UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2012, Vol. 5, No. 1

Asociación de polimorfismos de nucleótido simple y de haplotipos para el gen de la Leptina con la ganancia de peso en la raza bovina blanco orejinegro usando técnicas bayesianas

Association of single nucleotide polymorphism and haplotypes of the Leptine gen with the weight gain in a Creole colombian bovine breed using Bayesian techniques

Hugo Andrés Gutiérrez Rojas[a]
hugogutierrez@usantotomas.edu.co

Ricardo Efrain Camacho Quiroga[b]
recamachoq@unal.edu.co

Resumen

Este artículo expone una metodología bayesiana para el análisis de asociación de polimorfismos de nucleótido simple (SNP) y de haplotipos con una característica de interés en un contexto de producción animal. En la primera etapa del análisis, se propone un modelo lineal bayesiano para clasificar los SNPs que tienen efecto sobre el promedio del valor genético de la variable respuesta. En una segunda etapa, después de la identificación de los haplotipos compatibles con los genotipos de influencia en la primera etapa, se discute la aplicación de un modelo lineal general y de un modelo de regresión logística en la identificación de los haplotipos que presentan una mayor asociación con el aumento del valor genético. En ambas etapas, se siguen metodologías bayesianas y cuando es pertinente se incluyen métodos de simulación de Monte Carlo para generar cadenas de Markov cuya distribución estacionaria corresponda a la distribución posterior condicional de los parámetros de interés. La aplicación práctica está supeditada al área de producción animal en una raza bovina criolla colombiana, denominada como raza blanco orejinegro (BON).

***Palabras clave*:** análisis bayesiano, haplotipos, MCMC, SNPS.

Abstract

This paper presents a Bayesian methodology for association study of single nucleotide polymorphism (SNP) and haplotypes with a special interest in animal

[a]Docente Investigador. Centro de Investigaciones y Estudios Estadísticos. Universidad Santo Tomás.
[b]Médico Veterinario. FEDEGAN.

production. In the first stage, we propose a Bayesian linear model to select the SNPs that have an effect on the average of the genetic value in the response variable. In a second stage, after the identification of haplotypes compatible with genotypes of influence in the first stage, we discuss the application of a general linear model and a logistic regression model in order to identify those haplotypes having a higher association with the increasing of genetic values. In both stages, Bayesian methodologies are used when appropriate and Monte Carlo simulation methods are implemented in order to generate Markov Chains whose stationary distribution corresponds to the conditional posterior distribution of the parameters of interest. The practical application is subject to animal production in a Colombian bovine breed.

***Key words*:** Bayesian Analysis, Haplotypes, MCMC, SNPs.

1. Introducción

En términos de producción animal en el contexto bovino, la calidad de la carne y de la canal se consideran como características fenotípicas en donde el medio ambiente ejerce un efecto mayor sobre la genética del individuo o la población a mejorar. Algunos genes explican la importancia de la variabilidad en características de composición y calidad cárnica, como lo es el caso del gen que codifica para la hormona leptina (LEP), que por su función biológica, es la responsable de la variabilidad en la deposición del tejido adiposo (Soria & Corva 2004). La leptina es una pieza clave en el complejo mecanismo de regulación del apetito y en el metabolismo energético de varias especies animales. También puede afectar la captación de nutrientes, la cantidad de grasa y su velocidad de acumulación en el animal y sus receptores son candidatos potenciales para el desarrollo de marcadores genéticos en un programa de mejoramiento, ya que esta disminuye la eficiencia de la utilización de energía.

El gen LEP bovino fue mapeado en el cromosoma 4 región q32 (Ji et al. 1998). Presenta 3 exones y 2 intrónes, con las regiones codificantes ubicadas en los exones 2 y 3 (Guerra & Navarro 2005) que corresponden alrededor de 18,9 kb del gen. El primer y segundo intrón tienen cerca de 14 y 1,7 kb respectivamente. El gen tiene 650kb, mientras que su ARNm posee 4.5kb. La organización exón-intrón de LEP se conserva entre humanos y bovinos. En diversos estudios con bovinos se han encontrado distintos polimorfismos en la secuencia codificante y promotora de LEP y polimorfismos en el gen receptor de la Leptina. Además, algunos de estos polimorfismos han sido asociados a variaciones en composición de la canal, calidad de la carne y crecimiento muscular (Soria & Corva 2004).

En este trabajo de investigación, se consideró la raza Blanco Orejinegro (BON), la cual es originaria de la zona comprendida entre Santafé de Antioquia al norte y Popayán, al sur a lo largo del río Cauca y en el departamento del Huila, Viejo Caldas y Antioquia y actualmente en el piedemonte llanero. Se ubica generalmente entre altitudes que van desde los 800 hasta los 1.800 msnm entre 18 y 24°C. Las *heredabilidades* para cada una de las características de peso han sido estimadas

de moderadas a altas, con valores al nacimiento de 0,36, al destete 0,48, a los 16 meses de 0,38 y con una ganancia de peso de 0,36 de heredabilidad (Martínez & Escobedo 2003).

Los polimorfismos de nucleotido simple (SNPs) son la forma más común de variación genética (Jurinke et al. 2006). El método más directo para detectar dichos marcadores es la secuenciación de segmentos de ADN, previamente amplificados por Reacción en Cadena de la Polimerasa (PCR), de varios individuos que representen la diversidad de la población. Se diseñan cebadores para amplificar fragmentos de ADN fundamentalmente de genes de interés en secuencias reportadas en bases de datos públicas. Los SNP pueden aparecer tanto en regiones fuera de los genes (que no afecten la producción o función de alguna proteína) como en un gen específico, donde pueden ubicarse en regiones codificantes (relacionados con cambios en la cantidad de proteínas producidas) o no codificantes (que afectan solo la secuencia de aminoácidos por una variación en el marco de lectura) (Taylor 1997).

Por otro lado, Pierce (2005) afirma que un conjunto específico de SNPs y variantes génicas observadas en un único cromosoma se denomina haplotipo y, como se encuentran ligados físicamente, tienden a heredarse en conjunto. Lo anterior explica que una herencia diferente del haplotipo está asociada con una mutación. Por tanto, después de conocer cuáles son los SNPs significativos en el modelo, resulta de gran interés estudiar cuáles de los haplotipos compatibles con el genotipo de los SNPs tienen una mayor asociación con la ganancia de peso del animal.

