Universidad de Valladolid
Campus de Palencia

ESCUELA TÉCNICA SUPERIOR

DE INGENIERÍAS AGRARIAS

Máster en Ingeniería Agronómica

SELECCIÓN DE UN SUBCONJUNTO DE LOCI ALTAMENTE INFORMATIVO PARA LA ASIGNACIÓN DE MUESTRAS DE ORIGEN BOVINO A SUS RAZAS CORRESPONDIENTES

Alumno/a: Fernando Bueno Gutiérrez

Tutor/a: Jesús Ángel Baro de la Fuente

TRABAJO DE FIN DE MASTER

MAYO 2015

☐ Copia para el tutor/a

*Quisiera mostrar mis agradecimientos al **Dr. Jesús Ángel Baro de la Fuente**, tutor de este estudio, por su esencial ayuda, la cercanía de su trato y en general, por su implicación en mi formación.*

*Quisiera agradecerle también al **Dr. Carlos Enrique Carleos Artime**, profesor de la Universidad de Oviedo, por su valiosa contribución en el manejo de R y su interés por el estudio.*

*Por último, mis más profundos agradecimientos a mis **padres** por su incesante apoyo.*

A la memoria de Cristino

Abreviaturas

ADN - ácido desoxirribonucleico (*DNA - desoxyribonucleic acid*)

AFLP- polimorfismos en la longitud de fragmentos amplificados (a*mplified fragment length polymorphism*)

AVIL- Avileña-Negra ibérica

ARCA- Sistema Nacional de Información de Razas Ganaderas

ASTV- Asturiana de los Valles

AUC- área debajo de la curva (a*rea under curve*)

BLUP- mejor predictor lineal insesgado (*best linear unbiased prediction)*

BRUP -Bruna dels Pirineus

BRSW – Brown Swiss

CP(s)-componente(s) principal(es)

CV- validación cruzada (*cross validation*)

DOP- Denominación de Origen Protegida

HWE - equilibrio de Hardy-Weinberg (*Hardy-Weinberg equilibrium*)

EE.UU- Estados Unidos

FAO- organización para la agricultura y la alimentación (*food and agrocultural organization*)

Feagas- Federación Española de Asociaciones de Ganado Selecto

FLCH - Fleckvieh

GRR- Representación gráfica de los errores de relación (*graphical representation of relationship errors*)

GUER – Guernesey

GWAS- estudio de asociación del genoma completo (g*enome-wide association study*)

IA- inseminación artificial

IBD- idéntico por descendencia (i*dentical by descendt*)

IBS- idéntico por estado (i*dentical by state*)

IGP- Indicación Geográfica Protegida

DL- desequilibrio de ligamiento

EL – equilibrio de ligamiento

LG- Libro Genealógico

MAGRAMA- Ministerio de Agricultura, Alimentación y Medio Ambiente

MLR- regresión lineal múltiple (*multiple linear regression*)

MORU- Morucha

Ne- tamaño efectivo de población (*effective population number*)

OLS- mínimos cuadrados ordinarios (*ordinary least squares*)

OR- razón de momios (o*dd ratio)*

PCA- análisis de componentes principales (p*rincipal components analysis*)

PCR- reacción en cadena de la polimerasa (*polymerase chain reaction)*

PCR- regresión de componentes principales (*principal components regression*)

QTL- locus de un carácter cuantitativo (*quantitative trait locus*)

PIRE- Pirenaica

PLS- mínimos cuadrados parciales (p*artial least squares*)

PLS-LDA- mínimos cuadrados parciales para análisis discriminante (p*artial least squares, discriminant analysis*)

PLSR- regresión por mínimos cuadrados parciales *(partial least squares regression)*

RETI -Retinta

RGAL- Rubia gallega

ROC- característica operativa del receptor (r*eceiver operating characteristic*)

SIMM– Simmental

SITRAN- Sistema Integral de Trazabilidad Animal

SNP - polimorfismo de nucleótido único (*single nucleotide polymorphism*)

UE- Unión Europea

VD(´s)- variable dependietne(s)

VE- Varianza Explicada

VI(´s)- variable(s) independiente(s)

VIF- factor de inflación de la varianza (*variance inflation factor*)

Glosario de fórmulas

A: número de componentes

$\mathbf{A^T}_{\mathbf{k\ x\ n}}$**:** matriz transpuesta de A

$\mathbf{b_{pls}}$: coeficientes de regresión PLS

Bpls $_{\mathbf{k\times m}}$: matriz de coeficientes de regresión de PLS. Por la que se multiplica a las variables originales para obtener Y.

$\boldsymbol{\beta_o}$: intercepto en la ecuación de regresión

$\boldsymbol{\beta_1}$: pendiente en la ecuación de regresión

$\mathbf{c_{lda}}$: proyección de los valores x en la dirección w_{lda}

D: matriz diagonal en PCR

E $_{\mathbf{nxp}}$: matriz de errores aleatorios al predecir X

F $_{\mathbf{nxq}}$: matriz de errores aleatorios al predecir Y

$F_{ST:}$ efecto de subpoblaciones con respecto a la población total

J (w): distancia entre las proyecciones medias.

K: número de predictores

m: número de variables respuesta

n: número de observaciones

$\mathbf{N_e}$: tamaño efectivo de la población

$\mathbf{N_f}$: número efectivo de hembras

$\mathbf{N_m}$: número efectivo de machos

$\mathbf{P}_{k \times a}$: matriz de cargas de PLS . Por la que se multiplica para obtener los nuevos valores.

$\mathbf{Q}_{m \times a}$: matriz de cargas para Y. Por la que se multiplica a Y para obtener las nuevas puntuaciones del variable respuesta.

S: matriz de covarianzas en PCA

ST: matriz de dispersión total

SW: matriz de dispersión dentro-de-clases

SB: matriz de dispersión entre-clases

$\mathbf{S}_{i:}$ dispersión dentro de la clase I

$\mathbf{T}_{n\times a}$: matriz de puntuaciones de PLS (nuevos valores)

T: suma de los eigenvalores

$\mathbf{U}_{n\times a}$: matriz de puntuaciones de PLS para la variable respuesta Y. Nuevos valores por nuevos pesos

$\boldsymbol{\mu}_{i:}$ vector medio de la clase i en el espacio x

$\tilde{\mu}$ i : vector medio de la clase i en el espacio y

$\vec{v}$ i: iº eigenvetor extraído

w: vector de coeficientes. Multiplicando su transpuesta por la izquierda a la matriz de predictores, obtenemos el vector de la variable respuesta.

W*: proyección óptima en LDA

$\mathbf{w}_{lda:}$ proyecciones de LDA

$\mathbf{X}$ **ó** $\mathbf{A}_{n\times k}$: matriz de predictores

x_i: observación i, de la que se conocen p predictores

$Y_{n \times m}$: matriz de predictores. Generalmente: $Y_{n \times 1}$

$\hat{Y}$: variable respuesta estimada, en término matricial

$\widehat{Yo}$: variable respuesta estimada, álgebra lineal

Z: matriz de nuevas variables en PCR

λ_1: eigenvalor del iº eigenvector extraído

ÍNDICE

ÍNDICE DE CONTENIDO

Índice de tablas

Índice de figuras

1. RESUMEN

La asignación de productos de origen animal a sus razas de origen ha cobrado, hoy en día, un gran interés con la aparición de las Denominaciones de Calidad y desde 2013 con la aparición de los Sellos de Raza. Como consecuencia de esto, lograr un método que haga de la prueba de asignación algo viable, supondría indudables ventajas.

La simplificación y el abaratamiento de la prueba sería la clave para su aprovechamiento, lo cual, entre otras cosas, reduciría con mucho las opciones de fraude en la industria agroalimentaria. Adicionalmente, dada la creciente preocupación por la biodiversidad, la posibilidad de dar un valor añadido a los productos en base a su origen podría suponer la conservación de los sistemas de explotación tradicionales, y con ellos de las razas autóctonas.

Frente a este objetivo, aparece la dificultad de distinguir entre las razas autóctonas españolas cuya diferenciación se ha dado únicamente a lo largo de los últimos siglos. Desde que empezó a investigarse en materia de trazabilidad varios métodos han fallado a la hora de asignar individuos a sus razas. No obstante, las investigaciones más recientes ya no van dirigidas a asignar bien los individuos a sus razas, sino a lograrlo con el menor número posible de marcadores. Este nuevo enfoque, con vistas a hacer más rentables las pruebas de genotipado, hará que la aparición de los SNPs como nuevos marcadores moleculares que secuencian a nivel de nucleótido, supongan un verdadero avance en estas investigaciones. Por otro lado, a la dificultad de diferenciar entre razas más o menos emparentadas se suma la complicada estructura interna de los datos genómicos.

Al trabajar con datos genómicos, dadas las muy extensas bases de datos y la compleja relación entre las variables, habitualmente son necesarios una serie de análisis genéticos para eliminar aquellos predictores cuya información perdería validez al analizar futuras muestras.

En el presente estudio se pretende identificar los SNPs que tienen un peso mayor en la diferenciación de una serie de razas de bovino consideradas. Para ello se realiza una análisis genético en base a los criterios de Desequilibrio de Ligamiento, Equilibrio de Hardy Weinberg, frecuencia del alelo menor y porcentaje mínimo de genotipado, y posteriormente se lleva a cabo un proceso de selección de los predictores. Para esta última fase se ha considerado una técnica estadística que se ofreció en 2009 como una alternativa para la selección de variables en este tipo de problemas. El análisis estadístico mediante mínimos cuadrados parciales (PLS) permite tratar con la multicolinealidad tan fuerte de los datos genómicos. En este estudio se combina esta técnica con un análisis lineal discriminante (LDA), que permite reorientar los datos de forma que se maximice la separación entre individuos pertenecientes a grupos distintos.

La metodología empleada permite lograr un porcentaje de aciertos superior al 95% para las 11 razas del estudio usando únicamente 132 marcadores SNP. Esto demuestra la eficacia de PLS-LDA como método para lograr una reducción de la

dimensión y solucionar el problema de la asignación. Adicionalmente, en el estudio se identifican algunas regiones cromosómicas en las que podrían estar contenidos los caracteres más asociados con la diferenciación de las razas.

Se presume que serán investigaciones como la del presente estudio las que permitirán que se abarate el precio de las pruebas de origen de los alimentos, contribuyendo así a lograr una industria agroalimentaria de mejor calidad e incentivando la biodiversidad.

2. INTRODUCCIÓN

1. Justificación

1.1. Justificación en el mercado

En la década de los 90 comenzó a cobrar importancia el lugar de origen de los productos alimentarios. El auge de la globalización lleva en Europa a una reapreciación de los productos locales. De manera que en la última década se han empleado diversas estrategias para promover el consumo de estos productos fabricados en el entorno del consumidor (Goodman. 2004), (Murdoch. 2000). Así mismo, se han promovido en Europa las figuras de calidad: IGP (Identificación Geográfica Protegida), DOP (Denominación de Origen Protegido) y EEU (Regulación de la Unión Europea) contempladas en el Reglamento (CEE) Nº 2081/92 con el objetivo de potenciar la diversidad en la producción agraria y proteger al consumidor de productos indeseados y de fraudes o imitaciones (Ilbery & Kneafsey. 2000).

La normativa en identificación, registro y etiquetado se ha visto modificada en los últimos años por la siguiente legislación:

LEGISLACIÓN COMUNITARIA

- **Reglamento (UE) nº 653/2014** del Parlamento Europeo y del Consejo del 15 de mayo de 2014, por el que se modifica el Reglamento (CE) nº 1760/2000 en lo referente a la identificación electrónica de los animales de la especie bovina y al etiquetado de la carne de vacuno.
    - **Reglamento (CE) Nº 1760/2000** del Parlamento Europeo y del Consejo del 17 de julio de 2000, que establece un sistema de identificación y registro de los animales de la especie bovina y relativo al etiquetado de la carne de vacuno y de los productos a base de carne de vacuno y por el que se deroga el Reglamento (CE) Nº 820/97.

- **Reglamento (UE) nº 1151/2012** del Parlamento Europeo y del Consejo del 21 de noviembre de 2012 sobre los regímenes de calidad de los productos agrícolas y alimenticios. Este reglamento se completó en 2014 con los siguientes Reglamentos Delegados y de Ejecución:
    - **Reglamento Delegado (UE) nº 664/2014** de la Comisión del 18 de diciembre de 2013, por el que se completa el Reglamento (UE) nº 1151/2012 del Parlamento Europeo y del Consejo en lo que se refiere al establecimiento de los símbolos de la UE para las denominaciones de origen protegidas, las indicaciones geográficas protegidas, y las especialidades tradicionales garantizadas; así como en lo que atañe a determinadas normas sobre la procedencia y procedimiento, y determinadas disposiciones transitorias adicionales.

- **Reglamento Delegado (UE) nº 665/2014** de la Comisión del 11 de marzo de 2014, que completa el Reglamento (UE) nº 1151/2012 del Parlamento Europeo y del Consejo en lo que atañe a las condiciones de utilización del término de calidad facultativo «producto de montaña».
- **Reglamento de Ejecución (UE) nº 668/2014** de la Comisión del 13 de junio de 2014, que establece las normas de desarrollo del Reglamento (UE) nº 1151/2012 del Parlamento Europeo y del Consejo sobre los regímenes de calidad de los productos agrícolas y alimenticios.

- **Directiva 2000/13/CE** del Parlamento Europeo y del Consejo del 20 de marzo de 2000 relativo a la aproximación de las legislaciones de los Estados miembros en materia de etiquetado, presentación y publicidad de los productos alimenticios.

LEGISLACIÓN NACIONAL

- **Real Decreto 505/2013** del 28 de junio, por el que se regula el uso voluntario del logotipo "Raza Autóctona" en los productos de origen animal, que permite reconocer los productos procedentes de estas razas de ganado en el etiquetado de los mismos.

  Este logotipo que recibe también el nombre de “Sello de Raza” es compatible con las otras figuras de calidad, ya sean IGP (Indicación Geográfica Protegida), DOP (Denominación de Origen Protegida), sello de ganadería ecológica o integrada, pliegos de etiquetado facultativo de carne de vacuno, o marcas de calidad y garantía.

  El uso del logotipo afecta a todo tipo de productos independientemente de que sean frescos o estén transformados, siempre que provengan de animales de razas autóctonas. De manera que el sello puede aparecer en cualquiera de los siguientes productos: carne, leche, huevos, derivados o incluso productos no alimenticios como la lana.

  Lo expuesto en este Real Decreto no exime del cumplimiento de los requisitos y condiciones exigidos por la normativa en materia de propiedad industrial, así como de lo dispuesto en los artículos 29, 30, 42 y 56 del Reglamento (UE) n° 1151/2012 del Parlamento Europeo y del Consejo del 21 de noviembre de 2012 sobre los regímenes de calidad de los productos agrícolas y alimenticios.

  Esta iniciativa forma parte del plan de desarrollo del Programa Nacional de Conservación, Mejora y Fomento de las Razas Ganaderas cuyas prioridades estratégicas son la utilización sostenible y las vías alternativas de rentabilidad para las razas y sus productos, además del desempeño de actividades para la difusión y divulgación.

El Ministerio de Agricultura, Alimentación y Medio Ambiente cede el uso de este logotipo a las asociaciones de criadores de animales de razas autóctonas oficialmente reconocidas; y son éstas las encargadas de verificar la pertenencia de los animales a las razas autóctonas mediante la supervisión del proceso.

Las características del logotipo «Raza Autóctona» son específicas de cada especie y quedan definidas en este Real Decreto.

Existe un logotipo genérico para todas las especies y productos diseñado para su promoción y logotipos específicos para cada especie con fines a su comercialización, que va acompañado del nombre de la raza correspondiente al pie del mismo.

- **Real Decreto 698/2013** del 20 de septiembre, por el que se modifica el Real Decreto 2129/2008, de 26 de diciembre, por el que se establece el programa nacional de conservación, mejora y fomento de las razas ganaderas.

- **Real Decreto 2129/2008** del 26 de diciembre, por el que se establece el Programa nacional de conservación, mejora y fomento de las razas ganaderas, así como las líneas de ayudas previstas específicamente para las razas autóctonas españolas.

Considerando lo anterior, la identificación específica mediante logotipos resulta muy recomendable a la hora de comercializar productos procedentes de animales de razas autóctonas.

Este estudio tiene como objetivo adaptar el consumo de productos de origen animal a estas nuevas exigencias del mercado, aumentando la confianza del consumidor con productos que efectivamente sean seguros y fiables y asegurando una trazabilidad de los mismos.

1.2. Justificación tecnológica

La aparición de la genética como una nueva ciencia que trata de explicar el “funcionamiento” de los seres vivos provocará rápidamente un gran giro, fundamentalmente en la Medicina, pero también en la Agronomía, que buscará en las moléculas de la herencia la respuesta a algunas de las cuestiones que tenía por resolver.

De acuerdo con los fundamentos de la genética cualquier célula de un individuo tiene el mismo contenido genético y además éste permanece prácticamente invariable a lo largo del tiempo. Esto permite que podamos conocer el contenido genético de un

animal analizando simplemente un pelo o cualquier otra muestra celular, con la única condición de que posea el núcleo en buen estado.

Será, sin embargo con la aparición de la genómica como una parte de la genética especializada en el estudio de los genomas, que la producción animal va a encontrar una importante fuente de mejora. El alcance de esta nueva ciencia en el campo de la producción animal es hoy en día muy grande.

Los genomas están constituidos por ADN (ácido desoxirribonucleico), el cual se localiza fundamentalmente en los cromosmas en el caso de las especies animales. El ADN no determina el comportamiento del animal, aunque si buena parte del mismo; pero sobre todo determina buena parte de sus condiciones físicas y productivas.

El fenotipo de un animal determina la totalidad del aspecto del mismo y también su producción. Este fenotipo será lo que los ingenieros agrónomos intenten mejorar y está determinado por dos componentes fundamentalmente: componente genético y componente ambiental. Mediante la genómica se puede llegar a conocer el factor genético o, al menos, una parte del mismo; de manera que, en última instancia, podremos estimar qué proporción de la variación lograda del fenotipo se debe a cada uno de los dos componentes, logrando así una economía de esfuerzos y recursos.

El material genético de los animales se parecerá más conforme más emparentados estén. Por tanto, entre razas muy alejadas las diferencias en términos genéticos serán considerablemente mayores que entre individuos de una misma raza. O yendo más allá, si analizamos dos animales de distintas razas resulta de interés estimar qué proporción de esa variación se debe a la diferenciación entre razas y qué proporción se debe a la diferenciación entre individuos. Por último, esa diferencia será mayor entre individuos no emparentados. Por tanto, la identificación de animales haciendo uso de la genómica también va a tener una enorme utilidad en pruebas de paternidad y en la elaboración de árboles genealógicos.

En genómica se estudia la estructura, función e interrelación entre los diferentes genes y el genoma en su conjunto. En muy poco tiempo se ha pasado de la identificación de nucleótidos de ADN a la secuenciación completa del genoma en diferentes organismos. En los últimos tiempos la investigación ha llevado a la elaboración de mapas genómicos en los que se representa la distancia entre las diferentes secuencias génicas dentro de un cromosoma. Por último, las investigaciones más recientes están enfocadas a la comprensión del funcionamiento celular a nivel de ADN. Esto incluye la regulación de genes y la producción de proteínas.

Uno de los principales desafíos en la biología moderna es la investigación de la base genética que está detrás de las diferencias genéticas, tanto entre especies como dentro de una especie. Hoy se sabe que los fundamentos de esas diferencias se basan en los genes que tienen influencia en los rasgos, así como en los mapas de ligamiento asociados.

Los genes ya conocidos sirven como referencia para la localización de otros

genes, y con carácter especial en las especies zootécnicas se definen los QTLs (q*uantitative trait loci*) que llevan a la identificación de rasgos asociados con la producción animal.

La genómica aplicada a la producción animal tiene multitud de aplicaciones. En la industria cárnica se puede ahora predecir la calidad de la carne recurriendo a una base de datos genómicos. Por ejemplo, los productores de carne pueden comprobar el porcentaje de pureza mediante un análisis de sangre o buscar la presencia de determinados SNPs asociados con la alta calidad de la carne. En la actualidad existen herramientas de búsqueda de genomas de libre acceso que permiten obtener una visualización detallada a nivel de nucleótido (Bozeman. 2014). Estos buscadores proporcionan al usuario una interpretación sencilla de los datos sobre los que se hace la consulta.

En las granjas de selección se utiliza la información genómica para detectar individuos portadores de genes asociados con resistencia a enfermedades, que posteriormente se seleccionan como reproductores para así desarrollar estirpes resistentes a determinadas enfermedades. Además, en el campo de la salud los animales de granja son de gran utilidad en el trasplante de órganos, resultando esta otra causa fundamental para identificar los animales mejor dotados. Los xenotransplantes tienen, sin embargo, algunas limitaciones importantes en la actualidad. Al incorporar tejido orgánico procedente de otra especie se rompe la barrera inmunológica y el portador podría verse afectado por determinados tipos de xenovirus, e incluso podría transmitir estos al medio social. Por otra parte, los xenotransplantes deben ir siempre acompañados de inmunosupresores para evitar que el cuerpo rechace del órgano implantado, y esto también suele plantear problemas.

La genómica como herramienta de identificación de los mejores reproductores ha supuesto, sin embargo, un avance muy importante ya que presenta una serie de ventajas frente a las herramientas más convencionales de selección. Más concretamente, la selección genómica ha resultado ser una alternativa muy sólida frente al BLUP (*best linear unbiased prediction*), que ha sido la herramienta de selección favorita por los mejoradores durante casi dos décadas. El BLUP es un modelo lineal mixto para estimar efectos aleatorios que fue desarrollado en 1950 por Charles Roy Henderson. Este modelo empezó a utilizarse en reproducción animal en 1991 como test de progenie para estimar la calidad de los reproductores (Robinson. 1991). Para implementarlo, se introducen en el modelo datos de parientes y se estima el potencial de los animales como reproductores a partir de caracteres reproductivos y relacionados con la salud. Los defensores de esta nueva técnica aseguran la superioridad de la genómica frente al BLUP fundamentalmente como estimador para los siguientes caracteres:

- Caracteres con baja heredabilidad, para identificar la ligera influencia genética.
- Caracteres que sean difíciles o costosos de medir. Por ejemplo, la eficiencia digestiva.
- Caracteres cuya mejora no podría ser comprobada hasta después del sacrificio,

por ejemplo, las características de la canal.

- Caracteres cuyas características fenotípicas son difíciles de cuantificar (por ejemplo, la textura de la carne).

Un grupo muy considerable de investigadores sostiene, sin embargo, que la genómica no podrá ser nunca un sustituto del BLUP. El valor de capacidad del reproductor que se obtiene con un BLUP, puesto que depende de la información que vamos introduciendo en el modelo, aumenta con la edad del animal a medida que vamos disponiendo de datos de producción tanto propios del animal como de su progenie. Por este motivo, decimos que el BLUP es un buen estimador a partir de que el reproductor ha cumplido una cierta edad. Cuando al cabo de un número suficiente de años disponemos de una larga serie de datos productivos, podemos lograr una capacidad predictiva aproximada del 80%. En los casos en los que, como en bovino de leche, los reproductores tienen un número muy elevado de descendientes (del orden de miles), si esperamos lo suficiente, podemos llegar incluso a tener con un modelo BLUP una capacidad predictiva superior al 99%. La genómica, sin embargo, tiene una capacidad predictiva menor (de entorno al 70%), pero ésta es independiente de los datos productivos o de cualquier otra información de que podamos disponer. De hecho, la capacidad predictiva que logramos con una prueba genómica es la misma para un animal de edad avanzada que para un feto varios meses antes de que nazca. Para realizar la prueba genómica en fetos de unos cinco meses de gestación, se extrae una muestra biológica de la placenta para evitar dañar al feto. Mediante dicha extracción obtendremos información del progenitor, puesto que la placenta pertenece biológicamente al feto y no a la madre.

Disponer de una información algo menos fidedigna, pero con una antelación mucho mayor permite adelantar considerablemente el ritmo de mejora. En la mayoría de los casos suele ser necesario que el reproductor cumpla cuatro años de edad para que la capacidad predictiva obtenida con BLUP supere a la capacidad predictiva de la prueba genómica. Hay que tener en cuenta que cuatro años puede ser mucho tiempo en términos de mejora, especialmente ahora que el ritmo de mejora crece de manera exponencial gracias a la implantación de la IA, el lavado de embriones, la genómica y otros avances en la selección.

De todo lo anterior se podría deducir que quizá la mejor opción para estimar los índices de los reproductores, sea la combinación de ambas técnicas. Considerando en un modelo ambos tipos de información, se puede obtener una estimación más fiable que con la prueba genómica y en un tiempo considerablemente menor que con el modelo BLUP. Son muchos los investigadores que sostienen que en los próximos años los índices de los reproductores se calcularan combinando ambas técnicas.

La genómica tiene, sin embargo, además de la estimación del potencial de los reproductores, otras utilidades en el campo de la producción animal, entre las que se

destacan las siguientes:

- Permite identificar con una antelación mucho mayor a los individuos con características más sobresalientes, lo cual favorece un mayor aprovechamiento de su potencial. De igual manera, la genómica, puede ayudar a detectar posibles futuros reproductores cuyas cualidades posiblemente se habrían visto mermadas al llegar a la fase adulta.
- Sirve para evaluar los programas de mejora que se hayan venido realizando.
- Permite realizar acoplamientos correctivos a nivel individual.
- Es una herramienta eficiente para detectar problemas de consanguinidad y de pérdida de biodiversidad.
- Puede utilizarse para cuantificar el verdadero efecto ambiental conseguido mediante, por ejemplo, cambios en el manejo.
- Permite identificar las mejores razas para un determinado sistema de producción, así como caracterizar especies y razas ya sea por interés productivo o de conservación.

Entre sus múltiples usos, la genómica se va a utilizar como un indicador individual de los animales puesto que presenta múltiples beneficios con respecto a los métodos tradicionales de identificación:

- La información es más exacta puesto que está libre de la influencia ambiental. Por tanto, dará igual en que época del año se identifique al animal.
- Puesto que se trata de otra fuente de información completamente ajena al resto, la identificación genética puede servir para valorar la eficiencia de otros métodos de identificación animal tradicionales tales como los crotales o los bolos ruminales.
- La forma en que viene dada la información permite una valorización oficial de los animales. Además, resulta sencilla la comparación con animales de otros países, al poderse estandarizar los resultados y automatizar las bases de datos.
- Los resultados pueden ser comparados con muestras biológicas de cualquier tipo, siempre que se trate de células con núcleo.
- La identificación genética se puede realizar independientemente de cuál sea la edad o el sexo del animal, pudiéndose incluso realizar antes del parto. Identificar animales sobresalientes a una edad temprana, o mejor aún antes del parto permite aumentar enormemente el ritmo de mejora.
- Permite identificar cualquier subproducto o pieza de origen animal con la misma

seguridad con la que se identificaría al propio animal, independientemente de cuál sea el estado taxonómico y de cuánto tiempo haya transcurrido desde el sacrificio.

- Las muestras son duraderas y resistentes, puesto que, aunque con una cierta pérdida de calidad, el ADN puede aguantar hasta 120º en medios de conservación son apropiados. Esta cualidad ha permitido el análisis forense en muestras quemadas, así como el análisis de muestras biológicas después de ser cocinadas.

- Permite detectar otros datos de interés como pueden ser la predisposición a enfermedades, el grado de diferenciación de las razas o individuos, la consanguinidad, el grado de parentesco, e incluso, permite diferenciar entre secuencias génicas IBS (*identical by state*) y IBD (*identical by descendt*). Decimos que dos secuencias son IBD cuando además de coincidir el alelo, éste proviene de un parental común, mientras que dos muestras son IBS cuando simplemente comparten el mismo alelo. Estudiar los animales según IBD permite detectar las mutaciones recientes. Se entiende por mutaciones recientes las que hayan podido producirse en un espacio de tiempo relativamente breve, que en términos de historia evolutiva de las especies, puede tratarse del tiempo transcurrido en algunos cientos de generaciones. Sin embargo, es este tipo de mutaciones las que en definitiva van a permitir diferenciar entre individuos más o menos sobresalientes dentro de los pertenecientes a una población concreta.

- Al contrario que con una evaluación morfológica del animal. Mediante un análisis genético, se utiliza la información, no sólo de los genes con capacidad de expresión, sino también de los que actúan de manera silenciosa, es decir, los que afectan a la producción pero no suponen ninguna alteración física.

- Permite distinguir entre alelos dominantes y recesivos, lo cual resulta esencial cuando nuestro interés consiste en incrementar determinadas frecuencias alélicas.

La identificación genética va a permitir, por un lado, disponer de una información mucho más completa y útil del animal (Williams., *et al*. 2009) y por otro, va a acoger un nuevo uso más simple, que es el de la asignación de animales a las correspondientes razas, mediante una interpretación más sencilla de los chips de genotipado. Es en este tipo de información donde se va a intentar reducir en la medida de lo posible el coste del genotipado, haciendo posible que la genómica llegue a un mayor número de animales y que su utilidad pueda ser observada a nivel de raza o de población.

Estos chips de genotipado son superficies sólidas, generalmente de vidrio o plástico, a las que se incorpora una colección de fragmentos de ADN que se utiliza para analizar y monitorizar la expresión diferencial de genes. El contenido genético se distribuirá en unos pocillos cuya posición es conocida. Para su lectura se mide el nivel

de hibridación existente entre una sonda específica “probe” y una molécula diana “target”, y después se marca con la correspondiente intensidad de fluorescente para poder distinguir y percibir ese nivel de expresión. De manera que se consigue leer determinadas porciones del genoma en dos tipos de plataformas fundamentalmente: microsatélites y SNP (*single nucleotide polymorphism*), como veremos más adelante.

El tamaño medio de estos chips está entre los 500-1000 kb para el escaneado de todo el genoma. Para este estudio se utiliza BeadChip de 700 kb. El diseño de estos chips de alta densidad estará adaptado a las diferentes especies y más particularmente a las razas o variedades características del entorno ganadero para el que se diseñara el chip. En el caso del BeadChip utilizado se diseñó para razas de Estados Unidos. Si consideramos las diferencias genéticas que pueda haber entre dichas razas y las españolas, podremos asumir que haya muchos pocillos cuya información no sea de nuestro interés. Siendo ésta una primera razón para lograr una simplificación del chip. Aunque probablemente la razón principal sea el escaso número de loci cuya información resulta realmente clave para asignar animales a las distintas razas.

En este tipo de investigaciones resulta de enorme utilidad, no tanto una reducción del tamaño del chip conforme se va avanzando en identificación de SNPs clave, sino el seguir utilizando los mismos chips de alta densidad pero limitando el análisis a unas pocas posiciones de interés para cada caso. Surge así un interés por lograr una “reducción de la dimensión” de los chips de genotipado en investigaciones como la del presente estudio. Se entiende por reducción de la dimensión, no a una reducción del tamaño del chip, sino a una disminución del número de variables (pocillos) a considerar. De hecho, dicha reducción es uno de los objetivos principales de este estudio. Concretamente, en este estudio se trata de seleccionar un subconjunto de loci lo más reducido posible que permita la asignación de animales a alguna de las principales razas de España y a cuatro de las razas europeas más características. Mediante la selección de dicho subconjunto se habrá satisfecho el objetivo principal de este trabajo: conseguir que el análisis genómico para la asignación de animales a las razas del estudio sea más barato, rápido y sencillo.

Este estudio será además de utilidad en cuanto a que facilita la identificación de las diferencias genéticas fundamentales entre las principales razas de aptitud cárnica europeas y españolas; así como para contrastar estas diferencias con las diferencias entre individuos, o estimar el grado de diferenciación de estas razas con respecto a las europeas.

Todo ello contribuye, no sólo a conservar el patrimonio genético, sino que así mismo, se presenta como un incentivo para incrementarlo al darse un valor añadido a los animales pertenecientes a las razas autóctonas.

Se mencionan a continuación algunas de las ventajas que puede suponer un abaratamiento en la utilización de estos chips de genotipado:

- Un precio asequible hará más viable la identificación de animales y productos derivados de los mismos.

- Un precio reducido hará posible el genotipado en un mayor número de ganaderías y las bases de datos podrán disponer de información de más y más diversos individuos.

- Los productos en el mercado tendrán una mayor trazabilidad y una mejor calidad.

- El éxito de un genotipado a gran escala facilitará enormemente futuras investigaciones en el campo de la genómica, lo cual muy probablemente se traduzca en otros beneficios como, por ejemplo, incrementos de la productividad.

- Un bajo coste del genotipado se traduce además en una mayor agilidad a la hora de actualizar las bases de datos permitiendo una adaptación en consonancia con el dinamismo de las razas.

1.3. Justificación en el marco ganadero

Este apartado se ha estructurado en tres puntos fundamentales:

a) Favorecimiento de las razas autóctonas

b) Preocupación por un excesivo mestizaje

c) El problema de la consanguinidad en los programas de selección

a) Favorecimiento de las razas autóctonas

La creciente preocupación por la sostenibilidad de los sistemas de explotación, así como el impacto ecológico, el bienestar animal, la mala imagen de algunos sistemas de explotación intensivos y la aparición de algunas enfermedades como el AIV (virus de la gripe aviar) o la encefalopatía espongiforme bovina han acabado llevando a un favorecimiento de los sistemas de explotación más tradicionales y en especial de la explotación de razas autóctonas. Como reacción a esto, la UE y la FAO (*food and agrocultural organization* – Organización de las Naciones Unidas para la alimentación y la agricultura) han recurrido a las razas locales para solucionar estos problemas logrando además una mejora de la calidad de la carne. Más aún, se convierte en una prioridad para la FAO mantener los recursos genéticos teniendo en cuenta que la

variación genética de caracteres puede ser útil para conservar determinadas características productivos en los diferentes entornos.

Estudiar la diversidad genética permite conocer la estructura de poblaciones y los cambios que hayan podido tener lugar a lo largo de su historia y evolución.

b) Preocupación por un excesivo mestizaje

La producción ganadera en España está marcada por un sistema de explotación extensivo con producción de carnes de calidad. Por este motivo, resulta de especial interés concienciar el consumidor de las diferencias de calidad existentes entre unos y otros sistemas de producción; así como apostar por la certificación de los productos de origen animal.

En bovino extensivo el número de razas importadas es considerable y además la variedad de razas autóctonas es muy elevada, por tanto el número de posibles cruces es enorme, como se aprecia en el siguiente gráfico elaborado por SITRAN (Sistema Integral de Trazabilidad Animal del MAGRAMA) (Ministerio de Agricultura, Alimentación y Medio Ambiente), buena parte de la población corresponde hoy a animales cruzados:

FUENTE: SITRAN, Diciembre de 2014

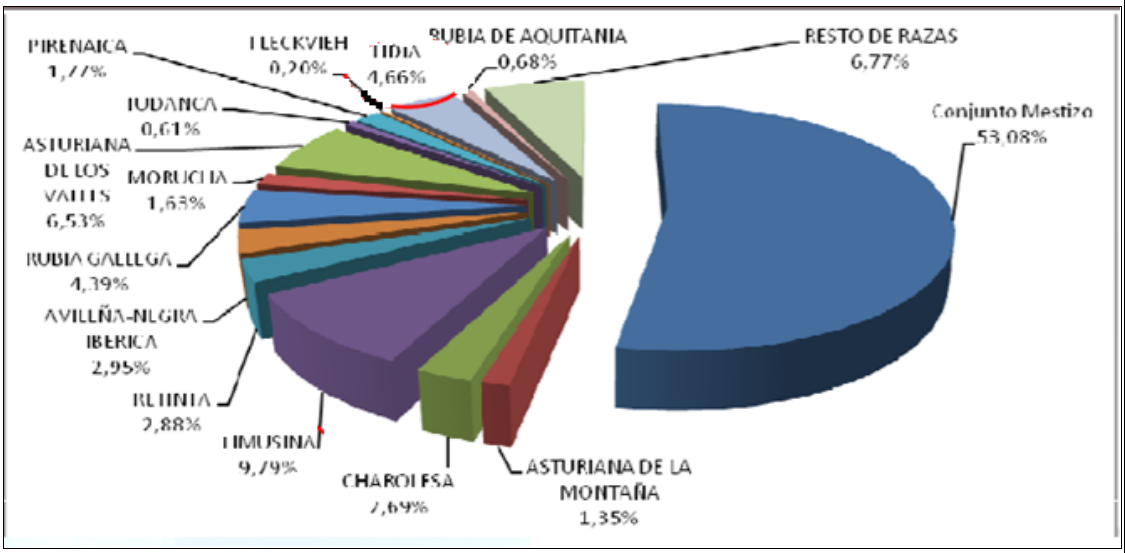


Ilustración 1. Censo de las principales razas.

El conjunto mestizo es, con mucho, el más abundante. Además, según SITRAN esto es así en todas las comunidades autónomas excepto en Asturias (Asturiana de los

Valles: 66,8%) y Navarra (Pirenaica: 52,05%).

El excesivo censo del conjunto mestizo contrasta con el objetivo anteriormente mencionado de apostar por una producción de carne de calidad y hace evidente la necesidad de tener un cierto control de los cruces y de promover la explotación de las razas de fomento. Adicionalmente, como consecuencia de los cruzamientos absorbentes, el número de razas en peligro de extinción en bovino de carne en España es cuanto menos alarmante.

Se expone a continuación una tabla con las 31 razas de bovino de carne que en la actualidad están en peligro de extinción en España según datos del MAGRAMA. Se ha marcado con fondo morado las razas en peligro de extinción que se han considerado en este estudio:

FUENTE: Adaptado de MAGRAMA

Razas en Peligro de Extención	Región principal	Razas en Peligro de Extención	Región principal
Albera	Cordillera Pirenaica	Marismeña	Huelva
Alistana-Sanabresa	Zamora	Menorquina	Menorca
Asturiana de la Montaña	Asturias	Morucha (variedad Negra)	CyL y Cáceres
Avileña-Negra Ibérica (variedad Bociblanca)	Muy distribuida	Monchina	Cantabria y País Vasco
Berrenda en Colorado	Andalucía	Murciana-Levantina	Murcia y Valencia
Berrenda en Negro	Muy distribuida	Negra Andaluza	Andalucia
Betizu	País Vasco	Pajuna	CyL
Blanca Cacereña	Cáceres	Palmera	La Palma
Bruna dels Pirineus	Cataluña	Pasiega	Cantabria y País Vasco
Cachena	Galicia	Sayaguesa	Zamora
Caldelá	Galicia	Serrana Negra	Soria
Canaria	Canarias	Serrana de Teruel	Teruel
Cárdena Andaluza	Andalucia	Terreña	Álava, Vizcaya
Frieiresa	Orense y Zamora	Tudanca	Cantabria
Limiá	Galicia	Vianesa	Ourense
Mallorquina	Mallorca		

Tabla 1: Razas autóctonas españolas en peligro de extinción

Como puede observarse, de las 31 razas que figuran en peligro de extinción, dos de ellas, Avileña-Negra ibérica (variedad Bociblanca) y Morucha (variedad Negra), no son razas como tal, sino que son variedades de dos de las razas que figuran en el catálogo de razas autóctonas españolas. Ambas razas se han incluido en este estudio,

aunque no se ha hecho diferenciación entre variedades.

Hoy en día, dada la creciente preocupación por la biodiversidad, sabemos de la importancia que tiene conservar la pureza de las razas que constituyen el patrimonio genético nacional, pero además sabemos que una cabaña ganadera con un alto censo del conjunto mestizo es, generalmente, una ganadería más envejecida, tal y como se puede ver en la siguiente ilustración:

FUENTE: SITRAN, Diciembre de 2014

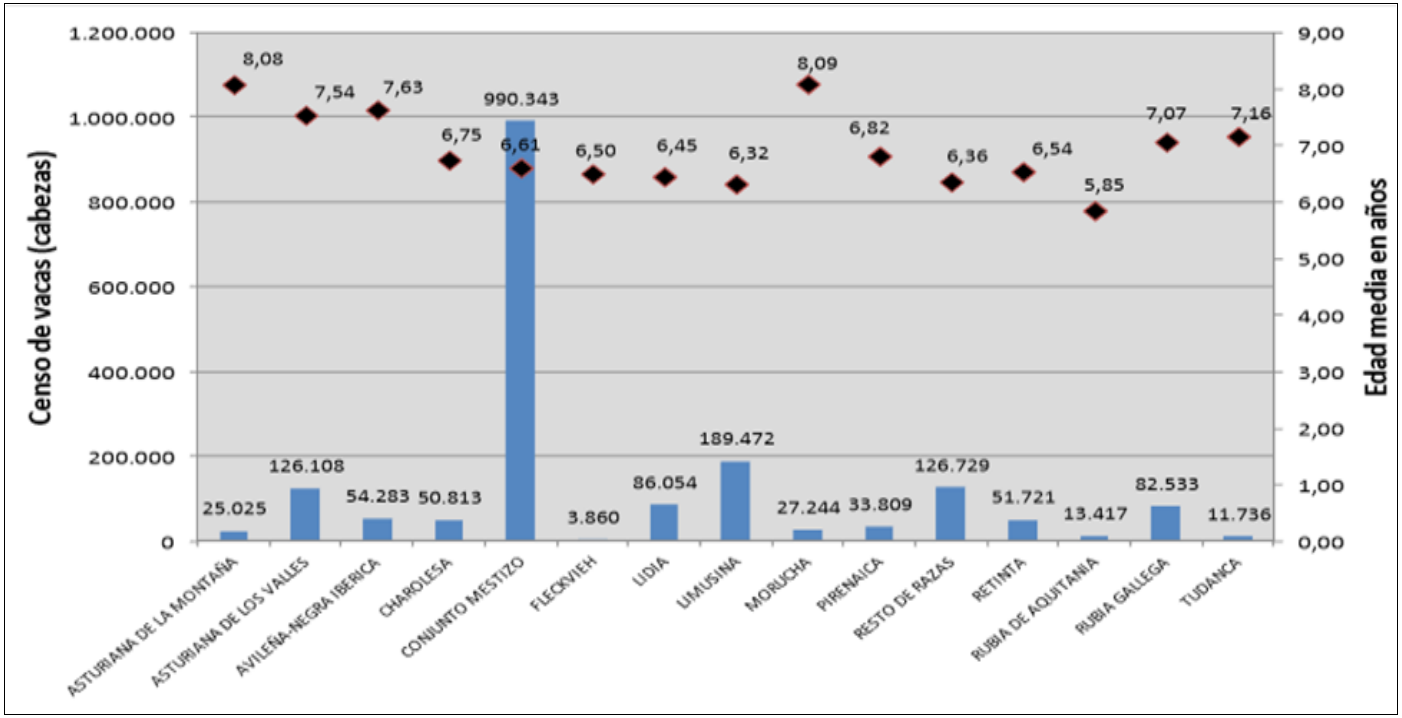


Ilustración 2. Edad media (años) y censo (cabezas)

El envejecimiento de la cabaña ganadera es perjudicial para el sistema de explotación en cuanto a que reduce drásticamente el ritmo de la mejora genética y la eficacia de los programas de selección. Por otro lado, hay que tener en cuenta que cuando el consumidor adquiere una pieza de carne, éste espera que la pieza tenga el sabor y las características esperadas, es decir, que las cualidades sean las mismas en cada una de las muestras que compra con el mismo nombre. Un cruce descontrolado de las diferentes razas haría que esto fuera inviable. Cuando se lleva a cabo un cruzamiento, se aparean dos animales de razas diferentes y la descendencia compartirá características de ambos linajes. De hecho, el descendente tendrá una proporción mayor de heterocigotos con respecto a los progenitores. Aunque los cruzamientos se han utilizado habitualmente para obtener individuos más resistentes a determinadas enfermedades, un exceso de cruzamientos podría resultar contraproducente o incluso conducir a la extinción de las razas puras.

Una distinción entre animales puros y cruzados permitirá una conservación del acervo genético de las razas, así como combatir la introgresión de genes propios de razas foráneas (Ibeagha-Awemu., *et al*. 2004). La introgresión es la transferencia de genes entre diferentes razas o especies que se da como consecuencia de un proceso de hibridación interespecífica seguido de un retrocruzamiento. El retrocruzamiento o "cruzamiento prueba" es el cruzamiento de un híbrido con un individuo perteneciente a la raza de alguno de sus parentales. Aunque la introgresión en sí misma no es ni positiva ni negativa para la producción, lo cierto es que la altera de una manera en cierto modo descontrolada. Además, la introgresión a largo plazo conduce a la desaparición de las razas puras, que constituyen en sí mismas un banco para la biodiversidad.

Se entiende por animales puros aquellos que están registrados en un Libro Genealógico de una raza oficial. La diferenciación entre razas queda justificada por los progresivos procesos de selección o por adaptación a un determinado entorno; si bien, en la práctica el concepto de raza es más conceptual que biológico. Los animales puros tienen la ventaja de que cuentan con un pedigree que sirve como guía para evaluar el valor zootécnico del animal. Sin embargo, éste pasaría a tener una importancia secundaria en caso de que se le realizase una prueba de progenie al animal.

c) El problema de la consanguinidad en los programas de selección

Para entender el problema de la consanguinidad, como un nuevo problema consecuencia de los intesivos programas de selección, debemos antes definir lo que es el N_e (tamaño efectivo de la población). N_e es una medida de diversidad genética dentro de una población (Wright. 1931), que puede entenderse como el número de genomas diferentes que se pueden hacer en la historia de dicha población. Este número es especialmente importante en poblaciones que han sido sometidas a un fuerte proceso de selección.

En una población que haya sido muy seleccionada, como pueda ser, por ejemplo, la cabaña ganadera de raza Holstein en España, que hoy cuenta con cerca de 6 M de individuos, podemos encontrarnos con que el número de posibles genomas diferentes

sea muy inferior al tamaño de la población. De hecho se estima que N_e para la raza Holstein en España es de entorno a 700. Una diferencia tan significativa entre el tamaño de la población y N_e manifiesta una relación muy estrecha de parentesco entre los individuos, lo cual habitualmente se traduce en un aumento de la consanguinidad. En el caso particular de la población de Holstein en España, el valor de N_e da a entender que si analizásemos los datos de todos los animales de todas las generaciones de individuos, únicamente encontraríamos 700 animales con ambos progenitores desconocidos. Denominamos a estos animales "fundadores".

Los problemas de consanguinidad y pérdida de biodiversidad están hoy en día muy latentes en las especies zootécnicas y están íntimamente ligados a Ne. En las razas autóctonas de bovino de carne también se ha observado una disminución de Ne. El sentido común y las investigaciones realizadas, nos llevan a suponer que esto es consecuencia de la progresiva implantación de la IA (inseminación artificial) y los intensos programas de mejora que se han llevado cabo a lo largo de las últimas dos décadas. Si bien, es razonable pensar que mejorar la identificación de los animales es una buena manera de combatir estos problemas, teniendo en cuenta que es muy probable que los programas de mejora vayan a continuar en las próximas décadas.

2. Antecedentes

2.1. Investigaciones en el ámbito del estudio

- Identificación de animales mediante marcadores

Entendemos por marcador un parámetro medible que puede tomar diferentes formas o valores que estimamos pueden servir como indicadores para conocer si nos encontramos o no ante una determinada circunstancia.

Cuando analicemos un marcador, éste podrá tomar diferentes formas y nuestro propósito será encontrar la relación más directa posible entre alguna de sus formas y un determinado fenotipo.

En sus comienzos, la identificación de animales se basaba en marcadores que eran, fundamentalmente, parámetros etnológicos como el color y la forma. Posteriormente se utilizaron polimorfismos proteico-enzimáticos capaces de identificar las diferencias en las proteínas de los individuos, por ejemplo, grupos sanguíneos; después, los avances en inmunología permitieron distinguir en base al tipo y a los niveles de determinados anticuerpos y de reacciones acopladas.

Ya en las últimas décadas, con el desarrollo de la genética, se hizo posible la identificación mediante marcadores que correspondían a secuencias de ADN, tanto para genes que codifican proteínas como para las muy variadas moléculas de ADN que no se

traducen a proteínas.

Se entiende por marcadores moleculares determinadas secuencias de ADN que por su pequeño tamaño y, fundamentalmente, por su localización tienen una gran especificidad, es decir, tienen una alta capacidad para ser exclusivas de un determinado grupo de individuos. Dadas estas características, los marcadores moleculares permiten diferenciar entre razas y entre individuos en base a su genoma.

Para obtener información genética de un individuo y así poder analizar la muestra mediante marcadores moleculares es necesario realizar una prueba genómica que consiste en una secuenciación del ADN del individuo. En los últimos años hemos podido observar un espectacular desarrollo de las tecnologías de secuenciación de nueva generación y hoy se puede secuenciar a un ritmo antes impensable (Bennett. 2004). En la actualidad se pueden hacer genotipados de baja densidad que permiten obtener miles de genotipos de manera simultánea en cerdos (Ramos., *et al*. 2009) y en vacas (Matukumalli., *et al*. 2009).

- Asignación de animales a una población

Los primeros estudios referentes a la asignación de animales a las diferentes poblaciones se llevaron a cabo con microsatélites (Koskinen. 2003). En un principio se trataba de estudios de genética de poblaciones, por ejemplo, para evaluar el intercambio genético entre diferentes poblaciones (Cegelski., *et al*. 2003), cuantificar la migración (Castric & Bernatchez. 2003) y detectar estructuras poblacionales escondidas (Peter., *et al*. 2006).

Sin embargo, este tipo de estudios basados en las pruebas genómicas tardaron poco en extenderse a otros campos, dando lugar a dos líneas de investigación fundamentalmente:

- Identificación individual y tests de paternidad, (Koskinen. 2003), (Liron., *et al*. 2004), (Werner., *et al*. 2004), (Heaton., *et al*. 2005), (Ayres. 2005).

- Trazabilidad de animales y de sus productos, analizando en su caso diferentes generaciones de individuos (Dalvit., *et al*. 2007), (Dalvit., *et al*. 2008), (Negrini., *et al*. 2008b),

De estas, las investigaciones del segundo tipo son aún muy escasas a pesar de que hoy en día son de especial interés dadas las crecientes exigencias en seguridad alimentaria.

- Investigaciones en materia de “reducción de la dimensión”

En las investigaciones anteriormente mencionadas, referentes a asignación de individuos a sus poblaciones va a aparecer un nuevo tipo de problema que es el de la “reducción de la dimensión”. Reducir a la mínima expresión el subconjunto de loci a analizar, pero al mismo tiempo seguir obteniendo resultados aceptables, va a ser uno de los objetivos principales de este tipo de investigaciones.

Para poder reducir la dimensión nos interesará encontrar las variantes (marcadores) más informativas. Llegados a este punto, conviene definir como sería el perfecto marcador. El marcador perfecto sería aquel para el que todos los individuos con un fenotipo determinado, por ejemplo, individuos pertenecientes a una raza, presentasen la misma forma del marcador, y que, a su vez, sólo los individuos de esa raza presentasen esa forma del marcador.

Para valorar la eficacia de un marcador molecular en base a estos criterios, se emplean los conceptos de sensibilidad y especificidad, que son indicadores de probabilidad. Ambos toman valores de 0 a 1, y los valores serán mayores conforme más se acerquen a la estimación perfectamente fiable. En el caso de asignación de individuos a sus razas, diremos que un marcador tiene una sensibilidad de uno si todos los individuos de una raza poseen la misma copia del mismo; mientras que diremos que tiene una especificidad de uno si sólo los individuos de dicha raza poseen dicha copia. Los valores intermedios de sensibilidad corresponden al porcentaje de verdaderos positivos. Definimos como verdaderos positivos a los individuos que, presentando el carácter considerado (pertenecer a la raza, por ejemplo), tienen la copia del marcador que es supuestamente exclusiva. Llamamos a la copia supuestamente exclusiva “copia de referencia”. Mientras que los valores intermedios de especificidad corresponden al porcentaje de verdaderos negativos, que serán los individuos que no presentando el carácter, no muestran la copia de referencia.

