UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

DOCTORADO EN CIENCIAS DE LA SALUD Y DE LA PRODUCCION ANIMAL

MAPEO DE VARIACIÓN EN EL NÚMERO DE COPIAS EN EL GENOMA, ASOCIADO AL CONTEO CELULAR SOMÁTICO EN BOVINOS DE RAZA HOLSTEIN

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTORADO EN CIENCIAS DE LA SALUD Y DE LA PRODUCCION ANIMAL

PRESENTA:
MC MARINA DURÁN AGUILAR

PhD FELIPE DE JESÚS RUÍZ LÓPEZ
CENTRO NACIONAL DE INVESTIGACIÓN EN FISIOLOGÍA Y MEJORAMIENTO ANIMAL.
INSTITUTO NACIONAL DE INVESTIGACIONES FORESTALES AGRÍCOLAS Y PECUARIAS.
PhD CARLOS GUSTAVO VÁSQUEZ PELAEZ
PhD EVERARDO GONZÁLEZ PADILLA
FACULTAD DE MEDICINA VETERINARIA Y ZOOTECNIA. UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

CUAUTITLAN IZCALLI, MÉXICO ENERO 2017

DEDICATORIAS

A DIOS:

Por todas las innumerables bendiciones que he recibido y por permitirme llegar hasta aquí

A MIS PADRES:

Marina y Francisco: Porque a pesar de todo, siempre me apoyaron, me ayudaron y porque gracias a ustedes soy quien soy

A MI HIJO:

Roberto Dante: Por ser mi más grande motivo para continuar y mi vida entera

A MI ESPOSO:

Roberto: Por tu incansable apoyo en todo y tu amor incondicional

A MIS HERMANOS:

Liliana, Francisco y Alejandro: Por ser una parte importante en mi vida, y porque siempre me ayudan cuando lo necesito

A MI FAMILIA:

Abuelitos, Tíos, primos y sobrinos, gracias por quererme tanto, gracias porque todos me ayudaron siempre y me apoyaron. En especial a mis tías “Pilla”, “Güeris”, “China”, mi tío “Chino” y mi comadrita “Yoyis”, que han sido mis Ángeles en este camino.

AGRADECIMIENTOS

Al Dr. FELIPE RUIZ: Gracias por creer en mí, por su amistad y por el invaluable apoyo recibido para la realización de este proyecto.

Al Dr. SERGIO ROMAN: Gracias porque sin ti, este proyecto no hubiera sido posible y por tener confianza en mí para su realización.

A la Dra. MARIA STRILLACCI y al Dr. ALESSANDRO BAGNATO por todas las facilidades recibidas para la realización de este trabajo.

A la Facultad de Estudios Superiores Cuautitlán de la UNAM, por haberme dado la oportunidad de realizar mis estudios de posgrado.

Al Instituto Nacional de Investigaciones Forestales Agrícolas y Pecuarias (CENID-FISIOLOGÍA-INIFAP) y a la Asociación Holstein de México, por el apoyo recibido y la confianza para utilizar la información.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por el apoyo recibido a través de la beca 42771 para la realización de mis estudios.

AL Dr. GERMINAL CANTÓ: Por tu apoyo, tu guía y por todos estos años de amistad

ÍNDICE

ÍNDICE DE FIGURAS Y CUADROS

ABREVIATURAS

GWAS	Estudios de asociación del genoma (por sus siglas en inglés)
CNV	Variaciones en el número de copias
SNP	Polimorfismos de un solo nucleótido
HMMC	Modelo de Cadenas Ocultas de Markov
PCA	Análisis de componentes principales

RESUMEN

La mastitis clínica (MC) es una de las enfermedades más importantes, comunes y costosas que afectan la glándula mamaria (Bloemhof*et al.,* 2009; Cha *et al.,* 2011) e implica pérdidas importantes en la industria lechera a nivel mundial.

Existe una fuerte correlación genética entre mastitis y la calificación lineal de células somáticas (SCS) que sugiere que esta última característica está asociada a la resistencia a mastitis. Por ejemplo, Carlén *et al*. (2004) encontraron una correlación entre mastitis y SCS de 0.70. Debido a esto, se han realizado múltiples esfuerzos para mejorar la resistencia a mastitis, pero la baja heredabilidad de ésta característica hace que el proceso no sea tan efectivo como se quisiera.

Los estudios de asociación del genoma (GWAS por sus siglas en inglés) analizan las variaciones en las secuencias de DNA a lo largo de todo el genoma (Bush y Moore 2012) y su asociación con fenotipos y secuencias de polimorfismos comunes, las cuales pueden jugar un papel en la explicación del desarrollo de enfermedades (Forer et al. 2010).

Por otro lado, de acuerdo a diversos estudios, variaciones estructurales como las variaciones en el número de copias (CNV) afectan la expresión del gen y están relacionados con el inicio de muchas enfermedades (Bae et al. 2010). Hay indicios de que CNV aparecen en humanos, primates, roedores, moscas, perros, pollos, cerdos y ganado (Fadista et al. 2008, 2010; Fontanesi et al. 2010; Graubert et al. 2007; Wain et al. 2015; J. Wang et al. 2015). Recientemente, estudios de secuenciación del genoma demostraron que la mayoría de las bases que varían entre los genomas humanos residen en los CNV (Forer et al. 2010).

El uso de CNV como marcadores para explicar la variación genética de SCS no ha sido explorado hasta ahora y dada la importancia de ésta variable por su relación con la mastitis, consideramos este trabajo como la base en la detección de regiones genómicas asociadas a esta enfermedad.

Primero se realizó un estudio de asociación que contenía la información de 777,952 polimorfismos de un solo nucleótido (SNP) en 254 individuos más la información de 758 hembras con 77,000 SNP que fueron imputados a 777,952 SNP, obteniendo una base de datos final de 911 animales con 554,000 SNP, que se asociaron a SCS con un GWAS. En este trabajo se reportaron 286 SNP distribuidos en 16 cromosomas, de los cuales 154 se encontraron en 50 genes.

Posteriormente se realizó un análisis con un modelo univariado (software Golden Helix SVS8) y uno con el HMMC (software PennCNV) en donde encontramos 24 y 47 CNVR asociadas con los valores genéticos para SCS (EBV_SCS) respectivamente.

El análisis que se realizó, permitió la asociación de CNVR a 18 genes candidatos (TERT, NOTCH1, SLC6A3, CLPTM1L, PPARa, BCL-2, ABO, VAV2, CACNA1S, TRAF2, RELA, ELF3,

DBH, CDK5, NF2, FASN, EWSR1 and MAP3K11) que están clasificados en el mismo grupo funcional, asociados a “stress”, “muerte celular” “inflamación” y “respuesta inmune”.

Los CNV identificados en este estudio constituyen el primer mapa de CNVR asociados a SCS en ganado, proveyendo información nueva para los siguientes estudios de asociación con características de interés económico y de salud.

ABSTRACT

Clinical Mastitis (MC) is one of the most common, expensive and important diseases that affect the mammary gland (Bloemhof*et al.,* 2009; Cha *et al.,* 2011). Which implies important loses in milk industry worldwide.

There is a strong genetic correlation between mastitis and SCS of 0.70 which suggests that SCS may be linked to mastitis resistance. Because of this, efforts have been made to improve mastitis resistance, but the low heritability of this characteristic makes this process not to be as effective as desired.

The GWAS measures and analyzes sequence variations of DNA throughout the entire genome (Bush y Moore 2012). GWAS have a identified associations among various phenotypes and sequences of common polymorphisms which may play a role in the development of diseases (Forer *et al.,* 2010).

On the other hand, according to diverse studies, structural variations such as CNV affect the expression of the gene and are related to the start to many diseases (Bae et al. 2010).

There are indications that CNV appear in humans, primates, rodents, flies, dogs, chicken, pigs and cattle (Fadista et al. 2008, 2010; Fontanesi et al. 2010; Graubert et al. 2007; Wain et al. 2015; J. Wang et al. 2015). Recently, sequence studies of genome probe that most of the bases that vary between human genomes reside in the CNV (Forer et al. 2010).

The use of CNV as markers to explain the genetic variation of SCS has not been explored so far and given the importance of this variable due to its relation to mastitis we consider this work as the base of identification of genomic regions associated to this disease.

First, an association study that contained the information of 777,952 SNP in 254 individuals plus the information of 758 females with 77,000 SNP which where imputed to 777,952 SNP, was realized getting a final data base of 911 animals whit 554,000 SNP, which where associated to SCS. This work reported 286 SNP distributed in 16 chromosomes of which 154 are in 50 genes.

Subsequently, an analysis was performed with a univariate model (software Golden Helix SVS8) and one with the HMMC (Software PennCNV), where 24 and 47 CNVR associated with the EBV_SCS were found, respectively.

The analysis performed allowed the association of CNVR to 18 candidate genes XX which are classified in the same functional group associated to “Stress”, Cell dead”, “Inflammation” and “immune response”.

The identified CNV in this study constitute the first map of CNVR associated to SCS in cattle, providing new information for future studies of association with characteristics of economic and health interest.

ANTECEDENTES

MASTITIS

La mastitis clínica y subclínica (MC y MS) son unas de las enfermedades más importantes, comunes y costosas que afectan la glándula mamaria. Está caracterizada por signos visibles como inflamación y puede ser causada por patógenos (Bloemhof *et al*., 2009; Cha *et al.,* 2011). Está definida como una inflamación de la glándula mamaria resultante de la introducción y multiplicación de microorganismos patógenos. Los principales organismos incluyen *Staphylococcus aureus*, *Streptococcus agalactiae* y coliformes. Estos patógenos pueden causar enfermedades clínicas con cambios en la composición de la leche, un incremento en el conteo celular somático (CCS) e incluso muerte (Bjørg Heringstad *et al*., 2000).

La susceptibilidad de la glándula mamaria bovina a mastitis, depende del proceso de involución del tejido glandular y de la exposición a diferentes factores de tipo fisiológico, genético y ambiental (Sordillo y Streicher, 2002). Los traumatismos mecánicos, térmicos y químicos predisponen a la glándula a infecciones intramamarias (Zhao y Lacasse 2008).

Se considera mastitis subclínica cuando hay inflamación de la glándula mamaria sin anormalidad visible de la leche o de la ubre (Radostits *et al*., 1992), pero está asociada con el incremento al CCS y tiene consecuencias negativas (Hagnestam-Nielsen *et al*., 2009). El CCS puede ser utilizado como un indicador de mastitis y como una medida de respuesta a la infección (Heringstad *et al*., 2006). La selección para bajas calificaciones lineales de células somáticas puede disminuir la incidencia de la mastitis clínica y proveer beneficios económicos a través de los premios para calidad de leche (Schutz, 1994).

El control de la mastitis es de primordial importancia, sin embargo, las medidas preventivas están asociadas a costos extra, por lo tanto, deben ser evaluadas contra las pérdidas económicas causadas por mastitis (Hagnestam-Nielsen *et al*., 2009).

El CCS se utiliza como una medida de calidad de la leche, es un método para determinar la existencia o no de mastitis. Las células somáticas son células del cuerpo presentes a bajos niveles en leche normal, altos niveles de estas células son indicativos de

una infección intramamaria; en general, se considera un nivel alto de mastitis si el conteo celular supera los 500,000/ml en tanque, aunque en algunos países se considera un nivel alto 300,000/ml (Radostits *et al*., 1992)

La resistencia a una enfermedad infecciosa o la ausencia de susceptibilidad puede estar definida como la habilidad de la respuesta inmune de un animal para evadir la replicación de patógenos después del establecimiento de la infección. Esto implica que los animales tienden a variar en potencial genético de inmuno-competencia (Knap y Bishop 2000). La respuesta a la mastitis tiene muchos genes que controlan esta característica en muchos loci (Sender *et al*., 2013).

La susceptibilidad genética a mastitis involucra interconexiones de mecanismos biológicos que activan y regulan los diferentes niveles de respuesta inmune (Schukken *et al*., 2011), se incrementa lentamente como una respuesta correlacionada a la selección para producción de leche, las correlaciones genéticas indican que se puede hacer una disminución en la mastitis clínica al seleccionar para reducir el CCS (Schutz 1994; Sender *et al*., 2013).

La heredabilidad para mastitis es de 0.03 y la de CCS se estima en 0.01 (de Haas *et al*., 2008). Las correlaciones genéticas entre CCS y mastitis clínica van de 0.55 a 0.93 y para CCS y mastitis subclínica de 0.55 a 0.98 (de Haas *et al*., 2008)

Heringstad *et. al.,* (2006) encontraron que la correlación genética entre los primeros días de prueba y SCS fue de 0.78; mientras que Carlén *et al*., (2004) encontraron una correlación entre mastitis y SCS de 0.70.

LA GENÓMICA EN GANADO HOLSTEIN

La raza Holstein es la más alta productora de leche; la mayor parte de su producción fue obtenida por el uso de la inseminación artificial y por la selección del ganado elite para el uso de sementales, misma que fue basada en los valores genéticos estimados por pruebas de progenie (Seroussi *et al*., 2010). La selección para características cuantitativas económicamente importantes en animales y plantas, está basada tradicionalmente en los valores fenotípicos del individuo y sus parientes (Meuwissen *et al*., 2001). La selección

genómica es una herramienta basada en la estimación de los valores genéticos para características cuantitativas a través del uso de marcadores moleculares que por su densidad y ubicación expliquen una alta proporción de variación genética aditiva, lo que hace posible predecir el mérito genético usando solo la información proveniente del genoma con una elevada precisión (Seroussi *et al*., 2010).

El genoma bovino está constituido de 29 autosomas y los cromosomas sexuales con un tamaño de genoma estimado de alrededor de 2.87 Gbp. (Bae *et al*., 2010). En los últimos años, la genómica en bovinos ha progresado rápidamente, generando herramientas valiosas, incluyendo un mapa compuesto de bases y conjuntos independientes del genoma (El Consorcio de la secuenciación y análisis del genoma bovino, 2009). Estos recursos proveen evidencia preliminar de variaciones en la secuencia de ADN específicas de rumiantes en genes asociados con características productivas y respuestas inmunes.

Debido a la importancia de la genómica en los bovinos, el consorcio de la secuenciación y análisis del genoma bovino, ha decodificado la información completa del genoma bovino y ha reportado que un mínimo de 22,000 genes están incluidos en el genoma del ganado (Bae *et al*., 2010).

Las variaciones genómicas han sido estudiadas en laboratorios citogenéticos por un largo periodo gracias a la disponibilidad de tecnologías sencillas para la secuenciación de ADN (Ácido Desoxirribonucleico).

POLIMORFISMOS DE UN SOLO NUCLEÓTIDO (SNP)

La comunidad que investiga la genética del ganado ha migrado a polimorfismos de un solo nucleótido (SNP por sus siglas en inglés). Los SNP son polimorfismos en una única base y se definen como una forma abundante de variación genómica, que se distingue ya que como requisito, las variaciones raras para el alelo menos abundante deben tener una frecuencia mayor al 1% (Brookes, 1999). Su descripción estadística consiste en estimar las frecuencias alélicas y genotípicas respectivamente (Iniesta, 2005).

Los SNP se han convertido en el criterio principal de valoración de la variación genética en el ganado. Esto después de que se produjera la primera versión del mapa de

SNP en ganado y el arreglo para genotipificar BovineSNP50 (Liu *et al*., 2010). El entendimiento de las bases biológicas de los efectos genéticos juega un papel importante en el desarrollo de nuevas terapias farmacológicas (Bush y Moore, 2012)

En la actualidad existen diferentes plataformas para identificar marcadores moleculares tipo SNP en altas densidades; lo que ha popularizado su uso en estudios de genética animal (Wang *et al*., 2010). Se denominan marcadores de alta densidad cuando la distancia obtenida entre cada marcador es menor a 500 Kb, debido a que a distancias mayores el desequilibrio de ligamiento desaparece (McKay *et al*., 2007).

Las evaluaciones del mérito genético basadas en SNP se convirtieron en realidad en 2009 dando lugar a una aceleración en el mejoramiento de las razas de leche y carne (Matukumalli *et al*., 2009). La investigación genética molecular en ganado y cultivos, tiene la expectativa de que la información a nivel de ADN dará lugar a una ganancia genética más rápida que la obtenida basándose en datos fenotípicos únicamente (Meuwissen *et al*., 2001).

ESTUDIO DE ASOCIACIÓN EN EL GENOMA (GWAS)

Los estudios de asociación del genoma (GWAS por sus siglas en inglés) miden y analizan las variaciones en las secuencias de DNA a lo largo de todo el genoma (Bush y Moore 2012). Los GWAS han identificado asociaciones entre varios fenotipos y secuencias de polimorfismos comunes, las cuales pueden jugar un papel en el desarrollo de enfermedades (Forer *et al*., 2010).

El Instituto Nacional de Salud (National Institute of Health) define al GWA como un estudio de variación genética común a través de todo el genoma humano, diseñado para identificar asociaciones genéticas con características observables (Pearson y Manolio 2008).

El GWAS utiliza los factores de riesgo genéticos para hacer predicciones e identificar los apuntalamientos biológicos de susceptibilidad a enfermedades para el desarrollo de nuevos tratamientos y estrategias de prevención (Bush y Moore 2012).

Las características cuantitativas generalmente son analizadas utilizando una regresión lineal bajo el supuesto de que la característica se distribuye normal, la varianza

dentro de cada grupo es la misma y los grupos son independientes. Se han reconocido algunos factores que influencian los resultados de asociación, tales como edad, sexo, etc. Un previo Análisis de Componentes Principales (PCA por sus siglas en inglés) puede identificar y corregir la influencia de estos factores (Scherer, 2014a).

