UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA

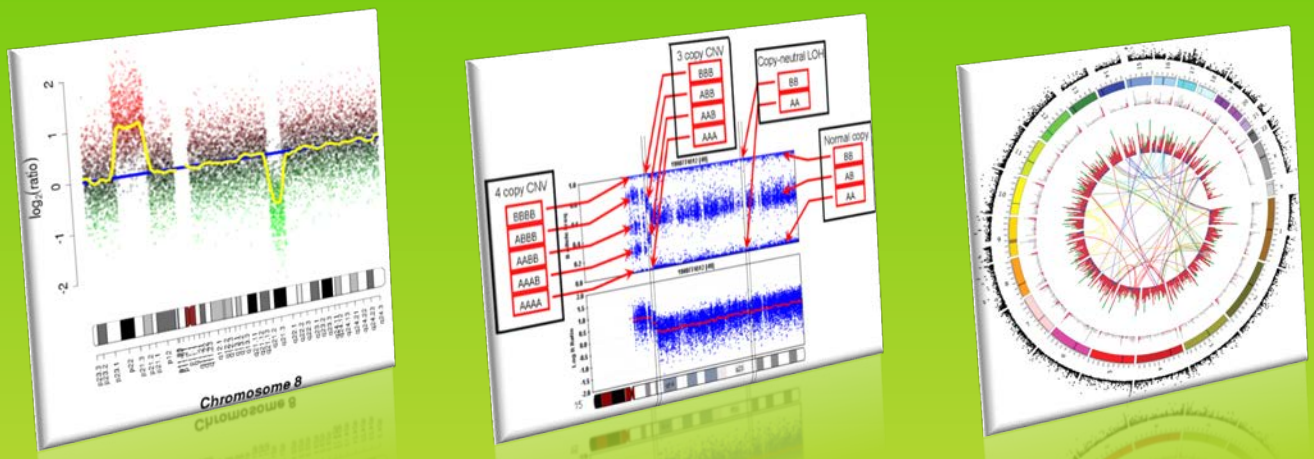
INSTITUTO DE INGENIERIA

Maestría y Doctorado en Ciencias e Ingeniería (MyDCI)

Algoritmos Bioinformaticos para la Detección de Variaciones Estructurales en el Genoma Completo del Ganado Bovino, Utilizando SNPs de Alta Densidad.

Tesis para obtener el grado de:

DOCTOR EN CIENCIAS



Presenta:

RICARDO SALOMON TORRES

Director de Tesis:

DR. RAFAEL VILLA ANGULO

Mexicali, B.C., Agosto del 2014

Dedicatoria

Para Sandra, mi amada Esposa.
(Proverbios 31:10-31)

Agradecimientos

Salmo 23

El Señor es mi pastor; nada me faltará.
En lugares de delicados pastos me hará descansar;
Junto a aguas de reposo me pastoreará.
Confortará mi alma;
Me guiará por sendas de justicia
por amor de su nombre.

Aunque ande en valle de sombra de muerte,
no temeré mal alguno, por tú estarás conmigo;
Tú vara y tu cayado me infundirán aliento.
Aderezas mesa delante de mí en presencia
de mis angustiadores;

Unges mi cabeza con aceite; mi copa está rebosando. Ciertamente el bien y la misericordia me
seguirán todos los días de mi vida,
y en la casa de Jehová moraré por largos días.

Agradecimientos

A Dios: Por los planes que tienes para mí, planes de bienestar y no de calamidad, a fin de darme un futuro y una esperanza. (*Jeremías 29:11*).- Gracias.

A mi Esposa: Por la paciencia y el amor que me has tenido, te amo!.- Gracias.

A mi Familia: Mis padres por enseñarme el buen camino y a mis hermanos por estar siempre conmigo, los amo!- Gracias.

Al Dr. Rafael Villa: Por compartir conmigo su amistad, sus conocimientos, su experiencia y por ser un hombre con gran calidad humana, nunca cambie Dr. Villa.- Gracias.

A mis tutores: Miembros del comité de tesis, por sus observaciones y por avalar este trabajo.

A CONACyT: Por financiar todos mis estudios doctorales.- Gracias.

Al Mayor Cab.DEM. Jorge Contreras: Por haber sido el primer escalón de esta cima.- Gracias.

A todos aquellos: A quienes en su momento fueron de gran apoyo para mí, a quienes de alguna manera directa o indirectamente tuvieron que ver con mi formación, a quienes un día fueron y seguirán siendo mis amigos.- Gracias.

Resumen

I. Identificación de CNVs. en el Genoma bovino.

Las variaciones del número de copias (CNVs) son una gran fuente de variación estructural del genoma. Estas pueden ser consideradas como marcadores que indican rasgos económicos y fenotípicos significativos, que tienen efectos funcionales sobre la expresión génica o la susceptibilidad a enfermedades en los genomas de los mamíferos. La tecnología más utilizada para detectar estas variaciones son los microarreglos de genotipado de polimorfismos de nucleótido simple (SNP chips). Recientemente, el genoma bovino ha sido objeto de muchos estudios sobre las tecnologías de microarreglos. Para este estudio, se inspeccionaron las CNVs en doce vacas de la raza Holstein de México, mediante el microarreglo Affymetrics Axiom Genome-Wide BOS 1 Array, que captura 648,315 SNPs y que proporcionan una gran cobertura para los estudios de genoma completo. Se aplicaron los dos algoritmos más utilizados para el descubrimiento de CNVs: PennCNV y QuantiSNP. Se encontraron un total de 56 regiones de variaciones del número de copias (CNVRs), lo que representa el 0.33% del genoma bovino (8.46Mb). Estas CNVRs varian en tamaño desde 1.5 Kb hasta 970.8 Kb, con un promedio de 151 Kb. Estas involucran a 103 genes, y al compararlas con estudios previos de CNVRs, se traslapan un 28%. De las 56 CNVRs encontradas, 20 son nuevas CNVRs no reportadas anteriormente. En este estudio presentamos el primer análisis genómico de CNVs en ganado Mexicano, utilizando datos de SNPs de alta densidad. Nuestros resultados proporcionan una nueva referencia para futuros estudios sobre la variación y asociación genómica entre CNVs y fenotipos, especialmente en el ganado mexicano.

Resumen

II. Análisis de Variaciones Estructurales en el Ganado Bovino.

Las variaciones estructurales genómicas representan una importante fuente de variación genética en los genomas de mamíferos, por lo tanto, con frecuencia se relacionan con las expresiones fenotípicas. En este trabajo, se analizaron ~770,000 genotipos de polimorfismos de nucleótido simple de 506 animales de 19 razas de ganado. Se definió una simple variación estructural basada en LD, y se realizó un análisis de todo el genoma. Después de aplicar algunos filtros de control de calidad, para cada raza y cada cromosoma se calculó el desequilibrio de ligamiento (r^2) de corto alcance ($\leq$ 100Kb). Ordenamos los pares de SNPs por distancia y obtuvimos un conjunto de medias de LD (llamada la media esperada) utilizando bloques de 5Kb. Se identificaron 15,246 segmentos de al menos 1 Kb, entre las 19 razas, que consisten en conjuntos de al menos 3 SNPs adyacentes de manera que, para cada SNP, la r^2 dentro de sus vecinos en un rango de 100Kb, hacia el lado derecho, fueran todos mayor que, o todos más pequeños que la media esperada correspondiente, y su valor-P fuera significativo después de una corrección de multiples pruebas de Benjamini-Hochberg. Además, para tener en cuenta sólo a las regiones distribuidas homogéneamente hemos considerado sólo SNPs que tienen al menos 15 vecinos SNPs dentro de 100kb. Definimos tales segmentos como variaciones estructurales. Al agrupar todas las variaciones a través de todos los animales en la muestra se definieron 9,146 regiones, con un total de 53,137 SNPs; que representan el 6.40% (160.98 Mb) del genoma bovino. Las variaciones estructurales identificadas cubren 3,109 genes. Un análisis de agrupación demostró la relación de las razas dado a la región geográfica en la que están evolucionando. En resumen, presentamos un análisis de variaciones estructurales en base a la desviación esperada del LD de rango corto entre los SNPs en el genoma bovino. Con una intuitiva y simple simple definición basada únicamente en datos SNPs fue posible discernir la cercanía de las razas debido a la agrupación por región geográfica en la que se están desarrollando.

Abstrac

I. Identification of CNVs in the Bovine Genome.

Copy Number Variations (CNVs) are a great source of genomic structural variation. They can be considered as markers indicating significant phenotypic and economic traits, having functional effects on gene expression or susceptibility to diseases in the mammalian genomes. The most commonly used technology to detect these variations is Single Nucleotide Polymorphism genotyping arrays (SNPchips). Recently, the bovine genome has been the subject of many studies on microarray technologies. For this study we inspected CNVs in twelve Mexican Holstein cows using the Affymetrics Axiom Genome-Wide BOS 1 Array, which assays 648,315 SNPs providing a great coverage for Genome-wide studies. We applied the two most widely used algorithms for the discovery of CNVs: PennCNV and QuantiSNP. We found a total of 56 Copy Number Variation Regions (CNVRs), representing 0.33% of the bovine genome (8.46Mb). These CNVRs range from 1.5 Kb to 970.8 Kb with an average of 151 Kb. They involve 103 genes, and when compared to CNVRs previously reported, they overlap 28%. From the 56 found CNVRs, 20 are novel CNVRs, not reported previously. In this study we present the first genomic analysis of CNVs in Mexican Cattle, using high density SNPs data. Our results provide new reference for future genomic variation and association studies between CNVs and phenotypes, especially in Mexican Cattle.

Abstrac

II. Structural Variations Analysis in Cattle Genome.

Genomic structural variations represent an important source of genetic variation in mammal genomes, thus, they are commonly related to phenotypic expressions. In this work, ~770,000 single nucleotide polymorphism genotypes from 506 animals from 19 cattle breeds were analyzed. A simple LD-based structural variation was defined, and a genome-wide analysis was performed. After applying some quality control filters, for each breed and each chromosome we calculated the linkage disequilibrium (r^2) of short range (≤100Kb). We sorted SNP pairs by distance and obtained a set of LD means (called the expected means) using bins of 5Kb. We identified 15,246 segments of at least 1 Kb, among the 19 breeds, consisting of sets of at least 3 adjacent SNPs so that, for each SNP, r^2 within its neighbors in a 100Kb range, to the right side of that SNP, were all bigger than, or all smaller than, the corresponding expected mean, and their *P*-value were significant after a Benjamini-Hochberg multiple testing correction. In addition, to account just for homogeneously distributed regions we considered only SNPs having at least 15 SNP neighbors within 100kb. We defined such segments as structural variations. By grouping all variations across all animals in the sample we defined 9,146 regions, involving a total of 53,137 SNPs; representing the 6.40% (160.98 Mb) from the bovine genome. The identified structural variations covered 3,109 genes. Clustering analysis showed the relatedness of breeds given the geographic region in which they are evolving. In summary, we present an analysis of structural variations based on the deviation of the expected short range LD between SNPs in the bovine genome. With an intuitive and simple definition based only on SNPs data it was possible to discern closeness of breeds due to grouping by geographic region in which they are evolving.

Lista de Tablas

Lista de figuras

Contenido

Contenido

Contenido

Capítulo I

Introducción

1.1 Biología molecular.

La Biología molecular es la disciplina científica que tiene como objetivo el estudio de los procesos que se desarrollan en los seres vivos desde un punto de vista molecular. Esta disciplina concierne principalmente al entendimiento de las interacciones de los diferentes sistemas de la célula, lo que incluye las relaciones entre el ADN con el ARN, la síntesis de proteínas y como todas estas interacciones son reguladas para conseguir un correcto funcionamiento de la célula. Los métodos que emplea esta relativa nueva ciencia son fundamentalmente los mismos que la Biofísica, Bioquímica y la Biología. Utiliza los análisis químicos cuantitativo y cualitativo, conocimientos de la Química Orgánica, Biología de microorganismos y de virus, etc., pero revisten especial importancia los nuevos métodos microanalíticos, tantos físicos como químicos [1] [2].

1.1.1 Conceptos Básicos de Biología Molecular.

1.1.1.1 Las Células.

Todos los organismos complejos y simples se componen por células, que a su vez se pueden integrar por organelos, y estos organelos en moléculas. La teoría celular se articula de tres partes: 1) todos los seres vivos están constituidos por una o más células; 2) las células son las unidades básicas de la estructura y función en un organismo; 3) las células provienen solo de la reproducción de las células existentes.

Existen dos clases básicas de células, las procariotas y las eucariotas. Las células eucariotas se distinguen por su tamaño y los tipos de estructuras internas, u orgánulos que contienen. Las células procariotas están estructuradas de manera más simple y están representadas por bacterias y algas. Ambas contienen una región nuclear que alberga material genético de la célula, pero el de las procariotas, se encuentra en una región mal delimitada que carece de una membrana para separarlo del citoplasma circundante. En cambio las eucariotas poseen una región bien delimitada llamada núcleo. Las moléculas más importantes en las células son las proteínas y los ácidos nucleicos[3].

1.1.1.2 Las Proteínas.

Todos los organismos vivos están compuestos en gran parte de las proteínas. Hay varios tipos de proteínas: 1) las proteínas estructurales que forman parte de una estructura celular, 2) enzimas que catalizan casi todas las reacciones bioquímicas que ocurren dentro de una célula, 3) proteínas reguladoras que controlan la expresión de genes o la actividad de otras proteínas, y 4) las proteínas de transporte, que llevan otras moléculas a través de la membrana celular o en todo el cuerpo. Una proteína está compuesta por cadenas de aminoácidos. Cada aminoácido se organiza alrededor de un átomo central de carbono, conocido como el carbono alfa. Otros componentes de un aminoácido incluyen un grupo amino (NH2), un grupo carboxilo (COOH), y una cadena lateral (grupo R). Es la cadena lateral que distingue a los aminoácidos entre sí. Existen en total veinte aminoácidos y todos los seres vivos están formados por diversas combinaciones de los aminoácidos. Estructuralmente, las proteínas son cadenas polipeptídicas, donde sus aminoácidos están unidos entre sí por enlaces peptídicos. Un enlace peptídico se forma por la unión del extremo carboxilo de un aminoácido en el extremo amino del aminoácido contiguo, con una molécula de agua liberada en el proceso. La secuencia de los residuos en un polipéptido se denomina la estructura primaria de una proteína. Además, las proteínas se pliegan en realidad en tres dimensiones, lo que resulta en estructuras secundarias, terciarias, y cuaternarias. La estructura secundaria de una proteína se forma a través de la tendencia del polipéptido a la bobina o pliegues debido a enlaces de H entre los grupos-R. Estos polipéptidos unidos forman un mayor nivel de embalaje llamado la estructura cuaternaria. La secuencia lineal de residuos (estructura primaria) de

una proteína, determina su estructura tridimensional y ésta, la función de una proteína[4].

1.1.1.3 Ácidos Nucleicos.

Los ácidos nucleicos son los responsables de codificar la información necesaria para producir proteínas y también los responsables de pasar esa receta a las generaciones siguientes. Existen dos tipos de ácido nucleico: 1) Ácido Ribonucleico (ARN) y 2) el Dexorribonucleico (ADN). La secuencia polipeptídica que forma la estructura primaria de una proteína está relacionada directamente con la secuencia de la información en la molécula de ARN, misma que a su vez, es una copia de la información en la molécula del ADN.

Una molécula de ADN consta de dos cadenas de moléculas más simples. Cada hebra tiene un esqueleto que consiste en repeticiones de la misma unidad básica. La unidad básica de ADN está formada por una molécula de azúcar, 2'-desoxirribosa, unido a un residuo de fosfato. Hebras de ADN tienen una orientación determinada por la numeración de los átomos de carbono, que por convención, se inicia en el extremo 5' y termina en el extremo 3'. Una secuencia de ADN de una sola cadena, por lo tanto siempre se escribe en esta canónica dirección 5' – 3'. En el ADN existen cuatro tipos de bases: adenina (A), guanina (G), citosina (C), y timina (T). Bases que tienen dos anillos de átomos de carbono y nitrógeno, como la adenina y la guanina, son llamadas purinas. Las pirimidinas son bases que tienen un anillo de átomos de carbono y nitrogeno, como la citosina y timina. Un fragmento corto de una molécula de ADN, típicamente entre cinco y 50 pares de bases de longitud, se denomina un oligonucleótido.

Las moléculas de ARN son similares a las moléculas de ADN, con las siguientes diferencias de composición y estructurales básicos: 1) el componente de azúcar del ARN es la ribosa en lugar de desoxirribosa; 2) en el ARN, la timina (T) se sustituye por uracilo (U), que también se une con adenina; 3) el ARN no forma una doble hélice. Hélices híbridos de ADN-ARN a veces se producen, o partes de una molécula de ARN pueden unirse a otras partes de la misma molécula por complementariedad. La estructura tridimensional de ARN es mucho más variada que la de ADN[4].

1.1.1.4 Dogma Central de la Biología Molecular.

Las moléculas de ADN son responsables de codificar la información necesaria para construir cada proteína o molécula de ARN que se encuentran en un organismo. En este sentido, el ADN se refiere a veces como "el diseño de la vida." La información fluye a partir del ADN a través de ARN y por lo tanto a la proteína se describe por el así llamado dogma central de la biología molecular, que incluye las cuatro etapas principales siguientes:

(1) La información contenida en el ADN se duplica a través del proceso de replicación.

(2) El ADN dirige la producción de ARN mensajero codificado (ARNm) a través de un proceso llamado transcripción.

(3) En las células eucariotas, el ARNm a continuación, se procesa y migra desde el núcleo hasta el citoplasma de la célula.

(4) En la etapa final del proceso de transferencia de información, ARN mensajero lleva la información codificada a las estructuras de síntesis de proteínas llamadas ribosomas.

1.1.1.5 Los genes y el Código Genético.

En las células eucariotas de un organismo, cada célula contiene más moléculas de ADN. Cada molécula de ADN forma un cromosoma. El juego completo de cromosomas dentro una célula se llama genoma. El número de cromosomas en un genoma es característico de una especie en particular. Esto puede ser en realidad un nombre inapropiado, ya que se ha sugerido que el ADN basura puede de hecho realizar funciones no reconocidas y valiosas. Un gen es una secuencia de ADN que contiene la información necesaria para construir una proteína o una molécula de ARN. Las longitudes de los genes varían, pero los genes humanos normalmente contienen 10,000 pares de bases. Los puntos inicial y final de los genes pueden ser reconocidos por ciertos mecanismos celulares. El mecanismo por el cual los genes especifican la secuencia de aminoácidos en una proteína se denomina código genético. Para ser más específicos, una tripleta de nucleótidos se utiliza para especificar cada aminoácido. Tal tripleta es conocida como codón. Las tripletas de nucleótidos se denotan usando ARN en lugar

de bases de ADN, ya que, es ARN que proporciona el enlace entre el ADN y la síntesis de proteínas.

Considerando los cuatro tipos de bases, el número total de posibles combinaciones de bases dentro de las tripletas de nucleótidos es 64. Sin embargo, estas combinaciones 64 sólo pueden referirse a los veinte aminoácidos que se producen en realidad. Por lo tanto, existe redundancia en la codificación, y varias tripletas diferentes corresponderán al mismo aminoácido. Por ejemplo, tanto AAG y el código de AAA para la lisina. Por otra parte, tres de los posibles codones (UGA, UAG y UAA) no codifican para ningún aminoácido y se utilizan en lugar de señalar el final de un gen. Tal redundancia es en realidad una característica valiosa del código genético, lo que hace que sea más robusto en el caso de los pequeños errores en el proceso de transcripción[4].

1.1.1.6 Transcripción y Expresión de Genes.

La transcripción es el proceso de síntesis de ARN utilizando genes como plantillas. Un gen es expresado cuando a través del proceso de transcripción, su codificación se transfiere a una molécula de ARN. Para iniciar un proceso de transcripción, la doble hélice de ADN se "descomprime", en el sitio promotor de un gen. El sitio de promotor es una región en el lado 5' de la hebra de ADN. Una vez que la doble hélice de ADN se ha abierto en este punto de partida, la hebra de ADN sirve como cadena molde. Una molécula de ARN está constituido por unión de ribonucleótidos complementarios a la cadena molde hasta que se cumpla el codón STOP. El proceso de composición siempre construye moléculas de ARNm desde el extremo 5' al extremo 3'. Este ARN resultante se denomina ARN mensajero (ARNm). Dado que las dos hebras de la hélice de ADN originales eran también complementarias, la nueva molécula de ARNm tendrá la misma secuencia de ribonucleótidos que la cadena de ADN no utilizado, con la base U sustituido por T. La selección de la plantilla a partir de las dos cadenas disponibles en el par de ADN original varía de un gen a gen, como señalado por la ubicación del sitio del promotor para cada gen. Después del proceso de transcripción, el ARNm se transportará a las estructuras celulares llamadas ribosomas para guiar la fabricación de proteínas.

El proceso de transcripción descrito anteriormente es solo para las células procariotas. Para las eucariotas, muchos genes están compuestos de partes alternantes llamadas intrones y exones. Después de la transcripción, los intrones son separados de los exones, lo que significa que sólo los exones participarán en la síntesis de proteínas. Debido a los cambios que resultan a través del corte y empalme de intrones y exones, a la totalidad de genes que se encuentran en el cromosoma se les suele llamar ADN genómico, y las secuencias empalmadas de exones se llaman secuencias codificantes. Puede obtenerse ADN complementario (ADNc) por un proceso de transcripción inversa de ARNm que se transforma de nuevo en ADN.

Dentro de una secuencia de ADN o de ARN, las bases pueden ser analizadas en formas alternativas para generar grupos de codones. Por ejemplo, en la secuencia TAATCGAATGGGC, bases adyacentes podrían agruparse como codones TAA, TCG, AAT, GGG, omitiendo la última C. También sería posible ignorar la inicial T, que producen codones AAT, CGA, ATG, GGC. Sin embargo, otra lectura marco produciría codones ATC, GAA, TGG, a través de la omisión de las dos bases iniciales (TA) y las dos bases finales (GC). Un proceso de lectura selecciona uno de estos enfoques y analiza las bases en una secuencia que comienza con el codón de inicio, contiene un número integral de los codones, y no incluye ningún codón de parada dentro de la secuencia [5].

1.1.1.7 Traducción y Síntesis de la Proteína.

Una vez que el proceso de transcripción ha generado ARNm adecuadamente codificado, se inicia el proceso de traducción que sintetiza proteínas, el cual tiene lugar dentro de las estructuras celulares llamadas ribosomas. Otro tipo de ARN, es el ARN de transferencia (ARNt), hace la conexión entre un codón y el aminoácido correspondiente. Cada molécula de ARNt tiene en un lado, un anticodón que tiene alta afinidad por un codón específico y, en el otro lado, un sitio de unión de aminoácidos que se une fácilmente a su correspondiente aminoácido. El ácido amino unido cae en su lugar justo al lado del aminoácido anterior en la cadena de proteína que se está formando. Una enzima adecuada cataliza entonces la adición de este aminoácido actual a la cadena de proteína, separándola de los ARNt. De esta manera, una proteína se construye parte por parte. Cuando aparece un codón de parada, sin ARNt asociado a este, el

ARN mensajero se libera y puede volver a usarse para sintetizar nuevamente la proteína, o degradarse por mecanismos celulares en ribonucleótidos, que luego serán reciclados para formar otro ARN[4].

1.1.1.8 Genotipo y Fenotipo.

Genomas que pertenecen a la misma especie varían ligeramente de organismo a organismo en un fenómeno conocido como la variación del genoma (o variación genética). Esta variabilidad en los genomas es responsable de la evolución y la diversidad de un organismo. Algunas variaciones del genoma son exclusivas de un organismo, mientras que otros se pasan de generación en generación a través de la reproducción las células. Una célula puede contener un solo conjunto de cromosomas (el estado hoploide) o dos juegos de cromosomas (el estado diploide); en este último caso, cada cromosoma está representado por dos copias. Una excepción es el par de cromosomas sexuales, que dibuja una copia del padre y otra de la madre. Mucha investigación actual se centra en descubrir las relaciones genotipo-fenotipo, como por ejemplo las que se denominan relaciones entre los enfermedades y su base genómica. El genotipo se refiere al maquillaje genético de un individuo, mientras que las características externas del individuo son su fenotipo. El fenotipo se expresa, forma, desarrolla y funciona sobre la base de la información proporcionada y codificada por el genotipo[4].

1.1.2 Extracción de ADN.

La extracción del ADN es un proceso vital con muchas aplicaciones científicas. En la investigación y en la medicina, sus usos incluyen secuenciación, detección de virus y bacterias, investigación de enfermedades y trastornos con base genética. En el campo de la ciencia forense, se utiliza para la identificación de los muertos, así como el análisis de la escena del crimen. La extracción de ADN es muy simple y sus técnicas de extracción varían de acuerdo al organismo. Dado que todos los organismos vivos contienen ADN, las opciones disponibles para la toma de la muestra son infinitas y la elección por lo general se relación directamente con el tema objeto de estudio.

Una vez que la muestra es obtenida, se debe lisar para obtener el material genético que se encuentra dentro de las células individuales. Dependiendo del fabricante de los químicos para tales efectos, se siguen sus instrucciones como si fuera recetario de cocina, que a grandes rasgos funciona de la siguiente manera: en pequeños tubos se coloca la muestra, y un detergente especialmente diseñado y

una enzima llamada proteinasa K, se encargaran de romper la estructura celular del material. El detergente corroe la membrana celular de la muestra y la membrana nuclear que rodea al material genético de la célula. Cuando estas membranas son alteradas, la proteinasa K degrada a una proteína llamada histona, que envuelve al ADN. Después una solución de sal concentrada se añade para agrupar a la proteína indeseada y a los restos celulares. Se centrifuga el tubo, en donde el la fuerza centrífuga deja al ADN difundido en una capa de la solución por encima del exceso de material más pesado. A continuación el ADN se retira y se coloca en otro tubo. Se agrega alcohol isopropilico y se mezcla cuidadosamente. Este proceso hace que el ADN de la solución se agrupe en filamentos visibles. Entonces, el material se coloca nuevamente en la centrifuga para forzar a las hebras de ADN a juntarse. El alcohol se retira y el ADN se deja precipitarse. Una vez completado el proceso, la muestra de ADN resultante esta lista para verificar su pureza y ser almacenada para cualquiera de sus muchos propósitos[6].

1.1.3 Técnica de PCR.

La reacción en cadena de la polimerasa, conocida como PCR por sus siglas en inglés, es una técnica de biología molecular desarrollada en 1986 por Kary Mullis, cuyo objetivo es obtener un gran número de copias de un fragmento de ADN particular, partiendo de una mínima cantidad. Esta técnica sirve para amplificar un fragmento de ADN y su utilidad es que tras la amplificación resulta mucho más fácil identificar con una muy alta probabilidad virus o bacterias causantes de una enfermedad, la identificación de cadáveres y realizar investigación científica con el ADN amplificado. Esta técnica se fundamenta en la propiedad natural de las Polimerasas para replicar hebras de ADN, para lo cual se emplean ciclos de altas y bajas temperaturas alternadas para separar las hebras de ADN recién formadas entre si tras cada fase de replicación y después, se deja que las hebras de ADN vuelvan a unirse para poder duplicarlas nuevamente[7].

1.1.4 Método delta-delta-Ct.

Los resultados del PCR en tiempo real se basan en la detección y cuantificación de los marcadores fluorescentes a lo largo de la reacción del PCR. Esto permite conocer la cantidad de fluorescencia emitida durante la fase exponencial de la reacción, donde un aumento significativo del producto de PCR se correlaciona con la cantidad inicial del ADN de estudio. Para obtener estos resultados, los valores del ciclo de umbral (Ct por sus siglas en inglés) deben ser obtenidos. Los valores de umbral son determinados por la identificación del ciclo en el cual la emisión de la intensidad del marcador fluorescente se eleva por encima del ruido de fondo en la fase exponencial de la reacción del PCR. En otras palabras, el

valor Ct está representado por el ciclo en el cual la producción de fluorescencia cruza el umbral establecido. Para nuestro estudio en particular, el método del ciclo comparativo de umbral relativo ($2^{-\Delta\Delta Ct}$) fue utilizado para cuantificar el número de cambios en las copias por comparación del valor ΔCt (ciclo de umbral(Ct) de la región del blanco menos la región de control Ct) de las muestras. El promedio del valor Ct de tres replicaciones para cada muestra fue calculado y normalizado contra un gen de control con la asunción de la existencia de dos copias de segmento de ADN en la región de control[8].

1.2 Las Variaciones Estructurales en el Genoma.

1.2.1 Introducción.

Las variaciones genéticas son el motor de la evolución, y por tanto el origen de nosotros mismos. Estas son las responsables de las diferencias en el aspecto de los individuos que van desde el color del cabello a la altura o la armonía de un rostro. También son los responsables del origen de ciertas enfermedades o de la predisposición a otras. Debido a todo esto es muy importante saber qué tipo de variaciones genéticas existen y qué incidencia tienen.

Hasta ahora se creía que las variaciones genéticas se debían a errores en el sistema genético o mutaciones conocidas como polimorfismos de nucleótido simple o SNP por sus siglas en inglés [9]. En este caso se produce la sustitución de una sola base por otra en la secuencia del ADN. La reorganización de secuencias de material genético, juegan un papel más importante que los que lo provoca una sola variación en una base (SNPs). Esta reorganización implica cambios en la estructura del ADN como son inserciones, borrados, inversiones y translocaciones de segmentos en el genoma[10].

Las variaciones del número de copias (CNVs) de ADN genómico son eventos que resultan de una falla en la maquinaria de replicación (división celular) y reparación del genoma o que también puede ser heredadas de un ancestro, en el cual se originó la CNV por algún problema de este mismo tipo. El estudio de las CNVs tienen una enorme importancia, debido a que pueden jugar un papel importante en el surgimiento de enfermedades complejas como el cáncer, sida o el mal de Alzheimer, entre otras.

1.2.2 Tipos de Variaciones Genéticas.

El éxito en el mapeo genético es contar con información detallada sobre la variación genética en los genomas. Esta variación genética indica que las

diferentes copias homologas de los cromosomas pueden tener diferentes secuencias de ADN en regiones muy específicas.

En las células el ADN sufre cambios químicos frecuentemente, especialmente cuando se está replicando. La mayoría de estos cambios son reparados inmediatamente y los cambios que no logran repararse son llamados mutaciones. El término mutación se usa a menudo para referirse al evento que crea a una nueva variante y no a la propia variante. Estas mutaciones pueden ser producidas por una exposición prolongada a la radiación, ciertos químicos o virus y a la herencia transmitida a los descendientes. Toda la nueva variación genética en los genomas se produce como resultado de estas mutaciones y estas pueden surgir de novo durante el proceso de la meiosis, lo que significa que están presentes en los hijos y no en los padres o que pueden también ser heredadas de los padres. Generalmente se reserva el término de mutación para una ocurrencia de novo y después nos referimos a ella como una nueva variante genética.

