

Universidad Autónoma de Baja California
Instituto de Ingeniería

Maestría y Doctorado en Ciencias e Ingeniería



Título:

"Herramienta bioinformática para la administración de hatos ganaderos incluyendo mejora genética basada en selección genómica"

TESIS PARA OBTENER EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
NATANIEL ANGULO CUEVAS

DIRECTOR DE TESIS:
DR. RAFAEL VILLA ANGULO

Mexicali, B.C.

Agosto 2019

AGRADECIMIENTOS

A mi Familia: Mis padres por la educación que me brindaron y a mis hermanos por su apoyo.

A mi novia. A ti Daniela Villanueva Uribe por la motivación y ánimos aun en los momentos menos gratos.

Al Dr. Rafael Villa Angulo: Por darme la oportunidad de formar parte de su grupo de investigación, por compartir su conocimiento conmigo y por su constante apoyo para lograr este objetivo.

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

A la Universidad Autónoma de Baja California y su Instituto de Ingeniería: Por aceptarme en su programa de Maestría y Doctorado en Ciencias e Ingeniería.

A CONACyT: Por el apoyo económico a través de su programa de Becas Nacionales.

A mis compañeros: que siempre estuvieron con la disposición de ayudar y por la convivencia.

RESUMEN

La Estimación de los Valores de Crianza tradicional combina solo datos fenotípicos y de pedigrí, hoy en día, con la inclusión de marcadores moleculares, tales como los SNPs (Single Nucleotide Polymorphism, por sus siglas en inglés), en la integración del análisis genético en los esquemas de selección, ya es posible integrar datos genotípicos para realizar estas estimaciones. Avances en las tecnologías de genotipificación de salida masiva para bovinos llevaron al desarrollo de los SNPchips, de los cuales hay de diferentes capacidades. Para la Estimación de los Valores Genómicos de Crianza (GEBV, por sus siglas en inglés) existen diversos métodos basados en el contexto del Modelo Lineal de Mezclas (Linear Mixed Models), en los que encontramos el GBLUP (Genomic Best Linear Unbiased Prediction) el cual es un método que utiliza relaciones genómicas para realizar dichas estimaciones. Para esto, es usada una matriz de relaciones genómicas, estimada de marcadores moleculares (como SNPs). Esta matriz define la covarianza entre individuos en un grupo de estudio basada en similitud observada a nivel genómico, en vez de la similitud basada en pedigrí. En este trabajo de tesis se reporta la implementación y pruebas de una herramienta bioinformática que utiliza el método GBLUP para estimar los valores genómicos de ganado bovino. Esta herramienta es parte integral del primer sistema tecnológico para la implementación de un programa de mejora genética basado en Selección Genómica, en el Estado de Baja California.

ÍNDICE

AGRADECIMIENTOS	I
RESUMEN.....	II
1. INTRODUCCIÓN	1
1.1 ANTECEDENTES.....	1
1.2 SELECCIÓN GENÓMICA.....	2
1.2.1 <i>Genómica</i>	2
1.2.2 <i>Genotipo y Fenotipo</i>	3
1.2.3 <i>Selección asistida por marcadores</i>	4
1.2.4 <i>Avances de la tecnología de genotipificación de salida masiva para bovinos</i>	7
1.2.5 <i>Selección Genómica</i>	11
1.3 MÉTODOS ESTADÍSTICOS UTILIZADOS PARA ESTIMAR VALOR GENÓMICO DE CRIANZA.....	14
1.4 PLANTEAMIENTO DEL PROBLEMA.....	15
1.5 OBJETIVO Y METAS.....	18
1.5.1 <i>Objetivo</i>	18
2. MODELO ESTADÍSTICO PARA ESTIMACIÓN DE VALORES GENÓMICOS DE CRIANZA.....	20
2.1 MÉTODO BLUP (BEST LINEAR UNBIASED PREDICTION).....	21
2.2 MÉTODO GBLUP (GENOMIC BEST LINEAR UNBIASED PREDICTION).....	23
2.2.1 <i>Matriz de Relaciones Genómicas</i>	23
2.2.2 <i>Estimado del Valor Genómico de Crianza (GEBV)</i>	25
2.3 INTERPRETACIÓN PRÁCTICA DEL ESTIMADO VALOR GENÓMICO DE CRIANZA (GEBV).....	26
2.4 CÁLCULO DE HEREDABILIDAD.....	27
2.5 CÁLCULO DEL COEFICIENTE DE ENDOGAMIA.....	27
2.6 CÁLCULO DE LA PRECISIÓN DE LOS VALORES GENÓMICOS DE CRIANZA.....	28
3. IMPLEMENTACIÓN.....	30
3.1 HERRAMIENTA BIOINFORMÁTICA PARA LA ADMINISTRACIÓN DE HATO GANADEROS.....	31
3.2 DATOS NECESARIOS PARA EL CÁLCULO DE LOS VALORES GENÓMICOS DE CRIANZA.....	39
3.2.1 <i>Matriz de incidencia de Genotipos de los marcadores (SNPs)</i>	39
3.2.1.1 <i>Matriz de Frecuencia de Alelo Menor</i>	40
3.2.2 <i>Fenotipos</i>	41
3.2.3 <i>Pedigrí</i>	42
3.3 MATRIZ DE RELACIONES GENÓMICAS UTILIZANDO FRECUENCIAS DE LOS ALELOS EN AMBIENTE R.....	43
3.4 INTERFAZ GRÁFICA IMPLEMENTADA.....	48
3.4.1 <i>La opción Calcular</i>	48
3.4.2 <i>Seleccionar Archivos</i>	49
3.4.3 <i>Mostrar Cálculos</i>	50
4. PRUEBAS DEL MODELO Y ANÁLISIS DE RESULTADOS.....	55
5. CONCLUSIONES Y TRABAJO FUTURO	65
5.1 CONCLUSIONES.....	65
BIBLIOGRAFÍA	67

