APPLICATION OF NEXT GENERATION SEQUENCING (NGS) FOR SCREENING SINGLE NUCLEOTIDE POLYMORPHISMS (SNP) RELATED TO GROWTH TRAITS OF OTTER CLAM (*Lutraria rhynchaen,* Jonas 1844)

Trieu Anh Tuan[1*], Thai Thanh Binh[2]
[1]Hung Vuong University, [2]Fisheries and Technical, Economic College

ARTICLE INFO

Received: 06/12/2022

Revised: 05/4/2023

Published: 13/4/2023

KEYWORDS

Otter clam
Growth
Sequencing
Screening
SNP

ABSTRACT

Otter clam farming in Van Don, Quang Ninh has been encountering difficulties because of poor quality of breeds and more frequent disease outbreaks. On that basis, it is necessary to conduct genetic studies to improve the source of breeds, especially the combination of quantitative inheritance and molecular genetics to select desirable traits by molecular indicator. In this study, we employed next-generation sequencing of gene technology NovaSeq6000 to sequence and screen single nucleotide polymorphism (SNPs) that relates to growth traits in otter clam. The results was screened shows that total 1,470,534 SNPs to 2 groups, in which 703.228 SNPs only appeared in fast-growing group, 394.723 SNPs only appeared in slow-growing group, and 372.583 SNPs appeared in both groups. Through screening and genetic linkage analysis, 20 potential SNPs for fast-growing group and 12 potential SNPs for slow-growing group were achieved. The study suggests that SNP Screen should be employed as a molecular indicator method to serve geoduck breed selection in Vietnam.

ỨNG DỤNG GIẢI TRÌNH TỰ GENE THẾ HỆ MỚI (NGS) ĐỂ SÀNG LỌC ĐA HÌNH NUCLEOTIDE ĐƠN (SNP) LIÊN QUAN ĐẾN TÍNH TRẠNG TĂNG TRƯỞNG Ở TU HÀI (*Lutraria rhynchaen*, Jonas 1844)

Triệu Anh Tuấn[1*], Thái Thanh Bình[2]
[1]Trường Đại học Hùng Vương, [2]Trường Cao đẳng kinh tế, Kỹ thuật và Thủy sản

THÔNG TIN BÀI BÁO

Ngày nhận bài: 06/12/2022

Ngày hoàn thiện: 05/4/2023

Ngày đăng: 13/4/2023

TỪ KHÓA

Tu hài
Tăng trưởng
Giải trình tự
Sàng lọc
SNP

TÓM TẮT

Nghề nuôi tu hài (*Lutraria rhynchaena)* tại huyện Vân Đồn, tỉnh Quảng Ninh hiện nay gặp nhiều khó khăn do chất lượng con giống thấp, dịch bệnh thường xuyên xảy ra. Trên cơ sở đó, cần có những nghiên cứu nhằm nâng cao chất lượng con giống; đặc biệt là cần có sự kết hợp giữa di truyền số lượng và di truyền phân tử để chọn lọc được tính trạng mong muốn thông qua các chỉ thị phân tử. Trong nghiên cứu này, chúng tôi ứng dụng công nghệ giải trình tự gen thế hệ mới NovaSeq6000 để giải trình tự và sàng lọc các đa hình nucleotide đơn (SNPs) liên quan đến tính trạng tăng trưởng ở tu hài. Kết quả đã sàng lọc được tổng số 1.470.534 SNPs cho cả hai nhóm tu hài, ở nhóm tu hài tăng trưởng nhanh xác định được 703.228 SNPs, ở tu hài tăng trưởng chậm xác định được 394.723 SNPs và xác định được 372.583 SNPs xuất hiện ở cả hai nhóm tu hài. Qua sàng lọc và phân tích liên kết thu được 20 SNP tiềm năng cho nhóm tu hài tăng trưởng nhanh và 12 SNP tiềm năng cho nhóm tu hài tăng trưởng chậm. Các SNP được sàng lọc bước đầu là chỉ thị phân tử phục vụ cho các chương trình chọn giống tu hài ở Việt Nam trong tương lai.

