Identify and characterize the SNP markers for traceability of scallop spiny lobster \textit{Panulirus homarus} (Decapoda, Palinuridae) in Vietnam

N N T Nhon\textsuperscript{1}, T Q Sang\textsuperscript{2}, N M Chau\textsuperscript{1} and D T Binh\textsuperscript{2}

\textsuperscript{1}Research Institute of Aquaculture, Nha Trang, Vietnam
\textsuperscript{2}Institute for Biotechnology and Environment, Nha Trang, Vietnam

Email: binhdt@ntu.edu.vn

Abstract. Lobster is an aquatic species of ecological and economic importance. In Vietnam, ornate spiny lobster (\textit{Panulirus ornatus}) and scallop spiny lobster (\textit{P. homarus}) are among the main species of aquaculture in Vietnam. Demand for lobster products domestically and internationally is increasing, as in the US, Japan and China markets. Today, consumers are more aware of product brands, food safety and hygiene, so product traceability is essential to protect consumers and businesses. The DNA markers are widely used for commercially traded seafood traceability. Among that, Single Nucleotide Polymorphisms (SNPs) have been shown to be robust markers, and offer high reliability. EzRAD sequencing is applied to detect potential SNPs that are characteristic for \textit{P. homarus} populations in Vietnam and Sri Lanka. Through 211 valid SNPs, we detected and 12 SNP\textsuperscript{S} occurring in Vietnam population, and designed 8 primer pairs to amplified 80-130 bp SNPs products. PCR optimization allows us to select 2 SNPs serving the traceability of Vietnamese \textit{P. homarus}. Further need to more population sampling, and whole genome sequencing in order to develop SNPs panel to seeking the origin of lobster seafood products.

1. Introduction
Lobsters are a highly valued species in the seafood industry, especially in Asia, Europe, and America [1,2]. There are about 7 species of lobster recognized in Vietnam, of which ornate spiny lobster (\textit{Panulirus ornatus}) and scallop spiny lobster (\textit{P. homarus}) are the two most widely cultured species [3,4]. Like most crustaceans, lobsters have a complex metamorphic life cycle, and a long planktonic larval stage [5,6]. That could lead to high connectivity in local populations, even on inter-oceans scales. Consequently, wild-caught lobsters can come from a variety of sources.

Over the past decades, lobster farming has developed rapidly, mainly on family scale, lack of planning and management in Vietnam. The process of artificial reproduction is not yet closed, the breeds are mainly in the wild, or illegally imported of unknown origin [3].

Along with the development of the society, the traceability of fishery products has become very urgent, in order to protect consumers and businesses. Traceability has been successfully implemented on many aquatic species [7], and initially conducted in lobsters [8].

In addition to traditional techniques, modern molecular tools are increasingly used in food traceability [9–11], identification of exotic species and their hybrids [12], and wildlife conservation as well [13].

In this study, we detected validated SNPs from \textit{P. homarus} populations in Vietnam and Sri Lanka, and on that SNPs panel, selected potential SNPs for use in traceability of Vietnamese lobsters.
2. Material and methods

2.1. Sampling and EzRAD libraries preparation

Wild scallop spiny lobsters (*P. homarus*) were collected in Vietnam and Sri Lanka, with 30 individuals per location. In Vietnam, the samples collected in the Binh Thuan, Khanh Hoa, Binh Dinh, Quang Ngai provinces, and Da Nang city, which represent the distribution area in the Central region, each province chooses 6 individuals. Due to geographical distance, lobster samples were collected only at one location (Negombo) in Sri Lanka. The sampling location is shown in Figure 1. Lobster samples were obtained directly from lobster fishermen. After precise identification based on morphological characteristics, a piece of muscle or swimming leg was stored in 95% molecular-grade ethanol, then transferred to the laboratory and used for further analysis.

DNA isolation was carried out using the Qiagen DNeasy Blood and Tissue kit (Hilden, Germany), and high-quality extracts were confirmed using gel electrophoresis, stained with ethidium bromide.

