MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1002–1011 doi: 10.1111/1755-0998.12503

Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations

J. M. YÁÑEZ,*,† S. NASWA,‡ M. E. LÓPEZ,†,§ L. BASSINI,†,§ K. CORREA,*,† J. GILBEY,¶ L. BERNATCHEZ,** A. NORRIS,†† R. NEIRA,†,§ J. P. LHORENTE,† P. S. SCHNABLE,‡‡,§§ S. NEWMAN,‡ A. MILEHAM,¶¶ N. DEEB,‡ A. DI GENOVA***,††† and A. MAASS***,†††,‡‡‡
*Faculty of Veterinary and Animal Sciences, University of Chile, Av. Santa Rosa 11735, Santiago, Chile, †Aquainnovo, Talca 60, Puerto Montt, Chile, ‡Genus plc, 100 Bluegrass Commons Blvd. Suite 2200, Hendersonville, TN 37075, USA, §Faculty of Agricultural Sciences, University of Chile, Av. Santa Rosa 11315, Santiago, Chile, ¶Marine Scotland Science, Freshwater Fisheries Laboratory, Faskally, Pitlochry, PH16 5LB, Scotland, UK, **Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC G1V 0A6, Canada, ††Marine Harvest, Kindrum, Fanad, C. Donegal, Ireland, ‡‡Data2Bio LLC, Ames, IA 50011, USA, §§Department of Agronomy, Iowa State University, Ames, IA 50011, USA, ¶¶Genus plc, 1525 River Road, DeForest, WI 53532, USA, ***Fondap Center for Genome Regulation, Av. Blanco Encalada 2085, 3rd floor, Santiago, Chile, †††Mathomics Bioinformatics Laboratory, Center for Mathematical Modeling and Center for Genome Regulation, University of Chile, Av. Blanco Encalada 2120, 7th floor, Santiago, Chile, ‡‡‡Department of Mathematical Engineering, University of Chile, Av. Blanco Encalada 2120, 5th floor, Santiago, Chile*

Abstract

A considerable number of single nucleotide polymorphisms (SNPs) are required to elucidate genotype–phenotype associations and determine the molecular basis of important traits. In this work, we carried out *de novo* SNP discovery accounting for both genome duplication and genetic variation from American and European salmon populations. A total of 9 736 473 nonredundant SNPs were identified across a set of 20 fish by whole-genome sequencing. After applying six bioinformatic filtering steps, 200 K SNPs were selected to develop an Affymetrix Axiom® myDesign Custom Array. This array was used to genotype 480 fish representing wild and farmed salmon from Europe, North America and Chile. A total of 159 099 (79.6%) SNPs were validated as high quality based on clustering properties. A total of 151 509 validated SNPs showed a unique position in the genome. When comparing these SNPs against 238 572 markers currently available in two other Atlantic salmon arrays, only 4.6% of the SNP overlapped with the panel developed in this study. This novel high-density SNP panel will be very useful for the dissection of economically and ecologically relevant traits, enhancing breeding programmes through genomic selection as well as supporting genetic studies in both wild and farmed populations of Atlantic salmon using high-resolution genomewide information.

Keywords: genomic selection, next-generation sequencing, pseudo-tetraploid, *Salmo salar*, SNP array

Received 12 October 2015; revision received 11 January 2016; accepted 16 January 2016

Introduction

The elucidation of genotype–phenotype association, dissection of the molecular basis of traits of ecological and economic importance in both wild and farmed Atlantic salmon and the implementation of genomic-enabled prediction of genetic merit require a large number of high-quality single nucleotide polymorphisms (SNPs) that segregate in multiple populations. Thus, the development and characterization of a dense SNP genotyping array will contribute to a better understanding of genome biology and complex traits in fish and shellfish species (Yáñez *et al.* 2015). From a genetic improvement perspective, the use of a dense SNP panel to assist Atlantic salmon breeding programmes has the potential to accelerate genetic progress for traits, which cannot be directly measured in selection candidates, for example disease resistance and carcass quality traits (Yáñez & Martínez 2010; Fernández *et al.* 2014; Ødegård *et al.* 2014; Yáñez *et al.* 2014a). Dense SNP panels can also be used for the identification of genomic regions under nat-

Correspondence: José M. Yáñez, Fax: + 56 22 978 55 51; E-mail: jmayanez@uchile.cl

ural or artificial selection during adaptation to different environments or domestication and so help in the detection of genetic factors involved in economically or ecologically important traits in fish populations (López *et al.* 2015).

Dense SNP arrays have been developed for several terrestrial domestic animal species, including cattle, pigs and chickens, for example (Matukumalli *et al.* 2009; Ramos *et al.* 2009; Groenen *et al.* 2011). Recently, a 57 K SNP array has been developed and characterized for rainbow trout (*Oncorhynchus mykiss*) (Palti *et al.* 2015). 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 16.5 K Illumina iSelect beadarray was developed (Kent *et al.* 2009). However, no more than 40% of the putative SNPs included on this array could be validated (i.e. SNPs showing detectable polymorphism at a population level) or were useable in further genetic studies (Gidskehaug *et al.* 2011). A key reason for the difficulty in generating a useful SNP genotyping array here was probably the tetraploid status of about one-third of the Atlantic salmon genome (Gidskehaug *et al.* 2011; Bourret *et al.* 2013). A new version of this SNP chip, now just 6 K SNPs, has been developed by the same group and used to generate a linkage map for Atlantic salmon (Lien *et al.* 2011). This platform has also been used by other research groups to perform various types of genetic studies (e.g. population genomics studies, QTL mapping and detection of selection signatures) in wild and farmed populations (Karlsson *et al.* 2011; Gutierrez *et al.* 2012, 2014, 2015a,b; Bourret *et al.* 2013; Ozerov *et al.* 2013; Johnston *et al.* 2014). While representing a valuable tool, the number of markers included in this platform is insufficient to efficiently capture the levels of linkage disequilibrium needed to detect genomic regions affecting quantitative traits (Gutierrez *et al.* 2015b) or to implement genome-enabled predictions (Dominik *et al.* 2010) in breeding populations from different origins.

