Marine Biotechnology (2020) 22:109–117
https://doi.org/10.1007/s10126-019-09935-5

ORIGINAL ARTICLE

High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (*Oreochromis niloticus*)

José M. Yáñez[1,2] · Grazyella Yoshida[1,3] · Agustín Barria[1,4] · Ricardo Palma-Véjares[5,6] · Dante Travisany[5,6] · Diego Díaz[5,6] · Giovanna Cáceres[1] · María I. Cádiz[1] · María E. López[1,7] · Jean P. Lhorente[3] · Ana Jedlicki[1] · José Soto[8] · Diego Salas[8] · Alejandro Maass[5,6]

Received: 4 April 2019 / Accepted: 19 November 2019 / Published online: 14 January 2020

Abstract

Nile tilapia (*Oreochromis niloticus*) is the second most important farmed fish in the world and a sustainable source of protein for human consumption. Several genetic improvement programs have been established for this species in the world. Currently, the estimation of genetic merit of breeders is typically based on genealogical and phenotypic information. Genome-wide information can be exploited to efficiently incorporate traits that are difficult to measure into the breeding goal. Thus, single nucleotide polymorphisms (SNPs) are required to investigate phenotype–genotype associations and determine the genomic basis of economically important traits. We performed de novo SNP discovery in three different populations of farmed Nile tilapia. A total of 29.9 million non-redundant SNPs were identified through Illumina (HiSeq 2500) whole-genome resequencing of 326 individual samples. After applying several filtering steps, including removing SNP based on genotype and site quality, presence of Mendelian errors, and non-unique position in the genome, a total of 50,000 high-quality SNPs were selected for the development of a custom Illumina BeadChip SNP panel. These SNPs were highly informative in the three populations analyzed showing between 43,869 (94%) and 46,139 (99%) SNPs in Hardy-Weinberg Equilibrium; 37,843 (76%) and 45,171(90%) SNPs with a minor allele frequency (MAF) higher than 0.05; and 43,450 (87%) and 46,570 (93%) SNPs with a MAF higher than 0.01. The 50K SNP panel developed in the current work will be useful for the dissection of economically relevant traits, enhancing breeding programs through genomic selection, as well as supporting genetic studies in farmed populations of Nile tilapia using dense genome-wide information.

Keywords SNP · *Oreochromis niloticus* · Next-generation sequencing · Illumina · Genomic selection

✉ José M. Yáñez
jmayanez@uchile.cl

[1] Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile, Santiago, Chile

[2] Núcleo Milenio INVASAL, Concepción, Chile

[3] Benchmark Genetics Chile, Puerto Montt, Chile

[4] Present address: The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

[5] Centro para la Regulación del Genoma, Universidad de Chile, Santiago, Chile

[6] Centro de Modelamiento Matemático UMI CNRS 2807, Universidad de Chile, Santiago, Chile

[7] Present address: Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

[8] Grupo Acuacorporacion, Internacional (GACI), Cañas, Costa Rica

Introduction

The study of phenotype–genotype association, the identification of the genomic basis of economically important traits and the implementation of genomic predictions in farmed fish require a considerable number of highly informative single nucleotide polymorphisms (SNPs) that preferably segregate in multiple populations. Thus, the discovery and characterization of dense SNP panels will help to better understand complex trait architecture and genome biology in farmed fish (Yáñez et al. 2015). From an animal breeding perspective, the use of a high number of SNP markers, to support Nile tilapia genetic improvement programs, has the potential to speed up genetic gains for traits which, by their nature, cannot be directly recorded in selection candidates,

e.g., carcass quality and disease resistance traits (Ødegård et al. 2014; Yáñez and Martinez 2010; Yáñez et al. 2014). Dense SNP panels can also allow the determination of genomic regions underlying selection and adaptation to different environmental conditions during the domestication process in farmed fish populations (Gutierrez et al. 2016; López et al. 2019a, b).

The discovery of SNP markers in fish species of commercial interest has recently increased due to the availability of high-quality reference genomes, as it is the case for Atlantic salmon (*Salmo salar*) (Lien et al. 2016), rainbow trout (*Oncorhynchus mykiss*) (Berthelot et al. 2014), and channel catfish (*Ictalurus punctatus*) (Liu et al. 2016). This information has facilitated the development of dense SNP panels, which are being currently available for different fish species including Atlantic salmon (Houston et al. 2014; Yáñez et al. 2016), rainbow trout (Palti et al. 2015), and channel catfish (Liu et al. 2014; Zeng et al. 2017). These genomic resources have been used to carry out several studies aiming at identifying the genetic architecture of economically relevant traits in fish by means of genome-wide association studies for traits such as growth (Gutierrez et al. 2015; Tsai et al. 2015; Yoshida et al. 2017; Li et al. 2018; Reis Neto et al. 2019), disease resistance (Correa et al. 2016, 2017; Rodríguez et al. 2019; Yáñez et al. 2019; Barria et al. 2019), and carcass quality (Gonzalez-Pena et al. 2016). These SNP panels have also been used to test different approaches for the implementation of genomic predictions in Atlantic salmon (Ødegård et al. 2014; Sae-Lim et al. 2017; Bangera et al. 2017; Correa et al. 2017) and rainbow trout (Vallejo et al. 2016, 2017, 2018; Yoshida et al. 2018a, b). This methodology has allowed increasing significantly the accuracy of selection for disease resistance to a wide variety of pathogens, when compared with conventional pedigree-based genetic evaluations (Vallejo et al. 2016, 2017, 2018; Yoshida et al. 2018a, b), and therefore, increasing the selection response. However, one of the most remarkable examples in the use of genomic data, impacting a commercially important trait in aquaculture species, is the case of a major quantitative trait locus (QTL) for resistance to infectious pancreatic necrosis virus (IPNV) in Atlantic salmon (Houston et al. 2008; Moen et al. 2009). The use of marker-assisted selection (MAS) for this QTL have decreased the IPN outbreaks to near zero, in Norwegian Atlantic salmon populations (Norris 2017).

