UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERÍA

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

“Estudio de Caracterización de Polimorfismos de Nucleótido Simple No Sinónimos en el Genoma Bovino”

TESIS PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA
PATRICIA CARVAJAL LÓPEZ

DIRECTOR
RAFAEL VILLA ANGULO

CODIRECTOR
VICTOR MANUEL GONZÁLEZ VIZCARRA

Mexicali, B. C. **Enero/2013**

AGRADECIMIENTOS

Al Instituto de Ingeniería de la Universidad Autónoma de Baja California y el CONACYT por la oportunidad brindada…

A mi tutor, sinodales, maestros y compañeros por sus enseñanzas, tiempo y dedicación para que pudiese completar este nuevo paso en mi formación profesional…

A mi familia y amigos por su cariño, comprensión y paciencia…

Y a todos los que me apoyaron en este proceso…

¡GRACIAS!

ÍNDICE

FIGURAS

GRÁFICAS

TABLAS

I. INTRODUCCIÓN

1.1 ANTECEDENTES

El secuenciamiento o secuenciación de genomas de especies agropecuarias da pie a comprender patrones evolutivos y hacer investigación sobre mejora genética de rasgos productivos. En años recientes, se han hecho estudios de caracterización genética (conocidos también como genotipificación o genotipado) en organismos modelo y especies como aves, caninos, equinos, porcinos y bovinos [1]. En el 2009, el Consorcio de Análisis y Secuenciado del Genoma Bovino publicó la anotación del genoma bovino Btau 4.0, dando a conocer una longitud estimada del genoma de 2.87 giga pares-base, y aproximadamente 22,000 genes codificantes[2] [3]. Zimm et al publicó también en el 2009 el ensamble alterno UMD 3.1 del genoma bovino [4]. Ese mismo año el Consorcio del Mapa de Haplotipos Bovino (Bovine HapMap) usó como base la secuencia de una hembra Hereford y secuencias comparativas de seis razas adicionales para desarrollar pruebas en 37,470 polimorfismos de nucleótido simple (acrónimo en inglés: SNP, "Single Nucleotide Polymorphism"). Las proporciones intragénicas, intrónicas y exónicas de los SNPs fueron 63.74, 34.90 y 1.35 % respectivamente. Algunos estudios iniciales realizados en base a esta información fueron correlación entre razas, desequilibrio ligado, tamaño de población efectiva,

taza de mutación, heterocigocidad en pares base de nucleótidos, barrido selectivos, variantes del número de copias ("CNV", o "Copy Number Variation" por su significado en inglés), y grado de diferenciación entre subpoblaciones [5]. Las plataformas de arreglos o matrices de SNPs a gran escala se han usado ampliamente como paneles de referencia estándar en estudios genómicos. La información arrojada de dichos paneles llevaron a estudios de asociación de genoma completo (GWA, o "Genome-Wide Association Study" por sus siglas en inglés) e identificación de locus de carácter cuantitativo (QTL por su acrónimo en inglés "Quantitative Trait Loci") [6]. Se requieren paneles de densidades de más de 580,000 SNPs a lo largo de todo de genoma bovino, para generar estudios profundos de GWA, y realizar ciertos análisis como la caracterización de estructura de bloques de haplotipos [7]. El año 2010 estuvieron disponibles al público paneles de mayores densidades. La casa comercial Illumina ofreció el panel "Bovine BeadChip Genotyping Array" el cual cubre 777,962 SNPs a lo largo de todo el genoma; por su parte la casa Affymetrix ofreció el panel "Axiom Genome-Wide Bos 1 Array" con 648,874 SNPs [8]. Estos paneles permitieron ampliar la cobertura y mejorar la calidad de estudios y evaluaciones genómicas previas.

En contraste con estudios de rasgos en humanos, los modelos de especies agropecuarias se enfocan mayoritariamente en el fenotipo (que es la manifestación externa de un conjunto de caracteres hereditarios que dependen tanto de los genes como del ambiente) y el mérito genético (el conjunto de características de mayor interés económico, para las cuales un individuo transmite genes a su progenie). Típicamente, un estudio de asociación de genoma completo (GWA) registra un rasgo de interés en un grupo de individuos, después, dichos individuos son "genotipados" con un panel de SNPs para detectar asociaciones estadísticas entre el rasgo de interés

y el polimorfismo. Los marcadores validados del estudio de GWA en especies agropecuarias son usados para seleccionar mejores animales a través de selección asistida por marcadores ("Marker Assisted Selection", o "MAS" por su acrónimo en inglés). Uno de los dos tipos de "MAS" utiliza mutaciones causales que fueron localizadas en genes o regiones regulatorias (por ejemplo, SNPs en regiones codificantes); el segundo "MAS", utiliza SNPs que están en desequilibrio ligado con locus de carácter cuantitativo (QTL) [9].

El descubrimiento de asociación de genes y polimorfismos se realiza típicamente de dos formas: la primera es realizada "en base a su rol en la fisiología del rasgo"; el otro enfoque utiliza marcadores genéticos para asignar genes a una ubicación cromosómica y así estudiar rasgos de interés [9].

Uno de los objetivos principales de los estudios genéticos es entender el papel que desempeñan los polimorfismos y su relación con determinados fenotipos [10]. Dentro de las variaciones presentadas en un genoma, se cree que las variantes que tienen mayor impacto en salud y/o fenotipo son las ubicadas dentro de regiones codificantes, como los polimorfismos de nucleótido simple no sinónimo (nsSNP: "non synonym Single Nucleotide Polymorphism"), ya que estos polimorfismos representan un potencial cambio en la estructura de una proteína, y dada la variación estructural, dan pie a un posible cambio en la función de ésta [11].

Existen en la actualidad diversas formas de estimar el potencial impacto a nivel funcional que representan los polimorfismos de nucleótido simple no sinónimo (nsSNP); puede ser analizado por homología de secuencias, homología de estructura proteica, y/o incluir anotación. Para obtener predicciones de la función de proteínas, se diseñaron métodos como SIFT [12], SNAP [13], y PANTHER PSEC [14] que utilizan homología de secuencia; SNPs3D [15], PMUT[16], TopoSNP [17], y

nsSNPAnalyzer [18], que utilizan homología de secuencia y estructura; y PolyPhen [19], que usa secuencia, estructura, y anotación.

En la búsqueda de genes involucrados con determinados rasgos, la asociación genotipo-fenotipo no conduce directamente a la identificación de estos. Una alternativa para trabajar con esta problemática, es la selección de genes candidatos involucrando una selección *a priori* de genes para análisis de asociación basados en funciones biológicas representativas documentadas [20].

Los polimorfismos de nucleótido simple no sinónimo ocurren con una menor frecuencia que los sinónimos [21]. Así mismo, se ha estimado que del 20% al 30% de los nsSNPs alteran la función proteica. Debido a lo antes mencionado, los nsSNPs pasan a ser de sumo interés, por su potencial efecto en la función de la proteína codificada y su consecuente influencia en fenotipos de interés [22]. Tomando ventaja de la información arrojada por una de las más recientes tecnologías de genetopificación en ganado bovino, el panel Illumina SNPchip HD que genotipifica mas de 700,000 polimorfismos, las técnicas existentes de predicción de estructuras secundarias de proteínas, la anotación del genoma bobino UMD 3.1, entre otras herramientas:

1.2 OBJETIVO DEL PROYECTO

El presente estudio propone un procedimiento de detección y análisis de polimorfismos de nucleótido simple no sinónimo, habilitando la caracterización *a priori* de información disponible antes de estudios de asociación.

II. MARCO TEÓRICO

2.1 BOVINOS, ORGANISMO DE ANÁLISIS

La evolución humana ha ido de la mano junto con la domesticación del ganado. Al emerger la civilización moderna esta fue acompañada por la adaptación, sistematización y crianza de animales en cautiverio. La domesticación del ganado inicio hace algunos 8,000 a 10,000 años. Desde entonces se han establecido más de 800 razas de ganado, esto representa un importante patrimonio para la humanidad y un recurso que le permite a la ciencia la comprensión de la genética de rasgos complejos. A la fecha, entre vacas, ovejas y cabras, hay aproximadamente 3.4 miles de millones de cabezas de ganado domesticado. Esto provee una fuente mayoritaria de proteínas para 6.6 miles de millones de humanos en el mundo. Cerca de tres cuartos de las tierras de cultivo del mundo producen forraje apto para la crianza de estos rumiantes. Los rumiantes son valorados también por sus pieles, pelo y capacidad de trabajo. Estos animales son fuente de alimento y proporcionan energía como animales de tiro.

Dada entonces la importancia de los bovinos en la sociedad surgen las siguientes necesidades de información y tecnología:

1) Tener poblaciones de animales con mediciones fenotípicas precisas y sus muestras de ADN genómico.

2) Marcadores de polimorfismos de ADN posicionados en todo el genoma.

3) Tecnología de laboratorio para realizar análisis de una alta densidad de marcadores.

4) Algoritmos y métodos para el análisis de datos genotípicos y fenotípicos.

2.2 SECUENCIACIÓN DEL GENOMA Y MAPA DE HAPLOTIPOS (HAPMAP) EN LOS BOVINOS

A través de centros de investigación y universidades, varios países unieron esfuerzos y el 2004 generaron dos consorcios con el fin de secuenciar el genoma y crear un mapa de haplotipos (conocido como HapMap) del ganado bovino.

- The Bovine Genome Sequencing and Analysis Consortium.
- The Bovine HapMap Consortium.

Las metas comunes de ambos consorcios fueron: Contribuir a la salud humana. Ofrecer mejor conexión entre el ADN de los humanos y otras especies. Expandir el entendimiento de los procesos evolutivos. Mejorar la salud animal con la aplicación de métodos basados en genoma. Detección de genes que afectan las características de rendimiento productivo (ejemplo: calidad y cantidad de leche producida). Estudiar la diversidad genética del organismo. Selección y predicción genómica.

En abril del 2009, inició una secuencia de publicaciones sobre los resultados de estos trabajos. Se publicó la secuencia del genoma bovino junto con más de veinte trabajos de investigación sobre la caracterización de la estructura genética de la especie. El artículo publicado por "The Bovine Genome Sequencing and Analysis Consortium* en la revista Science [2], dicta que *"En la anotación Btau 4.0, se colocó el 95% del total de la secuencia genómica en los 30 cromosomas. El tamaño estimado del genoma es de 2.87 Gpb. En consenso, el conjunto de genes codificadores de proteínas del Btau 3.1, consiste en 26,835 genes con una validación del 82%. Con esta base, se estima que el genoma bovino contiene al menos 22,000 genes codificadores de proteínas. El genoma bovino y sus recursos asociados facilitará la identificación de novedosas funciones y sistemas regulatorios de gran importancia en mamíferos, y podrá proveer herramientas para mejora genética en las industrias lechera y de carne."* Así mismo, "The Bovine HapMap Consortium" publica en la misma revista [5], *"Un ensamblaje de secuencias de lecturas shotgun de una vaca Hereford y secuencias comparativas muestreadas a partir de seis razas adicionales fueron usadas para desarrollar sondas de búsqueda de 37,470 SNPs en 497 individuos provenientes de 19 razas de diversas procedencias geográficas y biológicas. 14 B.taurus (n = 376), 3 B. indicus (n = 73), dos razas híbridas (n = 48) dos individuos respectivamente de las razas Bubalus quarlesi & Bubalus bubalis (divergente de Bos taurus). La proporción de SNPs en regiones intergénicas, intrónicas y exónicas fue de 63.74, 34.9, y 1.35%, respectivamente, similar a su representación en el genoma. Algunos estudios realizados fueron: Desequilibrio ligado (asociación no aleatoria de dos alelos en dos o más loci; Tamaño de población efectiva (Ne); Tasa de mutación de población (θ); Heterozigocidad de pares de nucleótidos (π); Métodos para detectar selección genómica en el ganado"*

Figura 1 - Science, vol. 324, pp. 522-528, April 24, 2009.

2.3 ADN, PROTEÍNAS Y POLIMORFISMOS DE NUCLEÓTIDO SIMPLE NO SINÓNIMO

La unidad fundamental de todo ser vivo es la célula. Dentro del núcleo de la célula se encuentra el ácido desoxirribonucleico, conocido como ADN. Esta molécula contiene toda la información que nos define en el aspecto biológico como especie y a su vez como individuos, posee también la información necesaria para generar proteínas que harán diversas funciones en el organismo. El ADN se compone de dos cadenas de nucleótidos enlazadas en una estructura de doble hélice. Los nucleótidos están formados por las bases nitrogenadas adenina, timina, guanina o citocina (A, T, C, G), una pentosa llamada desoxirribosa y un grupo fosfato. Ambas cadenas de la doble hélice se mantienen unidas gracias a la formación de puentes de hidrógeno entre las bases AT o CG. La cadena de ADN (Fig. 2) está compuesta por aproximadamente tres miles de millones de pares de bases (pb). La información en el ADN está dividida

en unidades discretas llamadas genes que contienen de 50000 a 100000 nucleótidos. Los humanos tenemos de 20000 a 25000 genes. A su vez la cadena de ADN es dividida en 23 pares de paquetes discretos llamados cromosomas, de los cuales uno de sus elementos proviene del padre y el otro de la madre.

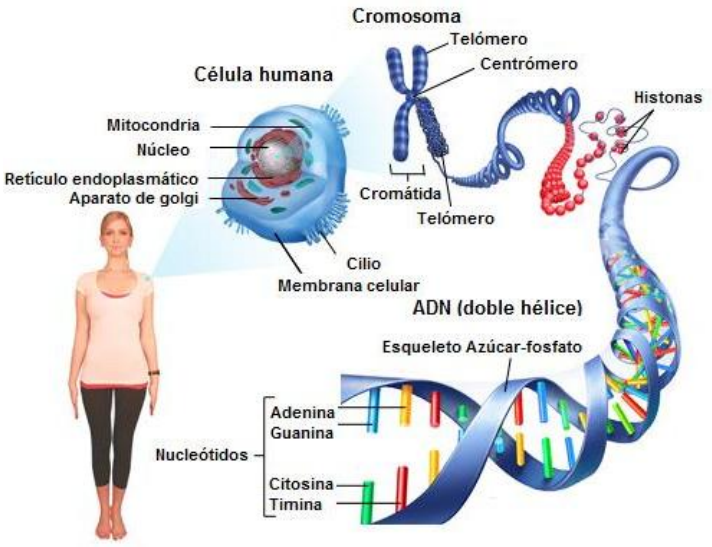


Figura 2 - La cadena de ADN.

Fuente: Portal Virtual Medical Centre, Australia [23].

No obstante la secuenciación de diversos genomas, se desconoce todavía mucha información sobre el ADN y sus funciones. Por el momento se conoce la función de solo el 5% del genoma humano, del cual el 1.5% pertenece a la función más estudiada que es la codificación de proteínas a partir de los genes codificadores. El 3% a 3.5% restante pertenece a regiones reguladoras de genes, entre otras funciones.

Las proteínas son las moléculas de trabajo de la célula y realizan actividades programadas que son codificadas por los genes. Estas son monómeros lineales que contienen de diez a varios miles de aminoácidos unidos por enlaces polipéptidos. Según su función las proteínas son (Fig. 3):

Tabla 1 - Clasificación de proteínas según su función.

Tipo de proteína	Función
Estructurales	Proveen rigidez estructural a la célula.
De transporte	Controlan el flujo de materiales a través de las membranas celulares.
Reguladoras	Actúan como sensores e interruptores para controlar las actividades de las proteínas la función de los genes.
De señalamiento	Transmiten señales externas al interior o exterior de la célula.
Motoras	Provocan Movimiento.
Catalizadoras	Aceleran una reacción química.

