Tạp chí Công nghệ Sinh học **15**(3): 471-480, 2017

PHÂN TÍCH HỆ PHIÊN MÃ VÀ SÀNG LỌC MỘT SỐ GEN GIẢ ĐỊNH LIÊN QUAN TỚI TÍNH TRẠNG TĂNG TRƯỞNG Ở TÔM SÚ (*PENAEUS MONODON*)

Nguyễn Hải Bằng[1], Phạm Quang Huy[2], Trần Xuân Thạch[2], Nguyễn Giang Thu[3], Nguyễn Thị Minh Thanh[2], Nguyễn Thị Hoa[2], Hà Thị Thu[2], Nguyễn Thị Tuyết Nhung[2], Nguyễn Cường[2], Nguyễn Hữu Ninh[4], Đồng Văn Quyền[2], Chu Hoàng Hà[2], Đinh Duy Kháng[2,] ✉

[1]*Trường Đại học Y Dược Hải Phòng*
[2]*Viện Công nghệ sinh học, Viện Hàn lâm khoa học và Công nghệ Việt Nam*
[3]*Vụ Khoa học công nghệ và Môi trường, Bộ Nông nghiệp và Phát triển nông thôn*
[4]*Viện nghiên cứu nuôi trồng thủy sản III, Bộ Nông nghiệp và Phát triển nông thôn*

✉ Người chịu trách nhiệm liên lạc. E-mail: khangvspt@ibt.ac.vn

Ngày nhận bài: 13.12.2016
Ngày nhận đăng: 10.3.2017

TÓM TẮT

Tôm sú (*Penaeus monodon*) là loài thủy sản nuôi trồng đem lại nguồn lợi lớn cho quốc gia. Trong những năm gần đây, xuất khẩu tôm sú có thể đạt gần một tỷ USD/năm. Tuy nhiên, các dữ liệu về hệ gen và hệ phiên mã của tôm sú còn hạn chế khiến cho việc nghiên cứu phục vụ cho việc chọn tạo giống với những tính trạng quan trọng như tăng trưởng nhanh, kháng bệnh còn gặp nhiều khó khăn. Giải trình tự và phân tích hệ phiên mã tôm sú sẽ cung cấp các dữ liệu quan trọng cho công tác chọn giống tôm sú. Trong nghiên cứu này, từ gói dữ liệu giải trình tự thế hệ mới mô cơ và mô gan tụy tôm sú thu nhận từ vùng biển Bắc Trung Bộ Việt Nam, chúng tôi đã đánh giá, tiền xử lý và lắp ráp *de novo* hệ phiên mã, tinh sạch và thu được 17.406 unigene với kích thước trung bình là 403,06 bp, N50 là 402 bp. Toàn bộ các unigene trong hệ phiên mã tinh sạch được chú giải với 4 cơ sở dữ liệu khác nhau và đã sàng lọc được 51 unigene liên quan đến tính trạng tăng trưởng. Phân tích biểu hiện cho thấy 16.148 unigene có sự biểu hiện khác biệt giữa mô cơ và mô gan tụy. Những kết quả này sẽ là nguồn dữ liệu hữu ích về hệ phiên mã tôm sú và có thể được áp dụng cho nhiều nghiên cứu tiếp theo đặc biệt trong việc sàng lọc các chỉ thị phân tử liên kết với những tính trạng có ý nghĩa kinh tế quan trọng ở tôm sú.

Từ khóa: *Hệ phiên mã, tính trạng tăng trưởng, tôm sú Penaeus monodon, unigene*

MỞ ĐẦU

Tôm sú (*Penaeus monodon*) là loài thủy sản mang lại giá trị kinh tế lớn, hiện nay đang được nhiều nước chú trọng phát triển như Thái Lan, Việt Nam, Hàn Quốc, Đài Loan, Malaysia, Indonesia, Ấn Độ (Rosenberry, 2004). Nghề nuôi tôm sú có ưu thế lớn với các nước này vì đó là nguồn tài nguyên bản địa có thể nuôi và khai thác lâu dài, đóng góp quan trọng vào vấn đề an toàn lương thực, xóa đói giảm nghèo và phát triển kinh tế xã hội của mỗi nước. Chiến lược phát triển lâu dài của toàn khu vực là có được ngành sản xuất tôm sú bền vững, hạn chế tối thiểu các tác động tiêu cực đến môi trường sinh thái. Nền tảng cho chiến lược phát triển này là phát triển nguồn tôm bản địa với các chương trình nhân giống khoa học để nâng cao tỷ lệ sống và sự tăng trưởng. Để đạt được mục đích này, việc nghiên cứu cấu trúc và chức năng của toàn bộ hệ gen tôm sú là một vấn đề khoa học cơ bản có định hướng ứng dụng hết sức quan trọng.

Nghiên cứu hệ gen tôm sú sẽ cung cấp thông tin chính xác cho việc xác định các tính trạng quan trọng như tính trạng tăng trưởng, tính kháng bệnh, tính chống chịu với điều kiện môi trường, các tính trạng liên quan đến chất lượng tôm. Do kích thước hệ gen tôm sú rất lớn, khoảng 2,17 Gb (You *et al.,* 2010) nên việc giải mã toàn bộ hệ gen tôm sú đòi hỏi thời gian và tốn nhiều kinh phí. Vì vậy, để có thể từng bước khai thác các thông tin cần thiết từ hệ gen tôm sú phục vụ thực tiễn sản xuất thì việc giải mã từng phần hệ gen như giải mã hệ phiên mã, giải mã từng phân đoạn trong hệ gen có định hướng sử dụng kỹ thuật GBS (Genome typing by Sequencing) với phương pháp xác định trình tự gen thế hệ mới (NGS) là cách tiếp cận thông minh và khả thi.