Nile tilapia (*Oreochromis niloticus*) is among the most important freshwater species farmed worldwide. Several selective breeding programs have been established for this species since the 1990s, allowing to genetically improve important commercial traits and expand tilapia farming across the globe (Gjedrem and Rye 2018). To date, the most widespread tilapia strain is the genetically improved farmed tilapia (GIFT) (Webster and Lim 2006), being farmed in Latin America, Asia and Africa. It has been shown that the selection response for growth rate reached up to 15% per generation after six generations of selection (Ponzoni et al. 2011), demonstrating the feasibility to improve this trait by means of artificial selection. However, and despite the large number of genetic programs and the advantages of Nile tilapia farming (e.g., fast growth and high adaptability), there are scarce studies on the application of genomic technologies for mapping variants associated with desired traits and enhancing selection through genomic predictions in comparison with other aquaculture species. Consequently, up to date, genetic improvement programs mainly rely on traditional pedigree-based breeding approaches.

The objective of this study was to perform a large-scale de novo SNP discovery using whole-genome resequencing of hundreds of Nile tilapia individuals from three different farmed populations and develop a 50K SNP panel, to be further used in the determination of the genetic basis of complex traits and genomic selection in this species.

Materials and Methods

Populations

The principal aim of the present study was to discover and characterize a highly informative 50K SNP panel for farmed Nile tilapia populations. Thus, we included animals from three different commercial breeding populations established in Latin America, originated from admixed stocks imported from Asia. We used 59 samples from a breeding population belonging to AquaAmerica, Brazil (POP A), and 126 and 141 samples from two breeding populations belonging to Aquacorporación Internacional, Costa Rica (POP B and POP C) (Neira et al. 2016). The three breeding populations are directly or indirectly related to the GIFT, which is the most spread Nile tilapia strain used for farming purposes worldwide. The GIFT strain was initially established in the Philippines by the crosses between four farmed Asian strains originally from Israel, Singapore, Taiwan and Thailand and four wild strains from Egypt, Senegal, Kenya and Ghana. The POP A breeding population was established based on GIFT animals which were introduced to Brazil from Malaysia for multiplication and farming purposes in 2005. The POP B breeding population is a mixture of the original Asian farmed populations from Israel, Singapore, Taiwan and Thailand present in the Philippines in the late 1980s, which gave origin to the GIFT strain. The POP C breeding population represents a combination of genetic material from the best available stocks corresponding to GIFT (generation 8) and two original African strains founding GIFT. The three populations have been genetically improved for growth rate for more than 8 generations in total, using genetic evaluations based on the best linear unbiased predictor.

Whole-Genome Resequencing

Tissue samples from the 326 fish were obtained by partial fin-clipping of fish anesthetized using benzocaine. Sampling was carried out in accordance with commercial practices and norms held by the two companies, Aquacorporación Internacional and AquaAmerica, which provided the samples. Genomic DNA was extracted from fin-clip samples using the DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. Whole-genome resequencing was performed on each of the individual DNA samples by multiplexing five to six barcoded samples per lane of 150 bp paired-end in an Illumina HiSeq 2500 instrument.

SNP Discovery and Annotation

We used the assembly ASM185804v2 (GenBank accession GCF_001858045.1) of *O. niloticus* as a reference genome for SNP discovery. This assembly consists of 1010 Mb of total sequence comprising 2990 contigs with a contig N50 of 3.09 Mb. Sequences from all samples were evaluated using FastQC to assess base quality and primer adapter contamination. Burrows–Wheeler Aligner (BWA-MEM) (Li and Durbin 2009) was used to map the reads of each sample to the reference genome. Briefly, BWA-MEM starts a local alignment between fragments of the read to the reference genome and extends it until the read is completely mapped. To avoid invalid flags in further analysis, reads without a pair were discarded from the output using SAMtools (Li et al. 2009). In order to obtain a high-quality BAM file, all duplicated reads were masked as such using Picard (http://broadinstitute.github.io/picard). For variant calling, we used the standard protocol implemented in the Genome Analysis Toolkit (GATK), version 3.5.0. All high-quality BAM files for each sample obtained previously were assessed at the SNP calling step and summarized into a single Genotyped Variant Calling Format file (VCF) containing all data. Each SNP was categorized as being either homozygous or heterozygous for the alternative (ALT) allele (i.e., the non-reference (REF) allele). To call a sample homozygous for an ALT allele at a given site, the most common ALT allele variant confidence divided by the total sample reads (QD) was at least 10 (QD > 10). This was done to normalize ALT alleles in zones with high-density depth and poor-quality calls. Only bi-allelic SNPs were preselected in posterior filters. The final VCF file was annotated using Variant Effect Predictor (VEP v92.1) in offline mode using the cached Orenil1.0 genome database and the gff file GCF_001858045.1_ASM185804v2_genomic.gff.

