UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERÍA

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

"Imputación de Genotipos para la implementación de Selección Genómica en Hatos lecheros de Baja California"

TESIS PARA OBTENER EL GRADO DE:

Maestro en Ciencias

PRESENTA

Mirvana Elizabeth González Macías

DIRECTOR

Dr. Rafael Villa Angulo

Mexicali, B. C. ***Junio/2012***

AGRADECIMIENTOS

En primer lugar te agradezco a ti Dios, por ayudarme a terminar este proyecto, gracias por darme la fuerza y el coraje para hacer este sueño realidad, por estar conmigo en cada momento de mi vida. Gracias por llenarme de hermosas bendiciones.

A mi familia, por su apoyo y compresión en todo momento. Ellos son mi pilar y mi fuerza motriz, AMO A TODOS !!. A mis pequeños sobrinos LALO y GABY, sus risas me hacen crecer y llenarme de energía.

A mi director de tesis y amigo Dr. Rafael Villa Angulo, por su confianza depositada en mí, por toda la paciencia y su valioso tiempo, por sus valiosas enseñanzas, apoyo incondicional, por transmitirme ese entusiasmo y satisfacción de lograr las cosas, a no darnos nunca por vencidos.

Agradezco al Dr. Carlos Villa Angulo por su tiempo y apoyo incondicional, por ser tan cuidadoso con esas pequeñas cositas que hacen que el escrito se lea y entienda mejor. Sobre todo por enseñarme que todo sale mejor si te relajas.

A la Dra. Adriana Morales Trejo, por su apoyo y observaciones, sobre todo por ser parte del comité.

Al Dr. Victor Manuel González Vizcarra, por haber aceptado ser parte del comité, por su apoyo y recomendaciones las cuales me hicieron crecer.

Agradezco el apoyo brindado por CONACYT, para la realización de mis estudios de Maestría en Ciencias.

A todos y cada uno de mis maestros, compañeros y amigos:
GRACIAS POR SU PACIENCIA!!

ÍNDICE

I INTRODUCCION 1

1.1 Situación actual de la industria lechera en la región de Baja California 1

1.2 Mejoramiento genético de ganado bovino 4

1.2.1 Estimación del valor genético de un animal 7

1.2.2 Métodos de selección de ganado bovino 9

1.2.2.1 Selección tradicional 9

1.2.2.2 Selección basada en marcadores 11

1.2.2.3 Selección genómica 12

1.3 Tecnologías de genotipado de salida masiva 15

1.3.1 SNPchips 16

1.4 Imputación de genotipos 20

1.4.1 El algoritmo Beagle 24

1.4.2 El algoritmo Impute 24

1.5 Planteamiento del problema de tesis 25

1.6 Objetivos del trabajo de tesis 27

1.6.1 Objetivo general 27

1.6.2 Objetivos específicos 27

II IMPUTACIÓN DE GENOTIPOS EN GANADO BOVINO LECHERO 28

2.1 Descripción de los datos 28

2.1.1 Razas y número de animales seleccionados. 29

2.1.2 Cromosomas y número de Marcadores SNP utilizados en el experimento 29

2.2 Preprocesamiento de los genotipos 31

2.3 Imputación en la raza Hereford 31

2.4 Imputación en la raza Holstein 33

2.5 Imputación en la raza Jersey 35

2.6 Imputación en la raza Brown Swiss 38

III ANÁLISIS Y DISCUSIÓN DE RESULTADOS 41

3.1 Análisis de la imputación con el algoritmo Beagle. 41

3.2 Análisis de imputación con el algoritmo Impute 42
3.3 Análisis del impacto de la imputación en la estimación de los GEBVs. 43
3.4 Análisis de optimización de costos por imputación..... 45
IV CONCLUSIONES Y TRABAJO FUTURO..... 47
4.1 Conclusiones 48
4.2 Trabajo futuro 49
BIBLIOGRAFÍA..... 51
ANEXOS..... 54

LISTA DE FIGURAS

Figura 1. Producción de leche en los últimos diez años en el estado de Baja California... 2
Figura 2. Mejoramiento genético en ganado bovino. 5
Figura 3. Efecto epigenético. 6
Figura 4. Método tradicional de selección. 10
Figura 5. Marcadores SNP. 13
Figura 6. Selección Genómica. 14
Figura 7. Ejemplo de una TGSM. 15
Figura 8. Diagrama esquemático de un SNPchip. 16
Figura 9. Imputación fusionando dos conjuntos de genotipos. 21
Figura 10. Imputación por paneles de referencia de haplotipos.. 22
Figura 11. Exactitud de imputación en el cromosoma 1 de la raza Holstein.. 34
Figura 12. Gráfica de la exactitud global en el genoma completo de la raza Holstein..... 35
Figura 13. Exactitud de imputación en el cromosoma 1 de la raza Jersey.. 37
Figura 14. Gráfica de la exactitud global en el genoma completo de la raza Jersey. 38
Figura 15. Exactitud de imputación en el cromosoma 1 de la raza Brown Swiss. 39
Figura 16. Gráfica de la exactitud global en el genoma completo de la raza Brown Swiss. 40
Figura 17. Gráfica global de exactitudes de las tres razas del ganado lechero, Holstein Jersey y Brown Swiss. 42

LISTA DE TABLAS

Tabla 1. SNPChips comercializados por Illumina In, para muestrear al ganado bovino. 18
Tabla 2. Lista de algoritmos que pueden ser utilizados para la imputación de genotipos. 23
Tabla 3. Número de animales por raza, en los experimentos. 29
Tabla 4. Número de SNPs por cromosoma del panel de baja densidad (Bovine6k) y el panel de alta densidad (Bovine50k). 30
Tabla 5. Resultados de imputación en el cromosoma 1 con el algoritmo Beagle, incrementando el porciento de valores a imputar del 1% al 6%. 32
Tabla 6. Resultados de imputación en el cromosoma 6 con el algoritmo Beagle, incrementando el porciento de valores a imputar del 1% al 6%. 32
Tabla 7. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Holstein... 34
Tabla 8. Suma global de exactitudes en el genoma completo de la raza Holstein. 35
Tabla 9. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Jersey..... 36
Tabla 10. Suma global de exactitudes en el genoma completo de la raza Jersey. 37
Tabla 11. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Brown Swiss. 39
Tabla 12. Suma global de exactitudes en el genoma completo de la raza Brown Swiss.. 40
Tabla 13. Costo del conjunto de SNPchips necesarios para genotipar distinto número de animales 45

Resumen

La imputación de genotipos es el proceso de predecir valores de marcadores que no han sido directamente capturados en un ensayo de genotipado. Los algoritmos de imputación hacen uso de técnicas estadísticas para capturar patrones en conjuntos de haplotipos de referencia, y aumentan el poder estadístico de los estudios al predecir los valores de genotipos no observados. En este proyecto se aplicó y evaluó imputación de un panel de ~6,000 a uno de ~50,000 polimorfismos de nucleótido simple, en las razas Holstein, Jersey y Brown Swiss. El mejor algoritmo de imputación obtuvo una exactitud de imputación de 94.02% para la raza Brown Swiss, 93.14% para la raza Holstein y 92.20% para la raza Jersey. Estos porcentajes de exactitud permitieron hacer una proyección con el 95% de precisión en la estimación de los valores genómicos de los animales, con motivos de implementación de un programa de mejora genética basado en Selección Genómica. La aportación principal de este trabajo de tesis fue la definición de un procedimiento de optimización económica para implementar programas de mejora genética; realizando genotipado en conjuntos pequeños de marcadores, y mediante imputación prediciendo los valores de conjuntos más grandes. Con este se logra disminuir el costo más de 51%, comparado con hacer genotipado en los conjuntos grandes de genotipos. Esta estrategia habilita la transferencia de tecnología de Selección Genómica para aplicaciones masivas en los hatos lecheros del estado de Baja California.

ABSTRACT

Abstract

Genotype imputation is the processes of predicting marker values not directly captured in a genotyping assay. Imputation algorithms work by using haplotype patterns in a reference panel, and increase the power of genetic association studies, by predicting unobserved genotypes. In this dissertation project imputation from a panel of ~6,000 to a panel of ~50,000 single nucleotide polymorphisms was applied and evaluated, in animal samples from Holstein, Jersey, and Brown Swiss cattle breeds. The best imputation algorithm produced an imputation accuracy of 94.02% in Brown Swiss, 93.14% in Holstein, and 92.20% on Jersey. These accuracy percentages permitted to make predictions of the Genomic Value of the animals with a precision of 95%, and enabled the implementation of a genetic improvement program based on Genomic Selection. The main contribution of this work was the definition of a procedure for optimizing the genotyping cost in genetic improvement programs. It recommends to genotype using low density panels and to apply imputation for predicting genotype values for higher density panels. The procedure guarantees a reduction of more than 51% of the cost compared to genotyping directly high density panels. It enables the transference of Genomic Selection technology to be applied massively in dairy herds of Baja California.

I INTRODUCCION

1.1 Situación actual de la industria lechera en la región de Baja California

En México, la industria del ganado bovino representa, en términos monetarios, el 47.3% de la producción nacional pecuaria [1]. De este porcentaje, el 43.7% proviene del ganado utilizado para la producción de leche, siendo la tercera actividad pecuaria más importante, por encima de la producción de cerdo y huevo. Sin embargo, a pesar de que en México contamos con un importante número de cabezas de ganado bovino dedicado a la producción de leche, nos situamos entre las naciones con menor productividad [2]. Países como Canadá y Argentina que poseen una cantidad mucho menor de cabezas de ganado lechero, han reportado una producción mayor en los últimos años. En general, los factores que han limitado los esfuerzos por incrementar la producción son de índole económico, de infraestructura, genéticos y ambientales; los cuales han provocado que la producción sea insuficiente, por lo que se ha tenido que importar una gran cantidad de

leche, principalmente de países como Estados Unidos de América, Nueva Zelanda y Uruguay [1].

En la región de Baja California, la ganadería de leche ha presentado, históricamente problemas de producción, principalmente en las zonas áridas, tales como los municipios de Mexicali y San Felipe. Y, a pesar del esfuerzo que han realizado los ganaderos de la región por mantenerla como una industria necesaria para el desarrollo de la región, la producción ha sufrido una disminución paulatina y constante en los últimos años [3]. La figura 1 muestra como la producción ha venido disminuyendo en los últimos 10 años.

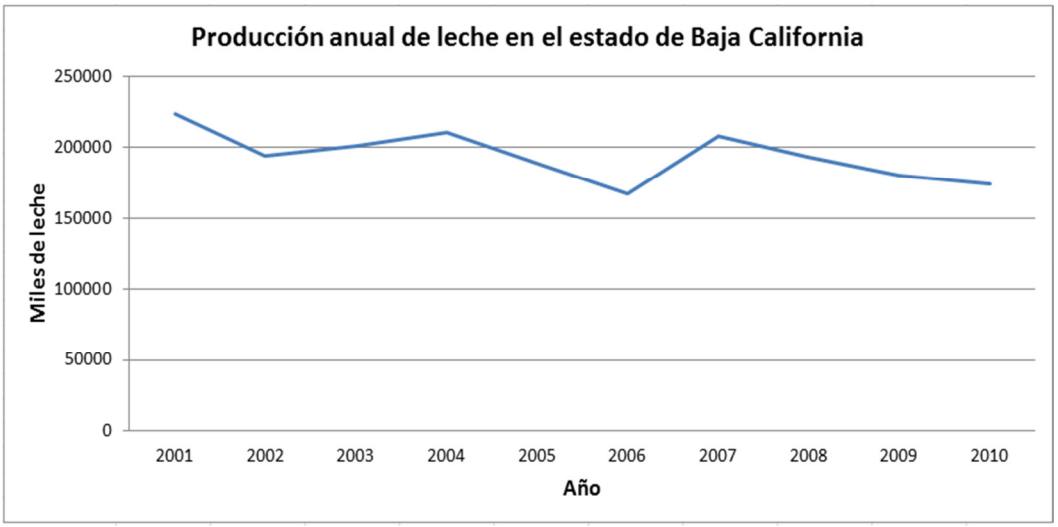

Figura 1. Producción de leche en los últimos diez años en el estado de Baja California (la información fue obtenida de la página de internet de la SAGARPA, en su sección de Servicios de Información Agroalimentaria Pesquera, abril de 2012).

El año pasado, 2011, en adición a la disminución de la productividad lechera manifestada en los últimos años, se sumó un alto desabasto de forraje para alimentación

del ganado, generando que las crisis en el sector lechero se agudizara. Esto provocó la pérdida de más de 10,000 cabezas de ganado lechero, que fueron vendidas para alimentar otras [4]. En un esfuerzo por frenar la agudización de la crisis en el sector lechero, miembros de la Unión Ganadera del Estado y directivos de la Comisión de Agricultura, Ganadería, Asuntos Pecuarios y Pesca del Congreso llevaron a cabo varias mesas de trabajo, acordando buscar y aplicar medidas para rescatar la industria lechera de Baja California y evitar que la crisis actual haga desaparecer esta industria en el estado [5].

Sin duda, la solución al problema de la industria lechera es compleja, y no surgirá de forma inmediata, debido a que, no son por si solos los factores ambientales, de alimentación y de genética propia del ganado los que están provocando el decline de la producción. Históricamente, en el sector productivo lechero regional no se tiene una cultura de implementación de planes integrales de organización y seguimiento para la producción a nivel de manejo de los animales, del medio ambiente que rodea los hatos, de medicina preventiva (vacunas, desparasitación, etc.), de aspectos de bioseguridad (tapetes sanitarios, manejos de desechos, etc.), ni de implementación de estrategias de mejora genética continua y controlada; los cuales son cruciales para garantizar la optimización en la producción de la industria ganadera.

