UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS

DEPARTAMENTO DE GENÉTICA

TESIS DOCTORAL

Gestión de la diversidad genética en programas de conservación utilizando datos de genotipado masivo

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Fernando Gómez Romano

Directores

Jesús Fernández Martín
Beatriz Villanueva Gaviña
Fernando Gómez Romano

Madrid, 2015

UNIVERSIDAD COMPLUTENSE DE MADRID

Gestión de la diversidad genética en programas de conservación utilizando datos de genotipado masivo

Fernando Gómez Romano

Tesis doctoral
2015

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE BIOLOGÍA

Departamento de Genética

Gestión de la diversidad genética en programas de conservación utilizando datos de genotipado masivo

Memoria para optar al grado de Doctor en Biología
Mención europea
Fernando Gómez Romano

VºBº Director	VºBº Directora	El Doctorando
Jesús Fernández Martín	Beatriz Villanueva Gaviña	Fernando Gómez Romano

Madrid, abril de 2015

Esta Tesis Doctoral se realizó en el Instituto de Investigación y Tecnología Agraria y Alimentaria (INIA) de Madrid, y ha sido financiada con una beca predoctoral FPI y con los proyectos CGL2009-13278-C02-02 y CGL2012-39861-C02-02 (Plan Estatal de I+D+i).

Los trabajos descritos en los Capítulos 2 y 3 se realizaron en parte durante estancias predoctorales en *The Roslin Institute* de la *University of Edinburg*h (Edimburgo, Reino Unido) y en el *Institut für Nutztierwissenschaften* de la *Universität für Bodenkultur* (Viena, Austria), respectivamente.

Los capítulos de esta Tesis se corresponden con las siguientes publicaciones científicas:

1. GÓMEZ-ROMANO F, VILLANUEVA B, DE CARA MAR & FERNÁNDEZ J (2013). Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. *Genetics Selection Evolution* **45**:38−45.
2. GÓMEZ-ROMANO F, VILLANUEVA B, FERNÁNDEZ J, WOOLLIAMS JA & PONG-WONG R. The use of genomic coancestry matrices in the optimisation of contributions for maintaining genetic diversity at specific regions of the genome. *Genetics Selection Evolution* (en revisión).
3. GÓMEZ-ROMANO F, SÖLKNER J, VILLANUEVA B, MESZAROS G, DE CARA MAR, PÉREZ O'BRIEN AM & FERNÁNDEZ J. The use of identical by descent segments for maintaining genetic diversity. *Journal of Animal Breeding and Genetics* (en preparación).

ÍNDICE

RESUMEN

Gestión de la diversidad genética en programas de conservación utilizando datos de genotipado masivo

El mantenimiento de la diversidad genética existente en una población es uno de los principales objetivos de los programas de conservación, tanto de especies salvajes como de especies domésticas. Esto es así porque la capacidad de respuesta de la población a nuevas condiciones ambientales, de la que va a depender su posibilidad de supervivencia, es proporcional a la magnitud de dicha diversidad. A más corto plazo, otro de los objetivos de los programas de conservación es controlar la tasa a la cual aumenta la consanguinidad para evitar así la pérdida de eficacia biológica de la población, un fenómeno conocido como depresión consanguínea.

Hoy en día se acepta que el método más eficiente para controlar la pérdida de variabilidad genética y el aumento de la consanguinidad es el método de Contribuciones Óptimas (OC, del inglés “Optimal Contributions”). Este método optimiza las contribuciones (número de hijos) de todos los reproductores potenciales a la siguiente generación con el objetivo de minimizar el parentesco promedio. El elemento fundamental en el que se basa el método OC es la matriz de parentesco, que contiene los coeficientes de parentesco entre todos los candidatos. Tradicionalmente, dicha matriz de parentesco se ha obtenido a partir de registros genealógicos pero también es posible obtenerla a partir de marcadores moleculares. Estudios previos han demostrado que cuando la densidad de marcadores (ej., marcadores tipo microsatélite) es relativamente baja, el uso del parentesco molecular es de limitado valor para el mantenimiento de la diversidad genética en comparación con el uso del parentesco genealógico.

El desarrollo de las plataformas de genotipado masivo de polimorfismos de un solo nucleótido (SNPs) llevado a cabo en los últimos años hace necesaria la revaluación

del beneficio de utilizar marcadores genéticos en la gestión de poblaciones. De hecho, la información molecular contenida en los paneles densos de SNPs aporta varias ventajas frente a la información genealógica: i) puede obtenerse para todo tipo de poblaciones y a partir de diferentes tipos de muestras; ii) permite conocer la proporción exacta del genoma que es compartida por dos individuos en lugar de la proporción promedio esperada que aporta la información genealógica; y iii) permite calcular el parentesco en regiones específicas del genoma. Al permitir diferenciar el grado de relación entre pares de individuos con el mismo parentesco genealógico, se puede esperar que la eficiencia en la gestión de una población para mantener variabilidad genética optimizando contribuciones aumente al sustituir el parentesco genealógico por el molecular, siempre que éste esté calculado con un número elevado de marcadores (parentesco genómico).

En esta tesis se ha evaluado el uso de información genómica para la gestión de diversidad genética en programas de conservación, empleando el método OC para minimizar el parentesco genómico en la descendencia. Cuando los apareamientos se producen al azar la consanguinidad promedio en una generación es igual al parentesco promedio de la generación previa, por lo que la minimización del parentesco lleva implícitamente asociado un control de la consanguinidad en la población.

La eficiencia del uso de parentesco genómico en la gestión de poblaciones para mantener diversidad genética depende del desequilibrio de ligamiento existente entre los marcadores y el resto de loci donde se quiere mantener diversidad. A su vez, el desequilibrio de ligamiento depende de la densidad de marcadores y del censo efectivo de la población (N_e). Así pues, a la hora de desarrollar nuevos chips de SNPs para especies en programas de conservación, es importante conocer cual sería la densidad de SNPs necesaria para obtener al menos la misma diversidad que la obtenida utilizando información genealógica, en la gestión de poblaciones. En esta tesis, se demuestra, a

través de simulación por ordenador, que una densidad de $3N_e$ SNPs/Morgan es suficiente para mantener la misma cantidad de diversidad genética (medida como heterocigosidad esperada) utilizando el parentesco molecular que el genealógico en la optimización de contribuciones (Capítulo 1). Densidades más altas de SNPs llevarían a una ventaja del parentesco molecular frente al genealógico. La densidad de los chips ya desarrollados actualmente para diferentes especies (fundamentalmente para especies domésticas) es suficientemente alta como para que el parentesco molecular sea una potente herramienta para la conservación de diversidad genética. En el Capítulo 1 también se demuestra que el beneficio que se obtiene por aumentar la densidad por encima de alrededor de 500 SNPs/Morgan, es muy pequeño en cuanto a la variabilidad mantenida.

Como ya se ha mencionado anteriormente, otra de las ventajas de utilizar el parentesco genómico es que nos permite enfocar la gestión en el mantenimiento de la diversidad genética en regiones específicas del genoma. En particular, puede ser de interés minimizar la pérdida de la diversidad genética en una región específica del genoma. En esta tesis, se demuestra, a través de simulación por ordenador, que el método OC, utilizando programación semidefinida y el parentesco, calculado con SNPs mapeados en regiones específicas, es muy eficiente a la hora de mantener (e incluso aumentar) la diversidad (medida como heterocigosidad esperada) en dichas regiones. También permite restringir la consiguiente pérdida de diversidad en el resto del genoma. Sin embargo, aunque los niveles de diversidad en el resto del genoma son mayores cuando se incluye la restricción que cuando no se incluye, la tasa de parentesco observada fue mayor que la restricción impuesta. Esto nos ha llevado a concluir que es necesario refinar la teoría de las contribuciones cuando se utilizan matrices genómicas,

para asegurar que las restricciones están debidamente consideradas en el modelo (Capítulo 2).

En los Capítulos 1 y 2 de esta tesis, el coeficiente de parentesco molecular se ha definido como la probabilidad de que dos alelos escogidos al azar de cualquier locus, uno de cada individuo, sean iguales. Es decir, este parentesco refleja tanto identidad por descendencia (IBD) como identidad en estado (IBS), a diferencia del parentesco genealógico que solo refleja IBD. Una medida alternativa de parentesco molecular, que refleja mejor IBD, es aquella basada en segmentos IBD, que son regiones de SNPs consecutivos que son iguales en dos individuos. En el Capítulo 3 se ha evaluado el uso del parentesco calculado a partir de segmentos IBD en el mantenimiento de diversidad genética, utilizando datos genómicos procedentes de poblaciones de vacuno de tres razas austriacas (entre 219 y 465 individuos por raza, genotipados para cerca de 40.000 SNPs). Esta medida de parentesco se ha comparado con el parentesco genómico mencionado previamente y con el parentesco genealógico. El censo efectivo obtenido a partir las tasas de los tres tipos de parentesco fueron muy similares en las tres poblaciones y de pequeña magnitud (estimas en los rangos 26 − 52, 61 − 93 y 84 − 112 para las tres razas). Este resultado destaca la importancia de la aplicación de estrategias activas de gestión para controlar el aumento de la consanguinidad y parentesco y la pérdida de la diversidad genética en las razas de especies ganaderas, incluso si los tamaños poblacionales son razonablemente grandes.

Uno de los principales problemas que presenta el cálculo del parentesco IBD es la necesidad de conocer las fases gaméticas de los SNPs utilizados, ya que, dadas las técnicas empleadas para la obtención de los genotipos, estas fases son desconocidas. En el Capítulo 3 se demuestra, a través simulación por ordenador, que la necesidad de estimar las fases gaméticas para obtener el parentesco basado en segmentos IBD (en

cada una de las generaciones de gestión de la población) no lleva consigo una pérdida en la diversidad mantenida. Esto es incluso cierto cuando las poblaciones son muy pequeñas, tal y como suelen ser aquellas objeto de un programa de conservación.

SUMMARY

Management of genetic diversity in conservation programmes based on massive genotyped data

The maintenance of genetic diversity is one of the main objectives of conservation programmes, for both wildlife and domestic species. This is because the ability of the populations to adapt to new environmental conditions, on which their probability to survive depends, is proportional to the amount of diversity. In the short term, another objective of conservation programmes is to control the rate of inbreeding in order to avoid inbreeding depression (i.e., loss of biological fitness).

Nowadays it is commonly accepted that the most efficient method to control the loss of genetic diversity and the increase of inbreeding is the Optimal Contributions method (OC). This method optimises the contributions (number of offspring) of all potential breeders to the next generation with the aim of minimising the average coancestry. The central element of the OC method is the coancestry matrix, which contains the coancestry coefficients between all breeding candidates. Traditionally, this coancestry matrix has been computed from genealogical records but it is also possible to compute it from molecular information. Previous studies have shown that when the marker density is relatively low (e.g., microsatellite markers), the usefulness of molecular coancestry is lower than that of genealogical coancestry for maintaining genetic diversity.

The development of high-throughput massive genotyping methods for single nucleotide polymorphisms (SNPs) markers that has been conducted in recent years has led to the need for a reassessment of the efficiency of molecular markers for

maintaining diversity through the genetic management of populations. The molecular information obtained from dense SNP panels provides several advantages compared with genealogical information: i) molecular information can be obtained from all kind of populations and from different types of samples; ii) molecular information provides the proportion of the genome that is shared by two individuals rather than the expected average relationship that is provided by genealogical information; and iii) molecular information allows us to calculate coancestry at specific regions of the genome. Given that molecular coancestry discriminates between pairs of individuals with the same genealogical coancestry coefficient, it can be expected that the use of molecular coefficients instead of genealogical coefficients will increase the efficiency of population management for maintaining genetic diversity, provided it is calculated with a large enough number of markers (genomic coancestry).

In this thesis the use of genomic information for managing genetic diversity in conservation programmes has been evaluated when emplying the OC method to minimise the genomic coancestry. With random mating, the average inbreeding in a particular generation is equal to the average coancestry between individuals of the previous generation. Therefore, minimising coancestry leads implicitly to the control of inbreeding in the population.

The efficiency of using genomic coancestry in population management for maintaining genetic diversity relies on the existing linkage disequilibrium between markers and loci where diversity is required to be maintained. In turn, linkage disequilibrium depends on marker density and population effective size (N_e). Thus, when new SNP chips are developed with the aim of maintaining diversity, it is

important to know the SNPs density required to obtain at least the same diversity levels than those obtained when genealogical information is used. In this study, it is demonstrated, through computer simulations, that a density of $3N_e$ SNPs/Morgan is enough for maintaining the same amount of genetic diversity (measured as expected heterozygosity) using genomic coancestry than that maintained using genealogical coancestry (Chapter 1). Higher SNP densities lead molecular coancestry to outperform genealogical coancestry. The SNP density of the chips currently available in commercial SNP panels, mainly for farm animal species with high commercial value is high enough for genomic coancestry be a powerful tool for maintaining genetic diversity. In Chapter 1 is also shown that the advantage obtained by increasing the density above 500 SNPs/Morgan is very small in terms of the variability maintained.

Another advantage of using genomic coancestry coefficients is the possibility of focusing the genetic management on maintaining genetic diversity at specific genomic regions. In particular, it can be of interest to minimise the loss of genetic diversity in a specific region of the genome. In this thesis it is shown, through computer simulations, that the OC method is very efficient in maintaining (or even increasing) diversity (measured as expected heterozygosity) in specific regions of the genome when using semidefinite programming and a coancestry coefficient computed using the SNPs mapped on those regions. The method also allows us to efficiently restrict the loss of diversity in the rest of the genome. However, although the level of diversity kept in the rest of the genome is higher when including a restriction on the rate of coancestry in the optimisation procedure, the observed coancestry rate exceeded the value of the restriction. This result has led to the conclusion that it is necessary to refine the theory

of genetic contributions when using genomic matrices in order to ensure that restrictions are properly considered in the model (Chapter 2).

In Chapters 1 and 2, the molecular coancestry coefficient has been defined as the the probability that two alleles taken at random from each individual at the same locus are equal. Thus, this coancestry coefficient reflects both identity by descent (IBD) and identity by state (IBS) in contrast with the genealogical coancestry that only reflects IBD. An alternative measure of molecular coancestry that is likely to better reflect IBD is that based on IBD segments (regions of consecutive SNPs shared by two individuals). In Chapter 3, the application of coancestry coefficients based on IBD segments for maintaining genetic diversity has been evaluated using genomic data from three populations of three different Austrian cattle breeds (from 219 to 465 individuals by breed, genotyped for around 40,000 SNPs). This coancestry measure has been compared with the genomic coancestry coefficient previously described and with the genealogical coancestry. The effective population size obtained from the rates of the three types of coancestry very similar and of small magnitude (estimated ranges were 26 − 52, 61 − 93 and 84 − 112 for the three breeds). This result highlights the importance of implementing active management strategies to control the increase of inbreeding and coancestry and the loss of genetic diversity in livestock breeds, even when the population size is reasonably large.

One of the main problems to be solved when calculating coancestry based on IBD segments is the need of estimating the gametic phases of the SNPs used that given the techniques used to obtain the genotypes, are unknown. In Chapter 3 it is shown, through computer simulations, that using estimates of gametic phases for computing

coancestry based on IBD segments (in each of the management generations) does not lead to any loss in the diversity maintained. This has proven to be true even when the size of the population is very small that is the usual situation in conservation programmes.

INTRODUCCIÓN GENERAL

El cambio en las condiciones ambientales a las que se ve sometida una población es un proceso continuo y de la capacidad de respuesta a dicho cambio van a depender sus posibilidades de supervivencia. La diversidad genética, que puede definirse como la variedad de alelos, genotipos y haplotipos presentes en una población, refleja su potencial para evolucionar. Debido a esto, el mantenimiento de diversidad genética es uno de los objetivos principales de un programa de conservación (FRANKHAM et al., 2002), tanto de poblaciones salvajes como de poblaciones domésticas. A más corto plazo, es necesario también controlar la consanguinidad y evitar así problemas de depresión consanguínea (pérdida de eficacia biológica como consecuencia de la consanguinidad), que es debida principalmente a la expresión de alelos recesivos deletéreos (ROFF, 1997).

Las tres medidas más habituales de diversidad genética son i) la heterocigosidad esperada (EH), que es la heterocigosidad que estaría presente en una población en equilibrio de Hardy-Weinberg con las mismas frecuencias alélicas que la población de interés; ii) la heterocigosidad observada (OH), que es la proporción de individuos heterocigotos en la población; y iii) la diversidad alélica (AD), que es el número de alelos segregando en los individuos de una población (TORO et al., 2009). Estas definiciones refieren a un solo locus pero pueden promediarse para todos los loci del genoma.

En poblaciones pequeñas, tal y como suelen ser aquellas objeto de un programa de conservación, el principal factor de disminución de la diversidad es la deriva genética. La pérdida de diversidad genética debida a deriva depende del tamaño de la población, más exactamente del censo efectivo de la población (N_e), que es el número de individuos de una población ideal teórica con la misma tasa de pérdida de variabilidad genética que la observada en la población de interés (WRIGHT, 1938). Por lo

tanto, la estrategia más adecuada para conservar la diversidad es maximizar N_e, lo que es equivalente a minimizar la tasa de parentesco (Δf), debido a la relación inversa que existe entre ambos ($\Delta f = 1/2N_e$) (FALCONER & MACKAY, 1996). Sin embargo, la depresión consanguínea no depende del coeficiente de parentesco (f), sino del de consanguinidad (F), por lo que la estimación y control de ambos parámetros es de importancia fundamental en los programas de conservación.

El coeficiente de consanguinidad de un individuo se define como la probabilidad de que porte, en un locus elegido al azar, dos alelos idénticos por descendencia (IBD), es decir, que sean dos copias de un mismo alelo ancestral (MALECOT, 1948). Este parámetro tiene una relación directa con la heterocigosidad observada, ya que $F = 1 - OH$. Por su parte, el coeficiente de parentesco se define como la probabilidad de que dos alelos de un locus particular escogidos al azar, uno de cada individuo, sean idénticos por descendencia (MALECOT, 1948). En este caso, se puede demostrar que $EH = 1 - f$.

El control de las tasas a las cuales aumentan el parentesco (Δf) y la consanguinidad (ΔF) es fundamental no solamente en programas de conservación sino también en programas de selección. De hecho, una gran cantidad de investigación se ha realizado en las últimas décadas en el contexto de los programas de mejora genética animal sobre distintos métodos para controlar ΔF y evitar las consecuencias negativas de la consanguinidad. En realidad, la diferencia fundamental entre los programas de selección y de conservación está en el énfasis relativo que se da a la respuesta a la selección y a la tasa de consanguinidad (o de parentesco). En programas de selección, el objetivo es maximizar la respuesta genética para caracteres de interés económico, imponiendo una restricción a la tasa de consanguinidad, mientras que en programas de conservación, el objetivo principal es minimizar la tasa de parentesco imponiendo o no una restricción a la respuesta para un determinado carácter que haga valiosa a la

población. Así pues, las estrategias empleadas para gestionar el nivel de diversidad y el control de la consanguinidad son válidas tanto para programas de mejora genética como para programas de conservación.

Hay dos tipos de decisiones que se tienen que tomar en la gestión genética de una población: las decisiones relativas a la elección de los individuos que van a contribuir a la siguiente generación (decisiones de selección) y aquellas relativas a cómo se aparean los individuos seleccionados (decisiones de apareamiento). A corto plazo, las decisiones tomadas en cuanto a qué animales seleccionar son las únicas que influyen en el control de la pérdida de variabilidad genética. No obstante, las estrategias de apareamiento resultan útiles para controlar la tasa de consanguinidad (TORO et al., 1988; WOOLLIAMS, 1989; SANTIAGO & CABALLERO, 1995; CABALLERO et al., 1996).

Muchos de los métodos descritos para controlar la consanguinidad y el parentesco (y por lo tanto, la pérdida de la variabilidad genética) comenzaron a desarrollarse en el contexto de la mejora genética animal. En un principio las tasas de respuesta genética y de consanguinidad fueron consideradas por separado (TORO & PÉREZ-ENCISO, 1990; VILLANUEVA et al., 1994) pero posteriormente, se propusieron métodos para tratar ambas tasas simultáneamente al elegir los individuos seleccionados. En concreto, estos métodos consisten en optimizar las contribuciones de los candidatos a la selección de manera que la respuesta sea máxima restringiendo la tasa de consanguinidad simultáneamente (MEUWISSEN, 1997; GRUNDY et al., 1998; MEUWISSEN & SONESSON, 1998; GRUNDY et al., 2000). En la actualidad se acepta que estos métodos de optimización de contribuciones (OC) son los que obtienen un mejor resultado tanto en programas de selección (VILLANUEVA et al., 2004) como en programas de conservación (SONESSON & MEUWISSEN, 2000; FERNÁNDEZ & CABALLERO, 2001; FERNÁNDEZ et al., 2003; VILLANUEVA et al., 2004). Cuando el objetivo es conservar la

diversidad genética, los métodos OC nos proporcionan el número óptimo de descendientes con que debe contribuir cada candidato a reproductor para minimizar el parentesco global en la siguiente generación. Si los apareamientos subsiguientes son al azar, la consanguinidad promedio en una determinada generación será igual al parentesco promedio de la generación anterior (FALCONER & MACKAY, 1996). Por tanto la minimización del parentesco lleva implícitamente asociado un control de la consanguinidad.

La herramienta fundamental para optimizar las contribuciones de los padres potenciales a través de los métodos OC es la matriz de parentesco, que contiene los coeficientes entre todos los candidatos a producir la siguiente generación. Tradicionalmente, los coeficientes de parentesco y de consanguinidad se han obtenido a partir de registros genealógicos (WRIGHT, 1922; EMIK & TERRILL, 1949). La información genealógica nos permite calcular los niveles de consanguinidad esperada para cada individuo y el parentesco esperado entre ellos (MALECOT, 1948, FALCONER & MACKAY, 1996). Sin embargo, la obtención de registros genealógicos no siempre es posible (por ejemplo, en especies salvajes o en razas ganaderas criadas en régimen extensivo) y, cuando lo es, está sujeta a errores. Por ejemplo, en poblaciones de especies ganaderas, donde la obtención de registros genealógicos es una práctica común, OLIEHOEK & BIJMA (2009) estimaron que la tasa de error en la asignación de padres es del orden de un 10%. Estos autores también investigaron la pérdida de eficiencia de los métodos OC en el mantenimiento de la diversidad genética cuando existen errores en los datos genealógicos. Los resultados de sus simulaciones indicaron que cuando la tasa de error es alta (por ejemplo, cuando la tasa de error en la asignación del padre está por encima de un 35%) estos métodos llevan a una diversidad menor que la obtenida

simplemente igualando las contribuciones de los candidatos (un método que no requiere información genealógica).

Frente a la información genealógica, la información molecular aporta una serie de ventajas de especial importancia para su empleo en el mantenimiento de diversidad. La información molecular puede obtenerse para todo tipo de poblaciones, en un procedimiento estandarizado y a partir de prácticamente cualquier muestra de origen animal. Esto es particularmente importante en poblaciones silvestres, para las que en muchas ocasiones es muy complicado o imposible identificar de manera fiable las relaciones genealógicas entre sus individuos (GARANT & KRUUK, 2005; KELLER et al., 2011). Otra clara ventaja de la información molecular respecto a la genealógica es que permite conocer la proporción exacta de alelos compartidos por dos individuos frente al valor promedio esperado proporcionado por la información genealógica.