SNP Filtering and Validation

Population genetics analyses and filtering steps, including Hardy–Weinberg equilibrium (HWE), minor allele frequency (MAF), and observed and expected heterozygosities (H_O and H_E, respectively), were carried out using VCFtools (Danecek et al. 2011) and Plink (Purcell et al. 2007). An initial common quality control (QC) for the three populations was performed using VCFtools software. The genotypes were filtered to remove indels and sequence alterations. Markers with genotype quality less than 0.15 (−minGQ < 0.15) and sites with quality less than 40 (−minQ < 40) were also excluded. Further, QC filters for each population were applied separately: we discarded loci with missing genotype rate greater than 0.6, MAF less than 0.01, deviation of HWE (p value < 1e−06), and an Illumina score lower than 0.8 (Table 1). SNPs were also filtered based on Mendelian error using genotypes from 8 trios (sire, dam, and offspring) from POP B, in which markers with less than one Mendelian error were retained. In addition, SNP probes were aligned to the Nile tilapia reference genome to retain markers which have a unique position in the genome assembly (GenBank accession GCF_001858045.1) generated by the University of Maryland and the University of Stirling (Conte et al. 2017) using the following procedure: (i) SNP probes of 121 bp were built using flanking SNP sequences (60 bp upstream and 60 bp downstream of each SNP); (ii) each probe was aligned to the reference genome by means of BLASTN (version 2.3.0+), using the following parameters: word size of 11 (−w) and minimum e value of e−40 (e); (iii) all hits were evaluated, tolerating up to two mismatches, and no gaps were allowed; and (iv) probes having a unique location in the genome were retained. All this procedure was achieved using in-house Python scripts. Furthermore, SNPs with MAF > 0.05 in the three commercial populations were prioritized. Finally, SNPs with an even distribution along the genome were selected. This was done by choosing SNPs from windows of equal size across various chromosomes of the genome using THIN < 9 kb command. When selecting SNPs from windows, higher preference was given to common SNPs between population POP B and POP C. A custom 50K SNP Illumina BeadChip was designed and printed using the final list of SNPs generated. The SNPs included in the chip were tested and validated in 1238 genotyped fish in a GWAS and genomic selection study for growth and fillet traits in one the breeding Nile tilapia populations used in the present study (Yoshida et al. 2019b).

Results

SNP Discovery

Whole-genome resequencing of 326 fish yielded a mean of 76.9 (SD = 65.0) millions of raw reads per fish, with a minimum and maximum of 20.6 and 545.6 millions of raw reads per fish, respectively. Quality-controlled reads were aligned to the Nile tilapia reference genome, and an average of 76.3

Table 1 Summary of the results from SNP discovery and quality control filtering for SNP selection in 326 whole-genome sequenced individuals from three farmed Nile tilapia (*Oreochromis niloticus*) populations

Chr	Initial number of SNP	First common filter[a]	Population-specific filters[b]			Last common filter[c]
			POP A	POP B	POP C	
Mito	1671	1596	0	0	0	0
LG1	779,881	708,381	6768	31,144	18,046	2045
LG2	891,811	797,928	7765	32,143	12,664	1881
LG3a	626,179	539,885	6583	17,175	9991	714
LG3b	3,491,391	2,943,480	22,758	47,028	29,881	1823
LG4	1,099,494	983,024	11,401	38,514	23,132	2150
LG5	792,857	706,930	6497	26,718	11,529	1811
LG6	1,247,549	1,112,177	10,300	47,240	23,576	2560
LG7	1,326,771	1,191,205	16,277	61,818	24,935	3549
LG8	834,050	752,753	10,587	35,798	17,751	1803
LG9	842,634	757,126	7171	27,404	13,135	1518
LG10	767,756	689,181	8389	29,290	13,115	1774
LG11	904,270	813,207	8839	31,506	16,411	2074
LG12	1,178,720	1,051,496	12,065	39,106	18,552	2181
LG13	797,314	722,353	7152	29,682	18,623	1819
LG14	960,008	861,238	11,750	39,640	17,386	2176
LG15	991,645	895,658	10,517	32,442	19,090	2002
LG16	1,319,450	1,181,919	12,268	49,954	25,188	2447
LG17	1,023,476	919,757	9226	40,200	25,573	2502
LG18	1,111,651	988,338	10,630	39,619	18,910	2082
LG19	624,640	563,099	7781	26,643	15,163	1799
LG20	762,408	685,589	7586	27,546	13,409	2013
LG22	1,202,188	1,062,702	13,440	42,665	18,451	2016
LG23	1,229,543	1,099,828	11,769	43,157	25,489	2610
US	5,149,044	4,386,247	24,031	50,640	31,645	2651
Total	*29,956,401*	*26,415,097*	*261,550*	*887,072*	*461,645*	*50,000*