Hệ phiên mã là tập hợp tất cả các phân tử RNA trong cơ thể sinh vật có khả năng mã hóa protein (Brown, 2002), là cầu nối từ thông tin trình tự hệ gen đến chức năng của hệ protein. Chính vì vậy phân tích hệ phiên mã sẽ giúp chúng ta thu được những kết quả sâu hơn khi phân tích chức năng của protein tương ứng. Sự ra đời của công nghệ giải trình tự thế mới (NGS) đã tạo điều kiện thuận lợi để thu nhận và khai thác thông tin về hệ gen và hệ phiên mã của sinh vật (Wang *et al.*, 2009). RNA-seq (RNA sequecing) là công nghệ giải trình tự thế hệ mới với đối tượng là RNA. RNA-seq sẽ giúp các nhà nghiên cứu có thể tìm hiểu sâu hơn thông tin liên quan trình tự hệ phiên mã và phân tích chức năng gen. Bằng phương pháp tính toán số lượng trình tự thu được từ RNA-seq, người ta có thể đánh giá được mức độ biểu hiện gen. Đây là phương pháp có khả năng thay thế được phương pháp micro-array truyền thống (Wang *et al.*, 2009). Hiện nay trên thế giới, nghiên cứu hệ phiên mã được chia làm 2 hướng: i) đối với đối tượng đã có dữ liệu tham chiếu cần sử dụng phương pháp re-sequencing; ii) với những dự án thực hiện trên những loài chưa có dữ liệu tham chiếu cần tiếp cận theo phương pháp lắp ráp *de novo* (Rismani-Yazdi *et al*., 2011; Rismani-Yazdi *et al*., 2012; Guo *et al*., 2014; Li *et al*., 2014; Liu *et al*., 2014).

Do chưa có hệ phiên mã tham chiếu, nên đối với loài tôm sú *Penaeus monodon*, chúng tôi đã tiến hành nghiên cứu ứng dụng công nghệ giải trình tự thế hệ mới để giải trình tự hệ phiên mã tôm sú. Trong nghiên cứu này, từ dữ liệu giải trình tự hệ phiên mã tôm sú thu được từ mô cơ và mô gan tụy, chúng tôi tiến hành lắp ráp *de novo,* chú giải và phân tích biểu hiện nhằm xây dựng bản đồ hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon* và sàng lọc các gen giả định liên quan tới tính trạng tăng trưởng.

VẬT LIỆU VÀ PHƯƠNG PHÁP

Mẫu tôm sú tươi được thu nhận từ vùng biển Bắc Trung Bộ (Nghệ An) được kiểm tra bằng Nested-PCR để loại bỏ các mẫu nhiễm bệnh (WSSV, MBV, TSV, IHHNV, IHHNV, YHV). Các mô gồm mô cơ, mô gan tụy được tách riêng từ mỗi mẫu tôm. RNA tổng số được tách chiết từ mỗi mẫu theo phương pháp Trizol (Chomczynski, Mackey, 1995). mRNA được tinh chế bằng hạt từ gắn Oligo(dT) (Life Techologies). Bộ sinh phẩm Truseq strand mRNA library preparation kit (Illumina) sử dụng để tạo thư viện cDNA. Chất lượng của thư viện cDNA được kiểm tra bằng thiết bị Bioanalyzer sử dụng High Sensitivity Chip (Agilent Technologies). Giải trình tự được tiến hành trên máy giải trình tự gen thế hệ mới Illumina MiSeq. Dữ liệu thu từ máy giải trình tự được lưu trữ theo định dạng FASTQ. Đây là định dạng chuẩn dùng để lưu trữ dữ liệu trình tự bao gồm điểm chất lượng của máy đọc trình tự thế hệ mới (NGS).

Phương pháp tiền xử lý dữ liệu thô

Dữ liệu trình tự đọc thô được đánh giá chất lượng và tiền xử lý bằng phần mềm FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) và Trimmomatic (Bolger *et al*., 2014) (parameters: ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:70) để thu được bộ dữ liệu trình tự đọc tinh sạch. Sau quá trình tiền xử lý, chúng tôi tiếp tục sử dụng FastQC để đánh giá lại chất lượng và kiểm tra khả năng tiền xử lý.

Phương pháp lắp ráp *de novo* hệ phiên mã

Dữ liệu trình tự đọc tinh sạch từ mô cơ và mô gan tụy được lắp ráp *de novo* bằng phần mềm Trinity phiên bản trinityrnaseq_r20140717 (Haas *et al.,* 2013) với tham số mặc định (k_{mer} = 25-mers) thu được hệ phiên mã thô. Để có thể loại bỏ tối đa những trình tự có chất lượng lắp ráp không tốt, chúng tôi tiến hành ánh xạ dữ liệu trình tự đọc tinh sạch vào hệ phiên mã thô bằng phần mềm RSEM 1.2.15 được tích hợp vào Trinity script align_and_estimate_abundance.pl (http://trinityrnaseq.github.io/), từ đó tính toán được số lượng trình tự đọc sử dụng để lắp ráp nên mỗi transcript trong hệ phiên mã thô theo điểm số FPKM (Fragments Per Kilobase of Exon Per Million Fragments Mapped). Những transcript có điểm số FPKM nhỏ hơn 5 sẽ bị loại bỏ khỏi kết quả lắp ráp. Một vấn đề khác có trong dữ liệu hệ phiên mã thô đó là có rất nhiều transcript giống nhau gây nên sự dư thừa dữ liệu, chúng tôi sử dụng đoạn mã Perl tự viết (https://namason.com/code/) để gộp transcript dài nhất trong mỗi nhóm (cluster) transcript định nghĩa bởi Trinity (c*g*), transcript dài nhất này được gọi là unigene. Thông qua 2 bước tinh sạch này, chúng tôi thu được hệ phiên mã tinh sạch bao gồm toàn bộ unigene để sử dụng cho các phân tích tiếp theo.

Nhằm đánh giá chất lượng lắp ráp, dữ liệu trình tự đọc tinh sạch được ánh xạ ngược trở lại vào hệ phiên mã tinh sạch bằng phần mềm Bowtie2 và SAMtools (Li *et al.,* 2009; Langmead, Salzberg, 2012).