Fuente: Loadish, Molecular Cell Biology [24].

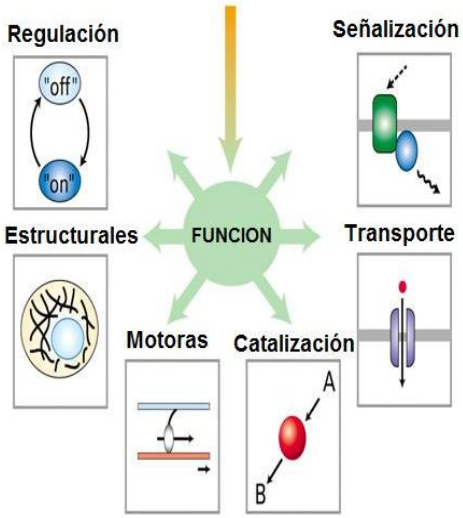


Figura 3 - Las funciones de las proteínas.

Fuente: Loadish, Molecular Cell Biology [24].

Los bloques de construcción de las proteínas son los aminoácidos. Hay 20 distintos aminoácidos que varían en forma, tamaño, capacidad de diversos tipos de enlaces, hidrosolubilidad o hidrofobisidad y composición química. Las proteínas de cualquier tipo de células eucariotas (entre otros organismos) están conformadas del

mismo conjunto de 20 aminoácidos. Los 20 aminoácidos conocidos son los siguientes:

Tabla 2 - Aminoácidos

Arginina y lisina	Carga positiva.
Ácido aspártico y Ácido glutámico	Carga negativa.
Histidina	Positiva o sin carga.
Asparagina y Glutamina	Sin carga pero con cadenas laterales polares que contienen capacidades de enlaces de hidrogeno.
Serina y Treonina	Similar a la anterior, pero con grupos de hidroxilos.
Alanina, Valina, Leucina, Isoleucina y Metionina	Consisten en su totalidad de hidrocarbonos, excepto por el átomo de sulfuro en la metionina y son no polares.
Fenilalanina, Tirosina y Triptófano	Son cadenas laterales aromáticas "de bulto".
Cisteína	Puede oxidar para formar un enlace covalente disulfuro
Glicina	El aminoácido más pequeño, tiene un átomo de hidrogeno en su grupo R.
Prolina	Se "dobla" para formar un anillo covalente a un átomo de nitrógeno (grupo amino) atado al C∞.

Fuente: Lehninger Principles of Biochemistry 4th edition [25].

Una proteína funciona adecuadamente únicamente cuando su estructura final es la correcta. Las proteínas tienen cinco niveles estructurales. Menciono aquí los primeros dos (Fig. 4). El nivel primario, que es la simple cadena de aminoácidos. El nivel secundario, consiste en varios arreglos "espaciales" resultantes de los pliegues (dobleces) de segmentos de la cadena polipéptida. Hay dos estructuras secundarias principales, las hélices alfa y las láminas beta.

Las proteínas son sintetizadas a partir del ADN, el cual se "transcribe" hacia ARN pre-mensajero, se depura para dejar únicamente regiones codificantes (exones) en la cadena dando pie al ARN mensajero, para que esta cadena posteriormente se "traduzca" en secuencias de tres en tres nucleótidos (llamados codones), a la cadena final de aminoácidos que da origen a las proteínas.

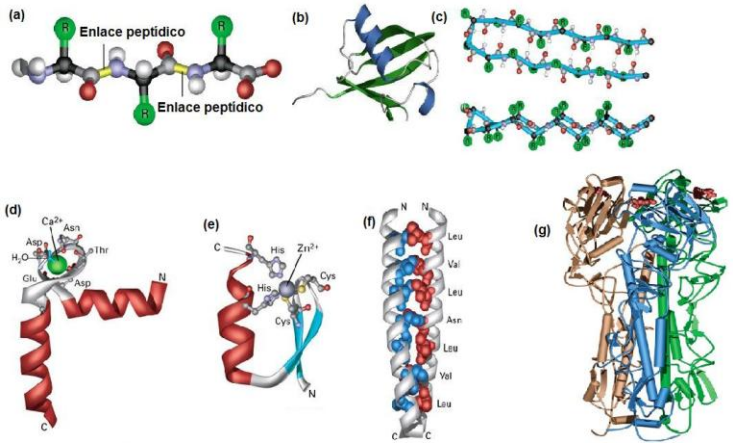


Figura 4 - Niveles estructurales de las proteínas.

(a) Estructura primaria, (b) estructura secundaria, lamina beta, (c) estructura secundaria, hélice alfa, (d) estructura terciaria, hélice-bucle-hélice, (e) estructura terciaria, dedo de zinc, (f) estructura terciaria, hélice superenrollada, (g) estructura terciaria, dominios funcionales. Fuente: Loadish, Molecular Cell Biology [24].

El conjunto de procedimientos por los cuales se codifica este intercambio de nucleótidos a aminoácidos se le denomina "código genético". El codón de inicio indica el punto inicial a partir del cual la secuencia de ARN mensajero se debe empezar a traducir en aminoácidos. El codón de inicio en los organismos con células eucariotas es el codón AUG. El codón de paro no se traduce en un aminoácido, sólo indica la detención de la síntesis de la cadena de ARN mensajero. Existen tres codones de paro: AUG, UAG y UAA. La traducción de codones de bases a aminoácidos se encuentra descrita en la figura 5.

Las diferencias genéticas que existen entre individuos de la misma especie son llamadas polimorfismos (ver figura 6), estas variaciones ocurren en la secuencia de un lugar determinado del ADN entre los individuos de una población. La mayoría de estas variaciones se presenta en forma de inserciones o borrados de tamaño pequeño, o en forma de polimorfismos de nucleótido simple. Un polimorfismo de nucleótido simple (*SNP* por su acrónimo en inglés) se refiere a cualquier posición del genoma en la cual se segregan dos o más distintos nucleótidos en una población. Uno de los

enfoques actuales en la investigación es el uso de polimorfismo de nucleótido simple para identificar variaciones responsables a nivel genómico de diversas variantes fenotípicas, incluyendo la susceptibilidad a enfermedades.

Segunda Letra

Primera letra	U	C	A	G	
U	UUU, UUC: Fenilalanina UUA, UUG: Leucina	UCU, UCC, UCA, UCG: Serina	UAU, UAC: Tirosina UAA, UAG: Código de parada (stop codon)	UGU, UGC: Cisteína UGA: Codigo de parada (**) UGG: Triptófano	U C A G
C	CUU, CUC, CUA, CUG: Leucina	CCU, CCC, CCA, CCG: Prolina	CAU, CAC: Histidina CAA, CAG: Glutamina	CGU, CGC, CGA, CGG: Arginina	U C A G
A	AUU, AUC, AUA: Isoleucina AUG: Metionina (Iniciación)	ACU, ACC, ACA, ACG: Treonina	AAU, AAC: Asparagina AAA, AAG: Lisina	AGU, AGC: Serina AGA, AGG: Arginina	U C A G
G	GUU, GUC, GUA, GUG: Valina	GCU, GCC, GCA, GCG: Alanina	GAU, GAC: Acido Aspartico GAA, GAG: Acido Glutámico	GGU, GGC, GGA, GGG: Glicina	U C A G

Figura 5 - El código genético.

Fuente: Portal biología.edu.ar [26].

Los *SNP* son los polimorfismos más recurrentes en el genoma y ocurren en promedio un *SNP* cada kilobase entre los dos cromosomas de un individuo (en el caso de los humanos).

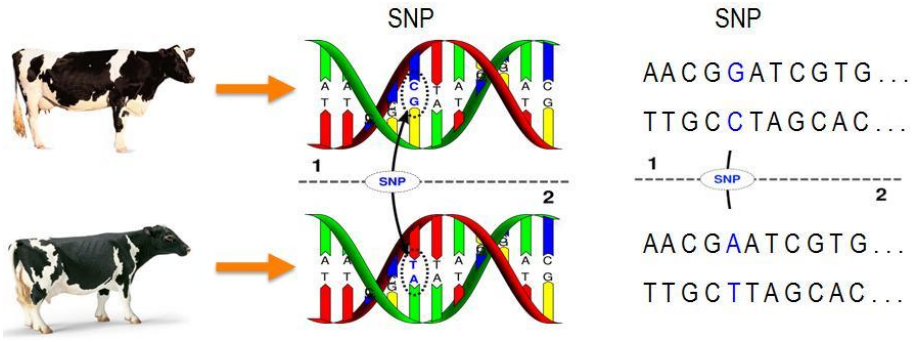


Figura 6 - Polimorfismo de Nucleótido Simple, SNP

Un polimorfismo de nucleótido simple, SNP, es una posición especifica en el ADN, que contiene un valor de alelo distinto, en al menos 5% de la población.

Fuente: Portal AbacusBio LTD [27].

Una consecuencia de la existencia de un SNP en la cadena de ADN, es que puede o no producir un cambio en la traducción de las secuencia de ARN mensajero a la cadena de aminoácidos.

Se les llama polimorfismos sinónimos, silenciosos o neutrales a los polimorfismos que debido al fenómeno de "degeneración del código genético" (varios codones codifican un mismo aminoácido) no modifican a la cadena de aminoácidos resultante en el proceso de traducción. En contraparte, los polimorfismos no sinónimos causan una modificación en la cadena de aminoácidos traducida por el ARN mensajero. El efecto potencial de esta modificación de cadena es el cambio en la secuencia de aminoácidos de una proteína, afectando así este cambio su estructura, y debido a que la forma tridimensional de una proteína está dictada por la secuencia de aminoácidos, su función también es afectada. Un ejemplo del tipo de efectos de estas variaciones es la enfermedad "anemia falciforme", donde la sustitución de A por T en el codón 6-Glu de la hebra no transcrita (GAG), o de T por A en la hebra transcrita, conduce a un codón GUG en el ARN mensajero, que codifica Val en lugar de Glu. Esta única mutación (hemoglobina S, $\alpha_2\beta^s_2$) es la causa de dicha enfermedad.

2.4 TECNOLOGÍAS DE GENOTIPIFICACIÓN DE POLIMORFISMOS DE NUCLEÓTIDO SIMPLE

El descubrimiento y caracterización de miles de marcadores SNPs en el genoma bovino y otros genomas de especies pecuarias trajo como consecuencia una reducción en los costos de genotipado. La casa comercial Illumina en Estados Unidos, puso a la venta a mediados del 2009 el SNPchip50 para capturar un conjunto de 55,609 polimorfismos a lo largo de los 30 cromosomas de los bovinos. Los estudios de validación de este chip muestrearon 565 individuos de 19 razas distintas, ubicadas alrededor de mundo. El SNPchip50, en solo unos meses se convirtió en el estándar para realizar estudios genéticos de las razas de ganado *B.taurus[6]* . El 2010 Illumina ofreció el SNPchip HD. Muestrea más de 777,000 SNPs a lo largo de todo el genoma, con intervalos de 3.47Kb en promedio. La validación se hizo en 639 individuos en las razas: 20 *B. taurus*, 3 *B. indicus* y 4 razas híbridas [28].

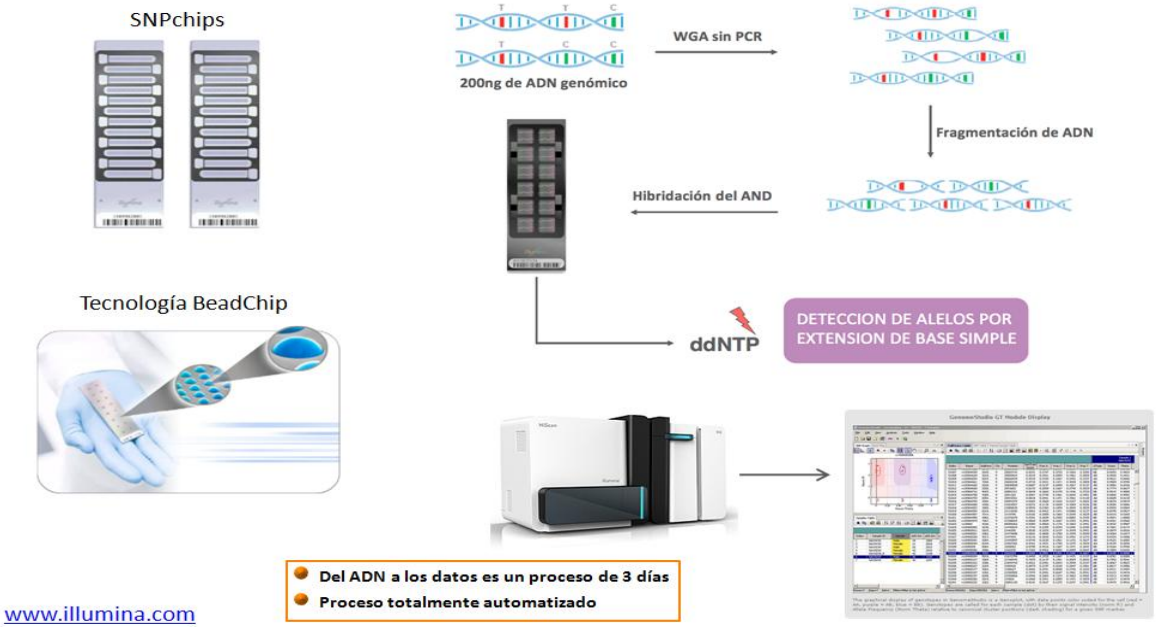


Figura 7 - Proceso automático de detección de múltiples SNPs

Proceso automático de detección de múltiples SNPs a través de la tecnología BeadChip ofrecida por la casa comercial Illumina. Fuente: Portal Illumina.com [29] .

2.5 HERRAMIENTAS ADICIONALES

2.5.1 LENGUAJE DE PROGRAMACIÓN PERL

Perl es el lenguaje computacional más usado en análisis de información biológica, es de libre acceso, multiplataforma, permite conexión directa a bases de datos y es extensible (hay más de 500 módulos de aplicación). Este lenguaje es muy adecuado para indagar información de búsqueda de patrones de cadenas y es muy potente en búsqueda de expresiones regulares. El módulo de aplicación BioPerl es una de las herramientas más potentes de este lenguaje en el ámbito biológico y bioinformático [30]. Este módulo de libre acceso, es capaz de ejecutar análisis y procesar resultados de operaciones tan complejas como búsquedas especializadas de secuencias y proteínas entre otras; ejecutar traducciones, BLAST, ClustalW; acceso a bases de datos especificas tales como GenBank, SwissProt, Kegg, NCBI, etc.

2.5.2 PREDICCIÓN DE LAS ESTRUCTURAS SECUNDARIAS DE PROTEÍNAS.

Jpred 3 es una herramienta de predicción de estructuras secundarias de proteínas. El servidor Jpred corre bajo el algoritmo de predicción Jnet el cual combina múltiples redes neuronales [31]. Este algoritmo provee predicciones de tres estados, ya sea: hélice α, lamina β, o bobina; presenta una precisión de 81.5%. Utiliza secuencias de aminoácidos como información de entrada, estas pueden ser sometidas de manera única o múltiple, y los resultados de predicción son obtenidos posteriormente vía correo electrónico o vía web en forma de texto, HTML, PostScript, PDF o editor JalView [32]. Cuando se hace predicción de una sola cadena de aminoácidos, Jpred 3 utiliza PSI-BLAST con el marco de referencia UniRef30. Después por redundancia, se filtra el alineamiento utilizando un modelo escondido de Markov (HMM). Los perfiles de una matriz de puntajes de posición específica

(PSSM) y el perfil del HMM son usados como entrada. Cuando se procesan múltiples secuencias solo se procesa el perfil del HMM.

