DEVELOPMENT OF A LOW-DENSITY COMMERCIAL GENOTYPING ARRAY FOR THE WHITE LEGGED SHRIMP, *LITOPENAEUS VANNAMEI*

D.B. Jones[1], K.R. Zenger[1,2], M.S Khatkar[2,3], H.W. Raadsma [2,3], H.A.M. van der Steen[4], J. Prochaska[4,5] , S. Forêt[6] and D.R. Jerry[1,2]

[1] Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and Engineering, James Cook University, Queensland, Australia
[2] ARC Hub for Advanced Prawn Breeding, James Cook University, Townsville QLD, Australia
[3] Sydney School of Veterinary Science, Faculty of Science, The University of Sydney, Camden, NSW, Australia
[4] Global Gen, Jalan Raya Narogong Km 14, Desa Cikiwul Bantar Gebang Bekasi, 17310 Indonesia
[5] Amity Aquaculture, LLC. 1603 Capitol Ave., Suite #310 A317, Cheyenne, WY 82001 USA
[6] ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD, Australia

SUMMARY

The Pacific whiteleg shrimp, *Litopenaeus vannamei*, is the most farmed shrimp species globally. The development of high quality genomic resources including a dense array of genetic markers and genetic maps are pivotal to integrating genomic selection in this species. We describe the development and utility of an Illumina low-density single nucleotide polymorphism (SNP) array (Infinium ShrimpLD-24 v1.0) which is now commercially available. These resources set the foundation for investigating the architecture of complex traits and genomic selection.

INTRODUCTION

The whiteleg shrimp, *Litopenaeus vannamei*, is an intensively farmed species with global production exceeding 3 million tonnes annually (GLOBEFISH 2016). Current breeding programs for *L. vannamei* use traditional phenotypic selection to produce shrimp with enhanced growth and that exhibit-lowered susceptibility to various viral pathogens like Taura syndrome virus (TSV) and White spot syndrome virus (WSSV). While this traditional approach has been moderately successful in producing more productive shrimp strains, genetic progress using multi-trait phenotypic selection in *L. vannamei* is in some cases significantly impeded by an unfavourable genetic correlation between growth and disease, as well as a poor correlative response in susceptibility to multiple diseases (Gitterle *et al.* 2007, Huang *et al.* 2012, Gjedrem 2015). *L. vannamei* is an aquaculture species that would benefit substantially from the integration of genomic information into traditional breeding programs, particularly for disease and growth traits. Recent increased research effort has yielded a number of genome-wide SNP and genome map resources for *L. vannamei* (Ciobanu *et al.* 2010, Du *et al.* 2010, Yu *et al.* 2015). However, none have yet to be made commercially available. Herein, we present a large transcriptome sequence reference assembly with utility for mining over 26,662 high quality SNP markers and a commercially available Illumina Infinium ShrimpLD-24 v1.0 genotyping array with 8,967 SNPs for *L. vannamei.*

MATERIALS AND METHODS

Sequencing, assembly and annotation

To enable the identification and development of genome-wide Type I SNPs, high-quality total RNA was extracted from the pleopod tissue of 30 *L. vannamei* individuals (provided by Global Gen, Indonesia) using TRIZOL® Reagent (Life Technologies). Equimolar pooled RNA was converted to cDNA using the Mint cDNA synthesis kit (Evrogen) and sequenced using an Illumina GA-IIX at 76 bp paired-end resulting in approximately 25 gigabases of paired-end EST sequence data (~10x genome coverage). Sequences were screened using the software Seqclean (https://sourceforge.net/projects/seqclean/) and MOTHUR (Schloss *et al.* 2009). The cleaned sequence data was assembled using Velvet V1.0 (Zerbino *et al.* 2008) and OASES (Schulz *et al.* 2012). Transcript assemblies were conducted at kmer lengths of k39, k41, k43, k45, k47, k49, k51 and k53 before being clustered together at a 90% sequence identify threshold using the software CD-HIT (Li *et al.* 2006). Assembly of the cleaned-up sequence data produced 76,963 contigs (N50 = 2,375 bp and average contig length = 1,429 bp).

SNP Discovery and Filtering

Genome-wide SNPs were identified within SAMTOOLs (Li *et al.* 2009). The varFilter option in SAMTOOLs was employed to filter SNPs, keeping only the most informative (i.e. minor allele frequency (MAF) >0.25, read depth >10 reads, minor allele reads >2, SNP mapping quality >25, flanking sequence quality >25). Any SNP identified within 50 bp of a candidate SNP was excluded to ensure a conservative flanking region for probe design. SNPs with the highest MAF and read depth were submitted for assay development analysis using Illumina's Assay Design Tool (ADT) and included if their ADT score was greater than 0.7. To ensure no unintentional duplicate SNPs were included on the array, probes for each SNP were mapped to the initial assembly using NOVOCRAFT (Novocraft Technologies) and only the probes that mapped uniquely were included.