Figure 1. Sampling location (green circles) of *Panulirus homarus* (drawing image) from Vietnam and Sri Lanka.

Purified DNA from each lobster individual was simultaneously digested with two restriction enzymes: MboI and Sau3AI (NEB). Cleaned digestions were inserted directly into the Illumina TruSeq nano DNA library Prep kit following the Sample Preparation v2 Guide starting with the “Perform End Repair” step for one-third volume reactions (Supplement S1 [14]). Digested libraries were end-repaired, 350 bp size-selected by SP bead. The 3’ ends of selected libraries were then adenylated and Illumina adapters were ligated to the digested genomic DNA samples. PCR reactions were performed using a total volume of 15 μl including 1.5 μl Illumina PCR Primer Cocktail, 6 μl
Illumina Enhanced PCR Mix, 1.875 µl ddH₂O and 5.625 µl DNA libraries under the following temperature program: initial denaturation at 95°C for 3 min, followed by 8 cycles of 98°C for 20s, 60°C for 15s and 72°C for 30s. Final extension was done at 72°C for 5 min and the soaking temperature was set to 4°C. PCR products were purified using SP Beads (1:1), and quantified using qPCR. DNA libraries were sequenced as paired-end 100 bp runs on HiSeq 2500/4000 system (Illumina) in Texas A&M University Corpus Christi Genomics Core Laboratory, USA.

2.2. SNPs discovery and filtering
At first, raw FastQ files were trimmed using Trimmomatic v0.3 [15] to simultaneously remove Illumina adapter sequences, and any bases that had a quality score (Q-score) of less than 10 [16].

De novo analysis was conducted following approaches: 1) Stacks v2.4 [17,18] is performed with six main steps: First, the readings are separated and checked for quality by the process_radtags program. The next three steps include: building loci (ustacks), creating a catalog (catalog) of loci (cstacks) and matching loci to directory (sstacks). In step 5, the gstacks program is conducted to assemble and fuse contigs, detect variation locations in the population and determine the genotype of each individual. In step 6, based on the individual genotype in the population, the SNP data are identified and exported as vcf format; 2) In order to achieve the reference genome with more length and number of contigs than that obtained from Stacks analysis. De novo reference assembly was done using Rainbow v2.0.2 [19] and CD-HIT v4.6.1 [20,21] based on overall sequence similarity (90% by default).

SNP calling was performed using population of Stack pipeline [17,18] with default parameters. Raw SNP files (from Stacks) were concatenated into a single variant call format (VCF) file using VCFtools v0.1.16 [22]. The raw SNPs were then filtered with VCFtools. LD was measured as the squared pairwise correlation coefficient between loci (r²) calculated using the ‘LD’ function in the R package ‘genetics’ [23] Selected outlier clusters (SOC) and Compound outlier clusters (COC) were identified by LD network analysis using R package ‘LDna’ [24], optimal value of φ and |E|min parameter and LD threshold was set up for SOC. LD network were constructed using the R package ‘igraph’ [25]. All loci putatively identified by either programs were removed from the dataset to generate a panel of neutral SNPs, which were used for further analyses.

2.3. Population structure analysis
Population structure was investigated through discriminant analysis of principal components (DAPC) using the R package Adegenet v.1.4-1. This analysis assigns individuals to clusters and selects the best number of clusters based on Bayesian information criterion (BIC) [26].

2.4. SNPs selection for lobster traceability
The neutral SNP data of the P. homarus populations were arranged based on the genotype of each individual according to sampling sites (Vietnam and Sri Lanka) using the package “vcfR” [27] of R software.

Population-specific SNP panel from the Stacks reference genome were aligned to the dDocent reference genome to verify and redefine the SNP sites. The SNP-containing contigs in Vietnamese population retrieved from two genomes were assembled using the Cap3 tool [28]. The primer pair is designed to amplify the DNA fragment containing putative SNP using the AS-PCR (Allele-Specific PCR assay) method [29], based on standard parameters such as primer length (18-22 nucleotides), melting temperature (50-65°C), GC ratio (50-60%), hairpin temperature (<20°C) and specific SNP position (3 'head').