DOI: https://doi.org/10.34238/tnu-jst.7056

[*] Corresponding author. *Email: tuantrieuanh85@gmail.com*

1. Giới thiệu

Tu hài (*Lutraria rhychaena*, Jonas 1844) là động vật thân mềm hai mảnh vỏ, là loài thủy sản nuôi có giá trị kinh tế cao. Ở Việt Nam, tu hài được nuôi phổ biến ở Vân Đồn – Quảng Ninh và Nha Trang – Khánh Hòa. Hàng năm nhu cầu con giống cho người nuôi khoảng 100 triệu con giống. Tuy nhiên, chất lượng nguồn con giống là vấn đề rất được quan tâm. Nguồn giống tu hài do di nhập từ nước ngoài vào nước ta không được kiểm soát, thường có chất lượng thấp đã gây hậu quả nghiêm trọng trong quá trình nuôi, ảnh hưởng không nhỏ đến sự phát triển bền vững nghề nuôi tu hài. Chính vì vậy, việc nghiên cứu sản xuất giống tu hài có chất lượng cao là việc làm cần thiết đối với các nhà chọn giống. Chọn giống tu hài theo phương pháp hiện đại là tìm kiếm các gen chức năng dựa trên cơ sở hệ gen tham chiếu của tu hài đã được công bố [1]. Việc khai thác và ứng dụng hiệu quả hệ gen tham chiếu của tu hài vào quá trình nghiên cứu là điều rất cần thiết để nâng cao chất lượng di truyền cho tu hài, đồng thời tuyển chọn được những gen liên quan đến tính trạng có giá trị kinh tế, đây là hướng chọn giống mới hiện nay, góp phần nâng cao hiệu quả của quá trình chọn giống.

Đa hình nucleotide đơn được xác định do sự biến đổi trình tự DNA xảy ra khi một nucleotide (A, T, C hoặc G) trong trình tự bộ gen bị thay đổi. Sự phân bố của các SNP (Single Nucleotide Polymorphism) trong hệ gen không đồng nhất. SNP thường xảy ra với tần số khác nhau trong các phần nhiễm sắc thể khác nhau và trong các vùng không mã hóa thường cao hơn so với các vùng mã hóa [2], [3]. Hầu hết các SNP đều ở dạng hai alen và liên quan đến sự thay thế Cytosine (C) với Thymine (T), đây là biến thể phong phú nhất trong chuỗi DNA giữa các cá thể trong một quần thể [4]. Những biến đổi của SNP có thể làm thay đổi tính trạng và được sử dụng để đánh giá đa dạng di truyền và tiến hóa. Trung bình ở bộ gen gà, cứ 225 bp sàng lọc được 1 SNP; đối với bộ gen người, khoảng 1.250 bp sàng lọc được 1 SNP [5]. Ở vùng gen mã hóa các SNP có thể tác động đến 50 tính trạng quan tâm, do đó rất hiệu quả trong việc xác định mối tương quan giữa SNP và tính trạng nào đó [6]. Lần đầu tiên ứng dụng SNP trong nghiên cứu chọn giống tu hài được tiến hành. Ở các loài thủy sản khác, các SNP được ứng dụng để sàng lọc các gen tiềm năng liên quan đến tính trạng tăng trưởng ở cá hồi *Salvelinus alpinus* [7], cá chẽm [8], tôm thẻ chân trắng *Penaeus (Litopenaeus) vannamei* và tôm sú *P. monodon* [9], tôm càng xanh [10] và cá tra [11]. Hiện nay có rất nhiều phương pháp để xác định các SNP [12], tuy nhiên sử dụng công nghệ giải trình tự gen thế hệ mới được coi là phương pháp tối ưu vì giải quyết được đồng thời nhiều vấn đề và sàng lọc được lượng lớn các SNP phục vụ cho việc phân tích di truyền.

Trong nghiên cứu này, chúng tôi tiến hành giải trình tự gen thế hệ mới và sàng lọc chỉ thị SNP cho hai nhóm tu hài và marker SNP sàng lọc được sẽ cung cấp nguồn cơ sở dữ liệu quan trọng phục vụ cho công tác chọn giống ở tu hài liên quan đến tính trạng tăng trưởng trong thời gian tới.

2. Phương pháp nghiên cứu

2.1. Vật liệu nghiên cứu

2.1.1. Đối tượng nghiên cứu

Tu hài vòi trắng *(Lutraria rhychaena,* Jonas 1844*)* được nuôi tại Trung tâm giống thủy sản lợ mặn đặt tại Vân Đồn, Quảng Ninh trực thuộc Trường Cao đẳng Kinh tế, Kỹ thuật và Thủy sản – Từ Sơn – Bắc Ninh.

2.1.2. Mẫu nghiên cứu

Tổng số 60 mẫu tu hài nghiên cứu, trong đó có 30 mẫu tu hài tăng trưởng nhanh và 30 mẫu tu hài tăng trưởng chậm được cung cấp từ kết quả nghiên cứu của đề tài "ứng dụng di truyền số lượng và di truyền phân tử chọn giống tu hài tăng trưởng nhanh", theo hợp đồng số 01/2017/HĐ-TS-CNSH. Mẫu thu của tu hài tăng trưởng nhanh được lựa chọn từ 30 gia đình có giá trị chọn

giống ước tính EBV (Estimated Breeding Value) cao và mẫu thu của tu hài tăng trưởng chậm được lựa chọn từ 30 gia đình có giá trị EBV thấp, giá trị chọn giống được mô tả tại Bảng 1.

