過程的基因。同時，爲了了解蝦體處在緊迫狀況時其相關的基因表現，作者也一併製作血細胞在四種不同緊迫環境時的 cDNA 基因庫，四種緊迫環境分別是（1）將蝦隻養殖在 35℃的水中一個小時、（2）被白點症病毒感染的蝦隻、（3）被黃頭病毒 (Yellowhead virus，YHV) 感染的蝦隻及（4）被哈威弧菌感染的蝦隻；在這四個基因庫中，核醣體蛋白質相關的 ESTs 在健康蝦體的血細胞基因庫中數量約佔有 8.2%，但在被白點症病毒感染的血細胞基因庫中則佔有 23.9%，被哈威弧菌感染的蝦體的血細胞基因庫中也有 17.6%；也就是說，相較於健康蝦體的血細胞，被白點症病毒及哈威弧菌感染的血細胞，其核醣體蛋白質的表現量增加了，與 2002 年 Rojtinnakorn *et al.* (Rojtinnakorn *et al.*, 2002)的研究有相同的結果。在這四個基因庫裡，與免疫相關基因的 ESTs 總數量並未有很大的差異性，但是其對應的基因與個別基因相關 ESTs 的數量卻有所變化，如抗微生物相關的分子在健康蝦體的血細胞中佔有 3%的比例，但在被白點症病毒感染的蝦體的血細胞中卻有 6.3%，被哈威弧菌感染的蝦體的血細胞中則有 7.2%。

2011 年，Jung *et al.* 利用 454 pyrosequencing 的次世代定序技術，研究淡

水長臂大蝦 (*Macrobrachium rosenbergii*) 的轉錄基因體，及定義與成長相關的基因(Jung *et al.*, 2011)。此研究製作雌蝦的肌肉及卵巢組織，和雄蝦的精巢及眼柄組織的 cDNA 基因庫，共得到 787,731 條短序列片段，利用 Newbler 2.5.3 (Roche) 將短序列片段進行 *de novo* assembly，得到 8,411 條 contigs 及 115,123 條 singletons，藉由序列同源性比對的方式尋找與重組序列擁有高度序列相似性的已知序列，3,757 條 contigs 及 21,965 條 singletons 與 NCBI nr database 中的已知序列有高度相似性。利用 Gene Ontology (GO) 及 Kyoto Encyclopedia of Genes and Genomes (KEGG) 分類有註解的 assemblies，在 GO molecular function 的分類下，許多重組序列表現 binding 或 catalytic function 的功能，在 GO biological process 的分類下，參與 cellular process 及 metabolic process 的重組序列占多數，此分析結果與前人對於其他甲殼類的研究相同 (Leekitcharoenphon *et al.*, 2010,Leelatanawit *et al.*, 2009,Leu *et al.*, 2011,Tassanakajon *et al.*, 2006)。經由 KEGG analysis 則發現重組序列參與代謝及肌肉收縮的生理途徑，如 metabolic pathways、oxidative phosphorylation 及 biosynthesis of alkaloids derived from histidine and purine 等。接著利用 InterProScan 搜尋重組序列是否具有蛋白質的功能模組(domain)，多數重組序列擁有的功能模組特徵為 RNA recognition motif、actin-like families 及 zinc finger domain 等。檢查轉錄基因體中，其分子功能可能與生長相關的重組序列有 cyclophilin、LIM domain protein 及 farnesoic acid O-methlytransferase (FAMeT) 等基因。

因此，本論文的目標，是應用目前新發展的核酸序列分析工具與序列組裝方法，以相近物種的基因體與轉錄體序列資料為參照，由白蝦 ESTs 定序資料建構轉錄體，並以解析其基因表現概況。

第二章 材料方法

2.1 材料介紹

2.1.1 白蝦(*Litopenaeus vannamei*)的 EST (Expressed Sequence Tags)

用於實驗的白蝦 (*L. vannamei*) 序列，皆取自 NCBI dbESTs (Database of Expressed Sequence Tags)。白蝦 ESTs 序列檔案 (檔案建立日期:2011/3/17) ，共包含 161,241 條，來自 84 個 cDNA 基因庫，序列長度的分布範圍爲 19bps ~ 2,143bps，平均長度爲 494bps (表格 2-2)。此十六萬條 ESTs 即爲本研究的起始資料。白蝦 ESTs 序列主要來自於六個 cDNA 基因庫，分別來自健康成蝦的眼柄、鰓、血細胞、肝胰腺、淋巴器官及神經索等六種組織，（表格 2-1，(O'Leary *et al.*, 2006))。

表 2-1 白蝦 EST 序列的主要 cDNA 基因庫概況，數據來源: NCBI Unigene Library Browser, http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=6689&CUTOFF=1

Library ID*	***Library Description***	***ESTs #***
Lib.22684	*Litopenaeus vannamei* eyestalk cDNA library	29,575
Lib.22686	*Litopenaeus vannamei* hemocyte cDNA library	27,369
Lib.22685	*Litopenaeus vannamei* gills cDNA library	24,296
Lib.22688	*Litopenaeus vannamei* lymphoid organ cDNA library	24,214
Lib.22687	*Litopenaeus vannamei* hepatopancreas cDNA library	22,272
Lib.22689	*Litopenaeus vannamei* nerve cord cDNA library	20,179
	Total	**147,905**

註：＊：Library ID 是指一批 ESTs 被提交至 dbEST 時，由 NCBI 給定的 ID。

表 2-2 白蝦 EST 序列的資料概述

	Numbers	Min. length(bps)	Max. length(bps)	Ave. length(bps)
ESTs of ***Litopenaeus vannamei***	161,241	19	2,143	494

2.1.2 水蚤轉錄基因體

水蚤 (*Daphnia pulex*) 是一種常見於淡水池塘的小型浮游生物，屬甲殼綱（Crustacea）橈腳目（Copepoda），水蚤是淡水生態系統中的重要物種之一，擁有以下特性:生活史短，族群數龐大，在田野及實驗室中都易於繁殖生長，在食物網的位置爲藻類的初級消費者，魚類的基礎飼料，同時對環境毒物的敏感度高，且能根據所處環境不同發展出不同表型的特性，所以常被用於研究生態學及演化學的研究；新近被 NCBI 列入「生物醫學研究模式物種」。轉錄基因體是以次世代定序法（next generation sequencing technology, NGS），取得大量短序列後經序列組裝流程而產生，是甲殼綱中第一個擁有基因體序列的物種(Colbourne *et al.*, 2011)。水蚤的轉錄基因體檔案版本爲 v1.1 (檔案建立日期:2011/4/13) ，共包含 30,907 個水蚤基因，基因長度的分布範圍爲 150bps ~ 24,144bps，平均長度爲 1,061bps (表格 2-3)。

表 2-3 水蚤轉錄基因體的序列資料概述

	Numbers	Min. length(bps)	Max. length(bps)	Ave. length(bps)
Genes of ***Daphnia pulex***	30,907	150	24,144	1,061

2.1.3 果蠅蛋白質體

果蠅 (*Drosophila melanogaster*) 自十九世紀初，開始成爲實驗研究物種。果蠅的全基因體定序在 2000 年完成(Adams, 2000)，在所有完成全基因體定序的模式物種中，果蠅與白蝦的親緣性僅次於水蚤，但是果蠅經過長年的研究，研究者透過果蠅了解許多生物體系的運作與調控。FlyBase (McQuilton *et al.*, 2012)是果蠅研究的整合平台計畫，資料詳盡，本研究所使用的果蠅全蛋白質序列是來自 FlyBase 的資料庫，取得的檔案版本爲 5.38 (檔案建立日期:2011/6/29) ，共包含 23,711 條蛋白質序列，序列長度的分布範圍爲 11 a.a ~ 22,971 a.a，平均長度爲 632 a.a (表 2-4)。

表 2-4 果蠅蛋白質序列的資料概述

	Numbers	Min. length(a.a)	Max. length(a.a)	Ave. legnth(a.a)
Proteins of *Drosophila melanogaster*	23,711	11	22,971	632

2.1.4 重組序列之策略及工具

ESTs 通常會經過序列組裝的流程以期重建原始的轉錄基因序列，因白蝦還未擁有完整的基因體序列，因此必須採取 *de novo* assembly 的方式，重新組裝；所謂的 *de novo* assembly，乃是利用不同的 ESTs 擁有相同的序列片段作爲依據，進而將 ESTs 根據重疊的部分合併連結，獲得一條較長的序列。MIRA 是眾多可執行 *de novo* assembly 的程式之一(Chevreux *et al.*, 2004)，其餘常用的 assembler，還有 CAP3、PHRAP 及 NEWBLER 等，MIRA 適用於由 Sanger、454 或 Solexa 等定序技術所產生的序列片段，其進行序列組裝的策略是 iterative multipass strategies，序列一開始以嚴苛的參數進行組裝，得到由軟體編輯過的結果，之後將結果以較不嚴苛的參數再次進行組裝，重複此過程直到組裝完成，或是參數過於寬鬆。此外，MIRA 以

序列的 high confidence region (HCR)爲主體，序列依照彼此 HCR 擁有相同或重疊的區間，進行合併延伸，減少因序列兩端的錯誤讀取，導致組裝錯誤的產生。由於 MIRA 擁有上述特性，且程式有持續更新，因此我們選擇 MIRA 爲白蝦 ESTs 進行組裝，其使用的參數如表 2-5:

表 2-5 MIRA assembler 進行 *de novo* assembly 時所使用的參數

--job=denovo,est,normal,sanger	(進行標準的 EST 組裝流程)
COMMON_SETTINGS	
-SK:mmhr=20 -SK:mnr=yes -SK:not=8	(使用 8 顆核心並遮蔽重複序列)
-CO:asir=yes	(使用較低的 gap 扣分標準，組裝 contigs)
-OUT:orc=no:org=no:ora=no:ors=no:rtd=yes	(輸出的格式設定)
SANGER_SETTINGS	
-CL:qc=no:cpat=no	(不經過 guality trimming)
-AL:egp=no	(不使用額外的 gap 處罰)
-LR:wqf=no -AS:epoq=no	(略過 quality file)
-OUT:stsip=no	(輸出設定)

爲增加重組序列的準確性及減少重組序列的重複性，我們將 Jain *et al.* 提出的概念應用於序列組裝的策略中(Jain *et al.*, 2007)；Jain *et al.*藉由物種的 draft genome 提供完整、未被序列間隔 (sequence gaps) 打斷的基因序列，將 ESTs 與 draft genome 進行序列比對，得到 ESTs 最適對應的基因序列，依據對應的基因將 ESTs 進行分組，之後將同一組內的 ESTs 以 *de novo* assembly 的方式進行組裝。因此，我們利用相近物種的 draft genome 作爲參考基因體，將白蝦 ESTs 與參考基因體進行序列比對，藉由對應的基因將白蝦 ESTs 進行分組，之後在組內進行 *de novo* assembly。

2.2 白蝦 ESTs (Expressed Sequence Tags) 組裝流程

到目前爲止，白蝦的全基因體序列並未有初步的解析，爲了更加瞭解白蝦基因表現概況與特定組織中的高表現基因群，我們必須將原本分散的ESTs加以重組。在此，希望透過親緣相近的全定序物種之基因體爲參考母體，來增加重組的可靠性與正確度。因此，本研究選用目前唯一已完成基因體定序的甲殼類物種－水蚤(*D. pulex*) 爲參考；將白蝦(*L. vannamei*) 的 ESTs 對水蚤的轉錄基因體進行序列比對 (使用參數:BLASTX，E-05)，再依照與水蚤基因的相近程度，將對應到同一個水蚤基因的白蝦 ESTs 歸類爲同組，透過 MIRA assembler 來進行組裝。無法在此階段與其他序列合併組裝的 ESTs，則與親緣較遠的果蠅 (*D. melanogaster*) 蛋白質體進行序列相似性比對 (使用參數:BLASTX，E-05)，同樣地將對應到同一果蠅蛋白質的白蝦 ESTs 歸類爲同組，再藉由 MIRA assembler 的協助，將之組裝成 contigs。最後，收集餘下未組裝成 contigs 的 ESTs，以 MIRA 進行 *de novo* assembly 的步驟 (使用參數如表 2-5)。最後，將這些完成組裝的 assemblies (contigs + singletons)，以序列相似性進行比對，找出白蝦重組序列在 nr database 及 Pfam database 中高度相似性的同源序列，援用最佳對應的序列之敘述與其功能註解，做爲白蝦重組序列的註解(圖 2-1)。

