TAP CHI SINH HOC 2015, 37(2): 220-227
DOI: 10.15625/0866-7160/v37n2.6427

PHÂN TÍCH HỆ GEN CHỨC NĂNG TỪ MÔ THẬN CÁ TRA (*Pangasianodon hypophthalmus*) NUÔI Ở ĐIỀU KIỆN MẶN: LẮP RÁP, CHÚ GIẢI, PHÂN TÍCH CHỈ THỊ SNP

Nguyễn Minh Thành[1]*, Võ Thị Minh Thư[1], Hyungtaek Jung[2], Peter Mather[2]

[1]Trường Đại học Quốc tế, ĐHQG HCM, *nmthanh@hcmiu.edu.vn
[2]Queensland University of Technology (QUT)

TÓM TẮT: Cá Tra là đối tượng thủy sản nước ngọt quan trọng có giá trị kinh tế ở Đồng bằng sông Cửu Long. Nghiên cứu của chúng tôi áp dụng kỹ thuật giải trình tự Ion Torrent nhằm xây dựng cơ sở dữ liệu EST từ mô thận của cá tra nuôi ở độ mặn 9 ppt. Kết quả giải trình tự đạt được 2.623.929 đoạn trình tự có chiều dài trung bình là 104 bp sau khi sàng lọc loại bỏ các đoạn trình tự có chất lượng thấp. Các đoạn trình tự được lắp ráp thành contig sử dụng các phần mềm lắp ráp CLC Genomic Workbench, Trinity và Velvet/Oases, trong đó CLC là chương trình lắp ráp tối ưu nhất. Kết quả lắp ráp sử dụng CLC đạt được 29.940 contig và xác định được 5.710 gen giả định khi so sánh với cơ sở dữ liệu của NCBI. Ngoài ra nghiên cứu của chúng tôi cũng phát hiện được số lượng lớn SNP. Kết quả nghiên cứu của chúng tôi là cơ sở dữ liệu chi tiết về hệ gen chức năng của cá tra cho đến thời điểm hiện tại.

Từ khóa: Pangasianodon hypophthalmus, hệ gen chức năng, mô thận, tính trạng chịu mặn

MỞ ĐẦU

Cá tra (*Pangasianodon hypophthalmus*) là đối tượng thủy sản nước ngọt có giá trị kinh tế cao ở Đồng bằng sông Cửu Long (ĐBSCL). Năm 2014, sản lượng cá tra đạt hơn 1,1 triệu tấn và kim ngạch xuất khẩu ước tính đạt khoảng 1,77 tỷ USD [28]. Chương trình chọn giống cá tra do Viện Nghiên cứu Nuôi trồng Thủy sản II thực hiện tạo ra giống cá tra có tốc độ tăng trưởng nhanh và tỷ lệ phi lê cao, đáp ứng sự phát triển vượt bậc của nghề nuôi cá tra trong những năm qua [25, 26]. Tuy nhiên, nghề nuôi cá tra đang đối mặt với nhiều thách thức mới, trong đó sự xâm nhập mặn ngày càng lan rộng ở nhiều vùng của ĐBSCL do tác động của biến đổi khí hậu là vấn đề cần quan tâm. Điều này cho thấy nhu cầu con giống cá tra có khả năng chịu mặn trở nên cấp thiết để thích nghi với vùng nuôi bị nhiễm mặn. Phương pháp chọn giống MAS (marker-assisted selection) dựa vào các chỉ thị phân tử và gần đây là phương pháp chọn giống GS (genomic selection) là những phương pháp chọn giống hiện đại có thể nâng cao hiệu quả chọn giống trong thời gian ngắn [3]. Để có thể ứng dụng phương pháp chọn giống hiện đại, việc xây dựng cơ sở dữ liệu thông tin di truyền của cá tra liên quan đến tính trạng chịu mặn là bước đi cần thiết đầu tiên. Tuy nhiên, cơ sở dữ liệu ở mức độ phân tử đối với cá tra còn rất hạn chế. Hiện nay chỉ có các công bố sử dụng chỉ thị microsatellite nghiên cứu quần đàn cá tra tự nhiên và gia hóa [9, 20, 21] và nghiên cứu định danh các loài cá da trơn bằng mã vạch DNA [31]. Kỹ thuật giải trình tự gen thế hệ mới đã mở ra nhiều cơ hội nghiên cứu hệ gen DNA (genome) và hệ gen chức năng RNA (transcriptome) dễ dàng hơn và đã được ứng dụng nghiên cứu hệ gen cho hơn 30 đối tượng thủy sản có giá trị kinh tế [18]. Trong đó nghiên cứu hệ gen chức năng RNA đơn giản hơn, giúp hiểu biết chi tiết các chức năng sinh học ở mức độ phân tử và có thể xác định được các gen tiềm năng liên quan đến tính trạng quan tâm [29].