2.5.3 BASES DE DATOS

KEGG (Kyoto Encyclopedia of Genes and Genomes) es una base de datos que alberga información sobre las funciones y utilidades de alto nivel de sistemas biológicos, como la célula, organismos y ecosistemas, a un nivel molecular, dados las grandes escalas de información generadas por los últimas tecnologías de salida masiva. Esta base de datos nos permite acceso en un mismo portal a información sobre, genes, genomas, ortologías, rutas biológicas (pathways), fármacos y compuestos relacionados con dichas rutas, entre otros. Así mismo, proporciona su propio módulo de programación (API), herramienta Perl que permite obtener información de esta base de datos en una manera más sencilla en el caso de múltiples búsquedas [33].

III. METODOLOGÍA

3.1 DETECCIÓN Y SELECCIÓN DE LOS nsSNPs DE ANÁLISIS.

La detección de nsSNPs se basó en la información provista por el SNP chip BovineHD de Illumina. El archivo de salida del equipo de genotipado masivo, es un archivo en formato estándar CSV. La información provista por el archivo de salida contiene: nombre asignado al polimorfismo (que permite identificar también si es un SNP no antes analizado por el panel SNPchip 50), cromosoma en el que está posicionado, coordenada del polimorfismo (locus), y secuencia fuente (cadena de flanqueo anterior + [valor del alelo "Wild Type"/ valor variante del alelo] + cadena de flanqueo posterior), ver figura 8.

Name	Chr	Coord	SourceSeq
ARS-BFGL-BAC-10172	14	6371334	CTCAGAAGTTGGTCCCCAAAGTATGTGGTAGCACTTACTTATGTAAGTCATCACTCAAGT[A/G]ATCCAGAATATTCTTTTAGTAATATTTTTGTTAATATTGAAATTTTTAAAACAATTGAAA

Figura 8 - Renglón de información del archivo de salida de la máquina de genotipado masivo. Fuente: USDA.

Ya que este archivo contiene 783,163 renglones (uno por SNP), por motivos de simplificación y organización de la información, se separó toda la información en 30 distintos archivos (uno por cada cromosoma).

Se inspeccionaron todos los alelos (valores basales de los SNPs) y sus cadenas de flanqueo (oligonucleótidos) anterior y posterior, leyéndolos en sus seis posibles alternantes, positivo en 1^{ra}-2^{da}-3^{ra} y negativo o reverso en 1^{ra}-2^{da}-3^{ra} base de inicio en el codón, en sus posiciones (locus) específicas. Este primer paso se realizó tomando cada uno de los polimorfismos y analizando las secuencias de nucleótidos antes mencionadas (cadena de flanqueo anterior + alelo + cadena de flanqueo posterior) con valor de base en su forma común ("wild type") y variante, en cada una de las seis alternantes. Este procedimiento genera doce secuencias a ser analizadas para cada SNP, 6 alternantes Wild Type, y sus respectivas variantes. Posteriormente, cada alternante es traducida de cadena de nucleótidos a cadena de aminoácido, se define al SNP como no sinónimo cuando se presentan diferencias entre las secuencias de aminoácidos traducidos de los valores Wild Type y variantes (ver figura 9).

PI, Común	CTCAGAAGTTGGTCCCCAAAGTATGTGGTAGCACTTACTTATGTAAGTCATCACTCAAGT AATCCAGAATATTCTTTTAGTAATATTTTTGTTAATATTGAAATTTTTAAAACAATTGAAA
Traducción	LRSWSPKYVVALTYVSHHSSNPEYSFSNIFVNIEIFKTIE
PI, Variante	CTCAGAAGTTGGTCCCCAAAGTATGTGGTAGCACTTACTTATGTAAGTCATCACTCAAGT GATCCAGAATATTCTTTTAGTAATATTTTTGTTAATATTGAAATTTTTAAAACAATTGAAA
Traducción	LRSWSPKYVVALTYVSHHSSDPEYSFSNIFVNIEIFKTIE

Figura 9 - Detección de nsSNPs. Cadenas alternantes y traducción de secuencias.

Una vez detectados los SNPs no sinónimos se procede a ubicarlos dentro de la anotación del genoma UMD 3.1. La anotación del genoma completo se encuentra contenida en un formato llamado gff, que es un registro de extensiones que puede ser usado para identificar subcadenas o secuencias biológicas [34]. Los campos utilizados de este archivo describen: Nombre de la secuencia, característica (gen, exón, CDS, codón de inicio, codón de paro), inicio y fin (figura 10).

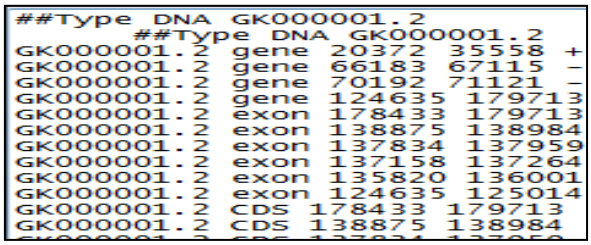


Figura 10 - Información de la anotación UMD 3.1 extraída del formato gff.

Fuente: The Bovine Genome Sequencing and Analysis Consortium.

Una vez ubicados los nsSNPs dentro de la anotación, se detectó sobre qué campo característico recae el SNP: gen, exón, CDS, codón de inicio o codón de paro. Una vez ubicados los SNPs dentro de estos campos, el criterio de selección de los nsSNPs previamente detectados fue el de elegir todos los valores que recaen en “Gen” y”otra clasificación” simultáneamente.

La información de entrada/salida y procesos de esta primera fase son descritos en las figuras 11, 12 y 13.

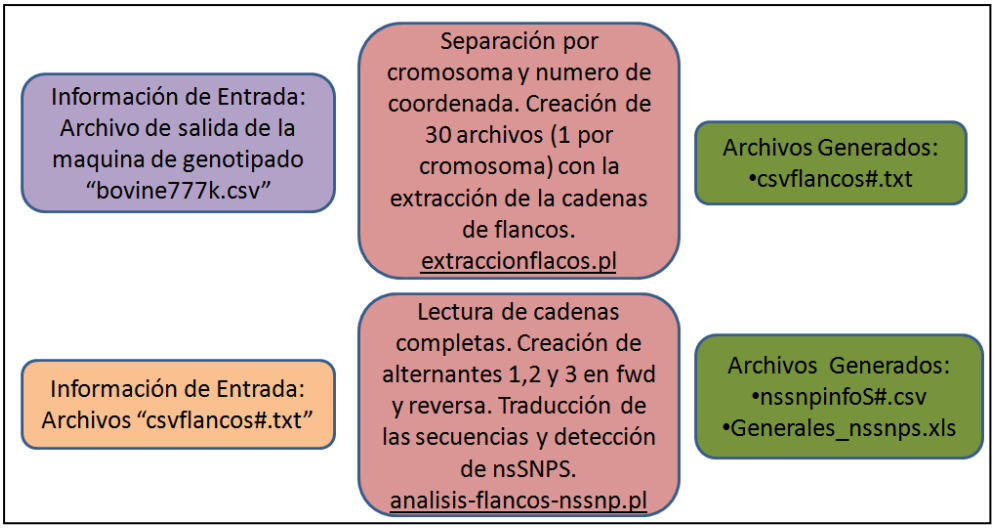


Figura 11 - PASO 1

Formación de secuencias de análisis a partir de los "flanking sequences", valores común (wild type) y variante de todas los SNPs. Genera todos los posibles alternantes Positivo o Fwd 1, 2, 3 y Reversa 1, 2, 3. Detección de nsSNPs VI.3.2

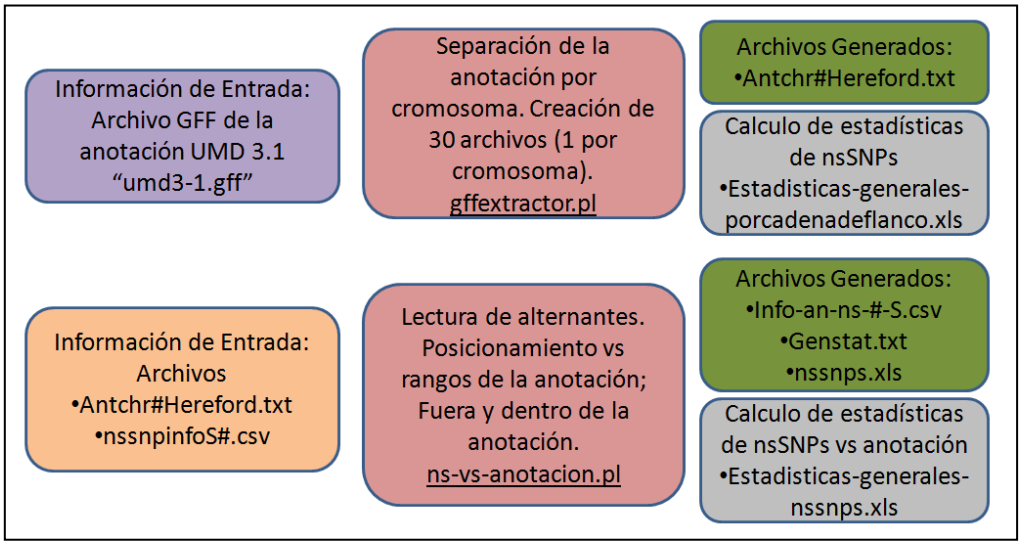


Figura 12 - PASO 2

Posicionamiento de todos los nsSNPs dentro del genoma UMD 3.1

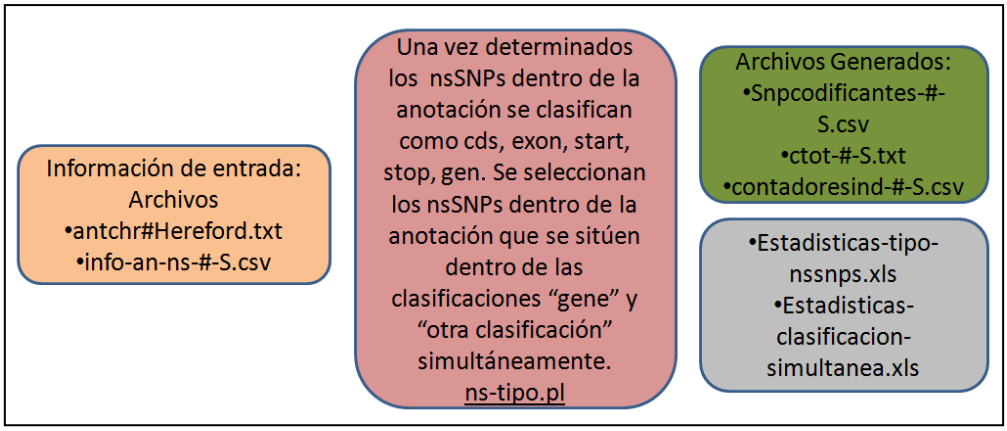


Figura 13 - PASO 3

Selección de nsSNPs de estudio. Todos los valores que recaen en "Gen" y "otra clasificación" simultáneamente.

3.2 DETECCIÓN DE LOS nsSNPs QUE AFECTAN ESTRUCTURA SECUNDARIA.

Una vez seleccionados los nsSNPs de análisis, el siguiente paso del procedimiento propuesto es detectar cuáles de estos polimorfismos causan un cambio en la estructura secundaria que generan las secuencias de aminoácidos "wild type" con respecto a sus mismas secuencias en forma variante. Dadas las ventajas que ofrece el sistema JPRED 3, como lo son, sometimiento múltiple de cadenas, relativa alta velocidad de procesamiento y alta precisión de predicción (81.5%), se determinó utilizar este sistema para la tarea de predicción de estructuras secundarias de las secuencias de aminoácidos. El ingreso de información se realizó por medio del portal *www.compbio.dundee.ac.uk/www-jpred/advanced.html*, perteneciente a la universidad Dundee del Reino Unido (Figura 14).

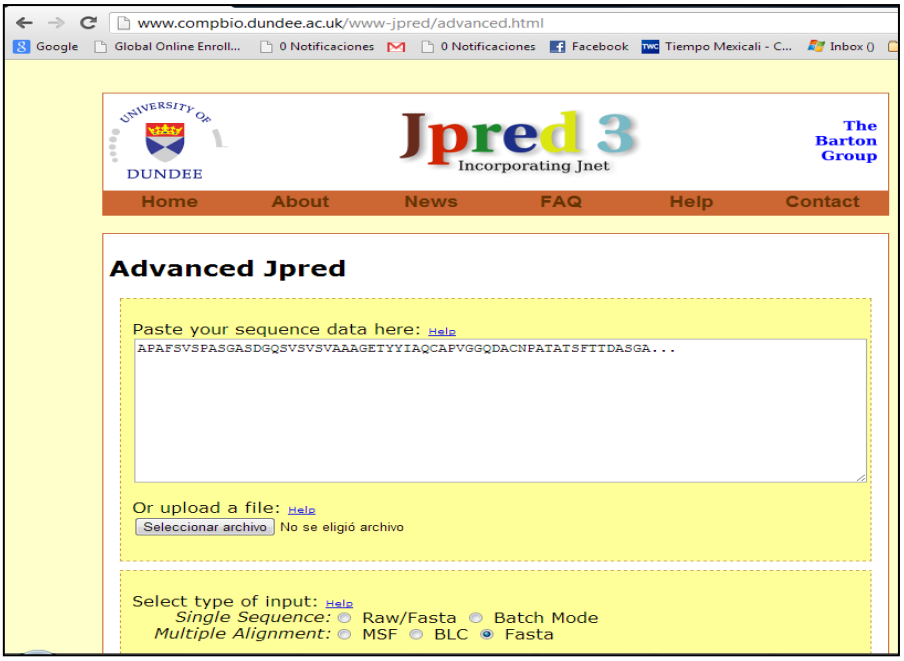


Figura 14 - Jpred 3

Fuente: Portal Jpred 3 de la Universidad de Dundee, Reino Unido [32].

La forma de ingreso de secuencias se realiza transfiriendo archivos FASTA de no más de 200 secuencias por archivo. Posteriormente, los resultados de predicción se obtuvieron vía correo electrónico en forma de texto y fueron procesados para detectar los cambios de estructuras secundarias.

La información de entrada/salida y procesos de esta segunda fase son descritos en las figuras 15 y 16.