Bảng 1. *Giá trị chọn giống ước tính cho tính trạng tăng trưởng giữa hai nhóm tu hài*

Giá trị	Nhóm tăng trưởng nhanh	Nhóm tăng trưởng chậm
Số lượng mẫu (con)	30	30
Số lượng gia đình (gia đình)	30	30
Khối lượng trung bình (g)	56,4 ± 0,20	38,5 ± 0,18
EBV trung bình	32,6 ± 28,3	-21,7 ± 27,2

Ghi chú: Số liệu biểu diễn ở dạng trung bình ± độ lệch chuẩn.

Nhóm tăng trưởng nhanh gồm 30 cá thể, khối lượng trung bình 56,4 g/con và giá trị EBV trung bình đạt 32,6. Nhóm tăng trưởng chậm gồm 30 cá thể, khối lượng trung bình 38,5 g/con và giá trị EBV trung bình đạt -21,7 (Bảng 1).

2.2. Phương pháp nghiên cứu

2.2.1. Tách chiết DNA

DNA được tách chiết từ 50 mg mẫu mô cơ màng áo của 60 cá thể tu hài bằng cách sử dụng bộ kit DNAeasy tissue extraction kit (Qiagen, Đức) theo hướng dẫn của nhà sản xuất.

DNA được kiểm tra bằng kỹ thuật điện di trên gel agarose 1,5%. Các băng DNA có kết quả sáng rõ, không bị đứt gãy được lựa chọn để tiến hành thí nghiệm. Hàm lượng DNA được xác định bằng bộ đo huỳnh quang Qubit 2.0 theo hướng dẫn của nhà sản xuất, nồng độ đảm bảo tối thiểu 3 ng/µl được sử dụng cho phân tích EzRAD.

2.2.2. Xây dựng thư viện DNA phục vụ giải trình tự hệ gen cho hai nhóm tu hài

Tiến hành gộp 30 mẫu DNA của tu hài tăng trưởng nhanh, gộp 30 mẫu DNA của tu hài tăng trưởng chậm. Sử dụng 2 enzyme cắt giới hạn *Mbo*I và *Sau*3AI để cắt đồng thời hai mẫu tu hài tăng trưởng nhanh và tăng trưởng chậm sau khi gộp. Thư viện DNA của 2 nhóm tu hài được xây dựng bằng kỹ thuật ezRAD [13].

Khuếch đại thư viện DNA: Thư viện DNA được khuếch đại với thể tích 15 µl, thành phần của phản ứng PCR gồm 6 µl Enhanced PCR Mix, 1,5 µl Primer Cocktail, 5,625 µl DNA và 1,875 µl H_2O. Chu trình nhiệt của phản ứng như sau: Biến tính ban đầu ở 94^{o}C trong 3 phút, tiếp theo là 7 chu kì ở 98^{o}C trong 20 giây, 60^{o}C trong 15 giây, 72^{o}C trong 30 giây và kéo dài ở 72^{o}C trong 5 phút, lưu giữ ở 4^{o}C.

Sản phẩm PCR được kiểm tra trên gel Agarose 1,5% nhuộm với ethidium bromide và được tinh sạch bằng Sample Purification Beads. Thư viện DNA có dải kích thước dài từ 350 – 550 bp, trong đó kích thước của adapter được chèn vào là 120 bp và trình tự của mẫu trung bình là 380 bp. Các thư viện DNA được bảo quản trong điều kiện giữ lạnh ở -20^{o}C đến -80^{o}C cho đến khi giải trình tự.

Giải trình tự Genomic của tu hài được thực hiện bằng máy Illumina NovaSeq6000, đọc kết quả được thực hiện tại Trung tâm Công nghệ Gennomic, Trường Đại học Tổng hợp Deakin, Geelong, Australia.

2.2.3. Giải trình tự và lắp ráp hệ gen tham chiếu cho tu hài tăng trưởng nhanh và tăng trưởng chậm

Cơ sở HIMB Core chạy một bộ lọc kiểm soát chất lượng tiêu chuẩn và phân tích cú pháp do máy Illumina đọc được thành các tệp FASTQ được sắp xếp theo chỉ mục. Sử dụng phần mềm Trimmoatic v0.33 để loại bỏ các adapter và các đoạn trình tự có chất lượng thấp (Q< 10) [14]. Số liệu được xử lý theo quy trình của dDocent v2.24 [15].