Chr chromosome, *Mito* mitochondria, *US* unplaced scaffolds

[a] First common quality control filtering using all populations together, based on excluding SNPs by genotype quality < 15 and minimum site quality < 40

[b] Population-specific quality control filtering based on removing SNPs with missing genotypes > 0.60, minor allele frequency (MAF) < 0.01, Hardy–Weinberg equilibrium (HWE) *p* value < 1e−06, Illumina score < 0.8, and at least one Mendelian error in eight trios from POP B

[c] Last common quality control filtering using all populations together, based on retaining SNPs with unique position in the genome, prioritizing SNP with MAF > 0.05 in the three commercial populations and evenly distributed across the genome

(SD = 64.6) million reads per fish, with a minimum and maximum of 20.5 and 543.1 million reads per fish, respectively, could be confidently and uniquely mapped to a single position in the genome and used for SNP discovery. Thus, the mean coverage for each fish was 8.7*x* (SD = 8.9*x*), with a minimum and maximum of 2.1*x* and 65.7*x* coverage per fish, respectively. After variant discovery phase, 38,454,404 sequence variants were identified across the set of 326 individuals. A total of 29,956,401 non-redundant SNPs were identified across the set of 326 fish, and 26,415,097 (88.17%) of these SNPs passed the genotype quality (GQ < 0.15) and minimum quality (minQ < 40) filters (Table 1). After discarding 1596 SNPs from mitochondria, specific QC filters were applied for each population separately, removing SNPs based on missing genotypes > 0.60, MAF < 0.01, HWE *p* value < 1e−06, at least one Mendelian error assessed in trios from POP B, and non-unique position of SNP probes in the Nile tilapia reference genome. A total of 261,550, 887,072, and 461,645 SNPs were retained after the filtering steps mentioned above for POP A, POP B, and POP C, respectively. From all these high-quality SNP variants, only 31,694 were common between the three populations and 238,025 SNP variants were common between the two high-priority populations (POP B and POP C). After applying THIN < 9 kb command in order to select SNPs as

evenly distributed along the genome as possible, only 16,275 SNPs were common between the three populations, which were used as the base SNPs for the final list. The gaps, to have a mean of one SNP every 9 kb, were filled with additional 33,769 SNPs common between POP B and POP C to reach a total of 50,044 SNPs. Out of these 50,044 SNPs, 44 SNPs from short unplaced scaffolds were removed. A custom 50K SNP Illumina BeadChip was designed and printed using the final list of SNPs generated here.

SNP Distribution and Annotation

To determine the distribution of SNPs in the Nile tilapia genome we identified their chromosome and position into the public GenBank accession assembly GCF_001858045.1 (Conte et al. 2017). The SNPs covered 1.01 Gb of the total assembly length and averaged one SNP every 9 kb. A total of 47,349 SNPs (94.70%) were located in chromosomes, and 2651 SNPs were located into unplaced scaffolds. After SNP annotation, we found that most of the uniquely anchored SNPs were located in introns (57.81%). Furthermore, a total of 12.2%, 11.97%, 7.16%, and 0.63% were located at downstream, upstream, intergenic, and exon regions, respectively. The remaining SNPs were found in splice acceptor, splice donor, splice site, 3′ UTR, and 5′ UTR regions. The Pearson correlation coefficient (r) between the number of SNPs within each chromosome and the total chromosome size in terms of Mb is 0.95 (p value < 2.24e−11). The relationship between the number of SNPs per chromosome and the total chromosome length in Mb is shown in Fig. 1. Thus, the discovered SNPs present an even distribution across the chromosomes on the Nile tilapia genome assembly (Fig. 2).

SNP Validation and Population Segregation

We also performed comparisons between different populations in terms of population genetic estimates using the 50K SNP validation panel. In this regard, the percentage of SNP segregating in HWE was 99%, 98% and 94% out of the 50K SNP panel for POP A, POP B and POP C, respectively. Furthermore, these SNPs showed 76% and 87%, 89% and 93%, and 90% and 93% of MAF > 0.05 and MAF > 0.01 for POP A, POP B and POP C, respectively (Table 2). The distribution of MAF values across SNPs ranged from 0.04 to 0.50 with a mean MAF value of 0.24 ± 0.12 (Fig. 3). The average observed and expected heterozygosity (H_O and H_E) were evaluated in each population (Table 2). Although the H_O values were very similar among populations, POP A and POP B expressed the lowest (0.20) and the highest (0.25) H_O values, respectively, suggesting that these populations are the least and the most genetically diverse populations in the present study. In the three populations, H_O diverged considerably from H_E, resulting in a heterozygote deficiency compared to HWE expectations.

