Tạp chí Công nghệ Sinh học **19**(2): 271-277, 2021

LẮP RÁP VÀ CHÚ GIẢI HỆ GEN TÔM THẺ CHÂN TRẮNG (*LITOPENAEUS VANNAMEI*) BỊ NHIỄM VIRUS ĐỐM TRẮNG Ở VIỆT NAM

Nguyễn Văn Tụng[1], Nguyễn Thị Kim Liên[1,✉], Dương Chí Thành[1], Nguyễn Thu Hiền[1], Nguyễn Ngọc Lan[1], Nguyễn Thị Thanh Ngân[1], Nguyễn Huy Hoàng[1], Trịnh Thị Trang[2], Nguyễn Hữu Ninh[3], Nguyễn Hữu Hùng[3]

[1]*Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[2]*Học viện Nông nghiệp Việt Nam, Bộ Nông nghiệp và Phát triển nông thôn*
[3]*Viện Nghiên cứu Nuôi trồng thủy sản 3, Bộ Nông nghiệp và Phát triển nông thôn*

✉Người chịu trách nhiệm liên lạc. E-mail: ntkimlienibt@gmail.com

Ngày nhận bài: 28.02.2020
Ngày nhận đăng: 18.12.2020

TÓM TẮT

Tôm thẻ chân trắng Thái Bình Dương (*Penaeus vannamei* hoặc *Litopenaeus vannamei*) là loài tôm có nguồn gốc từ Nam Mỹ, đang là một trong những đối tượng tôm nuôi quan trọng có giá trị kinh tế cao ở Việt Nam và nhiều nơi trên thế giới. Trong hai thập kỷ gần đây, bệnh do virus đốm trắng (white spot syndrome virus - WSSV) gây ra ảnh hưởng nghiêm trọng đến ngành nuôi tôm với tỷ lệ gây chết có thể lên đến 100% sau 3 đến 10 ngày phát bệnh. Giải trình tự và lắp ráp hệ gen là một bước quan trọng để cung cấp thông tin di truyền và nghiên cứu các cơ chế phân tử ở các loài có giá trị kinh tế cao. Do đó nghiên cứu này tiến hành giải trình tự, lắp ráp và chú giải hệ gen của tôm thẻ chân trắng nhiễm virus đốm trắng tại Việt Nam. Dữ liệu hệ gen của tôm thẻ chân trắng nhiễm virus đốm trắng được lắp ráp bằng công cụ SOAPdenovo2 thu được hệ gen có kích thước khoảng 1,67Gb với 3.180.049 scaffold với N50 là 616 bp, từ đó dự đoán được 187.948 gen. Trong đó có 133.548 gen được chú giải trên cơ sở dữ liệu UniProtKB/Swissprot và 33.611 gen được chú giải trên cơ sở dữ liệu NT. Đây là những kết quả ban đầu có ý nghĩa quan trọng cho các nghiên cứu tiếp theo về tính trạng kháng bệnh virus đốm trắng của tôm thẻ chân trắng ở Việt Nam.

Từ khóa: *lắp ráp de novo, Litopenaeus vannamei, SOAP denovo2, tôm thẻ chân trắng, virus đốm trắng*

ĐẶT VẤN ĐỀ

Tôm thẻ chân trắng Thái Bình Dương (*Penaeus vannamei* hoặc *Litopenaeus vannamei*) là một trong những loài giáp xác được nuôi rộng rãi nhất trên thế giới do năng suất cao và yêu cầu về nồng độ muối trong môi trường nuôi thấp (Zhou *et al*., 2012). Sản lượng tôm thẻ chân trắng chỉ đứng sau sản lượng tôm sú nuôi trên thế giới, điểm đặc biệt của loài tôm này là tăng trưởng nhanh, tính thích nghi môi trường tốt, yêu cầu về nguồn dinh dưỡng trong thức ăn thấp. Ngoài ra, vào mùa mưa độ mặn và nhiệt độ thường xuống thấp nhưng tôm thẻ chân trắng lại thích ứng tốt với các mô hình nuôi có độ mặn từ 0 - 40%. Khoảng hai thập kỷ gần đây, các bệnh do virus gây ra diễn biến ngày càng phức tạp, đe dọa nghiêm trọng ngành nuôi tôm *(*Escobedo-Bonilla *et al., 2008;* Lightner *et al*., 1997; Naylor *et al*., 2000; Valles-Jimenez *et al*., 2004), trong đó, bệnh do virus đốm trắng gây ra là nguy hiểm nhất, tỷ lệ tôm chết lên đến 100% sau 3 đến 10 ngày phát bệnh gây thiệt hại kinh tế lớn (’t Hoen *et al*., 2008). Virus đốm trắng (white spot syndrome virus - WSSV) có bộ gen lớn (~300 kb) (van Hulten *et al*., 2001; Yang *et al*., 2001) và có phạm vi vật chủ rộng, bao gồm hầu hết các loài giáp xác và cả côn trùng thủy sinh (Tan and