Phân tích *de novo* trong Stacks được thực hiện với 6 bước chính: Đầu tiên, các đoạn đọc được phân tách và kiểm tra chất lượng bởi chương trình process_radtags. Ba bước tiếp theo bao gồm: Xây dựng loci (Ustacks), tạo danh mục (Catalog) các loci (Cstacks) và khớp loci với danh mục (Sstacks). Ở bước thứ 5, chương trình stacks được tiến hành để lắp ráp và hợp nhất các contig,

phát hiện các vị trí biến thể trong quần thể và xác định kiểu gen của mỗi cá thể. Ở bước thứ 6, dựa trên kiểu gen cá thể trong quần thể, các dữ liệu SNP được xác định và xuất dưới định dạng vcf hoặc genpop.

Dựa trên hệ gen tham chiếu, các đoạn đọc được tạo thành từ hai nhóm tu hài sinh trưởng nhanh và sinh trưởng chậm được dóng hàng bằng công cụ BWA v0.7.12-r1044 [16]. Đánh số chỉ mục (index) cho các nucleotid cần so sánh với hệ gen tham chiếu thực hiện bằng phầm mềm SAMtools [17].

2.2.4. Phát hiện chỉ thị SNP cho tu hài sinh trưởng nhanh, sinh trưởng chậm

Dữ liệu SNP thô được lọc bằng phần mềm vcftools v.0.1.13 qua các thông số minor Allele Frequency (mAF > 0,05), minimum Mean Depth (mean DP ≤ 10) [18], kiểu SNP (thay thế vị trí Nucleotide), HWE (Hardy-Weinberg Equilibrium) với giá trị P – value < 0,001, Mean quality score (Q)> 30, và max-missing = 0,95 (tỷ lệ phát hiện SNP = 95% trên tất cả các cá thể). Tiếp đó, phần mềm vcffilter sẽ lọc lấy các SNP dựa vào Allele Balance (AB) >0,3, Mean mapping quality (MQM/MQMR) (0,9 – 1,05), tần số thay thế allele (PAIRED/PAIREDR) (0,05 – 1,75). Dữ liệu SNP được kiểm tra rad_haplotyper để loại bỏ các biến dị và xác định 1 SNP/contig thu được SNP đặc trưng loài (Validated SNP). Xác định SNP, xác định indel do thêm đoạn (Insertion), mất đoạn (Deletion) trình tự được tiến hành khi dóng hàng các contig dựa vào sai khác trên ít nhất bốn trình tự đọc [19]. Xử lý số liệu thô được thực hiện theo quy trình Stacks [20], [21] trên hệ điều hành Ubuntu Server 15.10.

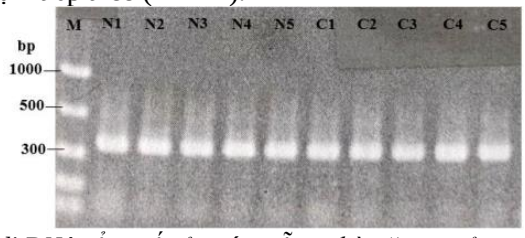
2.2.5. Kiểm định outlier loci và sàng lọc SNP tiềm năng ở nhóm tu hài tăng trưởng nhanh, tăng trưởng chậm

Outlier loci được xác định là những vị trí loci có giá trị khác biệt so với phần còn lại của bộ gen (Neutral loci). Để xác định các oulier loci được tiến hành bằng phương pháp linkage disequibrium (LD). LD được đo bằng hệ số tương quan cặp bình phương giữa các locus (r2) bằng cách sử dụng hàm 'LD" trong gói dữ liệu "genetic" thuộc phần mềm R [22]. Sử dụng LD network (Phần mềm R) để xác định Selected outlier clusters (SOC) và Compound outlier clusters (COC). Giá trị tối ưu 'LDna" [23] với các tham số φ và |E| min và ngưỡng LD được thiết lập cho các SOC. Mạng LD được xây dựng bằng gói dữ liệu "igraph" trên phần mềm R [22]. Các loci được xác định là outlier được loại bỏ để tạo bộ dữ liệu SNP trung tính. Sử dụng chương trình Lositan để xác định các oulier loci, các outlier loci được xác định bằng cách so sánh phân bố loci quan sát ở mức độ tin cậy 99% và 1% (với giá trị FDR < 0,05). Các loci bên ngoài khoảng tin cậy 99% và 1% được xác định là outlier loci.

3. Kết quả và bàn luận

3.1. Tách chiết DNA, xây dựng thư viện DNA cho nhóm tu hài tăng trưởng nhanh và tăng trưởng chậm

Kết quả điện di DNA tổng số cho thấy các dải băng DNA sáng rõ và không bị đứt gãy đảm bảo đủ điều kiện thực hiện thí nghiệm tiếp theo (Hình 1).