Phương pháp chú giải và phân loại unigene trong hệ phiên mã

Chú giải chức năng cho các unigene trong hệ phiên mã đòi hỏi phải sử dụng những thuật toán tìm kiếm tương đồng trên các cơ sở dữ liệu protein quan trọng. Chúng tôi sử dụng công cụ BLAST+ với chương trình BLASTx để so sánh toàn bộ unigene lên các cơ sở dữ liệu NCBI non-redundant protein (Nr, http://www.ncbi.nlm.nih.gov/) và Swiss-Prot (http://www.expasy.ch/sprot) với tham số E-value là 1e-6. Kết quả chú giải từ Ngân hàng gen (vùng lựa chọn Nr) sau đó được phần mềm Blast2GO sử dụng để lấy ra mã Gene Ontology (GO) riêng biệt cho mỗi unigene. Toàn bộ unigene trong hệ phiên mã sẽ được ánh xạ vào các mã GO và phân loại dựa vào 3 hạng mục: quá trình sinh học, thành phần tế bào và chức năng phân tử. Trong nghiên cứu này chúng tôi tập trung vào nghiên cứu sàng lọc unigene tiềm năng liên quan tới tính trạng tăng trưởng.

Phương pháp phân tích biểu hiện hệ phiên mã

Một trong những ứng dụng quan trọng của giải trình tự RNA-seq là phân tích biểu hiện. Chúng tôi tiến hành đo mức độ biểu hiện cho từng unigene trong hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon* bằng phần mềm RSEM (RNA-seq by expectation maximization) để tiến hành ước lượng số lượng unigene biểu hiện theo từng mô (Li, Dewey, 2011). Trình tự đọc được từ mỗi thư viện giải trình tự được ánh xạ ngược trở lại vào bộ dữ liệu unigene tinh sạch bằng "script run_RSEM_align_n_estimate.pl" với tham số mặc định, sau đó tính toán điểm số biểu hiện cho mỗi thư viện giải trình tự bằng "script merge_RSEM_frag_counts_single_table.pl". Bước cuối cùng, chúng tôi sử dụng câu lệnh "run_DE_analysis.pl" được tích hợp sẵn trong gói công cụ EdgeR và được thực thi trên môi trường ngôn ngữ thống kê R (Robinson *et al*., 2010) để tiến hành phân tích biểu hiện khác biệt. Tham số độ tin cậy FDR (False discovery rate) được cài đặt là FDR ≤ 0,001 và giá trị tuyệt đối |log2(Độ sai khác)| ≥ 2 là những tham số được sử dụng để xác định mức độ biểu hiện giữa các thư viện trình tự đọc. Toàn bộ những câu lệnh và script được sử dụng ở trên đều được tích hợp trong bộ phần mềm Trinity (Haas *et al*., 2013).

KẾT QUẢ VÀ THẢO LUẬN

Kết quả tiền xử lý dữ liệu

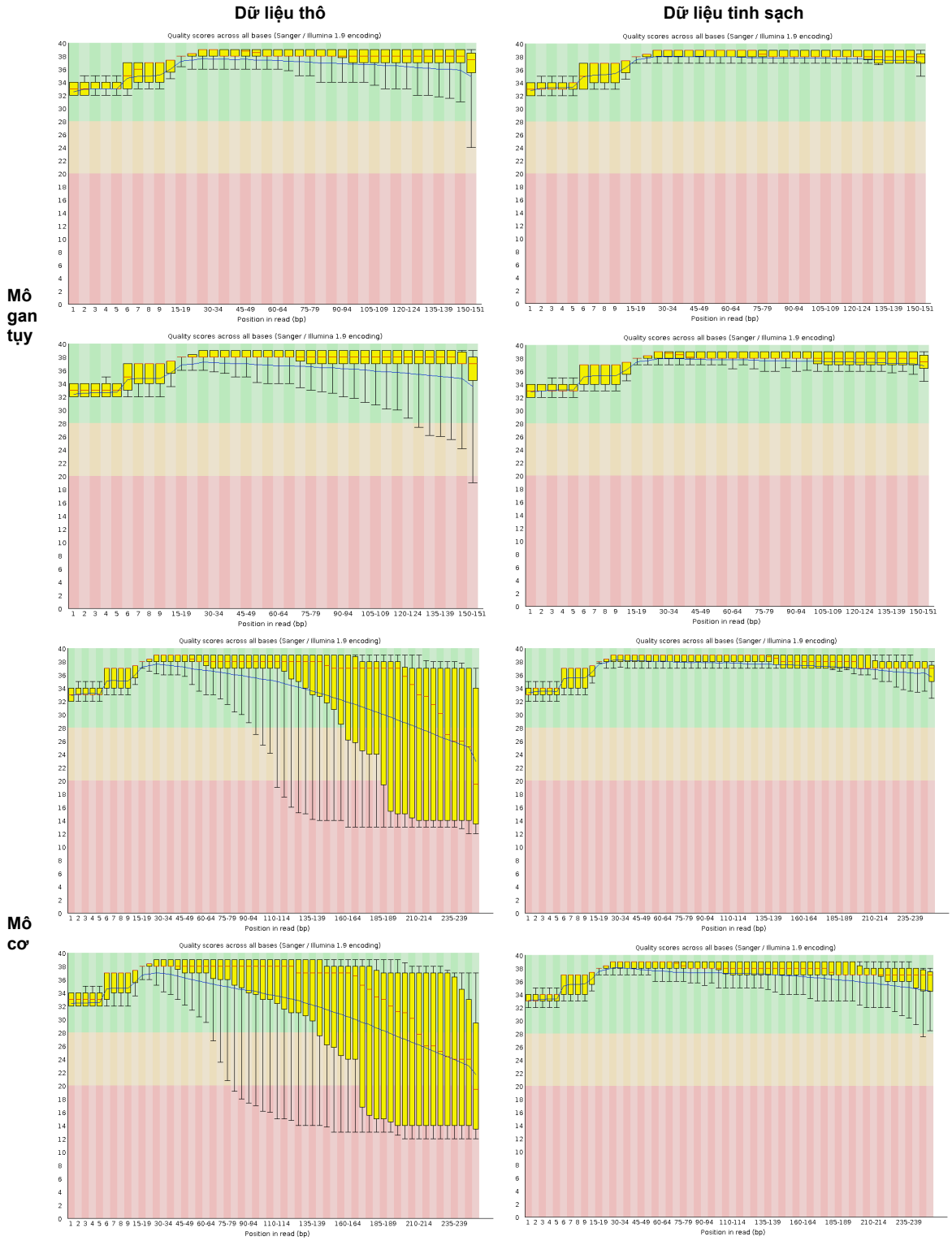
Dữ liệu trình tự đọc thô được đánh giá chất lượng bằng phần mềm FastQC (v0.11.2) và được xử lý loại bỏ đoạn trình tự thừa và chất lượng thấp bằng phần mềm Trimmomatic (v0.32), kết quả thu được với chất lượng thấp nhất với QC là 30 và độ dài trong khoảng từ 70 đến 151 bp đối với mô gan tụy và từ 70 đến 251 bp đối với mô cơ . Kết quả chi tiết và chất lượng của trình tự đọc trước và sau khi xử lý được thể hiện ở bảng 1 và hình 1.