PCR reactions were optimized with the temperature gradient from 52-62°C, a step of 2°C using specimens of P. homarus from Vietnam and Sri Lanka. PCR reactions were performed using a total volume of 25 µl with 12.5 µl Master mix (GoTaq® Green master mix, Promega, US)), 1 µl each primer (10 mM), 5 µl DNA template and distilled water to the final volume. PCR products were run on 1.5% agarose gel for confirmation of equal length against an appropriate size marker.
3. Results and discussion

3.1. Discovering and detecting and SNP markers

RAD sequencing efforts for *P. homarus* in Vietnam and Sri Lanka (18 individuals each) generated a total of 278,971,023 paired-end, 151 bp reads, which, when filtered and aligned to create a catalogue, resulted in a total of 1,762,040 RAD tags used to generate a de novo reference of 263,980,396 bp. Using the Freebayes tools for SNP calling, the initial dataset consisted of 394,723 raw SNPs. Following extensive filtering, the final dataset consisted of 386 putative SNPs in 27 individuals (Table 1). LD analysis detected 175 outlier loci; these loci were removed from the SNP panel and the 211 remaining loci were assumed to be neutral.

**Table 1.** Information on individuals removed at each step of SNPs filtering for *P. homarus*. Mac (Minor allele count), AF (Allele Frequency), Maf (Minor allele frequency), Indv. (Individuals).

| Parameters          | Missing loci 0% | Missing indiv. | Mac <3 | AF < 0.05, AF > 0.464 | Maf (0.01-0.05) | HWE < 0.001 | Rad haplotype | LD |
|---------------------|----------------|----------------|--------|------------------------|-----------------|-------------|---------------|----|
| No of Indiv.        | 36             | 27             | 27     | 27                     | 27              | 27          | 27            | 27 |
| No of SNPs          | 6424           | 6424           | 5748   | 1660                   | 881             | 604         | 386           | 211|

3.2. Population structure

DAPC analysis using neutral SNP panels detected two minor overlapping clusters, representing each of the two countries (Figure 2).

**Figure 2.** First principal component resulting from the discriminant analysis of principal components (DAPC) using the neutral SNP panel. The red group represents 18 individuals from Vietnam, while the blue group represents 18 Sri Lanka individuals.

The neutral SNP panel of *P. homarus* was mapped using the 2 references genomes of Stacks and dDocent (10,141,132 bp, 32,802 RAD tag with max length 240 bp), resulting in max 290 bp containing SNP contigs. SNPs were then screened, and 12 putative SNPs were detected for the Vietnamese *P. homarus* population (Table 2). Among that, only 8 SNPs can be used for primer designed with tentative PCR products from 80-200 bp (Table 2).
Table 2. Information for putative SNPs and designed primers for *P. homarus* traceability in Vietnam.