Hình 1. *Kết quả điện di DNA tổng số của các mẫu tu hài tăng trưởng nhanh (giếng N1 – N5), tăng trưởng chậm (giếng C1-C5), trong đó giếng M: Marker (thang chuẩn DNA).*

Sau khi tiến hành PCR, sản phẩm được điện di trên gel agarose 1,5%. Kết quả thu được dải băng gọn, sáng rõ. Thư viện DNA của nhóm tu hài tăng trưởng nhanh và tăng trưởng chậm được tinh sạch bằng SPB để loại bỏ các trình tự có độ dài trên 550 bp và nhỏ hơn 350 bp. Trình tự DNA có độ dài trong khoảng từ 350 – 550 bp được giữ lại và xác định hàm lượng DNA. Thư viện DNA của hai nhóm tu hài đảm bảo đủ điều kiện để giải trình tự.

3.2. Giải trình tự và xây dựng hệ gen tham chiếu cho hai nhóm tu hài

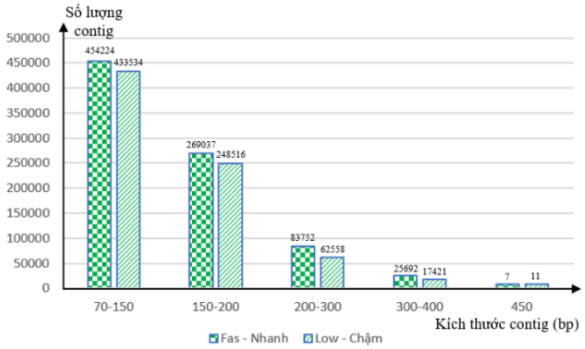
Kết quả giải trình tự và xây dựng hệ gen tham chiếu cho nhóm tu hài tăng trưởng nhanh và tăng trưởng chậm được mô tả qua Bảng 2.

Bảng 2. *Tóm tắt kết quả xử lý giải trình tự*

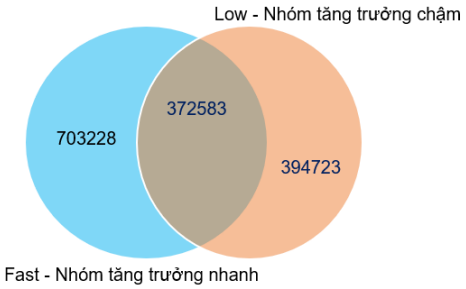
Chỉ số phân tích	Giá trị	
	Nhóm tăng trưởng nhanh	**Nhóm tăng trưởng chậm**
Độ bao phủ giải trình tự	200	200
Số lượng đoạn trình tự (read) sử dụng để kết nối contig	212.340.022	212.980.396
Tổng số contig	832.712	762.040
Kích thước hệ gen tham chiếu tạm thời (Mb)	434	429

Ở tu hài tăng trưởng nhanh, sau khi giải trình tự và phân tích đã thu được 276.023.281 đoạn trình tự thô (raw read), trong đó có 230.479.439 (chiếm 83,5%) đoạn trình tự mang barcode. Ở tu hài tăng trưởng chậm, thu được 279.136.083 đoạn trình tự thô (raw read), trong số đó có 230.845.540 (chiếm 82,7%) đoạn trình tự barcode. Sau khi tinh sạch, loại bỏ các đoạn adapter và các đoạn trình tự có chất lượng thấp ở hai nhóm tu hài, chúng tôi thu được 212.340.022 (76,8%) đoạn trình tự có chất lượng tốt ở tu hài tăng trưởng nhanh và 212.980.396 (76,3%) đoạn trình tự có chất lượng tốt ở tu hài tăng trưởng chậm. Đoạn trình tự có chất lượng tốt ở tu hài tăng trưởng nhanh và tăng trưởng chậm được sử dụng để kết nối contig và thiết lập hệ gen tham chiếu tạm thời, với dung lượng dữ liệu 200GB. Từ 212.340.022 đoạn trình tự ở tu hài tăng trưởng nhanh và 212.980.396 đoạn trình tự ở tu hài tăng trưởng chậm sau khi kết nối thu được lần lượt 832.712 và 762.040 constig với sự phân bố kích thước các contig được thể hiện ở hình 2. Contig có chiều dài từ 70 - 150 bp chiếm số lượng lớn nhất (454224 và 433534 contig), tiếp đến là các contig có chiều dài từ 150 - 200 bp (269037 và 248516 contig), contig có chiều dài 200 - 300 bp (83752 và 62558 contig), contig có chiều dài 300 - 400 bp (25692 và 17421 contig) và contig có chiều dài từ 450 bp trở lên (7 và 11 contig) chiếm số lượng thấp nhất (Hình 2).

Từ các đoạn đọc chất lượng cao, việc lắp ráp đã thu được hệ gen tham chiếu với số lượng đoạn contig cao, tỷ lệ % A+T và G+X là tương đương nhau (50:50).