Shi, 2011; Wang *et al*., 2000). Mầm bệnh lây truyền theo cả chiều dọc từ bố mẹ sang con và theo chiều ngang từ các loài giáp xác (cua, tép, chân chèo) nhiễm WSSV trong ao nuôi, do tôm ăn thức ăn nhiễm virus, do nguồn nước có WSSV và do tôm khoẻ ăn tôm chết nhiễm WSSV trong ao nuôi. Khi bùng phát dịch bệnh đốm trắng sẽ gây thiệt hại rất lớn cho người nuôi tôm cũng như ngành thuỷ sản. Tôm chân trắng được xác định là một trong hai đối tượng tôm nuôi nước lợ chủ lực của nước ta, nhu cầu giống tôm chân trắng kháng bệnh đốm trắng ngày càng tăng về số lượng và chất lượng, do đó việc nghiên cứu chọn tạo giống ở cấp độ phân tử là cần thiết qua đó chủ động phát triển đàn tôm thẻ chân trắng bố mẹ chất lượng cao có khả năng kháng bệnh đốm trắng tại Việt Nam.

Những năm gần đây, sự phát triển mạnh mẽ của các công nghệ giải trình tự gen thế hệ mới (NGS: Next-Generation Sequencing) và sự lớn mạnh của lĩnh vực liên ngành Tin sinh học khiến việc lắp ráp và chú giải hệ gen đã trở thành phương pháp nghiên cứu phổ biến. Các thuật toán phổ biến được sử dụng để xử lý loại dữ liệu này là sử dụng đồ thị de Bruijn và OLC (overlap-layout-consensus) (Flicek and Birney, 2009; Miller *et al*., 2010; Schatz *et al*., 2010) đi kèm theo đó là những công cụ lắp ráp như SOAP denovo2 (Luo *et al*., 2012), Platanus (Kajitani *et al*., 2014), Ray-assembler (Boisvert *et al*., 2010), Hipmer (Georganas *et al*., 2015). Hiện nay trên thế giới đã có những nghiên cứu lắp ráp hệ gen sinh vật, bao gồm cả các loài giáp xác qua đó góp phần cung cấp hiểu biết về dữ liệu trình tự hệ gen sinh vật (Song *et al*., 2016; Yuan *et al*., 2017). Đối với tôm thẻ chân trắng, năm 2015 Yu cùng cộng sự đã lắp ráp *de novo* hệ gen của loài tôm này (Yu *et al*., 2015) với kích thước hệ gen khoảng 2,3 Gb. Năm 2019, nhóm nghiên cứu Xiaojun Zhang đã lắp ráp hệ gen này sử dụng đồ thị Bruijin mờ (fuzzy Bruijn graph - FBG) thu được hệ gen có kích thước 1,66 Gb (Zhang *et al*., 2019). Đồng thời đã có nghiên cứu phân tích hệ gen biểu hiện của tôm nhiễm virus đốm trắng nhằm nghiên cứu tương tác giữa virus đốm trắng và tôm, qua đó có những hiểu biết về cơ chế tác động của virus này đến hệ miễn dịch của vật chủ (Chen *et al*., 2013). Những hiểu biết ở mức độ phân tử về hệ gen của tôm *L. vannamei* nhiễm virus đốm trắng rất hữu ích trong việc nghiên cứu tương tác tôm thẻ chân trắng và virus cũng như cung cấp đầy đủ hơn thông tin di truyền về đối tượng tôm này. Sự khác biệt giữa hệ gen *L. vannamei* nhiễm virus với hệ gen tôm khỏe mạnh có thể được chỉ ra bằng hai phương pháp chính. Phương pháp tiếp cận thứ nhất dựa trên việc gióng hàng các đoạn đọc ngắn tạo ra khi giải trình tự với hệ gen tham chiếu (alignment-based approach). Đây là phương pháp được sử dụng phổ biến hiện nay để chỉ ra khác biệt giữa dữ liệu "re-sequencing" với hệ gen tham chiếu. Tuy nhiên, phương pháp này có thể tồn tại một số hạn chế như hệ gen tham chiếu lắp ráp chưa hoàn chỉnh (Meyer *et al*., 2013), đột biến cấu trúc tồn tại trong hệ gen của đối tượng cần nghiên cứu (Sudmant *et al*., 2015), lỗi giải trình tự và đa hình nucleotide đơn (SNP) trong đoạn đọc (Iqbal *et al*., 2012) làm ảnh hưởng đến kết quả gióng hàng. Phương pháp tiếp cận thứ hai dựa trên việc so sánh kết quả lắp ráp hai hệ gen (*de novo* assembly-based approach). Trong phương pháp này, các đoạn đọc ngắn của đối tượng nghiên cứu được lắp ráp *de novo* thành contig hoặc scaffold rồi so sánh với hệ gen tham chiếu. Mặc dù chưa được ứng dụng rộng rãi nhưng đây được coi là phương pháp lý tưởng để phát hiện sự khác biệt giữa hai hệ gen (Chaisson *et al*., 2015; Xiao *et al*., 2016).