Recently, a high-density SNP genotyping array has been developed and validated in European Atlantic salmon populations containing about 132 K usable SNPs (Houston *et al.* 2014). The utility of this platform has not been validated in populations of North American origin or aquaculture strains present in Chile, the second largest Atlantic salmon producer country in the world (FAO 2014). The aim of this study was to perform a *de novo* SNP discovery, taking genome duplication and across-continent genetic variation into account and to develop and validate a high-density SNP panel to be used in the genetic dissection of complex traits in Atlantic salmon.

Materials and methods

DNA sequencing

Thirteen fish from Chilean commercial populations with European origin (Scottish and Norwegian origin) and seven fish with North American origin were used for DNA preparation. Samples were obtained by partial fin-clipping of fish anesthetized using benzocaine. After sampling, fish were placed back into the same tank or cage of origin. The sampling procedure to obtain fin clips from farmed fish was approved by The Comité de Bioética Animal, Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile (Certificate N° 29–2014). Genomic DNA was extracted from fin-clip samples using the DNeasy Blood & Tissue Kit (QIAGEN). Whole-genome sequencing (WGS) was performed on each of the 20 individuals (Macrogen, Korea) multiplexing two barcoded samples per lane of 100 bp paired-end in Illumina HiSeq2000 machine.

SNP discovery

DNA sequence analysis, including SNP discovery, was conducted by Data2Bio LLC (Ames, IA, USA). The preliminary assembly ASM2.1 from ICSASG (Davidson *et al.* 2010) was used as the reference genome for SNP calling. This assembly consisted of 3247 megabases of total sequence comprising 864 862 contigs with a contig N50 of 19 339 kb. Low-quality bases were trimmed from raw reads and each read examined in two phases. In the first phase, reads were scanned starting at each end and nucleotides with quality values lower than the threshold removed. The remaining nucleotides were then scanned using overlapping windows of 10 bp, and sequences beyond the last window with average quality value less than the specified threshold were truncated. The trimming parameters were specified using the trimming software, LUCY2 (Li & Chou 2004).

Trimmed reads were aligned to the reference genome using Bowtie2 (Langmead & Salzberg 2012) with default sensitivity parameters for paired-end (PE) fragments allowing a maximum fragment size of 1000 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, high-confidence 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. Reads that passed the filtering criteria were used for subsequent analyses.

SNPs were first identified within each sample (i.e. 20 independent SNP calling runs), and animals were categorized as being either 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 of 40 (≤3% error rate) were considered (Ewing & Green 1998; Ewing *et al.* 1998). 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, we used the following criteria: (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 stringent SNP filtering process was used to remove: (i) tri-allelic sites and variants caused by alignment errors, (ii) SNPs that were not polymorphic in any of the sequenced animals (thereby reducing the chance of selecting SNPs that were simply sequence errors in the reference genome), (iii) SNPs that were close to (within 35 bp) another polymorphism, (iv) A/T and C/G SNPs because they require double space on the array, (v) SNPs with excessive read counts (>15X, given that the median count per SNP per fish was >15), to reduce the chance that the variant was in a repetitive region and (vi) SNPs with excessive numbers of heterozygous genotypes among the 20 fish samples (observed/expected heterozygous frequency >1.5), because they had a higher probability of being paramorphisms (Emrich *et al.* 2004). Furthermore, we favoured SNPs segregating in our priority populations (representing Chilean farmed populations with Norwegian origin). SNP sequences were aligned to the reference genome to select an even distribution of variants across the genome and scored using Affymetrix criteria.

SNP validation

We designed and generated an Affymetrix Axiom® myDesign Custom Array including 200 K SNPs. The SNPs printed on this array were tested and validated in 480 fish from different origins (Table 1).

Given that the main objective of this study was to find markers useful in Chilean farmed Atlantic salmon populations, we included fish from commercial Chilean populations originating from stocks imported from Norway and Scotland, in the SNP validation step. A Chilean farmed population with North American origin (Canada) was also included to assess the number of SNPs useable in a divergent population. Furthermore, wild Scottish and Canadian populations were included to determine SNP allele frequencies in two natural populations.

Table 1 Genotyped samples from different populations of Atlantic salmon. Number of samples genotyped (n_g), condition (farmed or wild), origin and institution which provided them for the validation of the 200 K Affymetrix Axiom® myDesign Custom SNP Array

Name	n_G	Condition	Origin	Institution
Farmed A	93	Chilean farmed	Norway	Salmones Chaicas – AquaChile
Farmed B	86	Chilean farmed	Norway	AquaChile
Farmed C	78	Chilean farmed	Norway	Salmones Chaicas
Farmed D	40	Irish farmed	Norway	Marine Harvest Ireland
Farmed E	40	Chilean farmed	North America	Aquainnovo
Farmed F	50	Chilean farmed	Scotland	Salmones Camanchaca
Wild A	46	Wild	North America	Université Laval
Wild B	47	Wild	Scotland	Marine Scotland Science

We used 257 samples from three Chilean farmed populations: named Farmed A (n = 93), Farmed B (n = 86) and Farmed C (n = 78), all of Norwegian origin; 40 samples from an Irish farmed population with Norwegian origin (named Farmed D); 40 samples from a Chilean farmed population with North American origin (named Farmed E); 50 samples from a Chilean farmed population with Scottish origin (named Farmed F); 46 samples from a wild North American population (named Wild A); and 47 samples from a wild Scottish population (named Wild B).