Es de crucial importancia el establecimiento de programas integrales a ser implementados en los hatos lecheros del estado, para mejorar la producción ganadera; en especial la mejora genética del ganado beneficiando con programas de reproducción las características productiva de interés económico, así como la resistencia a efectos de estrés calórico y/o una baja susceptibilidad a enfermedades de predisposición genética debe

estar aunada a un programa de salud animal, de bioseguridad y de registro productivo adecuado.

En resumen, un plan estratégico adecuado para mejorar la producción de leche en la región de baja california debe incluir junto a los aspectos antes mencionados, un plan adecuado de mejora genética continua, que garantice la selección de ganado genéticamente óptimo para garantizar la producción.

A continuación se describen algunos aspectos importantes sobre mejoramiento genético de ganado bovino.

1.2 Mejoramiento genético de ganado bovino

El mejoramiento genético con fines de aumentar la productividad del ganado bovino comenzó a lograrse mucho antes de que el mecanismo hereditario fuera entendido. Siempre se ha buscado maximizar la probabilidad de que los padres transmitan sus características productivas a sus crías. Históricamente, todas las estrategias de mejora genética están basadas en la premisa de realizar la mejor selección de los animales a ser utilizados para reproducción [6,7].

Figura 2. Mejoramiento genético en ganado bovino. Haciendo cruzamiento de los mejores sementales con las mejores vacas se ha logrado que los padres trasmitan sus características productivas a sus crías.

Los inicios de los métodos tradicionales y modernos de selección datan de los 1700's, cuando el Ingles Robert Bakewell estableció las primeras estrategias para crianza de ganado bovino de carne [8]. Bakewell definió el ganado que consideró ideal (ejemplo: machos con buenos aplomos, bloques bien formados y de maduración rápida, entre otras características), y con esto propuso el objetivo de la crianza buscando obtener el ganado ideal. Su estrategia fue cruzar lo mejor con lo mejor sin importar la relación sanguínea entre los animales, cosa que no siempre lo condujo a los mejores resultados. Los métodos modernos de crianza están soportados en dos enseñanzas básicas del trabajo de Bakewell: 1) es de crucial importancia definir un objetivo o meta específica en un programa de crianza, y 2) las mediciones de mérito genético; tal como las diferencias esperadas de progenie, DEPs [9], cuentan en toda la información de rendimiento de los animales emparentados.

Asignar un valor cuantitativo a la genética de un animal no es un trabajo sencillo debido a que se sabe que muchos de los rasgos fisiológicos observables (fenotipos) del animal, de sus propiedades bioquímicas, de su morfología, de su desarrollo, comportamiento y producción, son epigenéticos. Esto es, son regulados por la interacción de los genes con el medio ambiente. Lo que genera, que según los factores de manejo, la raza, la región geográfica, la época del año y el estado de lactancia, entre otros; sumado a la naturaleza dominante de algunos genes que posee el animal, al efecto aditivo de genes que actúan de forma independiente y al efecto epispástico de genes que actúan de forma interactiva para controlar los rasgos del animal, influya de forma determinante en el rendimiento productivo.

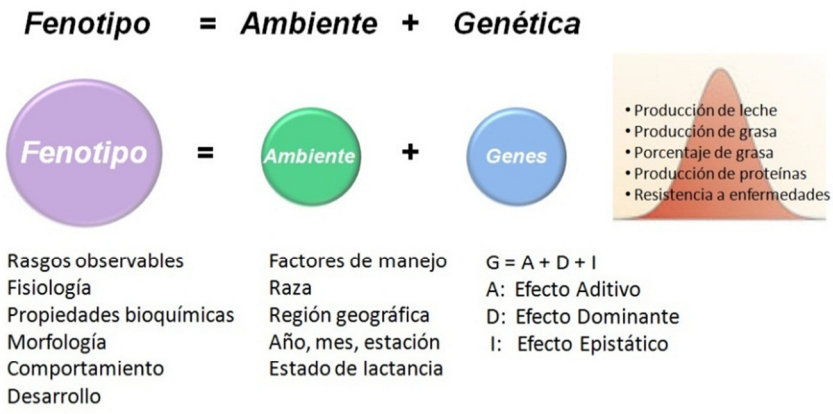

Figura 3. Efecto epigenético. La mayor parte de los rasgos de importancia económica en el ganado bovino son epigenéticos. La mayor parte de los fenotipos son regulados por la interacción de los genes con el medio ambiente.

A medida que se ha ido entendiendo el fenómeno epistático en el contexto productivo de los animales, se ha ido mejorando la técnica para signarles un valor

genético. De hecho, una fuente invaluable de información que a contribuido al avance de los programas de mejora fue la creación, hace más de 100 años, de las asociaciones de razas, con el objetivo de generar registros de manadas y recopilar de forma organizada información de pedigrí y de rendimiento del ganado. Esta información condujo eventualmente a experimentos efectivos de crianza selectiva, con el fin de incrementar la productividad, salud y conformación. Muchas razas de ganado bovino son ahora sujetas a selección de múltiple rasgos usando estimaciones del mérito genético, producido por métodos estadísticos y computacionales complejos [10,11]. Por ejemplo, el mérito genético para la entrega de leche en ganado lechero Holstein incrementa actualmente a una tasa de 1%, o 21 galones por año.

A continuación se presenta una descripción de cómo se realiza la estimación de valor genético de un animal. Posteriormente se presenta una revisión de los métodos de selección de animales dentro de programas de mejora genética.

1.2.1 Estimación del valor genético de un animal

El valor genético de crianza de un animal (EBV, "Estimated Breeding Value", por sus siglas en inglés) es estimado para cada rasgo de importancia económica. Métodos estadísticos avanzados evalúan de forma simultánea a sementales y vacas, integrando una gran gama de información fenotípica y de pedigrí, tal como: datos sobre los padres y/o hijos y hermanos, distribución de las hijas en el hato, longitudes de lactancia, número de récords de lactancia por hija y diferencia genéticas entre miembros del hato, entre otros. Estos métodos son ejecutados en computadoras de alta velocidad para estimar de forma

precisa el mérito genético de un animal (Ej. BLUP [12] es el método más utilizado actualmente para evaluar sementales). En este proceso deben ser identificadas todas las posibles fuentes de variación ambiental, y cada animal debe ser evaluado en relación a otros animales criados bajo un manejo igual o similar. De esta forma, el mérito genético para un animal específico es calculado como la desviación de su valor genético de crianza estimado con respecto a la media en su hato, o en hatos con condiciones de manejo similar [13].

El EBV representa la contribución genética en el récord del animal. Por ejemplo; un EBV = +2000 lbs. para una vaquilla, nos indica que bajo las mismas condiciones de manejo, cuando esta vaquilla sea madura, se espera que produzca 2000 lbs. más que una vaquilla con un EBV = 0. La mitad del EBV de un animal es utilizado para predecir su habilidad de transmitir su valor genético a sus crías (PTA, "Predicted Transmmiting Ability", por sus siglas en ingles). Por ejemplo; un PTA = +1000 lbs. para la vaquilla del ejemplo anterior, nos indica que bajo las mismas condiciones de manejo, se espera que una hija madura de esta vaca, produzca 1000 lbs. más que una vaca con PTA = 0. En el caso de animales jóvenes sin récord, se obtiene el promedio de la habilidad transmisora de sus padres (PA, "Parent Average", por sus siglas en ingles). Esto es:

PA = ½ PTA del padre + ½ PTA de la madre

La certeza de la estimación del mérito genético se calcula una vez teniendo el rendimiento del animal, o sus crías, según el caso. Se utiliza el récord del animal y su información de pedigrí. Con esto se mide la precisión de PTA, que es el cuadrado del grado de correlación de los valores de rendimiento real y los estimados de PTA. Se puede

estimar el valor de certeza para animales jóvenes sin récord utilizando la certeza de sus padres:

Certeza = (certeza padre + certeza madre)/4

Este es un valor estimado que utiliza la información de los padres pero no garantiza que las crías hereden los rasgos genéticos deseados de sus padres.

1.2.2 Métodos de selección de ganado bovino

Los métodos de selección de ganado dentro de los programas de mejora genética han evolucionado con la tecnología y la experiencia acumulada en los últimos 300 años. Los métodos difieren básicamente en la forma de obtener el valor genético de los animales. Estos podemos clasificarlos como sigue: método tradicional, método asistido por marcadores y métodos de selección genómica. A continuación se hace una descripción de cada uno de estos.

1.2.2.1 Selección tradicional

El método tradicional de crianza selectiva utiliza información fenotípica y de pedigrí para calcular el valor genético de los animales. El análisis de la información es realizado con modelos sofisticados de selección cuantitativa. Este método ha rendido frutos satisfactorios, aun cuando no se sabe con exactitud cuáles son los genes que han incrementado o decrementado su frecuencia en las poblaciones de ganado. Generalmente se describe el método de selección como una "caja negra" debido a que no se sabe

exactamente cuales genes han sido afectados. La implementación de programas de pruebas efectivas de progenie y de sementales élites bien validados se ha convertido en el método que mejores resultados arroja en programas de mejora genética. Varios avances han hecho que éste método básico haya mejorado de una forma sustancial; la adopción de comparaciones con animales contemporáneos y posteriormente del modelo animal para realizar la evaluación genética son ejemplo de ello. En décadas recientes, los productores y genetistas han reconocido la importancia de la crianza en la mejora de la salud, reproducción y longevidad, por lo que rasgos de salud, reproducción y vida productiva se han convertido en una parte muy importante para la selección de ganado.

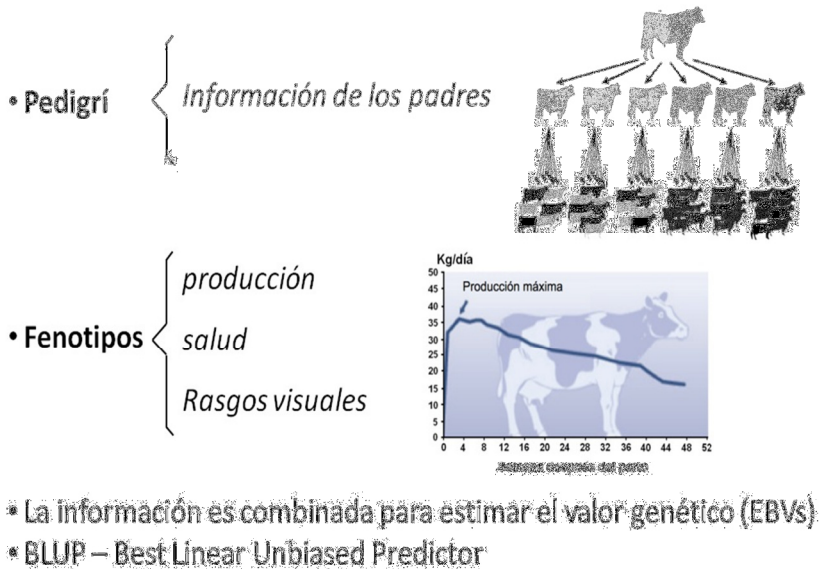

Figura 4. Método tradicional de selección. Este está basado en el mérito genético, calculado de información fenotípica y de pedigrí.

1.2.2.2 Selección basada en marcadores

Con el método de selección tradicional, que utiliza información fenotípica y de pedigrí, se logra tener una efectividad del 30 – 40% en las predicciones. Esto se debe en gran parte a que, aun cuando logramos la mejora genética al ver un aumento en nuestra producción, el conjunto de genes que han sido favorecidos (o desfavorecidos) en nuestro programa de selección sigue siendo una caja negra.

Un adelanto crucial en la integración del análisis genético en los esquemas de selección fue la inclusión de los marcadores moleculares, con los cuales se integra de forma específica el efecto de los genes en el rendimiento productivo del animal. Un marcador de ADN es una posición, o región, dentro de un cromosoma, físicamente identificable y cuya heredabilidad puede ser monitoreada. La función inicial y básica de los marcadores es identificar la expresión de genes relacionados con las características de importancia económica, y utilizar esta información para hacer una estimación más acertada de los EBVs.

Aun cuando la idea de usar marcadores de ADN para incrementar la tasa de ganancia genética en el ganado lechero, ha existido por décadas [14,15], la adopción de la selección asistida por marcadores (MAS, "Marker Asisted Selection", por sus siglas en ingles) en la industria lechera había estado limitada, hasta hace muy poco tiempo. Existían deferentes razones para esto; para muchos rasgos cuantitativos, de importancia económica, tales como los rasgos de producción y salud, en el ganado lechero, una gran cantidad de locus afectan a cada rasgo, donde cada locus captura una proporción limitada de la varianza genética total [16,17,18,19]. Consecuentemente, solo una ganancia

relativamente pequeña era posible con el número limitado de marcadores que se tenían disponibles, y el costo de genotipado de estos era alto. En adición, la complejidad para el cálculo de los valores genéticos de cría incluyendo la información de los marcadores era una barrera más para la aplicación de la selección asistida por marcadores [20].

1.2.2.3 Selección genómica

La selección genómica (SG) es una forma de hacer selección asistida por marcadores genéticos (MAS), en la cual los marcadores están distribuidos a lo largo de todo el genoma, y todos los Locus de Características Cuantitativas (QTLs) están en desequilibrio ligado con al menos un marcador [21]. La Selección Genómica está revolucionando a la industria del ganado. En esta, la toma de decisiones de selección de ganado para su reproducción está basada en el valor genómico de cría de los animales (GEBV, "Genomic Estimated Breeding Value", por sus siglas en ingles).

Los GEBVs son calculados como la suma de los efectos de una cantidad grande de marcadores moleculares, distribuidos a lo largo de todo el genoma. Los genotipos y/o haplotipos de los marcadores involucrados capturan los efectos de todos los QTLs que contribuyen a la variación de un rasgo de valor económico [22,23,24].