Se utiliza el término marcador genético, o simplemente marcador, para describir a la información genética observada a nivel molecular en un lugar en particular, que permita identificar las diferencias genéticas entre individuos. Las variaciones genéticas que se presentan en una región codificante de genes pueden originar en la proteína codificada por ese gen, un mal funcionamiento y que las células que dependen de esta proteína no funcionen correctamente, provocando problemas en los tejidos u órganos. A estas condiciones relacionadas con mutaciones o variaciones genéticas, se les llama desordenes o enfermedades genéticas. Un locus de susceptibilidad a alguna enfermedad indica que un gen o un locus genético especifico, tiene una variación genética asociada con una enfermedad.

Existen varios tipos de variaciones genéticas entre las cuales podemos mencionar a:

Polimorfismo de Nucleótido Simple (SNP). Es el tipo más simple de marcador genético. La estructura de doble hélice del ADN requiere que cada cromosoma tenga bases complementarias en cada lugar, como se observa en la Figura No. 1, para cada uno de los dos cromosomas, que muestra un SNP en un par no idéntico para los cromosomas homólogos. Los SNPs juegan un papel muy importante en el mapeo genético moderno, puesto que se distribuyen a lo largo del genoma y el costo de una genotipificación completa con SNPs es muy inferior a una secuenciación completa, lo que los hace muy atractivos para los estudios genéticos. Los SNPs ocurren uno cada 300 pares de bases en promedio, a lo largo de aproximadamente 10 millones de SNPs en el genoma humano. Los SNPs que se encuentran en regiones no codificantes, hasta el momento no se ha demostrado que tengan efectos genéticos directos sobre fenotipos o enfermedades, pero los SNPs dentro de una región codificadora, pueden provocar severos daños a los fenotipos[11].

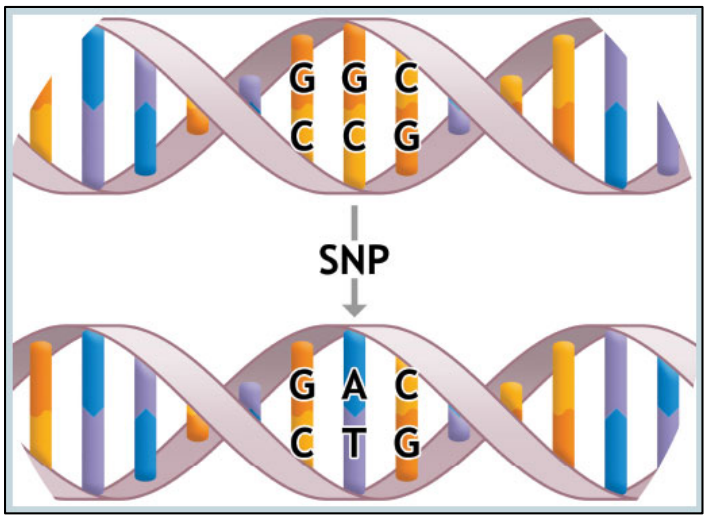


Figura No. 1.- SNPs

Indeles. Son pares de bases adicionales (entre 1 y 1000 bases par, ver Figura No. 2) que pueden ser insertados o borrados entre dos pares de bases especificas en una secuencia de ADN. Este tipo de variantes se diferencian de los SNPs, en que estos últimos únicamente sustituyen un nucleótido por otro y los indeles insertan o borran cadenas de nucleótidos menores a un 1 Kbp[11].

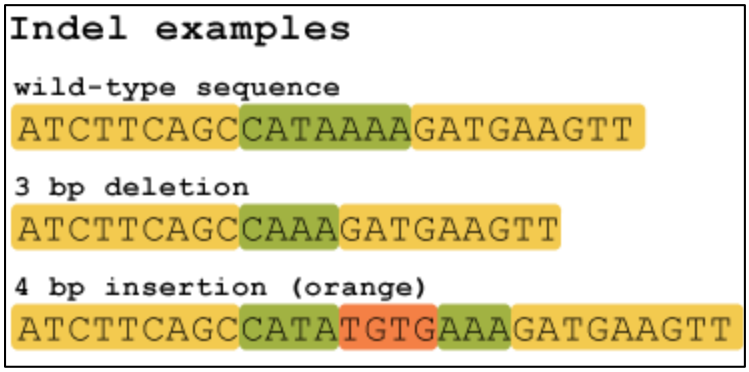


Figura No. 2.- Indeles.

Numero Variable de Repeticiones en Tandem (VNTRs). Otro tipo común de variación que consiste en secuencias de ADN específicas que se repiten inmediatamente adyacentes entre sí, un número variable de veces. (ver Figura No. 3). Los microsatélites (SSR o STR por sus acrónimos en inglés para Simple Sequence Repeat y Short Tandem Repeat) son una importante clase de VNTR que tienen un tamaño que varía de 1 a 6 pares de bases que se repiten. Debido a que el número de repeticiones de pares de bases se secuencias pueden variar ampliamente de una persona a otra, los microsatélites son excelentes marcadores para distinguir a una persona de otra. Estos son ampliamente utilizados en pruebas forenses y de paternidad[11].

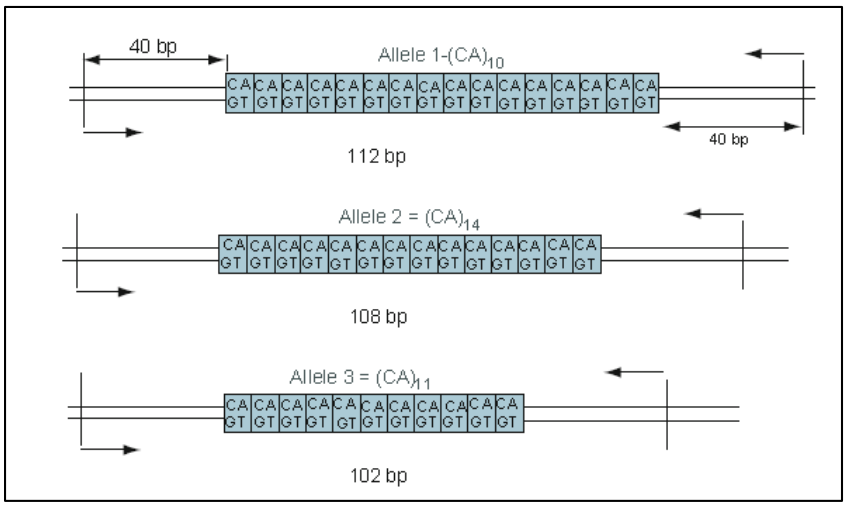


Figura No. 3.- Ejemplo de tres diferentes alelos de Short Tandem Repeat (STR)

Variantes Estructurales. Estos incluyen muchos tipos de reordenamientos en la estructura de los cromosomas, tales como duplicaciones, inserciones, translocaciones, inversiones y borrados de material genético. Por lo general cuando surgen de novo estas variantes, es durante la formación del ovulo y el espermatozoide. Los borrados y duplicados que involucran segmentos de ADN mayores o iguales a 1 Kbp se definen generalmente como Variaciones del Número de Copias (CNVs), por que en tales casos, los individuos pueden tener demasiadas copias (más de 2) o demasiado pocas (1 ó 0) del gen o segmento cromosómico (ver Figura No.4). Las CNVs pueden ser asociadas con algunas enfermedades como el cáncer. Para este tipo de enfermedad, al existir duplicados en algunas secuencias en donde algunos genes sintetizan proteínas, como por ejemplo los oncogenes, que incrementan la división celular, al existir un mayor número de oncogenes, reacciona con mayor frecuencia el mecanismo de activación tumoral. De existir borrados en las secuencias (siguiendo el caso de alguna enfermedad cancerígena), por ejemplo en la región involucrada con las inactivaciones de genes supresores de tumores, esto provocaría la proliferación de las células cancerígenas[11].

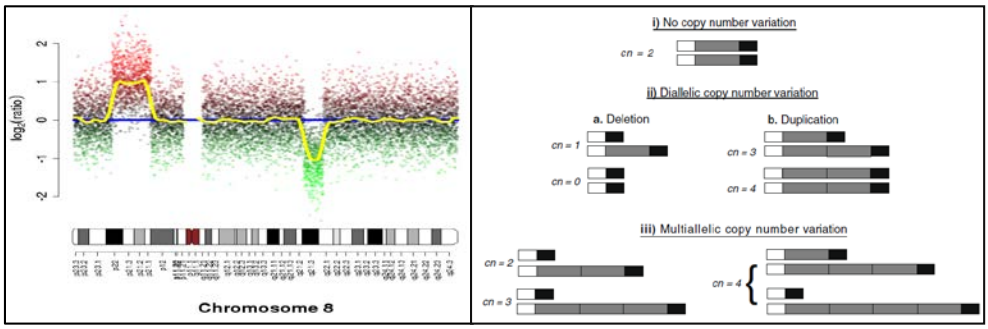


Figura No. 4.- Ejemplos de Variaciones del Número de Copias (CNVs).

Barrido Selectivo (Selective Sweep). En genética de poblaciones y evolución, es la eliminación o reducción de la variabilidad genética entre los nucleótidos vecinos a una mutación, como resultado de un proceso reciente y positivo de selección natural. Este ocurre cuando se produce una mutación que incrementa la eficacia biológica de un organismo en relación a sus congéneres de la misma población (Figura No. 5). La selección favorecerá a aquellos individuos con la mayor aptitud y en el transcurso de las generaciones el nuevo alelo mutante incrementara su frecuencia en la población[12].

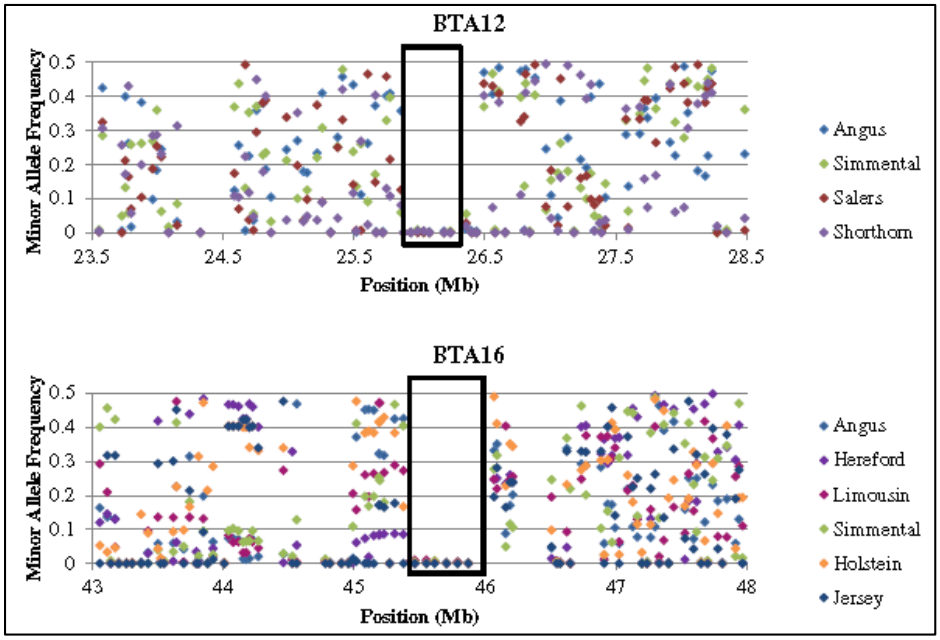


Figura No. 5.- Ejemplos de Selective Sweeps en ganado bovino.

Bloques de Haplotipos. Son una secuencia de ADN en la que una fracción de los haplotipos observados, están representados al menos n veces en la muestra. Los bloques de haplotipos son un conjunto de alelos estrechamente vinculados en un cromosoma que con el tiempo evolutivo, tienden a ser heredados juntos. Los límites de estos bloques, se relacionan estrechamente con la recombinación. Recientes estudios sugieren que el genoma humano está organizado en bloques de haplotipos[13]. Ver Figura No. 6.

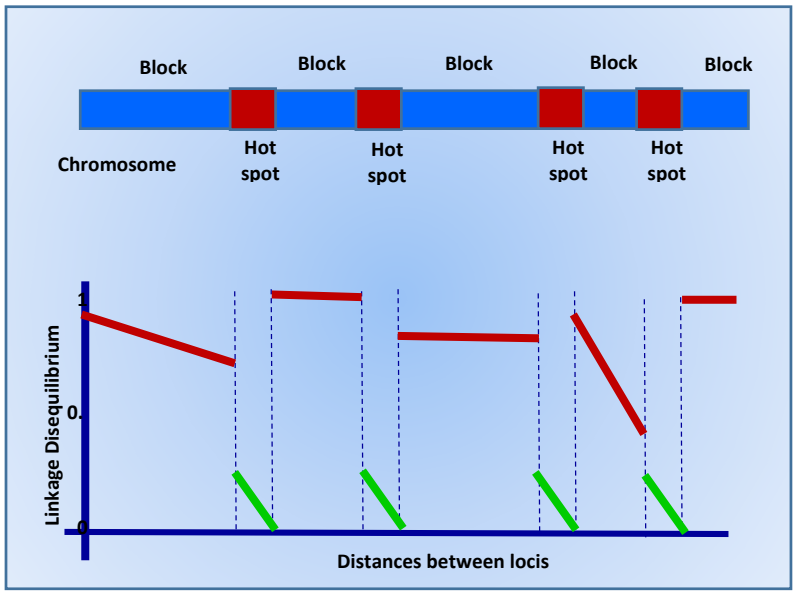


Figura No. 6.- Bloques de Haplotipos.

Segmentos Duplicados (SDs).- Es un tipo de mutación que causa redundancia en los genomas. Estos son originados por un error en la recombinación donde dos cromosomas se cortan en diferentes lugares y las piezas se vuelven a unir de tal manera que la secuencia de ADN entre los dos lugares se incluyen dos veces, teniendo como resultado, que el cromosoma de un organismo contenga dos copias casi idénticas de secuencia, ver Figura no. 7. Una secuencia de ADN es considerada SD cuando su longitud es mayor a 1 Kbp y cuando su secuencia repetida es idéntica al menos un 95%[14].

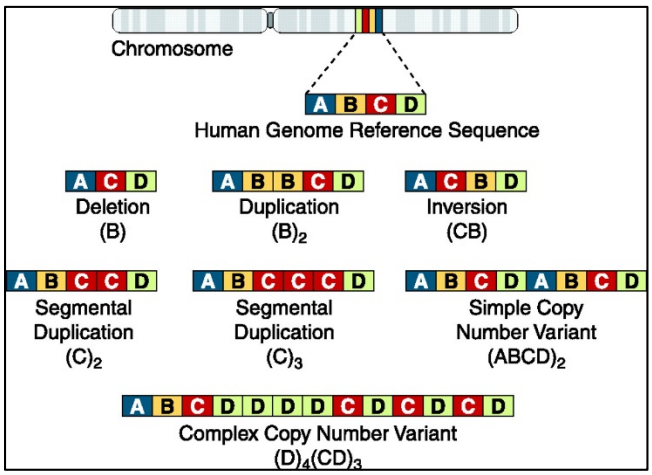


Figura No. 7.- Segmentos Duplicados.

1.3 Desequilibrio Ligado y Equilibrio de Hardy-Weinberg.

En esta sección veremos dos conceptos importantes, relacionados con los estudios de asociación basados en la población, que son: El Desequilibrio Ligado (LD) y el Equilibrio de Hardy-Weinberg (HWE). Ambos conceptos se basan en el componente genético de nuestros datos y tanto el LD como el HWE son medidas de asociación alélica. La diferencia entre ellos es que el LD es una medida de asociación alélica entre dos sitios en el genoma, mientras que el HWE es una medida de asociación alélica entres dos cromosomas homólogos en un solo sitio.

1.3.1 Patrones de Desequilibrio Ligado (LD) y Haplotipos.

En la meiosis, un alelo normalmente se transmite junto con los alelos que se encuentran alrededor de él, lo cual conduce a una asociación o correlación entre un loci que está cerca el uno del otro, tal fenómeno es llamado Desequilibrio Ligado (LD). La distribución del LD en todos los genomas es un tema de gran interés debido a su papel fundamental en el mapeo de genes, la recombinación y su historia[15].

Asumiendo dos alelos A y B en dos loci con frecuencias πA y πB en una población respectivamente, esperamos que la frecuencia del haplotipo AB sea $\pi A \pi B$ si los dos loci son independientes. Si la frecuencia del haplotipo AB en una población no se ajusta a lo siguiente: $\pi AB = \pi A \pi B$, entonces los dos loci se encuentran en LD. Existe una amplia variedad de pruebas estadísticas para medir el LD, la más simple de las tales es: $D = \pi AB - \pi A \pi B$, donde los valores de D son seriamente afectados por las frecuencias alélicas. La normalización de D es la forma más común de abordar la dependencia de D en las frecuencias de los alelos marginales. La D' (D prima) de Lewontin[16], es una normalización de la D, donde $|D'| = 1$ nos indica que existe un perfecto LD entre dos loci.

El cálculo de la D' está dado por:

$$D' = \begin{cases} \dfrac{D}{\min(\pi A \pi b, \pi a \pi B)} & D > 0 \\ \\ \dfrac{D}{\min(\pi A \pi B, \pi a \pi b} & D < 0 \end{cases}$$

Donde πa y πb son las frecuencias de los alelos a y b, que son la contraparte de A y B en los mismos loci respectivamente. El inconveniente obvio de D' es que

las propiedades de su muestreo son poco conocidas cuando |D'| < 1, además la estimación de D' es fuertemente inflada cuando tenemos muestras pequeñas, especialmente en las variantes raras. Actualmente la medida más utilizada entre locis bialelicos es la r^2, cuyos valores reflejan la cantidad de información proporcionada sobre los demás:

$$r^2 = \frac{D^2}{\pi A \pi a \pi B \pi b}$$

En el caso de la $r^2 = 1$, es un indicativo de que existe un perfecto LD, lo que significa que un marcador provee una completa información acerca de su correlación con otro. La r^2 provee de un método en caso de que el tamaño de la muestra se pequeño, y se define asi:

$$r^2 corrected = \frac{r^2 computed - \frac{1}{n}}{1 - \frac{1}{n}}$$

La recombinación en el sentido genético se define como la unión de dos cadenas de ADN, una del lado de la madre y otra del lado paterno. Esto ocurre a nivel cromosomas de los padres, cuya información genética será heredada a los hijos. Los estudios de mapeo genético, son a menudo desarrollados por los investigadores en regiones de cromosomas que fueron identificados a través de análisis de LD. Estos estudios identifican la ubicación de un gen específico candidato, con mayor precisión.

Con la terminación del proyecto HapMap (proyecto internacional desarrollado para crear un mapa de haplotipos del genoma humano), se han proporcionado los patrones de LD a escala muy fina del genoma humano[15] y se han definido rasgos en el genoma como: en primer lugar, se descubrió que el LD varia notablemente en escalas de 1 – 100 Kb, siempre es discontinuo y se compone por estructuras de tipo de bloques. En segundo lugar, se plantea la diversidad de haplotipos únicamente a través de la mutación en el genoma cuando la recombinación está ausente. Esto es, cuando los SNPs generados en la misma

rama de la genealogía se encuentran en perfecto (LD |D'| = 1), mientras que los que ocurrieron en diferentes ramas tienen poca o ninguna correlación. En tercer lugar, a pesar de que las diferentes poblaciones tienen diferentes frecuencias de haplotipos, tanto comunes como raros son generalmente compartidos entre las poblaciones. Cuarto, algunos SNPs localizados en puntos de recombinación, tienen muy bajo LD con sus SNPs vecinos, y finalmente, el LD se correlaciona con muchas características genómicas tales como la tasa de recombinación, la tasa de mutación, el contenido de Guanina + Citosina (GC), la variación en la secuencia, la composición de la repetición y la longitud de un cromosoma. Un estudio reveló que en cierta medida, el LD en el genoma humano se refleja en la distribución geográfica de los haplotipos, mismos que pierden diversidad a medida que aumentan su distancia con Africa[17].

1.3.2 Equilibrio de Hardy-Weinber (HWE).

Conocido también como Principio o Ley de Hardy-Weinber, establece que la composición genética de una población permanece en equilibrio mientras no actué la selección natural, ni ningún otro factor y no se produzca ninguna mutación. Es decir, la herencia Mendeliana, por si misma no engendra cambio evolutivo. Por lo tanto el HWE es una expresión de la noción de que una población está en equilibrio genético.

Mientras que el LD se refiere a la asociación alélica a través de sitios en un solo homologo, HWE denota la independencia de alelos en un solo sitio entre dos cromosomas homólogos.

HWE, afirma que bajo ciertas condiciones, después de una generación de apareamiento al azar, las frecuencias delos genotipos de un locus individual se fijaran en un valor de equilibrio en particular. También especifica que esas frecuencias de equilibrio se pueden representar como una función sencilla de frecuencias alélicas en ese locus. En el caso más sencillo, con un locus con alelos *A* y *a*, con frecuencias alélicas de *p* y *q* respectivamente, el HWE establece que la frecuencia genotípica para el homocigoto dominante *AA* es p^2, la del heterocigoto *Aa* es *2pq* y la del homocigoto recesivo *aa* es q^2[18], ver Figura No.8. Los tres genotipos *AA-Aa-aa* se definen en la siguiente ecuación:

$$p^2 + 2pq + q^2 = (p + q)^2 = 1.$$

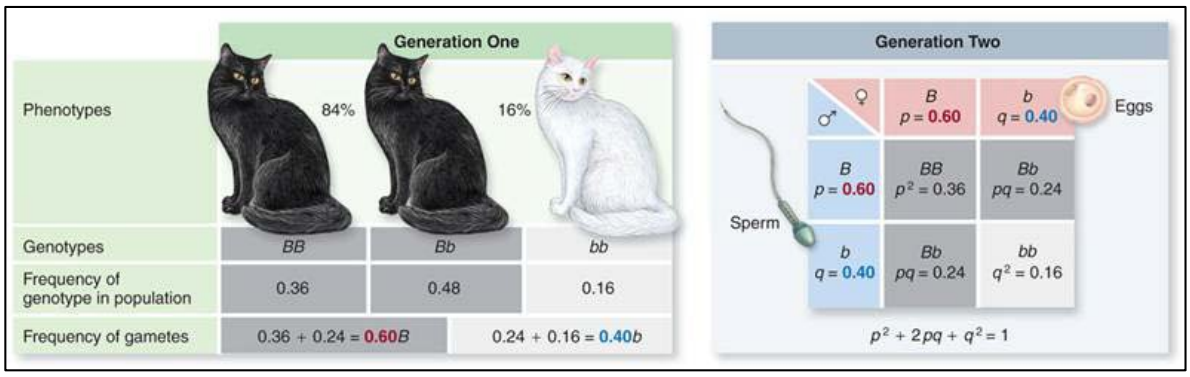


Figura No. 8.- Equilibrio de Hardy-Weinber.

1.4 Estadística Bioinformática.

Los actuales desarrollos en la biotecnología han llevado a un crecimiento exponencial en los estudios sobre el análisis del ADN en organismos, sus funciones, interacciones entre sus biomoléculas y sus rutas metabólicas. Con estas nuevas tecnologías, también se ha incrementado la información en las bases de datos, mismas que demandan de nuevas herramientas para almacenar, analizar e interpretar los grandes volúmenes de información. Debido a que la información es compleja y cruda, se requiere de la normalización en la información por medio de métodos estadísticos, para al final obtener datos que sean muy fáciles de interpretar. Recientemente ha surgido una nueva ciencia llamada Bioinformática Estadística que tiene como objetivo analizar la información en las bases de datos, para mejorar la comprensión de los complejos fenómenos biológicos a través del desarrollo de nuevas metodologías, mediante el uso de métodos estadísticos. Esta ciencia interdisciplinaria requiere de la aplicación de conocimientos matemáticos y estadísticos en las ciencias biológicas. La investigación en bioinformática básicamente consiste en el desarrollo de nuevos métodos, modificación de los métodos disponibles, combinación de varias herramientas informático-biológicas y de la aplicación de métodos estadísticos[19]. La bioinformática se ha convertido en el campo más prominente de las estadísticas y esto se comprueba al observarse el gran crecimiento que existe en el número de publicaciones sobre la investigación utilizando métodos bioinformáticos. Entre los muchos beneficios en la aplicación de la bioinformática estadística podemos mencionar:

- Ayuda a la identificación de nuevos genes y proporciona información acerca de su funcionamiento en diferentes condiciones.

- Facilita el estudio a mayor precisión a los investigadores sobre las enfermedades complejas como el cáncer.
- Proporcionar información sobre los patrones de actividades de genes y ayuda a la clasificación de enfermedades en base a sus perfiles genéticos en lugar de clasificarlos en base los órganos donde se encuentra el tumor o células enfermas.
- La bioinformática estadística ayuda a comprender las correlaciones entre las respuestas terapéuticas a las drogas y los perfiles genéticos, lo que ayuda a desarrollar fármacos que contrarrestan las proteínas producidas por genes específicos, para reducir o eliminar sus efectos.
- Ayuda a estudiar la correlación entre los tóxicos y los cambios en los perfiles genéticos de las células expuestas a dichas sustancias toxicas.

En los siguientes párrafos veremos la descripción de algunos métodos aplicados en la bioinformática estadística.

1.4.1 Análisis de Componentes Principales (PCA).

El tema central de las estadísticas se basa en la idea de que se cuenta con un conjunto de datos y se desea analizar ese grupo de datos en términos de las relaciones entre los distintos puntos de ese conjunto de datos. Una de las formas de buscar una relación entre un conjunto de datos es por medio de un Análisis de Componentes Principales (PCA), que es una manera práctica para identificar patrones en los datos y expresarlos de tal forma que resalten sus similitudes y diferencias. Dado que los patrones de datos pueden ser difíciles de encontrar en los datos de alta densidad, el PCA puede representarlos de forma gráfica. Una de sus ventajas principales es que una vez que haya encontrado estos patrones, permite comprimir la información, reduciendo el número de dimensiones sin mucha pérdida de información.

El procedimiento que utiliza un PCA, es tener definidos vectores de datos de x dimensiones, después se resta la media a cada dato de cada una de las x dimensiones, con esto obtenemos datos ajustados. A los datos ajustados de los vectores se les calcula una matriz de covarianzas y a esta matriz de covarianzas le calculamos los vectores propios (eigenvectores) y sus valores propios (eigenvalores). Mediante este proceso de tomar los vectores propios de la matriz de covarianza, tenemos que ha sido capaz de extraer las líneas que caracterizan los datos. El siguiente paso es la elección de los componentes y la formación de un vector de características, y es aquí donde entra el concepto de compresión de

datos y reducción de dimensionalidad. Al observar los valores en los vectores propios y en los valores propios, podemos constatar que los valores propios son bastante diferentes a los vectores propios, de hecho, el vector propio con el valor propio más alto es el componente principal del conjunto de datos, que es la relación más significativa entre las dimensiones de los datos. En general, una vez que se encuentran los vectores propios de la matriz de covarianza, el siguiente paso es ordenarlas por valor propio, del más alto al valor más bajo. Esto le da a los componentes su orden de importancia. Se pueden ignorar los componentes de menor importancia, debido a que los valores más altos, capturan la mayor variabilidad. Finalmente tendremos dos opciones para representar valores, uno de ellos es graficar con los dos primeros componentes principales en las dimensiones de los ejes “x” y “y”, o graficar únicamente el mayor, en una dimensión en cualquiera de los dos ejes.

Mediante este procedimiento básicamente hemos transformado nuestros datos, expresándolos en términos de variabilidad de patrones entre ellos, donde los patrones son las características que describen más estrechamente las relaciones entre los datos. Este análisis es muy útil, porque ahora nos ha clasificado nuestro conjunto de datos en una combinación de atributos de cada una de nuestros valores en los vectores originales, y al final con esta información podemos deducir su tendencia, ya sean agrupándose o alejándose de sí mismos[20].

1.4.2 Modelos Ocultos de Markov (HMM).

Es un modelo estadístico en el que se asume que el sistema a modelar es un fenómeno aleatorio dependiente del tiempo, para el cual la distribución de probabilidad del valor de una variable aleatoria depende de su valor presente, pero el cual al ser un proceso estocástico, la probabilidad condicional sobre el presente, pasado y futuro del sistema son totalmente independientes. Su objetivo es determinar los parámetros desconocidos (u ocultos, de ahí su nombre) de dicha a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo análisis sucesivos, por ejemplo en aplicaciones de reconocimiento de patrones.

En un modelo de Markov normal, el estado es visible directamente para el observador, por lo que las probabilidades de transición entre estados son los únicos parámetros. En un HMM, el estado no es visible directamente, sino que lo son las variables influidas por el estado. Cada estado tiene una distribución de probabilidad sobre los posibles símbolos de salida. La secuencia de símbolos generados por un HMM proporciona cierta información acerca de la secuencia de estados[21].

Los HMM son especialmente aplicados al reconocimiento de patrones en formas temporales, como el reconocimiento del habla, de rostros, de escritura, gestos y movimientos corporales, reconocimiento óptico de caracteres, traducción automática, entre otros. En la segunda mitad de la década de los 80's, los HMM comenzaron a ser aplicados en análisis de secuencias biológicas, particularmente de ADN, y actualmente en campos como la bioinformatica y Genomica, se han convertido en modelos muy eficaces en la predicción de regiones codificadoras de proteínas, en el modelado de secuencias de familias de proteínas o de ADN, en la predicción de elementos de estructuras secundarias en secuencias primarias de proteínas, en la identificación variaciones del número de copias en los genomas, etc.

1.4.3 Prueba t-test.

Esta se utiliza para determinar si existe alguna diferencia significativa entre las medias de dos grupos, siendo estas las medias de dos poblaciones independientes y normales, y se asume que las variables dependientes tienen una distribución normal. Esta prueba estadística regresa un valor "t", que conceptualmente el valor-t representa el número de unidades estándares que están separando a las medias de los dos grupos. Una vez que se haya determinado un valor t, es posible encontrar un valor p (p-value) que se asocie con una tabla de valores de la distribución t de student. Si el valor p calculado es menor al límite elegido por significancia estadística (usualmente los niveles de significancia o nivel del valor alfa, son valores de 0.10, 0.05 o 0.01, que estamos dispuestos a aceptar), entonces la hipótesis nula se rechaza en favor de la hipótesis alternativa. Esta prueba está definida por la siguiente ecuación:

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Donde $\bar{x}$ es la media muestral, μ_0 es un valor específico, *S* es la desviación estándar muestral y *n* es el tamaño de la muestra[22].