Trục tung của các biểu đồ trong Hình 1 thể hiện điểm chất lượng giải trình tự (quality score). Điểm chất lượng càng cao thể hiện nucleotide tại vị trí đó được giải trình tự chính xác càng cao. Hình nền của biểu đồ được phân thành các màu sắc khác nhau dựa theo trục tung của biểu đồ tương ứng với chất lượng giải trình tự cao (màu xanh lá cây), chất lượng giải trình tự trung bình (màu tím nhạt), chất lượng giải trình tự kém (màu tím).

Phần mềm Trimmomatic được sử dụng để loại bỏ dữ liệu trình tự đọc có chất lượng kém với tham số như sau: tất cả các trình tự đọc có điểm chất lượng nhỏ hơn 30 (QC < 30) và đoạn trình tự có kích thước nhỏ hơn 70 bp sẽ được loại bỏ. Hình 1 (dữ liệu tinh sạch) cho thấy tất cả các đoạn trình tự đều có điểm chất lượng tốt và nằm trong vùng an toàn (vùng màu xanh của biểu đồ). Những kết quả ở Bảng 1 và Hình 1 cho thấy dữ liệu trình tự đọc đạt tiêu chuẩn để tiến hành các bước phân tích tiếp theo.

Bảng 1. Thống kê số lượng, độ dài trình tự đọc theo từng mô .

Mô	Tham số	Trước khi tiền xử lý	Sau khi tiền xử lý	% số đoạn trình tự giữ lại
Mô cơ	Tổng số đoạn trình tự	12.312.819	8.533.944	69,31%
	Độ dài đoạn trình tự	35 - 251 bp	70 - 251 bp	
Mô gan tụy	Tổng số đoạn trình tự	20.512.979	17.964.211	87,57%
	Độ dài đoạn trình tự	35 - 151 bp	70 - 151 bp	
Tổng số đoạn trình tự chất lượng cao của 2 mô	26.498.155 (80,72%)			



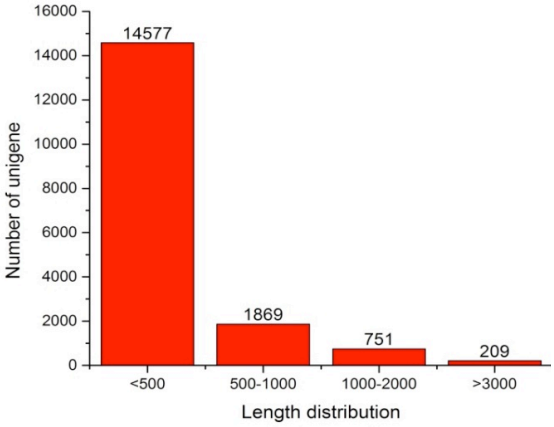
Hình 1. Kết quả đánh giá chất lượng dữ liệu trình tự đọc thô và dữ liệu trình tự đọc tinh sạch ở các mô.

Kết quả lắp ráp hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon*

Dữ liệu trình tự đọc thô sau khi tiền xử lý được lắp ráp bởi phần mềm Trinity thu được hệ phiên mã thô bao gồm 157.995 transcript, trải qua 2 bước loại bỏ những transcript lắp ráp kém chất lượng hoặc những transcript giống nhau, chúng tôi thu được hệ phiên mã tinh sạch với 17.406 unigene (độ dài nhỏ nhất là 201 bp, độ dài lớn nhất là 12.392 bp) với chỉ số N50 là 402 bp và độ dài trung bình là 403,06 bp (Bảng 2). Mặc dù số lượng transcript của hệ phiên mã thô giảm đi trong quá trình tinh sạch để đạt được tập unigene của hệ phiên mã tinh sạch, tỷ lệ % trình tự đọc tinh sạch ánh xạ ngược trở lại hệ phiên mã thô và hệ phiên mã tinh sạch lần lượt là 67,60 % và 64,05 %) (Bảng 2). Phân bố độ dài unigene trong hệ phiên mã tinh sạch được thể hiện như trong Hình 2, chiếm phần lớn là độ dài dưới 500 bp (83,74 % tổng số unigene). Từ 3 tiêu chí là N50, số lượng trình tự đọc sử dụng cho lắp ráp hệ phiên mã và phân bố độ dài unigene trong hệ phiên mã tinh sạch cho thấy chất lượng lắp ráp *de novo* là tương đối tốt.

