MÁSTER INTERUNIVERSITARIO EN MEJORA GENÉTICA ANIMAL Y BIOTECNOLOGÍA DE LA REPRODUCCIÓN

Genomic coancestry and inbreeding in a farmed population of turbot (*Scophthalmus maximus*)

Tesis de Máster

Valencia, Julio 2018

Elisabeth Morales González

Directoras

Prof. Beatriz Villanueva Gaviña

Dra. María Saura Álvarez

Acknowledgments

First of all I would like to thank my supervisors, Prof. Beatriz Villanueva and Dr. María Saura, for the continuous support during the development of this thesis, for their patience, motivation and for all the things they taught me during this year. Their guide helped me at all times during the research and writing of this Master thesis.

I would also like to thank all the members of the Animal Breeding Department of INIA who gave me the opportunity to join their team and use their research facilities. In particular, I thank Drs. Jesús Fernández, Almudena Fernández for their help and advice.

I am grateful to my office mates Alejandro Saborío, Beatriz Delgado and Latifa Ouatahar, for their friendship, the shared hours and for enduring that I could not be more than two hours in silence, and to Silvia García-Ballesteros, Gabriel Molina, Adrián López and Oscar González for the moments of coffee, laughter and advice.

I would like to thank Prof. Agustín Blasco for the good organization of the Master in Animal Breeding and Reproduction Biotechnology, and also to all the teachers who collaborated in this Master for all the knowledge they shared. I also want to thank the classmates of the Master for the hours helping each other and sharing knowledge and notes.

And to my parents, for being there in all the important moments of my life, for the unconditional support and love that they always gave me.

I thank Dr. Santiago Cabaleiro (CETGA) and Prof. Paulino Martínez and collaborators, (University of Santiago de Compostela) for providing the data analysed and the physical map used in this thesis. I thank also Prof. Miguel Ángel Toro (UPM) for useful discussions. The research leading to these results has received funding from the European Union's Seventh Framework Programme (KBBE.2013.1.2- 10) under grant agreement n° 613611.

Communications to conferences

The results obtained in this thesis have led to two communications in two different scientific congresses:

Morales-González, E., Saura, M., Fernández, A., Fernández, J., Cabaleiro, S., Martinez, P. and Villanueva, B. 2018. Efficiency of different genomic coancestry matrices to maximize genetic variability in turbot selective breeding programs. 69th Annual Meeting of the European Federation of Animal Science (EAAP), Dubrovnik, Croatia, August 27-31. Accepted for oral presentation.

Morales-González, E., Saura, M., Fernández, A., Fernández, J., Cabaleiro, S., Martinez, P. and Villanueva, B. 2018. Evaluation of different genomic coancestry matrices to maintain genetic variability in a turbot selected population. XIX Reunión Nacional de Mejora Animal, León, Spain, June 14-15. Oral presentation.

Also, during the development of my thesis I have contributed to other two communications in scientific congresses:

Villanueva, B., Saura, M., Caballero, A., Santiago, E., **Morales, E.,** Fernández, A., Fernández, J., Cabaleiro, S., Martínez, P., Millán, A., Palaiokostas, C., Kocour, M., Houston, R., Prchal, M., Bargelloni, L., Kostas, T. 2018. The importance of ensuring genetic variability when establishing selection programmes in aquaculture. AQUA 2018 (World Aquaculture Society), Montpellier, France, August 25-29.

Saura, M., Caballero, A., Santiago, E., **Morales, E.,** Fernández, A., Fernández, J., Cabaleiro, S., Martínez, P., Millán, A., Palaiokostas, C., Kocour, M., Houston, R., Prchal, M., Bargelloni, L., Kostas, T. and Villanueva, B. 2018. The importance of ensuring genetic variability when establishing selection programmes in aquaculture. XIX Reunión Nacional de Mejora Animal, León, Spain, June 14-15. Oral presentation.

CONTENTS

RESUMEN

El rodaballo (*Scophthalmus maximus*) es un pez plano de gran valor comercial y para el cual actualmente existen varios programas de mejora genética, cuyo principal objetivo es aumentar el crecimiento. Estos programas tienen un gran potencial, dado que la alta capacidad reproductiva de los peces permite aplicar una selección intensa y, por tanto, obtener altas respuestas a la selección. Sin embargo, esto también puede conducir a altas tasas de consanguinidad, con la consiguiente reducción de la variabilidad genética y un mayor riesgo de depresión consanguínea. Por lo tanto, el control de la consanguinidad y de la pérdida de variabilidad genética en los programas de selección del rodaballo, es fundamental para garantizar su sostenibilidad. Sin embargo, hasta la fecha, no existen estudios que evalúen los niveles de consanguinidad en poblaciones comerciales de esta especie. El método más eficiente para controlar la pérdida de variabilidad genética y el aumento de la consanguinidad es el método de Contribuciones Óptimas (CO). Este método se basa en la optimización de las contribuciones de los candidatos para minimizar el parentesco de los individuos seleccionados y así minimizar la consanguinidad esperada en la siguiente generación. Por lo tanto, el elemento central del método de CO es la matriz de parentesco. Tradicionalmente, esta matriz se ha obtenido a partir de registros genealógicos, pero el desarrollo reciente de herramientas genómicas permite estimarla con mayor precisión. De hecho, se han desarrollado distintas medidas de parentesco (y consanguinidad) genómico, pero se desconoce su eficacia relativa a la hora de mantener la variabilidad genética. Los objetivos de este estudio fueron: i) obtener estimas de parentesco y consanguinidad genómicas en una población cultivada de rodaballo; ii) evaluar la eficiencia de diferentes matrices de parentesco para retener variabilidad genética cuando se utiliza CO; y iii) determinar los patrones de consanguinidad genómica en el genoma del rodaballo. Se utilizaron genotipos de 18,097 SNPs disponibles para 1,391 peces pertenecientes a 36 familias de hermanos completos. Se compararon seis matrices de parentesco genómico: i) θ_{SIM}, basada en la proporción de alelos compartidos entre dos individuos; ii) $\theta_{L\&H}$, basada en el exceso de homocigosis; iii) θ_{VR1}, basada en la matriz de relaciones genómicas de VanRaden (método 1); iv) θ_{VR2}, basada en la matriz de relaciones genómicas de VanRaden (método 2); v) θ_{YANG}, basada en la matriz de relaciones genómicas de Yang; y vi) θ_{SEG}, basada en la proporción de segmentos idénticos por descendencia compartidos por dos individuos. La magnitud de los diferentes coeficientes de parentesco difirió considerablemente (el parentesco promedio fue 0.78 para θ_{SIM}, 0.00 para $\theta_{L\&H}$, θ_{VR1} y θ_{VR2} y θ_{YANG} y 0.13 para θ_{SEG}) y la

causa de estas diferencias se debe fundamentalmente al momento en el cual se estableció la población de referencia. Sin embargo, las correlaciones entre los distintos coeficientes de parentesco fueron altas (≥ 0.7). Todos los coeficientes resultaron tener capacidad para discriminar diferentes grados de relación (padres-hijos, hermanos completos, medios hermanos y peces menos relacionados), y f_{VR} y f_{YANG} fueron los que más se aproximaron a los valores esperados de parentesco obtenidos a partir de registros genealógicos. La variabilidad genética retenida en los candidatos seleccionados cuando se utilizaron las diferentes matrices de parentesco en CO fue similar en términos de heterocigosis esperada y porcentaje de alelos segregantes (> 99% en todos los casos). Sin embargo, esto se logró seleccionando un número muy diferente de individuos en cada caso. En particular, con $\theta_{L\&H}$ y θ_{SEG} solo se seleccionaron el 9% y el 13% del número inicial de candidatos, respectivamente, en comparación con el 47-85% de los candidatos seleccionados cuando se utilizaron θ_{SIM}, θ_{VR1}, θ_{VR2} y θ_{YANG}. Estas diferencias pueden explicarse por el hecho de que $f_{L\&H}$ and f_{SEG} fueron los coeficientes que presentaron las varianzas más altas en el grupo de individuos 'menos relacionados' (el grupo más numeroso). Por el contrario, θ_{SIM}, que fue la matriz que condujo al mayor número de candidatos seleccionados, también fue la matriz con menor capacidad para diferenciar relaciones entre individuos.

Palabras clave: consanguinidad, Contribuciones Óptimas, parentesco, rodaballo**.**

ABSTRACT

Turbot (*Scophthalmus maximus*) is a flatfish of great commercial value for which several genetic breeding programmes are currently underway whose main objective is increasing growth rate. These programmes have a great potential given that the high fecundity of fish allows to apply high selection intensities and therefore, to obtain high selection responses. However, this can also lead to high rates of inbreeding, with the consequent reduction of genetic variability and increased risk of inbreeding depression. Thus, the control of inbreeding and the loss of genetic variability in turbot selection programmes, is fundamental to ensure their sustainability. However, to date, there are no studies evaluating inbreeding levels in commercial populations of this species. The most efficient method to control the loss of genetic variability and the increase of inbreeding is the Optimal Contributions method (OC). This method is based on optimising the contributions of candidates to minimise group coancestry of the selected breeders and thus minimise the expected inbreeding in the next generation. Therefore, the central element of the OC method is the coancestry matrix. Traditionally, this has been computed from pedigree data but the recent development of genomic tools allows to use genome-wide information to estimate it with higher precision. In fact, several measures of genomic coancestry (and inbreeding) have been developed but their relative efficiency for maintaining genetic variability is unknown. The aims of this study were to i) obtain estimates of genome-wide coancestry and inbreeding coefficients in a turbot commercial population; ii) evaluate the efficiency of different genome-wide coancestry matrices in retaining genetic variability when using OC; and iii) determine the patterns of genomic inbreeding across the turbot genome. Genotypes of 18,097 SNPs were available for 1,391 offspring belonging to 36 full-sib families. Six different genome-wide coancestry matrices were compared: i) θ_{SIM}, based on the proportion of alleles shared by two individuals; ii) $\theta_{L\&H}$, based on the excess of SNP homozygosity; iii) θ_{VR1}, based on the genomic relationship matrix of VanRaden (method 1); iv) θ_{VR2}, based on the genomic relationship matrix of VanRaden (method 2); v) θ_{YANG}, based on the genomic relationship matrix of Yang; and vi) θ_{SEG}, based on the proportion of identity by descent segments shared by two individuals. The magnitude of the different coancestry coefficients differed greatly (average coancestry was 0.78 for θ_{SIM}, 0.00 for $\theta_{L\&H}$, θ_{VR1} and θ_{VR2} and θ_{YANG} and 0.13 for θ_{SEG}) and that was mainly due to differences in the time where base populations were set. However, the correlations between them were high (≥ 0.7). All coefficients had a good ability to discriminate different degree of relationships (parent-

offspring, full-sibs, half-sibs and less related fish), and f_{VR} and f_{YANG} were those that more approximated to the expected values derived from pedigree data. The genetic variability retained in the selected candidates when using the different coancestry matrices in OC was similar in terms of expected heterozygosity and percentage of alleles that remained segregating (> 99 % in all cases). However, this was achieved by selecting very different number of individuals. In particular, with $\theta_{L\&H}$ and θ_{SEG} only 9% and 13% of the initial number of candidates were selected, respectively, in comparison with 47-85% of candidates selected when using θ_{SIM}, θ_{VR1}, θ_{VR2} and θ_{YANG}. These differences can be explained by the fact that $f_{L\&H}$ and f_{SEG} were the coefficients presenting the highest variances for the group of 'less related' individuals (the most numerous group). Conversely, θ_{SIM}, which led to the highest number of candidates selected, was also the matrix with lower ability to differentiate relationships.

Keywords: coancestry, inbreeding, Optimal Contributions, turbot.

INTRODUCTION

Fisheries and aquaculture are important sources of food, nutrition, income and livelihoods for millions of people around the world. Fish represents one of the most commercialized food products and thanks to the intense growth of aquaculture, the world fish supply reached a historical maximum of 20 kg per capita in 2014 (FAO, 2016). In fact, while capture fishery production has been relatively static since the late 1980s, the percentage of fish produced for human consumption from aquaculture increased from 7% in 1974 to 26% in 1994 and to 39% in 2004. Currently, aquaculture provides half of all fish for human consumption and represents thus a promising option for nutrition and food security of a world growing population that is expected to reach 9.7 billion people by 2050 (FAO, 2016).

A major challenge and opportunity for increasing competitiveness of the aquaculture industry comes from the application of efficient breeding programmes. When compared with livestock species, genetic breeding programmes in aquaculture species are relatively recent and scarce. In fact, only about 10% of the current global aquaculture production is based on genetically improved stocks. This is despite the fact that the potential of selection programmes in aquaculture species is considerable, given the high reproductive capacity of these species that permit applying high selection intensities (a reduced number of breeders is enough for producing a whole generation) and therefore high genetic gains (Gjerde *et al*., 1996). For a key trait such as growth, genetic gains per generation from selection reported in the literature average 13% and show that growth rate could be doubled in six generations of selection (Gjedrem and Rye, 2016). Successful examples include the Atlantic salmon national selective breeding programme initiated by AKVAFORSK in Norway at the beginning of the 1970s (reviewed by Gjedrem, 2010) and the international collaborative project to improve the genetic performance of farmed Nile tilapia (Gjedrem, 2012), a project commonly known as the Genetic Improvement of Farmed Tilapias (GIFT). It has been documented that genetically improved salmon from the Norwegian breeding programme grow twice as fast as wild Atlantic salmon and require 25 % less feed (Thodesen and Gjedrem, 2006). The GIFT project showed that five generations of traditional selection based on a synthetic farmed population of Nile tilapia gradually increased the body weight at harvest by 67–88% (Bentsen *et al*., 2017).

Although the reproductive potential of fish species can lead to high selection responses, it can also lead to high rates at which inbreeding increases if a small number

of individuals contribute largely to subsequent generations. Increases in the rate of inbreeding (ΔF) would have negative consequences such as reductions of genetic variability and increased risk of inbreeding depression. The risk of high inbreeding can be even more important when using selection methods that do not make use of family information; i.e., mass selection (Villanueva *et al*., 1996). In fact, the magnitude of estimates of recent effective population size (N_e, a parameter directly related to ΔF given that $N_e = 1/2\Delta F$) found in commercial fish populations (Su *et al*., 1996; Pante *et al*., 2001; Gallardo *et al*., 2004; Brown *et al*., 2005; Yáñez *et al*., 2014) is generally lower than the critical value of 50 individuals recommended to avoid inbreeding depression and retain fitness in the short-term (Frankham *et al*., 2002). Thus, the control of the rate of inbreeding, necessary in any selection programme, gains in importance in aquaculture species.

