MASTER INTERUNIVERSITARIO EN MEJORA GENÉTICA ANIMAL Y BIOTECNOLOGÍA DE LA REPRODUCCIÓN

Exploiting genomic data to estimate effective population size from linkage disequilibrium and to identify genomic regions involved in sex determination in Spanish Atlantic salmon

Tesis de Master
Valencia, Julio 2016

Amel Chtioui

Director:
Beatriz Villanueva
Codirector:
María Saura

Acknowledgments

This Master Thesis was carried out in the Departamento de Mejora Genética Animal of INIA and was funded with a grant from CIHEAM.

I would like to express my sincere gratitude to my advisors Drs Beatriz Villanueva and Maria Saura for the continuous support in my Master study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me all the time in the research and writing of this thesis. I could not have imagined having better advisors and mentors for my Master study.

Besides my advisors, I would like to thank Prof Agustín Blasco for organizing well the Master of Animal Breeding and Biotechnology of Reproduction. Thanks to all the professors who collaborated in this Master for their insightful comments and encouragement, but also for their hard questions which incented me to widen my research from various perspectives.

My sincere thanks also goes to all members of the Animal Breeding Department of INIA who provided me an opportunity to join their team and use their research facilities. In particular, I am grateful to Drs Ana Fernández and Almudena Fernández for their help and their advices. I would also like to thank Profs Paloma Morán and Armando Caballero from the University of Vigo for very useful discussions.

I am grateful to all the people I met during these two years. Thanks for all the love and the respect that they have given me.

Kenza, thank you for sharing with me the moment of joy and pain, for your love and support. I am deeply grateful to you for what you did for me.

Mom, you have given me so much, thanks for your faith in me, and for teaching me that I should never surrender.

Daddy, you always told me to "reach for the stars." I think I got my first one. Thanks for inspiring my love for science.

My sister and my brothers, thank you for providing me with unfailing support and continuous encouragement.

Amel

Elle qui croyait par son éducation, par ses études, tout savoir de la vie animale découvre sur ces arpents de terre que la vie naturelle est un bien meilleur professeur parce qu'elle ne donne pas la même réponse à toutes les questions et qu'elle laisse le savoir germer et mûrir comme tout ce qui est vrai et vivant.

J.M.G. Le Clézio
Prix Nobel de littérature 2008

Communications to conferences

The results obtained in this thesis have led to four communications in three different congresses (two international and one national):

Chtioui A, Villanueva B, Morán P, Kent MP, Saura M (2016). Effective population size estimated from genomic data in Spanish Atlantic salmon. 4th International Symposium on Genomics In Aquaculture (GIA2016), Athens, Greece, April 20-22. Oral presentation.

Chtioui A, Villanueva B, Morán P, Kent MP, Saura M (2016). Linkage disequilibrium patterns in Spanish Atlantic salmon obtained from genomic data. 4th International Symposium on Genomics In Aquaculture (GIA2016), Athens, Greece, April 20-22. Poster.

Chtioui A, Villanueva B, Morán P, Kent, MP, Saura M (2016). Estimating linkage disequilibrium and effective population size from genomic data in Spanish Atlantic salmon populations. XVIII Reunión Nacional de Mejora Genética Animal, Valencia, Spain, June 2-3. Poster.

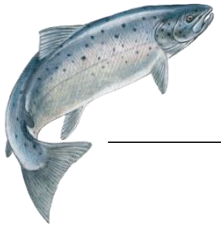
Saura M, **Chtioui A,** Fernández AI, Morán P, Kent MP, Villanueva B (2016). Exploiting genomic data of Spanish Atlantic salmon to identify genes involved in sex determination and to estimate effective population size. 35th International Society for Animal Genetics Conference (ISAG), Salt Lake City, Utah, USA, 23-27 July. Oral presentation.

Contents

Resumen

Los avances en la genómica del salmón atlántico de los últimos años han llevado al desarrollo de chips densos de polimorfismos de un solo nucleótido (SNP), abriendo así nuevas oportunidades para la investigación de la dinámica genética de las poblaciones. En particular, estos chips pueden utilizarse para inferir el tamaño efectivo ancestral y actual de las poblaciones y para detectar genes que afectan a los caracteres del ciclo biológico de la especie. En este estudio se ha utilizado el panel de 220K SNP de Affymetrix (Aquagene/CIGENE) para caracterizar los patrones de desequilibrio de ligamiento, para estimar el tamaño efectivo a partir dichos patrones y para identificar regiones genómicas que controlan la determinación del sexo en poblaciones españolas de salmón atlántico. Estas poblaciones sufren las condiciones más extremas del rango de distribución de la especie en el mundo. Se genotiparon muestras de seis ríos (Miño, Ulla, Eo, Sella, Urumea y Bidasoa) que cubren toda el área de distribución de la especie en España. Después del control de calidad, se disponía de 187 muestras y más de 150.000 SNPs para los análisis. El desequilibrio de ligamiento, medido como el coeficiente de correlación entre pares de SNPs al cuadrado, fue relativamente elevado entre marcadores cercanos (> 0,5 para marcadores separados 0.005 Mb), aunque disminuyó rápidamente al aumentar la distancia entre marcadores (disminuyendo un 90% a 0.3 Mb). El análisis de estructura poblacional reveló que la población española de salmón atlántico actual está compuesta por dos principales grupos ancestrales. Estos incluyen un grupo Atlántico con individuos de los ríos Miño y Ulla (metapoblación atlántica), y un grupo Cantábrico con individuos de los ríos Sella, Urumea y Bidasoa (metapoblación cantábrica). El río Eo se sitúa entre ambos grupos. La metapoblación atlántica mostró mayores niveles de desequilibrio de ligamiento y un tamaño efectivo más bajo que la metapoblación cantábrica. Las estimas de tamaño efectivo estaban comprendidas entre aproximadamente 500 (Miño) y 1200 individuos (Bidasoa) hace 50 generaciones y desde 100 y 400 individuos hace seis generaciones. A través de un estudio de asociación del genoma completo, se encontró que 317 SNPs se asociaron significativamente con la determinación del sexo. Esto se tradujo en nueve regiones QTL putativas en seis cromosomas (Ssa02, Ssa06, Ssa09, Ssa10, Ssa21 y Ssa22). La anotación funcional reveló que las regiones más interesantes relacionados con la diferenciación sexual y el desarrollo sexual se encontraban en los cromosomas Ssa02 y Ssa06 porque pueden contener los siguientes genes: (i) el gen *sdY*, que es el gen determinante del sexo ya identificado en 15 especies de salmónidos (Ssa02, a 36 Mb, muy cerca de una de nuestras regiones 2); (ii) el gen *gata4*, que codifica un factor de transcripción implicado en el desarrollo de las gónadas (Ssa06, a 63

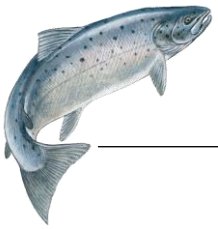
Mb); (iii) el gen *rspo1*, que produce una proteína activadora secretada en el ovario (Ssa06, a 63 Mb); y (iv) el gen *esr1*, un gen receptor de estrógeno, que es esencial para el desarrollo sexual y la función reproductiva de las hembras (Ssa06, a los 66 Mb, cerca de otra de nuestras regiones). Hasta donde sabemos, este es el primer estudio que identifica regiones genómicas asociadas a la determinación del sexo en esta especie e investiga la magnitud del tamaño efectivo de la población a partir de datos genómicos.



Summary

Advances in Atlantic salmon genomics in recent years have led to the development of high-density single nucleotide polymorphisms (SNP) chips opening thus new opportunities for investigating the genetic dynamics of populations. In particular, they can be used to infer ancestral and current population sizes and for detecting genes affecting life history traits. In this study, the 220K high-density Affymetrix SNP genotyping array (Aquagene/CIGENE) has been used to characterise patterns of linkage disequilibrium, to estimate effective population size from these patterns, and to identify genomic regions involved in sex determination in Spanish Atlantic salmon populations. These populations suffer the most extreme conditions of the distribution range of the species in the world. Samples from six rivers (Miño, Ulla, Eo, Sella, Urumea and Bidasoa) covering all the distribution area of the species in Spain were genotyped. After quality control, 187 fish and more than 150,000 SNPs were available for the analyses. Linkage disequilibrium, measured by the squared correlation coefficient between SNP pairs, was found to be relatively high between close markers (> 0.5 for markers 0.005 Mb apart), although it declined rapidly with increasing distance (decreasing by 90% at 0.3 Mb). The population structure analysis revealed that two main ancestral clusters compose the current Spanish population. These clusters include an Atlantic group with individuals from rivers Miño and Ulla (Atlantic metapopulation), and a Cantabric group with individuals from rivers Sella, Urumea and Bidasoa (Cantabric metapopulation). River Eo fell between both groups. The Atlantic metapopulation showed higher levels of disequilibrium and a smaller population size than the Cantabric metapopulation. Across rivers, estimates of effective population size ranged from about 500 (Miño) to 1200 (Bidasoa) individuals 50 generations ago, and from 100 to 400 individuals six generations ago. Through a genome-wide association study a total of 317 SNPs were found to be significantly associated with sex determination. This translated into nine putative QTL regions on six chromosomes (Ssa02, Ssa06, Ssa09, Ssa10, Ssa21 and Ssa22). The functional annotation revealed that the most interesting regions associated with sex differentiation and sexual development were in chromosomes Ssa02 and Ssa06 because they may contain the following genes: (i) the *sdY* gene, which is the master-sex-determining gene already identified in 15 species of salmonids (Ssa02, at 36 Mb, very close to one of our region 2); (ii) the *gata4* gene, which encodes a transcription factor involved in gonads development (Ssa06, at 63 Mb) ; (iii) the *rspo1* gene, which produces a secreted activator protein in ovary (Ssa06, at 63 Mb); and (iv) the *esr1* gene, an estrogen receptor gene which is essential for sexual development and

reproductive function in females (Ssa06, at 66 Mb, close to another of our regions). To our knowledge, this is the first study identifying genomic regions associated to sex determination in this species and investigating the magnitude of effective population size from genomic data.



General Introduction

Advances in salmon genomics in recent years have led to the development of high-density single nucleotide polymorphism (SNP) chips opening thus new opportunities for investigating the evolutionary history of populations. In particular, they can be used to infer ancestral and current population sizes and for detecting genes affecting life history traits. In this thesis, the 220K high-density Affymetrix SNP genotyping array (Aquagene/CIGENE) has been used to characterize linkage disequilibrium patterns, to estimate recent and ancestral effective population size, and to identify candidate genes involved in sex determination in Spanish Atlantic salmon populations.

Atlantic salmon biology

Atlantic salmon (*Salmo salar*) belongs to the family *Salmonidae* that comprises three subfamilies: *Coregoninae* (white fish), *Thymallinae* (graylings) and *Salmoninae* (char, trout and salmons) (Behnke, 2002). The common ancestor of salmonids experienced a genome duplication event between 20 and 120 million years ago (Allendorf and Thorgaard, 1984), and the extant species may be considered pseudotetraploid as they are in the process of reverting to a stable diploid state by losing segments of the genome, by gene silencing or by divergence such that duplicated genes have different patterns of expression (Bailey *et al.*, 1978).

Atlantic salmon is distributed naturally in a large number of waterways (Figure 1) that flow into the European and American coasts of the North Atlantic, from about 41ºN latitude in both continents to the Ungava Bay in Quebec and the Pechora River in northwest Russia (Sedgwick, 1982).

Figure 1. Natural distribution of Atlantic salmon in the world and main migratory paths. (Source: Álvarez *et al.*, 2010).

Atlantic salmon is an anadromous species (Figure 2). Fish begin their lives in freshwater, where the young grow to several cm in length, and then migrate to the sea, where they grow more rapidly and become sexually mature after one or several years. Spawning occurs between November and December, depending on the latitude. The female selects a site which is often at the tail end of a large pool. It is important that the water is flowing steadily through clean loose gravel. This ensures a free flow of oxygen for the fertilised eggs to develop. A female may lay 1,500 eggs or more for each kg of body weight. Pea sized orange eggs are deposited in riverbed gravel in the autumn, and hatch in early spring. The partly transparent alevin remains hidden in the riverbed gravels, feeding from the attached yolk sac. Alevins emerge from the redd about four to six weeks after hatching, when they are about two cm in length. At this stage, they have all eight fins that will be used to maintain their position in the fast flowing streams and manoeuvre about in the water. The fry grow in the river over a period of one year or more, depending mainly on the temperature of the river. They eventually reach a length of five to eight cm before transforming into parrs. Parr remain in the river for one to four or five years, depending on water temperatures and food availability.

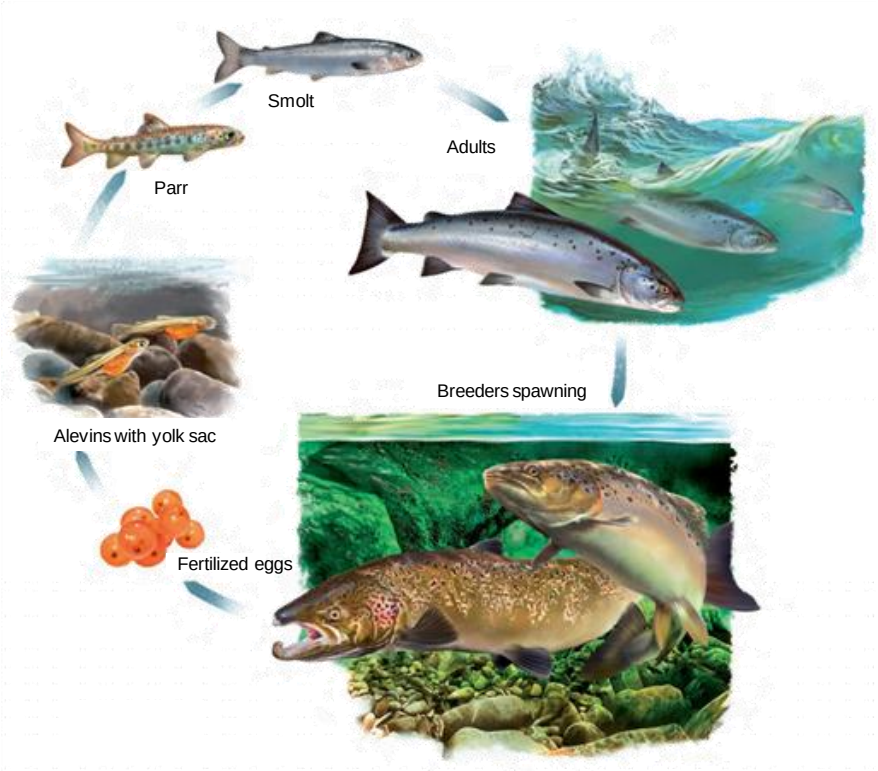


Figure 2. Atlantic salmon life cycle. (Source: Álvarez *et al*., 2010).