Bảng 2. Thống kê kết quả số lượng và đặc điểm unigene lắp ráp trong hệ phiên mã tinh sạch từ mô cơ và mô gan tụy tôm sú *Penaeus monodon.*

Các thông số của thống kê	Hệ phiên mã thô	Hệ phiên mã tinh sạch
Số lượng unigene	157.995	17.406
Kích thước hệ phiên mã (bp)	51.854.174	7.015.641
N50 (bp)	314	402
Độ dài trung bình các unigene (bp)	328,20	403,06
Số đoạn trình tự đọc tinh sạch ánh xạ ngược trở lại hệ phiên mã (Tỷ lệ)	17.913.904 (67.60%)	16.971.031 (64.05%)
Unigene ngắn nhất (bp)	201	201
Unigene dài nhất (bp)	12.392	12.392



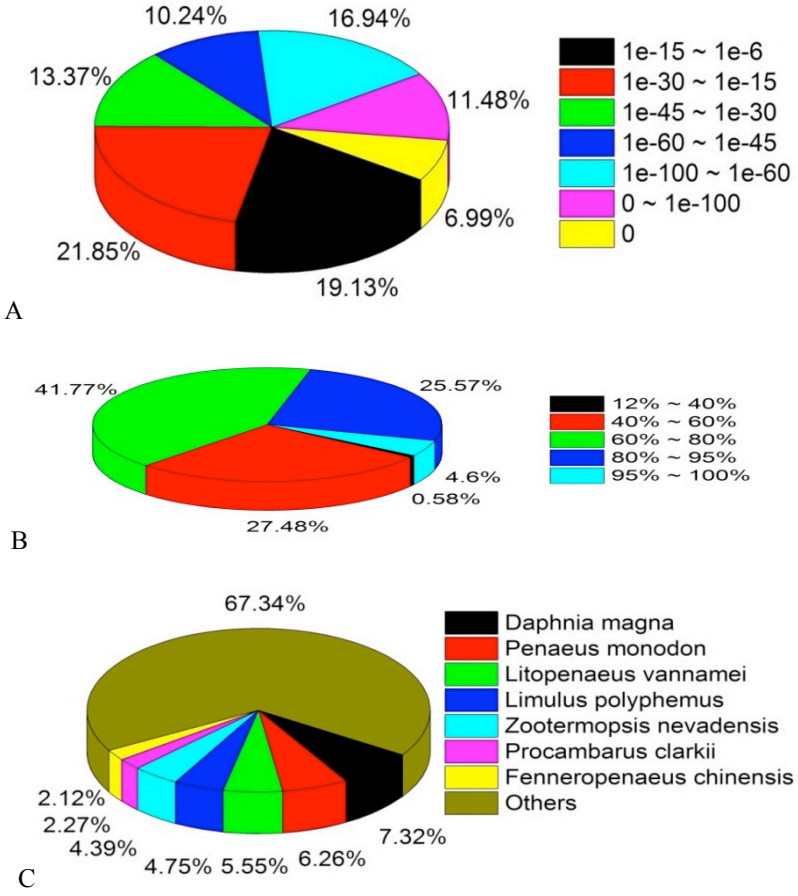
Hình 2. Phân bố độ dài toàn bộ unigene trên hệ phiên mã tinh sạch

Chú giải chức năng hệ phiên mã từ từ mô cơ và mô gan tụy tôm sú *Penaeus monodon*

Quá trình chú giải chức năng bằng BLASTX cho kết quả 1.950 (11,20%) unigene được tìm thấy trên cơ sở dữ liệu nr-NCBI với tham số E-value 1e-6, vì không có hệ gen tham chiếu tôm sú nên sẽ có một lượng lớn unigene không thể chú giải chức năng. Số lượng unigene không được chú giải trong nghiên cứu của chúng tôi có thể là những trình tự transcript mới và đặc hiệu với *Penaeus monodon*. Thêm vào đó, còn có một lý do khác giải thích cho tỷ lệ chú giải chức năng thấp là do các trình tự unigene sau khi lắp ráp có độ dài khá ngắn. Phân bố E-value của các kết quả chú giải chức năng trong nr-NCBI của các unigene cho thấy 59,03% kết quả có giá trị trong khoảng 0 –> 1.0e-30 và 45,66% số lượng trình tự có điểm số E-value cao và tin cậy (E-value < 10^{-45}) (Hình 3A). Những kết quả như vậy đã khẳng định giá trị và độ tin cậy của kết quả lắp ráp *de novo* hệ phiên mã trong nghiên cứu này. Bên cạnh đó, phần lớn các trình tự chú giải trong nr-NCBI của các unigene (71,94%) có độ tương đồng (similarity) lớn hơn 60% và 30,17% số lượng trình tự có độ tương đồng lớn hơn 80% (Hình 3B). Sau khi tìm kiếm tương đồng bằng BLASTX, chúng tôi thống kê phân bố loài trong bộ kết quả tin cậy nhất (E-value thấp nhất) và được thể hiện như trong Hình 3C. Trong kết quả này, loài *Daphnia magna* chiếm số lượng kết quả nhiều nhất với tỷ lệ 7,32%. Trong khi đó kết quả

ứng với tôm sú *Penaeus monodon* là 6,26% và tôm thẻ chân trắng *Litopenaeus vannamei* là 5,55%. Điều này có thể lý giải do dữ liệu về hệ gen tôm trên cơ sở dữ liệu nr-NCBI còn quá ít.

Bên cạnh việc được chú giải bằng cơ sở dữ liệu nr-NCBI, 17.406 unigene của hệ phiên mã tinh sạch lắp ráp từ mô cơ và mô gan tụy của tôm sú *Penaeus monodon* còn được chú giải bằng các cơ sở dữ liệu Swiss-Prot, Gene Ontology và KEGG. Tổng số 1957 unigene đã được chú giải từ những cơ sở dữ liệu này (Bảng 3).