En una situación en la que nos interesa reducir el número de marcadores, trataremos de encontrar aquellos que tengan los valores de sensibilidad y especificidad más elevados. Es decir, los que más se aproximen a la situación idílica de sensibilidad y especificidad de uno, tal y como se puede apreciar en la siguiente ilustración:

FUENTE: Andrea. 2009

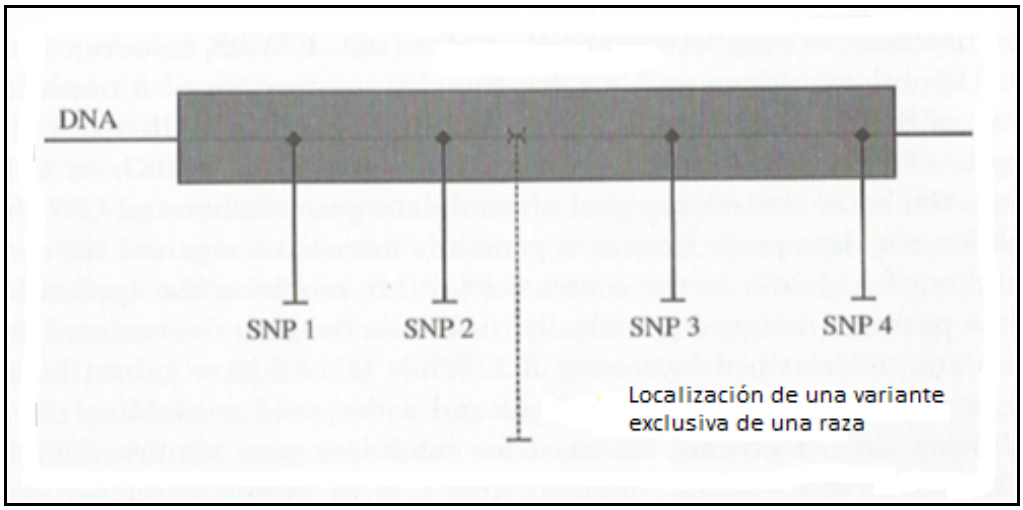


Ilustración 3. Marcador SNP

Para encontrar una variante exclusiva, plenamente informativa, la raza en cuestión debería ser el resultado de un proceso de selección mediante el cual un individuo que hubiese sufrido una mutación diese lugar a una nueva variante. Posteriormente, una progresiva diferenciación permitiría que la variante pudiera ser definida como raza.

Hay que tener en cuenta que este posible origen para una raza es contrario al proceso por el cual en la actualidad se generan la mayoría de razas, que es mediante cruzamiento y migración. Estos dos procesos, ambos basados en la mutación, pueden dar lugar a nuevas razas mediante la mezcla de material genético procedente de diferentes poblaciones de individuos.

- Estudios de caracterización de razas en España

El creciente interés por la caracterización de razas ha venido acompañado de una serie de investigaciones con este propósito en España. La caracterización de las razas españolas resulta particularmente complicada dado que las diferencias genéticas son reducidas; lo cual se debe, fundamentalmente, a que la diferenciación entre las razas se ha producido a lo largo de los últimos siglos. Como consecuencia, hasta el momento han fallado diversas investigaciones que tenían como objetivo diferenciar a los animales según su raza de origen.

En una de las primeras investigaciones al respecto (Martín-Burriel,. *et al*. 1999), se cuantifica la diversidad genética en 6 razas autóctonas españolas, entre las que se incluyen Asturiana de los Valles, Asturiana de las Monatañas, Grupo de Morenas del

Nor-Oeste, Pirenaica, Menorquina y la raza de Lidia. Mediante diferencial de frecuencias alélicas, se calculan las distancias genéticas y se obtiene un dendrograma. El estudio concluye que la diferenciación más acusada se da entre las razas de Lidia y Menorquina. Investigaciones como ésta permiten conocer el origen reciente de las razas autóctonas.

La caracterización genética de razas españolas ha llegado las especies zootécnicas más importantes en nuestro país:

- En porcino se ha caracterizado la raza ibérica (Martínez., *et al*. 2000)
- En caprino, la raza Blanca Andaluza (Martínez., *et al*. 2004)
- En ovino, ocho razas europeas entre las que se incluía la raza española Rubia del Molar (Pariset., *et al.* 2006)
- Incluso se ha estudiado la caracterización genética del camello canario (Schulz., *et al.* 2010).

Las investigaciones también han llegado a razas de perros, (Méndez., *et al*. 2010) logran una caracterización genética del perro de agua cantábrico; y al ganado equino (Tupac-Yupanqui., *et al*. 2010) caracterizan el Caballo Monchino.

Volviendo al bovino de carne, se ha estudiado la variabilidad genética en algunas de las razas en peligro de extinción: Tudanca (Saínz., *et al*. 2011), raza de Lidia (Cortés., *et al*. 2011), raza Pasiega (Celorio., *et al*. 2011), Asturiana de las Montañas (Baro., *et al*. 2012). Además de en algunas razas autóctonas de fomento, como Asturiana de los Valles (Carleos., *et al*. 2009).

También se han llevado a cabo algunas investigaciones en las que se caracterizaban grupos de razas y se estimaba la variabilidad y distancia genética entre las mismas. En (Beja-Pereira., *et al*. 2003) se caracterizan quince razas ibéricas y tres razas francesas haciendo uso de PCA (principal components analysis). Más adelante estudiaremos detalladamente ésta y otras técnicas estadísticas que son útiles en la diferenciación de individuos. En (Martín-Burriel., *et al.* 2007) se realiza un estudio para algunas de las razas autóctonas en peligro de extinción. En (Martín-Burriel., et al. 2011b) se lleva a cabo un estudio filogenético para estimar la diversidad genética global en las razas de la península ibérica. Por último, en (Cañas-Álvarez., *et al*. 2014b) se estudia la variabilidad genética en 5 de las razas de este estudio así como la relación entre las mismas.

Entre las razas del estudio (Cañas-Álvarez., *et al*. 2014b), se encuentran: Asturiana de los Valles, Avileña-Negra Ibérica, Bruna dels Pirineus, Pirenaica y Retinta.

De dicho estudio se pueden extraer las siguientes conclusiones principales:

- La heterocigosidad es de entorno al 0.31, similar a otras razas Europas.

- El grado de diferenciación genética es de pequeño a moderado, con un coeficiente estimado de F_{ST} (efecto de subpoblaciones con respecto a la población total) de entre 0.024 y 0.065.
- La diferenciación entre las razas quedaría de la siguiente manera sobre un dendrograma:

Fuente: Cañas-Álvarez., *et al*. 2014a

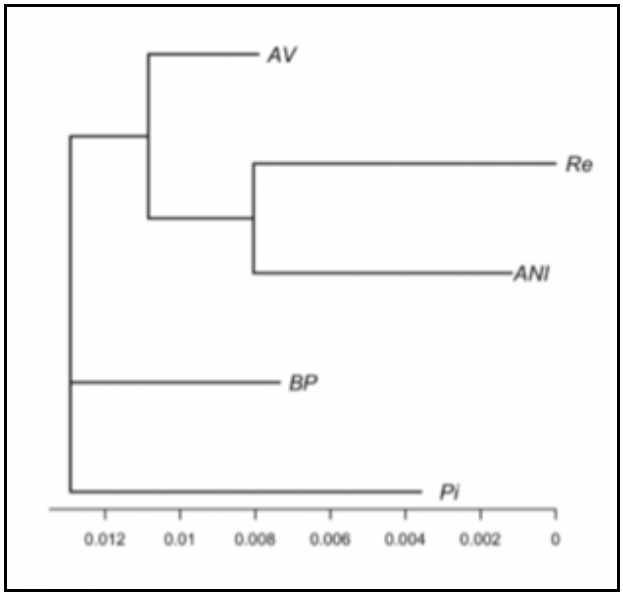


Ilustración 4. Dendrograma basado en la Distancia de Nei.

Razas: Asturiana de los Valles (AV), Avileña – Negra Ibérica (ANI), Bruna dels Pirineus (BP), Pirenaica (Pi) y Retinta (Re)

La distancia entre cada una de las razas representada en el dendrograma corresponde a la distancia de Nei, la cual refleja el número de sustituciones nucleotídicas en el ADN.

2.2. Técnicas estadísticas empleadas

- Probabilísticos y deterministas

Las investigaciones relacionadas con la trazabilidad y la genealogía se han enfocado utilizando tanto criterios probabilísticos como deterministas (Ciampolini., *et al*. 2006), (Dalvit., *et al*. 2008), (Maudet., *et al*. 2002).

Conviene recordar que, en matemáticas son modelos determinísticos aquellos donde donde para determinadas entradas se obtienen invariablemente determinadas salidas y no existe ni incertidumbre ni azar. Los modelos probabilísticos, por el contrario, son aquellos en los que si hay incertidumbre puesto que se consideran ciertos efectos aleatorios, por tanto, a un valor de entrada no le corresponde necesariamente un valor único y explicable en la salida.

En mejora genética animal, los deterministas se basan en la detección de alelos específicos de razas, fundamentalmente genes del color de la capa y su alcance se limita a hacer discriminación entre clusters de razas (Casellas., *et al*. 2004), (Maudet., *et al*. 2002). Por otro lado, las investigaciones probabilísticas usan fundamentalmente microsatélites multialélicos. Aunque en un principio el objetivo era la asignación de individuos a clusters de razas emparentados en mayor o menor medida (Baudouin & Cornuet. 2004) , (Corander., *et al*. 2003), más adelante se aspirará a hacer diferenciaciones más precisas, habitualmente, entre-clases y dentro-de-clases.

- Métodos de frecuencia, bayesianos y de verosimilitud

Las investigaciones encaminadas a asignar animales a sus razas de origen se servirán de criterios como la verosimilitud, frecuencia y métodos bayesianos, mostrando cada uno de ellos ventajas e inconvenientes.

Los primeros estudios se basaban en los métodos de frecuencias y calculaban la probabilidad de que un genotipo perteneciera a una población en base a unos genotipos de pertenencia conocida (Paetkau., *et al*. 1995). Después, en (Rannala., *et al*. 1997) se hace uso de los métodos bayesianos y frecuencias alélicas de las poblaciones para así determinar la significación con la que se puede afirmar que un animal pertenece a una población; y en (Pritchard., *et al*. 2000) se utilizan métodos bayesianos para inferir la estructura de una población.

- Estudios estadísticos con PLS

La herramienta de PLS (*partial least squares*) se ha utilizado para la extracción de variables latentes altamente informativas y data sus orígenes en 1966, cuando Herman

Wold publicó dos procedimientos alternativos para resolver por mínimos cuadrados.

En (Garthwaite. 1994) se resolvió un problema en el que el número de predictores superaba por mucho al número de observaciones. Mediante un método basado en PLS para una única respuesta se seleccionó una serie de predictores.

La eficacia de los coeficientes de regresión de PLS ha sido comparada con VIP (*variable importance in the projection scores*) (Wold. 1995). VIP permite expresar de una manera resumida la contribución que una variable determinada hace al modelo. Si un predictor tiene un coeficiente relativamente bajo (en valor absoluto) y un valor VIP también reducido, entonces se trata de un predictor del que se puede prescindir sin notar apenas ninguna diferencia en el modelo.

Los estudios que utilizan PLS para resolver problemas de predictores, se han extendido ya al campo de la mejora genética animal. En (Moser., et al. 2007) se utiliza PLS para estimar valores medios de cruce evitando tener que recurrir a QTLs o análisis de pedigree.

- Estudios estadísticos con PLS-LDA

LDA es una generalización del discriminante lineal de Fisher que es una técnica que hace uso de la estadística, el reconocimiento de patrones y el aprendizaje de máquinas para encontrar la combinación lineal de caracteres que permita separar dos o más clases de objetos. Los fundamentos de LDA fueron desarrollado por Fisher en 1973, aunque fue a partir de 1979 cuando se empezó a utilizar de una manera más generalizada, a raíz de los trabajos realizados por Dillon. En genética, la mayoría de los estudios que utilizan LDA se han llevado a cabo en el campo de la medicina (Jieping., *et al*. 2004).

PLS-LDA, es la combinación de esta técnica con PLS, y se ha venido usando desde principios de siglo fundamentalmente en el campo de la medicina. Un primer estudio (Barker & Rayens. 2003) utilizó PLS-DA con datos genómicos de cáncer de mama. Después se ha utilizado PLS-LDA en clasificación de tumores, receptor de estrógeno en tumores negativos y positivos, y para comparar situaciones antes y después de tratamientos con quimioterapia. Además, en otro estudio (Pérez-Enciso & Tenenhaus. 2003) se concluye que PLS-LDA es una herramienta poderosa y simple para análisis de datos en genómica y proteómica.

3. Estado del arte

Se comentan a continuación algunas de las investigaciones recientes más relevantes en materia de reducción de la dimensión para la asignación de animales a sus razas.

En un primer estudio en el que se utiliza la genómica como herramienta para asignar individuos a sus razas, se pretende asignar correctamente 396 animales a 16 razas italianas de origen (Negrini., *et al*. 2007). El estudio concluye que son necesarios 141 marcadores AFLP para lograr un porcentaje de aciertos del 93%. Para resolver, se prueba un método bayesiano y otro basado en la máxima verosimilitud. De las pruebas se deduce que los métodos bayesianos son superiores en tres aspectos fundamentales: (i) el porcentaje de aciertos es mayor (93% vs 81%); (ii) en una de las razas (Rmagnola) los porcentajes de aciertos son de 91% vs 45%; (iii) también se obtienen mejores resultados al probar con individuos no usados en el diseño del modelo. Este estudio sirve como aliciente para usar métodos bayesianos frente a métodos de verosimilitud en problemas de asignación. Adicionalmente, el estudio sirve como paradigma en cuanto a que se aporta un número de marcadores que se intentará reducir en futuras investigaciones. Sin embargo, también hay que tener en cuenta que el precio del análisis con marcadores AFLP es muy sustancial, y es por esta razón que la aplicabilidad del método empleado fue limitada.

Unos meses después, en (Negrini., *et al*. 2008b) se estudia la asignación de individuos a 24 razas de bovino europeas analizando los datos en forma de SNPs. En primer lugar, se pretende distinguir el país de origen de la muestra analizada y después se trata de determinar si la muestra pertenece o no a un grupo razas englobadas dentro de una IGP (Indicación Geográfica Protegida). Concretamente, el estudio permitía averiguar si una muestra nueva cumplía los requisitos raciales para pertenecer a alguna de las siguientes IGPs: “Pure Highland Beef”, ‘‘Vitellone dell’Appennino Centrale”, “Ternera de Navarra”, y “Boeuf de Chalosse”. Acorde con los resultados obtenidos en la investigación anterior, en este estudio se emplean métodos bayesianos, y son necesarios únicamente 90 marcadores polimórficos para lograr también un 93% de aciertos. Se puede intuir una mejora considerable con respecto al estudio anterior, puesto que en este caso se trata de marcadores SNP, cuyo precio de análisis es menor, y además, el número de marcadores necesarios se reduce considerablemente. Sin embargo, también hay que tener en cuenta que en este caso no se trata de asignar individuos a sus razas, sino a una IGP, que habitualmente, tienen un mayor margen de variabilidad genética.

En otra investigación con las mismas razas y el mismo panel de 90 SNPs (Negrini., *et al*. 2008a), se compara la eficiencia de un método bayesiano y un método de frecuencias para la selección de los predictores y se deduce que el método bayesiano resulta mejor para la fase de entrenamiento (96% de aciertos vs 85%), mientras que el de frecuencias es más eficiente para la fase de prueba. Además, se concluye que se pueden obtener mejores resultados combinando ambos métodos.

Posteriormente, en un estudio de asignación realizado con cerdos (Ramos., *et al*. 2011), se apuesta por una mayor precisión y se consigue un 99,2% de aciertos con 193 marcadores SNPs. En este estudio se pretendía asignar 151 individuos a 5 de las razas más importantes. Hay que tener en cuenta que el grado de variabilidad entre las diferentes razas va a influir de manera muy directa en el número de marcadores que van a ser necesarios para lograr una diferenciación entre las mismas. En el caso de este estudio las diferencias genéticas entre las razas son bastante sustanciales.

Finalmente, en (Martínez-Camblor., *et al*. 2014) se pretende diferenciar entre diferentes líneas de la raza de Lidia. En este estudio los autores sostienen que los algoritmos computacionales complejos no son la mejor opción para trabajar con las enormes bases de datos genómicas, debido a que se puede generar una pérdida estadística. Por este motivo, se opta por usar una aproximación clásica y se obtiene finalmente una diferenciación con un 99.94% de aciertos usando 3000 microsatélites multialélicos. Es llamativo de este estudio el hecho de que se empleen marcadores micosatélites, pero al mismo tiempo, es perfectamente razonable puesto que como se comenta en el siguiente apartado, los microsatélites son más eficientes cuando se trata de diferenciar entre individuos de una misma raza. Recordamos que en esta investigación, se pretende diferenciar entre diferentes líneas de una misma raza (raza de Lidia). Entre las técnicas empleadas está el método estadístico de regresión logística, diferentes técnicas de búsqueda de datos y aprendizaje de máquina; mientras que será mediante un método de máxima probabilidad que se obtuvo el mejor resultado. Se implementa una regresión logística binomial para cada una de las razas estudiadas, obteniendo para cada individuo una probabilidad de pertenencia a cada raza, y quedando éste asignado a la raza en la que la probabilidad era mayor. Sin embargo, el método empleado en dicho estudio se basaba en la modelación de la segregación de alelos y el coste computacional sería excesivo a la hora de manejar los datos de genotipados masivos de SNPs. A este problema habría que añadir el exceso de parámetros que tendría el modelo.

4. Tipos de marcadores

Para la realización de este tipo de investigaciones, se han utilizado varios tipos de marcadores moleculares: microsatelites (Dalvit., *et al*. 2008), AFLPs (*Amplified Fragment Length Polymorphism*) (Negrini., *et al*. 2007) y SNPs (Heaton., *et al*. 2005), (Negrini., et al. 2008b). En los tres casos se trata de polimorfismos. Un polimorfismo es una variante genética que se da, al menos en un 1% de la población y que nos va a permitir diferenciar entre razas y entre individuos. Los AFLP son polimorfismos amplificados, los microsatélites son polimorfismos de varios nucleótidos, y los SNPs, como su nombre indica, tienen un único nucleótido. A continuación veremos detalladamente cada uno de ellos.

- **Marcadores AFLP**

Los marcadores AFLPs deben su descubrimiento a la técnica de PCR (*polymerase chain reaction*). La PCR es una tecnología muy empleada en biología molecular para amplificar copias de ADN en diferentes órdenes de magnitud, generando de miles a millones de copias de una secuencia particular de ADN. La ventaja fundamental es que al disponer de la muestra amplificada resulta mucho más fácil su identificación y análisis. El procedimiento se basa en las propiedades naturales de las ADN polimerasas.

Mediante la aplicación de ciclos de incrementos de temperatura se incentiva la replicación de las hebras de ADN. Una vez replicadas, se deja reposar la muestra para que vuelvan a unirse antes de volver a replicar.

Para obtener los AFLP se utilizan encimas de restricción para digerir el ADN y posteriormente se fijan unos adaptadores a los extremos adhesivos de esos fragmentos de restricción. Por último, mediante unos cebadores complementarios a la secuencia de los adaptadores se selecciona un subconjunto y se amplifica.

Los marcadores AFLP resultan muy útiles para la detección de la variabilidad genética ya que se analiza la presencia o ausencia de los diferentes sitios de restricción. Sin embargo, fundamentalmente debido a lo laborioso de su obtención, su utilización es cada vez menor.

Los AFLPs son marcadores multialélicos dominantes. Son multialélicos en cuanto a que el número de formas posibles que puede tomar es superior a dos. Los marcadores multialélicos tienen la ventaja de que pueden proporcionar más información al dar una respuesta no binaria. Es decir, en un ejemplo básico de tres posibles alelos y analizando un locus determinado, podremos decir de un individuo, que tiene el alelo A, el B, o el C. Sin embargo, estos marcadores se pueden utilizar también como marcadores bialélicos, simplificando la respuesta al establecer uno de los alelos como copia de referencia. Es decir, para un locus podríamos decir del individuo que tiene el alelo A o que no lo tiene. Por otra parte, los marcadores AFLP son dominantes en cuanto a que no permiten diferenciar entre homocigotos dominantes y heterocigos, por tanto, sólo podremos hacer la estimación de las frecuencias si asumimos HWE (*Hardy-Weinberg equilibrium*). (El concepto de HWE se explica en el apartado de Análisis Genético.)

- **Marcadores Microsatélites**

Se trata de fragmentos de 2 a 6 pares de bases de tamaño que se repiten de manera consecutiva a lo largo del genoma. Cada alelo o copia es una posible combinación de los pares de bases que constituyen los fragmentos, de tal manera que se pueden identificar distintos alelos para cada fragmento. Los fragmentos que se analizan al hacer un genotipado con microsatélites, tienen la peculiaridad de que se encuentran muchas veces repetidos a lo largo del genoma, habitualmente del orden de centenares de veces; por tanto, la principal utilidad de los microsatélites no estriba en cual sea la copia observada en un determinado locus en un individuo, sino en cuantas veces se encuentre repetida esa copia a lo largo del genoma del individuo.

El principal inconveniente de los microsatélites y en especial de su aplicación en este tipo de estudios es que su análisis suele resultar excesivamente caro para la caracterización de razas. En caracterización de razas, especialmente en vacuno de carne, el número de nucleótidos cuya lectura es realmente necesaria para resolver el problema es muy reducido (Ruzzante., *et al*. 2001). Por tanto, lo más habitual será que si usamos

una colección de microsatélites, estemos analizando un número de nucleótidos considerablemente mayor de lo estrictamente necesario. Estaremos "derrochando" buena parte del análisis.

Los microsatélites presentan, sin embargo, algunas ventajas frente a otros marcadores moleculares:

- Son muy variables entre individuos, por lo que tienen enorme utilidad en los tests de paternidad.

- Tienen una gran relevancia en la diferenciación de especies ya que la ISAG (Sociedad Internacional de Genética Animal) cuenta con una larga lista de los microsatélites más informativos.

Los microsatélites, como los AFLP, son marcadores multialélicos; pero en el caso de los microsatélites, se trata de marcadores codominantes, es decir, permiten distinguir entre homocigotos dominantes y heterocigotos. Por lo anterior, la información disponible será mucho mayor. Al hacerse posible esta distinción, no sólo sabremos si la muestra a analizar presenta o no la copia de referencia, sino además cuantas veces la tiene. En el caso de los organismos diploides, puesto que la muestra consiste en pares de cromosomas homólogos, el número máximo de copias posibles en un locus es dos. Por tanto, para un único alelo obtenemos tres posibles respuestas, que son: "no tiene la copia de referencia" (individuo homocigoto), "tiene una copia" (heterocigoto), o "tiene dos veces la copia de referencia" (homocigoto del alelo de referncia). Precisamente por ser multialélicos y además codominantes, los microsatélites son los marcadores moleculares que permiten obtener una mayor información. Sin embargo, a la hora de elegir el marcador a utilizar deben considerarse además otros criterios, como veremos a continuación.

- **Marcadores SNPs**

Estos polimorfismos de un sólo nucleótido tienen la ventaja de que se dan con una frecuencia predecible en una población. Además, puesto que se secuencia a nivel de base nitrogenada, son bialélicos ya que, salvo que haya mutaciones, tendremos, o una de las dos bases pirimidina, o una de las dos bases purina. Por tanto, la información de que disponemos será más simple.

Los SNPs presentan las siguientes ventajas frente a los otros tipos de marcadores moleculares:

- Las tasas de mutación son más pequeñas

- Se logra un genotipado más robusto y una interpretación más fácil de los datos (Krawczak. 1999)

- Permiten una interpretación estandarizada al poderse expresar el ADN de manera digital (Fries & Durstewitz. 2001)
- El genotipado se puede automatizar (Lindblad-Toh., *et al*. 2000)
- Permiten utilizar la información, no sólo los exones, sino también de los intrones, que son nucleótidos que nos manifiestan.
- Apenas se producen cambios de un generación a la siguiente
- Permiten buscar variaciones genéticas dentro de una población

Podemos decir que los SNPs son la unidad fundamental de variación genética debido a su abundancia (Heaton., *et al*. 2005), estabilidad genética y capacidad de ajuste a los análisis automatizados (Lindblad-Toh., *et al*. 2000).

En la actualidad se han utilizado SNPs como marcadores de muy diversos rasgos (Weston., *et al*. 1997) y para estudios de ligamiento (Hamada., *et al*. 2005). También sabemos que los SNPs, puesto que (al igual que los microsatélites), son marcadores codominantes, resultan especialmente eficientes para analizar individuos diploides como en el caso de la especie bovina, puesto que pueden darse tres posibles combinaciones (homocigoto, homocigoto del alelo de referencia, o heterocigoto).

Otra ventaja fundamental de utilizar este tipo de marcadores es que buena parte de la información genómica hoy en día disponible viene referida a SNPs. Por tanto, utilizar SNPs desde un principio, posiblemente permita un mayor aprovechamiento de la información disponible y, al mismo tiempo, permita incorporar la información a las bases de datos.

La utilización de los marcadores SNPs, puesto que proporcionan una lectura mucho más sencilla, se ha visto favorecida con el surgimiento de los denominados “chips ubicuos”. Los sistemas ubicuos son aquellos que pueden estar prácticamente en cualquier parte, gracias fundamentalmente a su pequeño tamaño. Esta capacidad para localizarse en los diferentes lugares, naturalmente, viene acompañado de nuevos usos. La aparición de estos nuevos usos de los sistemas ubicuos emergentes (en este caso un chip de SNPs), permitirá que su lectura sea de interés en diferentes contextos, y por tanto, serán más los laboratorios que puedan proporcionar una lectura de los mismos.

Este concepto de ubicuidad, rompe en gran medida con otro concepto cada vez más habitual que es el de los ASICs (circuito integrado para aplicaciones específicas - *application-specific integrated circuit*). Los ASICs son sistemas que se desarrollan para cumplir con una función específica muy concreta. Estos sistemas, que se van imponiendo cada vez más, a medida que avanza la miniaturización y las herramientas de diseño, suelen traducirse en un incremento del coste de adquisición de los sistemas, la imposibilidad de reparar el sistema por partes y, en definitiva, una menor versatilidad. Mientras que la ubicuidad de los chips, suele traducirse en una reducción de los costes y

una mayor aplicabilidad, al simplificarse los envíos a los laboratorios; los ASICs se muestran como algo completamente contrario esto.

Los marcadores SNPs comparten con los micosatélites que son codominantes, pero, contrariamente a los mismos, los SNPs son bialélicos. Los marcadores bialélicos, pueden proporcionar una información más limitada, puesto que sólo pueden diferenciar entre dos grupos de sujetos: los que tienen la copia de referencia y los que no. Por otro lado, los SNPs presentan una importante ventaja con respecto a los marcadores multialélicos de número de alelos desconocido, como por ejemplo los AFLP. Mientras que para estos últimos, el espacio de búsqueda de variables es infinito, para los SNPs el espacio de búsqueda es finito y toma valor uno.

Fundamentalmente desde el proyecto genoma de bovino (2009) se han detectado multitud de SNPs asociados con determinadas razas (Amaral., *et al*. 2009), (Ramos., *et al*. 2009), (Van Bers., *et al*. 2010), (Wiedmann., *et al*. 2008) e incluso algunos ya se han validado.

En la actualidad la información disponible del genoma de la vaca es abundante aunque cabe esperar que futuras investigaciones sirvan para disponer de una información mucho mayor.

Algunas de las características más importantes de genoma de la vaca son que está compuesto por unos $3 \cdot 10^9$ pares de bases repartidos en 29 cromosomas y que corresponde a uno de los más grandes de los estudiados hasta ahora. También, que contiene entorno a 22000 genes y 14000 de los cuales son comunes en todos los mamíferos; y que entorno al 80% de sus genes están presentes también en la especie humana.

5. Novedades de este estudio

Este estudio comparte la característica habitual de los GWAS (*Genomic Wide Association Study*) en cuanto a que se pretende identificar secuencias de información derivadas de individuos no emparentados para medir unos rasgos, la raza en este caso. Dicho de otra forma, se trata de una búsqueda de polimorfismos candidatos o asociados con una característica. Si bien, contrariamente a la mayoría de estas investigaciones, en este estudio no se busca la relación entre múltiples polimorfismos y una característica medible, sino que se eligen los loci en función de su capacidad de asignación a una de las razas concretamente. Asimismo, las investigaciones en el campo de la asignación de individuos se diferencian de las investigaciones en el ámbito de la selección genómica en que, en el caso de los problemas de asignación, conocemos la variable que se desea estimar (raza de pertenencia) y podemos evaluar los resultados de la estimación.

El estudio también es novedoso en cuanto a que emplea SNPs como marcadores con las correspondientes ventajas arriba mencionadas, y en cuanto a que emplea el método estadístico de PLS para lograr una reducción de la dimensión. Sin embargo, lo

más novedoso de este estudio probablemente sea la combinación de PLS con la técnica de LDA (*linear discriminat analysis*). El motivo por el que se decidió combinar estas técnicas es, precisamente, la peculiaridad del problema planteado: no queremos extraer las variables latentes más informativas sino seleccionar las variables originales que mejor desempeñan dicha función predictiva, pero, además, teniendo en cuenta que no se predice una variable continua sino una categoría, y que nos encontramos ante un problema de asignación no binaria.

La motivación y el interés de este estudio también son originales en cuanto a que se realiza sobre un conjunto de razas que, se estima, cubren buena parte de la variabilidad de las razas españolas, y en cuanto a que se consideran también otras razas europeas, lo cual permitirá la estimación del grado de diferenciación entre las razas españolas y las europeas.

3. OBJETIVOS

Este estudio tiene tres objetivos principales:

1. Comprobar si la combinación de las técnicas de mínimos cuadrados parciales y análisis linear discriminante resulta eficiente en la asignación de individuos a sus poblaciones usando marcadores SNP.

2. Seleccionar un subconjunto reducido de loci que permita asignar correctamente nuevas muestras a alguna de las razas estudiadas.

3. Identificar las regiones cromosómicas que pudieran tener una mayor relevancia en la diferenciación de razas.

4. MATERIAL Y MÉTODOS

1. Razas del estudio

1.1. Elección de las razas

Para la elaboración de este estudio se ha optado por elegir siete de las razas autóctonas de aptitud cárnica más representativas. Adicionalmente, se ha decidido completar el estudio incluyendo cuatro razas europeas pensando que esto permitiría llegar a más conclusiones. Con este objetivo y para poder hacer más comparaciones, dos de las razas europeas son de aptitud lechera y una es de doble aptitud. La única raza europea de aptitud cárnica considerada debe su inclusión en el estudio a ser una de las razas más extendidas por todo el mundo.

1.1.1. Elección de las razas españolas

La aparición de dos nuevos problemas en el ganado bovino de aptitud cárnica, como son la pérdida de biodiversidad y el aumento de la consanguinidad, lleva a los ingenieros agrónomos a planteare la idea de hacer un sondeo para comprobar cuál es, efectivamente, el grado de consanguinidad. Para ello, se opta por realizar estudios con grupos de razas representativas que permitan obtener unos resultados fiables en sus correspondientes rebaños, pero que al mismo tiempo esos resultados puedan ser extrapolados a algunas de las otras razas más minoritarias de los sistemas ganaderos nacionales. De esta manera, surge un interés por identificar un grupo de razas autóctonas que sea representativo a nivel nacional, capaz de abarcar las diferencias tan significativas que se dan, tanto en producción como en características de la canal, en las ganaderías españolas. De hecho, la elección de las razas está también fundamentada en las diferencias que hay entre las razas en cuanto a tamaño de las cabañas, ambientes de cría y grado de cruzamiento dentro y fuera de las razas.

Así mismo, se buscaba un grupo de razas que fuesen competitivas en cuanto a producción de carne, pero que al mismo tiempo apoyasen el sistema de explotación tradicional español que se caracteriza, fundamentalmente, por su carácter extensivo , un clima extremo, un terreno agreste y una escasez de pastos. También era conveniente que las razas tuvieran un desarrollo medio-alto en cuanto a control e implantación tecnológica y que tuvieran un gran potencial de modernización. De este modo, se dispondrá de datos e información más abundante y fiable, mejorando los recursos disponibles para las diferentes investigaciones; y se garantizará una mayor utilidad de las mismas investigaciones, al realizarse sobre razas cuyos sistemas de explotación se espera que se mantengan.

Por todo lo anterior, las razas elegidas para el estudio fueron (ordenadas de mayor a menor censo):

- Asturiana de los Vales (ASTV)
- Rubia Gallega (RGAL)

- Avileña-Negra Ibérica (AVIL)
- Morucha (MORU)
- Retinta (RETI)
- Pirenaica (PIRE)
- Bruna del Pirineus (BRUP)

Estas razas coinciden con las razas autóctonas de mayor censo en España, a excepción de Bruna dels Pirineus que como muestra la siguiente ilustración es superada por Asturiana de las Montañas (no considerada en este estudio):

FUENTE: SITRAN (Diciembre de 2014)

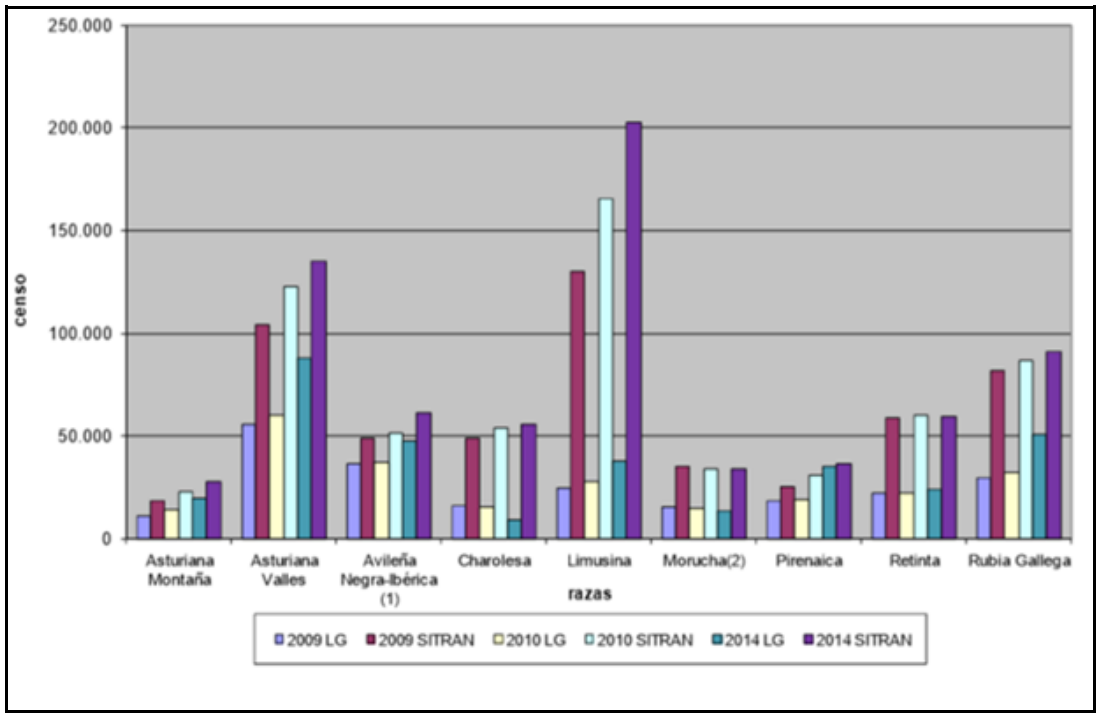


Ilustración 5. Censo de las principales razas

El gráfico muestra el censo actual de las principales razas autóctonas en relación con las dos principales razas de importación (Limusín y Charolés). Se aprecia además una falta de concordancia entre los censos de ARCA (Sistema nacional de información de razas ganaderas): 30,9% nodrizas está en LG (Libro Genealógico); y SITRAN, según el cual sólo un 46% de las declaradas están en LG.

La elección de la raza Bruna dels Pirineus está fundamentada en su censo (aún elevado), en la pureza de sus rebaños y el elevado grado de aislamiento de los mismos; y, fundamentalmente, porque recientemente se han hecho investigaciones con esta raza.

Las razas españolas seleccionadas están incluidas en el Catálogo Oficial de Razas de Ganado, donde aparecen todas como Razas Autóctonas de Fomento, a excepción de Bruna dels Pirineus que figura en el Grupo de Razas Autóctonas en Peligro de Extinción.

Puesto que las razas elegidas representan el 72% del censo nacional, el número de ganaderías para las que se podrá efectuar la prueba genómica será muy elevado. La prueba servirá para dar un valor añadido a los productos derivados de animales puros, lo que se pretende sirva para incrementar el porcentaje de animales puros frente a mestizos. Además, dada la envergadura del estudio a nivel nacional (siete razas principales con ganaderías bastante entremezcladas) se podría llegar a resultados interesantes en lo referido a interacción entre rebaños y fenómenos de migración, selección, mutación y deriva genética.

Justificación: elección de las razas

La elección se justifica, fundamentalmente, por las siguientes razones:

- Las razas son representativas: En conjunto, representan un 72% del censo nacional autóctono de bovino (MAGRAMA)

- La base de datos es avanzada. En todas ellas se empezó a hacer BLUP en la década de los 90.

- Las razas presentan diferencias significativas en producción (Piedrafita., *et al*. 2003) y en características de la canal (Gil., *et al*. 2001).

- Recientemente se han realizado estudios con estas razas. Lo que permite un mayor aprovechamiento de la información obtenida.

- Para todas ellas existe una entidad competente oficial con suficiente fuerza como para ser capaces de proporcionar información abundante y fiable.

Por otro lado, las razas elegidas cumplen con el perfil para el que la FAO

propone elaborar planes de conservación. A saber, (FAO - 2007):

- Razas autóctonas adaptadas a un determinado ambiente específico
- Razas locales de alta productividad o con productos bien diferenciados
- Razas únicas desde el punto de vista genético
- Razas de gran belleza
- Razas importantes desde el punto de vista histórico

Justificación: representatividad de las razas

En este apartado se justifica la elección de las 7 razas autóctonas seleccionadas en base a su nivel de representatividad en España y en la Península Ibérica.

Para comprender la distribución de las razas en España resulta de utilidad analizar lo que ocurre en la península puesto que los Pirineos suponen un grado de aislamiento considerable.

En la península Ibérica hay 51 razas oficialmente reconocidas, 38 de las cuales son españolas. En (Martín-Burriel., *et al*. 2011a) se genotiparon 40 de estas razas con 19 marcadores microsatélites y llegaron a las siguientes conclusiones:

- Se observaron valores de heterocigosidad que oscilan entre 0.596 y 0.787. Correspondiendo a una heterocigosidad media, aunque con gran variabilidad entre unas razas y otras.

- Algunas razas tenían individuos correspondientes a dos clusters distintos; aunque no fue el caso de las razas de este estudio, con un mayor grado de diferenciación.

- Se concluye que la variabilidad genética está más correlacionada con la zona productiva que con la diferenciación morfológica, lo cual es una prueba más de lo eficaz que resulta un análisis a nivel genético frente al morfológico.

- El relativo aislamiento genético es consecuencia de las barreras geográficas, la deriva genética y la adaptación al ambiente y al sistema productivo.

- Las razas Europeas han tenido una marcada influencia de las razas del Oriente Próximo (Troy., *et al*. 2001), (Beja-Pereira., *et al*. 2006).

- La mezcla entre razas debe ser tomada en consideración antes de invertir en la asignación de individuos a razas.

La diferenciación genética puede expresarse con un gráfico en forma de estrella. Se ha usado la Distancia de Reynolds (Reynolds., *et al*. 1983) o Distancia de Parentesco (Martín-Burriel., *et al*. 2011a) para representar la distancia genética entre las razas ibéricas. Este método, más concretamente, utiliza un modelo basado en SMM (tasas de mutación por pasos - *stepwise mutation model*) (Ohta., *et al*. 1973).

La siguiente ilustración muestra la relación genética entre 40 razas oficiales de la península Ibérica de las 51 registradas:

FUENTE: adaptado de Martín-Burriel., *et al*. 2011a

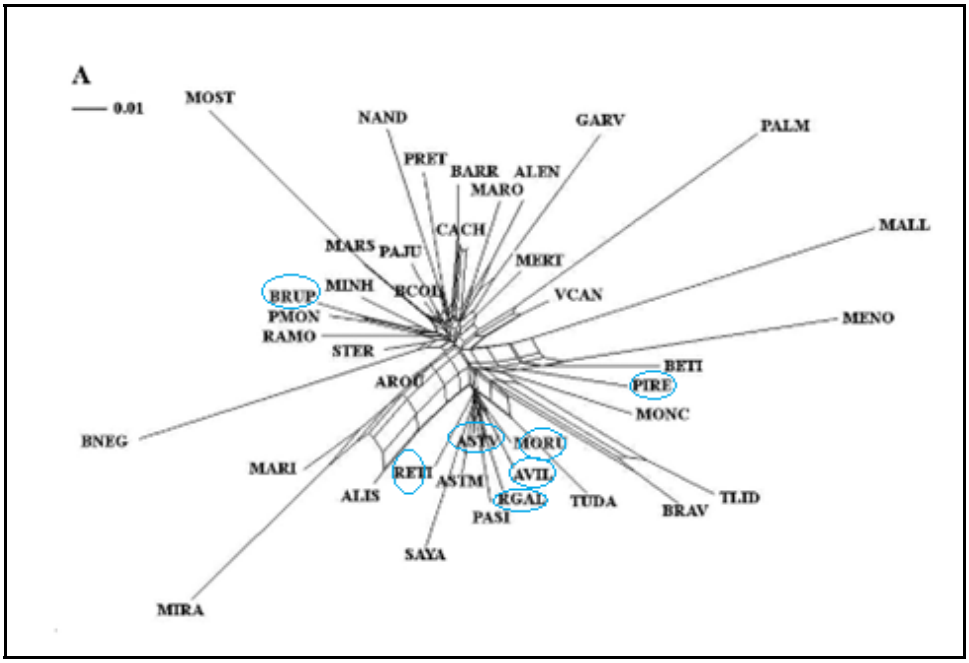


Ilustración 6. Relación genética entre razas de la Península ibéricas

Título original: Neighbor-joining tree summarizing the Reynolds distances among 40 native cattle from Spain and Portugal

Leyenda:

Alistana (ALIS), Asturiana de los Valles (ASTV), Asturiana de las Montañas (ASTM), Avileña (AVIL), Berrenda en Colorado (BCOL), Berrenda en Negro (BNEG), Betizu (BETI), Bruna dels Pirineus (BRUP), Mallorquina (MALL), Marismeña (MARS), Menorquina (MENO), Monchina (MONC), Morucha (MORU), Mostrenca (MOST), Negra Andaluza (NAND), Pajuna (PAJU), Parda de Montaña (PMON), Pasiega (PASI), Pirenaica (PIRE), Retinta (RETI), Rubia Gallega (RGAL), Sayaguesa (SAYA), Serrana de Teruel (STER), Toro de Lidia (TLID), Tudanca (TUDA), Vaca Canaria (VCAN), Vaca Palmera (PALM), Alentejana (ALEN), Arouquesa (AROU), Barrosã (BARR), Brava de Lide (BRAV), Cachena (CACH), Garvonesa (GARV), Marinhoa (MARI), Maronesa (MARO), Mertolenga (MERT), Minhota (MINH), Mirandesa (MIRA), Preta (PRET), and Ramo Grande-Azores (RAMO).

Como se puede apreciar, las razas del estudio se encuentran repartidas en tres grupos, siendo Bruna dels Pirineus (BRUP) la más alejada del resto. Se aprecia además que precisamente gracias BRUP nuestra selección de razas abarca prácticamente todo el espectro de variabilidad genética de las razas españolas.

A continuación, se muestra un gráfico en el que se puede contrastar la información anterior con la formación de cuatro clusters fundamentales de razas:

FUENTE: adaptado de Martín-Burriel., *et al*. 2011a

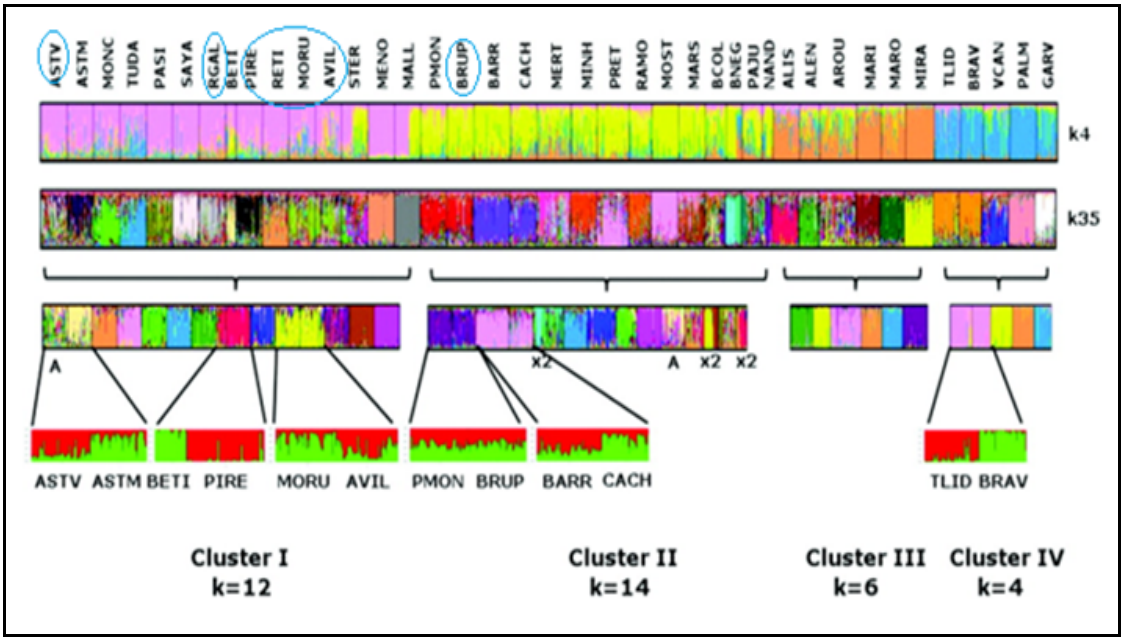


Ilustración 7. Representación cluster de las razas de la península ibérica

Título original: Genetic structure of 40 Spanish and Portuguese autochthonous cattle breeds

Leyenda:

Alistana (ALIS), Asturiana de los Valles (ASTV), Asturiana de las Montañas (ASTM), Avileña (AVIL), Berrenda en Colorado (BCOL), Berrenda en Negro (BNEG), Betizu (BETI), Bruna dels Pirineus (BRUP), Mallorquina (MALL), Marismeña (MARS), Menorquina (MENO), Monchina (MONC), Morucha (MORU), Mostrenca (MOST), Negra Andaluza (NAND), Pajuna (PAJU), Parda de Montaña (PMON), Pasiega (PASI), Pirenaica (PIRE), Retinta (RETI), Rubia Gallega (RGAL), Sayaguesa (SAYA), Serrana de Teruel (STER), Toro de Lidia (TLID), Tudanca (TUDA), Vaca Canaria (VCAN), Vaca Palmera (PALM), Alentejana (ALEN), Arouquesa (AROU), Barrosã (BARR), Brava de Lide (BRAV), Cachena (CACH), Garvonesa (GARV), Marinhoa (MARI), Maronesa (MARO), Mertolenga (MERT), Minhota (MINH), Mirandesa (MIRA), Preta (PRET), and Ramo Grande-Azores (RAMO).

Al igual que en la ilustración anterior se aprecia que cinco de las siete razas de este estudio son parecidas genéticamente entre sí; que ASTV es incluida dentro del mismo cluster, pero que la diferencia con las anteriores es prácticamente la mayor posible dentro del cluster. También se aprecia que BRUP pertenece a otro cluster, aunque éste sea el más cercano. Quedando los otros dos cluster para las razas portuguesas.

1.1.2. Elección de las razas foráneas

Incluir razas foráneas en el estudio permitirá comparar la variabilidad genética así como

estimar la proporción de esa variación que está ligada a la geografía, o en qué medida el asentamiento de las razas en la península ibérica lleva a una diferenciación.

Las razas europeas consideradas en este estudio son las siguientes: Brown Swiss (BRSW) (Suiza), Fleckvieh (FLCH) (Alemania), Guernsey (GUER) (isla de Guernsey, Francia), y Simmental (SIMM) (Suiza).

Esta selección de razas foráneas es de enorme utilidad para este estudio y para las diversas investigaciones que se están haciendo en genética de poblaciones con razas europeas. La diversidad de sus orígenes y la gran diferencia tanto genética como morfológica que existe entre las mismas, convierte a este grupo de razas en un buen referente de la situación en que se encuentran las razas europeas.

Como puede verse, dos de las razas (BRSW y SIMM) son originarias del mismo país, pero es de interés considerarlas a ambas puesto que BRSW es una raza fundamental de aptitud lechera, mientras que SIMM es una de las razas de aptitud cárnica más distribuidas por el mundo.

Incorporar la raza GUER de aptitud lechera es también de enorme interés dado su gran porcentaje de pureza consecuencia del aislamiento en la isla francesa de Guernsey. Pero también por la peculiaridad de su leche y la gran exportación de animales de esta raza a regiones tan lejanas como EE.UU (Estados Unidos).

FLCH, por su parte, es otra raza que conviene tener en cuenta, puesto que es un buen ejemplo de la adaptabilidad de una raza (SIMM) cuando los objetivos de producción toman caminos diferentes. SIMM, en origen vaca lechera, se fue seleccionado para aptitud cárnica, mientras que FLCH se seleccionó, fundamentalmente en el sur de Alemania y en Austria para doble aptitud.

1.2. Características de las razas elegidas

1.2.1. Razas autóctonas

Características generales

En cuanto a las razas autóctonas, en los siete casos se trata de razas seleccionadas para producción de carne y altamente adaptadas al ambiente de sus respectivas zonas de explotación.

Algunas características que comparten estas razas son su rusticidad, su buena aptitud maternal y su capacidad lechera. En cuanto a los datos genealógicos de estas razas han pasado, generalmente, por las siguientes fases:

- Algunas de estas razas tienen datos en el LG anteriores a 1950; siendo los registros más antiguos los correspondientes a Avileña-Negra ibérica, con datos

desde 1933.

- En 1970 con la Resolución de la Dirección General de Ganadería se reglamenta la información para la mayoría de las razas.
- Más tarde, en 1975 el Ministerio de Agricultura atribuye la realización de los Libros Genealógicos a las asociaciones nacionales de Ganado Vacuno Selecto.
- En 1977 se vuelven a actualizar los datos con la Resolución de la Dirección General de la Producción Agraria de 28 de febrero de 1977.
- Por último, en 1980 se modifican algunas de las denominaciones de raza, aprobándose en su caso una nueva reglamentación específica del Libro Genealógico.

Todas estas razas pertenecen a alguna Asociación, Federación, o Confederación con diferentes competencias según el caso: desde la elaboración y control del Libro Genealógico, control del rendimiento cárnico, y coordinación del Programa de Mejora; hasta el testaje de sementales jóvenes. Estas entidades colaboran para determinados trabajos con el MAGRAMA; o con determinados departamentos de las Universidades Españolas.

Estas razas españolas, también se han visto afectadas con el desarrollo del proyecto europeo Gene2Farm (Enero de 2012 a Diciembre de 2015), que va encaminado a responder las necesidades de la industria cárnica de vacuno, y en especial de los ganaderos y del consumidor final. La definición formal del proyecto Gene2Farm, por parte de la Comisión Europea es que se trata de un “sistema europeo de nueva generación para el manejo y la mejora de los sistema ganaderos de bovino de carne”. El proyecto basa su capacidad de actuación en el nuevo conocimiento científico y busca lograr una mayor sostenibilidad y rentabilidad de los sistemas ganaderos europeos de vacuno de carne.

Feagas (Federación Española de Asociaciones de Ganado Selecto) colabora directamente con este proyecto y es líder de tarea en el paquete de trabajo WP6: Aplicación de modelos estadísticos a diferentes situaciones en razas específicas, dirigido por Trygve Roger Solberg. WP6 tiene como objetivo coordinar todos los módulos de la selección de paquetes de genómica de nueva generación. Las tareas fundamentales son llevar a cabo pruebas en situaciones específicas ocasionadas en la industria, revisar la actuación y hacer las modificaciones que sean necesarias.

Los otros paquetes de trabajo de Gene2Farm son:

- WP1: Nuevos modelos estadísticos, para la selección del genoma, desarrollo de software y pruebas, dirigidos por Theo Meuwissen
- WP2: Información sobre el genoma y los paneles de SNP, dirigido por Giulietta Minozzi.