Así pues, el parentesco molecular (definido como la probabilidad de que dos alelos escogidos al azar de cualquier locus, uno de cada individuo, sean iguales) puede sustituir al parentesco genealógico en la gestión de poblaciones, con el objetivo de obtener una mayor eficiencia a la hora de mantener la variabilidad genética. Sin embargo, dicha eficiencia depende del desequilibrio de ligamiento (LD) existente entre los marcadores y el resto de loci en el genoma donde se quiere mantener la diversidad genética. A su vez, el LD depende de la densidad de marcadores utilizada y del N_e histórico de la población. La densidad tiene que ser suficientemente alta para que la gestión de poblaciones basada en información molecular sea eficiente. De hecho, utilizando una densidad relativamente baja de marcadores de tipo microsatélite, FERNÁNDEZ et al. (2005) concluyeron que el uso del parentesco molecular era de limitado valor para el mantenimiento de diversidad genética en comparación con el uso del parentesco genealógico.

En los últimos años se ha producido un rápido desarrollo de la genómica, que ha permitido la creación de paneles densos de polimorfismos de un solo nucleótido (SNPs) en multitud de animales de granja (ganado vacuno, ovino, porcino, avícola y caballar y salmón). La densidad de SNPs es muy superior a la disponible para cualquier otro tipo de marcador molecular utilizado previamente. Esto ha permitido incrementar el LD y con ello aumentar la eficiencia de los marcadores moleculares en el mantenimiento de diversidad (BJELLAND et al., 2013). Con un número tan alto de SNPs (decenas o cientos de miles), es muy probable que cada uno de los loci donde nos interesa mantener la diversidad esté en LD con al menos uno de los marcadores del panel. DE CARA et al. (2011) demostraron que el uso del parentesco genómico calculado a partir de paneles densos de SNPs para minimizar el parentesco global resulta más eficaz que el uso de genealogías en el mantenimiento de la diversidad genética, medida como EH. Hay que recordar que el LD entre marcadores y loci en el resto del genoma es inversamente proporcional a N_e, por lo que la eficacia de un panel determinado de marcadores será mayor en poblaciones pequeñas (tal y como suelen ser las poblaciones objeto de conservación).

El desarrollo de los primeros paneles densos de SNPs se llevó a cabo en el marco del Proyecto Genoma Humano (THE 1000 GENOMES PROJECT CONSORTIUM, 2010) y posteriormente se comenzaron a desarrollar para las especies ganaderas de mayor importancia económica. La densidad de los paneles de SNPs disponibles puede variar mucho dependiendo de la especie de la que se trate, y está relacionada principalmente con su importancia económica (SMOUSE, 2010). El mayor desarrollo en una especie ganadera a la hora de desarrollar y utilizar paneles de alta densidad de SNPs se ha conseguido para el ganado vacuno, para el que se han llegado a incluir hasta 777,962 SNPs en el Infinium BovineHD BeadChip. Sin embargo, para especies cuya explotación

no genera un rendimiento económico tan alto su desarrollo ha sido más limitado. En todo caso, el desarrollo de la tecnología en las especies de mayor rendimiento produce un abaratamiento de los costes de los chips, por lo que se espera que su uso en un número cada vez mayor de especies será posible en un plazo de tiempo relativamente corto.

Cuando se desarrolla un nuevo panel de SNPs es importante, a fin de optimizar el gasto en su obtención, determinar el orden de magnitud del número de SNPs necesario para conseguir el objetivo planteado, cualquiera que éste sea. Por ejemplo, si el objetivo es simplemente determinar paternidades, la densidad requerida será mucho menor que si el objetivo es obtener estimas genómicas de valores mejorantes. En el caso que nos ocupa, la pregunta sería cuál es el número mínimo de marcadores necesario para mantener eficientemente la diversidad existente en la población de interés (SMOUSE, 2010). En concreto, a nivel práctico, sería muy valioso conocer la densidad de SNPs necesaria para obtener al menos la misma diversidad que la obtenida utilizando información genealógica en la gestión de poblaciones. En el Capítulo 1 se investiga, a través de simulación por ordenador, cómo la densidad de SNPs y N_e afectan a la efectividad de utilizar el parentesco molecular para mantener diversidad a través de la optimización de contribuciones. El objetivo último sería determinar la densidad mínima de SNPs necesaria para mantener al menos la misma diversidad genética (medida como EH) que la obtenida usando parentesco genealógico, dependiendo del N_e histórico de la población.

Otra de las ventajas que aportan los coeficientes de consanguinidad y de parentesco moleculares es que nos permiten investigar patrones de diversidad a lo largo del genoma, lo que no es posible con los coeficientes genealógicos que, como ya se ha mencionado anteriormente, representan valores esperados promedio para todo el

genoma. Así pues, la gestión de poblaciones utilizando el parentesco molecular permite enfocar el mantenimiento de la diversidad genética a regiones específicas del genoma. Esto puede ser deseable en regiones con interés específico (por ejemplo, la región MHC, Complejo Mayor de Histocompatibilidad, que está implicada en la resistencia a enfermedades) o en zonas cercanas a loci seleccionados, donde la diversidad ha disminuido como consecuencia de la selección. El uso del parentesco molecular obtenido utilizando sólo los marcadores mapeados en dichas regiones, nos permitiría gestionar la diversidad de manera independiente. Sin embargo, la optimización de contribuciones basada exclusivamente en minimizar la pérdida de diversidad en unas regiones concretas, puede llevar a un aumento en la pérdida de la diversidad considerable en otras regiones del genoma. En el Capítulo 2 se estudia, a través de simulación por ordenador, la efectividad del uso de paneles densos de SNPs cuando se optimizan las contribuciones para maximizar la diversidad genética en regiones específicas del genoma, mientras que se impone una restricción a la consiguiente pérdida de diversidad en el resto del genoma. También se estudia la posibilidad de maximizar la diversidad global en todo el genoma imponiendo una restricción particular en regiones específicas. Para ambas tareas se ha utilizado un algoritmo de optimización basado en programación semidefinida.

En los trabajos mencionados anteriormente, los coeficientes de consanguinidad y de parentesco moleculares utilizados reflejan tanto IBD como IBS (identidad en estado), a diferencia de los coeficientes genealógicos que reflejan únicamente IBD. Una medida alternativa de consanguinidad molecular, que refleja mejor IBD, es aquella basada en tramos de homocigosidad o ROH (del inglés “Runs Of Homozigosity”), que son largos fragmentos de loci homocigotos consecutivos. Si estos fragmentos son suficientemente largos, la probabilidad de que las dos copias del tramo presentes en un individuo hayan

sido heredadas por sus padres a partir de un ancestro común es muy alta (GIBSON et al., 2006). En este caso, el coeficiente de consanguinidad molecular se define como la proporción del genoma de un individuo que se encuentra formando parte de estos tramos de homocigosidad. Esta medida de consanguinidad ha sido ampliamente estudiada y utilizada en humanos (BROMAN & WEBER, 1999; GIBSON et al., 2006; MCQUILLAN et al., 2008; KIRIN et al., 2010; KU et al., 2011; HERRERO-MEDRANO et al., 2013) y en diferentes especies ganaderas (FERENČAKOVIĆ et al., 2011; PURFIELD et al., 2012; BJELLAND et al., 2013; FERENČAKOVIĆ et al., 2013a; SILIÓ et al., 2013; SAURA et al., 2013; SCRAGGS et al., 2014; SAURA et al., 2015). De la misma manera, se puede calcular un parentesco basado en los segmentos genómicos que comparten una pareja de individuos, a los que denominaremos segmentos IBD en esta tesis (GUSEV et al., 2009). Existen varios estudios que utilizan esta nueva estima de parentesco genómico con objetivos tales como detectar señales de selección natural (ALBRECHTSEN et al., 2009; CAI et al., 2011; HAN & ABNEY, 2013), inferir la historia demográfica de poblaciones (CAMPBELL et al., 2012; GUSEV et al., 2012; PALAMARA et al., 2012; RALPH & COOP, 2013) y estimar la heredabilidad (PRICE et al., 2011; ZUK et al., 2012; BROWNING & BROWNING; 2013). Sin embargo, su uso en el mantenimiento de la diversidad genética (DE CARA et al., 2013a) o en el control de la consanguinidad (PRYCE et al., 2012) no ha sido muy explorado.

Un inconveniente de utilizar el coeficiente de parentesco obtenido a partir de los segmentos compartidos es que es necesario estimar las fases gaméticas de los genotipos de los SNPs que, dada la técnica de genotipado masivo utilizada actualmente, son desconocidas. Existe un gran número de métodos que permiten inferir las fases gaméticas de los genotipos de los SNPs (BROWNING & BROWNING, 2011), pero no existen estudios previos que hayan investigado la eficiencia de la gestión de poblaciones

que utiliza este tipo de parentesco molecular cuando las fases son estimadas. En el Capítulo 3 se lleva a cabo este estudio, a través de simulaciones por ordenador, en las que se utilizan datos reales de genotipado masivo de individuos pertenecientes a tres razas de ganado vacuno austriaco para crear las poblaciones base.

OBJETIVOS

OBJETIVO GENERAL

Evaluar la utilidad de los paneles densos de SNPs (polimorfismos de un solo nucleótido) para mantener la diversidad genética en programas de conservación.

OBJETIVOS ESPECÍFICOS DE CAPÍTULO

Capítulo 1

1.1. Evaluar, mediante simulación por ordenador, el efecto del tamaño efectivo de la población y de la densidad de marcadores SNP sobre la eficiencia del parentesco genómico para mantener la diversidad genética cuando se utiliza la metodología de Contribuciones Óptimas.

1.2. Determinar la densidad mínima de SNPs necesaria para mantener, mediante la minimización del parentesco genómico, al menos la misma diversidad genética que aquella mantenida al minimizar el parentesco genealógico.

Capítulo 2

2.1. Evaluar, mediante simulación por ordenador, la efectividad de utilizar paneles densos de SNPs cuando se optimizan las contribuciones con el objetivo de i) minimizar la pérdida de diversidad genética en regiones específicas del genoma, restringiendo simultáneamente la pérdida de diversidad en el resto del genoma; o ii) maximizar la diversidad genética en todo el genoma, restringiendo simultáneamente la pérdida de diversidad en regiones específicas.

Capítulo 3

3.1. Estimar las tasas de parentesco genómicas (y los correspondientes tamaños efectivos) en tres poblaciones de ganado vacuno y comparar dichas estimas con aquellas obtenidas a partir de las genealogías.

3.2. Evaluar, mediante simulación con ordenador, la eficiencia de las estimas genómicas de parentesco basadas en segmentos IBD en el mantenimiento de la diversidad genética cuando las fases gaméticas tienen que ser estimadas.

CAPÍTULO 1:

Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size

Introduction

With the growing availability of genomic tools, animal genetic studies are evolving with a wide and increasing diversity of applications. In recent years, genome-wide markers have been increasingly used in selection programmes of farm animals (GODDARD, 2012) but much less attention has been given to its application in conservation programmes. One straightforward application of genomic tools in conservation programmes is to use information from single nucleotide polymorphism (SNP) panels to increase the accuracy of estimated genetic relationships between individuals (SANTURE et al., 2010; ENGELSMA et al., 2011) which would improve the efficiency of strategies aimed at managing genetic diversity.

Management of populations under conservation programmes are usually aimed at maintaining the maximum possible genetic diversity (usually measured as expected and observed heterozygosity and sometimes also as allelic diversity) and avoiding high levels of inbreeding. This can be achieved by optimising contributions of potential parents through the minimisation of their global coancestry (MEUWISSEN, 1997; GRUNDY et al., 1998; CABALLERO & TORO, 2000). With a limited number of microsatellite-type markers, FERNÁNDEZ et al. (2005) concluded that the exclusive use of molecular information to compute coancestry coefficients for the optimization process was of limited value to maintain genetic diversity compared to genealogical information. However, recently DE CARA et al. (2011) showed that with high-density panels of markers, the expected and observed heterozygosities maintained were higher when using molecular coancestry than when using genealogical coancestry.

The benefits of using marker information to maintain diversity at ungenotyped loci across the whole genome depend on the amount of linkage disequilibrium (LD)

between these loci and the markers used to manage the population, which in turns depends on effective population size (N_e) and marker density (d). In endangered populations, N_e is usually low and, therefore, LD is expected to be high. This enhances the potential benefits of molecular approaches to maintain genetic diversity. However, the density of available SNPs panels differs largely among species (SMOUSE et al., 2010). While high-density panels containing tens or hundreds of thousands of SNPs have been developed in the last years for farm animal species (e. g. cattle, sheep, swine, chicken, horse and salmon), this is not the case for other species for which the economic benefit of using SNPs panels could be more limited. However, as the technology becomes cheaper, arrays will be developed for non-commercial species (GENOME 10K COMMUNITY OF SCIENTISTS, 2009). Therefore, it is essential to determine the order of magnitude of the minimum SNP density required to maintain a significant percentage of the existing diversity through population management (SMOUSE et al., 2010).

The aims of this study were to (i) investigate, through computer simulations, how N_e and SNPs density affect the performance of molecular coancestry to maintain genetic diversity when used in the optimisation of contributions; and (ii) determine the minimum SNP density required to maintain at least the same genetic diversity with molecular coancestry than with genealogical coancestry.

Methods

Populations at mutation-drift equilibrium with LD between loci were generated through computer simulations. These populations were subsequently managed over ten generations based on genealogical or molecular information (see below). A large number of scenarios with different population sizes and numbers of markers per chromosome were considered.

Generation of the base population

In order to generate a base population at mutation-drift equilibrium, 5,000 discrete generations of random mating were simulated. Four different population sizes (N_e = 20, 40, 80 or 160 individuals, half of each sex) were considered. Sires and dams were sampled with replacement and N_e was kept constant across generations. Note that under this regime N_e equals census size (N). The genome was composed of 20 chromosomes of 1 Morgan (M) each. Two types of biallelic loci (marker and non-marker loci) were simulated. Marker loci were used for management (see below) and non-marker loci were used for measuring diversity. The number of non-marker loci per chromosome was always 1,000 but different densities were considered for marker loci (d = 10, 30, 50, 100, 500, 1,000 and 2,000 SNPs per chromosome). All loci were equidistant and marker loci were interspersed between the non-marker loci in such a way that they covered the whole chromosome evenly. All loci were fixed for allele 1 at the initial generation (t = –5,000). The mutation rate per locus and generation was μ = 2.5 x 10^{-3} for both types of loci. The number of new mutations simulated each generation was sampled from a Poisson distribution with mean $2N_e n_c \mu n_l$ where n_c is the number of chromosomes and n_l is the total number of loci (markers and non-markers) per chromosome. Mutations were then randomly distributed across individuals, chromosomes and loci and they switched allele 1 to allele 2. If a mutation occurred at a position where a previous mutation had already occurred, this allele was allowed to return to its previous state (i.e., 1) instead of choosing another position for the mutation. This rate of reverse mutation was however very low. Individuals were mated at random. When generating the gametes, the number of crossovers per chromosome was drawn from a Poisson distribution with a mean equal to 1. Crossovers were randomly distributed without interference. For all scenarios considered, the population reached

mutation-drift equilibrium after 5,000 generations. We assessed this equilibrium by checking that the mean heterozygosity measured at non-marker loci was stabilized. The population at this point is referred to as the base population (t = 0).

In order to investigate the generality of the results for different mutation rates when creating the base population, some additional scenarios were run assuming a lower mutation rate (2.5 x 10^{-5}). In these scenarios, N_e was 1,000 and marker densities were d = 2,500, 3,000 and 3,500 SNPs/M. The higher values for N_e and d with μ = 2.5 x 10^{-5} were chosen to achieve a reasonable number of segregating loci at t = 0.

Management

Management of the population was carried out for ten discrete generations. Population size was kept constant across generations and equal to its size at t = 0 (i.e., N = 20, 40, 80 or 160 individuals for the high mutation rate scenarios and 1000 for the low mutation rate scenarios). The management method followed the strategy of minimising global coancestry. Thus, the contribution of each individual (i.e., the number of offspring that each individual leaves for the next generation) was optimized by minimising the following expression:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{c_i c_j f_{ij}}{(2N)^2},$$

where c_i is the contribution of individual i and f_{ij} is the coancestry between individuals i and j. The optimization was subjected to the following restrictions: both the sum of contributions of females and sum of contributions of males were equal to N (i.e., $\sum_{i=1}^{N_f} c_i = \sum_{i=1}^{N_m} c_j = N$, where N_f and N_m are the numbers of female and male candidates, respectively), and c_i is an integer number ≥ 0 (FERNÁNDEZ & TORO, 1999).

Coancestry coefficients (f_{ij}) were calculated either from molecular or genealogical data. The molecular coancestry coefficient between individuals *i* and *j* was computed as $f_{ij} = \frac{1}{L}\sum_{l=1}^{L}\left[\left(\sum_{k=1}^{2}\sum_{m=1}^{2} I_{lk_{(i)}m_{(j)}}\right)/4\right]$, where *L* is the number of SNPs and $I_{lk_{(i)}m_{(j)}}$ is the identity of the k^{th} allele of individual *i* with the m^{th} allele of individual *j* for SNP *l* and takes the value of 1 if both alleles are identical and zero if they are not (NEJATI-JAVAREMI et al., 1997). Molecular coancestry coefficients used in the optimisation across generations were calculated using only the marker loci segregating at $t = 0$. Genealogical coancestries were calculated assuming that individuals at $t = 0$ were unrelated and not inbred. All optimizations were performed using a simulated annealing algorithm as described in FERNÁNDEZ & TORO (1999).

In addition, an extra set of simulations was carried out in which genotypes for the non-marker loci (i.e., loci targeted to minimise the loss of diversity) were assumed to be known and used in the optimisation. These extra simulations provided the upper limit of the diversity level that could be maintained using molecular information. In these scenarios (NM), molecular coancestry coefficients were calculated from the non-marker loci and thus, management was based on the same loci that those used to measure diversity. In all simulated scenarios, once the contributions were decided, matings between individuals were performed at random.

Measured variables

Expected (EH) and observed (OH) heterozygosities and allelic diversity (AD) were measured over all non-marker loci and evaluated across the ten generations of management for each simulated scenario. For a single locus, EH (also called gene diversity) was calculated as $\mathrm{EH} = 1 - \sum_{i=1}^{2} p_i^2$, where p_i is the frequency of allele *i*, OH was calculated as the proportion of heterozygous individuals and AD was calculated as

the number of different alleles at the locus. These three variables were then averaged over all non-marker loci. The correlation between molecular and genealogical coancestry coefficients was also calculated across generations.

Linkage disequilibrium was measured at $t = 0$ as the average squared correlation coefficient between adjacent pairs of SNPs (HILL & ROBERTSON, 1968) which can be expressed as $r^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{D_{ij}^2}{(1-p_i)(1-p_j)}$, where p_i is the frequency of allele i at the first locus, p_j is the frequency of allele j at the second locus and D_{ij} is the difference between the observed haplotype frequency and the expected frequency under linkage equilibrium (p_ip_j).

The results presented are averages over 50 replicates. A new base population at mutation-drift equilibrium was simulated for each replicate and the same base population was used for both management methods (genealogical and molecular).

Results

As expected, the distribution of allelic frequencies at $t = 0$ was U-shaped (results not shown). The percentage of segregating markers at $t = 0$ ranged from 48% ($N_e = 20$) to 99% ($N_e = 160$). As expected, the amount of LD at $t = 0$ before management began, increased with increasing d and with decreasing N_e (Figure 1.1). Values for r^2 ranged from 0.13 to 0.30 for $N_e = 20$ and from 0.02 to 0.10 for $N_e = 160$ in scenarios with the highest mutation rate. Notwithstanding, the increase observed in r^2 when increasing d was small for high densities of SNP. For the scenarios with the lowest mutation rate and N_e =1,000, r^2 ranged from 0.065 to 0.072. These levels of LD are in the same range as those obtained in previous studies considering similar population parameters (SOLBERG et al., 2008; HARRIS & JOHNSON, 2010).

Figure 1.1. Average linkage disequilibrium (r^2) between adjacent markers at the initial generation (t = 0) for different marker densities (d) and effective population sizes (N_e).

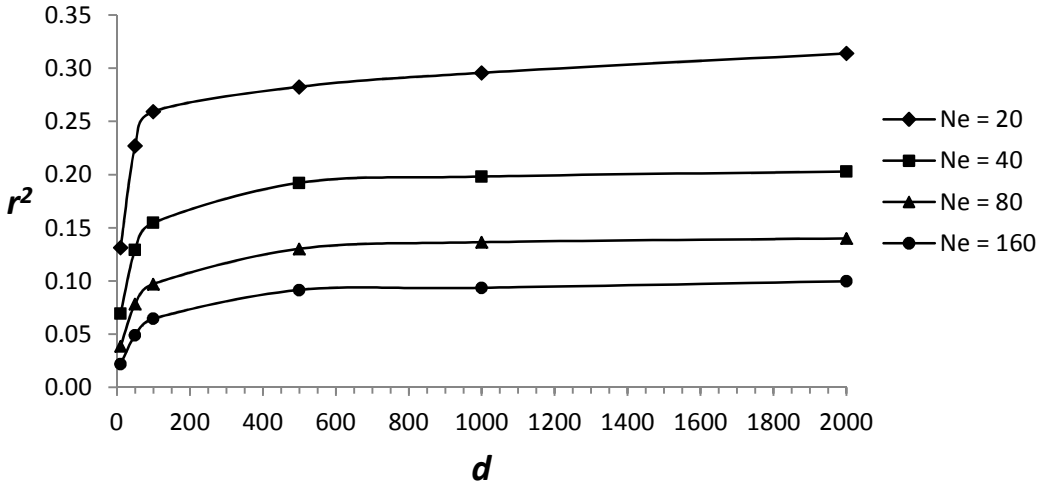


Table 1.1 shows EH values calculated when the optimization was performed with molecular or genealogical information. Results obtained using genealogical data are expressed as deviations from those obtained using molecular data. OH values (not shown) were always higher than EH values across all the scenarios simulated. The mean difference between both measures of diversity was 3% and the maximum difference, reached with the lowest initial N_e (N_e = 20) and the lowest marker density (d = 10 SNPs/M), was 8.4%. Thus, deviations from Hardy-Weinberg equilibrium (α = (EH – OH)/EH) were always negative and ranged from –0.006 to –0.092. This was mainly due to sampling (WANG, 1996) and to a lesser extent to the management strategy implemented, which resulted in lower levels of genetic relationships that those expected using random contributions.

As expected, the initial heterozygosities (EH and OH) were higher in scenarios with higher N_e (Table 1.1). These scenarios also maintained a higher amount of diversity across generations than those with lower N_e. For instance, using genealogical

Table 1.1. Expected heterozygosity over generations (*t*) obtained for management based on molecular or genealogical data for different marker densities (*d*; SNPs/M) and effective population sizes (N_e).

		d = 10		*d* = 100		*d* = 500		*d* = 1,000		*d* = 2,000		NM	
N_e	*t*	EH_M[†]	EH_{M-G}[‡]	EH_M	EH_{M-G}	EH_M	EH_{M-G}	EH_M	EH_{M-G}	EH_M	EH_{M-G}	EH_M	EH_{M-G}
20	0	0.136	+0.000	0.136	+0.000	0.136	+0.000	0.135	+0.000	0.137	+0.000	0.136	+0.000
	1	0.131	–0.003	0.135	+0.001	0.135	+0.001	0.135	+0.001	0.137	+0.002	0.143	+0.009
	2	0.126	–0.006	0.133	+0.000	0.134	+0.001	0.134	+0.002	0.135	+0.002	0.146	+0.014
	3	0.122	–0.009	0.131	+0.000	0.132	+0.002	0.132	+0.002	0.134	+0.002	0.148	+0.017
	4	0.118	–0.011	0.129	+0.000	0.131	+0.002	0.131	+0.002	0.132	+0.003	0.149	+0.020
	10	0.100	–0.019	0.118	–0.001	0.122	+0.003	0.122	+0.004	0.124	+0.004	0.152	+0.033
160	0	0.378	+0.000	0.378	+0.000	0.378	+0.000	0.378	+0.000	0.378	+0.000	0.378	+0.000
	1	0.371	–0.007	0.377	–0.001	0.378	+0.000	0.378	+0.001	0.379	+0.001	0.390	+0.012
	2	0.366	–0.011	0.375	–0.002	0.378	+0.001	0.378	+0.001	0.379	+0.002	0.395	+0.018
	3	0.362	–0.015	0.374	–0.003	0.377	+0.001	0.378	+0.002	0.379	+0.002	0.399	+0.022
	4	0.357	–0.019	0.372	–0.004	0.377	+0.001	0.378	+0.002	0.378	+0.003	0.401	+0.025
	10	0.336	–0.036	0.364	–0.009	0.373	+0.001	0.375	+0.003	0.377	+0.005	0.413	+0.040

[†]EH_M = expected heterozygosity obtained when management is based on molecular data.