La revolución de la selección genómica empezó con dos desarrollos claves. El primero fue la reciente secuenciación del genoma bovino [25], lo cual habilitó el descubrimiento de miles de marcadores de ADN, en forma de polimorfismo de nucleótido simple, SNPs [26]. El descubrimiento y caracterización de miles de marcadores SNPs en el genoma

bovino y otros genomas de especies pecuarias trajo como consecuencia una reducción dramática en los costos de genotipado.

El segundo desarrollo fue la demostración de que es posible tomar decisiones de selección con alta precisión cuando los valores genéticos de los animales son predichos usando una cantidad densa de marcadores [27].

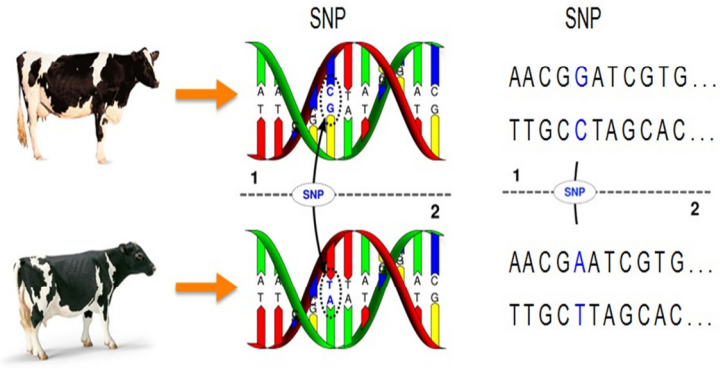

Figura 5. Marcadores SNP. La forma más común de polimorfismos en los genomas son Polimorfismos de Nucleótido Simple.

La tecnología de selección genómica involucra tres pasos principales; 1) captura de las muestras y obtención del ADN genómico, 2) Aplicación de SNPchips para obtención de los genotipos y 3) Análisis bioinformático de la información para obtener los GEBVs. La captura de la muestras pueden ser hechas de esperma, sangre, saliva, segregación nasal o una pinchada en oreja, entre otras. Estas muestras deben ser transportadas al laboratorio para su procesado químico de extracción de ADN. El ADN debe ser ADN genómico completo y purificado. El ADN es amplificado y depositado en SNPchips. Se realiza una reacción de hibridización del ADN al SNPchip para posteriormente correr una reacción de extensión en la cual el valor alélico de cada

posición polimórfica queda visible por elementos fluorescentes. Por medio de un escáner laser son medidos los valores de los alelos en lo polimorfismos y se genera una base de datos con la información genética de cada uno de la animales de la muestra.

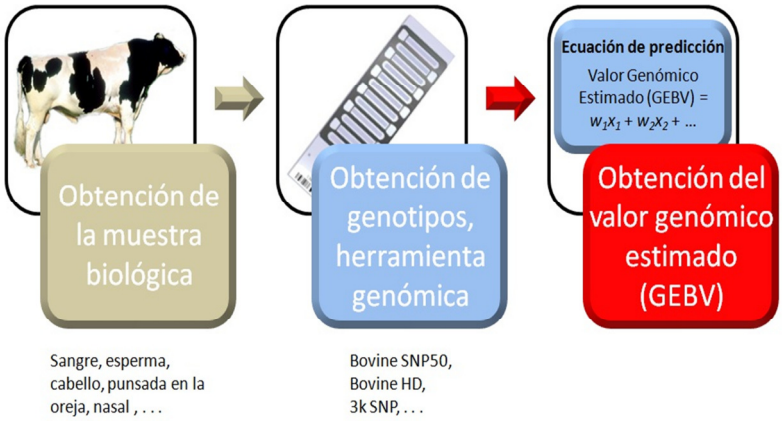

Figura 6. Selección Genómica. La selección genómica involucra tres pasos principales: 1) captura de la muestras y obtención de ADN genómico, 2) extracción de genotipos y 3) Análisis de la información.

Métodos bioinformáticos de análisis son aplicados a los datos para hacer estudios de asociación de genes con los rasgos de importancia, con enfermedades y para obtener los GEBVs. Toda esta información resultante es integrada a los criterios para seleccionar los sementales y las vacas que serán utilizadas para generar las nuevas crías en el hato.

En la siguiente sección se describe a más detalle las tecnologías de genotipado masivo que hacen uso de SNPchips para muestrear la genética de los animales.

1.3 Tecnologías de genotipado de salida masiva

Los polimorfismos de nucleótido simple son el tipo de variación genética más común en los genomas, millones de estos has sido detectados y caracterizados en distintos organismos, tales como los humanos y bovinos [26,28]. Existen métodos biomoleculares que pueden identificar de forma directa e independiente la información genotípica y haplotípica de los SNPs. Sin embargo, debido a que son de alto costos y consumen mucho tiempo, estos son usados solo para cantidades pequeñas de muestras [29]. Para muestras a gran escala (cientos y miles de individuos), se han desarrollados las Tecnologías de Genotipado de Salida Masiva (TGSM); con las cuales miles de SNPs son capturados de forma simultánea en un solo ensayo [30,31,32].

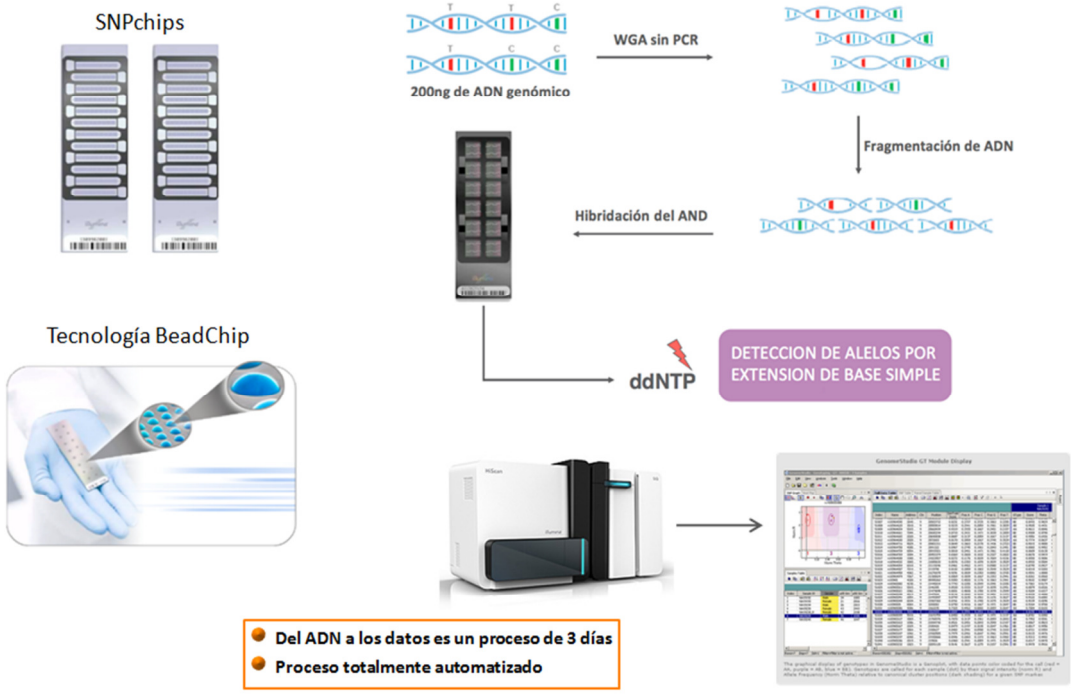

Figura 7. Ejemplo de una TGSM. Ensayo *Infinitum* de la empresa Illumina. Puede muestrear miles de SNPs en distintos individuos de forma simultánea.

Estas tecnologías están permitiendo la implementación de estudios de asociación de genoma completo a gran escala para detectar variaciones genéticas comunes, que tienen un rol significativo en la identificación de locus asociados con enfermedades, proteínas, biomarcadores, y respuestas farmacogenómica [33]. Las TGSMs están basadas en el uso de SNPchips para la captura de los genotipos. A continuación se describe a más detalle los SNPchips.

1.3.1 SNPchips

Un SNPchip es un arreglo matricial en el cual cientos de miles de pruebas (oligos) son colocadas en un chip (laminilla) habilitando la medición de miles de SNPs de forma simultánea [34]. El valor de los alelos en la posición de un SNP es capturado hibirdizando un oligo de DNA complementario en el sitio del SNP. La figura 8 muestra de forma esquemática la arquitectura de un SNPchip.

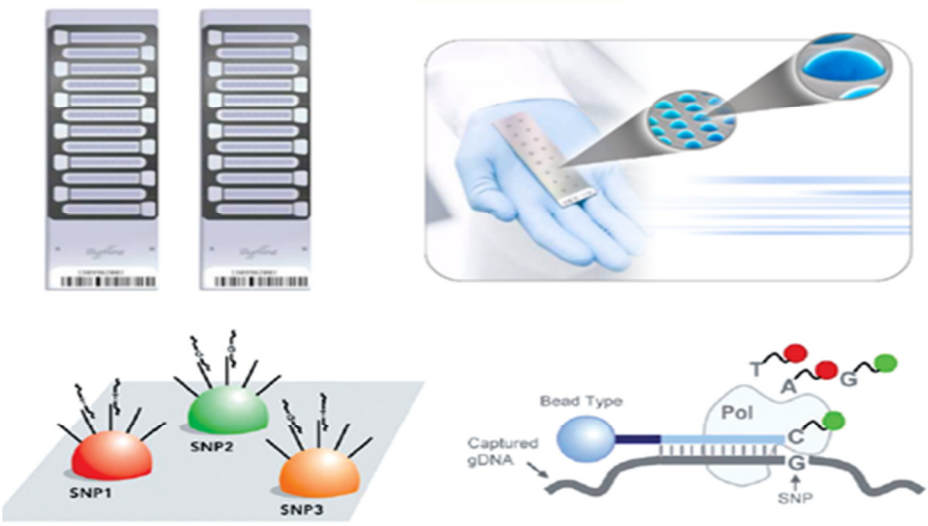

Figura 8. Diagrama esquemático de un SNPchip. La estructura física y la reacción química general que se efectúa en el chip.

Actualmente, en el mercado existen diferentes versiones de SNPchips para genotipar ganado bovino, entre las cuales encontramos los de la empresa Illumina, que capturan 777962, 54609 y 6909 SNPs a lo largo de los 30 cromosomas del genoma bovino. En especial el de 50000 SNPs (Bovine SNP50) fue utilizado para caracterizar la estructura genética del genoma bovino [35], y es actualmente el estándar para realizar selección genómica, debido a su costo y a que nos permite realizar estudios de asociación de genoma completo [36].

El desarrollo de las TGSM y el aumento de la capacidad de los SNPchips han permitido la implementación de Selección Genómica dentro de los programas de mejora genética [19,21]. Con ésta, es posible obtener el valor genómico de cada animal con respecto a la característica de importancia que deseamos mejorar en nuestro hato (ej: producción de leche, taza de preñez, peso al nacer, susceptibilidad a cierto padecimiento, etc.), y seleccionar ganado genéticamente óptimo para la producción, garantizando la transmisión genética adecuada de acuerdo a las características propias de medio ambiente y recursos de las regiones donde se sitúan los hatos lecheros. Los países desarrollados han adoptado rápidamente esta tecnología, y proyectan grandes avances en sus programas de mejora genética [21].

Un aspecto fundamental es revisar los costos de la tecnología; el SNPchip BovineHD captura 777962 SNP, muestrea 8 animales y tiene un costo, hasta el 2009, de 250€ por animal, mientras que el chip BovineSNP50 captura 54609 SNP, muestrea 24 animales y el costo por animal es de 164€. El BovineLD captura 6909 polimorfismos, muestrea 24 animales el costo por animal es de 80€ [37]. En la tabla 1 se observa la

descripción de los SNPchips, especializados en ganado Bovino, de la empresa Illumina Inc. Es importante tomar en cuenta que los costos que se muestran en la tabla son solo los costos del chip. Para estimar el costo total de implementar esta tecnología es necesario adherir el costo que involucra obtener la muestra de cada animal en el campo, el transporte de la muestra a un laboratorio de biología molecular, la extracción y purificación del ADN genómico, el costo por genotipar y el costo del análisis de la información.

Tabla 1. SNPChips comercializados por Illumina In, para muestrear al ganado bovino.

SNPchip	**Cantidad de SNPs**	**Costo**	**Animales por chip**	**DNA requerida**	**Nombre del Ensayo**	**Instrumento**
BovineHD	777962	2,000 €	8	200 ng	Infinitum HD	iScan o HiScanSQ
SNP50	54609	3,936 €	24	200 ng	GoldenGate	iScan o HiScanSQ
BovineLD	6909	1,920 €	24	200 ng	Infinitum HD	iScan o HiScanSQ

Sería de gran impacto adoptar en la región de Baja California la tecnología moderna (Selección Genómica), para garantizar el progreso de la industria pecuaria, en general, y lechera en particular. Sin embargo, el factor más importante que limita la trasferencia de esta tecnología es el factor económico; debido a que la situación actual de los ganaderos productores de leche es complicada. Por ejemplo: para un hato con 200

animales, el costo para genotipar con el SNP50, que es el estándar para implementar selección genómica seria de ~35,424 €, solo en costo de los chips. Esta cantidad no puede ser soportada por los ganaderos de la región.

El problema del costo de la tecnología no es solo de la región de Baja California, igualmente lo es para el resto de los ganaderos del mundo. De esto que ha sido necesario trabajar en estrategias de optimización de costos para poder llevar esta tecnología a ser implementada de forma masiva y con costos soportables por los ganaderos. Es en este rubro de optimización donde está centrado este trabajo de tesis.

Una alternativa recomendada para bajar los costos es realizar el genotipado solo de los sementales y las vacas elite (vacas con producción alta, y/o con hijas productivas y con características de salud y conformación deseables). Así puede implementarse una estrategia de reproducción para las vaquillas de reemplazo, utilizando el valor genómico del ganado elite del hato. Una ventaja que hará que los costos disminuyan es la próxima inclusión (a partir del 2013) del valor genómico, basado en genotipos, de todo el semen que se vende de forma internacional. Sin embargo, aun con estas estrategias, los ganaderos de la región de Baja California aún no están en posición económica de invertir en un proyecto de Selección Genómica, a menos que los costos sean reducidos aún más.