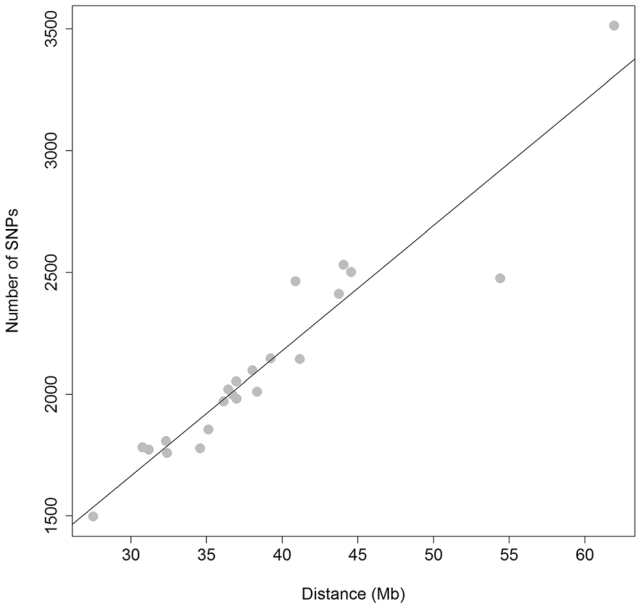

Fig. 1 Relationship between the number of SNPs and chromosome length. Scatter plot of the number of SNPs per chromosome and the total chromosome length in Mb according to the assembly GCF_001858045.1. The correlation coefficient (r) between the number of SNPs and chromosome size is 0.95

Discussion

The application of molecular markers into breeding programs has been widely spread along terrestrial and aquaculture species. Dense SNP panels have been shown to facilitate genome-scale studies by allowing the simultaneous evaluation of thousands of SNPs in commercially important fish species, such as Atlantic salmon (Houston et al. 2014; Yañez et al. 2016) and rainbow trout (Palti et al. 2015). These markers have facilitated the discovery of genetic variants associated with important commercial traits and also the evaluation and implementation of genomic selection in fish species (Correa et al. 2015; Palaiokostas et al. 2016; Bangera et al. 2017; Vallejo et al. 2018). However, and despite Nile tilapia is widely produced in several countries, with the existence of more than 20 breeding programs (Neira 2010), there are still scarce studies aiming at the application of genome-wide SNP information for the identification of quantitative trait loci, and the evaluation and practical implementation of genomic predictions in this species. The SNP discovery strategy used here allowed us to identify and develop a high-quality 50K SNP panel that can be reliably used to genotype different populations of farmed Nile tilapia with a GIFT origin. The GIFT strain is the most spread Nile tilapia strain used for farming purposes worldwide (Ponzoni et al. 2011). The results from the study on the

Fig. 2 The 50K SNP distribution in Nile tilapia genome. The *x*-axis represents the physical distance along each chromosome, split into 1 Mb windows. The different colors correspond to the number of SNPs per window (ranging from 0 to > 80)

segregation of SNPs between different populations indicate that the SNP panel developed in the present study would be useful for genetic studies across populations, although the performance of this set of markers would slightly decrease when used in POP A. This is most likely due to the genetic differentiation between populations, which might be associated with their distinct origin (founder effect) and independent genetic selection by more than ten generations, in some cases. The emphasis placed on including SNPs segregating in POP B and POP C may have caused ascertainment bias, which most likely contributed to the lower diversity observed in the POP A. In addition, there are differences in the number of SNPs with MAFs higher than 0.05 and 0.01 for POP A compared to POP B and POP C. Therefore, these considerations must be taken into account when using the current SNP panel in farmed or wild Nile tilapia populations with different origins.

A recent study has shown the development of a 58K SNP array for Nile tilapia by means of SNP discovery performed using whole-genome resequencing data of 32 fish from one commercial population (Joshi et al. 2018). In this previous study, 40,549 (69.35%) out of 58,466 SNPs were retained after filtering by MAF ≤ 0.05. In our study, between 37,843

Table 2 Descriptive results of population genetic estimates and statistics for three different populations of farmed Nile tilapia using the 50K SNP validation panel

Population	HWE[a]		MAF > 0.05[b]		MAF > 0.01[c]		H_O	H_E
	n	%	*n*	%	*n*	%		
POP A	46,139	99.07	37,843	75.69	43,450	86.9	0.2011	0.2843
POP B	45,757	98.25	44,696	89.39	46,570	93.14	0.2497	0.3130
POP C	43,869	94.20	45,171	90.34	46,570	93.14	0.2463	0.3243