1.4.4 Prueba X^2 de Pearson (Chi Square).

Esta prueba es utilizada para saber si la frecuencia observada obtenida a partir de un experimento, está lo suficientemente cerca de la frecuencia esperada,

indicando con esto en que medida existen diferencias entre ambas frecuencias. Esta prueba esta definida por la siguiente ecuación:

$$x^2 = \sum_i \frac{(observada_i - teorica_i)^2}{teorica_i}$$

Donde $observada_i$ i =1,2, . . . n, denota la frecuencia observada de los objetos y $teorica_i$ i = 1,2, . . . n, es la frecuencia esperada de los objetos, n es el tamaño de la muestra[19].

1.5 Pruebas de Hipótesis y Procedimientos de Comparación Múltiple.

Algunos métodos estadísticos tradicionales establecen pruebas que describen la asociación entre un genotipo y un rasgo. Estos métodos generalmente requieren de una comparación de la prueba estadística de distribución, con el fin de cuantificar la probabilidad de ver lo que hacemos o algo más extremo, bajo el supuesto de que la hipótesis nula es verdadera. Nuestra decisión de aceptar o rechazar la hipótesis nula, se basa en una comparación de una cantidad (llamada el valor p), a un umbral basado en un nivel determinado previamente de error aceptable. Por ejemplo, en los estudios de asociación basados en la población, por lo general, se pretende probar la presencia de asociaciones entre el rasgo y cada uno de los múltiples genotipos a través de varios SNPs y loci de genes.

1.5.1 Pruebas de Hipótesis.

La prueba de hipótesis es otra parte importante de la inferencia estadística. En el caso de la estimación, nos encontramos con el valor esperado del parámetro de la población desconocida, que puede ser un único valor o un intervalo, mientras que en la prueba de hipótesis, decidimos si el valor dado de un parámetro es en realidad el valor verdadero del parámetro. En las pruebas de hipótesis, nos formulamos dos o más hipótesis y ponemos a prueba sus puntos fuertes mediante el uso de la información proporcionada por la muestra. Por lo tanto, sobre la base de una muestra de la población, se decide cuál de las dos o más hipótesis complementarias es verdadera.

Definimos una hipótesis como una declaración acerca de un parámetro de la población. Una simple hipótesis específica complemente la distribución,

mientras que una hipótesis compuesta no especifica completamente la distribución. Por lo tanto, una hipótesis simple no solo especifica la forma funcional de la distribución de la población subyacente, si no que también especifica los valores de todos los parámetros. La hipótesis se clasifica como una hipótesis nula o hipótesis alternativa. Las hipótesis nula y alternativa son hipótesis complementarias. En general, la hipótesis nula, denotada por *Ho*, es la afirmación de que se supone inicialmente para ser la verdad. La hipótesis nula puede ser considerada como una afirmación de "ningún cambio, efecto o consecuencia". Esto sugiere que la hipótesis nula debe tener una afirmación en la que no existe ningún cambio en la situación, no hay diferencia en las condiciones, o ninguna mejora se ha reclamado. La hipótesis alternativa es la afirmación, denotada por *Ha*. Por ejemplo, en el sistema judicial de USA, una persona se presume inocente hasta que se demuestre que es culpable. Esto significa que la hipótesis nula es que la persona es inocente, mientras que la hipótesis alternativa es que la persona es culpable[19].

La forma de la hipótesis nula está dada por

Ho: parámetro de la población o la forma = valor de la hipótesis, donde el valor de la hipótesis se especifica por el investigador o el problema en sí.

La hipótesis alternativa se puede tomar de una de las 3 formas siguientes:

Ha: parámetro de la población > valor hipotético, esta forma es llamada alternativa unilateral a la derecha.

Ha: parámetro de la población < valor hipotético, a esta forma se le conoce como alternativa unilateral a la izquierda.

Ha: parámetro de la población ≠ valor hipotético, a esta forma se le llama alternativa a doble cara.

La forma de la hipótesis alternativa debe ser decidida de acuerdo con el objetivo de la investigación.

Una prueba de hipótesis es un método para decidir si la hipótesis nula es rechazada o aceptada sobre la base de un dato de la muestra. Las pruebas de hipótesis tienen dos componentes: 1.- una prueba estadística de la muestra como punto de referencia, se utiliza para tomar la decisión. 2.- la prueba de hipótesis divide las posibles opciones de la prueba estadística en dos subconjuntos: una región de aceptación de la hipótesis nula y una región de rechazo de la hipótesis nula. Una región de rechazo consta de los valores de la prueba estadística que conducen al rechazo de la hipótesis nula. La hipótesis nula se rechaza si y solo si

el valor numérico de la prueba estadística, calculada sobre la base de la muestra, cae en la región de rechazo. Además del valor numérico de la prueba, es recomendable el cálculo de un valor p, que es una probabilidad calculada por la prueba estadística. Si el valor p es menor que o igual al nivel de significancia, se decidirá de antemano, el rechazo de la hipótesis nula[18].

En el proceso de comprobar una hipótesis estadística, encontramos cuatro situaciones que determinan si la decisión es correcta o errónea. Estas cuatro posibles opciones se muestran en la Tabla No.1, en las que dos situaciones dan lugar a decisiones equivocadas. Un error de tipo I se produce cuando se rechaza Ho y que en realidad es verdadera. Un error tipo II se produce cuando aceptamos Ho y en realidad es falsa.

Decisión Tomada	**Situación Real**	
	Ho es verdadera	*Ho* es Falsa
Aceptamos *Ho*	Decisión correcta	Error Tipo II (Falso Negativo)
Rechazamos *Ho*	Error Tipo I (Falso Positivo)	Decisión correcta

Tabla No. 1.- Pruebas de hipótesis.

1.5.2 Significancia Estadística.

En la estadística, un resultado es estadísticamente significativo cuando no es probable que se haya realizado debido al azar. Una diferencia estadísticamente significativa", únicamente representa que hay evidencias estadísticas de que existe una diferencia. El nivel de significancia de una prueba, se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando esta es verdadera (error tipo I o falso positivo). La decisión se toma a menudo utilizando el *Valor-P*, si este valor es inferior al nivel de significación, entonces la hipótesis nula es rechazada. Cuanto menor sea el *Valor-P*, más significativo será el resultado[23].

El nivel de significancia estadística, se representa comúnmente por el símbolo griego α (alfa). Son comunes los niveles de significancia 0.05, 0.01 y 0.001. Si un contraste de hipótesis proporciona un *Valor-P* inferior a α, la hipótesis nula es rechazada, siendo tal resultado denominado estadísticamente significativo.

Cuanto menor sea el nivel de significancia, más fuerte será la evidencia de que un hecho no se debe a una mera coincidencia al azar.

1.5.3 El Valor-P.

En contrastes de hipótesis en estadística, el *Valor-P*, se define como la probabilidad, de obtener un resultado al menos tan extremo, como el que realmente se ha obtenido (valor estadístico calculado), suponiendo que la hipótesis nula es cierta. Es fundamental tener en cuenta que el *Valor-P*, está basado en la asunción de que la hipótesis nula es verdadera. El *Valor-P* es un valor de probabilidad, que oscila entre 0 y 1, por lo que los valores muy próximos a 1, no rechazan la hipótesis nula y sus valores muy aproximados a 0, rechazan la hipótesis nula.

Se rechaza la hipótesis nula si el *Valor-P* asociado al resultado observado es igual o menor al nivel de significancia establecido, convencionalmente en 0.05 ó 0.01. Si el *Valor-P* es inferior al nivel de significancia, nos indica que lo más probable es que la hipótesis de partida sea falsa[23]. Sin embargo. También es posible que estemos ante una observación atípica, por lo que estaríamos cometiendo el error estadístico de rechazar la hipótesis nula, cuando esta es cierta, basándonos en que hemos tenido la mala suerte de encontrarnos con una muestra atípica. Este tipo de errores se pueden reparar rebajando el *Valor-P*. Un *Valor-P* de 0.05 es usado en investigaciones habituales, mientras que *Valores-P* de 0.01 se utilizan en investigaciones médicas, en las que cometer un error puede acarrear consecuencias más graves. También se puede tratar de subsanar el error, aumentando el tamaño de la muestra, lo que resude la posibilidad de que el valor obtenido sea casualmente raro.

1.5.4 Procedimientos de comparaciones múltiples.

Estos procedimientos de múltiples comparaciones o pruebas múltiples, normalmente se refieren a comparaciones de dos grupos, tales como un grupo de tratamiento y un grupo de control. Estas comparaciones múltiples surgen cuando un análisis estadístico abarca una serie de comparaciones formales, con la presunción de que la atención se centrara en las diferencias más fuertes entre todas las comparaciones que se hacen. Estas comparaciones se realizan cuando se tiene en cuenta a un conjunto de inferencias estadísticas simultáneamente o se infiere un subconjunto de parámetros seleccionados sobre la base de valores observados. Los errores en la inferencia estadística, (incluyendo los intervalos de confidencia), que incluyen en sus parámetros de poblaciones correspondientes o las pruebas de hipótesis que rechazan incorrectamente la hipótesis nula, es más probable que ocurran fallas o errores cuando se tiene en cuenta el conjunto como

un todo. Varias técnicas estadísticas se han desarrollado para evitar que esto suceda, lo que permite niveles de significancia para las comparaciones individuales y múltiples, para ser comparados directamente. Estas técnicas generalmente requieren un umbral de significación mayor para comparaciones individuales, a fin de compensar el número de inferencias que se están realizando.

Hay básicamente tres tipos de pruebas de comparación múltiple[19]: *single-step*, *step-down* y *step-up*. En la prueba *single-step*, cada hipótesis se evalúa utilizando un valor crítico que es independiente de los resultados de las otras pruebas de hipótesis. En el procedimiento *step-down*, el rechazo de una hipótesis en particular se basa no solo en el número total de hipótesis, sino también en el resultado de las pruebas de otras hipótesis, mientras se mantiene la tasa de *Error Tipo I* bajo control. En las pruebas *step-down*, las pruebas de hipótesis se realizan secuencialmente, empezando con las hipótesis correspondientes a las pruebas estadísticas más significantes. Con esta prueba, se puede detener el método de comparación, debido a que si una de las hipótesis es aceptada, el resto de las hipótesis son aceptadas automáticamente. Para las comparaciones *step-up*, las hipótesis corresponden a las mínimas pruebas estadísticas significativas analizadas sucesivamente, y una vez que una de las hipótesis es rechazada, todas las hipótesis restantes son rechazadas automáticamente.

1.5.5 Ajustes de comparaciones múltiples.

En la mayoría de la literatura, sobre métodos para ajustar las comparaciones múltiples, describen el control de al menos uno de los dos tipos de errores siguientes: la *family-wise error rate* (FWER) y la *false Discovery rate* (FDR).

Asumamos que por cada gen se realiza una prueba estadística para la expresión diferencial, si fijamos el nivel de significancia de $\alpha = 0.05$, en promedio uno de cada 20 genes que en realidad no se expresan diferencialmente, mostraran un *Valor-P* inferior a α solo por casualidad. Debido a la gran cantidad de genes representados en un microarreglo, esto puede conducir a un gran número de falsos positivos. Por esta razón, su sugiere el uso de *family-wise error rate* (FWER) y este se define como la probabilidad de que el conjunto seleccionado de los genes contiene por lo menos un falso positivo. Un procedimiento de múltiples pruebas, se dice que proporciona un fuerte control de la FWER, si controla la FWER para cualquier combinación de hipótesis nulas falsas o verdaderas. Si los Valores-P para la prueba estadística *T1…Tn* de *n* genes están

disponibles, un simple ajuste que le da un fuerte control de la FWER, es la corrección de Bonferroni, donde cada prueba es controlada a un nivel igual a α. Esto significa que para cualquier simple prueba, la probabilidad de rechazar incorrectamente la hipótesis nula, su tasa de error tipo I, es menor o igual a α. El ajuste de Bonferroni para múltiples comparaciones, involucra el usar simplemente un $\alpha' = \alpha/n$, en lugar de α para el nivel de cada prueba, donde *n* es el número de comparaciones a ser desarrolladas[18].

Para muchas aplicaciones, sin embargo, es control de la FWER es demasiada conservadora, con el peligro de que muchos genes interesantes se estén perdiendo. Como los microarreglos se utilizan a menudo para mostrar los genes candidatos, que pueden ser luego validados a través de nuevos experimentos, el investigador puede estar dispuesto a aceptar una cierta fracción de falsos positivos y esto cabe en el concepto de *false Discovery rate* (FDR), donde para este caso existe el ajuste de Benjamini – Hochber (B-H). Iniciemos por considerar probar la serie de hipótesis nulas independientes dadas por *H1….Hm*, y suponiendo que los *Valores-P* resultantes están dados por *p1….pm* y supongamos que queremos controlar la tasa de falsos descubrimientos en un nivel *q*. Para el ajuste B-H, supongamos por ejemplo, que queremos probar la asociación entre cada uno de diez SNPs y la presencia de una enfermedad. Por simplicidad asumamos que cada SNP está en un gen separado y nuestras hipótesis son independientes. Además supongamos que estamos interesados principalmente en los principales efectos de 10 SNPs y no en sus interacciones. Para este caso, para cada SNP*i* = 1….10 y los *Valores-P* deben estar ordenados de menor a mayor. El ajuste B-H comparara el índice *i* del *Valor-P* ordenado de menor a mayor, contra el valor obtenido por $\alpha^* = 0.05(i/10)$ para cada *i* y B-H rechazara la hipótesis nula, cuando α* sea menor al *Valor-P*.[18,23]

1.6 Recursos Bioinformáticos.

Con la conclusión del Proyecto del Genoma Humano, provoco la proliferación de recursos bioinformáticos on-line disponibles para toda la comunidad científica. Descubrir, localizar y aprender a utilizar las nuevas aplicaciones, supone un costo, sobre todo en términos de tiempo, donde la mayoría de los investigadores no lo pueden asumir. Por esto surge la necesidad de organizar los recursos existentes para facilitar lo más posible estas tareas de búsqueda. Con el número creciente día tras día de recursos bioinformáticos, en los últimos años se han desarrollado herramientas que permiten la búsqueda e indexación automática de recursos bioinformáticos a partir de publicaciones científicas[24]. Esto con el fin de evitar la pérdida de tiempo al buscar el

recurso adecuando en la red. En la investigación biomédica, cada vez es más frecuente que los recursos generados por los investigadores (bases de datos, software y otros) se pongan a disposición de toda la comunidad científica, con el fin de acelerar el avance a favor de la ciencia. A continuación haremos una breve descripción de las bases de datos consultadas, así como el software bioinformático utilizado.

1.6.1 El Lenguaje Perl.

Perl esta implementado como un intérprete escrito en lenguaje C, junto con una gran colección de módulos escritos en Perl y C. Estructuralmente, Perl está basado en un estilo de bloques y fue ampliamente adoptado por su destreza en el procesado de texto y no tener ninguna de las limitaciones de los otros lenguajes de script.

El intérprete tiene una arquitectura orientada a objetos y todos sus elementos están representados en el intérprete como estructuras C. La ejecución de un programa Perl se puede dividir en dos fases: tiempo de compilación y tiempo de ejecución. En el tiempo de compilación el intérprete analiza la sintaxis (parsea) el texto del programa en un árbol sintáctico. En tiempo de ejecución, el programa se ejecuta siguiente el árbol. El texto es parseado solo una vez y el árbol sintáctico es optimizado antes de ser ejecutado, para que la fase de ejecución sea relativamente eficiente.

Se ha utilizado Perl desde los primeros días de la Web para escribir guiones (scripts) CGI. Es una de las “tres Pes” (Perl, Python y PHP), que son los lenguajes más populares para la creación de aplicaciones Web. Muchos sitios Web con alto tráfico como amazon.com y tycketmaster.com utilizan Perl extensamente.

Perl se usa a menudo como un lenguaje pegamento, debido a que liga sistemas e interfaces que no fueron diseñados específicamente para interoperar, y para la operación con datos, convirtiendo o procesando grandes volúmenes de datos, siendo estas las principales fortalezas de Perl. Para mayor información sobre el lenguaje Perl visitar www.perl.org.

1.6.2 El Lenguaje R.

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, que nació como resultado de la implementación GNU del premiado lenguaje S. R y S-Plus, son los lenguajes

más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, bioinformática y matemáticas financieras. Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. Gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C, C++ o fortran. Los usuarios más avanzados pueden también manipular los objetos directamente de R, desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux. R proporciona un amplio abanico de herramientas estadísticas como modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación, agrupamiento, etc., y por supuesto, la obtención de una gran variedad de graficas partiendo de la información que estamos utilizando. R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. R también puede utilizarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas comerciales como MATLAB. Para mayor información sobre el lenguaje Perl visitar http://www.r-project.org/.

1.6.3 Algoritmos para la detección de CNVs.

Existe una gran variedad de software especializado para la detección de Variaciones Estructurales, mismas que se encuentran disponibles para diversas plataformas. Los hay en ambiente comercial y libre. Para el ambiente comercial podemos mencionar a GenomeStudio Software, de la compañía Illumina, Genotyping Console Software, de la compañía Affymetrix, SNP & Variation Suite (SVS), de la empresa Golden Helix, entre los más conocidos. Para los algoritmos de ambiente libre existe una gran diversidad, con características y requerimientos de entrada muy específicas, donde para nuestro caso en particular, requerimos que trabaje con SNPs-Chips y con su Señal de Intensidad (SI). De acuerdo con un estudio, estos algoritmos presentan sus ventajas y desventajas, por lo que a la hora de elegir el apropiado, debemos considerar algunos puntos importantes en el software, como el que sea más referenciado por sus buenos resultados, que presente buen soporte técnico, que funcione en varias plataformas, entre otros puntos. Dentro de estas características, los algoritmos más utilizados y eficaces son PennCNV y QuantiSNP, mismos que consideran los valores de la Señal de Intensidad para el Logaritmo de Radio R (LRR) y la Frecuencia del Alelo B (BAF), implementando un HMM similar. La salida

estándar de estas herramientas es una lista de los eventos detectados y breves resúmenes estadísticos utilizados para la comprobación de la calidad. Los datos con mucho ruido en la señal, a menudo causan predicciones con falsos positivos[25].

QuantiSNP[26] implementa un Marco Objetivo de Bayes (Objetive Bayes Framework), un método de remuestreo para establecer parámetros, aplica el método de máxima probabilidad marginal a los datos de entrenamiento para establecer otros parámetros. PennCNV[27], escribe la emisión de probabilidades del LRR y BAF dentro de la misma función de probabilidad y estima el modelo de parámetros para la maximización de probabilidad de observancia de los datos de entrenamiento. Los parámetros de probabilidad en emisión y transición son reparados en el HMM cuando se analizan los diferentes datos.

1.6.4 Bases de Datos Públicas.

Derivadas de la secuenciación del proyecto del genoma humano, existen tres servidores internacionales que albergan los datos del análisis del genoma humano: EMBL-Bank del EBI europeo, DDBJ (DNA Data Bank of Japan) en el CIB/NIG y GenBank en el NCBI de USA. Estas tres plataformas están conectadas en alianza para asegurar la disponibilidad de las secuencias al público en general y ninguna revista científica o análisis de diagnóstico molecular puede describir resultados de una secuencia de nucleótidos, proteínas o su interpretación sin hacer referencia a los depósitos de una de estas tres principales bases de datos o sus derivados. Asimismo estas bases de datos cuentas con una gran variedad de herramientas para consultar la información contenida en ellas.

Cuando se secuencian otros genomas, nacen sus propias bases de datos como por ejemplo el Proyecto de Anotación del Genoma del Arroz en USA o la Base de Datos del Genoma del Arroz en China. Para nuestro caso en particular, utilizamos como referencia la base de datos del Genoma Bovino, que apoya a la investigación del ganado bovino ofreciendo herramientas y datos del genoma bovino en la anotación UMD31.

Otros recursos disponibles, que ofrecen servicios gratuitos de consulta y comparación de información entre diversos genomas son: BioMart y Browse Genes by Organisms, donde es posible identificar el contenido de genes de regiones específicas de ADN, así como una descripción detallada de cada gen.

Otra importante herramienta bioinformática es el portal de DAVID[28], mismo que realiza una consulta sobre el enriquecimiento funcional de los genes y CNVs, utilizando la Ontologia de Genes (GO)[29] y la Enciclopedia de Genes y Genomas Kyoto (KEGG)[30]. También existen Bases de Datos especializadas en QTL's como la Bovine QTLdb[31] y la de Variaciones Estructurales Genómicas dbVar de CNBI, por mencionar algunas.

1.7 Microarreglos.

Todos los organismos vivos contienen ADN, una molécula que codifica toda la información necesaria para el desarrollo y el funcionamiento de un organismo. Encontrar y descifrar la información codificada en el ADN y la comprensión de como tal molécula simple puede dar lugar a la diversidad biológica increíble de la vida, es una meta compartida por todos los investigadores de los seres vivos.

Un microarreglo de ADN consiste en un gran número de moléculas de ADN ordenadas sobre un sustrato solido de manera que formen una matriz de secuencias en dos dimensiones. Estos fragmentos de material genético pueden ser secuencias cortas llamadas oligonucleótidos o de mayor tamaño, ADNc (ADN complementario, sintetizado a partir de ARNm), o también productos de PCR fragmentos de ADN de una sola hebra inmovilizados en el soporte, se les denomina a menudo "sondas". Los acidos nucleicos de las muestras a analizar se marcan por diversos métodos (enzimáticos, flourescentes, etc.) y se incuban sobre el panel de sondas, permitiendo la hibridación (reconocimiento y unión entre moléculas complementarias) de secuencias homologas. Durante la hibridación, las muestras de material genético marcadas, se unirán a sus complementarias inmovilizadas en el soporte del chip, permitiendo la identificación y cuantificación del ADN presente en la muestra (mutaciones, patógenos, etc.). Con posterioridad, el escáner y las herramientas informáticas nos permiten interpretar y analizar los datos obtenidos.

Los microarreglos proporcionan una visión sin precedentes en la biología del ADN, y por lo tanto una rica forma de examinar los sistemas vivos. Todas las células del cuerpo humano contienes el mismo ADN y hay cientos de diferentes tipos de células que expresan cada una, una configuración única de los genes. En este sentido el ADN podría ser descrito como existente en algunos números de estados. Los microarreglos son una herramienta utilizada para leer los estados del ADN y estos han tenido un efecto transformados en las ciencias biológicas. En el pasado los biólogos tenían que trabajar muy duro para generar pequeñas cantidad de datos que podrían ser utilizados para explorar una hipótesis con una observación a la vez. Con el advenimiento de la

tecnología de microarreglos, los experimentos individuales generan miles de datos u observaciones. La naturaleza altamente paralela de los microarreglos que se utilizan para realizar las observaciones biológicas, provoca que la mayoría de los experimentos generen más información de la que el investigador podía interpretar. Debido a que desde el punto de vista estadístico, cada gen medido en un microarreglo, es una variable independiente en un gran experimento paralelo. Para aprovechar el exceso de información en los datos de microarreglos, se han creado repositorios donde los investigadores pueden depositar sus experimentos, para que los datos estén disponibles para toda la comunidad científica, como son Expresion Genica Omnibus (GEO) y ArrayExpress por mencionar algunos.

1.7.1 Tipos de Microarreglos.

Existe una gran diversidad de microarreglos, todos con características muy especiales de acuerdo a su función específica, pero podemos clasificarlos como se observa en la Tabla No.2.

Clasificación del microarreglo según:	Características	Nombre
Material inmovilizado.	- Oligonucleotidos. - ADNc. -Proteinas. -Tejidos.	Gene Chips. cDNA Array. Protein Chip. Tissue Chip.
Diseño del arreglo.	- Personalizado. - Industrial.	Arrayers. Cassettes
Fabricacion	Densidad de Integración	- Alta Densidad - Baja Densidad
Impresión.	- Generacion de la sonda	“in situ” “depositadas”
Soporte/Tipo de union.	- No poroso / covalente. - Poroso / No covalente.	Soporte rigido (cristal y plástico). Membranas.
Aplicación.	- Secuenciacion por hibridación. - Deteccion de cambios en la expresión génica. - Cuantificacion de la expresión génica.	- Chips de secuenciación. - Chips de hibridación comparativa. - Chips de expresión.

Tabla No. 2.- Tipos de Microarreglos.

1.7.2 Microarreglos de ADN.

Un microarreglo de ADN (Chip de ADN o biochip) es una colección de puntos de ADN microscópicos unidos a una superficie sólida. Los científicos usan micromatrices de ADN para medir los niveles de expresión de un gran número de genes simultáneamente o determinar el genotipo de múltiples regiones de un genoma. Cada pozo en el arreglo contiene una pequeña secuencia específica de ADN, conocida como sonda (oligo), que pueden ser de una sección corta de un elemento de un gen u otro ADN que se utiliza para hibridizar una muestra (llamada blanco) de ADNc o ARNc. La hibridación de la sonda-blanco se detecta generalmente y se cuantifica por la detección del etiquetado de fluoroforo, plata o por el marcado de los blancos de quimioluminiscencia, para determinar la abundancia relativa de secuencias de ácidos nucleicos en el objetivo[32].

En los microarreglos estándar, las sondas se sintetizan y se unen a una superficie solida mediante un enlace covalente a una matriz química. Esta superficie solida puede ser de vidrio o de un chip de silicio, en cuyo caso así se conocen coloquialmente los chips de la compañía Affymetrix. Existe otra plataforma de microarreglo, como la tecnología que emplea la compañía Illumina, misma que utiliza perlas microscópicas, en lugar de un soporte sólido.

Los microarreglos de ADN se pueden utilizar para medir los cambios en los niveles de expresión de genes, para detectar polimorfismos de nucleótido simple (SNPs), para detectar el ADN (Hibridación Genómica Comparativa - CGH) o para determinar el genotipo o una resecuenciación especifica. También difieren en su fabricación, funcionamiento, precisión, eficiencia y costo. Sus aplicaciones podemos verlas en la Tabla No. 3.

Aplicación o Tecnología	Descripción
Expresión de Genes	Los niveles de expresión de miles de genes se controlan simultáneamente para estudiar los efectos de determinados tratamientos, enfermedades y etapas de desarrollo sobre la expresión génica.
Hibridación Genómica Comparativa	Evalúa el contenido del genoma en diferentes células o los organizas que se encuentren estrechamente relacionados.
GeneID	Pequeños microarreglos que comprueban los Id de los organismos en los alimentos, microplasmas en cultivos celulares o patógenos para la detección de enfermedades, utilizando la combinación de las tecnologías de PCR y microarreglo.
Inmunoprecipita-cion de la	Las secuencias de ADN fijadas a una proteína en particular, se pueden aislar mediante inmunoprecipitacion de la proteína (ChIP), después estos fragmentos pueden ser hibridizados en un

cromatina en ChIP	microarreglo, que permita la determinación del sitio de unión en todo el genoma.
DamID	Analoga a la del ChIP, regiones genómicas vinculadas por una proteína de interés pueden ser aisladas y utilizadas para sondas de microarreglos, para determinar la ocupación del sition de unión. A diferencia del ChIP, DamID no requiere anticuerpos pero hace uso de la metilación de adenina cerca de sitios de unión de la proteína para amplificar selectivamente dichas regiones, introducidas mediante la expresión de cantidades diminutas de proteínas de interés, fusionada a una bacteriana adenina metiltransferasa de ADN.
SNPs	Identifica los Polimorfismos de Nucleótido Simple entre los alelos o dentro de las poblaciones. Este tipo de microarreglo tiene varias aplicaciones con el resultado de genotipificación de SNPs, como análisis forenses, medición de predisposición a las enfermedades, identificación de drogas-candidatos, evaluación de mutaciones germinales en los individuos o mutaciones somaticas en cáncer, evaluación de pérdidas de la heterogosidad y análisis de desequilibrio ligado.
Deteccion de Empalme Alternativo	Utiliza una gama de diseño de unión de exones que utiliza sondas específicas para los sitios de empalme esperados o potenciales de exones predichos para un gen. Se utiliza para ensayar la expresión de formas de empalme alternativas de un gen. Los arreglos de exones tienen un diseño diferente, empleando sondas diseñadas para detectar cada exón individual para para genes conocidos.
Microarreglo de Fusion de Genes	Puede detectar transcripciones de fusión, por ejempl, a partir de especímenes de cáncer. El principio detrás de todo esto es la construcción de microarreglos de empalme alternativo. Su estrategia de diseño en oligos, permite mediciones combinadas de uniones de transcripción quiméricos con mediciones de exón-sabio de parejas de fusión individual.
Arreglo Tiling	Consiste en sondas superpuestas diseñadas densamente para representar una región genómica de interés, a veces tan grandes como todo un cromosoma humano. Su propósito es detectar empíricamente la expresión de las transcripciones o formas empalmadas alternativamente, que no pueden haber sido conocidas con anterioridad.

Tabla No. 3.- Microarreglos de ADN.

1.7.3 Microarreglos de Proteínas.

Un microarreglo de proteínas es un método de alto rendimiento, utilizado para el seguimiento de las interacciones y las actividades de las proteínas, y para determinar su función a gran escala. Su principal ventaja radica en el hecho de que un gran número de proteínas pueden ser rastreadas en paralelo. Este chip consta de una superficie de soporte tal como un portaobjetos de vidrio, membrana

de nitrocelulosa, perla o placa de microtitulación, al que se enlaza un conjunto de proteínas de captura. Las moléculas sonda, normalmente marcados con un colorante fluorescente, se añaden a la matriz. Cualquier reacción entre sonda y la proteína inmovilizada, emite una señal fluorescente que es leído por un escáner laser. Estos chips son rápidos, automatizados, económicos y altamente sensibles, consume pequeñas cantidades de muestra y reactivos[33,34].

Hay tres tipos de microarreglos de proteínas que se utilizan actualmente para estudiar las actividades bioquímicas de las proteínas, estos se describen en la Tabla No.4.

Aplicación o Tecnología	Descripción
Microarreglos Analíticos	En esta técnica una biblioteca de anticuerpos, aptameros o aficuerpos se adapta sobre la superficie soporte. Estos se utilizan como moléculas de captura, ya que cada una, atrapa específicamente a una proteína en particular. El arreglo sondea con una solución de proteína compleja, tal como un lisado celular. El análisis de las reacciones de unión resultantes, utiliza diversos sistemas de detección, que pueden proporcionar información sobre los niveles de expresión de determinadas proteínas en la muestra. Este tipo de microarreglo es especialmente útil en la comparación de la expresión de proteínas en diferentes soluciones. Por ejemplo, la respuesta de las células a un factor en particular puede ser identificado mediante la comparación de los lisados de células tratadas con sustancias especificas o cultivados bajo ciertas condiciones con lisados de células de control. Otra aplicación es la identificación y el perfilado de tejidos enfermos.
Microarreglos de Proteínas Funcionales	Conocidos también como matrices de proteína objetivo, se construyen mediante la inmovilización de grandes cantidades de proteínas purificadas y se utilizan para identificar proteína-proteína, proteina-ADN, proteina-ARN, proteina-fosfolipido y las interacciones de moléculas de proteina pequeña, de ensayo de actividad enzimática y para la detección de anticuerpos y demostrar su especificidad. Se diferencian de los microarreglos analíticos, en que los microarreglos de proteínas funcionales están compuestos de matrices que contienen proteínas funcionales de larga duración o de los dominios de la proteina. Estos chips de proteínas se utilizan para estudiar las actividades bioquímicas de todo el proteoma en un solo experimento.
Microarreglos de Proteínas de Fase Inversa (RPPA)	Este tipo de microarreglos implican muestras complejas, como lisados tisulares. Se aíslan de las células de diversos tejidos de interés y se lisan. El lisado se pone sobre el microarreglo y se sondea con anticuerpos contra la proteína blanco de interés. Estos anticuerpos se detectan normalmente con quimioluminiscncia, flurescencia o ensayos colorimétricos. Los péptidos de referencia se imprimen en las diapositivas para permitir la cuantificación de proteínas delos lisados de muestra. Estos microarreglos permiten

	la determinación de la presencia de proteínas alteradas u otros agentes que pueden ser el resultado de la enfermedad. Específicamente las modificaciones post-traduccionales, que típicamente están alterados como resultado de la enfermedad se pueden detectar utilizando RPPA.