Con el objetivo principal de bajar los costos de la tecnología de selección genómica, se han desarrollado los métodos de *imputación*; con los cuales se infieren los valores de genotipos y haplotipos de SNPs que no han sido muestreados en ensayo de genotipado [38]. Esto ofrece la opción de utilizar paneles de baja densidad (ej. BovineLD, que

muestrea 6k), que son los más económicos, e imputar los genotipos de paneles de alta densidad (ej. Bovine SNP50, que muestrea 50K).

En la siguiente sección se describen algunos métodos de imputación.

1.4 Imputación de genotipos

Imputación de genotipos es el proceso de inferir o predecir genotipos que no han sido directamente capturados en un ensayo de genotipado [39]. La imputación puede ser usada para diferentes propósitos. Dos de las aplicaciones más utilizadas son:

1. Inferencia de genotipos faltantes en un panel, al fusionarlo con otro panel que ha sido genotipado en un ensayo distinto.
2. Inferencia de genotipos faltantes en un panel, utilizando un panel completo de referencia de haplotipos de mayor densidad; muestreado en individuos distintos de la misma raza.

En la figura 9 se muestra la fusión de información de diferentes fuentes para estudios combinados. La meta es asignar genotipos en dos muestras que han sido genotipadas en dos ensayos distintos [40].

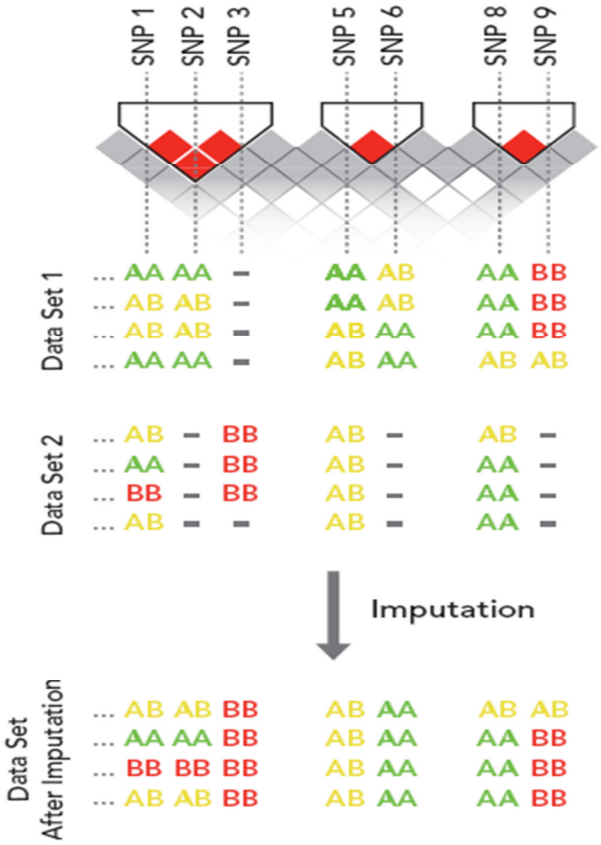

Figura 9. Imputación fusionando dos conjuntos de genotipos. La imputación provee un valor de probabilidad para cada uno de los tres posibles genotipos (AA, Aa y aa) y decide sobre el genotipo más probable.

Para la imputación basada en paneles de referencia de haplotipos, el razonamiento biológico es que los SNPs pueden ser modelados como mosaicos de haplotipos que contendrán valores iguales debido a que son idénticos por descendencia. Se infieren los haplotipos de los genotipos existentes en la muestra, y se busca la mejor alineación con el panel de referencia. Los genotipos faltantes son imputados, basados en los haplotipos con mejor alineación [41] . La figura 10 muestra un ejemplo de este tipo de imputación.

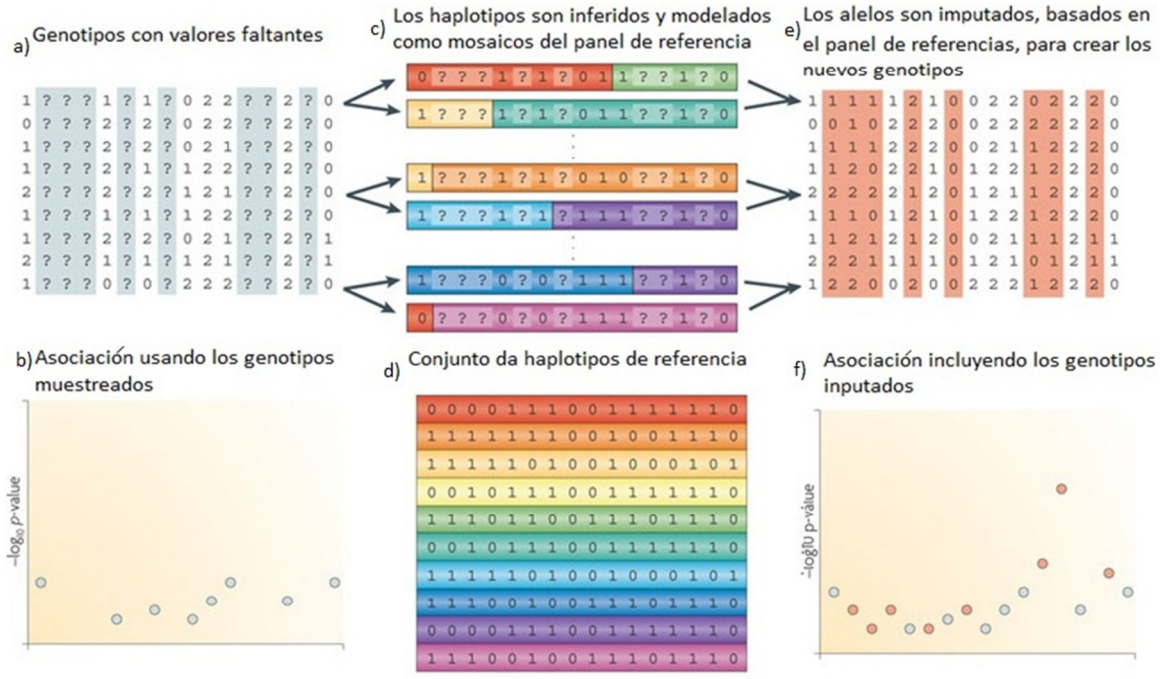

Figura 10. Imputación por paneles de referencia de haplotipos. Se obtienen nuevos genotipos y una mejor asociación incluyendo los genotipos imputados.

Una definición formal del problema de imputación se expresa como sigue:

Asumiendo que se cuenta con los datos de *L* SNPs autosomales bi-alelicos, y que los dos alelos en cada SNP han sido codificados con 0 y 1:

- Sea *H* un conjunto de haplotipos en *L* SNPs, de un conjunto de individuos de cierta raza.

- Sea *G* el conjunto de genotipos de los *L* SNPs en *K* individuos de la misma raza, distintos a los individuos del conjunto anterior, donde

- $G_i = (G_{i1}, \ldots, G_{iL})$ denota los genotipos del i_{th} individuo.

- Los genotipos individuales son ya sea observados, $G_{ik} \in$ existente, o no observados, G_{ik} = ? = faltante.

- El objetivo es predecir, utilizando el grupo completo de haplotipos del primer conjunto, los genotipos de aquellos SNPs que no han sido genotipados en el segundo conjunto.

Existen distintos algoritmos de imputación, [42]. Cada algoritmo tiene diferentes ventajas y desventajas, por lo que se recomienda usar más de uno y comparar resultados antes de reportar los resultados. A continuación se describen los dos algoritmos más utilizados en la actualidad.

Tabla 2. Lista de algoritmos que pueden ser utilizados para la imputación de genotipos, junto con las versiones y los sitios de los que los puedes obtener. Solo los dos primeros fueron considerados para este proyecto.

Software	**Versión**	**URL**	**Ref.**
BEAGLE	3.0.2	http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html	6.7
IMPUTE	0.5.0	http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html	1
MACH	1.10.16	http://www.sph.umich.edu/csg/abecasis/MACH/	8
PLINK	1.05	http://pngu.mgh.harvard.edu/~purcell/plink/	9
BIMBAM	0.99	http://stephenslab.uchicago.edu/software.html	2
FAMHAP	18	http://famhap.meb.uni-bonn.de/	10

1.4.1 El algoritmo Beagle

Es un modelo empírico debido a que no tiene parámetros que deban ser estimados. Está basado en el modelo gráfico de un conjunto de haplotipos, y trabaja de forma iterativa ajustando el modelo al conjunto actual de haplotipos estimados, y después reestimando los haplotipos de cada individuo, dado el modelo ajustado [41,43].

El método es ejecutado en dos pasos:

1. Un árbol bifurcante que describe a un conjunto de haplotipos, es construido, de izquierda a derecha en el conjunto de haplotipos. Una vez construido, cada tallo es asignado con un peso dado por el número de haplotipos que pasan por el.
2. El árbol es refinado para producir una caracterización más parsimoniosa de los datos.

La probabilidad de los genotipos faltantes es calculada del modelo ajustado en la última iteración.

1.4.2 El algoritmo Impute

El algoritmo *Impute* está basado en el modelado del vector de genotipos G_i, de cada individuo, utilizando un Modelo Oculto de Markov (HMM) condicional en H y un conjunto de parámetros; para inferir los haplotipos de SNPs no muestreados en el ensayo. El modelo puede ser escrito como sigue:

$$P\left(G_i \middle| H, \theta, \rho\right) = \sum_{z} P\left(G_i \middle| Z, \theta\right), P\left(Z \middle| H, \rho\right)$$

Dónde:

$Z = \{Z_1, \ldots, Z_L\}$, con $Z_j = \{Z_{j1}, Z_{j2}\}$ y $Z_{jk} = \{1, \ldots, N\}$. Z representa el par de haplotipos, para el SNP j, proveniente del panel de referencia, que están siendo copiados para formar el vector de genotipos.

$P(Z|H,\rho)$, está definido por una cadena de Markov y modela como el par de haplotipos copiados del panel de referencia cambia a lo largo de la secuencia debido a un mapa de recombinación (ρ) definido a lo largo de todo el genoma.

$P(G|Z,\theta)$, genera la variación de los vectores de genotipos observados con respecto a los haplotipos copiados del panel de referencia a través de una tasa de mutación(θ) [41].

1.5 Planteamiento del problema de tesis

Dada la situación actual de crisis del sector lechero en la región de Baja California, investigadores de la UABC, en coordinación con ganaderos de la región han acordado generar estrategias de apoyo para la solución al problema. Entre las metas establecidas se ha propuesto diseñar un plan maestro de mejora genética basado en tecnología de Selección Genómica, con el cual se generarán nuevos linajes de ganado genéticamente más apto y productivo para las condiciones de la región (Linajes que hemos denominado "Ganado Cachanilla o Vaca Cachanilla"). Como primer paso se acordó implementar un procedimiento piloto en un hato lechero de ~100 cabezas (hato perteneciente al Instituto de Investigación en Ciencia Veterinarias de la UABC) para hacer la caracterización inicial a nivel de salud animal, de variables productivas y de genética del ganado, con el

fin de implementar un plan de mejora genética utilizando selección basada en marcadores genéticos en el proceso de reproducción. Esto permitirá seleccionar ganado genéticamente óptimo para la producción de leche, garantizando la transmisión genética adecuada de acuerdo a las características propias de medio ambiente y recursos de la región.

Un factor crucial para que este proyecto de mejora genética pueda ser implementado de forma masiva en los hatos de la región de Baja California, será la adecuada optimización de los recursos. En primera instancia, es necesario hacer una elección precisa de la tecnología a ser implementada, la cual minimice costos y ofrezca una resolución adecuada los resultados. La opción que se ha tomado es realizar genotipado utilizando el SNPchip BovineLD, que muestrea ~6000 SNPs a lo largo del genoma bovino, por lo que será necesario realizar imputación de genotipos para obtener la resolución de 50,000 SNPs necesario para realizar Selección Genómica. De esto que es necesario realizar un estudio para encontrar cual es el mejor método de imputación para las razas que se trabajan en la región de Baja California.

En la siguiente sección se presentan los objetivos general y específicos de este proyecto de tesis.

1.6 Objetivos del trabajo de tesis

1.6.1 Objetivo general

El objetivo general de este trabajo de tesis es encontrar el algoritmo más adecuado para realizar imputación de un panel de baja densidad (~6,000 SNPs) a un panel de mayor densidad (~50,000 SNPs) en genotipos de ganado bovino lechero típico de la región de Baja California. Y, habilitar el uso de Selección Genómica para la implementación de estrategias de mejora genética.

1.6.2 Objetivos específicos

- Obtener genotipos utilizando los paneles BovineLD y Bovine SNP50 en las razas Holstein, Jersey, Brown Swiss y Hereford, que son las más utilizadas en la región.
- Analizar la exactitud de imputación de distintos algoritmos.
- Encontrar el algoritmo de imputación más adecuado, para realizar inferencia de un panel de BovineLD 6K a uno de Bovine SNP50 50k, para el ganado típico de la región de Baja California.

II IMPUTACIÓN DE GENOTIPOS EN GANADO BOVINO LECHERO

2.1 Descripción de los datos

Para realizar un experimento de imputación es necesario genotipar una muestra de animales. El experimento se realizó de la siguiente forma: primeramente se seleccionaron tres de las razas de ganado lechero y una de ganado de engorda, más utilizadas en la región y en el país. Debido a que no se cuenta con genotipos de ganado local, se eligieron animales, de las razas seleccionadas, pertenecientes al Consorcio Internacional para la Generación de un Mapa de Haplotipos del Genoma Bovino [Bovine HAPMAP, 2009]. Es importante mencionar que los animales genotipados son de Estados Unidos de Norte América y Canadá, y principalmente el ganado Holstein que se tiene en México, es descendiente de este ganado; por lo que la genética se consideró equivalente para este ejercicio. Igualmente es importante mencionar que puede existir diferencia en la genética del ganado local debido a su adaptación a las condiciones geográficas y climatológicas de la región; hecho que será corroborado cuando se cuente con genotipos del ganado local.

