Development of a 200K SNP Array for Atlantic Salmon: Exploiting Across Continents Genetic Variation

J.M. Yáñez[1,2], S. Naswa[3], M.E. López[2], L. Bassini[1,2], M.E. Cabrejos[2], J. Gilbey[4], L. Bernatchez[5], A. Norris[6], C. Soto[7], J. Eisenhart[8], B. Simpson[8], R. Neira[1,2], J.P. Lhorente[1], P. Schnable[9,10], S. Newman[3], A. Mileham[11], N. Deeb[3]

[1]Aquainnovo, Puerto Montt, Chile, [2]University of Chile, Santiago, Chile, [3]Genus plc, Hendersonville, TN, USA, [4]Marine Scotland Science, UK, [5]IBIS, Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Canada, [6]Marine Harvest, Dublin, Ireland, [7]Camanchaca, Puerto Montt, Chile, [8]GeneSeek, Lincoln, NE, USA, [9]Data2Bio LLC, Ames IA USA; [10]Iowa State University, Ames IA USA; [11]Genus plc, DeForest, WI, USA

ABSTRACT: The dissection of the molecular basis of relevant traits in farmed Atlantic salmon and the implementation of genomic selection schemes require a considerable number of single nucleotide polymorphisms (SNPs). In this study we performed a de novo SNP discovery and developed an Affymetrix Axiom® myDesign Custom Array, taking genome duplication and across continents genetic variation into account. 9,736,473 non-redundant SNPs were identified across a panel of 20 fish by whole genome sequencing. After applying six filtering steps, 200K SNPs were selected and genotyped in 480 fish representing wild and farmed fish from Europe, North America and Chile. Nearly 79.6% (159,099) SNPs had probes belonging to good quality categories according clustering properties. This array provides the platform for the dissection of economically important traits, assisting breeding programs through genomic selection and genetic studies in wild populations using high-resolution genome-wide information.

Keywords: single nucleotide polymorphisms; *Salmo salar;* Tetraploid

Introduction

The dissection of the molecular basis of economically important traits in farmed Atlantic salmon and the implementation of genomic selection schemes require a considerable number of high quality single nucleotide polymorphisms (SNPs) that preferably segregate in multiple populations. High-density SNP arrays have been developed for several domestic animal species, including cattle, sheep, horses, pigs and chickens (Goddard and Hayes (2009); Groenen et al. (2011)). The use of this information to assist Atlantic salmon breeding programs can accelerate the genetic progress for traits which cannot be directly measured in the selection candidates, (e.g. disease resistance and carcass quality traits) (Yáñez and Martínez (2010)). The International Collaboration to Sequence the Atlantic salmon genome (ICSASG) has facilitated the identification of a large number of SNPs (Davidson et al. (2010)) and a 16.5K Illumina iSelect bead-array was developed (Kent et al. (2009)). However, no more than 35% of the putative SNPs included on this array were validated and useable in further genetic studies. Much of the difficulty in SNP validation here is likely due to the tetraploid status of about one-third of the Atlantic salmon genome (Bourret et al. (2013); Gidskehaug et al. (2011)). Recently, a high-density SNP genotyping array has been developed and validated in European Atlantic salmon populations containing around 132K SNPs (Houston et al. (2014)). The utility of this platform has not yet however been validated in populations of North American origin and aquaculture strains of fish in present in Chile. The objective of this study was to perform a de novo SNP discovery, taking the genome duplication and across continents genetic variation into account, to develop a high-density SNP array to be used in the genetic dissection of complex traits and selective breeding of Atlantic salmon.

Materials and Methods

DNA Sequencing. We performed whole genome sequencing (WGS) of 20 individuals from seven different commercial populations of Atlantic salmon: **A**, **B**, **C**, **D**, **E**, **F**, and **G** (2 to 3 fish per population). This step was carried out multiplexing two bar-coded samples per lane of 100bp paired-end Illumina HiSeq2000.