The coefficient of inbreeding of an individual (F) is defined as the probability that the two alleles of a locus taken at random from the same individual are identical by descent (IBD); i.e., both are copies of an allele carried by a common ancestor of the parents of the individual (Malécot, 1948). It expresses the degree of relationship between the individual's parents. Thus, another coefficient that is closely linked with F is the coancestry coefficient between two individuals (f) that is defined as the probability that two alleles at a given locus taken at random from two different individuals are IBD. By definition, both F and f refer to a base population that is assumed to be composed by non-inbred and unrelated individuals. The rate at which f increases per generation (i.e., the rate of coancestry or Δf) has a direct relationship with N_e given that $N_e = 1/2\Delta f$ (Falconer and Mackay, 1996). Alternatively, as mentioned above $N_e = 1/2\Delta F$. The value of F is subject to individual mating decisions (i.e., whether or not mattings between relatives are avoided) and then genetic variability is better given by f. In any case, both measures of N_e are equivalent with random mating (Caballero and Toro, 2000) or with non-random mating if the level of non-randomness is constant across generations (Villanueva *et al*., 2010). Also, inbreeding depression depends on F, not on f. Thus, both f and F are very important for managing populations.

Coancestry and inbreeding coefficients are directly related with the expected (H_E) and the observed (H_O) heterozygosity (Toro *et al*., 2009), which are commonly used to

quantify genetic diversity. The relationship of f with H_E (expected heterozygosity under Hardy-Weinberg equilibrium) is given by $H_E = 1 - f$, and the relationship of F with H_O (the number of heterozygous individuals divided by the sample size) is given by $H_O = 1 - F$ (Frankham *et al*., 2002).

The most efficient method to control the loss of genetic variability and the increase of coancestry and inbreeding is the Optimal Contributions method (OC). The method was developed in the 90s in the context of genetic breeding programmes and provides optimal contributions for all breeding candidates for maximising gains obtained through selection while restricting at the same time the increase in coancestry and inbreeding (Meuwissen, 1997; Grundy *et al*., 1998; Woolliams *et al*., 2015). However, the application of OC for simply minimising coancestry is straightforward (Fernández *et al*., 2003; Villanueva *et al.*, 2004). The objective function can be modified to accommodate the conservation aim that would be to minimise rates of coancestry and inbreeding. In this scenario, the method optimises the contributions of candidates to minimise the group coancestry of the selected breeders and thus minimise inbreeding in the next generation. Therefore, the central element of OC is the coancestry matrix ($\boldsymbol{\theta}$) or equivalently, the additive genetic relationship matrix ($\mathbf{A}$), as $\mathbf{A} = 2\boldsymbol{\theta}$.

Traditionally, f and F have been estimated from pedigree data. However, the new techniques of large scale genotyping have allowed the detection of a large number of single nucleotide polymorphisms (SNPs) that can be used to obtain genomic estimates of these coefficients with a higher degree of accuracy. In particular, next-generation sequencing (NGS) has greatly reduced the cost of nucleic acid sequencing, and therefore also genetic marker discovery. This has opened new opportunities for rapid generation of genome-wide genetic marker datasets, either through SNP arrays or through genotyping by sequencing (GBS) techniques (Davey *et al*., 2011). GBS is based on Restriction-Site Associated DNA Sequencing (RAD sequencing or RAD-Seq) that combines the use of genome complexity reduction with Restriction Enzymes (REs) and the high sequencing output of NGS technologies. RAD-Seq was first described by Baird *et al*. (2008). This technique is widely used in fish because the low economic value of individual fish requires a trade-off between this value and genotyping costs. In addition, this technique has the advantage that does not need previous information about the SNPs (a

characteristic very useful for non-model species), in contrast to SNP arrays. Depending on the number and type of REs, different RAD-Seq variants are available, including 2b-RAD-Seq. Briefly, this method consists on using REs that cut the genomic DNA at both sides of the recognition site at a fixed distance, producing short genomic DNA fragments of identical size (33–36 bp). These fragments are subsequently sequenced on next-generation platforms (Wang *et al*., 2012; Robledo *et al*., 2017). In comparison with other RAD-based techniques, an advantage of 2b-RAD is that it facilitates the sampling and sequencing of identical sites across individuals. An illustration of the RAD-Seq method is given in Figure 1.

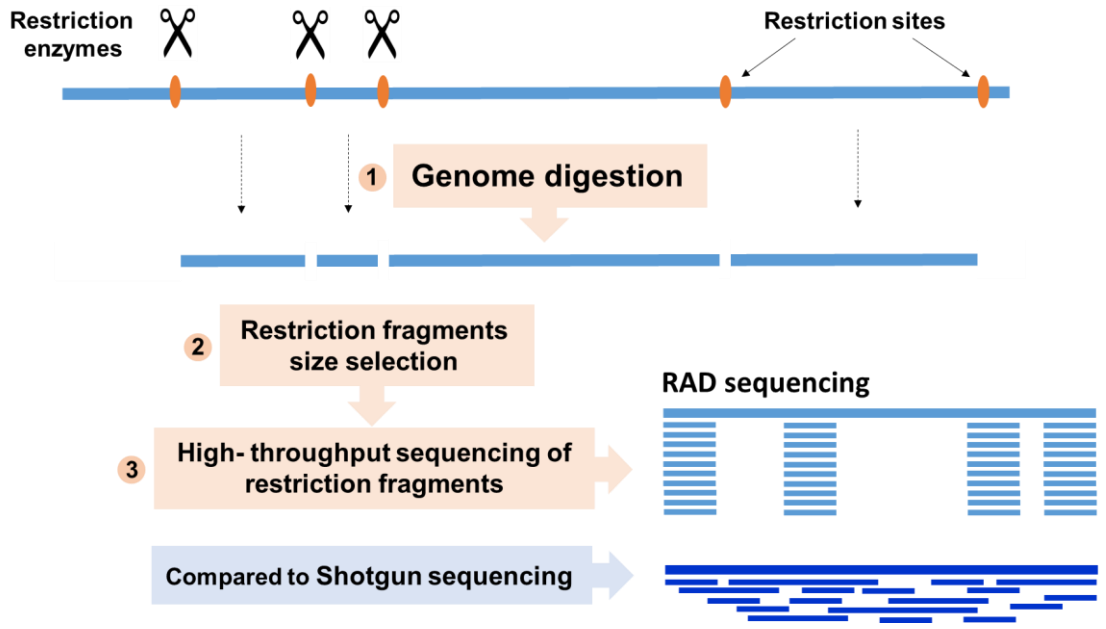

Figure 1. Graphical illustration of the RAD sequencing method. Results of the distribution and coverage of the fragments sequenced with RAD compared to Shotgun sequencing is also represented. This comparison evidences a reduced representation of the genome (with higher coverage) in the case of RAD sequencing.

Genomic coefficients measure the proportion of loci that two particular individuals have in common (*f*) or the proportion of homozygous genes (*F*) directly while pedigree-based coefficients give only expectations of these proportions that can differ from the exact proportions. In fact, using genome-wide coefficients in OC has been proved to lead to higher diversity maintained than pedigree-based coefficients (de Cara *et al*., 2011; Gómez-Romano *et al*., 2013). Other advantages of using molecular measures are that they permit to investigate patterns of *f* and *F* across the genome (Kleinman-Ruiz *et al*., 2016) and to incorporate similarity or autozygosity arising from very distant

common ancestors (Keller *et al*., 2011). A very useful application of genomic *F* is that it permits the detection of genomic regions responsible of inbreeding depression, which is not possible with pedigree-based *F* (Pryce *et al*., 2014; Saura *et al*., 2015). Finally, genomic *f* and *F* can be estimated in populations where pedigree recording is difficult or impossible.

Several measures of genomic coancestry (and inbreeding) have been developed but their relative efficiency when used in OC for maintaining genetic variability is unknown. The simplest measures are based on the proportion of alleles shared between two individuals (coancestry) or the proportion of homozygous genotypes in the individual (inbreeding). A second measure of coancestry (and inbreeding) is based on the deviations of the observed number of alleles shared between two individuals (or homozygous genotypes within an individual) from the expected numbers under Hardy-Weinberg equilibrium. The former attempts to correct for the homozygosity present in the base population (Toro *et al*., 2002, 2014). This second measure of genomic *F* was first proposed by Li and Horvitz (1953). Other measures of genomic *f* and *F* can be obtained from the different genomic relationship matrices (**G**) proposed by VanRaden (2008) and by Yang (2010) taking into account the fact that $\boldsymbol{\theta} = ½\mathbf{G}$. Finally, *f* and *F* can be obtained by considering segments rather than single points in the genome. Thus, genomic *F* can be obtained from runs of homozygosity or ROH (McQuillan *et al*., 2008), defined as long segments of consecutive homozygotes SNPs, and genomic *f* can be obtained from IBD segments, defined as segments of DNA that are found to be identical in two individuals (Gusev *et al*., 2009; de Cara *et al*., 2013; Gómez-Romano *et al*., 2016). An advantage of the segment-based inbreeding is that it allows to differentiate old from recent inbreeding according to the length of the segments (Keller *et al*., 2011; Pemberton *et al*., 2012; Bjelland *et al*., 2013,). Short ROH reflect old inbreeding while long ROH reflect inbreeding of more recent origin (Gusev *et al*., 2009).

Within aquaculture species, turbot (*Scophthalmus maximus*) is a flatfish of great commercial value. It belongs to the family *Scophthalmidae*, within the order *Pleuronectiformes* and displays a benthic lifestyle. It is naturally distributed along the entire European coast and Northwest Africa, including the Baltic and the Black Seas, as well as the Eastern and North-eastern part of the continental shelf of the Atlantic Ocean, reaching North Africa. Some studies conducted on wild populations show a fairly

homogeneous genetic structure with some local adaptations, as a result of gene flow among populations (Nielsen *et al.*, 2004).

Given its high commercial value, turbot was traditionally the target of extractive fishing. However, today captive breeding production almost doubles fishing production. Turbot aquaculture started in Europe in the 70´s but in recent years, it has spread to other continents (Asia and America) from fry imported from Europe. Currently, Spain is the leading European producer.

Nowadays, there are genetic breeding programmes well organized for turbot, with important technological facilities that help the management. In particular, there are three turbot breeding programmes in Europe, whose main objective is to increase growth rate (Janssen *et al*., 2017). However, the magnitude of preliminary estimates of recent N_e in commercial turbot populations are low (< 50 fish), probably due to important bottlenecks occurring when domestication started (Saura *et al*., 2018).

When compared with the genome of other vertebrates, the genome of turbot is relatively small. Its size is approximately 600 Mb (Figueras *et al*., 2016; Martínez *et al*., 2016), it is organized in 22 pairs of chromosomes and does not show chromosomal heteromorphism associated with sex (Bouza *et al*., 1994). In recent years, much progress has been made in the development of genomic tools for this species (Martínez *et al*., 2016) which have led to the development of genetic maps and to the sequencing and assembly of the complete genome (Figueras *et al*., 2016, Maroso *et al*., 2018).

The first genetic map of turbot included 242 microsatellites grouped into 26 linkage groups (LGs) (Bouza *et al*., 2007). Subsequent maps used a greater number of markers that were grouped in 24 LGs (Vera *et al*., 2011; 2013; Navajas-Pérez *et al*., 2012). By integrating all previous available information, Hermida *et al*. (2013) developed a consensus map that grouped the markers in 22 LGs, a number that corresponds to the number of chromosomes of this species. Wang *et al*. (2015) made use of massive sequencing techniques and constructed a high density consensus genetic linkage map using 6,647 SNPs which were also assigned to 22 LGs. These genetic maps have been used to identify loci affecting quantitative traits (QTLs) of interest including growth (Sánchez-Molano *et al*., 2011), resistance to diseases caused by bacteria (Millán *et al*.,

2011; Rodríguez-Ramilo *et al*., 2011), viruses (Rodríguez-Ramilo *et al*., 2014; Pereiro *et al*., 2016) or parasites (Pardo *et al.*, 2012; Rodríguez-Ramilo *et al*., 2013) as well as to identify candidate genes associated with sex (Martínez *et al*., 2009).

As mentioned above, the turbot genome has been sequenced and annotated recently (Figueras *et al*., 2016). In fact, turbot is the first vertebrate genetically sequenced in Spain and the assembled genome is one of the highest quality among aquaculture species (Martínez *et al*., 2016). Also, approximately 22,751 genes have been identified (Figueras *et al*., 2016). The complete sequencing of the genome of turbot represents a milestone both to understand the origin and diversification of flatfish and to investigate optimal designs of genetic selection programmes with proper control of inbreeding and therefore, loss of genetic variability in farm populations. Very recently, Maroso *et al*. (2018) have refined the turbot physical and genetic maps and improved the anchoring of the genome assembly from 80 to 97%. The consensus map comprehended 8,532 cM and 22 LGs averaging 387.9 cM (range 282.9cM – 588.7cM). After filtering, a set of 18,214 SNPs was identified that can be very useful for obtaining genomic measures of f and F that can be deployed for managing efficiently the current breeding programmes.

OBJECTIVES

The objectives of this thesis were:

1. Estimate coefficients of coancestry and inbreeding in a farmed population of turbot.

2. Evaluate the use of different matrices of genomic coancestry to maximize genetic variability using the Optimal Contributions method.

3. Determine the patterns of genomic inbreeding across the turbot genome.

MATERIAL AND METHODS

Animals

Data used in this study came from a challenge experiment carried out under the European project FISHBOOST (http://www.fishboost.eu/), currently under development. The general objective of the project is to increase the efficiency and profitability of European aquaculture by advancing selective breeding to the next level for each of the six main finfish species, including turbot. One of the main objectives is to improve disease resistance and, specifically, to determine the genetic basis of the major components of host response to pathogens that produces infectious diseases. To this end, CETGA (Aquaculture Cluster of Galicia) carried out a transmission experiment in turbot infected with the parasite *Philasterides dicentrarchi*. This ciliated protozoan causes scuticociliatosis, a disease that results in severe economic losses in the aquaculture industry.

The experiment included 1,440 fish challenged by cohabitation. These fish belonged to 36 full-sib families (i.e., 40 fish per family) created from 23 sires and 23 dams. The resulting full-sib families included 12 paternal half-sib families (11 males were mated with 1 female, 11 males with 2 females and 1 male with 3 females) and 11 maternal half-sib families (12 females were mated with 1 male, 9 female with 2 males and 2 females with 3 males). A full description of the experimental design is given in Anacleto *et al*. (2018).

The CETGA's broodstock used to create the 36 families were unrelated and came from different genetic selection programmes (Cabaleiro, 2017; personal communication) that had the same wild origin. In fact, the broodstock of CETGA constitutes a population representative of turbot from the Atlantic area (Maroso *et al*., 2018). Selective breeding has been practised for three to five generations (Janssen *et al*., 2017) and the main objective has been to increase growth.

Genotypic data

Genome-wide SNP data were available for all 1,440 fish and were provided by the Universidad de Santiago de Compostela. Genotyped fish included 1,394 offspring from the 36 full-sib families used in the experiment and their parents (the 23 sires and 23 dams). Genotypes were obtained using a 2b-RAD approach as described in Maroso *et al*. (2018). Briefly, after mapping to the reference genome of the turbot (Figueras *et al*., 2016)

and applying quality filters (Maroso *et al.*, 2018) an initial set of 25,511 SNPs was obtained. From them, only those present in 80% of parents and with a minimum coverage of 10x were retained. This set of SNPs was used as a reference to obtain the SNPs in the offspring. Markers showing Mendelian errors (offspring genotype being inconsistent with Mendelian transmission, given the parental genotypes), unmapped SNPs and those with MAF < 0.015 in the parental population and with extreme departures of Hardy-Weinberg equilibrium ($P < 0.001$) were removed. Also, for tags containing multiple polymorphisms only one SNP was retained. After quality control a total of 18,125 SNPs were retained.