Tabla No. 4.- Microarreglos de Proteinas.

1.7.4 Microarreglos de Tejidos (TMA).

Este tipo de microarreglo, son una innovación reciente en el campo de la patología. El método fue diseñado como una técnica de biología molecular de alto rendimiento para los investigadores que permite la evaluación de la expresión de genes candidatos relacionados a enfermedades o productos de genes simultaneas en cientos de muestras de tejido. Esta técnica permite a los patólogos llevar a cabo un análisis a gran escala mediante la inmunohostoquimica, hibridación in situ fluorescente (FISH), o ARN de hibridación in situ (ISH), en un menor tiempo y en costos menores. Esta tecnología no debe confundirse con microarreglos de ADN, donde cada pequeño punto representa un ADNc clonado único o de oligonucleitidos. En los microarreglos de tejidos, las manchas son más grandes y contienen pequeñas secciones histológicas de tejidos o tumores únicos[35].

Según ls composición de los microarreglos de tejidos, estos pueden definirse de acuerdo a la descripción de la Tabla No.5.

Microarreglos de Tejido	Descripción
TMA Multitumor.	Se componen de distintos tipos de tumores, se utilizan para filtrar alteraciones moleculares de interés. El primer ejemplo de TMA multitumor contenía 397 muestras de 17 tipos tumorales diferentes.
TMA de Progresión.	Se usan para el estudio de alteraciones moleculares en distintas etapas de un tumor en particular, por ejemplo cáncer de mama, vejiga urinaria, riñón y próstata. Casos muy particulares de su uso es el estudio del avance de un cáncer de próstata que incluya tejido normal o hiperplasico, carcicoma incidental, carcicoma localizado, carcicoma con crecimiento extraprostatico. Metástasis y recidivas.
TMA de Pronóstico.	Se constituye de tumores de los que se poseen datos clínicos y seguimiento. Con la ayuda de estos arrays, nuevos parámetros de pronósticos podrían ser identificados o se puede probar el valor de alteraciones moleculares para predecir la respuesta a la quimioterapia.

Tabla No. 5.- Microarreglos de Tejido.

1.7.5 Microarreglos de ADNc.

Un microarreglo es un conjunto de etiquetas cortas de secuencias expresadas (ESTs), a partir de una biblioteca de ADNc de un conjunto de locis conocidos (o parcilamente conocidos) de genes. Las ESTs son puestas en una placa de vidrio con cubierta antideslizante, que llegar a tener muchos miles de ESTs en una sola placa.

Su funcionamiento consiste en preparar un juego completo de las transcripciones de ARNm (transcriptoma) del tejido de un tratamiento o condición experimental, por ejemplo, los peces alimentados con una dieta alta el proteínas o en una persona con cáncer de mama. La transcripción reversiva del ADNc se prepara y se marca con tinte rojo fluorescente. Se construye una biblioteca de control a partir de una fuente no tratada, por ejemplo, una dieta de pescado estándar, o tejido de mama no canceroso; esta biblioteca se marca con un colorante fluorescente verde. Las bibliotecas experimentales y de control de hibridizan a la matriz. Un canal doble de laser excita el colorante correspondiente y la intensidad de fluorescencia indica el grado de hibridación que se ha producido. La relativa expresión génica se mide como la relación de las dos longitudes de onda fluorescentes. El aumento de expresión o regulación de los genes en el transcriptoma experimental respeto al control se visualizara como más marcado en color rojo, y la disminución de la expresión o la muestra de baja regulación, se mostrara en un color verde más claro. La intensidad del color es proporcional a la expresión diferencial. Cuando no existen cambios en la expresión, se le indicara con un color negro neutro[36]. Ver Figura No.9.

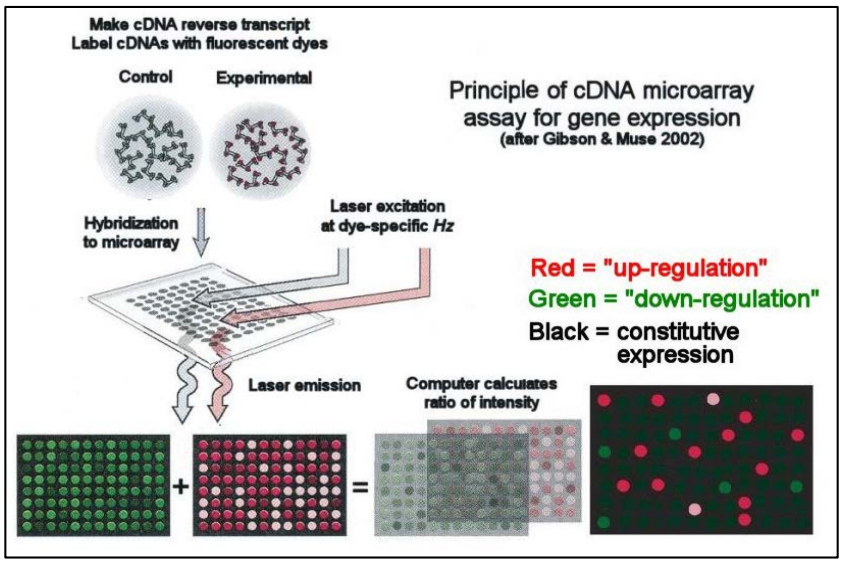


Figura No.9.- Expresión de genes en un microarreglo de ADNc.

1.7.6 Control de Calidad y Procesamiento de Datos en Microarreglos de SNPs

Después del proceso de genotipificación, al obtener lo que se conoce como los datos crudos, es necesaria una cuidadosa consideración de posibles errores en los datos, a través de la aplicación de medidas de control de calidad. La necesidad de este análisis previo, es muy relevante para garantizar la fidelidad de los resultados, particularmente para los Estudios de Asociación del Genoma Completo (GWAS) con SNPs, en donde existe una amplia gama de controles para el manejo de datos.

1.7.6.1 Microarreglos de SNPs.

Los microarreglos de SNPs son utilizados para leer las secuencias de un genoma en particular en determinadas posiciones. Este tipo particular de microarreglo es utilizado para detectar variaciones individuales y a través de poblaciones. Los oligonucleótidos son capaces de indicarnos polimorfismos, que podrían ser los responsables de variaciones genéticas dentro de una población o ser la fuente de susceptibilidad a distintas enfermedades genéticas e incluso ciertos tipos de cáncer.

La introducción de la tecnología de SNPs chips, junto con el éxito del consorcio internacional del Proyecto Mapas-Haplotipos (HapMap), ha dado lugar a una explosión de nuevos estudios de investigación que involucran la exploración del genoma completo (GWAS). Plataformas de genotipificación de SNPs de alto desempeño, ya proporcionan un

genotipificado simultáneo que va desde 500,000 hasta más de un millón de SNPs. Mientras que el genoma humano consta aproximadamente de $3x10^{9}$ bases par, la variabilidad en el genoma es capturado por un subconjunto de SNPs, en bloques de LD bien definidos. La evidencia actual, sugiere que con microarreglos a nivel del Genoma Completo de un millón de SNPs, son suficientes para caracterizar la variabilidad genética humana a través de una población.

La salida de una matriz de SNPs es un cuarteto de sonda para cada posible par de alelos para todos los SNPs en el chip. Este cuarteto de sonda consiste en un conjunto de medidas continuas: dos intensidades basadas en perfecta coincidencia, una para cada par de alelos y dos intensidades correspondientes para los desfases de sondas. Por ejemplo, para los dos alelos A y T, en cada sitio que tenemos medidas de la sonda de intensidad correspondientes a: 1) un complemento perfecto para A; 2) un complemento perfecto para T; 3) una falta de coincidencia para A y 4) una falta de coincidencia para T. La aplicación de un algoritmo produce una llamada al genotipo, por ejemplo: AA, AT o TT, sobre la base de estos datos. Tal algoritmo que muestra un rendimiento relativamente fuerte, se basa un modelo lineal ajustado robustamente. Mientras que la concordancia entre las llamadas a genotipo y los verdaderos genotipos subyacentes estén notablemente elevados utilizando un algoritmo sofisticado, una cierta tasa de error tiende a introducirse en esta etapa.

1.7.6.2 Errores de Genotipificación.

Un error de genotipificación se define como una desviación entre el verdadero genotipo subyacente y el genotipo que se observa a través de la aplicación de un enfoque de secuenciación. Estos errores se producen con mayor o menor frecuencia a través de las diferentes plataformas tecnológicas y surgen por diversas razones diferentes. El enfoque estadístico más común de detectar errores de genotipificación en los estudios basados en la población de individuos no relacionados está probado por una salida de HWE en cada uno de los SNPs investigados. Esto se puede realizar mediante una prueba X^2 (Chi Squared) o mediante una prueba exacta de Fisher para asociación[18].

1.7.6.3 Procesamiento de los Datos.

En el nivel más fundamental los microarreglos de datos son archivos de imagen de las señales fluorescentes en matrices. En su estado inicial

los datos digitales genotipificados, es la información que proporcionan las compañías que se dedican a la genotipificación de SNPs, mismas que mediante sus procesos internos extraen los datos del ADN, esta información no es útil para realizar análisis estadísticos y abordar cuestiones de interés. Los datos deben ser convertidos en patrones que sean entendibles para realizar comparaciones biológicamente significativas entre arreglos. Estos datos conocidos como datos crudos, deben pasar por un proceso conocido como "normalización", que es necesario para ajustar inicialmente estos valores y después de esto, poder explorar cualquier prueba estadística necesaria para la asociación de valores normalizados con fenotipos o al enfoque que se le esté dando al estudio. El término normalización se utiliza para describir el procesamiento de datos al cual se tiene la intención de transformar las señales de intensidad que son biológicamente comparables entre arreglos. Existe una gran cantidad de métodos propuestos para normalizar los valores que arrojan los microarreglos, uno de los más comunes para los SNPs chips, es la normalización de la Señal de la Intensidad mediante el Logaritmo de Radio R (LRR), con esta señal normalizada, más la Frecuencia del Alelo B (BAF), ya tenemos valores a escalas pequeñas que son capaces de indicar alteraciones fenotípicas en las muestras. Pero de igual manera, si se va a trabajar con la normalización de los genotipos, existen otras técnicas para imputar los valores iniciales, y obtener la información requerida. Después de que las señales en los microarreglos se han normalizado, es una práctica común, eliminar los datos de algunos conjuntos antes de los análisis estadísticos. Este proceso es conocido como filtrado de datos y se utiliza para excluir los datos que no cumplieron satisfactoriamente con el proceso de normalización, que traen errores de tecnología, o que sus valores no aportan la información suficiente para las siguientes pruebas estadísticas. Después de que los datos fueron normalizados y filtrados, nuestra información de microarreglos de SNPs, está lista para ser procesadamente por los análisis estadísticos, de acorde al estudio que estemos realizando como GWAS o LD. Preferentemente los análisis estadísticos deben proporcionar valores de significancia estadística (*Valores-P*), para que finalmente nuestros estudios sean corregidos por pruebas de comparación múltiple, como por ejemplo Benjamini o Bonferroni, que le vendrán a dar un riguroso soporte estadístico a nuestros resultados obtenidos[23].

Capítulo II

Identificación de Variaciones del Número de Copias en el Genoma Completo del Ganado Mexicano de la Raza Holstein Utilizando Microarreglos de SNPs en Alta Densidad

2.1 Resumen.

Las variaciones del número de copias (CNVs) son una gran fuente de variación estructural del genoma. Estas pueden ser consideradas como marcadores que indican rasgos económicos y fenotípicos significativos, que tienen efectos funcionales sobre la expresión génica o la susceptibilidad a enfermedades en los genomas de los mamíferos. La tecnología más utilizada para detectar estas variaciones son los microarreglos de genotipificado de polimorfismos de nucleótido simple (SNP chips). Recientemente, el genoma bovino ha sido objeto de muchos estudios sobre las tecnologías de microarreglos. Para este estudio, se inspeccionaron las CNVs en doce vacas de la raza Holstein de México, mediante el microarreglo Affymetrics Axiom Genome-Wide BOS 1 Array, que captura 648,315 SNPs y que proporcionan una gran cobertura para los estudios de genoma completo. Se aplicaron los dos algoritmos más utilizados para el descubrimiento de CNVs: PennCNV y QuantiSNP. Se encontraron un total de 56 regiones de variaciones del número de copias (CNVRs), lo que representa el 0.33% del genoma bovino (8.46Mb). Estas CNVRs varian en tamaño desde 1.5 Kb hasta 970.8 Kb, con un promedio de 151 Kb. Estas involucran a 103 genes, y al compararlas con

estudios previos de CNVRs, se traslapan un 28%. De las 56 CNVRs encontradas, 20 son nuevas CNVRs no reportadas anteriormente. En este estudio presentamos el primer análisis genómico de CNVs en ganado Mexicano, utilizando datos de SNPs de alta densidad. Nuestros resultados proporcionan una nueva referencia para futuros estudios sobre la variación y asociación genómica entre CNVs y fenotipos, especialmente en el ganado mexicano.

2.2 Introducción.

Las variaciones genéticas en los genomas de mamíferos tienen diferentes formas que van desde cambios de un solo nucleótido a grandes regiones de ADN. Una de las variaciones genéticas más importante es la variación del número de copias (CNVs), que se definen formalmente como una alteración genómica que contempla segmentos de ADN ≥ 1 Kb, y que pueden corresponder a borrados, duplicaciones, inserciones, inversiones y translocaciones [39]. La CNVs pueden llegar a abarcar grandes regiones de ADN, y estas cubren recientemente entre el 12-15% del genoma humano [40]. Con el tiempo pueden conducir a grandes efectos cambiantes a la estructura de los genes, la alteración de la regulación de los genes, la exposición de los alelos recesivos y otros mecanismos moleculares [41,42]. Variaciones fenotípicas causadas por CNVs, son objeto de estudio en muchos animales domésticos, en las que podemos mencionar: vacas [43-51], caballos [52], cerdos [53,54], ovejas [55], pollos [56], conejos [57], perros [58], entre otros. Hay varios métodos para identificar CNVs, a partir del cual podemos mencionar a escala del genoma completo: 1) los arreglos de hibridación genómica comparativa (aCGH), 2) los arreglos de genotipificado de SNPs y 3) la secuenciación de alto rendimiento (de nueva generación) [54]. Los microarreglos de genotipificado de SNPs son los más ampliamente utilizados. Tienen la ventaja de medir simultáneamente la intensidad total de la señal (Logaritmo de Radio R - LRR), la relación de la intensidad alélica (Frecuencia del Alelo B - BAF), identificar el número de copias de ADN y la pérdida de la copia neutral de heterogosidad (LOH) [59]. En el caso de la comunidad científica que investiga el ganado bovino, se ha centrado principalmente en el uso de Polimorfismos de Nucleótido Simple (SNPs) para estudios de variación en todo el genoma, convirtiéndose los arreglos de genotipificado de SNPs en la principal herramienta de investigación para la variación genética en el ganado bovino[45]. Una de las densidades disponibles comercialmente más alta, es el microarreglo de genotipificado para el ganado Axiom Genome-Wide BOS 1, de Affymetrix, Inc. (Santa Clara, CA), que captura 648,315 SNPs informativos, a través del genoma completo del ganado bovino. En este trabajo se realizó una detección de CNVs con datos de este microarreglo.

Tradicionalmente, para los estudios de detección de CNVs, los investigadores han utilizado solo un algoritmo para su detección, pero recientemente, se ha convertido en una práctica común el uso de 2 ó 3 algoritmos, con el fin de minimizar la detección de falsos positivos, asegurandose que al menos 2 de ellos coincidan y confirmen la misma posición o se traslapen en la misma CNV [46,47,60,61]. Actualmente, existen varios algoritmos comerciales y públicos para detectar CNVs, pero de los más ampliamente

utilizados podemos mencionar a PennCNV [62] y QuantiSNP [63]. Ambos algoritmos Inferien CNVs desde el LRR y el BAF, la implementación de un Modelo Oculto de Markov (HMM). QuantiSNP implementa un Marco Objetivo de Bayes. Se utiliza un método de remuestreo para establecer algunos parámetros en los antecedentes, y se aplica el método de probabilidad marginal máximo de los datos de entrenamiento para ajustar otros parámetros. En contraste, PennCNV, escribe las probabilidades de emisión de LRR y BAF en la misma función de probabilidad, y estima los parámetros del modelo mediante la maximización de la probabilidad de observar los datos de entrenamiento. Posteriormente, los parámetros de las probabilidades de transición y de emisión se fijan en el HMM al analizar diferentes datos [64]. En nuestro trabajo hemos utilizado estos dos algoritmos. Tomamos muestras de 12 vacas de la zona noroeste de México, y utilizamos el Axiom Genome-Wide BOS 1 Array para la obtención de los genotipos. Detectamos las CNVs utilizando PennCNV y QuantiSNP. En total se detectaron 56 Regiones putativas de Variaciones del Número de Copias (CNVR), que incluían a 106 genes. Estos genes están relacionados con el sistema neurológico, superficie célular de recepción, la reproducción, salud (mastitis), receptor de superficie celular ligado a la señal de transducción, entre otros. También se analizó la relación de nuestras CNVRs con Quantitative Trail Loci (QTLs) bovinos, reportados por una base de datos pública. Encontramos superposición con QTLs relacionados con la proteína de la leche, grasa de la leche, distocia, color de la carne, entre otros.

2.3 Materiales y Métodos.

2.3.1 Muestras de Animales y Genotipificado.

Se obtuvieron muestras de 12 vacas Holstein de hato de ganado experimental del Instituto de Investigación en Ciencias Veterinarias de la Universidad Autónoma de Baja California, en Mexicali, Baja California, México. Las vacas fueron seleccionados de tal manera que no se relacionaron en las últimas tres generaciones. El ADN genómico de los animales fue tomada de muestras de sangre.

Los procedimientos para la extracción y purificación de ADN fueron los establecidos por los protocolos de QIAGEN® (Purificación de ADN a partir de sangre o fluidos corporales y de purificación de ADN a partir de tejidos). Todas las muestras de ADN fueron analizadas mediante espectroscopia y la electroforesis en gel de agarosa, y se genotipo con el Axiom Genome-Wide BOS 1 Array (Affymetrix, Inc.) con una tasa promedio de llamadas a Genotipo para cada muestra individual del 99.7%. Los datos crudos de nuestro SNP chip se han depositado en la base datos pública de Expresión Génica Omnibus (http://www.ncbi.nlm.nih.gov/geo) bajo el número de acceso GSE54813.

2.3.2 Identificación de CNVs en el Ganado Bovino.

Con el fin de aumentar el nivel de confiabilidad en la detección de CNVs y para disminuir la tasa de falsos positivos se aplicaron dos de los algoritmos más precisos para la predicción de CNVs: PennCNV [62] y QuantiSNP [63]. El algoritmo de PennCNV requiere como entrada la intensidad de la señal (normalizada por el Logaritmo de Radio R, LRR), la Frecuencia Alélica B (BAF) ambos para cada marcador, y la distancia entre cada SNP (posiciones de pares de bases). El LRR y BAF se obtuvieron utilizando la directriz establecida en los protocolos de PennCNV-Affy para la detección de CNVs en arrays de Affymetrix SNP (http://www.openbioinformatics.org/penncnv/penncnv_ tutorial_ affy_gw6.html) y el uso de las Affymetrix Power Tools (APT) (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx). PennCNV también incluye un argumento llamado *GCmodel*, el cual utiliza un modelo de regresión para ajustar el alto contenido de GC, recupera las muestras afectadas por la "ola de genómica" [65]. El archivo del modelo de GC para este estudio se ha generado por medio un script en Perl que calcula el contenido de GC dentro de 1 Mb alrededor de cada marcador (500Kb cada lado), y las olas de genómicas se ajustaron utilizando el argumento *-gcmodel*. PennCNV fue ejecutado con la opción *-test*, teniendo en cuenta que no había ninguna relación entre las muestras y la información de trío/pedigrí no se incluyó. Se aplicó sólo a los 29 cromosomas autosómicos con la opción *-lastchr29*, usando los valores por defecto (desviación estándar LRR de 0.30, el BAF drift de 0.01 y el wavines factor de 0.05). QuantiSNP se ejecutó con las opciones *-isaffy* y *–levels* habilitados ya que utilizamos un microarreglo de Affymetrix. De la misma forma fue habilitada la opción *-gcdir* para llevar a cabo la corrección del logaritmo de Radio R, en los marcadores afectados por las ondas genómicas.

Para declarar una CNV como putativa consideramos que al menos 3 SNPs estuvieran adyacentes indicando una pérdida o ganancia, con una longitud total superior a igual a 1 Kb, que fuera detectada simultáneamente por los dos algoritmos en el mismo animal, ya sea en la misma posición o que se traslapen, y finalmente las Regiones de CNVs (CNVRs) se definieron sobre la base de los criterios utilizados en un estudio previo (Redon et al, 2006). Ver Figura No. 10.

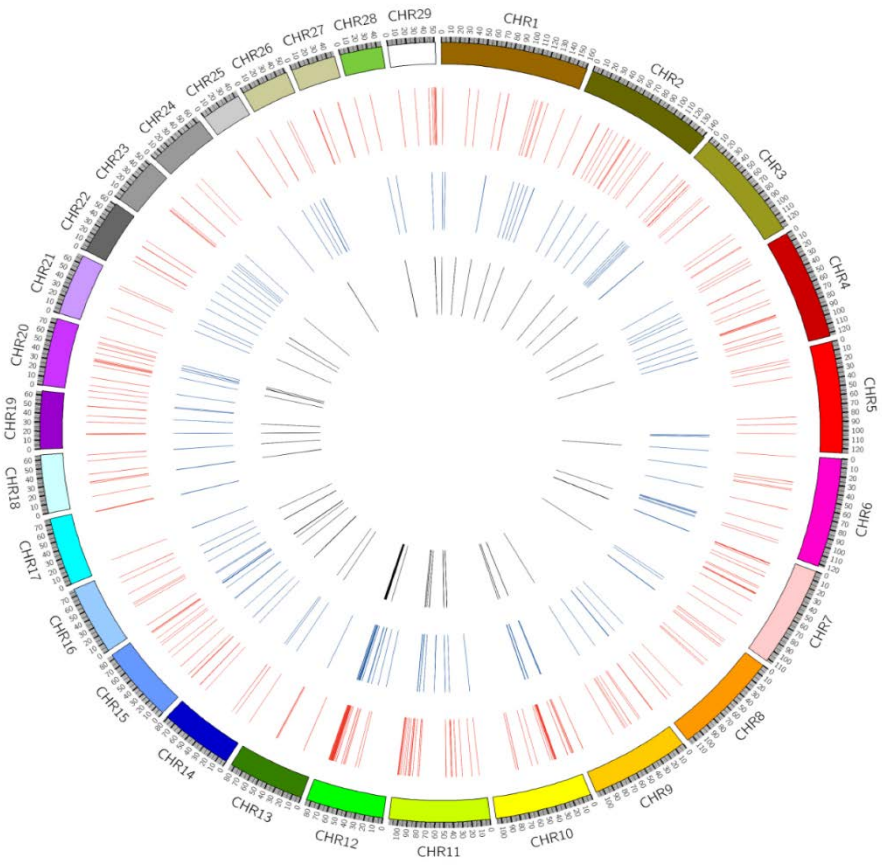


Figura No. 10: Las CNVs detectadas por los algoritmos PennCNV y QuantiSNP a través de los cromosomas autosomales. El primer círculo interno en líneas rojas, representa las 302 CNVs detectadas por el algoritmo QuantiSNP. El segundo circulo interno en color azul, representa las 155 CNVs identificadas por el algoritmo PennCNV. Y el tercer circulo interno en líneas negras, representa las 77 CNVs en las cuales ambos algoritmos coinciden.

2.3.3 Validación de las CNVs mediante qPCR.

Con el fin de confirmar la exactitud de nuestra predicción de CNVs, se utilizó un PCR en tiempo real (qPCR) para validar 7 CNVRs seleccionados de entre los 56 detectadas en este estudio. De las CNVRs seleccionados, 3 eran una simple copia duplicada (CNVRs 1, 16 y 35), 2 eran borrados de una copia (CNVRs 2 y 11) y 2 eran duplicaciones de copia dobles (CNVR 21 y 55). Para cada CNVR blanco, dos pares de cebadores fueron diseñados teniendo en cuenta los límites de cada CNVR. Los cebadores para PCR fueron diseñados utilizando la herramienta Primer-BLAST de NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome). (Ver Tabla No. 6 y para mayores

detalles ver Tabla S1 en Archivo Adicional No. 1). Todos los cebadores se diseñaron basándose en la secuencia de referencia Bos taurus UMD3.1.

No. CNVR	Chr.	Start(bp)	End(bp)	Length(bp)	Type CNV	Animal ID.	Forward (5'-3')	Reverse (5'-3')
1	1	1971535	2003454	31919	Single copy duplication	HOL-10	CGTGCTGTTTTGTCAGGTGG	TGGAGTGTGCTTAGCAGGTG
							CATGGAAAGCACTTGGTGCC	CTGTGTTTTGTGGTCTGCCG
2	1	34613754	34617353	3599	Deletion of one copy	HOL-9	GGAAACCAAAAGGATGTGCCC	TCCTACTCACAAGACTGCCCT
							AGTGGCAGGAATGGTCAACA	CACCATCCTTGTATAGGAGGTCAG
11	4	33547395	33556611	9216	Deletion of one copy	HOL-9	GCAGGTGTTTGGAATGGTGG	CTGTGGGCTCTTCATGTCCA
							GCATGTGCCAAAGTGTGGTG	GCAAGCTACCCAAGGTCACA
16	7	97442762	97444260	1498	Single copy duplication	HOL-12	TGCTGAGAGCATTGGAACGG	CACTTGGACACGTTGGCCT
							AGCACAGAAGGATACAGACGA	CTTACAATGGGGCCGTTCCA
21	11	39886500	39918145	31645	Double copy duplication	HOL-12	GCTAGTGGGTAGCTCAGCAG	GCTGATGCATCTGTTGCCAG
							CCACCTTTTTCAGGGTGGGA	CCAACCACGATGTGATGGGA
35	14	22348562	22393497	44935	Single copy duplication	HOL-10	GGGGTGGTCGTAACTTCCAG	ATGCCCTTAGACCACCAAGC
							GTCTCCAGGTGACAGAGCAC	TCCAGTTCCTAGACTCGGGG
55	28	22706652	22725331	18679	Double copy duplication	HOL-9	TGACCAGAGCTGTGTGGTTC	AGCTCATGGAGGATCCTGGT
							AGAACACACAGGCATGCTCA	ACCCTCAGTCTGTTGCTTGG
Control Gene - *BTF3*							TCCTGCCATTCCCTTCACAC	GAACCAGGAGAAACTCGCCA
							CAGAGCTTCAGCCAGTCTCC	CCACGCTGAGACAAAGCAAC

Tabla No. 6.- Información sobre las 7 CNVRs y los cebadores que fueron utilizados para el experimento con qPCR.

Las reacciones se realizaron por triplicado con un volumen de 20 µl en un LightCycler CFX connect, usando los siguientes reactivos: 10 µl de itaq^{tm} Universal SYBR Green supermix (Biorad), 1 µl de ADN (aproximadamente 50 ng), 1 µl (20 pM/µL) para ambos cebadores (adelante y atrás), 10 µl de Master

Mix (2x) y agua. La PCR corrió como sigue: 5 minutos a 95°C seguido de 40 ciclos a 95°C durante 10 segundos, y 60°C durante 10 segundos. La eficiencia se probó para cada par de cebadores en tres puntos de la disolución. Se utilizó el Factor de Transcripción Básico (BTF3) como un gen de control para comparar el número de copias en cada CNVR, como se realizó en estudios previos [45-47]. Se utilizó el método de ciclo umbral comparativo ($2^{-\Delta\Delta Ct}$) para cuantificar el número de cambios de las copias, mediante la comparación del valor ΔCt (ciclo umbral (Ct) de la región del blanco menos la región de control Ct) de las muestras con CNV a una ΔCt de un calibrador sin CNV [47,66,67].

Se calculó el valor promedio de Ct de tres repeticiones para cada muestra, una vez normalizado, se comparó contra el gen de control, con la suposición de la existencia de dos copias del segmento de ADN en la región de control. Para cada CNVR a validar, se calculó un valor de fórmula $2x2^{-\Delta\Delta Ct}$ para cada individuo [50]. El valor obtenido se utiliza para decidir si una CNVR estaba en un estado normal (sin CNVR, si el valor fue de alrededor de 2), un estado de ganancia (si el valor fue de aproximadamente 3 o superior), o un estado de eliminación (si el valor fue de cerca de 0 ó 1) [47]. Vea la Figura No. 13 y en la Tabla No. 7 se muestran los resultados de las pruebas de la qPCR.

No. CNVR	Chr	Start	End	Type	Average normalized value	Value expected	Validated by PCR
1	1	1971535	2003454	Gain	2.7	3	Yes
2	1	34613754	34617353	Loss	1.2	1	Yes
11	4	33547395	33556611	Loss	1.7	1	No
16	7	97442762	97444260	Gain	3.1	2	Yes
21	11	39886500	39918145	Gain	2.2	3	No
35	14	22348562	22393497	Gain	2.7	3	Yes
55	28	22706652	22725331	Gain	4.2	4	Yes

Tabla No. 7: Resultados del análisis cuantitativo del PCR en tiempo real de las 7 CNVRs confirmadas. Los valores esperados alrededor de 1, indican una copia perdida, alrededor de 3 indican 3 copias duplicadas y alrededor de 4, indican 4 copias duplicadas.