- WP3: Fenotipos de mejora y gestión de ganado, liderado por Giorgios Banos
- WP4: Estructura de la población, dirigido por Miguel Toro
- WP5: Integración y suministro de herramientas, dirigido por Alessandra Stella
- WP7: Formación, transferencia de conocimiento y la Sostenibilidad, dirigido por Toine Roozen.
- WP8: Gestión del proyecto, dirigido por John Williams

Características particulares de las raza autóctonas elegidas para el estudio

Se describen a continuación las características más importantes de las razas autóctonas elegidas para este estudio. La mayor parte de esta información procede del MAGRAMA y de Feagas.

- Rubia Gallega

Es una raza indígena que procede de *Bos taurus*. La selección llevada a cabo en esta raza se ha traducido en un aumento del formato, un mayor ritmo de crecimiento y un mejor índice de conversión con resultados muy satisfactorios. El éxito de esta selección ha llegado hasta el mercado en forma de comercialización de carne pura de esta raza. Los programas de mejora en esta raza han supuesto la implementación de la IA por un lado y por otro la valorización de ejemplares. Estos valores de calidad que se atribuyen a los animales cumplen con el protocolo de control de la capacidad y valor genético establecido en la Decisión 94/515/CE.

Es característico de esta raza su excelente crecimiento, su carácter musculoso y una estructura muy buena de la carne. El color de la piel de los animales es Rubia con manchas rosadas.

Entidad competente: Asociación Nacional de Criadores de ganado vacuno selecto de raza Rubia Gallega (ACRUGA)

Censo estimado (MAGRAMA): 94 682 cabezas de ganado. Los primeros registros de esta raza datan de 1969.

Características productivas:
Las vacas tienen precocidad sexual, alta fertilidad, un corto intervalo entre partos y una alta capacidad lechera que les permite mantener los bastante habituales partos gemelares.
Es la raza más prolífica de las razas autóctonas españolas. Se trata de una raza rústica, resistente y dócil, y con una larga longevidad (hasta 21 años).

Distintivos de calidad: La raza está integrada dentro de la indicación geográfica

protegida de “Ternera Gallega”.

Otros datos de interés:
Su censo ha caído un 4,8% desde 2009.
Es una de las razas más precoces con un 70,2% estimado de partos antes de los 3 años.
Raza muy utilizada en cruces con Morucha, Avileña-Negra ibérica y con mestizos en España. Además de con Holstein y razas cebuínas en Sur-América.

- Asturiana de los Valles

Raza que desciende del Tronco Cantábrico como se aprecia, por ejemplo, en el color castaño de la capa. Implantada en España desde 1950 viene de la raza originaria Pardo-Alpina (Austria). En sus orígenes fue una raza de triple aptitud.

Los toros son muy utilizados en cruzamiento industrial con vacas frisonas o de otras razas, ya sea con monta natural o mediante IA. Algunos individuos presentan un gen de hipertrofia muscular.

Entidad competente: Asociación Española de Criadores de Ganado Vacuno Selecto de la Raza Asturiana de los Valles (ASEAVA)

Censo estimado (MAGRAMA): 66162 cabezas de ganado.

Características morfológicas:
Presentan perfil de recto a ligeramente subconvexo y aspecto equilibrado. Se distinguen dos aptitudes distintas: culón (más musculada) y normal.

Características productivas:
Carne de calidad y alto índice de conversión (para tratarse de una raza criada en condiciones semi-extensivas). Raza muy adaptada a las zonas montañosas con una temperatura media anual de entorno a 9º-11ºC y 1000-1400 cc/m^2 de lluvia.

Distintivos de calidad:
Marca "Xata Roxa" integrada en la Indicación Geográfica Protegida de "Ternera Asturiana". Con los siguientes productos típicos: Ternera sacrificada con menos de 12 meses, y Añojo sacrificado entre los 12 y los 18 meses.

Otros datos de interés:
Baja proporción de individuos puros, tal y como reveló un estudio realizado por ASEAVA (Asociación Española de criadores de ganado vacuno selecto de la raza Asturiana) según el cual de los 99000 ejemplares sólo 44000 eran ser puros.
De las 51 razas oficiales de la península ibérica es la que mayor variedad alélica presenta (Martín-Burriel., *et al*. 2011a).

- Avileña-Negra Ibérica

Procede del Tronco Negro ibérico. Se fusionó con la Negra ibérica en 1980.

Entidad competente: AECRANI (Asociación Española de Criadores de ganado Vacuno Selecto de Raza Avileña-Negra Ibérica. 1971)

Censo estimado (MAGRAMA): 53 428 cabezas de ganado

Lo habitual es que esta raza se mantenga sin forraje, excepto en periodos de sequía.

Distintivos de calidad: Carne comercializada bajo la IGP Carne de Ávila que ampara, además de animales puros, los animales procedentes del primer cruce de Avileñas con toros de razas cárnicas. Este distintivo cuenta con el respaldo de la UE desde 1996.

- Morucha

Raza procedente de *Bos braquicerus africana*. Más concretamente tiene sus orígenes en el Tronco Negro ibérico.

Raza muy rústica con un alto índice de fertilidad y habilidad maternal. Probablemente la característica que más se haya usado como criterio de selección sea su facilidad de parto. Como inconvenientes, los animales alcanzan la edad de reproducción a una edad algo tardía.

Tradicionalmente se han usado los bueyes para manejo del ganado

Entidad competente: Asociación Nacional de Criadores de Raza Morucha Selecta.

Censo estimado (MAGRAMA): 40 026 cabezas de ganado

Características morfológicas:
Color, blanco y negro entremezclado dando un aspecto grisáceo; o bien sólo negro. Son animales de perfil recto.

Características productivas:
Es la segunda raza más prolífica de las autóctonas españolas. Produce una carne muy roja y sabrosa.

Distintivos de calidad:
DOP (Denominación de Origen Protegida): Carne de Morucha de Salamanca

- Retinta

Procedencia: Rama Roja Convexa (*Bos taurus turdetanus*) y las razas indígenas del sur de la península Ibérica.

Su distribución hacia el sur de la península Ibérica originó los tres ecotipos bien diferenciados: Colorada Extremeña, Rubia Andaluza y Retinta Andaluza. De la fusión de estos tres se obtuvo la versión mejorada que actualmente se explota. Es la principal raza bovina autóctona de la España seca.

Entidad competente: Asociación Nacional de Criadores de Ganado Vacuno Selecto Raza Retinta.

Censo estimado (MAGRAMA): 29 394 cabezas de ganado

Características morfológicas:
Capa única de color marrón. Son animales largos, de gran tamaño pero proporcionados, con una frente amplia y ligeramente subconvexa.

Características productivas:
Son animales con buenos rendimientos productivos y reproductivos, y una alta resistencia frente a parásitos. Aguantan bien las épocas de sequía y el calor. Además los animales se mantienen bastante bien con una nutrición pobre.
Habitualmente se explota en rebaños grandes, con un tamaño medio superior a 40 cabezas. Con frecuencia, comparten el terreno con otras especies y razas propias del sistema de dehesa, principalmente, la oveja Merina y el cerdo Ibérico.
En estos sistemas de explotación se requiere complementar con paja, heno y pienso, en las épocas de escasez de pastos.
La cubrición es mayoritariamente con monta natural estacional. Siendo la época de monta entre noviembre y junio.

Produce una carne roja, tierna, jugosa y posee una exquisita sapidez, con una baja relación de ácidos grasos saturados/totales.

Distintivos de calidad: las canales y piezas están identificadas con la Marca “Carne de Retinto”, propiedad de la Asociación Nacional de Criadores de Ganado Vacuno Selecto Raza Retinta.

- Pirenaica

Raza indígena de procedencia aún no confirmada. Posibles procedencias son las llamadas razas francesas “blondes”; pero también *Bos taurus europea*, *Bos taurus primigenius* o *Bos taurus brachyceros*.

Entidad competente: Confederación de Asociaciones de Criadores de Ganado Selecto de Raza Pirenaica (CONASPI), que consta en actualidad de siete asociaciones: ASPINA

(Navarra), ASGAPIR (Bizkaia), HEBE (Gipuzkoa), ARPIEL (Alava), ASAPI (Aragón), ASPIC (Cataluña), ASPICAN (Cataluña).
Censo estimado (MAGRAMA): 16 378 cabezas de ganado

Características morfológicas:
Son animales muy musculados, con línea dorso lumbar recta en la hembra y morro anguloso en los machos. Capa de color único que varía de blanco a un marrón claro.

Además, hay que destacar de esta raza su gran rusticidad, docilidad, facilidad para el parto y buena capacidad como nodriza.

Características productivas:
La raza ha demostrado ser prácticamente inmune a enfermedades de tipo genital. Está bien adaptada al pastoreo bajo condiciones climáticas extremas. Un parto al año y habitualmente fácil, tanto en hembras puras como en hembras procedentes de un cruzamiento industrial. Además tienen una gran habilidad maternal.
Puesto que son capaces de pastorear a gran altura, desempeñan un buen papel en la conservación de los recursos naturales.

La raza pirenaica está adaptada a un régimen mixto: pastoreo-estabulación de intenso componente forrajero. Es propio de esta raza el pastadero colectivo o comunal. Los tipos comerciales más característicos que produce son el añojo y vacuno mayor.

Distintivos de calidad: Participa en la Indicación Geográfica Protegida: “Ternera de Navarra” y en el Label Vasco de calidad “Euskal Okela”.

- Bruna dels Pirineus

Esta raza proviene de la unión entre una línea autóctona de Cataluña y la Parda Alpina (Suiza). Su morfología refleja de una manera clara esa pertenencia al Tronco Alpino. Aunque se ha incluido en esta lista, esta raza no figura en el catálogo oficial de razas autóctonas españolas, sino que figura en el catálogo de razas españolas en peligro de extinción.

En el desarrollo de esta raza, se ha tratado de potenciar la aptitud cárnica de la antigua Brown Swiss (década de los 50) pero evitando que se redujera el potencial lechero o la buena aptitud maternal.

Entidad competente: Federación de la Vaca Bruna dels Pirineus (FEBRUPI)

Censo estimado (MAGRAMA): 11 557 cabezas de ganado

Características morfológicas:
Son de color pardo y proporciones armónica, con tendencias longilíneas. Es una vaca rústica con precocidad mediana.
Es una raza musculada y con una fuerte estructura ósea.

Características productivas:
Sus terneros se caracterizan por tener un peso apropiado al parto y un buen crecimiento hasta el destete. Se utiliza también en cruzamiento industrial con toros de razas cárnicas. Lo habitual es que las ganaderías ocupen las zonas de valle y suban a la montaña en los meses de calor. Se producen fundamentalmente, dos tipos de terneros: De seis meses de edad (más habitual) y de 12-13 meses de edad (con unos 540 kg de PV). En el primer caso es habitual cebar luego los teneros mediante un sistema de explotación intensivo.

Comparaciones entre las razas

En este apartado se mencionan alguna de las diferencias fundamentales entre las razas autóctonas elegidas. La siguiente tabla muestra la distribución de estas razas en la península ibérica:

- Diferencias geográficas

La distribución de las razas en la península es la siguiente:

FUENTE: adaptado de FEAGAS

Raza	Zona geográfica
ASTV	N
RGAL	N-W
PIRE	N-E
BRUP	N-E
AVIL	Centro
MORU	Centro
RETI	Sur

Tabla 2. Distribución geográfica razas españolas

- Diferencias morfológicas

FUENTE: adaptado de MAGRAMA

	RGAL	ASTV	AVIL	MORU	PIRE	RETI	BRUP
Altura cruz machos (cm):	149	146	148	143	150	44	142
Altura cruz hembras (cm):	138	140	140	137	132	139	140
Peso machos (kg):	1,300	1,050	900	900	800	850	1,050
Peso hembras(kg):	600	600	500	500	525	650	600

Tabla 3. Diferencias morfológicas entre las razas autóctonas del estudio

FUENTE: Elaboración propia (Datos: FAO)

Raza	RGAL		ASTV		AVIL		MORU		PIRE		RETI	
	Machos	Hembras	Machos	Hembras	Machos	Hembras	Machos	Hembras	Machos	Hembras	Machos	Hembras
Altura media (cm)	149	138	146	140	148	140	147	137	150	132	144	139
Peso medio (kg)	1.300	575	1000	600	950	550	900	530	950	550	900	600

Tabla 4: Diferencias de tamaño entre las razas autóctonas del estudio

No se dispone de datos de tamaño para la raza BRUP

1.2.2. Razas foráneas

Características generales

Para tener una idea de como son las razas de la Europa continental, se han comparado estas razas con las británicas.

Las razas europeas son más alargadas, de madurez sexual más tardía, producen canales menos grasas y con un mayor rendimiento de piezas aprovechables, y tienen una mayor dificultad de partos cuando se cruzan, por ejemplo, con las razas británicas, que tienen una morfología más ancha.

Considerar algunas de estas razas, tan extendidas en los cinco continentes pero en particular en Europa, permitirá tener un impresión de hasta qué punto estas razas difieren de nuestras razas indígenas. Conocer con certeza las diferencias entre las dos líneas originarias de un cruce permitirá analizar de una manera más eficiente la efectividad del cruce y las posibilidades de futuros cruces.

Características particulares de cada raza

Simmental

A nivel internacional es de todas las razas elegidas, con mucho, la más conocida dada su enorme exportación y su antigüedad. Esta raza es originaria de algunas de las razas más importantes de Europa, como Fleckvieh, pero también Montbeliarde (Francia) y Razzeta d´oroppa (Italia).

SIMM es una raza en la que, desde el principio, se ha visto intensificada la selección por el rápido crecimiento de las novillas en condiciones adecuadas. Hoy en día sabemos además que SIMM es la raza con mayor variabilidad dentro-de-raza del mundo, ya que sus descendientes abarcan todo el espectro de producciones, desde las puramente lecheras hasta las puramente cárnicas.

Fleckvieh

Esta raza data sus orígenes de 1830, cuando fue importada de Suiza a la zona de Babaria y también a Austria. A partir de entonces, FLCH fue sometida a un progresivo proceso de selección para doble aptitud, hasta finalmente convertirse en una raza independiente casi un siglo después, en 1920. Posteriormente, esa raza se exportó a Francia y a Italia. Dada su demostrada capacidad como raza de doble aptitud, ha resultado de especial interés a la hora de establecer programas de selección. Como consecuencia del éxito de dichos programas, FLCH es hoy una raza muy modernizada.

Brown Swiss

Esta raza de aptitud lechera es la segunda más productiva del mundo con una media de 9000 kg/lactación de una leche que además tiene mejores propiedades (4% Gr; 3.5% Pr) para elaboración de quesos que la raza Holstein. La raza también se caracteriza por sus gestaciones largas, habituales partos gemelares, dócil temperamento y rusticidad.
Es una raza muy extendida en EE.UU. desde 1869, aunque la adaptabilidad ha hecho que estas difieran en mucho de las Brown Swiss originarias, que recibían el nombre de Schwyzer Braunvieh.

Guernsey

Esta raza es, fundamentalmente, conocida por la alta proporción de beta carotenos en su leche. La raza data sus orígenes de 1700 y ha sido sometida a un fuerte aislamiento, en primer lugar por su situación geográfica (isla francesa de Guernsey), y después por la prohibición de importar otras razas a la isla. Como resultado de esta diferenciación, algunos países, fundamentalmente EE.UU. Se interesaron por esta raza e importaron su leche y semen para inseminar. Estas vacas son fáciles de manejar y presentan algunas ventajas con respecto a otras razas, por ejemplo, tienen una enorme facilidad de parto, y son muy longevas.

Algunas federaciones de ganaderos de estas razas tienen, al igual que Feagas, participación directa en el proyecto Gene2Farm:

- The English Guernsey Cattle Society
- Swiss Brown Cattle Breeders´ Federation
- European Simmental Federation
- European Brown Swiss Cattle Federation

1.3. El problema de la consanguinidad en España

En un estudio referente al creciente problema de la consanguinidad en los rebaños de las siete razas autóctonas consideradas en este estudio, (Cañas-Álvarez., *et al*. 2014a) se concluye que la consanguinidad es significativa, llegando incluso a amenazar la variabilidad de las razas. Pero que, sin embargo, el grado de parentesco es, por lo general reducido, como consecuencia de un uso no excesivo de los machos en las explotaciones.

La explicación de algunos conceptos servirá para entender mejor estas conclusiones:

- Consanguinidad: Se entiende por consanguinidad al producto del cruzamiento entre dos individuos emparentados y permite cuantificar la homocigosis en relación a una población base.

- Coeficiente de consanguinidad: es una forma de expresar el grado de consanguinidad y explica la probabilidad de que los genes presentes en un locus de un individuo X sean idénticos por ascendencia.

- Parentesco Aditivo es, por el contrario, una medida de covarianza entre valores de individuos emparentados que permite expresar el porcentaje de genes compartidos por relación de parentesco. Si bien, lógicamente, esta covarianza será considerablemente mayor cuando dos individuos tienen consanguinidad. Lo que mide este coeficiente de Parentesco Aditivo es la proporción esperada de genes en común idénticos por ascendencia entre dos individuos X e Y.

- El coeficiente de parentesco, por su parte, es la correlación entre valores de cría de X e Y (Wright. 1921). Dicho de otra forma, el coeficiente de parentesco es la probabilidad de que dado un locus, dos individuos sean idénticos por descendencia.

Según dicho estudio, la situación de estas siete razas permite estimar el grado de consanguinidad en la mayoría de las razas del panorama nacional.

Otras conclusiones extraídas de investigaciones realizadas con las razas objeto de este estudio son las siguientes:

- La información disponible está en continuo aumento

- La inseminación artificial continúa sin extenderse, a excepción de Rubia Gallega

- En todas las razas ha aumentado el censo, especialmente en ASTV, BRUP, y RGAL con incrementos del 158 al 381%. Incrementos, muy probablemente, explicados por las ayudas gubernamentales en programas de selección (Gutiérrez., *et al*. 2003).

- La proporción de hembras por macho varía entre 10:1 (ASTV) y algo más de 30:1 (RGAL); y entre 14:1 y 22:1 para el resto de razas. Lo que se explica por la lenta implantación de la IA a excepción de RGAL. Así mismo, la elevada proporción de machos en ASTV probablemente se deba a que se trata de rebaños más aislados.

- El coeficiente de consanguinidad ha aumentado más en las razas de estudio que en otras razas minoritarias, puesto que, estas siete han sido sometidas a programas de selección más intensos. Para las siete razas del estudio la consanguinidad ha aumentado desde 1998 a 2009, entre 0.6% (BRUP) y 7.2% (RETI).

La consanguinidad es un problema en cuanto a que se traduce en un incremento del número de homocigotos recesivos que se pueden manifestar en la aparición de caracteres indeseables. Por otro lado, se debe tener en cuenta que ese coeficiente de consanguinidad, muy probablemente, esté subestimado debido a una pérdida de información en los pedigrees. Por tanto, este coeficiente pierde veracidad precisamente en las razas en las que se tiene menos información de pedigrees. Por otra parte, se plantea el interrogante de si ese incremento en los coeficientes se debe a un incremento de la consanguinidad o a una mayor disponibilidad de la información que permite detectar la consanguinidad.

El hecho de que, aunque los coeficientes hayan aumentado en todas las razas, lo hayan hecho especialmente en aquellas en las que se ha visto incrementada la información disponible, sugiere que la segunda hipótesis es la más acertada. Sin embargo, resulta preocupante el hecho de que, por ejemplo, en las razas irlandesas, en las que la información de pedigree se ha completado tanto como en estas razas españolas, el grado de consanguinidad no haya aumentado tanto.

Se ha detectado que el grado de consanguinidad no está relacionado con la proporción de individuos con consanguinidad. Por ejemplo, la raza pirenaica es la raza con mayor proporción de animales con consanguinidad, pero presenta un grado bajo de consanguinidad. De lo que se puede deducir que aunque los ganaderos han usado un número pequeño de reproductores, se ha realizado un esfuerzo para evitar la reproducción entre parientes cercanos. Otra explicación de este bajo grado de consanguinidad podría ser la existencia de un gran número de cabañas de multiplicación en esta raza (ver Tabla 5).

- El aumento de la consanguinidad ha ocasionado un descenso del tamaño efectivo de población, que en algunas de las razas es considerablemente menor de lo recomendado.

  El tamaño efectivo de la población queda definido por la expresión:

$$Ne = \frac{4\,NmNf}{Nm + Nf}$$

  , donde Nm y Nf son el número efectivo de machos y hembras, respectivamente , en condiciones de apareamiento aleatorio bajo la condición de que el número de apareamientos sea finito y no haya variación en el número de apareamientos por parte de ninguno de los dos sexos (Falconer & Mackay. 1996).

  La siguiente tabla muestra el N_e de las razas del estudio:

FUENTE: Cañas-Álvarez., *et al*. 2014a

Breed[1]	Year	Males	Females	N_e
AV	1998	1,152	8,064	4,032
	2003	1,792	15,501	6,425
	2009	2,444	20,884	8,752
ANI	1998	232	5,318	889
	2003	310	6,856	1,186
	2009	389	7,504	1,479
BP	1998	77	1,369	292
	2003	154	2,643	582
	2009	335	6,593	1,275
Mo	1998	159	1,588	578
	2003	167	3,355	636
	2009	125	1,947	470
Pi	1998	551	7,372	2,051
	2003	677	9,612	2,530
	2009	810	10,531	3,009
Re	1998	239	3,933	901
	2003	290	4,246	1,086
	2009	345	4,573	1,283
RG	1998	168	5,154	651
	2003	207	9,190	810
	2009	433	16,512	1,688

Tabla 5: Tamaño efectivo de la población.

Número de machos y hembras que parieron en los años 1998, 2003 y 2009, y N_e (tamaños efectivo de Población) correspondiente, en condiciones de apareamiento aleatorio.

Razas: Asturiana de los Valles (AV), Avileña–Negra Ibérica (ANI), Bruna dels Pirineus (BP), Morucha (Mo), Pirenaica (Pi), Retinta (Re) y Rubia Gallega (RG)

En la tabla se puede apreciar un N_e mucho mayor en Avileña y menor en Morucha y Rubia Gallega, lo cual se explica porque las ganaderías de Avileña están más aisladas y la IA ha tenido más aceptación en Rubia Gallega.

- El grado de parentesco es menor de lo indicado según el coeficiente de

consanguinidad, debido a un esfuerzo por parte de los ganaderos para evitar la reproducción entre animales emparentados. Esto se corrobora por el hecho de que, incluso en los rebaños en que se ha realizado una selección más intensa, se han utilizado machos ajenos a la explotación.

- Se ha detectado el fenómeno de cuello de botella en la mayoría de estas razas, favorecido por el hecho de que el número efectivo de ancestros ha disminuido.

  Este fenómeno de cuello de botella es perfectamente detectable porque el número efectivo de fundadores es superior al número efectivo de ancestros. Esto ocurre así especialmente en PIRE, RGAL y ASTV, aunque no en AVIL.

  Se entiende por número efectivo de fundadores al número de ancestros que explican el 50% de la variabilidad de la raza, que varía entre 17 (RGAL) y 89 (MORU), que son unos valores muy parecidos a los que encontramos en otras razas internacionales. Únicamente ASTV es una excepción en este sentido y muestra valores bastante más elevados. No obstante, dado que los valores de estos coeficiente están muy influenciados por la cantidad de información de que se disponga, se utiliza el parámetro de probabilidad de origen de gen, que considera la reducción en la variabilidad de una raza que se deriva de la utilización de un mayor o menor número de sementales. En este parámetro todas las razas, excepto BRUP, experimentan una ligera reducción.

  La probabilidad de origen de gen también ha aumentado desde 1995 (Gutiérrez., *et al*. 2003). Lo cual, sigue pudiéndose explicar por un incremento de información en los Libros Genealógicos.

- Hay un extenso intercambio de machos de los diferentes rebaños de una misma raza y ninguna explotación se ha aislado del resto en cuanto a intercambio de animales.

- El número de cabañas de multiplicación es elevado y/o se ha incrementado en los últimos años como consecuencia de la adecuación de los programas de selección.

  Uno de los principales problemas de los programas de selección es la aparición de los llamados “rebaños núcleos” en los que, por querer incrementar el valor genético de su rebaño, los ganaderos en ocasiones cometen el error de utilizar sus propios toros de manera excesiva. La no incorporación de sangre nueva a la ganadería acaba generando importantes problemas que podrían incluso extenderse luego al resto de las explotaciones en la segunda fase del programa en la que la selección se expande desde las granjas de selección al resto de las explotaciones. Es por tanto, de un interés especial detectar este tipo de problemas, razón por la cual se diseñan tablas como la que se muestra a continuación:

FUENTE: Cañas-Álvarez., *et al*. 2014a

		Type of herd[2]				
Breed[1]	Year	Nucleus	Multiplier 1	Multiplier 2	Commercial 1	Commercial 2
AV	1998	0	18.2 (72.02)	5.68	7.62 (81.59)	68.67
	2003	0	21.05 (75.56)	8.42	6.57 (84.84)	63.78
	2009	0	23.05 (77.70)	10.58	6.77 (84.19)	59.60
ANI	1998	0	10.14 (51.88)	3.32	7.34 (81.72)	79.20
	2003	0	10.36 (51.51)	4.33	9.12 (77.04)	76.20
	2009	0	13.64 (53.13)	5.34	9.42 (78.18)	71.59
BP	1998	0	7.89 (65.87)	0.88	4.97 (77.32)	86.26
	2003	0	11.70 (66.46)	5.34	12.48 (72.03)	70.48
	2009	0	39.85 (56.15)	3.42	20.54 (58.11)	36.19
Mo	1998	0	38.21 (45.90)	3.65	19.93 (46.99)	38.21
	2003	0	44.98 (38.21)	2.13	20.06 (42.04)	32.83
	2009	0	48.53 (35.64)	3.53	21.47 (39.2)	26.47
Pi	1998	0	21.41 (75.18)	19.19	5.66 (75.66)	53.74
	2003	0	22.62 (73.91)	18.59	5.98 (80.94)	52.81
	2009	0	23.49 (73.36)	18.68	6.19 (81.8)	51.64
Re	1998	0	18.22 (48.12)	3.78	18.44 (63.49)	59.56
	2003	0	17.82 (47.93)	3.94	17.07 (64.06)	61.16
	2009	0	18.77 (46.67)	4.61	16.72 (68.76)	59.90
RG	1998	0	5.1 (85.72)	7.22	1.19 (86.16)	86.49
	2003	0	4.66 (84.39)	5.97	1.49 (82.46)	87.88
	2009	0	5.33 (81.32)	6.42	1.98 (81.53)	86.27

Tabla 6: Granjas de selección, multiplicación y receptoras.

Leyenda:

Nucleous: Cabañas usando sus propios toros, y vendiendo sementales para reproducción; Multiplier 1: Cabañas usando sementales propios y comprados, y vendiendo sementales; Multiplier 2: Cabañas que sólo usan toros comprados y venden sementales; Commmercial 1: Cabañas usando sementales propios y comprados; Commercial 2: Cabañas que sólo usan toros comprados

Razas: Asturiana de los Valles (AV), Avileña–Negra Ibérica (ANI), Bruna dels Pirineus (BP), Morucha (Mo), Pirenaica (Pi), Retinta (Re) y Rubia Gallega (RG)

Un incremento en el número de ganaderías de multiplicación como el que se ha detectado en las razas BRUP, ASTV, AVIL y MORU, se traduce en un enriquecimiento del patrimonio genético y en una mayor variabilidad en los ganados de recepción, resultando esto clave para combatir la consanguinidad y el parentesco.

- El intervalo generacional en ASTV, AVIL BRUP, MORU y PIRE ha crecido ligeramente de 5.95 a 7.8 años para los hembras y de 4.7 a 7.6 años para los machos (Gutiérrez., *et al*. 2003), siendo considerablemente mayor en RG debido a la IA, que permite alargar la vida útil de los toros, y menor en Avileña como consecuencia de un reciente programa de selección.

- El número de descendientes por toro se ha visto incrementado en los últimos años y de igual manera el número de descendientes seleccionados, con valores que oscilan entre 8.3 (ASTV) y 44 (RGAL).
- El número de generaciones con genealogía conocida es considerablemente mayor que en otras razas más minoritarias y se ha visto considerablemente incrementado en las últimas dos décadas. Este incremento, muy probablemente, se deba a un propósito por incrementar los índices de calidad de los reproductores.

Toda esta información sirve para concienciar a los profesionales de la zootecnia de la necesidad de tener un cierto control sobre el fenómeno de consanguinidad, para asegurar una mejora en el futuro de la selección genética. Incentivar el genotipado de los animales mediante el abaratamiento de las pruebas genómicas, y en especial, promover el uso de los identificadores genéticos, resulta la mejor manera de seguir haciendo selección genética sin sufrir los efectos de la consanguinidad y la pérdida de biodiversidad.

2. Diseño experimental

El diseño experimental consta de las siguientes etapas:

1) Elección de los individuos.
2) Elección de un kit comercial de genotipado que se ajuste a nuestras muestras y acorde con el análisis que queremos efectuar.
3) Obtención del ADN y preparación de los datos genómicos.

2.1. Elección de los individuos

Para la elaboración de este estudio se han facilitado unos archivos de la base de datos genómicos de los proyectos Selgenbeef y Gene2Farm. Dado el alto coste del genotipado, en genómica es habitual aprovechar los datos para diversos estudios.

Dada la procedencia de los datos no se han podido elegir personalmente los individuos que formarán parte del estudio, aunque si se ha podido elegir el número de individuos y la relación entre los mismos, es decir, si interesaba que los individuos tuvieran un grado de parentesco o, si por en contrario se quería los individuos fueran

completamente independiente.

La utilización de estos datos en diversas investigaciones, algunas de las cuales de enorme relevancia en cuanto a valoración de los recursos genéticos nacionales, es el mejor indicador de la calidad de las muestras.

- Tamaño de la muestra

En estudios de poblaciones se suele considerar que 30 ó 40 individuos representativos son suficientes para caracterizar una población y realizar asignaciones probabilísticas de individuos problema con seguridad (Negrini., *et al*. 2008a). En este estudio se ha considerado una media de 49 individuos para cada una de las razas españolas y alguno más para las razas europeas.

Se han analizado un total de 726 muestras de 11 razas diferentes, con los siguientes animales genotipados:

FUENTE: ELABORACIÓN PROPIA

RAZA	ABREVIATURA	N.º de animales genotipados
Asturiana de los Valles	ASTV	50
Avileña-Negra ibérica	AVIL	48
Rubia Gallega	RGAL	48
Retinta	RETI	48
Bruna dels Pirineus	BRUP	50
Morucha	MORU	50
Pirenaica	PIRE	48
Simmental	SIMM	158
Flechviech	FLCH	69
Brown Swiss	BRSW	129
Guernsey	GUER	28

Tabla 7: Número de animales analizados en el estudio

De los datos se puede deducir que 342 de los individuos pertenecen a las siete razas españolas, y que 384 pertenecen a las cuatro razas europeas. Por otro lado, como en la mayoría de los estudios de genética de poblaciones, contamos con una proporción del 50% de machos y hembras. Concretamente hay 364 machos y 362 hembras. Hay un

macho de más en la raza Brown Swiss y otro en la raza Fleckvieh.

- Elección de los individuos

Se ha señalado la importancia de asumir o no que exista relación entre las muestras. Para elegir una u otra opción es necesario definir antes de que tipo de estudio se trata.

Estudios basados en poblaciones VS Estudios basados en familias

En función del objetivo del estudio y de los datos disponibles se diferencian dos tipos de estudios cuyas diferencias fundamentales son:

- En el caso de los "estudios basados en familias", las muestras se parecerán más. Este parecido recibe en estadística el nombre de "clustering" e implica una correlación entre individuos de una misma familia. Este fenómeno tiene su fundamento en la idea de que los individuos de un mismo cluster tienen un efecto ambiental parecido. Esto hace suponer que los parámetros a medir se parecerán más en individuos de un mismo cluster y que este efecto será más notorio cuanto más exclusivo sea el cluster. En términos generales, decimos que en estos casos se incumple la hipótesis de independencia. Sin embargo, existen otros casos en los que debemos desechar dicha hipótesis. Por ejemplo, deberemos desechar la hipótesis de independencia también en el caso de que estemos midiendo varias veces un mismo parámetro, llamando a lo mismo con distintos nombres, en cuyo caso podría parecernos que existe cluster donde no debería de haberlo. Cualquiera que sea el caso, en estas situaciones, debemos hacer un análisis de correlación para estimar los componentes de la varianza.

- Una segunda diferencia fundamental radica en el concepto de fase alélica, que se define como la alineación de los nucleótidos en uno sólo de los cromosomas homólogos. Ésta, habitualmente, es sólo observable en "estudios basados en familias". Esta diferencia hará que los métodos de análisis difieran de un tipo de estudio a otro.

De lo anterior se deduce que cuando queramos realizar estudios basados en familias utilizaremos, generalmente, individuos de parentesco conocido. En tal caso, tomaríamos de la base de datos individuos formando tríos, es decir, tomaríamos datos de un descendiente y de sus dos progenitores. Al final tendríamos, por tanto, dos tercios de individuos fundadores, que serían aquellos para los que se asume la hipótesis de independencia. Denominamos individuos fundadores a aquellos que, de según nuestros datos, tienen ambos padres desconocidos.

El caso contrario, es el de los "estudios basados en poblaciones". En estos casos, prescindiríamos de la información en forma de tríos y elegiríamos sólo individuos fundadores. Este es el caso del presente estudio, en el que es precisamente el parecido entre individuos de distintas familias, el que cobra una importancia mayor.
De la segunda diferencia fundamental se deduce que no podremos observar la fase alélica a menos que hagamos inferencia. Sin embargo, la fase alélica no es de especial interés en este estudio.

Para considerar nuestro estudio como un "estudio basado en poblaciones" es necesario asumir una hipótesis de independencia de las muestras mediante una estimación de la relación entre individuos.

Relación entre individuos

En este estudio vamos a asumir que los individuos no están relacionados. En caso de que dicha asunción fuese falsa, incurriríamos en una serie de errores y estaríamos haciendo un estudio basado en familias en lugar de un estudio basado en poblaciones. Desafortunadamente, en grandes poblaciones, habitualmente se da un cierto grado de parentesco y este es, generalmente, desconocido. Razón por la cual, se han desarrollado diversos métodos para determinar el grado de parentesco en grandes poblaciones, siendo GRR (*graphical representation of relationship errors*) (Abecasis., *et al*. 2001) uno de los más utilizados.

GRR se basa en la idea de que una pareja de individuos que mantienen una relación de parentesco recíproca, por ejemplo, gemelos, pares de descendientes y progenitores, o individuos no relacionados, compartirán el mismo número de alelos IBS. Estos alelos tienen la misma composición de ADN y pueden provenir o no del mismo ancestro, mientras que los alelos IBD si provienen del mismo ancestro. GRR comienza con la enumeración de los distintos alelos IBS compartidos entre los diferentes pares de individuos para cada una de las cuatro posibles relaciones y a lo largo de los distintos SNPs.

Nuestra asunción es que los individuos no están relacionados y por tanto sólo puede darse uno de los tipos posibles de relación recíproca, es decir, sólo podrán ser individuos no relacionados.

Para proceder con este análisis existen diferentes procedimientos. Quizá el más simple sea seleccionar aleatoriamente un individuo de cada familia y eliminar el resto de individuos que estén emparentados. Afortunadamente, los genotipos utilizados en este estudio se han usado previamente en otras investigaciones y se conoce la no relación significativa entre pares de individuos.

Precauciones en la elección de los individuos de la base de datos

Cuando se eligieron los animales a genotipar, se eligieron animales de diferentes ganaderías y se consultó la información del LG de cada raza, para asegurar que los individuos no estuvieran emparentados. No obstante, como ya se ha comentado, en poblaciones a gran escala es bastante habitual que exista un cierto grado de parentesco. Por tanto, se efectuaron pruebas de paternidad con microsatélites para poder asumir la hipótesis de independencia. Los microsatélites, como se ha explicado en el apartado de Introducción, son especialmente útiles para diferenciar entre individuos.

Para garantizar la homogeneidad de las muestras se han escogido animales de una edad parecida, sanos, de aspecto normal y de ambos sexos. Además, el conjunto de las las muestras se extrajo en un periodo corto de tiempo.

2.2. Elección de chip

Según algunas investigaciones previas (Reese., *et al*. 2010), considerando el genoma de la vaca (3 billones de pares de bases) para caracterizar una parte significativa de la variabilidad genética de una población, hace falta una colección de entorno a 1M de SNPs. Actualmente existen seis chips comerciales de SNPs para analizar todo el genoma de la vaca, con tamaños que oscilan entre 500 y 1000 kb y con dos opciones de plataforma: SNPs y Microsatélites. Para la elección de los loci que componen estas colecciones se utilizan criterios basados en la identificación de bloques bien definidos de desequilibrio de ligamiento.

FUENTE: /journal/v19/n3/fig_tab/ejhg2010182f1.html

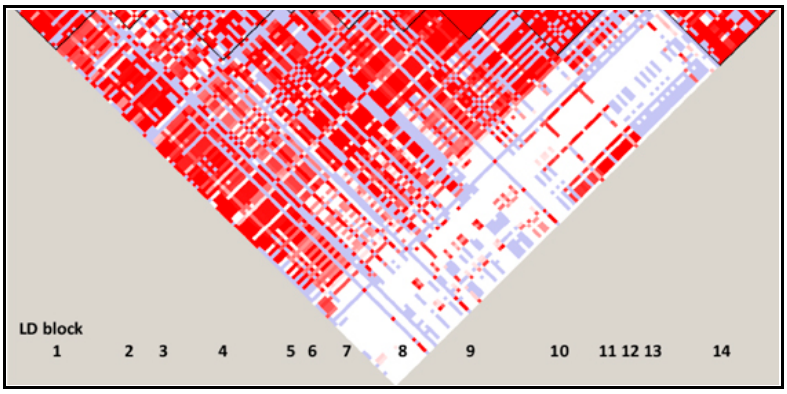


Ilustración 8. Ejemplo de diagrama de bloques de desequilibrio de ligamiento

Los datos proporcionados para este estudio se obtuvieron a partir del chip de genotipado Bovine HD (778 K) Beadchip de la empresa Illumina Inc, USA ®. (HD:

Alta densidad: *high density*). Este chip, permite leer 777962 Loci y se eligió por su adecuado tamaño para los proyectos Selgenbeef y Gene2Farm destinados a la realización de diversas investigaciones en materia de mejora genética animal en las principales especies zootécnicas. El proyecto Selgenbeef ha proporcionado los archivos con los datos genómicos de las razas españolas y el proyecto Gene2Farm para las razas europeas. En el apartado de (Introducción) se han comentado cuales son los principales ámbitos de trabajo del proyecto europeo Gene2Farm. La información del chip ha sido ya usada con éxito en las razas españolas consideradas en este estudio en una investigación llevado a cabo con microsatélites (Cañas-Álvarez., *et al.* 2014a)

Para este estudio se ha usado la información genética obtenida con este chip en forma de marcadores SNP. La información disponible corresponde a 702422 SNPs. No se disponía de información de los cromosomas sexuales ni del ADN mitocondrial.

2.3. Obtención del ADN y preparación de los datos genómicos

Aunque este estudio comienza con la manipulación de ficheros, se ha decidido incluir un breve resumen del procedimiento llevado a cabo hasta la obtención de los ficheros con la información genómica.

Para las muestras biológicas se extrajo sangre de la vena cava caudal según el procedimiento descrito en (Martínez., *et al*. 2000), (Sanz., *et al*. 2007), (Ginja., *et al*. 2010), posteriormente se extrajo el ADN y se purificó, quedando culminado lo que es el proceso de genotipado.

Posteriormente se elabora una genoteca y se secuencia el ADN. Una genoteca es una biblioteca genómica con las secuencias insertadas en vectores.

Para la secuenciación del ADN se siguió el protocolo descrito por Illumnia Inc ® en un laboratorio comercial (Xenética Fontao ®). Se describe brevemente:

- Se amplifican las muestras isotérmicamente para evitar una parcialidad alélica apreciable. En origen el tamaño de las muestras individuales es de entorno a 1000 ng.

- Después, se fragmenta mediante una encima controlada y se produce una precipitación de alcohol y una re-suspensión del ADN.

- Se templan con alguno de los loci específicos que dependen del tipo de gota. Habrá un tipo de gota por cada alelo de cada SNP).

- Después de la hibridación, la especificidad del alelo es conferida por una extensión de base enzimática.

- Después se manchan con fluorescente de distintas intensidades para posibilitar su lectura.

- Por último, se procesa la información mediante un software para un llamamiento automatizado de los genotipos.

Para la nomenclatura de los alelos se utilizó el formato estándar de proyectos de investigación en diversidad de bovino (EU RESGEN CT 98-118).

3. Análisis bioinformático

3.1. Introducción

En este estudio se parte de los siguiente datos:

- 702422 marcadores SNPs

- 726 individuos: 364 machos y 362 hembras. Todos ellos fundadores e hijos únicos, de modo que asumimos la hipótesis de independencia.

- 11 familias nucleares (o razas), a las que pertenecen los individuos.

Realizaremos un análisis bioinformático para seleccionar los SNPs que mejor permitan asignar futuras observaciones a cualquiera de las razas aquí consideradas, para ello haremos uso de tres paquetes estadísticos:

- PLINK (Purcell. 2007) v1.07 para efectuar un control de calidad en base a criterios genéticos

- R (R Core Team. 2014) v3.1.1 para la selección de los SNPs más informativos. Se han usado dos paquetes R: pls (Bjørn-Helge., *et al.* 2013) y plsgenomics (Strimmer. 2014)

- Statgraphics Centurion XVI/win ® para las representaciones cluster de las razas del estudio

3.2. Análisis genético

Software empleado - PLINK

Para el análisis genético se ha utilizado el software de libre acceso PLINK. PLINK está constituido por un conjunto de herramientas para el estudio de genomas completos. Las herramientas están diseñadas para desarrollar una larga serie de funciones y análisis a gran escala, de una manera computacionalmente eficiente. El software está integrado con otros programas de análisis genético, tales como gPLINK y Haploview, de modo que la transferencia de datos es directa.

Se resumen a continuación algunas de las tareas que puede desempeñar PLINK:

- Manejo de datos: En este estudio se hacen transformaciones entre diferentes formatos, por ejemplo, se obtienen los datos en forma binaria. Además se fusionan ficheros y se extraen subconjuntos de SNPs e individuos. El programa permite además desarrollar otras funciones, como por ejemplo, invertir el orden de las cadenas de ADN.

- Extraer resúmenes estadísticos para controles de calidad: En este estudio se obtienen frecuencias alelicas y de haplotipos, valores del estado de HWE, y se identifican los genotipos que son faltantes. Adicionalmente, PLINK permite obtener coeficientes de parentesco IBD e IBS para pares de individuos; y se pueden detectar transferencias no-mendelianas dentro de una familia y hacer comprobaciones de sexo contrastando la información de los SNPs en el cromosma-X.

- El software permite detectar estratificaciones dentro de una población. Se pueden realizar clusters de individuos según la jerarquía del ligamiento completo entre los mismos.

- Se pueden hacer pruebas básicas de asociación, como estudios de caso-control basados en pruebas de alelos estándar, diferentes tipos de tests (Test exacto de Fisher, prueba de tendencia de Cochran-Armitage o Mantel Haenszel), o modelos generales de dominancia/recesivo... También se pueden hacer pruebas de asociación para estudios basados en familias, determinar asociaciones e interacciones de caracteres cuantitativos...

- Es posible trabajar con fases alélicas o efectuar análisis del número de copias de una variante.

- Se puede hacer Meta-analysis, por ejemplo, se pueden combinar de una manera automática multitud de carpetas de resúmenes generados para millones de SNPs.

PLINK está especialmente indicado para investigaciones GWAS y nos va a permitir desarrollar completamente el análisis genético que es necesario realizar para

este estudio. Como ya se ha comentado, previamente a la selección de los mejores SNPs en base a su capacidad predictiva, es necesario realizar un análisis genético con el objetivo de eliminar los SNPs cuyas características genéticas no sean las recomendable para analizar la lectura de nuevas muestras. Con este fin, se establecen unos valores umbrales para los SNPs teniendo en cuenta los siguientes criterios:

- Criterio DL
- Criterio HWE, y errores de genotipado
- Criterio nivel de significación
- Criterio genotipo faltante

Estableceremos unos valores umbrales para cada criterio y seleccionaremos sólo los SNPs que satisfagan todos los valores umbrales. Sin embargo, con carácter previo a la aplicación de estos criterios de calidad, es necesario en la mayoría de los estudios GWAS (y este estudio no es una excepción), implementar un filtro de calidad informativa de los individuos analizados. Se identificarán los individuos cuyo genotipado no haya sido legible en más de un 10% de los SNPs. Para ello, usamos el siguiente comando:

```
--mind 0.1
```

Este valor de 0.1, que viene por defecto en el PLINK, está fijado de tal manera que sólo se eliminen los individuos con defectos graves de genotipado, permitiendo que sean realmente los filtros aplicados a los SNPs los que reduzcan el tamaño de la muestra. Eliminar demasiados individuos supondría renunciar a mucha información que si podría ser valiosa en otros loci y, en última instancia, no estaríamos amortizando correctamente el coste del genotipado de los animales.

Es importante señalar que la eliminación de los individuos mediante este filtro, se lleva a cabo con anterioridad a la aplicación de los criterios de calidad, con el objetivo de que los individuos que puedan tener defectos de genotipado no deterioren el análisis.

- **Criterio DL (desequilibrio de ligamiento)**

Para entender DL es necesario comprender antes lo que es el EL (equilibrio de ligamiento). Vamos a suponer la situación más sencilla en la que se dan 2 SNPs:

FUENTE: Devlin & Risch. 1995

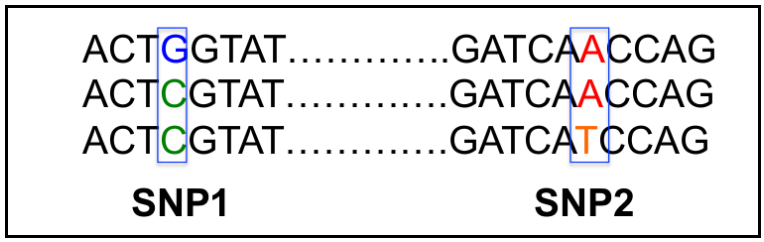


Ilustración 9: Ejemplo básico con 2 SNPs

Cada SNP tiene dos posibles alelos. En este caso el SNP1 puede tener una base citosinsa (C) o una base Guanina (G) y el SNP2 puede tener como alelos adenina (A) o tianina (T). Cada una de estas bases o alelos se dará en la población con una determinada frecuencia:

FUENTE: ELABORACIÓN PROPIA

SNP1		**SNP2**	
Alelo	**Frecuencia**	**Alelo**	**Frecuencia**
G	p1	A	q1
C	p2	T	q2

Tabla 8: Frecuencias alélicas para 2 SNPs

Como puede verse, cada alelo de cada SNP tiene una frecuencia y si tenemos 4 alelos en total, tendremos 4 frecuencias.

Un haplotipo es una colección de alelos. En el caso del ejemplo con 2 SNPs, los haplotipos estarán constituidos por dos alelos, uno por cada SNP. Los haplotipos son secuencias particulares de ADN en una agrupación de nucleótidos ligados en un cromosoma, que en principio se heredarán juntos.

En el ejemplo básico de 2 SNPs, los posibles haplotipos serán los siguientes:

FUENTE: ELABORACIÓN PROPIA

		SNP2	
	Alelo	A	T
SNP1	G	GA	GT
	C	CA	CT

Tabla 9: Posibles haplotipos dado un ejemplo con 2 SNPs

Para 2 SNPs bialélicos habrá cuatro posibles haplotipos. Las frecuencias de los haplotipos serán:

FUENTE: ELABORACIÓN PROPIA

haplotipo	frecuencia	haplotipo	frecuencia
GA	p11	**GT**	p12
CA	p21	**CT**	q22

Tabla 10: Frecuencias de los haplotipos en un ejemplo con 2 SNPs

La frecuencia de cada haplotipo dependerá de las frecuencias de los alelos que lo constituyen. En una situación de EL las frecuencias de los haplotipos corresponden a los productos de las frecuencias alélicas, tal que se cumple lo siguiente (Lewontin. 1988):

FUENTE: ELABORACIÓN PROPIA

frecuencia del haplotipo	producto de fecuencias alélicas
p11	p1q1
p12	p1q2
p21	p2q1
q22	p2q2

Tabla 11: Frecuencias de los haplotipos en la situación de equilibrio

Mientras que en la situación de DL, lo anterior se incumple y llamamos D a la diferencia con respecto a la situación de EL:

FUENTE: ELABORACIÓN PROPIA

		SNP2		
		1	2	
SNP1	1	p1q1 + D	p2q1 – D	p1
	2	p1q2 – D	p2q2 – D	p2
		q1	q2	1

Tabla 12: Frecuencia de los haplotipos en la situación de DL

Decimos que en la situación de DL la segregación es normal pero con un defecto de recombinantes.

De manera que podremos cuantificar el DL en función de esa diferencia de frecuencias. Las frecuencias de los haplotipos son observables, mientras que las frecuencias alélicas nos proporcionan los valores esperables bajo condiciones de equilibrio:

$$p_{11}\, p_{22} = (p_1\, q_1 + D\,)(p_2\, q_2 + D\,)$$

$$p_{12}\, p_{21} = (p_1\, q_2 - D\,)(p_2\, q_1 - D\,)$$

De lo que se deduce que:

$$D = p_{11}\, p_{22} - p_{12}\, p_{21}$$

En el caso de que para 2 SNPs las 2 frecuencias alélicas sean idénticas: $p_1=q_1=0.5$; $p_2=q_2=0.5$, diremos que los SNPs están en EL, puesto que:

$$P_{11} = p_1q_1 = 0.5 \text{ x } 0.5 = 0.25$$

$$P_{22} = p_2q_2 = 0.5 \text{ x } 0.5 = 0.25$$

$$P_{12} = p_1q_2 = 0.5 \text{ x } 0.5 = 0.25$$

$$P_{21} = p_2q_1 = 0.5 \text{ x } 0.5 = 0.25$$

$$D = (P_{11})(P_{22}) - (P_{12})(P_{21})$$

$$D = (0.25)\ (0.25) - (0.25)\ (0.25) = 0$$

Mientras que no se cumple la situación de EL, cuando:

$$P_{11} = p_1q_1 + D = 0.25 + D = 0.5$$

$$P_{22} = p_2q_2 + D = 0.25 + D = 0.5$$

$$P_{12} = p_1q_2 - D = 0.25 - D = 0$$

$$P_{21} = p_2q_1 - D = 0.25 - D = 0$$

$$D = (P_{11})(P_{22}) - (P_{12})(P_{21})$$

$$D = (0.5)\ (0.5) - (0)\ (0) = 0.25$$

En este caso, en la población sólo se darán los haplotipos homocigóticos. Sin embargo, tal y como se está planteando, ese valor D, de desequilibrio de ligamento podría tomar valores negativos. Por este motivo se hace una estandarización del valor, que llamamos D´. Para obtener D´ simplemente aplicamos:

$$D' = D / D_{max}$$

Donde:

$$D_{max}= \min [(p_1q_2) \text{ ó } (p_2q_1)]$$

Cuando D>0

$$D_{max}= \min [(p1q1) \text{ ó } (p2q2)]$$

Cuando D<0

Es decir, se toma en cada caso el producto menor de las frecuencias alélicas. De esta manera, D´ tomará valores entre 0 y 1. Sin embargo, el parámetro más utilizado para cuantificar el grado de DL es el coeficiente de correlación de Pearson (r), o más habitualmente, el cuadrado del mismo, para que el valor no se vea afectado por la arbitrariedad de signos.

Este coeficiente de correlación entre dos loci (r^2) dependerá de:

- La diferencia entre las frecuencias de los haplotipos y las frecuencia alélicas: $D_{A1B1} = frec(A_1B_1) - p_{A1}p_{B1}$

- El producto de todas las frecuencias alélicas para ese loci $p_Aq_ap_Bq_b$

$$r^2=\frac{D^2}{p_A q_a p_B q_b}$$

En estudios de genética, resulta conveniente comprobar el grado de significación de este desequilibrio. Para ello se estima chi-cuadrado (χ^2).