Después de una breve introducción, la sección 2 explica detalladamente el enfoque bayesiano que se utiliza en esta investigación. En esta sección se expone, en primer lugar, el modelo utilizado para la clasificación de los SNPs y, en segundo lugar, se discuten dos modelos que se pueden implementar en la práctica para realizar un análisis de haplotipos. La sección 3, correspondiente a la aplicación de esta metodología en el sector de producción animal, expone a profundidad los materiales y métodos utilizados en la recolección del material experimental, así como los resultados de las dos etapas del análisis bayesiano, que redundan en la escogencia de algunos polimorfismos que tienen efecto sobre el promedio del valor genético de la ganancia de peso en el individuo y en el análisis de los haplotipos compatibles con los anteriores genotipos. Por último, en la sección 4 se discuten los resultados a la luz de los procesos de producción animal. La implementación computacional de esta aplicación se realizó en la plataforma `WinBugs` (Lunn & Thomas 2000) y en el software de libre acceso `R` (R Development Core Team 2011). Los códigos computacionales utilizados se presentan en el apéndice.

2. Análisis bayesiano

En esta sección se abordarán brevemente las características más importantes del análisis bayesiano, el cual, además de especificar un modelo para los datos observados $\mathbf{Y} = (y_1, \ldots, y_n)$, dado un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, usualmente en forma de densidad condicional $p(\mathbf{Y} \mid \boldsymbol{\theta})$, supone que $\boldsymbol{\theta}$ es aleatorio

y que tiene un densidad *previa* $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$, donde $\boldsymbol{\eta}$ es un vector de hiper-parámetros. De esta forma, la inferencia concerniente a $\boldsymbol{\theta}$ se basa en una distribución *posterior* $p(\boldsymbol{\theta} \mid \mathbf{Y})$, bajo la cual es posible calcular una estimación puntual para el vector $\boldsymbol{\theta}$ dados los datos observados. Esta, dependiendo de la función de pérdida establecida en el estudio, está dada por alguna medida de tendencia central de la distribución $p(\boldsymbol{\theta} \mid \mathbf{Y})$. En particular, bajo la función de pérdida cuadrática, un estimador puntual del parámetro es la media de la distribución posterior. Es decir,

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \mathbf{Y}) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \mathbf{Y})\, d\boldsymbol{\theta} \tag{1}$$

Con fines de inferencia, también es posible calcular la región C de credibilidad del $100(1\text{-}\alpha)\,\%$, definida como

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C \mid \mathbf{Y}) = \int_C p(\boldsymbol{\theta} \mid \mathbf{Y})\, d\boldsymbol{\theta} \tag{2}$$

En este estudio se utilizarán estos dos criterios (estimaciones puntuales e intervalos de credibilidad) para decidir acerca de la inclusión de marcadores moleculares y/o patrones de herencia de alelos, definidos como factores en los modelos propuestos, así como la significación estadística y grado de asociación de los mismos con respecto a la ganancia de peso en la raza criolla.

2.1. Asociación de SNPs con los valores genéticos

En términos del modelamiento estadístico, la relación entre un vector de variables de interés $\mathbf{Y}$ y una matriz de variables auxiliares $\mathbf{X}$, es una de las herramientas estadísticas más utilizadas por los investigadores. Herramientas como la regresión simple, la regresión múltiple, el análisis de varianza y los modelos lineales generalizados forman parte del arsenal de opciones que la ciencia estadística ofrece a los usuarios que desean establecer relaciones de causalidad en el contexto propio de la investigación.

Como lo menciona Migon & Gamerman (1999), es muy útil adoptar la notación matricial para el desarrollo posterior del análisis bayesiano; entonces, se definen

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{X} = (1, \mathbf{x}_1, \ldots, \mathbf{x}_q) = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nq} \end{pmatrix}$$

y se supone que existe una relación de causalidad de parte de $\mathbf{X}$ reflejada en $\mathbf{Y}$ que puede ser descrita mediante el siguiente modelo probabilístico

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

en donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$ es el vector de parámetros de interés, de dimensión $q + 1$, y $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ es un vector aleatorio que sigue una distribución de

probabilidad normal multivariante con media nula y matriz de covarianzas $\boldsymbol{\Sigma}$. Antes de comenzar con la estipulación propia del análisis bayesiano, es necesario aclarar el papel que juegan las variables auxiliares en la inferencia estadística. En primer lugar, nótese que el interés particular recae en la distribución del vector de n variables aleatorias $\mathbf{Y} = (Y_1 \ldots, Y_n)'$, condicional a la matriz de variables auxiliares $\mathbf{X}$ e indexada por el vector de parámetros de interés $\boldsymbol{\beta}$ dada por $p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X})$.

Basado en lo anterior, y suponiendo que las variables de interés son intercambiables, entonces se asume que la verosimilitud para las variables de interés es

$$Y \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim Normal_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Al considerar que los parámetros son independientes *a previa* y que la distribución previa del vector de parámetros $\boldsymbol{\beta}$ es normal, la cual no depende de $\boldsymbol{\Sigma}$ y tiene su propia estructura de varianza, se tiene que

$$\boldsymbol{\beta} \sim Normal_{q+1}(\mathbf{b}, \mathbf{B})$$

Asimismo, la matriz de parámetros de dispersión $\boldsymbol{\Sigma}$ no depende de $\boldsymbol{\beta}$ y es posible asignarle la siguiente distribución previa

$$\boldsymbol{\Sigma} \sim Inversa - Wishart_v(\boldsymbol{\Lambda})$$

Nótese que la cantidad de parámetros individuales que se deben modelar crece a medida que el tamaño de muestra crece. Por otro lado, para encontrar las distribuciones posteriores que definan la estructura probabilística posterior, es necesario utilizar el condicionamiento posterior notando que

$$\begin{aligned} p(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) &= p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta}) p(\boldsymbol{\Sigma}) \end{aligned}$$

y para encontrar las distribuciones posteriores, se tiene que

$$p(\boldsymbol{\beta} \mid Y, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\beta}, \mathbf{Y}, \underbrace{\boldsymbol{\Sigma}}_{fijo})$$

y análogamente,

$$p(\boldsymbol{\Sigma} \mid Y, \boldsymbol{\beta}) \propto p(\boldsymbol{\Sigma}, \mathbf{Y}, \underbrace{\boldsymbol{\beta}}_{fijo})$$

Bajo este marco de referencia, es bien sabido que la distribución posterior del parámetro $\boldsymbol{\beta}$ condicionado a $\boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X}$ es

$$\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma} \sim Normal_{q+1}(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\mathbf{B}_q = \left(\mathbf{B}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$$
$$\mathbf{b}_q = \mathbf{B}_q\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}\right)$$

Por otro lado, la distribución posterior del parámetro $\boldsymbol{\Sigma}$ condicionado a $\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}$ es

$$\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \sim Inversa - Whishart_{v+q+1}(\mathbf{S}_{\boldsymbol{\beta}})$$

donde $\mathbf{S}_{\boldsymbol{\beta}} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(Y - X\boldsymbol{\beta})' + \boldsymbol{\Lambda}$. Luego, como se conocen las distribuciones condicionales posteriores para los parámetros de interés $\boldsymbol{\beta}$ y $\boldsymbol{\Sigma}$, es posible recurrir al algoritmo de Gibbs, en donde al fijar valores iniciales para los parámetros, se inicializa una cadena de Markov cuya distribución estacionaria es finalmente la distribución posterior conjunta para $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$.