Imputation

In order to infer haplotype phases (necessary to estimate the coefficient of coancestry based on shared DNA segments as indicated below) and impute missing genotypes, the software BEAGLE 4.1 was used (Browning and Browning, 2007; 2011). The software was run with the default parameters and taking into account the population structure (parents and offspring). This software performs both tasks (infer haplotype phases and impute missing genotypes) in a unified framework. Only imputed genotypes that had a high reliability (> 90%) were included in the analysis. Those that did not reach this threshold were definitively considered as missing data. Finally, samples with a call rate (number of called SNPs per sample over the total number of SNPs in the dataset) lower than 0.95 and SNPs with a call rate (number of called individuals per sample over the total number of individuals in the dataset) lower than 0.90 were excluded. After this imputation and filtering step, 1,437 individuals and 18,097 SNPs were available to carry out the analyses.

Genomic coancestry and inbreeding coefficients

Six different genome-wide coancestry and inbreeding coefficients were compared. They are described below.

1. f_{SIM}: SNP-by-SNP similarity between two individuals; i.e., the proportion of alleles shared by two individuals. Specifically, the coancestry coefficient between individuals *i* and *j* ($f_{SIM(i,j)}$) was computed as

$$f_{SIM\ (i,j)} = \frac{\frac{1}{S}\sum_{k=1}^{S}\left[\left(\sum_{m=1}^{2}\sum_{h=1}^{2} I_{km(i)h(j)}\right)\right]}{4}$$

where *S* is the number of SNPs for which individuals *i* and *j* had genotype and $I_{km(i)h(j)}$ is the identity of the m^{th} allele of individual *i* with the h^{th} allele of individual *j* for SNP *k* and takes the value of 1 if both alleles are identical and zero if they are not (Gómez-Romano *et al*., 2013).

2. $f_{L\&H}$: Coancestry coefficient based on the deviation of SNP homozygosity, computed as

$$f_{L\&H(i,j)} = \frac{\sum_{k=1}^{S} f_{SIM(i,j)k} - S + 2\sum_{k=1}^{S} p_k(1-p_k)}{2\sum_{k=1}^{S} p_k(1-p_k)}$$

where p_k is the allelic frequency of SNP *k* (Li and Horvitz, 1953). Note that $f_{SIM(i,j)}$ is the observed coancestry and that $S + 2\sum_{k=1}^{S} p_k(1-p_k)$ is the expected homozygosity in the base population.

3. f_{VR1}: Coancestry coefficient computed according to the first method of VanRaden (2008) for creating a genomic relationship matrix. The genomic relationship between individuals *i* and *j* was computed as

$$g_{(i,j)} = \frac{\sum_{k=1}^{S}(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2\sum_{k=1}^{S} p_k(1-p_k)}$$

where x_{ki} is the genotype of individual *i* for SNP *k* that was coded as 0, 1 or 2 for genotypes AA, AB and BB, respectively and p_k is the frequency of the allele of SNP *k* whose homozygote genotype is coded as 2. The coancestry coefficient between individuals *i* and *j* ($f_{VR1(i,j)}$) was obtained as $g_{(i,j)}/2$.

4. f_{VR2}: Coancestry coefficient computed according to the second method of VanRaden (2008) for creating a genomic relationship matrix. The genomic relationship between individuals *i* and *j* was computed as

$$g_{(i,j)} = \frac{1}{S}\sum_{k=1}^{S} \frac{(x_{ki} - 2p_k)(x_{kj} - 2p_k)}{2p_k(1-p_k)} \quad (1)$$

As with the previous method, the coancestry coefficient between individuals *i* and *j* ($f_{VR2(i,j)}$) was obtained as $g_{(i,j)}/2$.

5. f_{YANG}: Coancestry coefficient computed according to the method of Yang (2010) that also aims at creating a genomic relationship matrix. In this case, off-diagonal elements of the genomic relationship matrix are computed as in VanRaden's

second method (i.e., Equation 1), while diagonal elements are computed by considering that self-relationships are expected to be equal to 1 plus inbreeding:

$$g_{(j,j)} = 1 + \frac{1}{S}\sum_{k=1}^{S} \frac{x_{kj}^2 - (1+2p_k)x_{kj} + 2p_k^2}{2p_k(1-p_k)}$$

As with the previous methods, the coancestry coefficient between individuals *i* and *j* ($f_{YANG(i,j)}$) was obtained as $g_{(i,j)}/2$.

6. f_{SEG}: Coancestry coefficients based on IBD segments. In particular, the coancestry between individuals *i* and *j* ($f_{SEG(i,j)}$) was defined as the proportion of IBD segments shared by both individuals (both have identical SNP in the segment). Specifically,

$$f_{SEG(i,j)} = \frac{\sum_k^S \sum_{a_i=1}^2 \sum_{b_j=1}^2 (L_{seg_k}(a_i, b_j))}{4l}$$

where $L_{seg_k}(a_i, b_j)$ is the length of the k^{th} shared IBD segment seg_k measured over homologue *a* of individual *i* and homologue *b* of individual *j*, and *l* is the length of the genome covered by SNPs. Thus, estimation of f_{SEG} requires that phases of SNP genotypes are known. As previously stated these were obtained using the software BEAGLE 4.1. The criteria used to define an IBD segment were based on the distribution of the distance between two consecutive SNPs and the density of SNPs observed along the genome; and are given in the Results section.

Frequencies used in coefficients 2 – 5 were those of the parental population.

The relationships between individuals were divided into groups according to the estimated pedigree relationships as follows: i) parent-offspring; ii) full-sibs; iii) half-sibs; and iv) less related individuals.

Inbreeding coefficients (F_{SIM}, $F_{L\&H}$, F_{VR1}, F_{VR2}, F_{YANG} and F_{ROH}) were obtained from the diagonal of the corresponding coancestry matrix as $F_{(i)} = 2f_{(i,i)} - 1$. Segment-based inbreeding for individual *i* ($F_{ROH(i)}$) is defined as the proportion of the genome of individual *i* that was covered by long uninterrupted homozygous segments (Lencz *et al.*, 2007).

Estimates of coancestry and inbreeding are relative to an arbitrary base population in which the individuals are assumed to be unrelated and non-inbred. Methods 2 – 6

correct for the homozygosity and coancestry in the base population in an attempt to move from an IBS to an IBD scale (Toro *et al*., 2014).

Coefficients f_{VR1} and f_{VR2} (and corresponding F_{VR1} and F_{VR2}) were obtained using the software Gmatrix (Legarra, personal communication), f_{SIM}, $f_{L\&H}$ and f_{SEG} (and F_{SIM}, $F_{L\&H}$, and F_{ROH}) were obtained using our in-house Fortran code and f_{YANG} and F_{YANG} were obtained using the PLINK 1.9 software (Purcell *et al*., 2007).

Patterns of genomic inbreeding

Two approaches were used for investigating patterns of inbreeding across the genome. The first approach was based on F_{SIM} and consisted on dividing each chromosome into sliding windows of 40 SNPs (approximated length 1 Mb), and calculating the average F_{SIM} within each window, that was moved one SNP at a time (Weir *et al*., 2005; Engelsma *et al*., 2012). Finally, values were averaged across individuals. The second approach was performed on ROH-based positional inbreeding. This approach consisted on representing, for each SNP, the proportion of individuals for which the SNP was contained in a ROH (Doekes *et al*., 2018).

Optimisation of contributions

In order to evaluate the amount of genetic variability retained when using different coancestry matrices in the management of populations, the OC method was used (Fernández *et al*., 2003; Villanueva *et al*., 2004; Woolliams *et al*., 2015). The problem to be solved is concerned with the allocation of contributions of the candidates to produce the next generation so as to minimise the global coancestry, and it can be formulated as:

Minimise $\mathbf{c}^T\boldsymbol{\theta}\,\mathbf{c}$

subject to the following constraints:

$\mathbf{Q}^T\mathbf{c} \leq ½\,\mathbf{1}$

$c_i \geq 0$ for i = 1,…, *n* candidates

where **c** is the (*n* x 1) vector of solutions (i.e., contributions or proportions of offspring left by each candidate), $\boldsymbol{\theta}$ is the coancestry matrix, **Q** is a (*n* x 2) known incidence matrix indicating the sex of the candidates with 0's and 1's, and **1** is a (2 x 1) vector of ones. The first inequality ensures that half of the contributions come from males and half come from females. The problem was solved using a simulated annealing algorithm (Kirkpatrick *et*

al., 1983). Only integer solutions were allowed (for details, see Fernández and Toro, 1999). The number of contributing individuals was not restricted. Candidates considered in the optimisation were the offspring with known sex (1,152 out of the 1,391 genotyped offspring). Different coancestry matrices (θ_{SIM}, $\theta_{L\&H}$, θ_{VR1}, θ_{VR2}, θ_{YANG} and θ_{SEG}) were used in the optimisation. The amount of genetic variability retained when using each coancestry matrix was measured as the number of segregating SNPs and the expected heterozygosity (H_E) in the selected candidates.

RESULTS

The genome had a total length of ~ 524 Mb (Table 1) and was organised in 22 LGs that correspond to the 22 chromosomes of the turbot haploid karyotype. The average LG length was 24 Mb. We followed the LG nomenclature of Figueras *et al*. (2016) but note that LG18 is missing due to its merging with LG08 (i.e., LG08+LG18 is now LG08) in the most recent genetic map (Maroso *et al*., 2018). The average number of SNPs per LG was 824, ranged from 576 (LG22) to 1,131 (LG02) and, in general, increased with increasing LG length.

Table 1. Number of SNPs (n_{snp}), length (in Mb), density of SNPs (in SNP/Mb) and average distance between SNPs (*d*, in Mb) for each linkage group (LG).

LG	n_{SNP}	*Length*	*Density*	*d*
01	1,031	26.87	38.33	0.026
02	1,131	31.89	35.46	0.028
03	783	21.32	36.72	0.027
04	902	29.50	30.58	0.033
05	705	24.81	28.42	0.035
06	850	25.19	33.74	0.030
07	697	24.31	28.67	0.035
08	1,061	30.93	34.31	0.029
09	928	25.79	35.99	0.028
10	919	25.10	36.62	0.027
11	801	27.22	29.43	0.034
12	794	25.24	31.46	0.032
13	722	19.99	36.12	0.028
14	775	21.45	36.13	0.028
15	884	24.13	36.64	0.027
16	807	23.80	33.91	0.030
17	678	15.88	42.70	0.023
19	754	21.78	34.62	0.029
20	775	22.75	34.06	0.029
21	760	21.35	35.59	0.028
22	576	14.91	38.63	0.026
23	793	19.89	39.87	0.025
Total	18,125	524.10		

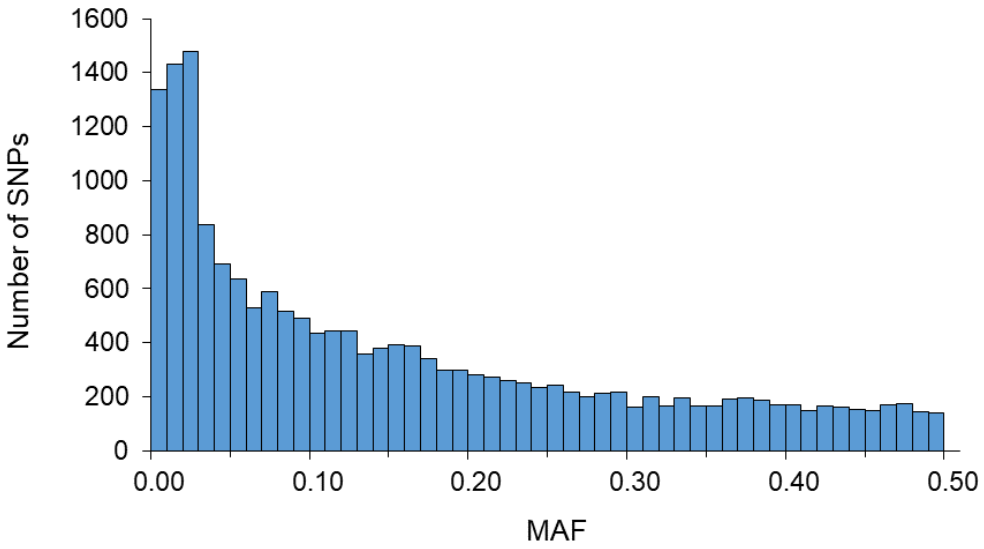

Figure 2. Distribution of the minimum allele frequency (MAF).

The distribution of MAF (Figure 2) indicated that almost 70% of the SNPs (12,466 SNPs) had a MAF ≥ 0.05 and only about 8% (1,338 SNPs) had a MAF < 0.01 (notice that monomorphic SNPs had been removed as indicated in Material and Methods). The average distance between adjacent SNPs was 0.029 Mb and ranged from 0.025 (LG23) to 0.035 (LG07). About 96% of adjacent SNPs were at distances ≤ 0.1 Mb (Figure 3). The average SNP density for the whole genome was 34.90 SNPs/Mb and ranged from 28.42 (LG05) to 42.70 (LG17) SNPs/Mb (Table 1).

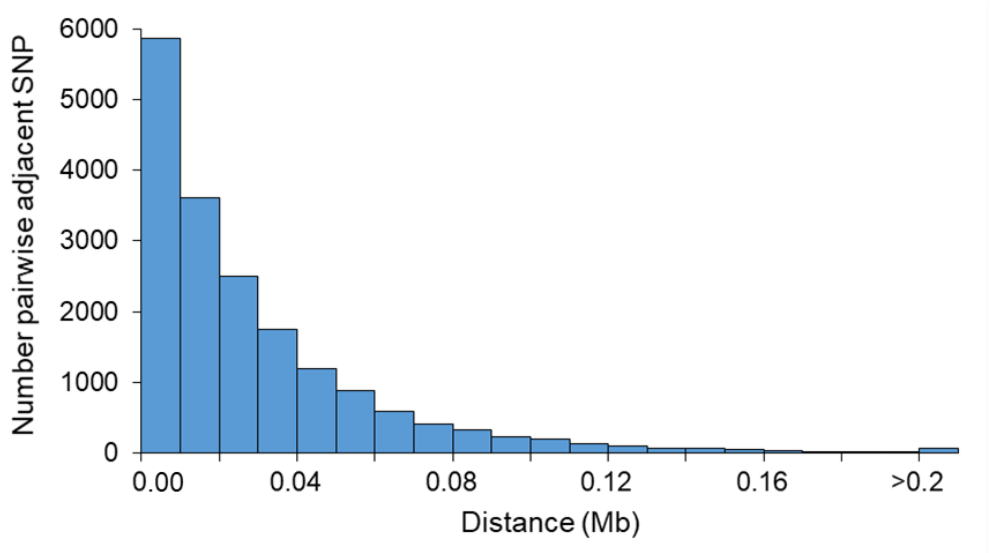

Figure 3. Distribution of the distance between adjacent SNPs (in Mb) in the genome.