PCA reduce el número de variables perdiendo la menor cantidad de información posible, realizándolo de tal forma que el primer factor se queda con la mayor proporción posible de variabilidad y así sucesivamente (Salinas, 2006)

Los GWAS también han demostrado interacciones gen-gen o modificaciones de la asociación de una variante genética a otra, y pueden detectar combinaciones de múltiples SNP dentro de un solo gen (Scherer 2014b).

Estas variaciones ocurren en la frecuencia del alelo entre subgrupos de la población. Usualmente examinando la distribución de las pruebas estadísticas generadas por los cientos de pruebas de asociación realizadas (Ej X2)

Asociaciones con 2 alelos de cada SNP son probadas de una manera relativamente sencilla, comparando la frecuencia de cada alelo en casos y controles, algunos análisis pueden incluir pruebas de diferentes modelos genéticos (aditivo, dominante o recesivo). Los modelos aditivos tienden a ser los más comunes, ya que se supone que cada copia de alelos incrementa el riesgo en la misma cantidad (Pearson y Manolio 2008).

En un valor de P convencional ($P<.05$ de nivel de significancia), un estudio de asociación de 1 millón de SNP puede mostrar 50,000 SNP “asociados” con la enfermedad, casi todos, falsos positivos. La manera más común de reducir el grado de falsos positivos es aplicando la corrección de Bonferroni, la cual consiste en dividir el valor de P entre el número de pruebas realizadas. Esta corrección supone asociaciones independientes para cada SNP con la enfermedad, aun cuando se conoce que los SNP tienen algún grado de correlación debido al desequilibrio de ligamiento (Pearson y Manolio 2008).

El número sin precedentes de comparaciones realizadas en estudios de asociación utilizando SNP ha llevado al reconocimiento de que una asociación no identificada inicialmente puede ser confiable hasta que haya sido replicada en uno o más estudios de tamaño adecuado. El proceso usualmente involucra un diseño multi-etapas, en el cual la

replicación se intenta para algunos de los SNP que fueron encontrados en el estudio original. (Hunter y Kraft 2007).

El principal inconveniente de trabajar con un número bajo de SNP es que el valor de P para la asociación de un SNP con un fenotipo dado en el estudio original, puede no ser lo suficientemente pequeño para que el SNP sea llevado a la segunda etapa del análisis, por lo que la asociación es desestimada, y se provoca un falso negativo (Hunter y Kraft 2007).

Para probar cada SNP en turno se supone que las variantes de la enfermedad pueden ser detectadas basadas en los efectos marginales. La parte principal involucra los valores de P. La interpretación de estos valores requiere conocimiento del poder de la prueba usada (Bush y Moore 2012).

Después del análisis de asociación se genera un valor de P, el cual Bush y Moore (2012) describen como: "Un valor de P es la probabilidad de ver una prueba estadística igual o mayor que la prueba estadística observada si la hipótesis nula es verdadera, esto se genera para cada prueba estadística. Esto significa que los valores bajos de P indican que, si no hay asociación, el cambio que se observa en estos resultados es extremadamente pequeño".

Las pruebas estadísticas generalmente son llamadas significativas cuando los valores de P están por debajo de cierto umbral, comúnmente es de 0.05, pero puede disminuir a 0.01 o menos. Esto aplica cuando se realiza una sola prueba estadística, pero GWAS estudia millones de pruebas.(Scherer 2014b).

GWAS tiene la habilidad de detectar, para una característica, pequeñas regiones del cromosoma que se ven afectadas en el análisis de ligamiento, esto provee evaluaciones más precisas del tamaño y la dirección de los efectos de los alelos (Abdel-Shafy *et al*., 2014)

Goddard y Hayes (2009) definieron los QTL como una medida de la característica que depende de la acción acumulativa de muchos genes y del medio ambiente, y que puede variar entre individuos en un rango determinado para producir una distribución continua de fenotipos.

Diferentes GWAS han detectado QTL para CCS, en los cromosomas 5, 6, 8, 11, 17, 18, 20 y 23 en diversas razas de ganado lechero (Lund *et al*., 2008; Meredith *et al*., 2013; Sender *et al*., 2013; Strillacci *et al*., 2014; Tiezzil *et al*., 2015).

VARIACIONES EN EL NÚMERO DE COPIAS (CNV)

Las variaciones estructurales genómicas entre especies pueden involucrar cambios tan pequeños como un solo nucleótido o tan grandes como segmentos de cromosoma visibles microscópicamente (Wang *et al*., 2010). No fue sino hasta hace poco que la variación genómica que involucra segmentos intermedios de ADN fue reconocida y llamada Variación en el Número de Copias (CNV, por sus siglas en inglés). Redon *et al*. (2006) definieron los CNV como un segmento de 1kb o mayor que presenta una variante en el número de copias con respecto al genoma de referencia. Éste tipo de variación genómica involucra inserciones submicroscópicas, eliminaciones, duplicaciones segmentales y cambios complejos desde 1 kb hasta 5 Mb de tamaño (Bae *et al*., 2010; Fadista *et al*., 2010; Hou *et al*., 2011; Wang *et al*., 2010). Se encontró que al menos el 20% del genoma era variable en el número de copias (Forer *et al*., 2010).

Anteriormente, se pensaba que los SNP eran la mayor fuente de variación entre individuos, pero ahora se conoce que la variación genética en genómica estructural, abarca más pares de bases (Fadista *et al*., 2010). De acuerdo a estudios, variaciones estructurales como los CNV afectan la expresión del gen y están relacionados con el inicio de muchas enfermedades (Bae *et al*., 2010). Hay indicios de que CNV aparecen a lo largo del genoma, no solo en humanos, también en primates, roedores, moscas, perros, pollos, cerdos, cabras y ganado (Fadista et al. 2008, 2010; Fontanesi et al. 2010; Graubert et al. 2007; Wain et al. 2015; J. Wang et al. 2015). Estudios de secuenciación del genoma demostraron que la mayoría de las bases que varían entre los genomas humanos residen en los CNV (Forer *et al*., 2010).

Los CNV pueden ser identificados usando varios planteamientos, incluyendo arreglos de hibridación genómica comparativa (aCGH, por sus siglas en inglés), arreglos SNP y secuenciación de próxima generación. Se han desarrollado diferentes algoritmos para

detectar CNV, algunos ejemplos de software que han sido basados en arreglos de SNP son: Partición CNV, QuantiSNP, PennCNV, Birdsuite, Golden Helix SVS y Cokgen (Hou *et al*., 2012).

Para la detección de CNV, se utiliza el Log R ratio (LRR) y la Frecuencia del Alelo B (BAF).

El valor del LRR representa la intensidad total de la señal y el valor del BAF es el balance alélico. La intensidad de la señal LRR es calculada como:

$$log2(\frac{Rsubject}{Rexpected})$$

Donde:

Rsubject: Es el valor de intensidad total de la señal observado por SNP para cada individuo

Rexpected: Es la interpolación de los puntos medios de dos cluster canónicos vecinos.

BAF es un estimado de la frecuencia relativa del alelo B de un locus en un individuo, en un rango de cero a uno (indica un genotipo heterocigoto, y 0 y 1 corresponden al genotipo homocigoto AA y BB respectivamente).

La frecuencia es obtenida por una interpolación lineal con las líneas D_1 y D_2 para un valor observado de θ de una muestra que está localizada entre dos clusters. Ambos valores, LRR y BAF, se pueden visualizar a través del agrupamiento en los clusters canónicos apropiados para cada SNP (Figura 1). BAF es calculado con respecto a frecuencias de alelos conocidas, asignadas a cada cluster canónico (BAF_{cc}:0,0.5,1):

$$BAF = \frac{D_1}{D_2} * BAF_{cc}$$

Donde:

D_1: Es la distancia de un valor de θ observado al punto medio de la solución cluster más cercana. El valor de θ entre los cluster canónicos vecinos son interpolados y la distancia entre ellos es medida como D_2 (Peiffer *et al*., 2006)

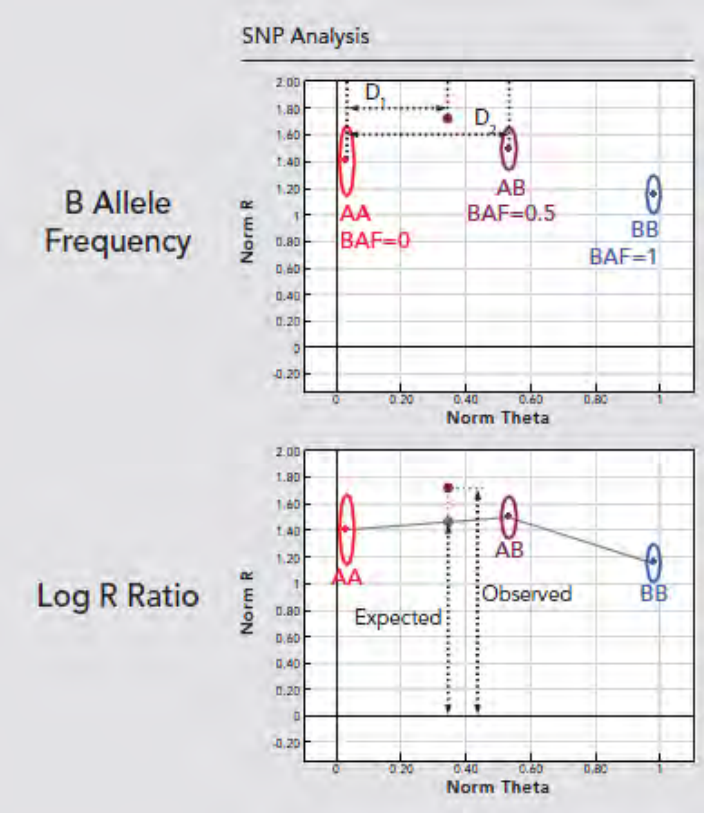
Diferentes algoritmos ayudan a la identificación genómica de CNV comparando el LRR de un genoma de referencia con el LRR de las muestras genotipadas. PennCNV y Golden Helix (Variation Suite 8.0, SVS8) son algunos de los softwares que pueden ser usados para la detección de los CNV.

Los métodos convencionales para la identificación de CNV en la plataforma de microarreglos de Illumina, involucran un examen de la intensidad de las señales, el cual identifica los cambios en el número de copias calculando la BAF por SNP en una ventana deslizante a lo largo del cromosoma. Tal es el caso del software PennCNV que incorpora múltiples fuentes de información, incluyendo la intensidad total de la señal y la relación de la intensidad alélica en cada marcador SNP (Wang *et al*., 2007). PennCNV también puede considerar la información de pedigrí (Hou *et al*., 2012). Este software trabaja en una base Linux, puede presentar gráficas mostrando la ubicación de los CNV y los SNP que se encuentran en ese CNV.

PennCNV utiliza un modelo de cadenas de Markov (HMMC) que integra múltiples fuentes de información para inferir los CNV de las muestras genotipificadas. La HMMC es una técnica estadística que supone que la distribución de la intensidad de un dato observado en un punto depende en un no observado (oculto) número de copias en cada locus, donde los elementos del estatus oculto siguen un proceso de Markov.

El software PennCNV incorpora diferente información (LRR y BAF) por cada SNP dentro de la HMMC, con la finalidad de detectar CNV y diferenciar las regiones de CNV neutrales (LOH) de las regiones normales. La distancia entre cada SNP determina la probabilidad de tener un cambio en el status del número de copia (CN por sus siglas en inglés).

Figura 1: Clusters de BAF y LogR Ratio



Seis estatus ocultos son identificados como se muestra en el cuadro 1 (Wang *et al.*, 2007)

Cuadro 1. Estatus ocultos y su correspondiente descripción de CN como es modelado en el software PennCNV (Wang *et al.*, 2007)

Status CN	CN Total	Descripción (para autosomas)	Genotipo de CNV
1	0	Delección de dos copias	Null
2	1	Delección de una copia	A, B
3	2	Status normal	AA, AB, BB
4	2	Copia-neutral con LOH	AA, BB
5	3	Duplicación en una copia	AAA, Aab, ABB, BBB
6	4	Duplicación en dos copias	AAAA, AAAb, AABB, ABBB, BBBB

Forer *et al.* (2010) desarrollaron una herramienta para el análisis de las CNV, llamada CONAN (por sus siglas en inglés, Copy Number Variation Analysis Tool), que define las regiones CNV (CNVR por sus siglas en inglés) como la unión de todos los CNV superpuestos entre los individuos. Estas regiones pueden ser divididas en sub-regiones. La frecuencia de una subregión CNV se define como el porcentaje de individuos que tienen una CNV dentro de sus límites (los límites de cada sub-región son aproximaciones a la posición física de su SNP.

El software Golden Helix SVS8 (SVS8), ofrece la posibilidad de procesar los datos puros de intensidades, para identificar regiones de CNV. El algoritmo del CNAM (Copy Number Analysis Method, por sus siglas en inglés) delinea los límites del CNV incluso en una sola sonda con niveles controlables de sensibilidad y tasa de falsos positivos.

El procedimiento para la detección de CNV con genotipos en SVS8 se describe a continuación:

El proceso de los datos puros y generación de los LRR: La "normalización" de los valores puros es necesaria debido a que la manufactura de los arreglos y algunos otros factores pueden influenciar la distribución de los valores de cada chip. La medida utilizada para identificar el estatus de los CNV es el LRR. Para su determinación se requiere un panel de referencia para determinar el "normal" o la línea base de la intensidad esperada para cada marcador. (Golden Helix, Bozeman, MT, www.goldenhelix.com)

Control de calidad: Este es el paso más importante en el análisis de CNV para reducir los falsos positivos y las asociaciones no replicables. El sesgo que se puede introducir en estos análisis puede venir de diferentes fuentes tales como la extracción de ADN, los tipos de células, las fluctuaciones en la temperatura e incluso los niveles de ozono en el laboratorio, debido a esto, se ha visto que las propiedades de cada muestra afectan la exactitud de la detección de los CNV, este software realiza varios filtros de calidad:

a) La derivada del Log Ratio (DLRS por sus siglas en inglés): es una medida de punto a punto de consistencia o ruido en los datos de LRR. Este valor esta correlacionado con una baja calidad del SNP "call rates" y sobre/baja abundancia de identificar segmentos de CN. Las muestras con altos valores

de DLRS tienden a tener señal pobre por lo que la precisión de la detección de los CNV es difícil para estas muestras.

b) Detección de ondas en los datos de LRR: El fenómeno de las ondas ocurre cuando los datos de LRR parecen tener un rango grande de patrón de onda después de graficar el genoma. Las ondulaciones parecen estar correlacionadas con el contenido de G-C de la región alrededor de la sonda. La metodología usada en SVS8 remueve las muestras con factores de onda extremos.

c) Análisis de componentes principales (PCA): Este procedimiento es necesario para detectar la presencia de efectos de grupo para la corrección de los datos de LRR antes de la detección de CNV.

Detección de CNV: SVS8 provee dos algoritmos de segmentación para la detección de CNV. El método univariado, el cual se utiliza para detectar CNV raros y/o largos, considera solamente una muestra a la vez: por el contrario, el método multivariado utiliza todas las muestras simultáneamente y es ideal para detectar CNV pequeños y comunes. El objetivo de este paso es determinar las regiones en el genoma donde la media del valor de LRR de una muestra dada difiere del valor de referencia. La media de un segmento LRR arriba de cero significa que hay un CN "ganancia", y la media de un segmento LRR debajo de cero significa que hay un CN "perdida".

Para analizar las variaciones en el número de copias se deben observar las medias de los LRR de los segmentos para encontrar las diferencias entre dos o más grupos (casos/controles, padres/crías, etc.). Esto se puede realizar de manera visual utilizando técnicas de graficado o estadísticas simples o análisis de regresión

VARIACIONES EN EL NÚMERO DE COPIAS (CNV) EN BOVINOS

La comunidad científica se ha enfocado en el uso de los SNP como la principal fuente de variación en ganado (Hou *et al.*, 2011). Aunque los SNP son más frecuentes, las CNV

involucran más secciones genómicas y potencialmente tienen más efectos, incluyendo cambios en la estructura del gen, dosificación, regulación en la alternancia del gen y exposición a alelos recesivos (Liu *et al.*, 2010).

Las CNV pueden representar variantes polimórficas benignas aun cuando en muchos otros casos estén asociados con la herencia Mendeliana en humanos y desordenes genéticos complejos (Fontanesi *et al.*, 2010). Varias publicaciones han revisado los efectos de las CNV en la expresión de genes y en enfermedades en humanos (Hou *et al.*, 2011). Las CNV existen en baja frecuencia; por lo que es conveniente analizar muchas muestras a fin de encontrar regiones comunes de CNV para análisis con varios fenotipos (Bae *et al*., 2010). Actualmente tenemos la posibilidad de buscar regiones genómicas que impactan la variación genética de importantes características fenotípicas (Fadista, *et al*., 2010). Poco se conoce acerca de cómo las CNV contribuyen a la variación fenotípica normal y a la susceptibilidad de enfermedades.

Muy pocos estudios han confirmado la presencia de CNV en ganado; Fadista *et al.* (2010), reportaron 304 regiones CNV en 20 animales de 4 razas de ganado usando arreglos de alta densidad aCGH con un espaciamiento de sonda promedio de 420 pb con respecto al último ensamblaje del genoma bovino. Evidencia adicional de la existencia de CNV proviene de datos de SNP bovinos, donde en un grupo inicial de 556 animales de 21 razas de ganado identificaron 79 candidatos de variantes de eliminación (Liu *et al*., 2010).

Fadista *et al.* (2010), sugieren que los CNV pequeños (<50kb) son mucho más frecuentes que los grandes, lo que concuerda con otros estudios de alta resolución. Si esto puede ser extrapolado para todo el genoma del ganado, el panel disponible comercialmente de Illumina 50 k SNP podría no ser suficiente para detectar el volumen de CNV existentes. Consecuentemente, el uso de los arreglos de alta densidad de SNP podría brindar la resolución adecuada para poder efectuar una mejor caracterización de los CNV en ganado.