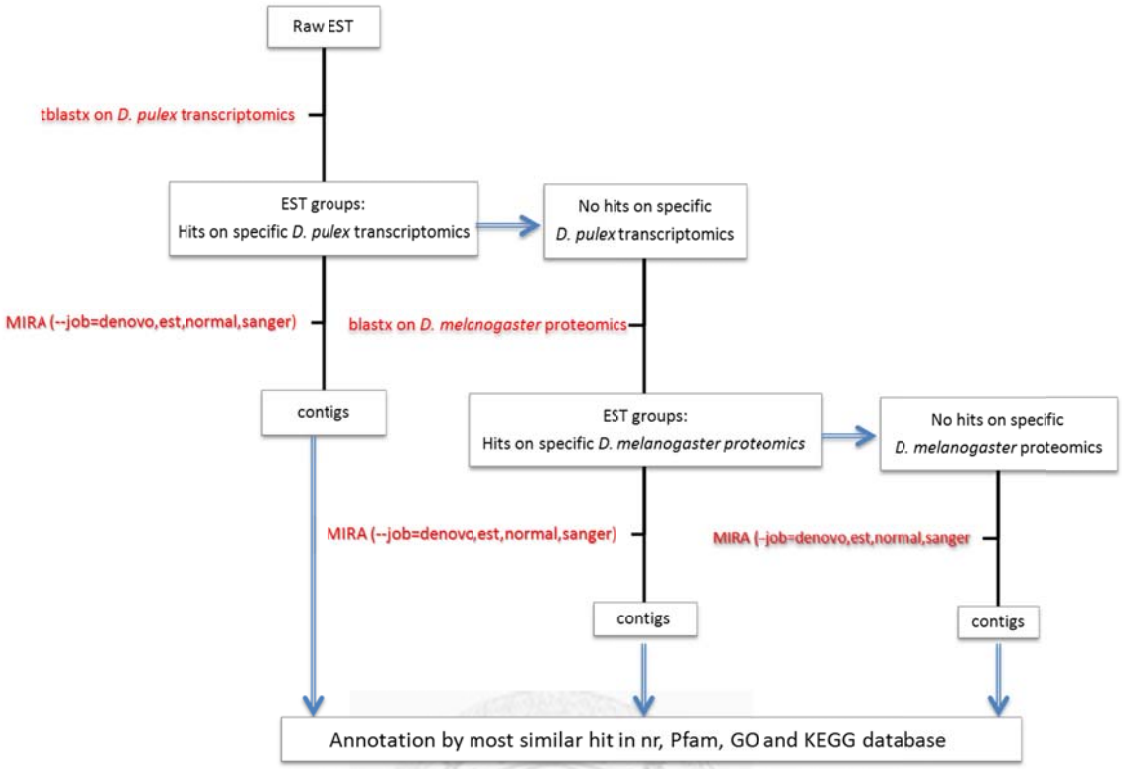


圖 2-1 白蝦 ESTs 組裝流程圖

2.2.1 與水蚤(*Daphnia pulex*)的序列相似性比較

使用 tblastx 將每一筆白蝦 ESTs 及水蚤轉錄基因體其各自 6 種可能的 ORF (Open Reading Frame) 互相比對，以 E-value < 10^{-5} 定義序列之間具有相似性。水蚤及白蝦同樣分屬甲殼綱，我們認爲兩物種間應有同源蛋白質的存在，同時考量到種間的差異，因此使用水蚤所有可能的轉譯序列爲參照。根據 tblastx 比對的結果，挑選 alignment 長度有達到 50 個胺基酸以上的 ESTs，認爲在此長度中至少包含一個 ORF，依據高度相似性的水蚤轉譯序列分群，將兩個 ESTs 以上的類群重組爲 contigs；在上述流程中不符合條件的 ESTs 都將歸於 Singleton I。詳細流程如圖 2-2 所示。

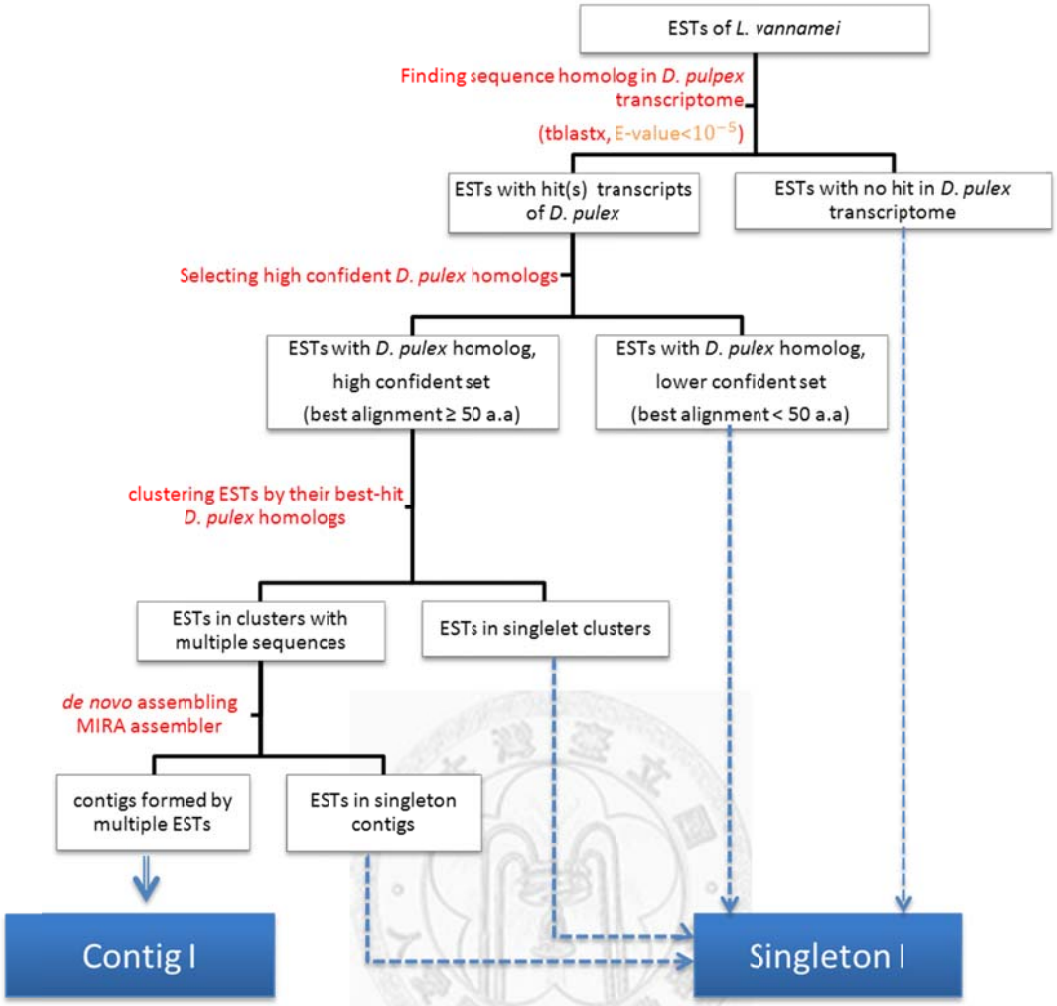


圖 2 -2 依水蚤轉錄基因體爲參照組裝白蝦 ESTs

2.2.2 以果蠅(*Drosophila melanogaster*) 的序列輔助建立

所有的白蝦 ESTs 經過利用水蚤序列的組裝流程（圖 2-2）後，未被納入 contigs 的序列歸入 Singleton I。爲了將更多的 EST 納入組裝 contig，我們利用 blastx 程式將 Singleton I 的序列與果蠅的蛋白質序列比對，依照相同的原則，也就是說，序列相近程度達到 E-value < 10^{-5} 且 alignment 長度達到 50 個胺基酸長度的 ESTs，依據其所對應的果蠅蛋白質歸類，對於所有具有兩筆白蝦序列以上的序列類群，透過 MIRA 程式組出 contig；在這過程中，不符合序列選取資格與無法形成 contig 的白蝦 ESTs 皆被納入 Singleton II 中（圖 2-3)。

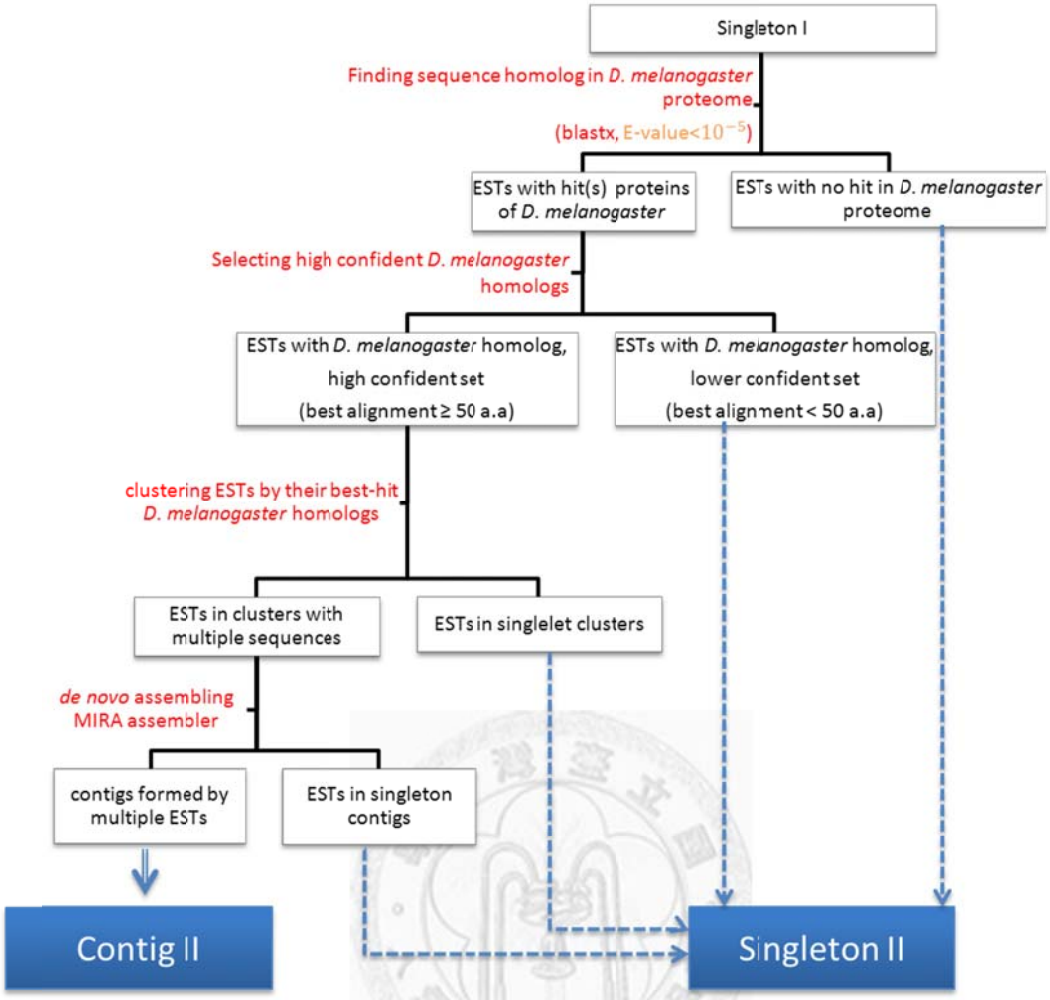


圖 2 -3 依果蠅蛋白質體爲參照組裝白蝦 ESTs

2.2.3 *De novo* Assembly

經過先前與水蚤轉錄基因體以及果蠅轉譯蛋白質體的相似性比對與組裝步驟之後，未能形成 contigs 的白蝦 ESTs 總集爲 Singleton II，直接以 MIRA 程式進行 *de novo* assembly 的步驟，得到 contigs 及無法被組裝成 contigs 的 ESTs，即爲 Singleton III (圖 2-4)。

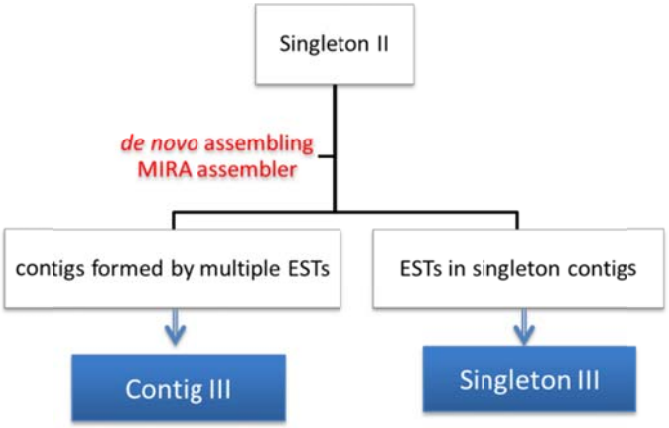


圖 2 -4 以 *de novo* assembly 組裝白蝦 ESTs

2.3 對重組序列進行可能的功能註解

2.3.1 重組序列與 nr 及 Pfam 兩資料庫的相似性比對

經過上述步驟後，可得到下列四群資料：

(1) 與水蚤轉錄基因體進行相似性序列比對後經過重組得到的 contigs，簡稱爲 Contig I。

(2) 與果蠅蛋白質體進行相似性序列比對後經過重組得到的 contigs，爲 Contig II。

(3) 經過 *de novo* assembly 組裝出的 contigs，爲 Contig III。

(4) 無法與任何 ESTs 合併組裝成 contig 的 Singleton，Singleton III。

將這四群資料整合後，即爲本研究所歸結的白蝦轉錄體組裝結果。藉由與已知功能的序列相互比對，可推測重組序列的作用；利用 blastx 程式將白蝦重組序列與 nr database 中的序列比對，以 E-value $< 10^{-5}$ 定義序列相近程度；接著再以 rpsblast 程式與 Pfam database 的序列比對，以 E-value $< 10^{-3}$ 定義相近程度，若有多條序列符合此條件，則選取 E-value 最小的序列做爲最適結果，借用其註解資料爲重組序列之可能註解。Pfam database 是一個蒐集蛋白質家族的資料庫 (Punta

et al., 2012)，蛋白質的功能乃由其所含有的功能模組 (domain) 所決定，一個有功能的蛋白質最少包含一個 domain，不同的 domain 組合使得蛋白質有不一樣的功能。如能了解序列中是否有 domain 的存在，可以幫助我們探知序列的可能作用。所以，藉由 Pfam database 的比對，可讓更多的重組序列擁有註解，而無法透過 nr 註解的 contigs，得以擁有註解，有助於後續的功能基因體分析。