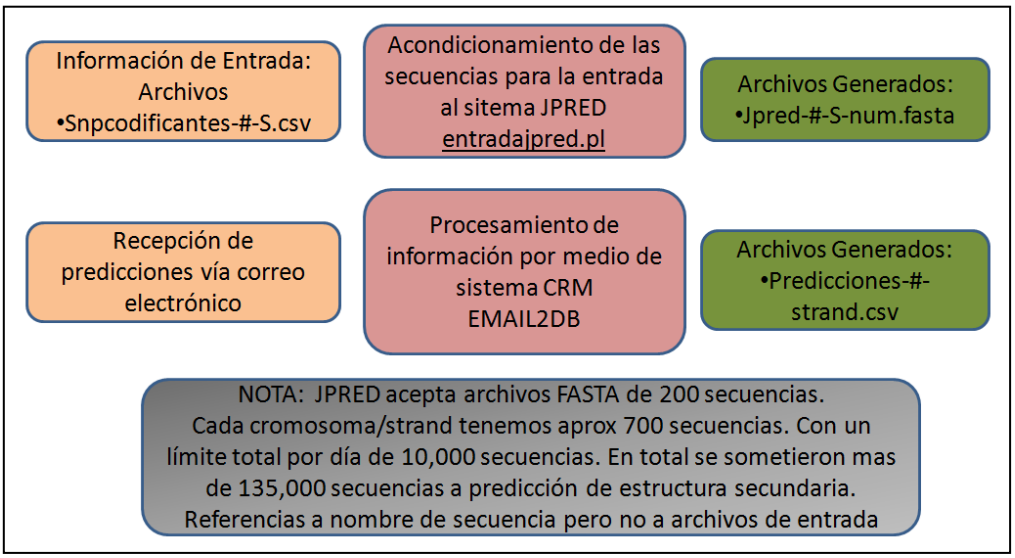


Figura 15 - PASO 4

Someter a predicción de estructuras secundarias todas las secuencias de análisis seleccionadas previamente en su forma común y variante, para todos los alternantes.

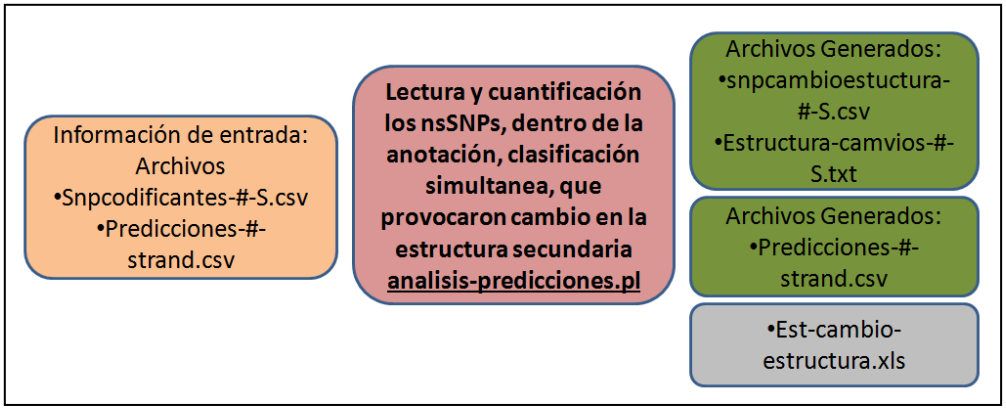


Figura 16 - PASO 5

Detección de cambio de estructuras secundarias a partir de las predicciones generadas.

3.3 ASOCIACIÓN CON RUTA BIOLÓGICA Y/O FÁRMACOS Y COMPUESTOS RELACIONADOS CON LA RUTA.

Concluida la fase de selección de los polimorfismos de análisis, la siguiente etapa del estudio permite asociar los polimorfismos con fenotipos. Esta asociación se hace automáticamente a través de programación Perl y programación Perl con interface KEGG [33]. Vía programación Perl, se obtienen los Números de Identificación NCBI de los genes (GeneID) a donde corresponden los polimorfismos previamente seleccionados. Obtenemos esta información a partir de la coordenada del polimorfismo y ubicándola en la anotación UMD 3.1. Por simplificación de proceso, una vez obtenidos estos identificadores de gen, se genera la lista específica de genes a ser buscados en la base de datos KEEG para obtención de fenotipo, ya que varios polimorfismos se encuentran ubicados dentro de un mismo gen (ver figuras 17 y 18).

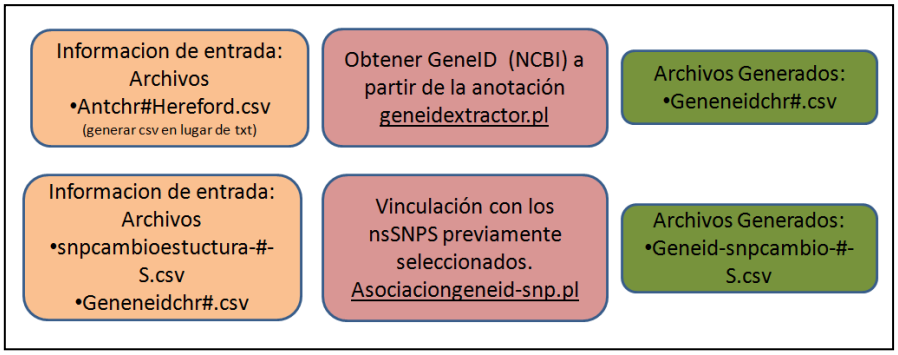


Figura 17 - PASO 6.1

Asociación de proteínas con los SNPs vía “Kegg” y“GO”.
(Obtención del identificador del gen y vinculación con los nsSNPS previamente seleccionados)

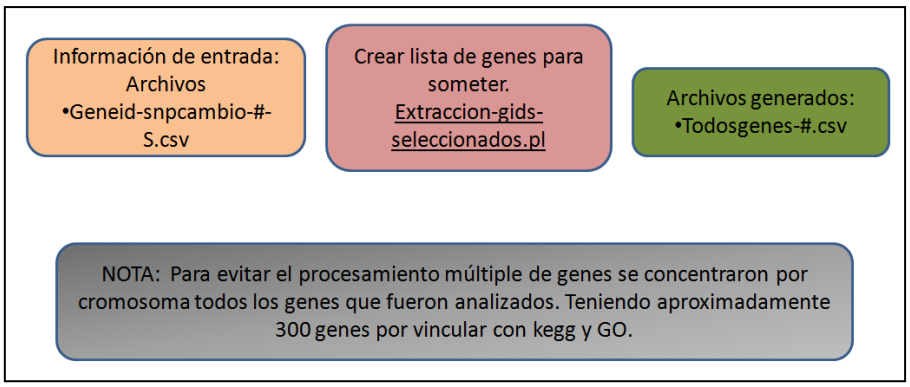


Figura 18 - PASO 6.2

Asociación de proteínas con los SNPs vía "Kegg" y"GO".
(Obtención de listas únicas por cromosoma de los genes que van a ser analizados)

Con las listas de los identificadores de gen "GeneID" generadas, por medio de la interface Perl:KEEG, se procede a la conversión al identificador de gen Kegg. Se realiza también la conversión al identificador de gen Uniprot [35]. Se detectan las rutas biológicas (Pathways) sobre las cuales actúa el gen. Posteriormente se identifican los motifs, fármacos y compuestos relacionados con la ruta. El identificador Uniprot permite que a través del EBI (European Bioinformatics Institute) se realice la búsqueda de ontologías del gen (GO term). Ver Figura 19, 20 y 21.

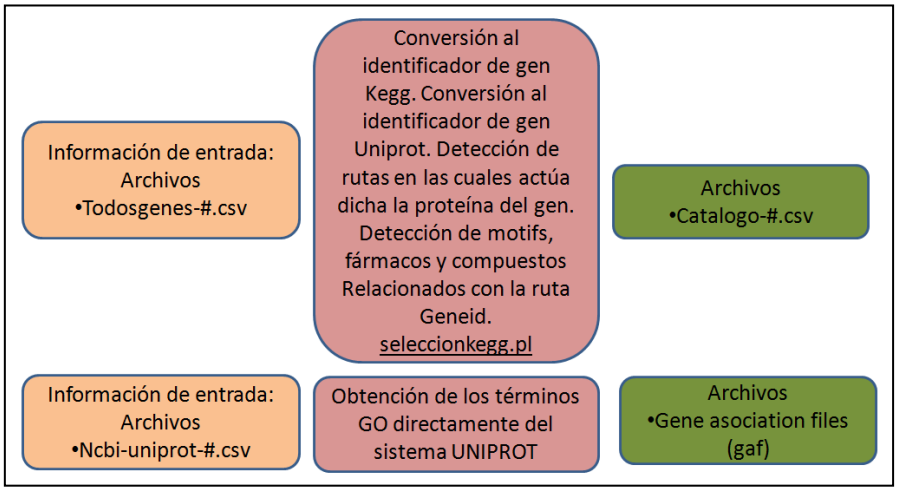


Figura 19 - PASO 6.3.

Asociación de proteínas con los SNPs vía "Kegg" y"GO".

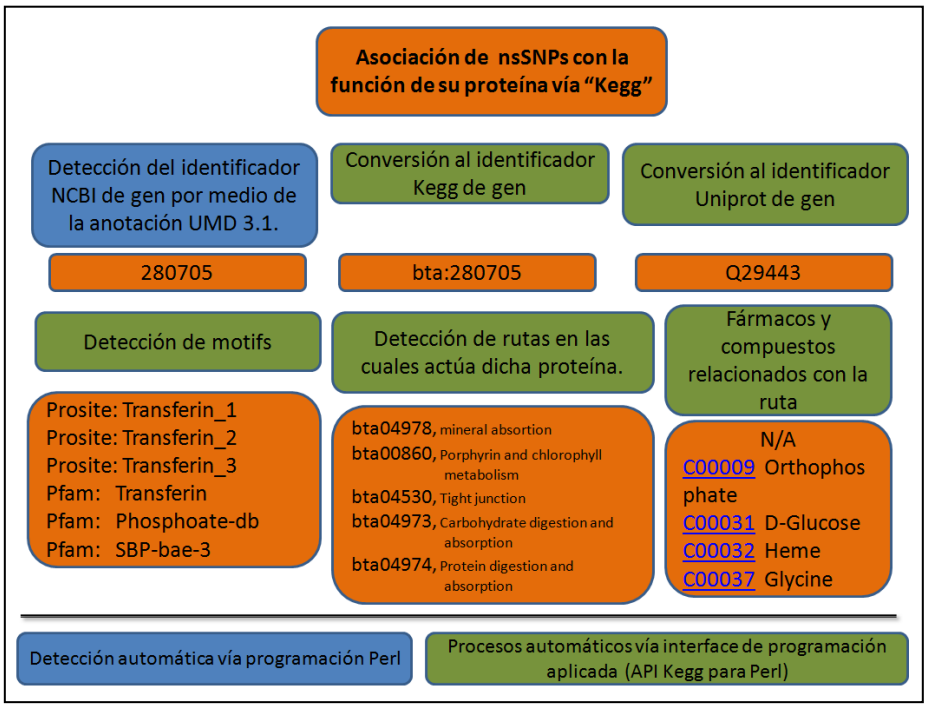


Figura 20 -Asociación Vía KEGG

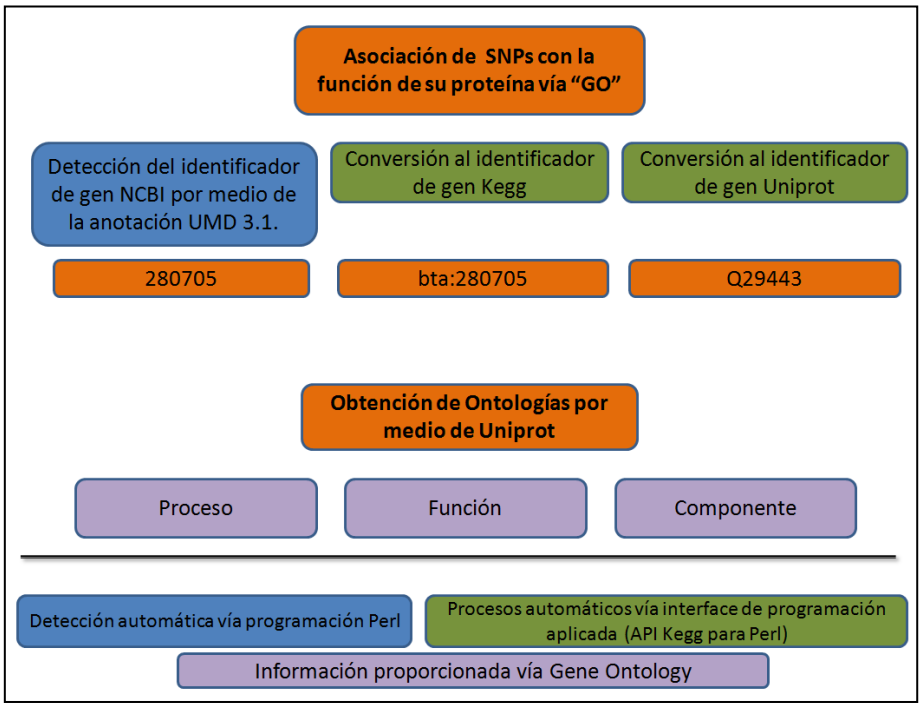


Figura 21 - Asociación con función Vía GO

IV. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Se desarrolló un procedimiento secuencial ("pipeline") que utiliza las 783,163 posiciones de los SNPs muestreadas por el panel Illumina BovineHD SNP chip a través de sus correspondientes secuencias de flanqueo, la anotación del genoma bovino UMD 3.1, y los valores de alelo común ("wild type" o alelo ancestral) y alelo variante derivados de genotipificar más de 150 animales de razas *B. taurus* y *B. indicus* y tres grupos externos [28], para formar seis hebras legibles (lecturas alternantes), realizar la traducción ADN-Aminoácido, y detectar los nsSNPs dentro de los 30 cromosomas del genoma bovino.

4.1 DETECCIÓN DE nsSNPs EN TODOS LOS POSIBLES ALTERNANTES

El SNP chip BovineHD contiene en promedio 26,105 nsSNPs por cromosoma, en total genotipa 783,163 SNPs en todo el genoma[28]. Dado el total de SNPs contenidos en el panel, el primer paso consistió en detectar cuántos de estos 783,163 son no sinónimos en sus distintos seis alternantes positivos o reversa (P1, P2, P3 y R1, R2, R3). Podemos observar en la Tabla 3 que en el genoma completo, más del 90% en las alternantes P1,P3,R1 y R3 resultaron tener traducciones no sinónimas,

mientras que en las alternantes P2 y R2, menos del 20% resultaron no sinónimas. Podemos observar que cuando el polimorfismo se encuentra en la parte central del codón, tiene una probabilidad aproximada a 0.2 ser un polimorfismo no sinónimo. Por otro lado cuando el SNP se encuentra en un extremo, independientemente del sentido de la lectura, tiene una probabilidad mayor al 90% de ser un polimorfismo no sinónimo. Esta variabilidad presenta un comportamiento esperado debido a que en el código genético existen aminoácidos que son codificados por más de un codón distinto, y la mayor variabilidad en dichos codones se encuentra en las posiciones de los extremos del codón.

Tabla 3- Promedio de nsSNPs por lectura a lo largo del todos los cromosomas.

	Alternante positivo			Alternante Reverso		
	1ra posición	2daposición	3ra posición	1ra posición	2daposición	3raposición
nsSNPs (promedio)	24609.63	4054.63	25535.10	24892.47	4334.33	25167.80
%	94.31	15.31	97.85	95.37	16.39	96.49

Las columnas representan la posición del nsSNP en el codón generador del aminoácido. Los renglones representan el promedio de nsSNPs en genoma completo, y el porcentaje de SNPs que fueron clasificados como nsSNPs.

El cromosoma uno resulto contener la más alta cantidad de nsSNPS en todo el genoma, 195457; mientras que el cromosoma 25 contuvo la menor cantidad, 53695 (La Tabla 3 muestra los promedios de nsSNPS encontrados, la Tabla 4 contiene la información de nsSNPs por cromosoma en todas sus alternantes). La razón de tener más polimorfismos en el primer cromosoma es debido a que el cromosoma 1 es en tamaño el más grande de los 29 cromosomas autosomales del organismo de estudio.