χ^2 es una distribución de probabilidad continua que está en función de un parámetro k, que depende de los grados de libertad. El grado de libertad de un conjunto de observaciones viene dado por el número de valores que pueden ser asignados de manera arbitraria. La distribución chi-cuadrado viene a ser la distribución de las desviaciones estándar de la muestra. Es decir, para obtener la distribución chi-cuadrado se extraerían todas las muestras posibles de una población normal y a cada muestra se le calcularía su varianza. El estadístico también puede venir dado por la siguiente expresión.

$$\chi^2 =\Sigma(obs-exp)^2 / exp$$

Esta expresión nos permite, en genética, determinar el grado de significación del desequilibrio. En ella, el valor esperado “exp” (expected) corresponde a una asociación aleatoria de los alelos.

También es habitual utilizar el coeficiente de correlación de Pearson para determinar esa significación:

$$\chi^2 = r^2 N$$

, donde N es el el número de cromosomas.

En las regiones en EL, los SNPs se heredarán por bloques, y por tanto, proporcionarán información redundante. Es decir, conociendo un SNP del bloque se conoce el resto. Para evita esta información redundante se utiliza *Snp tagging*, que consiste en seleccionar los SNPs que estén en perfecto DL, ($r^2 = 0$).

La recombinación es un intercambio genético entre cromosomas homólogos que impide que podamos asumir asociación perfecta entre diferentes puntos de un cromosoma. Cuando dos SNPs se encuentran en la situación de EL completo, decimos que no se han separado por recombinación y se cumple que: $D' = 1$ y $R^2 = 1$

Para conocer la tasa de recombinación entre dos loci tenemos en cuenta que, cuanto más alejados se encuentren dentro de un cromosoma, más susceptibles serán de sufrir el fenómeno de recombinación. Sin embargo, también hay que tener en cuenta que algunas regiones del cromosoma son más sensibles que otras.

Crterio DL - Análisis bioinformático

Para el análisis bioinformático, PLINK calcula el grado de ligamiento para cada par de loci. Con valores de r próximos a 1, asumimos que ambos loci están ligados, puesto que su tasa de recombinación es muy reducida. Por tanto, y puesto que queremos evitar la estructura en bloque de los datos, seleccionaremos los loci que tengan valores bajos de r^2 en todas las posibles parejas. Sin embargo, en esta herramienta de PLINK no se establece un valor r^2, sino que se fija un valor umbral VIF (*variance inflation factor*), que es inversamente proporcional a r^2. La relación entre ambos parámetros se expone más adelante en este mismo apartado. El término VIF se utiliza en estadística para cuantificar el grado de multicolinealidad en análisis de regresión por mínimos cuadrados ordinarios y es un indicativo de cuánto aumenta la varianza de un modelo como consecuencia de la colinealidad.

Estableceremos un valor máximo de VIF que se corresponderá con un valor mínimo de r^2, y eliminaremos los SNPs cuyo valor VIF quede por encima. Por tanto, conforme más alto fijemos dicho valor umbral VIF, más permisivos estaremos siendo. Visto de otra forma, habrá un mayor número de SNPs que consideremos independientes, y que cumplen que tienen una r^2 lo suficientemente alejada de 1.

Dada la enorme cantidad de loci, sería computacionalmente muy costoso determinar la r^2 para cada una de las posibles parejas, y por tanto, el criterio se aplica sobre ventanas de SNPs. Dichas ventanas se irán desplazando ordenadamente,

permitiendo una eliminación recursiva de los SNPs.

Utilizamos el siguiente comando:

```
--indep 50 5 2
```

Como vemos, al establecer el valor umbral se fijan tres parámetros:

- 1º parámetro (50): Tamaño de la ventana desplazable, expresado en número de SNPs consecutivos. Se ha fijado un valor de 50. Con este valor se garantiza que los SNPs situados más próximos guarden un mínimo de independencia. Recordemos que los SNPs más próximos son más propensos a formar bloques de ligamiento. Para completar el análisis y aplicar el criterio a toda la colección de SNPs distribuidos a lo largo del genoma, la ventana se irá desplazando recursivamente con un desfase que se define con el 2º parámetro del comando.

  Un tamaño excesivo de la ventana supondría que se eliminarían demasiados SNPs, algunos de los cuales podrían ser muy útiles de cara a la predicción.

- 2º parámetro (5): Este parámetro indica el desfase de la ventana recursiva medido en número de SNPs. Se ha fijado un valor de 5. Este valor significa que, considerando una ventana, la siguiente empezará 5 SNPs más allá de lo que empezó la primera. Entendemos por "más allá" al sentido con el que nos vamos alejando de lo que consideramos el primer SNP del cromosoma. Con un desfase de 5, se considerarán 10 ventanas antes de abandonar definitivamente la ventana inicial y empezar en la siguiente.

- 3º parámetro (2): El valor se corresponde con el umbral VIF, que queda determinado por la siguiente expresión:

$$VIF = \frac{1}{1 - R^2}$$

, donde R^2 corresponde a la r^2 definida con anterioridad en este apartado.

Este valor puede oscilar entre 1 y 200, que se corresponden, respectivamente, con una R^2 de 0 (independencia completa de cada SNP con el resto) y de 1 (dependencia completa). Por tanto, conforme más bajo sea el valor VIF más SNPs eliminaremos.

PLINK recomienda establecer valores umbrales VIF entre 1.5 y 2. Si el umbral se fijara demasiado bajo, se eliminarían demasiados SNPs, y no estaríamos

conservando necesariamente los que son mejores predictores. Por otro lado, si fuéramos excesivamente permisivos y fijáramos un valor umbral excesivamente alto, se eliminarían muy pocos SNPs. Para este estudio se ha fijado un valor VIF de 2, que se corresponde con una R^2 de 0.5. Obsérvese que este valor es el máximo dentro de los recomendados por PLINK y la razón es que se ha optado por elegir un valor VIF medio-alto, ya que pretendemos que la causa fundamental de eliminación de los SNPs no sea su nivel de independencia, sino su capacidad predictiva.

Una R^2 de 0.5 y un tamaño de ventana de 50 SNPs garantiza una adecuada eliminación de los SNPs ligados, al mismo tiempo que se asegura la conservación de los SNPs más válidos.

El margen de valores recomendados por PLINK varía, lógicamente, en función del número de variables. Si el número de variables consideradas es muy elevado, como es nuestro caso puesto que analizamos un genoma completo, deberíamos ser más exigentes. Sin embargo, en este estudio, con posterioridad al análisis genético, vamos a aplicar otro análisis estadístico y nos conviene ser más exigentes en este último. Por otro lado, el valor umbral de la R2 no podrá ser muy elevado puesto que dispondríamos de muchos SNPs para el siguiente análisis, y el programa estadístico que se utiliza para este último análisis tiene un tope de admisión de variables. Sabemos que dicho tope está entorno a las 100 000 variables.

Este filtro es el que va a eliminar una mayor proporción de SNPs y es el único que no se ejecuta simultáneamente al resto, sino que se aplica antes. El software creará dos archivos, uno con los SNPs a conservar y otro con los SNPs a eliminar. Aplicaremos el resto de filtros sobre el archivo de SNPs conservados. Eliminar los SNPs que tengan un fuerte EL permitirá reducir en gran medida el problema de la multicolinealidad (dependencia estadística de los predictores).

- **Criterio HWE (*Hardy-Weinberg equilibrium*) y errores de genotipado**

Decimos que se da HWE en poblaciones que cumplen unas determinadas condiciones, cuando tras una generación de apareamiento aleatorio, las frecuencias de los distintos genotipos (AA, Aa y aa) se mantienen constantes a lo largo de las generaciones y son las esperables según:

$$P = frec(AA) = p^2$$
$$H = frec(Aa) = 2pq$$
$$Q = frec(aa) = q^2$$

Esas condiciones son: los individuos son diploides, la población es grande, el cruzamiento es al azar, las generaciones no se solapan, la población está aislada, los

genes no tienen mutaciones, no existe selección y la frecuencias alélicas son independientes del sexo.

Efectivamente, estas condiciones son muy exigentes y HWE sólo se cumple en una situación idílica. Sin embargo, conocer el valor de HWE nos permite saber que loci son más susceptibles de tener un cambio en sus frecuencias alélicas. En este estudio nos interesa seleccionar los nucleótidos cuyas frecuencias alélicas sean poco cambiantes de una generación a otra.

HWE y Errores de genotipado

Determinar la situación de HWE nos permite detectar errores de genotipado. Llamamos error de genotipado a una desviación entre el genotipo real y el genotipo observado tras la secuenciación. Estos errores aparecen con una frecuencia cambiante en las diferentes plataformas tecnológicas y pueden deberse a diferentes razones. El procedimiento más habitual para la identificación de estos errores es el cálculo de HWE para cada uno de los SNPs estudiados. Esto se hace mediante el test de Chi-Cuadrado o el test de Fisher en estudios de asociación.

En investigaciones como las del presente estudio, la desviación de HWE entre individuos que pertenecen a una u otra categoría, puede deberse a una asociación entre los genotipos y la pertenencia a la categoría correspondiente. Dada esta posibilidad, algunos investigadores apuestan por calcular HWE para cada una de las categorías por separado. Si bien, esto puede llevar a malinterpretaciones, puesto que podríamos encontrarnos en situaciones en las que para los individuos de una categoría no se diera HWE, pero si se diera HWE al considerar todas las categorías a un mismo tiempo. Para solucionar esto se realizan pruebas de bondad de ajuste, mediante las que se identifica el modelo más probable de pertenencia a una categoría. Esto permite distinguir entre las desviaciones de HWE que se deben a una asociación con una categoría y las que, por el contrario, se deben a otro fenómeno como, por ejemplo, errores de genotipado.

Al hablar de HWE, debemos de tener en cuenta que una desviación de HWE posiblemente se deba a la existencia de subestructuras dentro de la población. En dicho caso, al rechazar los SNPs por el simple hecho de presentar desviaciones en HWE, podría darse el caso de que estuviéramos renunciando precisamente a los SNPs que permiten una detección de dichas subestructuras. Por este motivo resulta recomendable no basar la detección de los errores de genotipado únicamente en pruebas de HWE, sino contrastar también esta información con repeticiones de los genotipados. En la práctica, sin embargo, debido al coste de genotipado no es habitual seguir esta recomendación.

Este estudio tiene la ventaja de que los datos genómicos ya se han utilizado en otros estudios y se ha comprobado que no existe una asociación importante entre las desviaciones de HWE y la pertenencia a las categorías (razas). No obstante, se ha optado por ser poco exigentes en cuanto a lo que “desviación de HWE” se refiere y se van a eliminar únicamente los SNPs con desviaciones más significativas, e

interpretaremos que en dichos SNPs se han producido errores de genotipado.

Para calcular HWE para cada loci, PLINK utiliza Chi-Cuadrado de Pearson:

$$\chi^2 = \sum_{(i,j) \epsilon C} \frac{(Oij - Eij)^2}{E_{ij}}$$

, donde para cada individuo no-fundador:

Oij: Son los valores observados en cada loci

Eij: Son los valores esperados en cada loci

Para eliminar los SNPs que más fallan a HWE, se ha fijado el valor umbral de 0.001, que significa que se conservarán los SNPs cuyo P-valor para la hipótesis de igualdad Chi-Cuadrado de Pearson sea menor o igual a 0.001 ($P_{valor} \leq 0.001$). El valor fijado se corresponde con el valor que PLINK establece por defecto. Se utiliza el comando:

```
--hwe 0.001
```

Por definición, este criterio sólo se aplica sobre los individuos fundadores que, en nuestro caso, son el conjunto de las observaciones.

- **Criterio Nivel de significación**

Para resolver el problema de la asignación de individuos, los SNPs más informativos son aquellos cuyas frecuencias alélicas estén próxima al 50%. Si bien, tenemos diferentes razas, y fijar una frecuencia alélica próxima al 50% supondría renunciar, precisamente, a los SNPs que son de un mayor interés por tener un alelo ligado a, por ejemplo, una sóla de las raza. De manera que será la raza menos numerosa en nuestros datos, la raza Guernsey con 28 individuos (un 3.8 % del total) la que va a determinar el nivel de significación.

Fijando un nivel de significación maf < 0,038 (maf : *minor allele frecuency*), garantizaremos que se conserve cualquier SNPs que pudiera tener una forma alélica exclusiva de cualquiera de las razas del estudio. Por otra parte, nos interesa que el valor umbral sea alto para así eliminar los SNPs menos informativos. Utilizamos:

--maf 0.03

El comando hará que se conserven los SNPs cuyo alelo menos común se dé, al menos en un 3% de la población.

Por defecto, PLINK fija un valor de 0.01. No obstante, dada la gran cantidad de SNPs de que se dispone (ya que analizamos un genoma completo), conviene ser más rigurosos, primero para reducir el número de predictores y segundo para que el subconjunto seleccionado sea más informativo. Este filtro sólo se aplica sobre los individuos fundadores. En nuestro caso, sobre toda la población.

- **Criterio genotipo faltante**

También eliminaremos los SNPs cuya lectura no haya sido posible en un cierto número de individuos.

Este criterio, que en principio se define para identificar los SNPs más informativos, es en nuestro caso de especial interés puesto que no nos interesa disponer de valores desconocidos (valores <NA> en PLINK), puesto en el análisis estadístico que queremos hacer posteriormente, las variables quedarán definidas como vectores. Por tanto, prescindiremos de cualquier SNP que nos proporcione cualquier información desconocida. Aplicamos:

--geno 0.0

En otras palabras, la información de cada SNP puede ser desconocida en un 0% de las observaciones de que disponemos.

Una vez comentados los criterios que se van a tener en cuenta en este estudio, conviene comentar un 5º criterio que, aunque por diferentes razones no se considera en este estudio, es tan habitual como los anteriores en la mayoría de estudios GWAS.

- **Otros criterios no considerados en el análisis**

Error Mendeliano

Se utiliza el término *error de Mendel* para referirse a alelos que no se corresponden con ninguna de las formas alélicas de sus parentales. Es decir, el individuo no lo recibe por herencia. Por definición, este criterio sólo puede ser considerado en aquellos casos en que se analizan familias de individuos. No es el caso en el presente estudio.

Ante estos casos pueden ocurrir dos cosas, que la lectura de alguno de los parentales sea incorrecta, o que por el contrario, se trate de una mutación en ese alelo. Una manera de comprobar si, efectivamente, se trata de un error de Mendel es haciendo todas las combinaciones posibles de genotipos y comprobando que, al menos una, incumple la ley de Mendel.

En estudios con datos de más de una generación de individuos, podríamos usar PLINK para saber cuando los datos superan cierto porcentaje de error Mendeliano. En tal caso podríamos analizar por un lado las familias y por otro los SNPs.

- **Acerca de los criterios considerados en el estudio**

De los cuatro criterios a tener en cuenta en nuestro análisis de los SNPs, los 3 últimos (HWE, nivel de significación y genotipo faltante) se aplican de manera simultánea, mientras que DL se aplica con anterioridad, ya que PLINK creará dos archivos: uno para los SNPs eliminados y otro para los SNPs seleccionados.

La razón por la que PLINK trabaja creando dos archivos en el caso de este criterio, en lugar de uno sólo como en el resto de criterios, tiene que ver, por un lado con los distintos objetivos que puedan tener los diversos estudios GWAS que recurren a PLINK para hacer análisis genéticos; y por otro, con el gran número de SNPs que, generalmente, se eliminan en base a este criterio (habitualmente más de un 80% del total). Para el resto de los criterios, aplicar los filtros de manera simultánea tiene la ventaja de que no importará el orden de aplicación de los mismos, y al final del análisis conoceremos el número real de SNPs que fueron eliminados en base a cada criterio en particular, con independencia de que esos SNPs hayan superado o no los otros criterios. De lo anterior se deduce que un mismo SNP podría ser eliminado en más de uno de los filtros y, por tanto, no debería de sorprendernos que al sumar el total de SNPs eliminados y los SNPs seleccionados al final, el resultado fuese superior al número de SNPs originales, ya que algunos SNPs se contabilizan como eliminados en más de un filtro.

La aplicación de estos filtros permite garantizar que los SNPs que vamos a considerar como predictores cumplan lo siguiente:

- Criterio de HWE: Podemos suponer que los loci son heredables. Con lo cual, podremos realizar la prueba de igual manera en futuras generaciones. Además, conseguimos dejar fuera de la selección los SNPs para los que se haya podido producir un error de genotipado.

- Criterio de DL: En el caso de utilizar los datos de las lecturas de los SNPs como predictores en un modelo estadístico, al poner un límite al estado de ligamento, reduciremos considerablemente el problema de la varianza, que podría ocasionar una sobre-estimación de pequeñas fluctuaciones en los datos de dicho modelo.

- Criterio de nivel de significación: Eliminaremos una serie de SNPs que, por la baja frecuencia de su alelo más infrecuente serían poco informativos
- Criterio de genotipo faltante: Eliminaremos los SNPs que podrían proporcionarnos algún valor desconocido <NA>.

Por otra parte, el análisis genético permitirá comprobar la no presencia de individuos outiers. En nuestro caso, serían outliers las observaciones que para una serie de predictores tomen valores muy diferentes al resto de las observaciones.

- **Comandos utilizados en este análisis**

Para adaptar nuestros datos al análisis en PLINK ha sido necesario tener en cuenta algunos aspectos.

En todas las líneas de comandos utilizamos el comando *--cow.* Puesto que PLINK está enfocado fundamentalmente a investigaciones de genética humana. Deberemos especificar que nuestros datos corresponden, no a 23 pares de cromosomas, sino a 29. Para ello utilizamos el comando *--cow.* Éste es una abreviatura del comando *--chr-set 29 no-xy*, que especifica que el número de pares de autosumas diploides es de 29. Este número se pondría en negativo si los cromomas correspondiesen a especies haploides. El comando también sirve para especificar que no se consideran los cromosomas no autosómicos, es decir, los cromsomas sexuales (cromosoma-X y cromosoma-Y), ni tampoco se considera el ADN mitocondrial.

En las diferentes líneas de comandos utilizaremos los siguientes 3 comandos:

--recode

Este comando es necesario para escribir un archivo nuevo cuando éste tiene una estructura diferente a la anterior. Este comando está diseñado para trabajar con SNPs.

--recode12

Se asignarán los valores “1” y “2” a los alelos. Concretamente se asigna “1” al alelo más frecuente y “2” al alelo más infrecuente. No es necesario introducir el primer comando (*--recode*) cuando se usa (*--recode12*).

--tab

Con este comando se separará en el fichero de datos, mediante una tabulación, los valores asignados a los dos alelos correspondientes a un mismo locus. Esto facilita la lectura del genotipo, que podrá hacerse considerando una columna por cada loci o una columna por cada alelo.

Los comandos utilizados para este análisis son los siguientes:

Comando I

```
p-link --file nobin1crom1 --merge-list 00025-2list.txt --cow
--recode --out 3abrilm
```

Que se interpretaría como:

> Fusionamos todos los archivos .ped y .map que figuran en la lista "*00025-2list.tx*t", con los archivos "*nobin1crom1.ped*" y "*nobin1crom1.map*". Consideramos que hay 29 cromosomas y que no hay cromosomas sexuales ni ADN mitocondrial. Creamos un archivo de nombre "*3abrilm*". La lista *00025-2list.tx*t es un archivo de texto que contiene los nombres de todos los archivos *.ped* y *.map* ordenados en 2 columnas.

Comando II

```
p-link --file 3abril --mind 0.1 --out 3abrilm
```

Se interpretaría de la siguiente forma:

> Cogemos el archivo de nombre "*3abril*" y eliminamos los individuos cuyo genotipo falta en más de un 10% de los SNPs.
>
> Creamos un archivo de nombre "*3abrilm*"

Comando III

```
p-link --file 3abril --cow –indep 50 5 2
```

Cogemos el archivo de nombre "*3abrilm*" y diferenciamos entre SNPs dependientes e independientes (R^2<0.5). Consideramos que hay 29 cromosomas, y que no hay cromosomas sexuales. Creamos dos archivos con los nombres: plink.prune.in (SNPs dependientes) y plink.prune.out (SNPs independientes).

Comando IV

```
p-link --file 3abrilm --cow --extract plink.prune.in --geno 0.0 --maf 0.03 --hwe 0.001 --recode12 --tab --out 3abrilDone
```

Cogemos el archivo "*3abrilm*", consideramos los 29 autosumas, extraemos los SNPs que figuran en el archivo plink.prune.in, y eliminamos los SNPs que son faltantes en algún individuo, los que tienen una frecuencia del alelo más infrecuente inferior al 3% y los que incumplen la hipótesis de HWE con un p-valor inferior o igual a 0.001. Recodificamos los alelos con 1 y 2, separamos los valores asignados a los dos alelos de un mismo locus mediante una tabulación, y creamos un archivo de nombre "*3abrilDone*".

3.3. Selección en base a capacidad predictiva

3.3.1. Planteamiento del modelo

Una vez hemos aplicado los filtros en base a las características genéticas, disponemos de datos de 87918 SNPs. Nuestra intención ahora es seleccionar los mejores SNPs en base a su capacidad predictiva para el carácter población. Para ello, construimos un modelo estadístico en el que los datos genómicos con la información de los SNPs son visualizados como un subconjunto de coordenadas es un espacio multidimensional, en el que cada variable o SNP es un eje. De modo que tenemos una única VD (variable dependiente) que es la población de los individuos, a la que llamamos "*pobla*". Ésta es de tipo categórico con 12 posibles categorías, una por cada raza y otra para individuos que no pertenezcan a ninguna de las razas consideradas en el estudio. Pretendemos estimar "*pobla*" a partir de un subconjunto muy grande de VI´s (variables independientes) que son los datos genómicos.

Las VI´s son de tipo numérico, ya que vamos a suponer un modelo aditivo. En

un modelo aditivo tener 2 copias de la variante de referencia implica tener el doble de probabilidad de ser considerado como “positivo” o, en este caso, de ser considerado como “perteneciente” a una de las razas. En concordancia con la mayoría de investigaciones GWAS, se ha acordado que la variante de referencia en cada uno de los alelos sea la más frecuente de las dos posibles. Recordemos que, puesto que analizamos a nivel de nucleótido, cada SNP tiene sólo dos posibles variantes, bien una de las dos bases purina (adenina o guanina) o bien una de las dos bases pirimidina (timina o citosina).

Disponemos de 726 observaciones (n), una por cada animal genotipado. Cada observación nos aporta datos tanto de la VD como de VI´s. Puesto que conocemos el origen de los datos y estos han sido previamente analizados, sabemos que en todas las observaciones los datos son válidos para efectuar una regresión.

Para poder resolver como en GWAS con la metodología de caso/control, descomponemos nuestra VD “pobla” en una serie variables dummy. Las variables dummy, puesto que son de tipo binario, son útiles para simplificar la respuesta. Cada variable dummy irá asociada a una raza y tendrá dos posibles respuestas: la muestra pertenece (asignamos valor 1) o no pertenece (asignamos valor 0). Por tanto, harán falta tantas variables dummy como categorías haya menos una, es decir 11. Recordamos que una de las categorías de la VD quedaba reservada para individuos que no perteneciesen a ninguna de las razas consideradas.

A modo de resumen, nuestro modelo estará constituido por las siguientes variables: una VD “pobla” (o 11 variables dummy), y 87918 VI´s (p) de tipo numérico con tres posibles valores (0, 1 ó 2 copias del alelo más frecuente).

De lo que se deduce:

- Tenemos que buscar un método estadístico de análisis multivariante , puesto que tenemos varias VI´s.
- Encontramos dos tipos de datos: los datos genómicos son de tipo numérico y la información de pertenencia a las razas viene dada en forma de datos categóricos.
- Tenemos un sólo tipo de datos en las VI´s y un sólo tipo de datos en la VD.

Si pretendiéramos resolver por MLR (regresión lineal múltiple), el modelo adquiriría la siguiente forma:

$$Y = BA + E$$

, donde Y sería la variable dependiente, A la matriz de predictores (SNPs en nuestro caso), B la matriz de coeficientes de regresión, y E los errores o residuos.

Aclaración: En determinadas expresiones, la matriz A se expresa como matriz X.

Para estandarizar los datos, se substrae la media a cada uno de los valores de la matriz A. Después, aplicaríamos mínimos cuadrados para estimar los coeficientes de regresión.

$$\hat{B} = (A^T A)^{-1} A^T Y$$

Puesto que los datos están estandarizados, se cumple que:

$$A^T A = R$$

, donde R es la matriz de covarianzas de las variables independientes. Si bien, la complejidad de nuestro modelo nos impide resolver mediante MLR.

3.3.2. Definición del problema

Dadas las característica de nuestros datos se perciben cinco dificultades fundamentales en nuestro modelo:

i. El número de VI´s es muy elevado en comparación con el número de observaciones (n << p): 726 << 87918 . Lo cual impide resolver directamente la ecuación de regresión, al tener más incógnitas que ecuaciones.

ii. El número de VI´s es muy elevado, lo cual podría provocar un sobreajuste del modelo o “memorización de la máquina”.

iii. Dadas las características de los datos genómicos y puesto que los loci se asocian en bloques, podemos asumir que va a haber multicolinealidad entre las VI´s.

iv. Hay abundancia de ruido. En estadística se dice que en un modelo hay ruido cuando una parte considerable de las variables a tener en cuenta no aportan nada de información o aportan información irrelevante para la variable respuesta.

v. La respuesta, muy probablemente dependa de patrones complejos y cada uno de los factores debería de ser considerado en el análisis para lograr una acertada interpretación del modelo.

A continuación se explican los problemas que estas condiciones generan en nuestro modelo:

- **Muticolinealidad**

Decimos que un modelo presenta multicolinealidad cuando dos o más variables predictoras estén correlacionadas, de manera que una puede ser predicha a partir de la otra con un grado de acierto no trivial.

La multicolinealidad no impedirá que, al hacer regresión múltiple podamos predecir la variable respuesta, pero resultará imposible saber cuáles son efectivamente los predictores que afectan más a la respuesta, puesto que no estaremos estimando la capacidad explicativa de una variable, sino de un grupo de variables.

Cuando hablemos de multicolinealidad, debemos poner especial cuidado y no confundir este concepto con otro muy parecido, que es la colinealidad.

Decimos que hay colinealidad cuando la relación entre dos o más variables es la misma para todas las observaciones. Es decir, dos variables tendrán colinealidad si, al representar nuestros datos, en ambos casos se obtiene la misma separación entre las observaciones es exactamente la misma. Obviamente, esto implica que las variables estén correlacionadas y el coeficiente de correlación de Pearson será próximo a 1. El coeficiente de correlación de Pearson viene dado por la expresión:

$$\rho_{x,y} = \frac{\sigma xy}{\sigma x \sigma y}$$

, donde:

σxy es la covarianza entre x e y

σx, y σy son las varianzas de x e y, respectivamente

Imaginemos dos predictores "x_1" y "x_2".. Diremos que ambas variables tienen una colinealidad perfecta si para todas las observaciones existen unos parámetros λ_0 y λ_1, tal que:

$$x_{2i} = \lambda_0 + \lambda_1 x_{1i}$$

Mientras que diremos que en un modelo hay multicolinealidad cuando haya dos o más variables altamente relacionadas linealmente, con correlaciones de 1 ó -1. Matemáticamente, un grupo de variables son perfectamente multicolineares si existen una o más relaciones lineales, tal que:

$$\lambda_0 + \lambda_1 x_{1i} + ... + \lambda_k x_{ki} = 0$$

Donde λ son constantes con valores distintos de 0, y x_{ji} es la i observación de la variable j. Podemos explorar el efecto causado por la multicolinealidad estimando:

$$Y = B_o + B_1 X_{1i} + \ldots + B_k X_{ki} + E_i$$

Para hacer mínimos cuadrados es necesario invertir la matriz X para obtener X^TX.

Si hay una relación perfecta entre las variables predictoras, entonces el rango de X, y por tanto el de X^TX, será menor que k+1; y X^TX no será invertible, lo que supone que no puedan aplicarse mínimos cuadrados ordinarios. Recordemos que el rango de una matriz es la colección más larga de variables linealmente independientes y que k es el número de predictores.

Sin embargo, los casos de multicolinealidad perfecta no son habituales y más frecuentemente nos encontraremos con:

$$\lambda_0 + \lambda_1 x_{1i} + ... + \lambda_k x_{ki} + \vec{u}_1 = 0$$

En este caso no hay una relación perfecta entre las variables, aunque si la varianza de $\vec{u}_1$ es pequeña para algún grupo de λ, las variables serán perfectamente multicolineales.

La matriz X^TX tiene una inversa, pero esta está igualmente condicionada y no todos los algoritmos computacionales podrán calcularla. Es más, aun calculando la inversa, ésta será muy sensible a ligeras variaciones de los datos, debido a una magnificación de los errores y la transformación matricial podría ser poco exacta.

La multicolinealidad puede tener los siguientes efectos sobre el modelo:

- Al estar x_1 , x_2 correlacionadas tendremos en los datos de las observaciones una relación lineal estocástica. No tenemos datos en las observaciones para los que los cambios en x_1 no afecten a x_2, así que estimamos erróneamente los cambios en x_1 . Decimos que hay redundancia de los datos.
- En caso de producirse cambios pequeños en los datos los efectos sobre los coeficientes de regresión podrían ser exagerados.

- Resulta difícil separar los efectos de dos predictores fuertemente correlacionados. Al hacer el análisis estadístico y comprobar el efecto de una variable particular, fijamos los valores del resto de variables. Sin embargo, si la correlación de la variable es muy grande con alguna de las variables que hemos fijado, estaremos infravalorando el valor de la primera, puesto que de no haber fijado la otra variable, esta habría provocado un efecto añadido sobre la primera variable. En otras palabras, al fijar el resto de variables para probar el efecto de un predictor concreto estamos suprimiendo una interacción (varianza conjunta) que se daría de forma natural.

- Por otro lado, cuando una proporción muy grande de los predictores están correlacionados, o en el caso de que sean prácticamente las mismas variables con distinto nombre, no seremos capaces de valorar adecuadamente las diferencias entre los individuos, puesto que a menudo parecerán más semejantes de lo que en verdad deberían. Es más, podríamos agrupar incorrectamente a los individuos si, por ejemplo, dos individuos muy parecidos, sin embargo, divergieran en alguna de las variables redundantes.

- Además, favorece el sobreajuste en los modelos de regresión, como veremos a continuación.

De entre los distintos efectos, se deduce que el impacto fundamental de la multicolinealidad es que se reduce el tamaño efectivo de las poblaciones, y con ello, la potencia predictiva. Dichos efectos provocan alteraciones en los coeficientes β (pendiente estandarizada) y b (pendiente no estandarizada). Estos permanecerán insesgados pero serán menos acertados cuando haya multicolinealidad.

La multicolinealidad afecta además a los intervalos de confianza que, para ser significativos, tendrán que ser más amplios. Sin embargo, este fenómeno no sesga los resultados ni afecta a la R^2 o a los p-valores, sino que genera un error estándar importante de las variables independientes relacionadas. El objetivo fundamental de la regresión es obtener unos coeficientes que nos permitan extrapolar los resultados a nuevas muestras, pero si en las nuevas muestras la relación entre variables se ha visto alterada, la predicción vendrá acompañada de errores.

Hay que tener en cuenta que al filtrar los SNPs en base a su valor VIF se ha reducido ya en mucho la multicolinealidad.

Conviene aclarar que si la correlación entre los predictores en dos observaciones no se correspondiese con la correlación entre las variables respuesta, este no sería un problema de multicolinealidad, sino un error en los datos.

➢ Sobreajuste

En regresión, cuando queremos predecir nuevos valores, el sobreajuste es el efecto de

sobreentrenar un algoritmo de aprendizaje, y ocurre al emplear en el diseño del modelo un exceso de datos de resultado conocido. Podemos detectar un sobreajuste en el modelo cuando éste muestre un error aleatorio, o bien, como en el caso del estudio, cuando haya abundancia de ruido.

Esta circunstancia es habitual en modelos complejos, por ejemplo, cuando el número de parámetros supera en mucho al número de observaciones. El sobreajuste causa, generalmente, defectos de predicción puesto que tenderá a exagerar cualquier fluctuación de los datos por pequeña que sea. Al darse una alta colinealidad de los datos, podríamos considerar varias veces un mismo efecto.

Decimos que un modelo presenta una alta probabilidad de estar sobreajustado cuando el criterio usado para entrenar el modelo es distinto del empleado para estudiar su predicción. Además, sabemos que aunque el modelo se entrena para conseguir una máxima eficacia de predicción en la fase entrenamiento, es cuando se predicen observaciones ajenas al modelo cuando realmente podemos estimar la capacidad predictiva.

Habrá sobreajuste cuando el modelo en lugar de “aprender” estimando la tendencia de la fase de entrenamiento, “memorice” las soluciones. Esto generalmente, empieza a ocurrir desde el momento en que el número de parámetro supere al de observaciones, aunque puede ocurrir incluso antes. Memorizar supondría fallar drásticamente al intentar predecir a partir de datos no usados en la construcción del modelo.

De lo anterior se deduce que el nivel de sobreajuste no radica únicamente en el número de parámetros y observaciones sino también en la colinealidad existente entre esos parámetros, es decir, en la adaptación de los datos al modelo, y en la diferencia existente entre los errores del modelo y el ruido o errores de los datos. Un sobreajuste completo podría llegar a ocasionar que el coeficiente de determinación se redujera a únicamente el correspondiente a la fase de entrenamiento.

Pare evitar el sobreajuste del modelo intentaremos identificar las situaciones en que un mayor entrenamiento deja de resultar beneficiosos para el modelo. Algunas técnicas con este fin son: Validación cruzada, Regularización, Pronto-parado, Poda y Bayesiano anterior. Estas técnicas actúan básicamente de dos maneras: penalizando los modelos extremadamente complejos, o evaluando la capacidad de predicción usando datos no utilizados en el diseño del modelo.

En este estudio se utilizará Validación cruzada. De acuerdo con esta técnica, estimamos que se ha producido un sobreajuste del modelo cuando al incrementarse el error de validación, observamos que el error de la fase de entrenamiento decrece. Por tanto, podemos decir que el modelo ajustado con mayor capacidad de predicción será aquel en el que el error de validación es mínimo.

- **Ruido**

Normalmente un algoritmo de aprendizaje se entrena con la fase de entrenamiento. Mediante parcialidad inducida, si la fase de entrenamiento es suficientemente representativa, el modelo podrá predecir nuevos casos. Pero en caso contrario se producirá una relación causal. Si la fase de entrenamiento es representativa, conseguiremos una mayor eficiencia de predicción, al disponer de una fase de entrenamiento mayor, pero al mismo tiempo, estaremos errando más en la predicción real con nuevos individuos ajenos al modelo.

Clasificando la información de un modelo en relevante e irrelevante (ruido), y asumiendo igualdad de condiciones, ocurrirá que cuanto más difícil sea de hacer la predicción más ruido habrá que distorsione la capacidad predictiva real. Por eso, resulta muy interesante que el modelo sea capaz de detectar el exceso de ruido. De un modelo con esta capacidad decimos que es robusto.

- **Sesgo**

En estadística, se llama sesgo a la diferencia entre el valor estimado del predictor y la información que el modelo interpreta realmente de ese predictor. El sesgo básicamente es una asunción errónea en los algoritmos de aprendizaje. Un sesgo elevado puede ocasionar que el algoritmo subestime determinadas relaciones relevantes entre predictores y la variable respuesta, en lo que se denomina “ajuste pobre” del modelo.

De entre los distintos tipos de sesgo, el sesgo espectral es de especial interés en este tipo de estudios en los que la variable respuesta es categórica. Este sesgo es un tipo específico de sesgo en el muestreo y tiene que ver con la posibilidad de analizar animales que no sean representativos de su categoría (de su raza en este caso) y que por tanto, causaran errores a la hora de calcular los ratios de especificidad y sensibilidad.

En este análisis es de especial interés el sesgo de confusión. Este tipo de sesgo se produce cuando se está tratando de determinar cual es el verdadero factor de un determinado estado en la variable respuesta. Si uno de los factores está asociado con otro que sí está asociado a la variable respuesta, podemos errar al pensar que el primer factor es el que realmente está estrechamente relacionado con la respuesta. Es decir, el efecto de un factor confunde o distorsiona el efecto del otro. Esto es habitual en análisis de datos genómicos, donde los marcadores pueden estar íntimamente ligados y donde, posiblemente, un marcador tenga plena influencia sobre otros muchos.

El sesgo sistemático o no aleatorio, por su parte, se da cuando los resultados difieren de una manera sistemática de los valores originales. Se producen subestimaciones o sobrestimaciones de lo resultados. El sesgo sistemático es cualquier

influencia que distorsione la comparación o las conclusiones que podamos sacar de un grupo. Cuando el sesgo sistemático sea reducido diremos que el modelo es acertado.

- **Varianza**

Por último, la varianza es un error que viene como consecuencia de pequeñas fluctuaciones en los predictores que podrían afectar a la fase de entrenamiento. Una alta varianza puede ocasionar un sobreajuste del modelo, haciendo que el ruido influya sobremanera en la variable respuesta. Este error es frecuente en análisis con datos genómicos y está íntimamente ligado al concepto de equilibrio de ligamiento.

Dadas las características de los datos genómicos y la variable respuesta "población" o "raza", se estima que unos cuantos predictores serán suficientes para llevar a cabo una predicción acertada. Lógicamente, los SNPs que podrán desempeñar mejor la función predictiva serán aquellos que tengan los coeficiente de correlación más elevados. La actuación de esos coeficientes de regresión para seleccionar los mejores predictores puede ser después comprobada comparando los valor AUC (Área debajo de la curva) de las curvas ROC (característica operativa del receptor), en las que se enfrentan los valores de sensibilidad y especificidad (Martínez-Camblor., *et al*. 2013). Los conceptos de sensibilidad y especificidad se han explicado en el apartado de Introducción, cuando se habla de investigaciones en materia de "reducción de la dimensión".

Es característico de este estudio que, además de predecir la VD, queremos seleccionar las VIś que más ayuden a esta causa. Esta es otra razón fundamental por la que no podemos limitarnos a hacer MLR.

Al problema de selección de los mejores predictores se le puede dar un enfoque basado en una clasificación de los predictores en base al text exacto de Fisher, o como se ha hecho, escogiendo un subconjunto de variables latentes efectuando validación cruzada.

3.3.3. Propuesta de alternativas

Dadas las características de nuestro modelo y los problemas existentes, debemos seguir dos estrategias fundamentales:

- Puesto que sobran predictores: Extracción de variables latentes que resuman la información (PCA, PLS o PCR).

- Puesto que queremos predecir una categoría: Búsqueda de la proyección que más facilite la clasificación (LDA).

Con ambas estrategias se consigue una reducción de la dimensión, si bien, será

combinando ambas estrategias que encontraremos la solución más óptima a nuestro problema, como veremos a lo largo de este capítulo.

- **Extracción de variables latentes que resuman la información**

Extraeremos variables latentes para poder predecir nuevos valores de la variable respuesta en la que basamos la pertenencia de una observación a una u otra categoría. Podemos extraer dichas variables latentes mediante tres técnicas fundamentales: PCA (*principal components analysis*), PLS (*partial least squeare*) (Garthwaite. 1994) y PCR (Jolliffe. 1982) (*principal components regression*). En este capítulo describiremos las tres técnicas y propondremos la mejor alternativa.

La suposición básica de estos métodos es que el modelo depende de un número pequeño de variables instrumentales llamadas variables latentes. Las variables latentes son estimadas como combinaciones lineales de las variables observadas.

Lo que se consigue con estos métodos es descomponer la información de los predictores que corresponden a mediciones correlacionadas, en un nuevo subconjunto de variables ortogonales (no correlacionadas).

Estas variables, dependiendo del contexto pueden recibir el nombre de Componentes Principales (CPs), factores eigenvectores, vectores singulares o cargas. Además, a cada uno de las variables, se le asigna un vector de puntuaciones que corresponde a su proyección sobre los componentes. A diferencia de con MLR, con estas técnicas el número de observaciones puede ser menor que el número de variables independientes ($n < p$).

- **Búsqueda de la proyección que más facilite la clasificación**

Proyectaremos las variables de manera que se facilite la diferenciación entre grupos. Para ello usaremos LDA.

El objetivo principal de LDA es proyectar los datos multidimensionales en un espacio de menos dimensiones (mediante unos vectores nuevos), donde los datos puedan ser separados fácilmente en diferentes categorías. Las proyecciones de estos vectores formarán los nuevos ejes y en ellos las observaciones tomarán unos valores nuevos. De manera que sobre estas direcciones se maximiza la diferencia entre-clases y se reduce la diferencia dentro-de-clases. Sin embargo, nosotros no queremos esas dimensiones, puesto que aunque nos permitirían diferenciar, no conseguiríamos predecir nuevos valores de la variable respuesta. Por esa razón implementaremos LDA sobre los resultados obtenidos en (PCA, PLS o PCR).

Ambas estrategias buscan la combinación lineal de las variables que mejor explique los datos. Si bien, PCA, PLS y PCR pretenden maximizar la capacidad

predictiva, mientras que LDA pretende modelar las diferencias entre clases.

A continuación, veremos en que consistente LDA para proyectar los datos con vistas a clasificar los individuos en grupos y seguidamente detallaremos las tres técnicas: PCA, PLS y PCR.

3.3.4. LDA (análisis lineal discriminante)

Con LDA se pretende encontrar la combinación lineal de predictores que mejor permita separar dos o más clases de observaciones y posteriormente usar esa combinación como clasificador lineal o, más habitual, para hacer una reducción de la dimensión y posterior clasificación.

Los caracteres que determinan esa nueva representación en LDA son combinaciones lineales de las variables originales. La proyección óptima o transformación en LDA se obtiene maximizando el ratio de las distancias de entre-clases y dentro-de-clases, consiguiendo así una máxima discriminación.

Imaginemos que tenemos un subconjunto de p-dimensiones $x_1,x_2,...x_n$, de los cuales N_1 pertenecen a una clase w_1 y N_2 pertenecen a la clase w_2. Pretendemos encontrar un escalar proyectando las muestras x en una línea:

$$y = w^T x$$

De todas las líneas posibles, nosotros queremos encontrar las que maximicen la diferencia entre los escalares. La siguiente imagen muestra un ejemplo:

FUENTE: http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

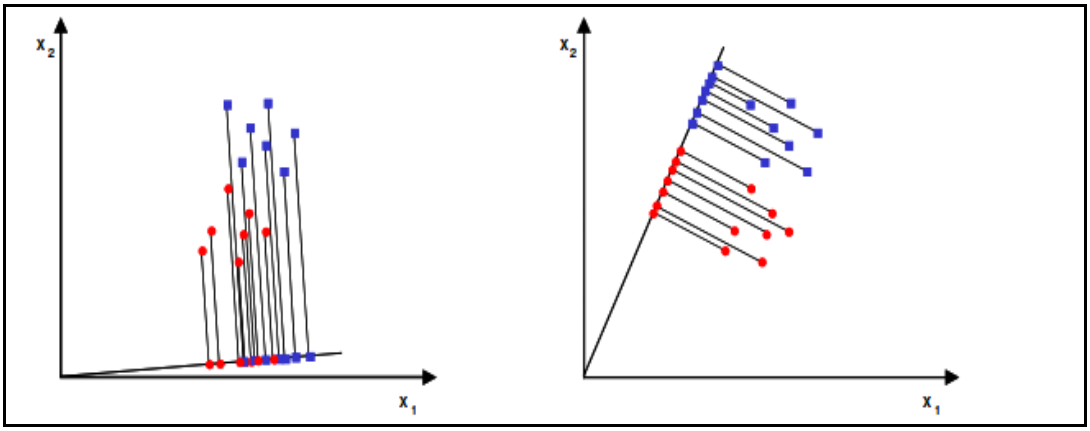


Ilustración 10: Ej. básico LDA

Para encontrar un buen vector de proyección necesitamos definir una medida de separación.

La media del vector en cada clase es equivalente al vector medio de cada clase en el espacio-X, y al vector medio de cada clase en el espacio-Y, como se aprecia en la siguiente expresión:

$$\mu_i = \frac{1}{N}\sum_{x\varepsilon wi} x \qquad = \qquad \tilde{\mu}\, i = \frac{1}{N}\sum_{y\varepsilon wi} y \;=\; \frac{1}{N}\sum_{x\varepsilon wi} W^T X = W^T \mu_i$$

Entonces podemos encontrar la distancia entre las proyecciones medias:

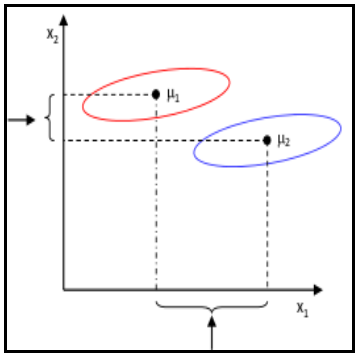
$$J(w) = |\, \tilde{\mu}_1 - \tilde{\mu}_2| = |\, W^T (\mu_1 - \mu_2)|$$

Sin embargo, la distancia entre las medias de las proyecciones no es una buena medida puesto que no se considera la desviación estándar entre clases:

Este eje diferencia mejor mejor entre clases

FUENTE: http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

Este eje, marca una diferencia mayor entre medias



Este eje, marca una diferencia mayor entre medias

Ilustración 11: Ej. básico LDA, no basta con la diferencia de medias

Ante este problema, Fisher (1936) propuso maximizar la diferencia entre medias pero normalizando antes con una medida de dispersión dentro de cada clase.

Para cada clase, definimos la dispersión con el equivalente de la varianza:

$$\widetilde{Si^2} = \sum_{y \epsilon wi} (y - \tilde{\mu}\, i)^2$$

Donde a la cantidad ($\tilde{S}1^2 - \tilde{S}2^2$) se la llama “dispersión dentro de cada clase” para los datos proyectados.

Entonces, el análisis discriminante de Fisher queda definido como la función lineal $W^T X$, que permite maximizar el siguiente criterio:

$$J(w) = \frac{(\mu 1 - \mu 2)^2}{\widetilde{S_1^2} + \widetilde{S_2^2}}$$

Por tanto, estamos buscando una proyección en la que los datos de una misma clase estén cerca unos de otros y, al mismo tiempo, las proyecciones de las medias de ambas clases estén lo más alejadas posible.

FUENTE: http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

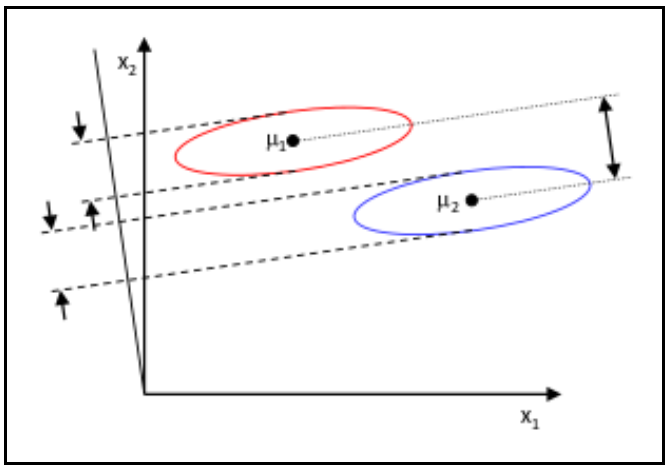


Ilustración 12: Ej. básico, Solución de Fisher

Para encontrar la w* óptima, expresamos J(w) como función de w. En primer lugar, definimos la medida de dispersión en el espacio x:

$$S_i = \sum_{x \epsilon wi} (X - \mu i) (X - \mu i)^T$$

Para dos clases $S_1 + S_2 = S_w$

Donde S_w es la matriz de dispersión dentro de una clase.

La dispersión en el espacio Y se puede después expresar como función de la dispersión en el espacio X.

$$\tilde{S}_i^2 = \sum_{y \epsilon wi} (y - \tilde{\mu} i)^2 = \sum_{x \epsilon wi} (W^T X - W^T \mu i)^2 =$$

$$= \sum_{x \epsilon wi} W^T (X - \mu i)(X - \mu i)^T W = W^T S_i W$$

$$\tilde{S}_1^2 + \tilde{S}_2^2 = w^T S_W w$$

De igual manera se puede expresar la diferencia entre las medias de las proyecciones en el espacio original.

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T\mu_1 - w^T\mu_2)^2 = w^T(\mu_1 - \mu_2)\text{-}(\mu_1 - \mu_2)^T\, w = w^T S_i w$$

Obsérvese que: $(\mu_1 - \mu_2)$ - $(\mu_1 - \mu_2)^T = S_B$

A esta S_B se la denomina “matriz dispersión dentro de las clases”. Puesto que S_B es el producto de dos vectores, su rango ≤1.

Finalmente, podemos expresar el criterio de Fisher en términos de S_W y S_B, como:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Para encontrar el valor máximo de J (w) derivamos e igualamos a cero:

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0$$

$$[w^T S_W w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_W w]}{dw} = O$$

$$[w^T S_W w]\, 2\, S_B\, w - [w^T S_B w]\, 2\, S_W\, w = 0$$

Dividiendo entre $[w^T S_W w]$:

$$\left[\frac{w^T S_W w}{w^T S_W w}\right] S_B\, w - \left[\frac{w^T S_B w}{w^T S_W w}\right] S_W\, w = 0$$

$$S_B\, w - J S_W\, w = 0$$

$$Sw^{-1} S_B\, w - Jw = 0$$

Solucionando el problema generalizado de eigenvalores ($Sw^{-1} S_B\, W = Jw$), se obtiene la expresión:

$$w^* = \arg\max \left[\frac{w^T S_B w}{w^T S_W w} \right] = Sw^{-1}(\mu_1 - \mu_2)$$

A esto se le conoce como el análisis discriminante de Fisher, aunque técnicamente no es una discriminación, sino la proyección de los datos sobre una sola dimensión.

Para varias clases (c clases), tendríamos que:

En lugar de una proyección "y", buscaríamos c-1 proyecciones $y_1, y_2 ... y_{c-1}$. como medias de los c-1 vectores de proyección (w_i) que estarían ordenados por columnas en una matriz de proyecciones $w = [w_1 | w_2 | ... | w_{c-1}]$:

$$y_i = w_i^T x \rightarrow y = w^T x$$

La dispersión entre clases se generaliza mediante la expresión:

$$S_w = \sum_{i=1}^{c} Si$$

, donde $Si = \sum_{x \varepsilon wi} (x - \mu i)\,(x - \mu i)^T$

Además tenemos que,

$$\mu_i = \frac{1}{N} \sum x \varepsilon w_i\, x$$

, y la dispersión entre clases se convierte en:

$$S_B = \sum_{i=1}^{c} N \quad (\mu i - \mu)(\mu i - \mu)^T$$

, donde

$$\mu = \frac{1}{N} \sum_{\forall} x_i \; x = \frac{1}{N} \sum_{i=1}^{c} Ni\mu i$$

A la matriz $S_T = S_B + S_W$, se la llama matriz de dispersión total.

Similarmente, se define el vector y las matrices de dispersión con:

$$\tilde{\mu}\, i = \frac{1}{N} \sum y \varepsilon w_i \, y$$

$$\tilde{\mu}\, i = \frac{1}{N} \sum_{\forall y} y$$

$$\tilde{S}_w = \sum_{i=1}^{c} \sum_{y \varepsilon wi} Y \quad (y - \tilde{\mu}\, i)(y - \tilde{\mu}\, i)^T$$

$$\tilde{S}_B = \sum_{i=1}^{c} N \quad (\tilde{\mu}\, i - \mu)(\tilde{\mu}\, i - \mu)^T$$

$$\tilde{S}_B = W^T S_W W$$

$$\tilde{S}_w = W^T S_B W$$

Recordamos que estamos buscando una proyección que maximice el ratio diferencia entre-clases entre diferencia dentro-de-clases. Puesto que la dimensión no es

escalar, ya que tiene c-1 dimensiones, usamos el determinante de la matriz de dispersión para obtener el escalar de la función objetivo.

$$J(W) = \frac{|\widetilde{S}B|}{|S\widetilde{W}|} = \left[\frac{W^T S_B W}{W^T S_W W}\right]$$

Y buscamos la proyección de la matriz w* que maximice dicho radio:

La proyección óptima de la matriz w* es aquella cuyas columnas son los eigenvectores con mayores eigenvalores en el siguiente problema generalizado de eigenvalor.

$$W^* = [w^*_1| w^*_2|...|w^*_{c-1}] \arg\max \left[\frac{w^T S_B w}{w^T S_W w}\right] (S_B - \lambda S_W)w^*i = 0$$

Advertencias:

- S_B es la suma de las c matrices de rango ≤ 1 y los vectores medios están constreñidos por:

$$\frac{1}{C}\sum_{i=1}^{c} \mu i - \mu$$

Entonces S_B tendrá un rango ≤ (c-1). Por tanto, sólo (C-1) de los eigenvalores de λ_i serán distintos de cero.