Hình 2. *Phân bố contig giữa các nhóm tu hài tăng trưởng nhanh và tăng trưởng chậm sau kết nối*



Hình 3. *Số lượng SNP ở nhóm tu hài tăng trưởng nhanh và tăng trưởng chậm*

3.3. Kết quả sàng lọc chỉ thị SNP ở nhóm tu hài sinh trưởng nhanh và sinh trưởng chậm

Dựa trên hệ gen tham chiếu của tu hài đã xây dựng, phát hiện và sàng lọc được tổng số 1.470.534 SNP, trong đó có 703.228 SNP xuất hiện ở nhóm tu hài tăng trưởng nhanh, 394.723 SNP xuất hiện ở nhóm tu hài tăng trưởng chậm và 372.583 SNP xuất hiện ở cả hai nhóm tu hài (Hình 3).

Kết quả sàng lọc SNPs trên hệ gen ở cả hai nhóm tu hài cho thấy, số lượng SNPs xuất hiện ở nhóm tu hài tăng trưởng nhanh cao gấp 1,78 lần (703.228 SNPs) so với tu hài tăng trưởng chậm. Với số lượng SNP sàng lọc được góp phần mở ra hướng khai thác dữ liệu SNP có ý nghĩa thực tiễn; trên cơ sở đó xác định được các chỉ thị SNPs có khả năng liên quan đến tính trạng tăng trưởng ở tu hài. Hiện nay, phần lớn hệ gen của các loài thủy sản có giá trị kinh tế chưa được giải mã, trong đó có các loài nhuyễn thể. Việc giải mã thành công hệ gen tham chiếu cho tu hài là cơ sở để xác định được các SNPs liên kết với các gen chức năng, từ đó góp phần mở ra tiềm năng ứng dụng phục vụ chọn giống đối với công tác chọn giống và nuôi trồng tu hài. Các chỉ thị SNPs sàng lọc được chiếm tới 90% sự khác biệt về di truyền giữa các cá thể [24]. Trong quá trình trao đổi chéo thường không tách rời các chỉ thị SNPs ra khỏi gen chức năng [11].

Bảng 3. *Thông tin về SNP sau các bước lọc ở hai nhóm tu hài*

Thông số lọc	Nhóm sinh trưởng nhanh		Nhóm sinh trưởng chậm	
	Số cá thể	Số SNP	Số cá thể	Số SNP
Độ nhiễu locus 90%	30	12728	30	6424
Tần số alen (Mac) <3	30	4844	30	5748
Tần số alen nhỏ (Maf) (0,01 - 0,05)	30	1467	30	881
Khuy INDV	30	12728	30	6424
Tần số alen nhỏ nhất < 0,05 và tần số alen lớn nhất > 0,464	30	2503	30	1660
Vị trí liên kết haplotype	30	663	30	386
HWE < 0,001	30	1467	30	604
Loại bỏ LD	30	525	30	211
Locus trung tính/ Locus khác biệt	**30**	**525/138**	**30**	**211/175**

Dữ liệu SNPs thô thu được từ hai nhóm tu hài được sàng lọc qua các bước bằng công cụ VCF tool để loại bỏ các biến thể. Số lượng các chỉ thị SNP ở hai nhóm tu hài sau khi phát hiện và tuyển chọn thu được 525 SNP ở tu hài tăng trưởng nhanh và 211 SNP ở tu hài tăng trưởng chậm. Kết quả chi tiết việc sàng lọc các SNP được mô tả ở (Bảng 3).

3.4. Kiểm định outlier loci và sàng lọc SNP tiềm năng ở nhóm tu hài tăng trưởng nhanh, tăng trưởng chậm

Kết quả kiểm định các outlier loci cho cả hai nhóm tu hài đã phát hiện được 138 outlier loci ở nhóm tu hài tăng trưởng nhanh và 175 outlier loci ở tu hài tăng trưởng chậm. Tiến hành loại bỏ các locus này, cuối cùng đã thu được 525 và 211 locus trung tính (Neutral loci). Việc chọn lọc các SNP tiềm năng từ các SNP trung tính ở hai nhóm tu hài được thực hiện nhờ sử dụng các công cụ tin sinh học trên hệ điều hành linu. Locus trung tính của cả hai nhóm tu hài được xây dựng dựa trên tham chiếu dDocent và Stacks. Đồng thời, tiến hành phân tích liên kết biến dị thay thế đồng hoán (Transition) và biến dị thay thế dị hoán (Tranversion), đã xác định được 4 biến dị transition (đồng hoán) với tổng số 340 SNP ở nhóm tu hài tăng trưởng nhanh và 150 SNP ở nhóm tăng trưởng chậm. Ngược lại, xác định được 8 biến dị tranversion (dị hoán) với tổng số 185 SNP ở nhóm tu hài tăng trưởng nhanh và 61 SNP ở nhóm tăng trưởng chậm. Việc kiểm định các neutral loci (loci trung tính) nhằm tăng độ chính xác của các SNP sàng lọc được [25].