2.4 Resultados.

2.4.1 Detección de CNVs en el Genoma Completo.

En este estudio se analizaron los datos de genotipos de 648,315 SNPs de 12 muestras de ganado mexicano de la raza Holstein y aplicamos los algoritmos PennCNV y QuantiSNP para la detección de CNVs. El primer algoritmo, PennCNV, detectó 155 CNVs, mientras que el segundo, QuantiSNP, detectó 302. Ambos algoritmos coincidieron en 77 CNVs detectadas en la misma posición y la misma muestra (ver Figura No. 10). Llamamos a estas CNVs putativas. Los 77 CNVs putativas fueron localizados en 22 cromosomas autosómicos en las 12 muestras (Tabla No.8 y para mayores detalles ver Tabla S2 en Archivo Adicional No. 1). El número promedio de CNV por muestra fue de 6.41% y el número medio de CNVs por cromosoma fue de 3.5. Se inspeccionaron las 77 CNVs que se traslapaban y definimos 56 Regiones de Variaciones del Número de Copias (CNVRs), que cubren el 0.33% (8.46 Mb) del genoma bovino (Tabla No. 9 y para mayores detalles ver Tabla S3 en Archivo Adicional No. 1). La longitud de las CNVRs varió de 1.5 Kb a 970.81 Kb con un tamaño promedio de 151.11 kb y una mediana de 51.6 Kb. La Figura No. 11 muestra la distribución de los tamaños de CNVRs, donde podemos notar que el tamaño más abundante oscila entre 10 y 50 Kb.

No.	Chr.	Start	End	Size (bp)	SNPs	Animals	Description
1	1	1971535	2003454	31920	5	HOL-6	Single copy duplication
2	1	1971535	2003454	31920	5	HOL-10	Single copy duplication
3	1	34613754	34617353	3599	3	Hol-9	Deletion of one copy
4	1	72480181	72559153	78973	9	HOL-6	Deletion of one copy
5	1	118627858	118649176	21319	4	HOL-2	Single copy duplication
6	1	153322793	153344166	21374	11	HOL-9	Single copy duplication
7	2	57747240	57750888	3649	3	HOL-9	Single copy duplication
8	2	124356632	124384127	27496	7	HOL-5	Deletion of one copy
9	2	124360687	124394326	33640	7	HOL-8	Deletion of one copy
10	3	18475187	18502234	27048	7	HOL-6	Single copy duplication
11	3	54578410	54820361	241952	17	HOL-3	Single copy duplication
12	4	2850063	2854906	4844	4	HOL-6	Deletion of one copy
13	4	33547395	33556611	9217	5	HOL-9	Deletion of one copy
14	4	106058968	106151480	92513	25	HOL-10	Single copy duplication
15	6	1065851	1093395	27545	6	HOL-6	Deletion of one copy
16	6	111951028	111966349	15322	7	HOL-9	Single copy duplication
17	7	10350329	11099631	749303	45	HOL-8	Single copy duplication
18	7	10587803	11040264	452462	31	HOL-5	Single copy duplication

19	7	97442762	97444260	1499	3	HOL-12	Single copy duplication
20	9	45449839	45464467	14629	5	HOL-4	Single copy duplication
21	10	1513111	1536142	23032	7	HOL-8	Deletion of one copy
22	10	1513111	1559113	46003	11	HOL-12	Single copy deletion
23	10	1524489	1559113	34625	9	HOL-11	Deletion of one copy
24	10	1524489	1559113	34625	9	HOL-10	Deletion of one copy
25	10	1524489	1559113	34625	9	HOL-12	Deletion of one copy
26	10	22443548	22510718	67171	5	HOL-12	Deletion of one copy
27	10	22443548	22455901	12354	4	HOL-5	Single copy duplication
28	10	27063526	27095789	32264	9	HOL-7	Deletion of one copy
29	11	39886500	39918145	31646	6	HOL-10	Double copy duplication
30	11	39886500	39918145	31646	6	HOL-12	Double copy duplication
31	11	47098882	47198702	99821	24	HOL-10	Deletion of one copy
32	11	47148092	47198279	50188	13	HOL-5	Deletion of one copy
33	11	82196097	82374760	178663	26	HOL-3	Deletion of one copy
34	11	88281698	88313232	31535	6	HOL-2	Deletion of one copy
35	11	91921125	92310890	389766	93	HOL-4	Single copy duplication
36	11	92104404	92340429	236026	70	HOL-4	Single copy duplication
37	12	57687270	57719962	32693	5	HOL-9	Deletion of one copy
38	12	70957641	71223810	266170	6	HOL-6	Single copy duplication
39	12	71383094	71604231	221138	4	HOL-6	Deletion of one copy
40	12	72184389	72263216	78828	7	HOL-9	Deletion of one copy
41	12	72407022	73327391	920370	9	HOL-12	Single copy duplication
42	12	73475758	73531526	55769	2	HOL-6	Full deletion
43	12	73475758	73721049	245292	4	HOL-10	Single copy duplication
44	12	74416147	75386962	970816	47	HOL-1	Single copy duplication
45	12	75363034	75367423	4390	3	HOL-9	Single copy duplication
46	12	75957893	76041136	83244	3	HOL-10	Single copy duplication
47	12	75993484	76729025	735542	13	HOL-6	Single copy duplication
48	14	2746730	2764767	18038	8	HOL-1	Single copy duplication
49	14	22348562	22393497	44936	9	HOL-10	Single copy duplication
50	15	5371776	5674210	302435	69	HOL-3	Deletion of one copy
51	15	47876389	47892054	15666	4	HOL-1	Deletion of one copy
52	15	47908320	47958138	49819	5	HOL-1	Deletion of two copies
53	15	70395589	70408555	12967	6	HOL-1	Deletion of two copies
54	15	80422173	80696516	274344	8	HOL-1	Deletion of one copy
55	15	80666234	80816572	150339	4	HOL-9	Deletion of one copy
56	15	80602302	80837237	234936	7	HOL-5	Single copy duplication
57	15	81018176	81075360	57185	6	HOL-9	Deletion of one copy
58	16	32899795	33028883	129089	26	HOL-2	Deletion of one copy
59	16	62025927	62179726	153800	31	HOL-4	Single copy duplication
60	18	27908043	28375735	467692	147	HOL-6	Single copy duplication
61	18	62340260	62533770	193510	4	HOL-10	Single copy duplication
62	19	16919639	17002661	83023	23	HOL-9	Single copy duplication
63	19	16919639	16956606	36968	14	HOL-6	Single copy duplication
64	19	39619531	39689562	70032	13	HOL-4	Single copy duplication
65	20	39355058	39389770	34713	7	HOL-11	Single copy duplication

66	20	55548628	55694959	146332	36	HOL-10	Deletion of one copy
67	20	55548628	55694959	146332	36	HOL-8	Deletion of one copy
68	20	55554731	55694959	140229	35	HOL-12	Deletion of one copy
69	20	60473408	60511209	37802	9	HOL-5	Single copy duplication
70	21	52306422	52328502	22081	6	HOL-8	Deletion of one copy
71	22	22392164	22397908	5745	5	HOL-6	Deletion of one copy
72	22	37186652	37231984	45333	7	HOL-11	Single copy duplication
73	23	29468816	29489851	21036	7	HOL-9	Deletion of one copy
74	25	36834075	36873499	39425	6	HOL-10	Single copy duplication
75	28	22706652	22725331	18680	4	HOL-9	Double copy duplication
76	28	22706652	22718082	11431	3	HOL-4	Double copy duplication
77	29	42455680	42514948	59269	7	HOL-6	Single copy duplication

Tabla No. 8.- Detalle de cada CNV detectada en este estudio.

CNVR	Chr	Start	End	Size	Type	frequency	GC percent inside
1	1	1971535	2003454	31919	Single copy duplication	16.66	42.82
2	1	34613754	34617353	3599	Deletion of one copy	unique	40.91
3	1	72455730	72559153	103423	Deletion of one copy	unique	45.05
4	1	118627858	118649176	21318	Single copy duplication	unique	33.99
5	1	153322793	153344166	21373	Single copy duplication	unique	44.46
6	2	57747240	57750888	3648	Single copy duplication	unique	38.22
7	2	124356632	124394326	37694	Deletion of one copy	16.66	45.04
8	3	18475187	18502234	27047	Single copy duplication	unique	38.75
9	3	54578410	54820361	241951	Single copy duplication	unique	40.27
10	4	2850063	2854906	4843	Deletion of one copy	unique	37.67
11	4	33547395	33556611	9216	Deletion of one copy	unique	34.78
12	4	106058968	106151480	92512	Single copy duplication	unique	41.71
13	6	1065851	1093395	27544	Deletion of one copy	unique	37.46
14	6	111951028	111967854	16826	Single copy duplication	unique	39.60
15	7	10350329	11126470	776141	Single copy duplication	16.66	38.93
16	7	97442762	97444260	1498	Single copy duplication	unique	39.02
17	9	45337323	45488218	150895	Single copy duplication	unique	40.35
18	10	1513111	1559113	46002	Deletion of one copy	33.33	41.42
19	10	22442126	22510718	68592	Deletion of one copy	16.66	40.90
20	10	27063526	27095789	32263	Deletion of one copy	unique	38.40
21	11	39886500	39918145	31645	Double copy duplication	16.66	36.54
22	11	47082297	47198702	116405	Deletion of one copy	16.66	41.86
23	11	82196097	82374760	178663	Deletion of one copy	unique	45.21
24	11	88281698	88313232	31534	Deletion of one copy	unique	41.63
25	11	91921125	92340429	419304	Single copy duplication	unique	43.21
26	12	57687270	57719962	32692	Deletion of one copy	unique	37.81
27	12	70957641	71223810	266169	Single copy duplication	unique	41.02

28	12	71383094	71604231	221137	Deletion of one copy	unique	40.23
29	12	72184389	72263216	78827	Deletion of one copy	unique	42.17
30	12	72407022	73327391	920369	Single copy duplication	unique	41.10
31	12	73475758	73721049	245291	Both	16.66	40.22
32	12	74416147	75386962	970815	Single copy duplication	16.66	41.10
33	12	75957893	76729025	771132	Single copy duplication	16.66	40.88
34	14	2746730	2764767	18037	Single copy duplication	unique	46.47
35	14	22348562	22393497	44935	Single copy duplication	unique	38.13
36	15	5371776	5674210	302434	Deletion of one copy	unique	36.90
37	15	47876389	47958138	81749	Deletion of two copies	unique	37.86
38	15	70395589	70408555	12966	Deletion of two copies	unique	37.77
39	15	80422173	80837237	415064	Both	25	38.25
40	15	81018176	81075360	57184	Deletion of one copy	unique	38.55
41	16	32899795	33028883	129088	Deletion of one copy	unique	42.24
42	16	62025927	62179726	153799	Single copy duplication	unique	39.57
43	18	27908043	28375735	467692	Single copy duplication	unique	38.85
44	18	62340260	62533770	193510	Single copy duplication	unique	57.95
45	19	16919639	17002661	83022	Single copy duplication	16.66	47.05
46	19	39619531	39689562	70031	Single copy duplication	unique	48.00
47	20	39355058	39389770	34712	Single copy duplication	unique	40.97
48	20	55548628	55694959	146331	Deletion of one copy	25	38.43
49	20	60473408	60511209	37801	Single copy duplication	unique	38.33
50	21	52306422	52328502	22080	Deletion of one copy	unique	37.00
51	22	22392164	22397908	5744	Deletion of one copy	unique	41.02
52	22	37186652	37231984	45332	Single copy duplication	unique	42.88
53	23	29468816	29489851	21035	Deletion of one copy	unique	37.93
54	25	36834075	36873499	39424	Single copy duplication	unique	46.89
55	28	22706652	22725331	18679	Double copy duplication	16.66	36.00
56	29	42455680	42514948	59268	Single copy duplication	unique	44.46
			Average	151111			**40.79**
			Total	**8462204**			

Tabla No. 9.- Características detalladas de las CNVRs en los cromosomas autosomales identificadas en este estudio.

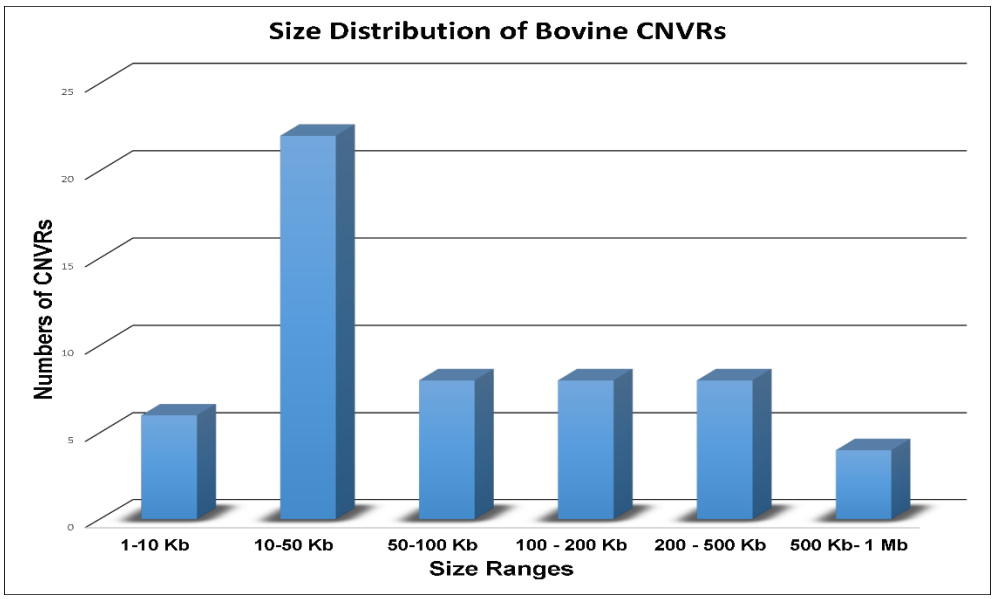
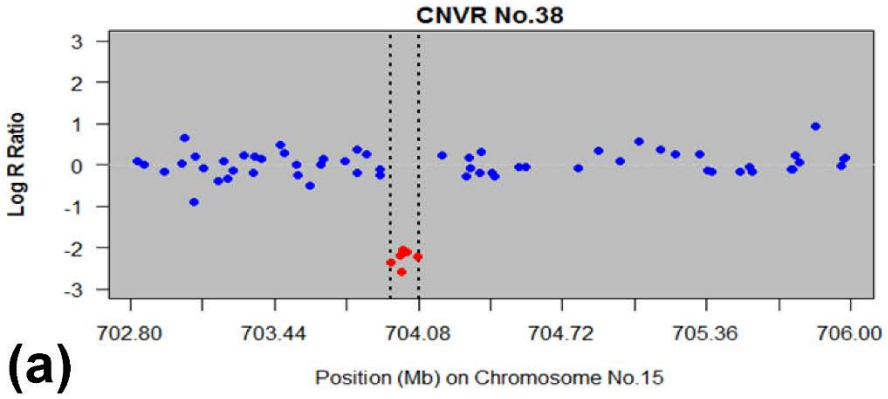


Figura No. 11.- Distribución de tamaños en las CNVRs detectadas.

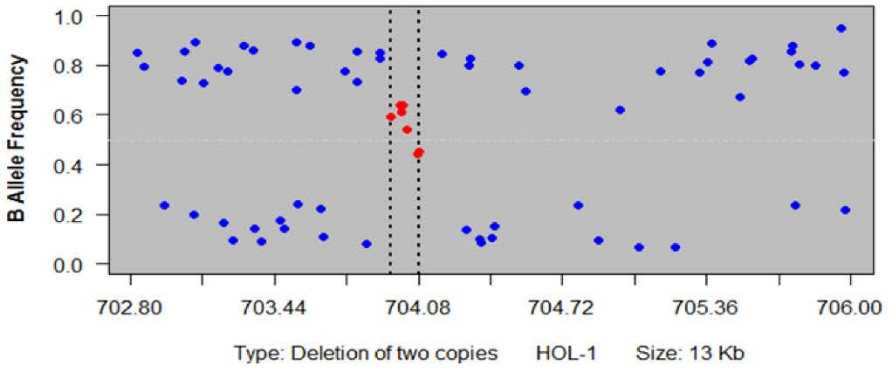
En la Tabla No. 10 se presentan las estadísticas de las CNVRs. Entre estas CNVRs, se encontraron 24 borrados, 30 inserciones, y 2 del Tipo borrado-inserción, de las cuales no se distribuyen de manera uniforme en todo el genoma. Entre los cromosomas, la proporción cubierta por CNVRs oscila desde 0.03 hasta 3.8%, dando como resultado el cromosoma 12 con la proporción más alta (3.8%), y el cromosoma 2 con la proporción más baja (0.03%). La mayor CNVR-Borrado fue de 0.30 Mb en el cromosoma 15 de la muestra número 3, mientras que el mayor CNVR-Inserción fue 0.97 Mb en el cromosoma 12 de la muestra número 1. El cromosoma con el mayor número de CNVRs fue el cromosoma 12, con 8, mientras que el cromosoma con el menor número de CNVRs fueron 9, 21, 23, 25, 28, y 29 con una CNVR por cromosoma. A partir de los 56 CNVRs, 20 fueron nuevas regiones de CNVs (no encontrada en estudios anteriores), lo que representa 35.71% del total de CNVRs detectadas. A partir de estos nuevas CNVRs, 10 fueron de borrados y 10 fueron de inserciones. En la siguiente sección se presenta el contenido de genes y el análisis funcional de todas las CNVRs. La Figura No. 12 muestra los valores del LRR y BAF a partir de dos CNVRs que se encuentran en los cromosomas 11 y 15.

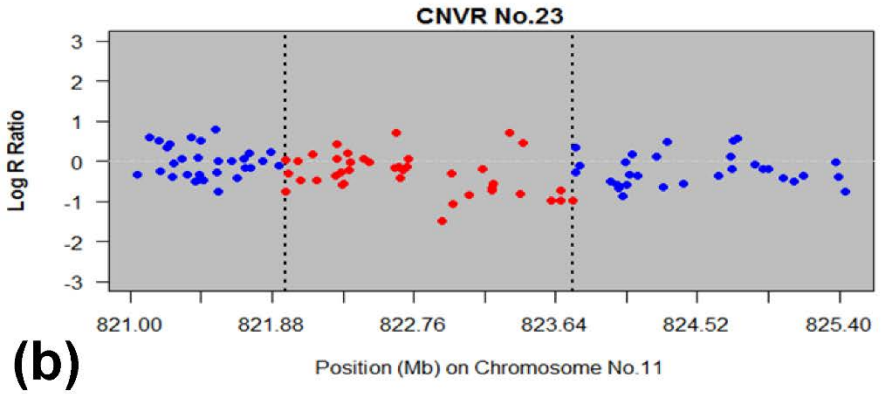
Type	CNVRs	Mean Size	Median Size	Size Range	CNVR content	Sequence covered	%CNVR
Loss	25	71,458	37,694	343 - 302,434	1,789,724		0.071
Gain	29	207,314	59,268	1,498 - 970,815	6,012,125	2,512,082,506	0.239
Both	2	330,377	330,377	245,291 - 415,064	660,355		0.026
All	56	151,052	51,593	343 - 970,815	8,462,204		0.337

Tabla No.10.- Características de las Regiones de CNVs, con tamaños en bases par (bp).


CNVR No.38
Log R Ratio
3
2
1
0
-1
-2
-3
702.80
703.44
704.08
704.72
705.36
706.00
Position (Mb) on Chromosome No.15

(a)


B Allele Frequency
1.0
0.8
0.6
0.4
0.2
0.0
702.80
703.44
704.08
704.72
705.36
706.00
Type: Deletion of two copies HOL-1 Size: 13 Kb


CNVR No.23
Log R Ratio
3
2
1
0
-1
-2
-3
821.00
821.88
822.76
823.64
824.52
825.40
Position (Mb) on Chromosome No.11

(b)

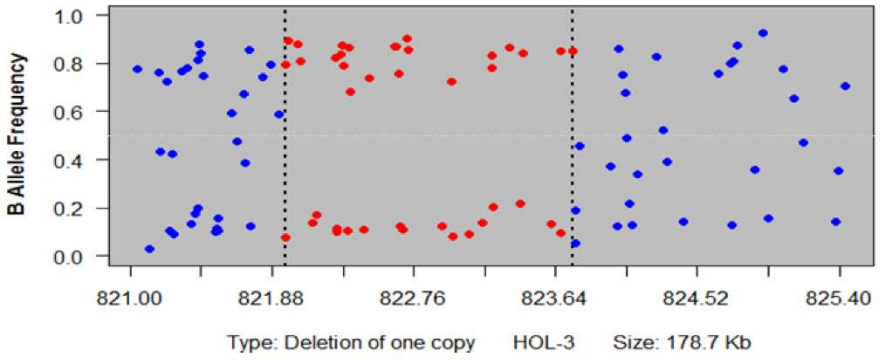

B Allele Frequency
1.0
0.8
0.6
0.4
0.2
0.0
821.00
821.88
822.76
823.64
824.52
825.40
Type: Deletion of one copy HOL-3 Size: 178.7 Kb

Figura No. 12: El LRR y el BAF muestran dos CNVRs. La combinación de ambos valores se utilizan para identificar CNVs. cada par de graficas muestra diferente posición en diferentes cromosomas. Los puntos en rojo dentro de la línea punteada representan una CNV, las líneas verticales punteadas indican los límites de cada región de CNV y los puntos en azul, representa los valores esperados en el cromosoma. (a) los valores bajos de LRR (menores que -1) y los valores del BAF agrupados alrededor de 0.5 indican un borrado de dos copias en una región del cromosoma no. 15. (b) algunos valores bajos (debajo de -1) en los LRR y la ausencia de valores alrededor de 0.5 indican una copia borrada en una región del cromosoma 11.

2.4.2 Contenido de Genes y Análisis Funcional.

Se utilizó la base de datos BioMart (http://www.biomart.org) para identificar el contenido de genes dentro de las regiones cubiertas por CNVRs y la base de datos RefGen (http://refgene.com) para encontrar la descripción para cada gen (ver Tabla No. 11 y para mayores detalles ver Tabla S4 en archivo adicional No. 1). Las CNVRs cubren un total de 103 genes, de los cuales 96 codifican proteínas, 2 son pseudo genes, 3 son ARNns y 2 son ARNmi. Estos genes se distribuyen dentro de 37 CNVRs (66%), mientras que el resto, 19 CNVRs (34%) no contiene ninguna anotación de genes.

No. CNVR	Chr	Start	End	Size	Type	Gene ID.	Associated Gene Name	Gene Biotype
1	1	1971535	2003454	31919	Single copy duplication			
2	1	34613754	34617353	3599	Deletion of one copy			
3	1	72455730	72559153	103423	Deletion of one copy	507784	*BT.87090*	protein coding
							no data	miRNA
4	1	118627858	118649176	21318	Single copy duplication	541277	*BT.70150*	protein coding
5	1	153322793	153344166	21373	Single copy duplication	131873	*COL6A6*	protein coding
6	2	57747240	57750888	3648	Single copy duplication			
7	2	124356632	124394326	37694	Deletion of one copy			
8	3	18475187	18502234	27047	Single copy duplication	388698	*FLG2*	protein coding
						788654	*HRNR*	protein coding
9	3	54578410	54820361	241951	Single copy duplication		*no data*	protein coding
						613313	*GBP4*	protein coding
						533657	*GBP6*	protein coding
							BT.37579	protein coding
							U2	snRNA
10	4	2850063	2854906	4843	Deletion of one copy			
11	4	33547395	33556611	9216	Deletion of one copy		*KIAA1324L*	protein coding
12	4	106058968	106151480	92512	Single copy duplication	615201	*CLEC5A*	protein coding

							MGAM	protein coding
						787789	TAS2R38	protein coding
13	6	1065851	1093395	27544	Deletion of one copy			
14	6	111951028	111967854	16826	Single copy duplication			
15	7	10350329	11126470	776141	Single copy duplication		*no data*	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
						390892	OR7A10	protein coding
						26333	OR7A17	protein coding
							no data	protein coding
							no data	protein coding
16	7	97442762	97444260	1498	Single copy duplication			
17	9	45337323	45488218	150895	Single copy duplication	64208	POPDC3	protein coding
						539988	BVES	protein coding
							BT.67847	protein coding
18	10	1513111	1559113	46002	Deletion of one copy		BT.32278	protein coding
19	10	22442126	22510718	68592	Deletion of one copy		BT.104278	protein coding
							no data	protein coding
20	10	27063526	27095789	32263	Deletion of one copy		*no data*	protein coding
							no data	protein coding
							no data	protein coding
							no data	miRNA
21	11	39886500	39918145	31645	Double copy duplication			
22	11	47082297	47198702	116405	Deletion of one copy		IGKC	protein coding
							BT.91557	protein coding
							IGKV5-2	protein coding
23	11	82196097	82374760	178663	Deletion of one copy			
24	11	88281698	88313232	31534	Deletion of one copy			
25	11	91921125	92340429	419304	Single copy duplication		*no data*	protein coding
						527409	MGC151949	protein coding
						281780	GGTA1	protein coding
26	12	57687270	57719962	32692	Deletion of one copy		*no data*	protein coding
27	12	70957641	71223810	266169	Single copy duplication			
28	12	71383094	71604231	221137	Deletion of one copy			
29	12	72184389	72263216	78827	Deletion of one copy		BT.63792	protein coding
30	12	72407022	73327391	920369	Single copy duplication		*no data*	protein coding
31	12	73475758	73721049	245291	Both		BT.61919	protein coding

32	12	74416147	75386962	970815	Single copy duplication		BT.82323	protein coding
33	12	75957893	76729025	771132	Single copy duplication			
34	14	2746730	2764767	18037	Single copy duplication		LY6K	protein coding
35	14	22348562	22393497	44935	Single copy duplication	517353	SNTG1	protein coding
36	15	5371776	5674210	302434	Deletion of one copy		DYNC2H1	protein coding
							BT.53810	protein coding
37	15	47876389	47958138	81749	Deletion of two copies		OR52N5	protein coding
							no data	protein coding
							no data	protein coding
							U6	snRNA
38	15	70395589	70408555	12966	Deletion of two copies			
39	15	80422173	80837237	415064	Both		*no data*	protein coding
							no data	protein coding
							no data	pseudogene
							OR5J2	protein coding
							OR8K5	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							OR8J1	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
							no data	protein coding
40	15	81018176	81075360	57184	Deletion of one copy	513633	OR5M3	protein coding
							OR5M11	protein coding
41	16	32899795	33028883	129088	Deletion of one copy	615715	EFCAB2	protein coding
42	16	62025927	62179726	153799	Single copy duplication	504287	SOAT1	protein coding
							AXDND1	protein coding
43	18	27908043	28375735	467692	Single copy duplication			
44	18	62340260	62533770	193510	Single copy duplication	507895	U2AF2	protein coding
						618485	CCDC106	protein coding
						789593	ZNF784	protein coding
							ZNF865	protein coding
							no data	pseudogene
							BT.44430	protein coding
						615123	SBK2	protein coding
							SSC5D	protein coding
						532809	NAT14	protein coding

							ZNF628	protein coding
						516212	ISOC2	protein coding
							SHISA7	protein coding
							ZNF581	protein coding
							ZNF580	protein coding
							U6	snRNA
45	19	16919639	17002661	83022	Single copy duplication	617930	ACCN1	protein coding
46	19	39619531	39689562	70031	Single copy duplication		BT.36663	protein coding
						787335	GPR179	protein coding
							SOCS7	protein coding
47	20	39355058	39389770	34712	Single copy duplication	525869	RAI14	protein coding
48	20	55548628	55694959	146331	Deletion of one copy		*no data*	protein coding
49	20	60473408	60511209	37801	Single copy duplication			
50	21	52306422	52328502	22080	Deletion of one copy			
51	22	22392164	22397908	5744	Deletion of one copy			
52	22	37186652	37231984	45332	Single copy duplication			
53	23	29468816	29489851	21035	Deletion of one copy		*no data*	protein coding
54	25	36834075	36873499	39424	Single copy duplication	617785	PILRA	protein coding
						505637	CNPY4	protein coding
							BT.59351	protein coding
55	28	22706652	22725331	18679	Double copy duplication		BT.102825	protein coding
56	29	42455680	42514948	59268	Single copy duplication		*no data*	protein coding
				8462204				

Tabla No. 11.- Genes que se encuentran dentro o que se traslapan en las Regiones de CNVs identificadas en este estudio.

Con el fin de analizar el enriquecimiento funcional de las CNVRs se realizaron búsquedas de ontología de genes [68] y la Enciclopedia de Kyoto de genes y genomas (KEGG) [69]. Ambos análisis se llevaron a cabo utilizando la herramienta bioinformática DAVID [70]. El análisis GO mostró términos de genes comunes entre los mamíferos, por ejemplo, la percepción sensorial, la cognición, receptor olfativo, proceso del sistema neurológico, la proteína del receptor acoplado a proteína G vía de señalización, y receptor de superficie celular ligado transducción de la señal. Mientras que el análisis vía KEGG reveló que estos genes están representados principalmente en la vía de transducción olfativa (Tabla No. 12 y para mayores detalles ver Tabla S5 en archivo adicional no. 1).

Category	Term	Count	%	P-Value	Benjamini
GOTERM_BP_FAT	sensory perception of smell	11	18,3	1,6E-7	2,1E-5
GOTERM_BP_FAT	sensory perception of chemical stimulus	11	18,3	4,1E-7	2,7E-5
GOTERM_BP_FAT	G-protein coupled receptor protein signaling pathway	13	21,7	2,6E-5	1,2E-3
GOTERM_BP_FAT	sensory perception	11	18,3	4,4E-5	1,5E-3
GOTERM_BP_FAT	cognition	11	18,3	1,2E-4	3,1E-3
GOTERM_BP_FAT	cell surface receptor linked signal transduction	15	25,0	2,2E-4	4,9E-3
GOTERM_BP_FAT	neurological system process	12	20,0	2,7E-4	5,1E-3
GOTERM_CC_FAT	plasma membrane	22	36,7	5,0E-4	3,3E-2
GOTERM_CC_FAT	intrinsic to membrane	27	45,0	7,0E-4	2,3E-2
GOTERM_CC_FAT	integral to membrane	26	43,3	1,2E-3	2,6E-2
GOTERM_MF_FAT	olfactory receptor activity	11	18,3	1,3E-7	1,2E-5
KEGG_PATHWAY	Olfactory transduction	11	18,3	1,2E-7	2,0E-6

Tabla No. 12.- Analisis de ontologia (GO) y de rutas metabolicas (KEGG) de los genes detectados en este estudio.

Se han comparado los CNVRs identificados en este estudio con QTLs de bovinos en la base de datos QTLdb [71] (Oct.31, 2013 (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index)) y se han encontrado un total de 34 CNVRs superposición. Estos QTLs se asocian a una amplia gama de características, incluyendo la longitud del cuerpo, proteína de leche, grasa de leche, altura, color de la carne, y algunos rasgos de susceptibilidad a la enfermedad como la distocia y mastitis clínica (Tabla No. 13 y para mayores detalles ver Tabla S6 en archivo adicional No. 1).