Prior to seaward migration, the fish undergo a preparatory smolting process that involves morphological, biochemical, physiological and behavioural changes and that preadapt them for life in high salinity water (Hoar, 1988; Høgåsen, 1998; Thorpe *et al*., 1998; Finstad and Jonsson, 2001). The morphological changes include a slimmer body form and alterations in body colouration (darkened fins, dark back, white belly and silver sides) that help to conceal the fish in the pelagic environment. Fully silver juvenile salmon migrating towards the sea are termed smolts during the freshwater portion of their journey, and post-smolts as soon as they enter the marine environment and until the end of the first winter in the sea (Allan and Ritter, 1977). The average total body length of wild smolts is usually between 10 and 20 cm, and they may weigh from 10 to 80 g (Allan and Ritter, 1977). In Spanish rivers, the river phase lasts between one and two years (Braña *et al*., 1995b).

Adult wild salmon grow rapidly in the sea. They travel great distances to reach feeding grounds in cold northerly waters. Their diet consists of small fish such as sand eels, krill and herring and crustaceans. The marine stage can last from one to several years, also depending on the latitude (Klemetsen *et al*., 2003). Salmon returning to the river after spending one year at sea are called grilses and those returning after two or more years are called multi-sea-winter (MSW) salmon. Spanish salmon spend at sea usually one or two years. Individuals spending three or more years at sea (Braña *et al.*, 1995b) are rarely recorded. Body length of grilses in Spain ranges between 50 and 75 cm and weight ranges between 1.0 and 4.5 kg. For 3SW salmon, body length and weight can reach about 110 cm and 13, respectively. After the stage at sea, salmon return to the river of birth to breed (homing behaviour). The return to the river occurs from March to October. Salmon remain in the river without feeding. Atlantic salmon is an iteroparous species as individuals can have a second reproductive cycle, although this is rare in Spanish rivers (< 5% of the population; Braña *et al*., 1995b).

Situation of Atlantic salmon in Spain

International organizations such as NASCO (North Atlantic Salmon Conservation Organization) agree that most salmon populations have suffered a major decline in recent times and encourage the conservation of the species (Parrish *et al*., 1998; World Wildlife Fund, 2001). The decline in numbers has been more important from the second half of the 20th century. This unfavorable evolution is evident in all countries and in most rivers. Extinction of some populations, declining catches in the rivers, changes in the structure of populations,

loss of usable area colonized by the species and withdrawal of populations towards the lower reaches of the watersheds are some recurring patterns worldwide (Álvarez *et al*., 2010).

Specifically in Spain, the decline suffered by Atlantic salmon in the last century has been dramatic. In some rivers the current accessible area is probably similar to that in the first half of the 20th century but in other rivers there has been a significant loss of such areas mainly due to the hydroelectric development experienced during the period 1955-1975 (Braña *et al*., 1995b; García de Leániz *et al*., 2001). Other factors, such as pollution and overfishing have contributed to exacerbate this decline. Figure 3 shows the reduction in number of catches of Spanish Atlantic salmon by recreational fishery across time.

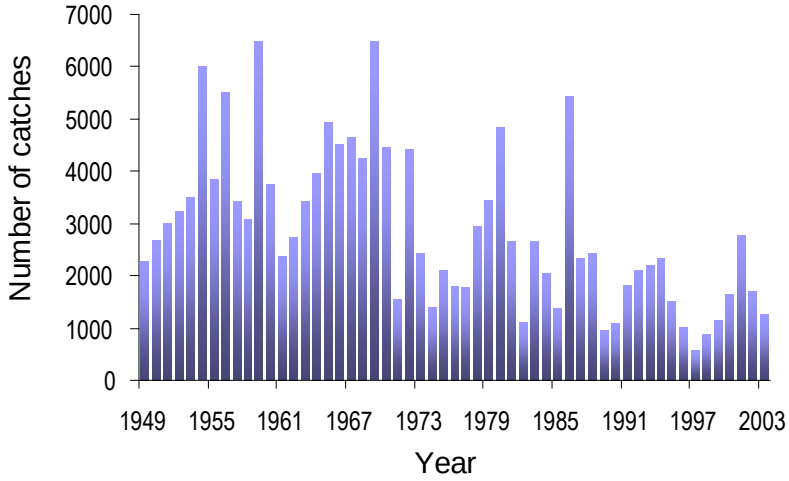


Figure 3. Number of catches of Atlantic salmon by recreational fishery in the main Spanish rivers from 1949 to 2003.

In the Iberian Peninsula, salmon are only found in rivers from the North and northwest entering the North Atlantic and the Cantabric Sea. Management of Atlantic salmon in Spain is carried out by the regional governments. With the aim of re-establishing the endangered populations, the regional governments of Galicia, Asturias, Cantabria, País Vasco and Navarra have implemented management plans since the early 1990s that include different actions. First, river accessibility has been improved by modifying existing, or installing new fish passes, some of which incorporate trapping facilities to monitor returning adult salmon. Second, restrictive regulations have been introduced for recreational fishery through the establishment of quotes and licences. Third, long-term supportive breeding programmes have been established. These programmes involve the capture of wild adults from the rivers, artificially mate them in hatcheries and the release the offspring into the rivers.

Currently, it is recognized that stocking must be carried out only with native individuals, in order to preserve local adaptations and population structures. However, historically this has not been always the case. As an example, between 1970 and 1990 intense stocking was carried out in Spanish rivers with imported salmon from northern Europe (Scotland, Ireland, Norway and Iceland). Although, in general, the impact of these practices was minimal in terms of increasing population sizes (Verspoor and Garcia de Leániz, 1997; Morán *et al*., 2005a; Morán *et al*., 2005b), there was some degree of introgression (altering thus the genetic structure of the native populations), particularly in the rivers further East of the Peninsula (Morán *et al*., 2005a; Campos *et al*., 2007).

Supportive breeding in recent years has been successful in terms of increasing the number of returning adults, that varies from less than 10% in Asturian rivers to 20-50% in rivers from Galicia and the Cantabric Sea (Álvarez *et al*., 2010). However, this increase in numbers might not necessarily entail an increase in effective population size (N_e). Therefore, it is important to perform genetic monitoring to determine the genetic impact of such practices on the natural populations.

Spanish salmon rivers under study

A detailed description of all salmon rivers in Spain can be found in Álvarez *et al.* (2010). Here, we give a brief description of those rivers considered in our study (Figure 4).

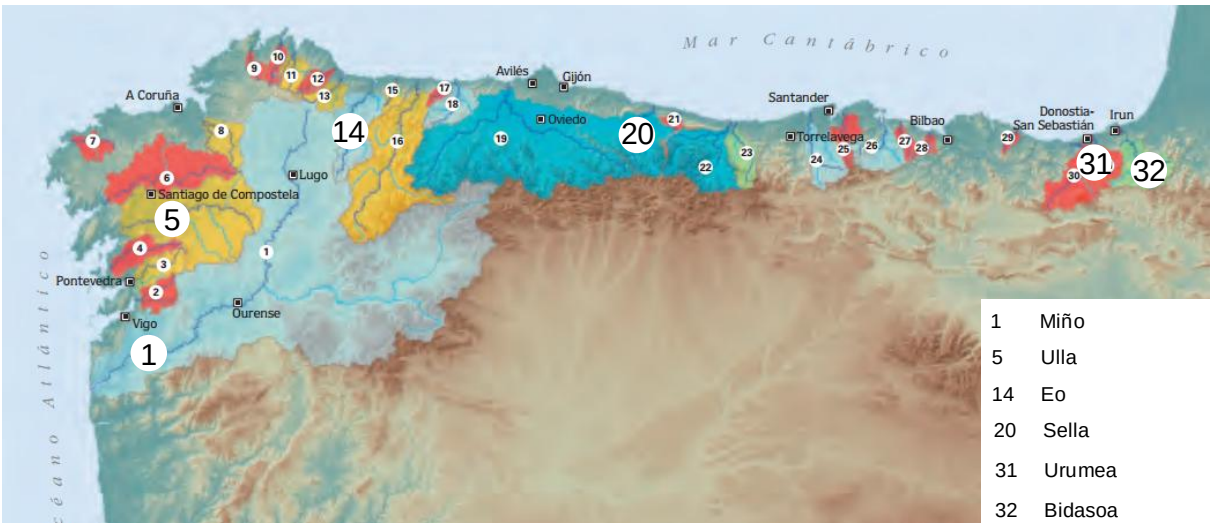


Figure 4. Geographical location of the rivers analysed in this study (1, 5, 14, 20, 31, 32).

Miño

River Miño is the largest Spanish salmon river and also the river having the greatest potential capacity, simply as a consequence of the extension of its hydrographic network. Currently, due to the hydroelectric development carried out in the past, the accessible area for spawning and juvenile development is reduced to about 6% of the potential area. In relative terms, this area is still considerable and the river Miño still preserves an important salmon population.

Ulla

River Ulla is the second salmon river in importance in Galicia after river Eo. Its basin covers 2,803 km^2. The main course is about 131 km long (102 km accessible). Again, the main impact on the river has been the hydroelectric development. The Portodemouros dam constructed in the late 1960s is located about 80 km from the intertidal limit. Although it leaves a considerable accessible stretch that allows the survival of the migratory species, the impact of this dam has been critical and has undermined the very significant salmon potential of the river. It destroyed or prevented the access to the best spawning and breeding areas, and also led to remarkable flow variations to the main course of the river. Pollution has also had an important effect, particularly in the late 1980s.

In recent years, the river has shown a clear recovery as a consequence of the management strategies initiated in the early 1990s that were based mainly on the stocking with native salmon (i.e., from the river Ulla itself). Initial results showed a very high percentage of stocked individuals (about 50% of the population), but this percentage has decreased through the progression of natural reproduction. The information collected at trapping facilities of this river confirms that the return of stocked fish contributes greatly to the increase in the number of salmon. In 2001, 44% of the returning salmon were stocked fish (Caballero, 2002).

Eo

River Eo is located in the Western Cantabric sector. Its length is 86 km and the extension of its basin is about 824 km^2. It represents the link between Asturian and Galician salmon populations. This river went through critical moments in the second half of the 1990s. This occurred in many other rivers, but in the Eo the problem was perhaps exacerbated by the difficulty of smolts for descending through the hydroelectric plant of Vina and Pe. Today its salmon population is showing evident signs of improvement. Catches have improved much

despite the restrictive measures introduced to control fishing pressure. In any case, these catches are still far from those of the recent past.

Sella

The basin of river Sella expands 1,278 km^2 and is the third in extension in the Cantabric side. Its Atlantic salmon population is the best preserved in Spain, although this is not clearly reflected in the ranking of annual catches.

Urumea

The basin of river Urumea falls in the Eastern Cantabric Sea. It is a small basin of 272 km^2. A characteristic of this basin is that water has been used as a driven force from the end of the 19^{th} century (for ironmongery) until the present (for hydropower generation). These infrastructures have been partly responsible of the decline of salmon populations. Factors such as overfishing or obstacles in the river have relegated the fish to areas closer to the estuary. In the early 20^{th} century, when dams were constructed, salmon went back only the final 15 km, losing the best breeding and spawning grounds. Industrial pollution also contributed to decimate the salmon population, so that in 1940 it became extinct. In the 1980s, the Provincial Council of Gipuzkoa undertook a reintroduction plan. Water quality and accessibility were also improved. From 1993 to 2006, 911 adult salmon had been controlled on their return. The species is now breeding in the river and is able to enter over 20 km upstream.

Bidasoa

River Bidasoa is the easternmost river of the Cantabric coast. Its 69 km includes the waters of a basin system of 750 km^2 of which 90% belongs to Navarre. When analyzing the evolution that has suffered the salmon population of this river over the last century it is evident that there has been: (i) a decrease in population size; (ii) a withdrawal of the population towards downstream (specifically, to the final 20 km of a total of 69 km); and (iii) a decrease in the average size of the fish, which is directly related to a higher presence of grilses in the age structure of the population (Álvarez *et al.,* 1995). The establishment of industries was moderate and concentrated in the lower course but hydroelectric development affected the entire river. Only in the final 33 km (area currently used by salmon), there are nine dams, although all of them are equipped with fish passes. The population decline started between the 19^{th} and the 20^{th} centuries and has continued until the present. In the middle of the 20^{th}

century, catches in favourable years ranged from 200 to 300 individuals. At the end of the century and nowadays catches only reach a few dozen of individuals.

The Regional Government of Navarra implemented a conservation program in 1988. Action measures coincide with those of the Galician program. During the first three years of the program, fertilized eggs originated in Iceland were imported. From 1992 imports of foreign eggs have been definitely abandoned and only native eggs are used for supportive breeding (Álvarez *et al.*, 1995). The results of the recovery plan for Atlantic salmon in the river Bidasoa in recent years demonstrated the success of stocking. However, recreational fishing is a problem for the age class structure. It has been suggested that the reduction in sea age and individual size is associated with recreational fishery, as the time in which salmon return to their river of origin is related to sea age (García de Leániz *et al.*, 1987; Braña *et al.*, 1995a). Multi-sea-winter salmon return to the river in early spring, while grilses return at the end of summer (Cross and Ward, 1980). In Spanish rivers, salmon is exploited during a few months of the year, typically from early spring to summer. This means that MSW salmon are captured before reproduction occurs, leading thus to selection for smaller body size and an associated reduction in fitness (Saura *et al*., 2010). Virtually, all MSW salmon are fished (Munárriz and Teniente, 2004). The few salmon that manage to escape are captured for the purpose of artificial spawning in the farm Mugaire Oronoz.

Genomics in salmon

The genome of the Atlantic salmon was selected to be the reference genome for salmonids, and an International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG) was established with researchers and funding agencies from Canada, Chile, and Norway (Davidson *et al*., 2010). Genetic (Danzmann *et al*., 2008; Lien *et al*., 2011) and physical maps (Phillips *et al*., 2009) have been developed. There is an extensive database for salmonid species including ETS sequences and information from Gene Ontology, metabolic pathways, SNP prediction, CDS prediction, orthologs prediction and several precalculated BLAST searches and domains (Di Genova *et al*., 2011). The Atlantic salmon genome is complex. It has essentially the same size as the human genome (~3 Gb) and a 60% repeat content which is among the highest found in vertebrates. It presents a high proportion of homologous blocks, representing 573 Mb with a sequence similarity >90 % (Lien *et al*., 2016).

The availability of a genome reference and next-generation sequencing technologies have facilitated the identification of a very large number of SNPs. With these developments, Atlantic salmon researchers have access to resources that are comparable to those available for major livestock species. Also, there have been at least three independent initiatives taken by groups from the world leading producers for this species (Norway, Chile and Scotland) to generate SNP arrays of different densities. In particular, SNP chips of 6K (Brenna-Hansen *et al*., 2012), 6.5K (Kent *et al*., 2009), 15K (Dominik *et al*., 2010), 130K (Houston *et al*., 2014) and 200K SNPs (Yáñez *et al*., 2014) and the 220K used here (Barson *et al*., 2015) have already been developed. These developments are greatly facilitating research conducted for understanding linkage disequilibrium patterns, estimating effective population sizes and detecting loci controlling traits of interest.