H_O observed heterozygosity, H_E expected heterozygosity

[a] SNPs in Hardy–Weinberg equilibrium

[b] SNPs with minor allele frequency > 0.05

[c] SNPs with minor allele frequency > 0.01

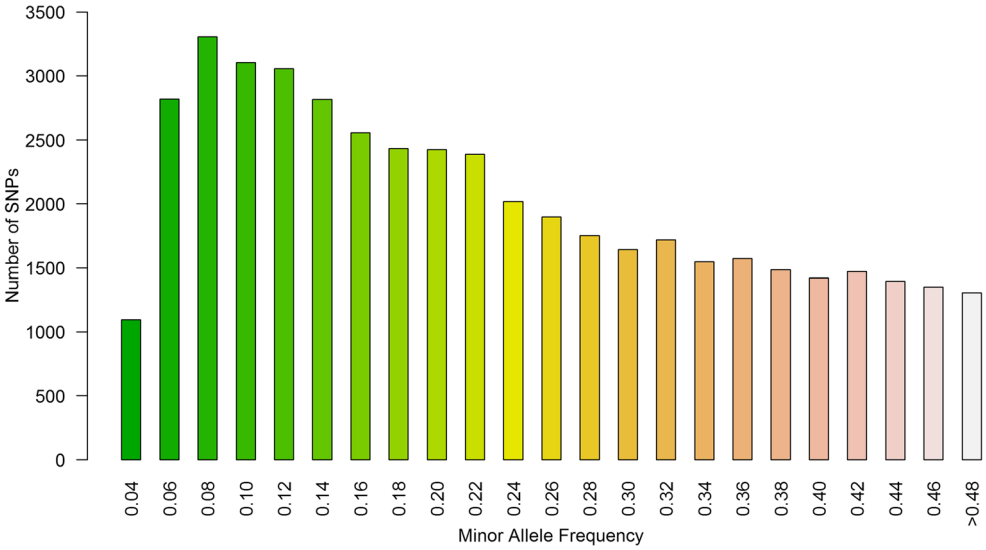

Fig. 3 Distribution of minor allele frequencies (MAFs) for the 50K SNP validation panel from 326 samples

(75.68%) and 45,171 (90.34) out of the 50K SNPs were retained after filtering by HWE and MAF ≤ 0.05, indicating a better proportion of SNP validated and a moderate variation (~ 15%) of availability of SNPs, depending on the target population. The latter is most likely due to ascertainment bias in SNP discovery and selection, and it has to be taken into account in further applications of this SNP panel in populations with different origins. When comparing the 50K SNP panel from our study against the 58K SNP array developed by Joshi et al. (2018), by means of aligning SNP probes to the reference genome in order to identify overlapping markers, we found that 100% of the SNPs were exclusive to each SNP panel. The high proportion of SNPs exclusive to each of the two SNP panels is likely due to variant sampling, and also potentially to the different genetic background of populations and design of the whole-genome resequencing experiments used for SNP discovery. The 50K SNP panel presented here was produced using whole-genome resequencing of 326 fish from three independent populations, which allowed us to have an initial list of 29.9 million putative SNPs, which was almost a three times larger when compared against the previous study from Joshi et al. (2018), in which 32 fish from a single population were whole-genome resequenced, generating 10.5 million putative SNPs for further filtering steps. More importantly, the results presented here indicate that currently available Nile tilapia SNP panels can be considered more as being highly complementary than redundant in terms of the genome variants represented.

The SNP resource presented here provide an excellent tool for the development of genome-scale studies of biologically and economically important traits. For instance, a recent genome-wide association study using a subset of 2.4 million SNPs derived from the 29.9 million SNPs available from the present study confirmed the anti-Müllerian hormone as a major gene associated with sex determination in different populations of farmed Nile tilapia (Cáceres et al. 2019). This information could assist future strategies aiming at generating monosex (all-male) Nile tilapia populations for farming purposes without using hormones, to better exploit the sexual dimorphism present in the species, in which male individuals grow faster than females (Baroiller and D'Cotta 2001). In addition, 50K SNP Illumina BeadChip developed in the present study will also allow the practical implementation of genomic predictions in Nile tilapia selective breeding programs, as it has been reported in a recent study in which an increase in the accuracy of EBVs has been demonstrated through the incorporation of genomic information into genetic evaluations for fillet traits (Yoshida et al. 2019b). Finally, this 50K SNP panel will also allow other kinds of population genetic studies in both farmed and wild populations of Nile tilapia using dense genome-wide information, as for example, has been recently done by the determination of the genetic structure and linkage disequilibrium in farmed populations using dense SNP genotypes (Yoshida et al. 2019a).

Conclusions

This paper describes the simultaneous discovery and validation of a 50K SNP panel in Nile tilapia through the use of whole-genome resequencing of hundreds of animals, for the development of a custom Illumina BeadChip genotyping tool. The 50K SNP panel presented here will provide an opportunity for the dissection of traits of biological and economic importance, such as growth, carcass quality, and disease resistance traits, through the application of genome-scale studies. Furthermore, it will allow increasing the response to selection for these traits by means of genomic selection in breeding programs. We believe that downstream applications of the important genomic resource developed here will help to

enhance Nile tilapia production by making it more efficient and sustainable.

Acknowledgments We would like to acknowledge the Aqua America and Aquacorporación Internacional for kindly providing the samples used in this work, and Gabriel Rizzato and Natalí Kunita from Aqua America and Diego Salas and José Soto from Aquacorporación International for their contribution of the samples from Brazil and Costa Rica, respectively.