The Farmed A, B and C populations derived from Mowi–Fanad strain (Farmed D) (Norris *et al.* 1999) and were introduced to Chile for aquaculture purposes during the 1990s (Solar 2009). The Farmed A and B populations belong to two different year-classes from a breeding nucleus established in 1997 in Puerto Montt, Chile, aimed at improving growth-related traits and more recently disease resistance traits (Yáñez *et al.* 2013, 2014b; Correa *et al.* 2015). The Farmed C population belongs to the broodstock from an Atlantic salmon farm located in the XII Region, Chile, which has been improved for growth for about three generations using phenotypic selection. The Farmed D population belongs to an Irish breeding programme established from genetic material derived from the Mowi strain, which was produced in the late 1960s in Norway and imported into Ireland between 1982 and 1986 (Norris

et al. 1999). The Farmed E population derived from a North American strain established in the 1950s from eggs obtained from the Gaspé Bay, Québec, Canada (Withler *et al.* 2005). Ova from this strain were imported to Chile from an aquaculture farm from the state of Washington, USA, between 1996 and 1998. The Farmed F population originated from individuals from the Loch Lochy region of Scotland and is characterized by rapid growth and high grilsing rate (Johnston *et al.* 2000). During the 1980s, ova from this strain were imported to Chile for aquaculture purposes. The Wild A population corresponds to individuals sampled from the St Jean River, which is the main affluent draining into Gaspé Bay, Québec, Canada (Dionne *et al.* 2008). The Wild B population corresponds to individuals sampled from the East coast of Scotland an area that lacks an aquaculture industry, thus minimizing the probability that the genetics of the wild population has been affected by farmed fish escaping from aquaculture facilities.

Samples from all farmed populations were obtained using the same procedure described in DNA Sequencing section. After sampling, recovered fish were placed back into the same tank or cage of origin. The sampling procedure to obtain fin clips from farmed fish was approved by The Comité de Bioética Animal, Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile (Certificate N° 29–2014). Samples representing Wild A population were collected as described by Dionne *et al.* (2008). Samples from Wild B population were collected using electrofishing followed by anaesthesia with tricaine and partial fin-clipping. The sampled fish were transferred to fresh water and allowed to fully recover before release back to the same location as capture. The work undertaken has been reviewed both by the Marine Scotland Ethics Review committee and by the United Kingdom Home Office (Project Licence 60/4251). All samples were collected with the permission of the relevant Fishery Trust or landowner.

Genotyping was performed by GeneSeek (Lincoln, NE, USA) following standard protocol for Axiom Affymetrix platform. Quality control of genotype data was carried out using *Axiom Genotyping Console* (AGT, Affymetrix) and *SNPolisher* (an R library developed by Affymetrix) following the *Best Practices* procedures recommended by Affymetrix (http://media.affymetrix.com/support/downloads/manuals/axiom_best_practice_supplement_user_guide. pdf). Population genetics analyses, including Hardy–Weinberg Equilibrium (HWE), minor allele frequencies (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).

Anchoring SNPs to reference Atlantic salmon genome

We located the SNPs in the genome assembly of Atlantic salmon (GenBank Accession no. GCA_000233375.4) produced by the ICSASG consortium (Davidson *et al.* 2010) using the following strategy: First, SNP probes of 71 base pairs of length were built using flanking SNPs sequence (35-bp upstream and 35-bp downstream of each SNP). Second, each probe was aligned to the reference genome with MEGABLAST (Altschup & Gish 1990) version 2.2.25 using as parameters a word size of 11 (-w), non-repeat-filter of query (-F F) and a minimum score of 70 (-s 70). Third, probes having a unique genomic location were used to assign SNPs coordinates to the ICSASG assembly. This step was achieved using local PERL scripts.

Results

SNP discovery and validation

WGS of 20 fish yielded an average of 214 732 882 reads per fish, representing an average of 21 688 021 152 base pairs per fish. Trimmed reads (99.2% raw reads) were aligned to the reference genome, and 57–64% of the trimmed reads per fish could be confidently and uniquely mapped to single positions in the genome and these were used for SNP discovery. Thus, approximately 2–4 million SNPs were identified per fish. In total 9 736 473 nonredundant SNPs were identified across the panel of 20 fish, and 2M (20%) of these SNPs was genotyped in all the fish. Over 60% of the SNPs were genotyped in at least 17 fish (see Fig. 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 Farmed A, Farmed B and Farmed C populations and applying Affymetrix scoring, 200 K SNPs were selected, printed and genotyped in 480 fish. DishQC (dish quality metric used to QC samples) was determined for all samples using *AGT*, and 413 samples with DishQC $\geq$ 0.82 selected. The genotype call rate was $\geq$ 97% for each selected sample. *SNPolisher* was used to cluster and classify SNPs according to their quality. Nearly 79.55% (159 099 of 200 K) SNPs had probes belonging to two good quality categories namely: (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