Từ việc phân tích các liên kết và sàng lọc đã thu được 100 SNP cho mỗi nhóm tu hài dựa trên sự ổn định của các nucleotide ở đầu 3'. Từ 100 chỉ thị SNP tiềm năng tiếp tục được sàng lọc dựa vào các tiêu chí là chỉ số QUAL (chọn chỉ số cao nhất), hệ số FPKM giữa tu hài tăng trưởng

nhanh và tăng trưởng chậm, chức năng được chú giải cho từng chỉ thị của gen quy định tính trạng tăng trưởng. Kết quả đã sàng lọc được 32 chỉ thị SNP tiềm năng nhất, trong đó có 20 SNP tiềm năng cho nhóm tu hài tăng trưởng nhanh và 12 SNP tiền năng cho nhóm tu hài tăng trưởng chậm (Bảng 4).

Bảng 4. *Vị trí contig và SNP ở hai nhóm tu hài*

STT	Nhóm sinh trưởng nhanh				Nhóm sinh trưởng chậm			
	No. Stacks contig	**Vị trí SNP**	**No. dDocent contig**	**SNP**	**No. Stacks contig**	**Vị trí SNP**	**No. dDocent contig**	**SNP**
1	4256*	14	1731*	T/A	42278	52	8465	T/C
2	89920*	43	5700*	G/C	**160179***	**69**	**17437***	**G/C**
3	23172*	46	12709*	T/A	**194687***	**70**	**17765***	**G/C**
4	13193*	49	3570*	T/G	**209666***	**75**	**33025***	**G/C**
5	**61085***	**52**	**3881***	**G/C**	206834*	78	14306	C/T
6	**103149***	**61**	**8058***	**T/C**	169228*	89	37770	T/C
7	14190	73	11689	A/G	19013	91	38550	T/C
8	6162	75	3765	A/C	**53272***	**94**	**12648***	**T/C**
9	**4649***	**88**	**7534***	**C/T**	**204613**	**96**	**8330**	**G/A**
10	**225978**	**89**	**1121**	**T/C**	**10876***	**130**	**1365***	**C/A**
11	168086	98	4771	A/G	**55129**	**132**	**36889**	**T/C**
12	120192	99	9875	C/T	54383*	146	55268*	G/T
13	200347	100	7897	A/G				
14	**38178***	**102**	**5226***	**G/A**				
15	**19584***	**104**	**14664***	**G/A**				
16	9952	112	3770	G/A				
17	86987	115	6904	T/C				
18	12694	117	10792	C/T				
19	**9980**	**121**	**2440**	**C/G**				
20	**84806**	**133**	**8290**	**T/C**				

*Ghi chú: * Kiểu biến dị thay thế dị hoán; Số bôi màu đậm là các SNP có thể thiết kế được mồi.*

4. Kết luận

Nghiên cứu đã xác định được kích thước hệ gen tham chiếu cho tu hài tăng trưởng nhanh là 434Mb và tu hài sinh trưởng chậm là 429Mb. Sàng lọc và phát hiện được tổng số 1097951 điểm SNP liên quan đến tăng trưởng ở cả hai nhóm tu hài, trong đó có 703228 điểm SNP ở tu hài tăng trưởng nhanh và 394723 điểm SNP ở tu hài tăng trưởng chậm. Từ 1097951 điểm SNP ở cả hai nhóm tu hài có 372.583 SNP xuất hiện chung ở cả hai nhóm tu hài.

Nghiên cứu đã sàng lọc được 32 SNP marker tiềm năng có liên quan đến tính trạng tăng trưởng ở tu hài, trong đó có 20 SNP ở tu hài liên quan đến tăng trưởng nhanh và 12 SNP ở tu hài liên quan đến tăng trưởng chậm.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] T. B. Thai, P. Y. Lee, M. H. Gan, C. M. Austin, L. J. Croft, A. T. Trieu, and H. M. Tan, "Whole Genmon Asembly of the Snout Otter Clam, *Lutraria rhynchaena*, Using Nanopore and lllumina Data, Benchmarked Against Bivalve Genome Assemblies," *Frontiers in Gentics*, vol. 10, p. 158, 2019.

[2] E. S. Lander, "The new genomics: global views of biology," *Science*, vol. 274, no. 5287, pp. 536-539, 1996.

[3] Z. Liu and J. Cordes, "DNA marker technologies and their applications in aquaculture genetics," *Aquaculture*, vol. 238, pp. 1-37, 2004.