[‡]EH_{M-G} = expected heterozygosity obtained when management is based on genealogical data, expressed as a deviation from EH_M.

data the percentage of EH maintained after ten generations of management for N_e = 20 and N_e = 160 was 88% and 98%, respectively. The levels of EH at t = 0 ranged from 0.136 (N_e = 20) to 0.378 (N_e = 160) (Table 1.1). This latter value is similar to that reported by ENGELSMA et al. (2010) for a comparable population.

The maintained EH decreased across generations in most scenarios, except when the SNP density was very high, in which case EH remained stable for several generations as previously reported and discussed by DE CARA et al. (2011). In the extreme case in which non-marker loci were used in the optimisation, EH even increased in the initial generations.

For low densities (generally, d < 500), management based on genealogical coancestry resulted in higher diversity than management based on molecular coancestry and the difference in EH between these two strategies increased across generations. For instance, with d = 10 and management based on molecular coancestry, the EH maintained at t = 10 was 74% (N_e = 20) and 89% (N_e = 160) of the initial EH. With management based on genealogical coancestry these figures increased to 88% (N_e = 20) and 99% (N_e = 160). However, with d = 500, the EH maintained with management based on molecular coancestry (90% and 99% of the initial EH for N_e = 20 and N_e = 160, respectively) exceeded that maintained with management based on genealogical coancestry (87% and 98% for N_e = 20 and N_e = 160, respectively). The advantage of using molecular coancestry increased with increasing d. In any case, differences in heterozygosity between management strategies using both types of information (molecular and genealogical) were generally small.

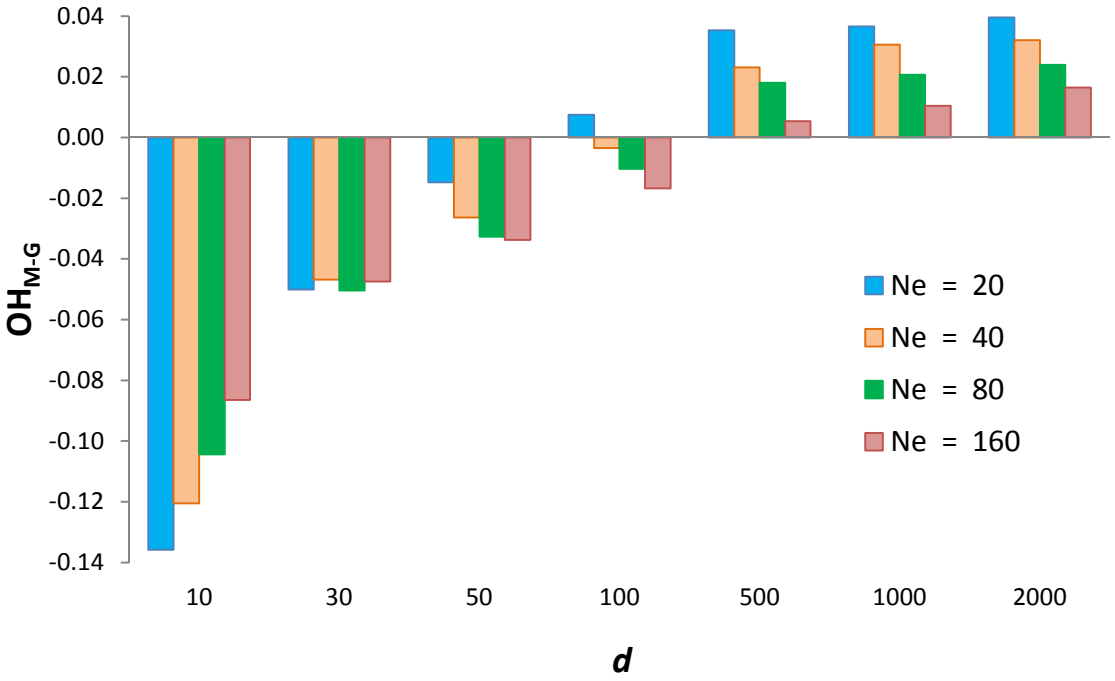
At t = 10, the proportion of EH maintained using marker information relative to that maintained using non-marker information (i.e., an ideal situation in which

management could be performed using the same loci in which diversity is measured) increased with increasing densities (Table 1.1). For example, for N_e = 20, the EH maintained using markers was 65% and 77% of that maintained using non-markers for d = 10 and d = 100, respectively. Corresponding figures for N_e = 160 were 80% and 88%. However, for d > 100 the benefit in maintained diversity from using more markers decreased. In fact, for $d \geq 500$ the increase in the proportion of EH maintained using non-markers instead of markers was practically negligible.

Figure 1.2 shows the difference in OH maintained at t = 10 between scenarios using molecular or genealogical coancestry for different N_e. With a low density (e.g., d = 10), as mentioned above, OH was lower than that obtained from genealogical-based management for all values of N_e and the difference between both management approaches decreased with increasing N_e. On the other hand, with $d \geq 50$, this difference increased with decreasing N_e.

For the smallest N_e considered, d = 100 (i.e., d = 5 N_e SNPs/M) was a sufficient density to reach higher levels of diversity with molecular than with genealogical coancestry. For N_e > 20, the density required to achieve these levels increased to 500 SNPs per chromosome. Given that scenarios with intermediate densities between d = 100 and d = 500 were not simulated, the number of markers required for achieving the same levels of heterozygosity from both types of management (i.e., based on molecular or on genealogical coancestry) was estimated for each N_e through linear interpolation, assuming that the change in performance from d = 100 to d = 500 was constant. Values obtained were equal to about 3 times N_e (ranging from 2.6 to 3.3N_e). This result, showing that a SNPs density of 3N_e SNPs/M is enough for molecular coancestry to equalise the performance of genealogical coancestry, was also valid for scenarios in

Figure 1.2. Difference between observed heterozygosity using molecular or genealogical coancestry (OH_{M-G}) at generation 10, for different marker densities (*d*; SNPs/M) and effective population sizes (N_e).



which the mutation rate used to generate the base population was 2.5 x 10^{-5} and N_e was increased correspondingly to N_e = 1,000. The difference in the OH maintained at t = 10 between scenarios using molecular or genealogical coancestry was –0.003, 0.000 and 0.002 for d = 2,500 (d/N_e = 2.5), 3,000 (d/N_e = 3.0) and 3,500 (d/N_e = 3.5), respectively. Thus, the result is general for different combinations of μ and N_e.

Genealogical coancestry was always more efficient in maintaining AD than molecular coancestry except for the scenario with the smallest N_e and the highest SNPs density (Figure 1.3). It is interesting to note that for a given d, the largest difference in AD between both management strategies (i.e., using genealogical or molecular coancestry) occurred at intermediate values of N_e.

Figure 1.3. Difference between allelic diversity using molecular or genealogical coancestry (AD_{M-G}) at generation 10, for different marker densities (*d*; SNPs/M) and effective population sizes (N_e).

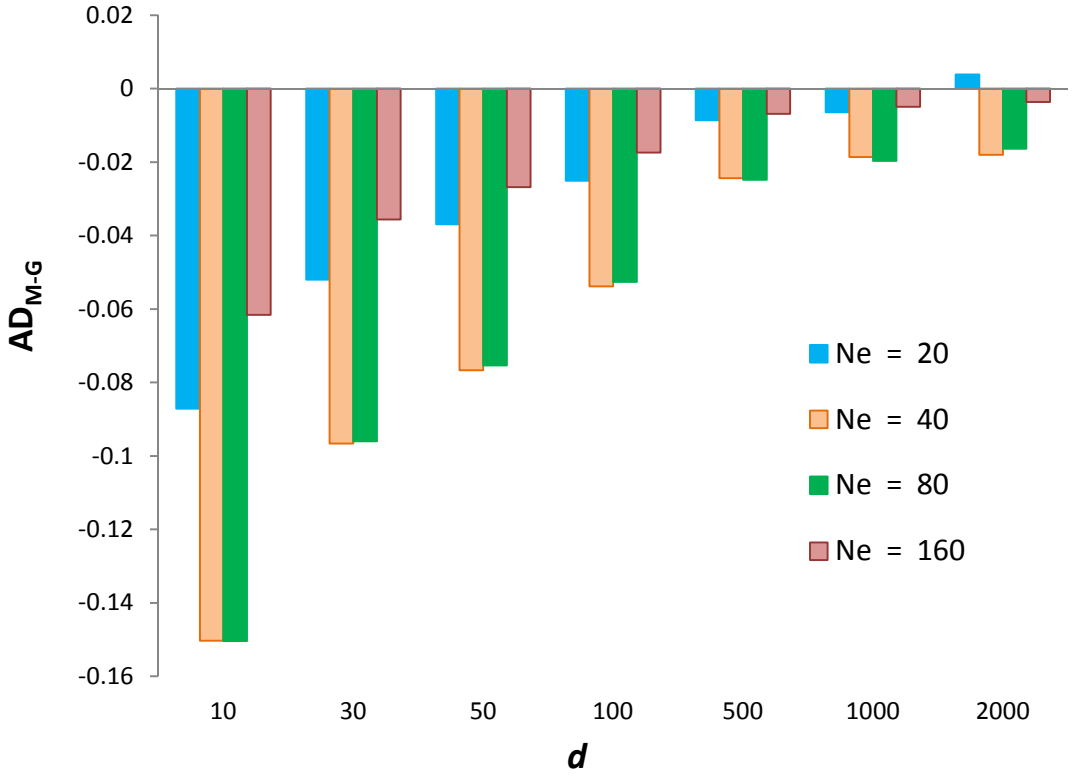


Table 1.2 shows the evolution across generations of the correlation between molecular and genealogical coancestries obtained when the management was based on molecular data. The correlation was highest with the smallest N_e and the highest *d*. In general, the correlation was very high (over 0.8) except in early generations for scenarios with a low *d*. Counter-intuitively, higher correlations between molecular and genealogical coancestry at $t = 0$ did not lead to smaller differences between both management methods.

Table 1.2. Correlation between molecular and genealogical coancestries across generations (*t*) for different marker densities (*d*; SNPs/M) and effective population sizes (N_e).

		N_e			
d	*t*	20	40	80	160
10	0	0.715	0.678	0.638	0.561
	1	0.832	0.833	0.826	0.790
	5	0.859	0.864	0.846	0.804
	10	0.863	0.869	0.854	0.797
100	0	0.829	0.822	0.814	0.799
	1	0.935	0.936	0.934	0.930
	5	0.954	0.953	0.947	0.937
	10	0.954	0.953	0.947	0.935
500	0	0.848	0.837	0.835	0.832
	1	0.950	0.949	0.949	0.948
	5	0.971	0.968	0.967	0.964
	10	0.970	0.969	0.966	0.963
1,000	0	0.849	0.845	0.837	0.836
	1	0.951	0.951	0.951	0.951
	5	0.973	0.971	0.970	0.968
	10	0.971	0.971	0.970	0.968
2,000	0	0.848	0.848	0.841	0.838
	1	0.952	0.951	0.953	0.936
	5	0.974	0.972	0.971	0.962
	10	0.973	0.972	0.971	0.964

Discussion

This study has investigated the effect of N_e and marker density on the efficiency of molecular coancestry when used in the optimization of contributions aimed at minimising the loss of genetic diversity. As expected, higher densities and lower N_e improved the performance of the management based on molecular coancestry. This was due to the higher LD created between marker loci and non-genotyped loci at which

diversity was measured. The density of SNPs required to maintain at least the same heterozygosity than that maintained using genealogical data was approximately $3N_e$ SNPs/M. The benefits of using molecular coancestry calculated with dense SNP data were small when compared to genealogical coancestry (a benefit of 3% in the most favourable molecular scenario). However, this represents an improvement over previous molecular approaches. Previous studies using microsatellites (FERNÁNDEZ et al., 2005; SANTURE et al., 2010) observed that management based on genealogical coancestry always outperformed that based on molecular coancestry with low density markers. It must be pointed out that, as shown by FERNÁNDEZ et al. (2005), the combined use of genealogical and molecular information could increase furthermore the precision of the coancestry coefficients and therefore their efficiency beyond that obtained with a single source of information.

Molecular coancestry coefficients have been calculated as the proportion of shared alleles between individuals. Many corrections aimed at making molecular coancestry closer to genealogical coancestry have been proposed and all assume that the initial allelic frequencies are known (TORO et al., 2002; VANRADEN, 2008). However, at least in our context of management aimed at maintaining the highest levels of diversity, there is no advantage in applying these corrections. DE CARA et al. (2011) showed that when the number of markers is sufficiently large, the use of molecular coancestry always maintains higher levels of diversity than genealogical coancestry. They also tested two of the estimators proposed but they did not improve the performance of uncorrected molecular coancestry.

Clearly, marker density requirements depend on the purpose for which these markers are used. In the context of genomic selection, SOLBERG et al. (2008) found that the accuracy of selection continued to increase with increasing marker density at least

up to $8N_e$, for scenarios with N_e = 100. However, the increase in SNPs density had diminishing returns in terms of accuracy. They showed that by doubling marker density from $1N_e$ to $2N_e$, the accuracy of estimated breeding values increased by 14%. This figure was reduced to only 2% when marker density was doubled from $4N_e$ to $8N_e$. This relatively small increase in accuracy appears to be insufficient to justify the increase in marker density, especially taking into account that with a density of $4N_e$ SNPs/M the accuracy had already reached 92% of the upper bound that could be obtained theoretically (SOLBERG et al., 2008). Similarly, in the context of conservation programmes, the increase in SNP density had diminishing returns in terms of the diversity maintained (Figure 1.2). In scenarios with d = 100, the EH maintained after 10 generations ranged from 77% (N_e = 20) to 88% (N_e = 160) of the upper bound (obtained when non-marker loci were used in the optimisation). These figures increased, respectively, to 80% and 90% when marker density was increased to 500 and then stayed nearly constant with higher densities. Thus, under the conditions studied here, a density of 500 SNPs/M could be considered as the most cost effective density, given that it makes it possible to maintain a substantial amount of heterozygosity with a relatively small number of markers. Most of the SNPs chips already available for farm animals (e.g. cattle, sheep, swine, chicken, horse and salmon) contain more than 500 SNPs/M and thus, they would be suitable for programmes aimed at the conservation of genetic resources that are based on the minimization of coancestry. Thus, when developing SNPs chips for a new species with this objective, the marker density should reach 500 SNPs/M. Since the costs of developing SNP chips are decreasing, it can be expected that SNPs chips with such densities will be available for all species of interest in a short-term horizon.

SOLBERG et al. (2008) concluded that a density of 800 SNPs/M was not sufficient to achieve the maximum possible accuracy for genomic breeding values. This density is considerably higher than that recommended here for the maintenance of diversity (d = 500). For other tasks associated with conservation genetic programmes such as the determination of relatedness between individuals, a density lower than 500 SNPs/M would be sufficient (SANTURE et al., 2010).

As mentioned above, the benefit of using management based on molecular coancestry relative to that based on genealogical coancestry increased with decreasing N_e, except when the density of SNPs was very low. This could be due to the fact that with a very low density, the level of LD between markers and non-genotyped loci is low even for the smallest N_e. However, larger sample sizes (i.e., larger N_e) can make the detection of groups of individuals with higher levels of genetic diversity possible. We observed that as d increased, the effect of N_e over the existing LD became more pronounced. The overall effect is that higher N_e lead to a substantial reduction in LD counteracting the beneficial effect of larger sampling sizes on the performance of the management based on molecular coancestry.

Allelic diversity has been considered as an alternative measure of genetic diversity, particularly from a long-term perspective because the limits to selection are determined by the initial number of alleles and because allelic diversity is more sensitive to bottlenecks than EH and, therefore, reflects better past fluctuations in population size (TORO et al., 2009). It should be noted that the optimisation method used here was originally developed to maximize EH and thus AD is maintained only indirectly (FERNÁNDEZ et al., 2004; LUIKART et al., 1998). Consequently, larger densities are required for molecular information to outperform genealogical management in terms of AD. In fact, this only occurred for d = 2,000 and N_e = 20. The fact that for any given

d, the performance of management based on molecular coancestry was less efficient for intermediate N_e could also be the consequence of the opposite effects of increasing N_e (i.e., reduced LD but increased sample size).

As expected, the correlation between molecular and genealogical coancestries increased with increasing density but this was not translated into an increased similarity in the diversity maintained with both approaches. At $t = 0$, all individuals were assumed to be unrelated and, thus, genealogical coancestry was uniform across the individuals. This led to an optimal solution that implied equalizing contributions of all individuals. However, molecular coancestry varied across pairs of individuals and the optimization method could find a combination of contributions that resulted in higher levels of EH even when some of the candidates did not contribute at all (about 60% of the individuals did not yield any offspring). The higher the number of markers, the higher was the variation in coancestry between pairs of individuals and the higher was the power of the method to discriminate between them. This led to a higher efficiency of molecular coancestry to maintain genetic diversity. Therefore, even with very high correlations between coancestries both management approaches produce different results.

Here, we investigated the benefits of using molecular SNPs data to maintain the levels of global diversity of a population. Another advantage of using molecular information is the possibility of maintaining diversity at specific genome regions especially at those responsible for adaptive variation. However, this could increase inbreeding and loss of diversity in the rest of the genome (ROUGHSEDGE et al., 2008). Thus, in this situation, it would be preferable to manage local and global diversity simultaneously by imposing restrictions on global coancestry while optimizing the local diversity maintained.

CAPÍTULO 2:

The use of genomic coancestry matrices in the optimisation of contributions for maintaining genetic diversity at specific regions of the genome

Introduction

It is generally accepted that controlling the rate of coancestry provides a general framework for managing genetic variability. Optimal Contribution (OC) methods (MEUWISSEN, 1997; GRUNDY et al. 1998) permit to determine the number of offspring that each breeding candidate should have to minimise coancestry. These methods were initially developed assuming the pedigree-based relationship matrix (**A**), that represents expected relationships assuming neutrality and does not take into account variation due to Mendelian sampling. Thus, although its use has proved to be efficient to manage diversity it has some limitations. For instance, individuals from the same (full-sib) family would inherit different set of alleles but they are assumed to be equally related. Additionally, since **A** does not consider variation between genomic regions, the optimisation of contributions would, in average, control the rate of coancestry to the chosen value, but some genomic regions may have substantially higher rates than those desired.

The management of genetic diversity can be improved if the **A** matrix is replaced by a realised relationship matrix calculated taking into account variation across animals of the same family and variation across genome regions. Such realised relationship matrices are now possible to be calculated because of the availability of high density SNP chips. Genotypes for hundreds or thousands SNPs across the genome are now commonly used to calculate relationship matrices in many species (VANRADEN, 2008; HAYES et al., 2009). These matrices have proved to satisfactorily manage global genetic diversity, outperforming matrices based on genealogical data (DE CARA et al., 2011; SAURA et al., 2013; GÓMEZ-ROMANO et al., 2013). Relationship matrices calculated from markers at particular genomic regions can also be used to minimise

variability loss at specific regions of the genome. However, this needs to be accompanied by constraints on coancestry in the rest of the genome. Otherwise, rates of coancestry, inbreeding and loss of variability could reach high values at positions away from the region where minimisation was targeted (ROUGHSEDGE et al., 2008).

The objective of this study was to assess, through computer simulations, the effectiveness of using dense SNP panels when optimising contributions i) for minimising the loss of genetic variability at specific genomic regions while restricting the overall loss in the rest of the genome; or ii) for maximising overall diversity while restricting the loss at specific genomic regions.

Methods

Optimisation of contributions for minimising the loss of genetic diversity

Let assume a set of N breeding candidates and let **c** be the vector of genetic contributions of the candidates to the next (offspring) generation. These contributions represent the fraction of the genetic material each candidate contributes to the gene pool. In diploid species each sex contributes half of the gene pool, so the genetic contribution of a given candidate ranges between [0:0.5]. Note that $c_i = 0$ indicates that the candidate i has no offspring and $c_i = 0.5$ indicates that all offspring are fathered (mothered) by i. Let **s** and **d** be vectors defining flags that indicate the sex of the candidates, with $s_i = 1$ if candidate i is a male and 0 if it is a female, and $\mathbf{d} = \mathbf{1} - \mathbf{s}$.

Optimisation problem 1

When the main breeding objective is to minimise the loss of genetic diversity, genetic contributions of candidates are optimised by minimising the expected average

level of coancestry in the offspring generation. Hence, the OC problem can be formulated as:

$$\text{Minimise} \quad \mathbf{c'Gc} \tag{1a}$$

$$\text{Subject to} \quad \mathbf{c's} = 0.5 \tag{1b}$$

$$\mathbf{c'd} = 0.5 \tag{1c}$$

$$c_i \geq 0 \tag{1d}$$

where **G** is the coancestry matrix containing coefficients of coancestry between all candidates in the population. Note that this differs from the formulation of MEUWISSEN (1997), GRUNDY et al. (1998) and PONG-WONG & WOOLLIAMS (2007) who used the numerator relationship matrix **A** which is twice **G** (i.e., $\mathbf{G} = \frac{1}{2}\mathbf{A}$). The constraints (1b-1d) are imposed in order to keep the solution for **c** within the valid range.

Matrix **G** can be computed from pedigree or molecular data. With the availability of dense SNP genotyping, it is also possible to obtain a **G** matrix that is related to specific regions of the genome. Hence, the optimisation problem can be implemented to minimise the loss of diversity at the whole genome or at specific regions of the genome by using the adequate **G** matrix (see below).

Optimisation problem 2

The optimisation problem can be further refined when the aim is to minimise the loss of diversity (overall or at specific regions) but also imposing extra constraint(s) so the expected level of coancestry in the offspring generation for one or more genome regions cannot be greater than a given predefined value (*K*). Hence, the OC problem can be reformulated by adding *m* extra constraint(s):

$$\text{Minimise} \quad \mathbf{c'Gc} \tag{2a}$$

$$\text{Subject to} \quad \mathbf{c'G_1c} \leq K_1 \tag{2b}$$

$$\mathbf{c'G_2c} \leq K_2$$

$$\vdots$$

$$\mathbf{c'G_mc} \leq K_m$$

$$\mathbf{c's} = 0.5 \tag{2c}$$

$$\mathbf{c'd} = 0.5 \tag{2d}$$

$$c_i \geq 0 \tag{2e}$$

where **G** is the matrix for the part of the genome where coancestry will be minimised (overall or regional) and $\mathbf{G}_j$ ($j = 1, \ldots, m$) is the coancestry matrix for region j where a restriction is imposed. The term K_j is the maximum expected level of coancestry to be allowed for region j. At a given generation, K_j can be calculated as $K_j = 1 - (1 - C_j)(1 - f_j)$, where f_j is the average coancestry at region j in that generation, and C_j is the targeted rate of coancestry for region j.

The implementation of both optimisation problems was carried out using a semidefinite programming (SDP) approach as described in PONG-WONG & WOOLLIAMS (2007). The detailed formulations are shown in Appendix 2.1. Thereafter, the reformulated problems were solved using the SDPA package (FUJISAWA et al., 2002).