Imputation

Before imputation, ~ 83% of the SNPs (15,131) had genotypes available for > 95% of the individuals. This percentage increased up to ~ 99% after imputation. Accordingly, before imputation, 82% of the individuals (1,179) had genotypes available for more than 95% of the SNPs and this percentage increased up to 99.9% after imputation. In total, the amount of genotypic data available with high reliability (> 90%) increased by > 13% by carrying out the imputation. Detailed information about the number of missing genotypes per individual and per marker before and after imputation is given in Table 2.

Table 2. Number (*N*) and percentage (%) of individuals and SNPs for different categories of call rates (in percentage) before and after imputation.

	Pre-imputation				Post-imputation			
	Individuals		SNPs		Individuals		SNPs	
Call Rate	*N*	%	*N*	%	*N*	%	*N*	%
>99	53	3.69	8,645	47.70	1,416	98.47	17,269	95.28
>98	379	26.36	12,291	67.81	1,434	99.72	17,637	97.31
>97	748	52.02	13,841	76.36	1,437	99.93	17,821	98.32
>96	987	68.64	14,593	80.51	1,437	99.93	17,917	98.85
>95	1,179	81.99	15,131	83.48	1,437	99.93	17,980	99.20
>90	1,409	97.98	16,466	90.85	1,438	100.00	18,097	99.85
>80	1,427	99.24	17,331	95.62	1,438	100.00	18,123	99.99
>70	1,430	99.44	17,711	97.72	1,438	100.00	18,124	99.99
>60	1,432	99.58	17,928	98.91	1,438	100.00	18,125	100.00

IBD segments and runs of homozygosity (ROH)

The criteria used to define a ROH and an IBD segment were the same and they were based on the results obtained when analysing the genomic information available. The minimum length of a ROH chosen was 0.4 Mb. It is expected that ROH of this length come from common ancestors born 50 generations ago, when assuming a recombination rate of 2.5 cM/Mb (Bouza *et al*., 2007) and the fact that ROH length (in cM) equals $100/2g$, where g are the number of generations in the past (Purfield *et al*., 2017). In order to avoid that sparsely covered genomic regions increase the length of a ROH and

artificially inflate estimates of *F* and *f*, the minimum density required was 1 SNP every 50 kb and the maximum distance allowed between two consecutive homozygous SNPs in a ROH was 0.1 Mb. Note that the minimum average density per LG (Table 1) was 28.4 SNP/Mb (i.e., 0.0284 SNP/kb or 1 SNP every 35.1 kb) and that more than 95% of SNP pairs were at distances shorter than 0.1 Mb (Figure 3). No heterozygous genotypes and a maximum of 1 missing genotype were allowed in a ROH. These criteria led to a minimum number of SNPs in a ROH of 9.

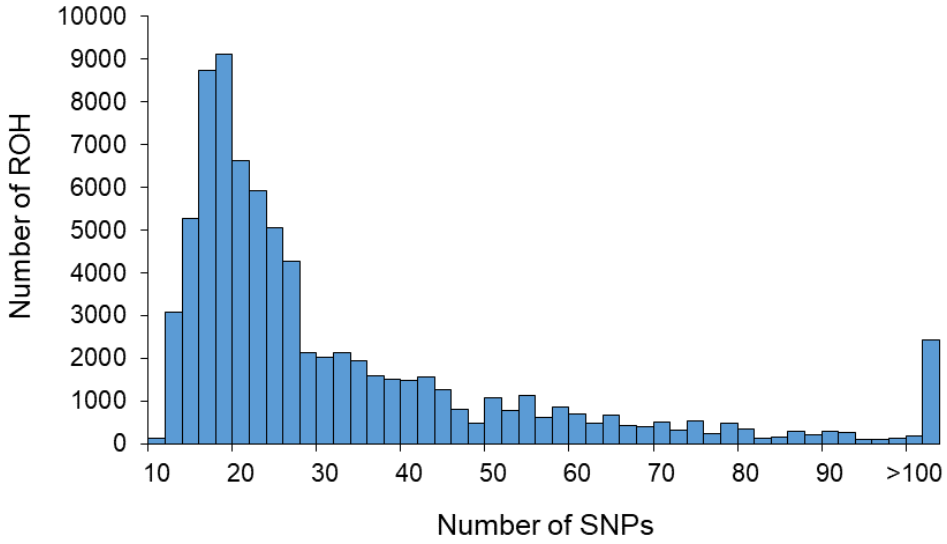

Figure 4. Distribution of the number of SNPs within a ROH.

The distribution of the number of SNP in a ROH is shown in Figure 4. A large proportion of ROH presented a low number of SNPs (66% of ROH had less than 30 SNPs). This proportion decreased as the number of SNPs within a ROH increased. This is in accordance with the pattern showed in Figure 5 that reflects a higher proportion of ROH of shorter length.

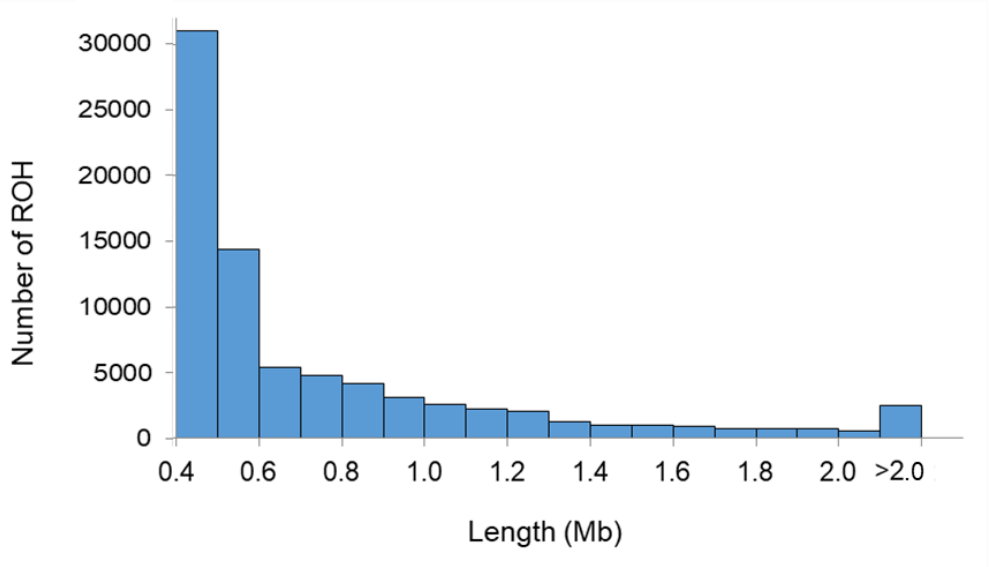

Figure 5. Distribution of the ROH length (in Mb).

The length of the ROH varied between 0.40 and 5.55 Mb. About 80% of the ROH detected in the entire genome of the offspring population had a length ≤ 1 Mb (Table 3), which corresponds to ancient inbreeding -approximately 50 generations ago, as previously mentioned- and only about 0.25% had length > 4 Mb (i.e. recent inbreeding, less than 5 generations ago). In the parents, the percentage of short ROH (≤ 1 Mb) increased to 91% and there were no ROH with a length > 5 Mb, thus indicating that most of the inbreeding in parents and offspring is ancestral.

Table 3. Number of ROH (N_{ROH}) and average number of ROH per individual (Ave_{ROH}) in parents and offspring populations by ROH length (l_{ROH}) category (in Mb).

	Parents		Offspring	
l_{ROH}	N_{ROH}	Ave_{ROH}	N_{ROH}	Ave_{ROH}
0.4 – 1.0	1,501	32.63	62,829	45.17
1.0 – 2.0	117	2.54	12,994	9.34
2.0 – 4.0	24	0.52	2,856	2.05
4.0 – 5.0	1	0.02	133	0.10
> 5.0	0	0.00	40	0.03

The distribution of ROH across the genome was not homogeneous (Table 4). Large ROH (> 5 Mb) were only observed in LG17 and LG20 and in the offspring population while ROH of 4 to 5 Mb were only detected in LG10, LG17, LG20 and LG23, also in the offspring population.

Table 4. Average number of ROH of different length (in Mb) in parents (P) and offspring (O) for the different linkage groups (LG).

	Length ROH category (Mb)									
LG	0.4−1.0		1.0−2.0		2.0−4.0		4.0−5.0		>5.0	
	P	O	P	O	P	O	P	O	P	O
1	1.13	2.27	0.13	0.55	0.09	0.12	0.00	0.00	0.00	0.00
2	0.87	2.22	0.02	0.58	0.00	0.07	0.00	0.00	0.00	0.00
3	1.11	1.56	0.28	0.56	0.00	0.05	0.00	0.00	0.00	0.00
4	2.04	3.18	0.09	0.49	0.00	0.00	0.00	0.00	0.00	0.00
5	2.24	2.38	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00
6	0.96	1.35	0.24	0.65	0.00	0.08	0.00	0.00	0.00	0.00
7	2.22	2.74	0.04	0.19	0.04	0.06	0.00	0.00	0.00	0.00
8	2.78	2.96	0.26	0.62	0.00	0.05	0.00	0.00	0.00	0.00
9	1.35	1.88	0.22	0.55	0.04	0.06	0.00	0.00	0.00	0.00
10	1.15	1.82	0.17	0.42	0.02	0.17	0.00	0.02	0.00	0.00
11	2.67	3.42	0.09	0.32	0.02	0.11	0.00	0.00	0.00	0.00
12	1.09	2.29	0.22	0.35	0.00	0.03	0.00	0.00	0.00	0.00
13	1.48	2.06	0.09	0.39	0.00	0.08	0.00	0.00	0.00	0.00
14	1.43	1.92	0.13	0.60	0.02	0.07	0.00	0.00	0.00	0.00
15	1.04	1.85	0.07	0.33	0.02	0.35	0.00	0.00	0.00	0.00
16	1.59	2.26	0.04	0.42	0.07	0.08	0.00	0.00	0.00	0.00
17	0.74	0.79	0.00	0.16	0.07	0.17	0.00	0.02	0.00	0.01
19	1.48	2.08	0.04	0.38	0.00	0.11	0.00	0.00	0.00	0.00
20	1.30	1.58	0.04	0.35	0.04	0.12	0.00	0.02	0.00	0.02
21	2.20	2.62	0.13	0.40	0.00	0.03	0.00	0.00	0.00	0.00
22	0.30	0.71	0.13	0.30	0.00	0.03	0.00	0.00	0.00	0.00
23	1.46	1.24	0.11	0.40	0.09	0.18	0.02	0.04	0.00	0.00
Total	32.63	45.17	2.54	9.34	0.52	2.05	0.02	0.10	0.00	0.03

The length used for the estimation of f_{SEG} and F_{ROH} was the length of the autosome covered by SNPs (i.e., the chromosome length minus the summed length of gaps longer than 100 kb).

The coefficients based on IBD segments provide a general idea of how inbreeding has evolved over generations. For instance, considering segments of less than 2 Mb would reflect the inbreeding about 10 generations ago, in this case $F_{ROH(10g)}$ = 0.109, while considering segments lower than 1 Mb, would reflect the inbreeding 20 generations ago, which corresponds to a value of $F_{ROH(20g)}$ = 0.072.

Estimates of genomic coancestry and inbreeding coefficients

Estimates of the different measures of genomic *f* and *F* are given in Table 5, for both parents and offspring. For coancestry coefficients, results from parents and offspring were very similar and therefore, we will focus on the results from the offspring. Estimates of f_{SIM} were much higher (mean = 0.78, SD = 0.02) than estimates for the other coefficients, in particular those for $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG}, which were close to zero. The high values for f_{SIM} can be explained by the fact that this coefficient reflects IBS rather than IBD. Alternatively, it can be explained by the fact that it is referred to very distant base population in which all alleles were unique. Given that the allelic frequencies used to compute $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG} were those in the parents, this generation can be considered as the reference population which explain their close to zero values. The estimate of f_{SEG} was also low but higher than that of $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG}. As the probability of inheriting the same segment increases when individuals are related, this coefficient reflects IBD. In particular, we used a minimum length of segments corresponding to a common ancestor 50 generations ago, as previously indicated.

Table 5. Mean, standard deviation (SD) and minimum (Min) and maximum (Max) values for the estimates of the different coancestry (*f*) and inbreeding (*F*) coefficients in parents and offspring.

	Parents				Offspring			
	Mean	SD	Min	Max	Mean	SD	Min	Max
f_{SIM}	0.776	0.026	0.740	0.912	0.777	0.018	0.740	0.927
$f_{L\&H}$	0.002	0.117	–0.157	0.608	0.004	0.081	–0.157	0.673
f_{VR1}	0.000	0.096	–0.123	0.748	0.002	0.062	–0.120	0.715
f_{VR2}	0.001	0.199	–0.174	1.210	0.002	0.064	–0.089	0.974
f_{YANG}	0.001	0.092	–0.081	0.649	0.002	0.063	–0.089	0.689
f_{SEG}	0.129	0.091	0.032	0.663	0.134	0.064	0.021	0.709
F_{SIM}	0.760	0.016	0.720	0.824	0.774	0.023	0.721	0.853
$F_{L\&H}$	–0.074	0.069	–0.263	0.196	–0.006	0.105	–0.265	0.350
F_{VR1}	–0.068	0.197	–0.272	0.431	–0.008	0.119	–0.272	0.431
F_{VR2}	–0.029	0.509	–0.503	1.421	0.000	0.345	–0.405	0.948
F_{YANG}	–0.004	0.092	–0.087	0.654	–0.006	0.069	–0.154	0.378
F_{ROH}	0.064	0.064	0.026	0.316	0.126	0.081	0.017	0.416

In the parents, the highest standard deviation was for f_{VR2} and the lowest was for f_{SIM}. The higher the variation, the higher the ability of a particular coefficient to differentiate relationships between individuals. The distributions of the different coancestry coefficients in the offspring showed that the lowest dispersion around the mean corresponded to f_{SIM}, and the highest dispersion corresponded to $f_{L\&H}$ (Figure 6).