Author Contributions J.M.Y. conceived of and designed the study, contributed to the analysis, and drafted the manuscript. G.Y. contributed to the analysis and writing. A.B. drafted the first version of the manuscript. G.C., M.E.L., and A.J. participated in the data collection, purification, and management of the samples for sequencing and genotyping. R.P., D.D., D.T., and A.M. assisted with the bioinformatics analysis and contributed to writing. J.P.L. participated in the design of the study and writing. JS and DS contributed to the collection of the samples and managmenent of populations from Costa Rica. All authors have reviewed and approved the manuscript.

Funding Information This study was partially funded from CORFO grant number 14EIAT-28667 from the Government of Chile. This work was supported by the Basal grant of the Center for Mathematical Modeling AFB170001 (UMI2807 UCHILE-CNRS) and the Center for Genome Regulation Fondap Grant 15090007 Powered@NLHPC. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).

Data Availability The sequence data used for SNP discovery and the 50K SNP chip developed here belongs to Aquainnovo/AquaChile, and it can be available upon reasonable request.

Compliance with Ethical Standards

Conflict of Interest Two commercial organizations (Aquainnovo and Illumina) were involved in the SNP identification and preparation of the manuscript. GMY and JPL were employed by Benchmark Genetics Chile during the course of the study.

References

Bangera R, Correa K, Lhorente JP, Figueroa R, Yáñez JM (2017) Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar). BMC Genomics 18:121

Baroiller JF, D'Cotta H (2001) Environment and sex determination in alligators. Comp Biochem Physiol C Toxicol Pharmacol 130:399–409

Barria A, Marín-Nahuelpi R, Cáceres P et al (2019) Single-step genome-wide association study for resistance to Piscirickettsia salmonis in rainbow trout (Oncorhynchus mykiss). G3 (Bethesda). https://doi.org/10.1534/g3.119.400204

Berthelot C, Brunet F, Chalopin D et al (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5

Cáceres G, López ME, Cádiz MI et al (2019) Fine mapping using whole-genome sequencing confirms anti-Müllerian hormone as a major gene for sex determination in farmed Nile tilapia (Oreochromis niloticus L.). G3 (Bethesda). https://doi.org/10.1534/g3.119.400297

Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD (2017) A high quality assembly of the Nile tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. BMC Genomics 18:341

Correa K, Lhorente JP, López ME et al (2015) Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (Salmo salar L.) chromosomes. BMC Genomics 16:854

Correa K, Lhorente JP, Bassini L et al (2016) Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (Salmo salar L.) using a 50K SNP genotyping array. Aquaculture. https://doi.org/10.1016/j.aquaculture.2016.04.008

Correa K, Bangera R, Figueroa R, Lhorente JP, Yáñez JM (2017) The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus rogercresseyi) resistance in Atlantic salmon (Salmo salar). Genet Sel Evol 49:15

Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Gjedrem T, Rye M (2018) Selection response in fish and shellfish: a review. Rev Aquac 10:168–179

Gonzalez-Pena D, Gao G, Baranski M, Moen T, Cleveland BM, Kenney PB, Vallejo RL, Palti Y, Leeds TD (2016) Genome-wide association study for identifying loci that affect fillet yield, carcass, and body weight traits in rainbow trout (Oncorhynchus mykiss). Front Genet 7:203

Gutierrez AP, Yáñez JM, Fukui S, Swift B, Davidson WS (2015) Genome-wide association study (GWAS) for growth rate and age at sexual maturation in Atlantic salmon (Salmo salar). PLoS One 10: e0119730

Gutierrez AP, Yáñez JM, Davidson WS (2016) Evidence of recent signatures of selection during domestication in an Atlantic salmon population. Mar Genomics 26:41–50

Houston RD, Haley CS, Hamilton A et al (2008) Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). 1115:1109–1115

Houston RD, Taggart JB, Cézard T et al (2014) Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). BMC Genomics 15:90

Joshi R, Árnyasi M, Lien S et al (2018) Development and validation of 58K SNP-array and high-density linkage map in Nile tilapia (O. niloticus). Front Genet 9:472

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li N, Zhou T, Geng X, Jin Y, Wang X, Liu S, Xu X, Gao D, Li Q, Liu Z (2018) Identification of novel genes significantly affecting growth in catfish through GWAS analysis. Mol Gen Genomics 293:587–599

Lien S, Koop BF, Sandve SR et al (2016) The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205

Liu S, Sun L, Li Y, Sun F, Jiang Y, Zhang Y, Zhang J, Feng J, Kaltenboeck L, Kucuktas H, Liu Z (2014) Development of the catfish 250K SNP array for genome-wide association studies. BMC Res Notes 7:135

Liu Z, Liu S, Yao J et al (2016) The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat Commun 7

López ME, Benestan L, Moore J, Perrier C, Gilbey J, di Genova A, Maass A, Diaz D, Lhorente JP, Correa K, Neira R, Bernatchez L, Yáñez JM (2019a) Comparing genomic signatures of domestication in two Atlantic salmon (*Salmo salar* L.) populations with different geographical origins. Evol Appl 12:137–156