La muestra en los animales, para extracción de ADN, fue de sangre en las vacas y semen en los toros. Las muestras fueron genotipadas con los paneles Bovine SNP50 y BovineHD. Los genotipos fueron pre-procesados para filtrar inconsistencias y errores de genotipado. En la raza de engorda se realizó imputación utilizado los genotipos del panel BovineHD, y en las razas lecheras se realizó imputación de un subconjunto de ~6,000 SNPs a los ~50,000 SNPs del panel Bovine SNP50, usando los dos algoritmos más referenciados en la literatura. Los resultados fueron analizados y se definió el algoritmo más adecuado para realizar imputación en ganado de la región, dentro del proyecto de la Vaca Cachanilla. En las siguientes secciones se describen los detalles de los experimentos.

2.1.1 Razas y número de animales seleccionados.

Las razas seleccionadas fueron Holstein, Jersey, Brown Swiss y Hereford. El número de animales por raza se muestra en la tabla 3.

Tabla 3. Número de animales por raza, en los experimentos.

RAZAS	ANIMALES
HOLSTEIN FRIESIAN	53
JERSEY	27
BROWN SWISS	24
HEREFORD	27
TOTAL	**131**

2.1.2 Cromosomas y número de Marcadores SNP utilizados en el experimento

El SNPchips Bovine SNP50 captura genotipos de 30 cromosomas del ganado bovino. Para este proyecto se consideraron solo los genotipos de los 29 cromosomas autosomales. En la tabla 4 se observa el número de SNPs que se muestreó en cada cromosoma, con

Bovine50k y el subconjunto de SNPs emulando un Bovine6K. En los anexos se observa el número de SNPs, después de haber realizado la imputación con el algoritmo Impute, en las razas Holstein, Jersey y Brown Swiss.

Tabla 4. Número de SNPs por cromosoma del panel de baja densidad (Bovine6k) y el panel de alta densidad (Bovine50k).

	Holstein			**Jersey**			**Brown Swiss**	
	Números de SNPS			Números de SNPS			Números de SNPS	
Cromosoma	**Bovine6k**	**Bovine50k**	**Cromosoma**	**Bovine6k**	**Bovine50k**	**Cromosoma**	**Bovine6k**	**Bovine50k**
1	400	1213	**1**	400	1003	**1**	400	947
2	349	1070	**2**	349	942	**2**	349	797
3	312	1027	**3**	312	789	**3**	312	715
4	307	972	**4**	307	820	**4**	307	688
5	307	955	**5**	307	803	**5**	307	635
6	311	1976	**6**	311	1566	**6**	311	1144
7	288	789	**7**	288	662	**7**	288	604
8	301	897	**8**	301	761	**8**	301	672
9	272	782	**9**	272	654	**9**	272	580
10	270	835	**10**	270	695	**10**	270	622
11	286	895	**11**	286	825	**11**	286	610
12	227	660	**12**	227	579	**12**	227	530
13	218	720	**13**	218	576	**13**	218	507
14	226	2195	**14**	226	1933	**14**	226	1819
15	225	630	**15**	225	539	**15**	225	441
16	209	615	**16**	209	534	**16**	209	422
17	195	636	**17**	195	542	**17**	195	464
18	178	536	**18**	178	460	**18**	178	383
19	183	527	**19**	183	477	**19**	183	349
20	207	629	**20**	207	496	**20**	207	459
21	186	504	**21**	186	429	**21**	186	361
22	169	518	**22**	169	474	**22**	169	369
23	152	473	**23**	152	397	**23**	152	356
24	178	516	**24**	178	462	**24**	178	347
25	142	1029	**25**	142	873	**25**	142	755
26	147	413	**26**	147	359	**26**	147	262
27	140	376	**27**	140	330	**27**	140	278
28	129	396	**28**	129	346	**28**	129	281
29	137	387	**29**	137	324	**29**	137	248

2.2 Preprocesamiento de los genotipos

Para garantizar la calidad global de las muestras y contar con un conjunto consistente de genotipos, se aplicaron filtros de control de calidad a los datos iniciales. Los filtros consistieron en la eliminación de todos los genotipos con >20% de valores faltantes, todos los SNPs que no cumplen con el principio de Hardy Weinberg, todos los animales con <98% de genotipos completos y todos los SNPs monomórficos para todas las razas. Igualmente se eliminaron los SNPs con una frecuencia de alelo menor (MAF) < 0.5 entre todas las razas [44].

2.3 Imputación en la raza Hereford

En el primer experimento se utilizó el panel BovineHD para realizar imputación en los cromosomas 1 y 6 de la raza Hereford. El cromosoma 1 fue seleccionado por ser el que contiene el mayor número de marcadores SNPs, y el cromosoma 6 fue seleccionado de forma aleatoria. El experimento consintió en lo siguiente:

De forma aleatoria, se removieron el 1% de los genotipos, asignando estos como faltantes, y se realizó su imputación utilizando el algoritmo Beagle. Este proceso fue repetido incrementando el porciento de valores faltantes del 1% hasta el 6%. El valor máximo de 6% en valores faltantes fue establecido debido a una restricción del algoritmo Beagle. Los resultados obtenidos en el cromosoma 1 se muestran en la tabla 5. Los resultados muestran que los valores imputados aumentaron su exactitud del 98.31% al 98.45 % a medida que el porciento de valores a imputar aumento del 1% al 6%, respectivamente.

Tabla 5. Resultados de imputación en el cromosoma 1 con el algoritmo Beagle, incrementando el porciento de valores a imputar del 1% al 6%.

	1%	2%	3%	4%	5%	6%
Alelos Faltantes	25106	50212	75318	100424	125531	150637
% Exactitud	**98.31**	**98.37**	**98.37**	**98.41**	**98.41**	**98.45**

Se repitió el experimento anterior para el cromosoma 6, observando que la exactitud varió del 98.33% al 98.40% cuando los valores a imputar variaron del 1% al 6%, respetivamente. Los resultados se muestran en la tabla 6.

Tabla 6. Resultados de imputación en el cromosoma 6 con el algoritmo Beagle, incrementando el porciento de valores a imputar del 1% al 6%.

	1%	2%	3%	4%	5%	6%
Alelos Faltantes	19178	38357	57535	76714	95893	115071
% Exactitud	**98.33**	**98.34**	**98.35**	**98.38**	**98.40**	**98.40**

Este experimento mostró que el algoritmo Beagle puede imputar un máximo de 6% de genotipos faltantes. Hecho que limita los alcances del proyecto de la Vaca Cachanilla, en el cual se pretende realizar genotipado con un panel de ~6,000 SNPs y realizar imputación a un panel de ~50,000 SNPs para hacer estimaciones del Valor Genómico de los animales. Debido a esto, los siguientes experimentos consistieron en realizar imputación utilizando el algoritmo Impute, el cual implementa la estrategia de usar un panel de referencia de mayor densidad y hace imputación a grupos de genotipos de menor densidad.

2.4 Imputación en la raza Holstein

El segundo experimento consistió en realizar imputación en la raza Holstein. De esta raza se contó con 53 animales, de los cuales se eligieron 33 animales para generar el panel de referencia de genotipos, y los 20 animales restantes se tomaron para generar el panel a inferir. Los genotipos se tomaron del panel BovineSNP50, de los cuales se infirieron haplotipos utilizando el algoritmo fastPHASE[44]. Se optó por generar un panel de referencia mayor, en el número de animales, que el panel a inferir, debido a que así lo específica el algoritmo.

El SNPchip Bovine6k muestrea 400 snps en el cromosoma 1, mientras el BovineSNP50 muestrea 1213, como se muestra en la sección (2.1.2). Dentro del panel de 50k, se seleccionaron de forma homogénea los 400 SNPs, los cuales se dejaron como genotipos fijos. Esto para emular que esos genotipos son del chip Bovine6k. Posteriormente, al resto del panel 50k, que son 813 SNPs, se les asignó como alelos faltantes, esto se tomó como el 100% de genotipos faltantes. Se realizó la imputación utilizando el algoritmo Impute, iniciando con el 100% de genotipos faltantes, y se fue disminuyendo de forma gradual a 80%, 60%, 40% y 20%. Se utilizó esta estrategia con el fin de estandarizar esta medida, y así poder comparar los resultados obtenidos con los de otras razas, cuyo número de animales es distinto.

Los resultados obtenidos en la imputación, se muestran en la tabla 7 y en la figura 11. Los resultados muestran que los valores imputados aumentaron su exactitud del 80.58% al 93.52 % a medida que el porciento de valores a imputar fue disminuyendo del 100% al 20%, respectivamente.

Tabla 7. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Holstein.

Valores para imputar	**20%**	**40%**	**60%**	**80%**	**100%**
Alelos Faltantes	163	325	488	650	813
% Exactitud	**93.52**	**90.50**	**86.86**	**83.97**	**80.58**

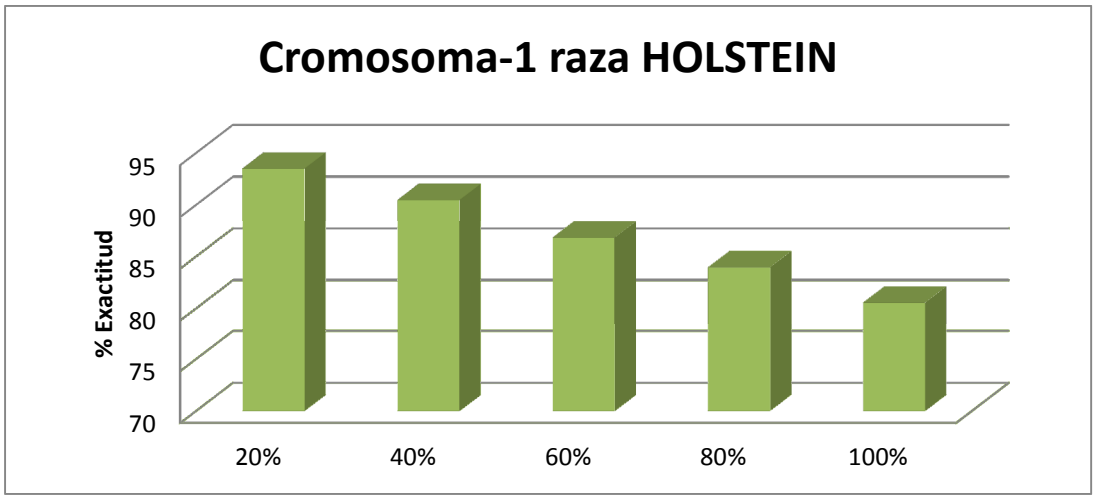

Figura 11. Exactitud de imputación en el cromosoma 1 de la raza Holstein. En la gráfica se observan cómo descendió del 20% al 100%, de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20%.

Este paso se repitió para los 29 cromosomas autosomales. La tabla 8 y figura 12 muestran las exactitudes globales del genoma completo, las tablas de las exactitudes de cada uno de los cromosoma se muestran en los anexos.

Tabla 8. Suma global de exactitudes en el genoma completo de la raza Holstein.

	20%	40%	60%	80%	100%
Suma Global de Exactitud	2701.02	2606.37	2508.61	2416.46	2322.34
% Global de Exactitud	**93.14**	**89.87**	**86.50**	**83.33**	**80.08**

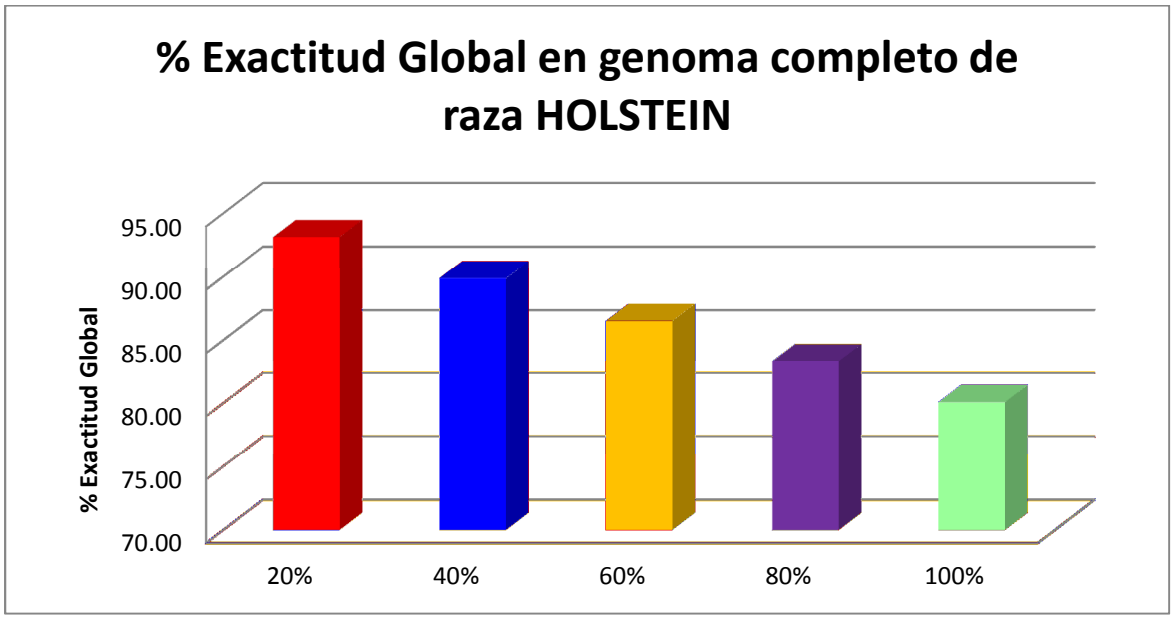

Figura 12. Gráfica de la exactitud global en el genoma completo de la raza Holstein.