Coancestry matrices

Different coancestry matrices were used in the optimisation of contributions. They included coancestry matrices computed from pedigree or from genomic information. Genomic matrices were calculated using a large amount of biallelic markers that mimicked SNPs and the allelic relationship method proposed by NEJATI-

JAVAREMI et al. (1997). For a given SNP the allelic relationship between two individuals is $(0.25)\sum_{i=1}^{2}\sum_{j=1}^{2}\delta_{ij}$, where δ_{ij} is the allele sharing status which equals to 1 if allele *i* from the first individual is identical to allele *j* from the second individual and 0 otherwise. The genomic coancestry between two individuals is the average value across all genotyped SNPs in the genome (for the whole genome matrix) or in the region(s) of interest (for a regional genomic matrix). Note that although the realised coancestry matrix calculated with SNPs information may be more precise than the traditional pedigree-based coancestry matrix, it would still represent estimates of the true relationships unless full sequences are available and used for calculating those relationships.

Simulations

Different management strategies aimed at minimising the loss of genetic diversity were compared using Monte Carlo simulations. The strategies differed in the type of information employed to compute coancestries to be used when optimising contributions. They also differed in the objective function to be minimised and the restrictions imposed during the optimisation.

The study considered populations of *N* animals (20 or 100) born per generation. The sex of the individuals was randomly assigned but ensuring that half were males and half were females. Each management scenario was replicated 100 times.

Genetic and population models

The genetic model assumed the genome divided into 20 chromosomes of one Morgan each. Each chromosome had n_{loci} biallelic loci equally spaced. The genotypes of $n_{loci}/2$ of them (those located at alternate positions) were assumed to be known and they

were used to create the genomic matrices implied in the optimisation of contributions. Thus, these $n_{loci}/2$ loci simulated per chromosome mimicked SNP markers. The remaining $n_{loci}/2$ loci per chromosome were used to assess the performance of the different management strategies.

Initially, a base population in mutation-drift equilibrium was generated. This ensured the existence of linkage disequilibrium between the SNPs and the non-marker loci. Details on how the base population was created are given in GÓMEZ-ROMANO et al. (2013). In brief, a historical population of size N was simulated for 10,000 generations of random mating. The historical population was initialised assuming that alleles at the $20n_{loci}$ simulated loci were fixed. Two different mutation rates were considered (μ = $2.5\text{x}10^{-3}$ and μ = $2.5\text{x}10^{-5}$) in order to mimic two different strengths of linkage disequilibrium between marker and non-marker loci. The last generation of this process was considered to be the base population (t = 0). In scenarios where μ = $2.5\text{x}10^{-3}$, n_{loci} was 2,000 and in those where μ = $2.5\text{x}10^{-5}$, n_{loci} was 60,000. These values for n_{loci} ensured that there were enough loci segregating at t = 0.

Thereafter this population was managed under different strategies for 10 generations. At each generation, the contributions of the potential parents were optimised according to the strategy used, and a generation of offspring of size N was created. In turn, they became the candidates for the next round. It should be noted that there were no mutations when creating the generations where management took place (i.e., after creating the base population).

Scenarios compared

Seven different management strategies (PED, MOL_{OVE}, MOL_{CHR}, MOL_{REG}, $\text{MOL}_{\text{OVE\_CON}}$, $\text{MOL}_{\text{CHR\_CON}}$ and $\text{MOL}_{\text{REG\_CON}}$) were considered (Table 2.1). The

management in strategies PED, MOL_{OVE}, MOL_{CHR} and MOL_{REG} was based on optimisation problem 1 and differed in the coancestry minimised; i.e., in the **G** matrix used in equation 1a. Strategy PED minimised pedigree-based coancestry (f_p), MOL_{OVE} minimised the overall (i. e., average for all markers in the genome) molecular

Table 2.1. Rates of coancestry minimised and restricted under each optimization strategy.

Strategy	Minimised	Restricted
PED	Rate of pedigree coancestry	–
MOL_{OVE}	Overall rate of molecular coancestry	–
MOL_{CHR}	Rate of molecular coancestry at chromosome 1	–
MOL_{REG}	Averaged rate of molecular coancestry across ten 10 cM regions located on different chromosomes	–
MOL_{OVE_CON}	Overall rate of molecular coancestry	Rate of molecular coancestry at each of ten 10 cM regions located on different chromosomes
MOL_{CHR_CON}	Rate of molecular coancestry at chromosome 1	Overall rate of molecular coancestry
MOL_{REG_CON}	Average rate of molecular coancestry across ten 10 cM regions located on different chromosomes	Overall rate of molecular coancestry

coancestry (f_{m_ove}), MOL_{CHR} minimised coancestry in an entire chromosome (chromosome 1) (f_{m_chr}) and MOL_{REG} minimised the average molecular coancestry across ten regions of 10 cM each located on a different chromosome (f_{m_reg}). The specific location of these ten regions was randomly choosen. For a given chromosome, the specific region was the same across replicates. Strategies MOL_{CHR_CON} and

MOL_{REG_CON} were based on optimisation problem 2 where the average coancestry in specific region(s) of the genome was minimised restricting simultaneously the coancestry in the rest of the genome ($f_{m\_ove\text{-}chr}$ and $f_{m\_ove\text{-}reg}$ for MOL_{CHR_CON} and MOL_{REG_CON}, respectively). The restriction applied on the rest of the genome was such that the intended rate of increase in $f_{m\_ove\text{-}chr}$ and $f_{m\_ove\text{-}reg}$ was either 1.0% or 0.1% per generation. Strategy MOL_{OVE_CON} was also based on optimisation problem 2 and implied minimising the overall molecular coancestry while imposing independent restrictions on the increase in coancestry at the ten regions located on different chromosomes. Note that the different genomic matrices required in the optimisation for the different strategies were calculated using the observed SNP genotypes. An extra scenario where contributions were randomly assigned (strategy RAN) was also considered for comparison.

Criteria of comparison

The rate at which genetic diversity is lost is given by the rate of coancestry. Thus, the main criteria for comparing different management strategies were the pedigree and genomic rates calculated at each generation. For the purpose of comparing strategies, genomic coancestry matrices were computed using the non-marker loci. The number of individuals that contributed to the offspring generation and the variance of contributions were also calculated each generation.

Results

Table 2.2 shows the percentage of individuals that produced offspring each generation and the variance of their contributions under the different management strategies. Results using $\mu = 2.5\text{x}10^{-3}$ or $\mu = 2.5\text{x}10^{-5}$ were very similar and only those for the latter are presented. The variance of contributions under random selection was

close to two, which is the theoretical expected value if contributions follow a Poisson distribution. The optimum solution for the strategy that minimised Δf_p (strategy PED) was that all individuals contribute and they do it equally (i.e., every candidate generates two offspring) each generation. This is because we assumed that individuals in the base population were all non-inbred and unrelated. However, when minimising rates of genomic coancestry not all individuals contribute to the offspring generation and those contributing left different numbers of offspring. This is a consequence of the differences that exist at the molecular level even for individuals with the same degree of pedigree relationship. The most extreme situation was found when minimising coancestry at specific regions of the genome (strategy MOL_{REG}) where the number of candidates contributing to the next generation was the lowest and the variance of their contributions was the highest. Under strategy MOL_{CHR} (results not shown), 4% to 15% less individuals contributed to the next generation than under strategy MOL_{REG}. On the other hand, the variance of contributions was 19% to 43% higher for MOL_{CHR} than for MOL_{REG}. In general, differences between both strategies were larger in early generations and for N = 100. As expected, when restrictions on the rate of coancestry were included in the optimisation the number of contributing candidates increased and the variance of contributions decreased. This was more pronounced for the most severe restriction. Except for strategy PED, increasing the population size from 20 to 100 led to a decrease in the percentage of individuals contributing to the next generation and to an increase in the variance of contributions.

Table 2.3 shows the average rates of pedigree and molecular coancestries for scenarios RAN, PED and MOL_{OVE}. When the contributions were assigned at random

Table 2.2. Percentage of individuals that contribute to the next generation (N_{cont}) and variance of contributions ($V(c)$) when applying different management strategies (RAN, PED, MOL_{OVE}, MOL_{REG}, MOL_{OVE_CON} and MOL_{REG_CON}) in populations of two different sizes (N = 20 and 100). Two different constraints (C) were imposed on Δf_{m_reg} or $\Delta f_{m\_ove\text{-}reg}$ when applying strategy MOL_{OVE_CON} and MOL_{REG_CON}, respectively. Mutation rate used to create base population was $\mu = 2.5 \times 10^{-5}$.

									MOL_{OVE_CON}				MOL_{REG_CON}			
	RAN[†]		PED		MOL_{OVE}		MOL_{REG}		C = 0.10%		C = 0.01%		C = 1.00%		C = 0.10%	
t	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$	N_{cont}	$V(c)$
								N = 20								
0	88.4	1.77	100	0.00	81.0	2.28	56.7	6.00	84.8	1.75	88.1	1.44	72.3	3.68	76.7	5.99
1	87.0	1.83	100	0.00	90.3	1.37	61.1	4.88	93.1	1.04	96.2	0.78	80.6	2.14	81.1	4.88
2	88.4	1.80	100	0.00	91.4	1.25	62.2	4.63	94.8	0.90	96.7	0.69	83.1	1.90	82.3	4.62
3	86.0	2.03	100	0.00	89.8	1.40	65.4	4.06	94.4	0.92	96.6	0.70	84.1	1.81	85.2	4.07
4	88.5	1.70	100	0.00	90.1	1.33	65.5	3.98	94.8	0.86	96.5	0.71	84.1	1.82	85.5	4.00
9	89.5	1.64	100	0.00	90.5	1.26	71.6	3.18	95.1	0.85	97.4	0.68	85.9	1.78	89.7	3.17
								N = 100								
0	85.6	2.10	100	0.00	54.3	6.33	34.0	13.67	55.4	6.12	56.0	5.83	38.1	11.48	51.9	7.12
1	86.4	1.97	100	0.00	60.1	5.12	38.4	11.49	61.4	4.78	61.4	5.05	40.2	10.43	56.7	5.96
2	87.1	1.95	100	0.00	61.4	4.88	39.9	10.54	63.8	4.43	64.6	4.15	40.9	10.28	57.8	5.47
3	86.1	2.07	100	0.00	61.3	4.52	42.0	9.88	64.7	4.29	66.1	3.97	43.4	9.51	58.1	5.43
4	87.9	1.96	100	0.00	65.0	4.43	44.6	9.10	65.8	4.08	66.5	4.04	44.0	9.08	60.6	5.00
9	86.9	1.99	100	0.00	81.0	2.28	51.3	6.97	67.5	3.74	69.5	3.54	50.3	7.31	63.7	4.39

[†]RAN: contributions are assigned at random; PED: contributions are optimised for minimising f_p; MOL_{OVE}: contributions are optimised for minimising f_{m_ove}; MOL_{REG}: contributions are optimised for minimising f_{m_reg}; MOL_{OVE_CON}: contributions are optimised for minimising f_{m_ove} imposing simultaneously a restriction on Δf_{m_reg}; MOL_{REG_CON}: contributions are optimised for minimising f_{m_reg} imposing simultaneously a restriction on $\Delta f_{m\_ove\text{-}reg}$.

Table 2.3. Rates of pedigree (Δf_p) and overall molecular (Δf_{m_ove}) coancestry across generations (t) when applying different management strategies (RAN, PED and MOL_{OVE}) in populations of two different sizes (N = 20 and 100) and using two different mutation rates (μ = $2.5x10^{-3}$ and μ = $2.5x10^{-5}$) to create the base population.

	Δf_p (%)						Δf_{m_ove} (%)					
	$\mu = 2.5x10^{-3}$			$\mu = 2.5x10^{-5}$			$\mu = 2.5x10^{-3}$			$\mu = 2.5x10^{-5}$		
t	RAN[†]	PED	MOL_{OVE}	RAN	PED	MOL_{OVE}	RAN	PED	MOL_{OVE}	RAN	PED	MOL_{OVE}
						N = 20						
1	2.46	1.28	2.47	2.45	1.28	2.69	2.47	1.32	0.17	2.57	1.31	0.20
2	2.40	1.30	1.79	2.39	1.30	1.97	2.47	1.25	1.32	2.45	1.31	1.01
3	2.44	1.30	1.73	2.43	1.30	1.89	2.34	1.24	1.29	2.30	1.32	1.08
4	2.52	1.30	1.70	2.52	1.30	1.89	2.55	1.30	1.40	2.48	1.32	1.05
5	2.46	1.30	1.75	2.45	1.30	1.88	2.40	1.35	1.50	2.48	1.30	1.10
10	2.39	1.30	1.81	2.39	1.30	1.85	2.36	1.28	1.47	2.42	1.35	1.07
						N = 100						
1	0.50	0.25	0.74	0.52	0.25	1.05	0.50	0.26	–0.16	0.55	0.22	–0.40
2	0.51	0.25	0.50	0.48	0.25	0.69	0.50	0.25	0.23	0.57	0.25	–0.16
3	0.49	0.25	0.49	0.46	0.25	0.65	0.50	0.25	0.28	0.44	0.19	–0.15
4	0.50	0.25	0.48	0.52	0.25	0.63	0.51	0.26	0.31	0.60	0.27	–0.05
5	0.50	0.25	0.47	0.51	0.25	0.64	0.50	0.26	0.34	0.60	0.25	–0.07
10	0.50	0.25	0.46	0.50	0.25	0.58	0.51	0.26	0.37	0.47	0.19	–0.03

[†]RAN: contributions are assigned at random; PED: contributions are optimised for minimising f_p; MOL_{OVE}: contributions are optimised for minimising f_{m_ove}.

(strategy RAN) or when minimising f_p (strategy PED) Δf_p was very similar to Δf_{m_ove}.Larger differences between both rates were however observed when contributions were optimised for minimising f_{m_ove} (strategy MOL_{OVE}). For μ = $2.5x10^{-5}$, Δf_{m_ove} was lower under MOL_{OVE} than under PED across all generations for both values of *N*. This low mutation rate implies strong LD between markers and non-marker loci (for μ = $2.5x10^{-5}$, the average squared correlation coefficient between adjacent pairs of SNPs (r^2) at *t* = 0 was 0.40 and 0.21 for *N* = 20 and 100, respectively) and thus, an efficient management. However, with a higher mutation rate ($\mu = 2.5x10^{-3}$) LD would be weaker (0.28 and 0.13 for *N* = 20 and 100, respectively) and the advantage of MOL_{OVE} over PED in terms of Δf_{m_ove} is only observed at early generations. In any case, the level of f_{m_ove} in all generations when applying MOL_{OVE} was lower than that observed when applying PED (data not shown). Counterintuitively, under strategy MOL_{OVE}, Δf_{m_ove} was slightly lower at *t* = 1with $\mu = 2.5x10^{-3}$ than with the lower μ and *N* = 20. This is probably due to the fact that the initial LD was high enough with the former value of μ and no extra benefits were observed for a lower value of μ.

Table 2.4 shows the rate of molecular coancestry separately for the targeted regions and for the rest of the genome under strategies MOL_{CHR} and MOL_{REG}. The OC method was efficient in avoiding the loss of diversity at the region(s) considered. In fact, the rates of coancestry at the targeted region(s) took even negative values at least in early generations. The OC method was more successful in reducing coancestry in a fraction of the genome when such a fraction included an entire chromosome than when included ten smaller regions located at different chromosomes although the proportion of the genome for which coancestry was minimised was the same (5%). In both cases, the success in retaining diversity at specific regions had undesired consequences in the rest of the genome given that the observed $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ were high. When an

Table 2.4. Rates of coancestry at specific genome regions (Δf_{m_chr} and Δf_{m_reg}) and at the whole genome except those regions ($\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$) across generations ($t$) when applying different management strategies (MOL_{CHR}, MOL_{CHR_CON}, MOL_{REG} and MOL_{REG_CON}) in populations of two different sizes (N = 20 and 100). Two different constraints (C = 1.0% and 0.1%) were imposed on $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ when applying strategies MOL_{CHR_CON} and MOL_{REG_CON}, respectively. The mutation rate used to create the base population was $\mu = 2.5 \times 10^{-5}$.

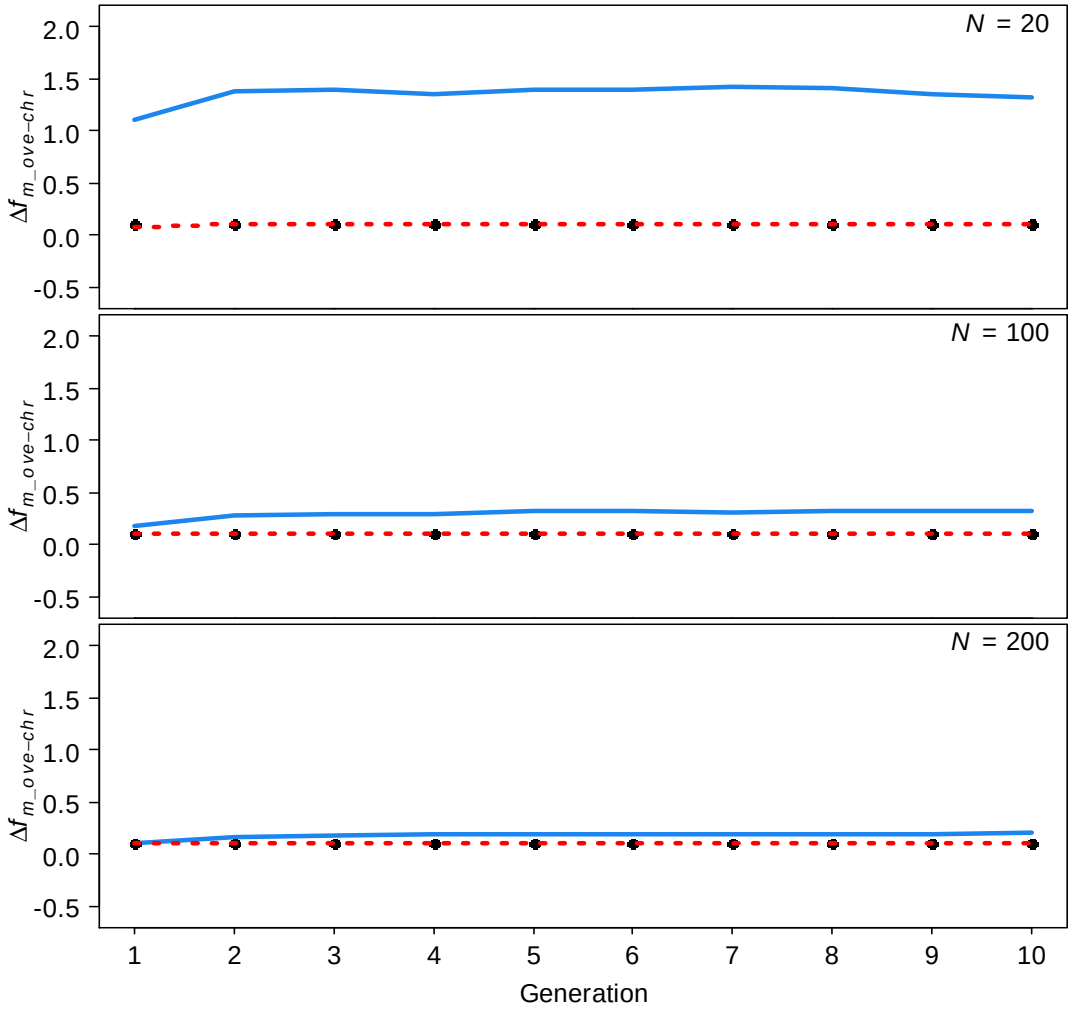
	MOL_{CHR}[†]		MOL_{CHR_CON}				MOL_{REG}		MOL_{REG_CON}			
			C = 1.0%		C = 0.1%				C = 1.0%		C = 0.1%	
t	Δf_{m_chr}	$\Delta f_{m\_ove\text{-}chr}$	Δf_{m_chr}	$\Delta f_{m\_ove\text{-}chr}$	Δf_{m_chr}	$\Delta f_{m\_ove\text{-}chr}$	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$
					N = 20							
1	–4.64	6.28	–4.33	2.27	–3.79	1.47	–2.18	3.86	–2.06	2.42	–1.78	1.67
2	–1.08	4.34	–1.09	2.44	–0.54	1.52	–0.39	3.22	–0.64	2.26	–0.16	1.59
3	–0.06	4.20	–0.28	2.32	–0.22	1.40	–0.33	3.31	–0.32	2.32	–0.05	1.61
4	–0.06	4.04	0.14	2.38	0.04	1.50	–0.25	3.05	0.23	2.46	0.09	1.59
5	0.30	3.82	0.22	2.37	0.14	1.48	0.10	2.98	0.25	2.30	0.23	1.67
10	0.42	3.28	0.37	2.40	0.35	1.40	0.54	2.81	0.35	2.25	0.40	1.45
					N = 100							
1	–4.69	3.43	–4.48	1.45	–4.18	0.67	–2.34	1.74	–1.93	1.30	–1.77	0.47
2	–1.83	1.70	–2.00	1.44	–1.94	0.65	–0.36	1.21	–0.40	1.16	–0.41	0.45
3	–0.81	2.17	–0.82	1.26	–0.86	0.34	–0.13	1.14	–0.17	1.12	–0.21	0.46
4	–0.67	1.32	–0.75	1.35	–0.91	0.52	–0.10	1.03	–0.07	1.05	0.17	0.43
5	–0.28	2.02	–0.36	1.23	–0.30	0.45	–0.10	1.00	0.10	1.03	0.17	0.45
10	–0.20	1.44	–0.17	1.16	–0.07	0.42	0.15	0.98	0.13	0.89	0.28	0.42

[†]MOL_{CHR}: contributions are optimised for minimising f_{m_chr}; MOL_{CHR_CON}: contributions are optimised for minimising f_{m_chr} while restricting $\Delta f_{m\_ove\text{-}chr}$; MOL_{REG}: contributions are optimised for minimising f_{m_reg}; MOL_{REG_CON}: contributions are optimised for minimising f_{m_reg} while restricting $\Delta f_{m\_ove\text{-}reg}$.

extra restriction was added for avoiding an excessive loss of diversity in the rest of the genome (strategies MOL_{CHR_CON} and MOL_{REG_CON}), the optimisation has still some success. Note that $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ were lower with than without the restriction and that the stricter the constraint the lower were $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$. However, the realised $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ were higher than the intended value (1.0% or 0.1%), particularly for the smallest population size considered (i.e., N = 20). In order to investigate if this unexpected result was a consequence of a failure in the optimisation process not finding a solution that meets the imposed restriction, we calculated the expected $\Delta f_{m\_ove\text{-}chr}$ by substituting the realised $f_{m\_ove\text{-}chr\ (t+1)}$ for the value of the objective function (**c'Gc**) that provides the optimal solution (i.e., $\Delta f_{m\_ove\text{-}chr\ (t+1)} = (\mathbf{c'}_{(t)}\mathbf{G}_{\boldsymbol{ov}\text{-}chr(t)}\mathbf{c}_{(t)} - f_{m\_ove\text{-}chr(t)})/(1 - f_{m\_ove\text{-}chr\ (t)})$). By doing so, we found that the expected $\Delta f_{m\_ove\text{-}chr}$ always met the restriction. This indicates that the solutions from the optimisation were valid in the sense that those found as optimum did, indeed, fulfil the restriction that the expected $\Delta f_{m\_ove\text{-}chr}$ should not be greater than 1.0% or 0.1% (i.e., the optimisation method performs well). This can be observed in Figure 2.1, where results for an additional population size (N = 200) was included to investigate the trend with respect to N. Figure 2.1 shows that the difference between expected and observed $\Delta f_{m\_ove\text{-}chr}$ clearly decrease with increasing N.