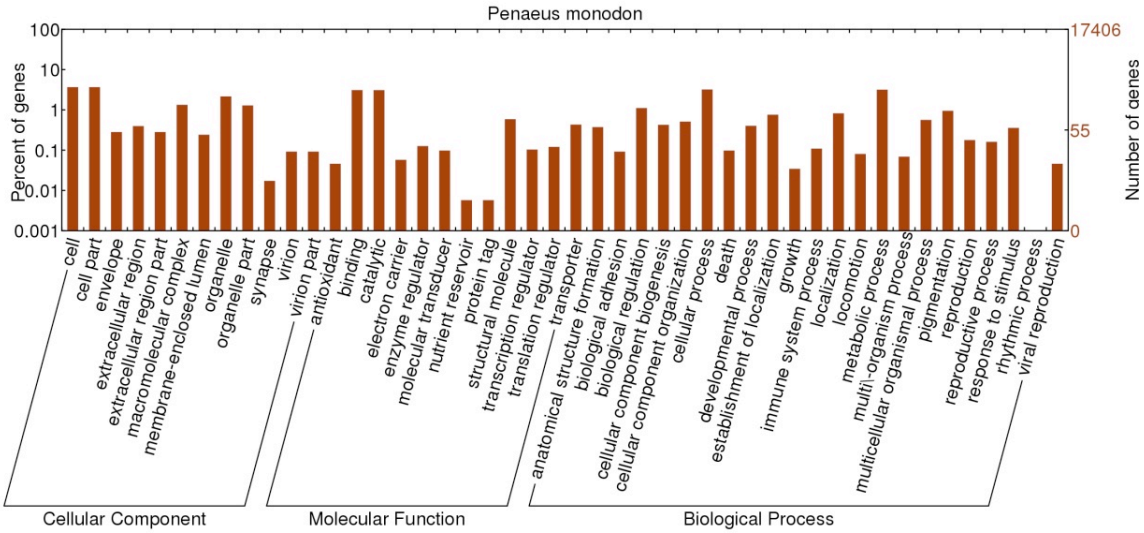


Hình 3. Thống kê kết quả chú giải trên cơ sở dữ liệu nr-NCBI, A: Thống kê phân bố giá trị E-value, B: Thống kê phân bố độ tương đồng, C: Thống kê phân bố loài trong bộ kết quả tin cậy nhất (E-value thấp nhất).

Bảng 3. Thống kê kết quả chú giải hệ phiên mã tôm sú trên các cơ sở dữ liệu.

Cơ sở dữ liệu	Số lượng unigene được chú giải
NR-NCBI	1.950
Swiss-Prot	939
KEGG	865
GO	1.119
Tất cả các cơ sở dữ liệu	1.957
Tổng số unigene	17.406
Tỷ lệ chú giải	11,24%

Bộ dữ liệu unigene tinh sạch sau khi được tìm kiếm tương đồng trên nr-NCBI sẽ được chú giải chức năng theo Gene Ontology (GO) và phân loại vào 3 thư mục: "quá trình sinh học" (Biological Process), "chức năng phân tử" (Molecular Function), "thành phần tế bào" (Cellular Component). Thông qua phần mềm Blast2GO, chúng tôi tiến hành chú giải chức năng trên ngân hàng Gene Ontology và thu được 1.119 unigene mang các mã chức năng Gene Ontology được phân vào 46 nhóm chức năng (Hình 4). Chú giải GO đã cung cấp thông tin tổng quan về chức năng hệ phiên mã thu được từ mô cơ và mô gan tụy tôm sú.



Hình 4. GO phân loại các trình tự lắp ráp. Tổng số 1.119 unigene đã được nhóm lại thành 3 nhóm GO chính: 'Biological Processes', 'Cellular Component', và 'Molecular Function'.

Sàng lọc các unigen liên quan đến trính trạng tăng trưởng từ hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon*

Hệ phiên mã được chú giải của tôm sú *Penaeus monodon* sẽ là nguồn tài nguyên quan trọng cho việc sàng lọc các gen ứng viên liên quan đến những tính trạng quan trọng của tôm sú, đặc biệt là khi so sánh với các phương pháp truyền thống trong việc phân lập các gen chưa biết trình tự bằng việc thiết kế mồi suy diễn (degenerate PCR). Bằng việc tổng quan tài liệu từ các công trình khoa học công bố thuộc lĩnh vực sinh học phân tử tôm, các nhà khoa học nhận thấy các gen ứng viên liên quan đến tính trạng tăng trưởng ở tôm thường được biểu hiện ở mô cơ và mô gan tụy (Jung *et al.,* 2013). Đây cũng chính là lý do chúng tôi đã sử dụng gói dữ liệu giải trình tự từ mô cơ và mô gan tụy của tôm sú *Penaeus monodon* phân lập được từ vùng biển Bắc Trung Bộ Việt Nam để lắp ráp *de novo* hệ phiên mã, chú giải chức năng và sàng lọc các unigene liên quan đến tính trạng tăng trưởng. Quá trình sàng lọc các unigene liên quan đến tính trạng tăng trưởng được thực hiện dựa trên các nguyên lý của Jung *et al.* (2013), đó là: (i) mối liên quan giữa các gen và tính trạng tăng trưởng đã được công bố trong nhóm giáp xác; (ii) các gen liên quan đến tính trạng tăng trưởng trong quá trình lột xác ở tôm; (iii) các gen phân giải và phát triển hệ cơ liên quan trong quá trình lột xác.