Mô thận là một trong các mô chính tham gia điều hòa áp suất thẩm thấu ở cá nước ngọt thích nghi với môi trường nước lợ mặn [14]. Vì vậy, nghiên cứu của chúng tôi lựa chọn mô thận để phân tích hệ gen chức năng liên quan đến tính trạng chịu mặn của cá tra bằng kỹ thuật giải trình tự gen thế hệ mới Ion Torrent. Các trình tự EST được kết nối thành contig bằng các phần mềm khác nhau và chú giải chức năng giả định. Các đoạn trình tự được so sánh với cơ sở dữ liệu của NCBI (National Center for Biotechnology Information) để xác định các

nhóm protein và gen tiềm năng ảnh hưởng đến khả năng chịu mặn của cá tra. Ngoài ra nghiên cứu cũng xác định được số lượng lớn chỉ thị phân tử SNP (single nucleotide polymorphism) có thể ứng dụng cho các nghiên cứu khác ở mức độ phân tử trên cá tra và cá da trơn.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Mẫu thí nghiệm

Nghiên cứu cá tra tăng trưởng được thực hiện tại Khu thí nghiệm Công nghệ sinh học, Trường Đại học Quốc tế. Cá tra giống (8-10g/con) được nuôi trong các bể composite 500L ở 4 độ mặn (6, 9, 12 và 15‰) và đối chứng (0‰) trong thời gian 6 tuần. Kết quả thí nghiệm cho thấy, cá tra thích nghi tốt ở độ mặn 9‰ dựa vào so sánh tốc độ tăng trưởng của cá nuôi ở điều kiện 9‰ không có sự khác biệt với tốc độ tăng trưởng của cá nuôi ở điều kiện nước ngọt. Vì vậy, chúng tôi thu mẫu mô thận từ cá tra nuôi ở độ mặn 9‰, bao gồm 3 cá thể tăng trưởng nhanh và 3 cá thể tăng trưởng chậm nhằm đa dạng hóa nguồn mẫu vật và tăng cơ hội phát hiện các đoạn gen hiếm liên quan đến khả năng chịu mặn của cá tra. Mẫu mô được bảo quản trong RNAlater cho đến khi tách RNA.

Tách RNA tổng số và phân tách mRNA

Mẫu được nghiền đồng nhất trong nitơ lỏng, xử lý trong TRIzol/Chloroform (Invitrogen) [2] để tách RNA tổng số. Chúng tôi sử dụng Turbo DNA-free kit (Ambion) để loại bỏ gDNA lẫn trong hỗn hợp RNA. Sau đó hỗn hợp RNA tổng số được tinh sạch bằng RNeasy mini kit (Qiagen). Sau khi tinh sạch, RNA tổng số được định tính và định lượng bằng Qubit 2.0 (Invitrogen) và Bioanalyser (Agilent). Trước khi tách mRNA, RNA tổng số từ nhiều cá thể được trộn lẫn nhau để tăng mức độ đa dạng của mRNA sau khi tách. mRNA được tách khỏi hỗn hợp RNA tổng số bằng Dynabeads mRNA purification kit (Invitrogen) theo hướng dẫn của nhà sản xuất. mRNA tiếp tục được định tính và định lượng bằng Bioanalyser.

Tổng hợp cDNA và giải trình tự bằng Ion Torrent

mRNA được cắt thành đoạn có kích thước 100-200 bp bằng Ion Total RNA-Seq kit (Life Technologies). Các đoạn mRNA được tinh sạch bằng RiboMinus Concentration Module (Invitrogen), sau đó được sử dụng làm khuôn mẫu để tổng hợp cDNA bằng Ion Total RNA-Seq kit (Life Technologies) theo hướng dẫn của nhà sản xuất. cDNA được định lượng bằng Qubit 2.0 và Bioanalyser. Nghiên cứu chuẩn bị các khuôn mẫu (template) bằng Ion OneTouch Template kit (Life Technologies) và sử dụng chip 316, hóa chất Ion PGM^{TM} 200 sequencing kit cho thiết bị Ion Torrent để giải trình tự. Giải trình tự thực hiện tại Molecular Genetics Research Laboratory của QUT, Brisbane, Ôxtrâylia.