In addition, another set of simulations were performed to further investigate the discrepancy between the observed and expected $\Delta f_{m\_ove\text{-}chr}$. For N = 20 and t = 0, the optimisation was run for obtaining the optimum **c** to produce t = 1. Then, 1,000 replicates of offspring generations (i.e., t = 1) were obtained using always the optimum **c**. This was performed three times using three different initial populations, i.e., three different populations at t = 0 (and three different optimum **c** vectors). Figure 2.2 shows

Figure 2.1. Expected (dotted lines) and observed (straight lines) rate of molecular coancestry computed for the whole genome except for chromosome 1 ($\Delta f_{m\_ove\text{-chr}}$, in %) in the offspring generation when the optimisation strategy was MOL_{CHR_CON} with a restriction on the rate of coancestry in the rest of the genome of 0.1% for three population sizes (N = 20, 100 and 200). The specific imposed restrictions are indicated as filled circles.



the results for the three sets. In all cases the observed value for $f_{m\_ove\text{-}chr}$ was higher than the expected value.

Table 2.5 shows the results for strategies MOL_{OVE} and MOL_{OVE_CON} where contributions were optimised for minimising overall coancestry (f_{m_ove}) with or without

a restriction on the increase of coancestry at specific regions of the genome (i.e., the strategy opposite to MOL_{REG_CON}). Minimising overall coancestry lead to only marginal increases in Δf_{m_reg}. Thus, when imposing a restriction on Δf_{m_reg} such restriction needed to be very strict (C = 0.01%) for observing a reduction in this rate of coancestry. These results show that minimising overall coancestry is efficient for maintaining diversity in the targeted regions since these specific regions harbour levels of diversity similar to the rest of the genome. Restricting Δf_{m_reg} was successful and did not affect the rates at the rest of the genome.

Figure 2.2. Distribution of observed average molecular coancestry in the offspring generation of three set of parents. For each set of parents 1000 offspring generations were created using the same parental optimised contributions. Population size was 20, and the optimisation strategy used was MOL_{CHR_CON} with a restriction on the rate of overall coancestry restriction of 0.1%. Dotted lines indicate the targeted coancestry for each set of parents.

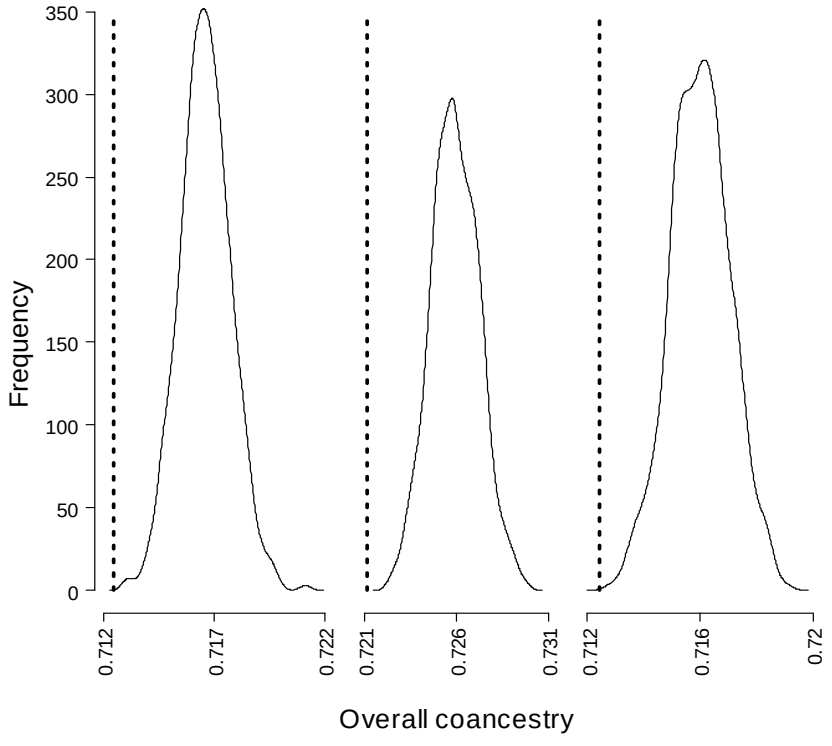


Table 2.5. Average rate of coancestry across specific regions (Δf_{m_reg}) and at the whole genome except those regions ($\Delta f_{m\_ove\text{-}reg}$) across generations ($t$) when applying two different management scenarios (MOL_{OVE} and MOL_{OVE_CON}) in populations of two different sizes (N = 20 and 100). Two different constraints (C = 1.0% and 0.10%) were imposed on Δf_{m_reg} when applying strategy MOL_{OVE_CON}. The mutation rate used to create the base population was $\mu = 2.5x10^{-5}$.

			MOL_{OVE_CON}			
	MOL_{OVE}†		C = 0.10%		C = 0.01%	
t	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$	Δf_{m_reg}	$\Delta f_{m\_ove\text{-}reg}$
			N = 20			
1	0.35	0.23	0.35	0.25	0.35	0.25
2	1.23	1.07	0.81	1.05	0.8	1.09
3	1.33	1.14	1.27	1.15	0.67	1.23
4	0.89	1.08	1.23	1.15	0.83	1.05
5	0.98	1.06	0.92	1.06	0.76	1.19
10	0.81	1.07	0.68	1.07	0.84	1.16
			N = 100			
1	–0.48	–0.51	–0.49	–0.49	–0.49	–0.47
2	–0.15	–0.16	–0.19	–0.11	–0.16	–0.12
3	–0.18	–0.07	–0.15	–0.07	–0.19	–0.06
4	–0.05	–0.06	–0.07	–0.09	–0.05	–0.06
5	–0.25	–0.03	–0.27	–0.02	–0.29	–0.07
10	–0.05	–0.03	–0.05	–0.03	–0.02	–0.03

†MOL_{OVE}: contributions are optimised for minimising f_{m_ove}; MOL_{OVE_CON}: contributions are optimised for minimising f_{m_ove} while restricting Δf_{m_reg}.

Discussion

This work has shown that OC methods that make use of molecular coancestry calculated from dense panels of biallelic molecular markers are efficient for minimising the loss of genetic variability at specific genomic regions. These methods are also efficient for restricting the increase in coancestry that occurs in the rest of the genome

when focusing on specific regions. By including extra constraints in the optimisation, OC methods were able to mitigate this negative effect and still maintain genetic variability at the specific regions at similar levels to those maintained when targeting only these specific regions. However, contrary to what is expected under current genetic contributions theory, the realised rate of coancestry resulted higher than the restriction imposed.

It is well known that in the absence of molecular or genealogical information, keeping equal numbers of males and females and constant census sizes (i.e., equalizing contributions) is the most appropriate procedure to avoid the loss of genetic diversity. In the present study, genealogical relationships between individuals of the base population were assumed to be unknown (and individuals were assumed to be unrelated and non-inbred) and thus, when minimising Δf_p the optimal solution was to equalize contributions. This occurred not only at the first generation but also in subsequent generations given that the population remains homogeneous at the genealogical level. On the other hand, for strategies using molecular coancestry equalizing contributions was never the optimal solution because marker genotypes allow us to distinguish genetic relationships between pairs of individuals with the same degree of genealogical coancestry. In fact, strategy MOL_{OVE} led to lower Δf_{m_ove} than strategy PED using less individuals with unequal contributions, especially for N = 100 and $\mu = 2.5\text{x}10^{-5}$. This implies that, in addition to maintaining a higher level of genetic diversity, the use of genomic coancestry could have some economic advantages when managing genetic conservation programmes as less animals need to be maintained (i.e., those animals not contributing to the next generation could be discarded) and maintenance costs could be reduced.

As indicated above, Δf_{m_ove} was clearly lower under strategy MOL_{OVE} than under strategy PED when $\mu = 2.5x10^{-5}$ was used to create the base population (Table 2.3). This is because strategy MOL_{OVE} is based on realised relationships between individuals while strategy PED is based on expectations. However, for a higher mutation rate ($\mu = 2.5x10^{-3}$) this only occurred at $t = 1$ for $N = 20$ and at $t \leq 2$ for $N = 100$. This can be explained by the fact that $\mu = 2.5x10^{-3}$ led to lower linkage disequilibrium between markers and loci where coancestry rate was measured than $\mu = 2.5x10^{-5}$. In any case, f_{m_ove} remained lower across all generations with MOL_{OVE} than with PED in all scenarios investigated and this agrees with previous results of GÓMEZ-ROMANO et al. (2013).

Strategies MOL_{CHR} and MOL_{REG} were clearly very efficient in reducing the rate of coancestry and therefore in maintaining diversity at a targeted region(s) (Table 2.4), especially when the region consisted of a single entire chromosome. The better performance of MOL_{CHR} in comparison to MOL_{REG} is due to the fact that the loci located at the one-chromosome region are more closely linked than those distributed across ten different regions. The vector of contributions that minimise molecular coancestry computed from one set of loci is likely to lead to the minimisation of molecular coancestry computed from a different set of loci that is strongly linked to the former. This would not be the case for sets of unlinked loci as those located on different chromosomes. Optimal contributions that minimise coancestry at one of the ten regions will not necessarily minimise coancestry in the rest of the regions. As the algorithm has to find an average solution for ten independent regions, this solution will probably not be optimal for each of them and this leads to a lower diversity maintenance when compared with the one-chromosome region scenario.

In general, the efficiency in reducing the rate of coancestry at specific regions was accompanied by a substantial increase in the rate of coancestry in the rest of the genome as it was previously described by ROUGHSEDGE et al. (2008). Our work shows that this undesired consequence can be mitigated by imposing a constraint on the rate of coancestry in the rest of genome (strategies MOL_{CHR_CON} and MOL_{REG_CON}). However, an unexpected result observed when minimising the rate of coancestry at specific regions but imposing simultaneously a restriction on the rate of coancestry in the rest of the genome (strategies MOL_{CHR_CON} and MOL_{REG_CON}) was that the realised $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ were always higher than the value of the imposed restriction (i.e., 1.0% or 0.1%), particularly for the lowest *N* value (Table 2.3 and Figure 2.1). This was despite the fact that the optimisation algorithm found the optimal solution that fulfils the restriction; i.e., that the expected $\Delta f_{m\_ove\text{-}chr}$ and $\Delta f_{m\_ove\text{-}reg}$ should not be greater than the restriction imposed. Therefore, **c'Gc** is a biased estimator of the mean coancestry in the next generation when using genomic coancestry matrices to perform the optimisations. ROUGHSEDGE et al. (2008) also showed a clear discrepancy between the observed and expected rates of molecular inbreeding. Thus, in both cases, the application of the theory developed using genealogical information to molecular estimates of coancestry and inbreeding failed in its expectation highlighting thus the need of revising the theory. It should be noted that the bias is higher when *N* is lower and that this is the situation where the need for managing genetic diversity is stronger.

Several optimisation methods have been proposed and implemented in the past for solving the OC problem. They mainly fall into three different categories: i) Lagrange multipliers (MEUWISSEN, 1997; GRUNDY et al., 1998); ii) genetic algorithms (CARVALHEIRO et al., (2010); and iii) semidefinite programming (SDP) approaches (PONG-WONG & WOOLLIAMS, 2007). The Lagrange multiplier approach is fast and very

efficient, but it does not guarantee that the optimum solution will always be found (PONG-WONG & WOOLLIAMS, 2007). Also, including extra constraints under the Lagrange multiplier approach requires major recalculation of the equations needed to find the optimum solution. Methodologies based in genetic algorithms are very flexible in terms of adding or removing constraints but the sampling approach in which the method is based means that the optimality of the final solution is not possible to be verified. Also, it can be computer intensive depending on the constraints included. On the other hand, the SDP approach guarantees that the solution found is the optimum. The method is also fast and flexible as extra constraints can be easily added to the optimisation. Also, general software packages for solving contribution problems with SDP are already available (FUJISAWA et al., 2002; BORCHERS, 1999; WU & BOYD, 2000; BENSON & YINYU, 2005).

The main limitation of the SDP methodology is that the constraints and objective functions need to be convex, which for the situation considered here means that the coancestry matrices need to be positive definite. Such property should hold when the genomic matrices are calculated using the method proposed by NEJATI-JAVAREMI et al. (1997) as done here. However, in practice, it is likely that there will be missing genotypes for a proportion of the SNPs and thus the coancestries between each pair of individuals can be calculated with a slightly different set of SNPs, which under certain situations, may results on the genomic matrix being non-positive definite. The problem could be solved by adding a very small quantity to all diagonal elements in the matrix, so it becomes positive definite. However, the consequences in the optimality of the solution when adding extra terms to the diagonal are yet to be quantified.

Another potential problem which may appear is that the SDP implementation requires the inverse of the genomic matrices (see Appendix 2.1). However, inversion

may not be possible especially when considering small genomic regions. For instance, if two sibs inherit from their common parent the same haplotype for the region in question, their relationship with the rest of the candidates will be the same and the resulting matrix will be non-invertible. Similarly, when the number of SNPs available for calculating the genomic matrix is smaller than the number of candidates, the matrix will also be non-invertible. A solution for this problem could be to use the generalised inverse of the genomic matrix or to add an extra term to the diagonal. Further studies are still required to determine the consequence of using generalized inverse matrices in this context.

Appendix 2.1: Formulation of the optimisation of genetic contributions to minimise the loss of diversity as a standard semidefinite programming

The optimisation of genetic contributions to minimise the increase of average coancestry (i.e., the loss of genetic diversity) is reformulated as a standard semidefinite programming using the same approach as that proposed by PONG-WONG & WOOLLIAMS (2007). However, small variations in the definition of genetic relationships and refinements in the optimisation problem mean that equations representing the standard semidefinite programming are slightly different to that reported by PONG-WONG & WOOLLIAMS (2007). The purpose of this appendix is to briefly describe the precise reformulation of the optimisation problem used in this study.

PONG-WONG & WOOLLIAMS (2007) showed that the problem of optimising contributions can be reformulated as a standard semidefinite programming, and thereby solved using such approach. Following the same notation of VANDENBERGHE & BOYD (1996), the standard form for a semidefinite programming problem is:

Minimise $\mathbf{a'x}$

Subject to $\mathbf{Y} \geq \mathbf{0}, \mathbf{Y} = \mathbf{Y_0} + \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Y_i} x_\mathbf{i}$

where **a** is the vector of 'cost', **x** is the vector of n variables to be optimised, x_i is the ith element of **x**, **Y** is a positive semidefinite matrix with n+1 affine matrices ($\mathbf{Y}_i$, i = 0, 1, 2, …, n). The matrix inequality $\mathbf{Y} \geq \mathbf{0}$ means that **Y** is positive semidefinite.

Optimisation problem 1

Following the same approach as PONG-WONG & WOOLLIAMS (2007), the reformulation of optimisation problem 1 is done by i) introducing an auxiliary variable *v* to serve as the upper limit of the objective function; ii) using the Shur complement to give a linear expression to quadratic constraint resulting from introducing; and iii) replacing the equality constraints for inequality ones. Hence, the problem 1 is reformulated as the optimisation of *v* and **c** to:

Minimise v

Subject to $\begin{bmatrix} \mathbf{G^{-1}} & \mathbf{c} \\ \mathbf{c} & v \end{bmatrix} \geq 0$

$\mathbf{c}'\mathbf{s} - 0.5 \geq 0$

$-\mathbf{c}'\mathbf{s} + 0.5 \geq 0$

$\mathbf{c}'\mathbf{d} - 0.5 \geq 0$

$-\mathbf{c}'\mathbf{d} + 0.5 \geq 0$

$\mathbf{c} \geq 0$

Then, matrix **Y** accounting for the six constraints is a block diagonal matrix of the form:

$$\boldsymbol{Y} = \begin{bmatrix} \begin{bmatrix} \mathbf{G^{-1}} & \mathbf{c} \\ \mathbf{c}' & v \end{bmatrix} & & & & & & \\ & [\mathbf{c}'\mathbf{s} - 0.5] & & & & & \\ & & [-\mathbf{c}'\mathbf{s} + 0.5] & & & & \\ & & & [\mathbf{c}'\mathbf{d} - 0.5] & & & \\ & & & & [-\mathbf{c}'\mathbf{d} - 0.5] & & \\ & & & & & [\mathrm{diag}(\mathbf{c})] \end{bmatrix}$$

with the (n + 2) affine matrices of **Y** equal to:

$$\mathbf{Y_0} = \begin{bmatrix} \begin{bmatrix} \mathbf{G^{-1}} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} & & & & & \\ & -0.5 & & & & \\ & & 0.5 & & & \\ & & & -0.5 & & \\ & & & & 0.5 & \\ & & & & & [\mathbf{0}] \end{bmatrix}$$

$$\mathbf{Y_i} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{I_i} \\ \mathbf{I_i'} & 0 \end{bmatrix} & & & & & \\ & s_i & & & & \\ & & -s_i & & & \\ & & & d_i & & \\ & & & & -d_i & \\ & & & & & [\mathbf{diag(I_i)}] \end{bmatrix}, i = 1, n$$

and

$$\mathbf{Y_{n+1}} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & [\mathbf{0}_{(nxn)}] \end{bmatrix}$$

where the size of the first block is (n+1)x(n+1), the next four are 1x1 and the last one is of size nxn. $\mathbf{0}_{(jxk)}$ are matrices/vectors of zeros of size jxk, $\mathbf{I}_i$ is the ith column of the identity matrix of size nxn and **diag(I$_i$)** is a diagonal matrix with diagonal equal to $\mathbf{I}_i$. All elements outside the block diagonal matrices are zero. Note that the formulation described above differs from that given in equation (8) of PONG-WONG & WOOLLIAMS (2007) by a constant value in first block of $\mathbf{Y_0}$**.** This is to account for the difference in the definition of relationship matrix (i.e., here the relation matrix contains coefficients of coancestry between individuals; whilst it is twice this value for PONG-WONG & WOOLLIAMS (2007).

Optimisation problem 2

Reformulation of problem 2 is similar as before, but the *m* extra constraints need to be added. Hence, using the Shur complement again, the formulation of (2) becomes:

Minimise v

Subject to $\begin{bmatrix} \mathbf{G^{-1}} & \mathbf{c} \\ \mathbf{c} & v \end{bmatrix} \geq 0$

$$\begin{bmatrix} \mathbf{G_j^{-1}} & \mathbf{c} \\ \mathbf{c} & K_j \end{bmatrix} \geq 0, j = 1, m$$

$\mathbf{c}$'$\mathbf{s}$ - $0.5 \geq 0$

-$\mathbf{c}$'$\mathbf{s}$ + $0.5 \geq 0$

$\mathbf{c}$'$\mathbf{d}$ - $0.5 \geq 0$

-$\mathbf{c}$'$\mathbf{d}$ + $0.5 \geq 0$

$\mathbf{c} \geq 0$

and the matrix **Y** is augmented to be:

$$\boldsymbol{Y} = \begin{bmatrix} \begin{bmatrix} \mathbf{G^{-1}} & \mathbf{c} \\ \mathbf{c}' & v \end{bmatrix} & \begin{bmatrix} \begin{bmatrix} \mathbf{G_1^{-1}} & \mathbf{c} \\ \mathbf{c}' & K_1 \end{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \begin{bmatrix} \mathbf{G_j^{-1}} & \mathbf{c} \\ \mathbf{c}' & K_j \end{bmatrix} \end{bmatrix} & & & & \\ & & [\mathbf{c}'\mathbf{s} - 0.5] & & & \\ & & & [-\mathbf{c}'\mathbf{s} + 0.5] & & \\ & & & & [\mathbf{c}'\mathbf{d} - 0.5] & \\ & & & & & [-\mathbf{c}'\boldsymbol{d} - 0.5] \\ & & & & & \end{bmatrix}$$

with the (*n+2*) affine matrices of **Y** equal to:

$$\mathbf{Y_0} = \begin{bmatrix} \begin{bmatrix} \mathbf{G^{-1}} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} & & & & & & & \\ & \begin{bmatrix} \begin{bmatrix} \mathbf{G_1^{-1}} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & K_1 \end{bmatrix} & \cdots & & & & & \\ \vdots & & \ddots & \vdots & & & & \\ & & \cdots & \begin{bmatrix} \mathbf{G_j^{-1}} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & K_j \end{bmatrix} & & & & \\ & & & & -0.5 & & & \\ & & & & & 0.5 & & \\ & & & & & & -0.5 & \\ & & & & & & & 0.5 \\ & & & & & & & [\mathbf{0}] \end{bmatrix}$$

$$\mathbf{Y_i} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{I_i} \\ \mathbf{I_i'} & 0 \end{bmatrix} & & & & & & & \\ & \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{I_i} \\ \mathbf{I_i'} & 0 \end{bmatrix} & \cdots & & & & & \\ \vdots & & \ddots & \vdots & & & & \\ & & \cdots & \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{I_i} \\ \mathbf{I_i'} & 0 \end{bmatrix} & & & & \\ & & & & s_i & & & \\ & & & & & -s_i & & \\ & & & & & & d_i & \\ & & & & & & & -d_i \\ & & & & & & & [\mathbf{diag}(\mathbf{I_i})] \end{bmatrix}, i = 1, n$$

and

$$\mathbf{Y}_{n+1} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} & & & & & & & \\ & \begin{bmatrix} \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} & \cdots & & \\ \vdots & & \ddots & \vdots & \\ & & \cdots & \begin{bmatrix} \mathbf{0}_{(nxn)} & \mathbf{0}_{(nx1)} \\ \mathbf{0}_{(1xn)} & 0 \end{bmatrix} \end{bmatrix} & & & & \\ & & & 0 & & & \\ & & & & 0 & & \\ & & & & & 0 & \\ & & & & & & 0 \\ & & & & & & & \begin{bmatrix} \mathbf{0}_{(nxn)} \end{bmatrix} \end{bmatrix}$$

Once the optimisation has been reformulated as an standard semidefinite programming problem, it can easily be solved using general purpose programmes already available. In this study, the software SDPA was used to solve the optimisation problem. It is important to notice that there is a slight difference in the definition for $\mathbf{Y}$ used here (i.e., $\boldsymbol{Y} = \mathbf{Y_0} + \sum_{i=1}^{n} \mathbf{Y}_i x_i$, adopted from VANDENBERGHE & BOYD (1996) and that used in SDPA (i.e., $\boldsymbol{Y} = -\boldsymbol{Y_0} + \sum_{i=1}^{n} \mathbf{Y}_i x_i$, (FUJISAWA et al., 2002)). Hence, from the practical point of view, the matrix $\mathbf{Y_0}$ described above needs to be multiplied by −1, before it is given as input to the SDPA programme.

CAPÍTULO 3:

The use of identical by descent segments for maintaining genetic diversity

Introduction

The maintenance of genetic diversity is one of the main objectives of genetic conservation programs. This objective is achieved by maximising the effective population size (N_e) or equivalently, by minimising the rate at which coancestry increases in the population. Also, it is generally accepted that Optimal Contribution (OC) methods (MEUWISSEN, 1997; GRUNDY et al., 1998; FERNÁNDEZ & CABALLERO, 2001) are the methods of choice for performing such minimisation. In particular, OC methods determine the number of offspring that each breeding candidate should have to minimise global coancestry in the following generation. Thus, the coancestry (f) coefficient is a key parameter in the genetic management of populations. Although, in principle, OC methods minimize the rate of coancestry, they also minimize the rate at which inbreeding increases in scenarios with random mating or in those where the level of non-randomness is constant across generations. This is important to note as inbreeding depression, which depends on the levels of inbreeding and not on coancestry, is also of paramount importance in conservation programs.