2.3.2 進行重組序列的 Gene Ontology(GO) 註解

Gene Ontology Database 是 Gene Ontology Project 的數據資料庫，Gene Ontology Project 的目的是建立一套有系統的語言，使各種資料庫對於基因產物的功能描述達到一致，減少研究者在搜尋上所耗費的時間，此計畫始於 1998 年，現今已包含數十個重要的資料庫(Ashburner *et al.*, 2000)。Gene Ontology 涵蓋了三個領域，分別是：cellular component，描述細胞的每個部份或胞外環境，molecular function，說明基因的分子功能，biological process，描述由一個以上的 molecular function 所完成的事件。此類型的資料庫的結構是一種 directed acyclic graph(DAG，有向無環圖)，也就是說從某一個節點(Node)出發之後，雖然會經過不同的邊(Edge)，但是無法回到原來的節點，在圖論(Topology)的領域裡，我們稱爲有向無環圖。所以，在 GO 的資料庫中，一個基因產物可能擁有來自同一領域的不同定義，也可能在不同的領域中擁有各自的定義，且越深一層的定義比上一層更加詳細精確。所以，在此我們利用 GO 的資料庫，來對白蝦的重組序列進行功能分類。透過先前所得到的 Pfam ID，再利用 Gene Ontology Database 中提供的 pfam2go 對應表(表 2-9)，我們就能得到在 Gene Ontology Database 中相對應的 GO 註解。

2.3.3 以定序方法研究白蝦基因表現概況（Expression Profiling）

如本章 2.2.1 所述，白蝦 ESTs 資料中包含六個定序數量超過 20,000 條 ESTs 的 libraries，分別來自六個組織：眼柄、鰓、血細胞、肝胰腺、淋巴器官及神經索。萃取上述六個組織的 RNA，移除基因體 DNA，製備各自的 cDNA 基因庫，透過

PCR Select cDNA Subtractive Hybridization Kit 得到 subtracted libraries，在六個基因庫中隨機選取 clones 進行定序得到 ESTs(O'Leary *et al.*, 2006)。本研究選定這六個基因庫，以其 EST 序列組成與表現差異分析，藉以反映各組織的功能特化。

2.3.4 各組織的基因表現概況

重組序列是我們利用 ESTs 所重建的可能基因構造，因此，組成重組序列的 ESTs 數量，與真正的基因表現量有高度的正相關性，根據重組序列的組成 ESTs 及基因庫中包含的 ESTs 數，可得到重組序列在不同基因庫中所擁有的 ESTs 數量。利用 Gene Ontology 的項目將基因庫中的重組序列進行分類，計算每個 GO 項目中其重組序列的 ESTs 數，之後將 ESTs 數除以此基因庫中擁有 GO 註解的總 ESTs 數，得到該 GO 項目在基因庫中的表現比例，藉由 GO 的項目及其表現比例，顯示不同基因庫(不同組織)的基因表現概況。依照 GO 包含的三個領域分別分析各組織的基因表現概況。

2.4 組織間表現基因的差異性

2.4.1 Digital Differential Display

爲了探討組織間的差異性，我們參考 NCBI 中 Digital Differential Display (DDD) 的作法。DDD 是一個可以比對多個基因庫中，其特定基因的表現是否在不同組織內，其整體 EST 的表現量是否有所差異的生物資訊工具。爲了消除不同基因庫內含不同量的 EST 所造成的差異，要先對所有的基因庫進行 TPM 的轉換，之後再利用 Fisher’s exact test，來檢定得出的結果是否有其統計上的顯著差異(Pontius *et al.*, 2002)，以眼柄組織爲例，將眼柄組織的 contigs 與鰓、血細胞、肝胰腺、淋巴組織及神經索的 contigs 表現量計算後，進行 Fisher’s exact test 檢定，以 p-value 來篩選，來找出在眼柄組織中的表現量與在其餘五個組織中的表現量有明顯差異的基因群。Fisher’s exact test 的計算方式如下所示(表 2-6)：

表 2-6 Digital Differential Dispaly 之 2*2 列聯表

	Selected library	Others library	Total
TPM value of selected contigs	A	B	A+B
TPM value of non-selected contigs	C	D	C+D
Total	A+C	B+D	A+B+C+D (=N)

p-value 的計算方程式如下:

$$p = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{A!\,B!\,C!\,D!\,N!} \quad (2)$$

在此選擇 p-value < 10^{-3} 的所有 contigs，就會得到特定組織中與其他五個組織表現有明顯差異的 contigs 列表。

2.4.2 以 Gene Ontology 來進行功能性分析

完成 DDD 的分析之後，將可篩選出特定表現模式的基因群，爲了更進一步瞭解這些基因群所隱喻的生物意義，我們將以 Gene Ontology 來進行細部的功能性分析。利用統計的方法檢定 GO 項目，由統計檢定值判斷 GO 項目是否具有統計上的意義，選擇其顯著差異性最高的項目描述基因群的生物功能，此過程稱爲 enrichment analysis (Stojmirovic and Yu, 2010)。本篇實驗選用 Fisher's exact test 進行檢定，將基因依其相關聯的生物特性的名稱(GO 項目)分類，之後藉由與背景值的比較，得到此項名稱的基因是否有較高量的表現(Huang da *et al.*, 2009)。以眼柄組織爲例，將眼柄組織與其他五個組織有明顯差異表現的 contigs 列表依據在 Gene Ontology Database 中所對應的類別歸類，歸類後的每個 GO 項目皆進行 Fisher's exact test，其 p-value 代表 GO 項目在眼柄組織的表現與在所有組織中的表現是否有明顯差異，Fisher's exact test 的計算方式如下(表 2-7):

表 2-7 GO enrichment 之 2*2 列聯表

	The selected library	Complement of the selected library	Total
The EST numbers of selected GO term	a	b	a+b
The EST numbers of complement of the selected GO term	c	d	c+d
Total number	a+c	b+d	a+b+c+d(=n)

p-vlaue 的計算方程式如下:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!} \quad (3)$$

在此選擇滿足 p-value < 10^{-3} 的 GO 項目，將其餘五個組織的 contigs 列表皆進行上述步驟，得到六個組織與其他組織表現有差異的 contigs 在 Gene Ontology Database 中的分布趨向。

2.4.3 Venn diagram & KEGG pathway enrichment analysis

Venn diagram，是一種展現不同集合之間的數學或邏輯上的圖像化呈現手法，在此，我們利用此一方式，來表現兩個基因庫之間的異同處。Kyoto Encyclopedia of Genes and Genomes (KEGG) database 是一個利用電腦呈現生物系統的大型資料庫，資料庫中提供基因體資訊、生化物質的結構與反應過程以及生物調控網路等資料，代謝、膜運輸、訊號傳導及細胞週期等都是一種生物調控網路，其中 PATHWAY database 即是提供圖像化生物調控網路的資料庫(Kanehisa and Goto, 2000)。選取欲進一步做分析的兩個 library，利用 Venn diagram 決定其交集及不重疊的 assemblies，接著以 blastx 程式對 KEGG PATHWAY database 進行序列相似性比對，以

E-value<10^{-3} 的標準，來找出與特定 KEGG ID 的對應關係，進而得知特定 contigs 所參與的生理代謝途徑，隨後進行 enrichment analysis，其計算方式如表 2-8 所示：

表 2-8 KEGG enrichment 之 2*2 列聯表

	The selected sets	Complement of the selected sets	Total
The contig numbers of selected pathway	α	β	$\alpha+\beta$
The contig numbers of complement of the selected pathway	γ	δ	$\gamma+\delta$
Total number	$\alpha+\gamma$	$\beta+\delta$	$\alpha+\beta+\gamma+\delta$ (=k)

p-value 的計算方程式如下:

$$p = \frac{(\alpha+\beta)!\,(\gamma+\delta)!\,(\alpha+\gamma)!\,(\beta+\delta)!}{\alpha!\,\beta!\,\gamma!\,\delta!} \quad (4)$$

選擇 p-value < 10^{-3} 的 pathway name；將其餘兩個集合的 assemblies 列表皆進行上述步驟，得到三個集合在 KEGG PATHWAY database 註解下的趨向。

2.5 建置白蝦 EST 資料庫

白蝦轉錄體資料庫 (http://ips.iis.sinica.edu.tw/lv) ，乃建構在雙顆四核心(2.5GHZ Intel Xeon)的伺服器主機上，配備有 20GB 的記憶體與 1T 的硬碟空間。使用的作業系統是 Linux Ubuntu 8.04，所使用網頁伺服器爲 Apache 2.2，關連性資料庫爲 Postgresql 8.3，並使用 Python 2.5 來撰寫網頁介面與協助結果呈現。

2.6 資料庫及程式清單

本研究所引用之資料來源、資料庫及應用程式網站如下表所示(表 2-9):

表 2-9 資料來源、資料庫及應用程式網站

檔案	版本 下載位址
白蝦 EST 檔案 Date: 2011/3/17	http://www.ncbi.nlm.nih.gov/nucest/
水蚤轉錄基因體	FilteredModelsv1.1.na.fasta.gz http://genome.jgi.doe.gov/
果蠅蛋白質體	dmel-all-translation-r5.38 ftp://ftp.flybase.net/releases/
NCBI nr Date:2011/6/11	ftp://ftp.ncbi.nih.gov/blast/db/
Pfam database Version:25	25 ftp://ftp.sanger.ac.uk/pub/databases/Pfam
pfam2go mapping file Date:2011/10/5	http://www.geneontology.org/GO.indices.shtml
Gene Ontology database	http://www.geneontology.org/
KEGG database Date:2011/6/18	http://www.kegg.jp/kegg/download/
BLAST 程式 Version:2.2.17	ftp://ftp.ncbi.nih.gov/blast/
MIRA assembler Version:3.2.1	http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title=Main_Page
Apache 2.2	http://www.apache.org/
postgreSQL 8.3	http://www.postgresql.org/
Python 2.5	http://www.python.org/

第三章 結果

3.1 白蝦的重組序列

根據重組序列流程，如 2.3 章節所述，我們共得到 4 組資料，(1) 與水蚤轉錄基因體進行相似性序列比對後經過重組得到的 contigs，此組資料稱爲 Contig I，共 3,361 條、(2) 與果蠅蛋白質體進行相似性序列比對後經過重組得到的 contigs，此組資料稱爲 Contig II，共 920 條、(3) 經過 *de novo* assembly 組裝出的 contigs，此組資料稱爲 Contig III，共 12,605 條以及 (4) 無法與任何 ESTs 合併組裝成 contig 的 Singleton，此組稱爲 Singleton III，共 20,515 條；4 組資料其包含的 ESTs 所佔有的比例如圖 3-1 所示；可看出 Contig III 包含的 ESTs 數目最多，約是全部 ESTs 數 161,241 的一半，其次是 Contig I，其包含的 ESTs 數約有 1/3 左右。

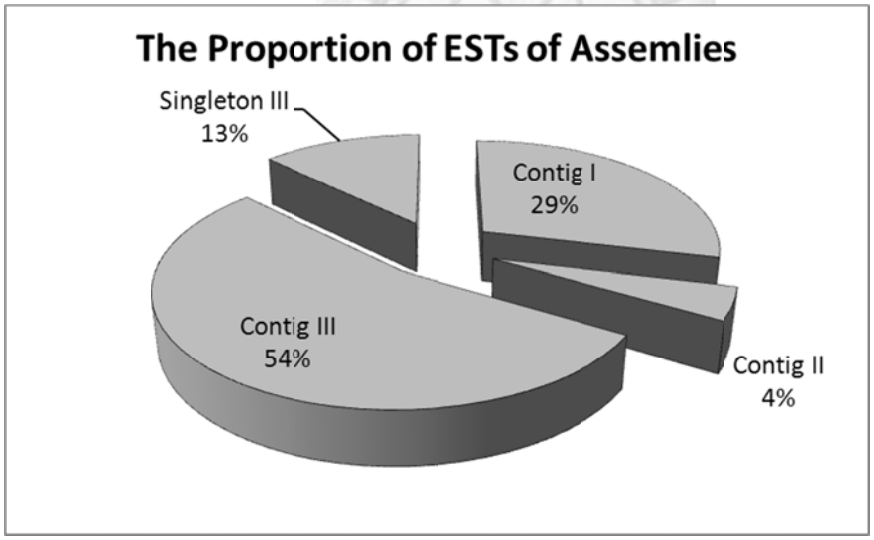


圖 3 -1 重組序列中的 ESTs 數其分佈比例

重組序列與原始的 ESTs 資料的數量、平均長度、最小及最大長度其資訊如表格 3-1 所示，經過組裝後的序列其平均長度皆有 600 bp 以上，最小長度也有 80 bp 以上；Contig III 的資料中最長的序列高達 4,501 bp，爲所有資料中最高；此外也