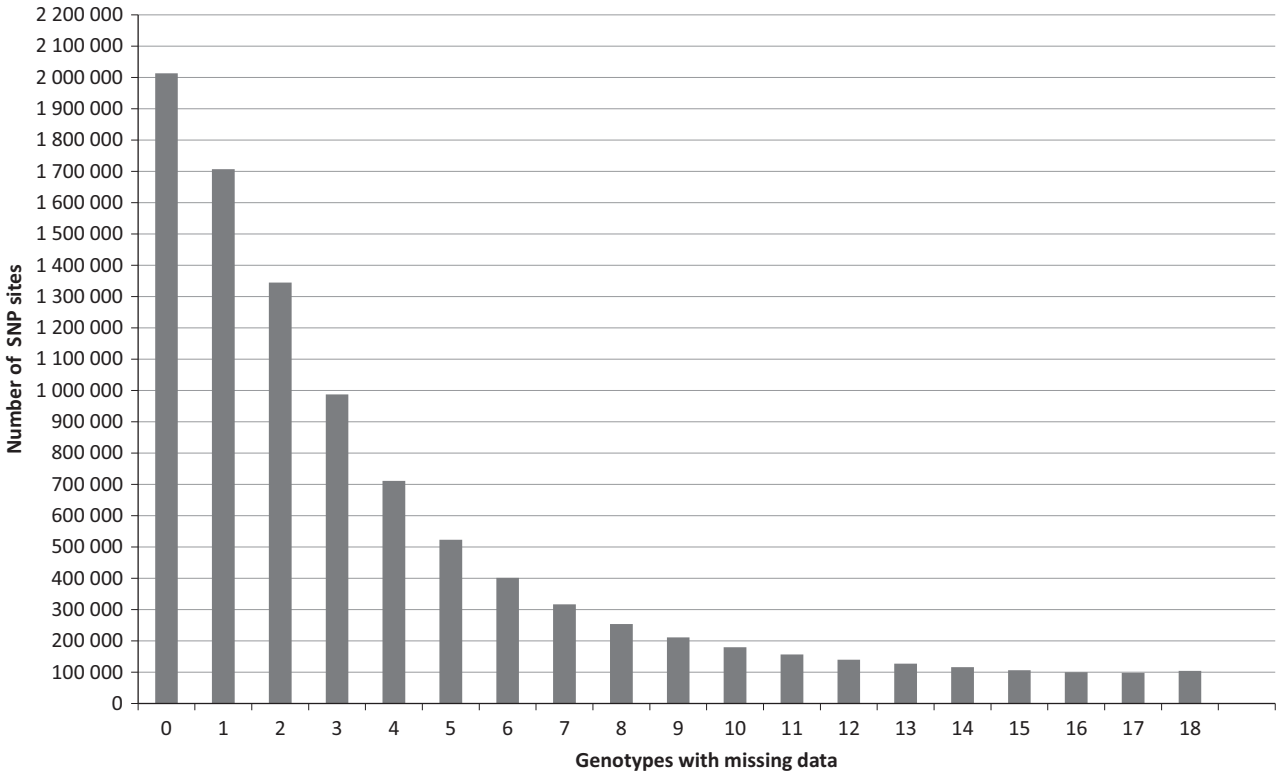

Fig. 1 Missing genotypes in samples sequenced. Histogram of number of whole genome sequenced samples (n = 20) with missing genotypes per SNP at a total of 9 736 473 SNP sites.

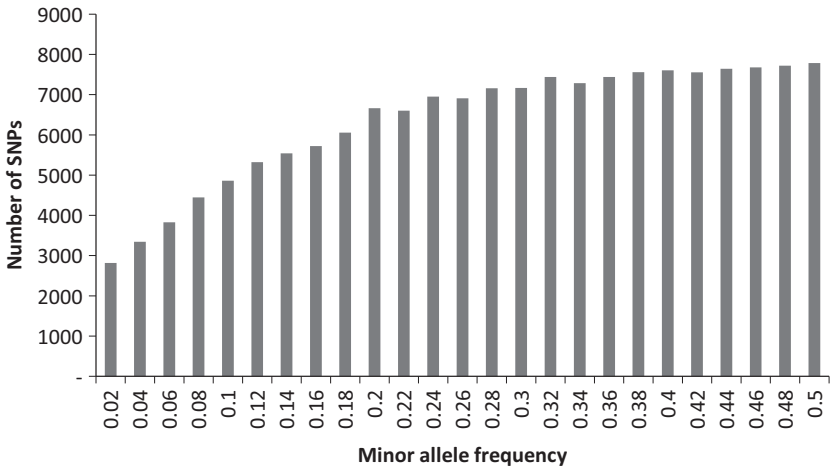

Fig. 2 Distribution of minor allele frequencies (MAFs). Distribution of MAFs for all high-quality SNPs (159 099) from the genotyped samples.

these good quality SNPs varied between ~0.001 and 0.5 with a median of 0.289 (Fig. 2).

Distribution of SNPs in the Atlantic salmon genome

In order to determine the distribution of SNPs in the Atlantic salmon genome, we anchored them to the public GenBank Accession assembly GCA_000233375.4 produced by the ICSASG consortium (368 060 contigs with a N50 57 618 kb and a total of 2.97 Gb of sequenced anchored to the 29 Atlantic salmon chromosomes). A total of 151 509 of the 159 099 (95.2%) validated SNPs had a unique location in this assembly. The SNPs cover 2.1 Gb of the total assembly length and averaged one SNP every 14 kb. We used gene models from the public database SalmonDB (Di Génova *et al.* 2011) to assay the distribution of SNPs over genomic regions. We found that 48.8%, 29.9%, 8.7%, 8.4%, and 2.2% of the uniquely