Trong nghiên cứu này, chúng tôi đã tiến hành ứng dụng công nghệ giải trình tự thế hệ mới để giải trình tự, lắp ráp và chú giải hệ gen của tôm thẻ chân trắng nhiễm virus đốm trắng qua đó cung cấp đầy đủ hơn thông tin di truyền ở cấp độ phân tử của loài tôm có giá trị kinh tế cao này.

DỮ LIỆU VÀ PHƯƠNG PHÁP

Dữ liệu

Mẫu tôm thẻ chân trắng bị nhiễm virus đốm trắng được cung cấp bởi Viện Nghiên cứu Nuôi trồng Thủy sản 3, Nha Trang, Khánh Hòa. DNA tổng số của tôm được tách chiết từ mô cơ bằng bộ kit QIAamp DNA Mini kit (QIAGEN, Hilden,

Đức) sau đó tiến hành giải trình tự bằng hệ thống đọc trình tự Illumina.

Đánh giá và xử lý dữ liệu

Dữ liệu trình tự thu được từ thiết bị đọc trình tự thế hệ mới được đánh giá và xử lý bằng công cụ FastQC và Trimmomatic (Bolger *et al*., 2014). Những đoạn trình tự có độ dài nhỏ hơn 36 bp hoặc chứa trên 10% nucleotide không xác định hoặc 4 nucleotide liên tiếp có điểm chất lượng trung bình nhỏ hơn 20 (QC<20) bị loại bỏ.

Lắp ráp và chú giải hệ gen

Dữ liệu sau khi xử lý được đưa vào lắp ráp để thu được các đoạn trình tự dài liên tục gọi là scaffold bằng phần mềm SOAP denovo2 (Luo *et al*., 2012) với giá trị k-mer tối ưu được xác định thông qua công cụ KmerGenie (Chikhi and Medvedev, 2014). Chất lượng lắp ráp được đánh giá thông qua các thông số như kích thước hệ gen, chỉ số N50 bằng phần mềm Quast (Gurevich *et al*., 2013). Các scaffold sau khi lắp ráp được so sánh với hệ gen tham chiếu của *L. vannamei* (ASM378908v1) bằng phần mềm MUMmer 3.0 (Kurtz *et al*., 2004). Tập hợp các scaffold sau khi lắp ráp có độ dài lớn hơn 200 bp được dự đoán gen bằng công cụ Augustus.2.5.5 (Stanke and Waack, 2003) và chú giải trên hai cơ sở dữ liệu NCBI NT (Pruitt *et al*., 2005) và UniProtKB/Swiss-Prot (Bairoch and Apweiler, 2000) với tham số E-value≤1e-5 bằng công cụ Blast+ (Camacho *et al*., 2009).