Lắp ráp các đoạn trình tự (de novo assembly)

Sau khi giải trình tự bằng thiết bị Ion Torrent, các đoạn trình tự được sàng lọc để loại bỏ các adapter, đoạn trình tự có chất lượng thấp và đoạn trình tự ngắn (<20 bp) thông qua máy chủ (server) của Ion Torrent. Kết quả giải trình tự được truy xuất ở định dạng FastQ và kiểm tra chất lượng bằng chỉ số Q >20. Sau đó các đoạn trình tự được kết nối (assembly) thành các đoạn contig dựa vào định dạng loài mới (*de novo*) chưa có genome tham khảo bằng phần mềm CLC Genomic Workbench (v6.0.4), Velvet/ Oases [23] và Trinity (r2013-08-14) [8]. Đối với phần mềm CLC, k-mer được sử dụng là 20 sau khi lắp ráp với nhiều k-mer khác nhau từ k=20 đến k=60. Tương tự, k-mer sử dụng cho phần mềm Velvet/Oases là 21 sau khi lắp ráp từ k=21 đến k=71. Các chỉ số được sử dụng để đánh giá phần mềm kết nối bao gồm số lượng contig, chiều dài contig N50, chiều dài trung bình của contig, và chiều dài của contig dài nhất. Nghiên cứu chỉ sử dụng kết quả kết nối từ phần mềm cho kết quả kết nối tốt nhất (cụ thể là CLC Genomic Workbench) cho các phân tích tiếp theo.

Chú giải các đoạn trình tự mRNA (annotation) và phân loại nhóm gen chức năng

Chúng tôi sử dụng công cụ BlastX để so sánh các contig với cơ sở dữ liệu KOG (eukaryotic orthologous groups) (giá trị E<1e^{-10}) nhằm phân chia các contig theo nhóm gen chức năng. Cơ sở dữ liệu KOG là một thành phần của cơ sở dữ liệu COG (clusters of orthologous

groups) [27].

Phân tích chỉ thị phân tử SNP

Chúng tôi sử dụng BWA [15] và SAMtools [16] để phát hiện SNP. SNP được xác định khi gióng hàng các contig và sự sai khác nucleotide được phát hiện trên ít nhất bốn trình tự (read) [6]. Tương tự sự thêm đoạn (insertion) hoặc mất đoạn (deletion) trình tự được xác định là indel khi gióng hàng các contig và phát hiện đoạn sai khác trên ít nhất bốn trình tự [6].

KẾT QUẢ VÀ THẢO LUẬN

Giải trình tự Ion Torrent và kết nối các đoạn trình tự (de novo assembly)

Kết quả giải trình tự mRNA của mô thận bằng Ion Torrent đạt được dữ liệu 378,14 Mbp với tổng số EST là 2.873.310 có độ dài trung bình là 140 bp. Sau khi sàng lọc loại bỏ các đoạn adapter, các đoạn trình tự chất lượng thấp và đoạn ngắn, dữ liệu còn lại đạt 272,73 Mbp, tổng số EST là 2.623.929 và độ dài trung bình là 104 bp (bảng 1). Độ dài trung bình của các EST của nghiên cứu này hoàn toàn tương tự với những công bố trước đây khi sử dụng các kỹ thuật nền khác nhau để giải trình tự DNA hoặc mRNA. Độ dài trung bình 104 bp giải mã từ Ion Torrent ngắn hơn so với các đoạn gen giải mã bằng 454 [11, 13] nhưng dài hơn các đoạn gen giải mã bằng Illumina [12, 17].

Bảng 1. Tóm tắt giải trình tự Ion Torrent

Chỉ số phân tích	Giá trị
Tổng số base (Mbp)	378,14
Tổng số base đạt chuẩn > Q20 (Mbp)	319,35
Số lượng đoạn trình tự (read)	2.873.310
Chiều dài trung bình các đoạn trình tự (bp)	140
Tổng số base sau khi sàng lọc (Mbp)	272,73
Tổng số đoạn trình tự sau khi sàng lọc sử dụng cho kết nối	2.623.929
Chiều dài trung bình các đoạn trình tự sau sàng lọc (bp)	104

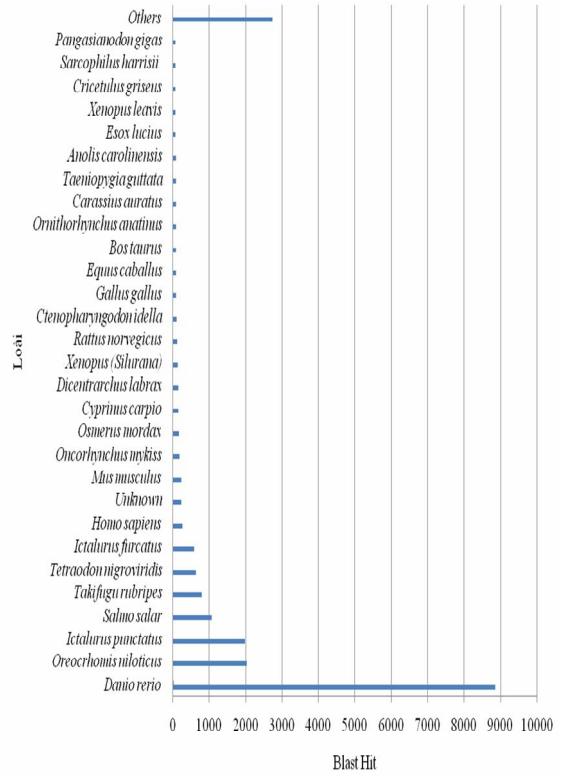
Bảng 2. Kết quả kết nối contig bằng các phần mềm chuyên dụng