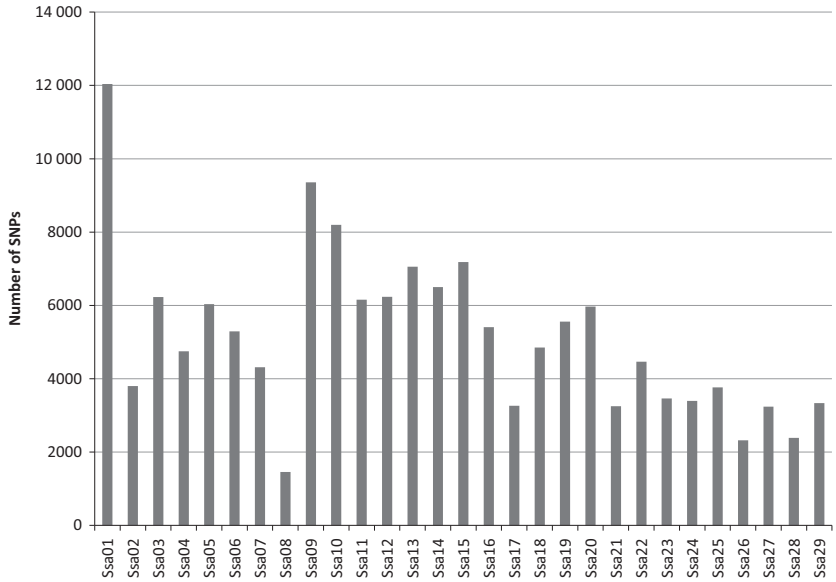

Fig. 3 Distribution of SNPs on the Atlantic salmon chromosomes. Chromosome sequences were produced by assembly version GCA_000233375.4 from the ICSASG consortium.

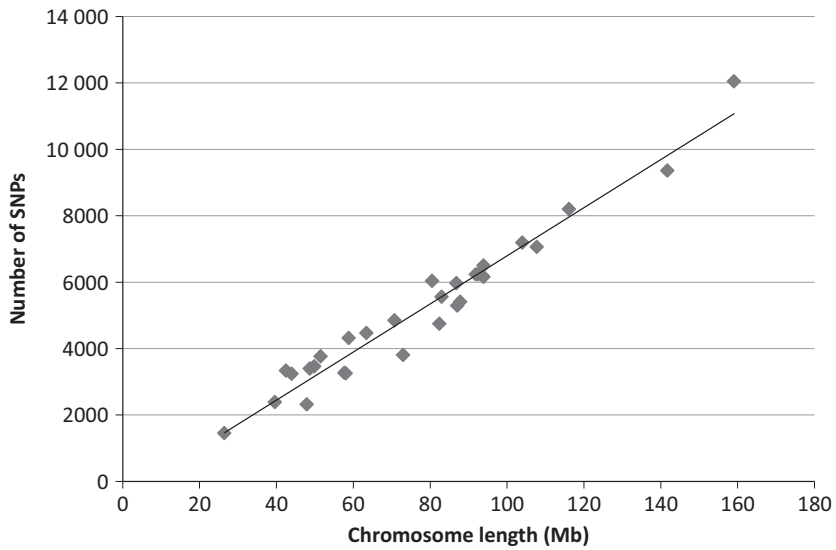

Fig. 4 Relationship between the number of SNPs and chromosome length. Scatter plot of the number of SNPs per chromosome and the total chromosome length in Mb according to the assembly GCA_000233375.4 produced from ICSASG. The correlation coefficient between the number of SNPs and chromosome size is $r = 0.98$.

placed SNPs were located in intergenic, intron, downstream, upstream, and exon regions, respectively. The remaining 2% were located in 3′UTR, 5′ UTR and splice site, splice donor and splice acceptor regions.

We also examined the distribution of SNPs across chromosomes by anchoring scaffolds to chromosomes according to the ICSASG reference genome. We found that 149 207 of the 151 509 SNPs (98.4%), both validated and had a unique position on the genome, could be assigned to chromosomes (Fig. 3).

Only 2302 (1.6%) SNPs had an unknown chromosomal position on the Atlantic salmon reference genome. The Pearson's correlation coefficient between the number of SNPs within each chromosome and total chromosome size in terms of Mb is $r = 0.98$ (P-value <2.2e-16). The relationship between the number of SNPs per chromosome and the total chromosome length in Mb is shown in Fig. 4. Thus, the validated SNP panel presents an even distribution across the chromosomes on the Atlantic salmon genome.

Population segregation of SNPs

We also performed comparisons between different populations in terms of basic statistics and population genetic estimates. In this regard, the percentage of SNPs segregating in HWE in all the populations was 95% of the 159 099 validated SNPs, except for the Chilean farmed population Farmed C in which 77% of SNPs were in HWE. The validated SNPs showed high levels of polymorphisms in farmed populations of Norwegian origin, with ~ 140 K, 141 K, 132 K and 136 K having a MAF

Table 2 Descriptive results of population genetic estimates and statistics for the different populations included in the validation of the 200 K Affymetrix Axiom® myDesign Custom SNP Array for Atlantic salmon

		HWE†		MAF >0.05‡		MAF >0.01§			
Population	n_{QC}*	*n*	%¶	*n*	%¶	*n*	%¶	H_0**	H_E††
Farmed A	91	151 248	95	140 202	88	148 116	93	0.3910	0.3832
Farmed B	85	151 261	95	141 685	89	149 220	94	0.3866	0.3828
Farmed C	74	121 827	77	132 798	83	142 793	90	0.4604	0.3817
Farmed D	40	151 375	95	136 086	86	144 487	91	0.3831	0.3747
Farmed E	37	150 750	95	42 225	27	61 168	38	0.3840	0.3505
Farmed F	43	151 382	95	115 861	73	135 835	85	0.3763	0.3660
Wild A	44	151 319	95	68 217	43	106 041	67	0.2888	0.2852
Wild B	41	151 403	95	132 388	83	142 578	90	0.3829	0.3764

*Number of samples which passed the QC.
†SNPs in Hardy–Weinberg Equilibrium.
‡SNPs with minor allele frequency >0.05.
§SNPs with minor allele frequency >0.01.
¶Percentage of SNPs out of 159 099 validated.
**Observed heterozygosity.
††Expected heterozygosity.