Linkage disequilibrium

Linkage disequilibrium (LD) refers to the non-random association of alleles at two or more loci which results in a higher frequency of certain haplotypes than expected by chance (Falconer and MacKay, 1996). The magnitude and extent of LD in a particular population is jointly affected by evolutionary forces (such as random drift, natural selection and mutation), molecular forces such as historical recombination events, and the population's breeding history including historical effective population sizes, intensity and direction of artificial selection, population admixture, and mating patterns (Du *et al*., 2007).

Several measures have been proposed to quantify LD between a pair of loci and all of them are related to the statistic D, which is defined as $D_{AB} = p_{AB} - p_A p_B$; i.e., the difference between the frequency of gametes carrying the pair of alleles A and B at two loci (p_{AB}) and the product of the frequencies of those alleles (p_A and p_B). Currently, the most commonly used measures of LD are the statistic D', the value of D standardized by the maximum absolute value it can attain given the allele frequencies (Lewontin, 1964) and r^2, the square of the correlation coefficient between two indicator variables – one representing the presence or absence of a particular allele at the first locus and the other representing the presence or absence of a particular allele at the second locus (Hill and Robertson, 1968). The r^2 measure is preferred over the D' measure as it is considerably more robust to allele frequency variation (Du *et al*., 2007). The D' measure tends to be strongly overestimated in small sample sizes and in the presence of rare or low frequency alleles.

A detailed knowledge of the magnitude and extent of LD within the population under study is required for assessing the feasibility of genome-wide associations studies (GWAS) and the accuracy of genome-wide selection and for inferring past and current effective population size. While the study of LD has a long history (Sved, 1971), research on the magnitude and extent LD in animal populations has recently become a focus area, particularly due to the advent of high throughput genotyping techniques and the development of high-density SNP chips in farm species. Dense SNP panels have been used to estimate genomic LD patterns in cattle (Farnir *et al*., 2000; Odani *et al*., 2006; Sargolzaei *et al*., 2008; Kim and Kirkpatrick, 2009; Qanbari *et al*., 2010b; Zhu *et al*., 2013; Edea *et al*., 2014), sheep (Meadows *et al*., 2008; Kijas *et al*., 2012; Al-Mamun *et al*., 2015), pig (Nsengimana *et al*., 2003; Amaral *et al*., 2008; Badke *et al*., 2012; Saura *et al*., 2015; Zanella *et al*., 2016), horse (Corbin *et al*., 2010; Do *et al*., 2014) and chicken (Rao *et al*., 2008; Qanbari *et al*., 2010a) breeds. Patterns of LD have been also investigated in aquatic species and in particular in Atlantic salmon using low-density SNP arrays (Gutierrez *et al*., 2015).

Effective population size

The effective population size (N_e) determines the rate of loss of neutral genetic variability and thus, it has a central role in evolution, ecology, animal breeding and conservation genetics. It is the size of an idealized population that would lose genetic diversity (or become inbred) at the same rate as the actual population (Frankham *et al*., 1995).

Wild Atlantic salmon populations have supported rod and net fisheries (e.g., Butler *et al*., 2009), and provided source stock for an aquaculture industry that produces over two million tonnes per year. However, wild populations of Atlantic salmon are considered to be endangered and in fact Atlantic salmon was included in the Red List of threatened species in Europe (Porcher and Baglinière, 2001). As mentioned above, historic angling records suggest the occurrence of a drastic decline in the Spanish Atlantic salmon census size during the second half of the 19th century, and therefore conservation plans were established. This decrease in census size may entail loss of genetic variation that can influence the dynamics and persistence of populations. Thus, maintaining N_e is central to the application of optimal conservation strategies.

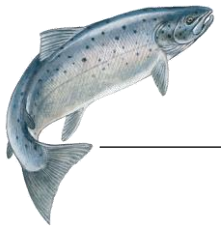
Demographic data from mating systems, pedigree or genetic data can be used to infer N_e. Unfortunately demographic data are generally difficult to collect in many wild populations and life-history information is often unavailable with sufficient precision to make a good estimate of N_e (Frankham, 1995). Thus, genetic methods have been developed (Wang, 2005; Luikart *et al.*, 2010) and their use has been potentiated in recent years with the increasing availability of large numbers of SNP markers. In particular, the single-sample method based on LD has become the method of choice as it leads to estimates of N_e of higher accuracy than other methods (e.g., the temporal method) and allow to infer N_e at any point in time. The LD method has been already used in salmon populations (Ribeiro *et al*., 2008; Palstra *et al*., 2009) using a limited number of microsatellite makers. Thus, it is of interest to take advantage of the large amount of genomic information currently available for obtaining estimates of N_e with potentially higher accuracy.

Sex determination in salmon

Sex determination and differentiation are fundamental biological processes that are mostly determined by a X/Y chromosomal system. As mentioned above, salmonids are descended from a common ancestor that underwent an autotetraploidization event. After a whole genome duplication, species could deal with sex determination by deleting one copy of the sex-determining gene, or by recruiting a duplicated transcription factor to become a novel sex-determining gene (Davidson, 2009). It is unknown which if any of these strategies salmonids adopted, but it appears that they all have primarily a genetic mechanism of sex determination with male heterogamety, although morphologically distinguishable sex chromosomes are not generally found. It is clear that salmonids are at an early stage in sex chromosome differentiation and therefore provide a good opportunity to study the evolution of sex determination. The reference salmonid genome sequence represents an important resource for research in this area (Davidson *et al*., 2010; Lien *et al*., 2016).

Several microsatellite genetic markers were identified as being linked to sex in Atlantic salmon, brown trout (*Salmo trutta*), rainbow trout (*Oncorhynchus mykiss*) and Arctic charr (*Salvelinus alpinus*) but markers differ across species (Woram *et al.*, 2003; Davidson *et al*., 2009). This led to the suggestion that there were either several sex-determining genes in salmonids or a single sex-determining gene that could move around the genome by transposition or translocation of a small chromosome arm (Woram *et al.*, 2003; Davidson *et*

al., 2009). The discovery of the master sex-determining gene in rainbow trout; i.e., the *sdY* gene (Yano *et al*., 2012), and its presence in males of many other salmonid species, including Atlantic salmon (Yano *et al.*, 2013), provides support for the hypothesis of a universal salmonid sex determination "jumping gene." Indeed, using microsatellites and a limited number of SNPs, Eisbrenner *et al.* (2014) identified three sex-determining loci in three different chromosomes (Ssa02, Ssa06 and Ssa03) in a Tasmanian Atlantic salmon population. The three loci showed evidence for the presence of *sdY*, which suggests that this gene might be moving across these three chromosomes. In any case, salmon sex determination is still very unclear and the use of genome-wide association studies (GWAS) with high-density SNP chips could help to identify the genes involved in this process.



Objectives

The main objectives of this thesis were:

1. To characterize genome-wide patterns of linkage disequilibrium in Spanish wild populations of Atlantic salmon.

2. To estimate the effective population size in the salmon populations from the genome-wide linkage disequilibrium measures obtained under Objective 1.

3. To identify genomic regions involved in sex determination in Spanish Atlantic salmon through a genome wide association study.

All objectives were achieved using a high-density SNP chip (220 K). **Chapter 1** deals with Objectives 1 and 2 and **Chapter 2** deals with Objective 3.

Chapter 1

Characterisation of linkage disequilibrium patterns and estimation of effective population size in Spanish populations of Atlantic salmon using genome-wide information

Introduction

The natural southern range of distribution of Atlantic salmon in Europe corresponds to rivers located in the North of Spain. In these rivers, salmon abundance has suffered a severe decline during the last century, mainly because of the construction of dams associated with the hydroelectric development, pollution and over-fishing (Braña *et al.*, 1995b).This decline has been observed in all species range and in other salmonid species (Law, 2000).

Since the late 1980s, conservation programmes have been put into practice in Spain focusing on three main actions: (i) improving salmon accessibility by constructing fish passes and trapping facilities; (ii) regulating angling by the concession of licenses and the setting of an angling period; and (iii) establishing supportive breeding programmes consisting of artificial spawning of salmon in farms and subsequent release of the offspring into the river.

It is currently accepted that in order to maintain the local adaptations and the population structure, stocking should be carried out with native individuals. In fact, the stocking carried out in the 1970s with foreign individuals, mainly from Scotland, but also from other European countries, such as Ireland, Norway and Island was unsuccessful in terms of significantly increasing the number of returning individuals (see Blanco *et al.*, 2005; Morán *et al.*, 2005b) but led to some introgression (Campos *et al.*, 2007), modifying (at least in part) thus the native genetic composition. Therefore, since the late 1980s supportive breeding is carried out only with native individuals and this has resulted in a significant increase in salmon in Spanish rivers and the recovery of some extinct populations (Caballero, 2002). Although the increase in the number of returning adults is evident, this occurrence might not necessarily be paired to an increase in effective population size (N_e).

Effective population size is an important parameter in animal genetics as it determines the rate at which genetic variability is lost. It can be defined as the size of an ideal population (i.e., one that meets all the Hardy-Weinberg assumptions) that would lose heterozygosity at a rate equal to that of the observed population. Thus, knowledge of N_e facilitates the design of efficient artificial selection schemes in animal breeding (e.g., Villanueva *et al.*, 2004) and the effective management of populations of endangered species (Frankham, 1995).

Traditionally, N_e has been estimated from demographic or pedigree data (Caballero, 1994). However, this information is usually unavailable or incomplete. In these cases, N_e can be estimated from genetic data (Wang, 2005). Until recently, most genetically based estimates of contemporary N_e have used the temporal method, which is based on the temporal changes in allele frequencies for samples of the same population collected at (at least) two different points in time (Nei and Tajima, 1981; Waples, 1989; Wang, 2001; Anderson, 2005). Nonetheless, the availability of high-density SNP chips has led to an increased interest on the method based on linkage disequilibrium (LD) (Nomura, 2008; Tallmon *et al*., 2008; Waples and Do, 2008; Pudovkin *et al*., 2009; Wang, 2009).

One of the main advantages of the LD method over the temporal method is that it uses much more information and therefore leads to estimates of higher accuracy (Waples and England, 2011; Waples and Do, 2010, Luikart *et al*., 2010). Also, the strength of LD at different genetic distances between loci can be used to infer N_e at any point in time from a single sample (Waples and Do, 2010), as LD between a pair of markers is determined by the product of N_e, recombination frequency and number of generations since mutation (Hill and Robertson, 1968).

Studies investigating the levels and strength of LD in Atlantic salmon are really scarce. Recently, Gutierrez *et al*. (2015) investigated the magnitude of LD and its extent and decay with distance in this species, using a 6.5K SNP array. Their results showed that although LD at short distances was high, it declined rapidly as distance increases. Given these results, they concluded that the density of the SNPs used should be increased in order to improve the power to detect association between traits of interest and markers.

The objectives of this study were to characterise LD patterns in wild Spanish populations of Atlantic salmon, using a high-density SNP panel, and to evaluate the evolution of N_e estimated from LD during the last century, where the drastic decline in numbers has occurred. Spanish populations have the particularity of suffering the most extreme conditions of the distribution range of the species in the world.

Material and Methods

Genotypic data

Data used in this study originated from 192 samples of Atlantic salmon from six Spanish rivers that cover all the distribution range of the species in the country. These rivers (see Figure 1 in the General Introduction) are: Miño (16 samples), Ulla (112 samples), Eo (16 samples), Sella (16 samples), Urumea (16 samples) and Bidasoa (16 samples). Information on sex and returning year is available for these samples (Table 1.1).

Table 1.1. Total number of individuals (N), number of males (N_m) and females (N_f) and river of origin used in this study. The returning year of the samples refer to the year when adult (mature) individuals returned to the river to spawn.

River	N	N_m	N_f	Returning year
Miño	16	6	10	2011
Ulla	18	8	10	2008
	46	35	11	2010
	15	7	8	2011
	33*	12	20	2012
Eo	16	1	15	2013
Sella	16	3	13	2012
Urumea	16	9	7	2011
Bidasoa	16	3	13	2013

*Sex was not recorded in one sample.

A custom 220K Affymetrix SNP genotyping array (Aquagen/CIGENE) was used to genotype samples according to the manufacturer's instructions with a GeneTitan genotyping platform (Affymetrix). The SNPs on this array were a subset of those included on the 930K XHD *Ssal* array (dbSNP accession numbers ss1867919552–ss1868858426) and were chosen for maximum informativeness on the basis of their SNPolisher performance (SNPolisher, version 1.4, Affymetrix), minor allele frequency (MAF) observed in aquaculture samples (MAF > 0.05) and physical distribution (Barson et *al.*, 2015). An Atlantic salmon physical map was provided by CIGENE. The number of SNPs mapped on each chromosome is given in Table 1.2.

Table 1.2. Chromosome length in Mb (L), number of SNPs per chromosome (N_{SNP}) and SNP density calculated as the number of SNPs per chromosome divided by the chromosome length.

Chromosome	L	N_{SNP}	Density
Ssa01	159.04	2,744	80.13
Ssa02	72.94	4,693	64.34
Ssa03	92.50	7,935	85.78
Ssa04	82.40	6,288	76.31
Ssa05	80.50	6,444	80.05
Ssa06	87.04	6,970	80.08
Ssa07	58.79	4,903	83.40
Ssa08	26.43	1,729	65.42
Ssa09	141.71	9,739	68.72
Ssa10	116.14	8,165	70.30
Ssa11	93.89	6,526	69.51
Ssa12	91.88	5,830	63.45
Ssa13	107.76	7,330	68.02
Ssa14	93.90	6,833	72.77
Ssa15	103.96	7,006	67.39
Ssa16	87.80	5,553	63.25
Ssa17	57.68	3,730	64.67
Ssa18	70.70	5,389	76.22
Ssa19	82.98	5,359	64.58
Ssa20	86.80	5,758	66.34
Ssa21	58.02	4,000	68.94
Ssa22	63.42	4,832	76.19
Ssa23	49.85	4,101	82.27
Ssa24	48.65	4,413	90.71
Ssa25	51.48	4,202	81.62
Ssa26	47.90	3,609	75.34
Ssa27	43.94	4,003	91.10
Ssa28	39.60	3,232	81.62
Ssa29	42.49	3,406	80.16

Raw data were imported into the Affymetrix Genotyping Console software (v4.1) genotype calling using the Axiom GT1 algorithm (Affymetrix). The R package SNPolisher (Affymetrix) was used for quality control (QC) of the SNPs, which were classified into different categories based on the clusterization and quality of the genotypes resulting from the hybridization signal. Hybridization signals are usually assessed by analysing cluster locations in order to generate the most accurate genotype calls. Given a population of samples that exhibit the three genotypes for every SNP, the software can automatically determine the cluster positions of the genotypes. If certain SNPs have one or two clusters that lack representation, the missing cluster positions can be estimated. Quality control categories (Figure 1.1) were: (i) *Poly High Resolution*, where three clusters corresponding to the three

possible genotypes of a polymorphic SNP are formed with good resolution; (ii) *No Minor Homozygote*, where only two clusters are formed with no samples of the minor homozygous genotypes; (iii) *Off-Target Variant*, where an additional low intensity cluster resulting from slight mismatches between the probe and the samples is observed; (iv) *Mono High Resolution*, in which only one cluster is formed with good resolution (this pattern corresponding to monomorphic SNPs); (v) *Other*, where the resultant SNP cluster pattern did not fall into any of the previous classes; and (vi) *Call Rate Below Threshold*, where genotype call rate (quality metric that indicates the reliability of each genotype call and ranges from 0 to 1) was under a predefined threshold. SNPs not included in the categories of *Poly High Resolution* and *Mono High Resolution* satisfying the corresponding call rate threshold (0.97 in this case) were discarded.