2.5 Imputación en la raza Jersey

El Tercer experimento consistió en realizar imputación en la raza Jersey. De esta raza se contó con 27 animales, de los cuales se eligieron 17 animales para generar el panel de referencia de genotipos, y los 10 animales restantes se tomaron para generar el panel a inferir. Se generó el panel de inferencia y se tomaron genotipos y haplotipos del panel Bovine SNP50.

El SNPchip Bovine6k muestrea en 400 snps en el cromosoma 1, mientras el BovineSNP50 muestrea 1003, como se muestra en la sección (2.1.2). Al igual que el experimento anterior, dentro del panel de 50k, se seleccionaron de forma homogénea los 400 SNPs, los cuales se dejaron como genotipos fijos. Posteriormente, **a**l resto del panel 50k, que son 603 SNPs, se asignó como alelos faltantes. Esto se tomó como el 100% de genotipos faltantes y se realizó la imputación utilizando el algoritmo Impute. Se realizó igualmente la imputación iniciando con el 100% de genotipos faltantes, y se fue disminuyendo de forma gradual hasta 20%, igual que en el experimento anterior.

Los resultados obtenidos en la imputación se muestran en la tabla 9. Los resultados muestran que los valores imputados aumentaron su exactitud del 93.74% al 82.44 % a medida que el porciento de valores a imputar fue disminuyendo del 100% al 20%, respectivamente.

Tabla 9. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Jersey.

Valores para imputar	**20%**	**40%**	**60%**	**80%**	**100%**
Alelos Faltantes	121	241	362	482	603
% Exactitud	**93.74**	**91.11**	**88.00**	**85.14**	**82.44**

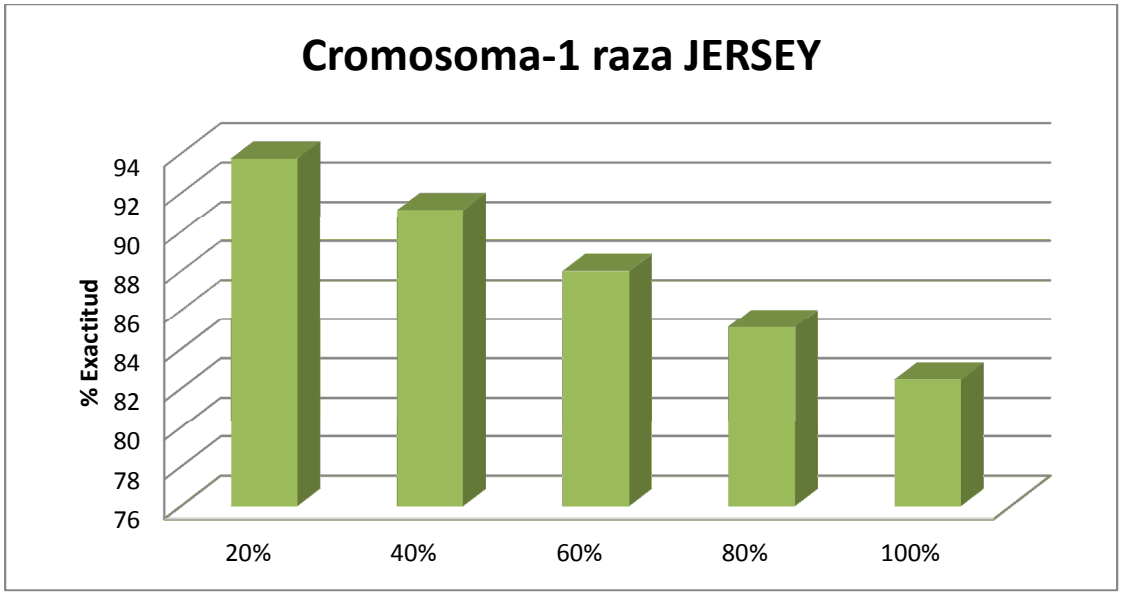

Figura 13. Exactitud de imputación en el cromosoma 1 de la raza Jersey. En la gráfica se observan cómo descendió del 20% al 100%, de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20%.

Este paso se repitió para los 29 cromosomas autosomales, las tablas de las exactitudes obtenidas en el genoma completo, se muestran en los anexos. La tabla 10 y gráfica 14 muestra las exactitudes globales para todo el genoma.

Tabla 10. Suma global de exactitudes en el genoma completo de la raza Jersey.

	20%	40%	60%	80%	100%
Suma Global de Exactitud	2673.72	2593.59	2509.80	2431.32	2351.57
% Global de Exactitud	**92.20**	**89.43**	**86.54**	**83.84**	**81.09**

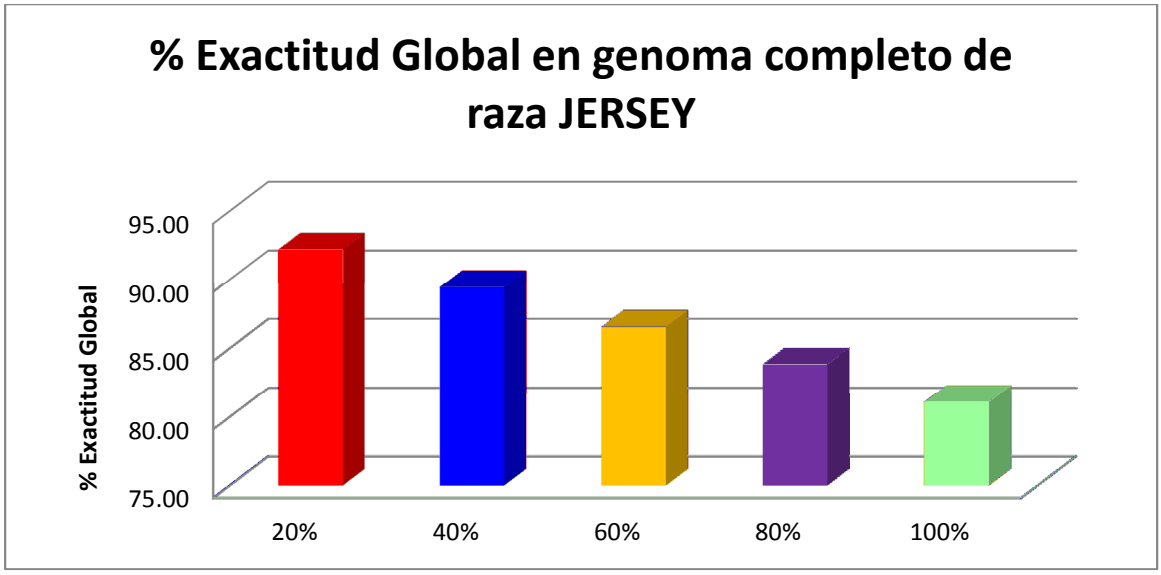

Figura 14. Gráfica de la exactitud global en el genoma completo de la raza Jersey.

2.6 Imputación en la raza Brown Swiss

El Cuarto experimento consistió en realizar imputación en la raza Brown Swiss. De esta raza se contó con 24 animales, de los cuales se eligieron 14 animales para generar el panel de referencia de genotipos y los 10 animales restantes se tomaron para generar el panel a inferir. Al igual que los experimentos anteriores, los genotipos se obtuvieron del panel Bovine SNP50, y los haplotipos fueron inferidos con el algoritmo fastPHASE.

El SNPchip Bovine6k muestrea en 400 snps en el cromosoma 1, mientras el BovineSNP50 muestrea 947, como se muestra en la sección (2.1.2). Del panel de 50k, se seleccionaron de forma homogéneamente los 400 SNPs, los cuales se dejaron como genotipos fijos, y se repitió el procedimiento de los experimentos anteriores.

Los resultados obtenidos en la imputación, se muestran en la tabla 11. Los resultados muestran que los valores imputados aumentaron su exactitud del 94.12% al

82.47 % a medida que el porciento de valores a imputar fue disminuyendo del 100% al 20%, respectivamente.

Tabla 11. Resultados de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20% en la raza Brown Swiss.

Valores para imputar	**20%**	**40%**	**60%**	**80%**	**100%**
Alelos Faltantes	109	219	328	438	547
% Exactitud	**94.12**	**91.08**	**88.18**	**85.26**	**82.47**

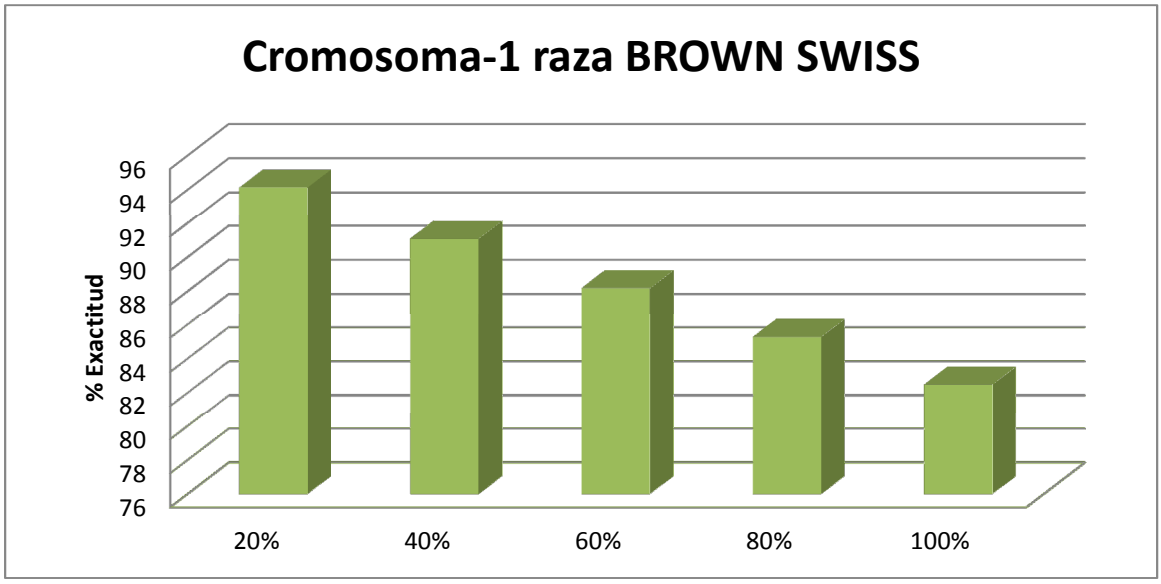

Figura 15. Exactitud de imputación en el cromosoma 1 de la raza Brown Swiss. En la gráfica se observan cómo descendió del 20% al 100%, de imputación en el cromosoma 1 con el algoritmo Impute, disminuyendo el porciento de valores a imputar del 100% al 20%.

Este paso se repitió para los 29 cromosomas autosomales las tablas, se muestran en los anexos.

Aplicando el algoritmo Impute obtuvimos las siguientes exactitudes de los 29 cromosoma autosomales en la tabla 12 se observa las exactitudes globales, y en la gráfica 16 se aprecia dichas exactitudes.

Tabla 12. Suma global de exactitudes en el genoma completo de la raza Brown Swiss.

	20%	40%	60%	80%	100%
Suma Global de Exactitud	2726.70	2645.75	2571.88	2493.58	2418.58
% Global de Exactitud	**94.02**	**91.23**	**88.69**	**85.99**	**83.40**

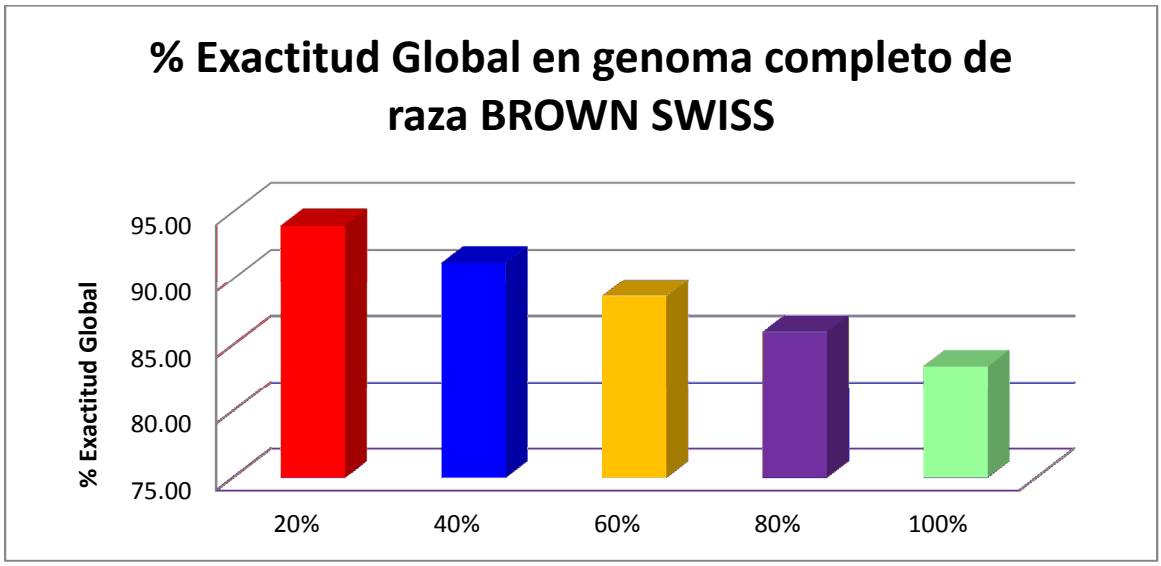

Figura 16. Gráfica de la exactitud global en el genoma completo de la raza Brown Swiss.

En el siguiente capítulo se realiza el análisis de la discusión de los resultados obtenidos con estos experimentos.