DNA Sequence data analysis. A preliminary assembly available from the ICSASG was used as the reference genome for SNP calling. Low quality bases were trimmed from raw reads. Trimmed reads were aligned to the reference genome using Bowtie2 with default sensitivity parameters as paired-end (PE) fragments allowing a maximum fragment size of 1,000 bp. If a pair of reads could not be aligned as fragments, each read was treated as a single-end (SE) read for alignment. From the Bowtie2 SAM output, confident and uniquely mapped reads were extracted allowing 2 mismatches for every 36 bp of read length and at most 5 bp tails for every 75 bp of read length.

SNP discovery. SNP discovery was conducted by Data2Bio LLC. SNPs were first discovered within each sample (i.e., we conducted 20 independent SNP calling runs) and animals were categorized as homozygous or heterozygous for the ALT allele (i.e., the non-REF allele). The first and last 3 bases of each read were ignored for SNP calling. Only polymorphic sites with PHRED scores ≥ 15 out of 40 ($\leq$ 3% error rate) were considered. To call a sample homozygous for an ALT allele at a given site the most common ALT allele must have been supported by at least 80% of all aligned reads and at least 3 reads must have supported this allele. To call a sample heterozygous for an ALT allele at a given site: i) Each of the two most common alleles must have been supported by at least 30% of aligned reads, ii) At least 3 reads must have supported each allele and iii) The sum of reads for the two most common alleles must have accounted for at least 80% of all aligned reads. Second, a SNP filtering process was conducted by removing: i) tri-allelic sites and variants caused by alignment errors, ii) non-polymorphic SNPs in the sequenced animals (thereby reducing the chance of selecting SNPs that are simply sequence errors in the reference genome), iii) SNPs that have evidence of a nearby

(within 35 bp) polymorphism, iv) A/T and C/G and their reverse complement SNPs because they require double space on the array, v) reads with excessive read counts (median count per fish >15) and vi) SNPs with excessive numbers of heterozygous fish among the 20 samples (observed/expected heterozygous frequency > 1.5), because they have higher probability to be paramorphisms (Emrich et al. (2004)). Furthermore, we retained those SNPs segregating in three top priority populations (**A**, **B** and **C**, representing Chilean adapted, European and North American populations). SNP sequences were also both aligned to the reference genome to select uniformly distributed variants and scored using Affymetrix criteria.

SNP Validation. We designed and generated an Affymetrix Axiom® myDesign Custom Array including 200K SNPs. The SNPs printed on this array were tested and validated in 480 fish from different origins. We used 257 samples representative of three Chilean farmed populations from the Chaicas breeding nucleus: named **Farmed A** (n=93), **Farmed P** (n= 86) and **Farmed Y** (n= 78); 40 samples from a farmed Irish population (named **Farmed F**); 40 samples from a farmed North American population (named **Farmed G**); 50 samples from a farmed Chilean population with Scottish origin (named **Farmed C**); 46 samples from a wild North American population from which one of the farmed strain originated (named **Wild G**); and 47 samples from a wild Scottish population (named **Wild C**). The analysis of genotyped data were carried out using *Axiom Genotyping Console* (AGT, Affymetrix) and *SNPolisher* (an R package developed by Affymetrix).

Results and Discussion

WGS of 20 fish yielded an average of 2 x 107,366,441 reads per fish, representing an average of 2 x 10,844,010,576 bp per fish. Trimmed reads (99.2% raw reads) were aligned to the reference genome. 57-64% of the trimmed reads could be confidently and uniquely mapped to single positions in the genome and these were used for SNP discovery. Approximately, 2-4 million SNPs were identified per fish. In total 9,736,473 non-redundant SNPs were identified across the panel of 20 fish. 2M (20%) of these SNPs were genotyped in all the fish. Over 60% of the SNPs were genotyped in at least 17 fish (see Figure 1). The average minor allele frequency (MAF) for the full set of SNPs was 0.17. After applying each of the six filters described above sequentially, a total of 2,095,989 (21.5%) SNPs remained and 443,241 SNPs presented no missing data across the panel of 20 sequenced fish. After retaining variants segregating in the A, B and C populations and applying Affymetrix scoring, 200K SNPs were selected, printed and genotyped in 480 fish. DishQC (dish quality metric used to QC samples) was determined for all the samples using *AGT*. 413 samples with DishQC >= 0.82 were selected. Genotype call rate was >=97% for each selected sample. *SNPolisher* was used to cluster and classify SNPs according to their quality. Nearly 79.55% (159,099 out of 200K) SNPs had probes belonging to two good quality categories viz. i) Poly-high-resolution (distinct clusters formed by homozygote and heterozygote samples and at least two occurrences of minor allele) and ii) No-minor-allele-homozygous (two distinct clusters with no minor allele homozygous samples). The minor allele frequency of these good quality SNPs varied between ~0.001 and 0.5 with a median of 0.289 (Figure 2).