Lopez MED, Linderoth T, Norris A et al (2019b) Multiple selection signatures in farmed Atlantic salmon adapted to different environments across Hemispheres. Front Genet 10:901

Moen T, Baranski M, Sonesson AK, Kjøglum S (2009) Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic

necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. BMC Genomics 10:368
Neira R (2010) Breeding in aquaculture species: genetic improvement programs in developing countries. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, vol 8. Leipzig, Germany
Neira R, García X, Lhorente JP et al (2016) Evaluation of the growth and carcass quality of diallel crosses of four strains of Nile tilapia (Oerochromis niloticus). Aquaculture 451:213–222
Norris A (2017) Application of genomics in salmon aquaculture breeding programs by Ashie Norris: who knows where the genomic revolution will lead us? Mar Genomics 36:13–15
Ødegård J, Moen T, Santi N et al (2014) Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front Genet 5:1–8
Palaiokostas C, Ferarreso S, Franch R, et al (2016) Genomic prediction of resistance to pasteurellosis in gilthead sea bream (Sparus aurata) using 2b-RAD sequencing. G3 (Bethesda) X:1–8
Palti Y, Gao G, Liu S et al (2015) The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Mol Ecol Resour 15:662–672
Ponzoni RW, Nguyen NH, Khaw HL et al (2011) Genetic improvement of Nile tilapia (Oreochromis niloticus) with special reference to the work conducted by the WorldFish Center with the GIFT strain. Rev Aquac 3:27–41
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
Reis Neto RV, Yoshida GM, Lhorente JP, Yáñez JM (2019) Genome-wide association analysis for body weight identifies candidate genes related to development and metabolism in rainbow trout (Oncorhynchus mykiss). Mol Genet Genomics 294:563–571
Rodríguez FH, Flores-Mara R, Yoshida GM et al (2019) Genome-wide association analysis for resistance to infectious pancreatic necrosis virus identifies candidate genes involved in viral replication and immune response in rainbow trout (Oncorhynchus mykiss). G3 (Bethesda). https://doi.org/10.1534/g3.119.400463
Sae-Lim P, Kause A, Lillehammer M, Mulder HA (2017) Estimation of breeding values for uniformity of growth in Atlantic salmon (Salmo salar) using pedigree relationships or single-step genomic evaluation. Genet Sel Evol 49:33
Tsai HY, Hamilton A, Tinch AE et al (2015) Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics 16:1–9
Vallejo RL, Leeds TD, Fragomeni BO et al (2016) Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. Front Genet 7:1–13
Vallejo R, Liu S, Gao G et al (2017) Similar genetic architecture with shared and unique quantitative trait loci for bacterial cold water disease resistance in two rainbow trout breeding populations. Front Genet 8:1–15
Vallejo RL, Silva RMO, Evenhuis JP et al (2018) Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: evidence that long-range LD is a major contributing factor. J Anim Breed Genet 135:263–274
Webster C, Lim C (2006) Tilapia: biology, culture, and nutrition
Yáñez JM, Martinez V (2010) Genetic factors involved in resistance to infectious diseases in salmonids and their application in breeding programmes. Arch Med Vet 42:1–13
Yáñez JM, Houston RD, Newman S (2014) Genetics and genomics of disease resistance in salmonid species. Front Genet 5:1–13
Yáñez JM, Newman S, Houston RD (2015) Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front Genet 6:1–3
Yañez JM, Naswa S, Lopez ME et al (2016) Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. Mol Ecol Resour 16:1002–1011
Yáñez JM, Yoshida GM, Parra Á, Correa K, Barría A, Bassini LN, Christensen KA, López ME, Carvalheiro R, Lhorente JP, Pulgar R (2019) Comparative genomic analysis of three salmonid species identifies functional candidate genes involved in resistance to the intracellular bacterium Piscirickettsia salmonis. Front Genet 10:665
Yoshida GM, Lhorente JP, Carvalheiro R, Yáñez JM (2017) Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (Salmo salar L.). Anim Genet 48:698–703
Yoshida G, Bangera R, Carvalheiro R et al (2018a) Genomic prediction accuracy for resistance against Piscirickettsia salmonis in farmed rainbow trout. G3 (Bethesda) 8:719–726
Yoshida G, Carvalheiro R, Rodríguez FH, Lhorente JP (2018b) Genomics single-step genomic evaluation improves accuracy of breeding value predictions for resistance to infectious pancreatic necrosis virus in rainbow trout. Genomics 111:127–132
Yoshida GM, Barria A, Cáceres G et al (2019a) Genome-wide patterns of population structure and linkage disequilibrium in farmed Nile tilapia (Oreochromis niloticus). Front Genet 10:745
Yoshida GM, Lhorente JP, Correa K et al (2019b) Genome-wide association study and cost-efficient genomic predictions for growth and fillet yield in Nile tilapia (Oreochromis niloticus). G3 (Bethesda) 9
Zeng Q, Fu Q, Li Y et al (2017) Development of a 690 K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence. Sci Rep 7:1–14