Cuadro 2: Número de CNVR detectadas en otros estudios utilizando plataformas de CGH, arreglos SNP y secuenciación de próxima generación.

Table 2. Number of overlapping copy number variant (CNV) regions detected in our study and previous studies using comparative genomic hybridization-based (CGH-based), SNP array, and whole-genome sequencing platforms

Method	Study	No. of breeds	CNV count	Overlapping
CGH-based studies	Liu et al., 2010	17	177	14
	Fadista et al., 2010	4	266	9
BovineSNP50 Beadchip	Bae et al., 2010	1	368	17
	Hou et al., 2011	21	682	25
	Hou et al., 2012	1	811	17
	Seroussi et al., 2010	1	278	9
	Jiang et al., 2012	1	99	6
	Jiang et al., 2013	1	367	21
	Cicconardi et al., 2013	5	326	27
Next-generation sequencing	Bickhart et al., 2012	3	1,265	27
	Zhan et al., 2011	1	520	12
	Stothard et al., 2011	2	790	3
	Boussaha et al., 2015	3	957	30

(Ben Sassi *et al*., 2016)

El cuadro 2 muestra una comparación de los CNV encontrados con diferentes plataformas y diferentes razas, excepto por la información de Prinsen *et al.*, (2016), lo que demuestra que la cantidad de CNV detectadas depende de la plataforma que se está utilizando, debido a que, hasta la fecha no existe información que relacione CNV con SCS, se decidió utilizar la plataforma de alta densidad para poder encontrar la mayor variación posible y con ello establecer las bases de asociación de CNV con SCS, porque como hemos observado, los trabajos anteriores solamente hacen la detección de CNV de manera general y las asociaciones que han encontrado no han sido dirigidas a enfermedades específicas.

El uso de CNV como marcadores para explicar la variación genética de SCS no ha sido explorada hasta ahora, y dada la importancia de ésta variable por su relación con la mastitis, consideramos este trabajo como la base en la detección de regiones genómicas asociadas a la enfermedad.

HIPÓTESIS

Los arreglos de alta densidad permiten identificar una mayor cantidad de regiones genómicas asociadas con Conteo Celular Somático.

Existen CNV en el genoma bovino que están relacionados con la resistencia a la mastitis en el ganado productor de leche de la raza Holstein.

OBJETIVOS

Identificar marcadores tipo SNP asociados a SCS

Identificar las variaciones en el número de copias asociadas a SCS

METAS

Identificar QTL asociados a SCS utilizando SNP imputados a alta densidad para la realización de un GWAS.

Identificar las regiones genómicas explicando la variación genética de SCS utilizando CNV en la población Holstein

TRABAJO 1

GWAS ON IMPUTED HIGH-DENSITY SNP GENOTYPES IN THE MEXICAN HOLSTEIN CATTLE POPULATION FOR SOMATIC CELL SCORE

ANIMAL GENETICS
International Society for Animal Genetics

GWAS ON IMPUTED HIGH-DENSITY SNP GENOTYPES IN THE MEXICAN HOLSTEIN CATTLE POPULATION FOR SOMATIC CELL SCORE

Journal:	*Animal Genetics*
Manuscript ID	AnGen-16-09-0283
Manuscript Type:	Short Communication
Date Submitted by the Author:	05-Sep-2016
Complete List of Authors:	Duran-Aguilar, Marina; Universidad Nacional Autonoma de Mexico, Facultad de Estudios Superiores Cuautitlán Strillacci, Maria; University of Milan, VESPA ROMAN-PONCE, SERGIO IVAN; Instituto Nacional de Investigaciones Forestales, Agricolas y Pecuarias, Centro Nacional de Investigación Disciplinaria en Fisiología y Mejoramiento Animal González Padilla, Everardo; Universidad Nacional Autonoma de Mexico, Departamento de Genética y Bioestadística, Facultad Medicina Veterinaria y Zootecnia Vásquez Peláez4, Carlos Gustavo; Universidad Nacional Autonoma de Mexico, Departamento de Genética y Bioestadística, Facultad Medicina Veterinaria y Zootecnia Ruiz López, Felipe J.; Instituto Nacional de Investigaciones Forestales, Agricolas y Pecuarias, Centro Nacional de Investigación Disciplinaria en Fisiología y Mejoramiento Animal Bagnato, Alessandro; University of Milan, Department of Health, Animal Science and Food Safety; University of Milan, Genomic and Bioinformatic Platform, c/o Filarete Foundation
Keywords:	Holstein, Imputation, Genome Wide Association Study, SNP, SCS

GWAS ON IMPUTED HIGH-DENSITY SNP GENOTYPES IN THE MEXICAN HOLSTEIN CATTLE POPULATION FOR SOMATIC CELL SCORE

Marina Durán Aguilar[1*], Maria Giuseppina Strillacci[2*], Sergio I. Román Ponce[3], Everardo González Padilla[4], Carlos G. Vásquez Peláez[4], Felipe J. Ruiz López[3], Alessandro Bagnato[2].

[1]Facultad de Estudios Superiores Cuautitlán. UNAM. Ave. 1o de Mayo S/N. Santa Maria las Torres, 54740 Cuautitlán Izcalli. México.

[2] Università degli Studi di Milano, Via Celoria 10, 20133, Milano, Italy.

[3] Centro Nacional de Investigación en Fisiología y Mejoramiento Animal. INIFAP. Km. 1 Carretera a Colón, Ajuchitlán, Querétaro, México. CP 76280

[4]Departamento de Genética y Bioestadística, Facultad Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Av. Universidad 300, México DF, 04510.

* equal contribution of authors

Corresponding author: Sergio I. Román Ponce, Centro Nacional de Investigación en Fisiología y Mejoramiento Animal. INIFAP. Km. 1 Carretera a Colón, Ajuchitlán, Querétaro, México. CP 76280.

roman.sergio@inifap.gob.mx

ABSTRACT

Mastitis is today considered one of the most significant diseases affecting dairy cattle. A genome wide association study (GWAS) was performed after imputation to Illumina Bovine HD SNP chip to identify Quantitative Trait Loci (QTL) associated with somatic cell score (SCS), an indicator of mastitis.

The association analysis was carried out on a total of 911 Mexican Holstein cattle using a final dataset of 541,915 SNPs. A total of 286 SNPs distributed on 16 autosomes reached genome-wide significance for association with SCS. Twenty-six distinct QTL regions (QTLR) were identified after defining haploblocks based on the linkage disequilibrium (LD) structure estimated for each SNP using the Pairwise LD approach. Neighbouring non-overlapping QTLRs of < 5Mb length were considered belonging to the same region. The largest frequency of significant SNPs was located in intronic positions of the *ATG4C* (BTA3), *EMOL1* (BT4) and *CHRM3* (BTA 28) genes. All these significantly associated SNPs, included in the respective haploblocks, are in high LD with each other with pairwise r^2 value of at least 0.95 and could be considered as markers for the mastitis resistance selection in dairy cattle.

Keywords

Holstein, Imputation, Genome Wide Association Study, SNP, SCS.

INTRODUCTION

Mastitis is today considered one of the most significant diseases affecting dairy cattle herds and, as many economically important quantitative traits in livestock, is under the control of multiple genes (Strillacci *et al.*, 2014; Sender *et al.*, 2013). With the availability of high throughput genotyping technologies, genome-wide association studies (GWAS) have become a useful tool for fine-scale quantitative trait loci (QTL) mapping (Sahana *et al., 2010*). The discovery of the genomic regions

that influence quantitative traits may provide an opportunity to gather new information about gene function, and lead to a better understanding of how genes interact to impact on physiology and on response to infectious diseases. In addition, detection of undisclosed loci contributing to the phenotypic variation of a trait may provide new markers that can be integrated into marker-assisted selection programs with the potential to have considerable impact on genetic gain and improvement for many traits of interest to breeders.

To reduce the cost of genotyping, the imputation strategy has been used to predict and assign genotypes that are not assayed by a genotyping chip (Meredith *et al.*, 2013). In this study, the genotype imputation from the Genseek genomic profiler to Illumina BovineHD Beadchip (777K SNP) was used to perform a GWAS for somatic cell score (SCS) as an indicator of mastitis in the Mexican Holstein cattle breed.

The Asociación Holstein de México (AHM) provided the genotypes and phenotypes for a total of 1002 samples. The Estimated Breeding Values (EBVs) for SCS (EBV_SCS) (mean -1.08 ± SD 1.4; minimum and maximum values of -5.63 and 4.52, respectively) is the official evaluation calculated based on an animal model by AHM (http://www.holstein.mx/QueToro.aspx).

Seven hundred and forty-eight females were genotyped with the medium-density customized GeneSeek Genomic Profiler HD Beadchip containing 76,883 SNPs. Two hundred and fifty-four individuals (53 males and 201 females) used a reference population (RP) in this study, were genotyped with the high-density Illumina BovineHD Beadchip comprising 777,962 polymorphic SNPs. The chromosomal position of the SNPs was determined according to the UMD3.1 assembly of the bovine genome.

A stringent quality control (QC) was carried out for each dataset separately using SVS (Golden Helix SVS 8.3.1) software (Golden Helix Inc., Bozeman, MT, USA). In detail, we excluded animals and SNPs with call-rates lower than 0.98, SNPs with deviations from the Hardy-Weinberg equilibrium ($p<0.00001$), SNPs with MAF <0.02, SNPs on not autosomal chromosomes, and all SNPs in the Geneseek Genomic Profiler HD Beadchip which were not present in BovineHD final

SNPs dataset. Genotype imputation was carried out by BEAGLE v3.3.2 software using a pre-phasing based approach (Browning and Browning, 2009).

The GWAS analysis was accomplished with a correlation/trend tests implemented in SVS software using only the SNPs with an accuracy imputation value (rsq) of at least 80% (the graphical representation of accuracies of imputated SNPs is reported in Figure S1). Stratification of population was considered during the association analysis through a PCA. In addition, a genome-wide Bonferroni multiple test correction set to 0.05 was applied.

A QTL region (QTLR) surrounding a detected significant association was defined based on the linkage disequilibrium (LD) structure for each SNP using the Pairwise LD approach (as measured by r^2) (Gabriel *et al.*, 2002). The haplotype blocks were also computed by SVS software and the haplotype frequencies for all 911 individuals were obtained with the standard expectation-maximization algorithm (EM). Neighbouring non-overlapping QTLRs of < 5Mb length were considered belonging to the same region.

Genes were annotated within QTLRs using the Table Browser of the USCS online database (https://genome.ucsc.edu/cgi-bin/hgTables). Gene ontology (GO) and pathways analyses, using the DAVID Bioinformatics Resources 6.8 (http://david.abcc.ncifcrf.gov/), were performed (with the high classification stringency option and the false discovery rate (FDR) correction) (Table S2).

After quality control and imputation procedures, a total 911 Holstein genotyped at 541,915 SNPs (corresponding to the 98% of imputed SNPs with an accuracy value of at least 0.80) (Figure S1a) was available for association testing. A total of 286 SNPs distributed on 16 autosomes showed significant genome wide association (range from 0.049 to 9.22E-08) with the EBV_SCS (Figure S2). Among the significant markers, 145 SNPs were annotated within 50 genes (details are reported in TableS1). The Table 1 reports the 26 QTLRs identified in Mexican Holstein cattle breed according to haploblocks definition; we found that the genome of our population was partitioned into 64,700 haploblocks. The highest number of significant SNPs (No. 35, of which 24 were included in different haploblocks), exceeding the significant threshold for genome-wide

significance signal, was found on BTA20 located at 41.88-42.95 Mb (QTLR_16). This QTLR maps within one of the previously significant genomic region associated with clinical mastitis identified by Tiezzi *et al.*, (2015). Other chromosome regions showing a higher number of significant SNPs (No. 24) is located and on BTA3 at 80.09-83.30 Mb (QTLR_3) and on BTA26 at 27.56-30.87 Mb (QTLR_22) (Table 1). This last region was significantly associated with the somatic cell count in Holstein cattle breed also for Longeri *et al.*, (2008).

The largest number of associated SNPs (No.18; p-values (Bonferroni correction) ranged from 3.11E-08 to 1.72E-09) of the QTLR_3 were located in intronic position of the *ATG4C* (*autophagy related 4C cysteine peptidase*) gene. The *ATG4C* is a member of a family of cysteine proteinases proposed to be involved in the processing of autophagy (Fernández and López-Otín, 2015) and together with other genes (i.e. *IL23R*) seems to be involved in the immune response against some infection disease, as those caused by *M. tuberculosis* (Daya *et al.*, 2015)

Other strongly associated SNPs are included in the QTLR_5 (p-values ranged from 2.34E-11 to 2.91E-11). These SNPs map within the *ELMO1* (*engulfment and cell motility 1*) gene that encodes for a member of the engulfment and cell motility protein family. These proteins, interacting with dedicator of cytokinesis proteins, promote phagocytosis and cell migration. Studies regarding this gene highlighted its role in facilitating intracellular bacterial sensing and the induction of inflammatory responses following infection disease after encountering commensals as well as pathogens (Das *et al.*, 2015).

In addition, within the intronic positions of the *cholinergic receptor muscarinic 3* (*CHRM3*) gene, map all the 4 significant SNPs that define the QTLR_26. This gene is expressed on macrophages, one of the key cells of inflammation, and contributes to their chemotaxis (Selivanova *et al.*, 2012). The activation of muscarinic receptors triggers Ca^{2+} influx causing smooth muscle contraction and glandular secretion. *CHRM3* is also one of the genes that resulted significantly affected by heat treatment of bovine mammary epithelial cells, as reported by Li *et al.* (2015).

The examination of the LD and the haplotypes frequencies calculated respectively among and for the SNPs defining haploblocks identified for the QTLR_3, QTLR_5 and QTLR_26 are shown in Figure 1. In detail, the haploblocks_1 and _2 (83.32–83.41 Mb) contain the most significant SNPs from the genome-wide association study mapping within the *ATG4C* gene (Figure 1a). The haploblock_3 (62.76–65.87 Mb) lies within the candidate gene *ELMO1*, while haploblocks_3 (12.76–12.78 Mb) overlap the candidate genes *CHRM3* (Figure 1b and Figure 1c). In addition, as shown, all significantly associated SNPs (in bold in table of the Figure 1a, 1b and 1c) included in the respective haploblocks are in high LD with each other with pairwise r^2 of at least 0.95.

These results show that the genomic regions evaluated in this study may contribute to a better understanding of the genetic control of the immune response to mastitis in dairy cattle and may allow the inclusion of more detailed QTL information in selection programs.

Figure legend

Figure 1. Linkage disequilibrium, Haploblocks and haplotype frequencies calculated for the significantly associated SNPs on BTAs 3, 4 and 26.

Supporting information

Figure S1 Graphical representation of accuracy values of imputated SNPs.

Figure S2 Manhattan plot of the GWAS for EBV_SCS.

Table S1 Significantly associated SNPs above the Bonferroni genome-wide threshold of 0.05.

Table S2 Annotation of genes mapping within QTLR associated with EBV_SCS.

Acknowledgements

We acknowledge the Cooperative Dairy DNA Repository and Holstein de Mexico A.C. for sharing genomic data. This work received funding from CONACYT, México (project "IDENTIFICACIÓN

DE VARIACIONES ESTRUCTURALES Y REGIONES GENÓMICAS ASOCIADAS A RESISTENCIA A MASTITIS EN BOVINOS PRODUCTORES DE LECHE¨. 176039).

Competing interests

The authors declare they have no competing interests.

References

Browning B.L. & Browning S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84. 210–223.

Das S., Sarkar A., Choudhury S.S., Owen K.A., Castillo V., Fox S., Eckmann L., Elliott M.R., Casanova J.E. & Ernst P.B. (2015). ELMO1 has an essential role in the internalization of Salmonella Typhimurium into enteric macrophages that impacts disease outcome. *Cellular and molecular gastroenterology and hepatology* 1. 311-324.

Daya M., Van der Merwe L., Van Helden P.D., Möller M. & Hoal E.G. (2015) Investigating the Role of Gene-Gene Interactions in TB Susceptibility. *PLoS One* 10(4): e0123970.

Fernández Á. F., & López-Otín C. (2015) The functional and pathologic relevance of autophagy proteases. *The journal of clinical investigation* 125, 33-41.

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R.,

Ward R., Lander E.S., Daly M.J. & Altshuler D. (2002) The structure of haplotype blocks in the human genome. *Science* 21, 2225-9.

Li L., Sun Y., Wu J., Li X., Luo M. & Wang G. (2015) The global effect of heat on gene expression in cultured bovine mammary epithelial cells. *Cell Stress and Chaperones* 20, 381–389.

Longeri M., Polli M., Strillacci M.G., Samore A.B. & Zanotti M. (2006) Short communication: quantitative trait loci affecting the somatic cell score on chromosomes 4 and 26 in Italian Holstein cattle. *Journal of Dairy Science* 89, 3175-7.

Meredith B.K., Berry D.P., Kearney F., Finlay E.K., Fahey A.G., Bradley D.G. & Lynn D.J. (2013) A genome-wide association study for somatic cell score using the Illumina high-density bovine beadchip identifies several novel QTL potentially related to mastitis susceptibility. *Frontiers in Genetics* 6, 4:229.

Sender G., Korwin-Kossakowska A., Pawlik A., Galal Abdel Hameed K. & Oprzadek J. (2013) Genetic basis of mastitis resistance in dairy cattle a review. *Ann Animal Science* 13, 663-673.