Additionally, the software Plink v1.09 (Purcell *et al.*, 2007) was used to perform further filters for the SNPs. In particular, the following SNPs were excluded: (i) unmapped SNPs, (ii) SNPs with Minor Allele Frequency (MAF) < 0.01, and (iii) SNPs that deviated significantly from Hardy Weinberg equilibrium ($p < 10^{-5}$). The final number of SNPs retained after filtering was 164,722. Detailed information on the number of SNPs removed at each step can be found in Table 1.3. Samples with a DQC score (chip-level quality metric) ≥ 0.82 and call rate > 0.97 were retained. The final number of samples retained after QC was 187.

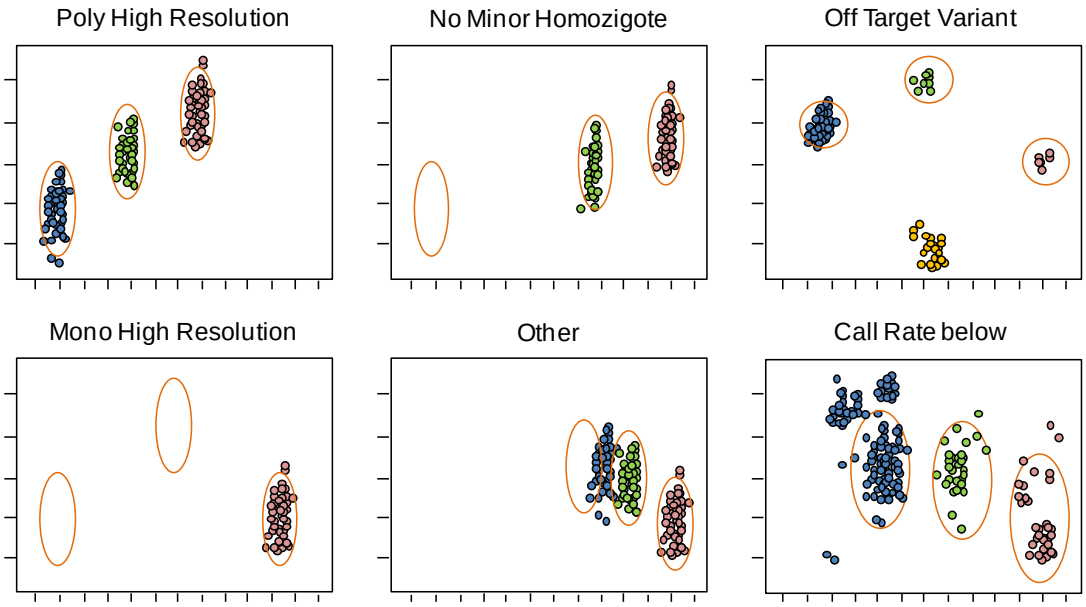


Figure 1.1. Illustration of quality control categories obtained from hybridization signal by the Axiom GT1 algorithm. (Source: Bassil *et al.*, 2015).

Table 1.3. Quality control criteria applied on genotypic data and number of SNPs discarded at each step.

Criterion	No. SNPs removed
No Minor Homozygote	30,160
OTV	1,236
Call rate below threshold	7,080
Other	9,379
Unmapped SNP	686
HWE ($p< 10^{-5}$)	450
MAF < 0.01	6,284
Total	55,278

Population structure

The population structure of the Spanish populations was assessed using the software Admixture 1.3 that implements a model-based clustering method that assumes *K* populations (where *K* is unknown) each of which is characterised by a set of allele frequencies at each locus (Zhou *et al.*, 2011). Individuals are assigned (probabilistically) to a single population, or jointly to two or more populations if their genotypes indicate that they are admixed. In order to choose the correct value of *K* (i.e. number of ancestral clusters), Admixture requires running several cases by changing the *a priori* value of *K*. The cross-validation procedure provides cross-validation errors for each value of *K*. The correct value for *K* is defined therefore by the test with the lowest value of cross-validation error.

Estimation of linkage disequilibrium

One of the most commonly used measures of LD is the squared correlation coefficient between SNP pairs (r^2). For two biallelic loci with alleles *A* and *a* at locus 1, and alleles *B* and *b* at locus 2, r^2 is computed as

$$r^2 = \frac{D^2}{p_A p_a p_B p_b},$$

where $D = p_{AB} - p_A p_B$, p_{AB} is the frequency of gametes carrying the pair of alleles *A* and *B* and p_A, p_a, p_B and p_b are the frequencies of alleles *A*, *a*, *B* and *b*, respectively (Hill and Robertson, 1968).

Pairwise r^2 between syntenic SNPs separated by up to a distance of 5 Mb were obtained using the software Plink v1.09 that implements an EM algorithm to estimate haplotype frequencies. The approach consists on setting a predefined distance bins (5 Mb in this study) and

estimating r^2 of all SNP pairs within the window that moves along the chromosome one SNP at a time. Given that LD is highly dependent of sample size, we applied the same sample size correction as that proposed by Weir and Hill (1980) for the estimation of N_e from LD measures (i.e., $r^2 - 1/N$, where N is sample size, see below).

In order to enable a clear presentation of results showing LD in relation to physical distance between markers, the chromosome length was divided into three distance classes: (i) 0.0 to 0.5 Mb, (ii) 0.5 to 1.0 Mb and (iii) 1.0 to 5.0 Mb. Distance bins of 0.005, 0.010 and 0.500 Mb were used for classes (i), (ii) and (iii), respectively, and average r^2 values for each bin were plotted against physical distance. To determine to which extent r^2 values are a function of distance between SNPs, background LD was calculated for a random set of non syntenic SNPs that included 1% of the SNPs per chromosome.

Estimation of effective population size

The well-known relationship between LD and N_e was derived by Sved (1971), and it was based on the work of Hill and Robertson (1968). However, Sved's equation did not account for mutation or for sample size. The consequence of ignoring mutation is that even with physical linkage the expected association between loci is incomplete and leads to an underestimation of r^2. The consequence of ignoring sample size is that the magnitude of r^2 can be overestimated in samples of small size as a consequence of spurious associations (Corbin *et al.*, 2012). Taking into account both mutation and sample size leads to a more general expression:

$$E[r^2] = (\alpha + 4N_e c)^{-1} + 1/N$$

where α is a parameter related to mutation, c is the recombination rate and the term $1/N$ is the adjustment for sample size. Note that $4N_e$ corresponds to the slope and α to the intercept of the regression equation. Parameter α can be set to a fixed value (one in the absence of mutation or two when mutation is considered) (Ohta and Kimura, 1971; McVean, 2002) but it can also be estimated from the data. Data were analysed here both estimating the parameter α and fixing its value to two.

For predicting N_e at any generation in the past, we have to consider r^2 between SNP pairs at a specific linkage distance in Morgans (d). This distance yields a prediction of the time since the gametes are expected to have coalesced $1/2c$ generations ago (Hayes *et al.*, 2003; De Roos *et al.*, 2008). Thus,

$$N_{e(t)} = (4d)^{-1}\left[(r_d^2 - N^{-1})^{-1} - \alpha\right],$$

where $N_{e(t)}$ is the effective population size *t* generations ago and r_d^2 is the mean value of r^2 for markers *d* Morgans apart. Based on this equation a non-linear least squares approach was implemented to statistically model the observed r^2.

In Atlantic salmon, generations overlap. In order to account for this, a number of consecutive cohorts equal to the generation interval (L = 3 in Spanish populations) need to be included in the data set to be analysed (Saura *et al.*, 2015). During the returning season in a particular year, salmon entering into the river represents the three cohorts.

With the aim to avoid dependence between LD and linkage distance estimated from the same data, it is recommended to use estimates of recombination rates for each chromosome from a different dataset for the same species (Saura *et al.*, 2015). Specifically, we used estimates from Lien *et al.* (2011). Therefore, for each SNP pair in LD separated by a particular physical distance (Mb), an equivalent linkage distance (Morgans) was calculated as the product of the recombination rate for a particular chromosome obtained from Lien *et al.* (2011) and the physical distance.

Results

Population structure

The cross-validation test revealed that the correct value for *K* was equal to two, i.e., two main ancestral clusters compose the current Spanish population of Atlantic salmon (Figure 1.2). These clusters include a clear Atlantic group with individuals from rivers Miño and Ulla (Atlantic metapopulation), and a Cantabric group with individuals from rivers Sella, Urumea and Bidasoa (Cantabric metapopulation). River Eo fell in the middle of both groups. In order to better understand the connection among populations, we also represented the results for *K* = 6. Here, results are represented as the proportion of each ancestral cluster in each river under study. Again, two main groups (green for Atlantic, yellow for Cantabric, Figure 1.3) can be observed, and the Atlantic group showed a higher purity or isolation (Figure 1.3.). However, when analysing the levels of inbreeding, similar observed (and expected) levels were found in these rivers, ranging from 0.62 (0.64) in Sella, Urumea and Bidasoa to 0.66 (0.64) in both Miño and Ulla.

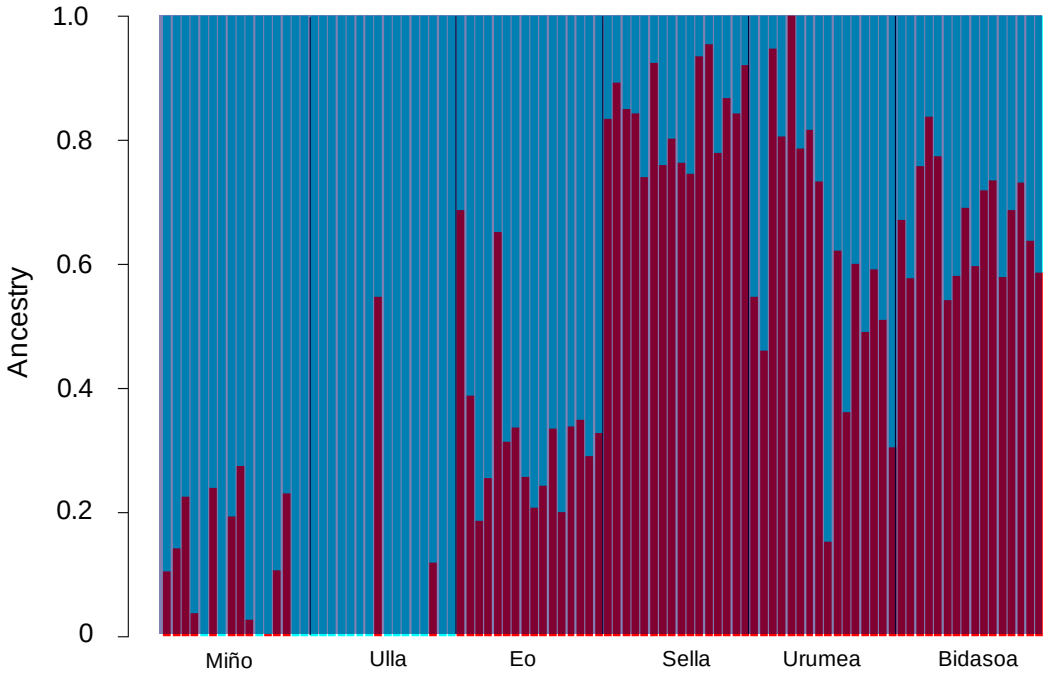


Figure 1.2. Proportion of each ancestral cluster (blue: Atlantic, red: Cantabric) within each individual (bars) from each river when *K* was set to 2.

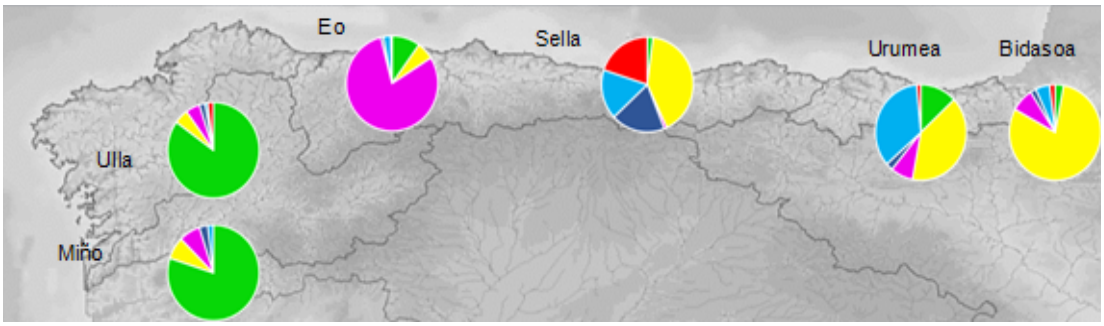


Figure 1.3. Proportion of each ancestral cluster (blue: Atlantic, red: Cantabric) within each population when *K* was set to 6.

Linkage disequilibrium

Figure 1.4 shows r^2 values against distance between SNPs for the river Ulla correcting (B) or not (A) for sample size. As mentioned above, river Ulla was represented by a higher number of individuals (108) than the other rivers (15 or 16 samples). We first obtained LD values for six subsamples of the Ulla population. LD for these subsamples was very similar but considerably higher than LD for the whole Ulla sample (Figure 1.4A). After correcting for sample size, however, subsamples and the whole sample of river Ulla showed very similar

patterns, allowing thus the comparison of LD across samples of different sizes (Figure 1.4B). Henceforth, the results presented include the correction for sample size.