- La proyección con más información de separación entre clases son los eigenvectores correspondientes a los eigenvalores mayores de $S_W^{-1} S_B$

- LDA puede ser derivado como el método de máxima verosimilitud para el caso de clases normales de densidades condicionales con matrices de igual covarianza.

- La separación que se establece en LDA está contrastada y es la misma que la que se obtendría con el modelo EDC (distancia Euclidea a los centroides).

Limitaciones de LDA:

- LDA puede producir, como mucho, (c-1) proyecciones de caracteres.
- LDA fallará si la información discriminatoria no está en la media sino en la varianza de los datos.
- LDA no nos proporcionará necesariamente las dimensiones más informativas para predecir Y.

3.3.5. Técnicas de extracción de variables latentes: PCA, PLS y PLS

A continuación veremos en que consisten las tres técnicas vistas para extraer variables latentes que resuman la información: PCA, PLS y PCR.

- **PCA (análisis de componentes principales)**

PCA es un procedimiento estadístico que, mediante una transformación ortogonal lineal convierte los datos de las variables predictoras (en principio correlacionadas) en datos de un nuevo subconjunto de variables independientes. Dicha transformación es un *endormorfismo* puesto que se conserva el producto escalar y, por tanto, los ángulos y distancias.

Se consigue adaptar los datos a un nuevo sistema de coordenadas de tal manera que la mayor de las varianzas, observando los datos desde cada una de las posibles perspectivas, queda plasmada en la primera coordenada, la segunda mayor varianza en la segunda coordenada y así sucesivamente.

Dicho de otra manera, PCA ajusta un elipsoide de n-dimensiones a los datos. En dicho elipsoide cada eje representa un CP. Los CPs son ortogonales puesto que son eigenvectores de la matriz de covarianzas, que es simétrica, como veremos más adelante.

Las siguientes ilustraciones son representaciones de dicho endormorfismo:

FUENTE:https://georgemdallas.wordpress.com

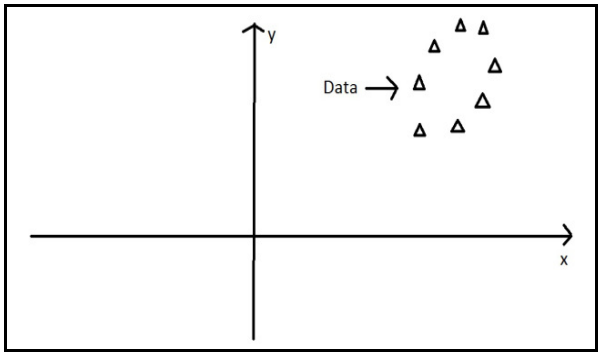


Ilustración 13: Ej. básico PCA, datos originales

FUENTE: https://georgemdallas.wordpress.com

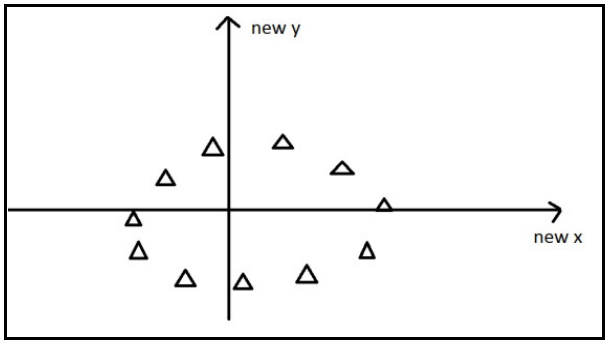


Ilustración 14: Ej. básico PCA, datos transformados

http:// completa: (https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/)

Como puede verse, los datos siguen siendo los mismo. Simplemente cambia que ahora los miramos desde otra perspectiva. Obtener los eigenvectores permite mover los datos de un juego de ejes a otro. Los nuevos ejes permiten ver la forma de los datos de una manera mucho más intuitiva.

- **PLS (mínimos cuadrados parciales)**

Los orígenes de PLS se remontan a un algoritmo desarrollado por Herman Wold para iteraciones no-lineales. Este tenía el objetivo de linealizar los modelos que fueran no lineales en sus parámetros. A este algoritmo se le conoce como NIPALS y fue adoptado en problemas de regresión sobredeterminada que antes se resolvía mediante PCR. Se consiguió así resolver este problema con un éxito mayor.

La regresión por mínimos cuadrados parciales (PLSR) se puede entender como una regresión ordinaria en la que a partir de una matriz de VI´s y otra de VD se trata de encontrar unas variables latentes. Conceptualmente es similar a ACP en cuanto a que reduce la dimensión de los datos X e Y para predecir Y de la mejor manera posible.

Tanto X como Y se descomponen progresivamente en puntuaciones y cargas que explican la mayor covarianza posible cov(X,Y). De modo que el método combina características de PCA (*principal components analysis*) y MLR. Las cargas y los pesos de los CPs pueden ser representados igual que en PCA y las VDs pueden ser estimadas con el coeficiente de confianza, como en MLR.

PLS nos es de utilidad en cuanto a que al simplificar la información de las VI´s, nos permitirá seleccionar los SNPs que mejor se adapten a dicha simplificación. Más concretamente, nuestro objetivo será seleccionar las variables de manera independiente para cada una de las posible respuestas teniendo en cuenta que $n << p$. Para ello se utilizan los coeficientes de regresión PLS que tendrán un valor absoluto mayor en las variables originales más relevantes. Si bien, existe otro método alternativo para seleccionar predictores con PLS, que es la “Puntuación del Predictor en un Resumen de la Importancia de las Proyecciones” (VIP), del inglés “*score of a predictor in a summary of the importance for the projections*”, que no debe confundirse con los valores VIP que se utilizan en el análisis genético para cuantificar el desequilibrio de ligamiento. Con el método de VIP se obtienen las variables latentes mediante un método distinto. Los valores para el método de VIP pueden ser calculados resumiendo la influencia de las variables sobre todas las dimensiones del modelo.

Cualquiera que sea el método a utilizar, PLS permitirá incluso resolver casos en lo que existen múltiples variables respuesta, permitiendo revelar las mejores variables en base a un único atributo relacionado con una de las variables respuesta.

Puesto que no se ve afectado por la colinealidad, y puesto que permite cuantificar la influencia del ruido, PLS es un modelo apropiado para analizar datos de procesos biológicos o para cualquier proceso cuyos datos tengan altas dimensiones (ya

sean datos de genómica, proteómica o peptidómica).

En este estudio se utilizan los coeficientes de regresión para seleccionar las mejores variables para cada una de las posibles categorías de la variable respuesta.

- **PCR (regresión de componentes principales)**

Esta técnica, al igual que PLS, se utiliza cuando se pretende analizar datos a partir de los cuales se quiere hacer una regresión múltiple, en los casos en que el problema de la multicolinealidad es importante.

Cuando hay multiclolinealidad, las estimaciones son insesgadas pero sus varianzas son grandes, de modo que las estimaciones podrían alejarse bastante de los valores verdaderos. Añadiendo un grado de parcialidad a las estimaciones, con PCR se reducen los errores estándar y se consigue obtener unas estimaciones más fiable. Se estima que la media de los errores al cuadrado de esta estimación sea menor que con mínimos cuadrados ordinarios.

Las m variables originales se transforman en un nuevo grupo de variables ortogonales y no correlacionadas que son los CPs extraídos de la matriz de covarianzas. Esta transformación permite ordenar posteriormente las variables según su capacidad explicativa. Para reducir la dimensión, simplemente eliminaremos los peores PCs.

Una vez eliminados, efectuamos una regresión lineal múltiple con el resto de los CPs, mediante mínimos cuadrados ordinarios (OLS). Puesto que las nuevas variables son ortogonales y cada par de variables es linealmente independiente, OLS es un método adecuado para ajustar el modelo. Una vez obtenidos esos coeficientes, los transformaremos para llevarlos a la misma escala que las variables originales. Llamaremos a estos nuevos coeficientes los estimadores de los Componentes Principales. El modelo de estimación será más robusto que con MLR. Una vez visto en que consisten estas tres técnicas, explicaremos detalladamente cada una de ellas. Empezaremos por PCA puesto que conocer esta técnica resulta básico para poder comprender las otras dos.

3.3.6. PCA (análisis de componentes principales)

Esta técnica es muy anterior a las otras dos, de hecho PLS y PCR son adaptaciones de PCA.

En primer lugar definiremos algunos conceptos clave:

- Coeficientes de regresión: Los coeficientes de regresión o pesos representan el cambio principal en la variable respuesta que pueda deberse a una unidad de cambio en la variable predictora bajo la condición de que se mantienen

constantes el resto de predictores del modelo. Este control estadístico de la regresión es importante porque permite estimar la importancia relativa de cada uno de los predictores sobre el resto.

En PCA tendremos dos vectores de pesos: uno a partir del cual obtenemos la variable respuesta mediante una combinación lineal de las variables predictoras, y otro a partir del cual obtenemos la variable respuesta a partir de las puntuaciones (que como veremos son los valores de los predictores al situarlos sobre los CPs). En el segundo caso, el producto de la matriz de puntuaciones por el vector de pesos nos daría el vector de la variable respuesta

- Eigenvectores: Son eigenvectores de una matriz cuadrada $A_{(m \times m)}$, aquellos vectores que al multiplicar por la izquierda a dicha matriz se obtiene un vector múltiplo de sí mismo. Si una matriz tiene algún eigenvector, decimos que es una matriz de transformación.

  Una matriz de transformación $A_{(m \times m)}$ tendrá m eigenvectores, puesto que los eigenvectores plasman los datos en un nuevo espacio con distintas dimensiones, pero el número de dimensiones seguirá siendo el mismo. Dadas las propiedades de los eigenvectores, estas nuevas dimensiones serán perpendiculares entre sí. Además, los distintos m eigenvectores recogen una cantidad muy diferente de la varianza de los datos originales, pudiéndose establecer una jerarquía de los eigenvectores en base a este criterio.

- Eigenvalores: Es eigenvalor de un eigenvector el múltiplo que antes hemos mencionado (el múltiplo del eigenvector que se obtiene al multiplicar una matriz de transformación por un egenvector). Los eigenvalores son mayores en aquellos eigenvectores con mucha varianza. Por tanto, los eigenvalores nos permiten ordenar los CPs según su capacidad explicativa. De hecho, la proporción de ese eigenvalor con respecto al total nos va a indicar la capacidad explicativa del correspondiente eigenvector. Si no hubiera varianza los eigevalores serían nulos.

  Puesto que los eigenvalores suelen tomar valores próximas a uno, habitualmente se toma la unidad como valor umbral y se descartan los eigenvectores cuyos eigenvalores son inferiores. Los eigenvalores representan la varianza o dispersión de los datos en la dirección definida por un eje.

- Matriz de covarianzas: Una matriz de covarianzas es una matriz cuadrada y simétrica que indica la correlación entre cada par de variables. La diagonal principal de la matriz de covarianzas corresponde a las varianzas de cada una de las variables. Cuando existe multicolinealidad perfecta el determinante de este valor es nulo y decimos que la matriz es singular y no transponible.

- Componente principal: Llamamos componentes principales a los eigenvectores de una matriz de covarianza. Cada componente principal es, bien una de las variables originales, o bien (más probablemente), una combinación lineal de algunas de las variables originales. Decimos que los CPs son la estructura fundamental de los datos, puesto que son las direcciones en las que hay una mayor varianza.

- Puntuaciones o *scores*: Las puntuaciones son los datos reales de las variables originales que han recibido un nuevo valor al estar representados sobre las nuevas variables. Constituyen el vector de pesos. Dispondremos de un score por cada observación y por cada nueva variable.

- Cargas o *loadings*: De lo anterior se deduce que para transformar los datos de las variables originales en puntuaciones hay que multiplicar esos datos originales por algún valor. Ese valor son las cargas. Dicho de otra manera, las cargas son los valores por los que se multiplica cada variable original estandarizada para obtener la puntuación en la nueva matriz transformada. Dicho de otra forma, son las marcas equivalentes sobre los componentes principales.

  De tal manera que las puntuaciones son Y veces las cargas, donde Y es el dato original. Por eso decimos que las nuevas variables (CPs) son combinación lineal de las anteriores.

Reducción de la dimensión

En análisis de regresión, un número muy grande de predictores será muy propenso a causar un sobreajuste del modelo y la ecuación de predicción no se podrá aplicar a nuevas muestras. Una solución a este problema, especialmente cuando además se da una fuerte correlación de las variables predictoras o cuando existe abundancia de ruido, es reducir estas a un número pequeño de componentes y haciendo después una regresión sobre estos.

Dicha reducción resulta muy útil para visualizar y procesar datos multidimensionales, reteniendo aún una parte muy importante de la información. Por ejemplo, seleccionando L = 2 y conservado sólo los dos primeros Cps, conseguiremos una interpretación más sencilla de los datos al representarlos sobre un plano bidimensional, que si los estuviéramos representando en un espacio multidimesional en el que la los datos estuvieran más dispersos. Si por el contrario, representáramos los datos en un plano bidimensional pero cuyos ejes los hubiéramos elegido aleatoriamente, habríamos simplificado la información pero obtendríamos una idea bastante alejada de la real.

Si cada una de las columnas proporciona una distribución gausiana de ruido idéntica e independiente, entonces las columnas de T, que son puntuaciones en el nuevo sistema de coordenadas, volverán a tener esa distribución de ruido. Esto ocurrirá con

independencia de cual sea la mariz de pesos W. Sin embargo, si concentramos buena parte de la varianza en los primeros CPs, y debido a que la varianza del ruido seguirá siendo la misma que antes, la proporción del efecto de ese ruido será menor. Los primeros componentes consiguen un mayor ratio de señal frente a ruido. Mientas que los últimos componentes estarán dominados, fundamentalmente, por el ruido. Por tanto, prescindir de ellos no supondrá una gran pérdida.

¿Como consigue ACP la reducción de la dimensión?

El número de componentes principales será menor o igual que el número de variables originales. La trasformación se hace de tal manera que el 1º CP es el que mayor varianza tiene, o lo que es lo mismo, expresa toda la variabilidad posible, y los siguientes CPs también expresarán toda la varianza posible, bajo la condición de que tienen que ser ortogonales al primero y a todos los anteriores.

Si una muestra de datos es visualizada como un subconjunto de coordenadas en un espacio multidimensional, PCA puede proporcionar una visión con menos dimensiones mediante una proyección o sombra del objeto multidimensional, visualizándolo desde la perspectiva que mejor explica el contenido. Esto se consigue usando sólo los primeros componentes principales. Dicho de otra manera, si alguno de los ejes es pequeño, entonces la varianza a lo largo de ese eje es también pequeña, y omitiendo ese eje y el correspondiente CP, perderemos sólo una pequeña parte de la información.

Con la transformación T = A W, cuya interpretación es que la matriz de puntuaciones es igual al producto de la matriz original de datos por la matriz de pesos, situaremos un vector x_i de un espacio original con p variables, en un nuevo espacio con p variables que no están correlacionadas. Sin embargo, no tienen que conservarse todos los CPs.

Al usar sólo los L primeros CP, obtenidos según los L primeros vectores de cargas, se proporciona la siguiente información truncada:

$$T_L=A*W_L$$

Mediante la construcción de esa matriz transformada de datos con sólo L columnas, las puntuaciones serán tales que se maximice la varianza de los datos originales, minimizándose el error.

Para encontrar los ejes del elipsoide, se realizan los siguientes pasos:

- Se dispone de n muestras de un espacio m-dimenional, con los vectores $x_1,x_2...x_m$ pertenecientes a R_m, y se calcular la media.

- Se substrae la media de cada variable para así centrar los datos alrededor del nuevo origen. Se obtiene así la matriz A.
- Se computa la matriz de covarianzas S y se calculan los eigenvalores $\lambda_1,...\lambda_m$, y eigevectores de dicha matriz, $\vec{v}$ i... $\vec{v}$ n.
- Se hace la transformación ortogonal de esos eigenvectores y se normalizan para transformarse en vectores unitarios.
- Una vez hecho esto, cada uno de los, mutuamente ortogonales, eigenvectores unitarios, se puede interpretar como un eje del elipsoide. Además, nos planteamos la cuestión de si existirá algún λ que sea mucho mayor que el resto.
- Ordenar los eigenvectores según la importancia de esos eigenvectores. La proporción de la varianza que representa cada eigenvector se puede calcular dividiendo el eigenvalor correspondiente entre la suma de eigenvalores.

Explicación matemática

Imaginemos una matriz A de número reales y simétrica, tal que $A^T = A$, entonces A es ortogonalmente diagonalizable, lo que significa que A tendrá n eigenvalores reales $\lambda_1,...\lambda_n$, y n eigenvectores $\vec{v}_1, ... \vec{v}_n$. Según el teorema espectral:

$$A\vec{v}_I = \lambda_i \vec{v}_I$$

Nuestra matriz A de datos genómicos contiene números reales pero no es simétrica. Sin embargo, sabemos que los productos AA_T y $A_T A$ si lo son. De manera que, aplicando el teorema espectral, tendremos:

$$(A^T A)\vec{v} = \lambda \vec{v}$$

Multiplicando en ambos términos de la ecuación por A, obtenemos:

$$AA^T(A\vec{v}) = \lambda (A\vec{v})$$

De lo que se deduce que el vector $A\vec{v}$ es un eigenvector de AA^T, con el eigenvalor λ, y que λ es igualmente, eigenvalor de $A^T A$. La única comprobación que deberíamos hacer es que $A\vec{v}$ no sea un vector cero, puesto que los eigenvectores no pueden tomar nunca valores nulos. Sin embargo, si $A\vec{v}$ fuera 0, entonces λ también lo sería, y esto no puede ser puesto que desde el principio especificamos que λ_1, debía

tomar valores distintos de cero. Pero además, teniendo un eigenvector $\vec{v}$ de $A^T A$, podemos obtener el eigenvector de AA^T multiplicando ambos términos de la ecuación por A; y de igual manera, en el sentido contrario, multiplicando $A_T \vec{w}$, en vez de $A \vec{v}$. Esto resulta muy útil en el caso de que n<<m. Por ejemplo, imaginemos $A_{500 \times 2}$, entonces hay una manera rápida de identificar los eigenvalores de AA^T, y ésta es encontrando antes los eigenvalores de $A^T A$, que es un matriz 2x2. Así, sabemos que los otros 498 eigenvalores toman valor cero. Entonces ahora nos preguntaremos cómo podemos saber que un eigenvalor de AA^T y $A^T A$ no es negativo.

Sabemos que el cuadrado de un vector($\vec{w}$ puede ser $\vec{w}\vec{w}$, pero también $\vec{w}_T \vec{w}$

Imaginemos que $\vec{v}$, es un eigenvector de $A_T A$ con el eigenvalor λ. Si computamos el cuadrado de $A\vec{v}$, tendremos:

$$(\| A\vec{v} \|^2 = (A\vec{v})_T (A\vec{v})$$

$$= \vec{v}_T (A_T A)\vec{v}$$

$$= \lambda \vec{v}_T \vec{v}$$

$$= \lambda \| \vec{v} \|^2$$

De donde se deduce que λ, puesto que se trata de una longitud, no puede ser negativo. Reemplazando A por A^T, se deduce lo mismo para A^T.

Concepto estadístico

Suponemos que hacemos n mediciones de una variable. Calculamos la media, según:

$$\mu_A = \frac{1}{n}(a_1 + ... + a_n)$$

y como medida de dispersión, calculamos la varianza:

$$\text{Var}(A) = \frac{1}{n-1}((a_1 - \mu_a)^2 + ... + (a_n - \mu_a)^2)$$

Si incluimos otra variable en el estudio, podemos calcular también las covarianzas de los distintos pares de variables. Esto nos permite estudiar la correlación entre las mismas mediante la siguiente:

$$cov(A,B) = \frac{1}{n-1}((a_1-\mu_a)(b_1-\mu_b)...(a_n-\mu_a)(b_n\mu_b))$$

Si cov(A,B) es negativa significa que cuando A crece, B decrece y viceversa. También sabemos que cov(A,B) = cov(B,A)

Así tenemos una medida de varianza de cada variable y una medida de covarianza por cada par de variables.

Llevado a un espacio multidimensional pode,mos escribir las medias de todas las variables en un vector perteneciente a R_m, según la expresión:

$$\vec{u} = \frac{1}{n}(x_1+...+x_n)$$

Es habitual re-centralizar los datos en R_m para que la media sea cero. Esto se consigue sustrayendo la media al vector x_i, para cada una de las observación.

Si $A_{m\,x\,n}$ es una matriz de datos, con la columna $x_i - \mu$, entonces:

$$A = [x_1 - \mu]...[x_n-\mu]$$

, y la matriz de covarianza $S_{m\,x\,m}$ se calculará según:

$$S = \frac{1}{n-1}BB_T$$

Para entender la matriz de covarianzas imaginemos que m = 2,y la siguiente representación de los n datos:

FUENTE: https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

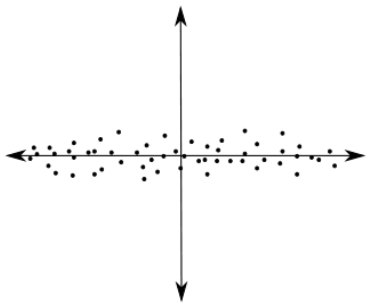


Ilustración 15: II Ej. básico PCA. Espacio bidimensional.

En el eje cartesiano se representan los datos en base a dos variables. En el eje horizontal se distribuyen los datos según los valores que tomen para una primera variable, y en el eje vertical en base a los valores para una segunda variable. En la variable considerada en el eje horizontal, los datos tienen mucha más varianza. Como puede apreciarse, el espectro de valores representados para nuestros datos es muy grande. Sin embargo, para la variable considerada en el eje vertical, la varianza es mucho menor. En la primera variable, puesto que hay tanta varianza, la entrada S_{11} de la matriz de covarianzas (cov A,A) tendrá valores altos, mientras que la matriz de covarianzas de la variable considerada ene el eje vertical S_{22}, tendrá valores pequeños. Además, a partir de la representación de las variable podemos llevarnos una idea aproximada de como es la covarainza entre el par de variables. Las matrices de covarianza de ambas variables, cuyos valores de la diagonal determinan como de grande es la covarianza, se denominan como S_{12} o S_{21} según cual sea la variable de referencia en la expresión del cálculo de covarianzas:

Se deduce que con ambas variables se obtiene el mismo valor de covarianzas a partir de las digonales, tal que $S_{12} = S_{21.}$

Puesto que se aprecia una escasa relación entre ambas variables, en cuanto a que valores altos en una variable no se corresponden necesariamente con valores altos en la otra y de igual maner para valores bajos o medios, podemos presuponer que $S_{12} = S_{21}$ también tomarán valores pequeños.

Sin embargo, en un segundo ejemplo vemos:

FUENTE:
https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction

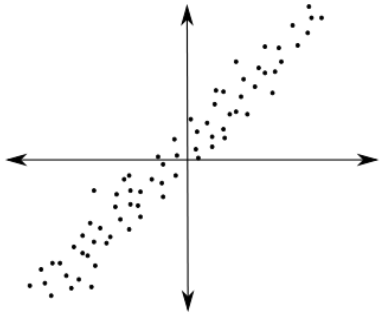


Ilustración 16: II Ej. básico PCA. Espacio bidimensional, varianzas parecidas

Las varianzas son muy parecidas en ambas variables. Además existe una clara correlación positiva. Los cuatro valores de la matriz de covarianza serán más parecidos. Destacamos que, aunque en ambos ejemplos la nube de puntos es idéntica, las matrices de covarianza son completamente distintas. Puesto que S es una matriz cuadrada y simétrica, puede ser ortogonalmente diagonalizada como en el teorema del espectro.

Aplicando el teorema y ordenando los eigenvalores de S, tal que: $\lambda_1 > \lambda_2 > ... > \lambda_m > 0$, con los correspondientes vectores ortogonales: $\vec{u}_1, \vec{u}_2 ... \vec{u}_m$, decimos que esos eigenvectores son Componentes Principales. Además, recordemos que un vector $\vec{u}_i$, siempre se puede sustituir con su negativo. También es de interés destacar que la traza de esa matriz S es la diagonal de la misma, que es precisamente la suma de todas las varianzas de las m variables y llamamos a esto la varianza total T de nuestros datos. Por otro lado, la traza de una matriz es igual a la suma de los eigenvalores. De modo que tenemos:

$$T = \lambda_1 + \ldots + \lambda_m$$

A modo de resumen:

Considerando la matriz de datos A con las columnas orientadas hacia cero (puesto que la media de la muestra se desplazó hacia el cero) interpretamos que cada una de las n filas representa una repetición diferente del experimento (una observaciones) y cada una de las p columnas proporciona un tipo particular de dato Por ejemplo, las columnas

podrían ser algo así como los resultados de un sensor particular, que hace un único tipo de medida estándar para todas la observación. La transformación queda definida por un subconjunto de p-dimensiones con vectores de pesos que proyectan cada fila de la matriz A hacia una nueva puntuación sobre el componente principal. De tal manera que esas puntuaciones de las nuevas variables expliquen la variabilidad de A de la manera más completa posible.

Consideramos la siguiente interpretación fundamental para PCA:

- En R_m, la dirección de $\vec{u}_1$ (dirección del primer componente principal) supone λ_1 del total de varianza T. Es decir explica la proporción (λ_i/T); y de igual manera el segundo componente principal (λ_2/T), y así sucesivamente.
- De manera que el vector $\vec{u}_1$ perteneciente a R_m, señala la dirección más significativa del subconjunto de datos.
- De entre las direcciones que son ortogonales a $\vec{u}_1$, $\vec{u}_2$ es la más significativa del subconjunto de datos.
- De entre las direcciones que son ortogonales a $\vec{u}_1$ y $\vec{u}_2$, $\vec{u}_3$ es la más significativa del subconjunto de datos y así sucesivamente.

Aplicaciones de PCA

Entre las aplicaciones que se pueden dar a estos CPs, la más importante es disponer de una interpretación más sencilla de nuestros datos, pero también, análisis de outliers, identificación de clusters de individuos o, en el caso del presente estudio, una reducción de la dimensión.

La reducción de la dimensión mediante variable latentes puede tener varios fines. Desde reducir la multicolinealidad y mejorar la capacidad predictiva a seleccionar un número reducido de variables, simplificando el análisis de los datos en futuras observaciones.

Conviene destacar la siguiente diferencia fundamental con OLS (Mínimos Cuadrados Ordinarios). Contrariamente a OLS, que minimiza los errores perpendicularmente a alguno de los ejes cartesianos, PCA minimiza los errores perpendicularmente a la línea del modelo.

Limitaciones de PCA

El análisis de componentes principales presenta algunas limitaciones que serán más o menos importantes según los objetivos del análisis:

- PCA es sensible a un escalar de los datos y no está consensuada cual es la mejor

manera de escalar los datos para obtener los mejores resultados.

- Los signos de los eigenvectores son completamente arbitrarios.
- Puesto que son ortogonales y completamente independientes, no se puede obtener un CP a través de otros CPs.
- La asunciones son las mismas que aquellas usadas en regresión lineal múltiple: linealidad, varianza constante o la no existencia de individuos ouliers, e independencia. Si bien, puesto que PC no proporciona intervalos de confianza, no se asume la hipótesis de normalidad.
- Al obtener finalmente los coeficientes de regresión debemos volver a ajustar esos valores a la escala de los datos originales.
- PCA sólo es aplicable en procesos lineales y estáticos.

3.3.7. Explicación de OLS (mínimos cuadrados ordinario)

Además de entender PCA, para comprender PCR o PLS, es necesario saber como se ajusta un modelo de regresión mediante OLS (mínimos cuadrados ordinarios). Conviene aclarar que OLS no es una técnica de extracción de variables latentes, sino que es el método más empleado en estadística para ajuste de modelos lineales.

Partiendo la ecuación básica de regresión:

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

OLS nos permite calcular los coeficientes β_0 y β_1. OLS considera la desviación entre el valor Y_i y su valor estimado:

$$Y_i = (\beta_o + \beta_1 X_i)$$

Concretamente, el método requiere considerar el cuadrado de la suma de las *n* desviaciones. Llamamos a este criterio Q:

$$Q = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

De acuerdo con OLS, los estimadores β_0 y β_1 serán, respectivamente, los valores b_0 y b_1 para los que se minimiza Q, dadas las observaciones (X_1,Y_1), (X_2,Y_2),...(X_n,Y_n).

Dichos valores b_0 y b_1 ,se pueden determinar de dos maneras, básicamente:

- Por tanteos
- Mediante un procedimiento analítico. Sólo cuando es modelo no es matemáticamente complejo.

En dicho procedimiento analítico, el valor mínimo de Q viene dado por el siguiente par de ecuaciones, que se denominan ecuaciones normales:

$$\sum Y_i = nb_0 + b_1 \sum Xi$$

$$\sum X_i Y_i = b_o \sum X_i + b_1 \sum X_i^2$$

Las ecuaciones normales se pueden despejar según:

$$b_1 = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2}$$

$$b_0 = \frac{1}{n}(\sum Yi - b1 \sum Xi) = \bar{Y} - b1\,\bar{X}$$

Donde $\bar{X}$ e $\bar{Y}$ son los valores medios de las observaciones X_i e Y_i, respectivamente.

Propiedades de la estimación por mínimos cuadrados:

El teorema fundamental de Gauss-Markov sostiene que bajo las condiciones del modelo de regresión arriba planteado, los estimadores b_0 y b_1, son insesgados y tienen varianza mínima entre todos los estimadores lineales insesgados. Puesto que b_0 y b_1 son insesgados, tenemos que:

$$E\{b_0\} = B0 \qquad E\{b_1\} = B1$$

De modo que ningún estimador tiende a sobrestimar o subestimar de manera sistemática. Además, el teorema sostiene que los estimadores b_0 y b_1 son más precisos que cualquier otro estimador que pertenezca a la clase de estimadores insesgados y que sean una función lineal de las observaciones $Y_1,...Y_n$. Los estimadores son una función lineal de Y_i. Considerando b_1, por ejemplo, tendremos que:

$$b_1 = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2}$$

Está demostrado que:

$$b_1 = \frac{\sum (Xi - \bar{X}) - Yi}{\sum (Xi - \bar{X})^2} = \sum K_i Y_i,$$

, donde:

$$K_i = \frac{Xi - \bar{X}}{\sum (Xi - \bar{X})^2}$$

Entonces, K_i es una constante conocida puesto que X_i lo es. De lo que se deduce que b_1 es una combinación lineal de Y_i, por tanto, es un estimador lineal.

Dadas las ecuaciones normales:

$$\sum Y_i = nb_0 + b_1 \sum Xi$$

$$\sum X_i Y_i = b_o \sum X_i + b_1 \sum X_i^2$$

Usando notación matricial y multiplicando en ambos lados de la ecuación por la transpuesta de X para poder multiplicar por el vector b, tendremos:

$$X^T X\, b = X^T Y$$

, donde b es el vector con los coeficientes de la ecuación de regresión por mínimos cuadrados.

Para obtener los coeficientes de regresión a partir de la ecuación anterior, multiplicamos por la izquierda en ambos términos de la ecuación por el producto $(X^T X)^{-1}$:

$$(X^TX)^{-1}\, X^TXb = (X^TX)^{-1}\, X^TY$$

Puesto que $(X^TX)^{-1}\, X^TX = I$; y $Ib = b$, el modelo se simplifica de la siguiente forma:

$$b = (X^TX)^{-1}\, X^TY$$

Que, naturalmente, genera los mismo coeficientes que:

$$b_1 = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2}$$

$$b_0 = \frac{1}{n}(\sum Yi - b1 \sum Xi) = \bar{Y} - b_1 \bar{X}$$

3.3.8. PLS (mínimos cuadrados parciales)

En primer lugar, en PLS al igual que en ACP, sustraemos la media a nuestros datos, tanto en la variable dependiente como en las independientes. Vamos a obtener una matriz de puntuaciones (T) con las nuevas puntuaciones sobre dimensiones ortogonales. Para ello, multiplicamos nuestra matriz de variables originales (X) por una matriz de pesos (W):

$$T_{(n \times c)} = X_{(n \times p)} W_{(p \times c)}$$

$T_{(n \times c)}$, es la matriz de puntuaciones que corresponde a la representación de los datos sobre los nuevos ejes.

$X_{(n \times p)}$, es la matriz con las variables originales.

$W_{(p \times c)}$, es una matriz de pesos.

Las columnas de W son vectores de peso para las columnas de X a partir de las

cuales se obtiene T. Estos pesos se calculan mediante un algoritmo de tal manera que cada uno de ellos maximice la covarianza entre la respuesta y las correspondientes puntuaciones. PCR, por el contrario, calcula esta matriz reflejando la covarianza entre sólo los predictores. Es decir, los métodos difieren en la forma de extraer las puntuaciones. Decimos que T contiene la descomposición en variables latentes para las n observaciones. Al haber multiplicado por unos pesos, es una combinación lineal de las variables originales obtenida de tal manera que no exista correlación entre las puntuaciones.

Una vez obtenidas los puntuaciones, consideramos el siguiente modelo de regresión lineal en dos ecuaciones:

$$Y_{(n \times q)} = T_{(n \times c)} Q^{T}_{(c \times q)} + F_{(n \times q)}$$

$$X_{(n \times p)} = T_{(n \times c)} P^{T}_{(c \times p)} + E_{(n \times p)}$$

$X_{(n \times p)}$, es la matriz con las variables originales

$Y_{(n \times q),}$ es la matriz original con la variable respuesta

$T_{(n \times c)}$, es la matriz de puntuaciones

$P_{(p \times c)}$, es una matriz de coeficientes de

$Q_{(q \times c)}$, es una matriz de de pesos o *loadings* para Y, y una matriz de coeficientes de T

$F_{(n \times q)}$, y $E_{(n \times p)}$ son matrices de error aleatorio.

Si T, P y Q satisfacen la ecuación, entonces para cualquier matriz se cumple singular $M_{(c \times c)}$, se cumple:

$$T^{*} = TM \qquad Q^{*} = Q(M\text{-}1)^{T} \qquad P^{*} = P(M^{-}1)^{T}$$

Además, considerando las otras ecuaciones vistas en este apartado y la ecuación general de regresión simple:

$$Y_{(n \times q)} = T_{(n \times c)} Q^{T}_{(c \times q)} + F_{(n \times q)}$$

$$T_{(n \times c)} = X_{(n \times p)} W_{(p \times c)}$$

$$Y = BX + E$$

, se obtiene la siguiente relación:

$$B = WQ^{T}$$

Entonces, tenemos:

$$Y_{(n \times q)} = T_{(n \times c)} Q^{T}{}_{(c \times q)} + F_{(n \times q)}$$

, conociendo Y y T, aplicamos Mínimos Cuadrados ($b = (X^{T}X)^{-1} X^{T}Y$), y tenemos :

$$Q = (T^{T} T)^{-1} T^{T} Y$$

Que es la simplificación obtenida de aplicar OLS en la regresión de Y sobre T.

Finalmente, podemos calcular el coeficiente de regresión, según:

$$Y = XB + F$$

$$B = WQ^{T} = W(T^{T}T)^{-1}T^{T}Y$$

, que nos permite calcular la variable respuesta estimada, con la siguiente expresión:

$$\hat{Y} = T(T^{T}T)^{-1}T^{T}Y$$

Dejando a un lado la notación matricial, tendremos:

$$\widehat{Yo} = \frac{1}{n}\left(\sum_{1=1}^{n} yi\right) + B^{T}\left(X_0 - \frac{1}{n}\sum_{1=1}^{n} xi\right)$$

Algoritmo de kernel

Este algoritmo permite convertir matrices largas de $X_{(N \times K)}$, o $Y_{(N \times M)}$ condensándolas en una matriz cuadrada pequeña de orde (k x k) llamada matriz de kernel $\mathbf{X^{T}YY^{T}X}$, y de tamaño igual al número de variables de X. Usando la matriz de kernel junto con las

pequeñas matrices de covarianza: $\mathbf{X}^T\mathbf{X}_{(K \times K)}$, $\mathbf{X}^T\mathbf{Y}_{(K \times M)}$ y $\mathbf{Y}^T\mathbf{Y}_{(M \times M)}$, se pueden estimar todos los parámetros para resolver una regresión por PLS.

En el algoritmo de kernel para PLS, el número de CPs aumenta gradualmente hasta que el modelo alcanza alguna dimensión óptima. Por ejemplo, podemos hacer Validación Cruzada para determinar la adecuación de un componente individual para entrar en el modelo, o para la comparación de modelos enteros de ciertas dimensiones. En este estudio, puesto que se trata de ordenar las variables originales, construiremos un modelo de una única dimensión.

3.3.9. PCR (regresión de componentes principales)

La regresión de Componentes Principales es un método de calibración para modelos multivariables en dos etapas:

■ En la primera etapa se efectúa un análisis de componentes principales sobre la matriz A, y esa matriz de predictores se convierte en una matriz de pesos y unas variables latentes.

Como ya se ha dicho, PCA crea nuevas variables ortogonales que son combinación lineal de las originales, tal que:

$$A = TP^T$$

Donde T es la matriz de puntuaciones o nuevas marcas, y P es la matriz de cargas. Esta descomposición plantea dos ventajas fundamentales: las nuevas variables son ortogonales y además, la inversa de T, que es necesaria para hacer MLE, ya no plantea el problema de disponer de variables correlacionadas. Puesto que se va a hacer PCA, en primer lugar se estandarizaran los datos de la matriz A y todos los cálculos se harán con los datos estandarizados. La ecuación fundamental de PCR es:

$$A^TA=PDP^T=ZZ^T$$

Donde D es la diagonal de la matriz de los eigenvalores de A^TA, P son los eigenvectores de dicha matriz, y Z una matriz de datos similar a A, pero sobre los CPs. P es ortogonal y por tanto $PP^T=I$

Hemos creado las nuevas variables Z como medias ponderadas de las variables originales A. Puesto que estas variables nuevas son CPs, las correlaciones de unos con otros son todas nulas. Se consigue detectar buena parte de la multicolinealidad en forma de eigenvalores pequeños.

Cuando hacemos regresión de Y sobre Z, la multicolinealidad ya no es un problema y podemos después devolver los resultados a la escala original de datos para obtener B.

Estas estimaciones estarán sesgadas pero muy probablemente ese sesgo se haya compensado con una pérdida de la varianza. En otras palabras, estimamos que la media de los errores al cuadrado de esta estimación sea menor que con mínimos cuadrados. Matemáticamente, la formula de estimación adquiere la siguiente forma, dadas las propiedades de PC:

$$A = (Z^T Z)^{-1} Z^T Y = DYZ^T Y$$

Para omitir un CP bastará con igualar a cero el correspondiente elemento de la matriz A.

Adviértase que la ecuación anterior es la ecuación de mínimos cuadrados ordinarios aplicada a un subconjunto distinto de variables independientes. Si recordamos, cuando vimos mínimos cuadrados ordinarios obteníamos que:

$$b = (X^T X)^{-1} X^T Y$$

■ En la segunda etapa, se realiza una regresión lineal múltiple entre los pesos obtenidos en el análisis de componentes principales y la matriz de respuestas Y.

Dada la capacidad explicativa de esas nuevas variables, se retendrán l CPs, donde l < min(n,p), quedando el modelo simplificado.

La posterior MLE es modelada según la expresión:

$$Y = TC + E = APC + E = AB + E$$

, cuyo coeficiente de regresión proviene de:

$$B = P\,(T\,^T T) - 1\,T\,^T Y$$

3.3.10. Elección del modelo – Comparaciones

◆ PLS vs OLS

La razón por la que se implementa PLS en lugar de OLS es que los subconjuntos de datos X e Y pueden contener un ruido aleatorio que debería ser excluido. Descomponer X e Y en una serie de variables latentes puede servir para asegurar que en la regresión se de efectivamente la variación real.

◆ PCR vs PLS

Las diferencias fundamentales entre PCR y PLS son las siguientes:

- Aunque en ambos casos se trata de encontrar los CP que mejor expliquen la variabilidad de los predictores, PCR crea CP sin considerar los valores de la variable dependiente. Por este motivo, decimos que PLS es un método supervisado. Como consecuencia de esto, para lograr una misma R^2, necesitaremos un mayor número de CPs al usar la técnica de PCR.

- PCR crea unos nuevos ejes de coordenadas con las direcciones ortogonales que mejor expliquen los datos de las VI´s, mientras que PLS no cambia los ejes de coordenadas, sino que reorienta los datos para que el sistema cartesiano explique sus tendencias de la mejor manera posible.

- En PLS la regresión y reducción de la dimensión se realizan simultáneamente, mientras que en PCR no lleva a cabo una reducción de la dimensión hasta después de haber hecho la regresión.

- Para un mismo efecto, PLS necesitará menos variable latentes, resultando ser un método más parsimonioso que PCR.

- Con PLS la interpretación de los vectores de cargas resulta más sencilla.

- Además, PLS tiene más capacidad a la hora de resolver problemas no lineales.

Por lo anterior y, fundamentalmente, porque permite una máxima correlación de los CPs con la VD, elegiremos PLS frente a PCR.

◆ PLS vs PCR para hacer LDA

En (Barker & Rayens. 2003) se sugiere que PLS debería usarse preferentemente a PCA

para hacer una reducción de la dimensión dirigida a hacer una discriminación, cuando se dispone de un subconjunto de entrenamiento.

- **PLS vs PCA**

La siguiente tabla muestra las principales diferencias de PLS con PCA:

FUENTE: ELABORACIÓN PROPIA

	PLS	PCA
Datos	Entrada X, salida Y	Entrada X
Objetivo	Regresión múltiple y conversión de datos	Comprensión de datos
"Regression variables"	Variables latentes	PCs
Selección de variables, y validación	Varianza acum, PRESS, RMSEC	Var. acumu plot de eigenvalor
Monitorización	SPE, RMSEP, residuos de Y, LV scores	SPE, puntuaciones de PC

Ilustración 17: Diferencias entre PLS y PCA

Además, contrariamente a PCA, PLS es un método de aprendizaje supervisado, en cuanto a que permite inferir una función a partir de los datos de entrenamiento. PCA, sin embargo, puesto que no es capaz de distinguir entre las diferentes clases de las observaciones, no permitirá estimar el porcentaje de fallos o aciertos para cada clase al realizar pruebas de asignación de muestras.

- **LDA vs PCA**

A diferencia de PCA, LDA no es una técnica interdependiente porque no distingue entre variables dependientes y variables independientes, a menos que se le especifique.

Por todo lo anterior, se ha decidido que el mejor método para resolver el problema es PLS-LDA.

3.3.11. PLS-LDA (mínimos cuadrados parciales con análisis lineal discriminante)

Como hemos visto, LDA es una herramienta estadística para hacer reducción de la dimensión y clasificación de las observaciones en grupo. Sin embargo, en determinadas aplicaciones, las proyecciones de las direcciones no se pueden considerar óptimas. Por otro lado, puesto que se sabía que las técnicas de PLS y CCA (correlación canónica) estaban racionados, y que por otro lado, CCA estaba relacionado con LDA, cabía esperar que PLS tuviera una conexión directa con LDA.

A raíz de esta asociación y para solucionar el problema de las soluciones no óptimas mediante la técnica de LDA, se combina el método de PLS con el algoritmo LDA en lo que se conoce como método de PLS-LDA.

PLS-LDA se formalizó en (Barker & Rayens. 2003) y es una de las múltiples aplicaciones de LDA que se han propuesto en los últimos años para resolver problemas que no tenían solución directa mediante LDA.

PLS-LDA busca las proyecciones de las direcciones de LDA pero, al mismo tiempo, usa la información de PLS.

Como hemos visto, LDA permite obtener la óptima solución J (w) de manera que se maximice la siguiente expresión:

$$J(w) = \left[\frac{W^T S_B W}{W^T S_W W}\right]$$

El paso fundamental de PLS-LDA es que se usan unos datos de entrenamiento X, y los correspondientes vectores Y para calcular las proyecciones de LDA w_{lda} y el valor c_{lda}, donde $c = Xw_{lda}$, que es la proyección de los valores de X en la dirección w_{lda}. Luego se utilizan c_{lda} para calcular los coeficientes de regresión b_{pls} para obtener finalmente el valor de X en las w direcciones. $X = X - c_{lda} (c_{lda} c_{lda}^T)^{-1} c_{lda}^T X$. Esto se repite hasta que se satisface el número de CPs requeridos para resolver el problema.

El algoritmo de PLS-DA puede entenderse como los parámetros relevantes de PLS, que se encontrarían mediante:

$$\text{argmin} \left(\left(\| Xbpls - clda \|^2 \right), s.t. bplsEk \right.$$

$$c_{lda} = \text{LDA}(X,y)$$

, donde:

- b_{pls} son los coeficientes de regresión PLS

- c_{lda} son las proyecciones de los valores x en las direcciones de LDA.

- LDA(X,y) es el algoritmo LDA para resolver problemas de proyección, mientras que X e y son los parámetros de entrada.

Obtendremos PLS-LDA de la siguiente manera a partir de los datos de entrada:

DATOS DE ENRADA: X, y, número de CPs de PLS, número de CP de PLS-LDA

DATOS DE SALIDA: la matriz de coeficientes $B_{lda\text{-}pls}$, y las direcciones de las proyecciones $W_{lda\text{-}pls.}$

1. $B_{lda\text{-}pls} = \nu\ W_{lda\text{-}pls} = \nu$

2. $(c_{lda}, W_{lda}) = LDA\ (X,y)$

3. $b_{pls} = PLS(X,c_{lda})$

4. $B_{lda\text{-}pls} = [B_{lda\text{-}pls}; b_{lda\text{-}pls}]$; $W_{lda\text{-}pls} = [W_{lda\text{-}pls}; W_{lda\text{-}pls}]$

5. $X = X - c_{lda}\ (c_{lda}\ c_{lda}{}^{T})^{-1}\ c_{lda}{}^{T} X$

6. Volver a paso 2 si el nº de iteraciones es menor que el nº de variables latentes; en caso contrario, ir al paso 7.

7. Salvar los coeficientes de la matriz $B_{lda\text{-}pls}$; y las direcciones de la proyección $W_{lda\text{-}pls}$

La idea original es usar $B_{lda\text{-}pls}$ como la dirección de la proyección del modelo. Entonces, las nuevas observaciones sólo necesitan proyectarse en la dirección $B_{lda\text{-}pls}$. Podemos usar este valor proyectado para clasificar los individuos en distintos grupos. La mejor proyección de la dirección puede localizarse entre dos direcciones de PLS-LDA correspondientes a las variables latentes 1ª y 2ª. Podemos asumir una combinación lineal entre los coeficientes de regresión y las proyecciones de la matriz $W_{lda\text{-}pls}$.

Para aplicar PLS-LDA, se seguirá el método utilizado en (Boulesteix., *et al*. 2014), que consiste en la reducción de la dimensión mediante PLS y análisis linear discriminante aplicado sobre los componentes de PLS.

3.3.12. Potencia predictiva del modelo

- **Modelo de validación**

El modelo de validación es el proceso en el cual se decide si los resultados numéricos que cuantifican las relaciones hipotéticas entre las variables obtenidas en el análisis de regresión proporcionan una descripción aceptable de los datos. Si se prescindiera de una

validación del modelo, se podría producir un sobreajuste del modelo o una actuación pobre del mismo. En el caso de los problemas de clasificación, podríamos incluso obtener unos resultados parecidos a los que obtendríamos mediante una asignación aleatoria.

- **Porcentaje de asignaciones correctas**

En problemas de asignación es habitual expresar los resultados en forma de porcentaje de asignaciones correctas.

Entenderemos como asignaciones correctas los casos en los que al predecir la variable respuesta de una observación dada, acertemos. Puesto que en nuestros datos de observaciones, no contamos sólo con valores de los predictores , sino que además conocemos la variable respuesta, podemos comparar este resultado, que sería el real, con el obtenido o estimado a partir del modelo. Claro, que sólo podemos calcular dicho porcentaje sobre individuos cuya pertenencia conocemos, que son a su vez, las observaciones cuyos datos usamos para construir el modelo. Por otro lado, para que los resultados obtenidos puedan ser extrapolados a nuevas observaciones, necesitaríamos contemplar también individuos ajenos al modelo. Razón por la cual se hace Validación Cruzada.

Validación Cruzada

Es una técnica de modelo de validación para valorar cuanto se ajusta el modelo a un subconjunto de datos independientes.

La idea básica es no incluir en el modelo un subconjunto de observaciones. De este modo, podrá comprobarse después como se comporta el modelo con ellas, con la ventaja de que conocemos la respuesta y podremos puntuar la actuación. De esta manera, se consigue que el porcentaje de asignaciones correctas sea sólo sobre individuos que no se usaron en la elaboración del modelo. Sin embargo, para que la información de todos los predictores siga siendo igualmente válida, la validación cruzada no hará un único modelo sino varios. De esta manera, en cada uno de los modelos una serie de observaciones independientes a la construcción del modelo actuarán como fase de prueba; mientras que el resto de las observaciones, que sí son consideradas en el modelo, constituirán una fase de entrenamiento específica para ese modelo. Las observaciones disponibles irán rotando de la fase de entrenamiento a la de prueba, con carácter aleatorio.

- **Pruebas de bondad de ajuste**

R-cuadrado (R^2)

En modelos de regresión se utiliza, fundamentalmente, la R^2 para definir la bondad de ajuste. R-cuadrado es el "porcentaje de varianza explicada" por el modelo y representa el porcentaje de variación en la variable dependiente que es explicado por las variables independientes. Dicho de otra manera, representa la proporción en la que la varianza de los errores es menor que la varianza de la variable dependiente.

La solución que ofrece mínimos cuadrados pasa por maximizar la R-cuadrado. Por otro lado, cuando descomponemos las variables en variables latente, el máximo valor de R-cuadrado se alcanzará cuando todos los CPs estén incluidos en el modelo. Sin embargo, un valor elevado de R-cuadrado no garantiza necesariamente que el modelo se ajuste bien a los datos, puesto que podría darse una malinterpretación de la función, o podría haber outlires que distorsionaran la estructura de los datos. Además, está el problema de que la R^2 tenderá a aumentar incluso cuando se añaden variables muy poco relevantes, y podríamos construir modelos poco parsimoniosos. Si bien, este problema se soluciona utilizando la R^2 -ajustada (R^2-adj), que penaliza las variables cuya aportación a la R^2 del modelo es poco significativa.

3.3.13. Implementación del modelo en R

En este apartado se diferencian cinco puntos fundamentales:

- Preparación de los datos

- Selección de variables PLS ("variable.selection")

- Análisis discriminante con PLS

- Potencia predictiva y porcentaje de aciertos

- Obtención de la R^2 ("plsr")

- **Preparación de los datos**

Se van a analizar los datos en R (R Core Team. 2014). R es un lenguaje de programación utilizado fundamentalmente para la elaboración de programas y aplicaciones en el campo de la estadística.

El análisis en R empieza con la importación del archivo PED extraído de PLINK. Puesto que PLINK no cuenta el número de copias de los alelos, sino que establece las clases categóricas "1" (si se trata del alelo más infrecuente), y "2" (cuando se trata del alelo más común), tenemos información de cada alelo; es decir, tenemos dos datos por cada SNP. Ambos datos aparecen separados por un espacio, por ejemplo "1 1", significa que el SNP presenta alelos infrecuentes en ambos cromosomas homólogos.