Chr	Start	End	QTL
1	1971535	2003454	Exterior_QTL
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	72455730	72559153	Production_Association
1	1.19E+08	1.19E+08	Production_Association

1	1.53E+08	1.53E+08	Reproduction_QTL
2	57747240	57750888	Milk_QTL
2	57747240	57750888	Milk_QTL
2	57747240	57750888	Production_QTL
2	57747240	57750888	Milk_QTL
3	18475187	18502234	Production_Association
3	18475187	18502234	Meat_Association
4	33547395	33556611	Reproduction_QTL
4	33547395	33556611	Production_QTL
4	33547395	33556611	Meat_QTL
4	33547395	33556611	Production_QTL
4	33547395	33556611	Production_QTL
4	1.06E+08	1.06E+08	Production_QTL
4	1.06E+08	1.06E+08	Meat_Association
6	1.12E+08	1.12E+08	Milk_QTL
7	97442762	97444260	Meat_QTL
7	97442762	97444260	Meat_QTL
7	97442762	97444260	Reproduction_QTL
7	97442762	97444260	Meat_QTL
7	97442762	97444260	Reproduction_QTL
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Production_Association
7	97442762	97444260	Production_Association
7	97442762	97444260	Production_Association
7	97442762	97444260	Meat_Association
7	97442762	97444260	Health_Association
9	45337323	45488218	Health_QTL
9	45337323	45488218	Meat_QTL
9	45337323	45488218	Milk_Association
9	45337323	45488218	Reproduction_Association
9	45337323	45488218	Reproduction_Association

10	1513111	1559113	Meat_QTL
10	1513111	1559113	Meat_QTL
10	1513111	1559113	Milk_QTL
10	1513111	1559113	Milk_QTL
10	27063526	27095789	Meat_Association
10	27063526	27095789	Meat_Association
10	27063526	27095789	Meat_Association
10	27063526	27095789	Meat_Association
10	27063526	27095789	Meat_Association
11	39886500	39918145	Milk_QTL
11	39886500	39918145	Milk_QTL
11	47082297	47198702	Production_QTL
11	47082297	47198702	Production_QTL
11	47082297	47198702	Reproduction_QTL
11	47082297	47198702	Production_QTL
11	82196097	82374760	Meat_QTL
11	88281698	88313232	Reproduction_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Milk_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Meat_QTL
11	88281698	88313232	Production_QTL
11	91921125	92340429	Production_QTL
11	91921125	92340429	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Meat_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	57687270	57719962	Production_QTL
12	70957641	71223810	Reproduction_QTL
12	71383094	71604231	Reproduction_QTL
12	72184389	72263216	Reproduction_QTL
12	72407022	73327391	Reproduction_QTL
12	73475758	73721049	Reproduction_QTL
12	74416147	75386962	Reproduction_QTL
12	75957893	76729025	Reproduction_QTL
14	2746730	2764767	Milk_QTL
14	2746730	2764767	Milk_QTL
14	2746730	2764767	Milk_QTL

14	2746730	2764767	Milk_Association
14	2746730	2764767	Milk_Association
14	2746730	2764767	Milk_Association
14	2746730	2764767	Milk_Association
14	2746730	2764767	Milk_QTL
15	80422173	80837237	Production_Association
15	80422173	80837237	Production_Association
15	80422173	80837237	Reproduction_QTL
15	80422173	80837237	Meat_QTL
15	80422173	80837237	Meat_Association
15	80422173	80837237	Meat_Association
15	80422173	80837237	Production_QTL
16	32899795	33028883	Health_Association
18	27908043	28375735	Meat_QTL
18	62340260	62533770	Meat_QTL
18	62340260	62533770	Reproduction_Association
18	62340260	62533770	Reproduction_Association
18	62340260	62533770	Reproduction_QTL
18	62340260	62533770	Reproduction_QTL
19	39619531	39689562	Reproduction_Association
19	39619531	39689562	Reproduction_Association
19	39619531	39689562	Milk_QTL
21	52306422	52328502	Reproduction_QTL
21	52306422	52328502	Reproduction_QTL
23	29468816	29489851	Meat_QTL
29	42455680	42514948	Meat_Association
29	42455680	42514948	Meat_Association
29	42455680	42514948	Meat_Association
29	42455680	42514948	Meat_Association
29	42455680	42514948	Meat_Association
29	42455680	42514948	Meat_Association

Tabla No. 13.- QTLs que se encuentran dentro o que se traslapan en las Regiones de CNVs identificadas en este estudio.

Por último, con el fin de confirmar la exactitud de nuestras CNVRs predichas, se seleccionaron 7 de las 56 CNVRs, y fue utilizado un PCR en tiempo real (qPCR) para validar experimentalmente. De los CNVRs seleccionados, 3 eran duplicaciones de copia única (CNVRs 1, 16 y 35), 2 eran borrados de una copia (CNVRs 2 y 11), y 2 eran copia duplicación doble (CNVRs 21 y 55). Para cada blanco CNVR, dos pares de cebadores fueron diseñados, teniendo en cuenta los límites de cada CNVR. Los cebadores para PCR fueron diseñados utilizando el Primer-BLAST de NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome). (Tabla No.6). Todos los cebadores se diseñaron basándose en la secuencia de referencia (Bos taurus UMD3.1) de

NCBI. Utilizando el método de ciclo umbral comparativo ($2^{-\Delta\Delta Ct}$) para cuantificar el número de cambios de las copias mediante la comparación del valor ΔCt, se confirmaron 5 CNVRs de 7 seleccionadas (ver Figura No. 13 y Tabla No.7).

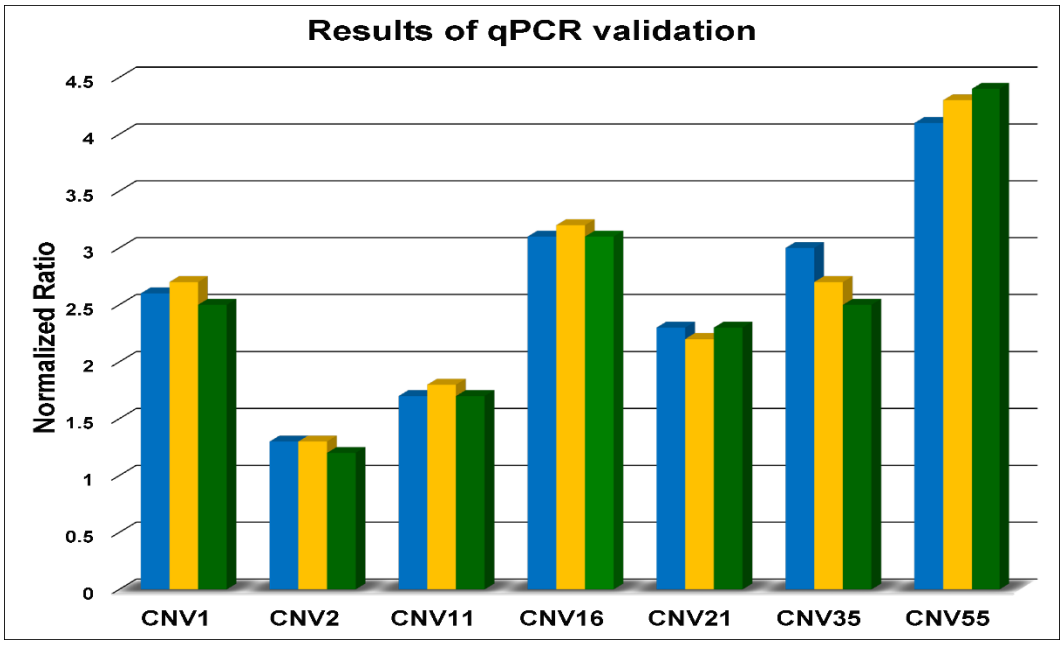


Figura No. 13: Valores de radios normalizados por el qPCR en tiempo real. El eje Y muestra los valores de radio obtenidos por qPCR y el eje X representa los números de CNVs que fueron validadas por triplicado en este estudio. Los valores alrededor de dos indican un estatus normal (No hay una CNV), alrededor de uno, indica una copia perdida o borrada, alrededor de tres indica tres copias ganadas y alrededor de cuatro indica cuatro copias ganadas.

2.5 Discusión.

Los datos del Axiom Genome-Wide BOS 1 array se normalizaron (véase la sección Métodos) y luego se aplicaron los algoritmos PennCNV y QuantiSNP para detectar CNVs. PennCNV identificó 155 CNVs mientras QuantiSNP identificó 302 CNV. En el análisis de traslape de las CNVs desde ambos algoritmos encontramos un 49.67% de traslapes de las detecciones de PennCNV a QuantiSNP, mientras que 25.49% corresponde de QuantiSNP a PennCNV. En total fueron 56 CNVs que coincidieron para ambos algortimos. El siguiente paso fue comparar nuestros resultados con los de estudios de CNVs anteriores detectados con diferentes tecnologías como la hibridación genómica comparativa (aCGH) [44,48], la secuenciación de próxima generación [72], y el SNP BeadChip 54Kb [45-47,49] y 777kb [50,51]. Todos los estudios utilizaron el

ensamble del genoma bovino Btau4.0, excepto los estudios en 777Kb y el nuestro, donde utilizamos el ensamble UMD3.1. Todos los datos en Btau4.0 fueron convertidos a UMD3.1 utilizando la herramienta de UCSC liftover [73] y el Servicio de Reasignación del Genoma de NCBI (Nov.5, 2013 (http://www.ncbi.nlm.nih.gov/genome/tools/remap), y finalmente, para ser coherente, comparamos sólo CNVRs de cromosomas autosómicos (Tabla No. 14). Para los estudios basados en aCGH, nuestros resultados se superponen con los reportados por Fadista [44] en un 16.07% (9 CNVRs) con una longitud de solapamiento total de 1.83Mb (21,63%). Para los reportados por Liu [48] Se encontró una superposición de 25% (14 CNVRs) con una longitud de solapamiento total de 3.93Mb (46.45%). Para el estudio basado en la secuenciación de siguiente generación reportado por Bickhart [72] encontramos una superposición de 19.64% (11 CNVRs) con una longitud total de superposición 2.82Mb (33.33%). Para un estudio basado en la tecnología de SNPs de 54Kb, reportado por Hou 2011 [45], encontramos una superposición de 26.78% (15 CNVRs) con una longitud total de superposición 3.99Mb (47.16%). Para el estudio publicado por Hou 2012b [49], encontramos una superposición de 19.64% (11 CNVRs) con una longitud de solapamiento total de 2.37MB (28.01%). Para el estudio publicado por Jiang [47], encontramos una superposición de 3.19% (3 CNVRs) con una longitud de solapamiento total de 0.66Mb (7.80%). Para que los resultados reportados por Cicconardi [46] encontramos una superposición de 26.78% (11 CNVRs) con una longitud de solapamiento total de 1.20Mb (14.18%). Por último, para los estudios que utilizan microarreglos de SNPs en alta densidad (777Kb), al comparar las CNVRs reportados por Hou 2012a [51] encontramos un 75% de solapamiento (42 CNVRs) que representa 7.59MB (89.72%), y cuando se comparan con los reportados obtenidos por Jiang 2013 [50], encontramos una superposición de 41.07% (23 CNVRs) que representa 1.31Mb (15.48%). De la comparación de nuestros resultados con estudios reportados previamente, vemos diferencias notables, sobre todo debido al uso de diferentes algoritmos, las diferentes

			Other Studies				Overlapped CNVRs of this study			
Study	**Methods**	**Algorithm used**	**Samples**	**Breds**	**CNVR**	**Length (Mb)**	**Count**	**Pecentage of count**	**Total length (Mb)**	**Percentage of length**
Fadista et al., 2010	aCGH	seg-MNT	20	4	233	14.25	9	16.07%	1.83	21.63%
Liu et al., 2010	aCGH	GIM; SW-ARRAY	90	17	142	35.83	14	25.00%	3.93	46.45%
Bickhart et al, 2012	Next-generation Seq	mrFAST	6	3	978	52.30	11	19.64%	2.82	33.33%
Hou et al, 2011ª	SNP chip (54k)	PennCNV	521	21	672	152.68	15	26.78%	3.99	47.16%

Hou et al, 2011b	SNP chip (54k)	PennCNV	472	1	462	86.05	11	19.64%	2.37	28.01%
Jiang et al, 2012	SNP chip (54k)	PennCNV; cnvPartition; GADA	2047	1	94	21.30	3	3.19%	0.66	7.80%
Cicconardi et al, 2013	SNP chip (54k)	PennCNV; QuantiSNP	2654	5	394	515.90	11	26.78%	1.20	14.18%
Hou et al, 2012	SNP chip (777K)	PennCNV	674	27	3438	146.90	42	75.00%	7.59	89.72%
Jiang et al, 2013	SNP chip (777K)	PennCNV	96	1	358	34.45	23	41.07%	1.31	15.48%
This Study	SNP chip (648K)	PennCNV; QuantiSNP	12	1	56	8.46				

Tabla No. 14: Comparación entre las CNVRs detectadas en este estudio contra estudios previamente reportados.

plataformas tecnológicas, y diferentes criterios de ajuste de parámetros. Un factor que contribuye a las diferencias puede ser la región geográfica en la que los animales han ido evolucionando. Nos dimos cuenta de que el tamaño de la muestra no influye en los resultados teniendo en cuenta que si lo nos comparamos con Jiang 2012 [47], que utilizaron 2,047 muestras y Cicconardi 2013 [46], con 2,654 muestras, encontramos coincidencia solo de 3.19% y 26.78%, respectivamente. Y al compararnos con Hou 2011 [45], que utilizo 521 muestras, encontramos 26.78% de traslapes. Por otra parte, al compararnos con Fadista 2010 [44] que utilizó 20 muestras de animales con tecnología de aCGH, encontramos que un 16.07% de CNVs se traslapan. Y al compararnos con Bickhart 2012 [72], que utilizo solo 6 muestras de animales y la tecnología de secuenciación de siguiente generación, encontramos un 19.64% de solapamiento de CNVs. El mayor número de coincidencias de nuestros resultados, fueron con los obtenidos como resultado del uso de tecnologías de alta densidad de genotipificado de SNPs [50,71], de los que tuvimos el 75% y el 41.07% de superposición, respectivamente. Una de las razones es que los SNPs de alta densidad pueden detectar pequeñas CNVs, que los detectados con menor densidad de SNPs, como se muestra por [45,47,50,51], cuyos estudios utilizando una densidad de 54k en SNPs, obtuvieron 682 y 94 CNVs, respectivamente, mientras que los últimos dos utilizando 777kb de densidad en SNPs se encontraron 3,346 y 358 CNVs, respectivamente.

El contenido de GC dentro de cada CNVR varía de 33.9% a 57.95%, lo que apoya la posición de que CNVRs son regiones con alto contenido de GC [44]. Las CNVs también se asocian con segmentos duplicados (SD), que se definen como una secuencia de ADN con una longitud ≥ 1Kb con al menos 90% de identidad de secuencia. Al comparar nuestros resultados con los SDs ya reportados [74] tenemos un 12.5% de traslapes.

Capítulo III

Análisis de Variaciones Estructurales en Alta Densidad Basadas en Desequilibrio Ligado en el Genoma del Ganado Bovino

3.1 Resumen.

Las variaciones estructurales genómicas representan una importante fuente de variación genética en los genomas de mamíferos, por lo tanto, con frecuencia se relacionan con las expresiones fenotípicas. En este trabajo, se analizaron ~770,000 genotipos de polimorfismos de nucleótido simple de 506 animales de 19 razas de ganado. Se definió una simple variación estructural basada en LD, y se realizó un análisis de todo el genoma. Después de aplicar algunos filtros de control de calidad, para cada raza y cada cromosoma se calculó el desequilibrio de ligamiento (r^2) de corto alcance ($\leq$ 100Kb). Ordenamos los pares de SNPs por distancia y obtuvimos un conjunto de medias de LD (llamada la media esperada) utilizando bloques de 5Kb. Se identificaron 15,246 segmentos de al menos 1 Kb, entre las 19 razas, que consisten en conjuntos de al menos 3 SNPs adyacentes de manera que, para cada SNP, la r^2 dentro de sus vecinos en un rango de 100Kb, hacia el lado derecho, fueran todos mayor que, o todos más pequeños que la media esperada correspondiente, y su valor-P fuera significativo después de una corrección de múltiples pruebas de Benjamini-Hochberg. Además, para tener en cuenta sólo a las regiones distribuidas homogéneamente hemos considerado sólo SNPs que tienen al menos 15 vecinos SNPs dentro de 100kb. Definimos tales segmentos como variaciones estructurales. Al agrupar todas las variaciones a través de todos los animales en la muestra se definieron 9,146 regiones,

con un total de 53,137 SNPs; que representan el 6.40% (160.98 Mb) del genoma bovino. Las variaciones estructurales identificadas cubren 3,109 genes. Un análisis de agrupación demostró la relación de las razas dado a la región geográfica en la que están evolucionando. En resumen, presentamos un análisis de variaciones estructurales en base a la desviación esperada del LD de rango corto entre los SNPs en el genoma bovino. Con una intuitiva y simple definición basada únicamente en datos SNPs fue posible discernir la cercanía de las razas debido a la agrupación por región geográfica en la que se están desarrollando.

3.2 Introducción.

Es muy conocido que entre los seres humanos, el 99.9% de la secuencia de ADN es idéntica [75], pero esa diferencia contribuye a las variaciones genéticas entre las personas. Los primeros estudios sobre el genoma humano se limitaban a los que solo podían ser identificados a través de un microscopio. Tales variaciones se definieron como variaciones estructurales, y contaban con una longitud de aproximadamente 3 Mb o más. Dado que las tecnologías evolucionan, los científicos fueron capaces de caracterizar las diferencias más pequeñas, así como abundantes alteraciones en las secuencias cortas de ADN, por lo general de menos de 1 Kb. En los últimos años, el análisis comparativo de secuencias ha revelado que la variación del ADN implica segmentos que son más pequeños que los reconocidos microscópicamente, pero más grandes que aquellos fácilmente detectados mediante análisis de secuencia convencional. Estas variaciones van desde ~1 Kb a 3 Mb y consisten en inserciones, barrados, inversiones y duplicaciones, que incluso pueden contener genes completos [76]. El impacto de estas variaciones puede variar desde ninguna diferencia observable a la interrupción del gen. Una variación estructural se define formalmente como una alteración genómica que contempla segmentos de ADN de mayores o iguales a 1 Kb y los cambios de este rango de magnitudes pueden ser submicroscópicos o microscópicos. Una variación estructural se designa como anormalidad estructural cuando se establece que puede, por sí misma, o en combinación con otros factores ambientales, ser la causa de una enfermedad genética o una expresión fenotípica. Las variaciones estructurales basadas en genotipos de datos han sido recientemente el foco de interés, sobre todo las del tipo variación del número de copias (CNV). El método más utilizado para la detección de CNVs se basa en dos medidas de la intensidad de la señal para cada SNP: el Logaritmo de Radio R (una medida normalizada de la intensidad de la señal total para dos alelos del SNP), y la frecuencia del alelo B (una medida normalizada de la relación de intensidad alélica de dos alelos). La combinación del Logaritmo de Radio R y la frecuencia del alelo B se utilizan para inferir los cambios en el número de copias en el genoma [77,78]. Otros estudios se han centrado en la viabilidad del desequilibrio ligado (LD) para la detección de CNVs [79-

81]. Un estudio que compara LD de regiones comunes de haplotipos entre diferentes poblaciones [81] informó de que un alto LD no es siempre una señal para una variación genómica, y un bajo LD puede estar implicado en las inserciones y borrados. Otro estudio de LD entre SNPs y CNVs reveló que las medidas de LD tradicionales son suficientes para SNPs que estén fuera de las CNVs, sin embargo la misma métrica inapropiadamente cuantificaría la covarianza para los SNPs dentro de una CNV [80]. En un reciente estudio, una CNV del tipo borrado en el cromosoma 6 bovino, se predijo a partir de su SNP vecino con un modelo de regresión múltiple, por Kadri et al. (2012) [82]. Ellos concluyen que el genotipo de una CNV de tipo borrado y su efecto de QTL putativo, se puede predecir con exactitud un máximo de 0.94 de SNPs a su alrededor. Esta alta precisión de predicción sugiere que la variación genética debido a una CNV con borrado simple, está bien capturado por grupos de SNP densos. Las variaciones estructurales han sido el foco de una serie de estudios recientes en el genoma del ganado bovino, especialmente CNVs [83-91]. Sin embargo no se han analizado las variaciones de otro tipo en lugar de CNV. Con el uso de alta densidad en marcadores de SNP uniformemente distribuidos en el genoma, es posible detectar regiones con desviación de LD significativo en comparación con el valor esperado, las cuales pueden ser interpretadas como variación genómica de rango corto, y podría ayudar en futuros estudios para evaluar la asociación con otro tipo de variaciones estructurales. La compañía Illumina, en colaboración con investigadores del genoma bovino, han desarrollado el panel BovineHD con 777,962 SNPs de genoma completo, por lo que es el panel de mayor densidad en el ganado. Matukumalli et al (2011) [92], genotipificaron un grupo de más de 500 animales de 19 razas y analizaron bloques de LD y segmentos de CNVs. En este artículo, hemos utilizado estos datos para inspeccionar la distribución de las variaciones estructurales basado en patrones de desequilibrio ligado (LD) de rango corto en el genoma completo de todas las razas de la muestra. En primer lugar, se presenta un análisis de la caída promedio del LD de rango corto a través de diferentes distancias; después, se propone una definición de variación estructural basada en el patrón promedio de patrones de LD de rango corto, y examinamos todo el genoma para estudiar la distribución de estas variaciones. Los métodos de agrupación nos permitieron diferenciar entre los grupos de las razas basados en las estadísticas de las variaciones estructurales. Finalmente, revisamos las coincidencias de las variaciones estructurales obtenidas con nuestro método y los diferentes tipos de variaciones reportadas previamente.

3.3 Materiales y Métodos.

3.3.1 Muestras de Animales y Descripción de los Datos.

Un conjunto de 777,962 SNPs capturados por el BovineHD Genotyping BeadChip fue proporcionado por Consorcio Internacional Bovino HapMap [93]. Los SNPs fueron asignados en el ensamble UMD3.1. Había 506 animales en la muestra. Se tomaron muestras de 19 razas diferentes (para más detalles véase Tabla No. 15). Todas las razas pertenecen a la subespecie de Taurus e Indicus de la especie Bos taurus. Se agruparon según su destino y el origen geográfico de la siguiente manera: las razas de carne británicas son Angus, Hereford y Angus Rojo. Razas de carne europeas son Charolais, Limousin, piedmontese y Romagnola. Razas británicas lecheras son Guernsey y Jersey. Razas lecheras europeas son Pardo Suizo, Holstein y Noruega Roja. El grupo Indicus incluye las razas Brahman, Nelore y Gir. Las razas africanas se forman por la N'Dama y Sheko y, por último, las razas híbridas (Taurus X Indicus) se forman por la Santa Gertrudis y Beefmaster.

Bos Taurus Taurus

Breed	Acronym	No. of animals	Country of sampling	Land of origin
Angus	ANG	27	USA and New Zeland	Scotland
Brown Swiss	BSW	24	USA	Switzerland
Charolais	CHL	22	USA	France
Guernsey	GNS	21	USA and UK	Channel Islands
Hereford	HFD	27	USA and New.	UK
Holstein	HOL	59	USA and New Zeland	Netherlands
Jersey	JER	32	USA and New Zeland	Channel Islands
Limousin	LMS	40	USA and France	France
Norwegian Red	NRC	17	Norway	Norway
N'Dama	NDA	24	Guinea	West Africa
Piedmontese	PMT	24	Italy	Italy
Red Angus	RGU	11	USA and Canada	Scotland
Romagnola	RMG	23	Italy	Italy
TOTAL		**351**		

Bos Indicus

Brahman	BRM	25	USA and Australia	USA
Gir	GIR	30	Brazil	India
Nelore	NEL	34	Brazil	India
TOTAL		**89**		

Bos Taurus x Bos Indicus				
BeefMaster	BMA	24	USA	USA
Sheko	SHK	18	Ethiopia	Ethiopia
Santa Gertrudis	SGT	24	USA	USA
TOTAL		**66**		

Tabla No. 15: Razas y número de animales en la muestra.

3.3.2 Filtros de Control de Calidad.

Se aplicaron filtros de control de calidad con el fin de eliminar los errores de tecnología del tipo "no call", para eliminar todos los genotipos que violaban el equilibrio de Hardy-Weinberg y los marcadores monomórficos (marcadores con MAF <0.05%) también fueron eliminados. Después de estos filtros de control de calidad se garantiza una calidad global en todas las muestras. Se trabajó, en promedio, entre las 19 razas, con un total de 523,651 marcadores que representa 67.31% de la información inicial (ver Tabla No. 16 y para mayores detalles ver tabla S2 en archivo adicional No. 2). Los datos pueden ser puestos a disposición de los investigadores que lo soliciten.

Chr	Angus	Brown Swiss	Charolais	Guernsey	Hereford	Holstein	Jersey	Limousin
1	32906	31871	35395	30214	34006	33943	28829	34669
2	28921	26569	30558	25985	30081	28004	24325	29155
3	25305	24706	27466	23858	27817	26161	21718	27274
4	25712	24335	26409	22984	26549	25718	23223	26712
5	23120	22140	26325	23463	26245	25192	22766	26451
6	25605	21698	28164	24328	26422	26525	23106	27225
7	23796	23177	25856	21919	24794	23028	21391	24205
8	19518	19294	22306	18310	18987	19929	18074	20858
9	22415	21818	24663	20421	22592	23134	20768	23550
10	23928	21996	24629	20741	23934	22961	20941	23658
11	23906	22009	25630	20605	23854	24574	22383	25075
12	18739	17848	19813	17727	20194	19582	17727	19746
13	14893	13875	15543	12893	15657	14875	12845	15835
14	14667	14071	16782	14422	15901	15293	14021	15846
15	17902	17421	19748	16829	19235	18688	16746	19291
16	17811	15705	18842	16247	18269	17261	16420	18228
17	17138	16056	17653	14886	17766	17000	15296	17722
18	14854	14377	15518	14491	15330	15554	13130	15916
19	14801	13517	15891	12705	15284	14603	13780	15010

20	16401	15611	17852	14898	16670	16136	13822	17252
21	15767	14899	16673	14017	15690	15608	14106	15805
22	14387	12928	15080	13289	15053	14153	13046	14840
23	12285	11118	12444	11295	12424	11901	10387	12309
24	13849	12485	14517	12490	13985	13814	12314	14134
25	10156	9421	10725	9516	10684	10357	9044	10206
26	11735	11226	12188	10886	12197	11903	10646	12195
27	10350	9546	10861	9588	10917	10420	9549	10484
28	10396	9895	11003	9817	10421	10417	9584	10400
29	11536	10055	11945	9786	11825	11416	9671	11521
TOTAL	**532799**	**499667**	**570479**	**488610**	**552783**	**538150**	**479658**	**555572**

Tabla No. 16: Resumen final de marcadores en nuestro estudio.

3.3.3 Inferencia de Haplotipos.

Se estimó el par de haplotipos para cada animal en la muestra utilizando el algoritmo BEAGLE3.3.2 [94], donde los modelos de vectores de genotipo en cada individuo utilizan un Modelo Oculto de Markov (HMM). El modelo se puede escribir de la siguiente manera:

$$P(G_i|H,\theta,\rho) = \sum_z P\,(G_i|Z,\theta), P(Z|H,\rho)$$

Dónde:

$$Z = \{Z_1, \ldots, Z_L\}, \text{ con } \; Z_j = \{Z_{j1}, Z_{j2}\} \;\; \text{y} \;\; Z_{jk} = \{1, \ldots, N\}.$$

Z representa el par de haplotipos, para el SNP *j*, procedentes de un panel de referencia, que está siendo copiado para formar el vector de genotipo.

P (Z | H, ρ), se define por una cadena de Markov, y se modela como el par de haplotipos copiado de los cambios del panel de referencia durante la secuencia debido a un mapa de recombinación (*ρ*) definido a lo largo de todo el genoma.

P (G | Z, θ), genera la variación observada de los vectores del genotipo con respecto a los haplotipos copiados de panel de referencia a través de una tasa de mutación (*θ*) [95].

3.3.4 Calculo del Desequilibrio Ligado.

El coeficiente de LD seleccionado para este estudio fue el coeficiente de correlación al cuadrado entre pares de SNPs (r^2) representa como:

$$r^2 = \frac{(p_{11} - p_1 q_1)^2}{p_1 q_1 p_2 q_2}$$

donde p_1 y p_2 son las frecuencias menores y mayores de los alelos en el SNP1 respectivamente, q_1 y q_2 son las frecuencias menores y mayores de los alelos en el SNP2 respectivamente, y p_{11} corresponde a la frecuencia observada entre los dos alelos de menor importancia en el mismo individuo a lo largo de toda la población . Además, una corrección para el tamaño de la muestra se aplicó a todos los valores calculados de r^2 utilizando la siguiente ecuación:

$$r^2 corregida = \frac{r^2 calculada - \frac{1}{n}}{1 - \frac{1}{n}}$$

donde *n* representa el número de haplotipos en la muestra [96]. (Para más detalles véase la Tabla S4 en archivo adicional No.2).

3.3.5 Definición de las Variaciones Estructurales Basadas en Desequilibrio Ligado de Rango Corto.

Se define una variación estructural basada en LD de la siguiente manera: en primer lugar, para cada raza y cada cromosoma se obtiene un conjunto de medias de LD, llamado medias esperadas, calculando el desequilibrio ligado (r^2) de corto alcance (≤ 100Kb), se ordenaron los pares de SNP por distancia, y se calculó la media para cada bloque de 5Kb. Entonces verificamos SNP por SNP en busca de segmentos de al menos 1 Kb, que contengan un grupo de al menos 3 SNPs adyacentes de manera que, para cada SNP, la r^2 dentro de sus vecinos en un rango de 100Kb hacia su flanco derecho de estos SNPs, todos son mayores que, o todos son menores que los medias correspondientes esperadas y sus *Valores-P* de la *prueba-t* para la igualdad de medias son significativas después de la corrección de múltiples pruebas de Benjamini-Hochberg. Además, tomamos en cuenta sólo por homogeneidad, las distribuciones de regiones de SNPs, que tengan al menos 15 SNPs vecinos dentro del rango de 100kb.

3.3.6 Corrección de Múltiples Pruebas.

Una corrección de múltiples pruebas de Benjamini y Hochberg [97] se aplicó a los *Valores-P* con el fin controlar la Tasa de Falsos Descubrimientos. Descrito de la siguiente manera: primero, todos los *Valores-P* se ordenan de menor a mayor. Denotando el más pequeño i-ésimo *Valor-P* por $p_{(i)}$, para cada *i* entre 1 y *m* (*m* es el número total de *Valores-P*), a continuación, a partir del *Valor-P* más grande $P_{(m)}$, comparar $P_{(m)}$ con 0.05 x *i/m*. Continuar siempre que $P_{(i)} > 0.05$ x *i/m*. Sea *k* el primer momento en que $P_{(k)}$ es inferior o igual a 0.05 x *k/m*, y declarar las diferencias correspondientes a los más pequeños *Valores-P k* como significantes.