Chỉ số phân tích	CLC	Trinity	Velvet/Oases
Tổng số contig	29.940	47.964	36.512
Tổng số base của contig	12.392.014	17.322.804	11.116.409
Số lượng contig ≥ 1.000 bp	6.089	744	1.172
Chiều dài contig N50 (bp)	417	371	372
Chiều dài trung bình (bp)	414	361	304
Chiều dài contig lớn nhất (bp)	3.462	2.571	14.498
Contig có ý nghĩa*	18.199 (60,78%)	27.137 (56,58%)	15.948 (43,68%)
Độ bao phủ (coverage) (x)	15,72	12,74	17,53

Contigs có giá trị E < $1e^{-5}$ khi so sánh với cơ sở dữ liệu NR (non-redundant) khi sử dụng BlastX.

Lựa chọn phần mềm kết nối phù hợp cho kết quả kết nối tin cậy là điểm then chốt trong phân tích hệ gen của các loài chưa có hệ gen tham chiếu. Phần mềm kết nối tối ưu là phần mềm sử dụng gần như hoàn toàn các đoạn trình tự để kết nối thành các contig [32]. Phần mềm Trinity đáp ứng được tiêu chí này khi sử dụng tổng số base lớn nhất (17.322.804 bp) và cho kết quả số lượng contig nhiều nhất (47.964 contig). Một điều cần lưu ý là phân tích hệ gen chức năng khác với phân tích hệ gen DNA. Một bản mã (transcript) có thể có nhiều phiên bản (variant) [7] và các đoạn trình tự có thể kết nối thành contig mặc dù các đoạn này không có nguồn gốc từ một gen [10]. Kết quả này sẽ không phù hợp với phân tích chú giải tiếp theo để tìm ra các gen chức năng. Vì vậy, tiêu chí số lượng contig lớn không phải là tiêu chí tối ưu để

lựa chọn phần mềm kết nối phù hợp. Theo quan điểm của tác giả Liu et al. (2013) [17] chiều dài contig N50 và chiều dài trung bình là tiêu chí chuẩn để đánh giá phần mềm kết nối. Phần mềm CLC cho kết quả phân tích đạt được các tiêu chí này (bảng 2). Ngoài ra phần mềm CLC cũng cho kết quả tỷ lệ contig tương đồng với các trình tự của cơ sở dữ liệu NR cao nhất (60,78%) khi sử dụng BlastX. Đây cũng là một tiêu chí sử dụng để đánh giá phần mềm kết nối [32]. Phần mềm CLC đạt được nhiều tiêu chí đánh giá phần mềm tin cậy so với Trinity và Velvet/Oases, vì vậy, kết quả kết nối từ phần mềm CLC được sử dụng cho các phân tích tiếp theo. Số lượng contig kết nối là 29.940, trong đó contig có chiều dài 300-600 bp là 26.115 (87,22%) và số lượng contig lớn hơn 1.500 bp là 259 (0,87%).