El enfoque bayesiano ofrece al investigador herramientas poderosas en términos del modelamiento de relaciones causales, aun cuando existan pocas observaciones. En este estudio, se desea llevar a cabo un estudio de asociación entre las variantes génicas de la leptina con los valores genéticos calculados para la ganancia de peso (entre el nacimiento y el destete del animal). Un modelo apropiado para determinar esta asociación es el modelo lineal general que, además de lo anterior, permite identificar aquellas variantes polimórficas que no ejercen ningún efecto estadístico sobre el valor genético de la muestra.

Por tanto, se define la variable respuesta como el valor genético individual y las variables explicativas del modelo como los SNPs del gen leptina que intervienen en el modelo mediante la creación de niveles dependientes del genotipo. Por ejemplo, si se ha detectado que el SNP 271 del contiguo (fragmento secuenciado de ADN) presenta tres genotipos, CC, CT y TT, entonces dicho SNP entra en el modelo mediante la creación de dos variables dicotómicas, similares a las que se podrían crear para un factor con tres niveles en un análisis de varianza. El anterior procedimiento se realiza con el fin de dar una mayor claridad del efecto del SNP sobre la variable respuesta en el modelo.

En términos de la aplicación práctica para este estudio, es plausible suponer que $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. De esta forma, se llega a establecer el siguiente modelo poblacional

$$E(Y_i \mid \boldsymbol{\beta}, \mathbf{X}) = \boldsymbol{\beta}\mathbf{x}_i' = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$$
$$Var(Y_i \mid \boldsymbol{\beta}, \mathbf{X}) = \sigma^2 \qquad i = 1, \ldots, n$$

Este modelo asume que los genotipos para cada locus polimórfico son fijos. Lo anterior es deseable porque este análisis no está basado en el modelamiento de la varianza al interior de cada SNP, sino en el efecto fijo que este posee sobre el valor genético individual. Por otro lado, este modelo no tiene en cuenta las posibles interacciones entre los polimorfismos puesto que se buscan patrones de herencia basados en el aporte de cada progenitor. Además, en esta propuesta se busca establecer cuáles son las variantes alélicas asociadas con un rasgo de importancia

zootécnica que se transmitan a una nueva generación y que ejercen un efecto positivo sobre el valor genético de la progenie, por lo cual no es necesario llevar a cabo un análisis que permita establecer los efectos epistáticos de los polimorfismos.

Por tanto, el análisis de asociación está supeditado a la significación estadística de la estimación del vector de coeficientes de regresión. Como regla, se tiene que si todos los niveles de un SNP no son significativos estadísticamente (en términos bayesianos, que el cero esté dentro del intervalo de credibilidad), entonces ese polimorfismo no se contemplará en análisis posteriores. El resultado de este modelo será un grupo de SNPs que servirán de insumo para realizar un posterior análisis de haplotipos.

2.2. Análisis de haplotipos

Luego de obtener los SNPs significativos mediante el modelo lineal anterior, es de interés conocer cuáles son los haplotipos que se asocian con un aumento en los valores genéticos de la variable respuesta. De esta forma, los posibles haplotipos resultantes de los SNPs pueden ser vistos como variables explicativas del fenómeno en estudio. De esta forma, y recordando que un genotipo tiene a lo más 2^r posibles haplotipos compatibles, se define h_{im} como el m-ésimo haplotipo que puede ser compatible o no con el i-ésimo individuo y se denomina M como el total de haplotipos compatibles. De esta forma, se constituyen unas nuevas variables explicativas cuya naturaleza es dicotómica. Sin embargo, a pesar de lo anterior, los haplotipos no constituyen una partición de la población y como consecuencia, estas nuevas variables dicotómicas no pueden ser tratadas como factores en un análisis de varianza.

En esta investigación consideramos dos maneras de estudiar esta asociación. La primera mediante un modelo lineal general, en donde el vector de la variable respuesta y la matriz de diseño están dadas, respectivamente, por:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{H} = \begin{pmatrix} 1 & h_{11} & \dots & h_{1M} \\ \vdots & \ddots & \vdots & \\ 1 & h_{n1} & \dots & h_{nM} \end{pmatrix}$$

Con base en lo anterior, al asumir que existe una relación causal de $\mathbf{H}$ en $\mathbf{Y}$, se supone el siguiente modelo de regresión

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{4}$$

en donde $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_M)'$ es el vector de parámetros de interés y $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ es el vector de errores que sigue una distribución de probabilidad normal con media nula y matriz de covarianzas $\boldsymbol{\Gamma}$. Por lo tanto,

$$\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\Gamma} \sim Normal_n(\mathbf{H}\boldsymbol{\alpha}, \boldsymbol{\Gamma})$$

Al considerar que los parámetros son independientes *a priori* y que la distribución previa del vector de parámetros $\boldsymbol{\alpha}$ es normal, la cual no depende de $\boldsymbol{\Gamma}$ y tiene su propia estructura de varianza, se tiene que

$$\boldsymbol{\alpha} \sim Normal_{M+1}(\mathbf{a}, \mathbf{A})$$

Igualmente, se asume *a priori* que la matriz de parámetros de dispersión $\boldsymbol{\Gamma}$ no depende de $\boldsymbol{\alpha}$ y es posible asignarle la siguiente distribución previa

$$\boldsymbol{\Gamma} \sim Inversa - Wishart_u(\boldsymbol{\Delta})$$

El análisis bayesiano en este modelo es similar al análisis llevado a cabo en la primera etapa. Después de la asignación de las distribuciones previas para los coeficientes de regresión y para la varianza del modelo, se encuentra que la distribución posterior condicional de los coeficientes de regresión es normal multivariante y la distribución de la matriz de varianzas es inversa-Whishart. Luego, una vez más, el análisis de asociación estará motivado por la significación estadística de la estimación del vector de coeficientes de regresión. De esta manera, si el cero está dentro del intervalo de credibilidad, entonces se afirma que el haplotipo no es significativo y no es posible concluir acerca de su asociación con el valor genético de la variable respuesta. Sin embargo, si los límites del intervalo de credibilidad son positivos, entonces se concluye a favor de la asociación positiva del haplotipo con la variable respuesta.