higher than 0.05 in Farmed A, Farmed B, Farmed C and Farmed D, respectively. Moreover, the number of SNPs having MAF levels higher than 0.01 increased by 5% for populations Farmed A, Farmed B, and Farmed D and 7% for population Farmed C, when compared to the number of SNPs showing MAF higher than 0.05 (See Table 2). Similar levels of variation were found in the wild population with Scottish origin (Wild B) where ~ 132 K SNPs had MAFs above 0.05; however, this value decreased slightly for farmed population with Scottish origin (Farmed F), which ~ 115 K SNPs with MAFs higher than 0.05. In addition, the proportion of SNPs having MAF values higher than 0.01 increased by 7% and 12% for populations Wild B and Farmed F, when compared to the number of SNPs showing MAF higher than 0.05 (See Table 2). For both wild and farmed populations of North American origin, the levels of polymorphisms were considerably lower for the SNPs on the array. Only ~ 42 K and 68 K SNPs had MAFs higher than 0.05 in Farmed E and Wild A populations. However, the number of SNPs having MAF values higher than 0.01 was increased by 11% and 24% for populations Farmed E and Wild A, when compared to the number of SNPs with MAFs higher than 0.05 (Table 2). When considering MAFs higher than 0.05 and 0.01, a total of 27 024 and 49 616 SNPs were shared between all the analysed populations, respectively.

The higher levels of variation detected in European strains can be attributable to the selection of SNPs on the array that were segregating in Chilean farmed populations with Norwegian and Scottish origins. The mean observed and estimated heterozygosity (H_O and H_E) was assessed in each population after removing markers with MAF <0.05. These values are shown in Table 2. Wild A and Farmed C populations expressed the lowest (28.8%) and the highest (46%) H_O values, respectively, suggesting that these populations are the least and the most genetically diverse populations in the present study. However, because emphasis was placed on retaining SNPs segregating in Chilean farmed populations with Norwegian and Scottish origins, it might be expected that populations with European origins would have the highest Ho values. For the Farmed C population, H_O diverged considerably from H_E, resulting in a heterozygote excess which is reflected in the increased proportion of SNPs showing departures from HWE (23%) in this particular population. The heterozygosity excess present in the Farmed C population may be due to recent crossing between different year-classes within this brood stock. All remaining populations had H_O values ranging from 37.6% to 39.1% indicating similar levels of genetic diversity between these populations for these markers. Moreover, the H_O and H_E values were consistent for these populations.

Comparison with currently available SNP panels

We compared the 151 509 validated SNPs with unique positions on the genome with SNPs currently available in two other salmon SNP arrays (Lien *et al.* 2011; Houston *et al.* 2014). These were the 5918 SNP chip developed by the Centre for Integrative Genetics (CIGENE) at the Norwegian University of Life Sciences (Lien *et al.* 2011) and the 281 346 SNP panel developed and published by the Roslin Institute at the University of Edinburgh

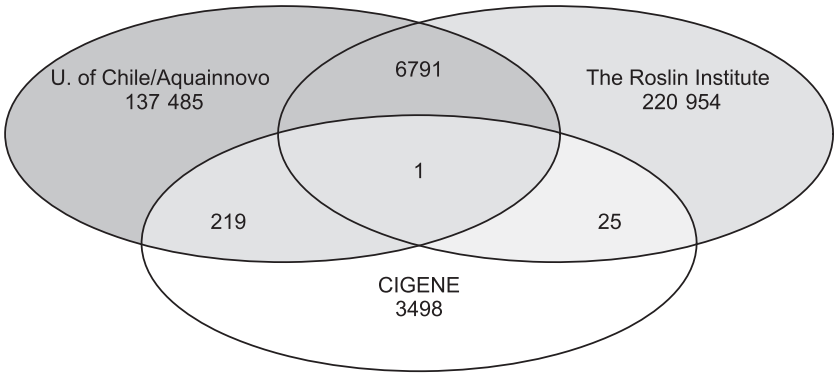

Fig. 5 Comparison in terms of SNP coordinates of available Atlantic salmon arrays. The Venn diagram shows the number of common and unique SNPs among the SNPs validated here with a known position on the genome (151 509 SNPs) developed in this study (University of Chile/Aquainnovo), the CIGENE (Lien *et al.* 2011) and the Roslin Institute (Houston *et al.* 2014) SNP arrays.

(Houston *et al.* 2014). All SNPs were placed on the ICSASG assembly version GCA_000233375.4 using the strategy previously described in the Material and Methods section. We placed 3990 and 234 590 SNPs with unique positions on the reference genome for the CIGENE and the Roslin Institute SNP panels, respectively. Figure 5 shows the distribution of SNPs between the three SNP panels. We observed that the majority of SNPs were exclusive to each SNP panel with 93.8%, 97.0% and 95.3% of SNPs exclusive to the CIGENE, the Roslin Institute and our SNP panel, respectively.

Discussion

The Atlantic salmon 200 K SNP panel characterized in the present study contained a high percentage (79.55%) of SNPs segregating in several populations with different origins (wild and farmed) and different ancestry (American and European). The SNP discovery strategy used here allowed us to efficiently identify a large number of high-quality SNPs which can reliably be genotyped in different populations of Atlantic salmon. Our approach accounted for the complexity of salmon genome due to ancestral tetraploidization events and improved on validation rates obtained in previous studies for this species. For example, less than 40% of SNPS could be validated on the Illumina iSelect SNP chip developed by the CIGENE group (Gidskehaug *et al.* 2011). Similar results for validation rates were obtained by different groups which used the same SNP chip for genetic studies in wild and farmed populations from different origins (Dominik *et al.* 2010; Lien *et al.* 2011; Bourret *et al.* 2013; Gutierrez *et al.* 2014). Sixty-six per cent of SNPs could be validated on a 200 K Affymetrix SNP array for farmed populations with Scottish and Norwegian origin, and wild population from Scotland, Ireland, Norway and Spain (Houston *et al.* 2014). Thus, the strategy used in the present study represents an efficient approach for high-throughput SNP discovery in a pseudo-polyploidy species.