可發現 ESTs 最短及最長的序列長度與 Singleton III 的相同，ESTs 中最短及最長的序列經過我們的組裝流程後，皆被歸於 Singleton。

表 3-1 EST 及重組序列的資料描述

	number	Avg. length	Min. length	Max. length
ESTs	161,241	494	19	2,143
Contig I	3,361	839	83	2,789
Contig II	920	712	99	2,199
Contig III	12,605	635	80	4,501
Singleton III	20,515	400	19	2,143

爲了了解四組資料其序列長度的分布，將三組 contig 資料及 singleton 資料分開作圖，如圖 3-2 所示，contig 的序列長度分布由 Contig I、Contig II 及 Contig III 三組資料決定，數據以堆疊直條圖呈現；singleton 的序列長度分布由 Singleton III 決定，數據以直條圖呈現，如圖 3.2 的內嵌圖。Contig I 的序列集中在 500～1200 bps 的範圍，Contig II 的序列集中在 600～800 bps 的範圍，Contig III 的序列集中在 200～1100 bps 的範圍，Contig II 的範圍最集中但其所占的比例最少；整體 contig 的序列集中在 300～900 bps，以 800 bps 的序列最多。Singleton III 的序列集中於 200～800 bps，以 200 bps 的序列最多。

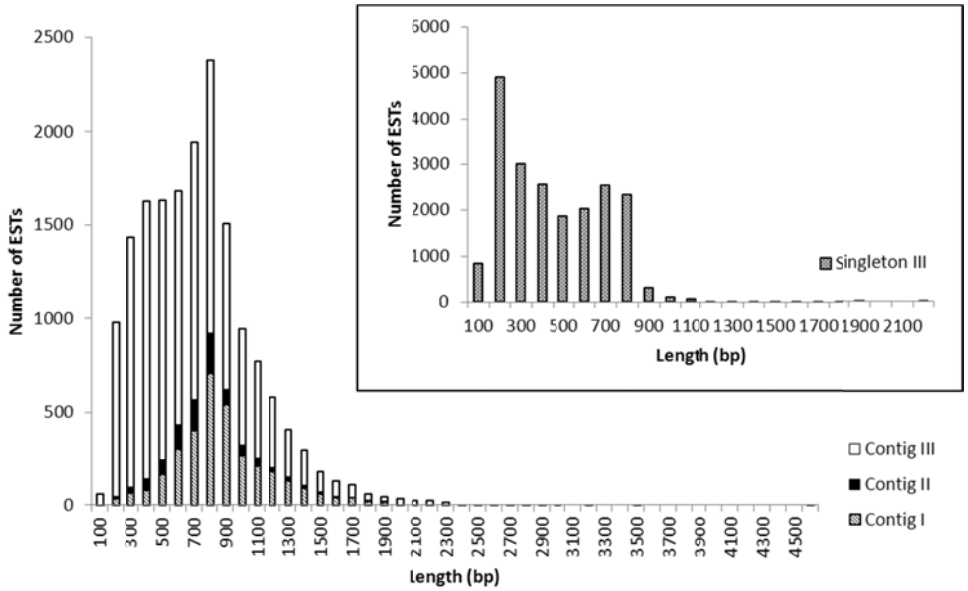


圖 3 -2 Contig I、Contig II 、Contig III 及 Singleton III 等四組重組序列的長度分布圖

3.2 註解重組序列

由 EST 序列衍生的重組序列其功能未明，本研究利用 blastx 程式，自 nr database 中搜尋是否有重組序列相似的同源序列對應，藉以註解個別序列的可能功能。同源序列相近性的判定標準設爲 E-value ＜ 10^{-5}，如果搜尋結果中出現多於一筆以上的同源序列對應，則僅取用最佳比對序列。本研究所得的重組序列資料，包括 Contig I，Contig II，Contig III,與 Singleton III，共計 37,401 筆，經過比對的結果顯示，有 11,565 筆資料可在 nr 資料庫中找到相似序列，約占全部重組序列的 30.92%。上述的 11,565 筆同源序列資料，根據其來源物種進行計算及歸類，結果如圖 3-3，約 32%的註解序列來自甲殼類（白蝦：7%，其他對蝦類：10%，對蝦以外的甲殼類：15%），40%來自甲殼類以外的節肢動物，其餘的 28%序列則來自其他物種。

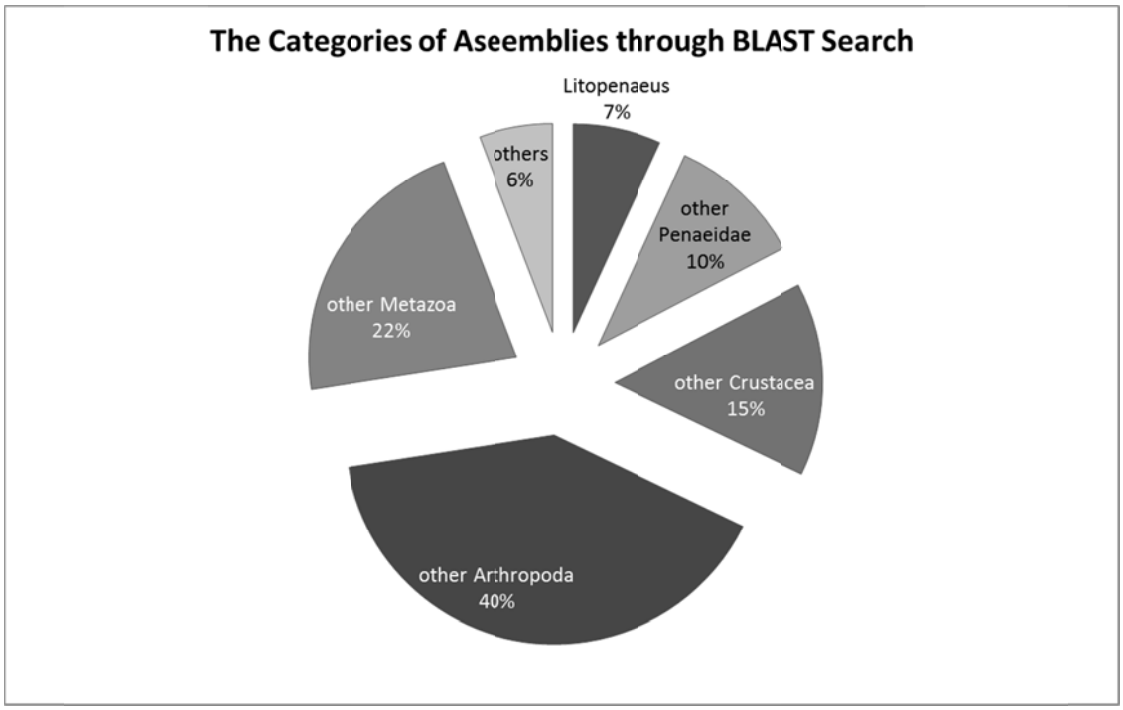


圖 3 -3 重組序列根據在 nr database 中所對應的序列其來源物種歸類

爲了得到更多的重組序列的功能註解，我們另外也以 rpsblast 程式，將重組序列對 Pfam database 進行比對，以得知這些序列是否具有已知的蛋白質模組特徵，符合模組特徵的判定標準爲 E-value ＜ 10^{-3}。在三萬七千筆重組序列中，經由 Pfam database 比對的註解程序後，共 15,398 筆資料能找到符合的模組特徵，約占全部重組序列的 41.17%。在 nr database 中有註解的 11,565 筆資料中，有 9,056 筆同時也有 Pfam 的蛋白模組特性；蛋白質模組的註解程序，讓我們多得到 6,342 筆重組序列的相關資訊。

3.3 組織的基因表現概況

檢視在白蝦 ESTs 序列資料的資訊，根據其 cDNA 基因庫來源組織型態分類，我們可得知目前所搜集的序列的整體概況。目前資料的組織來源的分佈比例如圖 3-4 所示；六個組織包含的 ESTs 數超過 10%，分別爲鰓（15％）、淋巴器官（15％）、血細胞（17％）、肝胰臟（14％）、眼柄（18％）及神經索（13％）；就註解的狀況來說，肝胰臟與眼柄的註解狀況最好，兩組織都有超過 70%的 ESTs 可經由

所屬的 contig 得到註解。此六組織各自包含的 ESTs 數、擁有註解的 ESTs 數、contigs 數及 singleton 數如表 3-2 所示。

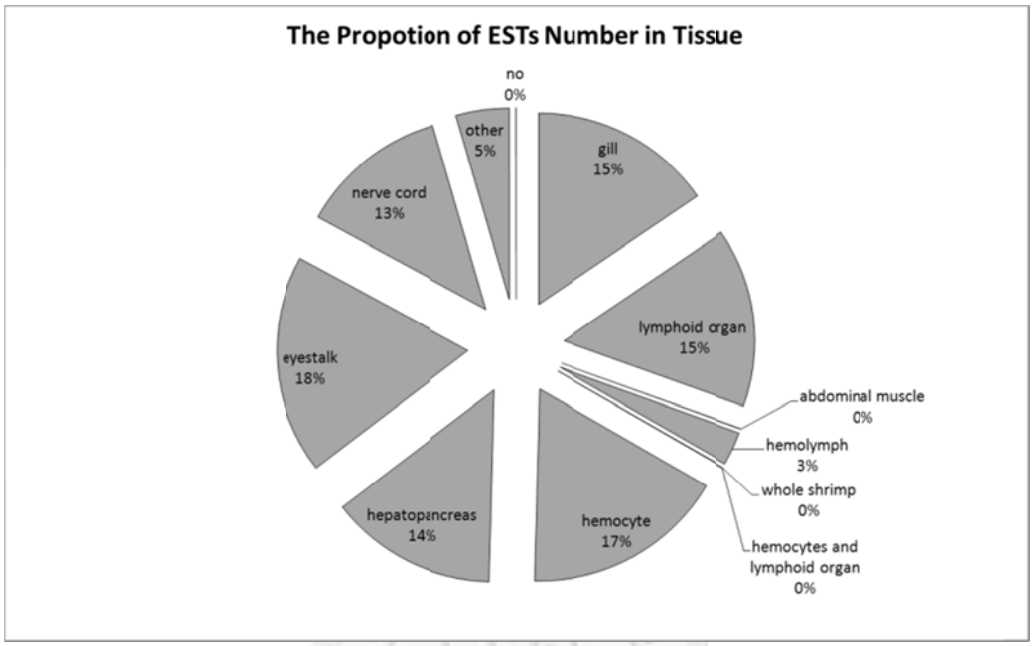


圖 3 -4 ESTs 的來源組織比例圖

表 3-2 六個組織的資料描述

	eyestalk	gills	hemocyte	hepatopancreas	lymphoid organ	nerve cord
EST	29575	24296	27369	22272	24214	20179
annotated ESTs (%)	20712 (70.03)	15166 (62.42)	17183 (62.78)	16641 (74.72)	16006 (66.10)	12374 (61.32)
contigs	3743	2905	2973	2825	2530	2836
singleton	1567	964	1025	1456	785	1126

經由重組序列的 ESTs 的組織來源解析，我們可以知道同屬於某單一重組序列的 ESTs，其在不同組織間的序列數量，進而估計個別重組序列在每一組織的表現狀況。將各重組序列透過對應的 Pfam domain，利用 Pfam database 與 GO database 的關聯對應表，引入相對應的 GO 註解，做爲重組序列的 GO 註解（參見本論文章節 2.3）。各組織的重組序列，依照 GO 的第一層分類的項目歸類，接著將擁有同一個 GO 註解的重組序列其包含的 EST 數量/組織中擁有 GO 註解的 EST 總數，用以顯示該 GO 項目在組織中所佔有的比例。以 GO 包含的三個主要類別的第一層分類（GO term, level 1)分別作圖。每個組織的分析則進一步依照重組序列的屬性 (contigs、singleton 及 all (contigs + singleton)) 分別呈現，以了解基因表現量是否因爲 singleton 而產生改變。由圖 3-5、3-6、3-7 的結果可看出 all 的趨勢與 contig 相似，因此 singleton 對於組織的基因表現概況的趨勢並未造成太大的影響。