Hình 1. Số lượng contig tương đồng với top 30 loài dựa trên phân tích BlastX

Phân tích so sánh EST

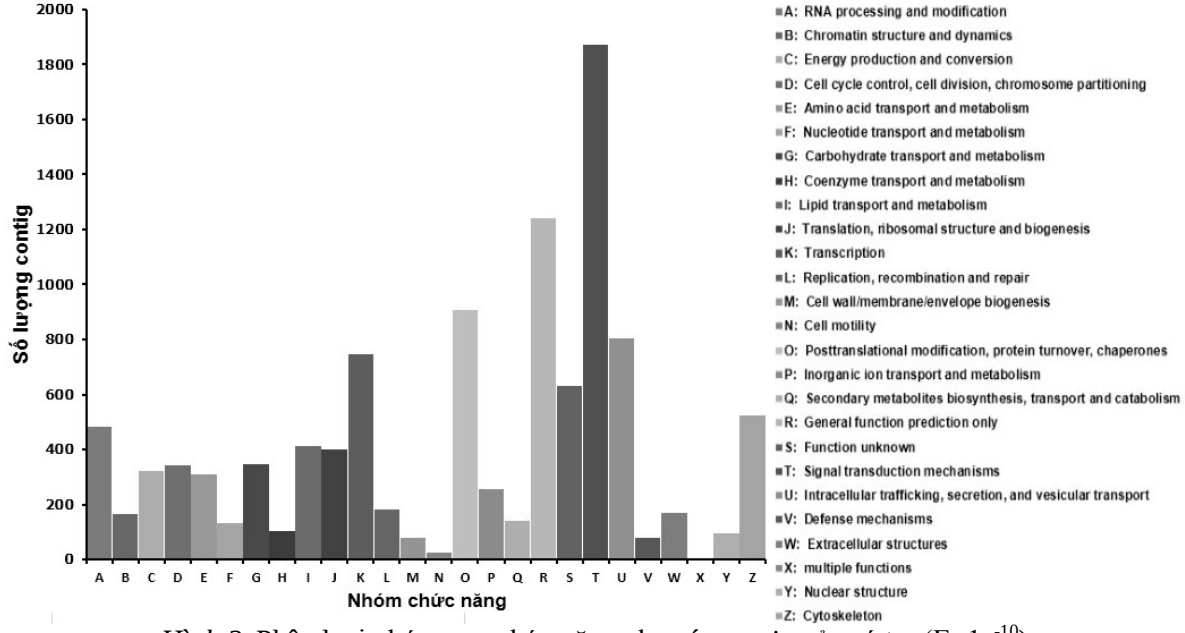
Tổng số 18.199 contig cá tra (60,78%) có trình tự nucleotide tương đồng với các trình tự được lưu trữ ở GenBank ($E<1e^{-5}$) khi sử dụng BlastX, trong đó 79,4% contig cá tra tương đồng với trình tự nucledotide của các lòai cá xương (hình 2). Kết quả này tương tự với những công bố trước đây khi nghiên cứu hệ gen chức năng trên các loài cá xương [22, 24]. Từ kết quả phân tích BlastX và loại trừ các trình tự được chú giải lặp lại cũng như protein của ribosome, số lượng contig được xem là gen giả định (putative gene) của hệ gen chức năng cá tra là 5.710 gen.

Các loài có mức độ tương đồng cao nhất với cá tra bao gồm zebrafish *Danio rerio* (8.856 contig), rô phi vằn *Oreochromis niloticus* (2.020 contig), cá nheo *Ictalurus punctatus* (1.984 contig), cá hồi đại dương *Salmo salar* (1.062 contig) và cá nóc *Takifugu rubripes* (785 contig) (hình 1). Đối với loài cá da trơn, cá nheo *Ictalurus punctatus* và *I. furcatus* có cơ sở dữ liệu gen rất lớn, bao gồm gần 500.000 EST [17, 30], 431.004 trình tự nucleotide và 17.204 trình tự protein được lưu trữ trên GenBank (truy cập ngày 29/5/2015). Tuy nhiên, nghiên cứu của chúng tôi chỉ có 1984 contig cá tra (9,3%) tương đồng với cá nheo *I. punctatus*. Ngoài ra nghiên cứu cũng xác định được tỷ lệ thấp contig cá tra (0,29%) tương đồng với trình tự cá tra dầu *Pangasianodon gigas*, là loài có quan hệ tiến hóa gần gũi với cá tra. Nghiên cứu hiện tại cũng không có contig tương đồng với loài nào thuộc giống *Pangasius*. Điều này có thể được giải thích là do số lượng rất hạn chế các đoạn trình tự của giống *Pangasinodon* và *Pangasius* có sẵn được lưu trữ trên GenBank (818 trình tự nucleotide và 618 trình tự protein, truy cập ngày 29/5/2015). Vì vậy, nghiên cứu của chúng tôi đạt được số lượng lớn các EST và được đăng ký lưu trữ trên GenBank với mã số SRP028517. Đây là nguồn EST phong phú cung cấp dữ liệu tham khảo cho các nghiên cứu tiếp theo trên cá tra ở mức độ phân tử và là cơ sở dữ liệu tin cậy trong so sánh hệ gen với các loài cá xương khác.