The current availability of high-density panels of SNP markers in most farm animal species permit us to obtain more accurate estimates of coancestry and inbreeding than those obtained traditionally from pedigree records. Rather than giving expected values as pedigree-based estimates do, genomic estimates reflect the realised proportions of homozygous loci in a particular individual (inbreeding) or the realised proportions of loci shared by two particular individuals (coancestry) (TORO et al. 2014). Different genomic estimates based on SNP genotypes have been proposed recently including i) estimates obtained on a SNP-by-SNP basis (NEJATI-JAVAREMI et al., 1997; TORO et al., 2002); and ii) estimates based on ROH (Run of Homozygosity) for

inbreeding (GIBSON et al., 2006) or on IBD (identical by descent) segments for coancestry (PRYCE et al., 2012; DE CARA et al., 2013a). The use of ROH for estimating inbreeding has been widely applied to different farm animal species such as pigs (BOSSE et al., 2012; PURFIELD et al., 2012; HERRERO-MEDRANO et al., 2013; SILIÓ et al., 2013; SAURA et al., 2015) and cattle (BJELLAND et al., 2013; FERENČAKOVIĆ et al., 2013a; 2013b; KIM et al., 2013; SCRAGGS et al., 2014). There are also examples of the use of IBD segments for estimating coancestry in real populations (ALBRECHTSEN et al., 2009; CAI et al., 2011; PRICE et al., 2011; CAMPBELL et al., 2012; GUSEV et al., 2012; PALAMARA et al., 2012; ZUK et al., 2012; RALPH & COOP, 2013; HAN & ABNEY, 2013; BROWNING & BROWNING; 2013) but research on the benefits of using this measure of coancestry for maintaining diversity is very limited (PRYCE et al., 2012). When the aim of the program is to maintain the highest levels of global neutral genetic diversity, DE CARA et al. (2011; 2013a) showed, through computer simulations, that the most powerful strategy was to minimize the molecular coancestry computed on a SNP-by-SNP basis and averaged across SNPs. However, this strategy also led to the accumulation of deleterious variants in the genome and, thus, to a decrease of the population's fitness. The use of coancestry calculated from long segments shared by individuals in the management of populations has been shown to provide a balanced solution between maintaining neutral variation and fitness levels (DE CARA et al., 2013b).

In order to obtain coancestry estimates based on IBD segments, phases of SNP genotypes need to be known. However, due to the characteristics of current genotyping techniques phases are not available and, therefore, they need to be inferred. Although there are many computational methods available to infer haplotypic phases from genotype data (BROWNING & BROWNING, 2011), the effect of inferring haplotype phases

on the accuracy of genomic coancestry based on IBD segments, and, by extension, on the efficiency of genetic population management based on this measure of coancestry is unknown.

The objective of this research was twofold. First, estimates of inbreeding and coancestry coefficients based on genealogical or genomic information for three Austrian dual-purpose cattle breeds (the Austrian Brown Swiss, Pinzgauer and Tyrol Grey breeds) were compared. Measures using molecular information included those obtained on a SNP-by-SNP basis and those calculated from ROH (inbreeding) or IBD segments (coancestry). Comparisons of the corresponding estimates of N_e were also performed. Second, we evaluated through computer simulations, the efficiency of using genomic estimates of coancestry based on IBD segments for managing the maintenance of genetic diversity when phases need to be inferred.

Material and methods

Data analyses of cattle breeds

Genotypic data from 465 Austrian Brown Swiss (BS), 219 Pinzgauer (PI) and 219 Tyrolean Grey (TG) bulls were available for the study. The three breeds are autochtonous traditional Austrian cattle breeds with different histories. The BS cattle breed census is relatively large (52,008 breeding females in 2013, according to the Domestic Animal Diversity Information System or DAD-IS, hosted by FAO) although it was established from a small population (SÖLKNER et al., 1998). The PI and TG breeds are smaller than BS (9,887 and 5,372 breeding females, respectively, in 2013 according to DAD-IS).

Genotyping was performed using two different SNP chips: the Illumina Bovine SNP50BeadChip v1 that contains 54,001 SNPs, and the Illumina BovineHD BeadChip that contains 786,799 SNPs. The number of animals genotyped with the 50K chip was 415, 101 and 99 for BS, PI and TG, respectively, and the number of animals genotyped with the HD chip was 50, 118 and 120 for BS, PI and TG, respectively. Only SNPs common for both chips were considered for the analyses. Unmapped SNPs and those mapped on sex chromosomes as well as monomorphic SNPs and those with more than 10% missing genotypes were excluded. The final number of SNPs used was 42,051, 40,418 and 43,015 for BS, GV and PI, respectively. After filtering the SNPs, animals with more than 5% missing genotypes for the remaining SNPs were also removed.

Two different genomic coancestry coefficients were obtained: i) coancestry coefficients obtained on a SNP-by-SNP basis; and ii) coancestry coefficients based on IBD segments. For a given SNP the SNP-by-SNP coancestry between individuals *i* and *j* was computed as $0.25\sum_{i=1}^{2}\sum_{j=1}^{2}\delta_{ij}$, where δ_{ij} is the allele sharing status which equals to 1 if allele *i* from the first individual is the same as allele *j* from the second individual and to 0 otherwise. Then, the total SNP-by-SNP coancestry between individuals *i* and *j* ($f_{MOL_{ij}}$) was computed as the average value across all SNPs in the genome. The coancestry coefficient based on IBD segments between individuals *i* and *j* ($f_{SEGi,j}$) was calculated as $f_{SEGi,j} = \sum_k \sum_{a_{i=1}}^{2} \sum_{b_{j=1}}^{2} L_{IBD_k}(a_i, b_j)/4L$, where L_{IBDk} (a_i,b_j) is the length of the *k-th* shared IBD segment measured over homologue *a* of individual *i* and homologue *b* of individual *j*, and *L* is the length of the genome (DE CARA et al., 2013a). The following criteria were used to define an IBD segment: i) no more than two SNPs with missing genotype in one or both of the individuals compared were permitted; ii) at most, one mismatch between homologues was allowed; iii) the minimum density was 1 SNP per 100 kb; iv) the maximum gap between consecutive SNPs was 1 Mb; and v) the

minimum length of the segment was 4 Mb. The last criterion to define an IBD segment was based on the work of FERENČAKOVIĆ et al. (2013b) who shown that when using the 50K chip, autozygosity is overestimated for segments shorter than 4 Mb. The inference of phases needed to obtain f_{SEG} was carried out using the BEAGLE software (BROWNING & BROWNING, 2007). Each chromosome was phased independently using 10 iterations and the default parameters.

Pedigree-based coancestry coefficient (f_{PED}) was obtained using the R´Tools software (PONG-WONG, personal communication) which is based on the algorithm of MEUWISSEN & LUO (1992). The genealogy was constructed with all the ancestors available for the genotyped individuals and included 5,642, 2,877 and 1,830 animals for BS, PI and TG, respectively. Values characterizing the quality of the genealogies, in terms of their 'depth' and completeness (maximum, complete and equivalent number of generations) for each breed were obtained from the software ENDOG (GUTIÉRREZ & GOYACHE, 2005).

Inbreeding coefficients computed on a SNP-by-SNP basis (F_{MOL}), and those based on ROH (F_{ROH}) or pedigree (F_{PED}) were also computed for each individual. For individual *i*, F_{MOL} was calculated as $F_{MOL_i} = 2f_{MOL_{ii}} - 1$ and F_{ROHi} was calculated as the length of the genome of *i* in ROH divided by the overall length of the genome (KELLER et al., 2011). The criteria to define a ROH were equivalent to those used to define an IBD segment. Genealogy-based inbreeding coefficients were obtained using the R´Tools software.

Rates of coancestry and inbreeding per year were calculated on a SNP-by-SNP basis ($\Delta F_{MOL(y)}$, $\Delta f_{MOL(y)}$) and based on ROH or IBD segments ($\Delta F_{ROH(y)}$, $\Delta f_{SEG(y)}$). Annual pedigree-based rates ($\Delta F_{PED(y)}$, $\Delta f_{PED(y)}$) were also computed. Rates of change in

f_{MOL}, f_{SEG} and f_{PED} per year ($\Delta f_{MOL(y)}$, $\Delta f_{SEG(y)}$ and $f_{PED(y)}$) were computed by regressing the natural logarithm of (1 □ f) for each pair of individuals on year of birth. The slopes of these regressions are approximately equal to $-\Delta f_{MOL(y)}$, $-\Delta f_{SEG(y)}$ and $-\Delta f_{PED(y)}$ (HINRICHS et al., 2007). Rates of change in coancestry per generation (Δf_{MOL}, Δf_{SEG} and Δf_{PED}) were calculated as $L\Delta f_{MOL(y)}$, $L\Delta f_{SEG(y)}$ and $L\Delta f_{PED(y)}$, where L is the generation interval that was calculated as the average age of the parents at the time of birth of their offspring. Equivalent rates of change in inbreeding per generation (ΔF_{MOL}, ΔF_{ROH} and ΔF_{PED}) were obtained in the same way. Finally, the effective population size was calculated from the rate of change in coancestry per generation as $N_{e\_f(MOL)} = 1/2\Delta f_{MOL}$, $N_{e\_f(SEG)} = 1/2\Delta f_{SEG}$ and $N_{e\_f(PED)} = 1/2\Delta f_{PED}$ and from the rate of change in inbreeding per generation as $N_{e\_F(MOL)} = 1/2\Delta F_{MOL}$, $N_{e\_F(ROH)} = 1/2\Delta F_{ROH}$ and $N_{e\_F(PED)} = 1/2\Delta F_{PED}$.

Computer simulations

Stochastic computer simulations were carried out to evaluate the efficiency in maintaining genetic diversity when using estimated phases to compute coancestry based on IBD segments (f_{SEG_E}) in comparison with using the true phases (f_{SEG_T}). Real genotypes of BS bulls were used to create the base population. The sex of the animals was randomly assigned but assuring that half of the animals were males and half females. Samples of two different sizes (N = 20 and 100) were taken at random from the whole BS genotypic data set (i.e., from the 415 genotyped animals) to establish the base population for each replicate. Phases at t = 0 were estimated using all 465 BS animals available (rather than just the 20 or 100 animals simulated) in order to increase the accuracy of haplotype phases estimates (BROWNING & BROWNING, 2011).

From the base population, 10 generations of management aimed at minimising the loss of genetic diversity were simulated. Full details on how generations were

created are given in GÓMEZ-ROMANO et al. (2013). The management objective was to optimize contributions of candidates for minimising overall coancestry under different strategies (see below). Once the contributions were optimized, matings between selected breeders were at random.

The OC problem was formulated as to minimize $\mathbf{c'\Phi c}$, subject to $\mathbf{c's} = 0.5$, $\mathbf{c'd} = 0.5$ and $c_i \geq 0$ where $\mathbf{c}$ is the vector of genetic contributions of the candidates, $\mathbf{\Phi}$ is the coancestry matrix, and $\mathbf{s}$ and $\mathbf{d}$ are vectors containing flags that indicate the sex of the candidates, with $s_i = 1$ if candidate i is a male and 0 if it is a female, and $\mathbf{d} = \mathbf{1} - \mathbf{s}$. The optimization method followed was the semidefinite programming approach described in detail in PONG-WONG & WOOLLIAMS (2007) and implemented using the software SDPA (FUJISAWA et al., 2002).

Management strategies differed in the type of coancestry matrix used in the optimization: i) $\mathbf{\Phi}$ containing coancestry coefficients based on IBD segments computed from real phases; or ii) $\mathbf{\Phi}$ containing coancestry coefficients based on IBD segments computed from estimated phases. Genomic coancestries were calculated using all SNPs that met the quality control criteria described previously (i.e., 42,051 SNPs). At $t > 0$ real known simulated phases or estimated phases were used. The accuracy of the reconstruction of haplotypes was evaluated through the switch error rate (SER) that is defined as the number of switches required to obtain the true haplotype phases from the estimated phases divided by the number of heterozygote loci in the genotype minus one (STEPHENS & DONNELLY, 2003).

In order to compare the performance of different strategies in maintaining diversity, average f_{PED}, f_{MOL}, f_{SEG_T} and f_{SEG_E} and their corresponding rates of increase per generation (Δf_{PED}, Δf_{MOL}, Δf_{SEG_T} and Δf_{SEG_E}, respectively) were calculated each

generation. The number of individuals that contributed to the offspring generation and the variance of contributions were also calculated each generation. Each management scenario was replicated 100 times and results were averaged across replicates.

Results

Table 3.1 summarises pedigree information and the number of genotyped individuals for each cattle breed as well as estimates of N_e obtained from genomic or

Table 3.1. Generation interval (L, in years), number of animals in the pedigree (N_{ped}), number of animals genotyped (N_{gen}), maximum (G_{max}), complete (G_{com}) and equivalent (G_{equ}) numbers of generations, and effective population size computed from rates of coancestry (N_{e_f}) or inbreeding (N_{e_F}) for the three cattle breeds. Rates of coancestry and inbreeding were computed from genomic data on a SNP-by-SNP basis (subscript *MOL*) or from IBD segments (subscript *SEG*) or from pedigree data (subscript *PED*).

	Breed		
	Brown Swiss	Pinzgauer	Tyrolean Grey
L	6.64	6.77	6.81
N_{ped}	5,642	2,877	1,830
N_{gen}	465	219	219
G_{max}	4.68	2.48	7.11
G_{com}	1.52	0.88	1.92
G_{equ}	2.50	1.44	3.46
$N_{e\_f(MOL)}$	26.24	83.93	60.68
$N_{e\_f(SEG)}$	26.52	98.47	63.29
$N_{e\_f(PED)}$	35.68	108.51	57.27
$N_{e\_F(MOL)}$	31.90	83.93	92.94
$N_{e\_F(ROH)}$	41.60	111.9	87.41
$N_{e\_F(PED)}$	51.57	98.47	83.43

genealogical data. The degree of completeness of each genealogy did not correlate with the number of individuals in it. For example, the smallest pedigree (corresponding to TG) showed the highest maximum, complete and equivalent number of generations.

The generation interval was very similar for the three breeds. The highest annual rates of increase in both coancestry and inbreeding were observed for the BS breed (data not shown). Consequently, BS showed the lowest N_e estimates. For a particular breed, similar estimates of N_e were obtained from the three different inbreeding measures. This was also the case when N_e estimates were obtained from coancestry coefficients. Except for genealogical estimates for PI, N_{e_f} was always lower than N_{e_F}.

Table 3.2. Intercept (*a*), regression coefficient (*b*) and correlation (*r*) between different coancestry (*f*) or inbreeding (*F*) coefficients for the three cattle breeds[†].

			Breed								
			Brown Swiss			Pinzgauer			Tyrolean Grey		
Regression of			*a*	*b*	*r*	*a*	*b*	*r*	*a*	*b*	*r*
f_{SEG}	on	f_{MOL}	0.67	0.36	0.98	0.66	0.39	0.94	0.66	0.37	0.99
f_{PED}	on	f_{MOL}	0.69	0.33	0.85	0.66	0.33	0.84	0.66	0.35	0.95
f_{PED}	on	f_{SEG}	0.06	0.91	0.85	0.02	0.9	0.91	0.02	0.92	0.96
F_{ROH}	on	F_{MOL}	0.71	0.32	0.95	0.7	0.34	0.85	0.71	0.3	0.94
F_{PED}	on	F_{MOL}	0.73	0.34	0.64	0.7	0.39	0.65	0.71	0.26	0.68
F_{PED}	on	F_{ROH}	0.06	1.00	0.62	0.02	0.91	0.70	0.02	0.83	0.70

[†]f_{PED}: pedigree-based coancestry; f_{MOL}: genomic coancestry computed on a SNP-by-SNP basis; f_{SEG}: genomic coancestry based on IBD segments; F_{PED}: pedigree-based inbreeding; F_{MOL}: genomic inbreeding computed on a SNP-by-SNP basis; F_{ROH}: genomic inbreeding based on IBD segments.

The correlations between different coefficients of coancestry (Table 3.2) were very high for the three breeds (they ranged from 0.84 to 0.99), especially those between the marker-based coefficients that were always higher than 0.90. The correlations between both marker-based coefficients of inbreeding were also very high (they ranged from 0.85 to 0.95). The lowest correlations were those between genealogical and marker-based inbreeding coefficients that ranged from 0.62 to 0.70.

When predicting a particular coancestry (inbreeding) measure from another measure, the ideal situation (i.e., the perfect predictor) would produce a regression line with intercept $a = 0$ and slope $b = 1$. Intercepts far from zero and slopes far from one were observed for the regressions of f_{PED} (F_{PED}) on f_{MOL} (F_{MOL}) for the three breeds. This is a reflection of the fact that SNP-by-SNP measures reflect both IBD and IBS while genealogical measures reflect only IBD. Contrarily, regressions involving pedigree and segment-based measures showed intercepts close to zero and slopes close to one. This indicates that the latter (f_{SEG} and F_{ROH}) reflect IBD well despite of being calculated from molecular information. Although f_{SEG} and F_{ROH} predict well f_{PED} and F_{PED}, respectively, the accuracy of the predictions was higher for coancestry than for inbreeding as reflected in the correlations shown in Table 3.2.

Table 3.3 shows the simulation results for the management strategy aimed at minimising the genomic coancestry based on IBD segments when they are obtained using true or estimated phases. For $N = 100$, Δf_{SEG_T} was negative in the first generation and close to zero (but still negative) in subsequent generations when using f_{SEG_T} or f_{SEG_E}. It is remarkable that, for this population size, f_{SEG} was lower at the end of the management period than in the base generation. Estimating phases rather than using the true phases when managing the population for minimising coancestry lead to negligible losses in diversity (0.05% at the end of the process). For $N = 20$, rates of coancestry

were always positive in all generations. At $t = 1$, Δf_{SEG_T} was lower when management relied on real phases than on estimated phases, as it was expected. Rates of coancestry using real or estimated phases were more similar at $t > 1$, which could be explained by the fact that the correlation between f_{SEG_T} and f_{SEG_E} (ρ) increases with increasing t and the SER value decreases with increasing t. After ten generations of management the average coancestry values were almost identical when using real or estimated phases (the difference was of 0.3%).

Table 3.3. Average coancestry coefficients (f_{SEG_T}, in %) and rates of coancestry (Δf_{SEG_T}, in %) based on true IBD segments when true or estimated phases are used in the optimization for populations of two different sizes (N). The correlation between f_{SEG_T} and f_{SEG_E} (ρ) and the accuracy of the phasing process (SER) are also presented.

			True phases		Estimated phases	
t	ρ	*SER*	f_{SEG_T}	Δf_{SEG_T}	f_{SEG_T}	Δf_{SEG_T}
			$N = 20$			
0	0.882	0.194	9.16	−	9.16	−
1	0.865	0.178	9.22	0.07	9.37	0.23
2	0.890	0.157	9.85	0.69	10.05	0.75
3	0.916	0.136	10.47	0.69	10.69	0.71
4	0.935	0.118	11.04	0.64	11.27	0.65
5	0.949	0.103	11.61	0.64	11.84	0.64
10	0.975	0.058	13.90	0.48	14.21	0.49
			$N = 100$			
0	0.989	0.033	7.30	−	7.30	−
1	0.996	0.013	4.95	−2.54	5.01	−2.47
2	0.996	0.012	4.78	−0.17	4.84	−0.17
3	0.996	0.011	4.62	−0.18	4.67	−0.17
4	0.996	0.011	4.48	−0.15	4.53	−0.15
5	0.996	0.010	4.36	−0.13	4.41	−0.13
10	0.996	0.010	3.94	−0.06	3.99	−0.06

Given the values of SER, the good performance of the management that used estimated phases is a reflection of the accuracy of the reconstruction. In our simulations, SER values always decreased over generations. At $t = 0$ SER was very low for $N = 100$ (3%) and much higher for $N = 20$ (20%), although this latter figure decreased up to 6% at $t = 10$ (Table 3.3).

Discussion

The first objective of this study was to compare the measure of coancestry based on IBD segments with other genomic measure (namely coancestry calculated in a SNP-by-SNP basis) and with the genealogical coancestry. Using data from three Austrian autochthonous cattle breeds, coancestry and inbreeding estimates based on IBD segments and ROH, respectively showed to be good estimators of genealogical coancestry and inbreeding. This could be a reflection of the fact that f_{SEG} is a measure of the realised IBD coancestry in contrast with f_{MOL}, that mixes both IBD and IBS coancestries and is then a poor estimator of f_{PED}. A second objective was to explore the suitability of coancestry coefficient based on IBD segments for maintaining diversity in conserved populations, taking into account the fact that calculation of this kind of coancestry requires estimating SNPs haplotypic phases in real scenarios. Simulation results showed that the levels of diversity achieved using f_{SEG} from inferred or true phases were very similar when optimising contributions.

Using data from the three cattle breeds we found that rates of change in coancestry and inbreeding and the corresponding N_e estimated from genomic (SNP-by-SNP or IBD segments) and pedigree information were very similar, as expected (SAURA et al., 2013). Therefore, when the aim is to describe the evolution of the population N_e any of the measures used here are equally useful. Except for pedigree-based estimates

for PI, values for N_{ef} were consistently lower than those for N_{eF}, which suggests that matings between close relatives have been avoided in the three breeds. The exception for PI could be simply a consequence of the quality of the pedigree information available for this breed that was clearly the lowest, as it is reflected in the maximum, complete and equivalent number of generations. The low N_e found for the three breeds in the study is compatible with heavy use of few sires in selection (BOICHARD et al., 1997). The even lower N_e of BS, despite of having a higher current population size, could be explained by the origins of the breed that was founded from a very small number of animals (SÖLKNER et al., 1998).

The slope (close to one) and intercept (close to zero) of the regression of pedigree-based coefficients on coefficients based on ROH (inbreeding) or IBD segments (coancestry) suggest that these molecular measures are good predictors of pedigree-based measures and that they reflect the IBD status well. These results are similar to those found in previous studies for coancestry (RODRÍGUEZ-RAMILO et al., 2015) and for inbreeding (KELLER et al., 2011; FERENČAKOVIĆ et al., 2013a). The advantage of IBD segment (ROH) estimates over pedigree-based estimates is that they reflect realised coancestries (inbreeding) rather than expected values.

Although several studies have indicated that it would be beneficial to employ molecular-based coancestries in the management of populations to reduce the loss of diversity in conservation and selection programs (e.g., DE CARA et al., 2011; ENGELSMA et al., 2011; SONESSON et al., 2012; GOMEZ-ROMANO et al., 2013; SAURA et al. 2013), there is an important issue that requires discussion. Strategies leading to higher genetic diversity can also lead to a decrease in fitness given that maintaining diversity implies maintaining deleterious alleles. In this context, previous studies showed that although the use of f_{MOL} computed from dense SNPs panels is the best strategy for maintaining

neutral diversity (DE CARA et al., 2011; GÓMEZ-ROMANO et al., 2013), it also leads to reduced values of fitness in the population (DE CARA et al., 2013a). Using f_{SEG} has been suggested to be a suitable strategy for achieving a balance between maintaining diversity and fitness (DE CARA et al., 2013b). The same conclusion can be extracted from the study of PRYCE et al. (2012) who found that the use of f_{SEG} calculated from relatively small segments was as useful as the use of f_{MOL} for controlling the rate of inbreeding in dairy herds, but led to a reduction of the occurrence of deleterious homozygotes due to recent inbreeding.