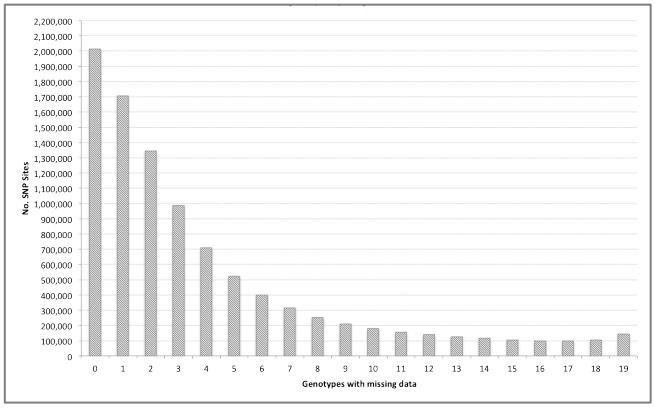

Figure 1: Histogram of number of whole genome sequenced samples (n = 20) with missing genotypes per SNP in a total of 9,736,473 loci.

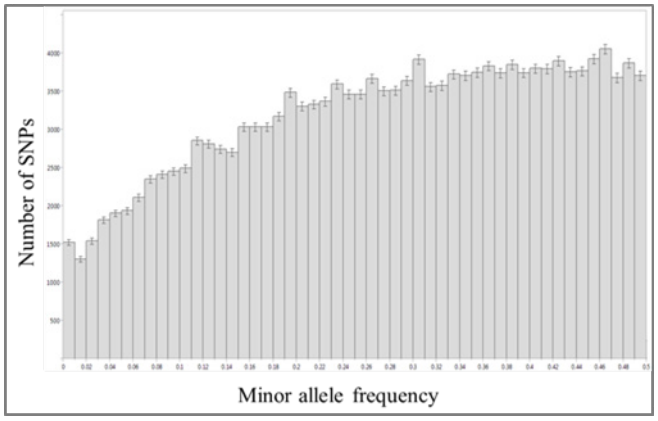

Figure 2: Distribution of minor allele frequencies of all good quality SNPs (159,099) from 413 samples

Conclusion

This paper describes the development of a high-density SNP genotyping array for Atlantic salmon and its validation in Chilean, European and North American populations, including fish from farmed and wild origin. This array provides a platform for the dissection of economically important traits for aquaculture, assisting breeding programs through genomic selection schemes and genetic studies in wild populations using high-resolution genome-wide information.

Literature Cited

Bourret, V., Kent, M.P., Primmer, C.R. et al (2010). Mol. Ecol., 22:532-551.

Davidson, W.S., Koop, B.F., Jones, S.J.M. et al. (2010). Genome Biol., 11:403

Emrich, S. J., Aluru, S., Fu, Y. et al. (2004). Bioinformatics, 20: 140–147.

Gidskehaug, L., Kent, M., Hayes, B.J. et al. (2011). Bioinformatics, 27:303-310

Goddard, M.E., and Hayes, B.J. (2009). Nat. Rev. Genet., 19: 381-391

Groenen, M.A.M., Megens, H., Zare, Y. et al. (2011). BMC Genomics, 12:274

Houston, R.D., Taggart, J.B., Cézard, T. et al. (2014). BMC Genomics, 15:90

Kent, M., Hayes, B.J., Xiang, Q. et al. (2009). Proc. 17th Plant Anim. Genome Conf., San Diego, CA.

Yáñez, J.M., and Martínez, V. (2010). Arch. Med. Vet., 42:1-13