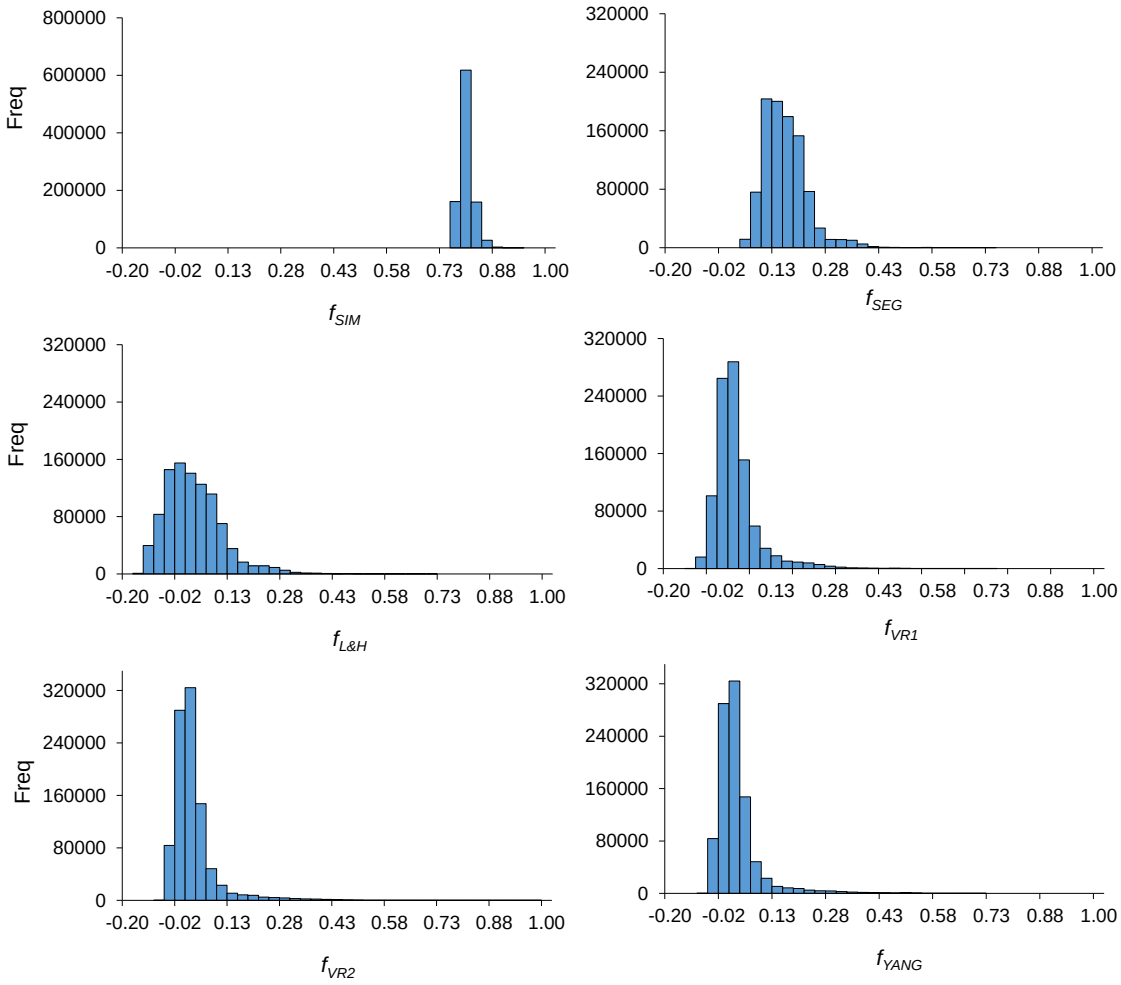

Figure 6. Distributions of the different estimates of coancestry coefficients in the offspring.

Results for inbreeding (Table 5) were similar to those obtained for coancestry. Note that in the estimates of coancestry coefficients, self-coancestries are also included. Here, again F_{SIM} presented the lowest standard deviation, while it was F_{VR2} the coefficient presenting the highest standard deviation both for parents and offspring (much higher than in the case of coancestry). This is also evident from Figure 7, where F_{VR2} shows a skewed right distribution with a high dispersion.

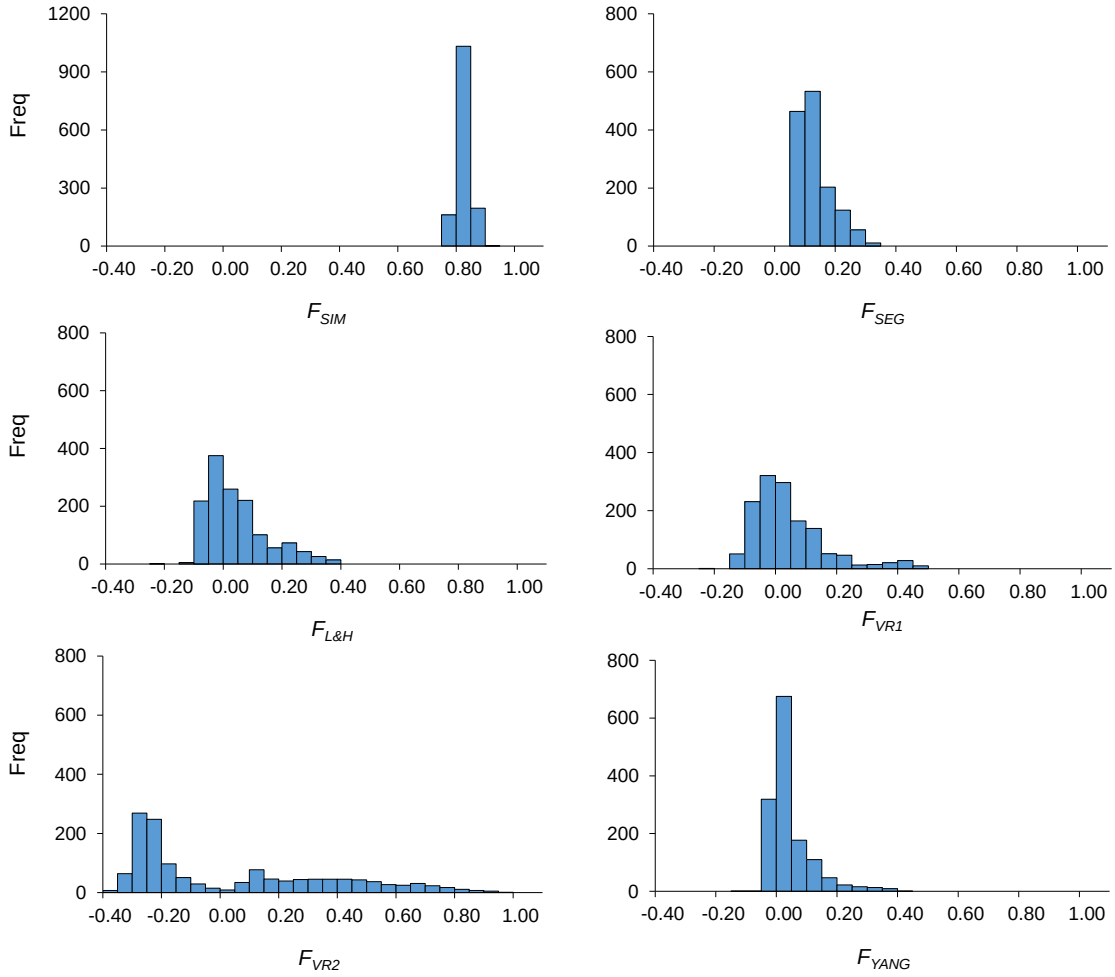

Figure 7. Distributions of the different estimates of inbreeding coefficients in the offspring.

In order to evaluate the ability to detect different degrees of relatedness for the six coancestry coefficients analysed, our data were split according to four degrees of relatedness, including parent-offspring, full-sibs, half-sibs and less related individuals (Table 6). All coefficients were able to discriminate different degrees of relationships, although f_{VR2} and f_{YANG} presented the values that more approximated to the expected values derived from pedigree-based coancestry (i.e., 0.250 for parent-offspring and full-sibs, 0.125 for half-sibs, Falconer and Mackay, 1996).

Table 6. Mean, standard deviation (SD) and minimum (Min) and maximum (Max) values of the estimates of the different coancestry coefficients for different degree of relationships. The number of individual-pairs (*N*) compared in each corresponding category is also indicated.

Relationship	*N*		Mean	SD	Min	Max
Parent-offspring	2,783	f_{SIM}	0.828	0.012	0.802	0.864
		$f_{L\&H}$	0.231	0.055	0.117	0.395
		f_{VR1}	0.228	0.067	0.107	0.424
		$f_{VR2}= f_{Yang}$	0.238	0.136	0.078	0.666
		f_{SEG}	0.325	0.042	0.244	0.446
Full-sibs	26,296	f_{SIM}	0.829	0.013	0.790	0.911
		$f_{L\&H}$	0.235	0.059	0.063	0.602
		f_{VR1}	0.232	0.065	0.058	0.512
		$f_{VR2}= f_{Yang}$	0.241	0.106	0.059	0.637
		f_{SEG}	0.328	0.047	0.177	0.567
Half-sibs	42,762	f_{SIM}	0.799	0.014	0.767	0.854
		$f_{L\&H}$	0.104	0.063	−0.040	0.348
		f_{VR1}	0.109	0.048	−0.029	0.303
		$f_{VR2}= f_{Yang}$	0.115	0.080	−0.023	0.411
		f_{SEG}	0.219	0.048	0.106	0.417
Less related	959,929	f_{SIM}	0.774	0.015	0.740	0.848
		$f_{L\&H}$	−0.008	0.067	−0.161	0.320
		f_{VR1}	−0.011	0.038	−0.120	0.213
		$f_{VR2}= f_{Yang}$	−0.011	0.033	−0.089	0.467
		f_{SEG}	0.124	0.051	0.021	0.396

Correlations between the different coancestry coefficients ranged between 0.72 and 1.00, being $f_{VR1,}$ f_{VR2} and f_{YANG} those showing the lowest correlations and the highest dispersion when regressed on the other coefficients (Figure 8). Two groups of coefficients can be observed: (1) $f_{SIM,}$ $f_{L\&H}$ and f_{SEG} with high correlations between them and low dispersion around the regression line; and (2) $f_{VR1,}$ f_{VR2} and f_{YANG} with generally lower correlations between them and with f_{SIM}, $f_{L\&H}$ and f_{SEG} and considerable dispersion around the regression line. Corresponding correlations for inbreeding evidenced more extreme results, ranging from −0.40 to 1.00 (Figure 9). Again, $F_{VR1,}$ F_{VR2} and F_{YANG} showed the lowest correlations and the highest dispersions. In particular, F_{VR2} was the only coefficient presenting negative correlations with the other coefficients and a low correlation (0.26)

with F_{YANG} (this was however expected given the different way of computing the diagonal by these methods). F_{VR1} presented also very low correlations with f_{SIM}, $f_{L\&H}$ and $f_{SEG.}$ In order to understand these extreme values (i.e., negative correlations), we re-estimated the correlations between inbreeding coefficients after removing the alleles at low frequency (MAF ≤ 0.05) (Table 7). Our results revealed a generalised increase in the correlations between $F_{VR1,}$ F_{VR2} and F_{YANG} with the other coefficients. This effect was more evident for the correlations involving F_{VR2}, thus indicating that this method is the one giving more weight to rare alleles.

Table 7. Correlations between different coefficients of genomic inbreeding using SNPs with MAF ≥ 0.05 (12,466 SNPs) and the offspring data.

	F_{SIM}	$F_{L\&H}$	F_{VR1}	F_{VR2}	F_{ROH}
$F_{L\&H}$	1.00				
F_{VR1}	0.53	0.53			
F_{VR2}	0.40	0.40	0.95		
F_{YANG}	0.92	0.92	0.77	0.71	
F_{ROH}	0.87	0.87	0.19	0.13	0.73

Patterns of genomic inbreeding

The average inbreeding within LGs ranged from 0.75 (LG22) to 0.80 (LG05) for F_{SIM} and between 0.08 (LG21) to 0.12 (LG11) for F_{ROH}. Substantial heterogeneity within LGs was observed according to the inbreeding patterns for F_{SIM} and F_{ROH} (Figure 10). The patterns of F_{SIM} differed from those of F_{ROH} although regions of increased inbreeding overlapped for both coefficients in particular cases (e.g., the peaks on LG14).

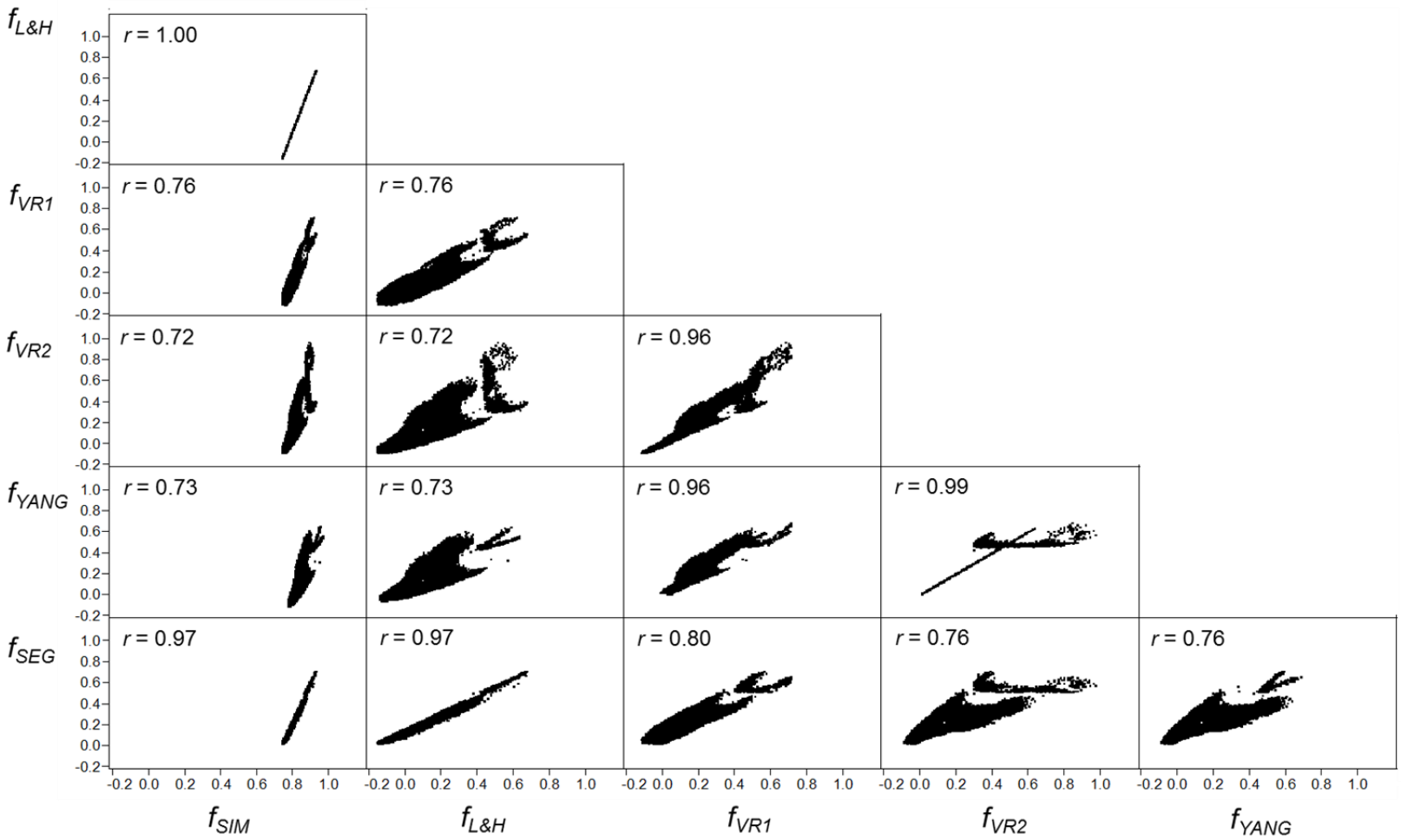

Figure 8. Linear regressions plots for each coancestry coefficient against each other and corresponding correlation coefficients. Note that the separated clouds correspond to self-coancestries.