Infinium Array Genotyping

To validate the performance of the Illumina ShrimpLD-24 v1.0 genotyping array, 1,134 female and 193 male parents of families (produced by Global Gen, Indonesia) were genotyped. To ensure all genotype calls were genuine and to identify aberrant SNP and DNA samples, strict data integrity was undertaken in GenomeStudio V2011.1 following methods outlined in Jones *et al.* (2013). Genotype reproducibility between batches was tested using 52 replicate samples and 26 replicate SNPs. SNPs with a MAF greater than 0.01 were considered polymorphic. SNPs were investigated for conformation to Hardy-Weinberg Equilibrium (HWE) and Mendelian Inheritance (MI) patterns.

To demonstrate the utility of the SNPs included on the Infinium ShrimpLD-24 v1.0 array, we generated a preliminary linkage map using 30 grand-maternal and 19 grand-paternal families containing 15 progeny on average. The linkage map was constructed in Carthagene V1.3 (de Givry *et al.* 2005) using an iterative *buildfw*, *annealing*, *flips 6* and *polish* method until the best map were produced. Finally, genomic relationship matrixes (GRMs) were calculated with subsets of SNPs and the full array to determine the minimum number of SNPs required for genomic selection (GS).

RESULTS AND DISCUSSION

Sequencing and assembly of transcripts

In total, over 25 Gb of sequence data (329 million raw EST sequences, 76 bp paired-end, ~15x genome coverage) was produced from an Illumina GA-IIx run. After sequence trimming, 19.7 Gb of high-quality data was retained. Assembly of remaining sequence data produced 76,963 contigs (N50 = 2,375 bp and the average contig length = 1,429 bp). The average read depth over all contigs was 210 reads with a median of 29. The assembled contig sequences and mapped raw reads have been submitted to GenBank (Accession number: SRP094129). This significant genomic resource enables the mining of over 17,000 additional SNPs not included within any commercial SNP array.

SNP discovery and filtering

From the assembled sequence dataset, 234,452 putative SNPs were identified *in-silico* before strict filtering parameters were applied. By filtering out all SNPs with a read depth less than 10 reads and a MAF of less than 0.25, a total of 26,662 high-quality SNPs were identified. A total of 1,142 SNPs did not return ADT values > 0.7 and 1,006 SNPs did not map to unique contigs and were removed. A further 7,003 SNPs were excluded due to being located within the flanking region of another SNP resulting in a final list of 9,447 high-value SNPs. Of these, the highest scoring 8,967 SNPs [8,616 novel; and 351 developed in Ciobanu *et al.* (2010) and mapped in Du *et al.* (2010)] were incorporated into the Illumina ShrimpLD-24 v1.0 array enabling high throughput, cost effective and accurate genotyping. The average MAF and ADT score of these high-value SNPs was 0.37 and 0.95 respectively. SNPs included on the custom array have been submitted to dbSNP on NCBI [ss2137297825-ss2137306471 (the current study); rs159816077-rs159831399 (Du *et al.* 2010); and rs142459135-rs142459627 (Ciobanu *et al.* 2010)]. The ShrimpLD-24 v1.0 array is available at https://www.illumina.com/products/by-type/microarray-kits/infinium-shrimp-ld.html.

Infinium array genotyping and validation

In total, 1,327 individuals were genotyped on the ShrimpLD-24 v1.0 array. From these samples, 70 (5.3%) individuals produced call rates of less than 90% and were removed from further analysis leaving 1,257 unique individuals to investigate SNP array performance. Analysis of the resulting genotypic data revealed that 6.0% of the SNPs did not amplify successfully (probe did not bind to the DNA) and 13.0% of the SNPs returned ambiguous clusters. From the resulting 7,259 SNPs, the SNP conversion and validation rates were 80.9% and 95.6% respectively (Table 1). Further filtering (i.e. excluding SNPs with a MAF < 0.01, SNP duplication, low call rates, or deviations from HWE or MI expectations) resulted in a final dataset of 6,379 high quality SNPs with an extremely high call rate (98.9%). The average minor allele frequency of these high-value SNPs was 0.37.

Table 1: SNP array performance indicating the number of SNPs retained throughout filtering.