En segundo lugar, si la variable respuesta es dicotómica, en el sentido de clasificación de los individuos en dos clases de valores genéticos, altos o bajos, es posible plantear un modelo de regresión logística que contemple la asociación de los haplotipos con los valores genéticos altos o bajos. Por consiguiente, si D_i es la variable binaria que clasifica a los individuos (la cual toma el valor uno si el individuo clasifica como de alto valor genético y cero, en otro caso) entonces el evento $D_i = 1$ ocurre con una probabilidad de éxito p_i. Por lo tanto, la probabilidad condicional a $\mathbf{H}_i = (1, h_{i1}, \ldots, h_{iM})$ para D puede expresarse como

$$p_i = Pr(D_i = 1|\mathbf{H}_i) = \frac{\exp\{\gamma_0 + \gamma_1 h_{i1} + \cdots, +\gamma_M h_{iM}\}}{1 + \exp\{\gamma_0 + \gamma_1 h_{i1} + \cdots, +\gamma_M h_{iM}\}}$$

Luego, al realizar la productoria sobre todos los individuos en la muestra, se concluye que la verosimilitud de esta regresión logística está dada por

$$\prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Manteniendo la distribución previa para el vector de coeficientes de regresión, definida como normal multivariante, no es posible realizar un análisis conjugado

que brinde estimaciones exactas, puesto que no se podrá encontrar una forma cerrada para la distribución posterior de los parámetros. Consecuentemente, se hace necesaria la incorporación de métodos de simulación de Monte Carlo, basados en cadenas de Markov, que permitan en cada iteración (o estado de la cadena) la selección de valores provenientes de las distribuciones condicionales posteriores.

Nótese que, definiendo el vector de coeficientes de regresión como $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_M)$, la verosimilitud está dada por

$$p(\mathbf{Y} \mid \boldsymbol{\gamma}, \mathbf{H}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{H}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{H}_i'\boldsymbol{\gamma})} \right)^{y_i} \left(1 - \left(\frac{\exp(\mathbf{H}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{H}_i'\boldsymbol{\gamma})} \right) \right)^{1-y_i} \tag{5}$$

Por otro lado, asumiendo que la distribución previa para $\boldsymbol{\gamma}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\gamma} \sim Normal_{M+1}(\mathbf{g}, \mathbf{G})$$

Entonces, la distribución posterior toma la siguiente forma

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{H}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{H}_i'\boldsymbol{\gamma})} \right)^{y_i} \left(1 - \left(\frac{\exp(\mathbf{H}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{H}_i'\boldsymbol{\gamma})} \right) \right)^{1-y_i} \times \exp\left\{ \frac{-1}{2} (\boldsymbol{\gamma} - \mathbf{g})' \mathbf{G}^{-1} (\boldsymbol{\gamma} - \mathbf{g}) \right\} \tag{6}$$

La anterior expresión no tiene una forma cerrada y no es sencillo simular observaciones y obtener inferencias posteriores. Sin embargo, con ayuda de la técnica del condicionamiento sucesivo y el algoritmo de Gibbs (Geman & Geman 1984, Gelfand & Smith 1990) es posible obtener observaciones provenientes de la distribución condicional posterior del parámetro $\boldsymbol{\gamma}_m$, para $m = 0, 1, \ldots, M$. De esta manera, el algoritmo de Gibbs generará muestras para el nuevo estado k-ésimo de la cadena. Es decir, para el m-ésimo coeficiente de la regresión, en la iteración k-ésima de la cadena, el procedimiento generará el valor $\boldsymbol{\gamma}_m^{(k)}$, desde la siguiente distribución posterior condicional

$$\boldsymbol{\gamma}_m^{(k)} \sim p(\boldsymbol{\gamma}_m | \boldsymbol{\gamma}_0^{(k)}, \boldsymbol{\gamma}_1^{(k)}, \ldots, \boldsymbol{\gamma}_{m-1}^{(k)}, \boldsymbol{\gamma}_{m+1}^{(k)}, \boldsymbol{\gamma}_M^{(k)})$$

Para simular valores de la anterior distribución condicional posterior, es posible utilizar el algoritmo DFARS (*Derivative Free Adaptive Rejection Sampling*, por sus siglas en inglés) (Gilks 1992), que constituye un método de muestreo por rechazo para distribuciones log-cóncavas cuando se utiliza el algoritmo de Gibbs. Sin embargo, existen otros métodos que pueden ser utilizados para obtener valores provenientes de la distribución posterior conjunto, por ejemplo, al definir nuevas variables de trabajo es posible aplicar una adaptación del algoritmo IRSL (*Iterative Reweighted Least Squares*, por sus siglas en inglés) (Gamerman & Lopes 2006,

p. 86), utilizado por West (1985) en el contexto de los modelos lineales generalizados. Por otro lado, también es posible utilizar métodos que aproximen la log-verosimilitud a una distribución normal, como los expuestos en Gelman et al. (2003, p. 422).

La convergencia de las cadenas resultantes, cuya distribución estacionaria está dada por la distribución posterior conjunta (6), puede ser constatada de distintas maneras. Dado que una gran parte del desarrollo de este proyecto está ligada a la programación e implementación de métodos de Monte Carlo para realizar inferencias posteriores de los parámetros de interés, se seguirá el razonamiento y recomendaciones de Gelman & Shirley (2010), que pueden ser resumidos en los siguientes ítems para cada parámetro de interés:

- Simulación de tres o más cadenas de forma paralela. Los valores iniciales de cada cadena deben estar dispersos entre sí.
- Descarte de la primera mitad de los valores generados en las cadenas. Esta etapa se conoce como *burning stage*.
- Una vez que las cadenas converjan, mezclar los tres conjuntos de valores generados por las cadenas. Esto garantiza, en primera instancia, que las cadenas no estén auto-correlacionadas.
- Además de realizar esta mezcla, descartar valores intermedios de las cadenas resultantes. Esta etapa se conoce como *thining stage*. Al final se recomienda almacenar una mediana cantidad de valores simulados.
- Comparación y contraste de los resultados con modelos simples que permitan examinar posibles discrepancias y corregir errores de programación.

Después de obtener las estimaciones de los coeficientes de regresión, es posible inferir acerca de la razón de Odds, para medir la magnitud de la asociación del haplotipo con la variable respuesta. Esta razón de Odds para el m-ésimo haplotipo se define como $\exp\{\gamma_m\}$. Si esta cifra es mayor que uno, entonces se concluye que hay fuertes indicios de que el haplotipo esté asociado con la clasificación de alto valor genético y si es menor a uno, hay indicios de que el haplotipo está asociado con la clasificación de bajo valor genético. Nótese que, siendo $\gamma_m^{(k)}$, el k-ésimo valor generado de la distribución posterior condicional del parámetro γ_m, y dado que las cadenas contienen, para cada iteración, todos estos valores generados, entonces es muy sencillo construir un intervalo de credibilidad para la razón de Odds.