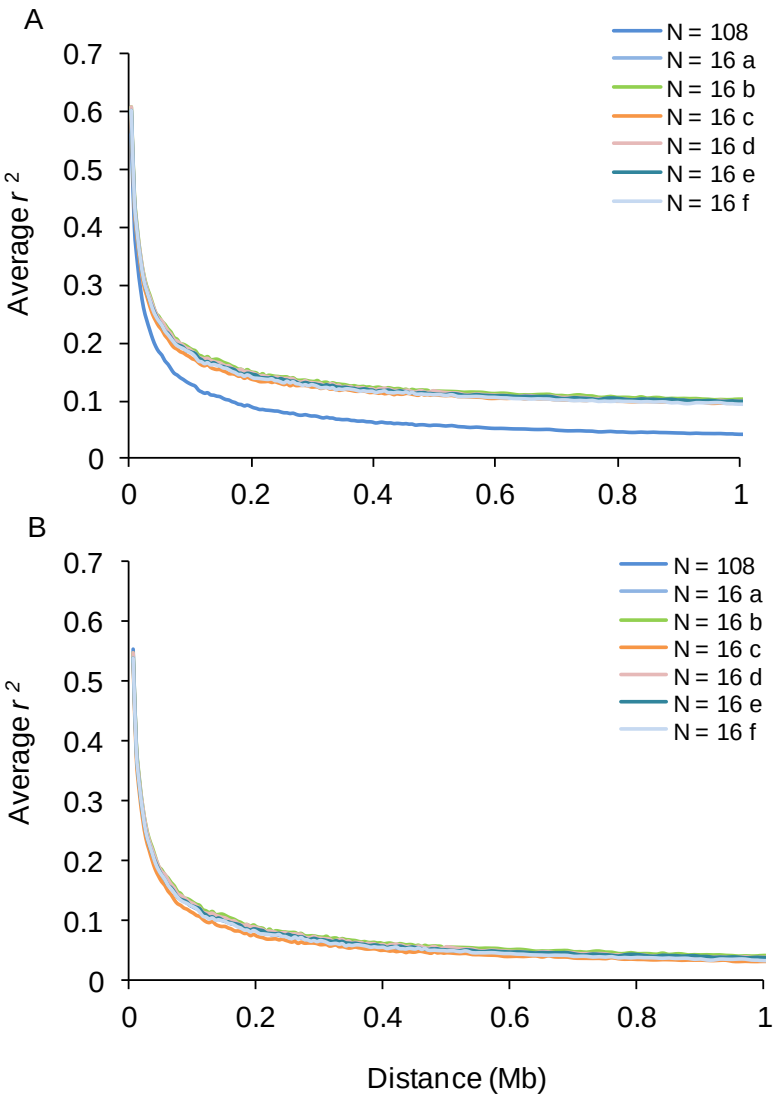


Figure 1.4. Average linkage disequilibrium measured as r^2 plotted against the average of the distance bins (Mb) for the complete sample of river Ulla (108 individuals) and different subsamples (16 individuals each, a-f), correcting (B) or not (A) for sample size.

As it can be observed in Figure 1.5B, the Atlantic metapopulation showed slightly higher levels of LD than the Cantabric metapopulation. LD was found to be relatively high between closely linked markers for each river (Figure 1.5A), and consequently for both metapopulations (Figure 1.5B). The average r^2 across rivers was 0.52 at 0.005 Mb (Table 1.4). However, it declined rapidly with increasing distance between SNP pairs. The average r^2 decreased by half over the first 0.05 Mb and by 90% at 0.3 Mb and it was then maintained for

longer distances. The average r^2 was reduced to non syntenic levels (0.063± 0.025) at distances less than 5 Mb.

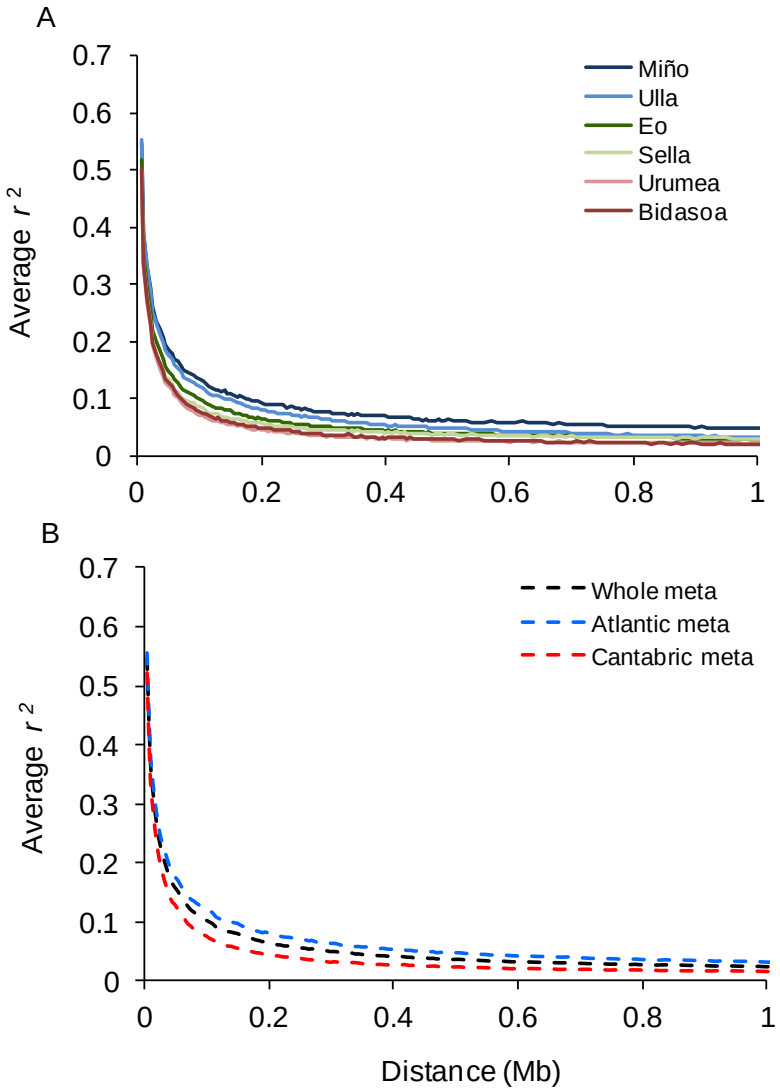


Figure 1.5. Average linkage disequilibrium measured as r^2 plotted against the average of the distance bins (Mb) for the different rivers (A) and metapopulations (B).

Table 1.4. Average r^2 estimated for SNP pairs at different distances (from zero and the distance indicated) for each river and averaged across rivers.

	Distance from zero to			
River	0.005 Mb	0.05 Mb	1 Mb	5 Mb
Miño	0.547	0.287	0.091	0.089
Ulla	0.554	0.282	0.078	0.075
Eo	0.518	0.250	0.030	0.064
Sella	0.499	0.231	0.060	0.058
Urumea	0.493	0.219	0.046	0.044
Bidasoa	0.500	0.229	0.052	0.050
Average	0.520	0.250	0.060	0.063

We also studied LD patterns for each chromosome. Figure 1.6 shows variation in the magnitude and extent of LD on different chromosomes for the case of river Ulla as an example. In general, Ssa13 showed the highest values of r^2 while Ssa08 showed the lowest values.

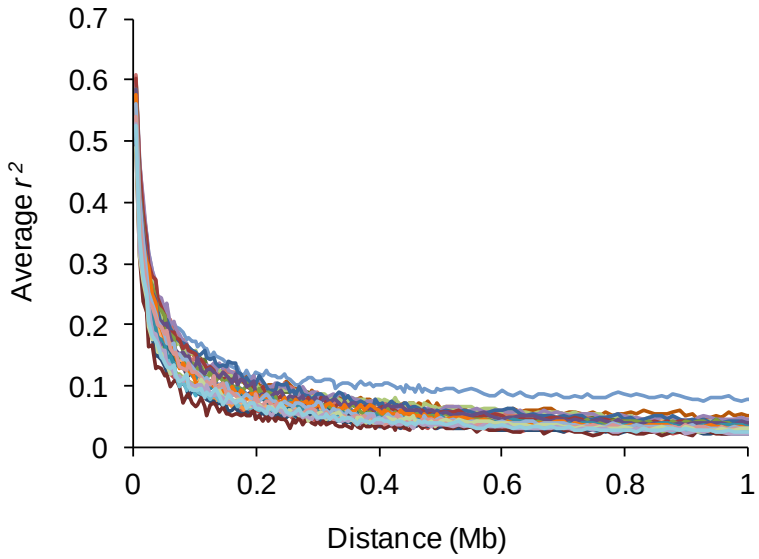


Figure 1.6. Average linkage disequilibrium measured as r^2 plotted against the average of the distance bins (Mb) for the different chromosomes for river Ulla.

Effective population size

The evolution of N_e was studied from the period between six and 50 generations ago. Note that the most recent sample belongs to the cohort returning in 2013 (i.e., six generations ago). Effective population size in the more recent generations was not estimated because r^2 between SNPs separated at the genetic distance corresponding to this time (i.e., physical distance of 5 Mb) was no longer a function of the distance, as showed by the background levels of LD.

Results are presented under a model where the parameter α was fixed to the theoretical value of two (i.e., accounting for mutation). When this parameter was estimated its value highly departed from the theoretical value, thus leading to unreliable estimates of N_e. Estimates of N_e six generations ago were three to five times lower than ancestral estimates corresponding to 50 generations ago (Figure 1.7). Values for N_e ranged from about 500±2.90 (river Miño) to 1200±10.79 (river Bidasoa) individuals 50 generations ago, and from 100±0.71 to 400±2.90 individuals six generations ago. Standard errors were low and similar for all rivers.

In order to investigate if there were differences in the rate of decrease of N_e before and after the decrease in census size due to the hydroelectric development that occurred during the 1950s, regressions of N_e on generation number before and after this point in time (about 25 generations ago), were carried out. The slopes of the regressions indicated that the rate of decrease in N_e was higher after such development started (b = 16) than in the previous period (b = 12).

In addition, an isolation-by-distance pattern was observed, both from LD patterns as well as from N_e estimates (see also Figures 1.4 and 1.5) and this is consistent with the occurrence of migration between salmon from close populations due to straying. Given than the population structure analysis indicated two main metapopulations, we estimated N_e for both. The estimated N_e for the Cantabric metapopulation was almost three times higher across time than N_e for the Atlantic metapopulation although the rate of decrease in N_e was faster for the former.

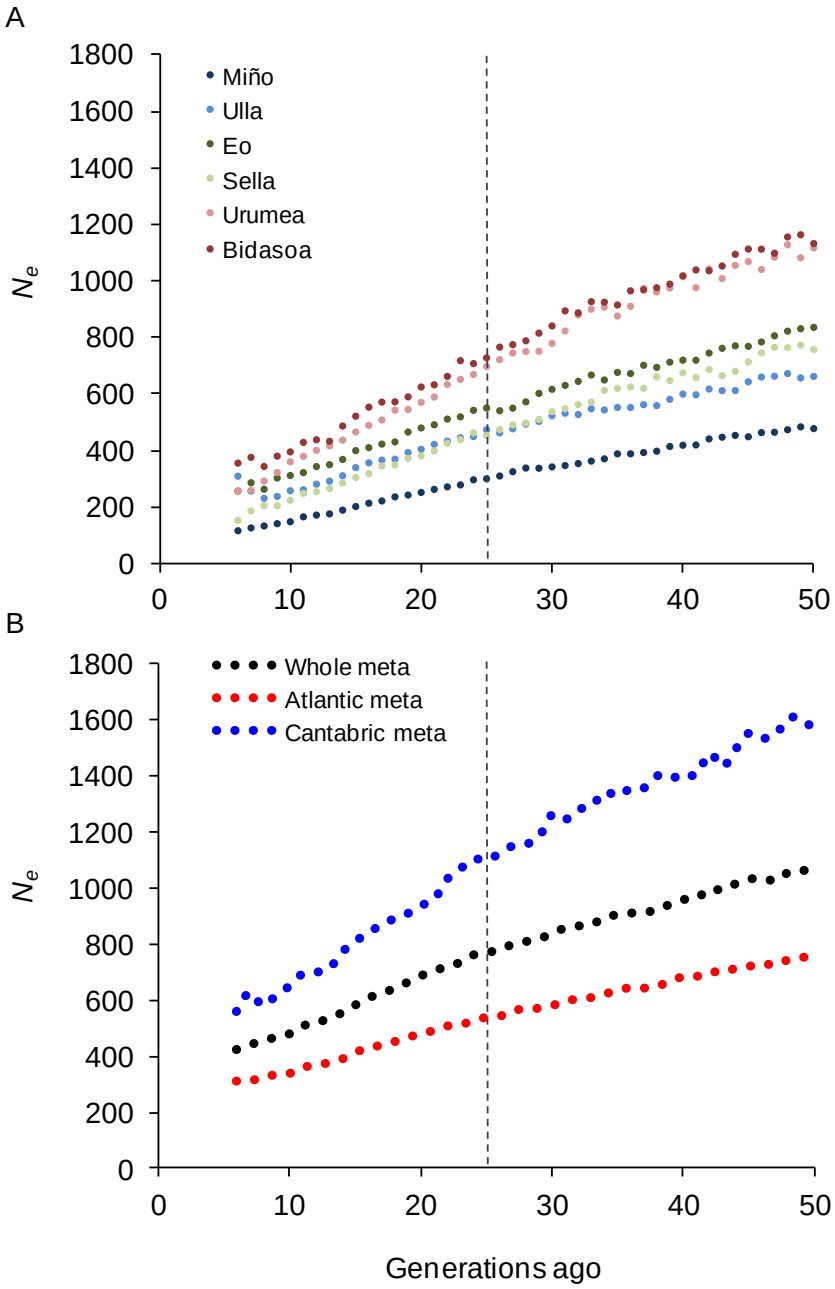


Figure 1.7. Estimates of N_e from LD measures across time (generations in the past) for the six rivers (A) and for the metapopulations (B). The broken line corresponds to the time when hydroelectric development started in Spain (about 25 generations ago).

Discussion

In this study, we have investigated LD patterns and the evolution of N_e from those patterns in Spanish wild populations of Atlantic salmon in the last 50 generations. Our results revealed that N_e six generations ago was over three times lower than 50 generations ago, before the decline in census size occurred in Spanish rivers in the second half of the 19th century mainly due to the hydroelectric development. Previous N_e estimates about two generations ago obtained for the river Eo with classical markers (microsatellites) were of the same order of magnitude but lower to those found here (Ribeiro *et al*., 2008) when comparing for the same years.

The high levels of LD at short chromosome distances and the fast decline as distance increases observed when using the 220K SNP array are consistent with the results from Gutierrez *et al*. (2015) who analysed data from a Tasmanian population of Atlantic salmon (originally from Canada) using a much less dense SNP array (6.5K). Their average r^2 between SNPs 0.005 Mb apart was 0.89. This value is higher than the value observed here in the Spanish populations (0.52 in average), even correcting by their sample size. This is probably due to the fact that the Tasmanian population was a farm artificially selected population. However, in both studies r^2 was reduced by half over the first 0.05 Mb, suggesting that, contrary to what authors suggested, the SNP density of the 6.5K chip could be enough for obtaining reliable levels of LD. Comparing to terrestrial farm animals such as cattle (Pérez O´Brien *et al*., 2014; Porto-Neto *et al*., 2014) or sheep (Kijas *et al*., 2014), the magnitude of r^2 at short distances is within the levels observed in the salmon populations, but its decline is faster in salmon and therefore it extends shorter distances than terrestrial farm species. This rapid decline in salmon could be explained because of the high fecundity typical of fish that would translate in a very high number of meiosis and recombinants. Data from tagging and recapture indicates that the rate of return of adults from juveniles stocked ranges between 0.5 and 17‰. This may imply a considerable number of returning adults, as the number of eggs released by a female ranges between 3,000 to 15,000 (Álvarez *et al*., 2010). Notwithstanding, given the results from this study that indicate that N_e is decreasing across time, this explanation is less plausible. Another possible explanation for this rapid decline on LD is the complexity of the salmon genome. The whole-genome duplication of the common ancestor of salmonids ~80 million of years ago is still in a reploidization process (Lien *et al*., 2016). This leads to a high proportion of repetitive sequences that makes difficult the annotation of SNPs

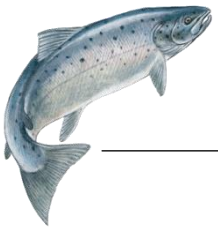
positions, given the early stage of development of the salmon genome sequence, whose second version has been recently uploaded to the NCBI database. The genetic flow that seems to happen between different populations could also be the cause of this rapid decline in LD with the distance between markers.