The results from the segregation of SNPs between different populations indicate that the SNP panel developed in the present study would be useful for genetic studies in European populations, although the performance of this set of markers would substantially decrease when is used in North American populations. This probably reflects the pronounced genetic differentiation between populations of Atlantic salmon of European and North American origin associated with their distinct evolutionary history and reduced gene flow between populations from different continents (McConnell *et al.* 1995; King *et al.* 2001; Bourret *et al.* 2013). The emphasis placed on including SNPs segregating in populations of Norwegian and Scottish origins may have caused an ascertainment bias which most likely contributed to the lower diversity observed in the wild population of North American origin as has been previously reported in a recent study (Mäkinen *et al.* 2015). In addition, there is a large difference in the number of SNPs with MAFs higher than 0.05 and 0.01 between Farmed E and Wild A, two populations with North American origins. A similar, but smaller difference was seen when comparing the two populations with Scottish origin (Farmed F and Wild B). The reduced variability in farmed populations when compared to wild populations from the same origin is expected due to sampling only a fraction of the population diversity when setting up aquaculture lines, as well as consequence of selective breeding on a reduced number of individuals. These considerations have to be taken into account when using the SNP panel presented here in wild and farmed populations of Atlantic salmon with North American origins.

The high proportion of markers exclusive to each of the three SNP panels analysed can be explained by the different strategies and populations origins of fish used for SNP discovery. While CIGENE SNP panel mainly used markers derived from sequencing genome complexity reduction (GCR) libraries and expressed sequence tags (ESTs) alignments (Lien *et al.* 2011), the Roslin Institute SNP array was constructed using

sequencing from reduced representation (RR-seq), restriction site-associated DNA (RAD-seq) and mRNA libraries (RNA-se1) (Houston *et al*. 2014). The SNP panel presented in the present study was generated using a WGS approach, which allowed us to have a larger bank of putative SNPs for further selection because it covered the whole genome. Furthermore, the CIGENE SNP chip was generated using markers mainly discovered and validated in fish from a commercial breeding population from Norway (Hayes *et al*. 2007; Lien *et al*. 2011) and the Roslin Institute SNP array included markers discovered and validated from farmed strains and wild fish from different European origins (Scotland, Norway, Ireland and Spain) (Houston *et al*. 2014). The SNP panel presented here was developed using samples from different Chilean commercial populations with different origins: North America, Scotland and Norway and these variants were validated on different wild and farmed populations from the same origins. Thus, our results indicate that currently available *Salmo salar* panels should be considered more as being complementary than redundant in terms of the number of represented SNPs.

Conclusion

This study describes the discovery of a high-density SNP genotyping panel for Atlantic salmon and its validation in Chilean, European and North American populations, including fish from wild and farmed origins. This novel SNP panel provides a platform for the dissection of traits of ecological and economic importance, the use of genomic selection in breeding programmes and genetic studies in wild populations using high-resolution genomewide information. Finally, our results indicate that the SNPs presented here are highly complementary and nonredundant with SNPs panels currently available for *Salmo salar*.

Acknowledgements

This study was partially funded through financial support from Genus, plc, and by grants from CORFO (11IEI-12843 and 12PIE17669), Government of Chile; Programa U-Inicia, Vicerrectoría de Investigación y Desarrollo, Universidad de Chile; Doctoral scholarships from CONICYT, Government of Chile; and CAPES, Government of Brazil. JM Yáñez would like to thank to María Eugenia Cabrejos for her valuable contribution during the early stage of the project and Carlos Soto for his kind contribution with samples from the Farmed F population.

Conflict of interests

Three commercial organisations (Aquainnovo, Genus plc and Data2Bio) were involved in the development of the SNP panel and preparation of the manuscript. L.Ba., J.P.L. and K.C. were employed in Aquainnovo, and S.Na., S.Ne., A.Mi. and N.D. in Genus plc, when this research was performed. This does not alter public accessibility to data from the SNP panel presented in this study.

References

Altschup SF, Gish W (1990) Basic Local Alignment Search Tool, 403–410.

Bourret V, Kent MP, Primmer CR *et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **22**, 532–551.

Correa K, Lhorente J, López M *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.

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

Davidson WS, Koop BF, Jones SJM *et al.* (2010) Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology*, **11**, 403.

Di Génova A, Aravena A, Zapata L *et al.* (2011) SalmonDB: a bioinformatics resource for *Salmo salar* and *Oncorhynchus mykiss*. *Database*, **2011**, bar050.

Dionne M, Caron F, Dodson JJ, Bernatchez L (2008) Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology*, **17**, 2382–2396.

Dominik S, Henshall JM, Kube PD *et al.* (2010) Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (*Salmo salar*) breeding population. *Aquaculture*, **308**, S56–S61.

Emrich SJ, Aluru S, Fu Y *et al.* (2004) A strategy for assembling the maize (*Zea mays* L.) genome. *Bioinformatics*, **20**, 140–147.

Ewing B, Green P (1998) Base-Calling of Automated Sequencer Traces, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Fernández J, Toro MÁ, Sonesson AK, Villanueva B (2014) Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. *Frontiers in Genetics*, **5**, 414.

Food and Agriculture Organization of the United Nations F (2014) *The State of World Fisheries and Aquaculture 2014*. Food and Agriculture Organization of the United Nations, Rome, Italy.

Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics (Oxford, UK)*, **27**, 303–310.

Groenen MAM, Megens H-J, Zare Y *et al.* (2011) The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*, **12**, 274.

Gutierrez AP, Lubieniecki KP, Davidson E a *et al.* (2012) Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (*Salmo salar*) using a 6.5K SNP array. *Aquaculture*, **358-359**, 61–70.