3. Aplicación en la raza BON

El material experimental del presente trabajo consta de un biotipo, al cual se le colectó la información de pesos al nacimiento, al destete y a los 16 meses de edad. Se tuvo en cuenta un grupo de 171 animales con edades entre 30 y 36 meses,

que se encuentran localizados en el Centro de Investigación San José del Nus en el departamento de Antioquia, en cercanías al municipio de Puerto Berrío en la región del Magdalena Medio a una altura promedio de 125 metros sobre el nivel del mar y una temperatura aproximada de 29 °C, y que nacieron en la ventana de observación que empezó en enero de 2008 y finalizó en julio de 2010.

El material genético se obtuvo por medio de extracción de ADN a partir de muestras de sangre de las unidades experimentales incluidas en la prueba. Una vez culminado el programa por el termociclador, se llevaron las muestras a una reacción de secuencia, la cual permite identificar los alelos polimórficos y variantes génicas. La secuencia de dichas regiones en poblaciones contrastantes, permitió mostrar aquellas variaciones alélicas implicadas en la presentación de la característica de mayor o menor crecimiento muscular. Posteriormente se realizó un alineamiento múltiple para comparar las secuencias obtenidas de la población seleccionada y se confrontaron con las bases de datos públicas para Bos taurus y Bos índicus, identificando posibles diferencias con las secuencias reportadas. Adicionalmente se identificaron los polimorfismos de un solo nucleótido, inserciones y deleciones por medio de los programas `PolyPhred 6.18` (Nickerson et al. 1997) y `PolyScan` (Chen et al. 2007).

Basado en las medidas de peso observado, se ajustó un modelo mixto animal (Henderson 1986) y se predijeron los valores genéticos[1] para cada uno de los 171 animales en consideración. Luego, se realizó un análisis descriptivo de estas predicciones (valores genéticos individuales) para identificar[2] los animales de mayor valor genético y menor valor genético, quienes serían finalmente los individuos considerados en el estudio. El grupo de animales con más alto valor genético estuvo conformado por 25 individuos, mientras que el grupo de animales con menor valor genético estuvo conformado por 30 individuos.

El propósito de esta aplicación es establecer, en una primera etapa, cuáles genotipos tienen un mayor efecto en el promedio del valor genético de la ganancia de peso. Luego, en una segunda etapa, utilizando solamente los genotipos significativos, se procederá a realizar un análisis de haplotipos a dos vías: la primera, mediante un modelo lineal general y la segunda mediante una modelo de regresión logística. Este análisis permitirá conocer qué genotipos y cuáles haplotipos son aquellos que están asociados con la ganancia de peso del individuo.

3.1. Primera etapa: identificación de SNPs

Después de realizar la secuenciación individual del gen leptina para la muestra, se encontraron los siguientes polimorfismos: C271T, con dos genotipos CC y CT; C428T, con dos genotipos CC y CT; T431C, con dos genotipos TC y TT; T443C, con dos genotipos TC y TT; y por último, C527T, con dos genotipos CC y CT. To-

[1]El modelo mixto se ajustó para un grupo de 3835 individuos, algunos de ellos ancestros de los animales incluidos en la muestra de tamaño 171.

[2]Esta discriminación se realizó utilizando un umbral bilateral de dos desviaciones estándar con respecto a la media.

dos estos SNPs se encontraron en regiones codificantes del gen Lep en el exón 3. La transición C271T produjo un cambio en el codón que codifica para el aminoácido Ala siendo reemplazado por Val, pero los demás polimorfismos representaban mutaciones sinónimas y sin efecto alguno sobre el péptido traducido. Las frecuencias genotípicas y alélicas para cada una de las anteriores variantes que se presentaron en la población muestreada se contemplan en la tabla 1.

Tabla 1: *Frecuencias genotípicas (columnnas 3, 4 y 5) y frecuencias alélicas (columnnas 6 y 7) para las variantes encontradas en la muestra de los individuos de la raza criolla*

Contiguo	SNP	CC	TT	CT	C	T
271	C¿T	0,49	0,00	0,51	0,75	0,25
428	C¿T	0,33	0,00	0,67	0,66	0,34
431	T¿C	0,00	0,45	0,55	0,27	0,73
443	T¿C	0,00	0,20	0,80	0,40	0,60
527	C¿T	0,31	0,00	0,69	0,65	0,35

Por tanto, se planteó el siguiente modelo que permite explicar la relación de los SNPs con el promedio del valor genético de la ganancia de peso en la raza criolla:

$$\begin{aligned} Y_i &= \beta_0 && \text{(término constante)} \\ &+ \beta_1 x_{1i} && \text{(para C271T)} \\ &+ \beta_2 x_{2i} && \text{(para C428T)} \\ &+ \beta_3 x_{3i} && \text{(para T431C)} \\ &+ \beta_4 x_{4i} && \text{(para T443C)} \\ &+ \beta_5 x_{5i} && \text{(para C527T)} \\ &+ \varepsilon_i \end{aligned}$$

En donde $E_(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$. Además, para $j = 1, \ldots, 5$, se tiene que

$$x_{ji} = \begin{cases} 1, & \text{si el individuo } i \text{ presenta el SNP } j; \\ 0, & \text{en otro caso.} \end{cases}$$

Para la implementación del análisis bayesiano, se consideraron distribuciones previas no informativas y planas para los coeficientes de regresión, β_j, mediante la asignación de distribuciones normales centradas en cero y con una gran varianza. De la misma forma, para la varianza del modelo, σ^2, se consideró una distribución previa no informativa de tipo inversa gama con parámetros de forma pequeño de escala grande. Los resultados se muestran en la tabla 2.

Los resultados muestran que el primer nivel de los polimorfismos C271T, C428T, y T443C, no son significativos en el modelo de ganancia de peso. Por tanto, estos polimorfismos no se tendrán en cuenta en el análisis posterior de variantes alélicas.

Tabla 2: *Estimación bayesiana de los parámetros del modelo de asociación de genotipos. DE es la desviación estándar, LI es el percentil 2.5 y LS es el percentil 97.5 de la distribución posterior de los parámetros*

Parámetro	Estimación	DE	LI	LS
β_0	37.71	18.52	14.76	61.11
β_1	-5.435	5.982	-13.05	2.123
β_2	-11.49	9.542	-23.63	0.773
β_3	-19.41	8.43	-30.13	-8.934
β_4	12.34	12.28	-3.18	27.91
β_5	-21.42	9.717	-33.79	-9.098
σ^2	288.0	61.01	191.7	431.1

Por otro lado, el primer nivel, TC, del polimorfismo T431C es significativo y tiene un efecto negativo sobre el promedio del valor genético para la ganancia de peso. También es claro que el primer nivel, CC, del polimorfismo C527T tiene un efecto significativo y negativo sobre el promedio del valor genético para la característica de interés. Luego, en términos prácticos se recomienda realizar una selección de individuos portadores de la variante genotípica TT, correspondiente al segundo nivel del polimorfismo T431C, y de la variante genotípica CT, correspondiente al segundo[3] nivel del polimorfismo C527T.