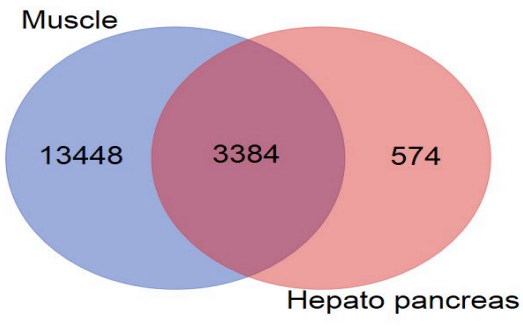
Từ hệ phiên mã lắp ráp và chú giải, chúng tôi sàng lọc được 51 unigene liên quan đến tính trạng tăng trưởng được phân bố trong 18 nhóm (Bảng 4). Có 8 nhóm gen được sàng lọc liên quan đến quá trình phân giải và phát triển của hệ cơ trong quá trình lột xác, đó là các nhóm gen: Actin, Profilin, Myosin, Alpha skeletal muscle, Calponin/calponintransgelin, Tropomyosin, Muscle lim protein and Lim domain binding, đây cũng là những gen đặc trưng cho mô cơ của tôm sú. Ngoài ra có 3 nhóm gen liên quan đến tính trạng tăng trưởng đặc trưng cho mô gan tụy đó là Alpha-amylase, Fatty acid binding protein, Cathepsin L; đây là những gen mã hóa cho những enzyme đóng vai trò quan trọng trong quá trình trao đổi vật chất ở tôm sú, đặc biệt là trong việc chuẩn bị nguồn vật chất cho chu kỳ lột xác tiếp theo ở tôm sú. Trong tương lai chúng tôi có dự định sẽ nghiên cứu mối liên quan giữa các gen ứng viên này với tính trạng tăng trưởng của tôm sú phân lập tại Việt Nam.

Bảng 4. Liệt kê 51 unigene liên quan đến tính trạng tăng trưởng.

STT	Các nhóm gen ứng viên	Unigene IDs
1.	Alpha-amylase	c83210_g1_i1, c44070_g1_i1, c50035_g1_i1, c61443_g1_i1
2.	Cathepsin L	c61287_g1_i1, c62382_g1_i2
3.	Cyclophilin	c19823_g1_i1
4.	Fatty acid-binding protein	c41270_g1_i1, c41041_g1_i1, c61108_g1_i1
5.	Fibrillarin	c43879_g1_i1
6.	Glyceradehyde-3-phosphate dehydrogenase (GAPDH)	c62621_g1_i1
7.	Profilin	c41374_g1_i1
8.	Growth hormone and insulin-like growth factor	c62969_g1_i1, c19902_g1_i1, c54868_g1_i1
9.	Secreted Protein Acidic and Rich in Cysteine (SPARC)	c60039_g1_i1
10.	Methyl farnesoate and farnesoic acid O-methyltransferase	c60754_g1_i1, c61318_g1_i2
11.	Ecdysteroid	c50607_g1_i1
12.	Calponin/calponintransgelin	c13961_g1_i1, c51091_g1_i1
13.	Tropomyosin	c165984_g1_i1, c54212_g1_i2
14.	Muscle LIM protein	c62133_g1_i1, c62133_g2_i1, c62133_g3_i1, c43449_g1_i1, c56823_g1_i1
15.	Alpha skeletal muscle	c41556_g1_i1, c37833_g1_i2, c53843_g1_i1, c53843_g2_i1
16.	Lim domain binding	c56793_g1_i2, c60234_g1_i2, c61458_g1_i2
17.	Actin	c62336_g3_i2, c106986_g1_i1, c166206_g1_i1, c53399_g1_i1, c151792_g1_i1, c175914_g1_i1
18.	Myosin heavy chain	c62492_g1_i1, c62492_g3_i1, c66492_g1_i1, c167495_g1_i1, c372_g1_i1, c20008_g1_i1, c22261_g1_i1, c32014_g1_i1, c43972_g1_i1

Phân tích biểu hiện hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon*

Ánh xạ dữ liệu trình tự RNA-seq được thực hiện với phần mềm RSEM (Li, Dewey, 2011) để từ đó tính toán được mức độ biểu hiện trên mỗi unigene đặc trưng cho từng mô. Kết quả ánh xạ cho thấy có 13.448 unigene biểu hiện đặc trưng cho mô cơ, 574 unigene biểu hiện đặc trưng cho mô gan tụy, 3.384 unigene biểu hiện ở cả mô cơ và mô gan tụy trong tổng số 17.406 unigene của hệ phiên mã tinh sạch (Hình 5). So sánh biểu hiện hệ phiên mã mô cơ vàmô gan tụy cho thấy có 16.184 unigene trong tập 17.406 unigene có biểu hiện khác biệt giữa 2 mô, được gọi là DEG (differentially expressed genes) với tham số độ tin cậy FDR $\leq$ 0,001. Trong số 16.184 unigene này chỉ có 1.400 unigene được chú giải, nguyên nhân là do thông tin về hệ gen của tôm sú đã được công bố là rất ít. Số lượng các unigene biểu hiện tăng và giảm giữa 2 mô cho thấy có 14.599 unigene biểu hiện tăng trong mô cơ so với mô gan tụy và 1.585 unigene biểu hiện tăng ở mô gan tụy so với mô cơ với giá trị tuyệt đối $|\log_2(\text{Độ sai khác biểu hiện})| \geq 2$.



Hình 5. Số lượng unigene biểu hiện đặc trưng ở mô cơ (muscle) và mô gan tụy (hepatopancreas) trong tập 17.406 unigene.

KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã lắp ráp *de novo* và phân tích hệ phiên mã từ mô cơ và mô gan tụy tôm sú *Penaeus monodon* thu được số lượng unigene của hệ phiên mã thô là 157.995 và hệ phiên mã tinh sạch là 17.046 unigene, chú giải được 1.957 unigene, cung cấp thông tin tổng quan về chức năng hệ phiên mã thu được từ mô cơ và mô gan tụy tôm sú. Đặc biệt chúng tôi đã sàng lọc được 51 unigene liên quan đến tính trạng tăng trưởng. Ngoài ra, phân tích biểu hiện cho thấy có sự khác biệt về biểu hiện của các unigene giữa 2 mô. Đây là những kết quả ban đầu góp phần hiểu biết tổng quan về hệ phiên mã từ mô cơ và mô gan tụy của tôm sú, từ đó làm cơ sở cho các nghiên cứu sâu hơn về hệ phiên mã của loài này, đặc biệt là những nghiên cứu về ánh xạ tính trạng hay chọn giống dựa trên các chỉ thị phân tử. Kết quả từ nghiên cứu khoa học công nghệ nền công bố ở đây tạo cơ sở định hướng ứng dụng lâu dài với hiệu quả kinh tế có thể tính đến trong những giai đoạn sau.