Selivanova P.A., Kulikov E.S., Kozina O.V., Trofimenko I.N., Freidin M.B., Chernyak B.A. & Ogorodova L.M. (2012) Differential expression of the β2-adrenoreceptor and M3-cholinoreceptor genes in bronchial mucosa of patients with asthma and chronic obstructive pulmonary disease. *Annals of Allergy, Asthma and Immunology* 108, 39-43.

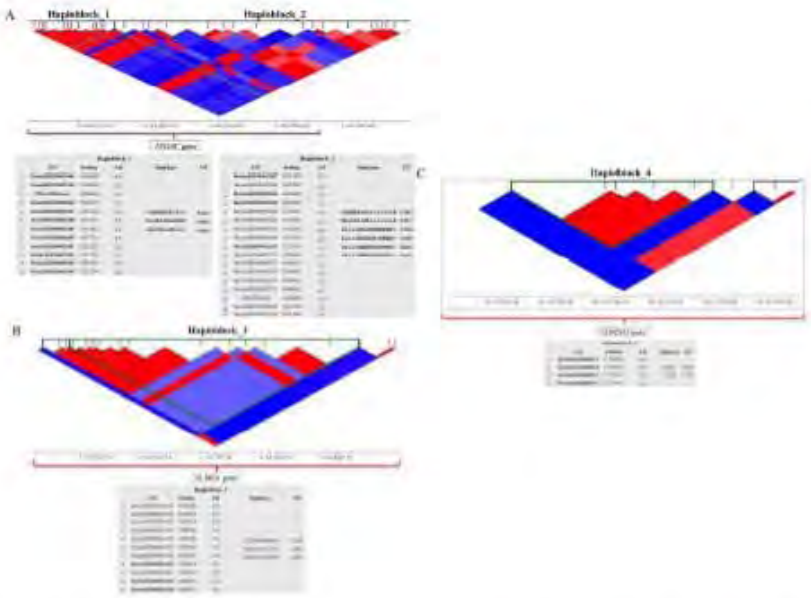
Strillacci M.G., Frigo E., Schiavini F., Samoré A.B., Canavesi F., Vevey M., Cozzi M.C., Soller M., Lipkin E. & Bagnato A. (2014) Genome-wide association study for somatic cell score in Valdostana red pied cattle breed using pooled DNA. *BMC Genetics* 15:106.

Tiezzi F., Parker-Gaddis K.L., Cole J.B., Clay J.S. & Maltecca C. (2015) A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. *PLoS One* 10(2): e0114919.

Table 1. QTLRs associated with EBV_SCS identified in Mexican Holstein cattle breed.

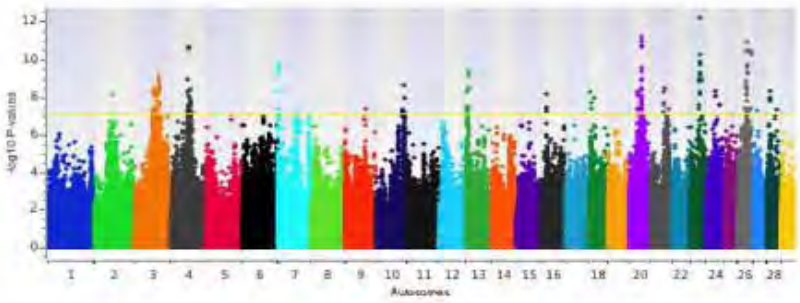
QTLR_ID	Bta	Start	End	Length	No. haploblocks in QTLR	No. SNPs in haploblocks	No. Significant SNPs in QTLR (in haploblock)*	Genes in QTLR
QTLR_1	3	60886435	61043744	157309	2	40	2	*LCT, UBXN4*
QTLR_2	3	67081990	67513292	431302	4	40	19 (10)	*UBP33, ZZZ3, AK5*
QTLR_3	3	80059276	83304366	3205090	4	33	26 (24)	*LEPR, LEPROT, DNAJC6, AK4, MIR101-1, JAK1, UBE2U PGM1, EFCAB7, ITGB3BP, ALG6, ATG4C*
QTLR_4	4	57002763	59933394	2930631	4	25	12 (11)	*ATRAID, RNF167*
QTLR_5	4	60785294	61536212	750918	3	31	9 (6)	*ELMO1, AOAH, MIR2887-1, MIR2887-2, MIR2904-1, MIR2904-2, MIR2904-3, EEPD1*
QTLR_6	4	66822211	67502414	680203	2	10	7 (6)	*SCRN1, PRR15, CHN2*
QTLR_7	4	70856071	70942348	46277	1	15	3	
QTLR_8	7	253267	354904	101637	3	11	10	*LOC100125913*
QTLR_9	7	59978956	60435725	456769	3	22	3	*PPP2R2B*
QTLR_10	10	91425685	95355512	3929827	2	19	13 (10)	*NRXN3, DIO2, TSHR, GTF2A1, STON2, SEL1L*
QTLR_11	13	715178	780158	64980	1	12	6 (4)	*LOC10473686*
QTLR_12	13	6976952	7929838	952886	3	12	13 (9)	*ISM1, TASP1, ESF1, NDUFAF5, SEL1L2, MACROD2*
QTLR_13	18	7792744	7819498	26754	1	2	1	*GCSH*
QTLR_14	18	18368327	18378072	9745	1	6	1	*ZNF423*
QTLR_15	20	29160751	29172374	11623	2	8	12 (8)	*HCN1*
QTLR_16	20	41887047	42965556	1078509	9	39	35 (24)	*PDZD2, DROSHA, CDH6*
QTLR_17	21	42323496	42326629	3133	1	3	1	*LOC104975398*
QTLR_18	21	47559273	47668689	109416	4	22	4	*SLC25A21*
QTLR_19	23	30185638	30193090	7452	1	8	1	*LOC539620*
QTLR_20	24	32537088	32541375	4287	1	3	3	
QTLR_21	24	45456676	45461016	4340	1	3	3	*SLC14A2*
QTLR_22	26	27569344	30875242	3305898	7	79	24 (18)	*SORCS1, XPNPEP1, ADD3*
QTLR_23	26	31464263	33658696	2194433	5	39	11 (6)	*RBM20, PDCD4, MIR6524, MIR4680, SHOC2, ADRA2A, TECTB, ACSL5, ZDHHC6, VTI1A*
QTLR_24	26	41757085	41822044	64959	1	30	7	
QTLR_25	26	46157097	46204880	47783	1	14	3 (2)	*ADAM12*
QTLR_26	28	12769898	12777912	8014	1	4	4	*CHRM3*

*No. Significant SNP up the Bonferroni threshold (<0.05)

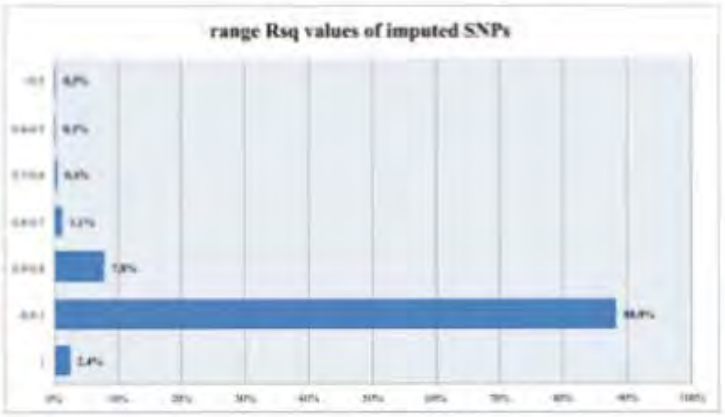


Linkage disequilibrium, Haploblocks and haplotype frequencies calculated for the significantly associated SNPs on BTAs 3, 4 and 26.

881x661mm (72 x 72 DPI)



881x661mm (72 x 72 DPI)



881x661mm (72 x 72 DPI)

TRABAJO 2

GENOME-WIDE ASSOCIATION STUDY FOR MILK SOMATIC CELL SCORE IN HOLSTEIN CATTLE USING COPY NUMBER VARIATION AS MARKERS

Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers

M. Durán Aguilar[1], S.I. Román Ponce[2], F.J. Ruiz López[2], E. González Padilla[3], C.G. Vásquez Peláez[3], A. Bagnato[4] & M.G. Strillacci[4]

1 Facultad de Estudios Superiores Cuautitlán, UNAM, Cuautitlán Izcalli, México
2 Centro Nacional de Investigación en Fisiología y Mejoramiento Animal, INIFAP, Auchitlán Querétaro, México
3 Departamento de Genética y Bioestadística, Facultad Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, México DF, México
4 Department of Veterinary Medicine (DiMeVet), University of Milan, Milan, Italy

Keywords
Copy number variants; GWAS; holstein; SNP; somatic cell score.

Correspondence
A. Bagnato, Department of Veterinary Medicine, Università degli Studi di Milano, Via Celoria 10, 20133 Milano, Italy.
Tel: +39 0250315740;
Fax: +39-02-50315746;
E-mail: alessandro.bagnato@unimi.it

Received: 23 February 2016;
accepted: 3 August 2016

Summary

Mastitis, the most common and expensive disease in dairy cows, implies significant losses in the dairy industry worldwide. Many efforts have been made to improve genetic mastitis resistance in dairy populations, but low heritability of this trait made this process not as effective as desired. The purpose of this study was to identify genomic regions explaining genetic variation of somatic cell count using copy number variations (CNVs) as markers in the Holstein population, genotyped with the Illumina BovineHD BeadChip. We found 24 and 47 copy number variation regions significantly associated with estimated breeding values for somatic cell score (SCS_EBVs) using SVS 8.3.1 and PennCNV-CNVRuler software, respectively. The association analysis performed with these two software allowed the identification of 18 candidate genes (*TERT, NOTCH1, SLC6A3, CLPTM1L, PPARx, BCL-2, ABO, VAV2, CACNA1S, TRAF2, RELA, ELF3, DBH, CDK5, NF2, FASN, EWSR1* and *MAP3K11*) that result classified in the same functional cluster. These genes are also part of two gene networks, whose genes share the 'stress', 'cell death', 'inflammation' and 'immune response' GO terms. Combining CNV detection/association analysis based on two different algorithms helps towards a more complete identification of genes linked to phenotypic variation of the somatic cell count.

Introduction

The most common and expensive disease in dairy cows affecting the mammary gland is the mastitis, an inflammation caused by pathogens. Because of the increased milk production and veterinary treatments, clinical mastitis cases can cost up to $200 per case, with a total estimated cost to the U.S. dairy industry of approximately $1.7–2 million dollars/year (Cha *et al.* 2011).

The susceptibility of the bovine mammary gland to mastitis largely depends on the involution process of the mammary gland tissue and on the exposure to different physiological, genetic and environmental factors (Sordillo & Streicher 2002).

Many efforts have been made to improve the genetic immune resistance to mastitis in dairy cows. The results are still very limited and obtained mainly through correlated traits as somatic cell count (SCC). The milk SCC, in fact, can be used as a predictor of mastitis susceptibility, as a moderate-to-high genetic correlations have been reported between clinical mastitis and SCC or its log transformation in somatic cell score (SCS) (Hinrichs *et al.* 2005).

In the recent past, a large number of studies have mapped QTL affecting mastitis SCC and SCS and are

  doi:10.1111/jbg.12238

reported in the QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index).

The availability of dense single-nucleotide polymorphism (SNP) arrays facilitates the identification of genomic regions associated with economically important traits in farm animals, thus allowing to better disclose QTLs and genetic variation for mastitis resistance. Recently, a class of structural variants, the copy number variants (CNVs), have been suggested as markers of genomic variation in complex disease (Redon *et al.* 2006). CNVs, in fact, are a genomic structural variation that, as SNP, is considered as an important marker of heritable genetic expression (Kijas *et al.* 2011).

Copy number variants are distributed over the whole genome in humans, domestic animals and other species; they are defined as large-scale genome mutations ranging from 50 bp to several Mb compared with a reference genome, which are presented as insertions, deletions and more complex changes (Mills *et al.* 2011). Although SNPs are more frequent, CNVs involve larger genomic regions that may affect gene structure and possibly determining a change in its expression and regulation (Hou *et al.* 2012a).

Several studies have shown that the CNVs are associated with residual feed intake variability in Holstein cows (Hou *et al.* 2012b) and with fertility in Israeli Holsteins (Glick *et al.* 2011). In addition, Xu *et al.* (2014) reported a genomewide CNV association analysis with milk production traits in Holstein, identifying 34 CNVs significantly associated with milk production traits, most of them overlapping known QTL.

Generally, QTL identifying studies are based on a very large number of individuals, as sample size is determinant to achieve reasonable power in a population-wide experimental design. The selective genotyping (Darvasi 1997) is an efficient method to identify chromosomal regions that harbour QTL by comparing marker allele frequencies from phenotypically extreme samples (samples that deviate the most from the mean of the phenotype). This approach allows maintaining the same statistical power for QTL detection, limiting the number of samples to genotype to those in the extreme high and low values, instead of genotyping the whole population as is done in population-wide experimental designs. Several studies based on SNP markers have addressed the feasibility and effectiveness of the selective genotyping method (also combined with the DNA pooling approach), to detect QTL associated with different traits (Fontanesi *et al.* 2007; Strillacci *et al.* 2014).

The use of CNVs as markers, to explain the genetic variation of SCS in milk, has not been explored so far. The purpose of this study was to identify genomic regions explaining the genetic variation of SCS using CNVs in the Holstein population using a selective genotyping approach. Two different algorithms were used to call CNVs in order to provide a cross-integration and validation of results and a clearer indication on the size of the CNVs identified.

Materials and methods

Sampling and genotyping

The SCS estimated breeding values (SCS_EBV) were obtained from the Mexican Holstein Association (http://www.holstein.com.mx/QueToro.aspx). To identify individuals with extreme high and low values, the entire database was ranked based on SCS_EBV values (1.38 mean ± 1.14 SD).

A total of 242 samples with available DNA, identified among individuals above and below 2 SD from the SCS_EBV values average, were selected and classified as follows: i) high phenotypes, a total of 102 samples with SCS_EBV mean 3.37 ± 0.523, and ii) low phenotype, a total of 140 samples with mean 1.67 ± 0.719.

SNP chip data, obtained from the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA, USA), were provided by the Genomic Improvement project of INIFAP and the Mexican Holstein Association.

Data editing

The log R ratio (LRR) and the B allele frequency (BAF) values were extracted using the Illumina BeanStudio software v.2.0 (Illumina Inc.). Samples with a call rate below 98% were excluded from the subsequent analyses, which were performed for the 29 autosomes.

The overall distribution of derivative log ratio spread (DLRS) values were evaluated using the SVS 8.3.1 software (Golden Helix Inc., Bozeman, MT, USA) to identify and filter outlier samples, as described by Pinto *et al.* (2011). To normalize the LRR values and then exclude the samples with extreme wave factors from the analysis, we used the wave correction algorithm, which corrects for the waviness contributed by GC content. In addition, batch effects in the LRR were corrected via numeric principal component analysis (PCA).

CNV detection

As suggested by different authors (Pinto *et al.* 2011) and applied in CNV mapping studies (Bagnato *et al.* 2015), because of intrinsic noisiness of CNV analysis, at least two algorithms should be used for the identification of CNVs. This strategy allows the integration and comparison of the CNV detection among different algorithms and may reduce the bias in detection (i.e. false negatives) proper of each algorithm. The possibility to identify false negatives may be relevant especially when running an association analysis between identified CNVs and traits of interest.

Two independent software based on different algorithms were here used to identify CNVs: i) the Copy Number Analysis Module (CNAM) provided by SVS 8.3.1 software (http://goldenhelix.com); ii) the Hidden Markov Model (HMM) by PennCNV software (http://penncnv.openbioinformatics.org/en/latest/).

For CNV calling using SVS 8.3.1 software, LRR values were employed under the univariate approach: this approach segments each sample independently. As suggested in the software manual, the options used were the following: i) univariate outlier removal; ii) maximum number of segments: search for up to 10 per 10 000 markers; iii) a minimum of 3 markers per segment; iv) a significance level of $p = 0.005$ for pairwise permutations ($n = 2000$). After segmentation analysis, all the segments were classified in three categories as losses, gains and neutral.

The individual-based CNV calling, based on LRR and BAF values for every SNP, was performed by PennCNV software using the default parameters of HMM (standard deviation (SD) of LRR <0.30 and BAF drift as 0.01).

CNV association with SCS_EBV

Linear regression in SVS 8.3.1 software was used to identify CNVs (detected by CNAM algorithm) associated with SCS_EBV with significance level of FDR >0.05, after classifying CNV calls in three state covariates, that is loss, neutral, gain (−1, 0, 1).

Instead, results of the CNV calling from PennCNV were utilized to perform an association analysis with SCS_EBV using the CNVRuler software (http://www.ircgp.com/CNVRuler/index.html), after the definition of CNV regions (CNVRs). In this study, the CNVRs were detected by merging overlapping CNVs by at least 1 bp identified across all samples, as described by Redon *et al.* (2006). The association analysis was performed between CNVRs and SCS_EBV, applying a linear regression model, with a minor allele frequency threshold value set to 0.02. The parameter of recurrence, set to 0.1 (default value), was applied to allow a more robust definition of regions. This option checks the density of regions of CNVs and trims the sparse area not satisfying the density threshold of 10%. Additionally, the 'Gain/Loss separated regions' option, which compiles the region based on the genotype (gain or loss of copy number), was applied.

Significant CNVs were detected when their false discovery rate (FDR)-adjusted p-values had a value of $p < 0.05$.

For a graphical visualization of the results, two separated Manhattan plots of associated CNVRs with SCS_EBV were created using the -log10 of the p-values resulting from the association analyses performed by SVS 8.3.1 and PennCNV-CNVRuler software.