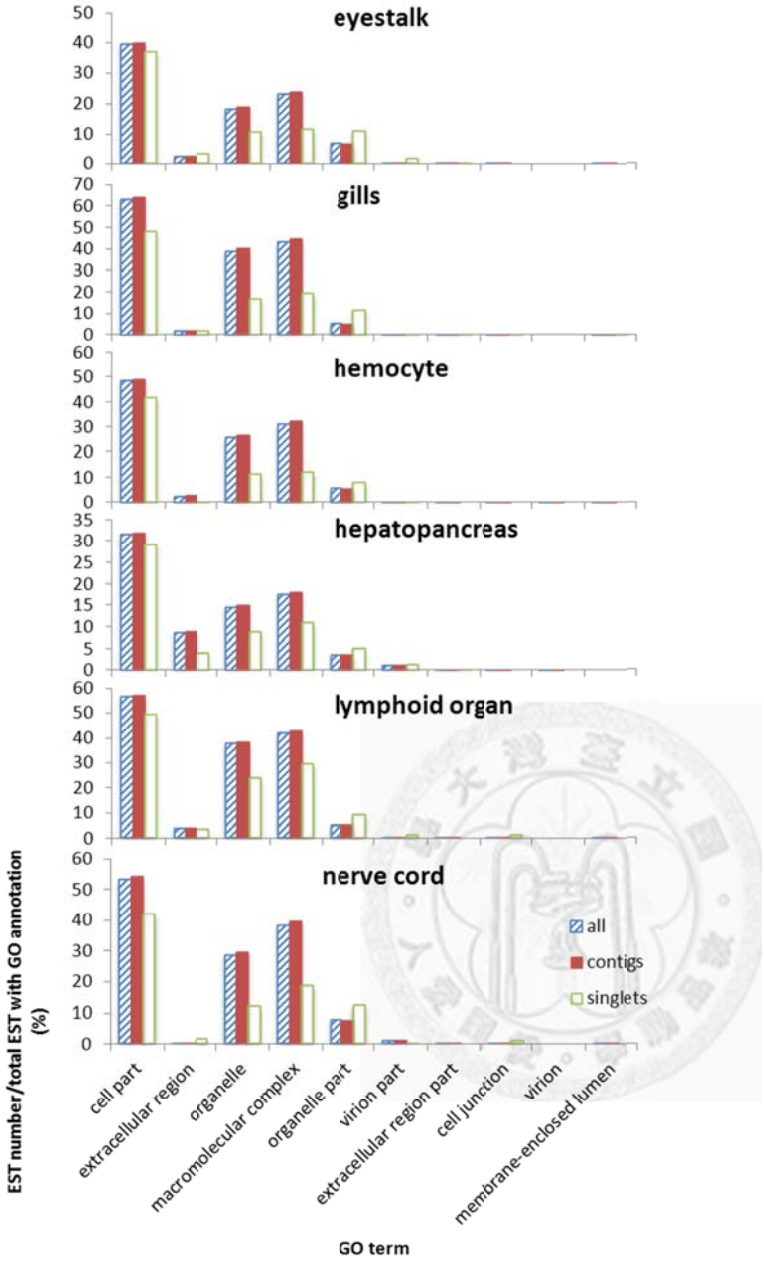


圖 3 -5 六個組織的重組序列，依照 cellular component 的第一層分類結果

六個組織在 cellular component 的表現概況呈現相似的趨勢(圖 3-5)，基因表現量高的項目分別爲 cell part、organelle 及 macromolecular complex，肝胰臟組織與其餘五個組織稍有不同，除了上述三個項目的基因表現量爲六個組織中最低之外，肝胰臟的重組序列在 extracellular region 有 10%左右的基因表現量，約爲最高基因表現量 cell part 的 1/3，高於其餘五個組織。

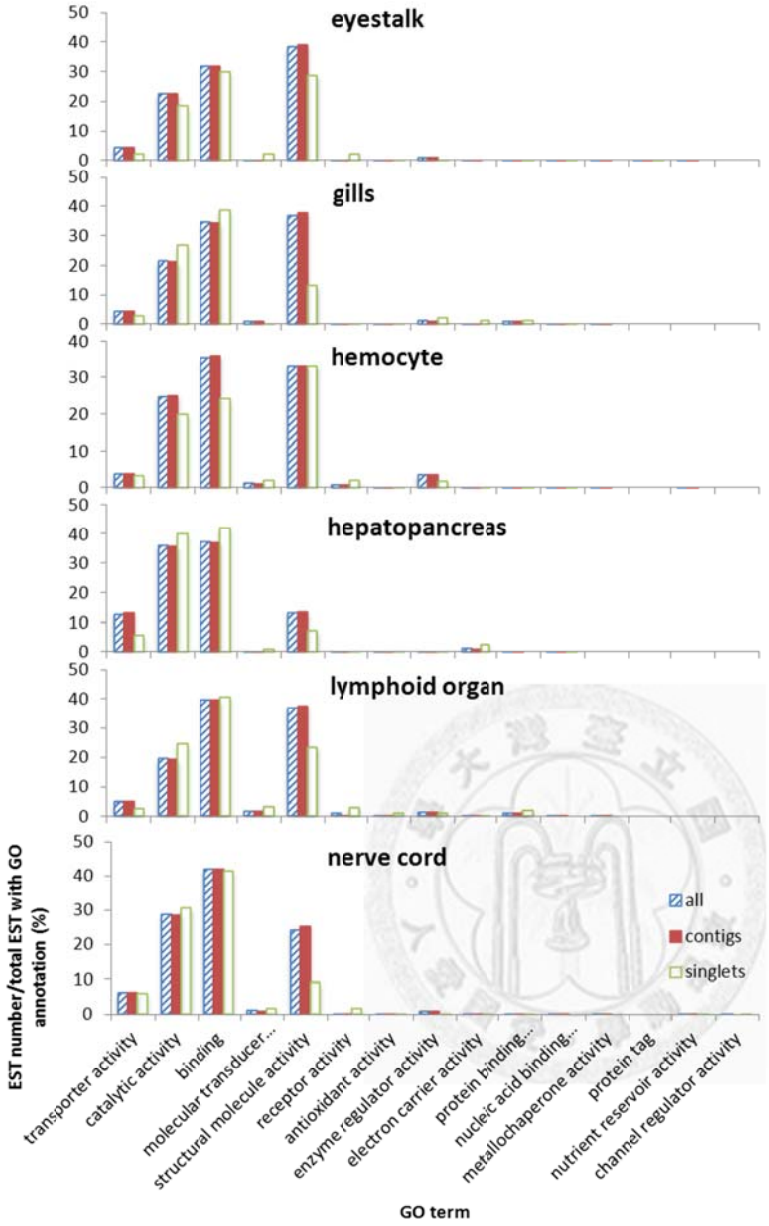


圖 3 -6 六個組織的重組序列，依照 molecular function 的第一層分類結果

六組織於 molecular function 的表現概況略有不同(圖 3-6)，基因表現量高的項目爲 transporter activity、catalytic activity、binding 及 structural molecular activity，眼柄及鰓兩組織，不論是在基因表現類型或是基因表現量，都呈現相似的結果，且基因表現量最高的項目是 structural molecular activity。血細胞及淋巴器官在基因表現量稍有差異，但整體分布趨勢相似，且兩個組織中，基因表現量最高的項目

皆爲 binding。神經索的重組序列，在 binding 項目的基因表現量最高，其次是 catalytic activity；肝胰臟的重組序列，在 catalytic activity 及 binding 兩項目的基因表現量最高，structural molecular activity 項目的基因表現量是六個組織中最低者，且六個組織中，僅肝胰臟的 transporter activity 項目的基因表現量超過 10%。

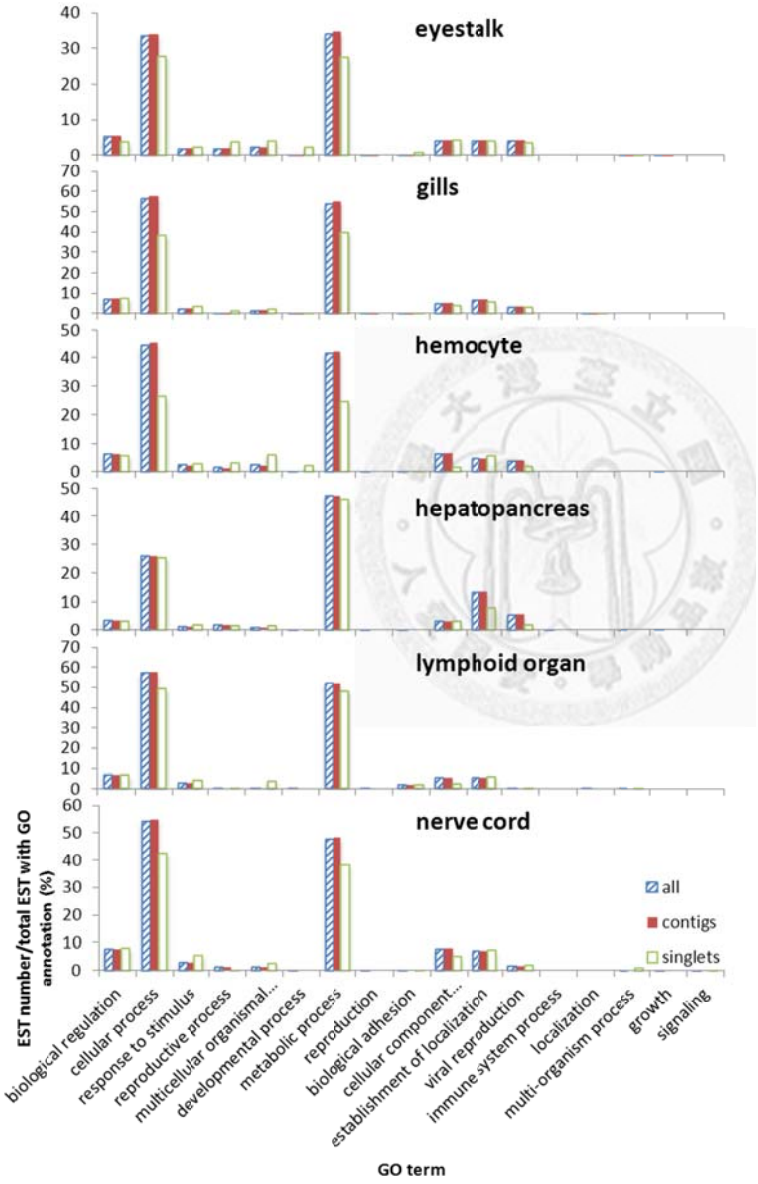


圖 3 -7 六個組織的重組序列，依照 biological process 的第一層分類結果

六組織在 biological process 都有兩個明顯的高基因表現量的項目(圖 3-7)，分別爲 cellular process 及 metabolic process，除肝胰臟以外，其餘五個組織中兩項目的基因表現量相近，肝胰臟的重組序列在 cellular process 項目的基因表現量約爲 metabolic process 項目的基因表現量的一半；其餘的項目在各組織中，都有零星的基因表現量，其中以肝胰臟在 establishment of localization 項目的基因表現量最多，約有 13%。

3.4 組織間表現基因的差異性

由 3.3 節中敘述的各組織之基因表現概況，略可得知各組織表現的基因功能類型之相似及相異程度，爲了更進一步了解組織間表現基因的差異，因此利用 Digital Differential Display(DDD) 的方法，以 Fisher’s exact test 進行統計檢定，以得到在不同組織之間有顯著差異表現的重組序列。在 p-value $< 10^{-3}$ 的條件下，眼柄和其餘五個組織有顯著差異表現量的重組序列爲 451 筆，鰓和其餘五個組織有顯著差異表現量的重組序列爲 530 筆，血細胞和其餘五個組織有顯著差異表現量的重組序列爲 572 筆，肝胰臟和其餘五個組織有顯著差異表現量的重組序列爲 732 筆，淋巴器官和其餘五個組織有顯著差異表現量的重組序列爲 590 筆，神經索和其餘五個組織有顯著差異表現量的重組序列 410 筆。

將上述的六批具有組織特異表現傾向的重組序列，分別進行 Gene Ontology enrichment analysis，也就是藉由統計方法，檢定出現在一個重組序列的子集合的 GO 項目，相對於整體重組序列而言，是否有較高的出現頻度。本研究以 Fisher’s exact test 爲量化 enrichment 的統計工具，E-value $< 10^{-3}$ 定義顯著表現的項目，選擇第二層 GO 項目進行分析，所得結果以 GO 的三個領域分別列表呈現 (表 3-3、3-4、3-5)，表中包含組織顯著表現的 GO 項目，與功能註解包含此 GO 項目的重組序列在組織中的比例，其比例是以 contigs 的數量做計算。

表 3-3 是六個組織的 cellular component GO 項目集中化程度的分析結果，在這些依照組織特異性表現的重組序列子集合中，血細胞及神經索兩個組織的重組序列子集合所包含的第二層 GO 項目，都沒有通過顯著性的檢定標準，也就是說，沒有任何第二層 GO 項目顯現集中的趨勢。眼柄及淋巴器官兩個組織則有較多顯著集中的 GO 項目，眼柄組織有六個，淋巴器官有五個，有五個項目相同且擁有相似的比例。

表 3-3 在 cellular component 中六組織顯著表現的 GO 項目及其比例 (比例的計算方式: 功能註解包含此 GO 項目的 contigs 數 / 組織中擁有 GO 功能註解的總 contigs 數)