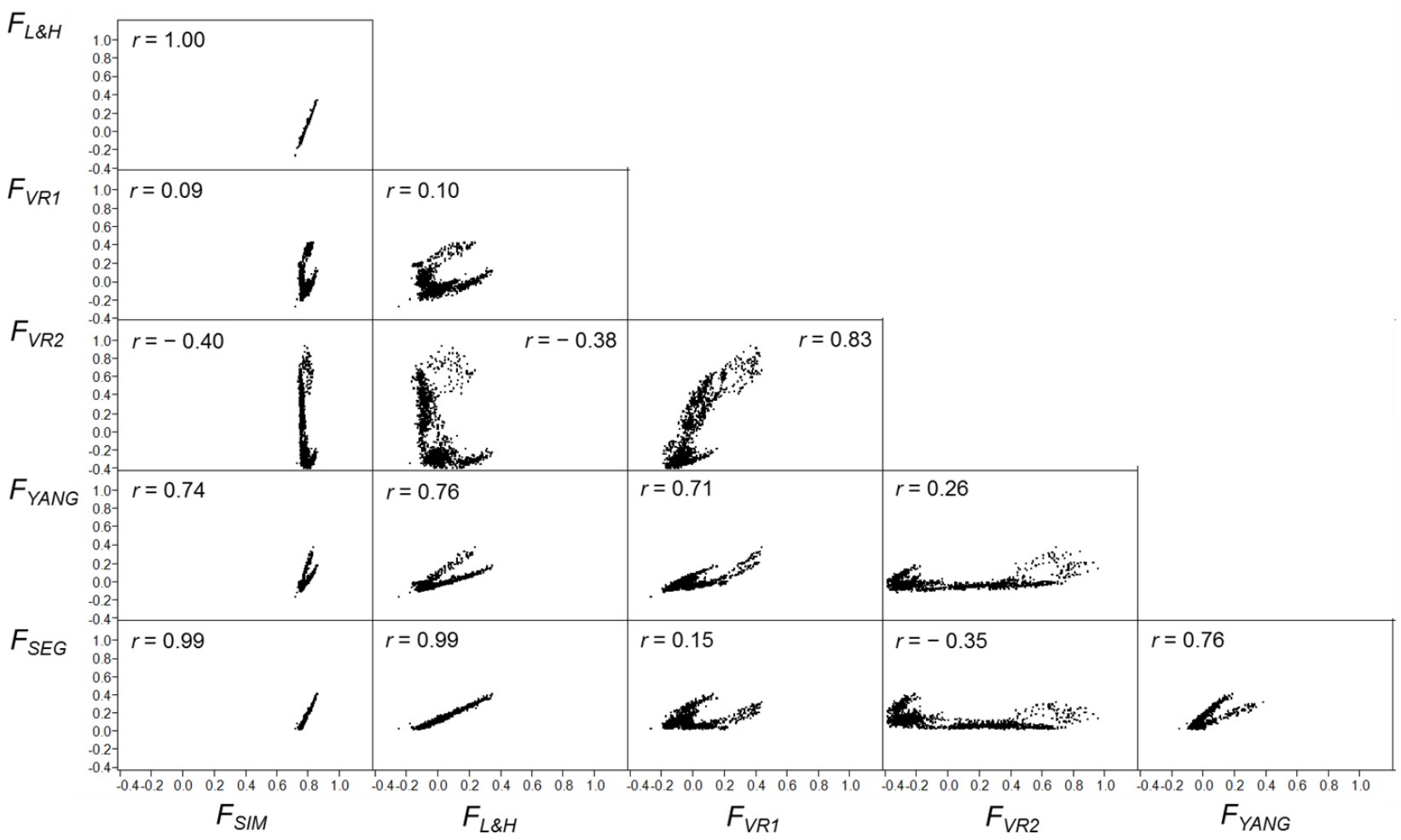

Figure 9. Linear regressions plots for each inbreeding coefficient against each other and corresponding correlation coefficients.

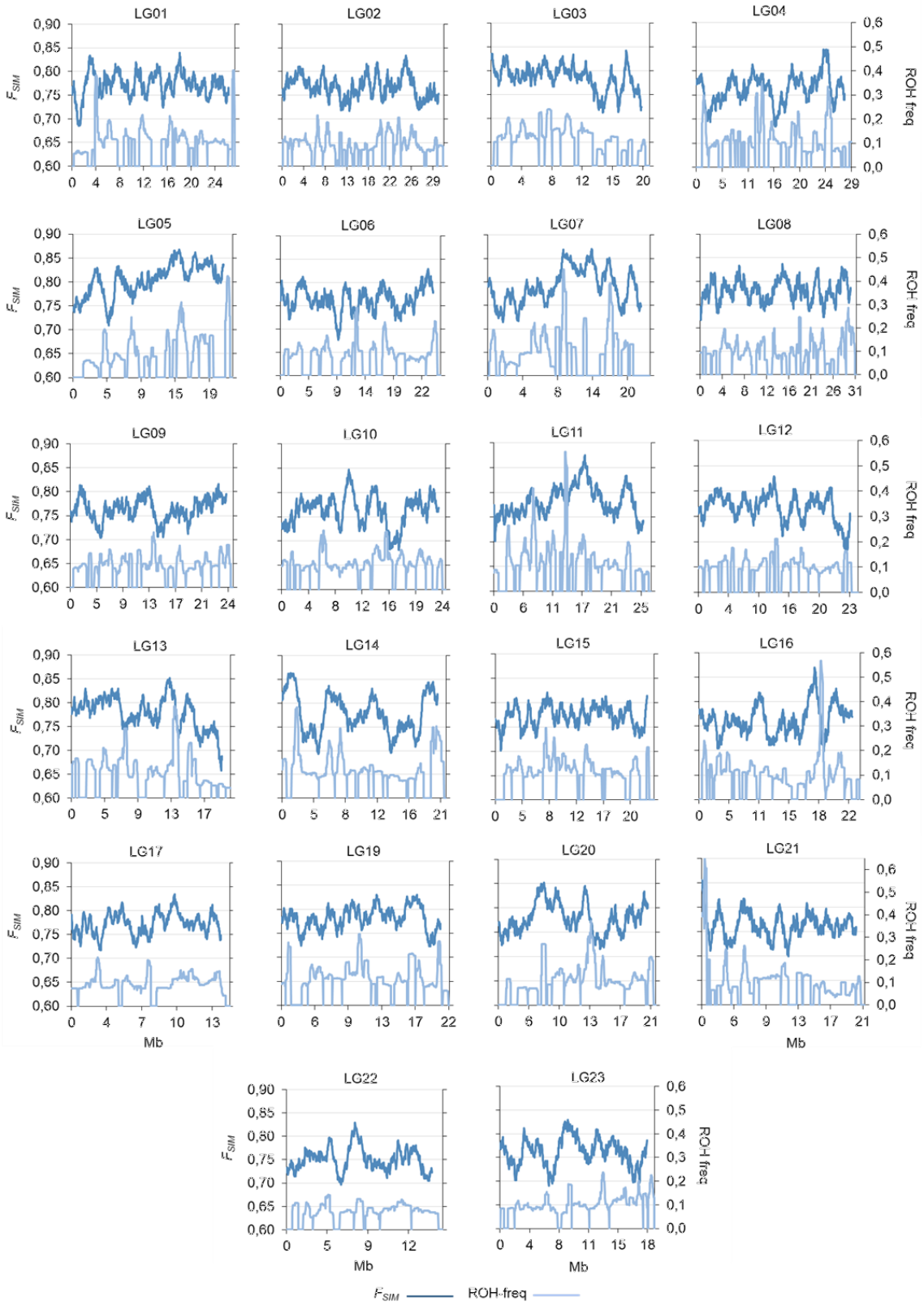

Figure 10. Patterns of genomic inbreeding measured as F_{SIM} (clear line) and as the frequency of individuals for which each SNP is contained in a ROH (ROH freq, dark line) for the different linkage groups (LGs).

Optimisation of contributions

The genetic variability retained in the selected candidates when using the different coancestry matrices in OC was similar in terms of expected heterozygosity and percentage of alleles that remained segregating (> 99 % in all cases) although the less fish selected the higher was the number of SNPs fixed (Table 8). However, the different matrices achieved similar variability by selecting very different number of individuals. Note that no restriction to the number of individuals selected was imposed. In particular, with $\theta_{L\&H}$ and θ_{SEG} only 9% and 13% of the initial number of candidates were selected, respectively, in comparison with 47-85% of candidates selected when using θ_{SIM}, θ_{VR1}, θ_{VR2} and θ_{YANG}. These differences can be explained by the fact that $\theta_{L\&H}$ and θ_{SEG} presented higher variances for coancestry coefficients estimated for the group of 'less related individuals' (Table 6). Conversely, θ_{SIM}, which led to the highest number of candidates selected, was also the matrix with lower ability to differentiate relationships. Similar numbers of males and females were selected in all cases (results not shown).

Table 8. Number of fish selected to contribute (N_{sel}), mean contribution (c) and corresponding standard deviation (SD), minimum and maximum contributions, and number (percentage) of SNPs segregating and expected heterozygosity (H_E) of the candidates selected after implementing OC using different coancestry matrices.

	N_{sel}	c	SD	Min	Max	SNP segregating		H_E
θ_{SIM}	976	0.001	0.002	0.000	0.017	18,097	(100.00%)	0.225
$\theta_{L\&H}$	94	0.011	0.008	0.000	0.042	17,921	(99.03%)	0.236
θ_{VR1}	639	0.002	0.001	0.000	0.007	18,087	(99.94%)	0.227
θ_{VR2}	636	0.002	0.001	0.000	0.007	18,096	(100.00%)	0.226
θ_{YANG}	544	0.002	0.003	0.000	0.027	18,052	(99.75%)	0.213
θ_{SEG}	152	0.007	0.006	0.000	0.030	18,041	(99.69%)	0.234

DISCUSSION

In this project, we have made use of new genomic tools recently developed for turbot in order to compare different estimators of coancestry and inbreeding based on genomic information. Different genomic coancestry matrices were evaluated in terms of their efficiency in retaining genetic variability when implementing the OC method. Our results revealed differences both in the magnitude and in the correlation between the different coancestry (and inbreeding) coefficients, that were mainly related with the reference population to which each coefficient referred to and to the weight given to rare alleles in the different measures. All coancestry matrices showed similar efficiency in retaining genetic variability when used in OC, although the number of candidates selected varied across greatly. This seemed to be explained by the different variance of the analysed coefficients which led to different abilities to detect differences in relationships between individuals.

The levels of the expected heterozygosity retained after the optimisation were low (between 0.21 and 0.24), although similar than those found in other commercial fish populations. For instance, in a farmed population of Atlantic salmon, Kijas *et al*. (2016) found an H_E of 0.20. This value was clearly lower than estimates for wild populations of the same species (0.31). Estimates of H_E obtained from microsatellites (and therefore not directly comparable to our estimate) for commercial populations of turbot (Coughlan, *et al.*, 1997; Bouza *et al.*, 2002; Exadactylos *et al*., 2007), Atlantic salmon (Skaala *et al.*, 2004) and carp (Ren *et al*., 2018) are also clearly lower than corresponding estimates for wild populations.

The low levels of H_E is in accordance with the low estimates of N_e obtained for commercial fish populations. Saura *et al.* (2018) have recently given estimates of N_e for turbot (actually for the same population studied here), gilthead seabream and carp of 28, 40 and 22 fish, respectively. Other estimates for gilthead seabream have ranged from 14 to 18 individuals between photoperiod-controlled broodstock groups (Brown *et al*., 2005). Estimates lower than 50 have been also obtained for commercial coho salmon (Gallardo *et al*., 2004; Yáñez *et al*., 2014) and rainbow trout (Su *et al*., 1996; Pante *et al*., 2001). These findings are in line with our results and highlight the necessity of broadening genetic diversity when base populations are built for starting breeding programmes in aquaculture.

Genomic coefficients of coancestry and inbreeding are very useful when genealogical information is not available, as is the case in many wild populations, for which it is difficult to register pedigrees, or the case of starting new selection programmes for aquaculture species. Genomic coancestry matrices, which are based on high density SNP information, can more accurately reflect the true relationships between individuals than the standard pedigree-based coancestry matrix because they take into account the variability among individuals with the same degree of relationship (e.g., full-sibs) due to Mendelian segregation of SNPs. Indeed, using the genomic coancestry matrix with sufficient marker densities in OC has been demonstrated to be more efficient in preserving genetic diversity than using pedigree-based relationships (de Cara *et al*., 2011; Gómez-Romano *et al*., 2013).

The physical map used in this work has also been obtained within the framework of the European project FISHBOOST. The battery of SNPs obtained through 2b-RAD-sequencing allowed a better integration of the genetic and physical maps available to date (Figueras *et al*., 2016) and the improvement of the anchoring of the genome assembly from 80 to 97% (Maroso *et al*., 2018). This physical map was integrated by 18,097 SNPs that translated in a SNP density comparable (or even higher) to predesigned commercial arrays (e.g., the Illumina PorcineSNP60 BeadChip). However, a disadvantage of GBS techniques is that they select only a small subset of sites along the genome and, although they are intended to be distributed homogeneously, this is not always achieved, which may be a possible source of bias affecting the magnitude of the coancestry and inbreeding coefficients based on IBD segments.

Our results evidenced differences in the magnitude of the different coefficients compared. In particular, f_{SIM} was much higher than the other coefficients, which was expected given that f_{SIM} reflects, by definition, the relationships caused by a common ancestor going back to a very distant base population in which all the alleles were unique. Indeed, depending on where we establish the base population, the magnitude of the coancestry and inbreeding coefficients is expected to change. The coancestry coefficient based on the proportion of shared IBD segments was lower than f_{SIM} but higher than $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG}, which can be explained by the fact that the base population for f_{SEG} was established 50 generations ago (by setting segments of minimum length of 0.4 Mb, as previously indicated). In contrast, the base population to which $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG}

referred to, was the parental population, as the allele frequencies in this population were those used to compute these coefficients. For this reason, their values were close to zero. Since individuals in the base population are assumed to be unrelated and not inbred, establishing the base population 50 generations ago seems to be more reasonable than establishing it at the present or at the theoretical ancestral time assumed by f_{SIM}. Thus, coefficients based on IBD segments may represent a good choice for estimating coancestry and inbreeding that range in a scale easy to interpret.

The negative estimates obtained for $f_{L\&H}$, f_{VR1}, f_{VR2} and f_{YANG} and for $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YANG} can be better interpreted in terms of Wright's (1921) original correlation concept of relatedness than in terms of Malécot's (1948) probability of IBD (Wang, 2014). Wright (1921) defined the inbreeding coefficient of an individual as the correlation between homologous genes of the two gametes (one from father and one from mother) uniting to form the individual, relative to the total array of such gametes in random derivatives of the reference population. Malécot's definition of F as the probability of IBD of the two homologous genes at a locus within an individual came later. Also, negative F values can have a biological meaning, signifying that the probability of the two homologous genes within an individual being IBD is smaller than that of two homologous genes drawn at random from the reference population (Wang, 2014). This contrasts with the possible values for f_{SIM}, f_{SEG}, F_{SIM} and F_{ROH} that will never be negative, because these coefficients express proportions.