Annotation

The full Ensembl v83 gene set (bovine UMD 3.1 assembly) for the autosomes was downloaded (http://www.ensembl.org/biomart/martview/76d1cab099658c68bde77f7daf55117e/).

In order to identify the genes located within the CNVRs, we created a consensus list (among CNVRs and the downloaded genes) using the BedTools software (Quinlan & Hall 2010).

Gene ontology (GO) and pathway analyses were performed using GenCLiP2.0, an online server for functional clustering of genes (http://ci.smu.edu.cn/GenCLiP2.0/analysis.php?random=new) accounting for false discovery rate.

Results and discussion

CNV calling and association analysis with SVS 8.3.1

The CNV detection was performed in 220 stringently quality-filtered samples: i) 88 high-phenotype samples (3.384 ± 0.511) and 132 low-phenotype samples (1.670 ± 0.726).

We identified a total of 5194 CNVs (covered at least by 3 SNPs) (Table 1) distributed on all autosomes, mainly on the BTA 12 ($n = 881$). Among the detected CNVs, the number of losses and gains were 5088 (98%) and 106 (2%) gains, respectively. The number of CNVs in all samples ranges from 11 to 42 (average of 25.12).

Overlapping CNVs across samples were summarized at the population level into 252 CNVRs (11 gains, 236 losses and 5 complex), with 62 singletons and 128

Table 1 Descriptive statistics for CNVs identified with PennCNV and SVS 8.3.1 software

Copy number[a]	Number of CNVs	Mean Length	Min Length	Max Length
SVS 8.3.1				
Loss	5088	318 123	1245	2 760 295
Gain	106	718 633	9245	2 805 791
Total	5194	518 378	1245	2 805 791
PennCNV				
0	2354	6447	1229	602 308
1	8121	5439	1112	1 248 573
3	1566	94 328.5	998	1 185 515
4	29	147 364	4044	724 916
Total	12070	90 138	998	1 248 573

[a]0 = homozygous deletion, 1 heterozygous deletion, 3 heterozygous duplication and 4 homozygous duplication.

CNVRs that comprise at least 5 CNVs (Table S1). CNVRs cover a total of 39.29 Mb of sequence, which corresponds to 1.5% of the Bovine UMD3.1 assembly.

Using a linear regression, 85 CNVs resulted to be associated with SCS_EBV (p-value <0.05 after FDR correction) (Figure 1a).

In order to better delineate the chromosomal regions resulting associated with the trait, the significant CNVs are grouped in 34 CNVR distributed on 17 autosomes, according to Redon *et al.* (2006)'s approach, using the BedTools software. Among those CNVRs, only the ones with CNV frequencies above 2% (CNVRs with at least one CNV identified in five samples) were retained and used to perform the annotation analysis. Based on UMB3.1 sequence assembly, 51 bovine genes were annotated within the significant CNVRs (Table 2).

CNV calling and association analysis with PennCNV

The use of the 'Filtering CNV calls by user-specified criteria' module of PennCNV allowed to identify low-quality samples and to eliminate them from further analysis. Of the 220 stringently quality-filtered samples, we then obtained a subset of samples (n = 124) with a maximum number of CNVs equal to 200: i) 49 high-phenotype samples (3.307 ± 0.385); ii) 74 low-phenotype samples (1.830 ± 0.080). This additional filtering is specifically required according to the PennCNV detection algorithm.

Overall, 12 070 CNVs distributed on all autosomes were then assessed, with an average per sample of 97.33 CNVs (ranging from 42 to 200) (Table 1).

Overlapping CNVs by at least one nucleotide were summarized to 1662 CNVRs (394 gains, 1215 losses and 53 complex), with 844 singletons and 408 CNVRs that comprise at least 5 CNVs (Table S2). The defined CNVRs cover 82.67 Mb of autosomal genome sequences, corresponding to 3.3% of the Bovine UMD3.1 assembly.

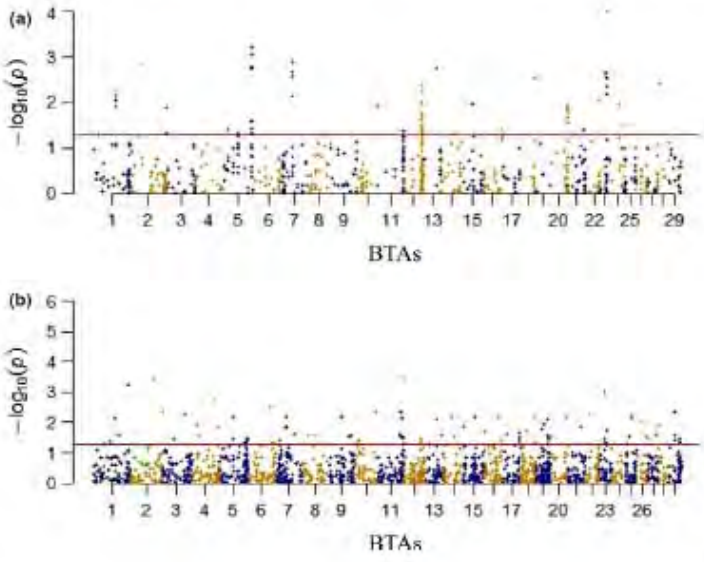


Figure 1 Manhattan plots of associated CNVs for SCS_EBV using SVS 8.3.1. (a) and PennCNV-CNVRuler (b).

Table 2 CNVRs significantly associated with SCS_EBV identified by SVS 8.3.1

CNVR_ID	CHR	START	END	LENGTH	STATE	FREQ	SNP Predictor	p-value	Genes within the significant CNVRs
CNVR_9_SVS	1	93957123	94357120	399997	loss	8	BovineHD0100026648	8,81E-03	
							BovineHD0100026649	7,11E-03	
							BovineHD0100026754	1,25E-02	
CNVR_23_SVS	2	46477034	46485436	8402	Loss	30	BovineHD0200013459	1,44E-03	
CNVR_39_SVS	3	7957960	7964523	6563	Loss	18	BTA-66943-no-rs	4,69E-02	
							BovineHD0300002600	1,27E-02	
CNVR_58_SVS	5	22514133	22563988	49855	Loss	25	BovineHD0500006525	3,94E-02	
CNVR_60_SVS	5	58966295	59255853	289558	Complex	5'	BovineHD0500035991	4,94E-02	
							BovineHD0500036000	5,32E-02	
CNVR_63_SVS	5	117246007	117651752	405745	Complex	179	BovineHD0500034077	4,79E-02	
							BovineHD0500034078	2,59E-02	
							BovineHD0500034128	2,50E-02	
							BovineHD0500034132	3,69E-02	
							BovineHD0500036296	3,63E-02	
							BovineHD0500034144	5,86E-04	
							BovineHD0500034145	8,88E-04	
							BovineHD0500034148	9,09E-04	
							BovineHD0500034150	8,98E-04	
							BovineHD0500034151	1,74E-03	
							BovineHD0500034152	1,80E-03	
							BovineHD0500034153	1,59E-03	
CNVR_83_SVS	7	42745346	42788788	43442	Loss	31	BovineHD0700012438	1,28E-03	OR2AK2
							BovineHD0700012440	2,65E-03	
							BovineHD0700012441	2,10E-03	
							BovineHD0700012444	2,65E-03	
							ARS-BFGL-NGS-23938	7,27E-03	
CNVR_122_SVS	10	87873996	87878635	4639	Loss	11	BovineHD1000024990	1,25E-02	
							ARS-BFGL-NGS-112168	1,13E-02	
CNVR_125_SVS	11	103644879	104195124	550245	Loss	140	BovineHD1100031771	5,78E-02	C9orf69, LHX3, QSOX2,
							BovineHD4100009284	5,39E-02	GPSM1, DNLZ, CARD9,
							BovineHD1100030807	4,21E-02	SNAPC4, SDCCAG3, PMPCA,
							BovineHD1100030845	5,43E-02	INPP5E, SEC16A, NOTCH1, EGFL7, bta-mir-126, AGPAT2, FAM69B
CNVR_134_SVS	12	70363408	72077746	1714338	Complex	159	BovineHD1200019362	9,96E-03	
							BovineHD1200028177	2,22E-02	
							BovineHD1200019797	4,63E-02	

(continued)

After PennCNV-CNVRuler analysis, a total of 47 CNVRs distributed on 18 autosomes were associated (p-value<0.05 after FDR correction) with SCS_EBV (Figure 1b). Table 3 reports the list of the 47 significant CNVRs and the 105 annotated genes.

Comparison of results obtained with SVS 8.3.1 and PennCNV software

To identify the CNVRs that fully overlapped each other among those identified within the two software, the Wain *et al.* (2009)'s approach was used in a BedTool software routine. The consensus CNVR set contained 265 regions.

After association analysis, only six CNVRs resulted associated with SCS_EBV for both analyses. These common regions were located on BTA1 (at 93.95 Mb), on BTA5 (at 58.96 Mb), on BTA5 (at 117.28 Mb), on BTA7 (at 42.73 Mb), on BTA12 (at 74.84 Mb) and on BTA23 (at 28.82 Mb).

The CNVR_11SVS on BTA12 comprised two different associated regions identified by CNVRuler (CNVR_22P and CNVR_23P). In addition, the associated CNVR _9SVS on BTA11 at 103.64–104.19 Mb lies in the proximity (approximately 100Kb) of the associated CNVR_19P.

The overlapping CNVRs between PennCNV and SVS8.3.1 did not contain any functional gene,

Table 3 CNVRs significantly associated with SCS_EBV identified by PennCNV-CNVRuler

CNVR ID	CHR	START	END	LENGTH	STATE	FREQ	pvalue	Genes within the significant CNVRs
CVNR_36_P	1	93954887	94357120	402234	Loss	8	7.20E-03	
CVNR_66_P	1	146975308	147110229	134922	Loss	86	5.50E-04	
CVNR_112_P	2	98480344	98490521	10178	Gain	6	3.46E-04	
CVNR_171_P	3	50167465	50191213	23749	Loss	10	3.70E-02	
CVNR_186_P	3	93310320	93315045	4726	Loss	7	5.71E-03	
CNVR_232_P	4	28744454	28751390	6937	Loss	7	1.14E-02	
CNVR_281_P	4	114419925	114514111	94187	Loss	7	1.43E-02	ABCB8, ASIC3, CDK5, SLC4A2, FASTK, TMUB1, AGAP3
CNVR_324_P	5	58966295	59168921	202627	Gain	6	3.68E-02	
CNVR_347_P	5	107628624	107660101	31478	Loss	12	4.52E-02	IQSEC3, SLC6A12
CNVR_375_P	5	117281795	117639815	358021	Loss	74	4.11E-02	
CNVR_379_P	5	118109413	118174364	64952	Loss	8	3.66E-02	TBC1D22A
CNVR_395_P	5	121149647	121183174	33528	Loss	9	3.60E-02	MOV10L1, bta-mir-2894, PANX2, TRABD
CNVR_471_P	7	15083922	15102276	18355	Loss	10	4.05E-02	
CNVR_502_P	7	42736530	42788788	52259	Loss	29	1.42E-02	OR2AK2
CNVR_503_P	7	42945525	43087430	141906	Loss	13	6.71E-03	OR2AJ1
CNVR_508_P	7	45487894	45538477	50584	Loss	21	1.57E-02	APC2, C19orf25, PCSK4, REEP6
CNVR_653_P	10	16816476	16844526	28051	Loss	3	4.27E-02	
CNVR_658_P	10	23540925	23635452	94528	Loss	3	3.94E-02	
CNVR_765_P	11	104295522	104764859	469338	Loss	76	7.07E-03	ABO, SURF6, MED22, RPL7A, SURF1, SURF2, STKLD1, REXO4, ADAMTS13, CACFD1, SLC2A6, TMEM8C, ADAMTSL2, FAM163B, DBH, SARDH, VAV2
CNVR_770_P	11	106158972	106415916	256945	Loss	103	3.21E-02	UAP1L1, SAPCD2, ENTPD2, NPDC1, FUT7, ABCA2, CLIC3, C9orf142, LCNL1, PTGDS, LCN12, C8G, FBXW5, TRAF2, EDF1, MAMDC4, PHPT1, C9orf172, RABL6, CCDC183, TMEM141, LCN8, LCN15, LCN10
CNVR_777_P	12	717729	731185	13457	Gain	4	3.19E-04	
CNVR_824_P	12	72432362	73015638	583277	Loss	57	3.69E-02	
CNVR_825_P	12	74840021	75238779	398759	Loss	75	4.59E-02	
CNVR_1048_P	16	70814352	71165517	351166	Loss	82	3.86E-02	SMYD2, RNPEP, ELF3, GPR37L1, ARL8A, PTPN7, LGR6, UBE2T, PPP1R12B, SYT2
CNVR_1062_P	17	15677009	15720425	43417	Loss	4	2.54E-02	INPP4B
CNVR_1090_P	17	70714297	70748407	34111	Loss	12	4.48E-02	EWSR1, GAS2L1, RASL10A, AP1B1
CNVR_1091_P	17	70794775	70817022	22248	Loss	7	3.25E-02	
CNVR_1092_P	17	70963787	71024477	60691	Loss	7	2.26E-02	NF2, CABP7, ZMAT5
CNVR_1134_P	18	27914135	28375996	461862	Gain	12	1.86E-02	
CNVR_1192_P	19	24548362	24571149	22788	Gain	4	1.75E-02	
CNVR_1210_P	19	37277118	37328651	51534	Loss	8	1.15E-02	DLX3, DLX4
CNVR_1124_P	19	51028723	51073939	45217	Loss	11	8.13E-03	
CNVR_1126_P	19	51365385	51514295	148911	Loss	10	4.98E-02	FASN, DUS1L, GPS1, RFNG, DCXR, RAC3, LRRC45, STRA13
CNVR_1231_P	19	52776058	52903129	127072	Gain	7	2.65E-02	
CNVR_1236_P	19	54639709	54687169	47461	Loss	15	3.52E-02	TMC8, TMC6
CNVR_1321_P	21	54162719	54196002	33284	Loss	9	1.44E-02	
CNVR_1345_P	22	20291128	20331448	40321	Loss	7	5.32E-03	
CNVR_1398_P	23	21694996	21702537	7542	Loss	5	3.40E-02	
CNVR_1399_P	23	25335659	25361041	25383	Loss	12	1.01E-03	
CNVR_1408_P	23	28828468	28849820	21353	Loss	22	4.64E-02	
CNVR_1417_P	23	34779270	34866601	87332	Gain	3	1.76E-02	
CNVR_1519_P	26	23347145	23380565	33421	Loss	7	1.00E-02	
CNVR_1549_P	26	51434163	51680135	245973	Loss	38	3.30E-02	IAKMIP3, DPYSL4, STK32C, LRRC27, PWWP2B
CNVR_1584_P	28	2263677	2271424	7748	Loss	4	1.21E-02	
CNVR_1617_P	29	27363231	27409510	46280	Loss	14	3.78E-02	
CNVR_1640_P	29	44416282	44502548	86267	Loss	14	4.69E-02	SSSCA1, FAM89B, EHBP1L1, KCNK7, MAP3K11, PCNXL3, SIPA1, RELA
CNVR_1648_P	29	47039694	47054342	14649	Loss	3	3.66E-02	TPCN2

identified CNVRs were not previously reported and may be thus population specific or not yet detected. A further evidence that significant CNVRs are true regions comes from the number of individuals defining them, spanning from 80 to 140 for SVS 8.3.1 and from 7 to 103 for PennCNV-CNVRuler.

We compared the identified associated CNVRs with the reported cattle QTL in the Animal QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index). Among the associated CNVRs, seven (SVS 8.3.1) and ten (CNVRuler) are regions overlapping the mapped QTL for SCS or for Mastitis, as reported for both software in Table 4 (clinical mastitis as CM; somatic cell count as SCC).

The Literature Mining Gene Network tool (provided by GenCLiP2.0), which searches for genes linked to keywords based on up-to-date literature profiling, revealed that 14 genes included within the significant CNVRs and two flanking genes (*BCL-2*, *PPARα*) have been associated mainly with the keywords 'stress', 'cell death', 'inflammation' and 'immune response', as reported in Figure 2.

The GO analysis performed for the gene included in the Figure 2 revealed that they are clustered into 19 groups of genes that were involved in a variety of cellular functions such as cell death, programmed cell death, tissue and organ development, and so on (Table S4 and Figure 3).

KEGG pathway analysis showed the involvement of several signal pathways, such as immune response, apoptosis and adipocytes signalling (Table S5 and Figure 4).

The annotation analyses has enabled the identification of genes encoding for proteins that may be involved in the phenotypic variation in the SCS_EBV and consequently in the mastitis resistance.

In particular, the association analysis performed with the SVS 8.3.1 allowed the identification of 7 candidate genes (*TERT*, *NOTCH1*, *SLC6A3*, *CLPTM1L*, *CACNA1S*, *PPARα* and *BCL-2*), while 11 candidate genes were found to be associated with CNV identified with PennCNV-CNVRuler analysis (*ABO*, *TRAF2*, *RELA*, *ELF3*, *DBH*, *CDK5*, *NF2*, *FASN*, *EWSR1*, *VAV2* and *MAP3K11*). Details on genes included in the networks and their function are included in Appendix S1 file.