Tabla 4 – Cantidad de nsSNPs por lectura en todos los cromosomas.

	Alternantes Positivas			Alternantes Reversas/Negativas			
	1ra posición	2da posición	3ra posición	1ra posición	2da posición	3ra posición	Total
Chr 1	44143	7776	45774	44615	8125	45024	195457
Chr 2	38053	6464	39455	38488	6907	38806	168173
Chr 3	33930	5629	35182	34310	5978	34646	149675
Chr 4	33252	5635	34526	33637	6114	34032	147196
Chr 5	33277	5503	34498	33610	5854	33881	146623
Chr 6	33671	5871	34924	34174	6333	34417	149390
Chr 7	31568	5206	32805	31977	5633	32310	139499
Chr 8	31943	5425	33049	32318	5789	32603	141127
Chr 9	29531	4972	30607	29863	5404	30126	130503
Chr 10	29045	4924	30201	29428	5155	29760	128513
Chr 11	30383	4886	31576	30780	5151	31147	133923
Chr 12	25090	4325	26032	25378	4608	25637	111070
Chr 13	22591	3456	23395	22810	3662	23065	98979
Chr 14	23558	3917	24429	23844	4275	24122	104145
Chr 15	23677	4013	24648	23993	4138	24317	104786
Chr 16	23069	3664	23926	23321	3946	23580	101506
Chr 17	21117	3377	21982	21431	3612	21675	93194
Chr 18	18507	2846	19200	18710	3018	18995	81276
Chr 19	18030	2589	18774	18240	2920	18514	79067
Chr 20	20372	3368	21167	20602	3583	20849	89941
Chr 21	20150	3150	20926	20397	3348	20700	88671
Chr 22	17159	2596	17775	17307	2908	17531	75276
Chr 23	14552	2193	15116	14749	2402	14992	63934
Chr 24	17708	2819	18356	17874	2968	18179	77904
Chr 25	12348	1662	12804	12406	1846	12629	53695
Chr 26	14501	2363	15034	14637	2397	14835	63767

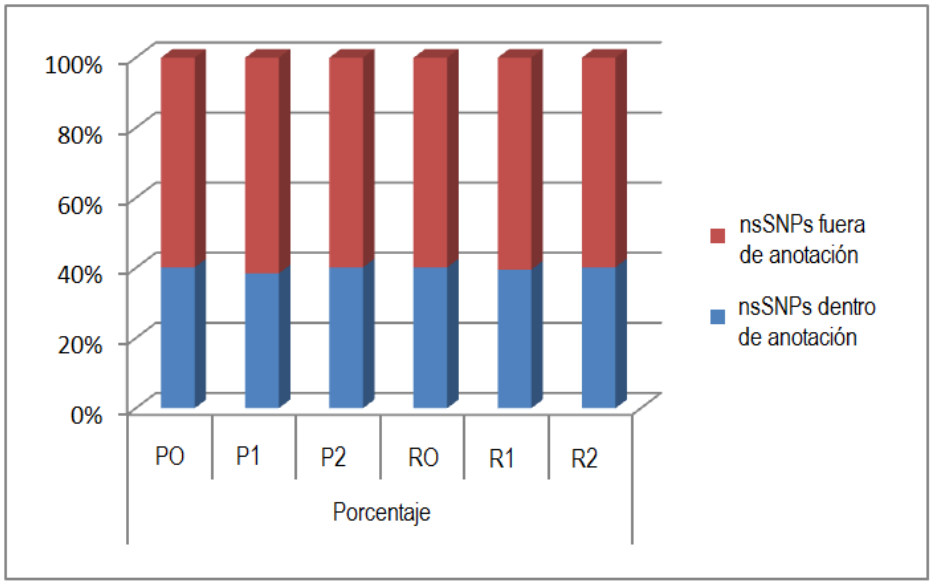
Chr 27	12576	2026	13021	12710	2217	12807	55357
Chr 28	12450	2005	12912	12558	2144	12738	54807
Chr 29	14094	2247	14585	14216	2387	14454	61984
Chr X	37944	6732	39373	38390	7253	38733	168425

Las columnas representan la posición del nsSNP en el codón que genera el aminoácido. Los renglones representan el número de nsSNPs por lectura, por cada cromosoma.

4.2 nsSNPs DENTRO DE LOS LÍMITES DE LA ANOTACIÓN DEL GENOMA BOVINO UMD 3.1

Los nsSNPs detectados, se posicionaron (coordenada o locus) dentro de los límites de la anotación UMD 3.1. El promedio de nsSNPs que recae en una o más categorías dentro de la anotación fue de 43,44%, mientras que el resto no recayó en ninguna categoría (ver Gráfica 1).

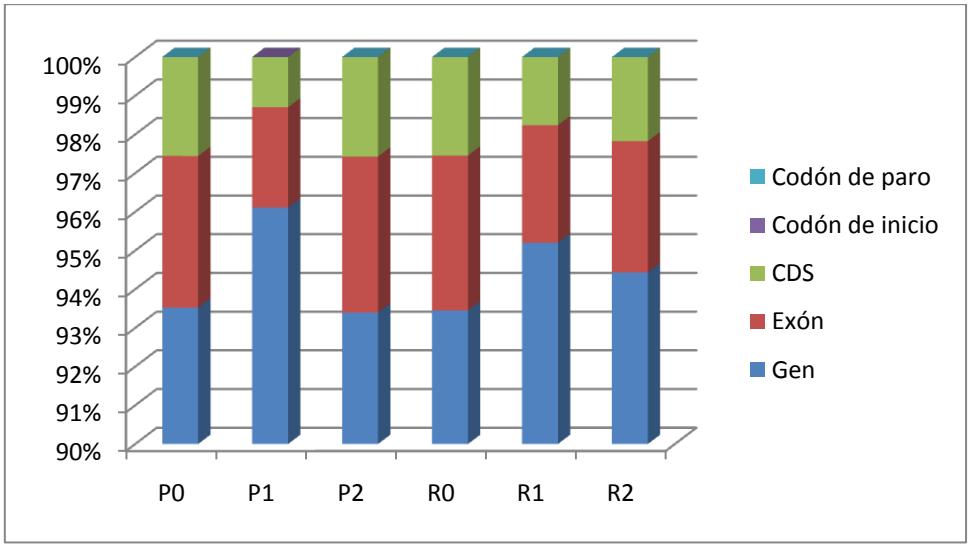
Gráfica 1 - Porcentaje de nsSNPs dentro de la anotación UMD3.1



Porcentaje de nsSNPs a través del genoma completo para todas las alternantes. P1, P2, y P3 corresponden a alternantes positivas; R1, R2, y R3 corresponden a alternantes negativas o reversa.

Se analizo la clasificación de cada nsSNP contenido en la anotación y se calculó la frecuencia de aparición por cada categoría. Resultó que, a través del genoma completo, para todos las 6 alternantes, 94.41% de los nsSNPs recaen en la categoría de Gen (Gene), 3.5% en exón, 2.15 en región codificadora (CDS), 0.0019% en codón de inicio (Start Codon) y 0.001% en codón de paro (Stop Codon) respectivamente (Gráfica 2). Las tablas 3 y 7 muestran la descripción de cantidades para cada cromosoma, la gráfica 2 muestra los porcentajes de nsSNPs que recaen dentro de la anotación UMD 3.1.

Gráfica 2 - Incidencia dentro de la anotación UMD 3.1



Porcentajes de nsSNPs encontrados en el genoma completo por alternante dentro de los rangos de la anotación.

Tabla 5 - Porcentajes por categoría de la anotación UMD 3.1 en todas las alternantes de los nsSNPs (Porcentaje total en genoma completo).

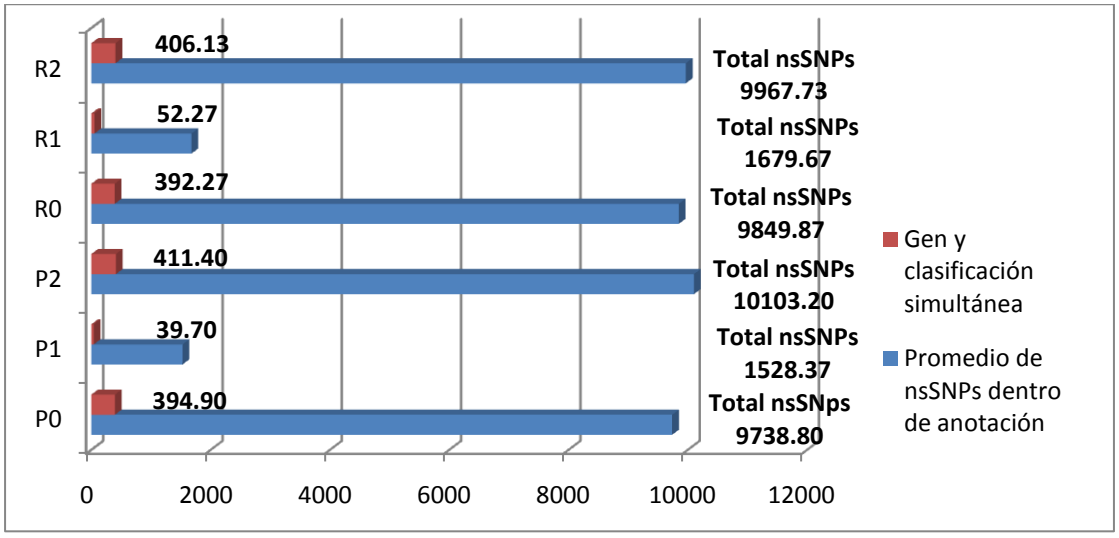
	Alternantes Positivas			Alternantes Reversas/Negativas		
	1ra posición	2da posición	1ra posición	2da posición	1ra posición	2da posición
Gen	7418.23 (93.33%)	1250.23 (95.07%)	7512.93 (93.28%)	6900.23 (93.20%)	1064.73 (95.99%)	7162.97 (93.17%)
Exón	320.63 (4.03%)	40.77 (3.16%)	326.83 (4.17%)	304.43 (4.23%)	29.57 (2.70%)	317.10 (4.24%)
CDS	209.37 (2.63%)	24.10 (1.83%)	214.07 (2.66%)	198.63 (2.68%)	14.93 (1.35%)	207.90 (2.70%)
Codón de inicio	0.133 (0.0017%)	0.000 (0.00%)	0.133 (0.0017%)	0.133 (0.0018%)	0.033 (0.003%)	0.133 (0.0017%)
Codón de paro	0.100 (0.0013%)	0.033 (0.0025%)	0.100 (0.0012%)	0.100 (0.0014%)	0.000 (0.00%)	0.0667 (0.0009%)
Gen + otra	300.53 (4.10%)	40.40 (3.27%)	312.40 (4.21%)	291.50 (4.28%)	28.47 (2.71%)	303.67 (4.29%)

Las columnas representan la posición del nsSNP en el codón que genera el aminoácido. Los renglones representan el porcentaje total en genoma completo que recaen dentro de cada una de las categorías de la anotación (Gen, Exon, CDS, codón de inicio, codón de paro, y Gen más otra categoría).

Los polimorfismos de selección para análisis fueron todos los nsSNPS detectados que recayeron dentro de los límites de la anotación UMD 3.1 en la clasificación "gene" y simultáneamente en alguna de las categorías adicionales. Este criterio de selección nos permitirá que los polimorfismos seleccionados estén localizados dentro de regiones codificantes, y así, que estos tengan efecto en la función de la proteína. La gráfica 3 muestra las cantidades promedio de nsSNPs encontrados en el genoma para todas las alternantes, y las cantidades promedio de nsSNPs que recayeron en categorías simultáneas de la anotación. Las alternantes P0, P2, R0 y R2 tienen aproximadamente 400 SNPs que recaen en clasificaciones

simultáneas, P1 y R1 tienen aproximadamente 45 polimorfismos en clasificación simultánea.

Gráfica 3 - Incidencia simultánea dentro de la anotación UMD 3.1: “gene” y "otra simultanea".



Se muestra los porcentajes de nsSNPs encontrados en el genoma completo por alternante dentro de los rangos de la anotación, y los porcentajes de nsSNPs que recayeron dentro de clasificaciones simultáneas dentro de la anotación.

En promedio, la cantidad máxima de nsSNPs encontrados dentro de los límites de la anotación perteneció a las alternantes P2, teniendo 10103.20 polimorfismos. El mínimo encontrado fue para las alternantes P1, resultando con 1528.37 nsSNPs. El porcentaje para clasificación simultánea varia del 2.60% (P1) al 4.07% (P2/R1). La Tabla 5 muestra los porcentajes por categoría de la anotación UMD 3.1 en todas las alternantes de de los nsSNPs (Porcentaje total en genoma completo). La tabla 3 también expone los porcentajes de polimorfismos que recaen en selección simultánea dentro de la anotación (último renglón de la tabla), esto es al seleccionar polimorfismos que están situados dentro de un gen y en una región codificante (exón, cds, codones de inicio o paro), nos aseguramos que nuestro

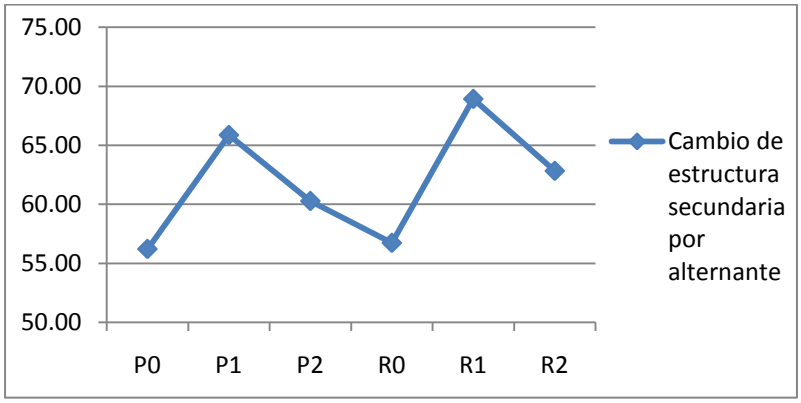
polimorfismo seleccionado previamente será traducido. Este grupo de polimorfismos fue seleccionado para realizar el resto del análisis.

4.3 nsSNPs QUE AFECTAN LA ESTRUCTURA SECUNDARIA EN EL OLIGONUCLEÓTIDO TRADUCIDO

Las cadenas de aminoácidos traducidas por medio de BioPerl, para cada polimorfismo previamente seleccionado en sus formas "wild type" o común y variante, se sometieron las a predicción de estructura secundaria por medio del sistema Jpred 3. Cabe mencionar que más de 135,000 secuencias de aminoácidos fueron sometidas a predicción de estructura secundaria en 675 distintos archivos FASTA, y que, dados los límites de procesamiento del servidor Jpred 3 de 10,000 secuencias o menos por día, el tiempo total de procesamiento en este paso para el estudio en particular fue de 15 días.

Una vez procesados los resultados de predicción, se detectaron las posibles variaciones de estructura por medio de programación en Perl. Posteriormente, se seleccionaron todos los nsSNPs que causaron un cambio a nivel de su estructura secundaria ya que hay una probabilidad más alta de que estos den origen a cambios en la función de la proteína, y eventualmente estén relacionados a un fenotipo específico. En promedio 67% de las alternantes P1 y R1 causaron cambios en sus estructuras secundarias. Las alternantes que causaron los menores porcentajes de cambios en estructura secundaria fueron las P0/R0. La gráfica 4 nos muestra las cantidades porcentuales por alternante de nsSNPs que causaron cambio en las estructuras secundarias. La Tabla 6 muestra los cambios de estructura secundaria por cromosoma/alternante.