SNP Exclusion Category	**# SNPs excluded**	**#SNPs remaining**
Total Number of SNPs:		8,967
Probe Didn't Bind	539	
Ambiguous Clusters	1169	
Number of SNPs producing genotypes (conversion rate):		7,259 (80.95%)
Monomorphic	318	
Number Validated SNPs (validation rate):		6,941 (95.62%)
HWE deviations (Heterozygous Excess / Deficit)	163	
Mendelian Inheritance Errors	399	
Number of SNPs with minimal errors:		6,379 (87.88%)
Mendelian Inheritance Errors (< 0.01), or MAF < 0.01	90	
Duplicated SNPs	43	
Call rate < 90%, or Only 2 Clusters	190	
Number of SNPs with no errors:		6,056 (83.43%)

A total of 52 replicate samples were included to evaluate array performance with concordance between replicate samples exceeding 99.9%. This provided strong support for highly reliable genotypic data across all validated SNPs. Furthermore, we reliably constructed a moderate density linkage map of 44 linkage groups containing 4,370 SNPs. These SNPs span 98.12% of the estimated genome size of 4619.3 cM at an average interval of 0.97 (map data to be revised and presented in subsequent publication). The number of markers placed within each linkage group ranged from 22 – 169 and linkage group distances ranged from 24.9 – 159.5 cM. By assigning positional information to these SNPs, not only we demonstrate their utility, but improve their value within ongoing studies.

In the current breeding program, 3,000 highly informative SNPs provided adequate power for accurate GRM calculations when compared to the 6,379 high quality filtered SNPs [Figure 1;

correlation value of $r^2 = 0.99$; see Khatkar et al. 2017 (these proceedings) for GS analysis]. The minimum number of SNPs for GRM analysis is also supported in similar studies of closed farm populations including Atlantic Salmon (Tsai *et al.* 2015).

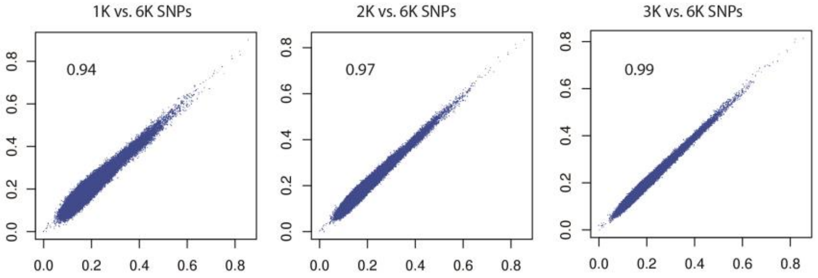

Figure 1: GRM comparisons of different subsets of SNPs.

The development and validation of a large EST-derived SNP resource is pivotal for ongoing research including identifying the major genes underlying important commercial traits, predicting production performance and developing genetic selective breeding programs for *L. vannamei*. If further SNPs are required these can be sourced from the SNP *in-silico* database. High SNP conversion rates are anticipated since the observed conversion rate within this array was > 80%.

REFERENCES

Ciobanu, D. C., J. W. M. Bastiaansen, J. Magrin, J. L. Rocha, D. H. Jiang, N. Yu, B. Geiger, N. Deeb, D. Rocha, H. Gong, B. P. Kinghorn, G. S. Plastow, H. A. M. Van Der Steen and A. J. Mileham (2010) Anim Genet **411**: 39-47.

de Givry, S., M. Bouchez, P. Chabrier, D. Milan and T. Schiex (2005) Bioinformatics **218**: 1703-1704.

Du, Z. Q., D. C. Ciobanu, S. K. Onteru, D. Gorbach, A. J. Mileham, G. Jaramillo and M. F. Rothschild (2010) Anim Genet **413**: 286-294.

Gitterle, T., H. Johansen, C. Erazo, C. Lozano, J. Cock, M. Salazar and M. Rye (2007) Aquaculture **272**, Supplement 1: S262.

Gjedrem, T. (2015) J Mar Sci Eng. **31**: 146.

GLOBEFISH, F. (2016). FAO GLOBEFISH Highlights, 1/2016: 48

Huang, Y. C., Z. X. Yin, S. P. Weng, J. G. He and S. D. Li (2012) Aquaculture 364: 111-117.

Jones, D. B., D. R. Jerry, S. Forêt, D. A. Konovalov and K. R. Zenger (2013) Mar Biotechnol **156**: 647-658.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009) Bioinformatics **2516**: 2078-2079.

Li, W. and A. Godzik (2006) Bioinformatics **2213**: 1658-1659.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks and C. J. Robinson (2009) Appl Environ Microbiol **7523**: 7537.

Schulz, M. H., D. R. Zerbino, M. Vingron and E. Birney (2012) Bioinformatics **288**: 1086-1092.

Tsai, H.-Y., A. Hamilton, A. E. Tinch, D. R. Guy, K. Gharbi, M. J. Stear, O. Matika, S. C. Bishop and R. D. Houston (2015) BMC Genomics **161**: 969.

Yu, Y., X. Zhang, J. Yuan, F. Li, X. Chen, Y. Zhao, L. Huang, H. Zheng and J. Xiang (2015) Sci Rep **5**: 15612.

Zerbino, D. R. and E. Birney (2008) Genome Res. **185**: 821-829.