Conclusions

The selective genotyping approach here used revealed to be efficient in identifying CNVs in the population and in associating them with the SCS_EBVs. The strategy here adopted to report CNVs mapped through the

Table 4 QTL mapped within significant CNVRs

CNVR_ID	Chr	Start	End	Length	Start_QTL	End_QTL	QTL trait_id
Significant CNVR_SVS 8.3.1							
CNVR_23_SVS	2	46477034	46485436	8402	45424584	52384967	CM (DYD) QTL #19007, QTL #19004, QTL #19005, QTL #19006
CNVR_83_SVS	7	42745346	42788788	43442	27358606	42831622	SCS QTL #2667
CNVR_140_SVS	13	53858853	53862891	4038	51062875	56847265	SCS QTL #2775
CNVR_213_SVS	23	25869447	26337243	467796	23274081	31653997	SCS QTL #2688
CNVR_214_SVS	23	28448873	28469826	20953	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CNVR_217_SVS	23	28828468	28849820	21352	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CNVR_240_SVS	28	10760635	10774825	14190	10665897	11438802	SCS QTL #16056
Significant CNVR_PennCNV							
CVNR_502_P	7	42736530	42788788	52259	27358606	42831622	SCS QTL #2667
CNVR_503_P	7	42945525	43087430	141906	42834942	50547685	SCC QTL #2698
CNVR_508_P	7	45487894	45538477	50584	42834942	50547685	SCC QTL #2698
CNVR_658_P	10	23540925	23635452	94528	22939631	40797089	SCC QTL #2701
CVNR_1134_P	18	27914135	28375996	461862	27863715	33011652	SCC QTL #4638
CVNR_1226_P	19	51365385	51514295	148911	51395368	51495967	SCS (DYD) QTL #32265
CVNR_1398_P	23	21694996	21702537	7542	21554613	22522198	SCS QTL #19986, #19991
CVNR_1399_P	23	25329895	25417035	87140	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CVNR_1408_P	23	28828468	28849820	21353	23274081	31653997	SCS QTL #2688
					27452360	31104253	SCS QTL #4989
CVNR_1648_P	29	47039694	47054342	14649	46178647	52998234	CM (DYD) QTL #19031

use of two different algorithms (CNAM and HMM) successfully reduced the false negative (and positives) that may be identified by only one approach.

Finally, this study is the first GWAS for SCS based on CNVs in Holstein cattle breed. Combining the CNV detection/association analysis using two software

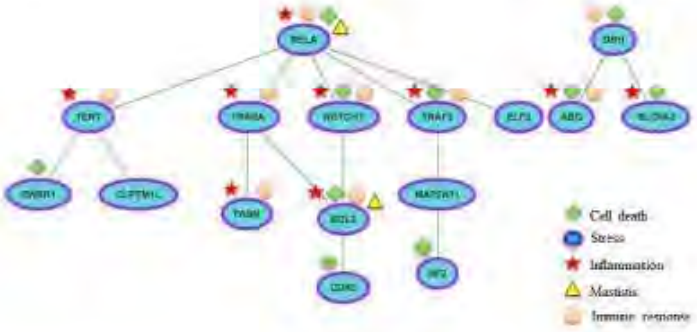


Figure 2 Candidate gene network.

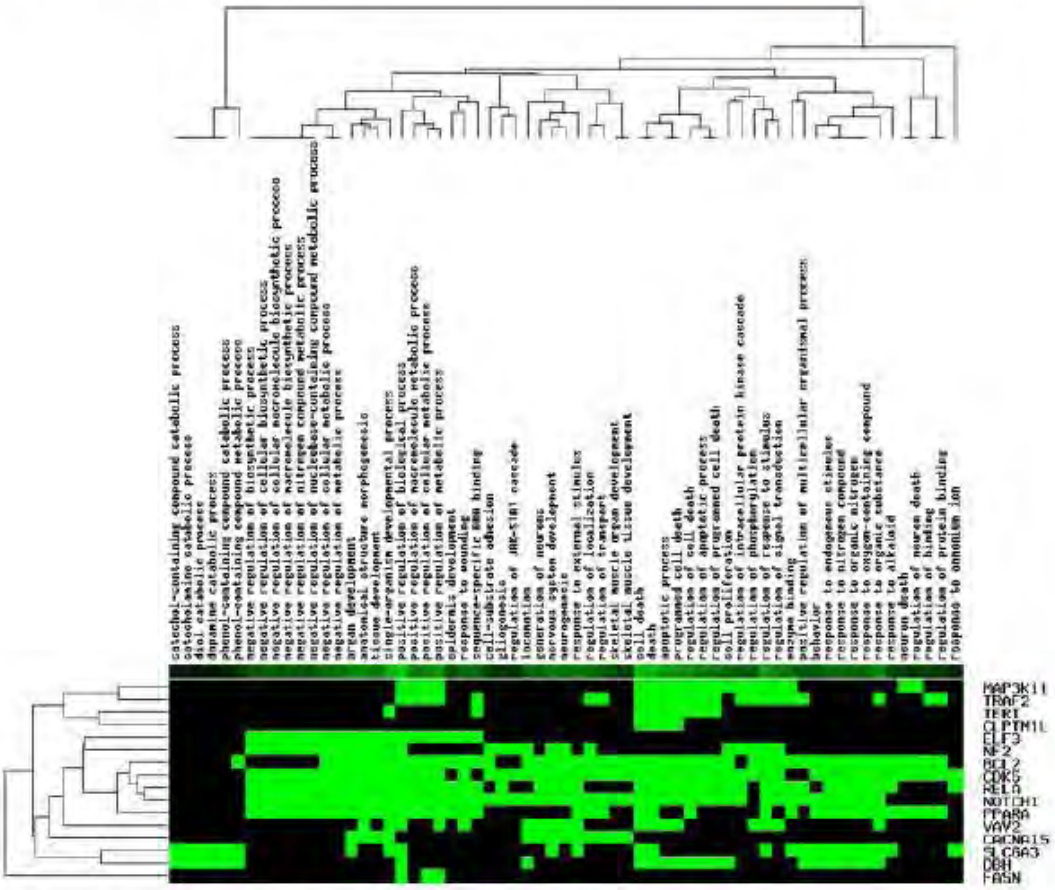


Figure 3 Cluster results of GO analysis for all genes included in significant CNVR (both software)

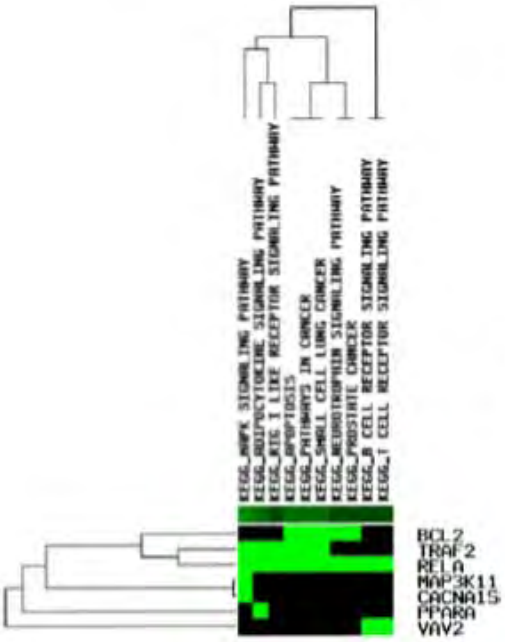


Figure 4 Cluster results for pathway analysis (both software).

allows a more complete identification of genes linked to phenotypic variation of the SCS trait, compared with those revealed using only one software.

Competing interests

The authors declare they have no competing interests.

Acknowledgements

We acknowledge the Council of Dairy Cattle Breeding and Holstein de Mexico A.C. for sharing genomic data. This work received funding from CONACYT, México (project 'IDENTIFICACIÓN DE VARIACIONES ESTRUCTURALES Y REGIONES GENÓMICAS ASOCIADAS A RESISTENCIA A MASTITIS EN BOVINOS PRODUCTORES DE LECHE', 176039).

References

Bagnato A., Strillacci M.G., Pellegrino L., Schiavini F., Frigo E., Rossoni A., Fontanesi L., Maltecca C., Prinsen R.T.M.M., Dolezal M.A. (2015) Identification and validation of copy number variants in Italian Brown Swiss dairy cattle using Illumina Bovine SNP50 Beadchip. *Ital. J. Anim. Sci.*, **14**, 2015.

Cha E., Bar D., Hertl J.A., Tauer L.W., Bennett G., González R.N., Schukken Y.H., Welcome F.L., Gröhn Y.T. (2011) The cost and management of different types of clinical mastitis in dairy cows estimated by dynamic programming. *J. Dairy Sci.*, **94**, 4476–4487.

Choi J.W., Chung W.H., Lim K.S., Lim W.J., Choi B.H., Lee S.H., Kim H.C., Lee S.S., Cho E.S., Lee K.T., Kim N., Kim J.D., Kim J.B., Chai H.H., Cho Y.M., Kim T.H., Lim D. (2016) Copy number variations in Hanwoo and Yanbian cattle genomes using the massively parallel sequencing data. *Gene*, **589**, 36–42.

Darvasi A. (1997) The effect of selective genotyping on QTL mapping accuracy. *Mamm. Genome*, **8**, 67–68.

Fontanesi L., Scotti E., Dolezal M., Lipkin E., Dall'Olio S., Zambonelli P., Bigi D., Davoli R., Soller M., Russo V. (2007) Bovine chromosome 20: milk production QTL and candidate gene analysis in the Italian Holstein-Friesian breed. Proceedings of the 17th ASPA Congress, Alghero, May 29-June 1, 2007, pp. 133–135.

Glick G., Shirak A., Seroussi E., Zeron Y., Ezra E., Weller J.I., Ron M. (2011) Fine mapping of a QTL for fertility on BTA7 and its association with a CNV in the Israeli Holsteins. *G3 (Bethesda)*, **1**, 65–74.

Hinrichs D., Stamer E., Junge W., Kalm E. (2005) Genetic Analyses of Mastitis data using animal Threshold models and genetic correlation with production traits. *J. Dairy Sci.*, **88**, 2260–2268.

Hou Y., Liu G.E., Bickhart D.M., Matukumalli L.K., Li C., Song J., Gasbarre L.C., Van Tassell C.P., Sonstegard T.S. (2012a) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics*, **12**, 81–92.

Hou Y., Bickhart D.M., Chung H., Hutchison J.L., Norman H.D., Connor E.E., Liu G.E. (2012b) Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. *Funct. Integr. Genomics*, **12**, 717–723.

Jiang L., Jiang J., Yang J., Liu X., Wang J., Wang H., Ding X., Liu J., Zhang Q. (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genom.*, **14**, 131.

Kijas J.W., Barendse W., Barris W., Harrison B., McCulloch R., McWilliam S., Whan V. (2011) Analysis of copy number variants in the cattle genome. *Gene*, **482**, 73–77.

Mills R.E., Walter K., Stewart C., Handsaker R.E., Chen K., Alkan C., Abyzov A., Yoon S.C., Ye K., Cheetham R.K., Chinwalla A., Conrad D.F., Fu Y., Grubert F., Hajirasouliha I., Hormozdiari F., Iakoucheva L.M., Iqbal Z., Kang S., Kidd J.M., Konkel M.K., Korn J., Khurana E., Kural D., Lam H.Y., Leng J., Li R., Li Y., Lin C.Y., Luo R., Mu X.J., Nemesh J., Peckham H.E., Rausch T., Scally A., Shi

X., Stromberg M.P., Stütz A.M., Urban A.E., Walker J.A., Wu J., Zhang Y., Zhang Z.D., Batzer M.A., Ding L., Marth G.T., McVean G., Sebat J., Snyder M., Wang J., Ye K., Eichler E.E., Gerstein M.B., Hurles M.E., Lee C., McCarroll S.A., Korbel J.O. 1000 Genomes Project. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Pinto D., Darvishi K., Shi X., Rajan D., Rigler D., Fitzgerald T., Lionel A.C., Thiruvahindrapuram B., Macdonald J.R., Mills R., Prasad A., Noonan K., Gribble S., Prigmore E., Donahoe P.K., Smith R.S., Park J.H., Hurles M.E., Carter N.P., Lee C., Scherer S.W., Feuk L. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, **29**, 512–520.

Quinlan A.R., Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Flegler H., Shapero M.H., Carson A.R., Chen W. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sasaki S., Watanabe T., Nishimura S., Sugimoto Y. (2016) Genome-wide identification of copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. *BMC Genet.*, **17**(1), 26.

Sordillo L.M., Streicher K.L. (2002) Mammary gland immunity and mastitis susceptibility. *J. Mammary Gland Biol. Neoplasia.*, **7**, 135–146.

Strillacci M.G., Frigo E., Canavesi F., Ungar Y., Schiavini F., Zaniboni L., Reghenzani L., Cozzi M.C., Samoré A.B., Kashi Y., Shimoni E., Tal-Stein R., Soller M., Lipkin E., Bagnato A. (2014) Quantitative trait loci mapping for conjugated linoleic acid, vaccenic acid and Δ(9)-desaturase in Italian Brown Swiss dairy cattle using selective DNA pooling. *Anim. Genet.*, **45**, 485–499.

Wain L.V., Armour J.A.L., Tobin M.D. (2009) Genomic copy number variation, human health, and disease. *Lancet*, **374**, 340–350.

Wieczorek D., Pawlik B., Li Y., Akarsu N.A., Caliebe A., May K.J., Schweiger B., Vargas F.R., Balci S., Gillessen-Kaesbach G., Wollnik B. (2010) A specific mutation in the distant sonic hedgehog (SHH) cis-regulator (ZRS) causes Werner mesomelic syndrome (WMS) while complete ZRS duplications underlie Haas type polysyndactyly and preaxial polydactyly (PPD) with or without triphalangeal thumb. *Hum Mutat.*, **31**, 81–89.

Wu Y., Fan H., Jing S., Xia J., Chen Y., Zhang L., Gao X., Li J., Gao H., Ren H. (2015) A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in Simmental cattle. *Anim. Genet.*, **46**, 289–298.

Xu L., Cole J.B., Bickhart D.M., Hou Y., Song J., VanRaden P.M., Sonstegard T.S., Van Tassell C.P., Liu G.E. (2014) Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genom.*, **15**, 683.

Xu L., Hou Y., Bickhart D.M., Zhou Y., Hay E.H.A., Song J., Sonstegard T.S., Van Tassell C.P., Liu G.E. (2016) Population-genetic properties of differentiated copy number variations in cattle. *Sci. Rep.*, **6**, 23161.

Zhang Q., Ma Y., Wang X., Zhang Y., Zhao X. (2015) Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. *Mol. Genet. Genomics*, **290**, 319–327.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 CNVRs identified by SVS 8.3.1 software.

Table S2 CNVRs identified by PennCNV software.

Table S3 Comparison of significantly associated CNVR found in this study with those identified in other researches.

Table S4 GO analysis results of candidate genes.

Table S5 KEGG pathway results of candidate genes.

Appendix S1 Candidate genes details and References.

DETECCIÓN DE CNV *DE NOVO*

INTRODUCCIÓN

Scharpf *et al.* (2012) definieron las CNV *De* Novo como los CNV en los hijos, que son diferentes en los padres y que tienen un papel funcional en las enfermedades.

Las CNV raras (<1%) en esquizofrenia, están relacionados con efectos importantes y tienden a ser removidas rápidamente de la población debido a selección natural, acorde con esto, tales CNV ocurren como mutaciones recurrentes llamadas *De Novo* (Kirov *et al.,* 2012).

Las CNV *De Novo* han resultado ser de interés en el sistema de tríos por tener un papel en el origen de las enfermedades (Scharpf *et al.,* 2012). Hay estudios que sugieren que las CNV *De Novo* pueden ser más informativas que las CNV (Kirov *et al.,* 2012).

Un estudio sobre esquizofrenia mostró que las CNV *De Novo* en pacientes sin historia familiar fue 8 veces más alta en casos que en controles (Kirov *et al.,* 2012).

Existen pocos métodos estadísticos para la detección de CNV *De Novo* (Scharpf *et al.,* 2012). Actualmente, solo se han realizado estudios de CNV *De Novo* en humanos y éstos no se encuentran muy bien caracterizados debido a limitaciones en el tamaño de las muestras y la fuente del material (Itsara *et al.* 2010).

Considerando la importancia potencial de los CNV De Novo en la explicación de las resistencia a la mastitis, se incluyó un apartado especial para investigar de manera preliminar la existencia de CNV De Novo en este proceso inflamatorio utilizando información genealógica (no siempre disponible en humanos) y con el uso del algoritmo HMMC (software PennCNV), ya que provee inferencia directa en las alteraciones *De Novo* en estudios con tríos, esto para identificar las alteraciones de CNV presentes en las crías y que difieren de los padres.

METODOLOGÍA

Se utilizó la información de 100 tríos (conformados por padre, madre e hija), 50 de ellos correspondientes al 20% con los valores más altos SCS (EBV_SCS) y los otros 50 correspondientes al 20% más bajo.

Después de los controles de calidad, permanecieron 47 tríos de la cola baja y 41 de la cola alta.

La estructura familiar puede utilizarse para generar CNV más precisas ya que podemos correlacionar la información de la CNV de miembros de la familia relacionados que tienen muchas probabilidades de compartir la misma región de la CNV (http://penncnv.openbioinformatics.org/en/latest/user-guide/denovo/).

Utilizando el software PennCNV (HMMC) con la opción "tríos" (que incluye la información de genealógica) se detectaron los CNV para cada trio para, posteriormente, utilizando únicamente los tríos en los que se detectaron CNV con la strata correspondientes a padre y madre normales con deleciones doble o sencilla en el hijo o con padre y madre normales con inserciones sencilla o doble en el hijo (códigos PennCNV 331, 332, 334 y 335) realizar el análisis con el *script* "*De Novo"* con la finalidad detectar los CNV que efectivamente sean De Novo.

RESULTADOS

Los análisis antes mencionados nos permitieron encontrar 8 CNV *De Novo* en 3 tríos diferentes que resultaron significativos en este trabajo.