ILUSTRACIONES

ILUSTRACIÓN 1. GENOTIPO Y FENOTIPO. LOS GENOTIPOS CORRESPONDEN A LA INFORMACIÓN GENÉTICA DE LOS DISTINTOS RASGOS, MIENTRAS QUE LOS FENOTIPOS CORRESPONDEN AL PRODUCTO FÍSICO O FISIOLÓGICO QUE PUEDE SER MEDIDO Y CARACTERIZAN LOS DISTINTOS RASGOS.	3
ILUSTRACIÓN 2. 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.	5
ILUSTRACIÓN 3. LA FORMA MÁS COMÚN DE POLIMORFISMOS EN LOS GENOMAS SON LOS SNPs.	8
ILUSTRACIÓN 4. RAZAS UTILIZADAS PARA LA VALIDACIÓN DE LOS 54,609 SNPs. SE TOMARON MUESTRAS DE 565 ANIMALES PERTENECIENTES A 21 RAZAS DE GANADO BOVINO Y SE REALIZARON LOS PRIMEROS ESTUDIOS DE ESTRUCTURA GENÉTICA DE GENOMA COMPLETO (IMAGEN TOMADA DE ILLUMINA.COM).	8
ILUSTRACIÓN 5. CARACTERÍSTICAS DEL SNPCHIP50, Y LA TECNOLOGÍA DE GENOTIPIFICACIÓN DE SALIDA MASIVA. ESTÁNDARES PARA REALIZAR GENOTIPIFICACIÓN MASIVA EN PROGRAMAS DE MEJORA GENÉTICAS ASISTIDA POR MARCADORES.....	9
ILUSTRACIÓN 6. PASOS DE LA 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.....	12
ILUSTRACIÓN 7. VENTANA PRINCIPAL DEL SOFTWARE IMPLEMENTADO.	31
ILUSTRACIÓN 8. SUBMENÚ DE LA OPCIÓN DE VACAS.....	32
ILUSTRACIÓN 9. SUBMENÚ DE LA OPCIÓN BECERRAS	32
ILUSTRACIÓN 10. VENTANA DE REGISTRO DE VACAS NUEVAS.	33
ILUSTRACIÓN 11. VENTANA DE REGISTRO DE PRODUCCIÓN.	34
ILUSTRACIÓN 12. REGISTRO DE LA VACA.	34
ILUSTRACIÓN 13. VENTANA DE MODIFICACIÓN DE REGISTROS O SU ELIMINACIÓN.	35
ILUSTRACIÓN 14. VENTANA PARA NUEVOS REGISTROS DE BECERRAS.....	35
ILUSTRACIÓN 15. SUBMENÚ DE LA OPCIÓN VER HATO.....	36
ILUSTRACIÓN 16. VENTANA DE DATOS DE VACAS.	36
ILUSTRACIÓN 17. VENTANA DATOS DE BECERRAS.....	37
ILUSTRACIÓN 18. VENTANA DE ANÁLISIS DE PRODUCCIÓN.....	37
ILUSTRACIÓN 19. SUBMENÚ DE LA OPCIÓN VALORES GENÓMICOS.	38
ILUSTRACIÓN 20. VENTANA GENERACIÓN DE ARCHIVOS.	38
ILUSTRACIÓN 21. TABLA DE GENOTIPOS VISTA EN UN EDITOR DE TEXTO.	40
ILUSTRACIÓN 22. TABLA DE GENOTIPOS VISTA EN UN EDITOR DE TEXTO, CODIFICADA EN 0, 1 Y 2.	41
ILUSTRACIÓN 23. TABLA DE FENOTIPOS VISTA EN AMBIENTE R.....	41
ILUSTRACIÓN 24. TABLA DE PEDIGRÍ VISTA EN AMBIENTE R.	42
ILUSTRACIÓN 25. VENTANA PRINCIPAL DE GEBV.EXE.	48
ILUSTRACIÓN 26. BOTONES EN EL CENTRO DE LA VENTANA PRINCIPAL DE GEBV.EXE.	48
ILUSTRACIÓN 27. VENTANA CALCULANDO DE GEBV.EXE.	49
ILUSTRACIÓN 28. VENTANA SELECCIONAR ARCHIVOS DE GEBV.EXE.....	50
ILUSTRACIÓN 29. SUBVENTANA SELECCIONAR ARCHIVOS DE GEBV.EXE.....	50
ILUSTRACIÓN 30. VENTANA VALORES GENÓMICOS DE CRIANZA DE GEBV.EXE.....	51
ILUSTRACIÓN 31. DESCRIPCIÓN DE VENTANA VALORES GENÓMICOS DE CRIANZA DE GEBV.EXE.	52
ILUSTRACIÓN 32. VENTANA PARA MOSTRAR GRÁFICAS DE GEBV.EXE.....	53
ILUSTRACIÓN 33. BARRA DE MENÚ EN LA VENTANA PRINCIPAL DE CRIANZA DE GEBV.EXE.....	54
ILUSTRACIÓN 34. CORRELACIÓN ENTRE LOS GEBVS RESULTANTES DE LOS MÉTODOS GEBVA Y GEBVB.	57
ILUSTRACIÓN 35. CORRELACIÓN ENTRE LOS GEBVS RESULTANTES DE LOS MÉTODOS GEBVA Y GEBVC	58
ILUSTRACIÓN 36. CORRELACIÓN ENTRE LOS GEBVS RESULTANTES DE LOS MÉTODOS GEBVB Y GEBVC.	58
ILUSTRACIÓN 37. CORRELACIÓN ENTRE LOS GEBVS RESULTANTES DE LOS MÉTODOS GEBVA Y A.	59