Gráfica 4 – Cambios de estructuras secundarias por alternante.



	Promedio	Total G+	Total SC
P0	56.35	394.9	222.5
P1	65.49	39.7	26
P2	60.14	411.4	247.4
R0	57.65	392.3	226.1
R1	69.32	52.27	36.23
R2	62.65	406.1	254.4

Cantidades porcentuales por alternante de nsSNPs que causaron cambio en las estructuras secundarias.

Tabla 6 - Cambios de estructura secundaria por cromosoma/alternante.

	Alternantes Positivas									Alternantes Reversas/Negativas								
	1ra posición			2da posición			3ra posición			1ra posición			2da posición			3ra posición		
	G+	SC	%	G+	SC	%	G+	SC	%	G+	SC	%	G+	SC	%	G+	SC	%
Chr 1	510	300	58.82	46	31	67.39	528	335	63.45	513	296	57.70	65	48	73.85	522	352	67.43
Chr 2	545	336	61.65	56	37	66.07	571	346	60.60	541	325	60.07	76	49	64.47	560	353	63.04
Chr 3	677	380	56.13	70	43	61.43	701	430	61.34	686	384	55.98	89	59	66.29	692	442	63.87
Chr 4	380	214	56.32	31	16	51.61	394	238	60.41	382	243	63.61	48	36	75.00	389	245	62.98
Chr 5	619	354	57.19	95	68	71.58	661	399	60.36	630	363	57.62	78	48	61.54	649	413	63.64
Chr 6	371	220	59.30	39	21	53.85	377	230	61.01	370	202	54.59	59	41	69.49	379	228	60.16
Chr 7	712	391	54.92	70	50	71.43	735	442	60.14	716	387	54.05	89	59	66.29	726	446	61.43
Chr 8	497	268	53.92	52	28	53.85	512	284	55.47	485	268	55.26	69	46	66.67	505	317	62.77
Chr 9	285	165	57.89	30	17	56.67	306	200	65.36	295	175	59.32	35	24	68.57	301	203	67.44
Chr 10	502	308	61.35	55	38	69.09	515	315	61.17	503	287	57.06	60	46	76.67	514	341	66.34
Chr 11	483	284	58.80	45	30	66.67	509	297	58.35	486	279	57.41	68	44	64.71	501	321	64.07
Chr 12	221	126	57.01	19	11	57.89	238	145	60.92	229	136	59.39	28	19	67.86	232	156	67.24
Chr 13	462	269	58.23	49	33	67.35	478	289	60.46	461	259	56.18	78	54	69.23	476	308	64.71
Chr 14	299	161	53.85	29	19	65.52	317	186	58.68	308	151	49.02	44	31	70.45	314	185	58.92
Chr 15	443	245	55.30	55	37	67.27	457	282	61.71	447	262	58.61	65	48	73.85	455	280	61.54
Chr 16	392	218	55.61	41	24	58.54	406	238	58.62	397	228	57.43	46	33	71.74	401	249	62.09
Chr 17	356	183	51.40	19	10	52.63	373	228	61.13	361	217	60.11	30	19	63.33	360	202	56.11
Chr 18	531	293	55.18	61	38	62.30	559	346	61.90	358	314	87.71	77	56	72.73	556	329	59.17
Chr 19	610	341	55.90	66	44	66.67	627	341	54.39	617	345	55.92	87	64	73.56	628	371	59.08
Chr 20	198	108	54.55	11	9	81.82	213	130	61.03	209	108	51.67	20	11	55.00	203	128	63.05
Chr 21	302	168	55.63	30	19	63.33	318	187	58.81	301	159	52.82	35	29	82.86	310	194	62.58
Chr 22	314	159	50.64	27	18	66.67	324	184	56.79	316	178	56.33	35	24	68.57	320	179	55.94
Chr 23	389	216	55.53	38	27	71.05	413	259	62.71	402	237	58.96	43	33	76.74	405	254	62.72
Chr 24	204	113	55.39	15	11	73.33	213	124	58.22	210	108	51.43	22	15	68.18	214	143	66.82
Chr 25	326	168	51.53	28	15	53.57	328	179	54.57	312	179	57.37	49	33	67.35	323	189	58.51
Chr 26	189	109	57.67	20	14	70.00	193	130	67.36	188	107	56.91	28	15	53.57	190	124	65.26
Chr 27	139	78	56.12	12	9	75.00	147	95	64.63	143	91	63.63	23	16	69.57	146	99	67.81
Chr 28	182	109	59.89	19	13	68.42	187	110	58.82	182	100	54.95	32	20	62.50	182	116	63.74
Chr 29	246	136	55.28	17	15	88.24	260	146	56.15	251	138	54.98	27	19	70.37	257	155	60.31
Chr X	463	256	55.29	46	35	76.09	482	307	63.69	469	258	55.01	63	48	76.19	474	311	65.61

Cada renglón representa un cromosoma. Cada cromosoma muestra las cantidades de seleccionadas en clasificación “simultanea” (G+), y la cantidad de cambios en estructura secundaria observados (SC).

Tabla 7 – Cantidad de nsSNPs por cromosoma que recaen dentro de las distintas categorías de la anotación UMD3.1

	Alternantes Positivas																		**Alternantes Reversas/Negativas**															
	1ra posición						**2da posición**						**3ra posición**						**1ra posición**						**2da posición**						**3ra posición**					
	G	**E**	**C**	**St**	**Sp**	**G+**	**G**	**E**	**C**	**St**	**Sp**	**G+**	**G**	**E**	**C**	**St**	**Sp**	**G+**	**G**	**E**	**C**	**St**	**Sp**	**G+**	**G**	**E**	**C**	**St**	**Sp**	**G+**	**G**	**E**	**C**	**St**	**Sp**	**G+**
Chr 1	16331	524	351	0	0	510	2743	49	26	0	0	46	16959	543	363	0	0	528	16521	528	348	0	0	513	2892	65	33	0	0	65	16658	537	359	0	0	522
Chr 2	15641	581	403	0	0	545	2557	58	32	0	0	56	16184	607	416	0	0	571	15791	574	396	0	0	541	2790	85	50	0	0	76	75991	596	410	0	0	560
Chr 3	14728	708	419	0	2	677	2385	71	31	0	0	70	15232	732	434	0	2	701	14869	718	424	0	1	686	2602	89	50	0	1	89	15013	724	431	0	2	692
Chr 4	15389	404	279	2	0	380	2521	35	22	1	0	31	15963	418	292	2	0	394	15546	402	279	2	0	382	2811	53	35	0	0	48	15701	413	287	2	0	389
Chr 5	13679	642	407	0	1	619	2167	97	51	0	0	95	14158	684	430	0	1	661	13815	651	407	0	1	630	2360	82	51	0	0	78	13914	672	425	0	1	649
Chr 6	12917	403	253	0	0	371	2082	40	15	0	0	39	13375	410	259	0	0	377	13107	400	252	0	0	370	2330	63	32	0	0	59	13200	409	261	0	0	379
Chr 7	12149	742	500	1	0	712	1850	73	41	0	0	70	12651	763	513	1	0	735	12312	742	500	1	0	716	2084	91	56	0	0	89	12472	755	504	1	0	726
Chr 8	12048	507	332	0	0	497	1907	53	23	0	0	52	12422	521	345	0	0	512	12124	495	329	0	0	485	2013	69	38	0	0	69	12274	515	342	0	0	505
Chr 9	10116	297	193	0	0	285	1594	31	12	0	0	30	10478	319	209	0	0	306	10282	308	201	0	0	295	1812	35	16	0	0	35	10330	313	209	0	0	301
Chr 10	12142	508	354	0	2	502	2024	55	25	0	0	55	12624	523	357	0	1	515	12304	515	342	0	2	503	2134	58	35	0	0	60	12453	522	355	0	2	514
Chr 11	12586	503	325	0	0	483	1909	47	22	0	0	45	13082	530	345	0	0	509	12738	506	330	0	0	486	2128	68	37	0	0	68	12937	522	340	0	0	501
Chr 12	7914	230	149	0	0	221	1326	21	10	0	0	19	8199	248	165	0	0	238	8016	239	159	0	0	229	1454	29	16	0	0	28	8118	242	159	0	0	232
Chr 13	9999	479	283	0	0	462	1507	50	21	0	0	49	10343	494	296	0	0	478	10096	477	280	0	0	461	1633	79	39	0	0	78	10197	493	293	0	0	476
Chr 14	8619	317	218	0	0	299	1381	29	13	0	0	29	8932	338	235	0	0	317	8714	328	225	0	0	308	1551	45	28	0	0	44	8789	335	235	0	0	314
Chr 15	9382	451	325	0	0	443	1514	55	30	0	0	55	9729	465	333	0	0	457	9487	455	324	0	0	447	1610	65	43	0	0	65	9632	462	329	0	0	455
Chr 16	9632	408	275	0	0	392	1528	42	19	0	0	41	10019	422	286	0	0	406	9772	411	280	0	0	397	1645	46	25	0	0	46	9866	417	284	0	0	401
Chr 17	8242	377	248	0	0	356	1207	19	14	0	0	19	8552	394	261	0	0	373	8361	382	257	0	0	361	1335	32	22	0	0	30	8446	381	252	0	0	360
Chr 18	8065	554	365	2	0	531	1181	62	35	1	0	61	8372	582	383	2	0	559	8156	561	369	2	0	358	1292	49	51	0	0	77	8293	577	383	2	0	556
Chr 19	9521	641	395	0	0	610	1314	81	52	0	0	66	9931	658	410	0	0	627	9647	648	401	0	0	617	1526	102	66	0	0	87	9821	659	407	0	0	628
Chr 20	6877	202	138	0	0	198	1092	11	4	1	0	11	7156	219	151	0	0	213	6967	215	147	0	0	209	1245	20	13	0	0	20	7073	209	143	0	0	203
Chr 21	7624	322	204	1	0	302	1142	31	17	0	0	30	7926	339	216	1	0	318	7711	320	204	1	0	301	1252	36	22	0	0	35	7840	331	212	1	0	310
Chr 22	8649	346	242	0	0	314	1220	27	14	0	0	27	8988	360	252	0	0	324	8725	351	246	0	0	316	1423	38	22	0	0	35	8863	355	249	0	0	320
Chr 23	6331	408	248	0	0	389	919	38	14	0	0	38	6566	432	266	0	0	413	6419	421	255	0	0	402	1032	44	20	0	0	43	6498	424	263	0	0	405
Chr 24	6387	210	137	0	0	204	957	15	7	0	0	15	6622	221	144	0	0	213	6453	218	139	0	0	210	1040	22	13	0	0	22	6556	222	144	0	0	214
Chr 25	6179	332	207	0	0	326	811	30	14	0	0	28	6422	334	208	0	0	328	6199	318	198	0	0	312	929	49	25	0	0	49	6323	329	207	0	0	323
Chr 26	6823	208	135	0	0	189	1062	20	7	0	0	20	7081	213	141	0	0	193	6898	207	139	0	0	188	1092	30	16	0	0	28	6984	210	140	0	0	190
Chr 27	3465	145	76	0	0	139	549	12	5	0	0	12	3610	151	80	0	0	147	3513	149	78	0	0	143	622	23	13	0	0	23	3551	152	79	0	0	146
Chr 28	6066	191	126	0	0	182	921	20	10	0	0	19	6268	196	130	0	0	187	6101	190	127	0	0	182	1026	35	18	0	0	32	6194	191	125	0	0	182
Chr 29	6043	258	183	0	0	246	941	17	12	0	0	17	6558	272	190	0	0	260	6093	262	185	0	0	251	990	27	19	0	0	27	6210	269	189	0	0	257
Chr X	12148	497	310	0	1	463	2086	48	27	0	0	46	12658	519	321	0	1	482	12352	501	314	0	1	469	2309	68	38	0	1	63	12459	511	318	0	1	474

V. CONCLUSIONES

El presente trabajo utilizó como base dos recursos principales; el primero fue una de las más recientes tecnologías en genotipificación masiva disponibles en el mercado, el SNPchipHD que caracteriza más de 700,000 polimorfismos de nucleótido simple a lo largo de todo el genoma bovino. El segundo recurso utilizado es la información de anotación de genoma completo en los bovinos. Ambos recursos están disponibles en la mayoría de organismos de interés agronómico y comercial, esto nos permite reproducir el procedimiento secuencial ("pipeline") en distintas especies de estudio que cuenten con proyectos de mapas de haplotipos (HapMap).

Las herramientas utilizadas para llevar a cabo el procedimiento son de libre acceso y están disponibles en línea.

La caracterización *a priori* de polimorfismos de nucleótido simple no sinónimos es llevada a cabo generalmente partiendo de un gen causal de un fenotipo, estudiando su secuencia y después localizando los SNPs responsables de dicho fenotipo.

La aportación principal del presente procedimiento propone el camino contrario; partir de cadenas de oligonucleótidos ya conocidas en polimorfismos específicos, y detectar los causales de variación de estructura a nivel proteína, para así proponer nsSNPs de estudio mediante comparación de secuencias. Esto nos permite una caracterización *a priori* de polimorfismos de nucleótido simple no sinónimo del organismo de estudio. Cabe mencionar que a la fecha no se han utilizado las tecnologías de genotipificación masiva en ninguna especie para la caracterización de nsSNPs.

Una de las ventajas que presenta el método desarrollado es la caracterización de polimorfismos no sinónimos en un genoma completo, definiendo la cantidad de polimorfismos candidatos a ser relacionados con enfermedades y fenotipos de forma anticipada, por lo que una vez identificada una enfermedad se tiene un conjunto de posibles polimorfismos a ser potencialmente relacionados con la misma.

Al realizan estudios evolutivos y de asociación en distintas especies, este método de caracterización de polimorfismos no sinónimos nos ayudará a identificar genes, segmentos de genes o codones individuales que presenten presión selectiva de cambio, ayudando a identificar patrones evolutivos que involucren selección positiva (la selección natural que favorece los cambios de aminoácidos) en sistemas de evolución rápida como lo son patógenos diversos. Otras aplicaciones de este procedimiento es que puede auxiliar en el desarrollo de herramientas de investigación de fármacos, en predicción de estructuras en proteínas e interpretación de secuenciado de genomas.

5.1 PRODUCTOS

1. La versión inicial de baja densidad del presente procedimiento secuencial (basada en el Illumina SNPchip50) fue propuesta en el Segundo Congreso Nacional de Estudiantes de Posgrado del Instituto de Ingeniería, presentando el artículo intitulado " Efecto de los polimorfismos no sinónimos en la estructura secundaria de las proteínas (Caso de estudio: Ganado bovino)" [36] y poster. El artículo se muestra en la sección anexa, "Producto 1".

2. Se presentó la versión de alta densidad (utilizando el SNPchipHD) del procedimiento en el congreso "Plant and Animal Genome XX" en la ciudad de San Diego el mes de enero del 2012, por medio del poster intitulado "An Approach for Genome-wide Characterization of nsSNPs using the Illumina BovineHD SNPchip" [37] . El resumen se muestra en la sección anexa, "Producto 2".