3.3.7 Análisis de Componentes Principales.

Vectores construidos con el número de variaciones estructurales por cromosoma se utilizaron para llevar a cabo un análisis de componentes principales (PCA) y buscar la diferenciación entre los subgrupos de ganado. Se utilizó el software R para realizar este análisis. La idea central de PCA es para reducir la dimensionalidad de un conjunto de datos que consta de un gran número de variables interrelacionadas, al tiempo que conserva tanto como sea posible la variación presente en el conjunto de datos. Esto se consigue mediante la

transformación de un nuevo conjunto de variables, los componentes principales (PCs), los cuales no están correlacionados, y están ordenados de tal forma que los primeros componentes capturan la mayor parte de la variación presente en todas las variables originales [98].

Formalmente, el PCA se define como una transformación lineal ortogonal que transforma los datos a un nuevo sistema de coordenadas, tales que la mayor varianza para cualquier proyección de los datos viene a establecerse en la primera coordenada (llamado el primer componente principal), la segunda mayor varianza en la segunda coordenada, y así sucesivamente. El PCA es teóricamente óptimo para la transformación de un conjunto de datos dados en términos de los mínimos cuadrados. El procedimiento para la obtención de los PCAs se puede resumir de la siguiente manera:

Dado un vector X^T de dimensiones n, $X^T = [x_1, x_2, ..., x_n]^T$, cuyo vector medio M y la covarianza C se describen a través de:

$$M = E(X) = [m_1, m_2, \ldots, m_n]^T$$

$$C = E[(X - M)(X - M)^T]$$

Calcular los valores propios (eigen valores) $\lambda_1, \lambda_2, ..., \lambda_3$, y los vectores propios (eigen vectores) $P_1, P_2, ..., P_d$; después organizarlos en función de su magnitud.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

Seleccionar los vectores propios d para representar las variables n, $d < n$. A continuación, el $P_1, P_2, ..., P_d$ son llamados los componentes principales.

3.4 Resultados.

3.4.1 Proporciones Polimórficas y Desequilibrio Ligado.

El conjunto de datos filtrados incluyó un promedio de 523,651 genotipos por raza, entre los 29 cromosomas autosómicos en 506 animales de 19 razas (véase la sección Métodos para la descripción de los filtros de control de calidad, ver

Tabla No. 16 y la Tabla S2 en el archivo adicional No. 2, para obtener más información de los marcadores totales en cada raza y cromosómica separado, y en la Tabla No. 15 para más detalles sobre razas descripciones y números de muestra). Para la caracterización de SNPs, se determinó la distribución de Frecuencias de Alelos Menores (MAF) en cada una de las 19 razas (ver Figura No. 14, Tabla No.17, para mayores detalles la Tabla S3 del archivo adicional No. 2 y las Figuras del archivo figure S1)

Breed	Samples	Markers	Mean MAF > (0.05)	Median MAF > (0.05)
Angus	27	532799	0.28	0.3
Brown Swiss	24	499667	0.27	0.27
Charolais	22	570479	0.28	0.3
Guernsey	21	488610	0.28	0.29
Hereford	27	552783	0.29	0.31
Holstein	59	538150	0.29	0.3
Jersey	32	479658	0.27	0.28
Limousin	40	555572	0.28	0.3
Norwegian Red	17	545844	0.28	0.29
N'Dama	24	415554	0.27	0.27
Piedmontese	24	569568	0.29	0.29
Red Angus	11	498781	0.29	0.32
Romagnola	23	551381	0.28	0.28
Bos taurus taurus	**27**	**522988**	**0.28**	**0.29**
Brahman	25	505846	0.24	0.22
Gir	30	402657	0.25	0.23
Nelore	34	420216	0.25	0.24
Bos taurus indicus	**29.67**	**442906**	**0.25**	**0.23**
Beefmaster	24	637713	0.29	0.29
Santa Gertrudis	24	622042	0.28	0.29
Sheko	18	562051	0.27	0.28
Btt x Bti (composite)	**22**	**607269**	**0.285**	**0.29**
AVERAGE SNPs		**523,651**		

Tabla No. 17: Promedios de la media y mediana en las 19 razas.

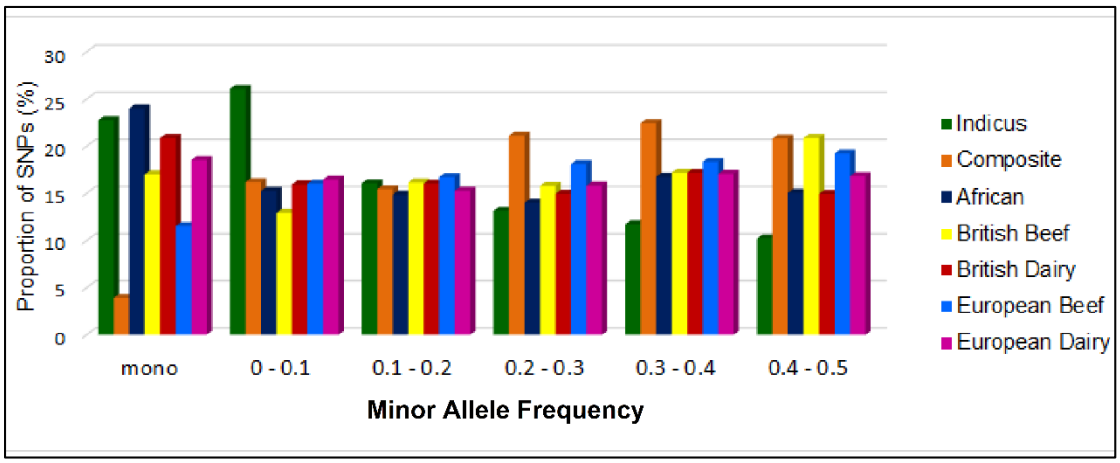


Figura no. 14: Distribución del MAF, proporciones promedio de frecuencias de SNPs por grupo de ganado.

La única raza adaptada al trópico *Bos taurus taurus*, N'Dama y dos razas de la *Bos indicus*, Gir y Nelore, tenían la menor proporción de SNPs polimórficos (67%, 71% y 74%, respectivamente). La raza Brahman puede tener una mayor tasa de polimorfismo que las razas indicus restantes, debido a que esa raza fue establecida mediante la importación principalmente toros y el retrocruzamiento para taurus derivadas de hembras. Las razas *Bos Taurus taurus* fueron generalmente moderadamente variables, con un promedio de 83% de los SNPs polimórficos. Los razas *Bos Taurus indicus* muestran sus SNPs polimórficos con el más bajo promedio con 81%. Figura No. 14 considera todos los SNPs (incluyendo SNPs monomórficos y polimórficos), pero para todos los posteriores análisis, los SNPs monomórficos fueron eliminados de este estudio debido a que no son informativos.

Se estimó el haplotipo par para cada animal en la muestra utilizando Beagle 3.3.2 [94], y se calculó el LD para cada raza, reportando el coeficiente de correlación, r^2. Los valores de LD se corrigieron para el tamaño de la muestra utilizando la ecuación descrita por Villa-Angulo et al. (2009) [96] y para cada SNP se calcularon todos los valores de LD a una distancia máxima de 100 Kb por pares.

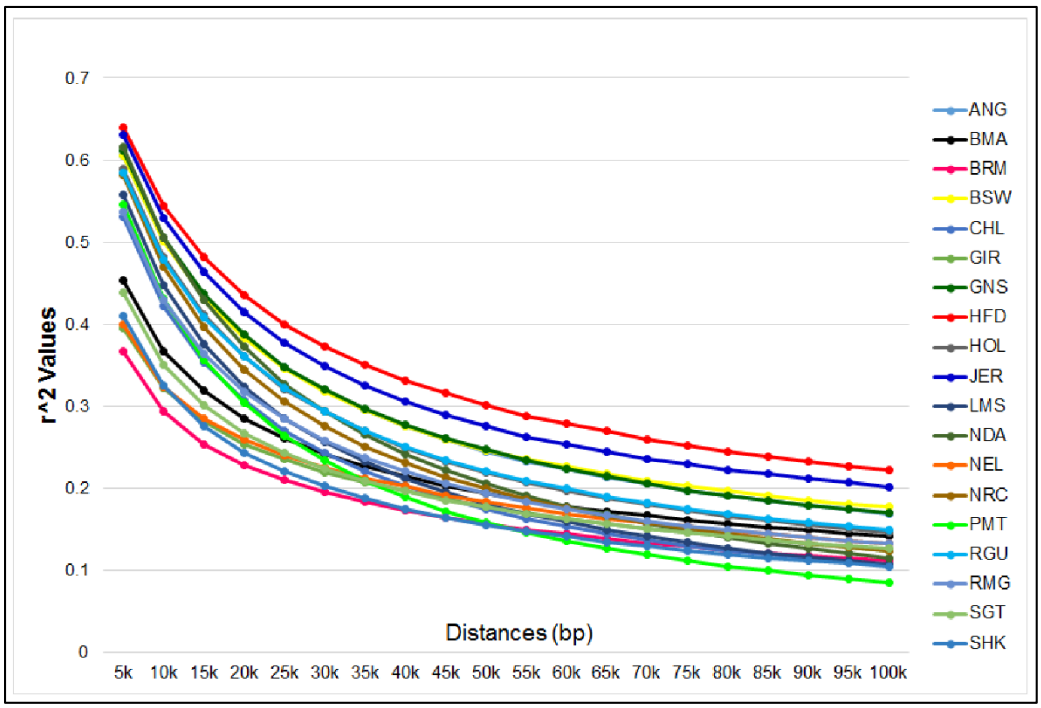


Figura no. 15: Desequilibrio Ligado, caída natural del LD en el genoma complete en todas las razas.

Los valores promedios se determinaron utilizando bloques de 5 Kb. Como se observa en la Figura No. 15, en distancias cortas (5 Kb) la raza con el promedio más bajo de LD fueron las razas *Bos taurus indicus*: Brahman, Gir, y Nelore, con 0.367, 0.395 y 0.401, respectivamente, mientras que las razas con mayor promedio de LD fueron Hereford, Jersey, y N'Dama, con 0.639, 0.630 y, 0.616, respectivamente. Por otro lado, en grandes distancias (100 Kb), las razas con los promedios más bajos de LD fueron piedmontese, Sheko y Charolais, con 0.085, 0.104 y 0.105, respectivamente, mientras que las razas con el promedio más alto de LD fueron Hereford, Jersey y Pardo Suizo, con 0.222, 0.201 y 0.177, respectivamente. (ver Tabla No. 18 y para mayores detalles ver Tabla S4 en archivo adicional No. 2).

BREED	5k	10k	15k	20k	25k	30k	35k	40k	45k	50k	55k	60k	65k	70k	75k	80k	85k	90k	95k	100k
PMT	0.547	0.431	0.357	0.305	0.264	0.235	0.21	0.19	0.173	0.158	0.146	0.136	0.126	0.119	0.111	0.105	0.099	0.094	0.09	0.085
SHK	0.411	0.325	0.276	0.243	0.22	0.202	0.187	0.175	0.163	0.155	0.147	0.14	0.134	0.129	0.123	0.119	0.115	0.111	0.107	0.104
CHL	0.531	0.422	0.355	0.307	0.271	0.244	0.221	0.203	0.187	0.174	0.162	0.153	0.145	0.137	0.13	0.124	0.119	0.114	0.11	0.105

LMS	0.558	0.447	0.376	0.325	0.286	0.256	0.232	0.212	0.196	0.182	0.169	0.159	0.149	0.142	0.134	0.127	0.122	0.116	0.112	0.107
BRM	0.367	0.293	0.254	0.228	0.21	0.195	0.183	0.173	0.164	0.156	0.149	0.144	0.138	0.133	0.129	0.124	0.121	0.117	0.114	0.111
NDA	0.616	0.507	0.429	0.372	0.326	0.293	0.265	0.241	0.222	0.206	0.191	0.179	0.168	0.157	0.149	0.141	0.133	0.127	0.121	0.115
NRC	0.581	0.469	0.397	0.345	0.305	0.276	0.251	0.23	0.213	0.199	0.186	0.176	0.166	0.158	0.15	0.144	0.138	0.132	0.128	0.123
GIR	0.395	0.322	0.281	0.254	0.235	0.219	0.206	0.195	0.185	0.177	0.17	0.163	0.156	0.151	0.145	0.14	0.137	0.132	0.128	0.124
SGT	0.438	0.35	0.301	0.268	0.243	0.224	0.209	0.196	0.185	0.176	0.169	0.162	0.156	0.15	0.145	0.14	0.137	0.133	0.13	0.126
NEL	0.401	0.322	0.284	0.259	0.239	0.225	0.212	0.201	0.191	0.183	0.175	0.169	0.163	0.158	0.153	0.148	0.143	0.139	0.136	0.132
RMG	0.536	0.429	0.364	0.319	0.284	0.259	0.238	0.22	0.206	0.194	0.183	0.175	0.167	0.161	0.154	0.149	0.145	0.14	0.136	0.132
BMA	0.454	0.368	0.319	0.286	0.261	0.242	0.227	0.213	0.202	0.193	0.185	0.178	0.171	0.166	0.161	0.156	0.152	0.148	0.145	0.141
HOL	0.589	0.481	0.411	0.361	0.321	0.293	0.269	0.249	0.232	0.219	0.207	0.197	0.188	0.18	0.173	0.166	0.161	0.155	0.15	0.146
RGU	0.584	0.478	0.408	0.359	0.321	0.293	0.268	0.249	0.233	0.22	0.208	0.198	0.188	0.181	0.174	0.168	0.162	0.157	0.152	0.147
ANG	0.605	0.503	0.434	0.386	0.347	0.319	0.295	0.275	0.259	0.245	0.233	0.223	0.213	0.205	0.197	0.19	0.184	0.178	0.174	0.168
GNS	0.611	0.506	0.437	0.388	0.348	0.32	0.297	0.277	0.261	0.247	0.234	0.223	0.214	0.206	0.197	0.191	0.185	0.179	0.174	0.17
BSW	0.606	0.501	0.433	0.384	0.346	0.317	0.295	0.276	0.26	0.246	0.235	0.226	0.216	0.209	0.202	0.196	0.19	0.185	0.181	0.177
JER	0.63	0.53	0.463	0.415	0.377	0.349	0.325	0.305	0.289	0.276	0.263	0.254	0.245	0.236	0.23	0.222	0.217	0.211	0.207	0.201
HFD	0.639	0.545	0.482	0.436	0.4	0.373	0.351	0.332	0.316	0.302	0.289	0.279	0.269	0.26	0.253	0.246	0.239	0.233	0.228	0.222

Tabla No. 18: Valores promedios obtenidos por r^2 para todas las razas (caída natural LD).

Al comparar estos resultados, con los valores de otro estudio en su caída natural de LD con los datos de densidad más baja (BovineSNP50) y utilizando las mismas muestras, las tendencias obtenidas son bastante similares, incluso cuando el estudio en mención solo considero las regiones de alta densidad de SNPs en los cromosomas 6,14, y 25 [96]. Cabe señalar que [todos o la mayoría] de los animales utilizados en este estudio fueron los mismos que en la investigación anterior, aunque el número de pares de marcadores situados a menos de 20Kb utilizando el BovineSNP50 era bastante pequeña.

3.4.2 Variaciones Estructurales.

Para cada raza y cada cromosoma se obtuvo un conjunto de medias de LD (medias esperadas) utilizando bloques de 5 Kb. Después, para cada SNP se calculó el LD con cada SNP polimórficos dentro de 100 Kb a la derecha de ese SNP. Las variaciones estructurales putativas se definieron en donde todas las mediciones de LD eran más grandes que, o menores que, todas sus correspondientes medias esperadas, para todos los pares de SNP dentro de una región de 100 Kb para al menos 3 SNPs polimórficos adyacentes. Figura No. 16 indica dos muestras declaradas como variantes estructurales y una indicando que no contiene variantes estructurales. En la parte superior de la figura; durante tres SNPs, se muestra los valores de LD con todos los SNPs dentro de 100 Kb a su derecha. Para estos tres SNPs, los valores de LD son todos mayores que su media

esperada, y fueron declaradas significativas después de correcciones de múltiples pruebas. En el medio de la Figura No. 16, los valores de LD indican otro ejemplo donde los valores de r^2 son más pequeños que las medias esperadas, y como en el ejemplo anterior fueron declaradas significativas después de la corrección. En la parte inferior, un ejemplo que indica que no contiene una variación estructural, porque su LD para algunos de los SNPs pares se alternan en su valores, que son mayores o menores a la caída natural del LD definido como promedio esperado (línea roja monótonamente decreciente).

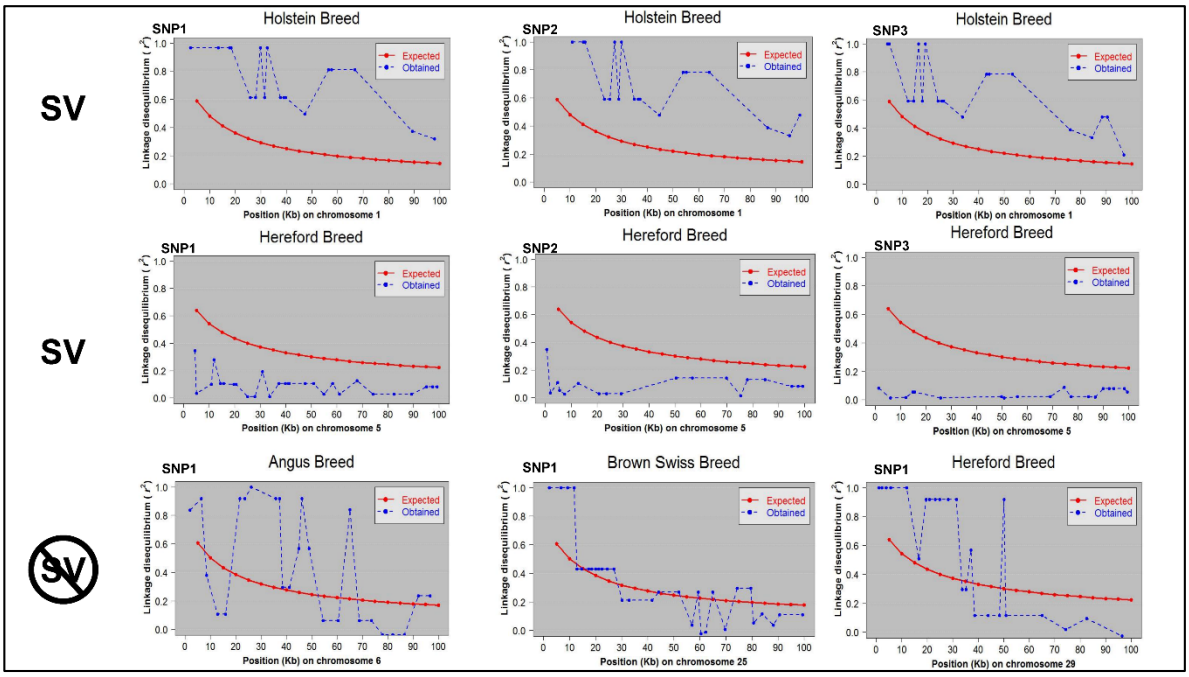


Figura no. 16: Dos ejemplos declarados como variación estructural y uno declarado como variación estructural no existente. En la parte de arriba tres SNPs adyacentes son inspeccionados por el LD con sus vecinos a su derecha en un rango de 100Kb. Todos sus valores de LD son mayor que la media esperada y fueron declarados significantes después de las correcciones de múltiples pruebas. En el medio, una muestra de cómo el LD es menor a la media esperada y como sus vecinos también lo son, por lo que son declarados significantes después de la corrección. En la parte de abajo, un ejemplo de lo que no es una variación estructural, por que el LD de alguno o varios SNPs, se intercalan entre el valor de la media esperada

Utilizando la definición anterior se exploraron todos los cromosomas autosómicos para cada raza. La Tabla No. 19 muestra un resumen de las variaciones estructurales que se encuentran en las 19 razas. El número total de variaciones encontradas entre las 19 razas fueron 15,246, pero al agrupar todas

las variaciones a través de todos los animales en la muestra, se definieron 9,146 regiones. El número total de SNPs que participan en las regiones fue de 53,137, que representan el 6.40% (160.98 Mb) del genoma bovino (véase la Tabla S7 en S7 archivo adicional No. 2 para los detalles del cromosoma, la posición inicial, la posición final, el tamaño en pares de bases, número de SNPs, número de razas y nombre de las razas involucradas en cada región). Las razas Hereford, Angus Rojo, y Jersey mostraron el mayor número de variaciones estructurales, con 1,444, 1,443, y 1,163, respectivamente, mientras que Nelore, Beefmaster y Gir mostraron el número más pequeño, con 296, 305 y 315 variaciones estructurales, respectivamente (véase Tabla No. 19)**.**

BREED	No. of SVs	Total Size (Mb)	Average Size (Kb)	Size Max (Kb)	Size Min (Kb)	No. of SNPs	Average SNPs per SV	Max SNPs per SV
ANG	1052	20.55	19.54	295.01	1.11	7676	7.29	93
BMA	305	5.08	16.67	197.73	1.15	1670	5.47	51
BRM	359	6.29	17.52	139.35	1.14	1871	5.2	35
BSW	1093	22.85	20.91	217.88	1.14	8030	7.34	100
CHL	748	12.09	16.16	218.68	1.00	4982	6.66	55
GIR	315	6.08	19.32	209.72	1.00	1756	5.57	41
GNS	1082	22.49	20.79	236.56	1.00	8044	7.43	62
HFD	1444	31.05	21.50	271.71	1.07	11363	7.86	87
HOL	855	14.15	16.56	220.68	1.28	5391	6.30	70
JER	1163	25.31	21.76	319.75	1.15	9001	7.73	102
LMS	849	13.08	15.41	191.31	1.00	5236	6.16	70
NEL	296	5.30	17.93	122.57	1.50	1560	5.27	29
NRC	1051	18.46	17.56	153.62	1.09	7223	6.87	62
NDA	1060	20.80	19.63	194.13	1.00	7242	6.83	48
PMT	727	10.82	14.89	166.70	1.06	4544	6.25	77
RGU	1443	29.59	20.50	280.74	1.06	11075	7.67	138
RMG	728	12.13	16.66	178.72	1.23	4715	6.47	78
SGT	336	6.39	19.02	396.03	1.17	1992	5.92	79
SHK	340	5.61	16.50	190.78	1.00	1733	5.09	38
TOTAL AVE.	**802.42**	**15.16**	**18.35**	**221.14**	**1.11**	**5531.78**	**6.49**	**69.21**

Tabla No. 19.- Estadísticas de variaciones encontradas en el genoma completo de todas de las razas.

La distancia promedio cubierta por las variaciones estructurales en el genoma completo en todas las razas fue de 15.16 Mb. La mayor región declarada como la variación estructural fue de la raza Santa Gertrudis, con 396 Kb de longitud. La región más pequeña fue localizada en las razas Charolais, Gir, Guernsey, Limuosine, N'Dama y Sheko con 1 Kb de tamaño. El promedio de SNPs por variación estructural fue de 6.49 SNPs. El promedio de las variaciones por Mb en el genoma completo fue de 0.30 y el promedio de las variaciones por cromosoma fue de 27.67 (véase la Tabla No. 20 y para mayores detalles la Tabla S5 el archivo adicional no. 2). El cromosoma con el promedio más alto de variaciones fue el cromosoma 5 con 0.46 variaciones por Mb, y el cromosoma con el promedio más bajo de variaciones fue el cromosoma 28, con 0.17 por Mb.

CHROM	ANG		BMA		BRM		BSW		CHL		GIR	
	Num.SVs	Av.per Mb	Num.SVs	Av.per Mb	Num.SVs	Av.per Mb	Num.SVs	Av.per Mb	Num.SVs	Av.per Mb	Num.SVs	Av.per Mb
1	75	0.47	15	0.09	33	0.21	79	0.50	41	0.26	12	0.08
2	65	0.47	22	0.16	24	0.18	70	0.51	59	0.43	22	0.16
3	58	0.48	12	0.10	18	0.15	47	0.39	33	0.27	11	0.09
4	54	0.45	13	0.11	11	0.09	41	0.34	34	0.28	15	0.12
5	93	0.77	18	0.15	29	0.24	88	0.73	48	0.40	29	0.24
6	53	0.44	10	0.08	24	0.20	124	1.04	44	0.37	20	0.17
7	54	0.48	23	0.20	21	0.19	38	0.34	31	0.28	23	0.20
8	28	0.25	14	0.12	8	0.07	35	0.31	37	0.33	7	0.06
9	50	0.47	18	0.17	13	0.12	56	0.53	43	0.41	18	0.17
10	30	0.29	14	0.13	18	0.17	38	0.36	34	0.33	9	0.09
11	41	0.38	16	0.15	15	0.14	86	0.80	30	0.28	18	0.17
12	44	0.48	10	0.11	18	0.20	46	0.50	40	0.44	11	0.12
13	31	0.37	5	0.06	7	0.08	20	0.24	16	0.19	8	0.09
14	43	0.51	18	0.21	27	0.32	25	0.30	33	0.39	14	0.17
15	34	0.40	14	0.16	8	0.09	35	0.41	18	0.21	8	0.09
16	37	0.45	15	0.18	7	0.09	59	0.72	33	0.40	22	0.27
17	13	0.17	12	0.16	6	0.08	23	0.31	18	0.24	9	0.12
18	22	0.33	6	0.09	12	0.18	24	0.36	16	0.24	6	0.09
19	25	0.39	11	0.17	5	0.08	16	0.25	7	0.11	5	0.08
20	35	0.49	5	0.07	9	0.12	18	0.25	17	0.24	6	0.08
21	25	0.35	15	0.21	11	0.15	25	0.35	14	0.20	14	0.20
22	27	0.44	5	0.08	13	0.21	18	0.29	21	0.34	4	0.07
23	17	0.32	1	0.02	2	0.04	7	0.13	12	0.23	4	0.08
24	28	0.45	4	0.06	11	0.18	17	0.27	23	0.37	2	0.03
25	19	0.44	2	0.05	2	0.05	14	0.33	12	0.28	4	0.09
26	23	0.45	1	0.02	0	0.00	10	0.19	17	0.33	4	0.08

27	11	0.24	4	0.09	1	0.02	13	0.29	5	0.11	5	0.11
28	8	0.17	1	0.02	2	0.04	8	0.17	3	0.06	2	0.04
29	9	0.17	1	0.02	4	0.08	13	0.25	9	0.17	3	0.06
TOTAL	**1052**	**11.58**	**305**	**3.26**	**359**	**3.77**	**1093**	**11.46**	**748**	**8.17**	**315**	**3.41**
AVERAGE	**36.28**	**0.40**	**10.52**	**0.11**	**12.38**	**0.13**	**37.69**	**0.40**	**25.79**	**0.28**	**10.86**	**0.12**

Tabla No. 20: Variaciones por cromosoma y promedios por megabase.

De la Tabla No. 19 se puede observar que las razas con el mayor número de SNPs que caen en SV (a partir de ahora y en abreviamos variación estructural como SV) fueron Hereford con 11363, Angus Roja con 11075, Jersey con 9,001, Guernsey con 8,044, Brown Swiss con 8030 y Angus con 7676, todos de la subespecie *Bos taurus taurus*. Mientras, Nelore, Beefmaster, Sheko, Gir, Brahman y Santa Gertrudis fueron las razas con el menor número de SNPs que caen en SV, con 1560, 1670, 1733, 1756, 1871 y 1992 respectivamente. Todas estas razas, pertenecen a la subespecie *Bos Taurus indicus* y grupos compuestos. El número promedio de SNPs que cae en las SV en todas las razas fue 5531,78. Las razas con el promedio más alto de SNPs por SV fueron Hereford, Jersey, Angus Roja, Guernsey, Pardo Suizo, Angus y Noruega Roja, todas del grupo de *Bos taurus*, mientras que las razas con el promedia más bajo de SNPs por SV fueron Sheko, Brahman, Nelore, Beefmaster, Gir, y Santa Gertrudis, del grupo *Bos indicus* y grupos compuestos.

Las 5 razas con SV que contienen el mayor número de SNPs (última columna de la Tabla No. 19) fueron Angus roja, Jersey, Pardo Suizo, Angus y Hereford, todos del grupo *Bos taurus*. Las razas asiáticas resultaron con la mejor cantidad de SNPs por SV.

De acuerdo con el número de variaciones estructurales encontradas para cada raza, (ver Figura No. 17), raza Hereford tiene el mayor número de variaciones, mientras que Nelore tiene el más pequeño. Hay más de tres veces de una diferencia entre las razas representadas en este estudio. Incluso si hay artefactos asociados a errores de montaje del genoma, todavía existe gran evidencia de inmensas diferencias entre razas en estos resultados.

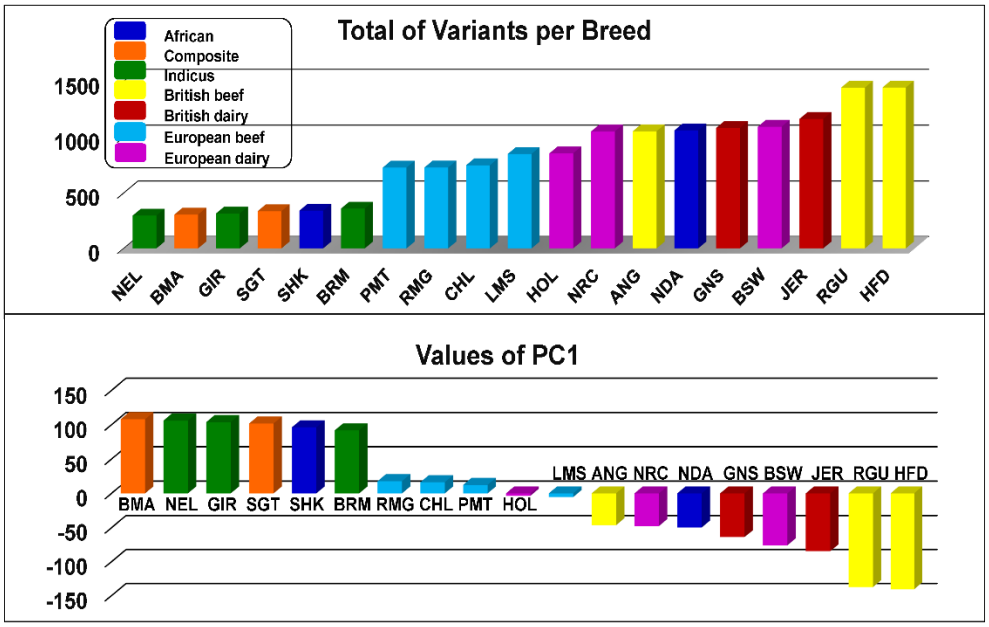


Figura no. 17: Numero de variaciones estructurales por raza y valores del componente principal uno.