The drawback of using f_{SEG} is that phases need to be estimated. There are many methods to estimate haplotypic phases from SNP genotypes, of which coalescent-based methods and hidden Markov models are the most frequently used (BROWNING & BROWNING, 2011). BEAGLE is a hidden Markov model based method that allows us to perform genome wide scale phasing without being highly time consuming. The switch error rate (or its complementary measure switch accuracy rate) is usually the most informative metric to measure haplotypic phase accuracy (BROWNING and BROWNING, 2011). The accuracy achieved in inferring phases depends on several parameters, among which the number of individuals available is one of the most important. SER values for different phasing methods in the literature are usually lower than 5% for populations of at least 1,000 individuals (BROWNING & BROWNING, 2011). In the present study we have considered much smaller populations (N = 20 and N = 100). Notwithstanding, for N = 100 figures for SER were similar to those previously described for larger populations. This was not the case for the N = 20 scenario for which SER was up to 20% in the first generation of management. However, in all cases SER decreased quickly across generations (0.058 at t = 10).

The relatively high accuracies of estimated phases led to high correlations (especially for N = 100) between true and estimated coancestry matrices and very similar results in maintaining diversity in the simulations performed. These results suggest that f_{SEG} calculated using inferred phases is an efficient coancestry measure to be used for maintaining diversity. Although phases in t = 0 still had to be estimated, starting from real genotypes of BS bulls provide the advantage of making less assumptions about the distance between markers, allelic frequencies or linkage disequilibrium than those if genotypes were simulated. This leads to a more realistic base population than that obtained when genotypes are simulated.

DISCUSIÓN GENERAL

El desarrollo de las plataformas de genotipado masivo de polimorfismos de un solo nucleótido (SNPs) en los últimos años ha llevado a la necesidad de realizar una re-evaluación de la utilización de marcadores genéticos en los programas de selección y de conservación. En esta tesis se han evaluado diferentes aplicaciones de estas herramientas genómicas en la gestión de la diversidad genética y en el control de la consanguinidad en programas de conservación. En el Capítulo 1 se demuestra que el uso de información molecular para minimizar la pérdida de diversidad mediante la optimización de contribuciones de los reproductores, permite mejorar los resultados obtenidos con la información genealógica. Resultados de esta investigación muestran que la densidad de SNPs necesaria para que la gestión de poblaciones que utiliza el parentesco molecular iguale la variabilidad genética mantenida a través de la gestión que utiliza el parentesco genealógico es $3N_e$ SNPs/M. En el Capítulo 2, se demuestra que la información genómica permite mantener (e incluso aumentar) la diversidad en regiones específicas del genoma de manera muy eficiente y restringir al mismo tiempo la pérdida de diversidad en el resto del genoma. En el Capítulo 3 se muestra que el parentesco genómico calculado a partir de segmentos IBD refleja efectivamente identidad por descendencia, con la ventaja, frente al parentesco genealógico, de permitir conocer la proporción concreta del genoma compartido por dos individuos en lugar de disponer solo de un valor esperado. Además, se demuestra que la necesidad de estimar las fases gaméticas para obtener el parentesco basado en segmentos IBD no lleva consigo una pérdida en la eficacia de la gestión de poblaciones que utiliza esta medida de parentesco a la hora de mantener la diversidad genética.

A lo largo de esta tesis se han empleado dos tipos de parentesco basados en información molecular: uno calculado SNP a SNP (f_{SNP}) y otro basado en segmentos

IBD (f_{SEG}). Mientras que f_{SEG} está en la misma escala que el parentesco obtenido a partir del pedigrí (f_{PED}), f_{SNP} está a una escala diferente y su magnitud es mucho mayor que la de f_{SEG} o f_{PED}. Ello se debe a que mientras que f_{SEG} básicamente refleja IBD, f_{SNP} no discrimina IBD de IBS. En el pasado, se desarrollaron una variedad de métodos para corregir por la identidad en la población base y poner así f_{SNP} en la misma escala que f_{SEG} o f_{PED} (TORO et al., 2002; TORO et al., 2014). Este tipo de correcciones son frecuentes por ejemplo, en el contexto de las evaluaciones genéticas (VANRADEN, 2008), donde frecuentemente parentescos obtenidos a partir de datos genómicos tienen que combinarse con parentescos obtenidos con datos genealógicos (no todos los candidatos han sido genotipados) y, por lo tanto, tienen que estar en la misma escala. El problema con estas correcciones es que están basadas en las frecuencias alélicas de la población base, que normalmente son desconocidas. Por ello, en la práctica, se utilizan frecuencias estimadas en la población actual. Esto no constituye ningún problema si el número de generaciones que han pasado desde la población base hasta la actualidad es relativamente pequeño y si N_e es grande. Cuando esto no es así, las estimas de parentesco corregidas pueden resultar muy sesgadas y fuera del espacio paramétrico (negativas). DE CARA et al. (2011) evaluaron dos de estas correcciones (LYNCH & RITLAND, 1999; OLIEHOEK et al., 2006) en el contexto de la conservación de la diversidad genética y encontraron que la utilización del parentesco corregido en la metodología de las contribuciones óptimas lleva a una heterocigosidad (esperada y observada) menor que la utilización del parentesco sin corregir (f_{SNP}). Además, DE CARA et al. (2013b) demostraron que la optimización de contribuciones utilizando f_{SNP} lleva a una diversidad mayor que aquella utilizando f_{SEG}. En cualquier caso, las correlaciones entre f_{SNP} y f_{PED} (TORO et al., 2002; LEROY et al., 2009; SAURA et al., 2013, CROS et al., 2014) y entre f_{SNP} y f_{SEG} (RODRÍGUEZ-

RAMILO et al., 2015) son altas, siempre que el número de marcadores sea suficientemente alto.

Las comparaciones llevadas a cabo entre las distintas medidas de parentesco (f_{PED}, f_{SNP} y f_{SEG}) provienen de simulaciones estocásticas donde se supuso un escenario neutral en el que los loci no están afectados por selección. Así pues, bajo esta suposición, no existe variabilidad en eficacia biológica. Sin embargo, bajo un escenario más realista, donde existe una acumulación de alelos deletéreos, las estrategias de manejo que llevan a una mayor diversidad pueden también llevar a una menor eficacia biológica de la población puesto que la presión de selección sobre los alelos deletéreos está disminuida (DE CARA et al., 2013a; 2013b). Recientemente, DE CARA et al. (2013a) han evaluado el uso del parentesco calculado a partir de marcadores SNP en un escenario de selección, suponiendo dos modelos mutacionales muy diferentes: el modelo clásico de Mukai (MUKAI et al., 1972) que supone que existen muchas mutaciones deletéreas de efecto pequeño y un modelo alternativo que supone pocas mutaciones de efecto grande (CABALLERO & KEIGHTLEY, 1994; GARCÍA-DORADO & CABALLERO, 2000). Bajo este último modelo, la optimización de contribuciones utilizando f_{SNP} no llevó consigo una pérdida de eficacia biológica porque al ser el efecto de las mutaciones suficientemente grande, éstas pueden ser purgadas por la selección. Sin embargo, bajo el modelo de Mukai, la mayor diversidad obtenida a partir de la optimización que utiliza f_{SNP} frente a la que utiliza f_{PED} fue acompañada por una pérdida mayor de la eficacia biológica. Con el objetivo de conseguir un balance entre la diversidad y la eficacia biológica mantenidas en la población se han propuesto diferentes estrategias (DE CARA et al., 2013a; 2013b). De todas ellas, la estrategia que proporciona un mejor balance es utilizar f_{SEG} pero calculada sólo con segmentos largos. La razón de ello la constituiría el hecho de que

la consanguinidad reciente (relacionada con segmentos largos) es más perjudicial que la consanguinidad ancestral (relacionada con segmentos cortos) puesto que la depresión consanguínea causada por esta última ha podido ser purgada. Esto se ha podido comprobar de manera teórica (GARCÍA-DORADO et al., 2008; 2012) y también en poblaciones reales, por ejemplo para el tamaño de camada en ratones (HOLT et al., 2005; HINRICHS et al., 2007) y para producción de leche en ganado vacuno (PRYCE et al., 2014). Utilizando sólo los segmentos largos en el manejo de las poblaciones, se podrían evitar ROH largos en la descendencia y, por lo tanto, la expresión de la depresión consanguínea.

Para especies domésticas de interés económico (aves, ganado vacuno, ovino, porcino y caballar así como en salmón) existen ya paneles densos de SNPs (decenas o cientos de miles de SNPs) a nivel comercial. Este no es el caso para especies para las que el beneficio económico del desarrollo de estos paneles no está tan claro. Sin embargo, es esperable que conforme la tecnología sea cada vez más barata, estos paneles densos de SNPs serán desarrollados para especies no comerciales pero de interés en cuanto a la conservación de los recursos genéticos. Por ello, es necesario determinar de antemano la densidad de SNPs requerida para poder realizar el manejo de las poblaciones de estas especies de forma eficiente con el objetivo de mantener su diversidad. PRYCE et al. (2012) compararon la eficacia de dos chips de diferente densidad (aproximadamente 3.000 y 50.000 SNPs) para controlar el incremento de consanguinidad, utilizando estrategias de apareamientos, en una muestra de toros Holstein. El rendimiento de ambos chips fue muy parecido, sugiriendo la posibilidad de que para este objetivo en concreto sería más razonable el genotipado de animales con el chip de menor densidad. Esto no es sorprendente en esta raza, que, a pesar de ser numéricamente muy grande (millones de animales repartidos por todo el mundo)

muestra un N_e relativamente bajo (del orden de 100 animales) tanto cuando éste se ha estimado a partir de datos genealógicos (BROTHERSTONE & GODDARD, 2005; HILL, 2010) como cuando se ha estimado a partir de datos genómicos (RODRÍGUEZ-RAMILO et al., 2015). En el Capítulo 1 de esta tesis, se ha demostrado que la densidad de SNPs necesaria para mantener diversidad genética también es relativamente baja. De hecho, por encima de 500 SNPs/M, el aumento en la densidad de marcadores tiene que ser muy grande para que se observe un aumento apreciable de la heterocigosidad mantenida. Este rendimiento decreciente observado al aumentar la densidad de los chips de SNPs también se ha observado en el contexto de las evaluaciones genómicas en cuanto a la precisión de las estimas de los valores mejorantes (SOLBERG et al., 2008). Sin embargo, en este caso la densidad de SNPs necesaria para obtener un beneficio por utilizar los marcadores en lugar de las genealogías es mayor que para mantener diversidad. SOLBERG et al. (2008) observaron que la precisión de los valores mejorantes seguía aumentando con una densidad de 800 SNPs/M (la máxima densidad considerada en el estudio).

La optimización de las contribuciones en esta tesis ha sido llevada a cabo para minimizar la tasa de parentesco y por tanto, para maximizar EH. Una estrategia alternativa sería optimizar las contribuciones para maximizar AD, que representa otra medida de diversidad genética. Es conocido que la respuesta a largo plazo depende fundamentalmente de la riqueza alélica inicial (el número de alelos por locus que segregan en la población al principio del proceso selectivo) (JAMES, 1971; HILL & RABASH, 1986) y por lo tanto, AD determina el potencial de adaptación a largo plazo de una población. Hay que hacer notar que FERNÁNDEZ et al. (2004) comprobaron, simulando marcadores de tipo microsatélite, que las estrategias que maximizan EH también mantienen niveles de AD tan altos como las estrategias que maximizan

especificamente AD. En el mismo sentido, en este estudio se muestra que el empleo de SNPs para mantener EH también mantiene altos niveles de AD. Por otra parte, un estudio reciente, realizado a través de un experimento con *Drosophila melanogaster* y simulaciones por ordenador, indica que poblaciones sintéticas obtenidas maximizando AD muestran una respuesta a la selección a corto plazo igual o mayor que aquellas obtenidas maximizando EH (VILAS, 2014). Sin embargo, este resultado puede estar condicionado por el número de loci controlando el carácter seleccionado y el número de marcadores usados para calcular la diversidad. Es necesario por tanto explorar con más profundidad si se debe priorizar AD o EH en los programas de conservación. En el caso concreto de los SNPs, que son marcadores bialélicos, esta medida de variación alélica, definida como el número total de alelos, puede ser cuestionada. Una alternativa, que merece ser investigada sería calcular los parentescos moleculares y la propia diversidad alélica a partir de haplotipos de un número variable de SNPs para aumentar el grado de polimorfismo (PÉREZ-FIGUEROA et al., 2012). La forma óptima de construir haplotipos de SNPs es, sin embargo, desconocida.

Además de la mayor cantidad de diversidad mantenida que se consigue utilizando medidas de parentesco molecular en la gestión de poblaciones, el uso de estas medidas aporta una ventaja adicional cuando se compara con el parentesco genealógico y ésta radica en que el número de individuos óptimo para producir la siguiente generación es menor tal y como se demuestra en el Capítulo 2. Cuando no se dispone de información (genealógica o molecular) previa sobre los individuos de la población de partida, se asume, tal y como ocurre en nuestras simulaciones, que los individuos no son consanguíneos ni están emparentados. En este escenario, la solución óptima cuando se utiliza el parentesco genealógico en la gestión de la población es la igualación de contribuciones (FERNÁNDEZ et al., 2003). Esta solución

sigue siendo la óptima en generaciones posteriores porque desde el punto de vista genealógico la población es homogénea (todos los individuos tienen el mismo parentesco promedio con los demás individuos de la población). Sin embargo, cuando se dispone de información molecular en la población de partida, las relaciones entre individuos se pueden estimar a partir de los genotipos de los marcadores, por lo que igualar contribuciones no es la mejor opción cuando se utiliza el parentesco molecular en la gestión. El uso de información molecular aporta entonces una ventaja económica, por ejemplo en programas de conservación *in vivo*, puesto que se consigue mantener al menos la misma cantidad de diversidad pero manteniendo menos individuos, lo que implica un ahorro en gastos de alimentación, sanidad e instalaciones.

A lo largo de esta tesis, una vez optimizadas las contribuciones de los posibles reproductores según el criterio particular usado en cada caso, los apareamientos se han realizado aleatoriamente. Sin embargo, se podrían plantear esquemas de apareamientos alternativos. *A priori* (al menos para una generación), la cantidad de diversidad genética transmitida no depende del tipo de apareamiento, pero este último si afecta los niveles de consanguinidad generados. El apareamiento entre individuos emparentados produce a corto plazo un rápido incremento en el coeficiente de consanguinidad por lo que, en la gestión de poblaciones animales, la idea general ha sido evitar en la medida de lo posible los apareamientos consanguíneos. Si bien es cierto que, a largo plazo, la tasa de consanguinidad es menor si se fuerza el apareamiento de parientes (WOOLIAMS & BIJMA 2000), la amenaza para la supervivencia de la población debida a la depresión consanguínea que se observa a corto plazo, hace que su uso no sea recomendable en la práctica.

Existe una gran cantidad de literatura científica sobre diferentes sistemas de apareamientos no aleatorios. La mayoría de ellos se propusieron en el contexto de los programas de selección con el objetivo de controlar el incremento de consanguinidad sin que esto llevara consigo una reducción importante de la ganancia genética (CABALLERO et al., 1996). Entre las estrategias propuestas destacan el diseño factorial (WOOLLIAMS, 1989), apareamientos de mínimo parentesco (TORO et al., 1988) y apareamientos compensatorios (CABALLERO et al., 1996). En un diseño factorial cada reproductor se aparea con más de un individuo del otro sexo. En el caso extremo (factorial completo), cada macho se aparea con todas las hembras y cada hembra con todos los machos. Para un número fijo de familias creadas y en comparación con el diseño jerárquico, los apareamientos factoriales crean un mayor número de familias de medios hermanos y un menor número de familias de hermanos completos, llevando así consigo una menor consanguinidad. El apareamiento compensatorio consiste en ordenar los individuos seleccionados de cada sexo en base a su parentesco promedio con el resto de los individuos de la población, apareando el macho más emparentado con la hembra menos emparentada y así sucesivamente. Este sistema tiene efecto en programas de mejora al mezclar individuos de las familias muy representadas debido a su elevado valor mejorante con individuos de familias menos representadas para paliar así el efecto de la selección (aumento del parentesco entre los individuos seleccionados). En programas de conservación este sistema no tiene efecto, ya que no existe selección artificial que favorezca el uso de animales muy emparentados para mejorar un determinado carácter. Por su parte, en el sistema de apareamientos de mínimo parentesco, tal y como como su nombre indica, se minimiza el parentesco promedio de los animales apareados. Así se garantiza la mínima consanguinidad promedio en la descendencia. SONESSON & MEUWISSEN (2000)

demostraron que en programas de mejora este sistema consigue, para la misma tasa de consanguinidad, una mayor respuesta genética que los apareamientos aleatorios. FERNÁNDEZ & CABALLERO (2001) también recomendaron el sistema de apareamientos de mínimo parentesco en el ámbito de la conservación, aunque las diferencias entre los distintos sistemas comparados fueron muy pequeñas. En los estudios que componen esta tesis, el énfasis se ha puesto en el mantenimiento de la diversidad más que en los niveles de consanguinidad. Es por ello que esperaríamos que, en consonancia con lo que ocurrió en los estudios anteriormente citados, las conclusiones generales se mantendrían con esquemas de apareamiento no aleatorio. En cualquier caso, DE CARA et al. (2013b) comprobaron que, una vez que las contribuciones óptimas han sido determinadas, la influencia del sistema de apareamiento, tanto en la diversidad genética mantenida como en los niveles de consanguinidad y eficacia biológica, es muy reducida. Existe además un problema con la implementación estricta de estos sistemas y es que podría ser difícil respetar las contribuciones óptimas obtenidas a partir de la minimización del parentesco promedio, dadas las restricciones fisiológicas de la mayoría de las especies que se gestionan. Una solución a este problema podría ser realizar todo el proceso (contribuciones y apareamientos) en un solo paso mediante la metodología denominada "Mate Selection" en la que se optimiza el número de hijos a obtener de cada uno de los apareamientos posibles (ALLAIRE, 1980; SMITH & ALLAIRE, 1985).

El desarrollo reciente de métodos de secuenciación de nueva generación (NGS, del inglés Next-Generation Sequencing) ha permitido obtener información mucho más precisa sobre el genoma que la obtenida a partir de los chips de SNPs. Mientras que estos últimos sólo incluyen las variantes más comunes, las secuencias incluyen también las variantes raras (aquellas que aparecen en una frecuencia menor

del 1% en la población), unas variantes que albergan una gran parte de la variabilidad genética de las poblaciones, así como todas las mutaciones causales (MACLEOD et al., 2014). Los métodos NGS (METZKER, 2010), desarrollados en un primer momento para su uso en humanos, se han comenzado ya a aplicar también en especies domésticas tales como aves (INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORCIUM, 2004) y ganado vacuno (ELSIK et al., 2009) y porcino (GROENEN et al., 2012) así como en especies salvajes (NARUM et al., 2013). A partir de NGS se puede incrementar rápidamente la cantidad de SNPs disponibles para una especie. Sin embargo, según se ha observado en el presente estudio, el aumento de la eficiencia en el mantenimiento de diversidad es cada vez menor cuando se alcanzan densidades altas de SNPs. En vista de estos resultados, podría ocurrir que el incremento de la densidad en especies para las que ya se han desarrollado chips densos de SNPs no se tradujese en una mejora para esta finalidad. En el mismo sentido, un estudio reciente realizado en el contexto de la mejora genética indica que no se observa un aumento apreciable en la precisión de los valores mejorantes genómicos cuando se utilizan secuencias en lugar de chips de SNPs comerciales en poblaciones de tamaño efectivo reducido (MACLEOD et al., 2014). Ello puede ser debido a que el LD ya es lo suficientemente alto entre SNPs y el resto de loci del genoma como para que la disponibilidad de secuencias mejore el rendimiento. En cualquier caso, el uso de secuencias permite detectar variantes estructurales (inversiones, inserciones, deleciones, duplicaciones y translocaciones) y reordenamientos de segmentos dentro de cromosomas, por lo que pueden ser una alternativa de gran utilidad en estudios relacionados con la caracterización y conservación de diversidad genética (ALLENDORF et al., 2010; NARUM et al., 2013), así como para detectar con mayor precisión regiones de homocigosidad (BOSSE et al., 2012). Una ventaja importante de las secuencias es que

permite evaluar directamente los polimorfismos presentes en regiones codificantes y que pueden estar relacionadas con procesos selectivos.

Hemos visto que la creciente disponibilidad de información molecular debida al desarrollo de la genómica en todo tipo de especies ha servido para desarrollar nuevas y potentes herramientas que tienen utilidad para el mantenimiento de la diversidad genética. Aunque todavía quedan aspectos de la implementación de las herramientas genómicas que deben ser exploradas con mayor profundidad antes de que puedan ser usadas de manera habitual en programas de conservación (por ejemplo, el control preciso de las restricciones a la tasa de parentesco deseado), en esta tesis hemos comprobado que la información molecular proveniente del genotipado masivo de SNPs puede reemplazar a la información genealógica y a los marcadores moleculares desarrollados previamente, en la optimización de contribuciones para minimizar la pérdida de la variabilidad genética.

CONCLUSIONES

1. Cuando se utiliza el parentesco molecular para optimizar las contribuciones, una densidad de $3N_e$ SNPs/Morgan es suficiente para mantener al menos la misma heterocigosidad esperada que la mantenida utilizando información genealógica. Esta densidad es inferior a la de los chips de SNPs ya desarrollados para especies de animales de granja y se espera que pronto se consiga para especies salvajes.

2. En general, para el desarrollo de nuevos paneles de SNPs con el objetivo de utilizarlos en el mantenimiento de la diversidad genética, se debería considerar adecuada una densidad cercana a los 500 SNPs/Morgan.

3. El uso del parentesco molecular en la optimización de contribuciones, conseguida a partir de programación semidefinida, permite focalizar el mantenimiento de la diversidad en regiones específicas del genoma. El método permite además incluir restricciones sobre la pérdida de variabilidad genética en otras regiones del genoma.

4. Existe la necesidad de refinar la teoría de la optimización de contribuciones cuando se utilizan matrices genómicas para incluir adecuadamente las restricciones sobre la tasa de parentesco en el modelo.

5. El parentesco calculado a partir de segmentos IBD es una herramienta útil para el mantenimiento de diversidad genética. La eficacia en el mantenimiento de la diversidad genética mediante la optimización de las contribuciones que utiliza este tipo de parentesco es muy similar cuando las fases gaméticas son conocidas sin error que cuando son éstas son inferidas utilizando los algoritmos actualmente disponibles.

CONCLUSIONS:

1. When using molecular coancestry in the optimisation of contributions, a density of $3N_e$ SNPs/Morgan is enough for maintaining at least the same expected heterozygosity than that maintained when using genealogical information. This marker density is lower than the density of most SNP chips already available for farm animals and it is expected to be achieved for wild species in the near future.

2. In general, when developing SNP chips for new species with the aim of maintaining genetic diversity, a marker density of around 500 SNPs/Morgan could be considered as adequate.

3. The use of molecular coancestry in the optimisation of contributions achieved through semidefinite programming, allows us to focus on maintaining diversity in specific regions of the genome. The method also allows to include restrictions on the loss of diversity in other genomic regions.

4. There is a need of refining the theory of genetic contributions when genomic matrices are used in order to adequately include restrictions on the rate of coancestry in the model.

5. The coancestry coefficient computed from IBD segments is a useful tool for the maintenance of genetic diversity. The efficiency in maintaining genetic diversity through the optimisation of contributions that uses this type of coancestry is very similar when true gametic phases are known without error or when they are inferred.

BIBLIOGRAFÍA

ALBRECHTSEN A, SAND KORNELIUSSEN T, MOLTKE I, VAN OVERSEEM HANSEN T, NIELSEN FC & NIELSEN R (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology* **33**:266–274.

ALLAIRE FR (1980). Mate selection by selection index theory. *Theoretical and Applied Genetics* **57**:267–272.

ALLENDORF FW, HOHENLOHE PA & LUIKART G (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**:697–709.

BENSON SJ & YINYU Y, (2005). DSDP5: Software for Semidefinite programming. *ACM ACM Transactions on Mathematical Software*. Disponible en *http://www.stanford.edu/~yyye/DSDP5-aper.pdf*.