3. Tesis para obtener el grado de Maestría en Ciencias con especialidad en Bioinformática.

5.2 TRABAJO FUTURO

Como continuación del estudio se trabaja en los siguientes proyectos:

1. Asociación con ruta biológica y/o fármacos y compuestos relacionados con la ruta. Estudio de Resultados.

2. Artículo sobre el procedimiento secuencial planteado y los resultados obtenidos; a ser sometido en revista indexada.

VI. REFERENCIAS BIBLIOGRÁFICAS

[1] Z.-Q. D. Bin Fan, Danielle M. Gorbach, Max F. Rothschild "Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals " *Asian-Aust. J. Anim. Sci.,* vol. 23, pp. 833-847, 2010.

[2] T. B. G. Sequencing, *et al.*, "The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution," *Science,* vol. 324, pp. 522-528, April 24, 2009 2009.

[3] Y. Liu, *et al.*, "Bos taurus genome assembly," *BMC Genomics,* vol. 10, p. 180, 2009.

[4] A. Zimin, *et al.*, "A whole-genome assembly of the domestic cow, Bos taurus," *Genome Biology,* vol. 10, p. R42, 2009.

[5] T. B. H. Consortium, "Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds," *Science,* vol. 324, pp. 528-532, April 24, 2009 2009.

[6] L. K. Matukumalli, *et al.*, "Development and Characterization of a High Density SNP Genotyping Assay for Cattle," *PLoS ONE,* vol. 4, p. e5350, 2009.

[7] R. Villa-Angulo, *et al.*, "High-resolution haplotype block structure in the cattle genome," *BMC Genetics,* vol. 10, p. 19, 2009.

[8] G. Rincon, *et al.*, "Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys," *Journal of Dairy Science,* vol. 94, pp. 6116-6121, 2011.

[9] M. E. Goddard and B. J. Hayes, "Mapping genes for complex traits in domestic animals and their use in breeding programmes," *Nat Rev Genet,* vol. 10, pp. 381-391, 2009.

[10] A. Vignal, *et al.*, "A review on SNP and other types of molecular markers and their use in animal genetics," *Genet Sel Evol,* vol. 34, pp. 275-305, May-Jun 2002.

[11] P. C. Ng and S. Henikoff, "Predicting the Effects of Amino Acid Substitutions on Protein Function," *Annual Review of Genomics and Human Genetics,* vol. 7, pp. 61-80, 2006.

[12] P. Kumar, *et al.*, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat Protoc,* vol. 4, pp. 1073-81, 2009.

[13] Y. Bromberg and B. Rost, "SNAP: predict effect of non-synonymous polymorphisms on function," *Nucleic Acids Research,* vol. 35, pp. 3823-3835, June 1, 2007 2007.

[14] P. D. Thomas and A. Kejariwal, "Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, pp. 15398-15403, October 26, 2004 2004.

[15] P. Yue, *et al.*, "SNPs3D: Candidate gene and SNP selection for association studies," *BMC Bioinformatics,* vol. 7, p. 166, 2006.

[16] C. Ferrer-Costa, *et al.*, "PMUT: a web-based tool for the annotation of pathological mutations on proteins," *Bioinformatics,* vol. 21, pp. 3176-3178, 2005.

[17] N. O. Stitziel, *et al.*, "topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association," *Nucleic Acids Research,* vol. 32, pp. D520-D522, January 1, 2004 2004.

[18] L. Bao, *et al.*, "nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms," *Nucleic Acids Research,* vol. 33, pp. W480-W482, 2005.

[19] I. A. Adzhubei, *et al.*, "A method and server for predicting damaging missense mutations," *Nat Methods,* vol. 7, pp. 248-9, Apr 2010.

[20] D. A. Magee, *et al.*, "A catalogue of validated single nucleotide polymorphisms in bovine orthologs of mammalian imprinted genes and associations with beef production traits," *Animal*, vol. 4, pp. 1958-70, Dec 2010.
[21] D. Chasman and R. M. Adams, "Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation," *Journal of Molecular Biology*, vol. 307, pp. 683-706, 2001.
[22] M. Lee, *et al.*, "Establishment of a pipeline to analyse non-synonymous SNPs in Bos taurus," *BMC Genomics*, vol. 7, p. 298, 2006.
[23] V.M.C. (2012, *Virtual Medical Centre*. Available: http://www.virtualmedicalcentre.com
[24] A. B. Harvey Lodish, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Lawrence Zipursky, and James Darnell, *Molecular Cell Biology*, Fifth Edition ed.: Macmillan Higher Education, 2004.
[25] M. M. C. David L. Nelson, *Lehninger Principles of Biochemistry*, 4th Edition ed.: Freeman, W. H. & Company, 2004.
[26] biología.edu.ar, "biología.edu.ar," 2012.
[27] A. Limited. (2012, *AbacusBio*. Available: http://www.abacusbio.com
[28] Illumina, "Data Sheet: BovineHD Genotyping Bead Chip," I. Inc., Ed., ed. San Diego, 2010.
[29] I. Illumina. (2012, *Illumina, Inc.* Available: www.illumina.com
[30] J. E. Stajich, *et al.*, "The Bioperl Toolkit: Perl Modules for the Life Sciences," *Genome Research*, vol. 12, pp. 1611-1618, October 1, 2002 2002.
[31] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 40, pp. 502-511, 2000.
[32] C. Cole, *et al.*, "The Jpred 3 secondary structure prediction server," *Nucleic Acids Research*, vol. 36, pp. W197-W201, July 1, 2008 2008.
[33] e. a. Shuichi Kawashima, "KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System," *Genome Informatics*, vol. 14, pp. 673-674, 2003.
[34] W. T. Sanger Institute. (2011, *Wellcome Trust Sanger Institute, Genome Research Limited*. Available: www.sanger.ac.uk/resources/software/gff/spec.html
[35] T. U. Consortium, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Research*, vol. 39, pp. D214-D219, January 1, 2011 2011.
[36] R. V. A. Patricia Carvajal López, Carlos Villa Angulo, Víctor M. González Vizcarra, "Efecto de los polimorfismos no sinónimos en la estructura secundaria de las proteínas (Caso de estudio: Ganado bovino)," presented at the Memorias del 2do Congreso Nacional de Estudiantes de Posgrado del Instituto de Ingeniería, Mexicali, Baja California. Mexico, 2010.
[37] L. K. M. Patricia Carvajal-Lopez, Curtis P. VanTassell, Rafael Villa-Angulo. (2012, *An Approach for Genome-wide Characterization of nsSNPs using the Illumina BovineHD SNPchip P0213*. Available: https://pag.confex.com/pag/xx/webprogram/Paper3999.html

VII. ANEXOS

PRODUCTO 1 - ARTÍCULO:

"Efecto de los polimorfismos no sinónimos en la estructura secundaria de las proteínas (Caso de estudio: Ganado bovino)".

Ver siguiente página.

Efecto de los polimorfismos no sinónimos en la estructura secundaria de las proteínas (Caso de estudio: Ganado bovino)

Patricia Carvajal López*, Rafael Villa Angulo, Carlos Villa Angulo, Víctor M. González Vizcarra
* patriciacarvajal.mxli@gmail.com

Abstract

One of the main areas of research in drug development is the study of the three-dimensional structure of proteins, since it defines its biological function. Recently, there has been an increased interest in finding the hereditary link of the protein three-dimensional conformation with the genetic variation of DNA, such information increases the statistical power in association studies of genes with disease. Non synonymous polymorphisms are a type of genetic variation that allow us to determine the primary, non epigenetic cause, of many changes in protein structure between groups of individuals in case and control. We present a method for identifying non-synonymous polymorphisms, based on genotypes obtained from commercial SNPchips. The method has been applied to cattle genotype information, captured with SNPchip50 from Illumina. This initial analysis will be used to develop a simulator of the possible evolution paths in organisms with high mutation rate (virus for example), and assist in the implementation of personalized medicine strategies.

Keywords: Non synonymous polymorphisms, genotype, SNPchips, protein structure.

Resumen

Una de las principales áreas de investigación en el desarrollo de medicamentos es el estudio de la estructura tridimensional de las proteínas, puesto que esta define su función biológica. Recientemente, se ha incrementado el interés por encontrar el vínculo hereditario de la conformación tridimensional de la proteína con las variaciones genéticas a nivel ADN, cuya información aumenta el poder estadístico en estudios de asociación de genes con enfermedades. Los polimorfismos no sinónimos son un tipo de variación genética que nos permite determinar la causa primaria, no epigenética, de muchos de los cambios en la estructura de las proteínas, entre grupos de individuos de caso y de control. En este trabajo presentamos un método para la identificación de polimorfismos no sinónimos, partiendo de genotipos capturado con SNPchips comerciales. El método ha sido aplicado a información genotípica del ganado Bovino, capturada con el SNPchip50 de la empresa Illumina. Este trabajo inicial será utilizado para desarrollar un simulador de posibles evoluciones de microorganismos con tasa de mutación alta (ej. virus), y asistir en la implementación de estrategias de medicina personalizada.

Palabras clave: Polimorfismos no sinónimos, genotipos, SNPchips, estructura de las proteínas.

1. Introducción

Una de las principales áreas de investigación en el desarrollo de medicamentos es el estudio de la estructura tridimensional de las proteínas, puesto que esta define su función biológica. Recientemente, se ha incrementado el interés por encontrar el vínculo hereditario de la conformación tridimensional de la proteína con las variaciones genéticas a nivel ADN, cuya información aumenta el poder estadístico en estudios de asociación de genes con enfermedades.

Las diferencias genéticas que existen entre individuos de la misma especie son llamadas polimorfismos (ver figura 1), variaciones en la secuencia de un lugar determinado del ADN entre los individuos de una población. La mayoría de estas variaciones se presenta en forma de inserciones o borrados de tamaño pequeño, o en forma de polimorfismos de nucleótido simple. Un polimorfismo de nucleótido simple (*SNP* por su acrónimo en inglés) se refiere a cualquier posición del genoma en la cual se segregan dos o más distintos nucleótidos en una población. Uno de los enfoques actuales en la investigación es el uso de polimorfismo de nucleótido simple para identificar variaciones responsables a nivel genómico de diversas variantes fenotípicas, incluyendo la susceptibilidad a enfermedades [1]. Los *SNP* son los polimorfismos más recurrentes en el genoma y ocurren en promedio un *SNP* cada kilobase entre los dos cromosomas de un individuo.

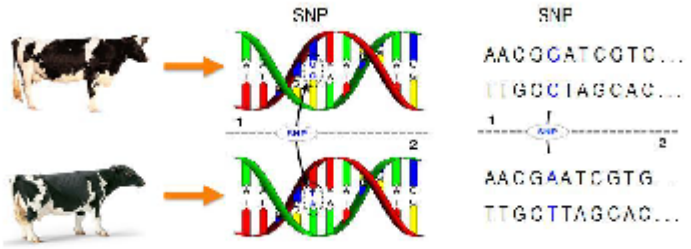


Figura 1. Un polimorfismo de nucleótido simple, SNP, es una posición especifica en el ADN, que contiene un valor de alelo distinto, en al menos 5% de la población.

Una consecuencia de la existencia de un SNP en la cadena de ADN, es que puede o no producir un cambio en la traducción de las secuencia de ARN mensajero a la cadena de aminoácidos. Se les llama polimorfismos sinónimos, silenciosos o neutrales a los polimorfismos que debido al fenómeno de "degeneración del código genético" (varios codones codifican un mismo aminoácido) no modifican a la cadena de aminoácidos resultante en el proceso de traducción. En contraparte, los polimorfismos no sinónimos causan una modificación en la cadena de aminoácidos traducida por el ARN mensajero [2]. El efecto potencial de esta modificación de cadena es el cambio en la secuencia de aminoácidos de una proteína, afectando así este cambio su estructura, y debido a que la forma tridimensional de una proteína está dictada por la secuencia de aminoácidos, su

función también es afectada. Un ejemplo del tipo de efectos de estas variaciones es la enfermedad "anemia falciforme", donde la sustitución de A por T en el codón 6-Glu de la hebra no transcrita (GAG), o de T por A en la hebra transcrita, conduce a un codón GUG en el ARN mensajero, que codifica Val en lugar de Glu. Esta única mutación (hemoglobina S, $\alpha_2\beta^s{}_2$) es la causa de dicha enfermedad.

En este trabajo presentamos un método para la identificación de polimorfismos no sinónimos, partiendo de genotipos capturado con SNPchips comerciales. El método ha sido aplicado a información genotípica del ganado Bovino, capturada con el SNPchip50 de la empresa Illumina. Este trabajo inicial será utilizado para desarrollar un simulador de posibles evoluciones de microorganismos con tasa de mutación alta (ej. virus), y asistir en la implementación de estrategias de medicina personalizada.

2. Material y Métodos

El análisis de variación en la cadena de aminoácidos provocados por polimorfismos no sinónimos, se realizó a partir de la secuencia genómica del cromosoma 6 del genoma bobino UMD3.1 obtenida de la base de datos *NCBI* [3] Esta secuencia genómica se comparó contra el archivo de polimorfismos del mismo cromosoma en la raza Limousine adquiridos de la prueba con el *SNPchip50*. Las diversas pruebas involucrando ambos archivos se programaron en *PERL* [4]. Así se sitúan los diversos *SNPs* dentro de la anotación y el genoma y se determina si los polimorfismos analizados son no sinónimos. En el caso afirmativo se simularon a nivel de estructura secundaria los cambios producidos con el simulador *scratch protein predictor* [5].

2.1. La anotación UMD 3.1 del genoma bobino

Ensamblar genomas completos conlleva modificaciones según lo requiera el análisis y actualizaciones de la información. La versión del genoma bovino utilizada en este análisis es la anotación UMD3.1. Fue publicada el 2009 y es de libre acceso a través de NCBI [6].

2.2. El código genético

El conjunto de procedimientos por los cuales se codifica el intercambio de nucleótidos (ARN mensajero) a aminoácidos se le denomina "código genético" [7]. El codón de inicio (AUG en células eucariotas) indica el punto inicial a partir del cual la secuencia de ARN mensajero se debe empezar a traducir en aminoácidos. El codón de paro no se traduce en un aminoácido, sólo indica la detención de la síntesis de la cadena de ARN mensajero. Existen tres codones de paro: AUG, UAG y UAA. La traducción de codones de bases a aminoácidos se encuentra descrita en la figura 2. Los 20 aminoácidos son: fenilalanina (PHE/F), serina, (Ser/S), tirosina (Tyr/Y), cisteína (Cys/C), leucina (Leu/L), prolina (Pro/P), histidina (His/H), triptófano (Trp/W), arginina (Arg/R), glutamina (Gln/Q), isoleucina (Ile/I), treonina (Thr/T), asparagina (Asn/N), lisina (Lys/K), metionina (Met/M/comienzo), valina (Val/V), alanina (Ala/A), ácido aspártico (Asp/D), glicina (Gly/G) y ácido glutámico (Glu/E). Basada en esta información se traduce la secuencia genómica seleccionada del genoma UMD3.1 en la cadena de aminoácidos final. Así mismo

se reemplazan los valores de los polimorfismos no sinónimos de frecuencia de alelo mayor y de frecuencia de alelo menor, y se hacen las traducciones con dichos remplazos.