Tissue	GO terms	Proportion (%)		P-value
eyestalk	non-membrane-bounded organelle	6.65	(30/451)	8.20E-09
	ribonucleoprotein complex	6.2	(28/451)	1.65E-07
	intracellular organelle	7.98	(36/451)	1.77E-07
	intracellular	5.32	(24/451)	5.09E-06
	intracellular part	10.42	(47/451)	0.000104
	protein-DNA complex	1.33	(6/451)	0.000522
gills	non-membrane-bounded organelle	4.72	(25/530)	8.45E-05
	ribonucleoprotein complex	4.34	(23/530)	0.000745
hemocyte				
hepatopancreas	membrane part	7.38	(54/732)	8.32E-07
lymphoid organ	non-membrane-bounded organelle	5.25	(31/590)	8.89E-07
	intracellular	4.75	(28/590)	7.48E-06
	ribonucleoprotein complex	4.92	(29/590)	1.64E-05
	intracellular organelle	5.93	(35/590)	0.000221
	intracellular part	9.32	(55/590)	0.000672
nerve cord				

表 3-4 是六個組織的 molecular function GO 項目集中化程度的分析結果，除肝胰臟外，structural constituent of cuticle 在其他五個組織的分析結果中，都被檢定爲顯著的 GO 項目；血細胞及神經索兩組織僅有此項目達到顯著水準，鰓則較上述兩組織多增加 structural constituent of ribosome；眼柄、肝胰臟及淋巴器官三個組織皆有四個以上的項目。眼柄及肝胰臟皆各有一個 GO 項目達到組織集中化的檢定標準，且涵蓋的重組序列達到此組織所有序列的 10%以上，在眼柄組織爲 structural constituent of cuticle，在肝胰臟則爲 hydrolase activity。

表 3-4 在 molecular function 中六組織顯著表現的 GO 項目及其比例 (比例的計算方式: 功能註解包含此 GO 項目的 contigs 數 / 組織中擁有 GO 功能註解的總 contigs 數)

Tissue	GO terms	Proportion (%)		P-value
eyestalk	structural constituent of cuticle	13.3	(60/451)	3.14E-34
	pattern binding	3.33	(15/451)	2.09E-08
	carbohydrate binding	3.33	(15/451)	2.84E-08
	structural constituent of ribosome	6.21	(28/451)	4.49E-08
gills	structural constituent of cuticle	4.72	(25/530)	1.19E-05
	structural constituent of ribosome	4.34	(23/530)	0.000247
hemocyte	structural constituent of cuticle	4.9	(28/572)	1.77E-06
hepatopancreas	hydrolase activity	12.70	(93/732)	7.52E-33
	pattern binding	4.64	(34/732)	2.68E-22
	carbohydrate binding	4.64	(34/732)	5.82E-22
	substrate-specific transporter activity	3.14	(23/732)	4.69E-07
lymphoid organ	structural constituent of ribosome	4.92	(29/590)	3.17E-06
	pattern binding	2.37	(14/590)	3.50E-06
	carbohydrate binding	2.37	(14/590)	4.57E-06

	structural constituent of cuticle	4.4	(26/590)	2.56E-05
	ion binding	3.56	(21/590)	0.000697
nerve cord	structural constituent of cuticle	4.39	(18/410)	0.000456

表 3-5 是六個組織在 biological process GO 項目集中化程度的分析結果，鰓、血細胞及神經索三個組織，在此 GO 類別中沒有顯著表現的 GO 項目，眼柄、肝胰臟及淋巴器官則有三個以上的項目，眼柄與淋巴器官的項目相似，眼柄較淋巴器官多一 multicellular organismal reproductive process 項目，而肝胰臟與眼柄則有兩個項目相異，除 macromolecule metabolic process 及 primary metabolic process 兩個項目相同外，眼柄組織顯著表現的項目還有 biosynthetic process 及 multicellular organismal reproductive process，肝胰臟顯著表現的項目則有 nitrogen compound metabolic process 及 cellular process involved in reproduction。

表 3-5 在 biological process 中六組織顯著表現的 GO 項目及其比例 (比例的計算方式: 功能註解包含此 GO 項目的 contigs 數 / 組織中擁有 GO 功能註解的總 contigs 數)

Tissue	GO terms	Proportion (%)		P-value
eyestalk	macromolecule metabolic process	13.3	(60/451)	5.01E-09
	primary metabolic process	14.19	(64/451)	2.75E-08
	biosynthetic process	7.54	(34/451)	2.39E-05
	multicellular organismal reproductive process	1.33	(6/451)	0.000178
gills				
hemocyte				

hepatopancreas	primary metabolic process	18.85	(138/732)	4.88E-28
	macromolecule metabolic process	16.67	(122/732)	2.67E-25
	nitrogen compound metabolic process	5.19	(38/732)	7.11E-07
	cellular process involved in reproduction	0.82	(6/732)	2.62E-06
lymphoid organ	biosynthetic process	7.29	(43/590)	5.97E-06
	macromolecule metabolic process	9.83	(58/590)	0.000197
	primary metabolic process	10.85	(64/590)	0.000265
nerve cord				

3.6 KEGG pathway analysis

由 GO enrichment analysis 的結果發現眼柄、肝胰臟及淋巴器官三個組織有較多集中化的GO項目，眼柄及淋巴器官兩組織的GO項目相似，肝胰臟則異於兩者；且根據 3.3 節的結果發現肝胰臟擁有最多的被註解 ESTs，在組織基因表現概況的分析中也發現異於其他組織的趨勢，因此我們挑選肝胰臟組織進行 KEGG pathway analysis，此外挑選眼柄組織作爲比較。

以文氏圖表現眼柄與肝胰臟組織其重組序列的關聯性，結果如圖 3-8 所示；眼柄組織本有 3,742 條重組序列，肝胰腺有 2,825 條，發現在兩個組織間有 1,002 條重組序列是相同的，因此扣除這相同的重組序列後，得到屬於眼柄組織且不屬於肝胰臟組織的重組序列，稱爲 eyestalk_only 共 2,740 條，屬於肝胰臟組織且不屬於眼柄組織的重組序列，稱爲 hepatopancreas_only 共 1,823 條及屬於眼柄組織且屬於

肝胰臟組織的重組序列，稱爲 intersection 共 1,002 條。

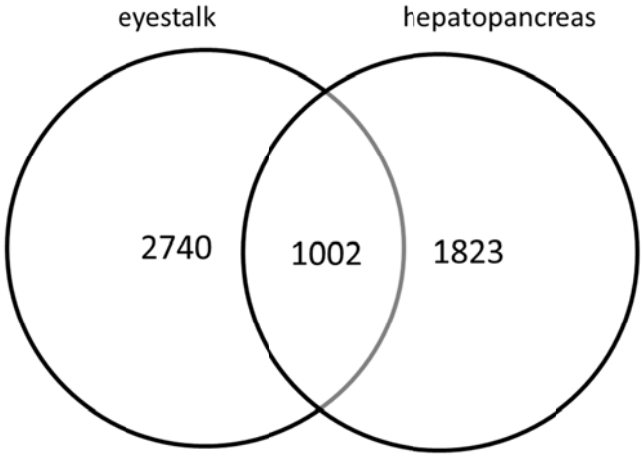


圖 3 -8 眼柄與肝胰腺的重組序列之關聯性

以 blastx 程式將上述的 eyestalk_only、hepatopancreas_only 及 intersection 三組序列資料，對 KEGG PATHWAY database 進行序列相似性比較，根據 E-value < 10^{-3} 決定序列具有相似性，同時根據對應的序列得到該序列參與的生理代謝途徑，三組序列資料的結果如表 3-6 所示；eyestalk_only 的重組序列數量最高，在 KEGG PATHWAY database 中找到最適對應的比例最低，intersection 重組序列數量最少，但找到最適對應的比例最高，爲 76%。

表 3-6 eyestalk_only、hepatopancreas_only 及 intersection 三組重組序列與 KEGG PATHWAY database 進行序列相似性比較的結果

	eyestalk_only	hepatopancreas_only	intersection
contigs	2740	1823	1002
E-value < 10^{-3}	1310 (48%)	1146 (63%)	759 (76%)
associated pathway	247	238	205

將上述的三份重組序列的列表分別進行 enrichment analysis，以 KEGG PATHWAY 爲分類依據，找出顯著集中出現的 PATHWAY 項目；集中程度以 Fisher's Exact Test 爲量化的統計工具，E-value < 10^{-3} 定義顯著表現的項目，以 KEGG PATHWAY database 作爲歸類依據，根據 P-value 進行排序，在此列出前十項顯著集中出現的 PATHWAY 項目，其結果如表格 3-7 所示。eyestalk_only 最顯著集中出現的 PATHWAY 項目爲 Tight junction ， hepatopancreas_only 最顯著集中出現的 PATHWAY 項目爲 Metabolic pathways，但 Metabolic pathway 屬於較廣泛的分類，因此選擇 Betalain biosynthesis，作爲此份重組序列最顯著集中出現的 PATHWAY，而 intersection 最顯著集中出現的則爲 Ribosome。Intersection 爲兩個組織皆有表現的基因，我們認爲這群在不同組織中表現的基因，應該是參與基本的生理功能，由上述結果發現最顯著出現的是 Ribosome。核醣體主要的功能爲合成蛋白質，而細胞內的各種功能都需要蛋白質的參與，因此，Ribosome 的確可能是在不同組織間，最顯著表現的基因群。hepatopancreas_only 的重組序列中，最顯著集中出現的 PATHWAY 爲 Betalain biosynthesis，但目前並未在蝦體中發現 betalain 的存在，因此我們須深入了解此基因群的 nr 及 Pfam 的註解，藉此確定僅出現於肝胰臟的重組序列是否真的參與 Betalain biosynthesis。eyestalk_only 的重組序列中，最顯著表現的基因群所參與的 PATHWAY 爲 Tight junction，但至今沒有研究說明在眼柄中有 Tight junction 的形成，因此必須進一步了解此基因群的 nr 及 Pfam 註解。

表 3-7 eyestalk_only、hepatopancreas_only 及 intersection 重組序列列表，經 enrichment 的步驟後，得到前 10 名生理代謝途徑名稱

eyestalk_only	**P-value**	**hepatopancreas_only**	**P-value**	**intersection**	**P-value**
Tight junction	2.41E-14	Metabolic pathways	1.13E-57	Ribosome	9.91E-75
Regulation of actin cytoskeleton	4.88E-13	Betalain biosynthesis	2.3E-44	Metabolic pathways	8.16E-45
Focal adhesion	5.83E-13	Isoquinoline alkaloid biosynthesis	2.3E-44	Oxidative phosphorylation	9.92E-38
Glycosphingolipid biosynthesis - ganglio series	5.67E-11	Riboflavin metabolism	4.2E-43	Parkinson's disease	3.15E-37
Glycosphingolipid biosynthesis - globo series	1.66E-10	Melanogenesis	4.98E-36	Huntington's disease	3.85E-33
GnRH signaling pathway	2.58E-10	Tyrosine metabolism	3.99E-35	Alzheimer's disease	1.63E-30
Leukocyte transendothelial migration	7.57E-10	Biosynthesis of secondary metabolites	1.59E-34	RNA transport	3.79E-13
Glycosaminoglycan degradation	3.02E-09	Protein digestion and absorption	3.24E-20	Cardiac muscle contraction	2.91E-12
Dilated cardiomyopathy	4.19E-09	Tuberculosis	1.72E-19	Proteasome	2.75E-11
Viral myocarditis	1.21E-08	Pancreatic secretion	7.25E-18	Carbon fixation in photosynthetic organisms	1.09E-10

第四章 討論

4.1 序列組裝之策略

目前白蝦的 ESTs 定序數量佔 NCBI dbEST 中甲殼綱（CRUSTACEA）ESTs 總數的 18.3 %，爲甲殼類 EST 定序數量最高的單一物種(Pontius *et al.*, 2002)；然而，目前白蝦基因體迄今尚未完成大規模的解序，因此本研究希望藉由重組大量的 ESTs，之後註解其可能的功能，藉此來對其基因有更多的瞭解。爲了更貼近原始的基因序列結構，我們藉由序列組裝的過程，將這些 ESTs 重新組合成較長的序列，由於白蝦爲非模式物種，同時也沒有已完成定序的相近物種可作爲參考基因體，因此只能經過 *de novo* assembly 的方式將 ESTs 重新組裝。不過 *de novo* assembly 會受到生物及非生物的因素影響，造成錯誤的序列結構產生。生物的因素，譬如基因體含有大量的重複性序列，可能造成我們無法解析這些 ESTs 之間的關係，同時不同的 ESTs 也可能因重複性序列而錯誤組裝。個體間可能存在的差異現象(polymorphism)及替代性剪接(alternative splicing)等，及其他基因體序列的混入等，也都可能造成 ESTs 在組裝過程中發生錯誤。而非生物的因素則爲定序時對序列的錯誤判讀，特別容易發生在序列的兩端，錯誤的判讀也可能導致序列組裝錯誤(Chevreux, 2005,Chevreux *et al.*, 2004)。爲了減少上述因素在組裝過程中造成的影響，本研究試著採用較相近的物種(水蚤與果蠅)作爲參考基因體，結合定址對映(Mapping)及 *de novo* assembly 的組裝策略，來解析白蝦基因體。其策略簡述如下：

1. 利用已完成大量定序的相近物種，作爲白蝦 ESTs 分類的參考基因體。首先，利用與白蝦同屬甲殼綱的水蚤之基因體進行第一次分類，以水蚤的轉錄基因群作爲參考基因體，利用 tblastx 的方式，找出白蝦 ESTs 所可能對應的基因序列。同時，以 E-value 值(E-05)及對應到的序列長度不少於 50 aa.