ILUSTRACIÓN 38. PRECISIÓN DE LAS MATRICES DE RELACIONES GENÓMICAS (G POR FRECUENCIAS OBSERVADAS, G POR REGRESIÓN, Y A DE PEDIGRÍ) PARA CALCULAR LA RELACIÓN ENTRE INDIVIDUOS.	60
ILUSTRACIÓN 39. CORRELACIÓN DE LOS RESULTADOS DE GEBVB Y LOS VALORES DEL FENOTIPO.	61
ILUSTRACIÓN 40. CORRELACIÓN DE LOS RESULTADOS DE GEBVA Y LOS VALORES DEL FENOTIPO.	62
ILUSTRACIÓN 41. GRAFICA DE CORRELACIÓN ENTRE LOS RESULTADOS DE GEBVA Y EL PAQUETE RRBLUP.	63
ILUSTRACIÓN 42. GRÁFICA DE CORRELACIÓN ENTRE LOS RESULTADOS DE GEBVA VS PAQUETE SYNBREED.	64

1. INTRODUCCIÓN

1.1 Antecedentes.

Las especies bovinas constituyen un grupo muy importante de animales, no solo por su posibilidad de explotación y aporte en la economía de muchos países, sino por ser de las primeras especies domesticadas por el hombre en la era Neolítica (~10,000 AC), hecho que le ha permitido acompañarlo en su evolución, y en gran parte de sus rutas migratorias, a través de la historia. El ganado bovino representa así junto con los cerdos, perros y gatos, una clase de mamíferos placentados que ha coevolucionado con los humanos, y su estudio permite no sólo ampliar el conocimiento de estas especies, sino brindar huellas importantes sobre la evolución e historia natural de las mismas [1].

La selección de animales para reproducción se ha venido practicando desde los inicios del proceso de domesticación; A principios la selección se basaba exclusivamente en la apariencia física y la capacidad productiva (rasgos fenotípicos) de los animales. Sin embargo, en los últimos dos siglos se ha venido implantando la

sistematización dirigida por parte de los productores, y se han diseñado programas de crianza de animales beneficiando funciones específicas, tales como producción de carne, de leche, facilidad de parto y longevidad, entre otras. En los últimos 50 años, con los avances en las técnicas de análisis genético, y más recientemente (últimos 5 años) análisis genómico, ha sido posible evaluar el mérito genético de los animales, lo que ha permitido tomar decisiones de selección más eficaces. La reciente inclusión de la evaluación genómica se proyecta como la tecnología más prometedora para implementar estrategias de selección y garantizar la capacidad productiva de las nuevas generaciones de ganado bovino [2].