| No | Stacks contig | Position dDocent contig | Position SNP | Primers | Size (bp) |
|----|---------------|-------------------------|--------------|---------|----------|
| 1  | 10876         | 135                     | 1361         | C/A     | X1361F 5’CACCACATATTGTCTCCTAAAA 3’ X1361R 5’GGTATGTTTGAAGGTTAGTGG3’ | 120 |
| 2  | 19013         | 91                      | 38536        | T/C     | X12648F 5’CTGCTTCACCATCTGACCCTCT3’ X12648R 5’TGAGATGAGGAGGAGGTG3’ | 125 |
| 3  | 42278         | 52                      | 8465         | T/C     | X35889F 5’GATCGGAGGATAAGCTGACAT3’ X35889R 5’AAACAACCACGTGACCCCAC3’ | 120 |
| 4  | 53272         | 65                      | 12648        | T/C     | X17437F 5’TACACGAAACAGCTGCGTG3’ X17437R 5’TCACTCGAACTATGAGGCA3’ | 80  |
| 5  | 54383         | 146                     | 45268        | G/T     | X8330F 5’CGTAAATTGTTGTCACATCC3’ X8330R 5’GGGCTAGATGAATGTATTC3’ | 111 |
| 6  | 55129         | 132                     | 35889        | T/C     | X14306F 5’CCACTCCTGATAGGGCTC3’ X14306R 5’GGAAGGGGATACACCCAGG3’ | 114 |
| 7  | 160179        | 85                      | 17437        | G/A     | X33025F 5’TTGAACCTCTGTCATAACCG3’ X33025R 5’TTATGGCGAAGCTCAGG3’ | 110 |
| 8  | 169228        | 89                      | 37737        | T/A     | X17765F 5’AGTAATGTCCTGAACTACAAGG3’ X17765R 5’ATCTGTAGGGCTTC3’ | 200 |
| 9  | 194687        | 73                      | 17765        | G/A     | X17765R 5’ATCTGTAGGGCTTC3’ | 114 |
| 10 | 204613        | 72                      | 8330         | G/A     | X8330F 5’CGTAAATTGTTGTCACATCC3’ X8330R 5’GGGCTAGATGATGGTTC3’ | 111 |
| 11 | 206834        | 78                      | 14306        | C/T     | X14306F 5’CCACTCCTGATAGGGCTC3’ X14306R 5’GGAAGGGGATACACCCAGG3’ | 114 |
| 12 | 209666        | 44                      | 33025        | G/C     | X33025F 5’TTGAACCTCTGTCATAACCG3’ X33025R 5’TTATGGCGAAGCTCAGG3’ | 110 |

Optimal results for 8 pairs of primers, 4 primer pairs (33025, 8330, 17765, and 17437) gave no PCR products; two primer pairs showed PCR product at all temperatures (35889), or at 54-58°C in both populations, two (14306 and 12648) showed high specificity when it only appeared in Vietnamese populations in all temperature ranges (54-62°C), and at 54-58°C, respectively (Figure 3).

**Figure 3.** The results optimized the PCR reactions of specific SNPs for *P. homarus* with 4 pairs of primers 35889 (A), 1361 (B), 14306 (C), and 12648 (D). M: DNA marker 100 bp, VN: Vietnam, SR: Sri Lanka.
Seafood traceability is growing demands from sociological, health and conservation perspective. Currently, the protection of the rights of consumers and producers and the prevention of fraud are becoming increasingly important and mandatory issues for the food industry. These problems are now more effectively addressed by enabling a traceability system that allows food network stakeholders, suppliers and users, on an informational basis, provided at each segment of the product value chain. Traceability has already been successfully implemented in the seafood industry [7,30,31]. Currently, there are several tools for traceability of food such as protein identification, metabolic reaction and DNA analysis [32]. Among that, DNA-technology, make it possible to identify species even in processed products [9,10,33].

Nowadays, consumers are growing concerned about ecological impacts and food traceability is clearly the going trend at any scales [8,31]. Lobster is a high commercial value species that is consumed in many high-income countries. Lobster traceability faces several challenges such as multiple small suppliers and finished products [8], which can be overcome by molecular markers. Among the variety of markers applied, SNP markers have proved as robust, and reliable tool for trace the species origin, determine the exotic species and hybrid [12], and population connectivity [34–36].