3.5 Discusión.

3.5.1 Cercanía entre las razas por su región geográfica.

Con el fin de investigar la cercanía de la variabilidad entre razas, dada la cantidad de variaciones estructurales identificadas, para cada raza se construyó un vector de 29 campos, donde cada campo contenía el número de SVs en cada cromosoma. Se aplicó un análisis de componentes principales [98] para estos vectores. Los resultados se muestran en la Figura No. 18, donde se observa cómo algunas razas muestran cercanía dado la región geográfica donde evolucionaron. Las razas de carne europeas: Piedmontese, Charolais, Limousin y Romagnola, por ejemplo, aparecen muy cerca, es posible que refleje la cercanía geográfica de las regiones evolución entre Italia y Francia. Las razas asiáticas (*Bos indicus*): Gir, Brahman y Nelore resultaron con cargas positivas si observan desde PC1 y se muestran muy próximas. De la misma manera las razas mixtas: Santa Gertrudis y Beefmaster resultaron con cargas positivas (por PC1), y se muestran muy cerca también. Algunas razas británicas: Angus, Angus Roja, Hereford, Jersey y Guernsey y las razas europeas: Pardo Suizo, Noruega Roja y Holstein, todos resultaron con cargas negativas, cuando se observan desde el eje PC1, y aparecen relativamente separadas de los otros grupos. Por otro lado, las dos razas del grupo africano, N'Dama y Sheko, aparecen relativamente lejos el uno del otro. Esto se

explica por el hecho de que Sheko es una raza *Taurus Indicus* mientras N'Dama es *Taurus taurus.*

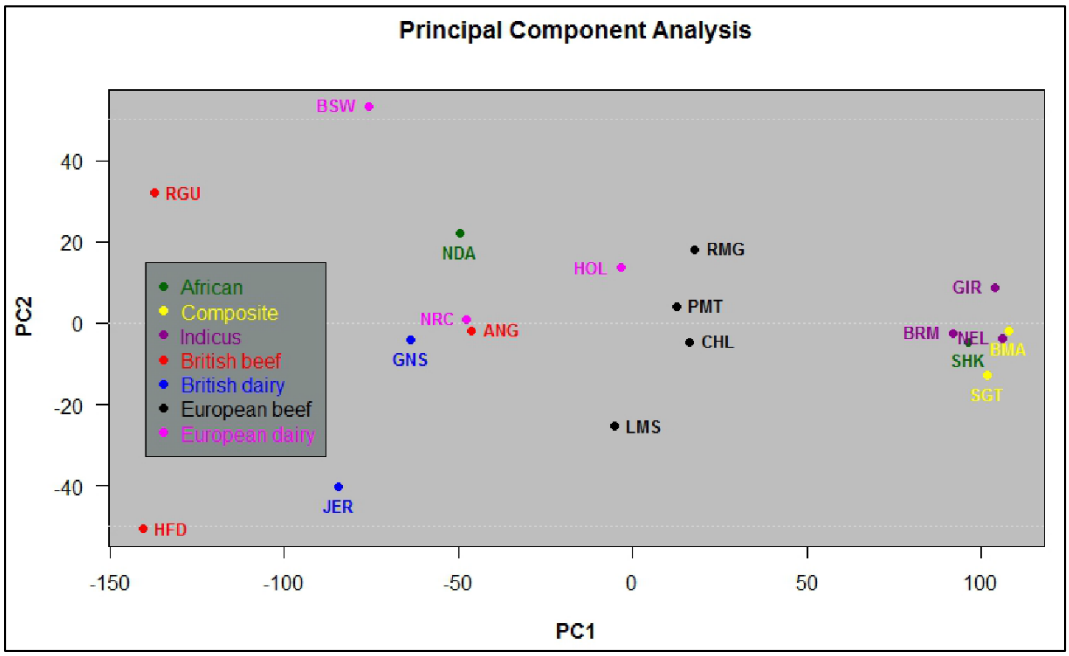


Figura no. 18: Análisis de componentes principales.

En apoyo de nuestros resultados que muestran una clara diferenciación regional entre grupos raciales, dada la cantidad de variaciones, se analizó el resultado de un estudio que se llevó a cabo en el cromosoma "Y" del ganado doméstico europeo [99], los autores encontraron una separación notable de un conjunto de haplotipos llamado Y1, que aparecen con una alta frecuencia en las razas de ganado que se encuentran en la región noroeste de Europa y las islas Británicas, y otro conjunto de haplotipos denominados Y2, con un carácter dominante en el ganado ubicadas en la región del Sur de Europa.

3.5.2 Comparación con Otras Variaciones Estructurales Reportadas.

Dado que encontramos el 6.4% del genoma involucrada con SVs, intuitivamente sugerimos superposición con otros tipos de variaciones ya informadas por otros grupos. A continuación, analizamos cuanta superposición existe entre nuestros resultados y las variaciones estructurales de tipo CNVs. Se comparó el número, la posición y la longitud de las variaciones estructurales. La Figura 19 presenta una comparación de nuestros resultados con seis estudios sobre CNVs reportados por Jiang et al. (2013) [100], Cicconardi et al. (2013) [85], Hou et al. (2011, 2012a y 2012b) [87,88,101] y Liu et al. (2010) [90]. Los tres primeros estudios

consideraron sólo las razas Holstein y Angus. Para la comparación se consideraron sus regiones de CNVs sólo en los cromosomas autosómicos. Para Jiang et al (2013) encontramos 113 superposiciones de 358 CNVRs, lo que representa el 31.56%. Para Cicconardi et al. (2013), encontramos 311 regiones se solapan, lo que representa el 78.73% de las 395 regiones reportados. Para Hou et al (2012a, 2012b, 2011), encontramos 178, 505 y 267 superposiciones de 462, 3438 y 672 variaciones reportadas, lo que representa el 38.52%, 14.68% y 39.73% de las regiones se solapan respectivamente. Finalmente, para Liu et al. (2010), se consideró que sólo las razas once que entre ambos estudios tienen en común. Hemos encontrado 36 regiones se solapan, lo que representa el 25,35% de las 142 regiones reportados.

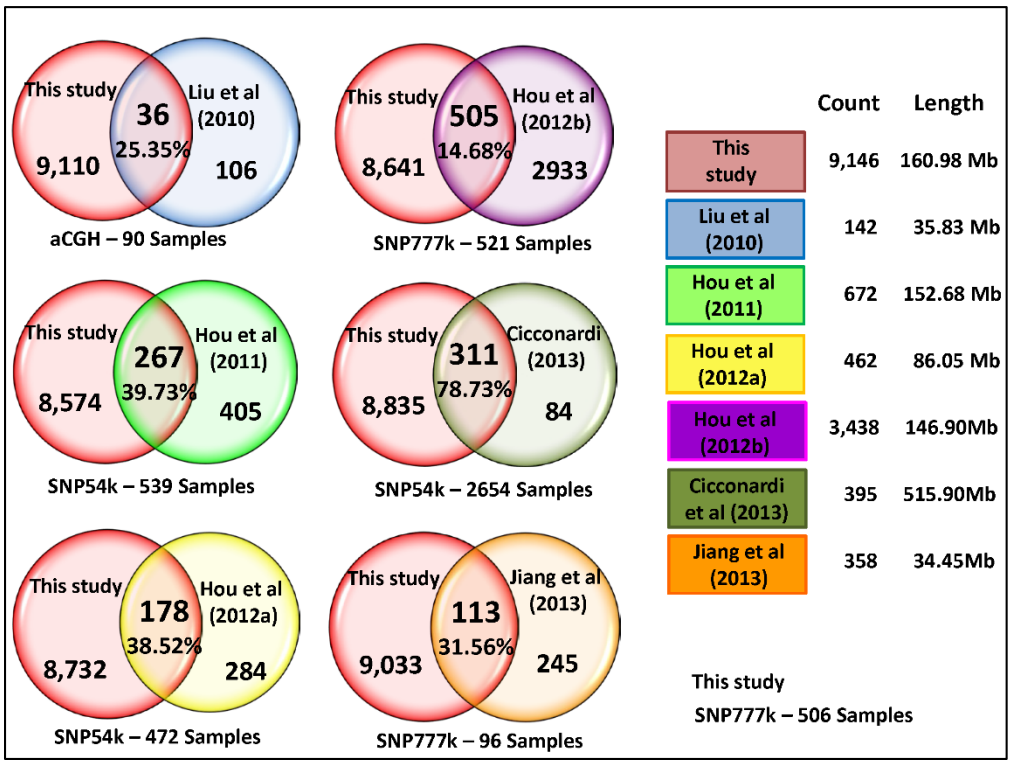


Figura no. 19: Comparación de nuestras variaciones estructurales detectadas contra otras variaciones reportadas.

A continuación, se realizaron búsquedas en NCBI [102], para los genes que participan en nuestras variaciones estructurales definidas. El mayor número de genes se encontró en las razas Angus Roja, Hereford y Jersey, con 490, 487 y 379 genes, respectivamente. El menor número de genes se encontró en las razas Beefmaster, Nelore y Santa Gertrudis, con 105, 115 y 126 genes,

respectivamente. (ver Tabla No. 21 y para mayores detalles ver Tabla S6 en archivo adicional No.2 el número de cromosomas, la posición, la posición final, el tamaño en pares de bases, el nombre del gen, el tipo de genes, y la descripción de genes empezar, Tabla No. 22 y Tabla S7 en el archivo adicional No.2 para el número de genes por raza, y Tabla No. 23 y Tabla S8 en el archivo adicional No.2 para el número de genes implicados con variaciones estructurales en cada cromosoma).

Chr	Start position	End position	Size in base pair	Name of Gene	Type of Gene	Description
1	975149	985262	10113	ITSN1	mRNA	intersectin 1 (SH3 domain protein)
1	1561012	1568515	7503	IL10RB	best RefSeq	interleukin 10 receptor. beta
1	2331080	2341224	10144	EVA1C	best RefSeq	eva-1 homolog C (C. elegans)
1	3556578	3568205	11627	TIAM1	mRNA	T-cell lymphoma invasion and metastasis 1
1	5890792	5912013	21221	GRIK1	best RefSeq	glutamate receptor. ionotropic. kainate 1
1	6385860	6391056	5196	C1H21orf7	mRNA	chromosome 1 open reading frame. human C21orf7
1	8808152	8819375	11223	ADAMTS5	best RefSeq	ADAM metallopeptidase with thrombospondin type 1 motif. 5
1	9305910	9328940	23030	CYYR1	best RefSeq	cysteine/tyrosine-rich 1
1	14781324	14795765	14441	NCAM2	mRNA	neural cell adhesion molecule 2
1	18737281	18788148	50867	C1H21orf91	best RefSeq	chromosome 1 open reading frame. human C21orf91
1	18980052	18981357	1305	CXADR	best RefSeq	coxsackie virus and adenovirus receptor
1	20650149	20680025	29876	USP25	mRNA	ubiquitin specific peptidase 25
1	24280892	24284868	3976	ROBO2	mRNA	roundabout. axon guidance receptor. homolog 2 (Drosophila)
1	24756916	24765965	9049	LOC781698	protein	putative Bcl-2 homologous antagonist/killer 2-like
1	26041310	26055067	13757	ROBO1	best RefSeq	roundabout. axon guidance receptor. homolog 1 (Drosophila)
1	28682147	28777565	95418	GBE1	best RefSeq	glucan (1.4-alpha-). branching enzyme 1
1	32402235	32411801	9566	CADM2	mRNA	cell adhesion molecule 2
1	34960825	34990108	29283	CHMP2B	best RefSeq	charged multivesicular body protein 2B
1	36768414	36799251	30837	EPHA3	best RefSeq	EPH receptor A3

Tabla No. 21.-Tipo, nombre y descripción de los genes encontrados en este estudio.

NAME OF BREED	NUMBER OF GENES
BEEF MASTER	105
SANTA GERTRUDIS	126
BRAHMAN	156
HOLSTEIN	273
NELORE	115
JERSEY	379
BROWN SWISS	376
ROMAGNOLA	241
HEREFORD	487
GIR	133
ANGUS	360
GUERNSEY	355
SHEKO	129
LIMOUSIN	277
CHAROLAIS	235
RED NORWEGIAN	344
N'DAMA	356
RED ANGUS	490
PIEDMONTESE	235
	5172

Tabla 22: Número de genes por raza.

CHROMOSOME	NUMBER OF GENES
1	179
2	183
3	195
4	136
5	205
6	144
7	162
8	101
9	135
10	139
11	186
12	62

13	86
14	88
15	104
16	111
17	64
18	115
19	129
20	71
21	99
22	81
23	43
24	64
25	64
26	59
27	23
28	31
29	50
	3109

Tabla No. 23: Número de Genes por cromosoma.

El último análisis fue observar dentro de nuestras variaciones estructurales definidas las regiones que fueron consistentes en todas las 19 razas. No encontramos regiones comunes para todas las razas. Sin embargo, hay 2 regiones en donde 17 razas coincidieron, 2 regiones donde 15 razas coincidieron, 2 regiones donde 14 razas coincidieron, 3 regiones donde 13 razas coincidieron y 13 regiones donde 12 razas coincidieron, incluso cuando no coincidan con la longitud total, al menos, compartían un segmento declarado variación. (Véase la Tabla No.24 y la Tabla S9 en el archiva adicional No. 2 para mayores detalles).

Index	Chr.	Start position	End position	Size in base pair	Number of SNPs.	Number of Breeds	Breeds
1	1	260275	301573	41298	10	1	ANG
2	1	902161	903971	1810	3	1	RGU
3	1	975149	985262	10113	4	1	RMG
4	1	1561012	1568515	7503	7	1	JER

5	1	1570091	1576880	6789	6	2	ANG-RGU
6	1	1798406	1821474	23068	6	1	SGT
7	1	1828653	1859264	30611	3	1	GNS
8	1	1865510	1883632	18122	5	4	CHL-HOL-JER-LMS
9	1	2331080	2341224	10144	6	2	JER-RGU
10	1	3230920	3240079	9159	5	2	ANG-PMT
11	1	3254174	3260878	6704	4	1	PMT
12	1	3556578	3568205	11627	3	1	RMG
13	1	4030192	4033310	3118	3	1	GNS
14	1	4564983	4572247	7264	3	1	NEL
15	1	5398072	5402036	3964	3	2	HOL-LMS
16	1	5890792	5912013	21221	4	1	JER
17	1	5914159	5932501	18342	11	1	HFD
18	1	5947212	5951378	4166	4	1	JER
19	1	6065203	6083856	18653	6	1	ANG

Tabla No. 24: Regiones de variaciones estructurales.

Capítulo IV

Conclusiones y trabajo futuro

4.1 Introducción.

Este proyecto de tesis se enfocó al estudio de las variaciones estructurales en el genoma completo del ganado Bovino, partiendo de 506 Genotipos del Consorcio Internacional Bovino HapMap y 12 muestras de cortesía de la compañía Affymetrix. Con ambas muestras, tuvimos la gran ventaja de trabajar con genotipos en alta densidad, utilizando los dos microarreglos más modernos para la genotipificación de SNPs en el ganado bovino como lo son: 1) el BovineHD Genotyping BeadChip de Illumina y 2) el Axiom Genome-Wide BOS 1 Array de Affymetrix, que son las dos compañías líderes del mercado en este tipo de tecnologías.

4.1.1 Genotipos del Axiom Genome-Wide BOS 1 Array de Affymetrix.

Con los genotipos de este microarreglo, logramos desarrollar el primer estudio de genoma completo en alta densidad de variaciones del número de copias en ganado mexicano de la raza Holstein. Utilizando los 648,315 SNPs que provee el Axiom Genome-Wide BOS 1 Array de Affymetrix de 12 vacas. Esta información fue aplicada a los dos algoritmos más utilizados para la detección de CNVs, los cuales detectaron en conjunto 56 CNVRs distribuidas a través de los 29 cromosomas autosomales y de estas, 20 fueron CNVRs nuevas aun no reportadas. Para comprobar la validación de nuestra detección, seleccionamos 7

CNVRs, para que fueran sometidas a pruebas por medio del qPCR, mismo que arrojó como resultado la comprobación de 5 CNVRs de 7.

El contar con SNPs en alta densidad, nos permite alcanzar una gran precisión en la identificación de CNVs y genes candidatos. Asimismo permiten un mapeo más fino y preciso en LD, lo que da a los investigadores una gran ventaja, al contar con información definida y validad, que servirá para futuros estudios de asociación.

Nuestros resultados proveen una nueva referencia para variación genómica y para los estudios de asociación entre CNVs y fenotipos, especialmente en ganado mexicano.

4.1.2 Genotipos del BovineHD Genotyping BeadChip de Illumina.

Con los genotipos de este microarreglo, presentamos los resultados de un trabajo de investigación, con una simple definición de variación estructural genómica, basada en la desviación esperada del desequilibrio ligado de rango corto entre SNPs. Desarrollando esta definición realizamos un análisis del genoma completo en el ganado bovino.

El número total de variaciones encontradas entre las 19 razas fue de 15,246. Estas al ser agrupadas definieron 9,146 regiones. El número total de SNPs involucrados en estas regiones fue de 53,137, representando el 6.40% (160.98 Mb) del genoma bovino. El número de genes que cobren estas variaciones fue de 3,109. Al compararnos con estudios de variaciones en CNVs previamente reportados, tuvimos traslapes que van desde un rango de 14.68% hasta 78.73%.

Un análisis de diferenciación basado en el número de variaciones estructurales del genoma completo, mostro gran diferencia entre algunas razas y revelo la cercanía de los grupos dada la región geográfica donde ellos evolucionaron. Finalmente, aun cuando existe un traslape de variaciones estructurales basadas en LD con CNVs, estas variaciones capturan diferentes patrones de variación geonómica, y futuros estudios serán necesarios para relacionar su asociación con enfermedades u otro tipo de rasgos fenotípicos.

4.2 Trabajo Futuro.

Con los resultados y conclusiones obtenidos de este trabajo de investigación y a fin de continuar explotando la información de genotipos con la que contamos para darle continuidad a este proyecto, tenemos el próximo trabajo a realizar:

- Con las muestras del HapMap trabajaremos en el análisis de un algoritmo que detecte las señales de selección (selective sweep) en el ganado bovino.

- Una vez definido este algoritmo, continuaremos con el desarrollo de una plataforma amigable, con la automatización del algoritmo para la detección de variaciones estructurales basadas en LD de rango corto, que permita a los usuarios la manipulación de los patrones de control, así como la obtención de diferentes salidas como gráficas, LD, PCAs, etc.

- Un estudio de asociación, que nos permita diferenciar las características del ganado local, obtenidas con los SNPs del microarreglo de Affymetrix, contra los SNPs obtenidos por el BovineHD Genotyping BeadChip de Illumina.

- Participar en el proyecto de selección y mejora genética de la Vaca Cachanilla.

Referencias Bibliográficas

1. Prigogine I (2012) El nacimiento del tiempo. Buenos Aires, Argentina: Fabula Tusquets editores.
2. Allison LA (2007) Fundamental molecular biology. Malden, MA: Blackwell Pub. xxi, 725 p. p.
3. Clark DP, Russell LD (2005) Molecular biology : made simple and fun. St. Louis, MO: Cache River Press. vii, 515 p. p.
4. Zhang A (2006) Advanced analysis of gene expression microarray data. New Jersey: World Scientific. xv, 339 p. p.
5. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, et al. (1996) A gene map of the human genome. Science 274: 540-546.
6. Surzycki S (2003) Human molecular biology laboratory manual. Malden, MA: Blackwell Pub. xiv, 229 p. p.
7. Bartlett JM, Stirling D (2003) A short history of the polymerase chain reaction. Methods Mol Biol 226: 3-6.
8. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408.
9. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420-426.
10. Conrad DF, Hurles ME (2007) The population genetics of structural variation. Nat Genet 39: S30-36.
11. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738.
12. Alba M (2001) Replicative DNA polymerases. Genome Biol 2: REVIEWS3002.
13. Laird NM, Lange C (2011) The fundamentals of modern statistical genetics. New York: Springer Science. xiv, 223 p. p.
14. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566-1575.
15. Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4: 587-597.
16. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77: 78-88.
17. Fusté MC (2012) Studies in population genetics. Croacia: InTech.
18. Lewontin RC (1964) The Interaction of Selection and Linkage. Ii. Optimum Models. Genetics 50: 757-782.
19. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38: 1251-1260.
20. Foulkes AS (2009) Applied statistical genetics with R : for population-based association studies. New York: Springer Verlag. xxiii, 252 p. p.

21. Mathur SK (2010) Statistical bioinformatics with R. Amsterdam ; Boston: Academic Press/Elsevier. xvi, 319 p., 318 p. of plates p.
22. Jolliffe IT (2002) Principal component analysis. New York: Springer. xxix, 487 p. p.
23. Fraser AM (2008) Hidden Markov models and dynamical systems. Philadelphia, PA: Society for Industrial and Applied Mathematics. xii, 132 p. p.
24. Rosner B (2011) Fundamentals of biostatistics. Boston: Brooks/Cole, Cengage Learning. xvii, 859 p. p.
25. Balding DJ, Bishop MJ, Cannings C (2007) Handbook of statistical genetics. Chichester, England ; Hoboken, NJ: John Wiley & Sons. p. p.
26. de la Calle G, Garcia-Remesal M, Chiesa S, de la Iglesia D, Maojo V (2009) BIRI: a new approach for automatically discovering and indexing available public bioinformatics resources from the literature. BMC Bioinformatics 10: 320.
27. Winchester L, Yau C, Ragoussis J (2009) Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 8: 353-366.
28. Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013-2025.
29. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.
30. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
32. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-360.
33. Hu ZL, Park CA, Wu XL, Reecy JM (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res 41: D871-879.
34. Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol Chapter 22: Unit 22 21.
35. Fuentes M (2013) Protein microarrays overview. Recent Pat Biotechnol 7: 83.
36. Sutandy FX, Qian J, Chen CS, Zhu H (2013) Overview of protein microarrays. Curr Protoc Protein Sci Chapter 27: Unit 27 21.
37. Jawhar NM (2009) Tissue Microarray: A rapidly evolving diagnostic and research tool. Ann Saudi Med 29: 123-127.
38. Tarca AL, Romero R, Draghici S (2006) Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol 195: 373-388.
39. Conrad DF, Hurles ME (2007) The population genetics of structural variation. Nat Genet 39: S30-36.
40. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704-712.

41. Bovine Genome S, Analysis C, Elsik CG, Tellam RL, Worley KC, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.
42. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10: 451-481.
43. Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, et al. (2010) Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. BMC Genomics 11: 673.
44. Fadista J, Thomsen B, Holm LE, Bendixen C (2010) Copy number variation in the bovine genome. BMC Genomics 11: 284.
45. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, et al. (2011) Genomic characteristics of cattle copy number variations. BMC Genomics 12: 127.
46. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, et al. (2013) Massive screening of copy number population-scale variation in Bos taurus genome. BMC Genomics 14: 124.
47. Jiang L, Jiang J, Wang J, Ding X, Liu J, et al. (2012) Genome-wide identification of copy number variations in Chinese Holstein. PLoS One 7: e48732.
48. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, et al. (2010) Analysis of copy number variations among diverse cattle breeds. Genome Res 20: 693-703.
49. Hou Y, Liu GE, Bickhart DM, Matukumalli LK, Li C, et al. (2012) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Funct Integr Genomics 12: 81-92.
50. Jiang L, Jiang J, Yang J, Liu X, Wang J, et al. (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. BMC Genomics 14: 131.
51. Hou Y, Bickhart DM, Hvinden ML, Li C, Song J, et al. (2012) Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. BMC Genomics 13: 376.
52. Doan R, Cohen N, Harrington J, Veazey K, Juras R, et al. (2012) Identification of copy number variants in horses. Genome Res 22: 899-907.
53. Wang J, Jiang J, Fu W, Jiang L, Ding X, et al. (2012) A genome-wide detection of copy number variations using SNP genotyping arrays in swine. BMC Genomics 13: 273.
54. Wang J, Wang H, Jiang J, Kang H, Feng X, et al. (2013) Identification of genome-wide copy number variations among diverse pig breeds using SNP genotyping arrays. PLoS One 8: e68683.
55. Liu J, Zhang L, Xu L, Ren H, Lu J, et al. (2013) Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. BMC Genomics 14: 229.
56. Crooijmans RP, Fife MS, Fitzgerald TW, Strickland S, Cheng HH, et al. (2013) Large scale variation in DNA copy number in chicken breeds. BMC Genomics 14: 398.
57. Fontanesi L, Martelli PL, Scotti E, Russo V, Rogel-Gaillard C, et al. (2012) Exploring copy number variation in the rabbit (Oryctolagus cuniculus) genome by array comparative genome hybridization. Genomics 100: 245-251.
58. Alvarez CE, Akey JM (2012) Copy number variation in the domestic dog. Mamm Genome 23: 144-163.
59. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-1148.

60. Zhao L, Bracken MB, Dewan AT (2013) Genome-Wide Association Study of Pre-Eclampsia Detects Novel Maternal Single Nucleotide Polymorphisms and Copy-Number Variants in Subsets of the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study Cohort. Ann Hum Genet.
61. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863-885.
62. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.
63. Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013-2025.
64. Xu Y, Peng B, Fu Y, Amos CI (2011) Genome-wide algorithm for detecting CNV associations with diseases. BMC Bioinformatics 12: 331.
65. Diskin SJ, Li M, Hou C, Yang S, Glessner J, et al. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 36: e126.
66. D'Haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50: 262-270.
67. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408.
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
69. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-360.
70. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
71. Hu ZL, Park CA, Wu XL, Reecy JM (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res 41: D871-879.
72. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, et al. (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22: 778-790.
73. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38: D613-619.
74. Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, et al. (2009) Analysis of recent segmental duplications in the bovine genome. BMC Genomics 10: 571.
75. Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, et al. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32: 135-142.
76. Wain LV, Tobin MD (2011) Copy number variation. Methods Mol Biol 713: 167-183.
77. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

78. Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013-2025.
79. Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. Hum Mol Genet 16 Spec No. 2: R168-173.
80. Wineinger NE, Pajewski NM, Tiwari HK (2011) A Method to Assess Linkage Disequilibrium between CNVs and SNPs Inside Copy Number Variable Regions. Front Genet 2: 17.
81. Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, et al. (2009) Genome-wide comparisons of variation in linkage disequilibrium. Genome Res 19: 1849-1860.
82. Kadri NK, Koks PD, Meuwissen TH (2012) Prediction of a deletion copy number variant by a dense SNP panel. Genet Sel Evol 44: 7.
83. Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, et al. (2010) Identification of copy number variations and common deletion polymorphisms in cattle. BMC Genomics 11: 232.
84. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, et al. (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22: 778-790.
85. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, et al. (2013) Massive screening of copy number population-scale variation in Bos taurus genome. BMC Genomics 14: 124.
86. Fadista J, Thomsen B, Holm LE, Bendixen C (2010) Copy number variation in the bovine genome. BMC Genomics 11: 284.
87. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, et al. (2011) Genomic characteristics of cattle copy number variations. BMC Genomics 12: 127.
88. Hou Y, Liu GE, Bickhart DM, Matukumalli LK, Li C, et al. (2012) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Funct Integr Genomics 12: 81-92.
89. Jiang L, Jiang J, Wang J, Ding X, Liu J, et al. (2012) Genome-wide identification of copy number variations in Chinese Holstein. PLoS One 7: e48732.
90. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, et al. (2010) Analysis of copy number variations among diverse cattle breeds. Genome Res 20: 693-703.
91. Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, et al. (2010) Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. BMC Genomics 11: 673.
92. Matukumalli LK, S. Schroeder SK, DeNise T, Sonstegard CT, Lawley M, et al. (2011) Analyzing LD blocks and CNV segments in cattle: Novel genomic features identified using the BovineHD BeadChip. Pub No 370-2011-002, Illumina Inc, San Diego, CA.
93. Bovine HapMap C, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, et al. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324: 528-532.
94. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81: 1084-1097.
95. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499-511.
96. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, et al. (2009) High-resolution haplotype block structure in the cattle genome. BMC Genet 10: 19.
97. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125: 279-284.
98. Jolliffe IT (2002) Principal component analysis. New York: Springer. xxix, 487 p. p.

99. Perez-Pardal L, Royo LJ, Beja-Pereira A, Chen S, Cantet RJ, et al. (2010) Multiple paternal origins of domestic cattle revealed by Y-specific interspersed multilocus microsatellites. Heredity (Edinb) 105: 511-519.
100. Jiang L, Jiang J, Yang J, Liu X, Wang J, et al. (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. BMC Genomics 14: 131.
101. Hou Y, Bickhart DM, Hvinden ML, Li C, Song J, et al. (2012) Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. BMC Genomics 13: 376.
102. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, et al. (2009) A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10: R42.

Anexos

Archivo Adicional No.1 (.XLSX).- Tablas contenidas dentro de este archivo adicional:

Tabla S1.- Resultados del análisis cuantitativo del PCR en tiempo real de CNVRs confirmadas.
Tabla S2.- Detalle de cada CNV detectada en este estudio.
Tabla S3.- Características de las Regiones de CNVs, con tamaños en bases par (bp).
Tabla S4.- Genes que se encuentran dentro o que se traslapan en las Regiones de CNVs identificadas en este estudio.
Tabla S5.- Análisis de ontología (GO) y de rutas metabólicas (KEGG) de los genes detectados en este estudio.
Tabla S6.- QTLs que se encuentran dentro o que se traslapan en las Regiones de CNVs identificadas en este estudio.

Archivo Adicional No.2 (.XLSX).- Tablas contenidas dentro de este archivo adicional:

Tabla S1.- Razas y número de animales en la muestra.
Tabla S2.- Resumen final de marcadores en nuestro estudio.
Tabla S3.- Promedios de la media y mediana en las 19 razas.
Tabla S4.- Valores promedios obtenidos por r^2 para todas las razas (caída natural LD).
Tabla S5.- Variaciones por cromosoma y promedios por Megabase.
Tabla S6.- Tipo, nombre y descripción de los genes encontrados en este estudio.
Tabla S7.- Número de genes por raza.
Tabla S8.- Número de Genes por cromosoma.
Tabla S9.- Listado de regiones de variaciones estructurales.

Archivo Figure S1 (.DOCX).- Figuras contenidas dentro de este archivo adicional:

Figura S1.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias por raza.
Figura S2.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias del grupo Africano.
Figura S3.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias del grupo Indicus.
Figura S4.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias del grupo Mixto.
Figura S5.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias del grupo de carne.
Figura S6.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias del grupo lechero.
Figura S7.- Distribución del MAF. Promedio de proporciones de SNPs de varias frecuencias por raza (tabla con porcentajes).