KẾT QUẢ VÀ THẢO LUẬN

Giải trình tự toàn bộ hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng

Giải trình tự toàn bộ hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng thu được dữ liệu bao gồm 348.908.913 đoạn đọc với độ dài đồng nhất là 150 bp. Sau khi loại bỏ những đoạn trình tự chất lượng kém thu được dữ liệu gồm 298.516.063 đoạn đọc, QC>30 đạt 93,70% (Bảng 1).

Bảng 1. Kết quả tiền xử lý dữ liệu.

Tổng số đoạn trình tự	298.516.063
Hàm lượng GC(%)	38
Q20 (%)	99,50
Q30(%)	93,70

Lắp ráp hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng và so sánh với hệ gen tham chiếu

Hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng được phần mềm KmerGenie ước lượng có kích thước là 1.994.848.115 bp và giá trị K-mer tối ưu là k=37. Lắp ráp bằng công cụ SOAPdenovo2 thu được hệ gen có kích thước 1.673.048.405 bp (bằng 82,87% kích thước ước đoán) với 3.180.049 scaffold có độ dài tối thiểu là 200 bp, trong đó có 280.126 scaffold có độ dài trên 1.000 bp với chỉ số N50 là 616 bp (Bảng 2).

Bảng 2. Kết quả lắp ráp hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng.

Kích thước hệ gen được lắp ráp (bp)	**1.673.048.405**
Tổng số scaffold	3.180.049
Scaffold dài nhất (bp)	137.569
Scaffold ngắn nhất (bp)	200
Số lượng scaffold ≥ 1000 bp	280.126
Số lượng scaffold ≥ 10000 bp	1.074
Số lượng scaffold ≥ 25000 bp	244
N50	616
N75	366
Tỉ lệ GC(%)	39,48

Các scaffold sau khi lắp ráp được so sánh với hệ gen tôm thẻ chân trắng tham chiếu có mã số ASM378908v1 bằng phần mềm MUMmer 3.0. Kết quả cho thấy hệ gen tôm thẻ chân trắng lắp ráp tại Việt Nam có chứa 23.790.445 điểm sai khác dạng thay thế, 1.421 đột biến thêm/bớt đoạn ngắn trong đó chủ yếu là các đoạn nhỏ hơn 50 bp.

Dự đoán gen và chú giải chức năng

Sử dụng phần mềm Augustus.2.5.5 để dự đoán gen thu được 238.558 đoạn gen, trong đó có 187.948 gen có độ dài lớn hơn 200 bp. Các đoạn gen có độ dài lớn hơn 200 bp được chú giải bằng cơ sở dữ liệu UniProtKB/Swissprot và NT (Bảng 3).

Bảng 3. Kết quả chú giải hệ gen tôm thẻ chân trắng.

Tổng số đoạn gen	**238.558**
Số đoạn gen có độ dài ≥ 200 bp	187.948
Số đoạn gen được chú giải trên cơ sở dữ liệu UniProtKB/Swissprot	133.548
Số đoạn gen được chú giải trên cơ sở dữ liệu NT	33.611

Kết quả có 133.548 gen được chú giải trên UniProtKB/Swiss-Prot chiếm tỉ lệ 71,06%. Đặc biệt, trong đó phát hiện 1 gen mã hóa E3 ligase WSSV222 của virus đốm trắng (có mã số trên GenBank là Q77J49.1). Đồng thời, những đoạn gen có độ dài trên 200 bp được chú giải trên cơ sở dữ liệu NT. Kết quả có 33.611 gen được chú giải trên NT chiếm tỉ lệ 17,88%, trong đó có 2 gen chưa rõ chức năng thuộc về virus đốm trắng chủng IN-06-I (có mã số trên GenBank là EF468498.1).