Correlations between the different coancestry coefficients were high, although coefficients based on genomic relationship matrices (i.e., f_{VR1}, f_{VR2} and f_{YANG}) showed an important dispersion around the regression line, in particular those correlations where f_{VR2} was involved (Figure 8). This effect was even more obvious in the case of inbreeding (Figure 9). Given the way in which f_{VR1} and f_{VR2} (and F_{VR1} and F_{VR2}) are computed, it seems as if the latter gives more weight to rare alleles. f_{YANG} has similar properties than f_{VR2}, with the only difference being that self-coancestries are computed more precisely with f_{VR2} (Eynard *et al*., 2016). The higher weight given to rare alleles in the VanRaden´s method (and in Yang´s method) seemed to be the cause of the low correlations and high dispersions observed in Figure 11(A). Removing SNPs with low MAF from the dataset, led to an increase in the correlations and a reduction in the dispersion from the regression line, as can be observed in Figure 11 where low MAF SNPs were sequentially removed.

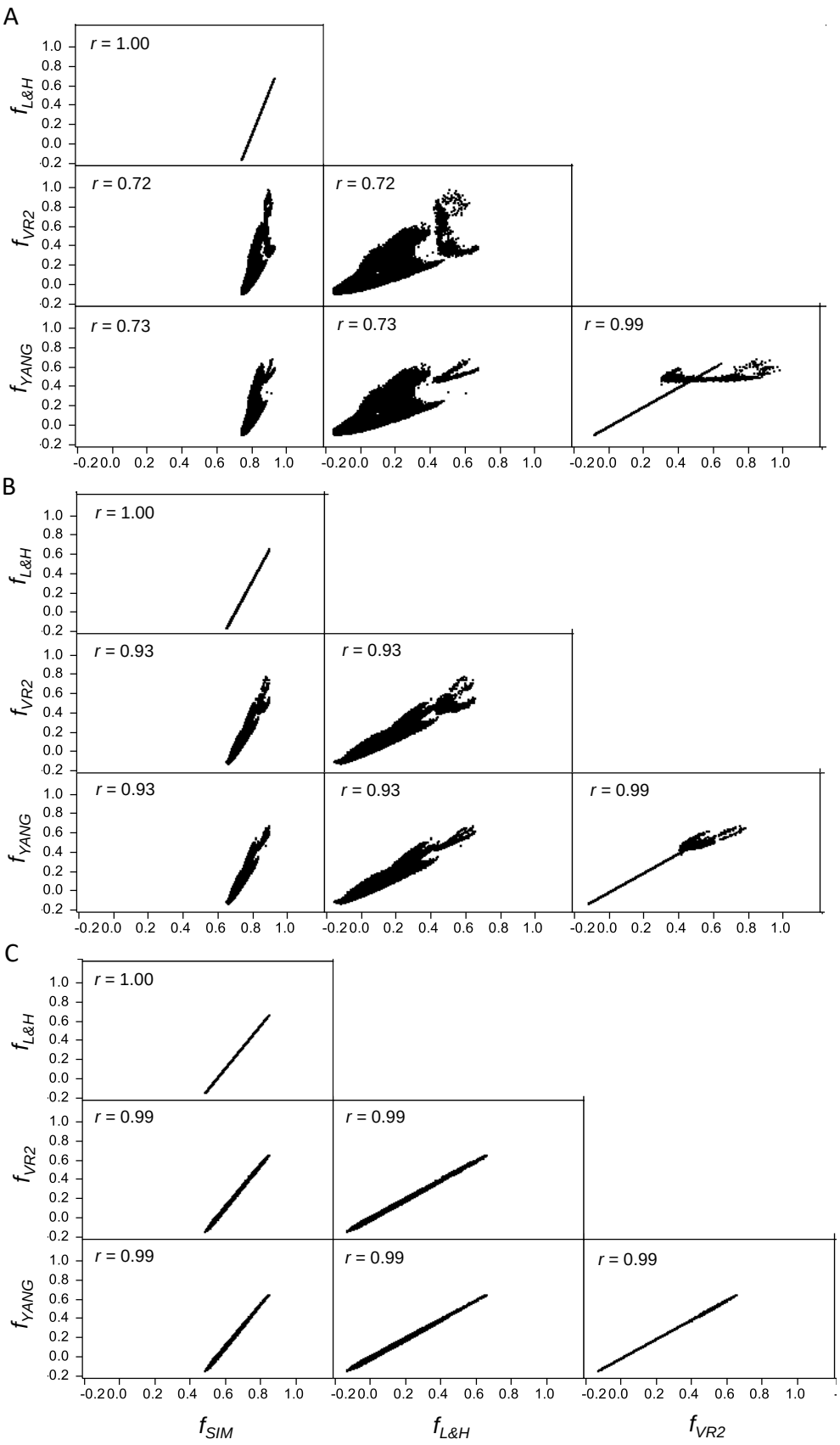

Figure 11. Linear regression plots for coancestry coefficients f_{SIM}, $f_{L\&H}$, f_{VR2} and f_{YANG} against each other and corresponding correlation coefficients when using SNPs with MAF > 0.00 (A), MAF ≥ 0.05 (B) and MAF ≥ 0.25 (C).

The coefficients based on the proportion of IBD segments, in particular the inbreeding coefficient F_{ROH}, has been widely used in human (e.g., Curik *et al*., 2014) and farm animal (e.g., Purfield *et al*., 2012; Bjelland *et al*., 2013; Saura *et al*., 2015; Rodríguez-Ramilo *et al*., 2015) studies. If these segments are long enough, it is likely that the two copies came from a particular ancestor (Gibson *et al*., 2006), thus reflecting (at least to some extent) IBD. It must be considered, however, that there is a certain limitation with segment-based coefficients since the results can vary considerably depending on the criteria used to define a segment and these criteria are, in a certain way, arbitrary. In fact, there is a lack of consensus regarding establishing these criteria to identify autozygosity (homozygosity in which the two alleles are IBD, that is, they are copies of an ancestral gene), differentiating non-autozygotic segments that are IBS from IBD segments (Peripolli *et al*., 2017; Zhang *et al*., 2015). Due to the wide variety of criteria used in the literature, one should be extremely cautious when comparing segment-based coefficients across studies. Here, we have tried to follow objective procedures to choose our criteria to identify IBD segments. The recombination rate, the size of the genome of the turbot and the number of generations in the past in which the base population is established to determine the minimum segment length were taken into consideration. Also, the density and distribution of the SNPs along the genome was taken into account to establish the minimum segment density and the maximum allowed distance between SNPs.

The fact that long ROHs (> 5 Mb) only appear in two chromosomes and only in some of the descendants (in certain families) and not in the parents, may confirm that the long segments reflect recent inbreeding and that this is low in our study. Alternatively, a segment-based IBD matrix can be used in Optimal Contributions Selection (OCS) to restrict the increase in recent inbreeding (Doekes *et al*., 2018). The rationale behind this approach is that recent inbreeding is more harmful than distant inbreeding, because the latter may have already been purged especially in functional genes (Ballou, 1997; Charlesworth and Willis, 2009). This agrees with the results found by Szpiech *et al*. (2013) that reported that individuals with a high ROH coverage had a higher fraction of deleterious variants occurring in long ROH, which is in agreement with the hypothesis that recent inbreeding enables rare deleterious variants to exist in homozygous form.

High correlations between coancestry coefficients computed from different matrices did translate in similar results from the optimisation. The genetic variability retained in the selected candidates when using the different coancestry matrices in OC was similar in terms of the expected heterozygosity and the percentage of segregating alleles. However, the different matrices achieved similar variability by selecting very different number of individuals. In particular, using $\theta_{L\&H}$ and θ_{SEG}, the method only selected 9% and 13% of the initial number of candidates, respectively, in comparison with 47-85% of candidates selected with the other matrices. These differences could be explained because $\theta_{L\&H}$ and θ_{SEG} presented higher variances for coancestry coefficients estimated for the group of 'less related individuals'. Conversely, θ_{SIM}, which selected the highest number of candidates, was also the matrix with lower ability to differentiate relationships, according to its variance. If we take into account that in a selection programme the resources to keep all the individuals we would like are limited, the matrix that selects the least number of candidates (that is $\theta_{L\&H}$) could be the best choice, however, we should be careful as we could create a bottleneck that could eventually give problems. These results agree with those by Eynard *et al*. (2016), since they obtained similar proportions of alleles segregating when none restriction was imposed on the number of selected candidates, however, these varied among methods when restrictions were placed.

In livestock breeding programmes, genomic selection (Meuwissen *et al*., 2001) is becoming a standard procedure for obtaining accurate estimates of the genetic merit of candidates for selection. When genomic inbreeding is controlled through OC, Sonesson *et al*. (2012) showed that genomic coancestry matrices should be used. The choice of the genomic coancestry matrix that is used in OC is important because it will have an impact not only on the diversity maintained, but also on the trajectory of the change in gene frequencies. Previous studies have evidenced that the optimisation of contributions that makes use of matrices based on allelic similarity (Nejati-Javaremi *et al*., 1997) such as θ_{SIM}, benefit solutions that lead the gene frequencies to 0.5, and therefore, to higher genetic variability. However, this is at the cost of changing the genetic composition of the population (Saura *et al*., 2008). On the contrary, the optimisation using the VanRaden's matrices (VanRaden, 2008) could lead to solutions that tend to maintain the gene frequencies although the genetic diversity would be lower. Thus, the choice of the

coancestry matrix used in the optimisation will depend on the emphasis given to each of these conservation aspects.

In aquaculture breeding, the high fecundity typical from fish facilitates obtaining thousands of offspring from one single couple, increasing the risk of high inbreeding rates. The genetic variability of the traits originally included in the breeding objective and those that will be included in the future will condition the success of the programmes in the future. Thus, the way in which the base population is built is fundamental. Traditionally, base populations in aquaculture have been created from a series of wild strains by sampling equal numbers of each strain. However, the increasing availability of genomic information in aquaculture species could help to optimally design base populations through the estimation of relationships within and between candidate strains, and thus optimise the percentage of individuals of each strain. Fernández *et al*. (2014) showed that selecting breeders through OC for the formation of the base population, gives up to 6% higher levels of phenotypic performance at the same level of global diversity than sampling equal numbers from each strain.

As previously mentioned, the genomic information allows the study of patterns of inbreeding and coancestry across the genome, facilitating more refined studies of regions associated with inbreeding depression (Pryce *et al*., 2014; Saura *et al.*, 2015) or affected by selection (Hohenlohe *et al*., 2010; Pemberton *et al*., 2012; Kleinman-Ruiz *et al*., 2016; Doekes *et al*., 2018). In this study, genomic patterns of inbreeding were characterised through similarity (F_{SIM}) and ROH-based (F_{ROH}) coefficients. The non-uniform distribution of both coefficients across the genome observed, reflects local genomic regions that have accumulated higher inbreeding. These regions may harbour genes affected by selection that could be further investigated.

In summary, this thesis has investigated, with empirical data from a commercial population of turbot, the properties of different estimators of coancestry and inbreeding based on genomic information. Coancestry matrices based on these estimators were also evaluated for managing inbreeding when implementing OC. Our results revealed that all the different matrices were efficient for controlling inbreeding and maintaining genetic diversity.

CONCLUSIONS

1. The levels of expected heterozygosity in the population of turbot analysed were low, although similar to those found in commercial populations of Atlantic salmon.

2. The magnitude of the different estimates of coancestry (and inbreeding) coefficients differed greatly. These differences can be explained because the assumed reference population in each case was set at a different generation in the past.

3. Correlations between the different coancestry and inbreeding coefficients were in general high and showed a low dispersion around the regression line, except for those involving coefficients computed using VanRaden and Yang´s methods. In particular, F_{VR2} showed low or even negative correlations with the other inbreeding coefficients. The explanation for these low correlations and high dispersions seem to be related with the higher weight given to rare alleles in VanRaden and Yang´s methods.

4. Genome-wide patterns of inbreeding across the genome were not uniform, based on both F_{SIM} and F_{ROH}. In line with genome-wide estimates, patterns based on F_{SIM} were of higher magnitude than patterns based on F_{ROH}. This information is useful for setting the basis for further studies on homozygosity mapping of genes.

5. The genetic variability retained in the selected candidates when using the different coancestry matrices in OC was similar both in terms of expected heterozygosity and percentage of alleles that remained segregating. However, the different matrices achieved similar variability by selecting very different number of individuals. This result could be explained by a differential ability of the coefficients evaluated to discriminate relationships in the group of ‘less related individuals’.

REFERENCES

Anacleto, O., Cabaleiro, S., Villanueva, B., Saura, M., Houston, R.D. *et al*. 2018. Genetic differences in host infectivity affect disease spread and survival in epidemics. Sci. Rep. (in press).

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L. *et al.* 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3, e3376.

Ballou, J.D. 1997. Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. J. Hered. 88, 169–178.

Bentsen, H.B., Gjerde, B., Eknath, A.E., de Vera, M.S.P., Velasco, R.R. *et al*. 2017. Genetic improvement of farmed tilapias: Response to five generations of selection for increased body weight at harvest in *Oreochromis niloticus* and the further impact of the project. Aquaculture 468, 206–217.

Bjelland, D.W., Weigel, K., Vukasinovic, N., Nkrumah, J.D. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J. Dairy Sci. 96, 4697–4706.

Bouza, C., Sanchez, L., Martínez, P. 1994. Karotypic characterization of turbot (*Scophthalmus maximus*) with conventional, fluorochrome and restriction endonuclease-banding techniques. Mar. Biol. 120, 609–613.

Bouza, C., Hermida, M., Pardo, B.G., Fernández, C., Fortes, G.G. *et al*. 2007. A microsatellite genetic map of the turbot (*Scophthalmus maximus*). Genetics 177, 2457–2467.

Bouza, C., Presa, P., Castro, J., Sánchez, L., Martínez, P. 2002. Allozyme and microsatellite diversity in natural and domestic populations of turbot (*Scophthalmus maximus*) in comparison with other Pleuronectiformes. Can. J. Fish. Aquat. Sci. 59, 1460–1473.

Brown, R.C., Woolliams, J.A., McAndrew, B.J. 2005. Factors influencing effective population size in commercial populations of gilthead seabream, *Sparus aurata*. Aquaculture 247, 219–225.

Browning, S.R., Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097.

Browning, S.R., Browning, B.L. 2011. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12, 703–714.

Caballero, A., Toro, M. 2000. Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genet. Res. 75, 331–343.

Charlesworth, D., Willis, J.H. 2009. The genetics of inbreeding depression. Nat. Rev. Genet. 10, 783–796.

Coughlan, J. P., Imsland, A. K., Galvin, P. T., Fitzgerald, R. D., Naevdal, G., Cross, T. F. 1998. Microsatellite DNA variation in wild populations and farmed strains of turbot from Ireland and Norway: a preliminary study. J. Fish. Biol. 52, 916–922.

Curik, I., Ferenčaković, M., Sölkner, J. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem. Livest. Sci. 166, 26–34.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499–510.

de Cara, M.A.R., Fernández, J., Toro, M.A., Villanueva, B. 2011. Using genome-wide information to minimize the loss of diversity in conservation programmes. J. Anim. Breed. Genet. 128, 456–464.

de Cara, M.A.R., Villanueva, B., Toro, M.A., Fernández, J. 2013. Using genomic tools to maintain diversity and fitness in conservation programmes. Mol. Ecol. 22, 6091–6099.