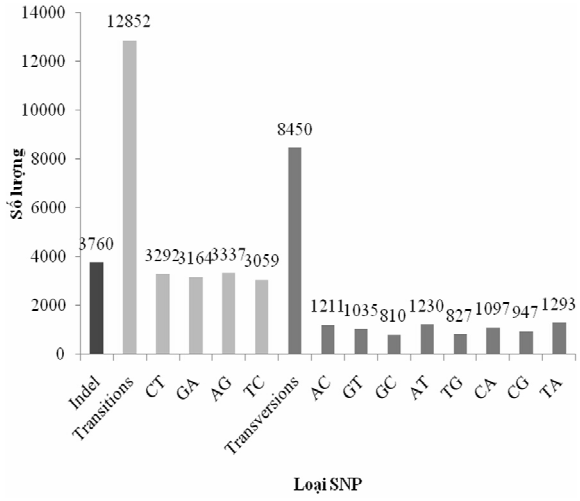
Phân loại nhóm gen chức năng

Định danh các nhóm gen chức năng cho các contig của cá tra sử dụng công cụ BlastX cho kết quả 10.769 contig (35,97%), tương đồng với các trình tự protein đã biết chức năng lưu trữ trên cơ sở dữ liệu KOG (hình 2). Các contig này

được phân loại thành 25 nhóm chức năng. Bên cạnh 1.874 contig không thể phân loại (general function prediction only và function unknown), nhóm chức năng tham gia quá trình tế bào và phát tín hiệu [5] chiếm số lượng lớn nhất bao gồm cơ chế truyền tín hiệu (signal transduction mechanisms) (1.872 contig), các cơ chế biến đổi sau dịch mã (post translational modification, protein turnover, chaperones) (909 contig), các quá trình bài tiết và vận chuyển nội bào (intracellular trafficking, secretion and vesicular transport) (803 contig) và chức năng phiên mã (transcription) (746 contig). Kết quả phân tích của nghiên cứu chúng tôi tương tự với nghiên cứu hệ gen chức năng của sò điệp khi phản ứng với thay đổi của môi trường [19] và nghiên cứu trên loài cá *Gillichthys mirabilis* sống ở môi trường có biên độ mặn rộng [4]. Tuy nhiên phân loại gen chức năng trong nghiên cứu của chúng tôi chỉ dựa trên chú giải các contig. Nghiên cứu biểu hiện gen là nghiên cứu tiếp theo cần thiết để khẳng định các gen chức năng trên cá tra.



Hình 2. Phân loại nhóm gen chức năng cho các contig của cá tra ($E<1e^{-10}$)



Hình 3. Số lượng SNP hoặc indel xác định từ hệ gen chức năng của cá tra

Phân tích các SNP giả định

Từ phương pháp gióng hàng các trình tự contig sau khi kết nối, nghiên cứu của chúng tôi phát hiện được 21.302 SNP giả định và 3.760 indel, bao gồm 12.852 SNP dạng transition và 8.450 SNP dạng transversion (hình 3). Tỷ lệ SNP cao nhất là C/T (15,5%) và A/G (15,7%). Tỷ lệ SNP thấp nhất là G/C (3,8%) và T/G (3.9%). SNP phát hiện từ hệ gen chức năng có nhiều ưu điểm hơn so với SNP ở vùng gen không mã hóa bởi vì các SNP này có thể liên kết với các gen chức năng [6]. Vì vậy SNP của hệ gen chức năng có thể được sử dụng để xác định sự sai khác về kiểu hình của một tính trạng quan tâm [1] cũng như giải thích sự thích nghi của vật nuôi với thay đổi môi trường [6]. Theo

nhóm tác giả Salem et al. (2012) [24] SNP giải thích 90% sự khác biệt di truyền giữa các cá thể, và quá trình trao đổi chéo trong phân bào giảm nhiễm rất hiếm khi tách rời chỉ thị SNP khỏi gen chức năng khi SNP được xác định nằm trên hoặc gần gen chức năng. Các SNP này có nhiều tiềm năng ứng dụng cho các đối tượng thủy sản bởi vì hệ gen của đa số loài thủy sản hiện nay chưa được giải mã hoàn toàn.

KẾT LUẬN

Đây là nghiên cứu chi tiết đầu tiên về hệ gen chức năng liên quan đến tính trạng chịu mặn của cá tra bằng kỹ thuật giải trình tự gen Ion Torrent. Nghiên cứu đạt dữ liệu 272,73 Mbp và 2.623.929 EST sau khi sàng lọc loại bỏ các đoạn trình tự có chất lượng thấp. Từ nguồn EST khổng lồ này, CLC là chương trình kết nối tối ưu cho kết quả kết nối thành 29.940 contig với 60,78% contig có trình tự nucleotide tương tự với các trình tự được lưu trữ ở GenBank và xác định được 5.710 gen giả định ở cá tra. Ngoài ra, nghiên cứu còn phân loại các contig thành 25 nhóm gen chức năng dựa trên cơ sở dữ liệu KOG. Nghiên cứu cũng phát hiện được số lượng lớn SNP có thể ứng dụng cho các nghiên cứu tiếp theo ở mức độ phân tử trên các tra. Nghiên cứu của chúng tôi đã xây dựng được cơ sở dữ liệu genome phong phú cho cá tra có thể sử dụng tham khảo cho nghiên cứu các đối tượng thủy sản khác có giá trị ở Việt Nam.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Quỹ phát triển khoa học và công nghệ quốc gia (NAFOSTED) trong đề tài mã số 106.99-2011.63.