Giải trình tự và lắp ráp hệ gen là một bước quan trọng để cung cấp thông tin di truyền và nghiên cứu các cơ chế phân tử ở các loài có giá trị kinh tế cao. Mặc dù *L. vannamei* là một trong những đối tượng tôm nuôi quan trọng ở Việt Nam và nhiều nơi trên thế giới, nhưng những nghiên cứu về hệ gen của loài tôm này chưa đầy đủ. Nghiên cứu trước đây cho thấy hệ gen của tôm thẻ chân trắng có nhiều đặc điểm đặc trưng, khó phân tích (Zhang *et al*., 2010). Ở nước ta đã có các nghiên cứu nhằm nâng cao chất lượng di truyền và kiểm soát dịch bệnh ở một số giống thủy sản đặc biệt là tôm sử dụng các kỹ thuật sinh học phân tử. Các nghiên cứu ở mức độ di truyền phân tử trên đối tượng thủy sản ở Việt Nam có thể kể đến như việc giải trình tự hệ transcriptome tôm sú và dự đoán những SSR tiềm năng (Nguyen *et al*., 2016), lắp ráp hệ gen cá tra *Pangasianodon hypophthalmus* và phân tích gen liên quan đến tăng trưởng (Kim *et al*., 2018). Tuy nhiên, hiện nay Việt Nam chưa có công bố nào nghiên cứu toàn bộ hệ gen tôm thẻ chân trắng, đặc biệt là tôm nhiễm virus.

Năm 2019, nhóm nghiên cứu Xiaojun Zhang đã lắp ráp hệ gen này sử dụng đồ thị Bruijin mờ (fuzzy Bruijn graph - FBG) thu được hệ gen có kích thước 1,66 Gb (Zhang *et al*., 2019). Nghiên cứu này đã lắp ráp hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng ở Việt Nam với kích thước hệ gen thu được sau khi lắp ráp là xấp xỉ 1,6 Gb. Kích thước hệ gen trong nghiên cứu của chúng tôi tương đương với kích thước hệ gen tôm thẻ chân trắng được công bố bởi Xiaojun Zhang và cộng sự (Zhang *et al*., 2019) nhưng nhỏ hơn kích thước hệ gen được công bố trước đó bởi Yu và cộng sự có kích thước 2,3 Gb (Yu *et al*., 2015). Theo Yu và cộng sự, hệ gen *L. vannamei* có nhiều đoạn trình tự lặp phức tạp. Những đoạn trình tự lặp lại này khiến quá trình lắp ráp trở nên khó khăn, các scaffold và contig được lắp ráp có độ dài không cao thể hiện qua chỉ số N50 nhỏ. Do đó, việc lắp ráp hoàn thiện hệ gen tôm thẻ chân trắng *L. vannamei* rất khó khăn nếu chỉ sử dụng dữ liệu được tạo ra bởi thiết bị giải trình tự thế hệ mới Illumina. Nhóm nghiên cứu của Xiaojun Zhang sử dụng kết hợp giữa dữ liệu đoạn ngắn của thiết bị Illumina với dữ liệu đoạn đọc dài hơn từ phương pháp giải trình tự PacBio đồng thời sử dụng thuật toán FDB thu được hệ gen có kích thước nhỏ hơn nhưng độ dài các scaffold lớn hơn (kích thước: 1,6 Gb, N50: 605,555 bp). Tuy nhiên, tất cả hệ gen được công bố bởi các nhóm nghiên cứu trên đều có kích thước nhỏ hơn kích thước ước lượng bằng phần mềm phân tích K-mer (2,6 Gb) và phương pháp đếm tế bào dòng chảy (2,45 Gb).

KẾT LUẬN

Trong nghiên cứu này, hệ gen tôm thẻ chân trắng nhiễm virus đốm trắng được lắp ráp có kích thước 1.673.048.405 bp, dự đoán được 187.948 gen có kích thước lớn hơn 200 bp. Các đoạn gen này được chú giải chức năng trên hai cơ sở dữ liệu UniProtKB/Swiss-Prot và NT. Kết quả có 133.548 gen được chú giải trên UniProtKB/Swiss-Prot, trong đó có 1 gen mã hóa E3 ligase WSSV222 của virus đốm trắng; có 33.611 gen được chú giải trên cơ sở dữ liệu NT, trong đó có 2 gen chưa rõ chức năng thuộc về chủng virus đốm trắng IN-06-I. Đây là những kết quả ban đầu cung cấp cái nhìn rõ hơn về hệ gen của tôm thẻ chân trắng nhiễm virus, cung cấp cơ sở khoa học cho các nghiên cứu sâu hơn về hệ gen và thông tin di truyền của loài tôm này.