3.2. Segunda etapa: análisis de variantes alélicas

Teniendo en cuenta el efecto que ejerce cada uno de los anteriores marcadores moleculares sobre la expresión diferencial de la característica de interés, estos serán incluidos en un análisis de haplotipos compatibles con los genotipos individuales. Notando como SNP3 al polimorfismo T431C y SNP5 al polimorfismo C527T, es posible que se presenten los siguientes genotipos basados en los cuatro posibles haplotipos[4]: T/T, C/C, T/C y C/T.

De esta manera, si un individuo posee TT en T431C y CC en C527T, entonces es compatible con el haplotipo T/C, únicamente. Sin embargo, si el individuo posee TT en T431C y CT en C527T, entonces será compatible con los haplotipos T/C y T/T, únicamente. Por otro lado, si el individuo posee TC en T431C y CT en C527T, entonces será compatible con los haplotipos T/C y C/T, únicamente. En la tabla 3 se presentan las frecuencias muestrales para los cuatro posibles haplotipos en T431C y C527T, que se calculan como el cociente entre la suma de los alelos presentes en los SNP3 y SNP5, provenientes de cada uno de los gametos, y la cantidad total de combinaciones posibles en la muestra.

[3]Estas recomendaciones se deben a que los primeros niveles de estos polimorfismos tuvieron un efecto significativo y negativo sobre el valor genético de la característica de interés.

[4]Nótese que si hay r heterocigotos en los loci, entonces el genotipo tendrá a lo más 2^r haplotipos compatibles con él.

Tabla 3: *Frecuencias de los haplotipos en la muestra de los individuos de la raza criolla*

Haplotipo	SNP3	SNP5	Gameto 1	Gameto 2	Suma	Frecuencia
h_1	T	T	0	8	8	0.072
h_2	C	C	0	0	0	0.000
h_3	T	C	55	17	72	0.654
h_4	C	T	0	30	30	0.272

De lo anterior, y considerando $\mathbf{H}_i$ como el vector de variables explicativas del valor genético de la ganancia de peso, se propone, en primera instancia, el siguiente modelo para explicar la asociación de la característica de interés con cada haplotipo.

$$\begin{aligned} Y_i = \alpha_0 & \quad \text{(término constante)} \\ + \alpha_1 h_{1i} & \quad \text{(para T/T)} \\ + \alpha_2 h_{2i} & \quad \text{(para C/C)} \\ + \alpha_3 h_{3i} & \quad \text{(para T/C)} \\ + \alpha_4 h_{4i} & \quad \text{(para C/T)} \\ + \epsilon_i & \end{aligned}$$

En donde $E(\epsilon_i) = 0$ y $Var(\epsilon_i) = \tau^2$. Además, para $m = 1, \ldots, 4$, se tiene que

$$h_{mi} = \begin{cases} 1, & \text{si el individuo } i \text{ es compatible con el haplotipo } m; \\ 0, & \text{en otro caso.} \end{cases}$$

Debido a que un mismo individuo puede ser compatible con más de un haplotipo al mismo tiempo, entonces algunos de estos haplotipos pueden encontrarse segregados en la población de la misma manera y como consecuencia, dado que $h_{mi} = 1$, para todo $i = 1, \ldots, n$, y para algunos m, entonces el haplotipo no tiene efecto en el modelo y debe ser eliminado. Este caso sucede con el haplotipo T/C, el cual es compatible con todos los individuos de la muestra. Por otro lado, dado que algunos haplotipos están ausentes, entonces no se pueden tener en cuenta en el modelo. Lo anterior sucede con el haplotipo C/C. Por consiguiente, nuestro modelo reducido es

$$Y_i = \alpha_0 + \alpha_1 h_{1i} + \alpha_4 h_{4i} + \epsilon_i$$

Al igual que en la etapa de la identificación de SNPs, se consideraron distribuciones previas no informativas y planas para los coeficientes de regresión, α_m. Para la varianza de los errores, τ^2, se consideró una distribución previa inversa gamma no informativa, con parámetro de escala grande y parámetro de forma pequeño. Los resultados de las estimaciones se muestran en la tabla 4, a partir de la cual es

posible concluir que los haplotipos T/T y C/T son estadísticamente significativos y tienen un efecto positivo para el promedio del valor genético de la ganancia de peso.

Nótese que el primer haplotipo, T/T, tiene un efecto positivo para el promedio del valor genético de la ganancia de peso y que esto coincide efectivamente con el análisis de genotipos de la primera etapa, puesto que se había identificado que los genotipos TC y CC, de los SNPs significativos, se encuentran en asociación con un bajo valor genético de la característica de interés, mientras que el haplotipo en cuestión es T/T que no coincide con los haplotipos generados por estos niveles, los cuales son T/C y C/C. De la misma manera, el cuarto haplotipo, C/T, tiene un efecto positivo para el promedio del valor genético de la ganancia de peso y esto concuerda con el análisis de genotipos de la primera etapa, puesto que C/T no coincide T/C y C/C.

Tabla 4: *Estimación bayesiana de los parámetros del modelo de asociación de haplotipos. DE es la desviación estándar, LI es el percentil 2.5 y LS es el percentil 97.5 de la distribución posterior de los parámetros*

Haplotipo	Estimación	DE	LI	LS
Intercepto	19.85	2.19	15.69	24.50
T/T	22.09	6.26	12.15	36.70
C/T	8.425	4.17	0.57	16.49
Varianza	6.012	0.15	5.74	6.31

Por último, y dado que en la muestra se seleccionaron individuos con valores genéticos extremos, se pueden diferenciar claramente dos grupos de individuos. el primero, correspondiente a aquellos animales con *alto valor genético* para la característica de interés, considerando un índice genético mínimo de 30. Por otra parte, los individuos que presentaron valores genéticos de menos de 18, en el índice genético, son clasificados en otro grupo, correspondiente al de *bajo valor genético*.