TÀI LIỆU THAM KHẢO

1. Bouck A., Vision T., 2007. The molecular ecologist's guide to expressed sequence tags. Mol. Ecol., 16: 907-924.
2. Chromczynski P., Mackey K., 1995. Short technical report. Modification of TRIZOL reagent procedure for isolation of RNA from Polysaccaride-and proteoglycan-rich sources. Biotechniques, 19: 942-945.
3. Dunham R. A., Taylor J. F., Rise M. L., Liu Z., 2014. Development of strategies for integrated breeding, genetics and applied genomics for genetic improvement of aquatic organisms. Aquaculture, 420-421: S121-S123.
4. Evans T. G., Somero G. N., 2008. A microarray-based transcriptomic time-course of hyper- and hypo-osmotic stress signaling events in the euryhaline fish *Gillichthys mirabilis*: osmosensors to effectors. J. Exp. Biol., 211: 3636-3649.
5. Franchini P., van der Merwe M., Roodt-Wilding R., 2011. Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis. BMC Res. Notes, 4: 59.
6. Gao Z., Luo W., Liu H., Zeng C., Liu X., Yi S., Wang W., 2012. Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS ONE, 7: e42637.
7. Garg R., Patel R. K., Jhanwar S., Priya P., Bhattacharjee A., Yadav G., Bhatia S., Chattopadhyay D., Tyagi A. K., Jain M., 2011. Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol., 156: 1661-1678.
8. Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol., 29: 644-652.
9. Ha H. P., Nguyen T. T. T., Poompuang S., Na-Nakorn U., 2009. Microsatellites revealed no genetic differentiation between hatchery and contemporary wild populations of striped catfish, *Pangasianodon hypophthalmus* (Sauvage 1878) in Vietnam. Aquaculture, 29: 154-160.
10. Haridas S., Breuill C., Bohlmann J., Hsiang T., 2011. A biologist's guide to *de novo*

genome assembly using next-generation sequence data: a test with fungal genomes. J. Microbiol. Methods, 86: 368-375.

11. Hou R., Bao Z., Wang S., Su H., Li Y., Du H., Hu J., Wang S., Hu X., 2011. Transcriptome sequencing and de novo analysis for Yesso Scallop (*Patinopecten yessoensis)* using 454 GS FLX. PloS ONE, 6: e21560.
12. Huang X.D., Zhao M., Liu W.G., Guan Y.Y., Shi Y., Wang Q., Wu S.Z., He M.X., 2013. Gigabase-scale transcriptome analysis on four species of Pearl oysters. Marine Biotechnology, 15: 253-64.
13. Jung H., Lyons R. E., Dinh H., Hurwood D.A., McWilliam S., Mather P.B., 2011. Transcriptomics of a giant freshwater prawn (*Macrobrachium rosenbergii*): *De novo* assembly, annotation and marker discovery. PLoS One, 6: e27938.
14. Laverty G., Skadhauge E., 2012. Adaptation of teleosts to very high salinity. Comp. Biochem. Physiol. A, 163: 1-6.
15. Li H., Durbin R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25: 1754-1760.
16. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25: 2078-2079.
17. Liu S., Zhang Y., Zhou Z., Waldbieser G., Sun F., Lu J., Zhang J., Jiang Y., Zhang H., Wang X., RajendranK.V., Khoo L., Kucuktas H., Peatman E., Liu Z., 2013. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC Genomics, 13: 595.
18. Liu S., Zhang Y., Sun F., Jiang Y., Wang R., Li C., Zhang J., (John) Liu Z., 2012. Functional genomics research in aquaculture: Principles and general approaches. In: Saroglia M., (John) Liu Z. (Eds.), Functional Genomics in Aquaculture. Wiley-Blackwell, pp. 1-40.
19. Meng X., Liu M., Jiang K., Wang B., Tian X., Sun S., Luo Z., Qiu C., Wang L., 2013. *De novo* characterization of Japanese scallop *Mizuhopecten yessoensis* transcriptome and analysis of its gene expression following cadmium exposure. PLoS ONE, 8: e64485.
20. Na-Nakorn U., Moeikum T., 2009. Genetic diversity of domesticated stocks of striped catfish, *Pangasianodon hypophthalmus* (Sauvage 1878), in Thailand: relevance to broodstock management regimes. Aquaculture, 297: 70-77.
21. Nguyen T. T. T., 2009. Patterns of use and exchange of genetic resources of the striped catfish *Pangasianodon hypophthalmus* (Sauvage 1878). Rev. Aquaculture, 1: 224-231.
22. Panhuis T. M., Broitman-Maduro G., Uhrig J., Maduro M., Reznick D. N., 2011. Analysis of expressed sequence tags from the Placenta of the live-bearing fish Poeciliopsis (Poeciliidae). J. Hered., 102: 352-361.
23. Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S. D., Mungall K., Lee S., Okada H. M., Qian J. Q., Griffith M., Raymond A., Thiessen N., Cezard T., Butterfield Y. S., Newsome R., Chan S. K., She R., Varhol R., Kamoh B., Prabhu A.L., Tam A., Zhao Y., Moore R. A., Hirst M., Marra M. A., Jones S. J., Hoodless P. A., Birol I., 2010. *De novo* assembly and analysis of RNA-seq data. Nat. Method, 7: 909-912.
24. Salem M., Vallejo R. L., Leeds T. D., Palti Y., Liu S., Sabbagh A., Rexroad III C. E., Yao J., 2012. RNA-seq identifies SNP markers for growth traits in rainbow trout. PLoS ONE, 7: e36264.
25. Sang N. V., Thomassen M., Klemetsdal G., Gjøen H. M., 2009. Prediction of fillet weight, fillet yield, and fillet fat for live river catfish (*Pangasianodon hypophthalmus*). Aquaculture, 288: 166-171.
26. Sang N. V., Klemetsdal G., Ødegård J., Gjøen H. M., 2012. Genetic parameters of