(150 bps) 來作爲篩選條件(Adams *et al.*, 1991)，希望可以鑑別出這些 ESTs 所真正對應的同源序列，再將這些 ESTs 做進一步的組合(Contig I)，使之更爲貼近真正的基因結構。對於那些沒有在水蚤轉錄基因體中找到最適對應的 ESTs，則進行第二次分析，以同屬於節肢動物門且爲重要模式生物的果蠅，作爲第二次的參考基因體。果蠅的基因體早於本世紀初被解序完成，且相關的基因表現及基因產物研究也十分豐富且詳盡，因此以 blastx 程式搜尋果蠅蛋白質體的同源基因，以 E-value 值(E-05)及 50 a.a (150 bps) 的長度作爲篩選條件，再次得到第二次分類的 ESTs，並將之組合(Contig II)(請參閱圖 2-2 及圖 2-3)。

2. 擷取兩次比對結果中有找出同源序列的 ESTs，分別經由 MIRA assembler 以 *de novo* 的方式重新組裝成爲較長的序列，將得到 contigs 及 singletons，之後，整合 singletons 及前兩次比對中皆未找到同源序列的 ESTs，再進行一次 *de novo* assembly，期能組出更爲完整的轉錄基因群組(Contig III) (請參閱圖 2-4)。

所以，藉由同源基因的比對，可將 ESTs 進行歸類及定位，提高所組裝出 contigs 的正確性。本研究選用 MIRA 作爲序列重組程式，乃是利用演算法的優勢，以序列中段即定序較準確的區域，作爲組裝過程的起始架構，減少因兩端區域的定序品質較差或是錯誤判讀等因素，所造成的錯誤序列組合，同時也希望藉由混合不同親緣關係之基因體比對策略，來提高重組序列的準確度及可信度。

4.2 重組序列之註解

經過重組後得到 16,886 條 contigs 及 20,515 條 singletons，其中 contigs 共包含 140,726 條 ESTs，爲原始資料 161,241 條 ESTs 的 87%，同時 singletons 佔有的比例僅剩 12.7%，與同爲白蝦之前人研究相比較，O’ Leary *et al.*(O'Leary *et al.*, 2006) 的研究共得到 13,656 條 ESTs 序列，其 contigs 包含 8,171 條(59.8%) ESTs，singletons

包含 5,484 條(40.2%) ESTs，Clavero-Sales *et al.*(Clavero-Salas *et al.*, 2007)的研究共得到 601 條 ESTs 序列，其 contigs 包含 404 條(67%) ESTs，singletons 包含 197 條(33%) ESTs。同時，也與其他對蝦的研究相比較，Tassanakajon *et al.*(Tassanakajon *et al.*, 2006)的研究共得到 10,100 條 ESTs 序列，其 contigs 包含 6,172 條 (61%) ESTs，singletons 包含 3,928 條 (39%) ESTs，Leu *et al.*(Leu *et al.*, 2007)的研究共得到 15,981 條 ESTs 序列，其 contigs 包含 7,723 條(48%) ESTs，singletons 包含 8,258 條(52%) ESTs。無論是同爲白蝦的先前研究或是其他對蝦之先前研究，皆僅以 ESTs 來進行序列重組，不引用參考基因體，從上述比較中發現，本實驗的 contigs 涵蓋更多的 ESTs 序列，同時無法與其他序列結合的下 ESTs (singletons)之數量也比較少。

基本上，這 16,886 條 contigs 是由章節 4.1 所述之策略所組裝而成的，由圖 3-1 的結果顯示，Contig I(以水蚤爲參考基因體)與 Contig II(以果蠅爲參考基因體)，爲透過參考基因體的序列比對分類後，重新組裝而成，兩組 contig 所包含的 ESTs 數，約佔全部 161,241 條 ESTs 的 1/3，其中 Contig I 涵蓋 46,471 條 ESTs 序列，較 Contig II 所涵蓋的 7,501 條 ESTs 序列爲多，顯示選擇與 ESTs 取樣來源越相近的物種，來作爲參考基因體時，其能找到同源基因的比例會越高，用於重組序列的 ESTs 數量也會增加，進而提高組裝的準確度。Contig III 包含 86,754 條 ESTs 序列，其數量爲三類 contigs 中最多者，顯示仍有超過半數的 ESTs 無法在這兩個參考基因體中找到同源序列，而僅能夠過直接進行 *de novo* assembly 的步驟，來對序列進行組裝。因 Contig III 及 Singleton III 是由沒有在水蚤轉錄基因體及果蠅蛋白質體中，找到同源序列的 ESTs 所組成，因此這兩份資料中也可能包含白蝦的特有基因。根據長度分佈圖(圖 3-2)，可看出 contigs 的序列大約集中分布於 200 ~ 1,200 bps，而以 800 bps 最多，同時 Contig I 及 Contig II 的最大長度皆小於 3,000 bps，但 Contig III 中卻有長度超過 4,000 bps 的重組序列，因 Contig III 是直接經過 *de novo* assembly 所產生，僅憑藉相同的排列，而將兩序列合併組裝，有可能因爲重複序列的原因，而容易組裝較長的序列，需要進一步的分析與實驗室的驗證，才能判別此類型的

重組序列存在與否。檢查此條序列，發現在我們的 nr 註解中，其同源序列爲 neuroblast differentiation-associated protein AHNAK，接著使用 ORFinder 尋找此重組序列可能的 open reading frame(ORF)，ORFinder 會針對序列的六個轉譯框架(Six Frames)進行分析，我們發現，在+2 轉譯框架(Frame)的結果中，顯示此序列包含 3 個可能的 ORF 區域，其中包含最長的 ORF 爲 2,574 bps，可轉譯出 857 aa。將 3 個可能的 ORF 的序列對 nr database 進行同源性搜尋，最長的 ORF 的序列與 neuroblast differentiation-associated protein AHNAK 有高度相似性，而+2 轉譯框架中的另外兩個 ORF，也顯示出與 neuroblast differentiation-associated protein AHNAK 有同源相似性，因此我們推測此序列可能參與 neuroblast differentiation 的過程。

藉由與已知基因的高度相似性可推測重組序列的基因功能，在眾多常用的已知序列基因資料庫中，我們選擇 nr 及 Pfam 兩個資料庫爲重組序列作註解；有 31%的重組序列在 nr database 中找到最適配對，在 Pfam database 中則有 41%找到最適配對，比較在兩資料庫中找到最適配對的重組序列，發現於 nr database 中找到最適配對的重組序列，有 78%也同時在 Pfam database 中找到最適配對，如圖 4-1 所示。Pfam database 是針對 domain 進行搜尋比對，因此較 nr database 對應到較多的結果，若結合兩個註解資料，大約有 48%的重組序列，可以得到註解，進而瞭解其可能牽涉的生理功能。

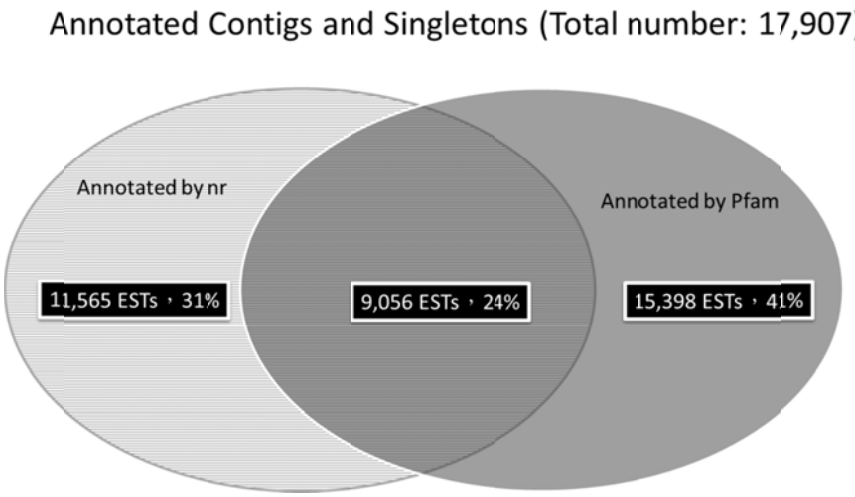


圖 4-1 擁有 nr 或 Pfam 註解的重組序列數與比例

4.3 不同組織間重組序列之比較

透過分析 NCBI 的白蝦 ESTs 資料群，可以發現有六個 libraries 的 ESTs 數皆超過 20,000 條，而且此六個 libraries 分別來自六個不同的組織。所以，我們比對各組織所擁有的 Contigs/Singleton，就能進一步探知不同組織間的異同之處。在此我們利用 Gene Ontology(GO)，透過 GO 的三大類別(Biological Process (BP), Cellular Component (CC), Molecular Function (MF))，來分類這些基因產物的之間的關係，了解各組織的基因在功能及表現量上的差異，其結果顯示雖然在表現量的趨勢上有所差異，但表現量高的分類項目(GO term)在六個組織中都相同，也許是因爲選擇 level 1 的項目作爲分類，因此差異比較不顯著。爲了進一步確定各組織的基因表現是否有特異性，我們運用 Digital Differential Display (DDD)的方式，以 Fisher-exact test 來計算，挑選出各組織與其他組織其表現量有顯著差異的重組序列，之後經由 functional enrichment 的方式，鑑別出各組織所擁有的特定表現基因，與其所牽涉的調控功能。Functional enrichment 乃是透過統計檢定的方法，如 Chi-square、Fisher’s exact test、Binomial probability 及 Hypergeometric distribution 等(Rivals *et al.*, 2007)，來對龐大的基因群進行分析，瞭解其中所隱含的生物意義。目前，在網

路上可找到許多提供 GO enrichment analysis 的工具或網站，DAVID 也是其中之一(Huang da *et al.*, 2009,Huang da *et al.*, 2007)。DAVID 的創建歷史較長久，且被許多文獻所引用，因此我們參考 DAVID 網站中的分析流程並選擇 Fisher’s exact test 作爲量化的方法，同樣以 GO 三大分類，來探索各個組織的特有基因表現與功能差異。其中，molecular function(MF)是我們較有興趣的類別，能夠明顯直接地了解基因的功能。在 MF 這個類別中，六個組織皆分別有其顯著表現的基因群。譬如眼柄、肝胰臟及淋巴器官等三個組織，其顯著表現的基因群功能較多樣(表 3-4)，如　而鰓、血細胞及神經索等三個組織，其顯著表現的基因群功能較少。顯示在一般情況，眼柄、肝胰臟及淋巴器官呈現的基因功能較多樣，可能與其反應外界刺激、能量調節與防禦免疫等機制有關。在表 3-4 的結果中，我們發現 structural constituent of cuticle 此 GO 項目，出現在除了肝胰臟以外的五個組織中。蝦子的鰓是被一層 cuticle 所包覆(Foster and Howse, 1978)，而 cuticle 中包含著許多的 cuticle protein，因此，在鰓組織其顯著表現的基因功能爲 structural constituent of cuticle，爲合理的結果，此外也曾有研究發現，在眼柄中有 cuticle proteins 的表現(Brady *et al.*, 2012)。淋巴器官、血細胞及神經索此三個組織，則未發現有相關文獻指出三組織中有 cuticle protein 的基因表現，因此，此結果應再進一步確認。

根據表 3-4 的結果，發現眼柄、肝胰臟及淋巴器官等三個組織，其顯著表現的基因群功能較多樣，且眼柄及肝胰臟兩個組織中，被註解的 ESTs 序列其比例超過 70% (表 3-2)，因此我們選取眼柄及肝胰臟兩組織之表現差異基因群，來進行 KEGG 資料庫的分析註解。先以文式圖(圖 3.8)區分各別出現於不同組織的重組序列，及同時出現於兩組織的重組序列，得到三組資料列表，分別爲只存在眼柄組織不存在肝胰臟的重組序列(eyestalk_only)、只存在肝胰臟組織不存在眼柄的重組序列(hepatopancreas_only)及同時存在眼柄與肝胰臟兩個組織的重組序列(intersection)。三組資料於 KEGG PATHWAY database 中搜尋同源序列後，透過 functional enrichment 的分析，檢視三組資料在 PATHWAY 的集中化程度。由表 3-7 的結果

可發現，intersection 的重組序列，最顯著表現的基因群所參與的 PATHWAY 爲 Ribosome，此基因群共包含 142 條 contigs，且幾乎全是核醣體蛋白，顯示在兩個不同組織其相同的生理作用中，轉譯作用最爲旺盛。Hepatopancreas_only 的重組序列中，最顯著表現的基因群所參與的 PATHWAY 爲 Betalain biosynthesis，betalain 爲一種出現在植物及高等菌類中的色素，目前並未在蝦體中發現，因此我們深入了解此基因群的 nr 及 Pfam 的註解，發現在此基因群中的序列皆爲血藍素(hemocyanin)，顯示在肝胰臟組織中存在許多血藍素，肝胰臟也被證實是負責合成血藍素的組織(王瑜琦, 2007)，因此，我們的實驗結果與前人研究的結果相吻合。eyestalk_only 的重組序列中，最顯著表現的基因群所參與的 PATHWAY 爲 Tight junction，此基因群中的基因大多爲 actin 及 myosin，actin 與 myosin 的交互作用參與許多細胞移動的過程，如肌肉收縮、細胞分裂、運送膜微囊(membrane vesicle)及細胞蠕動(cell crawling)等，所以此基因群是否參與 Tight junction 這個 PATHWAY，仍須進一步的實驗證實。