Lời cảm ơn: *Công trình nghiên cứu này được tài trợ kinh phí của Bộ Nông nghiệp và phát triển Nông thôn cho đề tài "Nghiên cứu tạo vật liệu ban đầu phục vụ chọn giống tôm thẻ chân trắng kháng bệnh đốm trắng".*

TÀI LIỆU THAM KHẢO

Bairoch A and Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res 28:* 45–48.

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol J Comput Mol Cell Biol 17:* 1519–1533.

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

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics 10:* 421.

Chaisson MJP, Wilson RK, Eichler EE (2015) Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet 16*: 627–640.

Chen X, Zeng D, Chen X, Xie D, Zhao Y, Yang C, Li Y, Ma N, Li M, Yang Q, et al. (2013) Transcriptome analysis of *Litopenaeus vannamei* in response to white spot syndrome virus infection. *PLOS ONE 8:* e73218.

Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics 30:* 31–37.

Escobedo-Bonilla CM, Alday-Sanz V, Wille M, Sorgeloos P, Pensaert MB, Nauwynck HJ (2008) A review on the morphology, molecular characterization, morphogenesis and pathogenesis of white spot syndrome virus. *J Fish Dis 31*: 1–18.

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nat Methods 6:* S6–S12.

Georganas E, Buluç A, Chapman J, Hofmeyr S, Aluru C, Egan R, Oliker L, Rokhsar D, Yelick K (2015) HipMer: an extreme-scale de novo genome assembler. In SC '15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp: 1–11.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics 29:* 1072–1075.

't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, Boer JM, van Ommen GJB, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res 36:* e141.

van Hulten MC, Witteveldt J, Peters S, Kloosterboer N, Tarchini R, Fiers M, Sandbrink H, Lankhorst RK, Vlak JM (2001) The white spot syndrome virus DNA genome sequence. *Virology 286:* 7–22.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet 44:* 226–232.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res 24:* 1384–1395.

Kim OTP, Nguyen PT, Shoguchi E, Hisata K, Vo TTB, Inoue J, Shinzato C, Le BTN, Nishitsuji K, Kanda M, et al. (2018) A draft genome of the striped catfish, *Pangasianodon hypophthalmus*, for

comparative analysis of genes relevant to development and a resource for aquaculture improvement. *BMC Genomics 19:* 733.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol 5:* R12.

Lightner DV, Redman RM, Poulos BT, Nunan LM, Mari JL, Hasson KW (1997) Risk of spread of penaeid shrimp viruses in the Americas by the international movement of live and frozen shrimp. *Rev Sci Tech Int Off Epizoot 16:* 146–160.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience 1:* 18.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res 41*: D64–D69.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics 95:* 315–327.

Naylor RL, Goldburg RJ, Primavera JH, Kautsky N, Beveridge MCM, Clay J, Folke C, Lubchenco J, Mooney H, Troell M (2000) Effect of aquaculture on world fish supplies. *Nature 405:* 1017–1024.

Nguyen C, Nguyen TG, Nguyen LV, Pham HQ, Nguyen TH, Pham HT, Nguyen HT, Ha TT, Dau TH, Vu HT, et al. (2016) *De novo* assembly and transcriptome characterization of major growth-related genes in various tissues of *Penaeus monodon*. *Aquaculture 464:* 545–553.

Pruitt KD, Tatusova T, Maglott DR (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res 33:* D501–D504.

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. *Genome Res 20:* 1165–1173.

Song L, Bian C, Luo Y, Wang L, You X, Li J, Qiu Y, Ma X, Zhu Z, Ma L, et al. (2016) Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. *GigaScience 5*.

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinforma Oxf Engl 19 Suppl 2:* 215-225.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature 526:* 75–81.

Tan Y, Shi Z (2011) Genotyping of white spot syndrome virus in Chinese cultured shrimp during 1998-1999. *Virol Sin 26*: 123–130.

Valles-Jimenez R, Cruz P, Perez-Enriquez R (2004) Population genetic structure of Pacific white shrimp (*Litopenaeus vannamei*) from Mexico to Panama: microsatellite DNA variation. *Mar Biotechnol NYN 6:* 475–484.

Wang YG, Lee KL, Najiah M, Shariff M, Hassan MD (2000) A new bacterial white spot syndrome (BWSS) in cultured tiger shrimp *Penaeus monodon* and its comparison with white spot syndrome (WSS) caused by virus. *Dis Aquat Organ 41:* 9–18.