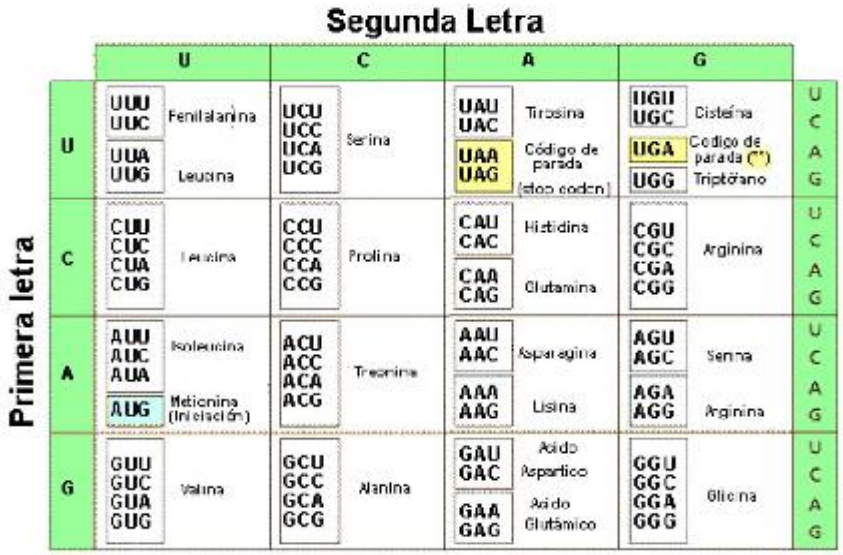


Figura 2. El código genético.

2.3. Genotipos del cromosoma 6 capturados en la raza Limousin con el *Bovine SNP chip 50*

La empresa Illumina, en colaboración con la USDA y las universidades de Missouri y de Alberta desarrollo el BovineSNP50 BeadChip, este hace pruebas para 54,609 pruebas de SNPs a través del genoma completo. Este habilita estudios de genética comparativa, entre otros [8]. En el 2009 se publicaron los resultados del estudio hecho en ganado bobino con el BivineSNP50 para comprobar la utilidad de la predicción genómica asociada. Se genotiparon 12,591 toros y vacas de distintas razas bajo esta plataforma y por condiciones de restricción se quedaron en caso de estudio 5,503 individuos [9].

2.4. Lenguaje de programación Perl

Perl es el lenguaje computacional más usado en análisis de información biológica, es de libre acceso, multiplataforma, permite conexión directa a bases de datos y es extensible (hay más de 500 módulos de aplicación). Este lenguaje es muy adecuado para indagar información de búsqueda de patrones de cadenas y es muy potente en búsqueda de expresiones regulares [10]. El procedimiento de determinación de las diversas cadenas de aminoácidos fue el siguiente:

1 Se programó un código en *Perl* para obtener todos los rangos en el cromosoma 6 a partir del archivo UMD3.1.gff. La información extraída se guardó en forma de texto.

2 Usando archivos de polimorfismos del cromosoma 6 para la raza Limousine, se desarrollo un programa en *Perl*. Con este se identificaron los SNPs heterocigotos, se obtuvieron las posiciones de los polimorfismos, las frecuencia de alelo mayor, frecuencia de alelo menor, y se genero un archivo de texto con dicha información.

3 Usando el archivo de información UMD3.1, se identificó la posición de los polimorfismos y sus frecuencias de alelo mayor y alelo menor. Con esta información se realizo un tercer programa que realiza lo siguiente:

4 Primero, ubica cuantos SNPs recaen dentro y fuera de los rangos establecidos en la anotación. Estos resultados se reflejan en la tabla 1 de la sección de Resultados.

5 Una de de las salidas que este programa genera es información sobre la ubicación de los polimorfismos, los cuales pueden caer en: *exón, CDS, gen, codón de paro o codón de inicio*. Los polimorfismos que caen dentro de estas categorías, fueron contados de forma individual. Los resultados se muestran también en la tabla 1 de la sección de resultados.

6 Otra salida de este programa indica las posiciones de cada polimorfismo que cayó dentro de las categorías. Esta información se guardó en un archivo de texto independiente.

7 Una salida adicional del programa describe: El número de rango, que tipo de información contiene, inicio de rango, final de rango, posición del primer SNP, frecuencia de alelo mayor, frecuencia menor, y así sucesivamente para cada SNP contenido dentro de las categorías descritas.

8 Analizando la salida anterior, se observó que el rango que contiene mayor cantidad de polimorfismos es el rango número 3,057. Se obtuvo un archivo en formato FASTA de la secuencia genómica perteneciente a este rango, de la base de datos *NCBI*.

9 Un cuarto programa en *Perl* genera los archivos FASTA con las secuencias de nucleótidos originales y las sustituciones de "frecuencia de alele mayor" y sustituciones de "frecuencia de alele menor".

10 Un quinto programa traduce las secuencias de nucleótidos a aminoácidos.

11 Por razones de límite de aminoácidos del simulador de estructuras, se obtuvieron secuencias de 100 aminoácidos, iniciando en la posición del primer cambio, menos dos aminoácidos, hasta donde ocurrió el primer cambio, mas 98 aminoácidos, para todas las cadenas de nucleótidos.

2.5. Simulador de estructuras secundarias *Scratch Protein Predictor*

La predicción de la estructura de una proteína ha sido un gran reto en la biología computacional. Se han creado diversas herramientas bioinformáticas disponibles en internet, que son de libre acceso. El simulador utilizado fue *"Scratch protein predictor"*, éste simulador integra métodos de aprendizaje computacional, patrones evolutivos, librerías fragmentadas de la base de datos *"protein data bank"* (PDB) y funciones energéticas para predecir características estructurales. Uno de los diez módulos de simulación contenidos en *Scratch protein predictor"*es el módulos SSpro8, este tiene una eficiencia estimada del 63%. Las clasificaciones de salida son: H: hélice alfa, G: hélice 3-10-, I: hélice pi (inusual), E: hebra extendida, B: puente beta, T: giro, S: doblez, C: *"the rest"[5]*. Bajo este modulo se sometieron diversas secuencias de aminoácidos de forma directa en la página web de este simulador, y posteriormente se recibieron las simulaciones vía correo electrónico.

3. Resultados y Discusión

3.1. Polimorfismos situados dentro de los rangos de anotación

El primer paso del análisis consistió en determinar si los *SNPs* del cromosoma 6 registrados para la raza Limousine se sitúan dentro de los múltiples rangos dictados en la anotación del genoma bovino (UMD3.1). Según es descrito por el estándar GFF en el tercer campo de información [11], dentro del archivo de anotación, las secuencias pueden recaer en: gen, exón, secuencia codificadora de ADN, codón de inicio o codón de paro. Este primer paso nos dio como resultado que: de 2,186 *SNPs* polimórficos, sólo 789 (36.14%) caen dentro de los rangos establecidos, y de estos 788 *SNPs* un 100% están situados dentro de genes (ver tabla1).

Tabla 1. Polimorfismos situados dentro de los rangos de anotación.

Polimorfismos fuera de rango	Polimorfismos dentro de rangos
1,394	788
63.85%	36.14%
Tipo de secuencia del rango	Cantidad – Porcentaje
Región codificadora de ADN	0 – 0%
Exón	0- 0%
Gen	789 – 100%
Codón de inicio	0 – 0%
Codón de paro	0 – 0%

3.2. Proteína de estudio

Una vez analizados los diversos locus de los polimorfismos y sus lugares dentro de la anotación, se determinó que el rango donde se situaron más *SNPs* fue el de las posiciones 33,473,221 al 33,809,586. Dentro de este rango cayeron 59 polimorfismos. De estos 59 polimorfismos, 33 SNPs provocaron variación en los aminoácidos de la cadena final en la

secuencia de frecuencia de alelo mayor, así también, estos mismos polimorfismos causaron 31 variaciones de aminoácidos en la secuencia de frecuencia de alelo menor (ver figura 2).

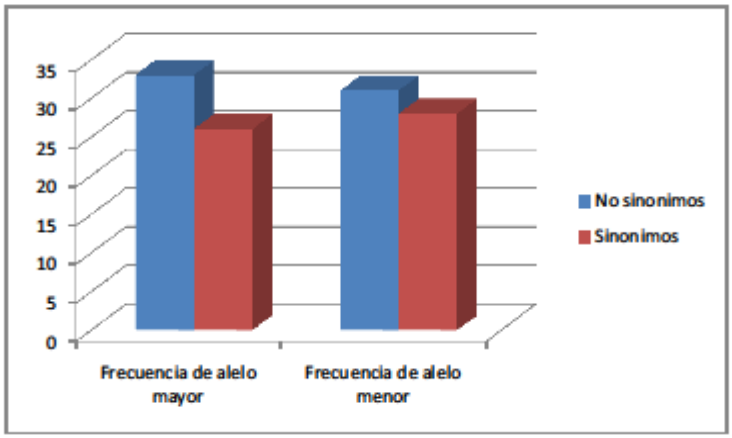


Figura 5. Polimorfismos sinónimos y no sinónimos.

3.3. Cambios en la estructura a nivel secundario

Utilizando el simulador *scratch protein predictor* se sometieron las secuencias de aminoácidos de la traducción de la secuencia original, la secuencia con las sustituciones de frecuencia de alelo mayor y la secuencia con las sustituciones de frecuencia de alelo menor (ver tabla 2). Se presentan aquí los cambios provocados por el primer polimorfismo no sinónimo en ambas cadenas. En el caso de frecuencia de alelo mayor el primer polimorfismo se sitúa en el aminoácido número 6,155, cambiando un triptófano de la secuencia original por una leucina. Este cambio de aminoácido no genero un cambio en la simulación de la estructura secundaria. Para la secuencia de frecuencia de alelo menor el primer polimorfismo está situado en el aminoácido número 5,904, cambiando un ácido glutámico por una glicina. El simulador presenta dos cambios en la estructura secundaria, originalmente se tienen 2 giros y en la sustitución se muestran dos "the rest" (según lo define el módulo SSpro8 del simulador *"scratch"*).

Tabla 2. Variación en las secuencias de aminoácidos y estructuras secundaria. Los cambios remarcados en gris. "FAMa", Frecuencia de alelo mayor. "FAMe" Frecuencia de alelo menor.

Secuencia	Secuencia de aminoácidos
Original	LFWMYFHSESITNKENCFFEVTCAILNIVPFTLRNYSSFYFTRNLSIPKI
FAMa	LFLMYFHSESITNKENCFFEVTCAILNIVPFTLRNYSSFYFTRNLSIPKIK
Original	GREGVKRLHVRLPFPAVFQQVNVEKCFTTTQHVIYSLAINHFFSFSETP
FAMe	GRGGVKRLHVRLPFPAVFQQVNVEKCFTTTQHVIYSLAINHFFSFTETP
Secuencia	SECUENCIA DE ESTRUCTURAS SECUNDARIAS
Original	CEEEEEECSEECCHTTCEEEEEEHEEEEECEEEEECCEEEEEETCCHHH
FAMa	CEEEEEECSEECCHTTCEEEEEEHEEEEECEEEEECCEEEEEETCCHHH

Original	CCTTEEEEEECCCCHEEEEEEHHHHHHHHHHHHHHHHHHHHHHEECCS
FAMe	CCCCEEEEEECCCCHEEEEEEHHHHHHHHHHHHHHHHHHHHHHEEECS

4. Conclusiones y Recomendaciones

En este trabajo, se demostró de manera didáctica que una sola variación en la secuencia de nucleótidos de las plantillas de ADN es suficiente para causar cambios en la estructura final de una proteína. Este análisis se sigue trabajando para obtener simulaciones de estructuras a nivel terciario, y también para simular cambios con sustituciones aleatorias o establecidas por el usuario.

Agradecimientos

A mis compañeros, M.C. Ulises Castro Peñaloza y M.C. José Barraza Beltrán. También al laboratorio de Bioinformática del Instituto de Ingeniería de la UABC.

Referencias

[1] Greg Gibson, S.V.M., *A Primer of Genome Science, Second Edition*, ed. A. Sinaur, Inc. Publishers. 2004.

[2] José Luque Cabrera, A.H., *Texto ilustrado de biología molecular e ingeniería genética: Conceptos, Técnicas y Aplicaciones en Ciencias de la Salud*, ed. Elsevier. 2006.

[3] Medicine, U.S.N.L.o. *National Center for Biotechnology Information,*. 2010; Available from: http://www.ncbi.nlm.nih.gov/.

[4] perl.org. *The Perl Programming Language*. 2010; Available from: http://www.perl.org/.

[5] J. Cheng, A.R., M. Sweredoski, P. Baldi, *SCRATCH: a Protein Structure and Structural Feature Prediction Server, Nucleic Acids Research*. 2005. **33**(Web Server Issue): p. 72-76.

[6] The Bovine Genome Sequencing and Analysis Consortium, C.G.E. and K.C.W. Ross L. Tellam, *The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution*, in *Science*. 2009.

[7] Jeremy M. Berg, J.L.T., Lubert Stryer, *Biochemistry*, ed. W.H.F.a. Company. 2002.

[8] Illumina, *BovineSNP50 Genotyping BeadChip*. 2010.

[9] Wiggans GR, S.T., VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF, Schenkel FS, Van Tassell CP., *Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada*. Journal of Dairy Science, 2009. **92**(7): p. 6.

[10] Jamison, D.C., *Perl Programming for Biologists*. 2003: John Wiley & Sons, Inc., Publication.

[11] Wellcome Trust Sanger Institute, G.R.L. *Wellcome Trust Sanger Institute, Genome Research Limited*. 2010; Available from: http://www.sanger.ac.uk/resources/software/gff/spec.html.

PRODUCTO 2 - RESUMEN:

"An Approach for Genome-wide Characterization of nsSNPs using the Illumina BovineHD SNPchip P0213".

08/01/13 Abstract: An Approach for Genome-wide Characterization of nsSNPs using the Illumina BovineHD SNP...

PLANT & ANIMAL GENOME XX
The Largest Ag-Genomics Meeting in the World.
January 14-18, 2012
San Diego, CA
www.intl-pag.org

Home/Search | Browse by Day | Browse by Type | Author Index | Poster Categories

P0213 **An Approach for Genome-wide Characterization of nsSNPs using the Illumina BovineHD SNPchip**

Patricia Carvajal-Lopez *, Universidad Autónoma de Baja California, Mexicali, Mexico*
Lakshmi K Matukumalli , NIFA, USDA, Washington, DC
Curtis P. VanTassell , USDA-ARS, Bovine Functional Genomics, Beltsville, MD
Rafael Villa-Angulo , UABC, Calexico, CA

SNPs are the most recurrent polymorphisms at the genome and they appear about one per each kilobase between both chromosomes of an individual. A principal approach of research investigation is the use of SNPs to identify genomic variation of diverse phenotypic traits. A method for identifying nsSNPs for any given organism, based on genotypes obtained from commercial SNPchips, is presented. Using the Illumina BovineHD SNPchip, which covers about 777,000 polymorphisms (an average of 26,105.43 per chromosome), nsSNPs were discovered by processing the SNP values and their corresponding flanking sequences. Translation was obtained analyzing the sequences in all forward and reverse strands, leading to the detection of nsSNPs. The non synonym sequences detected were located within the UMD 3.1 annotation; all sequences which fell into the "gene / every other" category were selected for further analysis. The second phase of the research was to detect all the previously selected sequences which lead to a change at the secondary structure of the protein product. Once the sequences that originated a secondary structure change were detected they were linked to their phenotype via GO and KEGG. An average of 39.79% of nsSNPs fell into the boundaries of the UMD3.1 annotation; out of this percentage, a 3.65% occurred within the "gene / every other" simultaneous selection of the annotation categories. 62.28% of these "gene / every other" - nsSNPs originate a change in their secondary structure.

Back to: Poster Abstracts

<< Previous Abstract | Next Abstract >>

https://pag.confex.com/pag/xx/webprogram/Paper3999.html 1/1