Variation in the magnitude and extent of LD can be affected by sample size. Here we have observed that LD was higher for the smaller sample sizes. Thus, although subsamples of 16 individuals from the whole sample from river Ulla gave similar patterns of LD across SNP distance, this pattern was different when the whole sample (108 individuals) was considered. When r^2 was corrected for sample size using the correction of Weir and Hill (1980), we observed that LD levels were comparable across rivers (Figure 1.3).

Conservation programs started in Spain in the late 1980s (about eight generations in the past), using native individuals to perform supportive breeding. Although the programs were successful in terms of increasing the number of returning adults, here we cannot provide any clear conclusion about the effect of the supporting breeding on N_e, as we only could trace back N_e up six generations ago, only two generations after the conservation programs were put into practice. The impossibility of estimating N_e in the most recent generations is due to the fact that the pairwise SNP distance corresponding to non-syntenic levels of r^2 (about 5 Mb) reflects N_e six generations in the past. This means that r^2 between SNPs separated at higher distances is no longer a function of distance, and therefore the method is not reliable for the most recent generations.

Estimating the slope of the regression of N_e on time before and after the period when hydroelectric development was more intense (25 generations ago) allowed us to conclude that such development had a negative impact on the N_e of the different populations. The combined effect of improving river accessibility (that facilitates the natural returning of adults) and supportive breeding practices (that improve survival increasing thus the number of individuals contributing to the next generation), may be translating in an increase in N_e in the most recent time. Samples from salmon returning nowadays would be very valuable for evaluating the real impact of the supportive breeding programmes.

Records of recreational fishery catches from 1950s to 1990s, although only tentative, indicate that the beginning of 1970s was an unfortunate year for rivers Ulla, Eo and Sella, while the early 1980s were bad for river Bidasoa (García de Leániz and Martínez, 1986). It is clear however that the demographic history of each river is different and that rivers have been subjected to different threats. Results from the analysis of population structure indicated that two ancestral nucleus (one Atlantic and one Cantabric) are composing these populations, showing a pattern of isolation by distance. Despite of the usually strong homing tendency resulting in genetic differentiation and local adaptations (Garcia de Leaniz *et al.*, 2007), some gene flow occurs between neighboring populations with straying rates ranging between 0 and 19%. Rivers from the Atlantic metapopulation (i.e., Miño and Ulla) were the most impacted by the construction of dams, reducing notably their accessible area to salmon and showing the lowest estimates of N_e. Salmon in these rivers are also adapted to waters from the Atlantic Ocean, with colder temperatures than the Cantabric Sea. This supports a certain degree of geographic isolation between both (Atlantic and Cantabric) metapopulations (Álvarez *et al.*, 2010). On the other hand, the Cantabric metapopulation showed the highest estimates of N_e. These rivers may receive natural migrations not only from Spanish rivers but also from close French rivers. In addition, the effect of stocking is not negligible and the stocking history with salmon from North Europe, may have contributed to an increase in N_e in these rivers compared to Atlantic rivers. This population admixture makes difficult to interpret ancestral N_e estimates, as the samples investigated may reflect the history of different lineages that are not pure any more. Despite of this, the population structure analysis reflects well the population history on the face of the isolation-by-distance pattern observed. In any case, the rate of decrease in N_e in the Atlantic populations was lower than that observed for the western populations, suggesting that the Atlantic metapopulation has been maintained more stable. This finding questions the supportive breeding that is currently performed; i.e., with salmon of the same river. Stocking with salmon of the same metapopulation (Atlantic or Cantabric) might be a more appropriate alternative for increasing N_e (and therefore genetic diversity) while maintaining local adaptations.



Chapter 2

Identification of genomic regions involved in sex determination in Spanish Atlantic salmon using genome-wide information

Introduction

Controlling the sex ratio is essential in fish aquaculture and conservation management. Therefore, in both contexts, understanding the genetic architecture of sex determination (SD) and differentiation is an important issue. Contrary to mammals and birds, which show conserved master genes responsible for gonad differentiation (GD), an enormous diversity of SD mechanisms has been reported in fish.

Although in all salmonid species investigated to date morphologically distinguishable sex chromosomes are not generally found, the SD system is often described as always being male heterogametic (XX/XY) (Davidson *et al*., 2009). However, this conclusion has been based only on the knowledge gained from the examination of a small subset of genera (*Oncorhynchus*, *Salmo* and *Salvelinus*), all belonging to the *Salmoninae* subfamily. In the majority of these species, the SD locus has been identified on different chromosomes (Woram *et al*., 2003; Li *et al*., 2011), leaving an open question about the uniqueness of a master SD gene in salmonids (Davidson *et al*., 2009).

Recently, the *sdY* gene (sexually dimorphic on the Y chromosome) was identified as a male-specific linked gene on the Y chromosome in rainbow trout (Yano *et al*., 2012) and subsequently it was found in most salmonids (Yano *et al*., 2013), strongly suggesting that the *sdY* may be the conserved master SD gene on this family (Lubieniecki *et al*., 2015). In Atlantic salmon, Woram *et al*. (2003) and Artieri *et al*. (2006) identified a SD gene located on Ssa02. However, recently Eisbrenner *et al*. (2014) analysed a Tasmanian farm Atlantic salmon population (originally from Canada) and found that the *sdY* gene suffers genomic instability, mapping on Ssa02 but also on Ssa03 and Ssa06.

In general, genetic mapping research on SD loci in salmonids, and particularly in Atlantic salmon, is surprisingly scarce and therefore very little is known. In fact, environmental factors have been suggested to play also an important role on fish SD and epigenetic mechanisms are becoming increasingly recognized for the establishment and maintenance of the GD pathways. Although major genetic factors are frequently involved in SD, this system resembles complex in fish (Martínez *et al*., 2014).

Given this background it is obvious the interest of investigating sex determination with high-density genomic information as this information allows a more detailed characterization of how many sex loci are involved in Atlantic salmon, and their location on the genome. Thus, the objective of this study was to identify genomic regions involved in sex determination in Spanish Atlantic salmon populations through a genome-wide association study (GWAS) exploiting the high-density 220K SNP chip (Aquagen/CIGENE) already developed for this species.

Material and methods

Genotypic data and phenotypes

Genotypic data used for this analysis were those previously described in Chapter 1. Briefly, 192 samples from six Spanish rivers (Miño, Ulla, Eo, Sella, Urumea, Bidasoa) that cover all the distribution range of the species in the country were available. Samples were genotyped with the Affymetrix SNP genotyping array (Aquagen/CIGENE). Sex was recorded from mature adults in trapping facilities during their return to the river to spawn. After quality filters the data set included 164,722 SNPs and 187 samples (104 females and 83 males).

Genome-wide association study

GWAS was carried out using the GenABEL software implemented in an R statistical environment (http://www.r-project.org). We focussed on detecting statistical associations between SNPs and sex by testing each SNP individually and correcting for population structure. To account for multiple testing, a very stringent *p* value is typically used (e.g., the Bonferroni correction). This reduces the occurrence of false positives, but at the expense of missing many real associations, especially if individual SNPs have a small effect on the trait (Yang *et al.*, 2010). Thus, we used an alternative approach (the quantile-quantile -QQ- plot approach) to overcome this problem. The QQ plot represents potential deviations of the observed and expected (from a theoretical χ^2-distribution) *p* values. SNPs deviating from the line X = Y are considered to be significant.

Those SNPs that showed a significant association with sex according to the QQ-plot were considered to be included in putative QTL regions. These regions were defined as segments of the genome contained at least two significant consecutive SNPs separated less than 1Mb.

Functional annotation

In order to identify the genes included in each region, the *Salmo salar* genome sequence available in the NCBI database was used. Genes that were not annotated in the Atlantic salmon sequence were identified through BLAST, which finds homologous sequences between species (mostly fish species in this case). With the objective of mapping genes of interest, the BLAST search was performed in two ways: (i) using the Atlantic salmon sequence against the NCBI database; and (ii) using homologous gene sequences from a different fish species close to Atlantic salmon against the Atlantic salmon genome. The first strategy provides the gene content of a specific region and the second strategy maps a specific gene.

Results

Three hundred and seventeen SNPs (Figure 2.1) showed significant associations according to the probability threshold obtained from the QQplot (-log(*p*-value) > 2.8, Figure 2.2).

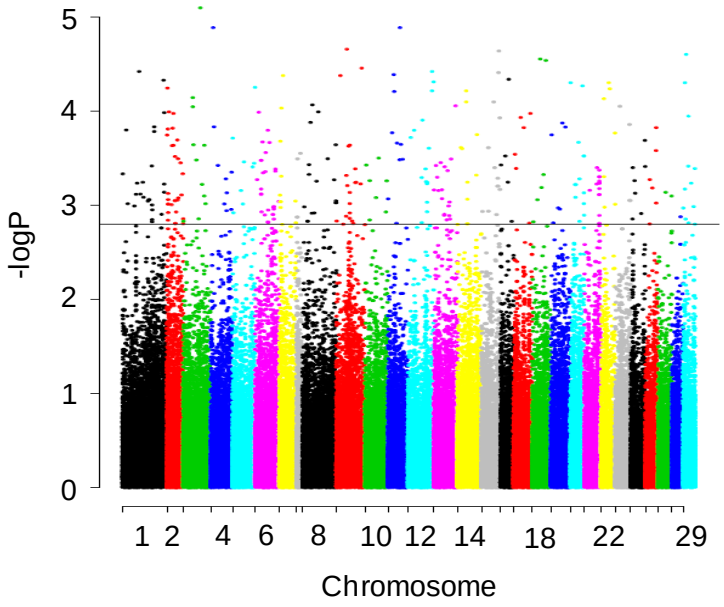


Figure 2.1. Manhattan plot representing the -logP associated to each SNP analysed.

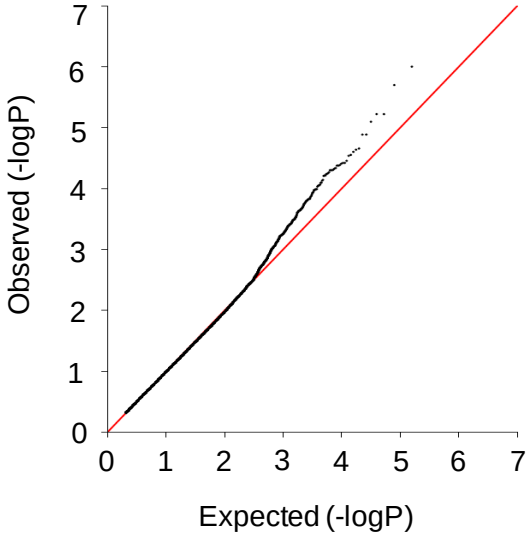


Figure 2.2. QQ plot resulting from the association analysis.

As explained above, those significant consecutive SNPs separated less than 1 Mb were considered as part of the same region. In total, nine regions were detected (Table 2.1) and they were located on Ssa02 (two regions), Ssa06 (one region), Ssa09 (one region), Ssa10 (two regions), Ssa21 (two regions) and Ssa22 (one region).

Table 2.1. Genomic regions potentially associated with Atlantic salmon sex determination.

Region ID	Chromosome	No. SNPs	Start bp	Stop bp	*p-value*
1	Ssa02	4	11633167	11713693	10^{-4}
2	Ssa02	3	31125898	31128201	10^{-5}
3	Ssa06	2	63726636	63743235	10^{-3}
4	Ssa09	3	137325858	137330086	10^{-4}
5	Ssa010	5	42962128	43063163	10^{-4}
6	Ssa010	2	76686499	76707523	10^{-4}
7	Ssa021	4	16086799	16108581	10^{-4}
8	Ssa021	3	54241448	54254620	10^{-4}
9	Ssa022	10	54691495	55593832	10^{-4}

The functional annotation revealed that the most interesting regions associated with sex differentiation and sexual development were in chromosomes Ssa02 and Ssa06 because they

may contain the following genes: (i) the *sdY* gene, which is the master-sex-determining gene already identified in 15 species of salmonids (Ssa02, at 36 Mb and very close to our region 2, Table 2.1); (ii) the *gata4* gene, which encodes a transcription factor involved in gonads development (Ssa06, at 63 Mb, region 3); (iii) the *rspo1* gene, which produces a secreted activator protein in ovary (Ssa06, at 63 Mb, region 3); and (iv) the *esr1* gene, an estrogen receptor gene which is essential for sexual development and reproductive function in females (Ssa06, at 66 Mb, close to region 3) (Figure 2.3). While genes in region 3 are annotated in the *Salmo salar* genome sequence, the sequence of the *sdY* gene is not, so we took the mRNA sequence of *sdY* from rainbow trout (annotated in NCBI) and performed a BLAST search against the *Salmo salar* sequence. We found that 34% of the *sdY* sequence aligned with 81% of identity with a region at 36 Mb in Ssa02, which is very close to one of the putative regions (31 Mb) identified in the GWAS. We also found 30-35% homology between this sequence and regions in other chromosomes (Ssa20, Ssa19 and Ssa07, in decreasing order of homology) (Figure 2.4), although these regions did not show any significant result in the GWAS.