Herein, we generated the SNP panel from \( P. ho\text{marus} \) populations in Vietnam, and Sri Lanka. Among that, we detected 12 putative SNPs, of which 8 SNPs positions allowed for primer designing. Although only 2 SNPs survived the PCR optimization steps to differentiate \( P. homarus \) from Vietnam and Sri Lanka, this is also the initial result showing the application of SNPs in lobster traceability. Growing demand and new sequence technology lead to the replacement of SNPs markers to Microsatellites for food traceability [37,38]. Eighteen highly informative SNPs located in different genes have been applied for bovine meat traceability in six European cattle breeds [37]. Around 109 out of the 163 parentage SNPs were discovered from different breeds of North American sheep, and successful applied for traceability in 96 families (sire, dam, and non-identical twin lambs [39]. For the striped catfish (\( Pangasianodon hypophthalmus \)), Vo et al (2018) [40] detected 2 out of 12 validate SNPs, which can trace the origin of Vietnamese catfish from other countries such as Cambodia, Thailand, and Bangladesh. According to Bylemans et al (2016) [38], SNPs can be used to detect subtle genetic differences between wild and captive-bred populations in the case of tracing captive-bred marine fish for fisheries management and wildlife forensics. Using two SNP panels (neutral versus outliers), Villacorta– Rath et al (2016) [41] can trace country of origin of Southern rock lobster (\( Jasus edwardsii \)) in Australia and New Zealand.

4. Conclusion
SNP markers are increasingly being applied in many aquatic species, and for diverse fields. SNP markers may not have high polymorphism like microsatellites, but are advantageous in number and distribution throughout the genome. Detecting polymorphic and population-specific SNPs is essential, especially for species of economic value such as lobsters. The genome of \( P. homarus \) needs to be investigated, and genotype examination throughout its distribution range.

References

[1] Phillips B F 2013 Lobsters: Biology, management, aquaculture and fisheries (UK: John Wiley & Sons, Ltd Wiley-Blackwell)
[2] Nagai Y, Fujikake N, Ohno K, Higashiyama H, Popiel H A, Rahadian J, Yamaguchi M, Strittmatter W J, Burke J R and Toda T 2003 Prevention of polyglutamine oligomerization and neurodegeneration by the peptide inhibitor QBP1 in Drosophila. Hum. Mol. Genet. 12 1253–9
[3] Tuan L A and Mao N D 2004 Present status of lobster cage culture in Vietnam. ACIAR lobster ecology workshop pp 24–30
[4] Hung L Van and Tuan L A 2017 Lobster sea-cage culture in Vietnam The Australian Centre for International Agricultural Research (ACIAR) pp 10–7
[5] Holthuis L B 1991 FAO species catalogue. Vol. 13. Marine lobsters of the world. An annotated and illustrated catalogue of species of interest to fisheries known to date. FAO Fish. Synopsis 125 1–292

[6] George R W and Holthuis L B 1965 A revision of the Indo-West Pacific spiny lobsters of the Panulirus japonicus group Zoölogische Verh. 72 1–36

[7] Angélica M, Díaz N F, Lamas C, Uribe C and Araneda C 2014 Traceability of mussel (Mytilus chilensis) in southern Chile using microsatellite molecular markers and assignment algorithms. Exploratory survey Food Res. Int. 62 104–10

[8] Cube Automation Food and Seafood Division 2015 Lobster Traceability Pilot Project Traceability implementation Suggested guidelines for the lobster industry

[9] Martins-lopes P, Gomes S, Pereira L and Guedes-pinto H 2013 Molecular Markers for Food Traceability Food Technol. Biotechnol. 9862 198–207

[10] Galimberti A, Mattia F De, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S and Labra M 2013 DNA barcoding as a new tool for food traceability Food Res. Int. J. 50 55–63

[11] Scarano D and Rao R 2014 DNA Markers for Food Products Authentication Diversity 6 579–96

[12] Ellis C D, Jenkins T L, Svanberg L, Eriksson S P and Stevens J R 2020 Crossing the pond: genetic assignment detects lobster hybridisation Sci. Rep. 1–7

[13] Iyengar A 2014 Forensic DNA analysis for animal protection and biodiversity conservation: A review J. Nat. Conserv. 22 195–205

[14] Peterson B K, Weber J N, Kay E H, Fisher H S and Hoekstra H E 2012 Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species PLoS One 7