Doekes, H.P., Veerkamp, R.F., Bijma, P., Hiemstra, S.J., Windig, J.J. 2018. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015. Genet. Sel. Evol. 50, 15.

Engelsma, K.A., Veerkamp, R.F., Calus, M.P.L., Bijma, P., Windig, J.J. 2012. Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. J. Anim. Breed. Genet. 129, 195–205.

Exadactylos, A., Rigby, M.J., Geffen, A.J., Thorpe, J.P. 2007. Conservation aspects of natural populations and captive-bred stocks of turbot (*Scophthalmus maximus*) and Dover sole (*Solea solea*) using estimates of genetic diversity. ICES J. Mar. Sci. 64, 1173–1181.

Eynard, S.E., Windig, J.J., Hiemstra, S.J., Calus, M.P.L. 2016. Whole‐genome sequence data uncover loss of genetic diversity due to selection. Genet. Sel. Evol. 48, 33.

Falconer, D.S., Mackay, F.C. 1996. Introduction to Quantitative Genetics. Longman Group Ltd, Harlow, Essex, England.

FAO. 2016. The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all, Rome, Italy.

Fernández, J., Toro, M.A. 1999. The use of mathematical programming to control inbreeding in selection schemes. J. Anim. Breed. Genet. 116, 447–466.

Fernández, J., Toro, M.A., Caballero, A. 2003. Fixed contributions designs vs. minimization of global coancestry to control inbreeding in small populations. Genetics 165, 885–894.

Fernández, J., Toro, M., Sonesson, A.K., Villanueva, B. 2014. Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. Front. Genet. 5, 414.

Figueras, A., Robledo, D., Corvelo, A., Hermida, M., Pereiro, P. *et al*. 2016. Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): A fish adapted to demersal life. DNA Res. 23, 181–192.

Frankham, R., Ballou, J.D., Briscoe, D.A. 2002. Introduction to Conservation Genetics. Cambridge University Press, Cambridge, United Kingdom.

Gallardo, J.A., García, X., Lhorente, J.P., Neira, R. 2004. Inbreeding and inbreeding depression of female reproductive traits in two populations of coho salmon selected using BLUP predictors of breeding values. Aquaculture 234, 111–122.

Gibson, J., Morton, N.E., Collins, A. 2006. Extended tracts of homozygosity in outbred human populations. Hum. Mol. Genet. 15, 789–795.

Gjedrem, T. 2010. The first family-based breeding program in aquaculture. Rev. Aquacult. 2, 2–15.

Gjedrem, T. 2012. Genetic improvement for the development of efficient global aquaculture: A personal opinion review. Aquaculture 344, 12–22.

Gjedrem, T., Rye, M. 2016. Selection response in fish and shellfish: a review. Rev. Aquacult. 10, 168–119.

Gjerde, B., Gjøen, H.M., Villanueva, B. 1996. Optimum designs for fish breeding programmes with constrained inbreeding mass selection for a normally distributed trait. Livest. Sci. 47, 59–72.

Gómez-Romano, F., Villanueva, B., de Cara, M.A.R, Fernández, J. 2013. Maintaining genetic diversity using molecular coancestry: The effect of marker density and effective population size. Genet. Sel. Evol. 45, 38.

Gómez-Romano, F., Villanueva, B., Sölkner, J., de Cara, M.A.R., Mészáros, G. *et al.* 2016. The use of coancestry based on shared segments for maintaining genetic diversity. J. Anim. Breed. Genet. 133, 357–365.

Grundy, B., Villanueva, B., Wooliams, J.A. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genet. Res. 72, 159–168.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D. *et al.* 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19, 318–326.

Hermida, M., Bouza, C., Fernández, C., Sciara, A.A., Rodríguez-Ramilo, S.T. *et al.* 2013. Compilation of mapping resources in turbot (*Scophthalmus maximus*): A new integrated consensus genetic map. Aquaculture 414, 19–25.

Hohenlohe, P.A., Phillips, P.C., Cresko, W.A. 2010. Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. Int. J. Plant Sci. 171, 1059–1071.

Janssen, K., Chavanne, H., Berentsen, P., Komen, H. 2017. Impact of selective breeding on European aquaculture. Aquaculture 472, 8–16.

Keller, M.C., Visscher, P.M., Goddard, M.E. 2011. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189, 237–249.

Kijas, J., Elliot, N., Kube, P., Evans, B., Botwright, N. *et al*. 2016. Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. Anim. Genet. 48, 237–241.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. 1983. Optimisation by simulated annealing. Science 220, 671–680.

Kleinman-Ruiz, D., Villanueva, B., Fernández, J., Toro, M.A., García-Cortés, L.A., Rodríguez-Ramilo, S.T. 2016. Intra-chromosomal estimates of inbreeding and coancestry in the Spanish Holstein cattle population. Livest. Sci. 185, 34–42.

Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E., Morgan, T.V. *et al.* 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. PNAS 104, 19942–19947.

Li, C.C., Horvitz, D.G. 1953. Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. 5, 107–117.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie, Paris, France.

Maroso, F., Hermida, M., Millán, A., Blanco, A., Saura, M. *et al*. 2018. Highly dense linkage maps from 31 segregant populations of turbot (*Scophthalmus maximus*) provide insights into recombination patterns and chromosome rearrangements throughout a new refined genome assembly. DNA Res. doi: 10.1093/dnares/dsy015.

Martínez, P., Bouza, C., Hermida, M., Fernández, J., Toro, M.A. *et al.* 2009. Identification of the major sex-determining region of turbot (*Scophthalmus maximus*). Genetics 183, 1443–1452.

Martínez, P., Robledo, D., Rodríguez-Ramilo, S., Hermida, M., Taboada, X. *et al*. 2016. Turbot (*Scophthalmus maximus)* genomic resources: application for boosting aquaculture production. In: MacKenzei, S., Jentoft, S. (eds.). Genomics in Aquaculture. Elsevier, London, United Kingdom. pp. 131–163.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M. *et al.* 2008. Runs of Homozygosity in European Populations. Am. J. Hum. Genet. 83, 359–372.

Meuwissen, T.H.E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75, 934–940.

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

Millán, A., Gómez-Tato, A., Pardo, B.G., Fernández, C., Bouza, C. *et al*. 2011. Gene expression profiles of the spleen, liver, and head kidney in turbot (*Scophthalmus maximus*) along the infection process with *Aeromonas salmonicida* using an immune-enriched oligo-microarray. Mar. Biotechnol. 13, 1099–1114.

Navajas-Pérez, R., Robles, F., Molina-Luzón, M.J., De La Herrán, R., Álvarez-Dios, J.A. *et al*. 2012. Exploitation of a turbot (*Scophthalmus maximus* L.) immune-related expressed sequence tag (EST) database for microsatellite screening and validation. Mol. Ecol. Resour. 12, 706–716.

Nejati-Javaremi, A., Smith, C., Gibson, J.P. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75, 1738–1745.

Nielsen, E.E., Nielsen, P.H., Meldrup, D., Hansen, M.M. 2004. Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. Mol. Ecol. 13, 585–595.

Pante, M.J.R., Gjerde, B., McMillan, I. 2001. Effect of inbreeding on body weight at

harvest in rainbow trout, *Oncorhynchus mykiss*. Aquaculture 192, 201–211.

Pardo, B.G., Millán, A., Gómez-Tato, A., Fernández, C., Bouza, C. *et al*. 2012. Gene expression profiles of the spleen, liver, and head kidney in turbot (*Scophthalmus maximus*) along the infection process with *Philasterides dicentrachi* using an immune-enriched oligo-microarray. Mar. Biotechnol. 14, 570–582.

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., Li, J.Z. 2012. Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. 91, 275–292.

Pereiro, P., Figueras, A., Novoa, B. 2016. Turbot (*Scopthalmus maximus*) vs VHSV (Viral Haemorrhagic Septicemia Virus): A review. Front. Physiol. 7, 192.

Peripolli, E., Munari, D.P., Silva, M.V.G.B., Lima, A.L.F., Irgang, R., Baldi, F. 2017. Runs of homozygosity: current knowledge and applications in livestock. Anim. Genet. 48, 255–271.

Pryce, J.E., Haile-Mariam, M., Goddard, M.E., Hayes, B.J. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet. Sel. Evol. 46, 71.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R. *et al.* 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Purfield, D.C., McParland, S., Wall, E., Berry, D.P. 2017. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. PLoS ONE 12, e0176780.

Purfield, D.C., Berry, D.P., McParland, S., Bradley, D.G. 2012. Runs of homozygosity and population history in cattle. BMC Genet. 13, 70.

Ren, W., Hu, L., Guo, L., Zhang, J., Tang, L. *et al*. 2018. Preservation of the genetic diversity of a local common carp in the agricultural heritage rice–fish system. PNAS. doi: 10.1073/pnas.1709582115.

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., Houston, R. 2017. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev. Aquac. doi: 10.1111/raq.12193.

Rodríguez-Ramilo, S.T., De La Herrán, R., Ruiz-Rejón, C., Hermida, M., Fernández, C. *et al*. 2014. Identification of quantitative trait loci associated with resistance to viral haemorrhagic septicaemia (VHS) in turbot (*Scophthalmus maximus*): A

comparison between bacterium, parasite and virus diseases. Mar. Biotechnol. 16, 265–276.

Rodríguez-Ramilo, S.T., Fernández, J., Toro, M.A., Bouza, C., Hermida, M. *et al*. 2013. Uncovering QTL for resistance and survival time to *Philasterides dicentrarchi* in turbot (*Scophthalmus maximus*). Anim. Genet. 44, 149–157.

Rodríguez-Ramilo, S.T., Fernández, J., Toro, M.A., Hernández, D., Villanueva, B. 2015. Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. PLoS ONE 10(4), e0124157.

Rodríguez-Ramilo, S.T., Toro, M.A., Bouza, C., Hermida, M., Pardo, B.G. *et al.* 2011. QTL detection for *Aeromonas salmonicida* resistance related traits in turbot (*Scophthalmus maximus*). BMC Genomics 12, 541.

Sánchez-Molano, E., Cerna, A., Toro, M.A., Bouza, C., Hermida, M. *et al*. 2011. Detection of growth-related QTL in turbot (*Scophthalmus maximus*). BMC Genomics 12, 473.

Saura, M., Caballero, A., Santiago, E., Morales, E., Fernández, A. *et al.* 2018. The importance of ensuring genetic variability when establishing selection programmes in aquaculture. XIX Reunión Nacional de Mejora Animal, León, Spain, June 14–15.

Saura, M., Fernández, A., Varona, L., Fernández, A.I., de Cara, M.A.R. *et al*. 2015. Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. Genet. Sel. Evol. 47, 1.

Saura, M., Pérez-Figueroa, A., Fernández, J., Toro, M.A., Caballero, A. 2008. Preserving population allele frequencies in ex situ conservation programs. Conserv. Biol. 22, 1277–1287.

Skaala, Ø., Høyheim, B., Glover, K., Dahle, G. 2004. Microsatellite analysis in domesticated and wild Atlantic salmon (*Salmo salar* L.): allelic diversity and identification of individuals. Aquaculture 240, 131–143.

Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E. 2012. Genomic selection requires genomic control of inbreeding. Genet. Sel. Evol. 44, 27.

Su, G.S., Liljedahl, L.E., Gall, G.A.E. 1996. Genetic and environmental variation of body weight in rainbow trout (*Oncorhynchus mykiss*). Aquaculture 144, 71–80.

Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Zöllner, S. *et al*. 2013. Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. 93, 90–102.

Thodesen, J., Gjedrem, T. 2006. Breeding programs on Atlantic salmon in Norway - Lessons learned. In: Ponzoni, R.W., Acosta, B.O., Ponniah, A.G. (eds.) Development of aquatic animal genetic improvement and dissemination programs: current status and action plans. WorldFish Center, Penang, Malaysia. pp. 22–26.

Toro, M.A., Barragán, C., Óvilo, C., Rodrigáñez, J., Rodríguez, C., Silió, L. 2002. Estimation of coancestry in Iberian pigs using molecular markers. Conserv. Genet. 3, 309–320.

Toro, M.A., Fernández, J., Caballero, A. 2009. Molecular characterization of breeds and its use in conservation. Livest. Sci. 120, 174–195.

Toro, M.A., Villanueva, B., Fernández, J. 2014. Genomics applied to management strategies in conservation programmes. Livest. Sci. 166, 48–53.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.

Vera, M., Álvarez-Dios, J.A., Fernández, C., Bouza, C., Vilas, R., Martínez, P. 2013. Development and validation of single nucleotide polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (*Scophthalmus maximus*) using high-throughput genotyping. Int. J. Mol. Sci. 14, 5694–5711.

Vera, M., Álvarez-Dios, J.A., Millán, A., Pardo, B.G., Bouza, C. *et al.* 2011. Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, *Scophthalmus maximus*, database. Aquaculture 313, 31–41.

Villanueva, B., Pong-Wong, R., Woolliams, J.A., Avendaño, S. 2004. Managing genetic resources in selected and conserved populations. In: Simm, G., Villanueva, B., Sinclair, K.D., Townsend, S. (eds.). Farm Animal Genetic Resources. BSAS Nottingham University Press, Nottingham, United Kingdom. pp. 113–132.

Villanueva, B., Sawalha, R.M., Roughsedge, T., Rius-Vilarrasa, E., Woolliams, J.A. 2010. Development of a genetic indicator of biodiversity for farm animals. Livest. Sci. 129, 200–207.

Villanueva, B., Woolliams, J.A., Gjerde, B. 1996. Optimum designs for breeding programmes under mass selection with an application in fish breeding. Anim. Sci. 63, 563–576.

Wang, J. 2014. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. J. Evol. Biol. 27, 518–530.

Wang, S., Meyer, E., McKay, J.K., Matz, M.V. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods 9, 808–810.

Wang, W., Hu, Y., Ma, Y., Xu, L., Guan, J. *et al*. 2015. High-density genetic linkage mapping in turbot (*Scophthalmus maximus L*.) based on SNP markers and major sex- and growth-related regions detection. PLoS ONE 10, e0120410.

Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., Hill, W.G. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 15, 1468–1476.

Woolliams, J.A., Berg, P., Dagnachew, B.S., Meuwissen, T.H.E. 2015. Genetic contributions and their optimisation. J. Anim. Breed. Genet. 132, 89–99.

Wright, S. 1921. Systems of mating. Genetics 6, 111–178.

Yang, J., Benyamin, B., Mcevoy, B.P., Gordon, S., Henders, A.K. *et al*. 2010. Common SNPs explain a large proportion of heritability for human height. Nat. Genet. 42, 565–569.

Yáñez, J.M., Bassini, L.N., Filp, M., Lhorente, J.P., Ponzoni, R.W., Neira, R. 2014. Inbreeding and effective population size in a coho salmon (*Oncorhynchus kisutch*) breeding nucleus in Chile. Aquaculture 420, S15–S19.

Zhang, Q., Calus, M.P.L., Guldbrandtsen, B., Lund, M.S., Sahana, G. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genet. 16, 88.