Gutierrez AP, Lubieniecki KP, Fukui S *et al.* (2014) Detection of quantitative trait loci (QTL) related to grilsing and late sexual maturation in Atlantic salmon (*Salmo salar*). *Marine Biotechnology (New York, NY)*, **16**, 103–110.

Gutierrez AP, Yáñez JM, Davidson WS (2015a) Evidence of recent signatures of selection during domestication in an Atlantic salmon population. *Marine Genomics*, doi:10.1016/j.margen.2015.12.007.

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

Hayes B, Laerdahl J, Lien S *et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

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.

Johnston IA, Alderson R, Sandham C *et al.* (2000) Muscle fibre density in relation to the colour and texture of smoked Atlantic salmon (*Salmo salar* L.). *Aquaculture*, **189**, 335–349.

Johnston SE, Orell P, Pritchard VL *et al.* (2014) Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **23**, 3452–3468.

Karlsson S, Moen T, Lien S, Glover KA, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, **11**(Suppl 1), 247–253.

Kent MP, Hayes B, Xiang Q, Berg PR, Gibbs RA, Lien S (2009) Development of 16.5 K SNP chip for Atlantic Salmon. XVII Plant and Animal Genome Conference. San Diego, CA, USA.

King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA (2001) Population structure of Atlantic salmon (Salmo salar L.): a range-wide perspective from microsatellite DNA variation. *Molecular Ecology*, **10**, 807–821.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li S, Chou H-H (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics (Oxford, UK)*, **20**, 2865–2866.

Lien S, Gidskehaug L, Moen T *et al.* (2011) A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics*, **12**, 615.

López ME, Neira R, Yáñez JM (2015) Applications in the search for genomic selection signatures in fish. *Frontiers in Genetics*, **5**, 1–12.

Mäkinen H, Vasemägi A, McGinnity P, Cross TF, Primmer CR (2015) Population genomic analyses of early-phase Atlantic Salmon (*Salmo salar*) domestication/captive breeding. *Evolutionary Applications*, **8**, 93–107.

Matukumalli LK, Lawley CT, Schnabel RD *et al.* (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, **4**, e5350.

McConnell SK, O'Reilly P, Hamilton L, Wright JM, Bentzen P (1995) Polymorphic microsatellite loci from Atlantic salmon (*Salmo salar*): genetic differentiation of North American and European populations. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 1863–1872.

Norris AT, Bradley DG, Cunningham EP (1999) Microsatellite genetic variation between and within farmed and wild Atlantic salmon (*Salmo salar*) populations. *Aquaculture*, **180**, 247–264.

Ødegård J, Moen T, Santi N *et al.* (2014) Genomic prediction in an admixed population of Atlantic salmon (*Salmo salar*). *Frontiers in Genetics*, **5**, 402.

Ozerov M, Vasemägi A, Wennevik V *et al.* (2013) Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (*Salmo salar* L.). *BMC Genomics*, **14**, 12.

Palti Y, Gao G, Liu S *et al.* (2015) The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources*, **15**, 662–672.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Ramos AM, Crooijmans RPMA, Affara NA *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE*, **4**, e6524.

Solar II (2009) Use and exchange of salmonid genetic resources relevant for food and aquaculture. *Reviews in Aquaculture*, **1**, 174–196.

Withler RE, Supernault KJ, Miller KM (2005) Genetic variation within and among domesticated Atlantic salmon broodstocks in British Columbia, Canada. *Animal Genetics*, **36**, 43–50.

Yáñez JM, Martínez V (2010) Factores genéticos que inciden en la resistencia a enfermedades infecciosas Genetic factors involved in resistance to infectious diseases in salmonids and their application in breeding programmes. *Archivos de Medicina Veterinaria (Valdivia)*, **13**, 1–13.

Yáñez JM, Bangera R, Lhorente JP, Oyarzún M, Neira R (2013) Quantitative genetic variation of resistance against *Piscirickettsia salmonis* in Atlantic salmon (*Salmo salar*). *Aquaculture*, **414–415**, 155–159.

Yáñez JM, Houston RD, Newman S (2014a) Genetics and genomics of disease resistance in salmonid species. *Frontiers in Genetics*, **5**, 1–13.

Yáñez JM, Lhorente JP, Bassini LN *et al.* (2014b) Genetic co-variation between resistance against both *Caligus rogercresseyi* and *Piscirickettsia salmonis*, and body weight in Atlantic salmon (*Salmo salar*). *Aquaculture*, **433**, 295–298.

Yáñez JM, Newman S, Houston RD (2015) Genomics in aquaculture to better understand species biology and accelerate genetic progress. *Frontiers in Genetics*, **6**, 128.

J.M.Y. conceived and designed the study, contributed to the analysis and drafted the manuscript. S.Na., M.E.L., K.C. and N.D. participated in the validation step. P.S, A.D.G and A.Ma. performed the bioinformatics analysis and contributed to writing. L.Be. and J.G. collected and provided the samples from the wild populations. M.E.L., A.N. and J.P.L. collected and provided the samples from the farmed populations. L.Ba. and A.Mi. participated in the purification and management of the samples for sequencing and genotyping. R.N., S.Ne., A.Mi. and N.D. provided guidance during the study and discussion. All authors have reviewed and approved the manuscript.

Data accessibility

Sequenced fish genomes were deposited in Short Read Archive (SRA) under BioProject SRP059652. The variants files for each sequenced fish, the genotyping matrix showing the genotypes of the 9 736 473 nonredundant SNPs across the twenty sequenced fish, the 200 K SNP list printed in the array and the 151.5 K validated SNPs, as well as the SNP data used for the comparison with previous SNP chips can be downloaded from SalmonDB database (Di Génova *et al.* 2011) http://salmondb.cmm.uchile.cl/download/Array-Aquainnovo-UChile/.