Hay que tener en cuenta que, en este caso, PLINK no ha distinguido entre fases,

y si en un SNP se dan ambas copias de alelos, interpretará que ese SNP es “1 2”, con independencia de a que cromosoma pertenezca cada uno. Dicho de otra manera, si existiera un SNP cuyos alelos fuesen “2 1”, ese SNP también recibiría en PLINK la notación “1 2”. Sin embargo, para evaluar la capacidad predictiva, consideramos las variables como vectores, y por tanto, deberemos tener un único valor por variable. De modo que hacemos la siguiente transformación, teniendo en cuenta que se cuenta el número de alelos comunes: alelos “2”:

SNPs con “2 2” → 2 (2 copias del alelo común: alelo “2”)

SNPs con “1 2” → 1 (1 copia del alelo común: alelo “2”)

SNPs con “1 1” → 0 (0 copias del alelo común: alelo “2”)

Para hacer estas transformaciones usamos AWK. AWK es un lenguaje de programación que apareció en 1977, aunque empezó a cobrar más importancia tras su última revisión en 1985 (Aho., *et al*. 1988). Este lenguaje se sigue utilizando en la actualidad para trabajar con datos basados en texto, fundamentalmente para manejo de ficheros y flujo de datos. En este caso vamos a realizar una transformación en un archivo .ped. Usamos el siguiente bloque de comandos:

```
awk '{gsub("2 2","2");print}' 3abrilDone.ped > 513poster_tmp.ped

awk '{gsub("1 2","1");print}' 513poster_tmp.ped > 513poster2_tmp2.ped

awk '{gsub("1 1","0");print}' 513poster2_tmp2.ped > 3abrilDone_tmp3.ped
```

Que, por ejemplo, en el primer caso, se interpretaría como: cogemos el archivo “*3abrilDone*” , que fue el último archivo creado con PLINK, y reemplazamos los “*2 2*” (“2” “espacio” “2”) por “2”. Creamos un archivo de nombre “*513poster_tmp.ped*”.

Una vez transformados los datos, los importamos en R mediante el siguiente comando:

```
Tmil<-read.csv("3abrilDone_tmp3.ped", header=FALSE, sep="\t")
```

Interpretaríamos lo siguiente:

Importamos el último de los archivos generados con AWK, eliminamos la cabecera de la tabla puesto que queremos que R asigne los nombres de las variables, de una manera ordenada, y consideraremos que el separador es "\t".

Ahora disponemos en R del marco de datos "*Nmil*" $_{513 \text{ x } 90712}$. Como es habitual en R, cada fila es una observación y cada columna es una variable. Se describen a continuación las variables que se corresponden con las columnas del marco de datos:

FUENTE: ELABORACIÓN PROPIA

1ª columna	ID familiar: variable población: variable respuesta categóricas
2ª C	ID individual
3ª C	ID maternal
4ª C	ID paternal
5ª C	Sexo del animal
6ª C a 90717º C	Genotipo: cada columna contiene información de un SNP

Tabla 13: Columnas del marco de datos en R

Las columnas 2ª a 5ª no son de interés para estudiar la capacidad predictiva, de modo que las eliminamos y obtenemos un marco de datos (513 x 90712), donde la 1ª columna contiene información de la raza, y el resto de columnas son variables continuas que contienen (0, 1 ó 2) para cada SNP.

Se muestra a continuación un fragmento de "*Nmil*":

FUENTE: ELABORACIÓN PROPIA

	V1	V2	V3	…	V90712
Obs 1	1	1	2		2
Obs 2	1	2	1		2
	…				
Obs 75	1	2	0		2
Obs 76	2	2	2		2
Obs 77	2	2	2		2
	…				
Obs 147	2	2	2		2
	…				
Obs 445	7	1	1		0
Obs 446	7	2	2		2
	…				
Obs 513	7	2	2		2

Tabla 14: Ejemplo de marco datos en R

Donde "*Obs*" son las observaciones y "V" las variables. Naturalmente, es de esperar encontrar que:

nº observaciones con "2" >> nº observaciones con "1" >> nº observaciones con "0"

Para los cálculos posteriores, descomponemos dicho marco en un vector "población" (513 x1) con la raza correspondiente, aunque de tipo continuo y una matriz "m" (513 x 87918) con las columnas correspondientes a los datos genómicos. A continuación, obtenemos un vector de tipo factorial "*pobla*", con los datos de "población". Una vez hecho esto, tenemos los datos preparados para el análisis.

- **Selección de variables PLS ("variable.selection")**

En primer lugar se seleccionan las mejores variables mediante PLSR. Para ello cargamos el paquete de R "plsgenomics" (Boulesteix., *et al*. 2014) y ejecutamos el comando *variable.selection*.

Al ejecutarlo se implementa una regresión por mínimos cuadrados parciales y se extrae un componente principal, que será el primero; es decir, el que explica un mayor porcentaje de la varianza.

Sabemos que cuando se extrae un Componente Principal las variables originales tendrán un coeficiente de ajuste con este, que serán los pesos correspondiente en una ecuación de regresión en la que el CP sería la variable respuesta.

Este coeficiente podrá ser negativo o positivo, pero es el valor absoluto el que va a determinar si una variable es o no relevante en la constitución del CP. De manera que, ordenando las variables según orden decreciente del valor absoluto de estos coeficiente, estaremos ordenando las variables según mayor a menor correlación con el primer componente principal.

Este orden de variables ha sido contrastado con otros métodos y coincide con el que se obtendría mediante el estadístico-F, y el T-test, si se usaran las mismas variables (Boulesteix. 2004).

Un inconveniente de esta función es que sólo resolverá completamente el problema cuando la clasificación que se desea efectuar sea binaria, es decir, cuando la variable respuesta tenga únicamente dos categorías: 1 y 2. Recordemos que con PLS, la variable repuesta es tenida en cuenta a la hora de extraer los componentes principales. Si bien, para este comando la variable respuesta podrá tener únicamente dos posibles clases. Así, las observaciones deberán tomar los valores categóricos 1 y 2, y será indiferente lo que represente cada uno. El CP tendrá la única función de distinguir las observaciones entre una y otra clase.

Para ejecutar el comando *variable.selection* hará falta:

- Una matriz X (n x p), con la información de las variables que consideramos como predictores. En nuestro caso con 87918 variables.
- Un vector Y (n x 1), con las clase a la que pertenecen las diferentes observaciones. Las clases van codificadas como “1” o como “2”.
- Un dígito (nvar) con el que indicamos cuántas variables queremos extraer (las mejores).

La matriz X contendrá todos los predictores y para cada uno de ellos, las observaciones toman los valores continuos 0, 1 o 2, según el caso. Para obtener el vector Y, creamos una variable dummy para cada raza; de manera que para la variable dummy de la raza r_i, las observaciones que pertenecen a la raza r_i toman valor “1”; y las que no pertenecen, toman valor “2”. Así, en la matriz constituida por las variables dummy, cada observación tomará un valor “1” (en la variable dummy de su raza); y un valor “2” en las otras 10 variables dummy. En cuanto a “nvar”, el número de variables a extraer será el mismo para cada raza y se irá probando con distintos valores hasta encontrar el número más óptimo de variables. Este número óptimo quedará determinado por el porcentaje de aciertos, como veremos más adelante.

Obtenemos un vectores de longitud (nvar) o “p” (si nvar = NULO), que contiene los índices de las variables a seleccionar. Las variables aparecen ordenadas de mejor a peor. Resulta fundamental aclarar que al ordenar las variables según el criterio de PLS, no se están considerando interacciones dentro del grupo de selección. Lo que quiere decir que si en lugar de seleccionar, por ejemplo, 500 variables seleccionáramos 1000,

las 500 primeras variables de la nueva selección serían las mismas, y en el mismo orden, que cuando seleccionamos 500. Visto con otro ejemplo, si eliminámaramos la mejor variable, la segunda mejor pasaría a ser la mejor, la tercera mejor pasaría a ser la segunda mejor, y así sucesivamente; el orden seguiría siendo el mismo. Por otro lado, por razones de computación, el comando *variable.selection* no utiliza el algoritmo de PLS, sino que el orden de las variables que se consigue es exactamente el obtenido usando los pesos de PLS que se obtendrían mediante la función *pls.regression*.

La función *pls.regression* implementa PLS para regresiones múltiples, usando el algoritmo Jong's SIMPLS (Jong. 1993). Debe de tenerse en cuenta que, las columnas o vectores con los datos, no deben estar centradas para tener media cero, sino que este ajuste se hará directamente al ejecutar el algoritmo.

Dada la definición de SIMPLS, los vectores de peso tienen longitud 1. Esto es así para que se satisfaga el criterio de optimalidad (Jong. 1993).

Mediante *variable.selection* extraeremos un único CP, que corresponde a la dimensión que mejor resume las VI´s. Una vez extraído, se seleccionan los SNPs que mejor se le ajusten, ya que estos serán los que tengan mayor capacidad predictiva. Como era de esperar, todos los SNPs se ajustan poco al CP, dado que p >> c. De hecho, llegados a este punto, conviene comentar el dogma central del modelo infinitesimal (Fisher. 1919). Este modelo tan importante en genómica , asume que los caracteres están determinados por un número infinito de loci aditivos no ligados con un efecto infinitesimal cada uno.

Los SNPs seleccionados tendrán, sin embargo, la mejor capacidad predictiva y hará falta un número menor de predictores para satisfacer un determinado baremo de aciertos en la predicción.

Ejecutaremos el comando *variable.selection* once veces (una por cada raza), obteniendo once listas de variables seleccionadas con los SNPs ordenados según su capacidad predictiva.

Puesto que nuestro interés es tener una única lista de variables, fusionamos las once listas en un único objeto “índices”. Además, puesto que alguna variable podría haber sido incluida en más de una lista de selección, como podría ser el caso de que un SNP fuese muy explicativo para más de una raza, eliminaremos las variables que pudieran estar repetidas.

Se han usado los siguientes comandos:

```
Nmarcadores = 20
```

Llamamos "nmarcadores" al número de predictores a seleccionar en cada raza. Al establecer "*nmarcadores*" de esta manera, resulta muy sencillo cambiar el valor para hacer las diferentes pruebas.

```
indices = unique ( c(sapply(1:11, function(i)
variable.selection(m,1+
(pobla==paste(i)),nvar=nmarcadores))))
```

Llamamos "*indices*" a un conjunto de once listas de valores únicos (*unique*) que corresponden a las mejores variables para cada una de las razas obtenidas mediante el comando *variable.selection* (predictor, respuesta, nvar). *variable.selection* se efectúa once veces, progresivamente desde la raza una hasta la once (*sapply*).

El predictor es en todos los casos la matriz "*m*" de predictores, mientras que la variable respuesta se corresponde, en cada caso, con una de las clases de la variable categórica "*pobla*". El nvar se corresponde con el valor de nmarcadores definido en la linea anterior y es también el mismo en todos los casos.

- **Análisis discriminante con PLS (PLS-LDA)**

Para que esas variables altamente informativas sean lo más útiles posible a la hora de asignar las observaciones a sus correspondientes razas, deberemos reorientar las proyecciones de las variables a tales efectos. Siguiendo así el método descrito en (Boulesteix. 2004) en el cual se implementa PLS para reducir la dimensión, y después de ejecuta LDA sobre los componentes de PLS.

Dicha reducción de la dimensión mediante PLS está, efectivamente, recogida en nuestro objeto "*índices*". De manera que aplicamos LDA con las variables que figuren en "índices". Para ello, con el mismo paquete de R "plsgenomics", ejecutamos el comando "*pls.lda*". Esto nos permitirá, a partir de la selección de predictores, encontrar las proyecciones de las direcciones que más faciliten la distinción entre razas.

La función "*pls.lda*" procede de la siguiente manera para predecir la clase de las observaciones de la fase de test:

- En primer lugar se corre el algoritmo SIMPLS sobre 'Xtrain' e 'Ytrain' para determinar los componentes PLS, basándonos únicamente en la fase de entrenamiento.

- Después, se proyectan los puntos de la fase de prueba sobre esos componentes y se realiza la clasificación aplicando LDA a esos nuevos componentes. Por su parte, el clasificador LDA también queda establecido teniendo en cuenta sólo los

datos de la fase de entrenamiento.

Para ejecutar "*pls.lda*" hace falta completar los siguientes argumentos:

- Xtrain: Hay que especificar una matriz (ntrain x p) que contiene los predictores para la fase de entrenamiento. Cada fila es una observación, y cada columna un predictor. En nuestro caso, y puesto que vamos a hacer validación cruzada, no vamos a predefinir una fase de entrenamiento y un fase de prueba, luego esta matriz contendrá todas las observaciones. Por otro lado, sólo utilizaremos los predictores que sean los mejores para cada raza ("*índices*"), y cuya cantidad podemos regular simplemente con cambiar "nvar" y volver a ejecutar *variable.selection*. Así, nuestra matriz Xtrain será de dimensiones (726 x nmarcadoresx11)

- Ytrain: Un vector de longitud ntrain en el que se dan las clases a las que pertenecen las observaciones de la fase de entrenamiento. Como en el caso de la matriz anterior Xtrain, al especificar Ytrain usamos todos las observaciones. Las clases irán codificadas del 1 al 11. Las dimensiones del vector son (726 x 1)

- ncomp: Si 'nruncv=0', 'ncomp' es el número de componentes latentes a considerar en la reducción de la dimensión por PLS. Mientras que si 'nruncv>0', se usa CV Boulesteix (2004) para seleccionar el mejor número de componentes. Puesto que queremos conocer la potencia predictiva de nuestras variables originales, utilizaremos todo su potencial discriminante. Por tanto, lo lógico sería usar todos los Cps, sin embargo, puesto que a partir de cierto número de CPs la ortogonalidad de los mismos empieza a ser cuestionable y pueden aparecer problemas de colinealidad, pondremos algún CP de menos (restando una unidad al producto de 11 razas por nmarcadores/raza) : ncomp=11*nmarcadores-1

- nruncv: Es el número de iteraciones de la VC que se efectúan para la selección de los componentes latentes. Si 'nruncv=0', no se efectuaría validación cruzada y se usarían 'ncomp' componentes latentes.

Efectuando este tipo de VC, los individuos que forman parte de la fase de entrenamiento y la fase de prueba irán rotando aleatoriamente con cada iteración, y los resultados variarán ligeramente. El número de iteraciones de VC, deberá ser tal que al volver a ejecutar de igual manera el comando, obtendríamos una diferencia no significativa. Esto dependerá entre otras cosas, del número de predictores, que en nuestro caso, es un número reducido. Después de probar, se ha decidido fijar el número de interacciones en 20 (nruncv=20).

Se ha usado el siguiente comando:

```
mi.pls.lda = pls.lda (m[indices], pobla,
ncomp=11*nmarcadores-1, nruncv=20)
```

Creamos el objeto "*mi.pls.lda*", que corresponde a la regresión PLS-LDA, en la que los predictores son aquellos que, perteneciendo a *m*, figuran en el conjunto de listas que hemos definido como "*índices*". La variable respuesta categórica es pobla (con los datos de raza o población), el número de componentes a extraer es 11 veces *nmarcadores* menos 1, y el número de iteraciones de validación cruzada es de 20. Recordamos que *nmarcadores* quedó definido en la primera línea de comandos de R, y que "*m*" es la matriz de predictores.

- **Determinar la proporción de aciertos**

Para determinar la proporción de aciertos se tienen en cuenta las mismas observaciones de la fase de entrenamiento que en la fase de pls.lda, aunque los resultados se podrán extrapolar a nuevas observaciones puesto que se ha realizado VC.

Los resultados obtenidos en la etapa anterior mediante pls.lda, son comparados con los datos correctos de la variable respuesta categórica "*pobla*".

Para expresar la proporción de aciertos se usa el comando:

```
mean(mi.pls.lda$predclass == pobla)
```

Se hace la media del número de valores que correspondiendo al vector que contiene las clases predichas de todas observaciones al obtener "*mi.pls.lda*", coinciden con los valores del vector "*pobla*".

Para especificar cuando podemos considerar que una colección de SNPs es válida para hacer predicciones con nuevas muestras, se ha decidido establecer como valor umbral un porcentaje de aciertos mínimo del 95% con 20 iteraciones de validación cruzada ("nruncv=20"). Para ello se ha tenido en cuenta (Fisher. 1954). Según Fisher, un experimento se puede considerar exitoso cuando falla menos de 1 vez de cada 20. Además se ha considerado que, hasta llegados a un punto, cuantas más iteraciones de VC se hagan, más fiables serán los resultados.

En el (Anexo II – Manipulación de los ficheros de datos) se da más información sobre otros comandos utilizados y sobre los diferentes bloques de comandos que se utilizan en este estudio.

- **Obtención de la R-cuadrado del modelo con PLS**

Podemos saber cual es la R-cuadrado de dicho subconjunto de variables usando el paquete de R "pls" (Bjørn-Helge., *et al.* 2013).

Con la función *plsr*, implementamos PLSR de la variable "*pobla*" sobre la matriz con los predictores seleccionados. Si bien, en este caso, consideraremos únicamente los datos de la fase de entrenamiento. Para ello, ésta debe quedar previamente establecida.

Para determinar las observaciones que formarán parte de esta fase se hará mediante un "muestreo sin repetición". Para ello, ordenamos todas las observaciones con un orden aleatorio y después seleccionamos las 400 primeras. De esta manera, contaremos con una fase de entrenamiento de 400 individuos y una fase de prueba de 326 observaciones, elegidos en ambos casos de manera aleatoria. Naturalmente, si sumamos los individuos de ambas fases, tenemos 726 individuos, que es el total de individuos analizados. Para los individuos de la fase de prueba, utilizaremos sólo los predictores, y la variable respuesta se usará únicamente para consultar el resultado y determinar la capacidad predictiva del modelo. No obstante, además de hacer esta diferenciación entre fase de entrenamiento y fase de prueba, hemos decidido aplicar el método LOO (*leave one out*) de VC, con los datos de la fase de entrenamiento para evaluar la capacidad predictiva del modelo. Esto nos permitirá probar la eficacia con individuos cuya información de la variable respuesta es considerada en otros modelos, pero no en el modelo particular en el que dichos individuos fueron probados.

El argumento debe tener la forma: respuesta ~ predictor. El predictor, puede ser más de una variable, separados por "+". Por tanto, aportaremos la forma: **y ~ X + Z**; o **y ~ m,** donde "m" es una matriz que contiene (X, Z).

Al contrario que el comando *variable.selection*, *plsr* si se puede ejecutar con variables multi-respuesta, aunque esta deberá tener la forma de varias respuestas binarias, razón por la cual construiremos unas variables dummy en la que las observaciones tomarán valor "1" si pertenecen a la raza correspondiente, y valor "0" en caso contrario. Recordamos que hace falta una variable dummy por cada variable. Para crear las variables dummy, lo hacemos a partir de la variable "*pobla*", que es de tipo factorial.

El valor de R-cuadrado se obtiene de la expresión:

$$1 - SEE/SST$$

, donde:

SST es la suma de los cuadrados de la respuesta corregida

SSE es la suma de los errores cuadrados del subconjunto de predictores.

En el caso de los datos de la fase de entrenamiento, este valor de R^2 se corresponde con la correlación entre los valores reales de Y, y los valores estimados de Y:

$$\text{corr}\ (Y, \hat{Y})$$

Aunque esto no es así en el caso de la fase de prueba. De hecho el valor de la R^2 , prácticamente, no tiene sentido al referirnos a la fase de prueba.

Para la obtención de la R2 en R se introducen una serie de comandos que se detallan en el (Anexo II)

5. RESULTADOS

Eficacia del genotipado

En este estudio se ha podido comprobar que la proporción de marcadores que pudieron ser leídos en más del 95% de las observaciones ha sido del 98,2519 %. Lo cual es una prueba de la eficacia del chip para genotipar las razas estudiadas. Adicionalmente, la prueba realizada para detectar individuos con problemas de genotipado ha revelado que a ningún individuo le falta información en más de un 10% de los SNPs, y por tanto, los 726 individuos analizados se han tenido en cuenta en el estudio.

Filtrado de los SNPs en base a criterios genéticos

Como se ha comentado, se establecen unos valores umbrales para unos criterios que nos van a permitir eliminar los SNPs que no cumplan unas determinadas características genéticas. Estos SNPs quedarán fuera de la selección puesto que la información que puedan proporcionar no es la más apropiada para hacer nuevas predicciones.

La siguiente tabla muestra los SNPs eliminados con cada criterio:

FUENTE: ELABORACIÓN PROPIA

	Nº total de SNPs leídos de BeadChip		702422	
Criterio	Valor Umbral	Orden aplicación	SNPs eliminados	% del total
LD	50 5 2	1º	579137	0.82
HWE	0.001	2º	14638	0.02
GENO	0.1	2º	0	0.00
MAF	0.01	2º	21268	0.03
		SNPs seleccionados	87918	0.13

Tabla 15: Eliminación de SNPs en base a criterios genéticos

De la tabla se deduce que se seleccionaron un total de 87918 SNPs para realizar con ellos las prueba estadística que nos permite determinar cuales son los mejores predictores. Este número se corresponde con un 13% del total de los SNPs que se leyeron en la prueba de genotipado, que recordamos que fueron 702422.

El criterio más eliminatorio, como es habitual en los estudios GWAS, fue el de DL. En base a este criterio se eliminaron 579137 SNPs, que son en torno al 82% de los marcadores que se leyeron. Con los criterios de HWE y MAF se eliminaron 14638 y 21268 SNPs, respectivamente; mientas que con el criterio de GENO no se eliminó ningún SNP. De hecho, probablemente ésto sea lo más llamativo de los resultados del análisis genético. A pesar de que se estableció un valor umbral muy riguroso (*--geno 0.0*) ningún SNP incumplió el criterio. De modo que queda demostrado que todos los SNPs que superaron el criterio de DL que, como ya se ha comentado se aplica con anterioridad al resto de los criterios, contenían información para todos los individuos. Dada la relevancia de este resultado, posteriormente se decidió probar a aplicar el filtro (*--geno 0.0*) directamente sobre el total de los SNPs que se leyeron en la prueba de genotipado. Es decir, prescindiendo de un filtrado previo en base a DL. En la prueba resultó que tampoco en esta situación se eliminaron SNPs en base al criterio GENO. Siendo ésta la mejor prueba posible de la eficacia del genotipado para las razas del estudio.

En cuanto al número tan reducido de SNPs eliminados con el criterio de HWE, los resultados sirven como prueba de que, para las diferentes poblaciones de *bos taurus*, se da una situación general de equilibrio.

El número de SNPs eliminados con el criterio de MAF es también reducido. Sin embargo, el resultado no es comparable con el de otros estudios puesto que no es habitual utilizar valores umbrales tan reducidos. Recordemos que en este estudio fijamos (*--maf 0.03)* debido a que disponíamos de un número de observaciones considerablemente más reducido para una de las razas. Esta raza era la raza francesa GUER, de la que sólo se analizan 28 individuos, que son sólo un 3.8% del total de individuos analizados.

El número de SNPs que se conservan tras la aplicación del criterio de DL es 123285, que son un 17.55% del total de SNPs que se analizaron. Como se ha comentado en el apartado de (Material y Métodos) la suma de los SNPs conservados y eliminados es superior al número de SNPs de partida puesto que en el caso de los filtros de HWE, GENO y MAF algunos SNPs cuentan como eliminados en más de uno de los filtro. Por tanto, no se cumple la siguiente igualdad:

SNPs conservados al final del análisis genético + SNPs eliminados según criterio de MAF + SNPs eliminados según criterio de HWE + SNPs eliminados según criterio de DL = SNPs leidos en el chip.

Para nuestros valores:

SNPs conservados al final del análisis genético (87918)

SNPs eliminados según criterio de MAF (21268)

SNPs eliminados según criterio de HWE (14638)

SNPs eliminados según criterio de DL (579137)

SNPs leídos en el chip (702422)

Vemos que no se cumple:

$$87918 + 21268 + 14638 + 579137 = 707961 > 707422$$

Descubrimos que hay 539 SNPs (707961 – 707422) que se eliminan por más de un criterio de los siguientes: HWE, GENO y MAF

Por otro lado, si se cumple que la suma de los SNPs eliminados en base a DL y conservados en base a este criterio equivalen al total de SNPs considerados:

$$123285 + 579137 = 707422.$$

En el primer sumando aparecen los conservados tras aplicación del criterio de DL (que van al fichero *plink.prune.out)* y en el segundo término aparecen los eliminados (*plink.prune.in).*

Con el objetivo de determinar si existía algún cromosoma para el que la situación de DL fuese significativamente diferente del resto, se ha determinado el número de SNPs eliminados en base a este criterio para cada cromosoma. En la siguiente tabla se muestra la proporción de SNPs que se conservaron tras a aplicación de este criterio:

FUENTE: ELABORACIÓN PROPIA

Cromosoma	1	2	3	4	5	6	7	8
	16,42	16,41	16,95	17,55	16,37	16,02	16,98	17,13
	9	10	11	12	13	14	15	16
	16,89	18,13	16,05	17,6	19,03	18,05	17,9	17,33
% de SNPs CONSERVADOS tras prueba DL	17	18	19	20	21	22	23	24
	17,78	18,2	19,7	17,32	17,12	18,53	20,83	17,55
	25	26	27	28	29			
	19,77	18,93	19,59	20,1	19,37			

Tabla 16: Porcentaje de SNPs conservados en base a DL en cada cromosoma

De la tabla se puede deducir que la situación de DL es bastante aproximada para todos los cromosomas. Adicionalmente, se puede observar que el DL es menor en los últimos cromosomas, como se aprecia por una proporción mayor de SNPs conservados y es mínimo en el cromosoma 23. También se concluye que en los primeros cromosomas el DL es mayor.

Selección de los SNPs mejores predictores mediante un análisis PLS-LDA

De entre los SNPs que se conservaron tras el análisis genético se seleccionaron posteriormente aquellos SNPs con mejor capacidad predictiva para la asignación de muestras a las razas del estudio y se comprobó que hacían falta 132 SNPs para lograr un porcentaje de aciertos superior al 95%. Estos 132 SNPs corresponden a los 12 mejores predictores para cada raza (12 mejores x 11 razas = 132). Los resultados se obtuvieron efectuando validación cruzada con 20 iteraciones.

Otros resultados obtenidos fueron:

- El número mínimo de SNPs para lograr un 90% de aciertos fue de 77. Es decir, los 7 SNPs más informativos para cada raza.
- El número mínimo de SNPs para lograr un 96% de aciertos fue de 275, correspondientes a los 25 SNPs más informativos para cada raza.

Adicionalmente, se ha comprobado que la potencia predictiva del modelo es mayor que la que se habría obtenido en el caso de haber usado el mismo número de predictores SNP pero elegidos éstos al azar de entre los que superaron los criterios genéticos. En la prueba realizada resultó que con 132 SNPs elegidos aleatoriamente de entre los 87918 que se conservaron tras el análisis genético, el modelo era capaz de asignar correctamente un 89,25 % de los individuos, frente al 95.32% que se lograba al elegir los predictores mediante el método de PLS-LDA. Así mismo, sabemos que el

número de SNPs necesarios para lograr un porcentaje de aciertos superior al 95% asciende a 195 en el caso de elegir aleatoriamente los predictores, tal y como se puede apreciar en la siguiente tabla comparativa:

FUENTE: ELABORACIÓN PROPIA

% mínimo de aciertos	nº mínimo de SNPs	
	Elegidos en base a PLS-LDA	Elegidos aleatorio
90	77	100
95	132	195
96	275	310

Tabla 17: Comparación con elección aleatoria de marcadores

Como puede deducirse de la tabla, cualquiera que sea el porcentaje mínimo de aciertos exigido, el método de selección mediante PLS-LDA ofrece una solución con un menor número de SNPs necesarios. La diferencia más significativa de SNPs necesarios en caso de seleccionar o no los predictores se da para un 95% mínimo de aciertos. Este porcentaje mínimo de aciertos es precisamente el nivel de exigencia en la predicción que se ha decidido considerar, tal y como se explica en el apartado de (Material y Métodos). Para lograr este porcentaje de aciertos hacen falta 132 SNPs en el caso de seleccionar los predictores mediante PLS-LDA y hasta 195 SNPs en el caso de no hacer selección. Es decir, aplicar el método planteado en este estudio permite reducir en un 47% el número de predictores.

Puesto que se exige un 95% mínimo de aciertos para evaluar si una selección de predictores es apta para resolver el problema de la asignación, podemos deducir de la (tabla 17 - Resultados) cual es el número mínimo de SNPs que son necesarios para resolver el problema de la asignación. Se puede concluir que el número mínimo de predictores SNP exigidos para este estudio es 132.

En el (Anexo I – Marcadores SNP seleccionados) se incluye una tabla con los 132 SNPs seleccionados y los genes que se han localizado en los respectivos loci. La selección de variables mediante el método de PLS-LDA ofrece una solución única. De modo que, bajo las condiciónes de que se apliquen los mismos criterios genéticos y de que las muestras sean representativas, se presupone que la selección de marcadores volverá a ser la misma para las once razas del estudio en caso de que se quisiera resolver con otras muestras. Este carácter de solución única que ofrece PLS-LDA es trivial para comprender la importancia de los resultados, puesto que resulta fundamental entender que existe una correlación directa entre los SNPs seleccionados en este estudio

y los loci que están más asociados con la diferenciación de las razas en general.

Zonas de interés en los cromosomas

Lo primero que hay que tener en cuenta es que, según el dogma del modelo infinitesimal, los SNPs tienen todos un efecto muy reducido sobre los caracteres causantes de la diferenciación entre las razas. Esta forma en que viene dada la información es propia de los caracteres cuantitativos. De estos caracteres habitualmente se deduce que dependen de un número muy elevado de predictores, todos ellos con un efecto infinitesimalmente pequeño. Frente a éstos están los caracteres cualitativos que, a diferencia de los cuantitativos, dependen de pocos caracteres con un gran efecto individual. Por este motivo, dados los resultados, tenemos motivos para pensar que los caracteres más implicados en la diferenciación de razas son, fundamentalmente, de tipo cuantitativo. Al ser el efecto de los caracteres tan reducido, la capacidad predictiva en caso de seleccionar los mejores predictores será sólo algo mayor que la que obtendríamos si eligiéramos los SNPs al azar.

De acuerdo con esto se han hecho diversas pruebas cambiando el valor umbral de DL y se ha comprobado que la selección final de SNPs cambiaba cuando alterábamos ese valor y que el cambio era más significativo conforme fijábamos valores de DL más alejados de los recomendados por PLINK. Esto ocurre así puesto que si somos excesivamente rigurosos, eliminamos muchos de los SNPs que son posibles buenos predictores. Sin embargo, tenemos motivos para pensar que en caso de no haber hecho un filtrado previo basado en criterios genéticos y de haber aplicado PLS-LDA directamente sobre los 702422 SNPs, no habríamos logrado una capacidad predictiva mayor. Los resultados no habrían sido mejores, fundamentalmente, por el problema de la multicolinealidad. Prácticamente sólo cambiaría que al prescindir del criterio de DL, la multicolinealidad tendría que ser resuelta en el análisis estadístico. Al calcular la matriz de correlaciones se analizaría la estructura interna de los datos (Martínez-Camblor., *et al*. 2014). Por otro lado, los SNPs que conservaríamos en caso de no aplicar el criterio de MAF tendrían, probablemente, información poco útil y no serían seleccionados en la prueba estadística. No se ha podido hacer dicha comprobación puesto que el software estadístico utilizado no permite trabajar a partir de un cierto número de predictores, que se estima está en torno a los 100000 y es necesario un filtrado previo en base a DL .

El dogma del modelo infinitesimal sugiere que los SNPs que seleccionemos como mejores predictores se encontrarán repartidos por todo el genoma. Tras representar el subconjunto de SNPs sobre el cariotipo con la ayuda de Ensembl comprobamos que esto ocurría así, tal y como se puede apreciar en la siguiente ilustración:

FUENTE: ELABORACIÓN PROPIA USANDO Ensembl (Cunningham., *et al*. 2015)

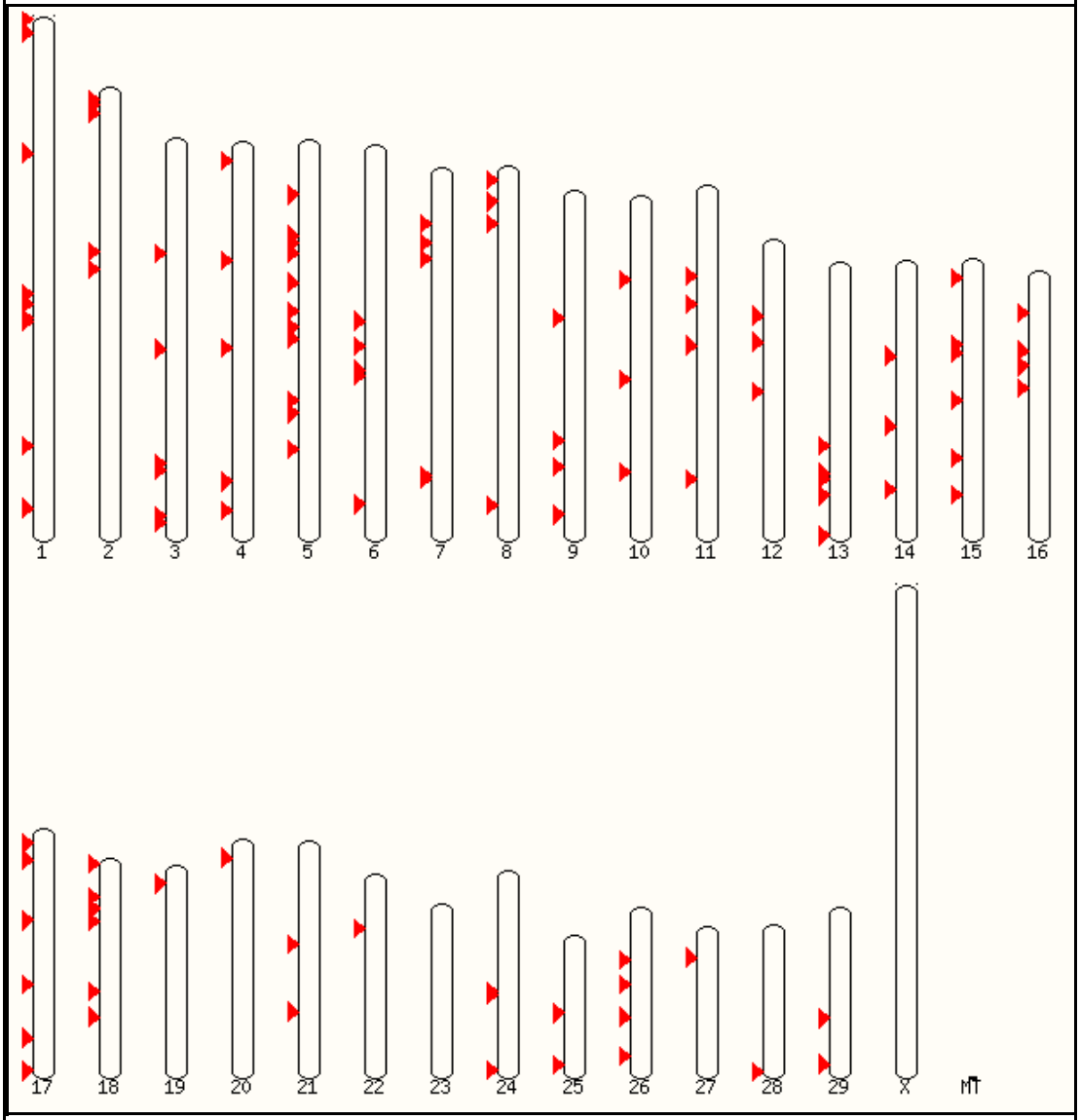


Ilustración 18: Representación de los SNPs seleccionados en el cariotipo

Es destacable de esta representación, además del reparto de los SNPs en los distinto cromosomas, el número de SNPs seleccionados que se encuentran en el cromosoma 5 (hasta 12 SNPs). Además sabemos que éstos se encuentran distribuidos por todo el cromosoma, exceptuando la sección central y los extremos.

Como era de esperar, en los cromosomas más cortos, puesto que se analiza un número menor de nucleótidos, la presencia de SNPs seleccionados es menor, aunque los cromosomas 17, 18 y 26 son excepciones a esto.

No se ha seleccionado ningún SNP en el cromosoma X, ni en el ADN

mitocondrial (MT) que aparece representado en la esquina inferior derecha de la imagen. Como ya se ha comentado, en este estudio no se disponía de información de ADN para estos casos.

A partir de esta representación y de la tabla que figura en el Anexo I y haciendo uso de las bases de datos de Ensembl (Cunningham., *et al*. 2015) y Genebank® (Benson., *et al*. 2004) hemos podido definir ocho regiones cromosómicas que son de especial interés en la diferenciación de razas. Se supone que estas regiones, puesto que concentran varios de los SNPs mejores predictores, son las que tienen mayor influenciar en la diferenciación de las razas. Por este motivo, es de especial interés conocer que genes se localizan en estas regiones.

Para definir las regiones de interés, se han considerado aquellas posiciones de los cromosomas en las que se concentran dos o más SNPs seleccionados. Se ha considerado que dos SNPs están lo suficientemente próximos como para constituir una región de interés cuando están separados por menos de 2M de pares de bases. Sin embargo, se ha sido más permisivo cuando se planteaba la opción de incluir un tercer o un cuarto SNP en alguna región que había sido ya previamente definida. La región de interés en la que figuran más SNPs es la región 2, en el extremo del cromosoma 2. En ella se han seleccionado hasta 5 SNPs bastante próximos.

Para especificar la posición aproximada en la que se localizan estas regiones de interés, se ha hecho referencia a la posición central del grupo de SNPs que constituyen la región.

Se ha considerado que lo más importante no son los genes que coinciden con las posiciones de los SNPs seleccionados, sino los que se localizan en estas regiones de interés. La información de los SNPs de las distintas regiones se ha extraído de la base de datos de Genebank. Se han identificado hasta 10 genes en estas regiones y a excepción del gen identificado en la primera región, se conoce la proteína para la que codifican los genes (Zimin., *et al*. 2009). Se ha añadido una descripción de estas proteínas cuando se habla de cada gen en particular. Dado el carácter complejo de la descripción, se ha decidido proporcionar la descripción original en inglés.

- **Región de interés 1: En la zona central del cromosoma 1**

En torno a la posición 1:91000000 se han seleccionado 3 SNPs muy próximos. La información más relevante aparece en la siguiente tabla:

Tabla 18: Región de interés 1

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs137370497	1:91013492	C/T		
rs42871708	1:91926201	A/G		
rs134158660	1:93205984	A/G	LOC101907397	Variante intrón

Sabemos que en esa región se da el gen LOC101907397, pero no se conoce la proteína para la que éste codifica.

El SNP cuya posición coincide con la del gen, es una variante intrón. Una variante intrón es una secuencia de nucleótidos que es retirada de un gen como consecuencia de una partición de la hebra de ARN durante la fase de maduración del producto final del ARN.

- **Región de interés 2: En el extremo superior del cromosoma 2**

Esta región es, de las 8 regiones de interés que se han definido, la que más SNPs contiene. Se seleccionaron hasta 5 SNPs a lo largo de una distancia de 17M de pares de bases, en torno a la posición 2:40000000, tal y como se puede apreciar en la siguiente tabla:

FUENTE: ELABORACIÓN PROPIA

Tabla 19: Región de interés 2

SNP	Posición	Alelo	Gen	Consecuencia funcional
rs109994609	2:3792579	C/T		
rs136309759	2:4121033	C/T	HS6ST1	variante intrón
rs136579166	2:5150666	A/C	MAP3K2	Codon sinónimo
rs134286256	2:49742611	C/T		
rs133243829	2:54944955	A/G	LRP1B	Variante intrón

Genes :

HS6ST1: heparan sulfate 6-O-sulfotransferase 1

MAP3K2: mitogen-activated protein kinase kinase kinase 2

LRP1B : low density lipoprotein receptor-related protein 1B

El SNP rs136579166 cuya posición coincide con la del gen MAP3K2 se ha identificado en Genebank como codon sinónimo. Un codon sinónimo es una sustitución evolutiva de una base por otra en un exón de un gen que codifica para proteína, de manera que el

aminoácido producido no se ve modificado.

- **Región de interés 3: En el cromosoma 3**

En torno a la posición 3: 113000000 se han seleccionado 3 SNPs muy próximos

FUENTE: ELABORACIÓN PROPIA

Tabla 20: Región de interés 3

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs42752357	3:115727220	C/T	CSMD2	variante intrón
rs135732729	3:114039417	A/G		
rs133790058	3:112311199	C/T		

Genes:

CSMD2: CUB y Sushi multiple domains 2

- **Región de interés 4: En el cromosoma 5**

En torno a la posición 5:29000000 se seleccionaron 3 SNPs muy próximos:

FUENTE: ELABORACIÓN PROPIA

Tabla 21: Región de interés 4

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs41604533	5:28834321	A/G	CSRNP2	
rs13324417	5:31096415	A/G		Variante intrón
rs42851619	5:34201257	A/C		

Genes:

CSRNP2: cysteine-serine-rich nuclear protein 2

RND1: Rho family GTPase 1

- **Región de interés 5: En el cromosoma 6**

En torno a la posición 6:60000000 se seleccionaron 4 SNPs muy próximos:

FUENTE: ELABORACIÓN PROPIA

Tabla 22: Región de interés 5

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs133208579	6:60530774	C/T		
rs135353878	6:67763484	C/T		
rs109094824	6:69666446	C/T		
rs137758365	6:71397773	A/G	PDGFRA	Codon sinónimo

Genes:

PDGFRA: platelet-derived growth factor receptor, alpha polypeptide

- **Región de interés 6: En el cromosoma 13**

En torno a la posición 13:69000000 se seleccionaron 2 SNPs muy próximos, tal y como muestra la siguiente tabla:

FUENTE: ELABORACIÓN PROPIA

Tabla 23: Región de interés 6

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs134759938	13:68278449	A/C	PPP1R16B	Variante intrón
rs132853415	13:70470318	C/T		Variante intrón

Genes:

PPP1R16B: protein phosphatase 1, regulatory subunit 16B) y TOP1 (topoisomerase (DNA) I)

- **Región de interés 7: En el cromosoma 15**

En torno a la posición 15:5900000 se seleccionaron 2 SNPs muy próximos:

FUENTE: ELABORACIÓN PROPIA

Tabla 24: Región de interés 7

SNPs	Posición	Alelo
rs42023581	15:5881099	G/T
rs109325605	15:60309469	A/G

- **Región de interés 8: En el cromosoma 18**

En torno a la posición 15:15000000 se localizaron 2 SNPs muy próximos.

FUENTE: ELABORACIÓN PROPIA

Tabla 25: Región de interés 8

SNPs	Posición	Alelo	Gen	Consecuencia funcional
rs110615481	18;15047605	A/G	VPS35	Variante intrón
rs135883818	18;15793170	C/T	ITFG1	Variante intrón

Genes:

VPS35: vacuolar protein sorting 35 homolog (S. cerevisiae)

ITFG1: integrin alpha FG-GAP repeat containing 1

La representación sobre el cariotipo permite distinguir claramente entre los cromosomas con mayor y menor importancia en el diferenciación de razas. Por ejemplo, a simple vista se aprecia que los cromosomas más relevantes son: 1, 3, 5, 15 y 17. Adicionalmente, se ha podido averiguar que los últimos cromosomas son menos importantes en la diferenciación de las razas. Los cromosomas del 19 al 29 parecen no contener información característica de las razas, a excepción del cromosoma 26.

Genes presentes en los nucleótidos seleccionados.

La siguiente tabla proporciona el nombre de los genes que se han localizado en las posiciones de los nucleótidos seleccionados. Se ha identificado un total de 66 genes distribuidos a lo largo de 24 cromosomas.

FUENTE: ELABORACIÓN PROPIA

Cromosoma	Genes
1	HTR3C, CLSTN2, C1H8orf42, LOC101907397, EPHA6 PWP2
2	MAP3K2, HS6ST1, LRP1B
3	AGBL4, CSMD2
4	VWC2 CNTNAP2, LOC101906647
5	FAM19A2, CSRNP2, MYO1A, RND1, BTA-2048, MRPS35
6	PDGFRA, RNF4
7	LOC101903762, DOT1L, EPOR
8	SCARA3
9	SLC22A1, LOC101905596, MAP7
10	SIPA1L1
11	NGS-119429, KDM3A
12	LOC101903856, MTUS2
13	TOP1, LOC526745, PPP1R16B
14	ASPH, EIF3H
15	LRRC4C, LOC101906823, DSCAML1
16	DNAH14, UCHL5
17	STX2, TMEM132C, SLC24A6
18	ITFG1, VPS35 GLG1, POP4, COX4I1, SIPA1L3, CYLD
21	FBXO22
22	ARPC4
24	DLGAP1, GATA6
25	SLC29A4, HS3ST4
26	NT5C2, LOC101908152, DOCK1, FAM196A, LOC522146
29	ANO1

Tabla 26: Genes localizados en las posiciones de los SNPs seleccionados

Los genes que se exponen en esta tabla son todos los que coinciden en las posiciones de los 132 SNPs seleccionados. Por tanto, en ella figuran los 10 genes que se han identificado en las 8 regiones que hemos definido como de interés y 56 genes más que se han identificado en las posiciones de algunos SNPs que se han seleccionado pero que se localizan en posiciones aisladas. Es importante comentar que ninguno de estos genes se identificó en más de una ocasión. Recordemos que un gen puede estar repetido y puede localizarse en multitud de regiones en el cariotipo, no es el caso de ninguno de los genes aquí contemplados.

Capacidad explicativa del modelo y comparación con un subconjunto aleatorio de SNPs entre los que superaron la prueba de análisis genético.

Se muestra a continuación una tabla con la R^2 de un modelo construido únicamente con 132 SNPs y en el que se implementa PLS. En ella se comparan los resultados de un modelo en el que los predictores son seleccionado mediante PLS-LDA, con otro en el que los predictores se eligen al azar.

Recordemos que se ha establecido ese número de SNPs puesto es el número mínimo de predictores para el que seleccionando los SNPs en base a PLS-LDA se consigue un porcentaje aciertos superior al 95%.

FUENTE: ELABORACIÓN PROPIA

RAZA	R2 SELECC 132	R2 132 RANDOM
ASTV	89.49	49.46
AVIL	89.12	62.3
BRUP	76.23	56.81
MORU	82.55	64.88
PIRE	89.25	67.74
RETI	83.64	69.15
RGAL	86.21	67.6
BRSW	79.16	79.03
FLCH	65.19	58.77
GUER	91.28	76.77
SIMM	88.61	81.74
R2 media	**83.7**	**66.75**

Tabla 27: Capacidad predictiva del modelo

En primer lugar, se aprecia en la tabla un incremento muy considerable de la capacidad explicativa del modelo cuando se seleccionan los SNPs en base al método de PLS-LDA. Adicionalmente, podemos presuponer que la diferencia habría sido incluso mayor en el caso de que los SNPs que se eligen aleatotriamente, se hubiesen elegido entre el total de SNPs leídos en el chip, sin haber aplicado previamente unos criterios genéticos.

Es destacable de esta tabla que la mayor varianza explicativa se consigue para la raza GUER, que es precisamente para la que menos muestras se tomaron, con sólo 28 individuos analizados (un 3.8% del total). Se consigue una R^2 mayor incluso que para la raza SIMM, que es la que cuenta con más observaciones (158 individuos, un 21,76% del total). Sin embargo, este resultado no debería de resultar tan sorprenderte puesto que en (Negrini., *et al*. 2008a) se concluye que para poder asignar correctamente muestras a una raza, debemos considerar en el análisis un mínimo de entre 20-30 individuos genotipados para esa raza. Por tanto, la raza GUER con 28 individuos, no debería de causar problemas en este sentido. En (Negrini., *et al*. 2008b) se estudia la variabilidad en 20 razas europeas y ocurre que para la raza Parthenaise, que contaba con sólo 14 observaciones, se consiguió un porcentaje de aciertos inferior al 35% cuando, sin embargo, se estaba logrando más del 90% de predicciones acertadas en el resto de las razas que si contaban con un número mayor de individuos analizados. Se concluye que, aunque con un número reducido de muestras para una raza, puede verse muy restringida la capacidad explicativa, no tiene porque ocurrir así cuando el número de muestras es reducido pero suficiente (caso de GUER en este estudio).

De la (tabla 27 - Resultados) también se deduce que para la raza BRUP, la capacidad explicativa es considerablemente menor que para el resto de razas españolas y esto, muy probablemente, se deba a sus diferencias genéticas con el grupo de razas españolas. Buena parte de los SNPs seleccionados para distinguir entre las razas españolas no aportan información acerca de la raza BRUP al ser ésta tan diferente del resto.

Diferenciación de razas

Para estudiar la variabilidad de forma intuitiva se han representado los animales en clusters de razas. Los cluster se han obtenido mediante PCA y LDA. Para un mismo número de variables (132 SNPs) los resultados son considerablemente mejores cuando seleccionamos los predictores en base a PLS-LDA que cuando los elegimos aleatoriamente, como veremos tanto en los clusters de CP como en los de LDA. Por un lado se compara el grado de diferenciación al seleccionar o no seleccionar los predictores para un cluster obtenido por PCA; y por otro, se hace la misma comparación entre dos clusters de razas obtenidos mediante LDA. Es importante aclarar que la diferenciación que en estos clusters se manifiesta, no tiene un grado alto de detalle. El

análisis de diferenciación de razas se ha mejorado mucho en los últimos años y tiene hoy en día una complejidad que escapa al alcance de este estudio. El análisis cluster aquí representado va enfocado a hacer las comparaciones pertinentes y a buscar una interpretación intuitiva de la diferenciación que se conseguiría al considerar únicamente los 132 SNPs seleccionados en este estudio.

Cluster mediante PCA

Se realizan dos clusters de razas. En un primer lugar eligiendo al azar los predictores y en un segundo lugar, seleccionando éstos mediante PLS-LDA. En ambos casos, el cluster se obtiene con la técnica de PCA extrayendo 2 CPs. Las observaciones quedan representadas en un eje cartesiano según el peso que tengan para cada uno de los CPs extraidos. Es lógico pensar que los individuos de una misma raza, puesto que se parecen más, tomarán pesos parecidos para ambos CPs y por tanto aparecerán cerca unos de otros al ser representados sobre el eje cartesiano.

El análisis está basado en la matriz de correlaciones. Generalmente observaremos que los datos tienen una mayor variabilidad para el 2º CP, puesto que el 1º CP tiende a parecerse al promedio de todas las variables consideradas, mientras que el 2º CP, al tener que ser perpendicular al 1º CP, suele tener una dirección más particular.

En una representación PCA, el centroide de un grupo de individuos (por ejemplo una raza) es el valor medio de los pesos de los CPs que toman los individuos de esta raza y tendrán tantas dimensiones como CPs se hayan considerado. En este caso, puesto que se han extraido 2 CPs, los centroides quedarán definidos únicamente por una ordenada y una abscisa. Puesto que el 1º CP tiene un mayor peso con respecto al total de variables y este se encuentra representado en el eje horizontal, asumiremos que si los centroides de dos razas se encuentran bastante alejados en el eje horizontal, las razas tienen una diferenciación genética considerable. Mientras que para un mismo alejamiento de los centroides pero caracterizado por una distancia fundamentalmente vertical, asumiremos que las razas se asemejan más (siempre y cuando los ejes estén en la misma escala y los datos estén estandarizados).