BJELLAND DW, WEIGEL KA, VUKASINOVIC N & NKRUMAH JD (2013). Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *Journal of Dairy Science* **96**: 4697–4706.

BOICHARD D, MAIGNEL L & VERRIER, É (1997). The value of using probabilities of gene origin to measure genetic variability in a population. *Genetics Selection Evolution* **29**:5–23.

BORCHERS B (1999). CSDP, A C library for semidefinite programming. *Optimization Methods and Software* **11**:613−623.

BOSSE, M, MEGENS HJ, MADSEN O, PAUDEL Y, FRANTZ LAF, SCHOOK LB, CROOIJMANS RPMA & GROENEN MAM (2012). Regions of homozygosity in the porcine genome: Consequence of demography and the recombination landscape. *PLOS Genetics* **8**:e1003100.

BROMAN KW & WEBER JL (1999). Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *The American Journal of Human Genetics* **65**:1493–1500.

BROTHERSTONE S & GODDARD M (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**: 1479–1488.

BROWNING SR & BROWNING BL (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *The American Journal of Human Genetics* **81**:1084–1097.

BROWNING SR & BL BROWNING (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**: 703–714.

BROWNING SR & BROWNING BL (2013). Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Human Genetics* **132**:129–138.

CABALLERO A & KEIGHTLEY P (1994). A pleiotropic non additive model of variation in quantitative traits. *Genetics* **138:**883–900.

CABALLERO A, SANTIAGO E & TORO MA (1996). Systems of mating to reduce inbreeding in selected populations. *Animal Science* **62**: 431-442.

CABALLERO A & TORO MA (2000). Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genetical Research (Cambridge)* **75**:331–343.

CAI Z, CAMP NJ, CANNON-ALBRIGHT L & THOMAS A (2011). Identification of regions of positive selection using Shared Genomic Segment analysis. *European Journal of Human Genetics* **19**:667–671.

CAMPBELL CL, PALAMARA PF, DUBROVSKY M, BOTIGUÉ LR, FELLOUS M, ATZMON G, ODDOUX C, PEARLMAN A, HAO L, HENN BM, BURNS E, BUSTAMANTE CD, COMAS D, FRIEDMAN E, PE'ER I & OSTRER H (2012). North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. *Proceedings of the National Academy of Sciences* **109**:13865–13870.

CARVALHEIRO R, DE QUEIROZ SA & KINGHORN B (2010). Optimum contribution selection using differential evolution. *Revista Brasileira de Zootecnia* **39**:1429−1436.

CROS D, SÁNCHEZ L, COCHARD B, SAMPER P, DENIS M, BOUVET JM & FERNÁNDEZ J (2014). Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. *Theoretical and Applied Genetics* **127**:981.

DE CARA MAR, FERNÁNDEZ J, TORO MA & VILLANUEVA B (2011). Using genomic wide information to minimize the loss of diversity in conservation programmes. *Journal of Animal Breeding and Genetics* **128**:456–464.

DE CARA, MAR, VILLANUEVA B, TORO MA & FERNÁNDEZ, J (2013a). Using genomic tools to maintain diversity and fitness in conservation programmes. *Molecular Ecology* **22**: 6091–6099.

DE CARA, MAR, VILLANUEVA B, TORO MA & FERNÁNDEZ J (2013b). Purging deleterious mutations in conservation programmes: combining optimal contributions with inbred matings. *Heredity* **110**: 530–537.

ELSIK CG, TELLAM RL & WORLEY KC (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**:522–528.

EMIK LO & TERRILL CE (1949). Systematic procedures for calculating inbreeding coefficients. *Journal of Heredity* **40**:51–55.

ENGELSMA KA, CALUS MPL, BIJMA P & WINDIG JJ (2010). Estimating genetic diversity across the neutral genome with the use of dense marker maps. *Genetics Selection Evolution* **42**:12.

ENGELSMA KA, VEERKAMP RF, CALUS MPL & WINDIG JJ (2011). Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. *Journal of Animal Breeding and Genetics* **128**:473–481.

FALCONER DS & MACKAY TFC (1996). Introduction to quantitative genetics. 4th Edition. Longman limited.

FERENČAKOVIĆ M, HAMZIC E, GREDLER B, CURIK I & SÖLKNER J (2011). Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. *Agriculturae Conspectus Scientificus (ACS)* **76**:325–329.

FERENČAKOVIĆ M, HAMZIĆ E, GREDLER B, SOLBERG TR, KLEMETSDAL G, CURIK I & SÖLKNER J (2013a). Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. *Jounal of Animal Breeding and Genetics* **130**:286–293.

FERENČAKOVIĆ M, SÖLKNER J & CURIK I (2013b). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genetics Selection Evolution* **45**: 42–51.

FERNÁNDEZ J & TORO MA (1999). The use of mathematical programming to control inbreeding in selection schemes. *Journal of Animal Breeding and Genetics* **116**:447–466.

FERNÁNDEZ J & CABALLERO A (2001). A comparison of management strategies for conservation with regard to population fitness. *Conservation Genetics* **2**:121–131.

FERNÁNDEZ J, TORO MA & CABALLERO A (2003). Fixed contributions designs vs. minimization of global coancestry to control inbreeding in small populations. *Genetics* **165**:885–894.

FERNÁNDEZ J, TORO MA & CABALLERO A (2004). Managing individuals' contributions to maximize the allelic diversity maintained in small, conserved populations. *Conservation Biology* **18**:1358–1367.

FERNÁNDEZ J, VILLANUEVA B, PONG-WONG R & TORO MA (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics* **170**:1313–1321.

FRANKHAM R, BALLOU JD & BRISCOE DA (2002). Introduction to conservation genetics. Cambridge, UK: Cambridge University Press.

FUJISAWA K, KOJIMA M, NAKATA K & YAMASHITA M (2002). SDPA (SemiDefinite Programming Algorithm) user's manual – version 6.00. *Research Reports on Mathematical and Computer Sciences Series B: Operations Research.*

GARANT D & KRUUK LE (2005). How to use molecular marker data to measure evolutionary parameters in wild populations. *Molecular Ecology* **14**:1843–1859.

GARCIA-DORADO A & CABALLERO A (2000). On the average degree of dominance of deleterious spontaneous mutations. *Genetics* **155**:1991–2001.

GARCIA-DORADO A (2008). A simple method to account for natural selection when predicting inbreeding depression. *Genetics* **180**:1559–1566.

GARCIA-DORADO A (2012). Understanding and predicting the fitness decline of shrunk populations: inbreedng, purging, mutation and standard selection. Genetics **190**:1461–1476.

GENOME 10K COMMUNITY OF SCIENTISTS (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *Journal of Heredity* **100**:659−674.

GIBSON J, MORTON NE & COLLINS A (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15**:789−95.

GODDARD ME (2012). Uses of genomics in livestock agriculture. *Animal Production Science* **52**:73−77.

GÓMEZ-ROMANO F, VILLANUEVA B, DE CARA MAR & FERNÁNDEZ J (2013). Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. *Genetics Selection Evolution* **45**:38−45.

GROENEN MA, ARCHIBALD AL, UENISHI H, TUGGLE CK, TAKEUCHI Y, ROTHSCHILD MF et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**:393–398.

GRUNDY B, VILLANUEVA B & WOOLLIAMS JA (1998). Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genetical Research, Cambridge* **72**:159–168.

GRUNDY B, VILLANUEVA B & WOOLLIAMS JA (2000). Dynamic selection for maximizing response with constrained inbreeding in schemes with overlapping generations. *Animal Science* **70**:373–382.

GUSEV A, LOWE JK, STOFFEL M, DALY MJ, ALTSHULER D, BRESLOW JL, FRIEDMAN JM & PE'ER I (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19**:318–326.

GUSEV A, PALAMARA PF, APONTE G, ZHUANG Z, DARVASI A, GREGERSEN P & PE'ER I (2012). The architecture of long-range haplotypes shared within and across populations. *Molecular Biology and Evolution* **29**:473–486.

GUTIÉRREZ JP & GOYACHE F (2005). A note on ENDOG: a computer program for analysing pedigree information. *Journal of Animal Breeding and Genetics* **122**:172–176.

HAN L & ABNEY M (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *European Journal of Human Genetics* **21**:205–211.

HARRIS BL & JOHNSON DL (2010). The impact of high density SNP chips on genomic evaluation in dairy cattle. *Interbull Bulletin* **42**:40–43.

HAYES BJ, BOWMAN PJ, CHAMBERLAIN AJ & GODDARD ME (2009). Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**:433−443.

HERRERO-MEDRANO JM, MEGENS HJ, GROENEN MAM, RAMIS G, BOSSE M, PÉREZ-ENCISO M & CROOIJMANS RPMA (2013). Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genetics* **14**:106–119.

HILL WG & ROBERTSON A (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**:226–231.

HILL WG & RABASH J (1986). Models of long term artifcial selection in finite population. *Genetical Research, Cambridge* **48**:41–50.

HILL WG (2010). Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**:73–85.

HINRICHS D, MEUWISSEN T, ØDEGARD J, HOLT M, VANGEN O & WOOLLIAMS JA (2007). Analysis of inbreeding depression in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding. *Heredity* **99**: 81–88.

HOLT M, MEUWISSEN THE & VANGEN O (2005). The effect of fast created inbreeding on litter size and body weights in mice. *Genetics Selection Evolution* **37**:523–537.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**:695–716.

JAMES JW (1971). Frequency in relatives for an all-or-none trait. *Annals of Human Genetics* **35**:47–49.

KELLER MC, VISSCHER PM & GODDARD ME (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**:237–249.

KIM ES, COLE JB, HUSON H, WIGGANS GR, VAN TASSELL CP, CROOKER BA, LIU G, YANG DA Y & SONSTEGARD TS (2013). Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. *PLOS ONE* **8**:e80813.

KIRIN M, MCQUILLAN R, FRANKLIN CS, CAMPBELL H, MCKEIGUE PM & WILSON JF (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**:e13996.

KU CS, NAIDOO N, TEO SM & PAWITAN Y (2011). Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics*, **129**:1–15.

LEROY G, VERRIER E, MERIAUX JC & ROGNON X (2009). Genetic diversity of dog breeds: within-breed diversity comparing genealogical and molecular data. *Animal Genetics* **40**:323–332.

LUIKART G, ALLENDORF FW, CORNUET JM & SHERWIN WB (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity* **89**:238–247.

LYNCH M & RITLAND K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**:1753–1766.

MACLEOD IM, HAYES BJ & GODDARD ME (2014). The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics* **198**:1671–1684.

MALÉCOT G (1948). Les mathématiques de l'hérédité. Masson et Cie, Paris.

MCQUILLAN R, LEUTENEGGER AL, ABDEL-RAHMAN R, FRANKLIN CS, PERICIC M, BARAC-LAUC L et al. (2008). Runs of homozygosity in European populations. *The American Journal of Human Genetics* **83**:359–372.

METZKER ML (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**:31–46.

MEUWISSEN THE & LUO Z (1992). Computing inbreeding coefficients in large populations. *Genetics Selection Evolution* **24**:305–313.

MEUWISSEN THE (1997). Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science* **75**:934–940.

MEUWISSEN THE & SONESSON AK (1998). Maximizing the response of selection with a predefined rate of inbreeding: overlapping generations. *Journal of Animal Science* **76**:2575–2583.

MUKAI T, CHIGUSA SI, METTLER LE & CROW JF (1972). Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics* **72**:335–355.

NARUM SR, BUERKLE CA, DAVEY JW, MILLER MR & HOHENLOHE PA (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**:2841–2847.

NEJATI-JAVAREMI A, SMITH C & GIBSON JP (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science* **75**:1738–1745.

OLIEHOEK PA, WINDIG JJ, VAN ARENDONK JAM & BIJMA P (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**: 483–496.

OLIEHOEK PA & BIJMA P (2009). Effects of pedigree errors on the efficiency of conservation decisions. *Genetics Selection Evolution* **41**:1–11.

PALAMARA PF, LENCZ T, DARVASI A & PE'ER I (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics* **91**:809–822.

PÉREZ-FIGUEROA A, RODRÍGUEZ-RAMILO ST & CABALLERO A (2012). Analysis and management of gene and allelic diversity in subdivided populations using the software program METAPOP. *Methods in Molecular Biology* **888:**261-75.

PONG-WONG R & WOOLLIAMS JA (2007). Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming. *Genetics Selection Evolution* **39:**3−25.

PURFIELD DC, DONAGH P, BERRY DP, MCPARLAND S & BRADLEY DG (2012). Runs of homozygosity and population history in cattle. *BMC Genetics* **13**:70.

PRICE AL, HELGASON A, THORLEIFSSON G, MCCARROLL SA, KONG A & STEFANSSON K (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics* **7**:e1001317.

PRYCE JA, HAYES BJ & GODDARD, ME (2012). Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *Journal of Dairy Science* **95**:377–388.

PRYCE JE, HAILE-MARIAM M, GODDARD ME & HAYES BJ (2014). Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genetics Selection Evolution* **46**:71.

RALPH P & COOP G (2013). The geography of recent genetic ancestry across Europe. *PLOS Biology* **11**:e1001555.

RODRÍGUEZ-RAMILO ST, FERNÁNDEZ J, TORO MA, HERNÁNDEZ D & VILLANUEVA B (2015). Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. *PLoS ONE* **10**:e0124157.

ROFF DA (1997). Evolutionary quantitative genetics. Chapman and Hall, New York.

ROUGHSEDGE T, PONG-WONG R, WOOLLIAMS JA & VILLANUEVA B (2008). Restricting coancestry and inbreeding at a specific position on the genome by using optimized selection. *Genetical Research, Cambridge* **90**:199–208.

SANTIAGO E & CABALLERO A (1995). Effective size of populations under selection. *Genetics* **139**:1013–1030.

SANTURE AW, STAPLEY J, BALL AD, BIRKHEAD TR, BURKE T & SLATE J (2010). On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology* **19**:1439–1451.

SAURA M, FERNÁNDEZ A, RODRÍGUEZ MC, TORO MA, BARRAGÁN C, FERNÁNDEZ AI & VILLANUEVA B (2013). Genome-wide estimates of coancestry and inbreeding in a closed herd of ancient Iberian pigs. *PLOS ONE* **8**:e78314.

SAURA M, FERNÁNDEZ A, VARONA L, FERNÁNDEZ AI, DE CARA MÁR, BARRAGÁN C & VILLANUEVA B (2015). Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. *Genetics Selection Evolution* **47**:1–9.

SCRAGGS E, ZANELLA R, WOJTOWICZ A, TAYLOR JF, GASKINS CT, REEVES JJ, DE AVILA JM & NEIBERGS HL (2014). Estimation of inbreeding and effective population size of full-blood wagyu cattle registered with the American Wagyu Cattle Association. *Journal of Animal Breeding and Genetics* **131:**3–10.

SILIÓ L, RODRÍGUEZ MC, FERNÁNDEZ A, BARRAGÁN C, BENÍTEZ R, ÓVILO C & FERNÁNDEZ AI (2013). Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. *Journal of Animal Breeding and Genetics* **130**:349–360.

SMITH SP & ALLAIRE FR (1985). Efficient selection rules to increase non-linear merit: application in mate selection. *Genetics Selection Evolution* **17**: 387-406.

SMOUSE PE (2010). How many SNPs are enough? *Molecular Ecology* **19**:1265–1266.

SOLBERG TR, SONESSON AK, WOOLLIAMS JA & MEUWISSEN THE (2008). Genomic selection using different marker types and densities. *Journal of Animal Science* **86**:2447–2454.

SÖLKNER J, FILIPCIC L & HAMPSHIRE N (1998). Genetic variability of populations and similarity of subpopulations in Austria cattle breeds determined by analysis of pedigrees. *Animal Science* **67**:249–256.

SONESSON AK & MEUWISSEN THE (2000). Mating schemes for optimum contribution selection with constrained rates of inbreeding. *Genetics Selection Evolution* **32**:231–248.

SONESSON AK, WOOLLIAMS JA & MEUWISSEN THE (2012). Genomic selection requires genomic control of inbreeding. *Genetics Selection Evolution* **44**:27.

STEPHENS M & DONNELLY P (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73**:1162–1169.

THE 1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**:1061–1073.

TORO MA, NIETO B & SALGADO C (1988). A note on minimization of inbreeding in small-scale selection programmes. *Livestock Production Science* **20**:317–323.

TORO MA & PÉREZ-ENCISO M (1990). Optimization of selection response under restricted inbreeding. *Genetics Selection Evolution* **22**:93–107.

TORO MA, BARRAGÁN C, ÓVILO C, RODRIGAÑEZ J, RODRÍGUEZ C & SILIÓ L (2002). Estimation of coancestry in Iberian pigs using molecular markers. *Conservation Genetics* **3:**309–320.

TORO MA, FERNÁNDEZ J & CABALLERO A (2009). Molecular characterization of breeds and its use in conservation. *Livestock Science* **120**:174–195.

TORO MA, VILLANUEVA B & FERNÁNDEZ J (2014). Genomics applied to management strategies in conservation programmes. *Livestock Science* **166**:48–53

VANDENBERGHE L & BOYD S (1996). Semidefinite programming. *SIAM Review* **38**:49–95.

VANRADEN PM (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**:4414–4423.

VILAS AM (2014). Caracterización y gestión de la diversidad genética en poblaciones estructuradas. Tesis de Doctorado. Universidad de Vigo. Departamento de Bioquímica, Genética e Inmunología.

VILLANUEVA B, WOOLLIAMS JA & SIMM G (1994). Strategies for controlling rates of inbreeding in MOET nucleus schemes for beef cattle. *Genetics Selection Evolution* **26**:517–535.

VILLANUEVA B, PONG-WONG R, WOOLLIAMS JA, AVENDAÑO S (2004). Managing genetic resources in selected and conserved populations. In: Farm animal genetic resources, BSAS Occasional Publication No. 30, pp.113-132 Edited by Simm G, Villanueva B, Sinclair KD, Townsend S, Nottingham University Press, Nottingham, UK.

WANG J (1996). Deviation from Hardy-Weinberg proportions in finite populations. *Genetical Research, Cambridge* **68**:249–257.

WOOLLIAMS JA (1989). Modifications to MOET nucleus breeding schemes to improve rates of genetic progress and decrease rates of inbreeding in dairy cattle. *Animal Production* **49**:1–14.

WOOLLIAMS JA & BIJMA P (2000). Predicting rates of inbreeding in populations undergoing selection. *Genetics* **154**:1851–1864.

WRIGHT S (1922). Coefficients of inbreeding and relationship. *The American Naturalist* **56**:330–338.

WRIGHT S (1938). Size of a population and breeding structure in relation to evolution. *Science* **87**:430–431.

WU SP & BOYD S (2000). DPSOL: A parser/solver for semidefinite programs with matrix structure. In: Advances in linear matrix inequality methods in control. Edited by El Ghaoui L, Niculescu SI, SIAM, pp. 79−91.

ZUK O, HECHTER E, SUNYAEV SR & LANDER ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**:1193-1198.

Lista de abreviaturas y símbolos

Lista de abreviaturas y símbolos empleados en la tesis

Abreviatura	*Significado*
OC	Optimización de contribuciones.
SNP	Polimorfismo de un sólo nucleótido.
N_e	Censo efectivo.
M	Morgan.
IBD	Identidad por descendencia.
IBS	Identidad en estado.
ROH	Tramos de homocigosidad.
EH	Heterocigosidad esperada.
OH	Heterocigosidad observada.
AD	Diversidad alélica.
f	Coeficiente de parentesco.
F	Coeficiente de consanguinidad.
Δf	Tasa de parentesco.
ΔF	Tasa de consanguinidad.
LD	Desequilibrio de ligamiento.
d	Densidad de marcadores.
N	Censo poblacional.
t	Generación.
μ	Tasa de mutación.
r^2	Coeficiente de correlación entre pares de SNPs al cuadrado.
RAN	Estrategia en la que las contribuciones se deciden al azar.
PED	Estrategia de optimización que minimiza la tasa de parentesco genealógico.
MOL_{OVE}	Estrategia de optimización que minimiza la tasa de parentesco molecular en todo el genoma.
MOL_{CHR}	Estrategia de optimización que minimiza la tasa de parentesco molecular en el cromosoma 1.
MOL_{REG}	Estrategia de optimización que minimiza la tasa de parentesco molecular en 10 regiones de 1 cM cada una, localizadas en 10 cromosomas diferentes.
MOL_{OVE_CON}	Estrategia de optimización que minimiza la tasa de parentesco molecular en todo el genoma mientras se restringe la tasa de parentesco en cada una de 10 regiones de 1 cM localizadas en 10 cromosomas diferentes.
MOL_{CHR_CON}	Estrategia de optimización que minimiza la tasa de parentesco molecular en el cromosoma 1 mientras se restringe la tasa de parentesco en el resto del genoma.

MOL_{REG_CON}	Estrategia de optimización que minimiza la tasa de parentesco molecular en 10 regiones de 1 cM cada una, localizadas en 10 cromosomas diferentes mientras se restringe la tasa de parentesco en el resto del genoma.
f_p	Parentesco genealógico.
f_{m_ove}	Parentesco molecular en todo el genoma.
f_{m_chr}	Parentesco molecular calculado en el cromosoma 1.
f_{m_reg}	Parentesco molecular promedio calculado en 10 regiones de 1 cM cada una, localizadas en 10 cromosomas diferentes.
$f_{m\_ove\text{-}chr}$	Parentesco molecular calculado en todos los loci excepto en aquellos localizados en el cromosoma 1.
$f_{m\_ove\text{-}reg}$	Parentesco molecular calculado en todo el genoma excepto en aquellos localizados una porción de 1M dividida en 10 regiones de 1 cM localizadas en 10 cromosomas diferentes.
Δf_p	Tasa de parentesco genealógico.
Δf_{m_ove}	Tasa de parentesco molecular en todo el genoma.
Δf_{m_chr}	Tasa de parentesco molecular en el cromosoma 1.
Δf_{m_reg}	Tasa de parentesco molecular calculado en una porción de 1M dividida en 10 regiones de 1 cM localizadas en 10 cromosomas diferentes.
$\Delta f_{m\_ove\text{-}reg}$	Tasa de parentesco molecular calculado en todo el genoma excepto en aquellos localizados una porción de 1M dividida en 10 regiones de 1 cM localizadas en 10 cromosomas diferentes.
$\Delta f_{m\_ove\text{-}chr}$	Tasa de parentesco molecular calculado en todo el genoma excepto en el cromosoma 1.
BS	Brown swiss, raza de vacuno austriaco.
TG	Tyrolean grey, raza de vacuno austriaco.
PI	Pinzgauer, raza de vacuno austriaco.
F_{MOL}	Consanguinidad molecular calculada SNP a SNP.
F_{ROH}	Consanguinidad molecular calculada a partir de tramos de homozigosidad.
F_{PED}	Consanguinidad genealógica.
f_{MOL}	Parentesco molecular calculado SNP a SNP.
f_{SEG}	Parentesco molecular calculado a partir de segmentos IBD.
f_{PED}	Parentesco genealógico.
f_{SEG_E}	Parentesco molecular calculado a partir de segmentos IBD utilizando fases gaméticas inferidas.

f_{SEG_T}	Parentesco molecular calculado a partir de segmentos IBD utilizando fases gaméticas conocidas.
Δf_{SEG_E}	Tasa de parentesco molecular calculada a partir de segmentos IBD utilizando fases gaméticas inferidas.
Δf_{SEG_T}	Tasa de parentesco molecular calculada a partir de segmentos IBD utilizando fases gaméticas conocidas.
SER	Tasa de error en la inferencia de fases gaméticas.
N_{e_f}	Censo efectivo calculado a partir de la tasa de parentesco.
N_{e_F}	Censo efectivo calculado a partir de la tasa de consanguinidad.