[15] Bolger A M, Lohse M and Usadel B 2014 Trimmomatic: A flexible trimmer for Illumina sequence data Bioinformatics 30 2114–20

[16] Toonen R J, Puritz J B, Forsman Z H, Whitney J L, Fernandez-Silva I, Andrews K R and Bird C E 2013 ezRAD: a simplified method for genomic genotyping in non-model organisms PeerJ 1 e203

[17] Rochette N C, Rivera-Colón A G and Catchen J M 2019 Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics Mol. Ecol. 28 4737–54

[18] Catchen J M, Amores A, Hohenlohe P, Cresko W and Postlethwait J H 2011 Stacks: Building and genotyping loci de novo from short-read sequences G3 Genes, Genomes, Genet. 1 171–82

[19] Chong Z, Ruan J and Wu C-I 2012 Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics 28 2732–7

[20] Li W and Godzik A 2006 Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences Bioinformatics 22 1658–9

[21] Fu L, Niu B, Zhu Z, Wu S and Li W 2012 CD-HIT: accelerated for clustering the next-generation sequencing data Bioinformatics 28 3150–2

[22] Banks E, Lunter G, Albers C A, Durbin R, Danecek P, Auton A, Handsaker R E, Abecasis G, McVean G, Marth G T, Sherry S T and DePristo M A 2011 The variant call format and VCFtools Bioinformatics 27 2156–8

[23] Warnes G 2019 Package ‘genetics’ 43 pp

[24] Kemppainen P, Hlaing T, Walton C, Somboon P, Knight C G, Mahanta J, Maung Maung Y N, Prakash A and Sarma D K 2015 Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure Mol. Ecol. Resour. 15 1031–45

[25] Csardi, G. and Nepusz, T. 2006 The igraph software package for complex network research, InterJournal, Complex Systems 1695.

[26] Jombart T, Sébastien D and François B 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations BMC Genet. 11 1–15

[27] Knaus B and Grünwald N 2017 VCFR: a package to manipulate and visualize variant call format
data Mol. Ecol. Resour. 17 44–53.

[28] Huang X and Madan A 1999 CAP3: A DNA Sequence Assembly Program Resource 868 Genome Research Genome Res. 868–77

[29] Wangkumhang P, Chaichoompu K, Ngamphiw C, Ruangrit U, Chanprasert J, Assawamakin A and Tongsima S 2007 WASP: A Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations BMC Genomics 8 1–9

[30] Maldini M, Nonnis Marzano F, Fortes G G, Papa R and Gandolfi G 2006 Fish and seafood traceability based on AFLP markers: Elaboration of a species database Aquaculture 261 487–94

[31] Hosch G and Blaha F 2017 Seafood traceability for fisheries compliance: country-level support for catch documentation schemes (FAO Fisheries and Aquaculture Technical Paper 619, Rome, Italy)

[32] Leal M C, Pimentel T, Ricardo F, Rosa R and Calado R 2015 Seafood traceability: Current needs, available tools, and biotechnological challenges for origin certification Trends Biotechnol. 33 331–6

[33] Jérôme M, Martinsohn J T, Ortega D, Carreau P, Verrez-Bagnis V and Mouchel O 2008 Toward fish and seafood traceability: Anchovy species determination in fish products by molecular markers and support through a public domain database J. Agric. Food Chem. 56 3460–9

[34] Kruck N C, Innes D I and Ovenden J R 2013 New SNPs for population genetic analysis reveal possible cryptic speciation of eastern Australian sea mullet (Mugil cephalus). Mol. Ecol. Resour. 13 715–25

[35] Bylemans J, Maes G E, Diopere E, Carvalho G R and Nielsen E E 2011 Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Resour. 11 Suppl 1 123–36

[36] Vo T T B, Nguyen H D, Nguyen T T H and Tran H T T 2018 Identification and analysis of snps in population of Vietnamese catfish (Pangasianodon hypophthalmus), using next generation sequencing and snp validation MOJ Curr Res Rev. 1 12–9