FUENTE:ELABORACIÓN PROPIA

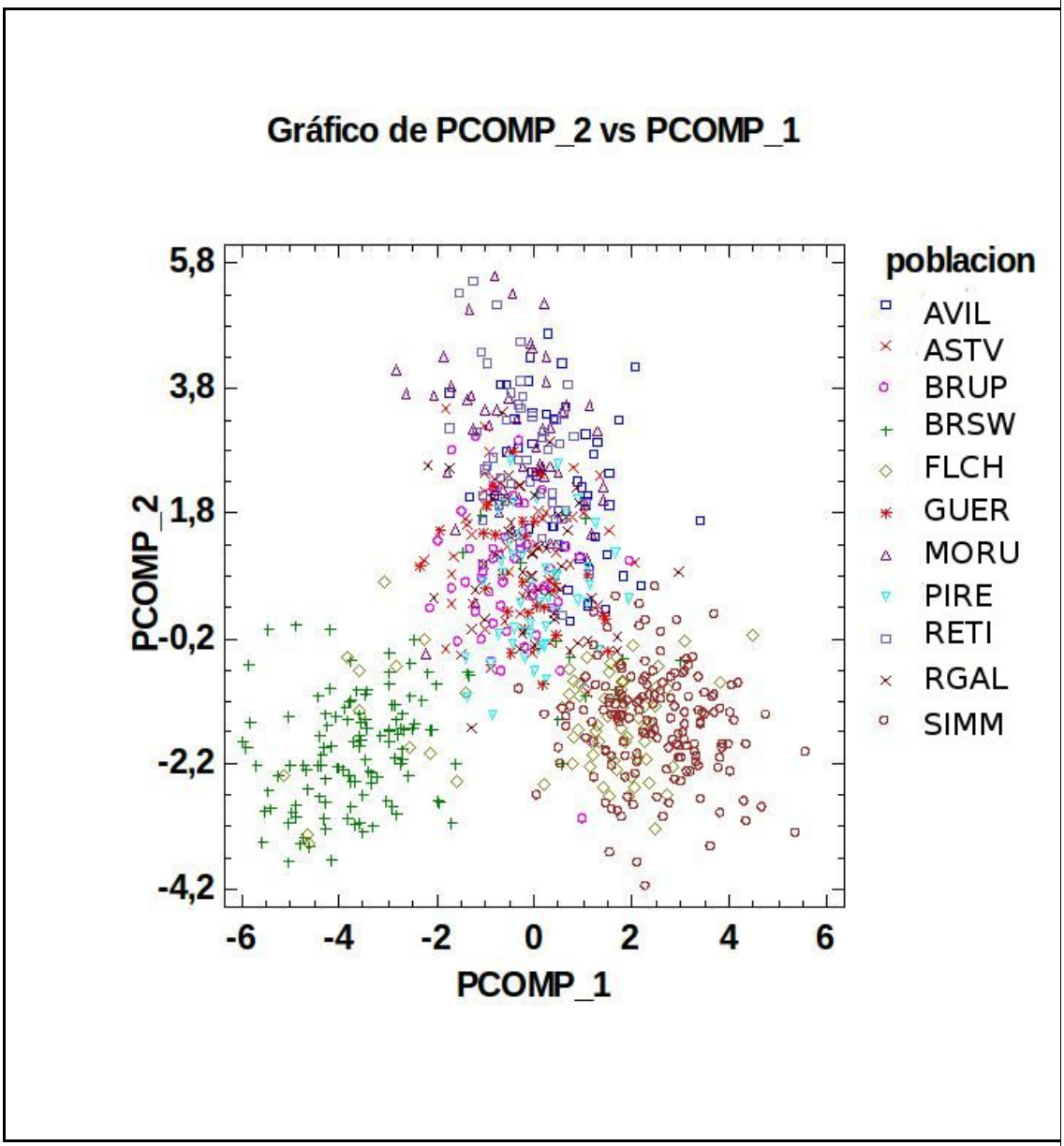


Ilustración 19: Representación Cluster mediante ACP con predictores no seleccionados

Las observaciones aparecen con una marca diferente según la raza a la que pertenecen, pero hay que tener en cuenta que esta variable no es considerada en el análisis sino que simplemente se utiliza como "tipo", para valorar la diferenciación obtenida. Por tanto, cuanto menos entremezclados estén los colores, mejor

diferenciación habremos logrado entre las razas.

Los animales de la raza SIMM ocupan la esquina inferior derecha. Lo cual indica que estos individuos son los más correlacionados con el 1º CP. De igual manera, podemos concretar que los individuos pertenecientes a AVIL y MORU están muy correlacionadas con el 2º CP puesto que sus individuos aparecen representados en la parte superior del eje cartesiano. Adicionalmente, observamos que los individuos de las distintas razas están, en general, muy entremezclados y no se consigue una buena diferenciación.

A continuación, vemos lo que ocurre cuando al efectuar el mismo análisis de Cps, utilizamos el mismo número de predictores pero estos han sido seleccionados mediante el método de PLS-LDA:

FUENTE: ELABORACIÓN PROPIA

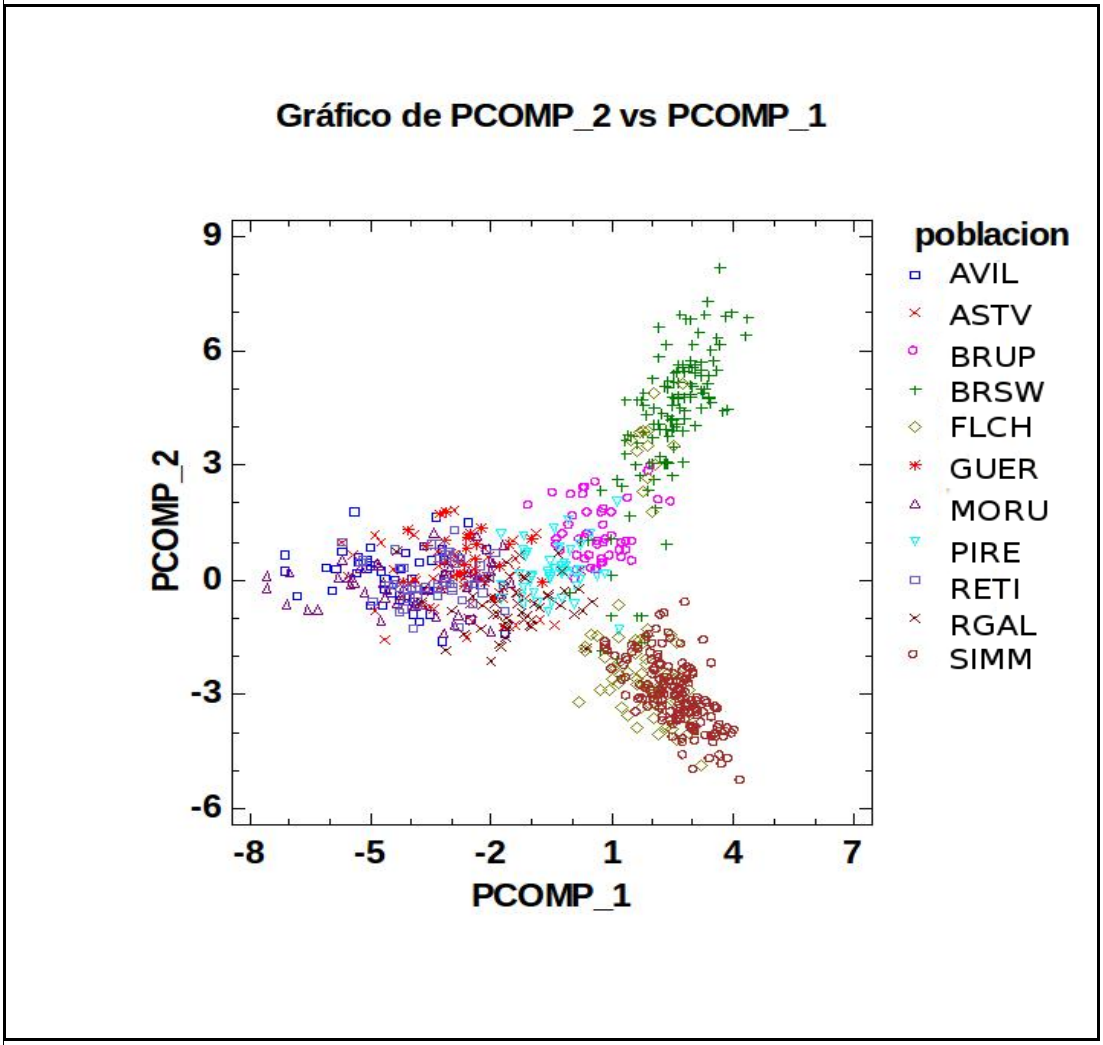


Ilustración 20: Cluster mediante ACP con predictores seleccionados

Como puede verse, el grado de diferenciación es considerablemente mayor en este caso. Sin embargo, sigue siendo difícil diferenciar entre los individuos de algunas parejas de razas. Por un lado vemos que las razas españolas AVIL y MORU tienen pesos mucho menores que el resto para el 1º CP y que los individuos de ambas razas tienen prácticamente los mismos pesos para los CPs, resultando prácticamente imposible su diferenciación. Los animales de la raza GUER, sin embargo, se diferencian de los de las razas AVIL y MORU en que tienen pesos algo mayores para ambos CPs. Las razas BRUP y PIRE son, por su parte, las razas más fácilmente distinguibles y bastará con unos pocos predictores para asignar correctamente las muestras a estas razas. Vemos que únicamente hay un individuo de la raza BRUP que podría asignarse a la raza europea BRSW.

Los animales de la raza SIMM son fácilmente distinguibles y se caracterizan por tomar valores muy altos para el 1º CP y muy bajos para el 2º. Sin embargo, algunos individuos de la raza FLCH podrían ser asignados erróneamente a la raza SIMM. De hecho, es muy llamativo de esta representación cluster la distribución de los individuos de la raza FLCH. Podemos observar que los individuos de esta raza se encuentran repartidos en dos grupos bien diferenciados. Por un lado, un grupo de individuos se asemeja a los individuos de la raza SIMM, y por otro, otro grupo de aproximadamente el mismo tamaño, se asemeja mucho a los individuos de la raza BRSW. En el caso de la raza FLCH podemos ver que la diferenciación no es adecuada, puesto que algunos de los individuos de esta raza se encuentran muy dispersos en el gráfico.

En estas representaciones cluster se observa que los datos se encuentran más dispersos en el eje vertical que en el horizontal y se corrobora el hecho de que habitualmente las observaciones muestran más variabilidad para el 2ºCP que para el 1º CP.

Análisis discriminante

El análisis mediante LDA permite una diferenciación mayor puesto que en este caso la variable raza no es simplemente un “tipo de observación” que es tenida en cuenta a la hora de colorear las observaciones según grupos, sino que en este caso se considera esta variable en el análisis. A la variable en base a la cual se va a intentar discriminar las observaciones se la denomina en LDA como “Factor de Clasificación”. El cluster se diseña de tal manera que se maximiza la diferenciación entre-clases frente a la diferenciación dentro-de-clases, donde las clases son las diferentes categorías que puede tomar el Factor de Clasificación.

Para la elección de las variables que van a aparecer en el eje cartesiano se lleva a cabo un procedimiento de selección hacia adelante. A las variables que se generan con LDA se las llama funciones. Este procedimiento comienza con un modelo en el que se incluyen todas las variables, en este caso los 132 SNPs con 3 posibles respuestas, y posteriormente se van eliminando una a una las variables que son poco significativas en el modelo provisional. En estas representaciones cluster, como en PCA, se representan únicamente las dos variables más importantes (Funciones) para así poder llevar los datos sobre un eje cartesiano.

FUENTE: ELABORACIÓN PROPIA

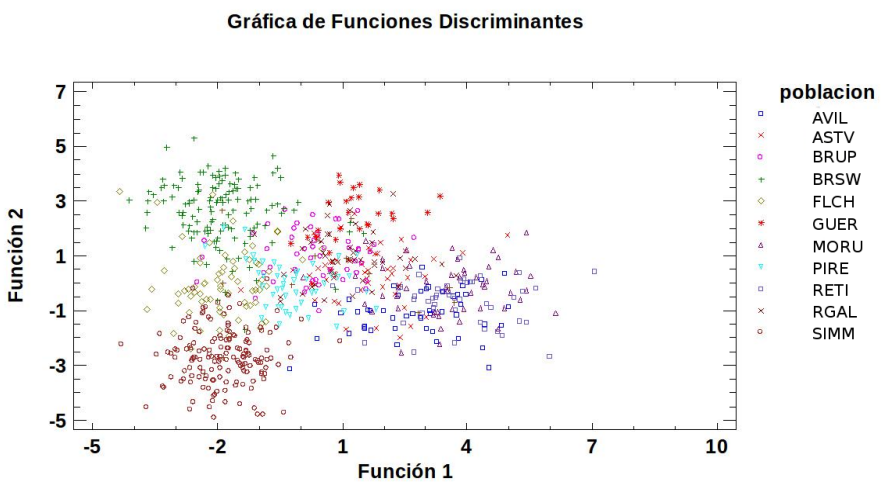


Ilustración 21: Cluster mediante LDA con predictores no seleccionados

Una prueba de que LDA es mejor herramienta para distinguir entre razas es la mejor diferenciación que se consigue para los individuos de la razas SIMM y BRSW, que aparecen en las esquinas izquierda-superior e izquierda-inferior de la ilustración. Sin embargo, las razas BRUP y PIRE aparecen en este caso más entremezcladas con las otras razas españolas. También observamos que las razas MORU y AVIL son, en este caso, muy difíciles de distinguir.

Vemos ahora lo que ocurre cuando hacemos LDA con 132 predictores seleccionados en base a PLS-LDA:

FUENTE: ELABORACIÓN PROPIA

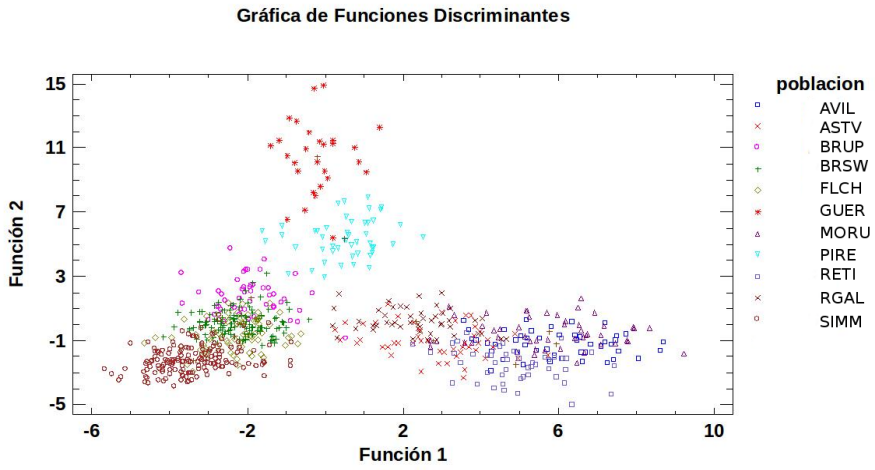


Ilustración 22: Cluster mediante LDA con predictores seleccionados

En el segundo caso, el análisis discriminante permite diferenciar mejor entre razas.

Resulta llamativo que la raza BRUP se parezca más a las razas europeas que a las españolas. Es destacable el grado de diferenciación de GUER tan marcado por el efecto isla. Por otro lado, el fuerte parecido entre RETI y AVIL coincide con los resultados expuestos en (Cañas-Álvarez., *et al*. 2014b), en su representación cluster mediante PCA de 5 de las razas españolas consideradas en el estudio. En dicha investigación no se incluía la raza MORU que, como hemos visto, es una de las razas más problemáticas dado su fuerte parecido con la raza AVIL y el máximo parecido se daba entre las razas RETI y AVIL.

De las dos representaciones cluster que tienen una mayor capacidad de diferenciación de razas, que son el ambos cluster (mediante PCA y LDA) cuando se seleccionan los predictores, se deduce que las razas más difíciles de distinguir son:

1) AVIL y MORU

2) BRSW y FLCH

Para estimar hasta que punto diferenciar entre estos pares de razas empeora la capacidad predictiva del modelo se ha calculado para el mismo número de marcadores, el porcentaje de aciertos en los siguientes casos:

A) Sin tener que hacer distinción entre las razas MORU y AVIL: el porcentaje de individuos correctamente asignados ascendió a 96.01

B) Sin tener que hacer distinción entre las razas BRSW y FLCH: la proporción de individuos correctamente asignados ascendió a 96.31

C) Sin tener que hacer distinción entre las razas MORU y AVIL, ni entre BRSW y FLCH: se logró un incremento en la predicción hasta llegar al 96.42 % de aciertos

Para hacer estas pruebas se ha considerado a los animales de una de las dos razas tan difíciles de distinguir como individuos pertenecientes a la raza próxima. Por ejemplo, hemos supuesto que todos los individuos, tanto los de AVIL como lo de MORU pertenecen a AVIL y que todos los individuos, tanto los de BRSW como los de FLCH pertenecen a la raza FLCH. De este modo se descarta la opción de errar en la asignaciones para estos pares de razas.

Adicionalmente, se ha determinado el porcentaje de aciertos al considerar sólo las razas españolas o sólo las razas europeas:

A) Considerando sólo las razas españolas: 97.95

B) Considerando sólo las razas europeas: 93.23

De los resultados se deduce que las razas europeas son más difícilmente distinguibles. Lo cual, en cierto modo coincide con lo expuesto en (Negrini., et al. 2008b), puesto que en la investigación se concluye que resulta mucho más difícil diferenciar entre individuos pertenecientes a las razas continentales (razas de Suiza en el caso de este estudio) que entre individuos de razas británicas, con porcentajes de aciertos del 77% frente a 97%, respectivamente.

En el caso de la comparación que se hace en este estudio entre el grado de diferenciación de las razas españolas y el de las europeas, hay que tener en cuenta que en ambos casos se ha reducido el número de observaciones con respecto a cuando considerábamos el conjunto de 11 razas. Para determinar el grado de diferenciación de las razas españolas se ha vuelto a realizar el análisis pero sin incluir en este caso los individuos pertenecientes a las razas europeas. Así mismo, para determinar la diferenciación entre las razas europeas, se han excluido del análisis los individuos pertenecientes a las razas españolas.

En el caso de las razas europeas, se pasa de disponer de 726 observaciones a 384 y en el caso de las españolas, el número desciende a 342.

Un mayor número de individuos a evaluar supone, naturalmente, un penalización de la capacidad predictiva. Por otro lado, a pesar de contar con menos observaciones en el caso de las razas españolas, la potencia predictiva ha aumentado desde el 95% al 97%. Una posible justificación es que en este caso no se le exige al

modelo que distinga entre las razas BRSW y FLCH, que como hemos visto, son razas muy difíciles de diferenciar.

6. DISCUSIÓN

1- Disminución en el número de predictores necesarios al aplicar el método de PLS-LDA – Comentarios (tabla 17 - Resultados).

Como puede verse en la (tabla 17 - Resultados) el número de SNPs que son necesarios para incrementar el porcentaje de aciertos es mayor conforme se exige un mayor porcentaje de aciertos. Dicho de otra manera, para pasar de un porcentaje de aciertos del 90% a un 91% harán falta muchos menos SNPs que para pasar de un 95% de aciertos a un 96%. Sin embargo, conviene considerar hasta que punto ésto es así para cada uno de los dos casos: cuando se seleccionan los SNPs en base a PLS-LDA y cuando no.

En el caso de la selección de SNPs mediante PLS-LDA 55 SNPs son suficientes para lograr incrementar la potencia predictiva en un 5% al pasar de un 90% de aciertos a un 95%. Obsérvese que 55 SNPs equivale a considerar únicamente 5 SNPs más para cada raza. Sin embargo, llegados a este grado de acierto del 95%, hacen falta 143 SNPs, correspondientes a 13 SNPs más para cada raza para lograr un único punto porcentual de potencia predictiva y llegar al 96% de aciertos. Este efecto tan exagerado se explica por la presencia de algunos individuos que son más difícilmente distinguibles que el resto y que requieren de la información de muchos más predictores para ser finalmente asignados a sus razas verdaderas.

Cuando se trata de llevar a cabo la asignación con pocos predictores puede ocurrir que estos individuos sean, a menudo, asignados a una raza próxima a la verdadera y al detectar ésto, puesto que conocemos respuesta verdadera (raza) también en el caso de los individuos de la fase de prueba, los individuos serán considerados por R como individuos mal asignados. Estos individuos difícilmente distinguibles presentan para algunos nucleótidos, copias alélicas que no son propias de los animales de su raza, sino de animales de alguna raza genéticamente próxima a ésta.

En el caso de los SNPs seleccionados aleatoriamente, sin embargo, el incremento de SNPs necesarios para conseguir más aciertos conforme nos vamos acercando a potencias predictivas más elevadas es menos patente. Mientras que para pasar de un 90 a un 95% de aciertos son necesarios 95 SNPs, que son casi 9 SNPs más para cada raza; para pasar de un 95% a un 96% son necesarios 115 SNPs, que son algo más de 10 SNPs más por raza. El hecho de que este efecto sea mucho mayor en el caso de la colección de SNPs seleccionados sugiere que el incremento de información que se logra al disponer de una colección de SNPs seleccionados es mayor al principio. Lo cual parece lógico puesto que siempre que añadimos un SNPs a la selección, incorporamos el mejor de los que no estaban seleccionados y la disponibilidad de SNPs va disminuyendo conforme vamos teniendo una selección mayor de predictores. De hecho, la diferencia en el número de SNPs necesarios al seleccionar mediante la técnica de PLS-LDA con respecto a cuando no se hace es más significativa cuando se consideran pocos SNPs. Para un 90% mínimo de aciertos hacen falta 23 SNPs más si no seleccionamos. Obsérvese que 23 SNPs son un 30% más de SNPs necesarios en el caso de no seleccionar. Para un 95% mínimo de aciertos hacen falta 63 SNPs más en caso de no usar el método. Lo cual se corresponde con un 47% más. Mientras que para lograr un

96% mínimo de aciertos hacen falta sólo 35 SNPs más, que son sólo un 12.7% más de SNPs que en caso de usar el método.

El hecho de que el incremento en la capacidad predictiva sea proporcionalmente mayor al considerar un 95% mínimo de aciertos que al considerar un 90% sugiere que los SNPs que se han añadido a la selección son todavía considerablemente mejores que los que quedan fuera de la selección. Los resultados obtenidos conducen a pensar que hasta que se alcanza un cierto número de SNPs, la diferencia entre la capacidad predictiva de un SNP considerado en la selección y otro no considerado es muy significativa, mientras que pasado ese punto, la diferencia empieza a ser significativamente menor. Los resultados obtenidos que se exponen en la (tabla 17 - Resultados) llevan a pensar que los 100-150 mejores SNPs en base a PLS-LDA son significativamente mejores predictores que el resto para el caso del estudio.

2- Comentarios acerca del número mínimo de SNPs que hacen falta para lograr una alta potencia predictiva – Comparación investigaciones

Otros trabajos han logrado un grado moderadamente alto de clasificación. En (Martínez-Camblor., *et al*. 2014) se hace uso de la regresión logística, minería de datos y aprendizaje de máquina, se seleccionan 3000 marcadores microsatélite en base al criterio del test exacto de Fisher para regresión logística y se consigue un 99.94% de aciertos en la asignación. Hay que tener en cuenta que dicho trabajo tenía la dificultad añadida de que se trataba de diferenciar entre 70 líneas de la raza lidia, siendo ésta precisamente la razón por la que se eligió usar marcadores microsatélite que, como se ha comentado, son especialmente útiles cuando se trata de encontrar diferencias entre individuos muy parecidos. Dicho esto, conviene mencionar aquí la considerable reducción del precio del análsis al considerar marcadores SNPs en lugar de marcadores microsatélite. Teniendo en cuenta que estos métodos se basaban en modelar la segregación de alelos, el coste computacional sería excesivo en caso de usar SNPs. Por otro lado, usar SNPs plantea otras ventajas además del precio, por ejemplo, para diferenciar entre distintas razas, en este trabajo se consiguen mejores resultados al usar marcadores SNP que los que se obtuvieron en un estudio con microsatélites (Troy., *et al*. 2001) en el que se pretendía diferenciar entre 10 razas.

Se comprueba que la combinación de usar SNPs y aplicar estadísticos resulta eficiente cuando se quiere hacer diferenciación entre razas, como ya se dedujo en (Negrini., *et al*. 2008a).

Los resultados obtenidos son comparables a los obtenidos en (Maudet., *et al*. 2002) con 6 razas francesas y a los que se obtuvieron en (Negrini., et al. 2008b), que consiguen una asignación correcta en un 96% de los individuos haciendo uso de estadística bayesiana y frecuencial. En dicho estudio, se consiguen uno resultados equiparables a los que aparecen en las investigaciones de (MacHugh *et al*. 1998) y (Troy., *et al*. 2001) que logran entorno a un 85% de aciertos con únicamente 90 SNPs en

lugar de los 19-23 microsatélites.

En (Negrini., *et al*. 2008b) se consigue un 90% de individuos correctamente asignados para 24 razas europeas usando también un panel de 90 SNPs, aunque en este caso se pretendía diferenciar entre animales correspondientes a distintas IGPs. Destacamos que en este estudio se consigue el porcentaje de aciertos del 90% con sólo 77 SNPs. Adicionalmente, queda demostrada la eficacia del método supervisado empleado en este estudio que, como se dedujo en (Negrini., e*t al*. 2008b) se comportan mucho mejor que los métodos no supervisados cuando se trata de resolver un problema de asignación. Por otro lado, PLS ha servido para resolver el problema del ruido estadístico tal y como se advirtió en (Palermo., *et al*. 2009) donde se deducía además, que se conseguían mejores resultados con PLS conforme más correlacionados estaban los predictores. Al combinar esta técnica con LDA, se soluciona el problema que se advertía en dicho artículo, según el cual para efectuar el método PLS era necesario que los predictores estuvieran fuertemente correlacionados.

Por último, mencionar que los resultados son algo peores que los obtenidos en razas porcinas. En (Ramos., *et al*. 2011) se consigue un 99% de aciertos con 192 SNPs. Si bien, hay que tener en cuenta que en ese caso de trataba únicamente de 7 razas. En cualquier caso, hay que tener presente que, aunque utilizando un método adecuado de selección de los predictores podemos conseguir incrementar en mucho la potencia predictiva del modelo, a la hora de comparar los resultados, estos dependerán en gran medida del grado de diferenciación entre las razas estudiadas.

3- Acerca del cromosoma 5-BTA

En este estudio se concluye que el cromosoma 5 es el más importante en la diferenciación de razas. Se han llevado a cabo diversas investigaciones sobre este cromosoma en *Bos Taurus* puesto que se ha identificado como especialmente relevante en varias investigaciones. En este cromosoma se han identificado varios QTLs relacionados con el crecimiento y con la producción de grasa. En (Rogberg-Muñoz., *et al*. 2010) se busca la asociación entre 4 microsatélites en el cromosoma 5 (5-BTA) con 2 genes de especial interés que se sabe pertenecen a este cromosoma. Esos genes son el gen factor 5 Myogenic y el gen IGF de factor 1 de crecimiento de tipo insulina. En dicho estudio se concluye que los QTLs no perdían importancia al analizarse individuos que se crían en pastos con un sistema de explotación tradicional, como pudiera ser el sistema de explotación típico español. Aunque no se han identificado esos 2 genes en las posiciones próximas a los SNPs seleccionados, los resultados obtenidos en (Rogberg-Muñoz., *et al*. 2010) por un lado, y los obtenidos en este estudio por otro, sugieren que posiblemente haya una relación entre los caracteres más implicados en la diferenciación de razas y los caracteres más relacionados con el crecimiento y con la producción de grasa. Adicionalmente, se han identificado en este cromosoma algunos QTLs asociados con el carácter peso al nacimiento en la raza Holstein (Gasparin., *et al*. 2005).

En un estudio sobre genética humana se eligió el cromosoma 5-BTA para compararlo con los cromosomas humanos 12 y 22. Este estudio sirvió para concluir que ha habido un reordenamiento entre los cromosomas de las especies humana y bovina.

4- Acerca de los genes más importantes en *Bos Taurus*

Dada la representación de los SNPs seleccionados en el cariotipo, se ha buscado si existía una relación entre alguno de estos y los genes que hoy son considerados como más importantes en genética de bovino. En (González-Rodríguez., *et al*. 2014) también se buscan estos genes de especial interés, principalmente los genes DGAT1, POLL y MSTN. DGAT1 es un gen relacionado con la producción de leche, POLL está relacionado con la presencia de cuernos en los ejemplares y MSTN está relacionado con la producción de carne. En dicha investigación se estudió la variabilidad de las razas y se identificó un marcador próximo al gen MSTN en el cromosoma 2. No ha sido así en el caso de este estudio.

5- Acerca de la capacidad explicativa del modelo– Comentarios (tabla 27 - Resultados)

ASTV es claramente la raza que se ve más favorecida al ser los predictores seleccionados. Para esta raza se pasa de explicar menos de un 50% a explicar casi el 90%. Otra raza española muy favorecida es RGAL. De entre las razas europeas, las raza para la que más se incrementa la capacidad predictiva al seleccionar los predictores es GUER.

Se supone que el incremento que se logra en la capacidad explicativa de GUER esté muy relacionado con el reducido número de muestras que se tomaron para esta raza. Al ser este número tan reducido, resulta bastante improbable que se seleccione algún SNP que sea exclusivo para esta raza, mientras que si seleccionamos los predictores con PLS-LDA estaremos forzando a que 1/11 de los predictores sean los más exclusivos de esta raza. Esto se corrobora por el hecho de que para las razas con mayor número de muestras se consigue un menor incremento de la capacidad explicativa. Por ejemplo, la raza SIMM con 158 muestras analizadas, aumenta sólo un 7% su capacidad explicativa y para la raza BRSW con (129 muestras), la capacidad explicativa prácticamente no aumenta. Además, las razas españolas, que tienen todas prácticamente el mismo número de muestras, aumentan la capacidad explicativa en proporciones parecidas. El mayor incremento en ASTV y RGAL seguramente se deba a una cuestion de “mala suerte” cuando se eligieron los SNPs aleatoriamente. Al seleccionar los predictores no sólo se consigue incrementar la capacidad explicativa en todas las razas, sino que adicionalmente se evita que alguna raza pueda quedar desfavorecida por la mala suerte en la elección de los predictores. Dicho de otra forma, con PLS-LDA los resultados no son sólo mejores, sino también más regulares.

Un aspecto a tener en cuenta es que el parecido entre algunas parejas de razas, fundamentalmente (AVIL y MORU) y (BRSW y FLCH) en cierto modelo ayuda a estas parejas de razas a diferenciarse del resto de razas. Los SNPs que se seleccionan mediante PLS-LDA para, por ejemplo, diferenciar a la raza AVIL del resto, dado el parecido de esta con MORU, proporcionarán también información útil para diferenciar a la raza MORU del resto de las razas. Este hecho no ayuda a distinguir entre individuos pertenecientes a MORU y a AVIL y por tanto no aumenta el porcentaje de aciertos, pero podría servir para aumentar la capacidad explicativa de ambas razas. De igual manera para las razas BRSW y FLCH. Sin embargo, este "favorecimiento mutuo" entre las razas próximas debe de tener un efecto muy pequeño, puesto que en la (tabla 27 - Resultados) no se aprecia una mayor capacidad explicativa para estos pares de razas. Probablemente este efecto se apreciaría más si todas las razas analizadas contaran con el mismo número de muestras, puesto que el hecho de que una raza tenga más individuos favorece mucho más a la capacidad explicativa que el "favorecimiento mutuo" que aquí se comenta.

6- Comentarios acerca de la diferenciación entre razas – Cluster de razas

Puesto que de los análisis clusters realizados en nuestro estudio, tanto mediante PCs como con LDA, los resultados son mejores cuando se seleccionan los predictores, cuando se comparen los resultados de este estudio con los de otras investigaciones haremos mención a los clusters de razas mediante PCs y LDA que se obtuvieron con la selección de predictores.

Se expone a continuación el cluster publicado en (Cañas-Álvarez., *et al*. 2014b):

FUENTE: Cañas-Álvarez., *et al*. 2014B

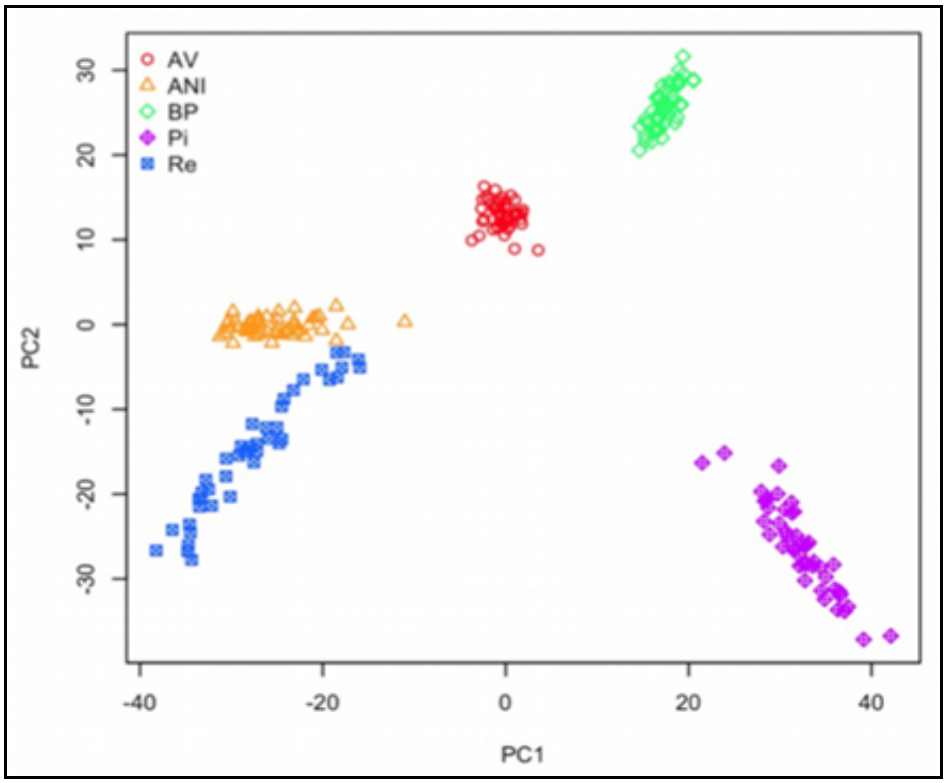


Ilustración 23: Análisis de Componentes Principales para 5 razas del estudio

Razas: Asturiana de los Valles (AV), Avileña–Negra Ibérica (ANI), Bruna dels Pirineus (BP), Pirenaica (Pi) y Retinta (Re)

En (Cañas-Álvarez., *et al*. 2014b) puesto que no se considera la raza MORU, resulta que las razas más parecidas son RETI y AVIL. El orden de parecido es: RETI y AVIL, ASTV, BRUP y PIRE. Recordemos que, asumiendo igualdad de escalas en los ejes, en una representación de 2 CPs en el que el 1º CP corresponde al eje horizontal, la distancia horizontal es más importante que la vertical en la diferenciación de razas. Por este motivo, se puede concluir que la raza más alejada del resto es PIRE. De nuestras representación cluster mediante LDA se concluye lo mismo (Ilustración 14 – resultados). Sin embargo, nuestra representación cluster según CPs contradice este resultado y estima que la diferenciación de la raza BRUP con respecto al resto es ligeramente mayor que en el caso de PIRE (Ilustración 12 - resultados). Esta segunda representación coincide con lo expuesto por (Martín-Burriel., *et al*. 2011a) en un

estudio realizado con 40 razas de la península ibérica. Conviene tener en cuenta que éste estudio es 3 años más antiguo que el estudio de (Cañas-Álvarez., *et al*. 2014b) y que las técnicas de diferenciación de razas han mejorado considerablemente desde 2011. Siendo una diferencia fundamental, por ejemplo, que en (Martín-Burriel., *et al*. 2011a) se utilizan marcadores microsatélites. Adicionalmente, respecto a nuestro estudio, concluíamos que cabe esperar que el cluster mediante LDA sea más acertado que el cluster según PCs.

El parecido entre las razas RETI y AVIL que tanto se aprecia en (Cañas-Álvarez., *et al*. 2014b) está también muy latente en nuestro estudio en ambas representaciones clusters. (Ilustraciones 20 y 22 – Resultados).

Dadas las características de las variables que se consideran en estas representaciones clusters, que son SNPs con 3 posible categorías (0, 1 o 2 copias) la interpretación de las distintas razas en los clusters carece en cierto modo de sentido. Es decir, si hubiera una raza que tuviera, por ejemplo, pesos muy elevados para ambos CPs apenas podríamos decir de ella algo al respecto.

7. CONCLUSIONES

Las conclusiones derivadas de este estudio son las siguientes:

1. El chip de genotipado utilizado ha logrado una lectura prácticamente completa de los nucleótidos para las razas de este estudio, y por otro lado, los individuos elegidos son representativos de sus razas.

2. Los marcadores SNP son eficientes cuando se trata de buscar las diferencias genéticas entre diferentes razas.

3. La realización de un análisis genético previo y la fijación de unos valores umbrales recomendados para los criterios de DL, HWE, MAF y GENO ha permitido simplificar en buena medida el posterior análisis estadístico.

4. El análisis genético sugiere que las poblaciones se encuentran en HWE y ha permitido conocer que el número de SNPs con MAF inferior al 3% es muy reducido.

5. PLS-LDA es un método muy competente para trabajar con datos genómicos y permite seleccionar los SNPs más informativos cuando se trata de solucionar un problema de asignación.

6. Se puede lograr un porcentaje de aciertos superior al 95% en la asignación de muestras a las razas del estudio leyendo únicamente los 12 mejores predictores para cada raza . Es decir, para las 11 razas del estudio, son necesarios 132 marcadores SNP.

7. Entre los SNPs que superaron los criterios genéticos, la diferencia de seleccionar los predictores con PLS-LDA a hacerlo al azar, supone una diferencia (para un porcentaje de aciertos > 95%) de 132 SNPs necesarios en el primer caso, frente a 195 cuando la elección es aleatoria.

8. El reducido número de SNPs que son necesarios para diferenciar entre razas tan semejantes entre sí como puedan ser las españolas es sólo otra muestra del potencial de las pruebas genómicas, de su utilidad a otros niveles y de su, cada vez mayor, rentabilidad.

9. Se han identificado hasta 66 genes en las posiciones en las que se encuentran los SNPs mejor predictores y se han distinguido ocho regiones de interés en el cariotipo que podrían ser las más relacionadas con los caracteres más particulares de las razas.

10. Se ha identificado el cromosoma 5-BTA como el que tiene un mayor peso en la diferenciación de las razas estudiadas.

11. Los resultados en las pruebas de asignación empeoran significativamente cuando se pretende diferenciar entre razas tan semejantes como lo puedan ser Morucha y Avileña-Negra Ibérica, o Brown Swiss y Fleckvieh

12. Se consigue la mayor varianza explicada para la raza Guernesey, que es precisamente la que cuenta con un número menor de observaciones.

13. Con los datos empleados para este estudio, cuesta más diferenciar entre las razas europeas que entre las razas españolas, aunque esto, posiblemente esté influenciado por el número de observaciones en uno y otro caso.

Las futuras investigaciones en este campo permitirán que la identificación de muestras procedentes de animales cruzados se convierta en algo asequible y practicable. Además, una investigación más exhaustiva de los SNPs seleccionados, así como de los QTLs próximos a ellos, permitirá conocer que rasgos o caracteres están más asociados con las diferentes razas.

8. BIBLIOGRAFÍA

1. [Abecasis., *et al*. 2001]. Abecasis, G.R., Cherny, S.S. Cookson, W.O. Cardon, L.R. (2001). GRR: graphical representation of relationship errors. Bioinformatics, 17, 742–743

2. [Aho., *et al*. 1988]. A. V. Aho, B. W. Kernighan, P. J. Weinberger, (1988). The AWK Programming Language. Addison-Wesley.

3. [Amaral., *et al*. 2009]. Amaral, A.J., Megens, H.J. Kerstens, H.H., Heuven, H.C., Dibbits, B., Crooijmans, R.P., Dunnen, J.T., Groenen, M.A. (2009). Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics. ;10:374. doi: 10.1186/1471-2164-10-374.

4. [Andrea. 2009]. Andrea, S. Foulkes, (2009). Applied Statistical Genetics with R. 1ª Edicion. Nueva York : Springer.

5. [Ayres. 2005]. Ayres, K.L. (2005). The expected performance of single nucleotide polymorphism loci in paternity testing. Forensic Science International 154, 167–72.

6. [Barker & Rayens. 2003]. Barker W, Rayens W.(2003) Partial least squares for discrimination.J. Chemom.;17: 166–173

7. [Baro., *et al*. 2012]. Baro, J.A., Carleos, C., Menendez-Buxadera, A., Rodriguez-Castañon., A., Cañon, J., 2012. Genetic variability underlying maternal traits of Asturiana de la Montaña beef cattle. Spanish Journal of Agricultural Research, 10: 69-73.

8. [Baudouin & Cornuet. 2004]. Baudouin, L., Piry, S. & Cornuet, J.M. (2004). Analytical Bayesian approach for assigning individuals to populations. Journal of Heredity 95, 217–24.

9. [Benson., *et al*. 2004]. Benson D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004). GenBank: update. Nucleic Acids Res., 32, D23–D26.

10. [Beja-Pereira., *et al.* 2003]. Beja-Pereira A1, Alexandrino P, Bessa I, Carretero Y, Dunner S, Ferrand N, Jordana J, Laloe D, Moazami-Goudarzi K, Sanchez A, Cañon J. 2003. Genetic characterization of southwestern European bovine breeds: a historical and biogeographical reassessment with a set of 16 microsatellites. J Hered. 2003 May-Jun;94(3):243-50.

11. [Beja-Pereira., *et al*. 2006]. Beja-Pereira, A., Caramelli, D., Lalueza-Fox, C. et al. (2006). The origin of European cattle: Evidence from modern and ancient DNA. Proceedings of the National Academy of Sciences of the United States of America 21, 8113–8.

12. [Bennett. 2004]. Bennett, S., (2004). Solexa Ltd. Pharmacogenomics 5,433 -438.

13. [Bjørn-Helge., *et al.* 2013]. Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland, (2013). pls: Partial Least Squares and Principal Component regression. R package version 2.4-3. http://CRAN.R-project.org/package=pls. Wold, J. Trygg, A. Berglund, H. Antti, Chemom. Intell. Lab. Syst. 58 (2001) 131–150

14. [Boulesteix. 2004]. A. L. Boulesteix, (2004). PLS dimension reduction for classification with microarray data, Statistical Applications in Genetics and Molecular Biology 3, Issue 1, Article 33.

15. [Boulesteix., *et al*. 2014]. Boulesteix., et al. (2014). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. Technical Report 171, Department of Statistics, LMU.

16. [Boulesteix., *et al*. 2014]. Anne-Laure Boulesteix, Sophie Lambert-Lacroix, Julie Peyre and Korbinian Strimmer, (2014). plsgenomics: PLS analyses for genomics. R package version 1.2-6. http://CRAN.R-project.org/package=plsgenomics

17. [Bozeman. 2014]. www.goldenhelix.com. Copyright © 2014 Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com.

18. [Cañas-Álvarez., *et al*. 2014a]. Cañas-Álvarez, J. J., González-Rodríguez, A., Martín-Collado, D., et al. (2014). Monitoring changes in the demographic and genealogical structure of the main Spanish local beef breeds. J. Anim. Sci.

19. [Cañas-Álvarez., *et al.* 2014b]. Cañas-Álvarez, J. J., González-Rodríguez, A., Munilla, S., et al. (2014). Genetic Diversity and Relationships among Spanish Beef Breeds Assessed by a Bovine High-density Chip 10th WCGALP.

20. [Carleos., *et al*. 2009]. Carleos, C., Baro, J.A., Villa, A., Cañon, J., 2009. Genetic parameter estimation with a mixed inheritance model in Asturiana de los Valles beef cattle breed. Archivos de Zootecnia, 58: 549-552.

21. [Casellas., *et al*. 2004]. Casellas, J., Jimenez, N., Fina, M., Tarres, J., Sanchez, A. & Piedrafita, J. (2004). Genetic diversity measures of the bovine Alberes breed using microsatellites, variability among herds and types of coat colour. Journal of Animal Breeding and Genetics 121, 101–10.

22. [Castric & Bernatchez. 2003]. Castric, V., Bernatchez, L. (2003). The rise and fall of isolation by distance in the anadromous brook charr Salvelinus fontinalis. Genetics, 163, 983–996.

23. [Cegelski., *et al*. 2003]. Cegelski, C.C., Waits, L.P. & Anderson, N.J. (2003). Assessing population structure and gene flow in Montana wolverines (Gulo

gulo) using assignment-based approaches. Molecular Ecology 12, 2907–18.

24. [Ciampolini., *et al*. 2006]. Ciampolini, R., Cetica, V., Ciani, E. et al. (2006). Statistical analysis of individual assignment tests among four cattle breeds using fifteen STR loci. Journal of Animal Science 84, 11–19.

25. [Corander., *et al*. 2003] Corander, J., Waldmann, P. & Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations. Genetics 163, 367–74.

26. [Cortés., *et al*. 2011]. Cortés O., I. Tupac-Yupanqui, M.A. García-Atance, S. Dunner, J. Fernández, J. Cañón. 2011. Paternal genetic variability into the lidia bovine breed. Archivos de Zootecnia, 60:417-420.

27. [Cunningham., *et al*. 2015]. Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos Garcín Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M.J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino and Paul Flicek. Ensembl (2015). Nucleic Acids Research 2015 43 Database issue:D662-D669

28. [Dalvit., *et al*. 2008]. Dalvit C., De Marchi M., Targhetta, C., Gervaso, M. & Cassandro, M. (2008). Genetic traceability of meat using microsatellite markers. Food Research International 41, 301–7.

29. [Dalvit., *et al*. 2007]. Dalvit, C., De Marchi, M., Cassandro, M. (2007). Genetic traceability of livestock products. A review. Meat Sci. 77:437-449.

30. [Devlin & Risch. 1995]. Devlin B., Risch N. (1995). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. Genomics 29 (2): 311-322.

31. [Falconer & Mackay. 1996]. Falconer, D.S. & Mackay, T.F.C. (1996). Introduction to Quantitative genetics. Longman Group, Essex, UK.

32. [Fisher. 1919]. Fisher R. A., 1919. XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 52: 399–433.

33. [Fisher. 1954]. Fisher, R.A. (1954). Statistical Methods for Research Workers.

Oliver and Boyd. ISBN 0-05-002170-2.

34. [Fries & Durstewitz. 2001]. Fries, R. & Durstewitz, G. (2001). Digital DNA signatures for animal tagging. Nature Biotechnology 19, 508.

35. [Garthwaite. 1994]. Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares, Journal of the American Statistical Association, 89 (425), 122-27.

36. [Gasparin., *et al*. 2005]. Gustavo Gasparin, Marcelo Miyata, Luiz Lehmann Coutinho, V, Mário Luiz Martinez, V, Marcos Vinícius G. Barbosa da Silva, Marco Antônio Machado, Ana Lúcia Campos, Luciana Correia de Almeida Regitano, V. (2005). Quantitative trait locus affecting birth weight on bovine chromosome 5 in a F2 Gyr x Holstein population. Genet. Mol. Biol. vol.28 no.4 São Paulo Oct./Dec

37. [Gil., *et al*. 2001]. Gil, M., Serra, X., Gispert, M., Oliver, M.A., Sañudo, C., Panea, B., Olleta, J.L., Campo, M.M. Oliván, M., Osoro, K. García-Cachán, M.D., Cruz-Sagredo, R., Izquierdo, M., Espejo, M., Martín, M., Piedrafita, J., (2001). The effect of breedproduction system on the myosin heavy chain 1, the biochemical characteristics and the colour variables of longissimus thoracis from seven Spanish beef cattle breeds. Meat Sci 58, 181-188. doi:10.1016/S0309-1740(00).00150-9

38. [Ginja., *et al.* 2010]. Ginja, C., L. Telo Da Gama, and M. C. Penedo. (2010b). Analysis of STR markers reveals high genetic structure in Portuguese native cattle. J. Hered. 101:201–210

39. [González-Rodríguez., *et al.* 2014]. A. González-Rodríguez, M.A.Toro,L. Varona, M. J Carabaño, J. J. Cañas-Álvarez,J. Altarriba,T.B R. da Silva, J. A. Baró, A. Molina,and C.Díaz. (2014). Genome-wide Analysis of Genetic Diversity in Autochthonous Spanish Populations of Beef Cattle. Proceedings, 10thWorld Congress of Genetics Applied to Livestock Production

40. [Goodman. 2004]. Goodman, D. (2004). Rural Europe redux. Reflections on alternative agro-food networks and paradigm change. Sociologia Ruralis, 44, 3–16.

41. [Gutiérrez., *et al*. 2003]. Gutiérrez, J.P., Altarriba, J., Díaz, C., Quintanilla, R., Cañón, J. Piedrafita, J. (2003). Pedigree analysis of eight Spanish beef cattle breeds. Genetics, Selection and Evolution, 35: 43-63.

42. [Hamada., *et al*. 2005]. Hamada, Watanabet, Chatani, K., Hayakawa, S., Iwamoto, M. (2005). Morphometrical comparison between Indian- and Chinese-derived rhesus macaques (Macaca mulatta). Anthropological Science. ;113:183–188.

43. [Heaton., *et al*. 2005]. Heaton, M.P., Keen, J.E., Clawson, M.L., Harhay, G.P., Bauer, N., Shultz, C., Green, B.T., Durso, L., Chitko-McKown, C.G. & Laegreid, W.W. (2005). Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. Journal of the American Veterinary Medical Association 226, 1311–4.

44. [Ibeagha-Awemu., *et al*. 2004]. Ibeagha-Awemu, E.M., Jann, O., Weimann, C. & Erhardt, G. (2004). Genetic diversity, introgression and relationships among West/ Central African cattle breeds. Genetics Selection Evolution 36, 673–90.

45. [Ilbery & Kneafsey. 2000]. Ilbery, B., & Kneafsey, M. (2000). Producer constructions of quality in regional speciality food production: A case study from South West England. Journal of Rural Studies, 16, 217–230.

46. [Jolliffe. 1982]. Jolliffe, Ian T. (1982). "A note on the Use of Principal Components in Regression". Journal of the Royal Statistical Society, Series C 31 (3): 300–303

47. [Jong. 1993]. SIMPLS: An alternative approach to partial least squares regression Chemometrics and Intelligent Laboratory Systems, Vol. 18, No. 3. (1993), pp. 251-263, doi:10.1016/0169-7439(93)85002-x by Sijmen de Jong

48. [Koskinen. 2003]. Koskinen, M.T. (2003). Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. Animal Genetics 34, 297–301.

49. [Krawczak. 1999]. Krawczak, M. (1999). Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis 20, 1676–81.

50. [Lewontin. 1988]. Lewontin R.C. (1988). On Measures of Gametic Disequilibrium. Genetics, 120(3): 849-852.

51. [Lindblad-Toh., *et al*. 2000]. Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., et al. (2000). Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat. Genet. 24 381–386.

52. [Liron., *et al*. 2004]. Liron, J.P., Ripoli, M.V., Garcia, P.P. & Giovambattista, G. (2004). Assignment of paternity in a judicial dispute between two neighbour Holstein dairy farmers. Journal of Forensic Science 49, 96–8.

53. [Martin-Burriel & Garcia-Muro. 1999]. Martin-Burriel, I., Garcia-Muro, E., Zaragoza, P. (1999). Genetic diversity analysis of six Spanish native cattle breeds using microsatellites. Anim Genet.;30:177-182.

54. [Martín-Burriel., *et al*. 2007]. Inmaculada Martín-Burriel, Clementina Rodellar,

Johannes A. Lenstra, Arianne Sanz, Carmen Cons, Rosarsio Osta, Miguel Reta, Santos De Argüello, Albina Sanz and Pilar Zaragoza. Genetic Diversity and Relationships of Endangered Spanish Cattle Breeds. Journal of heredity. 98, 687-691.

55. [Martín-Burriel., *et al*. 2011a]. Martín-Burriel, I., C. Rodellar, J. Cañón, O. Cortés, S. Dunner, V. Landi, A. Martínez Martínez, L.T. Gama, C. Ginja, M.C.T. Penedo, A. Sanz, P. Zaragoza, J.V. Delgado Bermejo (2011) Diversity of Iberian cattle Genetic diversity, structure and breed relationships in Iberian cattle. J Anim Sci, 89: 893-906

56. [Martín-Burriel., *et al*. 2011b]. Martín-Burriel, I., Rodellar, C, Cañon, J., Cortes, O., Dunner, S., Vincenzo, L., Martinez, A., Gama, L.T., Ginja, C., Zaragoza, P., Delgado, J.V., 2011. A global diversity and phylogenetic study of Iberian cattle using microsatellites. Journal of Animal Science, 89: 893-906.

57. [Martínez-Camblor., *et al*. 2014]. Martínez-Camblor P., Baro, J.A., Carleos, C., Cañon, J., (2014). Standard statistical tools for the breed allocation problem. Journal of Applied Statistics, 41(8), p. 1848-1856.

58. [Martínez-Camblor., *et al*. 2013]. Martínez-Camblor, P., Carleos, C., and Corral, N. (2013). General nonparametric ROC curve comparison, J. Korean Statist. Soc. 42, pp. 71–81.

59. [Martínez., *et al*. 2000]. Martínez, A.M., J.V. Delgado, A. Rodero and J.L. Vega-Pla. (2000). Genetic structure of the Iberian pig breed using microsatellites. Animal Genetics, 31: 295-301.