Xiao W, Wu L, Yavas G, Simonyan V, Ning B, Hong H (2016) Challenges, solutions, and quality metrics of personal genome assembly in advancing precision medicine. *Pharmaceutics 8*.

Yang F, He J, Lin X, Li Q, Pan D, Zhang X, Xu X (2001) Complete genome sequence of the shrimp white spot Bacilliform virus. *J Virol 75:* 11811–11820.

Yu Y, Zhang X, Yuan J, Li F, Chen X, Zhao Y, Huang L, Zheng H, Xiang J (2015) Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific white shrimp *Litopenaeus vannamei*. *Sci Rep 5:* 15612.

Yuan J, Gao Y, Zhang X, Wei J, Liu C, Li F, Xiang J (2017) Genome sequences of marine shrimp exopalaemon carinicauda holthuis provide insights into genome size evolution of Caridea. *Mar Drugs 15*.

Zhang X, Zhang Y, Scheuring C, Zhang HB, Huan P, Wang B, Liu C, Li F, Liu B, Xiang J (2010) Construction and characterization of a bacterial artificial chromosome (BAC) library of Pacific white shrimp, *Litopenaeus vannamei*. *Mar Biotechnol NYN 12:* 141–149.

Zhang X, Yuan J, Sun Y, Li S, Gao Y, Yu Y, Liu C, Wang Q, Lv X, Zhang X, et al. (2019) Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nat Commun 10:* 356.

Zhou J, Fang W, Yang X, Zhou S, Hu L, Li X, Qi X, Su H, Xie L (2012) A nonluminescent and highly virulent *Vibrio harveyi* strain is associated with "Bacterial white tail disease" of *Litopenaeus vannamei* shrimp. *PLOS ONE 7*: e29961.

Tạp chí Công nghệ Sinh học **19**(2): 271-277, 2021

GENOME ASSEMBLY AND ANNOTATION OF THE WHITE SPOT SYNDROME VIRUS - INFECTED PACIFIC WHITE SHRIMP (*LITOPENAEUS VANNAMEI*) IN VIETNAM

Nguyen Van Tung[1], Nguyen Thi Kim Lien[1], Duong Chi Thanh[1], Nguyen Thu Hien[1], Nguyen Ngoc Lan[1], Nguyen Thi Thanh Ngan[1], Nguyen Huy Hoang[1], Trinh Thi Trang[2], Nguyen Huu Ninh[3], Nguyen Huu Hung[3]

[1]*Institute of Genome Research, Vietnam Academy Science and Technology*
[2]*Vietnam National University of Agriculture, Ministry of Agriculture and Rural Development*
[3]*Research Institute for Aquaculture No. 3, Ministry of Agriculture and Rural Development*

SUMMARY

Pacific white shrimp (*Penaeus vannamei* or *Litopenaeus vannamei*) is native to South America, high economic value, and widely cultivated in the world and Vietnam. Over the last two decades, viral diseases have seriously threatened the shrimp aquaculture industry. Among the viral diseases, white spot syndrome virus (WSSV) is the most important viral pathogens of shrimp farming. WSSV causes a cumulative mortality can reach 100% within 3–10 days. Genome sequencing and assembly has been an important step for deciphering molecular mechanisms and accelerating genetic improvements of traits of interest in economically important species. This study aims at constructing and annotating the genome of white spot syndrome virus - infected Pacific white shrimp in Vietnam. The whole genome sequencing data was *de novo* assembled using SOAP denovo2 to obtained draft genome of WSSV- infected *L. vannamei* shrimp. The draft genome contained 3,180,049 scaffolds (genome size ~1.67 Gb) with the length arranging from 200 bp to 137,569 bp and with N50 as 616 bp. Applying gene prediction method, we have been able to identify 187,948 putative genes. The results have shown that 33,611 genes were annotate in NT database and 133,548 genes were annotated in UniProtKB/Swissprot database. These results are only the initial information about white spot syndrome virus - infected Pacific white shrimp but they are really important for future studies relating to white spot syndrome virus – resistance *L. vannamei* shrimp in Vietnam.

Keywords: *de novo assembly, Litopenaeus vannamei, SOAP denovo2, Pacific white shrimp, white spot syndrome virus*