economically important traits recorded at a given age in striped catfish (*Pangasianodon hypophthalmus*). Aquaculture, 344-349: 82-89.

27. Tatusov R. L., Fedorova N. D., Jackson J. D., Jacobs A. R., Kiryutin B., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4: 41.
28. Tổng cục Thủy sản, 2015. Tình hình sản xuất thủy sản năm 2014 (26/02/2015). http://www.fistenet.gov.vn/thong-tin-huu-ich/thong-tin-thong-ke/thong-ke-1/tinh-hinh-san-xuat-thuy-san-nam-2014/ (truy cập 29/5/2015)
29. Vasemägi A., Primmer C. R., 2005. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol. Ecol., 14: 3623-3642.
30. Wang S., Peatman E., Abernathy J., Waldbieser G., Lindquist E., Richardson P., Lucas S., Wang M., Li P., Thimmapuram J., Liu L., Vullaganti D., Kucuktas H., Murdock C., Small B.C., Wilson M., Liu H., Jiang Y., Lee Y., Chen F., Lu J., Wang W., Xu P., Somridhivej B., Baoprasertkul P., Quilang J., Sha Z., Bao B., Wang Y., Wang Q., Takano T., Nandi S., Liu S., Wong L., Kaltenboeck L., Quiniou S., Bengten E., Miller N., Trant J., Rokhsar D., Liu Z., the Catfish Genome Consortium, 2010. Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome Biology, 11: R8.
31. Wong L. L., Peatman E., Lu J., Kucuktas H., He S., Zhou C., Na-Nakorn U., Liu Z., 2011. DNA barcoding of catfish: species authentication and phylogenetic assessment. PLoS ONE, 6: e17812.
32. Zhou Y., Gao F., Liu R., Feng J., Li H., 2012. *De novo* sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. BMC Genomics, 13: 266.

A TRANSCRIPTOMIC ANALYSIS OF THE KIDNEY TISSUE OF TRA CATFISH (*Pangasianodon hypophthalmus*) REARED IN SALINE CONDITION: *DE NOVO* ASSEMBLY, ANNOTATION, SNP DISCOVERY

Nguyen Minh Thanh[1], Vo Thi Minh Thu[1], Hyungtaek Jung[2], Peter Mather[2]

[1]International University, VNU-HCM
[2]Queensland University of Technology (QUT)

SUMMARY

Pangasianodon hypophthalmus is a commercially important freshwater fish used in inland aquaculture in the Mekong Delta, Vietnam. The current study using Ion Torrent technology generated EST resources from the kidney for Tra catfish reared at a salinity level of 9 ppt. We obtained 2,623,929 reads after trimming and processing with an average length of 104 bp. *De novo* assemblies were generated using CLC Genomic Workbench, Trinity and Velvet/Oases with the best overall contig performance resulting from the CLC assembly. *De novo* assembly using CLC yielded 29,940 contigs, and allowing identification of 5,710 putative genes when comppared with NCBI non-redundant database. A large number of single nucleotide polymorphisms (SNPs) were also detected. The sequence collection generated in our study represents the most comprehensive transcriptomic resource for *P. hypophthalmus* available to date.

Keywords: *Pangasianodon hypophthalmus*, transcriptome, kidney, salinity tolerance.

Ngày nhận bài: 10-1-2015