Por tanto, se define D_i como la variable dicotómica para el individuo i ($i = 1, \ldots, n$), la cual es igual a uno, si el individuo presenta alto valor genético e igual a cero, si el individuo presenta bajo valor genético. De esta forma, suponiendo que un individuo es clasificado en el grupo de alto valor genético con probabilidad p_i, entonces es plausible considerar el siguiente modelo de regresión logística, al considerar a los haplotipos como variables dependientes.

$$p_i = Pr(D_i = 1|\mathbf{H}_i) = \frac{\exp\{\gamma_0 + \gamma_1 h_{1i} + \gamma_4 h_{4i}\}}{1 + \exp\{\gamma_0 + \gamma_1 h_{1i} + \gamma_4 h_{4i}\}}$$

En donde $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_4)'$ es el vector de parámetros del modelo reducido. Al realizar la productoria sobre todos los individuos en la muestra, se concluye que la verosimilitud de esta regresión logística está dada por la expresión (5).

En general, al asignar distribuciones previas planas y no informativas al vector de parámetros, y siguiendo la regla de bayes, se encuentra que la distribución posterior

no tiene una forma cerrada y por lo tanto, es difícil realizar inferencias conjugadas. Luego, usando los algoritmos de Monte Carlo, discutidos en la sección 4, es posible construir cadenas de Markov que generan, en cada nuevo estado, valores de los coeficientes de regresión provenientes de la distribución condicional del parámetro. Para esta aplicación, se consideraron distribuciones previas no informativas normales para los coeficientes de regresión, γ_m. Luego de realizar 1500 iteraciones, después de la convergencia de las cadenas, se tienen los resultados expuestos en la tabla 5 para la inferencia de los parámetros de la regresión logística.

Tabla 5: *Estimación bayesiana de los parámetros del modelo de regresión logística de haplotipos. DE es la desviación estándar, LI es el percentil 2.5 y LS es el percentil 97.5 de la distribución posterior de los parámetros*

Haplotipo	Estimación	DE	LI	LS
Intercepto	-0.149	0.227	-0.596	0.310
T/T	50.666	35.702	3.643	117.498
C/T	0.700	0.449	-0.183	1.630

De esta forma, derivado del análisis bayesiano, es posible construir la razón de Odds, para los haplotipos. Para realizar lo anterior, y teniendo en cuenta los valores simulados de cada coeficiente de regresión, se define para la k-ésima iteración, una nueva variable denominada $Odd_{h_j} = \exp\{\gamma_j^{(k)}\}$, en donde $\gamma_j^{(k)}$ es una realización de la distribución condicional del parámetro γ_j. Como lo afirma Iniesta & Moreno (2008), con estos valores, no solo se logra obtener estimaciones puntuales de la razón de Odds para cada parámetro, sino que también es posible construir intervalos de credibilidad al 95 %. La tabla 6 muestra dichas estimaciones.

Tabla 6: *Estimación bayesiana de las razones de Odds para los parámetros del modelo de regresión logística de haplotipos. DE, es la desviación estándar, LI, es el percentil 2.5 y LS es el percentil 97.5 de la distribución posterior de los parámetros*

Haplotipo	Estimación	LI	LS
T/T	1.009 E+22	38.23	1.069 E+51
C/T	2.014	0.83	5.10

Basado en lo anteriormente expuesto, se puede observar que existe una marcada asociación entre el haplotipo T/T con alto valor genético para ganancia de peso debido a que el valor de la estimación de la razón de Odds es, además de mayor a uno, muy grande. Por otro lado, a pesar de que la estimación de la razón de Odds para el haplotipo C/T no es tan alta como la anterior, el hecho de que sea mayor a uno muestra también una alta asociación con el valor genético deseado.

4. Discusión y conclusiones

A través de las técnicas aplicadas durante este trabajo, se pudieron establecer procedimientos que abren múltiples expectativas al mundo de la genética molecular animal con aplicación de la bioinformática y bioestadística, herramientas clave para el diagnóstico de asociación de rasgos de tipo genético con variables de importancia productiva, reproductiva, sanitaria y de calidad de los productos de origen animal.

Las observaciones indican que realmente existe la posibilidad de alterar la expresión de una característica afectada por un gen, lo que podría generar grandes variaciones por diferentes polimorfismos y cambios en el patrón de replicación, transcripción y por ende traducción de la proteína, que en algunos casos llegaría a ser económicamente importante por la introgresión de un marcador molecular al ser transmitido a su siguientes generaciones haciendo que exista la segregación de un rasgo deseado. Este panorama permite encontrar respuestas a incógnitas tales como: de qué manera los haplotipos y posibles patrones de herencia y de combinaciones alélicas en diferentes *loci* se logran transmitir a la progenie.

El uso de la secuenciación de genes como diagnóstico y su análisis, por medio de herramientas estadísticas eficientes, toma importancia en el estudio de características raciales donde los costos y el tiempo de análisis serán menores que en evaluaciones con datos fenotípicos por medio de modelos cuantitativos mixtos; sin embargo, es fundamental conocer y aplicar dichas herramientas tradicionales para establecer predicciones de algunos valores genéticos y calcular coeficientes de consanguinidad e índices para realizar programas de selección.

Por otra parte, durante la primera fase de este trabajo se logró identificar los marcadores moleculares tipo SNP que tenían un efecto directo sobre una característica expresada en el valor genético de la misma. Un vez fueron seleccionadas, se obtuvo información de las mejores variantes para cada uno de los genotipos polimórficos, donde para el SNP T431C el genotipo no deseado TC se encuentra en una frecuencia alta dentro de la población correspondiente al 55 %; sin embargo, para el marcador C527T, se halló que la variante CT era la de mejor elección ya que CC se encontraba en estrecha relación con valores genéticos bajos para la ganancia de peso, esto permite concluir que la población estudiada posee un mayor número de individuos portadores de este genotipo deseable en una frecuencia del 69 % compensando la deficiencia del anteriormente explicado; pero al observar las frecuencias alélicas se puede suponer que existe una mayor probabilidad de presentación del genotipo homocigoto para C debido a su distribución dentro de la población en un 65 % y en caso contrario, para el SNP T431C que exhibe una mayor probabilidad de obtener el genotipo homocigoto a T que es el deseable, por estar presente dicho nucleótido en el 73 % de todos los alelos del material experimental.

Ahora bien, tomando la información del segundo análisis se puede realizar un sistema de selección de individuos mejoradores de la raza, sacando provecho de los haplotipos asociados con efectos positivos sobre la variable valor genético de ganancia de peso, donde el haplotipo T/T posee una mayor prevalencia por en-

contrarse estos alelos en una mayor frecuencia que el haplotipo C/T. Este último, a pesar de su efecto positivo sobre la variable de interés, ejerce un menor efecto, el cual ha sido estimado en la razón de Odds, y la frecuencia de estos alelos es más baja, haciendo que la probabilidad de éxito disminuya.