Cuadro 3: CNV *De Novo*

Marcadores	Paternal_orig	Maternal_orig	P-value	TRIO
1108	235	0	3.62E-71	HOLUSAM000125997863 HOLMEXF000660300434 HOLMEXF000660314867
434	54	0	1.11E-16	HOLUSAM000125997863 HOLMEXF000660314788 HOLMEXF000660314989
134	38	0	7.28E-12	HOLUSAM000125997863 HOLMEXF000660314788 HOLMEXF000660314989
114	24	0	1.19E-07	HOLUSAM000125997863 HOLMEXF000660300434 HOLMEXF000660314867
147	23	0	2.38E-07	HOLUSAM000002297473 HOLMEXF000660300440 HOLMEXF000660314964
93	9	0	0.00390625	HOLUSAM000125997863 HOLMEXF000660314788 HOLMEXF000660314989
64	9	0	0.00390625	HOLUSAM000125997863 HOLMEXF000660300434 HOLMEXF000660314867
60	8	0	0.0078125	HOLUSAM000125997863 HOLMEXF000660314788 HOLMEXF000660314989

CONCLUSIONES

El resultado de estos análisis muestra que existen CNV *De Novo* en la población Holstein de México, sin embargo, esta información debe validarse por medio de PCR en tiempo real porque es posible que existan errores en la genotipificación que nos induzcan a falsos positivos. La validación no fue objetivo de esta tesis por lo que se deja para trabajos posteriores, pero era importante mencionar que es posible que existan estas estructuras en ganado Holstein en México.

BIBLIOGRAFÍA

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. *De Novo* rates and selection of large copy number variation. Genome Res. 20: 1469-1481.

Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, Moran J, Chambert K, Toncheva D, Georgieva L, Grozeva D, Fjodorova M, Wollerton R, Reesm E, Nikolov I, van de Lagemaat LN, Baye A, Fernandez E, Olason PL, Böttcher Y, Komiyama NH, Collins MO, Choudhary J, Stefansson K, Stefansson H, Grant SGN, Purcell S, Sklar P, O'Donovan MC, Owen MJ. 2012. *De Novo* CNV analysis implicates specific abnormalities of postsynaptic signaling complexes in the pathogenesis of schizophrenia.

Scharpf RB, Beaty TH, Schwender H, Younkin SG, Scott AF, Ruczinski I. 2012. Fast detection of de novo copy number variants from SNP arrays for case-parent trios. *BMC Bioinformatics* 13:330.

DISCUSIÓN GENERAL

Strillacci *et al*. (2014) encontraron 171 SNP relacionados significativamente con SCS, distribuidos en 24 cromosomas y de estos, 52 SNP se encontraron dentro de 36 genes; Meredith *et al*., (2013) encontraron 138 SNP asociados en 15 cromosomas. Para este trabajo se reportan 286 SNP distribuidos en 16 cromosomas, de los cuales 154 se encuentran en 50 genes, esta diferencia puede ser debida a la plataforma utilizada, ya que los autores citados utilizaron más animales, pero menos SNP.

Este trabajo definió como región cromosómica (CNVR) asociada a SCS y mastitis clínica aquellas que contenían al menos 3 SNP asociados a SCS.

Con el objeto de simplificar las comparaciones de los resultados de este trabajo con otros que reportaron SNP asociados a SCS, se detallarán los SNP significativos y los genes asociados reportados por otros autores, por cromosoma, en la inteligencia de que no ha habido ningún trabajo anterior que reporte CNVR asociadas a SCS,

Cromosoma 1

En este trabajo se encontraron (utilizando el modelo univariado (software SVS8)) una CNVR asociada a SCS que se traslapa con una de las dos CNVR encontradas con el algoritmo HMMC (software PennCNV); sin embargo, ninguno de los trabajos reportó alguna asociación de SNP para esta característica.

Cromosoma 2

Al igual que en el cromosoma 1, ningún trabajo reportó SNP asociados a SCS y en este trabajo se encontraron (utilizando el modelo univariado) una CNVR asociada a SCS y una CNVR con el algoritmo HMMC.

Cromosoma 3

Los genes AK5, NEGR1, LEPR y ATG4C son los que se reportaron en ente estudio asociados a SCS con los SNP BovineHD0300019924, BovineHD0300021327, BovineHD0300023057 y BovineHD0300023843 respectivamente.

Es importante mencionar que el gen *ATG4C* (*autophagy related 4C cysteine peptidase*), es miembro de la familia de las cistein proteinasas que se cree están involucradas en el proceso de autofagia (Fernández and López-Otín, 2015) y junto con otros genes (i.e. *IL23R*) parece estar involucrada en la respuesta inmune contra algunas enfermedades infecciosas como las causadas por *M. tuberculosis* (Daya *et al.,* 2015).

En este cromosoma se encontró una CNVR asociada a SCS identificada con el modelo univariado y 2 CNVR con HMMC.

Cromosoma 4

Este trabajo encontró que los SNP BovineHD0400016562, BovineHD0400016630, BovineHD0400016851, BovineHD0400017071, BovineHD0400017147 y BovineHD0400018332 están asociados a los genes ELMO1, AOAH, EEPD1, DPY19L1, NPSR1 y SCRN1.

Reviste importancia considerar que Los SNP que el gen *ELMO1* (*engulfment and cell motility 1*), se califica como un miembro de la familia de proteínas de motilidad celular. Estas proteínas interactúan con las proteínas citoquinesis, promueve la fagocitosis y la migración celular. Estudios encuentran que el rol de este gen es facilitar el reconocimiento intracelular de bacterias y la inducción de la respuesta inflamatoria seguida de la enfermedad infecciosa, lo que se relaciona con el proceso inflamatorio en la glándula mamaria (Das *et al.,* 2015).

A diferencia de los cromosomas anteriores, solo HMMC logró identificar dos CNVR asociadas a SCS.

Cromosoma 5

Aunque no existen informes previos de SNP asociados a SCS, en este trabajo se encontraron (utilizando el modelo univariado) tres CNVR asociada a QTL para SCS, de las cuales, 2 se traslapan con dos de las 5 CNVR identificadas con el HMMC.

Cromosoma 6

Al igual que Strillacci *et al.* (2014), en este trabajo no se encontró asociación con ningún SNP o CNVR.

Sin embargo, otros trabajos han reportado SNP asociados a SCS en ganado Holstein, tales como Sahana *et al*. (2013) quienes en un estudio sobre la confirmación y mapeo fino de mastitis clínica y QTL para SCS en ganado Holstein Nórdico y usando BovineSNP50 BeadChip, encontraron el número más alto de asociaciones significativas en el cromosoma 6, identificando un QTL para mastitis clínica a 83.37-88.89 Mb (UMD3.1 assembly); resultado que fue confirmado por Abdel-Shafy *et al.* (2014) en Ganado Holstein de Alemania y por Meredith *et al*. (2013).

Cromosoma 7

A diferencia de la literatura revisada, este trabajo encontró asociación del SNP BovineHD0700017213 que se encuentra dentro del gen PPP2R2B además de que se identificaron (utilizando el modelo univariado) una CNVR asociada a un QTL para SCS, misma que se traslapa con una de las 4 CNVR asociadas a QTL para SCS encontradas con HMMC.

Cromosoma 9

Aunque en este trabajo no se encontraron SNP o CNVR asociadas a SCS, se ha reportado que en el cromosoma 9 se encuentran los genes VNN1 y VNN2 que están relacionados con la resistencia a mastitis. Estos genes han sido asociados a su vez a los SNP BovineHD0900019961 (rs136413030), BovineHD0900019716 (rs109049649), BovineHD4100007550 (rs41662465) y Hapmap49339- BTA-84110 (rs41662464). Estos SNP

también fueron mapeados cerca del gen CTGF (connective tissue growth factor) (Jiang et al. 2012).

Adicionalmente, Lund *et al*. (2008) encontraron un QTL identificado para resistencia general a enfermedades (incluyendo mastitis clínica) y para SCS en este cromosoma.

Cromosoma 10

En este trabajo se encontraron (utilizando el modelo univariado) una CNVR asociada a SCS y dos CNVR con el algoritmo HMMC; sin embargo, ninguno de los trabajos reportó alguna asociación de SNP para esta característica.

Cromosoma 11

En este trabajo se encontraron (utilizando el modelo univariado) una CNVR asociada a SCS que se traslapa con una de las dos CNVR identificadas con el algoritmo HMMC; sin embargo, ninguno de los trabajos reportó alguna asociación de SNP para esta característica.

Cromosoma 12

Tres CNVR fueron identificadas con el modelo univariado asociadas a QTL para SCS, una de las cuales se traslapa con una de las 3 CNVR encontradas con HMMC, sin embargo, ningún trabajo revisado (incluyendo el nuestro con SNP) encontró asociación a SCS en este cromosoma

Cromosoma 13

Cuatro SNP (BovineHD4100010442 (rs41634068), BovineHD1300022626 (rs137320993), BovineHD1300022630 (rs109123247) y BovineHD1300022672 (rs41710487)) se encontraron cerca del gen ZNFX1 (X1-type zinc finger-containing) localizado en el cromosoma 13 (Strillacci et al. 2014).

Los genes ZNFX1, CTGF, TRIM21, CXCL2 y CXCL10 son expresados diferencialmente por las células epiteliales de la glándula mamaria, simulados con el polisacárido crudo de *E. coli* (Gilbert *et al*., 2013).

Jensen *et al*. (2013) estudiaron y compararon la respuesta transcripcional de los cuartos de la glándula mamaria no infectada adyacentes a cuartos infectados con *E. coli* y *S. aureus* en vacas Holstein. El CXCL2 resulto ser uno de los genes expresados diferencialmente entre los cuartos infectados con ambos patógenos, el CXCL10 es uno de los genes expresados diferencialmente en el control de los cuartos infectados con *S. aureus*, a 24 y 72 horas.

A diferencia de los estudios anteriores, los SNP que resultaron asociados en nuestro estudio fueron BovineHD1300001809, BovineHD1300002031, BovineHD1300002090 que corresponden a los genes ISM1, SEL1L2 y MACROD2 respectivamente.

Solo se identificó una CNVR con el modelo univariado asociada a un QTL para SCS.

Cromosoma 15

En este cromosoma los SNP BovineHD1500008135 (rs134980659), BovineHD1500008366 (rs41754552) y el SNP BovineHD1500008367 (rs110269361) se localizaron cerca del gen THY1 (Thy-1 cell surface antigen) (Strillacci et al. 2014).

THY1 es uno de los genes expresados diferencialmente entre los cuartos de las vacas infectadas con *E. coli* y *S. aureus* (Jensen *et al*., 2013). También Moyes *et al*. (2009) reportaron la THY1 sobre regulación en infecciones intramamarias con *S. uberis*.

Strillacci *et al*. (2014) encontraron que el número más elevado de SNP significativos asociados a SCS (14) fue en el cromosoma 15. Estos autores reportaron una región asociada a SCS localizada a 50.43- 51.63 Mb que no había sido reportada anteriormente en ganado (http://www.animalgenome.org/cgi-bin/QTLdb/index)

Este trabajo no encontró SNP ni CNVR asociados a SCS en este cromosoma.

Cromosoma 16

Una CNVR fue identificada con el modelo univariado asociada a QTL para SCS y una con HMMC, sin embargo, ningún trabajo revisado (incluyendo el nuestro con SNP) encontró asociación a SCS en este cromosoma.

Cromosoma 17

No se identificaron SNP asociados a SCS pero si se detectaron cuatro CNVR con HMMC asociadas a QTL para SCS.

Lo anterior, en contraste con lo reportado por Wang *et al.* (2012), quienes encontraron un SNP (BovineHD1700018352 (rs135157738)) asociado al gen que codifica para la deshidratasa serina (SDS). Este gen está incluido en el metabolismo de la glicina, serina y trionina, ha sido demostrado que la SDS es una de las enzimas que cambia significativamente en bovinos afectados con mastitis.

Cromosoma 18

En este trabajo se encontraron (utilizando el modelo univariado) una CNVR asociada a SCS y una CNVR con el algoritmo HMMC; sin embargo, ninguno de los trabajos reportó alguna asociación de SNP para esta característica.

Cromosoma 19

En este cromosoma se encuentra el gen SOCS3 (supresor del señalador de citocina) a 673,863 pb rio arriba del SNP BovineHD1900015066 (rs132720248). Este gen es importante para la homeostasis de la glándula mamaria, codifica para un inhibidor intracelular de citocinas, de esta manera, juega un papel importante en los pasos iniciales del reconocimiento de patógenos asociados a patrones moleculares (PAMP) de las células de inmunidad innata. Esto permite la iniciación y activación de la respuesta inmune innata y adaptativa (Heeg y Dalpke, 2003; Brenaut *et al*., 2014)

Seis CNVR que no incluyen a este gen fueron identificadas con el HMMC asociadas a QTL para SCS.

Cromosoma 20

Tiezzi *et al.,* (2015) Encontraron regiones genómicas asociadas a mastitis clínica, resultados que concuerdan con los encontrados en este trabajo, ya que detectamos 48 SNP (BovineHD2000008576, BovineHD2000008577, BovineHD2000008581,

BovineHD4100014625, BovineHD2000008582, BovineHD2000008583, BovineHD2000008584, BTA-90616-no-rs, BovineHD2000008585, BovineHD2000008590, BovineHD2000008592, BovineHD2000008708, BovineHD4100014715, BovineHD4100014716, BovineHD2000012031, BovineHD2000012032, BovineHD2000012036, BovineHD2000012039, BovineHD2000012045, BovineHD2000012047, BovineHD2000012163, BovineHD2000012167, BovineHD2000012168, BovineHD2000012180, BovineHD2000012221, BovineHD2000012222, BovineHD2000012223, BovineHD4100014721, BovineHD2000012224, BovineHD2000012225, BovineHD2000012256, BovineHD2000012284, BovineHD2000012285, BovineHD2000012286, BovineHD2000012287, BovineHD2000012288, BovineHD2000012289, BovineHD2000012290, BovineHD2000012291, BovineHD2000012294, BovineHD2000012295, BovineHD2000012296, BovineHD2000012297, BovineHD2000012298, BovineHD2000012299, BovineHD2000012301, BovineHD2000012305 y BovineHD2000012329) dentro de 3 genes.

Una CNVR fue identificada con el modelo univariado asociadas a QTL para SCS; Sin embargo, al utilizar el modelo HMMC no fue posible identificar ninguna CNVR

Cromosoma 21

Cuatro SNP (BovineHD2100012950, BovineHD2100013201, BovineHD2100013563 y BovineHD2100017704) se encontraron asociados a 4 genes (LOC104975398, KIAA0391, SLC25A21 y LOC101904984) en el presente trabajo.

Una CNVR fue identificada con el modelo univariado asociadas a QTL para SCS y una con HMMC.

Otro estudio (Sugimoto y Sugimoto 2012) asociaron al SNP BovineHD2100001405 (rs133992914) (no identificado en este estudio) con el gen IGF1R (insulin-like growth factor 1 receptor) que está involucrado con la inmunidad innata en bovinos. La región 5’UTR de IGF1R se asoció a mastitis en Bos taurus.

Cromosoma 22

Este trabajo identificó a una CNVR con el algoritmo univariado asociadas a QTL para SCS y una con HMMC, la primera asociada a los genes CHCHD6, TXNRD3, C3orf22, CHST13, UROC1, ZXDC, SLC41A3, ALDH1L1, KLF15, CCDC37.

Similar al presente estudio, Lund *et al*. (2008) Encontraron un QTL asociado a SCS localizado a 32.62-43.31 Mb. En este cromosoma se encontró el gen *cholinergic receptor muscarinic 3* (*CHRM3*), mapeado a 4 SNP significativos. Este gen es expresado en macrófagos, una de las células clave en el proceso inflamatorio y que contribuye a su quimiotaxis (Selivanova *et al.,* 2012). *CHRM3* también es uno de los genes que se ve afectado por el tratamiento de calor de las células epiteliales de la glándula mamaria bovina, según lo reportado por Li *et al*., (2015).

Otros estudios han reportado otros marcadores asociados a diferentes genes en este cromosoma. El SNP BovineHD2200003506 (rs110821186), se mapeo cerca del gen MYD88 (myeloid differentiation primary- response gene 88) a 11.72Mb, el cual juega un papel funcional en la transducción pro-inflamatoria de la molécula del lipopolisacarido (LPS) que es la responsable de la mayoría de los casos clínicos agudos de mastitis (Cates EA, *et al*., 2010). Estas regiones están localizadas en QTL que fueron mapeados, respectivamente, para mastitis clínica usando un análisis de ligamiento para SCS (Lund *et al*., 2008; Rupp 2003).

Cromosoma 23

A diferencia de otros estudios que no encontraron asociaciones en este cromosoma, este trabajo asoció a los SNP BovineHD4100016123, BovineHD2300008669, BovineHD2300008704, BovineHD2300009361, BovineHD2300009846, BovineHD2300009851, BovineHD2300009875, ARS-BFGL-NGS-6851, BovineHD2300009906, BovineHD2300009921, BovineHD2300015308, BovineHD2300009937, BovineHD2300009942, BovineHD2300009956 y BovineHD2300011292 con 5 genes (ZSCAN12, LOC539620, LOC104975673, LRRC16A, RNF144B).

Tres CNVR fueron identificadas con el modelo univariado asociadas a QTL para SCS, una de las cuales es exactamente la misma a una de las 4 CNVR encontradas con HMMC.

Cromosoma 24

En este trabajo se encontraron los SNP BovineHD2400008906, BovineHD2400008908, BTA-57840-no-rs, BovineHD2400012443, BovineHD2400012444 y BovineHD2400012445 asociados al gen SLC25A21.

Con el modelo univariado se pudieron identificar dos CNVR a QTL para SCS; Sin embargo, al utilizar el modelo HMMC no fue posible identificar ninguna CNVR, es importante mencionar que ningún otro trabajo revisado reportó asociación a esta característica.