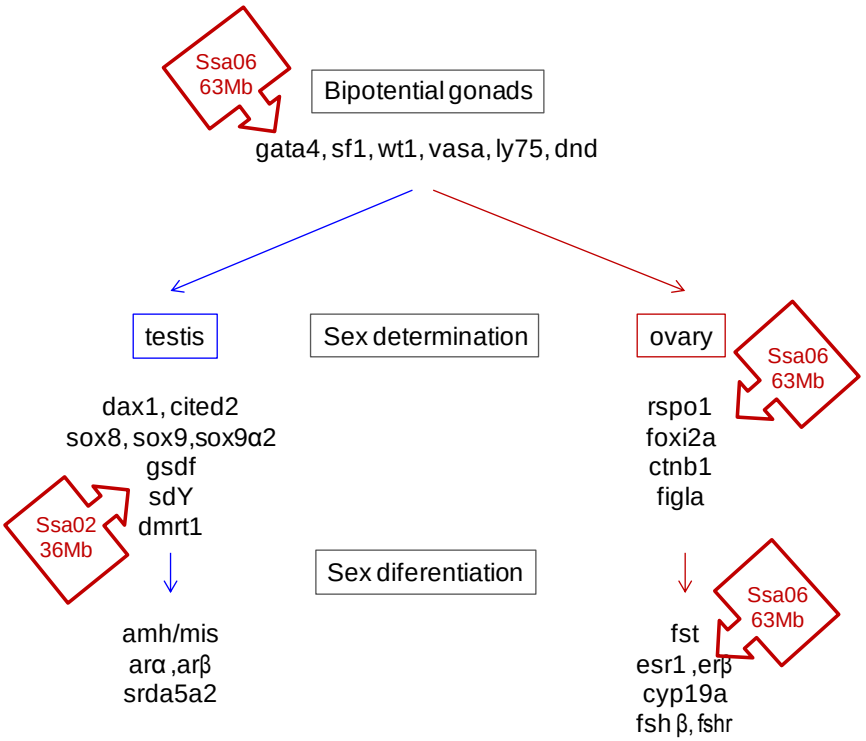


Figure 2.3. Genes identified in the GWAS and corresponding position in chromosomes involved in sex determination and sexual development routes.

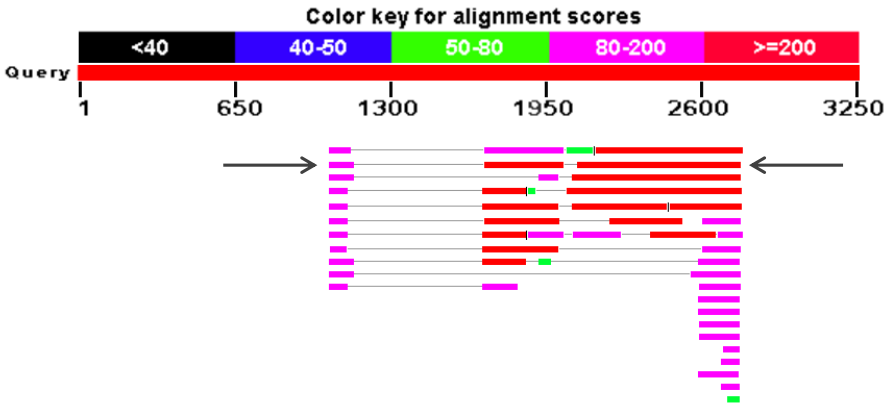


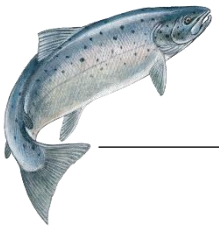
Figure 2.4. Homology of the *sdY* sequence from rainbow trout with different Atlantic salmon chromosomes (Ssa02 between arrows). Colour keys reflect alignment scores between both sequences, indicating that Ssa02 is the region showing the highest homology.

Discussion

In this study we have investigated candidate genes for sex determination in Atlantic salmon through a genome-wide association analysis using high-dense SNP information. Our results, although preliminar, are promising, as nine putative QTL regions have been identified. In salmonids, early maturation occurs differentially in males and females and is associated with reduced growth, increased disease susceptibility and worsening in organoleptic properties. Thus, in aquaculture production, females are preferred as they mature later than males (Davidson *et al*., 2009, Eisbrenner *et al*., 2014, Martínez *et al*., 2014). Also, information about sex-determining loci can help to separate males from females before they become fully developed facilitating thus the management of populations.

The discovery of the *sdY* gene, the sex-determining gene in rainbow trout (Yano *et al*., 2012), and its presence in many male salmonids although in different chromosomes (Yano *et al*., 2013) suggest that there is a common sex-determining gene that has the capability to move around the genome either by transposition or by translocation. Additional evidence for a salmonid-specific, sex-determining jumping gene came from the mapping of the sex-determining locus to three different chromosomes in Tasmanian male Atlantic salmon (Eisbrenner *et al*., 2014) and from a study by Lubiebiecki *et al*. (2015) based on BAC sequences containing *sdY*. However, these studies have some limitations because they were based on the amplification of exons from *sdY* and no new potential genes were investigated.

Here, for the first time we took benefit of a high-dense SNP chip to perform a more refined mapping of regions associated to sex determination and differentiation in Atlantic salmon. Even though the annotation of the *Salmo salar* sequence genome is scarce, there is the possibility of annotating the genome in a relatively easy way using a BLAST search, both directly (region on databases) or inversely (annotated regions from other species in the *Salmo salar* sequence). The signal detected in the genome-wide analysis supports that *sdY* is located in Ssa02, although the incipient status of the *Salmo salar* genome annotation as well the complexity of salmonids genome difficult our objective. Our results are preliminar and additional research is still ongoing, but the new genes related with sexual differentiation identified here indicate that genome-wide analysis is a useful tool for the identification of putative QTL regions associated to sex in Atlantic salmon.



General Discussion

This study made use of a high-density SNP array to explore LD patterns and to investigate the evolution of N_e across time in wild populations of Spanish Atlantic salmon, and to identify genomic regions associated to sex determination in this species. Our results revealed that LD measured by r^2, was high for markers close to each other and decayed rapidly with increasing inter marker distance. This is important because high LD levels at short distances are needed for powerful detection of SNPs associated to causal mutations. Our estimates of N_e six generations ago was three to five times lower (average N_e = 240) than 50 generations ago (average N_e = 830) and the rate of decrease in N_e after the hydroelectric development in the country was higher than in earlier times. This reflects the negative impact of such development on wild Atlantic salmon populations. Our results also showed that there was a pattern of isolation-by-distance, in agreement with gene flow between close rivers. Atlantic salmon Spanish populations seemed to be structured in two main ancestral clusters, including an Atlantic group and a Cantabric group, the former with a higher N_e. Through the genome-wide association study, nine putative QTL regions responsible of sex determination were identified on seven chromosomes.

The conclusions drawn from this study have clear practical applications, since they are not only useful for a better understanding of the dynamics of the species, but also for applying more appropriate and rational conservation and selection programmes. While our study has dealt with wild Atlantic salmon, all the techniques applied can be used in farm populations of the same species. In farmed Atlantic salmon, as well as in other aquaculture species, the control of the rate of inbreeding (or equivalently the control of N_e) is of paramount importance given that their high reproductive rate could imply the production of large populations from few sires and dams, compromising thus the viability of the populations in the long-term. In current selective breeding programmes for Atlantic salmon the control of inbreeding and N_e can be performed using pedigree data given that this information is generally recorded. A common practice in aquaculture is to rear families in separate tanks until the fish are large enough to be individually tagged. Using genomic information would not require maintaining families in separated tanks. Also, the LD method would allow to estimate N_e for a base population, which otherwise would be not possible. Knowledge on patterns of genome-wide LD is also of interest in farm populations for evaluating the power of GWAS aimed at detecting genomic regions associated to economically important traits and importantly for predicting the accuracy of genome-wide selection (Meuwissen *et al.*, 2001;

Nielsen *et al*., 2009; Sonesson and Meuwissen, 2009). The only study published so far on genome-wide LD in Atlantic salmon was carried out with farmed populations and showed similar LD patterns to those detected here (Gutierrez *et al.*, 2015), suggesting that the density of the 6.5K SNP chip captures the same levels of LD than the 220K SNP chip, with important economic implications. Additional research comparing results from different chip densities would be valuable.

Population structure results questions the supportive breeding that is currently being applied; i.e., with salmon from a single river. Stocking with salmon of the same metapopulation (Atlantic or Cantabric) might be a more appropriate alternative for increasing N_e and genetic diversity while maintaining local adaptations.

Te use of a high density SNP chip has allowed the estimation of N_e (not only recent N_e but also ancestral N_e) from LD measures with high accuracy (the standard errors were extremely low, as exposed in Chapter 1). In the past, the LD method led to inaccurate estimates of N_e because LD was computed from microsatellites, which are in much lower number within the genome. However, the LD method had some limitations: (i) it was not possible to estimate N_e in the most recent generations because r^2 between SNPs separated at the genetic distance corresponding to this time (i.e. physical distance of 5 Mb) was no longer a function of the distance, as showed by the background levels of LD; and (ii) LD is highly dependent of sample size. The latter limitation was however overcome by applying the same sample size correction as that proposed by Weir and Hill (1980) for the estimation of N_e from LD measures (i.e., $r^2 - 1/N$, where N is sample size). It would be interesting to compare the LD method used here with the sibship assignment method of Wang (2009) that infers N_e from the frequencies of a pair of sibs (sharing one or two parents) taken at random from the population. Comparative analyses of extensive simulated data and some empirical data indicate that the sibship assignment method could be more accurate and flexible than other methods such as the LD method and the temporal method (Wang, 2009).

Previous studies have used the LD approach for estimating N_e in salmon populations, although using a limited number of microsatellite markers. Ribeiro *et al*. (2008) estimated N_e of the Atlantic salmon inhabiting the river Eo using a panel of eight microsatellites and the LD method. Their estimate (around 165 individuals) is lower than the estimates obtained in

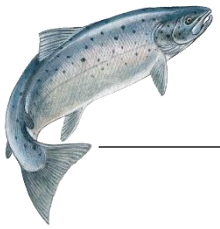
our study for this river for the same period, but their errors were much higher, implying that genome-wide LD allows now to obtain more accurate estimates of N_e that will provide valuable information both for conservation and selection purposes.

A detailed knowledge of genome-wide LD patterns within the population under study also facilities knowing the potential of GWAS for QTL mapping. Genome-wide association was first used in the analysis of human diseases and later extended to terrestrial farm species (cattle, dog, sheep, pig, horse and chicken) where dense SNP chips were available (review by Zhang *et al.*, 2012). However, until recently GWAS has not been applied to aquaculture species because of the lack of genome-wide polymorphic markers. Now that a large number of SNPs are available for a number of aquaculture species, GWAS is technically feasible. For farmed Atlantic salmon in particular, the 6.5K SNP was used to conduct GWAS for two important traits for aquaculture production: growth rate and age at sexual maturation (Gutiérrez *et al*., 2015; see also Gutiérrez *et al*., 2014).

Sex determination is also an important issue in farmed fish production. In salmonids, early maturation occurs differentially in males and females and is associated with reduced growth, increased disease susceptibility and changes in organoleptic properties. Thus, in salmon aquaculture, females are preferred as they mature latter than males. Sex determination in salmonids is still underexplored. Here, we have detected nine putative QTL regions on chromosomes Ssa02 (two regions), Ssa06 (one region), Ssa09 (one region), Ssa10 (two regions), Ssa21 (two regions) and Ssa22 (one region), that contain or are very close to annotated genes (in Atlantic salmon or close species) related to sex determination and sexual differentiation. Two of these were annotated as the *sdY* gene, the master male-specific sex-determining gene (Ssa02) and the *esr1* gene, the estrogen receptor gene, which is essential for sexual development and reproductive function in females (Ssa06). These findings validate previous results (Woram *et al*., 2003; Artieri *et al*., 2006; Yano *et al*., 2012; Eisbrener *et al*., 2014) and place GWAS as a useful tool for QTL mapping.

A consideration of multiple comparisons is an essential part for determining the statistical significance when testing thousands of SNPs in a single association study. Bonferroni adjustments are very conservative and thus, information can be lost (Johnson *et al*., 2010). False discovery rate may produce a high proportion of false positives (Tabangin *et al*., 2007),

and permutation testing is not always a viable option because it demands a heavy computation effort, and using approximations may lead to biased results. Here, a QQ plot was used to characterize the extent to which the observed distribution of the test statistic follows the expected (null) distribution after correcting by population structure. The probability threshold does not need to be fixed *a priori*. In any case, although a single GWAS is never concluding, the GWAS carried out in this thesis has given new insights on the genetic architecture of salmonids sex determination, supporting the hypothesis that there is a common sex-determining gene that has the capability to move around the genome.



References

Allan IRH, Ritter JA (1977). Salmonid terminology. *ICES Journal of Marine Science* 37: 293-299.

Allendorf FW, Thorgaard GH (1984). Tetraploidy and the evolution of salmonid fishes. In: Evolutionary genetics of fishes. Turner BJ (ed.), Plenum Press, New York, pp. 1-53.

Al-Mamun HA, Clark SA, Kwan P, Gondro C (2015). Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. *Genetics Selection Evolution* 47: 90.

Álvarez J, Antón A, Azpíroz I, Caballero P, Hervella F, De la Hoz J *et al.* (2010). Atlas de los ríos salmoneros de la península ibérica, Ekolur SLL, pp. 164.

Álvarez J, Lamuela M, Castién E (1995). Plan de recuperación del salmón Atlántico (*Salmo salar* L.) en el río Bidasoa (Navarra). Primeros resultados. In: Biología y conservación del salmón atlántico (*Salmo salar*) en los ríos de la región cantábrica. Braña F (ed.), ICONA, Madrid, pp. 163-185.

Amaral AJ, Megens HJ, Crooijmans RP, Heuven HC, Groenen MA (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 179: 569-579.

Anderson EC (2005). An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. *Genetics* 170: 955–967.

Artieri CG, Mitchell LA, Ng SH, Parisotto SE, Danzmann RG, Hoyheim B *et al*. (2006). Identification of the sex-determining locus of Atlantic salmon (*Salmo salar*) on chromosome 2. *Cytogenetic and Genome Research* 112: 152–159.

Bailey GS, Russell TM, Poulter TM, Stockwell PA (1978). Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proceedings of National Academy of Sciences USA* 75: 5575–5579.

Barson N, Aykanat T, Hindar K, Baranski M, Bolstad GH, Fiske P *et al.* (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature* 528: 405-408.

Bassil NV, Davis MT, Zhang H, Ficklin S, Mittman M, Webster T *et al*. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry Fragaria × ananassa. *BMC Genomics* 16:155.

Behnke RJ (2002). Trout and salmon of North America. Tomelleri J (ed.), Free Press, pp. 138.

Blanco G, Ramos MD, Vázquez E, Sánchez JA (2005). Assessing temporal and spatial variation in wild populations of Atlantic salmon with particular reference to Asturias (Northern Spain) rivers. *Journal of Fish Biology* 67: 169-184.

Braña F, Garrido R, Reyes-Gavilán LF, Toledo MM, Nicieza AG (1995a). Distribución del salmón Atlántico en la Península Ibérica. Localización en las cuencas fluviales y en el contexto de las comunidades de peces. In: Biología y conservación del salmón atlántico (*Salmo salar*) en los ríos de la región cantábrica. Braña F (ed.), ICONA, Madrid, pp. 13-25.

Braña F, Nicieza AG, Garrido R, Vauclin V (1995b). Caracterización de las poblaciones actuales y análisis de las tendencias de variación. In: Biología y conservación del salmón Atlántico (*Salmo salar*) en los ríos de la región Cantábrica. Braña F (ed.), ICONA, Madrid, pp. 27-66.