Es deseable que esta gama de conocimientos bioestadísticos bayesianos esté al alcance del sector agropecuario para realizar dichos macroanálisis de forma acertada, con el fin de generar gran cantidad de información útil dentro de un programa de selección de individuos superiores que podrán ser usados como material genético mejorador de una raza o en programas de cruzamiento, buscando la expresión de diferentes caracteres de importancia zootécnica.

Recibido: 10 de enero de 2012
Aceptado: 8 de mayo de 2012

Referencias

Chen, K., McLellan, M. D., Michael, L. D. & Ding, L. (2007), 'Polyscan: An automatic indel and snp detection approach to the analysis of human resequencing datag', *Genome Research* **17**, 659 – 666.

Gamerman, D. & Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman and Hall/CRC.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Society* **85**, 398 – 409.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003), *Bayesian Data Analysis*, 2 edn, Chapman and Hall/CRC.

Gelman, A. & Shirley, K. (2010), *Handbook of Markov Chain Monte Carlo*, CRC, chapter Inference from Simulations and Monitoring Convergence.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721 – 741.

Gilks, W. R. (1992), *Derivative-free Rejection Sampling for Gibbs Sampling*.

Guerra, C. & Navarro, P. (2005), 'Brown adipose tissue specific insulin receptor knockout shows diabetic phenotype without insulin resistance', *Journal of Clinical Investigation* **108**, 1205 – 1213.

Henderson, C. R. (1986), 'Estimation of variances in animal model and reduced animal model for single traits and single records', *Journal of Dairy Science* **69**(5), 1394–1402.

Iniesta, R. & Moreno, V. (2008), *Monte Carlo and Quasi-Monte Carlo Methods*, Springer Berlin Heidelberg, chapter Assessment of Genetic Association using Haplotypes Inferred with Uncertainty via Markov Chain Monte Carlo, pp. 529 – 535.

Iniesta, R. & Moreno, V. (2010), *BayHap: Bayesian analysis of haplotype association using Markov Chain Monte Carlo*.

Ji, S., Willis, G. M., Scott, R. R. & Spurlock, M. E. (1998), 'Partial cloning and expression of the bovine leptin gene', *Animal Biotechnology* **9**, 1–4.

Jurinke, C., Denissenko, M., Oeth, P., Ehrich, M., Dirk, v. B. & Cantor, C. (2006), 'A single nucleotide polymorphism based approach for the identification and characterization of gene expression modulation using massarray', *Mutation Research* **573**, 83–95.

Lunn, D. & Thomas, A. (2000), 'Winbugs a bayesian modelling framework: concepts, structure, and extensibility', *Statistics and Computing* **10**, 325 – 337.

Martínez, S. & Escobedo, M. (2003), 'Situación de los recursos zoogenéticos en colombia', *Ministerio de Agricultura y Desarrollo Rural* .

Migon, H. S. & Gamerman, D. (1999), *Statistical Inference: An Integrated Approach*, Arnold.

Nickerson, D. A., Tobe, V. A. & Taylor, S. L. (1997), 'Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence based resequencing', *Nucleic Acids Research* p. 27452751.

Pierce, B. A. (2005), *Genetics - A Conceptual Approach*, W. H. Freeman.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Soria, L. & Corva, P. (2004), 'Factores genéticos y ambientales que determinan la terneza de la carne bovina', *Archivos Latinoamericanos de Producción Animal* **12**, 73–88.

Taylor, G. (1997), *Laboratory Methods for the Detection of Mutations and Polymorphisms in DNA*, CRC Press.

West, M. (1985), *Bayesian Statistics 2*, Oxford University Press, chapter Generalized Linear Models: Outlier Acommodation, Scale Parameters and Prior distributions (with discussion), pp. 461 – 484.

A. Códigos computacionales

En la implementación de la primera etapa, se utilizó la siguiente sintaxis del sistema computacional `WinBugs` (Lunn & Thomas 2000).

```
model
{
#Creación de variables dicotómicas
for(i in 1:34)
{
X[i,1] <- 1.0
X[i,2] <- equals(g1[i],1)
X[i,3] <- equals(g1[i],2)
X[i,4] <- equals(g2[i],1)
X[i,5] <- equals(g3[i],1)
X[i,6] <- equals(g4[i],1)
X[i,7] <- equals(g5[i],1)
X[i,8] <- equals(g6[i],1)
}
#Verosimilitud for(i in 1:34)
{
y[i] ~ dnorm(mu[i], tau)
mu[i] <- inprod(X[i,], beta[])
}
#Previas no informativas
for(j in 1:8){beta[j]~dnorm(0.0, 1.0E-06)}
sigma2<-1/tau
tau ~ dgamma(0.01, 0.01)
}
```

Después de seleccionar los SNPs significativos, se decidió utilizar el paquete `BayHap` (Iniesta & Moreno 2010) para realizar inferencias acerca de la relación de los haplotipos con las características de interés: ganacia de peso y peso a los 16 meses. A continuación se presenta el código que se implementó para obtener dicho resultado.

```
require(BayHap)

#Lectura y adecuación de los datos
R_BonGH1 <- read.table("C:/.../R_LEPROMO_GP.txt", header=T)
attach(R_BonGH1)
names(R_BonGH1)
View(R_BonGH1)
data.orig<-data.frame(SNP1,SNP2,GP)
data<-setupData(data.orig,snps.name=c("SNP1","SNP2"),sep="/")
data
```

```
#Estimación bayesiana de las frecuencias de haplotipos
res.freq<-bayhapFreq(data=data,na.snp.action="keep",col.snps=1:2,sep="/",
                     total.iter.haplo=10000)
print(res.freq)

#Estimación bayesiana de los efectos de los haplotipo
#en relación a las variables de interés
res.q<-bayhapReg(formula=GP~haplotypes,data=data, family="gaussian",
                 t.model="additive",na.snp.action="keep",
                 freqmin=0.01,burn.in.haplo=5000,burn.in.pheno=5000,
                 total.iter=5000,devhaplo=0.1,dist=1,lag.number=10,
                 sign=0.05,file=TRUE,prior.val=haplo.prior(),verbose=2)
print(res.q)
```

Por último, también se utilizó el paquete `BayHap` para inferir acerca de la asociación de los haplotipos con el valor genético de los individuos en el estudio. El siguiente código muestra la sintaxis empleada.

```
#Estimación bayesiana de la asociación de los haplotipos
#con el valor genético
res.l<-bayhapReg(formula=VG~haplotypes,data=data,
                 family="binomial",t.model="additive",na.snp.action="keep",
                 freqmin=0.01,burn.in.haplo=5000,burn.in.pheno=5000,
                 total.iter=5000,devhaplo=0.1,dist=1,lag.number=10,
                 sign=0.05,file=TRUE,prior.val=haplo.prior(),verbose=2)
print(res.l)
```