[4] J. Fernández, M. A. Toro, A. K. Sonesson, and B. Villanueva, "Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress," *Front. Genet*, vol. 5, p. 414, 2014.

[5] Z. Liu, "Single nucleotide polymorphism (SNP)," In: *Liu, Z. (Ed.), Aquaculture Genome Technologies*. Blackwell, USA, 2007, pp. 59-72.

[6] N. D. Beuzen, M. J. Stear, and K. C. Chang, "Molecular markers and their use in animal breeding," *Veterinary Journal*, vol. 160, pp. 42-52, 2000.
[7] W. J. Tao and E. G. Boulding, "Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus L.*)," *Heredity*, vol. 91, pp. 60-69, 2003.
[8] Y. X. Xu, Z. Y. Zhu, L. C. Lo, C. M. Wang, G. Lin, F. Feng, and G. H. Yue, "Characterization of two parvalbumin genes and their association with growth traits in Asian seabass (*Lates calcarifer*)," *Anim Genet*, vol. 37, pp. 266-268, 2006.
[9] K. L. Glenn, L. Grapes, T. Suwanasopee, D. L. Harris, Y. Li, K. Wilson, and M. F. Rothschild, "SNP analysis of AMY2 and CTSL genes in Litopenaeus vannamei and Penaeus monodon shrimp," *Animal Genetics*, vol. 36, pp. 235-236, 2005.
[10] M. T. Nguyen, A. C. Barnet, P. B. Mather, Y. Li, and R. E. Lyons, "Correlation of SNP (Single Nucleotide Polymorphisms) in the crustacean hyperglycemic hormone genes with individual growth performance in giant freshwater prawn *Macrobrachium rosenbergii*," *Conference Proceedings National Fisheries Science*, Nong Lam University, 2011.
[11] M. T. Nguyen, T. M. T. Vo, H. Jung, and P. Mather, "A transcriptomic analysis of th kidney tisue of tra catfish *(Pangansianodon hypophthlmus)* reared in saline condition: *De no vo assembly, annotation, SNP discovery*," *Journal Biology*, vol. 37, no. 2, pp. 220-227, 2015.
[12] D. T. Nguyen, "DNA marker techniques in study and selection of plant," *Jounal Biology*, vol. 36, no. 3, pp. 265-294, 2014.
[13] R. J. Toonen, J. B. Puritz, Z. H. Forsman, J. L. Whitney, I. Fernandez-Silva, K. R. Andrews, and C. E. Bird, "ezRAD: a simplified method for genomic genotyping in non-model organisms," *PeerJ*, vol. 1, no. e203, 2013.
[14] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114-2120, 2014.
[15] J. B. Puritz, C. M. Hollenbeck, and J. R. Gold, "dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms," *PeerJ*, vol. 10, no. 2, 2014, Art. no e341.
[16] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows - Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754-1760, 2009.
[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, and N. Homer, "Genome Project Data Processing, the Sequence Alignment/Map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, no. 16, pp. 2078-2079, 2009.
[18] Z. Gao, W. Luo, H. Liu, C. Zeng, X. Liu, S. Yi, and W. Wang, "Transcriptome Analysis and SSR/ SNP Markers Information of the Blunt Snout Bream (*Megalobrama amblycephala*)," *PLOS ONE*, vol. 7, no. 8, 2012, Art. no. e42637.
[19] J. M. Catchen, A. Amore, P. Hohenlohe, W. Cresko, and J. H. Postlethwait, "*Stacks*: Building and Genotyping Loci *De Novo*," *From Short-Read Sequences*, vol. 1, no. 3, p. 11, 2011.
[20] N. C. Rochette, A. G. Rivera-Colón, and J. M. Catchen, "Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics," *Molecular ecology*, vol. 28, no. 21, p. 18, 2019.
[21] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *Inter Journal, Complex Syst*, 2006, Art. no. 1695.
[22] P. Kemppainen, C. G. Knight, D. K. Sarma, T. Hlaing, A. Prakash, and Y. N. Maung, "Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure," *Mol Ecol Resour*, vol. 15, pp. 1031-1045, 2015.
[23] T. Antao, A. Lopes, R. J. Lopes, A. Beja-Pereira, and G. Luikart, "LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method," *BMC bioinformatics*, vol. 9, p. 323, 2008.
[24] M. Salem, R. L. Vallejo, T. D. Leeds, Y. Palti, S. Liu, and A. Sabbagh, "RNA-Seq identifies SNP markers for growth traits in rainbow trout," *PLoS One*, vol. 7, no. 5, 2012.
[25] F. W. Allendorf, P. A. Hohenlohe, and G. Luikart, "Genomics and the future of conservation genetics," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 697-709, 2010.