利用 GO database 或 KEGG PATHWAY database 的資料，爲基因進行功能性分群的方法，其結果會受到所引用的註解資料庫的限制，包括基因在資料庫中的註解數量、註解的完整程度及註解的資訊是否正確(Khatri and Drăghici, 2005)，之後我們藉由統計方法了解其集中程度，此過程稱爲 functional enrichment。Functional enrichment 的結果會因爲下列因素而有所不同，如不同的統計方式、檢定值的設定、背景值的選擇、基因群的樣本大小及基因群在資料庫中的註解情形等(Huang da *et al.*, 2009)，因此，選擇的基因群越大且基因被註解的越多，其檢定的結果越能貼近真實的情形。本實驗中，重組序列的註解情形並不完整，所以在功能性分群及 functional enrichment 產生的結果，可能會偏離生物體中的真實情形。

第五章 結論

本篇論文之主旨，乃是利用線上公用資料庫上所提供的大量南美白蝦 (*L. vannamei)* ESTs，以相近物種爲參考基因體，輔以 *de novo* assembly 方式，組裝出 37,401 條序列 (20,515 條 singletons 及 16,886 條 contigs)，並進行後續的功能性比較分析。本研究顯示，利用參考基因體同源基因的概念，結合 *de novo* assembly 的方式，可以提高重組序列的正確性，進而重建白蝦轉錄基因體，並能涵蓋更多 ESTs，且降低 singletons 的形成。隨後以重組序列作爲註解的主體，以 ESTs 的對應數量 (Mapping to Contigs) 來代表重組序列的表現量，藉此可以進一步比較不同組織的基因表現，發現在一般情況下，眼柄、肝胰臟及淋巴器官三個組織，呈現的基因功能較多樣化，眼柄及肝胰臟兩個組織間都有共同表現的重組序列，大多是參與轉譯過程的基因，眼柄組織特有序列較集中表現 actin 及 myosin 兩個基因，肝胰臟組織的特有序列，則是表現血藍素 (hemocyanin)。此外，所建構出的轉錄體資訊，也能提供研究者作爲實驗依據，驗證本研究所發現的新基因，及其序列之組成及探知其可能的調控功能。同時，研究的成果，亦可作爲物種親緣性分析及後續蛋白質體研究之參考。但是，使用 ESTs 爲材料進行分析，因實驗技術的原因，有其先天性的限制。由於 EST 通常爲單端單次定序，除了無法提供較爲完整的基因序列外，其涵蓋範圍 (Coverage) 也不完全，同時同一鹼基的重複涵蓋次數 (Depth) 也不足，且因單次定序，所以造成整體的序列品質不佳，容易有較高的錯誤率，進而影響後續序列的重組。此外，表現量高的基因被選殖定序的機率較高，而低量表現的基因群往往數量偏低，甚至沒有資訊，因此這些表現量低的基因，以 ESTs 方式來進行研究的話，很容易被忽略及或做出不恰當的研究結論。然而，隨著定序技術的發展，次世代定序 (Next Generation Sequencing，NGS) 的出現，讓我們得以跳脫出原有技術的限制，以平行雙端的方式，快速獲得大量的序列，提升我們對轉錄體整體的涵蓋度與

更爲精準的表現概況。如此，將可協助我們更進一步地掌握白蝦的基因表現，瞭解環境變動與病原感染下，蝦體的調控機制與免疫防禦的運作模式，期能對於疾病的防治，及提升養殖成功率等方面，能有所助益。

參考文獻

1. Adams MD (2000) The Genome Sequence of *Drosophila melanogaster*. Science 287: 2185-2195.
2. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651-1656.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
4. Brady P, Elizur A, Williams R, Cummins SF, Knibb W (2012) Gene expression profiling of the cephalothorax and eyestalk in *Penaeus monodon* during ovarian maturation. Int J Biol Sci 8: 328-343.
5. Cheng W, Chen SM, Wang FI, Hsu PI, Liu CH, Chen JC (2003) Effects of temperature, pH, salinity and ammonia on the phagocytic activity and clearance efficiency of giant freshwater prawn *Macrobrachium rosenbergii* to *Lactococcus garvieae*. Aquaculture 219: 111-121.
6. Cheng W, Liu CH, Kuo CM (2003) Effects of dissolved oxygen on hemolymph parameters of freshwater giant prawn, *Macrobrachium rosenbergii* (de Man). Aquaculture 220: 843-856.
7. Chevreux B (2005) MIRA: An Automated Genome and EST Assembler: German Cancer Research Center Heidelberg.
8. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S (2004) Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. Genome Res 14: 1147-1159.
9. Clavero-Salas A, Sotelo-Mundo RR, Gollas-Galvan T, Hernandez-Lopez J, Peregrino-Uriarte AB, Muhlia-Almazan A, Yepiz-Plascencia G (2007) Transcriptome analysis of gills from the white shrimp *Litopenaeus vannamei* infected with white spot syndrome virus. Fish Shellfish Immunol 23: 459-472.
10. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva

EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011) The ecoresponsive genome of *Daphnia pulex*. Science 331: 555-561.

11. Dong B, Xiang JH (2007) Discovery of genes involved in defense/immunity functions in a haemocytes cDNA library from *Fenneropenaeus chinensis* by ESTs annotation. Aquaculture 272: 208-215.
12. FAO (2010-2012) Fisheries and Aquaculture Department.
13. Foster CA, Howse HD (1978) A morphological study on gills of the brown shrimp, *Penaeus aztecus*. Tissue Cell 10: 77-92.
14. Gross PS, Bartlett TC, Browdy CL, Chapman RW, Warr GW (2001) Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei*, and the Atlantic White Shrimp, *L. setiferus*. Dev Comp Immunol 25: 565-577.
15. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.
16. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
17. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35: W169-175.
18. Jain M, Shrager J, Harris EH, Halbrook R, Grossman AR, Hauser C, Vallon O (2007) EST assembly supported by a draft genome sequence: an analysis of the *Chlamydomonas reinhardtii* transcriptome. Nucleic Acids Res 35: 2074-2083.
19. Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. Brief Bioinform 1: 76-92.
20. Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, Mather PB (2011) Transcriptomics of a giant freshwater prawn (*Macrobrachium rosenbergii*): *de novo* assembly, annotation and marker discovery. PLoS ONE 6: e27938.
21. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.
22. Khatri P, Drăghici S (2005) Ontological analysis of gene expression data: current

tools, limitations, and open problems. Bioinformatics 21: 3587-3595.
23. Leekitcharoenphon P, Taweemuang U, Palittapongarnpim P, Kotewong R, Supasiri T, Sonthayanon B (2010) Predicted sub-populations in a marine shrimp proteome as revealed by combined EST and cDNA data from multiple Penaeus species. BMC Res Notes 3: 295.
24. Leelatanawit R, Sittikankeaw K, Yocawibun P, Klinbunga S, Roytrakul S, Aoki T, Hirono I, Menasveta P (2009) Identification, characterization and expression of sex-related genes in testes of the giant tiger shrimp *Penaeus monodon*. Comp Biochem Physiol A Mol Integr Physiol 152: 66-76.
25. Lehnert SA, Wilson KJ, Byrne K, Moore SS (1999) Tissue-specific expressed sequence tags from the black tiger shrimp *Penaeus monodon*. Mar Biotechnol 1: 465-476.
26. Leu JH, Chang CC, Wu JL, Hsu CW, Hirono I, Aoki T, Juan HF, Lo CF, Kou GH, Huang HC (2007) Comparative analysis of differentially expressed genes in normal and white spot syndrome virus infected *Penaeus monodon*. BMC Genomics 8: 120.
27. Leu JH, Chen SH, Wang YB, Chen YC, Su SY, Lin CY, Ho JM, Lo CF (2011) A review of the major penaeid shrimp EST studies and the construction of a shrimp transcriptome database based on the ESTs from four penaeid shrimp. Mar Biotechnol (NY) 13: 608-621.
28. Liao IC, Huang TL (1969) A preliminary report on artificial propagation of *Penaeus monodon*(Fabricius). Joint Commission on Rural Reconstruction (JCRR) Fisheries Series 8: 67-71.
29. McQuilton P, St Pierre SE, Thurmond J (2012) FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res 40: D706-714.
30. Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8: 6-21.
31. O'Leary NA, Trent HF, 3rd, Robalino J, Peck ME, McKillen DJ, Gross PS (2006) Analysis of multiple tissue-specific cDNA libraries from the Pacific whiteleg shrimp, *Litopenaeus vannamei*. Integr Comp Biol 46: 931-939.
32. Pongsomboon S, Wongpanya R, Tang S, Chalorsrikul A, Tassanakajon A (2008) Abundantly expressed transcripts in the lymphoid organ of the black tiger shrimp, *Penaeus monodon*, and their implication in immune function. Fish Shellfish Immunol 25: 485-493.
33. Pontius JU, Wagner L, Schuler GD (2002) UniGene: a unified view of the transcriptome. In: McEntyre J, Ostell J, editors. Bethesda (MD): National Center for Biotechnology Information (US).
34. Preechaphol R, Leelatanawit R, Sittikankeaw K, Klinbunga S, Khamnamtong B,

Puanglarp N, Menasveta P (2007) Expressed sequence tag analysis for identification and characterization of sex-related genes in the giant tiger shrimp *Penaeus monodon*. J Biochem Mol Biol 40: 501-510.

35. Preston N (1985) The combined effects of temperature and salinity on hatching success and the survival, growth, and development of the larval stages of *Metapenaeus bennettae* (Racek & Dall). Journal of Experimental Marine Biology and Ecology 85: 57-74.
36. Primavera JH (1998) Tropical shrimp farming and its sustainability. In: Silva SSD, editor. Tropical Mariculture. San Diego: Academic Press. pp. 257-289.
37. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-301.
38. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23: 401-407.
39. Robalino J, Almeida JS, McKillen D, Colglazier J, Trent HF, Chen YA, Peck MET, Browdy CL, Chapman RW, Warr GW, Gross PS (2007) Insights into the immune transcriptome of the shrimp *Litopenaeus vannamei*: tissue-specific expression profiles and transcriptomic responses to immune challenge. Physiological Genomics 29: 44-56.
40. Rojtinnakorn J, Hirono I, Itami T, Takahashi Y, Aoki T (2002) Gene expression in haemocytes of kuruma prawn, *Penaeus japonicus*, in response to infection with WSSV by EST approach. Fish Shellfish Immunol 13: 69-83.
41. Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends in Plant Science 8: 321-329.
42. Sathiyamoorthy S, In JG, Gayathri S, Kim YJ, Yang DC (2010) Generation and gene ontology based analysis of expressed sequence tags (EST) from a *Panax ginseng* C. A. Meyer roots. Mol Biol Rep 37: 3465-3472.
43. Selvam DG, Mujeeb Rahiman KM, Mohamed Hatha AA (2012) An investigation into occasional white spot syndrome virus outbreak in traditional paddy cum prawn fields in India. ScientificWorldJournal 2012: 340830.
44. Stojmirovic A, Yu YK (2010) Robust and accurate data enrichment statistics via distribution function of sum of weights. Bioinformatics 26: 2752-2759.
45. Supungul P, Klinbunga S, Pichyangkura R, Hirono I, Aoki T, Tassanakajon A (2004) Antimicrobial peptides discovered in the black tiger shrimp *Penaeus monodon* using the EST approach. Dis Aquat Organ 61: 123-135.
46. Supungul P, Klinbunga S, Pichyangkura R, Jitrapakdee S, Hirono I, Aoki T, Tassanakajon A (2002) Identification of immune-related genes in hemocytes of

black tiger shrimp (*Penaeus monodon*). Mar Biotechnol (NY) 4: 487-494.

47. Tassanakajon A, Klinbunga S, Paunglarp N, Rimphanitchayakit V, Udomkit A, Jitrapakdee S, Sritunyalucksana K, Phongdara A, Pongsomboon S, Supungul P, Tang S, Kuphanumart K, Pichyangkura R, Lursinsap C (2006) *Penaeus monodon* gene discovery project: the generation of an EST collection and establishment of a database. Gene 384: 104-112.
48. Yamano K, Unuma T (2006) Expressed sequence tags from eyestalk of kuruma prawn, *Marsupenaeus japonicus*. Comp Biochem Physiol A Mol Integr Physiol 143: 155-161.
49. Xiang J, Wang B, Li F, Liu B, Zhou Y, Tong W (2008) Generation and analysis of 10,443 ESTs from cephalothorax of *Fenneropenaeus Chinensis*. The 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, China. p. 74–80.
50. 王瑜琦 (2007) 白蝦免疫相關基因之表現及其受口服葡聚多醣的影響. 高雄市: 國立中山大學. 209 p.
51. 鄭金華 (2007) 無特定病原(SPF)白蝦繁養殖模式之開發與產業應用. 農業生技產業季刊. pp. 48-59.