60. [Martínez., *et al*. 2004]. A.M. Martínez, M.P. Carrera, J.M. Acosta, P.P. Rodríguez-Gallardo, A. Cabello, E. Camacho and J.V. Delgado. (2004). Genetic characterisation of the Blanca Andaluza goat based on microsatellite markers. South African Journal of Animal Science 2004. 34, 17-19.

61. [Matukumalli., *et al*. 2009]. Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., 2009. Development and characterization of a high density SNP genotyping assay for cattle. DOI: 10.1371/journal.pone.0005350

62. [Maudet., *et al*. 2002]. Maudet, C., Luikart, G. & Taberlet, P. (2002). Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. Journal of Animal Science 80, 942–50.

63. [Méndez., *et al*. 2010]. Méndez, S., S. Dunner, J.A. García, S. de Argüello, Mª J. Crespo, N. Chomón, L. A. Calderón, B. Sañudo, J. Cañón. 2010. Characterization of the Cantabric water dog. Archivos de Zootecnia, 60:405-408.

64. [Moser., *et al*. 2007]. Moser G, Crump RE, Tier B, Sölkner J, Zenger KR, Khatkar MS, Cavanagh JAL, Raadsma HW. (2007). Genome based genetic evaluation and genome wide selection using supervised dimension reduction based on partial least squares. Proc Assoc Advmt Anim Breed Genet. ;17:227–230.

65. [Murdoch. 2000]. Murdoch, J. (2000). Networks – A new paradigm of rural development? Journal of Rural Studies, 16, 407–419.

66. [Negrini., *et al*. 2007]. Negrini, R., Milanesi, E., Colli, L., Pellecchia, M., Nicoloso, L., Crepaldi, P., et al. (2007). Breed assignment of Italian cattle using biallelic AFLP markers. Animal Genetics, 38, 147–153.

67. [Negrini., *et al*. 2008a]. Negrini, R., Nicoloso, L., Crepaldi, P., Milanesi, E., Colli, L., Chegdani, F., Pariset, L., Dunner, S., Leveziel, H., Williams, JL., et al. (2008). Assessing SNP markers for assigning individuals to cattle populations. Animal Genetics 40, (1).: 18-26.

68. [Negrini., *et al*. 2008b]. Negrini, R., et al. (2008). Traceability of four European Protected Geographic Indication (PGI) beef products using Single Nucleotide Polymorphisms (SNP) and bayesian statistics. Meat Science, 80, 1212-1217.

69. [Ohta., *et al*. 1973]. Ohta, T., and M. Kimura. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22: 201–204.

70. [Paetkau., *et al*. 1995]. Paetkau, D., Calvert, W., Stirling, I. & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4, 347–54.

71. [Palermo., *et al*. 2009]. Palermo G, Piraino P, Zucht HD (2009) Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. Advances and applications in bioinformatics and chemistry: AABC 2: 57–70

72. [Pariset., et al. 2006]. Pariset, L., Cappuccio, I., Ajmone-Marsan, P., Bruford, M., Dunner, S., Cortes, O., Erhardt, G., Prinzenberg, E-M., Gutscher, K., Joost, S., Pinto-Juma, G., Nijman, I.J., Lenstra, J.A., Perez, T., Valentín, A., for the Econogene Consortium, 2006. Characterization of 37 breed-specific single-nucleotide polymorphisms in sheep. Journal of Heredity, 97: 531-534.

73. [Peter., *et al*. 2006]. Peter, C., Bruford, M., Peretz, T., Dalamitra, S., Hewitt, G. & Erhardt, G. (2006). The ECONOGENE Consortium Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. Animal Genetics 38, 37–44.

74. [Pérez-Enciso & Tenenhaus. 2003]. M Pérez-Enciso, M Tenenhaus (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach - Human genetics.

75. [Piedrafita., *et al*. 2003]. Piedrafita, J., Quintanilla, R., Sañudo, C., Olleta, J.L., Campo, M.M., Panea, B., Reinand, G., Turin, F., Jabets, Osoro, K. et al. (2003). Carcass quality of 10 beef cattle breeds of the Southwest of Europe in their typical production systems. Livest Prod Sci 82, 1-13. doi:10.1016/S0301-6226(03).00006-X.

76. [Pritchard., *et al*. 2000]. Pritchard, J., Stephens, M. Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

77. [Purcell. 2007]. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

78. [R Core Team. 2014]. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

79. [Ramos., *et al*. 2011]. Ramos M, Megens HJ, Crooijmans RPM, Schook LB, Groenen M. (2011). Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Anim Genet;42:613–620.

80. [Ramos., *et al*. 2009]. Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. ;4:e6524. doi: 10.1371/journal.pone.0006524.

81. [Rannala., *et al*. 1997]. Rannala, B. & Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America 94, 9197–221.

82. [Reese., *et al*. 2010]. Reese JT, Childers CP, Sundaram JP, Dickens CM, Childs KL, Vile DC, Elsik CG. (2010). Bovine Genome Database: supporting community annotation and analysis of the Bos taurus genome. BMC Genomics. 11:645. PMID: 21092105.

83. [Reynolds., *et al*. 1983]. Reynolds, J., Weir, B.S., and Cockerham, C., (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779.

84. [Rogberg-Muñoz., *et al.* 2010]. Rogberg-Muñoz, A.Prando, A.Melucci, L.Villegas-Castagnaso, E.E.Ripoli, M.V.Peral-García, P.Baldo, A.Añon, M.C.Givambattista, G. (2010). Possible association of bovine chromosome 5 markers with growth and fat traits in Hereford cattle raised under extensive conditions.Livestock science. p. 186 - 186.

85. [Ruzzante., *et al*. 2001]. Ruzzante, D.E., Taggart, C.T., Doyle, R.W., Cook, D. (2001). Stability in the historical pattern of genetic structure of Newfoundland cod (Gadus morhua). despite the catastrophic decline in population size from 1964 to 1994. Conservation Genetics, 2, 257–269.

86. [Saínz., *et al*. 2011]. Saínz, R., J.A. García, S. de Argüello, F. Barquín, Mª J. Crespo, N. Chomón, L. A. Calderón, J. Cañón. 2011. Genetic structure of the Tudanca bovine breed inferred from the genealogical information. Archivos de Zootecnia, 60:401-404.

87. [Sanz., *et al*. 2007]. Sanz, A., Martín-Burriel, I., Rodellar, C., Osta, R., Sanz, A., Abril, F. y Zaragoza, P. (2007). Caracterización genética de la población bovina Serrana Negra de Teruel. Arch. Zootec., 56: 461-465.

88. [Schulz., *et al*. 2010]. Schulz, U, Tupac-Yupanqui, I., Martínez, A., Méndez, S., Delgado, J.V., Gómez, M., Dunner, S., Cañón, J. 2010. The Canarian camel: a traditional dromedary population. Diversity, 2: 561-571.

89. [Strimmer. 2014]. Korbinian Strimmer. (2014). plsgenomics: PLS analyses for genomics. R package version 1.2-6. http://CRAN.R-project.org/package=plsgenomics

90. [Troy., *et al*. 2001]. Troy, C.S., MacHugh, D.E., Bailey, J.F., Magee, D.A., Loftus R.T., Cunningham, P., Chamberlain A.T., Sykes, B.C, Bradley, D.G. (2001). Genetic evidence for Near-Eastern origins of European cattle. Nature. 410, 1088–1091.

91. [Tupac-Yupanqui., et al. 2010]. Tupac-Yupanqui, I, S. Dunner, B. Sañudo, A. González, S. de Argüello, F. Barquín, Mª J. Crespo, N. Chomón, C. Cimadevilla, L. A. Calderón, M. Fernández Y J. Cañón. 2010. Genetic characterization of the Caballo Monchino breed and its relationships with other Spanish local equine breeds. Archivos de Zootecnia, 60: 425-428.

92. [Van Bers., *et al*. 2010]. Van Bers NEM, Oers K van, Kerstens HHD, Dibbits BW, Crooijmans RPMA, Visser ME & Groenen MAM. 2010. Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. Molecular Ecology 19 (Suppl. 1), 89–99.

93. [Werner., *et al*. 2004]. Werner, F.A.O., Durstewitz, G., Habermann, F.A., Thaller, G., Krämer, W., Kollers, S., Buitkamp, J., Georges, M., Brem, G., Mosner, J., et

al. (2004). Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. Anim Genet 35:44-49.

94. [Weston., *et al*. 1997]. Weston, A., Pan, C.F., Ksieski, H.B., Wallenstein, S., Berkowitz, G.S., Tartter, P.I., Bleiweiss, I.J., Brower, S.T., Senie, R.T., Wolff, M.S. (1997). p53 haplotype determination in breast cancer. Cancer Epidemiol Biomarkers Prev6: 105–112

95. [Wiedmann., *et al*. 2008]. Wiedmann, R.T., Smith, T.P., Nonneman, D.J. (2008). SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. ;9:81. doi: 10.1186/1471-2156-9-81.

96. [Williams., *et al*. 2009]. Williams, J.L., Dunner, S., Valentín, A., Mazza, R., Amarger, V., Checa, M.L., Crisà, A., Razzaq, N., Delourme, D., Grandjean, F., Marchitelli, C., García, D., Gomez, R.P., Negrini, R., Ajmone-Marsan, P., Levéziel, H., (2009). Discovery, characterization and validation of single nucleotide polymorphisms within 206 bovine genes that may be considered as candidate genes for beef production and quality. Animal Genetics, 40: 486-491.

97. [Wold. 1995]. Wold S. 1995. PLS for multivariate linear modeling. In: van de Waterbeemd H,editor. Chemometric Methods in Molecular Design. Vol. 2. Weinheim:Verlag Chemie.

98. [Wright. 1921]. Wright, S. (1921). Systems of mating II: the effects of inbreeding on the genetic composition of a population. Genetics 6, 124–143.

99. [Wright. 1931]. Wright, S. (1931). Evolution in Mendelian populations. Genetics 16:97–259. 1940. Breeding structure of populations in relation to speciation. Am. Nat. 74:232–248.

100. [Zimin., *et al*. 2009]. Zimin,A.V., Delcher,A.L., Florea,L., Kelley,D.R., Schatz,M.C., Puiu,D., Hanrahan,F., Pertea,G., Van Tassell,C.P., Sonstegard,T.S. Marcais,G., Roberts,M., Subramanian,P., Yorke,J.A. and Salzberg,S.L. (2009) A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10 (4), R42

ANEXO I – Marcadores SNP seleccionados

Tabla 28: Selección de SNPs elegidos con PLS-LDA para lograr un 95% de aciertos en la asignación de nuevas muestras pertenecientes a alguna de las razas del estudio

Orden de selección	Cromosoma	Nombre	Gen presente en la posición	Coordenadas
1	18	BTA-162568		
2	18	rs135883818	ITFG1	18:15793120-15793220
3	2	rs137019068		2:7946538-7946638
4	2	rs109994609		2:3792529-3792629
5	2	rs135907790		2:7877329-7877429
6	6	rs135353878		6:67763434-67763534
7	7	rs109590662	LOC101903762	7:27293434-27293534
8	6	rs109094824		6:69666396-69666496
9	2	rs136579166	MAP3K2	2:5150616-5150716
10	6	rs137758365	PDGFRA	6:71397723-71397823
11	17	rs109807781		17:73559702-73559802
12	6	rs133715317		6:52976666-52976766
13	1	rs137370497		1:91013442-91013542
14	14	rs134726463		14:69063979-69064079
15	16	rs136411542	DNAH14	16:28920461-28920561
16	1	rs42871708		1:91926151-91926251
17	1	rs135146392	HTR3C	1:83683031-83683131
18	3	rs135732729		3:114039367-114039467
19	13	rs132853415	TOP1	13:70470268-70470368
20	5	rs110493818	FAM19A2	5:51946335-51946435
21	22	rs41998944	ARPC4	22:16964284-16964384
22	18	rs110615481	VPS35	18:15047625-15047725
23	5	rs110811338		5:43142868-43142968
24	9	rs137704698	SLC22A1	9:97766189-97766289
25	5	rs42851619		5:34201207-34201307
26	5	rs41604533	CSRNP2	5:28834271-28834371
27	5	rs110311254	MYO1A	5:56722521-56722621
28	7	rs43505488	DOT1L	7:22648583-22648683
29	25	rs41587822	SLC29A4	25:39525334-39525434

Orden de selección	Cromosoma	Nombre	Gen presente en la posición	Coordenadas
30	18	rs41858612	GLG1	18:1968612-1968712
31	5	rs133244172	RND1	5:31096365-31096465
32	17	rs42637225		17:9716091-9716191
33	20	rs135086085		20:6129912-6130012
34	24	rs42341325		24:37008351-37008451
35	1	rs136694729		1:4843210-4843310
36	3	rs137099532	AGBL4	3:98016445-98016545
37	1	rs42965066	CLSTN2	1:129451549-129451649
38	8	rs133497496	SCARA3	8:10993288-10993388
39	7	rs137234102		7:93054442-93054542
40	8	rs135238577		8:17475732-17475832
41	5	rs133409493		5:60045735-60045835
42	17	rs135014254		17:28101016-28101116
43	28	rs134024805		28:44921892-44921992
44	29	rs110348532	ANO1	29:47874555-47874655
45	15	rs42023581		15:5881049-5881149
46	25	rs136336564	HS3ST4	25:23849857-23849957
47	11		NGS-119429	
48	1	rs135806460	C1H8orf42	1:148320216-148320316
49	12	rs133665556		12:23157746-23157846
50	12	rs43697950		12:45880850-45880950
51	4	rs133726612	VWC2	4:6001248-6001348
52	13	rs41696817		13:55529539-55529639
53	17	rs136504743		17:4920759-4920859
54	4	rs41598977	CNTNAP2	4:111609173-111609273
55	21	rs41974397	FBXO22	21:31739875-31739975
56	13	rs133154840	LOC526745	13:65481209-65481309
57	10	rs43633970		10:55453548-55453648

Orden de selección	Cromosoma	Nombre	Gen presente en la posición	Coordenadas
58	12	rs109536038	LOC101903856, MTUS2	12:31040526-31040626
59	10	rs132836789		10:25430629-25430729
60	11	rs136942751		11:35723176-35723276
61	3	rs135995556		3:99990452-99990552
62	6	rs133208579		6:60530724-60530824
63	13	rs109177675		13:82643381-82643481
64	9	rs42632453		9:83339351-83339451
65	5		BTA-2048	
66	16	rs134325498	UCHL5	16:12819524-12819624
67	4	rs137615910		4:35980169-35980269
68	5	rs110041718		5:16627856-16627956
69	15	rs41777688	LRRC4C	15:71370729-71370829
70	18	rs133405722	POP4	18:40370572-40370672
71	9	rs135586223	LOC101905596	9:38307597-38307697
72	24	rs133623896	DLGAP1	24:38326889-38326989
73	26	rs109932090	NT5C2	26:24079744-24079844
74	15	rs109325605		15:60309419-60309519
75	11	rs135689536	KDM3A	11:48273661-48273761
76	8	rs43577415		8:102605263-102605363
77	26	rs137794748	LOC101908152	26:45557716-45557816
78	3	rs109999726		3:34617313-34617413
79	7	rs135871529	EPOR	7:16998526-16998626
80	16	rs136962050		16:24251836-24251936
81	26	rs110881165	DOCK1, FAM196A	26:46986637-46986737
82	18	rs132939608	COX4I1	18:11803428-11803528
83	2	rs136309759	HS6ST1	2:4120983-4121083
84	12	rs133215161		12:46367164-46367264

Orden de selección	Cromosoma	Nombre	Gen presente en la posición	Coordenadas
85	15	rs110670191		15:26018717-26018817
86	29	rs136179661		29:34015152-34015252
87	8	rs42751059		8:4445482-4445582
88	5	rs134983051		5:93532325-93532425
89	11	rs137388076		11:27211163-27211263
90	3	rs42752357	CSMD2	3:112311149-112311249
91	1	rs134158660	LOC101907397	1:93205934-93206034
92	11	rs132965186		11:88727702-88727802
93	29	rs135380781		29:32989704-32989804
94	27	rs42111141		27:9523339-9523439
95	26	rs132771159		26:33784270-33784370
96	11	rs42126433		11:35338939-35339039
97	9	rs42762746		9:81813467-81813567
98	17	rs42851499	STX2	17:47325179-47325279
99	1	rs136398832	EPHA6	1:41193502-41193602
100	21	rs135876490		21:52108825-52108925
101	16	rs137417925		16:35677225-35677325
102	1	rs137570137		1:747523-747623
103	5	rs132762926		5:82617720-82617820
104	26	rs42937889	LOC522146	26:16517485-16517585
105	5	rs109076663	MRPS35	5:82527977-82528077
106	17	rs109188645	TMEM132C	17:49580280-49580380
107	13	rs134759938	PPP1R16B	13:68278399-68278499
108	7	rs110107876		7:94006520-94006620
109	2	rs134286256		2:49742561-49742661
110	13	rs41616451		13:63715216-63715316

Orden de selección	Cromosoma	Nombre	Gen presente en la posición	Coordenadas
111	14	rs109947702	ASPH	14:28714991-28715091
112	3	rs109739448		3:63446545-63446645
113	18	rs133391211	SIPA1L3	18:48103789-48103889
114	6	rs43496947	RNF4	6:108333931-108334031
115	15	rs109127570	LOC101906823	15:42836646-42836746
116	19	rs133600489		19:5960927-5961027
117	18	rs134922545	CYLD	18:19276264-19276364
118	4	rs137630132	LOC101906647	4:62500185-62500285
119	5	rs108949332		5:78740741-78740841
120	24	rs133648402	GATA6	24:34562092-34562192
121	2	rs133243829	LRP1B	2:54944905-54945005
122	14	rs136833003	EIF3H	14:49814485-49814585
123	17	rs110254328	SLC24A6	17:63571731-63571831
124	15	rs133118277	DSCAML1	15:28571512-28571612
125	3	rs133790058		3:115727170-115727270
126	1	rs135280580		1:86567029-86567129
127	14	rs109498677		14:71061116-71061216
128	1	rs43277995	PWP2	1:146850293-146850393
129	4	rs137810150		4:102781554-102781654
130	10	rs43129506	SIPA1L1	10:83392168-83392268
131	24	rs135089094		24:60794481-60794581
132	9	rs134387737	MAP7	9:75416244-75416344

ANEXO II – Manipulación de los ficheros de datos

1- Archivos de partida

En este estudio se ha partido de una colección de archivos *.map* y *.beagle*. Los datos de las razas españolas son propiedad del proyecto Selgenbeef mientras que los datos de las razas europeas pertenecen al proyecto europeo Gene2Farm. Estas colecciones de archivos se han facilitado para este estudio en la forma de un fichero de datos por cada raza y por cada cromosoma. No se incluía información del cromosoma X ni del ADN mitocondrial. En total, se ha dispuesto para este estudio de 319 ficheros *.map* y *.beagle* (11 razas x 29 cromosomas = 319).

A continuación se hacen algunos comentarios acerca de en que consistió la fase de manipulación de ficheros de este trabajo.

La siguiente imagen muestra ordenados algunos de los archivos *.map* que se han utilizado para este estudio:

FUENTE: ELABORACIÓN PROPIA

```
Archivo  Editar  Ver  Terminal  Pestañas  Ayuda
fern@fern-HP-15-Notebook-PC:~$ cd Documentos
fern@fern-HP-15-Notebook-PC:~/Documentos$ cd chromosomes
fern@fern-HP-15-Notebook-PC:~/Documentos/chromosomes$ ls
10.chrom10.map  1.chrom16.map  3.chrom21.map  5.chrom27.map  7.chrom5.map
10.chrom11.map  1.chrom17.map  3.chrom22.map  5.chrom28.map  7.chrom6.map
10.chrom12.map  1.chrom18.map  3.chrom23.map  5.chrom29.map  7.chrom7.map
10.chrom13.map  1.chrom19.map  3.chrom24.map  5.chrom2.map   7.chrom8.map
10.chrom14.map  1.chrom1.map   3.chrom25.map  5.chrom3.map   7.chrom9.map
10.chrom15.map  1.chrom20.map  3.chrom26.map  5.chrom4.map   8.chrom10.map
10.chrom16.map  1.chrom21.map  3.chrom27.map  5.chrom5.map   8.chrom11.map
10.chrom17.map  1.chrom22.map  3.chrom28.map  5.chrom6.map   8.chrom12.map
10.chrom18.map  1.chrom23.map  3.chrom29.map  5.chrom7.map   8.chrom13.map
10.chrom19.map  1.chrom24.map  3.chrom2.map   5.chrom8.map   8.chrom14.map
10.chrom1.map   1.chrom25.map  3.chrom3.map   5.chrom9.map   8.chrom15.map
10.chrom20.map  1.chrom26.map  3.chrom4.map   6.chrom10.map  8.chrom16.map
10.chrom21.map  1.chrom27.map  3.chrom5.map   6.chrom11.map  8.chrom17.map
10.chrom22.map  1.chrom28.map  3.chrom6.map   6.chrom12.map  8.chrom18.map
10.chrom23.map  1.chrom29.map  3.chrom7.map   6.chrom13.map  8.chrom19.map
10.chrom24.map  1.chrom2.map   3.chrom8.map   6.chrom14.map  8.chrom1.map
10.chrom25.map  1.chrom3.map   3.chrom9.map   6.chrom15.map  8.chrom20.map
10.chrom26.map  1.chrom4.map   4.chrom10.map  6.chrom16.map  8.chrom21.map
10.chrom27.map  1.chrom5.map   4.chrom11.map  6.chrom17.map  8.chrom22.map
10.chrom28.map  1.chrom6.map   4.chrom12.map  6.chrom18.map  8.chrom23.map
10.chrom29.map  1.chrom7.map   4.chrom13.map  6.chrom19.map  8.chrom24.map
10.chrom2.map   1.chrom8.map   4.chrom14.map  6.chrom1.map   8.chrom25.map
10.chrom3.map   1.chrom9.map   4.chrom15.map  6.chrom20.map  8.chrom26.map
10.chrom4.map   2.chrom10.map  4.chrom16.map  6.chrom21.map  8.chrom27.map
10.chrom5.map   2.chrom11.map  4.chrom17.map  6.chrom22.map  8.chrom28.map
10.chrom6.map   2.chrom12.map  4.chrom18.map  6.chrom23.map  8.chrom29.map
10.chrom7.map   2.chrom13.map  4.chrom19.map  6.chrom24.map  8.chrom2.map
10.chrom8.map   2.chrom14.map  4.chrom1.map   6.chrom25.map  8.chrom3.map
10.chrom9.map   2.chrom15.map  4.chrom20.map  6.chrom26.map  8.chrom4.map
11.chrom10.map  2.chrom16.map  4.chrom21.map  6.chrom27.map  8.chrom5.map
11.chrom11.map  2.chrom17.map  4.chrom22.map  6.chrom28.map  8.chrom6.map
11.chrom12.map  2.chrom18.map  4.chrom23.map  6.chrom29.map  8.chrom7.map
11.chrom13.map  2.chrom19.map  4.chrom24.map  6.chrom2.map   8.chrom8.map
11.chrom14.map  2.chrom1.map   4.chrom25.map  6.chrom3.map   8.chrom9.map
11.chrom15.map  2.chrom20.map  4.chrom26.map  6.chrom4.map   9.chrom10.map
11.chrom16.map  2.chrom21.map  4.chrom27.map  6.chrom5.map   9.chrom11.map
11.chrom17.map  2.chrom22.map  4.chrom28.map  6.chrom6.map   9.chrom12.map
11.chrom18.map  2.chrom23.map  4.chrom29.map  6.chrom7.map   9.chrom13.map
11.chrom19.map  2.chrom24.map  4.chrom2.map   6.chrom8.map   9.chrom14.map
11.chrom1.map   2.chrom25.map  4.chrom3.map   6.chrom9.map   9.chrom15.map
```

Ilustración 24: Ficheros *.map* utilizados en el estudio.

En la ilustración no aparecen los 319 archivos .map, pero se puede deducir el orden de los mismos. El comando *ls* de *Linux* muestra los archivos que hay dentro de un directorio. El directorio “chromosomes” en este caso.

- Ficheros *.map*

Los archivos .map son habituales para trabajar con datos genómicos y proporcionan la información de los marcadores moleculares en el software PLINK.

En estos ficheros cada línea corresponde a un marcador y debe tener exactamente 4 columnas, que deben ser las siguientes:

- Cromosoma
- Identificador SNP
- Distancia genética (morgans)
- Posición (unidades de pares de bases)

Se muestra a continuación un ejemplo de fichero *.map:*

FUENTE: ELABORACIÓN PROPIA

```
fern@fern-HP-15-Notebook-PC:~/Documentos$ head 1.chrom28.map
28      BovineHD2800000001      0       5302
28      BovineHD2800000003      0       11110
28      BovineHD2800000004      0       12640
28      BovineHD2800000007      0       21994
28      BovineHD2800000008      0       25713
28      BovineHD2800000009      0       26638
28      BovineHD2800000010      0       28194
28      BovineHD2800000011      0       29841
28      BovineHD2800000012      0       30475
28      BovineHD2800000013      0       31503
fern@fern-HP-15-Notebook-PC:~/Documentos$
```

Ilustración 25: Ejemplo de fichero *.map*

La imagen corresponde a la llamada en la terminal del archivo *1.chrom28.map*. Ponemos delante la palabra *head* para que únicamente aparezcan en pantalla las 10 primeras filas correspondientes a los 10 primeros SNPs. El archivo *1.chrom28.map*. contiene las cuatro columnas de información vistas para todos los marcadores en el

cromosoma 28 para la raza 1. No se dispone de la información de la distancia genética como se aprecia porque la columna 3 aparece con valor cero en todos los casos; aunque si se conoce la posición, que viene referenciada respecto de uno de los extremos del cromosoma. Como puede verse, los marcadores están ordenados según su posición desde un extremo hasta el otro. Los archivos *.map* suelen contener, en el caso del Beadchip utilizado, información de varias decenas de miles de SNPs. En los cromosomas más grandes como pueden ser el 1 o el 5 se han leído hasta 50000 SNPs, mientras que en los cromosomas más pequeños, como el 24 o el 25, se leen en torno a los 10000 SNPs.

La identificación del SNP que figura en pantalla no es la oficial sino la que se ha asignado para los proyectos Selgenbeef y Gene2Farm. La identificación oficial de los SNPs puede encontrarse en cualquiera de los muchos buscadores genéticos. En este estudio se han utilizado Genebank y Ensembl.

- Ficheros *.beagle*

Los archivos *.beagle* contienen la información genómica y pueden ser de dos tipos:

- Archivos *.beagle.* En estos archivos no se ha respetado la información de las fases alélicas.
- Archivos *.beagle.phased.* En estos archivos si se han respetado las fase alélicas y podemos conocer, además de los haplotipos, como son los alelos en cada uno de los cromosoma homólogos. Los archivos *.beagle* proporcionados para este estudio son de este tipo, aunque en nuestro caso no se va a utilizar la información de las fases alélicas.

A continuación se muestra un ejemplo de un archivo *.beagle.phased*

FUENTE: ELABORACIÓN PROPIA

```
fern@fern-HP-15-Notebook-PC:~/Documentos$ head Beagle.10.chromo28.pre_phase.bgl.
phased
I id col.3 col.4 col.5 col.6 col.7 col.8 col.9 col.10 col.11 col.12 col.13 col.1
4 col.15 col.16 col.17 col.18 col.19 col.20 col.21 col.22 col.23 col.24 col.25 c
ol.26 col.27 col.28 col.29 col.30 col.31 col.32 col.33 col.34 col.35 col.36 col.
37 col.38 col.39 col.40 col.41 col.42 col.43 col.44 col.45 col.46 col.47 col.48
col.49 col.50 col.51 col.52 col.53 col.54 col.55 col.56 col.57 col.58
# sampleID GLENLEA_PREMIER_900016       GLENLEA_PREMIER_900016  LES_JAONNETS_DAY
S_DREAMBOY_20075004     LES_JAONNETS_DAYS_DREAMBOY_20075004     LES_JAONNETS_IRI
S_IVAN_20055011 LES_JAONNETS_IRIS_IVAN_20055011 LES_JAONNETS_TEMPESTS_CONCORD_20
030014  LES_JAONNETS_TEMPESTS_CONCORD_20030014  MAPLE_LEAF_LINDON_SULTAN_910019M
APLE_LEAF_LINDON_SULTAN_910019  MEADOW_COURT_ELLYS_ERIC_20085003        MEADOW_C
OURT_ELLYS_ERIC_20085003        MYRTLES_SUMMET_II_CICERO_970004 MYRTLES_SUMMET_I
I_CICERO_970004 MYRTLES_SUMMETS_MAGIC_960004    MYRTLES_SUMMETS_MAGIC_960004   R
OSALEAS_RINGO_OF_LES_JAONNETS_930021    ROSALEAS_RINGO_OF_LES_JAONNETS_930021  R
OZELYN_PATMAR_JAY_GLACIER_602961        ROZELYN_PATMAR_JAY_GLACIER_602961      L
ES_JAONNETS_CARAS_AMIR_20075002 LES_JAONNETS_CARAS_AMIR_20075002        LES_JAON
NETS_CARAS_CONQUEROR_20055012   LES_JAONNETS_CARAS_CONQUEROR_20055012   LES_JAON
NETS_FASCINATIONS_SUNNYBOY_20040014     LES_JAONNETS_FASCINATIONS_SUNNYBOY_20040
014     LES_JAONNETS_PEDROS_WORKMAN_20055009    LES_JAONNETS_PEDROS_WORKMAN_2005
5009    MEADOW_COURT_ELLYS_EPIC_20065002        MEADOW_COURT_ELLYS_EPIC_20065002
MEADOW_COURT_RED_OAK_20010008   MEADOW_COURT_RED_OAK_20010008   MEADOW_COURT_ROS
ES_RONALD_20055005      MEADOW_COURT_ROSES_RONALD_20055005      MYRTLES_ROB_ROY_
20055004        MYRTLES_ROB_ROY_20055004        IDLE_GOLD_E_CHALLENGE_ET_604268I
DLE_GOLD_E_CHALLENGE_ET_604268  JENS_GOLD_C_BLUE_SPRUCE_68012543        JENS_GOL
D_C_BLUE_SPRUCE_68012543        LANGHAVEN_SPIDER_NOMAR_68013555 LANGHAVEN_SPIDER
_NOMAR_68013555 RIVERWOOD_TILLER_KHAN_605000    RIVERWOOD_TILLER_KHAN_605000   S
PRING_WAL_SHERBERTS_MINT_68017224       SPRING_WAL_SHERBERTS_MINT_68017224     B
eechgroves_Cornelius_43538      Beechgroves_Cornelius_43538     Easby_Jordans_Ro
cket_42352      Easby_Jordans_Rocket_42352      Tiresford_Pedro_43490   Tiresfor
d_Pedro_43490   Tiresford_Shaka_43551   Tiresford_Shaka_43551   Kenvin_Posh_Padd
y_43608 Kenvin_Posh_Paddy_43608
M BovineHD2800000001 A G A G G G A G A G G G G A G G G A G A G A G G G G G G
 G G G G G G G A A G A A G G G G A G A G A G G G G G
M BovineHD2800000003 A G A A A G G G G G A G G A G G G A A A G A A A A G A A G A
 A A G G A A G G G A G G A G A G G G A G G G G A A A G
M BovineHD2800000004 A G A A A G G G G G A G G A G G G A A A G A A A A G A A G A
```

Ilustración 26: Ejemplo de fichero de tipo *.beagle*

En la ilustración se muestra la información del 1º SNP del cromosoma 28 (BovineHD2800000001) para 56 individuos. Como puede verse, aparecen en la ilustración 58 columnas, de ellas las dos primeras contienen la información acerca de la identificación del SNP. Por una cuestión de espacio, las 58 columnas no aparecen una a continuación de la otra, sino que llegado al extremo de la pantalla, el programa continua colocando las columnas en la fila siguiente. Por debajo de la numeración de las columnas figura el nombre de los animales correspondientes. Aparecen 56 nombres de animales separados por tabulaciones.

Como puede verse, la información genómica que proporcionan los dos primeros marcadores (abajo en la ilustración) corresponde a dos alelos: Adenina (A) y Citosina (C). Como hemos dicho, los marcadores SNP proporcionan información bialélica.

La ilustración da una idea acerca de como es la variabilidad. Estos dos marcadores del cromosoma 28 tienen, una gran variabilidad, como puede verse por el hecho de que ambos alelos se repiten bastante en los individuos. No obstante, para trabajar con los datos en PLINK ha sido necesario transformar estos archivos .beagle.phased en archivos *.ped.* Para ello se pueden transformar los archivos en una hoja de cálculo y darle el nuevo aspecto deseado a la distribución de filas y columnas, o se puede utilizar AUK, que como se ha comentado es un lenguaje de programación especialmente indicado para manipulación de ficheros.

- Ficheros .*ped*

Los ficheros .*ped* deben su nombre al término *pedigree* y como los archivos .*beagle* contienen la información genómica. Los ficheros contienen básicamente la misma información que los archivos .*beagle* pero la distribución de filas y columnas es distinta. Como podemos ver:

Ilustración 27: Ejemplo de fichero *.ped*

Cada letra (alelo) en la ilustración corresponde a una columna pero como se ha comentado, por una cuestión de espacio, las columnas se siguen en la línea siguiente. En la ilustración se muestra parte del genotipo de un individuo perteneciente a la raza 1 de identificación 12M0842. El archivo *.ped* corresponde al cromosoma 28 y cada línea (cada animal) tendrá 12378 columnas, una por cada SNP analizado en el cromosoma 28. Puesto que se analizan muchos SNPs vemos que las cuatro bases nitrogenadas (alelos) se encuentran repetidos varias veces.

La distribución de las columnas es la siguiente.

- Identificación de familia. En nuestro caso se pone un número de 1 a 11 indicando la raza a la que pertenece la muestra.
- Identificación individual
- Identificación paterna
- Identificación materna
- Sexo: Es una variable categórica con 3 posibles valores: 1 para machos, 2 para hembras y 0 para individuos desconocidos. En este estudio no se tiene en cuenta el sexo de las muestras. De hecho, el estudio debe ser aplicable a piezas de alimentos.
- Fenotipo: el fenotipo corresponde en este caso a la información genómica.

*En el caso de este estudio, puesto que se trata de individuos fundadores las columnas de identificación de los parentales llevan siempre valor 0.

Dado el tamaño de los archivos es habitual utilizar un formato binario para reducir el tiempo de espera en determinados análisis. También es habitual utilizar la forma transpuesta de los ficheros. Los archivos *.tped* en lugar de tener un genotipo por fila, tienen un marcador por fila y cada columna corresponde a un animal.

Los archivos *.map* puesto que sólo contienen información de los SNPs son mucho más ligeros y su tamaño suele estar en torno a 1MB, dependiendo del tamaño del cromosoma. Los archivos *.ped*, sin embargo, contienen toda la información genómica y son mucho más pesados. Suelen pesar en torno a los 10 – 15 MB dependiendo del número de muestras analizadas en cada raza. Por otro lado, los archivos pueden provocar errores en determinados programas por su número tan elevado de caracteres.

2-Principales comandos utilizados en el manejo

Se explican a continuación algunos de los comandos más importantes de los diferentes bloques de comandos diseñados para este estudio.

BLOQUE DE COMANDOS EN PLINK PARA EL ANÁLISIS ESTADÍSTICO

Para realizar el análisis genéticos de los datos genómicos se eligen unos criterios de calidad en base a los cuales se reducirá la colección de SNPs. Se introducirán en la terminal unas órdenes en forma de comando para PLINK. Los comandos a introducir son fundamentalmente unos valores umbrales predeterminados para los criterios que se ha decidido considerar. Las líneas de comandos permiten utilizar varios comando de manera simultánea cuando este sea nuestro interés. Puesto que PLINK permite trabajar sin necesidad de una instalación, es necesario para cada línea de comando especificarle a la terminal que queremos que ejecute PLINK. En el caso de tres de las cuatro líneas de comandos utilizadas es necesario especificar cual es el fichero de entrada sobre el que se ejecutan las órdenes en forma de comando y cual es el nombre del nuevo archivo que se genera tras las ejecución. Para la tercera línea de comandos, sin embargo, no es necesario especificar un nombre para el archivo de salida, puesto que será el propio programa el que asigne los nombres. En el caso del comando fundamental de esta línea (*--indep*) no se genera un archivo único sino que se generan dos.

Cuando establecemos un nombre para el archivo de salida, éste tomará el nombre indicado con independencia de que ya haya otro archivo con ese nombre en la carpeta de destino. PLINK sobreescribirá el archivo y la versión anterior no será recuperable.

```
p-link --file nobin1crom1 --merge-list 00025-2list.txt --cow
--recode --out 3abril

p-link --file 3abril --mind 0.1 --out 3abrilm

p-link --file 3abril --cow –indep 50 5 2

p-link --file 3abrilm --cow --extract plink.prune.in --geno
0.0 --maf 0.03 --hwe 0.001 --recode12 --tab --out
3abrilDone
```

BLOQUE DE COMANDOS EN AWK PARA LA TRANSFORMACIÓN DEL FICHERO .PED

Para la manipulación de ficheros no es necesario utilizar ningún programa sino que será necesario con emplear un lenguage de programación. AWK se ha diseñado específicamente para manipulación de ficheros de tipo texto y flujos de datos. Será necesario especificar a la terminal que queremos usar AWK antes de introducir ningún comando. El comando gsub se refiere a transformaciones del tipo “cambio de texto”. Entre paréntesis se especifica cual es la trasformación que queremos efectuar, y en nuestro caso, especificamos que queremos que además de llevarse a cabo la transformación, esta quede impresa en un archivo de salida. Como en el caso de PLINK, es necesario especificar cual es el archivo de entrada y cual el de salida. Separaremos ambos archivos mediante “>”. El comando de salida será el último en escribirse en la línea de comandos.

```
awk '{gsub("2 2","2");print}' 3abrilDone.ped > 513poster_tmp.ped

awk '{gsub("1 2","1");print}' 513poster_tmp.ped > 513poster2_tmp2.ped

awk '{gsub("1 1","0");print}' 513poster2_tmp2.ped > 3abrilDone_tmp3.ped
```

COMANDO EN R PARA IMPORTAR EL FICHERO TRANSFORMADO .PED CREADO CON PLINK

Para la importación del fichero en R también es necesario especificar cual es el archivo a importar y cual es el nombre que le queremos dar al archivo tras su paso por R. En el caso de R el nombre del archivo de destino debe ir al principio en la línea de comandos.

```
Tmil<-read.csv("3abrilDone_tmp3.ped", header=FALSE, sep="\t")
```

BLOQUE DE COMANDOS EN R PARA LA SELECCIÓN DE LOS MEJORES PREDICTORES EN BASE A PLS-LDA

Para la implementación de PLS-LDA en R es necesaria una única línea de comandos, que sería la tercera del bloque. Sin embargo, es necesario especificar antes cuales son las variables independientes y cuales las dependientes, que en nuestro caso son las once variables dummy. La última línea de comandos proporcionará el resultado cruz-validado de la predicción.

```
Nmarcadores = 20

indices = unique ( c(sapply(1:11, function(i)
variable.selection(m,1+
(pobla==paste(i)),nvar=nmarcadores))))

mi.pls.lda = pls.lda (m[indices], pobla,
ncomp=11*nmarcadores-1, nruncv=20)

mean(mi.pls.lda$predclass == pobla)
```

BLOQUE DE COMANDOS PARA LA OBTENCIÓN DE LA CAPACIDAD EXPLICATIVA DEL MODELO (R^2)

```
o<-m[indices]
op<-as.matrix(o)
lidia <- cbind.data.frame(pobla)
idx <- sort(unique(pob))

dummy <- matrix(NA, nrow = nrow(pob), ncol =
length(idx))
for (j in 1:length(idx)) {
  dummy[,j] <- as.integer(pob == idx[j])
}

lidia$P<-dummy
lidia$Loc<-op
lidiaTrain <- idx[1:400,]
lidiaTest <- idx[401:726,]

library("pls")

lidia1 <- plsr(P ~ Loc, ncomp = 132, data = lidiaTrain,
validation = "LOO")

summary(lidia1)
```

En primer lugar se especifica que los predictores van a ser aquellos que perteneciendo a la matriz *m*, estén dentro de la lista de nombre "*índices"* anteriormente descrita. La lista *índice* contiene 11 predictores por el número fijado en *Nmarcadore*s y siempre toma los mejores predictores para cada raza. Con estos SNPs seleccionados se crea el objeto de R

"*o*". Para darle a este objeto forma de matriz utilizamos la función *as.matrix*. Creamos así la matriz "*op*". *op*, al igual que *m c*ontiene los datos genómicos y por tanto, los predictores, pero en este caso es más reducida. Mientras que la dimensión de *m* era (726 x 87918), la dimensión de *op* es (726 x 132).

Para considerar los predictores en un análisis así como la variable dependiente es necesario crear un marco de datos. Dicho marco de datos se puede crear a partir de cualquier objeto, ya sea el vector *pobla* o la matriz *op*, siempre y cuando las dimensiones se corresponden. Tenemos que dim (*pobla*) es 726 x 1, luego concuerda co*n op* (726 x 132). Creamos el marco de datos *lidia*, que pasará a tener toda la información necesaria para hacer la regresión y para calcular la capacidad explicativa del modelo.

Puesto que para calcular la verdadera capacidad explicativa del modelo es necesario probar la predicción con individuos que no se tuvieron en cuenta en la construcción del modelo, distinguiremos nuestras observaciones en dos fases, una fase de prueba (Test) y otra de entrenamiento (Train). Para que la asignación de las observaciones a una u otra fase sea arbitraria se reordenarán las columnas del vector *pobla*. Para ello utilizamos el comando de *R sort.* Utilizamos el comando *unique* para especificar que querems muestreo con repetición, es decir, si se extrae un individuo para una determinada fase, ese individuo ya no se tendrá en cuenta en los siguientes sorteos de individuos.

Posteriormente creamos una función en R para descomponer una variable categórica en variables dummy. La función de R implementada es:

```
dummy <- matrix(NA, nrow = nrow(pob), ncol =
length(idx))

for (j in 1:length(idx)) {

  dummy[,j] <- as.integer(pob == idx[j])
}
```

El resultado es una matriz de nombre *dummy* que tendrá tantas columnas como categorías haya en el vector *pob*. Es decir, la matriz *dummy* tendrá 11 columnas. Para especificar que queremos una columna por cada posible categoría construimos un bucle de ordenes con *for*.

El comando *for* reiterará la siguiente operación: "creame una columna binaria en

dummy cuando *idx* tome el valor [j].” La igualdad especificada en el comando *as.integer,* es parte del bucle *for*, que a su vez es parte de la función implementada para obtener *dummy* y tiene la siguiente forma:

```
dummy[,j] <- as.integer(pob == idx[j])
```

Esta linea se interpretaría como “Para cada caso pon un valor 1 en aquellas líneas en que el valor del vector *pob* es igual al número de la variable columna que se esté creando para la matriz *dummy* en esa iteración del bucle”. Dicha columna binaria es lo que se conoce como variable *dummy*, aunque en este caso se llama *dummy* al conjunto de variables dummy. Para referirnos a las filas utilizamos *idx*, puesto que como se explicó en una línea anterior, el objeto corresponde a las filas tomando un orden aleatorio.

En el bucle, el primer valor [j], lógicamente será el 1 y el último será aquel que no sea ninguno de los ya considerados y que sea el que más se aleja de 1. En este caso, el último valor de [j] sería 11, puesto que es el valor que más se aleja de 1 de todos los posibles valores que puede tomar *idx*. Cuando [j] sea igual a 11, se efectuará el último bucle con *for* y se creará una última columna binaria en la matriz *dummy*.

Una vez disponemos de las variables dummy especificamos que esta matriz *dummy* forma parte del marco de datos creado (*lidia*), puesto que va a ser la variable dependiente en la predicción. También especificamos que la matriz *op* forma parte del marco de datos y son los predictores. Utilizamos el símbolo del dollar ($) para considerar que una matriz forma parte de un marco de datos. La matriz *op* recibe el nombre de *Loc* (loci) en el marco de datos. *Loc* viene a ser como un grupo de columnas del marco de datos. La variable *población* recibe el nombre de *P* dentro del marco de datos *lidia*. Puesto que hemos ordenado las líneas con caracteres aleatorio podremos decir que las 400 primeras filas del marco de datos formen la fase de entrenamiento y el resto formarán la fase de entrenamiento. Ahora podemos cargar el paquete de R “pls” e implementar una regresión PLS con los datos de la fase de entrenamiento. El comando *summary* nos proporciona la capacidad explicativa para cada una de las categorías de la variable dependiente, y se obtendrá una tabla similar a la expuesta en (tabla 27 - Resultados).

3- Algunos ejemplo de la realización del análisis

A continuación se muestra un extracto de la información que apareció en la terminal tras insertar en PLINK la última línea de comandos que hemos especificado para este Anexo y que permite concluir el análisis genético:

FUENTE: ELABORACIÓN PROPIA

```
+++ PLINK 1.9 is now available! See above website for details +++

Writing this text to log file [ 3abrilDone.log ]
Analysis started: Sat May 16 23:58:31 2015

Options in effect:
        --file 3abril
        --cow
        --extract plink.prune.in
        --geno 0.0
        --maf 0.03
        --hwe 0.001
        --recode12
        --tab
        --out 3abrilDone

702422 (of 702422) markers to be included from [ 3abril.map ]
726 individuals read from [ 3abril.ped ]
0 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
0 cases, 0 controls and 726 missing
364 males, 362 females, and 0 of unspecified sex
Reading list of SNPs to extract [ plink.prune.in ] ... 123285 read
Before frequency and genotyping pruning, there are 123285 SNPs
726 founders and 0 non-founders found
14638 markers to be excluded based on HWE test ( p <= 0.001 )
        0 markers failed HWE test in cases
        0 markers failed HWE test in controls
Total genotyping rate in remaining individuals is 1
0 SNPs failed missingness test ( GENO > 0 )
21268 SNPs failed frequency test ( MAF < 0.03 )
After frequency and genotyping pruning, there are 87918 SNPs
After filtering, 0 cases, 0 controls and 726 missing
After filtering, 364 males, 362 females, and 0 of unspecified sex
Writing recoded ped file to [ 3abrilDone.ped ]
Writing new map file to [ 3abrilDone.map ]

Analysis finished: Sun May 17 00:04:30 2015
```

Ilustración 28: Obtención del fichero para importar en R

Como ya se ha comentado, el fichero de nombre *3abrilDone* es importado en R tras unas modificaciones sencillas en AWK. En la ilustración puede observarse que el fichero cuenta con 87918 SNPs tras haber concluido el análisis genético.

A continuación se muestra una ilustración correspondiente a un pantallazo del

programa R en ejecución. Se ha decidido mostrar esta ilustración porque en ella se muestran algunos aspectos que pueden resultar útiles para explicar como es la manipulación de los ficheros en R:

FUENTE: ELABORACIÓN PROPIA

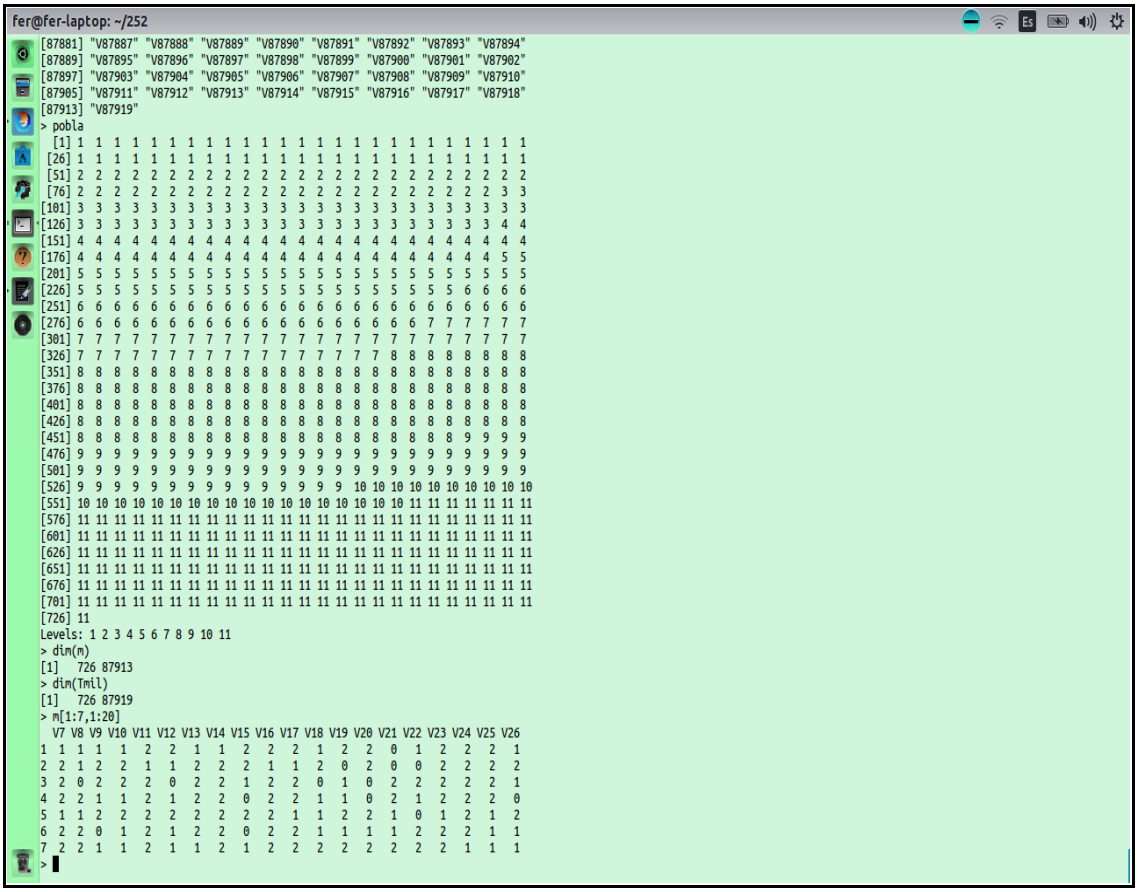


Ilustración 29: Muestra de manipulación de ficheros en R

En la ilustración se observa en la parte posterior las últimas columnas que aparecen en R cuando usamos el comando

```
names (m)
```

El comando *names (objeto)* permite conocer cual es el nombre de las columnas de una matriz o de un marco de datos. Como puede verse, en este caso no se ha puesto un nombre particular a las columnas, dado el elevado número de columnas. Pero se puede ver el nombre que proporciona R a las columnas de nombre desconocido que es

"Vnº de variable en la matriz". En segundo lugar en la ilustración se aprecia el vector población en el que a cada uno de los 726 individuos analizados se le asigna un número para la variable población según la raza a la que corresponda. Como puede verse, los individuos se encuentran ordenados según la raza de pertenencia. Por este motivo es necesario utilizar el comando *sort* antes de especificar las fases de prueba y entrenamiento, como se ha comentado.

Después aparecen en la ilustración las dimensiones de la matriz de *m* que contiene la información de los predictores y la dimensión del marco de datos *Tmil* que contiene además la columna población y 6 columnas que no son de interés para el estudio.

Adicionalmente, con el objetivo de facilitar la comprensión, se muestra en la ilustración un extracto de la variable *m*. El extracto corresponde a los primeros 20 SNPs de los primeros 7 animales. Como puede verse en prácticamente todos los SNPs que aparecen en la ilstración, la variabilidad es bastante considerable para estos 7 individuos.

Por último se ofrece un pantallazo de la la capacidad explicativa que se obtiene al llevar a cabo el procedimiento completo que en este Anexo se describe.

FUENTE: ELABORACIÓN PROPIA

```
> library(plsgenomics)
Loading required package: MASS
> nmarcadores = 12
> indices = unique ( c(sapply(1:11, function(i) variable.selection(m,1+(pobla==paste(i)),nvar=nmarcadores))))
> mi.pls.lda = pls.lda (m[indices], pobla, ncomp=11*nmarcadores-1, nruncv=20) # ana'lisis discriminante para asignar con
nruncv)
> mean(mi.pls.lda$predclass == pobla)
[1] 0.953168
```

Ilustración 30: Obtención del porcentaje de asignaciones correctas en R

Como puede apreciarse la media del porcentaje de aciertos para las razas analizadas es superior al 95%. Como puede verse por la nota que figura en la ilustración (separada en el comandopor #), los resultados dependerán del número de iteraciones de validación cruzada (nruncv), que se ha fijado en 20.