Lời cảm ơn: *Công trình này được thực hiện với sự tài trợ kinh phí của Bộ Khoa học và Công nghệ thông qua nhiệm vụ "Lập bản đồ gen tôm sú (Penaeus monodon)". Mã số nhiệm vụ: NVQG-2011/24.*

TÀI LIỆU THAM KHẢO

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114–2120.

Brown TA (2002) *Chapter 3. Transcriptomes and Proteomes. Genomes*, 2nd ed. Oxford: Wiley-Liss.

Chomczynski P, Mackey K (1995) Short technical report. Modification of the TRIZOL reagent procedure for isolation of RNA from Polysaccharide-and proteoglycan-rich sources. *Biotechniques* 19(6): 942-945.

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* 36: 3420–3435.

Guo Q, Ma X, Wei S, Qiu D, Wilson IW, Wu P, Tang Q, Liu L, Dong S, Zu W (2014) De novo transcriptome sequencing and digital gene expression analysis predict biosynthetic pathway of rhynchophylline and isorhynchophylline from *Uncaria rhynchophylla*, a non-model plant with potent anti-alzheimer's properties. *BMC Genomics* 15: 676.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8: 1494–1512.

Jung H, Lyons RE, Hurwood DA, Mather PB (2013) Genes and growth performance in crustacean species: a review of relevant genomic studies in crustaceans and other taxa. *Rev Aquac* 5: 77–110.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Li Q, Liu J, Zhang L, Liu Q (2014) De novo transcriptome analysis of an aerial microalga *Trentepohlia jolithus*: pathway description and gene discovery for carbon fixation and carotenoid biosynthesis. *PloS One* 9: e108488.

Liu S, Wei W, Chu Y, Zhang L, Shen J, An C (2014) De novo transcriptome analysis of Wing development-related signaling pathways in *Locusta migratoria* Manilensis and *Ostrinia furnacalis* (Guenee). *PloS One* 9: e106770.

Liu Y, Huang Z, Ao Y, Li W, Zhang Z (2013) Transcriptome Analysis of Yellow Horn (*Xanthoceras sorbifolia* Bunge): A Potential Oil-Rich Seed Tree for Biodiesel in China. *PloS One* 8.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139-140.

Rosenberry B (2004) *World shrimp farming 2004. In Shrimp News International*. San Diego, California, USA.

Sookruksawong S, Sun F, Liu Z, Tassanakajon A (2013) RNA-Seq analysis reveals genes associated with resistance to Taura syndrome virus (TSV) in the Pacific white shrimp *Litopenaeus vannamei*. *Dev Comp Immunol* 41: 523–533.

Wang S, Wang X, He Q, Liu X, Xu W, Li L, Gao J, Wang F (2012) Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant Cell Reports* 31: 1437–1447.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57–63.

Xue S, Liu Y, Zhang Y, Sun Y, Geng X, Sun J (2013) Sequencing and *De Novo* Analysis of the Hemocytes Transcriptome in *Litopenaeus vannamei* response to White Spot Syndrome Virus Infection. *PLoS One 8*: e76718.

TRANSCRIPTOME ANALYSIS AND SCREENING OF SOME GROWTH-RELATED PUTATIVE GENES OF BLACK TIGER SHRIMP (*PENAEUS MONODON*)

Nguyen Hai Bang[1], Pham Quang Huy[2], Tran Xuan Thach[2], Nguyen Giang Thu[3], Nguyen Thi Minh Thanh[2], Nguyen Thi Hoa[2], Ha Thi Thu[2], Nguyen Thi Tuyet Nhung[2], Nguyen Cuong[2], Nguyen Huu Ninh[4], Dong Van Quyen[2], Chu Hoang Ha[2], Dinh Duy Khang[2]

[1]*Hai Phong University for Medicine and Pharmacy*
[2]*Institute of Biotechnology, Vietnam Academy of Science and Technology*
[3]*Science Technology and Environmental Department, MARD*
[4]*Research Aquaculture Institute III, MARD*

SUMMARY

Black tiger shrimp (*Penaeus monodon)* is an aquaculture species with a great economic potential for our country. In the recent years, the export revenue from Black tiger shrimp has reached nearly a billion USD per year. Our national development strategy is to achieve stable, sustainable shrimp production with minimal negative environmental impact. A cornerstone for this strategy is the development of domesticated stocks of *P. monodon* and rational breeding programs for improved survival and growth. However, the genomic and transcriptomic data of Black tiger shrimp are not well documented until now. It makes us facing a lot of difficulties in the trait mapping and marker-assisted breeding for important traits, such as fast growth and disease resistance. Sequencing and analysis of *P. monodon* transcriptome will provide important data for shrimp breeding. In this study, NGS data from two transcriptome libraries of muscle and hepatopancreas tissues of *P. monodon* collected from North Central Coast of Vietnam were undergone pre-processing and *de novo* assembling. After transcript refinement, we obtained a final set of 17,406 unigenes (N50 of 402 bp, average length of 403.06 bp). Comparisons of the assembled unigenes against four public protein databases, a set of 51 unigenes related to growth were identified. The expression analysis revealed 16,184 unigenes differentially expressed in the two tissues. The new data obtained in this study provide a valuable information on the *P. monodon* transcriptome and play an important role for the further research, especially for screening important markers linked with economically important traits of Black tiger shrimp.

Keywords: *Black tiger shrimp Penaeus monodon, transcriptome, unigenes related to growth*