III ANÁLISIS Y DISCUSIÓN DE RESULTADOS

3.1 Análisis de la imputación con el algoritmo Beagle.

Para la imputación con el algoritmo Beagle en la raza Hereford, es interesante observar que la exactitud de imputación aumentó (del 98.31% al 98.45% en el cromosoma 1, y del 98.33% al 98.40% en el cromosoma 6) a medida que el porcentaje de valores a imputar fue incrementándose. Estas precisiones son arriba del 98%, lo que indica muy buena imputación, sin embargo este algoritmo, en su implementación actual solo puede imputar un máximo del 6% de valores faltantes. Esta es una limitación para ser usado en el proyecto de la Vaca Cachanilla, en el cual se pretende hacer imputación de un panel ~6,000 SNPs a uno de ~50,000 SNPs, para hacer estimación de los valores genómicos de los animales. Este algoritmo es recomendable solo cuando el porcentaje de valores a imputar es pequeño, lo cual garantiza una precisión arriba del 98%. Sin embargo, no ofrece los rangos de imputación que se requieren para el proyecto de la Vaca Cachanilla.

De esto que, en el resto de los experimentos se utilizó el algoritmo Impute, cuyos resultados se discuten a continuación.

3.2 Análisis de imputación con el algoritmo Impute

El algoritmo Impute fue usado para imputar haplotipos en las razas Holstein, Jersey y Brown Swiss. Este algoritmo no impuso ninguna restricción en el número de genotipos faltantes. La figura 17 muestra las exactitudes de imputación en el genoma completo para cada una las tres razas.

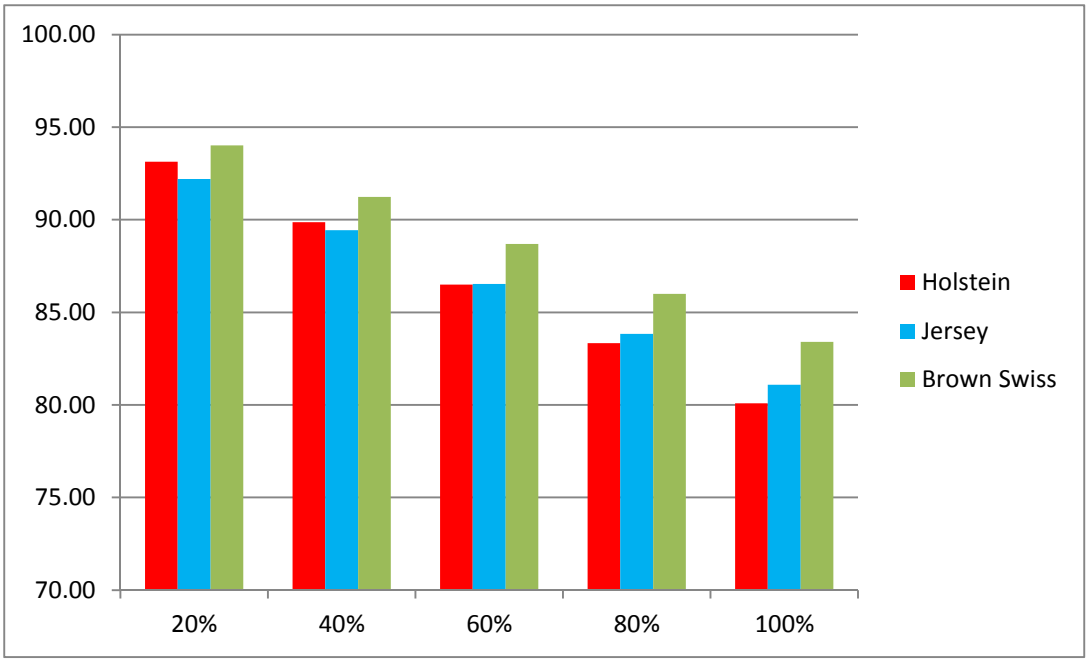

Figura 17. Gráfica global de exactitudes de las tres razas del ganado lechero, Holstein Jersey y Brown Swiss.

La raza con mayor exactitud fue la Brown Swiss, para la cual se obtuvieron exactitudes de imputación de 94.02% y de 83.40% cuando se asignó el 20% y el 100%

de valores faltantes, respectivamente. En el caso de las razas Holstein y Jersey, la exactitud de imputación tuvo un comportamiento interesante; Holstein resultó con exactitudes de imputaciones mayores, comparado con Jersey, cuando se asignaron proporciones pequeñas de valores faltantes. Como se observa en la figura 17, en Holstein se obtuvo una exactitud de 93.14% y de 89.87% cuando de asignaron el 20% y el 40% de valores faltantes, respectivamente, mientras que para Jersey se obtuvo 92.20% y 89.43% de exactitud para los mismos porcentajes de valores faltantes. Por otro lado, Jersey resultó con exactitudes mayores a las de Holstein cuando se asignaron proporciones grandes de valores faltantes. Para Jersey se obtuvo una exactitud de 83.84% y de 81.09.% cuando de asignaron el 80% y el 100% de valores faltantes, respectivamente, mientras que para Holstein se obtuvo 83.33% y 80.08% de exactitud para los mismos porcentajes de valores faltantes. Curiosamente, la exactitud de imputación es muy similar (86.50% vs 86.54%) para proporciones intermedias (60%) de valores faltantes.

3.3 Análisis del impacto de la imputación en la estimación de los GEBVs.

Un aspecto muy importante es el efecto que tendrá la exactitud de la información imputada en la estimación de los Valores Genómicos de los animales. En los experimentos realizados, la exactitud de las imputaciones varió con respecto al porcentaje de valores a imputar. Una pregunta que surge es si la exactitud de los valores genómicos, estimados con la información imputada, se verá reducida. Y si la variación de la exactitud de imputación hará que el valor genómico de los animales varié. Este hecho puede ser

comprobado de forma experimental si se estiman los valores genómicos de los animales, pero el realizar las estimaciones de los valores genómicos involucra la incorporación de historiales de producción, de pedigrí y de manejo, al igual que la aplicación de técnicas matemáticas y computacionales que no se abordan en este trabajo de tesis. Sin embargo, investigaciones previas han demostrado, con experimentos similares, que la exactitud de los valores genómicos de los animales no varía más del 5%, cuando estos son estimados de valores imputados. Habier et al. [45], demostraron, mediante simulación, que la precisión de los valores genómicos obtenidos de información imputada de un chip de baja densidad a un chip de alta densidad sufre una reducción promedio de 4.5%. En otro experimento, con datos reales, Weigel et al. [46], investigaron el efecto de la imputación de un panel de 3,000 SNPs (Bovine3K) a un panel de 50,000 SNPs (Bovine SNP50), en la estimación del valor genómico de animales de raza Jersey, encontrando que los valores genómicos son estimados con una exactitud del 95%. Los resultados encontrados tanto con simulación como con datos reales confirman una precisión del ~95% en la estimación de los valores genómicos. Estos resultados son una referencia confiable de que los valores genómicos de los animales de los experimentos realizados en este trabajo de tesis, al igual que los animales que serán utilizado en el proyecto de la Vaca Cachanilla tendrán valores genómicos con no menos del 95% de precisión, estimados de valores imputados del panel BovineLD al panel BovineSNP50.

3.4 Análisis de optimización de costos por imputación.

Sin duda la imputación es un recurso de optimización muy importante para poder aplicar la tecnología de Selección Genómica de forma masiva, llevándola a los hatos ganaderos de la región. El aspecto más importante para los ganaderos es el costo-beneficio de la tecnología, buscando siempre que el beneficio sea el mayor posible, al menor costo. Como se analizó en la sección anterior, genotipar con el SNPchip BovineLD, que es el más económico, y realizar imputación de los haplotipos para obtener los valores del Bovine SNP50, y estimar con estos los valores genómicos de los animales, es la mejor estrategia de optimización. La tabla 13 muestra los costos del conjunto de SNPchips que deben ser utilizados para genotipar hatos ganaderos con 50, 100, 150 y 200 animales.

Tabla 13. Costo del conjunto de SNPchips necesarios para genotipar distinto número de animales

No. animales	BovineHD	BovienSNP50	BovineLD
50	14,000 €	7,872 €	3,840 €
100	26,000 €	19,680 €	9,600 €
150	38,000 €	27,552 €	13,440 €
200	50,000 €	35,424 €	17,280 €

Como se observa en la tabla 13, si se realiza genotipado de 150 animales, por ejemplo, utilizando chips BovineSNP50, el costo de los chips sería de 27,552 Euros., con la ventaja de obtener el 100% de los genotipos reales de los animales. Pero, si se realiza el genotipado utilizando los chips BovineLD, que son más económicos, el costo de los chips sería de 13,440 Euros, con la premisa de realizar imputación para estimar los genotipos no capturado del BovienSNP50. El hecho de genotipar con el chip de baja

densidad y a través de imputación estimar el resto de la información, genera una reducción del costo en chips de 14,112 Euros, que equivales a una reducción del 51.2%.

Ahora, como se mencionó en la sección 1.3.1, es recomendable genotipar solo el ganado elite del hato, esto es: los sementales y las vacas más productivas y con mejores récords de salud y conformación, por lo que haciendo una buena selección, puede reducirse aún más los costos.

IV CONCLUSIONES Y TRABAJO FUTURO

En este proyecto de tesis de maestría, se propuso como objetivo general encontrar el algoritmo más adecuado para realizar imputación de un panel de baja densidad (~6,000 SNPs) a un panel de mayor densidad (~50,000 SNPs) en genotipos de ganado bovino lechero, típico de la región de Baja California. Y, con esto habilitar el uso de Selección Genómica para la implementación de estrategias de mejora genética. Para lograr este objetivo, se seleccionaron las razas Holstein, Jersey y Brown Swiss de ganado lechero y la raza Hereford de ganado de engorda, para realizar experimentos. Posteriormente se obtuvieron genotipos de grupos de animales de dichas razas con el panel BovineSNP50, de los cuales se extrajo los genotipos de un subconjunto de ~6,000 SNPs para emular el genotipado con el panel BovineLD. Se realizaron experimentos de imputación con dos algoritmos distintos y se documentaron los resultados. Se analizó el impacto de las exactitudes de imputación para hacer estimación del Valor Genómico de Crianza de los animales, basado en valores

imputados. Finalmente se analizó la reducción de costos en la etapa de medición genética y estimación de valores genómicos al utilizar imputación.

4.1 Conclusiones

De los dos algoritmos de imputación que se probaron, Beagle e Impute, Beagle tuvo mejor rendimiento, con exactitudes de imputación hasta de 98.40% en la raza Hereford, sin embargo tiene la restricción de imputar un máximo de 6% de la cantidad de genotipos existentes. Esto lo hace no adecuado dentro del proyecto de la Vaca Cachanilla, en el cual se requiere hacer imputación de un panel de ~6,000 SNPs a un panel de ~50,000 SNPs. Por otro lado, el algoritmo Impute tuvo un máximo de exactitud de imputación de 94.02% para la raza Brown Swiss, un máximo de 93.14% para la raza Holstein y un máximo de 92.20% para la raza Jersey. Una ventaja de este algoritmo es que no pone restricciones en la cantidad de genotipos y haplotipos a imputar. Y, con los porcentajes de exactitud obtenidos puede proyectarse que la precisión de la estimación de los valores genómicos será aproximada al 95%, la cual es suficiente para realizar selección genómica dentro de un programa de mejora genética.

La estrategia de realizar el genotipado con el panel BovineLD y realizar imputación para estimar el panel completo Bovine SNP50 para el proyecto de la Vaca Cachanilla, en el cual se genotiparan ~150 animales, ofrece ahorro en el costo de 13,440 Euros, que equivale al 51.2% del costo si se realiza el genotipado con el panel Bovine SNP50. Ahora, puede reducirse los costos aun más si se hace una selección adecuada del ganado elite del hato, lo que reduciría el número de animales a genotipar.

En resumen, el producto principal de este trabajo de tesis es la generación de la siguiente estrategia para optimizar costos en la implementación de programas de mejora genética basados en Selección Genómica, en los hatos de la región de Baja California:

1. Hacer una buena selección, en el hato, de aquellos animales con los mejores récords de producción, de salud y de conformación,
2. Los animales seleccionados conformaran el subgrupo elite que será utilizado para reproducción,
3. Realizar el genotipado del grupo elite utilizando el panel BovineLD, el cual captura ~6,000 SNPs a lo largo de todo el genoma bovino,
4. Realizar imputación del conjunto de ~6,000 SNPs provenientes del panel BovineLD, y obtener los valores de genotipos y/o haplotipos del panel Bovine SNP50, que captura ~50,000 SNPs, utilizando el algoritmo Impute.
5. Utilizar el panel Bovine SNP50 imputado, junto al récord de producción de los animales y su información de pedigrí, y estimar los valores Genómicos de Crianza para cada rasgo de interés económico que se desee mejorar en el hato,
6. Generar un ranking de los animales, por su Valor Genómico,
7. Seleccionar pares semental-vaca, con repetición en los sementales, cuidando no provocar un alto grado de consanguinidad en el hato, y generar las nuevas crías.

4.2 Trabajo futuro

Este proyecto se realizó como una investigación proyectiva para definir parte de la estrategia a utilizar para optimizar costos dentro del proyecto de la Vaca Cachanilla. Las

razas de ganando usadas en este proyecto son razas que se utilizan en la región, y en el país, sin embargo los animales muestreados son animales principalmente de Estados Unidos de Norteamérica. Esto se debió a que el Consorcio Internacional para la Generación del Mapa de Haplotipos del Genoma Bovino, ofreció la información de genotipos de los animales. Los resultados de imputación y su impacto en la estimación de los Valores Genómicos de los animales son utilizados como proyecciones cercanas a los valores reales de los animales de la región. El siguiente paso es realizar el genotipado de animales pertenecientes a los hatos de la región y ejecutar la estrategia de genotipado e imputación analizada en este trabajo. Comprobar de forma experimental el costo-beneficio de la selección genómica como método de mejora genética, y llevarla a todos los hatos de la región de Baja California. Para, con esto, contribuir en rescate de la industria lechera de la región.