Brenna-Hansen S, Li J, Kent MP, Boulding EG, Dominik S, Davidson WS, Lien S (2012). Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. *BMC Genomics* 13: 432.

Butler JRA, Radford A, Riddington G, Laughton R (2009). Evaluating an ecosystem service provided by Atlantic salmon and other fish species in the river Spey, Scotland: the economic impact of recreational rod fisheries. *Fisheries Research* 96: 259–266.

Caballero A (1994). Developments in the prediction of effective population size. *Heredity* 73: 657–79.

Caballero P (2002). Programas de recuperación del salmón atlántico (*Salmo salar* L.) en los ríos Ulla, Lerez y Miño. Consellería de Medio Ambiente, II Jornadas del salmon Atlántico en la Península Ibérica, Xunta de Galicia 83–116.

Campos JL, Posada D, Morán P (2007). Introgression and genetic structure in northern Spanish Atlantic salmon (*Salmo salar* L.) populations according to mtDNA data. *Conservation Genetics* 9: 157-169.

Corbin LJ, Blott SC, Swinburne JE, Vaudin M, Bishop SC, Woolliams JA (2010). Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Animal Genetics* 41: 8-15.

Corbin LJ, Liu AYH, Bishop SC, Woolliams JA (2012). Estimation of historical effective population size using linkage disequilibria with marker data. *Journal of Animal Breeding and Genetics* 129: 257–70.

Cross TF, Ward RD (1980). Protein variation and duplicate loci in the Atlantic salmon *Salmo salar* L. *Genetic Research, Cambridge* 36: 147-165.

Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Koop BF, Hoyheim B *et al.* (2008). Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). *BMC Genomics* 9: 557.

Davidson WS, Huang TK, Fujiki K, von Schalburg KRB, Koop BF (2009). The sex-determining loci and sex chromosomes in the family Salmonidae. S*exual Development* 3: 78–87.

Davidson WS, Koop BF, Jones SJM, Iturra P, Vidal R, Maass A *et al.* (2010). Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology* 11: 403.

De Roos APW, Hayes BJ, Spelman RJ, Goddard ME (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics* 179: 1503–1512.

Di Genova A, Aravena A, Zapata L, Gonzalez M, Maass A, Iturra P (2011). SalmonDB: a bioinformatics resource for *Salmo salar* and *Oncorhynchus mykiss*. Available at http://salmondb.cmm.uchile.cl/.

Do KT, Lee JH, Lee HK, Kim J, Park KD (2014). Estimation of effective population size using single-nucleotide polymorphism (SNP) data in Jeju horse. *Journal of Animal Science and Technology* 56: 28.

Dominik S, Henshall JM , Kube PD, King H, Lien S, Kent MP, Elliott NG (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: 56-61.

Du FX, Clutter AC, Lohuis MM (2007). Characterizing linkage disequilibrium in pig populations. *International Journal of Biological Sciences* 3: 166-178.

Edea Z, Dadi H, Kim SW, Park JH, Shin GH, Dessie T, Kim KS (2014). Linkage disequilibrium and genomic scan to detect selective loci in cattle populations adapted to different ecological conditions in Ethiopia. *Journal of Animal Breeding and Genetics* 358: 366.

Eisbrenner WD, Botwright N, Cook M, Davidson EA, Dominik S *et al.* (2014). Evidence for multiple sex-determining loci in Tasmanian Atlantic salmon (*Salmo salar*). *Heredity* 113: 86–92.

Falconer DS, Mackay TFC (1996). Introduction to quantitative genetics. 4th Edition. Longman Group Ltd, Harlow, Essex.

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B *et al.* (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* 10: 220-227.

Finstad B, Jonsson N (2001). Factors influencing the yield of smolt releases in Norway. *Nordic Journal of Freshwater Research* 75: 37–55.

Frankham R (1995). Conservation genetics. *Annual Review of Genetics* 29: 305–327.

García de Leániz C, Fleming IA, Einum S, Verspoor E, Jordan WC, Consuegra S *et al.* (2007). A critical review of adaptive genetic variation in Atlantic salmon: implications for conservation. *Biological Reviews* 82: 173–211.

García de Leániz C, Hawkins T, Hay D, Martínez JJ (1987). The Atlantic salmon in Spain. Atlantic salmon trust. Moulin, Pitchlory, Scotland.

García de Leániz C, Martinez JJ (1986). The Atlantic salmon in the rivers of Spain with particular reference to Cantabria. In: Atlantic salmon: planning for the future. Mills D, Piggins D (eds.), Croom Helm, London & Sydney, pp. 179–207.

García de Leániz C, Serdio A, Consuegra S (2001). Present status of Atlantic salmon in Cantabria. In: El salmón, joya de nuestros ríos. García de Leániz C, Serdio A, Consuegra S (eds.), pp. 55-82.

Gutiérrez AP, Lubieniecki KP, Fukui S, Withler RE, Swift B, Davidson WS (2014). Detection of quantitative trait loci (QTL) related to grilsing and late sexual maturation in Atlantic salmon (*Salmo salar*). *Marine Biotechnology* 16: 103–110.

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.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* 13: 635–43.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38: 226–31.

Hoar WS (1988). The physiology of smolting salmonids. *Fish Physiology* 11: 275–343.

Høgåsen HR (1998). Physiological changes associated with the diadromous migration of salmonids. In: Canadian Special Publication of Fisheries and Aquatic Sciences. Haynes RH, (ed.), pp. 1078–1081.

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

Johnson RC, Nelson GW, Troyer LJ, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11: 724.

Kent MP, Hayes B, Xiang Q, Berg PR, Gibbs RA, Lien S (2009). Development of a 6.5K SNP-Chip for Atlantic salmon. Plant & Animal Genomes XVII Conference, San Diego, CA.

Kijas JW, Porto-Neto L, Dominik S, Reverter A, Bunch R, McCulloch R *et al.* (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Animal Genetics* 45: 754-757.

Kim ES, Kirkpatrick RW (2009). Linkage disequilibrium in the North American Holstein population. *Animal Genetics* 40: 279-288.

Klemetsen A, Amundsen PA, Dempson JB, Jonsson B, Jonsson N, O'Connell MF, Mortensen E (2003). Atlantic salmon (*Salmo salar* L.), brown trout (*Salmo trutta* L.) and Arctic charr (*Salvelinusalpinus* L.): a review of aspects of their life histories. *Ecology of Freshwater Fish* 12: 1-159.

Law R (2000). Fishing, selection and phenotypic evolution. *ICES Journal of Marine Science* 57: 659–668.

Lewontin RC (1964). The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* 49: 49-67.

Li J, Phillips RB, Harwood AS, Koop BF, Davidson WS (2011). Identification of the sex chromosomes of brown trout *Salmo trutta* and their comparison with the corresponding chromosomes in Atlantic salmon *Salmo salar* and rainbow trout *Oncorhynchus mykiss*. *Cytogenetic and Genome Research* 133: 25–33.

Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS *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.

Lien S, Koop BE, Sandve RS, Miller JR, Kent MP, Nome T *et al.* (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533: 200-205.

Lubieniecki KP, Lin S, Cabana EI, Li J, Lai YYY, Davidson WS (2015). Genomic instability of the sex-determining locus in Atlantic salmon (*Salmo salar*). G3: *Genes, Genomes, Genetics* 5: 2513-2522.

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics* 11: 355–373.

Martínez P, Viñas AM, Sánchez L, Díaz N, Ribas L, Piferrer F (2014). Genetic architecture of sex determination in fish: applications to sex ratio control in aquaculture. *Frontiers in Genetics* 5: 340.

McVean GAT (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* 162: 987-91.

Meadows JRS, Chan EKF, Kijas JW (2008). Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genetics* 9:61.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. *Genetics* 157:1819–1829.

Morán P, Pérez J, Dumas J, Beall E, García-Vázquez (2005a). Stocking-related patterns of genetic variation at enzymatic loci in south European Atlantic salmon populations. *Journal of Fish Biology* 67: 185-199.

Morán P, Pérez J, García-Vázquez E (2005b). Genetic variation in endangered populations of Atlantic salmon (*Salmo salar* L.) from Northwestern Spain. *Journal of Fish Biology* 67: 206-212.

Munárriz P, Teniente J (2004). El salmón en Navarra. Informe ANAPAM, Navarra (www.anapam.org).

Nei M, Tajima F (1981). Genetic drift and estimation of effective population size. *Genetics* 98: 625–640.

Nielsen HM, Sonesson AK, Yazdi H, Meuwissen THE (2009). Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. *Aquaculture* 289: 259–264.

Nomura T (2008). Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications* 1: 462–474.

Nsengimana J, Baret P, Haley CS, Visscher PM (2004). Linkage disequilibrium in the domesticated pig. *Genetics* 166: 1395-1404.

Odani M, Narita A, Watanabe T, YokouchiK, Sugimoto Y, Fujita T *et al*. (2006). Genome wide linkage disequilibrium in two Japanese beef cattle breeds. *Animal Genetics* 37: 139-144.

Ohta T, Kimura M (1971). Linkage disequilibrium between two segregating population. *Genetics* 68: 571-580.

Palstra FP, O'Connell MF, Ruzzante DE (2009). Age structure, changing demography and effective population size in Atlantic salmon (*Salmo salar)*. *Genetics* 182: 1233–1249.

Parrish DL, Behnke RJ, Gephard SR, McCormick SD, Reeves GH (1998). Why aren't there more Atlantic salmon (*Salmo salar*)? *Canadian Journal of Fisheries and Aquatic Sciences* 55: 281–287.

Pérez O'Brien AM, Mészáros G, Utsunomiya YT, Sonstegard TS, Garcia JF, Van Tassell CP *et al.* (2014). Linkage disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium and high density SNP chip data and different minor allele frequency distributions. *Livestock Science*. Genomics Applied to Livestock Production 166: 121–132.

Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF *et al*. (2009). Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). *BMC Genetics* 46: 10.

Porcher JP, Baglinière JL (2001). Le Saumon atlantique. Atlas des poissons d'eau douce de France. Keith P, Allardi J (eds.), *Patrimoines Naturels* 47: 240-243.

Porto-Neto LR, Kijas JW, Reverter A (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. *Genetics Selection Evolution* 46: 22.

Pudovkin AI, Zhdanova OL, Hedgecock D (2009). Sampling properties of the heterozygote-excess estimator of the effective number of breeders. *Conservation Genetics* 11: 759–771.

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

Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H (2010a). Linkage disequilibrium reveals different demographic history in egg laying chickens. *BMC Genetics* 11: 103.

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H (2010b) . The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics* 41: 346-356.

Rao YS, Liang Y, Xia NM, Shen X, Du YJ , Luo CG, Nie QO, Zeng H, Zhang XQ (2008). Extent of linkage disequilibrium in wild and domestic chicken populations. *Hereditas* 145: 251- 257.

Ribeiro A, Morán P, Caballero A (2008). Genetic diversity and effective size of the Atlantic salmon *Salmo salar* L. inhabiting the River Eo (Spain) following a stock collapse. *Journal of Fish Biology* 72: 1933-1944.

Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR (2008). Extent of linkage disequilibrium in Holstein cattle in North America. *Journal of Dairy Science* 91: 2106-2117.

Saura M, Morán P, Brotherstone S, Caballero A, Alvarez J, Villanueva B (2010). Predictions of response to selection caused by angling in a wild population of Atlantic salmon (*Salmo salar*). *Freshwater Biology* 55: 923–930.

Saura M, Tenesa A, Woolliams JA, Fernández A, Villanueva B (2015). Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: an empirical case. *BMC Genomics* 16: 922.

Sedgwick SD (1982). The salmon handbook: The life and cultivation of fishes of the salmon family. Andre Deutsch, London, pp. 247.

Sved JA (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* 125: 141.

Sonesson AK, Meuwissen THE (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution* 41:37.

Tallmon DA, Koyuk A, Luikart G, Beaumont MA (2008). One Samp: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources* 8: 299–301.

Tabangin ME, Woo JG, Liu C, Nick TG, Martin LG (2007). Comparison of false-discovery rate for genome-wide and fine mapping regions. *BMC Proceedings* 1: S148.

Thorpe JE, Mangel M, Metcalfe NB, Huntingford FA (1998). Modelling the proximate basis of salmonid life-history variation, with application to Atlantic salmon, *Salmo salar* L. *Evolutionary Ecology* 12: 581-599.

Villanueva B, Pong-Wong R, Woolliams JA, Avendaño S (2004). Managing genetic resources in commercial breeding populations. In: Farm Animal Genetic Resources. BSAS Occasional Publication No. 30, Simm G, Villanueva B, Sinclair KD, Townsend S (eds.), Nottingham University Press, Nottingham, UK, pp. 113-132.

Verspoor E, García de Leániz CG (1997). Stocking success of Scottish Atlantic salmon in two Spanish rivers. *Journal of Fish Biology* 51: 1265-1269.

Wang J (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetic Research* 78: 243–257.

Wang J (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B* 360: 1395–409.

Wang J (2009). A new method for estimating effective population size from a single sample of multilocus genotypes. *Molecular Ecology* 18: 2148–2164.

Waples R, England PR (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics* 189: 633–44.

Waples RS (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121: 379–391.

Waples RS, Do C (2008). LD Ne: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* 8: 753–756.

Waples RS, Do C (2010). Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* 3: 244–262.

Weir BS, Hill WG (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* 95: 477–488.

World Wildlife Fund (2001). The status of wild Atlantic salmon: a river by river assessment. Washington, DC.

Woram RA, Gharbi K, Sakamoto T, Hoyheim B, Holm LE, Naish K *et al*. (2003). Comparative genome analysis of the primary sex-determining locus in salmonid fishes. *Genome Research* 13: 272–280.

Yáñez JM, Naswa S, López ME, Bassini L, Cabrejos ME, Gilbey J *et al*. (2014). Development of a 200K SNP Array for Atlantic salmon: Exploiting across continents genetic variation. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production.Vancouver, Bc, Canada, August 17-22.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al*. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42:565–569.

Yano A, Guyomard R, Nicol B, Jouanno E, Quillet E, Klopp C *et al*. (2012). An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*. *Current Biology* 22:1423–1428

Yano A, Nicol B, Jouanno E, Quillet E, Fostier A, Guyomard R *et al*. (2013).The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evolutionary Applications* 6: 486–496.

Zanella R, Peixoto JO, Cardoso FF, Cardoso LL, Biegelmeyer P, Cantao ME *et al.* (2016). Genetic diversity analysis of two commercial breeds of pigs using genomic and pedigree data. *Genetics Selection Evolution* 28: 24.

Zhang H, Wang Z, Wang S, Li H (2012). Progress of genome wide association study in domestic animals. *Journal of Animal Science and Biotechnology* 3: 26

Zhou H, Alexander D, Lange K (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing* 21 : 261-273.

Zhu M, Zhu B, Wang YH, Wu Y, Xu L, Guo LP *et al.* (2013). Linkage disequilibrium estimation of Chinese beef Simmental cattle using high-density SNP panels. *Journal of Animal Science* 26: 772-779.