Cromosoma 26

Longeri *et al.* (2008) también encontraron SNP asociados a SCC en Holstein.

A diferencia de lo reportado anteriormente, en este cromosoma no se identificaron CNVR con el modelo univariado, pero dos CNVR fueron identificadas con el HMMC asociadas a QTL para SCS.

Cromosoma 28

Una CNVR fue identificada con el modelo univariado asociadas a QTL para SCS y una con HMMC. Sin embargo, ningún otro trabajo revisado reportó asociación de SNP a esta característica.

Cromosoma 29

Al igual que en el cromosoma 26, no se reporta ninguna CNVR con el modelo univariado, sin embargo, tres CNVR fueron identificadas con el HMMC asociadas a QTL para SCS. Ningún otro trabajo revisado reportó asociación de SNP a esta característica.

CNV ENCONTRADOS

Aunque otros trabajos han reportado un mayor número de CNV como, Prinsen *et al*. (2016), quienes, utilizando la información de 1104 bovinos de la raza Suizo Pardo (nuestra muestra total fue de 242 bovinos Holstein); identificaron CNV utilizando los mismos algoritmos (Univariado (SVS8) y HMMC (PennCNV)), encontraron 398 CNVR cubriendo 3.92% del genoma con el modelo univariado (nuestro trabajo reportó 252 CNVR correspondiente al 1.5% del genoma), mientras que con HMMC identificó 5,578 CNVR (1,662 CNVR cubriendo el 3.3% del genoma), cubriendo el 7.68% del genoma. El tamaño de los CNV va de 1,244pb a 1,381.355. El consenso detectó 563 CNVR en común; un total de 775 genes fueron identificados, de los cuales 537 tienen información funcional.

Al comparar los resultados antes mencionados con los presentados en este trabajo, se podría inferir que no solo la plataforma es importante al momento de realizar los análisis, este trabajo reporta únicamente los CNV asociados a SCS, a diferencia de (Princen *et al*., 2016) que reporta todos los CNV encontrados en su estudio; con el modelo univariado encontramos 85 CNVR asociados a SCS y con HMMC 47 CNVR, éstas fueron concordantes para ambos algoritmos en los cromosomas 1,2,3,5,7,10,11,12,16,18,21 y 23; SVS también detectó en los cromosomas 13, 20, 22, 24 y 28, mientras que PennCNV detectó en los cromosomas 4, 17,19,26,28 y 29.

El presente trabajo muestra resultados importantes de asociaciones a genes relacionados con la inmunidad de los animales, esto nos podría llevar a poder hacer selección en un futuro hacia animales que contengan CNV asociados con las características de inmunidad, no solo relacionadas a mastitis.

Como sabemos, el mejoramiento genético para características productivas ha tenido un desarrollo exponencial en las últimas décadas, esto con la ayuda de la inclusión de la información genómica a las evaluaciones, sin embargo, la parte de inmunidad y algunas otras características están cobrando importancia dado que se vuelven un costo importante para el productor de leche y un problema de salud pública, de aquí la importancia del

desarrollo de técnicas que aporten información para la selección de animales potencialmente resistentes a las enfermedades más importantes en la producción

CNV ASOCIADOS A SCS

El análisis que realizamos, permitió la asociación de CNVR a 18 genes candidatos (TERT, NOTCH1, SLC6A3, CLPTM1L, PPARa, BCL-2, ABO, VAV2, CACNA1S, TRAF2, RELA, ELF3, DBH, CDK5, NF2, FASN, EWSR1 and MAP3K11)

La selección asistida por marcadores involucra elegir individuos basados en su genotipo en un locus específico. Este enfoque es conocido por ser particularmente benéfico cuando las características de interés son costosas o difíciles de medir (Boichard *et al*., 2006).

La selección genómica (GS), definida por (Meuwissen *et al*., 2001), es una herramienta importante para el mejoramiento genético del ganado que permite estimar los valores genómicos directos (DGV) de candidatos utilizando mapas de marcadores sin necesidad de registros fenotípicos de los animales (Meuwissen *et al*., 2001).

El QTL se relaciona con susceptibilidad/resistencia a enfermedades infecciosas y con las variaciones de las características productivas, está caracterizado por heterogeneidad genética y herencia multifactorial, involucrando genes polimórficos de diferentes vías alternativas. Con la accesibilidad del arreglo de SNP, el análisis de ligamiento y el estudio de Asociación (GWAS) han sido utilizados frecuentemente para estudios del componente genético en características complejas, y utilizando los SNP se han podido identificar los CNV.

Este trabajo ha mostrado que los genes encontrados tienen pocas coincidencias con los encontrados en otras poblaciones, esto puede ser debido a los factores medio ambientales, y al hecho de que estamos utilizando la información de hembras y de machos (y en la mayoría de los artículos revisados utilizan solo sementales), por lo que podríamos considerar que los análisis deben realizarse para poblaciones definidas dado que no podemos suponer que los resultados serán similares para diferentes poblaciones.

CONCLUSIONES GENERALES

Las regiones genómicas identificadas en este estudio, contribuyen a un mejor entendimiento de la parte genética relacionada a la respuesta inmune de la mastitis, en ganado Holstein, y podría permitir hacer más eficientes los programas de mejoramiento.

Este estudio mostró evidencia de asociación entre SNP y SCS, así como CNV y SCS en varios cromosomas.

Las regiones genómicas identificadas en este estudio pueden ayudar a identificar genes candidatos potencialmente responsables de la resistencia y/o susceptibilidad a mastitis.

Los CNV identificados en este estudio constituyen el primer mapa de CNVR asociados a SCS en ganado, proveyendo información nueva para los siguientes estudios de asociación con características de interés económico y de salud.

FINANCIAMIENTO

Este proyecto forma parte del proyecto número 176039 “Identificación de variaciones estructurales y regiones genómicas asociadas a resistencia a mastitis en bovinos productores de leche” aprobado para su financiamiento en la convocatoria 2012-01 del Fondo SEP – CONACYT.

También se incluyen los proyectos: “Estudio de la diversidad genética de ganado Holstein determinada con base en información genómica en México.” No. Sinaso: 1523542158 y el de “Incorporación de información genómica a los procesos de evaluación genética del ganado productor de leche.” No. Sinaso: 1056821832

BIBLIOGRAFÍA GENERAL

Abdel-Shafy H, Bortfeldt RH, Reissmann M, Brockmann GA. 2014. Short Communication : Validation of Somatic Cell Score-Associated Loci Identified in a Genome- Wide Association Study in German Holstein Cattle. *J Dairy Sci* 97(4):2481-24866.

Abdel-Shafy H, BortfeldtRH, Tetens J, Brockmann GA. 2014. Single Nucleotide Polymorphism and Haplotype Effects Associated with Somatic Cell Score in German Holstein Cattle. *Genetics Selection Evolution* 46(1): 35.

Bae JS, Cheong HS, Kim LH, Gung SN, Park TJ, Chung JY, Kim JY, Pasaje CFA, Lee JS, Shin HD. 2010. Identification of Copy Number Variations and Common Deletion Polymorphisms in Cattle. *BMC genomics* 11: 232.

Ben Sassi N, González-Recio O, de Paz-Del Rio R, Rodríguez-Ramilo ST, Fernández AI. 2016. Associated Effects of Copy Number Variants on Economically Important Traits in Spanish Holstein Dairy Cattle. *Journal of Dairy Science* 99(8): 6371–80.

Bloemhof S, de Jong G, de Haas Y. 2009. Genetic Parameters for Clinical Mastitis in the First Three Lactations of Dutch Holstein Cattle. *Veterinary Microbiol.* 134(1–2): 165–71.

Brookes AJ. 1999. The Essence of SNPs. Gene 234: 177–86.

Bush WS, Jason HM. 2012. Genome-Wide Association Studies. *PLoS Computational Biology* 8(12).

Carlén EE, Strandberg M, Roth A. 2004. Genetic Parameters for Clinical Mastitis, Somatic Cell Score, and Production in the First Three Lactations of Swedish Holstein Cows. *J Dairy Sci* 87(9): 3062–3070.

Cates EA, Connor EE, Mosser DM, and Bannerman DD. 2010. Functional Characterization of Bovine TIRAP and MyD88 in Mediating Bacterial Lipopolysaccharide-Induced Endothelial NF- κB Activation and Apoptosis. *Comp Immunol Microbiol Infect Dis* 32(6): 1–14.

Fadista JB, Thomsen L, Holm E, Bendixen C. 2010. Copy Number Variation in the Bovine Genome. *BMC genomics* 11: 284-288.

Fontanesi L, Matelli PL, Beretti F, Riggio V, Dall"Olio S, Colombo M, Casadio M, Russo V,

Portolano B. 2010. An Initial Comparative Map of Copy Number Variations in the Goat (Capra Hircus) Genome. *BMC genomics* 11(1): 639.

Forer L, Schonherr S, Weissensteiner H, Haider F, Kluckner T, Gieger C, Wichmann HE, Specht G, Konenberg F, Kloss-Brandstatter A. 2010. CONAN: Copy Number Variation Analysis Software for Genome-Wide Association Studies. *BMC bioinformatics* 11: 318-323.

Gilbert FB, Cunha P, Jensen K, Glass GF, Foucres G, Robert-Granie C, Rupp R, Rainerd P,. 2013. Staphylococcus Aureus or Escherichia Coli Agonists of the Innate Immune System Differential Response of Bovine Mammary Epithelial Cells to Staphylococcus Aureus or Escherichia Coli Agonists of Th.44:40.

de Haas Y, Ouweltjes N, ten Napal J, Windig JJ, de Jong G. 2008. Alternative Somatic Cell Count Traits as Mastitis Indicators for Genetic Selection. *J Dairy Sci* 91(6): 2501–2511.

Hagnestam-Nielsen CU, Emanuelson B, Berglund B, Strandberg E. 2009. Relationship between Somatic Cell Count and Milk Yield in Different Stages of Lactation. *J Dairy Sci* 92(7): 3124–33.

Heringstad B, Gianola D, Chang YM, Odegard J, Klemestsdal G. 2006. Genetic Associations between Clinical Mastitis and Somatic Cell Score in Early First-Lactation Cows. *J Dairy Sci* 89(6): 2236–2244.

Heringstad B, Klemetsdal G, Ruane J. 2000. Selection for Mastitis Resistance in Dairy Cattle: A Review with Focus on the Situation in the Nordic Countries. *Livestock Production Science* 64(2–3): 95–106.

Hou Y, Liu GE, Bickhart DM, Carfdone MF, Wang K, Kim E, Matukumalli LK, Ventura M, Song J, VanRaden PM, Sonstegard TS, van Tassell CP. 2011. Genomic characteristics of cattle copy number variations. *BMC genomics* 12(1): 127.

Hou Y, Bickhart, Chung H, Hutchinson JL, Norman HD, Connor EE, Liu GE. 2012. Analysis of Copy Number Variations in Holstein Cows Identify Potential Mechanisms Contributing to Differences in Residual Feed Intake. *Functional & Integrative Genomics* 12 (4): 717–723.

Hou, Y, Bickhart DM, Hvinden, LM, Li C, Song J, Boichard DA, Fritz S, Eggen A, DeNise S,

Wiganns GR, Sonstegard TS, Van Tessell CP, Liu GE. 2012. Fine Mapping of Copy Number Variations on Two Cattle Genome Assemblies Using High Density SNP Array. *BMC genomics* 13: 376.

Hunter D J, Kraft P. 2007. Drinking from the Fire Hose--Statistical Issues in Genomewide Association Studies. *The New England journal of medicine* 357: 436–39.

Iniesta R, Guinó E, Moreno V. 2005. Análisis Estadístico de Polimorfismos Genéticos En Estudios Epidemiológicos. Gaceta Sanitaria 19(4): 333–341.

Jensen K, Günther J, Talbot R, Petzl W, Zerbe H, Schuberth HJ, Seyfert HM, Glass EJ. 2013. Escherichia Coli- and Staphylococcus Aureus - Induced Mastitis Differentially Modulate Transcriptional Responses in Neighbouring Uninfected Bovine Mammary Gland Quarters. *BMC Genomics*. 16;14:36.

Jiang L, Sørensen P, Thomsen B, Edwards SM, Skarman A, Røntved CM, Lund MS, Workman CT. 2012. Gene Prioritization for Livestock Diseases by Data Integration. *Physiol Genomics* 1;44(5): 305–17.

Knap PW, Bishop SC. 2000. Relationships between Genetic Change and Infectious Disease in Domestic Livestock. *Occ. Publ. Br. Soc. Anim. Sci.* 27: 65–80

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, Gasbarre LC, Lacalandra G, Li RW, Matukumalli LK, Nonneman D, Regitano LC, Smith TP, Song J, Sonstegard TS, Van Tassell CP, Ventura M, Eichler EE, McDaneld TG, Keele JW. 2010. Analysis of Copy Number Variations among Diverse Cattle Breeds Analysis of Copy Number Variations among Diverse Cattle Breeds.*Genome Res.* 20(5): 693–703.

Lund MS, Guldbrandtsen B, Buitenhuis AJ, Thomsen B, Bendixen C. 2008. Detection of Quantitative Trait Loci in Danish Holstein Cattle Affecting Clinical Mastitis, Somatic Cell Score, Udder Conformation Traits, and Assessment of Associated Effects on Milk Yield. *Journal of dairy science* 91(10): 4028–36.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, Van Tassell CP. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *Plos One;* 4(4).

McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Dias Neto E, Gill CA, Gao C, Mannen H, Stothard P, Wang Z, Van Tassell CP, Williams JL, Taylor JF, Moore SS. 2007. Whole Genome Linkage Disequilibrium Maps in Cattle. *BMC genetics* (10);8: 74.

Meredith BK, Berry DP, Kearney F, Finlay EK, Fahey AG, Bradley DG, Lynn DJ. 2013. A Genome-Wide Association Study for Somatic Cell Score Using the Illumina High-Density Bovine Beadchip Identifies Several Novel QTL Potentially Related to Mastitis Susceptibility. *Frontiers in Genetics* 4(11): 1–10.

Meuwissen TH, HayesBJ, Goddard ME. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157(4):1819-29.

Pearson TA, Manolio TA. 2008. How to Interpret a Genome-Wide Association Study. *Jama* 299(11): 1335–44.

Peiffer D, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. 2006. High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping. *Genome Research* 16(9): 1136–48

Prinsen RTMM, Strillacci MG, Schiavini F, Santus E, Rossoni A, Maurer V, Bieber A, Gredler B, Dolezal M, Bagnato A. 2016. A Genome-Wide Scan of Copy Number Variants Using High-Density SNPs in Brown Swiss Dairy Cattle. *Livestock Science* 191(9): 153–60

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global Variation in Copy Number in the Human Genome. *Nature* 444(7118): 444–54.

Salinas PH, Albornoz VJ, Reyes PA, Erazo BM, Ide VR. 2006. Analisis de Componentes Principales. In *Revista Chilena de Obstetricia Y Ginecolog*, 71 (1);17-25.

Scherer A. 2014a. *Clinical Next-Gen Sequencing Analysis*. Golden Helix Inc

Scherer A. 2014b. *GWAS Golden Helix, Inc*

Schukken YH, Günther J, Fitzpatrick J, Fontaine MC, Goetze L, Holst O, Leigh J, Petzl W, Schuberth HJ, Spika A, Smith DG, Quesnell R, Watts J, Yancey R, Zerbe H, Gurjar A, Zadoks RN, Seyfert HM. 2011. Host-Response Patterns of Intramammary Infections in Dairy Cows. *Veterinary Immunology and Immunopathology* 144(3–4): 270–89

Schutz MM. 1994. Genetic Evaluation of Somatic Cell Scores for United States Dairy Cattle. *J. Dairy Sci.* 77(7): 2113–29

Sender G, Korwin-Kossakowska A, Pawlik A, Hamed GA Oprzadek J. 2013. Genetic Basis of Mastitis Resistance in Dairy Cattle – A Review / Podstawy Genetyczne Odporności Krów Mlecznych Na Zapalenie Wymienia – Artykuł Przeglądowy. *Annals of Animal Science* 13(4): 663–73.

Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, Ezra E, Zeron Y. 2010. Analysis of Copy Loss and Gain Variations in Holstein Cattle Autosomes Using BeadChip SNPs. *BMC genomics* 11(1): 673

Sordillo LM, Streicher KL. 2002. Mammary Gland Immunity and Mastitis Susceptibility. *Journal Mammary Gland Biology and Neoplasia* 7(2):135-46.

Strillacci MG, Frigo E, Schiavini F, Samoré AB, Canavesi F, Vevey M, Cozzi MC, Soller M, Lipkin E, Bagnato A. 2014. Genome-Wide Association Study for Somatic Cell Score in Valdostana Red Pied Cattle Breed Using Pooled DNA. *BMC Genetics* 15(1): 106.

Sugimoto M, Sugimoto Y. 2012. Variant in the 5' Untranslated Region of Insulin-Like Growth Factor 1 Receptor Is Associated With Susceptibility to Mastitis in Cattle. G3 (Bethesda) 2(9): 1077–84.

Tiezzi F, Parker-Gaddis KL, Cole B, Clays JS, Malteca C. 2015. A Genome-Wide Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. *PLoS ONE* 10(2): 1–15.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, BucanM. 2007. PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data. *Genome Res.* 17(11): 1665–74.

Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N. 2010. An Initial Map of Chromosomal Segmental Copy Number Variations in the Chicken. *BMC genomics* 11: 351.

Zhao X, Lacasse P. 2008. Mammary Tissue Damage during Bovine Mastitis: Causes and Control. *Journal of anim sci* 86(13 Suppl): 57–65