BIBLIOGRAFÍA

1. CANILEC (2009) Bovino y sus derivados. Mexico DF.
2. USDA (2011) Dairy: World Markets and Trade.
3. SIAP (2012) Leche de bovino, produccion, precio y valor. In: SAGARPA, editor. Servicio de informacion agroalimentaria y pesquera. Mexico DF.
4. Vargas E (2011) Pierde B.C. 10 mil cabezas de ganado lechero. Noticias Ensenadanet. Ensenada.
5. Garcia RG (2011) En crisis la industria lechera de BC. La Voz de la Frontera. Mexicali.
6. Andrabi SMH, Moran C (2007) Selection of Dairy Cow Bulls for Artificial Insemination. INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY 9: 4.
7. Syrstad O, Ruane J (1998) Prospects and strategies for genetic improvement of the dairy potential of tropical cattle by selection. Trop Anim Health Prod 30: 257-268.
8. Newman S (2008) Genetics Research at Fort Keogh: A Historical Perspective. USDA.
9. Greiner SP (2009) Understanding Expected Progeny Differences (EPDs). Virginia Cooperative Extension: Virginia Tech. pp. 6.
10. Robinson GK (1991) That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science 6: 18.
11. Wilcox CJ, Webb DW (1992) Genetic Improvement of Dairy Cattle. EDIS Gainsville, University of Florida IFAS extension 11.
12. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423-447.
13. Long TE (2003) Estimating Genetic Merit. Swine Genetics NSIF, Purdu University: 4.
14. Smith C (1967) Improvement of metric traits through specific genetic loci. Animal Production 9.
15. Soller M, Beckmann JS (1983) Genetic polymorphism in variety identification and genetic improvement. Theoretical and Applied Genetics 67: 9.
16. Shrimptom AE, Robertson A (1988) The isolation of polygenic factors controlling bristlc score in Drosophila Melanoganster. Genetics 148: 14.
17. Hayes B, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol 33: 209-229.
18. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, et al. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40: 198-203.
19. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, et al. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92: 16-24.
20. Hayes BJ, Goddard ME (2008) Technical note: prediction of breeding values using marker-derived relationship matrices. J Anim Sci 86: 4.
21. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92: 433-443.
22. Seidenspinner T, Bennewitz J, Reinhardt F, Thaller G (2009) Need for sharp phenotypes in QTL detection for calving traits in dairy cattle. J Anim Breed Genet 126: 455-462.
23. Abdel-Azim G, Freeman AE (2002) Superiority of QTL-assisted selection in dairy cattle breeding schemes. J Dairy Sci 85: 1869-1880.
24. Heyen DW, Weller JI, Ron M, Band M, Beever JE, et al. (1999) A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiol Genomics 1: 165-175.

25. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.
26. Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, et al. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5: 247-252.
27. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 11.
28. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.
29. Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56: 303-320.
30. Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. Methods Mol Biol 513: 19-39.
31. Dearlove AM (2002) High throughput genotyping technologies. Brief Funct Genomic Proteomic 1: 139-150.
32. Brennan MD (2001) High throughput genotyping technologies for pharmacogenomics. Am J Pharmacogenomics 1: 295-302.
33. Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2: 41-49.
34. Rapley R, Harbron S (2004) Molecular Analysis and Genome Discovery; Chichester, editor: John Wiley & Sons Ltd.
35. Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, et al. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324: 528-532.
36. Illumina I (2010) BovineSNP50 Genotyping BeadChip. Illumina, Inc.
37. O. González Recio JAJ-MyRA La selección genómica aplicada a un programa de mejora en vacuno de leche. Frisona Española 177.
38. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387-406.
39. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-913.
40. Illumina I (2009) Imputation Estimates Genotypes at Un-genotyped Loci. Illumina, Inc.
41. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499-511.
42. Ellinghaus D, Schreiber S, Franke A, Nothnagel M (2009) Current software for genotype imputation. Hum Genomics 3: 371-380.
43. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210-223.
44. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.
45. Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182: 10.

46. Weigel KA, De los Campos G, Vazquez AI, Rosa GJM, Gianola D, et al. (2010) Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J Dairy Sci 93: 12.

ANEXOS

Número de SNPS por cromosoma del panel de baja densidad (Bovine6k) y el panel de alta densidad (Bovine50k). En las razas Holstein, Jersey Y Brown Swiss, después de haber utilizado el algoritmo Impute.

	Holstein			Jersey			Brown Swiss	
	Números	de SNPS		Números	de SNPS		Números	de SNPS
Cromosoma	Bovine6k	Bovine50k	Cromosoma	Bovine6k	Bovine50k	Cromosoma	Bovine6k	Bovine50k
1	385	1213	**1**	375	1003	**1**	388	947
2	342	1070	**2**	332	942	**2**	329	797
3	311	1027	**3**	295	789	**3**	304	715
4	304	972	**4**	292	820	**4**	276	688
5	305	955	**5**	278	803	**5**	285	635
6	305	1976	**6**	233	1566	**6**	294	1144
7	287	789	**7**	258	662	**7**	272	604
8	300	897	**8**	269	761	**8**	282	672
9	271	782	**9**	256	654	**9**	262	580
10	269	835	**10**	253	695	**10**	258	622
11	284	895	**11**	263	825	**11**	265	610
12	222	660	**12**	214	579	**12**	208	530
13	215	720	**13**	198	576	**13**	201	507
14	222	2195	**14**	177	1933	**14**	189	1819
15	225	630	**15**	206	539	**15**	213	441
16	208	615	**16**	199	534	**16**	200	422
17	191	636	**17**	176	542	**17**	182	464
18	178	536	**18**	158	460	**18**	163	383
19	182	527	**19**	169	477	**19**	176	349
20	204	629	**20**	197	496	**20**	185	459
21	180	504	**21**	173	429	**21**	182	361
22	168	518	**22**	152	474	**22**	160	369
23	152	473	**23**	144	397	**23**	141	356
24	177	516	**24**	169	462	**24**	167	347
25	142	1029	**25**	125	873	**25**	122	755
26	147	413	**26**	137	359	**26**	146	262
27	140	376	**27**	115	330	**27**	130	278
28	129	396	**28**	111	346	**28**	124	281
29	136	387	**29**	127	324	**29**	136	248

Análisis completo de los 29 cromosomas autosomales de las razas HOLSTEIN, JERSEY Y BROWN SWISS.

HOLSTEIN

	Alelos Faltantes					**% Exactitud**				
Cromosoma			Alelos	Faltantes		20%	40%	60%	80%	100%
1	163	325	488	650	813	93.52	90.50	86.86	83.97	80.58
2	145	288	433	577	721	93.84	90.45	86.80	83.58	80.58
3	143	286	429	572	715	93.79	90.10	86.89	83.22	80.09
4	133	266	399	532	665	94.05	90.84	87.68	84.46	81.17
5	130	259	389	518	648	95.22	92.07	88.68	85.41	82.04
6	333	666	999	1332	1665	75.27	72.17	71.49	71.47	70.99
7	100	200	301	401	501	94.89	92.00	89.03	85.71	82.26
8	119	238	358	477	596	93.93	90.71	87.47	84.22	80.81
9	102	204	306	408	510	94.38	90.79	87.76	84.86	81.48
10	113	226	339	452	565	93.42	90.09	86.73	84.18	80.44
11	122	244	365	487	609	94.27	91.07	87.66	84.31	81.17
12	87	173	260	346	433	94.85	92.08	88.62	85.47	82.06
13	100	201	301	402	502	94.03	90.91	87.38	83.95	80.66
14	394	788	1181	1575	1969	92.54	88.56	83.70	78.55	74.38
15	81	162	243	324	405	94.15	90.88	87.66	84.42	81.35
16	81	162	244	325	406	93.68	89.99	86.63	83.36	80.47
17	88	176	265	353	441	93.73	90.47	87.06	83.66	80.66
18	72	143	215	286	358	92.55	89.04	85.81	82.83	79.67
19	69	138	206	275	344	94.02	90.89	87.83	84.85	81.54
20	84	169	253	338	422	93.84	90.45	87.11	83.98	80.76
21	64	127	191	254	318	94.36	91.32	87.87	84.67	80.92
22	70	140	209	279	349	93.96	89.56	86.11	82.90	79.47
23	64	128	193	257	321	93.93	90.65	86.93	83.62	80.24
24	68	135	203	270	338	93.60	90.58	87.26	84.18	80.96
25	177	355	532	710	887	93.06	89.27	85.01	80.64	75.98
26	53	106	160	213	266	94.26	90.71	87.30	84.00	80.90
27	47	94	142	189	236	95.07	91.98	88.54	85.01	82.15
28	53	107	160	214	267	95.19	91.26	87.72	84.36	81.57
29	50	100	150	200	250	87.89	86.58	83.24	79.87	76.5

JERSEY

	Alelos Faltantes					% Exactitud				
Cromosoma		Alelos	Faltantes			20%	40%	60%	80%	100%
1	121	241	362	482	603	93.74	91.11	88.00	85.14	82.44
2	119	237	356	474	593	93.72	90.94	88.21	85.20	82.15
3	95	191	286	382	477	94.57	91.48	88.36	86.22	83.48
4	103	205	308	410	513	93.73	91.14	87.62	84.80	81.79
5	99	198	298	397	496	92.58	89.53	86.60	83.76	80.90
6	251	502	753	1004	1255	78.95	78.65	77.34	76.17	74.45
7	75	150	224	299	374	94.43	92.13	89.97	87.81	85.47
8	92	184	276	368	460	78.67	78.37	76.07	75.26	75.15
9	76	153	229	306	382	90.83	87.88	85.04	82.55	79.76
10	85	170	255	340	425	93.21	90.86	87.83	84.91	82.10
11	108	216	323	431	539	93.49	90.33	87.28	84.15	80.87
12	70	141	211	282	352	92.90	89.79	86.57	83.86	81.46
13	72	143	215	286	358	93.02	89.96	87.26	84.60	81.84
14	341	683	1024	1366	1707	90.24	85.89	82.15	78.16	75.19
15	63	126	188	251	314	92.74	89.92	87.13	84.02	81.80
16	65	130	195	260	325	91.92	88.92	85.56	82.66	79.91
17	69	139	208	278	347	94.19	91.28	88.55	85.85	82.75
18	56	113	169	226	282	94.14	91.33	89.23	86.47	83.47
19	59	118	176	235	294	93.39	90.00	87.18	84.76	81.80
20	58	116	173	231	289	93.35	90.44	87.52	85.26	82.04
21	49	97	146	194	243	91.86	89.57	86.86	84.61	82.33
22	61	122	183	244	305	93.31	90.27	87.52	84.55	81.60
23	49	98	147	196	245	92.24	88.83	85.00	82.34	79.22
24	57	114	170	227	284	92.60	89.59	86.50	83.96	81.30
25	146	292	439	585	731	92.67	88.95	84.79	80.69	76.48
26	42	85	127	170	212	94.63	91.92	89.31	86.12	83.02
27	38	76	114	152	190	94.45	91.62	88.87	85.59	83.08
28	43	87	130	174	217	94.72	92.11	89.24	85.98	82.05
29	37	75	112	150	187	93.43	90.78	88.24	85.87	83.67

BROWN SWISS

	Alelos Faltantes					% Exactitud				
Cromosoma		Alelos	Faltantes			20%	40%	60%	80%	100%
1	109	219	328	438	547	94.12	91.08	88.18	85.26	82.47
2	90	179	269	358	448	94.05	91.53	88.90	86.07	83.61
3	81	161	242	322	403	94.45	91.57	88.64	86.19	83.38
4	76	152	229	305	381	92.98	90.22	87.67	85.60	82.87
5	66	131	197	262	328	94.14	91.73	89.40	86.73	84.24
6	167	333	500	666	833	91.40	87.83	84.38	80.76	79.31
7	63	126	190	253	316	94.43	92.13	89.97	87.81	85.47
8	74	148	223	297	371	91.84	89.88	87.92	85.42	82.39
9	62	123	185	246	308	94.61	92.30	89.54	87.04	84.20
10	70	141	211	282	352	93.98	91.14	88.48	85.97	83.24
11	65	130	194	259	324	94.07	91.39	88.81	86.14	83.63
12	61	121	182	242	303	93.71	91.22	88.92	85.89	83.70
13	58	116	173	231	289	94.78	91.82	88.89	86.20	83.35
14	319	637	956	1274	1593	91.15	84.38	82.92	78.56	74.60
15	43	86	130	173	216	94.52	92.24	89.61	87.18	84.56
16	43	85	128	170	213	94.92	91.96	89.52	87.33	84.78
17	54	108	161	215	269	94.46	91.64	89.09	86.68	84.23
18	41	82	123	164	205	93.81	91.23	88.57	86.13	83.56
19	33	66	100	133	166	95.17	92.79	90.50	87.97	85.68
20	50	101	151	202	252	93.96	91.14	88.57	86.13	83.50
21	35	70	105	140	175	95.18	92.68	90.11	87.54	85.21
22	40	80	120	160	200	96.18	93.38	91.24	87.26	84.81
23	41	82	122	163	204	94.20	91.54	88.86	86.30	83.36
24	34	68	101	135	169	95.36	93.00	90.32	87.84	85.61
25	123	245	368	490	613	91.88	87.89	84.88	80.39	77.01
26	23	46	69	92	115	94.67	92.43	90.11	87.98	85.87
27	28	55	83	110	138	93.12	90.91	88.36	86.64	84.02
28	30	61	91	122	152	93.73	91.03	88.77	86.09	83.75
29	22	44	67	89	111	95.83	93.67	90.75	88.48	86.17