1.2 Selección Genómica.

1.2.1 Genómica

La Genómica es el conjunto de ciencias y técnicas dedicadas al estudio integral del funcionamiento, el contenido, la evolución y el origen de los genomas (la totalidad de la información genética que posee un organismo o una especie en particular) [3]. Existen muchas áreas relacionadas con la genómica que se han estado desarrollando a lo largo de los años, algunas de las más importantes por su potencial tanto económico como social y ambiental son la medicina genómica, la genómica forense, la genómica ambiental, la genómica industrial, y en especial la genómica agropecuaria, que es el área de aplicación de este trabajo de tesis.

El desarrollo de la Genómica Agropecuaria podría detonar en México la sustentabilidad del sector agropecuario. En especial la mejora genética de las especies, basada en el análisis de su genoma, cambiará los paradigmas de

producción y conservación de recursos naturales. En México ya se están haciendo algunos esfuerzos por adoptar las tecnologías genómicas para la mejora de la producción de leche y carne proveniente de la especie bovina, tal es el caso del proyecto de mejora genética basado en técnicas de Selección Genómica, que se desarrolla actualmente por investigadores de los Institutos de Investigaciones en Ciencias Veterinarias e Ingeniería de la Universidad Autónoma de Baja California; con el cual se pretende generar nuevos linajes de ganado bovino mejor adaptados a las condiciones climáticas de las zonas áridas de Baja California, así como más resistentes a los patógenos propios de la región, y más productivos [4].

1.2.2 Genotipo y Fenotipo.

El genotipo de un animal representa el gen o grupo de genes responsable de un rasgo en particular. En un sentido más general, el genotipo describe todo el grupo de genes que un individuo ha heredado. Como contraste, el fenotipo es el valor que toma un rasgo; en otras palabras, es lo que puede ser observado o medido. Por ejemplo, el fenotipo puede ser rasgos físicos y/o fisiológicos tales como las enfermedades, el color de piel, la producción individual de leche de una vaca, la facilidad de parto, etc. [5].

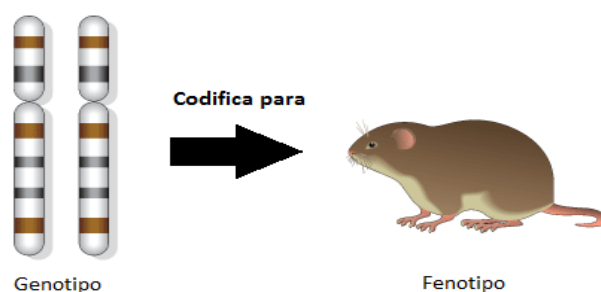


Ilustración 1. Genotipo y Fenotipo. Los genotipos corresponden a la información genética de los distintos rasgos, mientras que los fenotipos corresponden al producto físico o fisiológico que puede ser medido y caracterizan los distintos rasgos.

Existe una diferencia importante entre genotipo y fenotipo. El genotipo es esencialmente una característica fija del organismo; permanece constante a lo largo de la vida del animal y no es modificado por el medio ambiente. Cuando solamente uno o un par de genes son responsables por un rasgo, el genotipo permanece generalmente sin cambios a lo largo de la vida del animal (p. ej., color de pelo). En este caso, el fenotipo otorga una buena indicación de la composición genética del individuo. Aun así, para algunos rasgos, el fenotipo cambia constantemente a lo largo de la vida del individuo como respuesta a factores ambientales. En este caso, el fenotipo no es un indicador confiable del genotipo. Esto generalmente se presenta cuando muchos genes se involucran en la expresión de un rasgo, tal como producción de leche [5].

1.2.3 Selección asistida por 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. Se sabe que un buen balance entre el manejo adecuado de la crianza y la selección minuciosa de animales con características fenotípicas mayores en los rasgos que deseamos ha conducido a lograr progreso, pero de igual forma se sabe que muchos de los rasgos observables (fenotipos) de fisiología del animal, de sus propiedades bioquímicas, de su morfología, de desarrollo, comportamiento y producción, son epigenéticos. Esto es, son regulados por la interacción de los genes con el medio

ambiente. Lo que hace 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 epistático de genes que actúan de forma interactiva para controlar los rasgos del animal, influya de forma determinante en el rendimiento productivo.

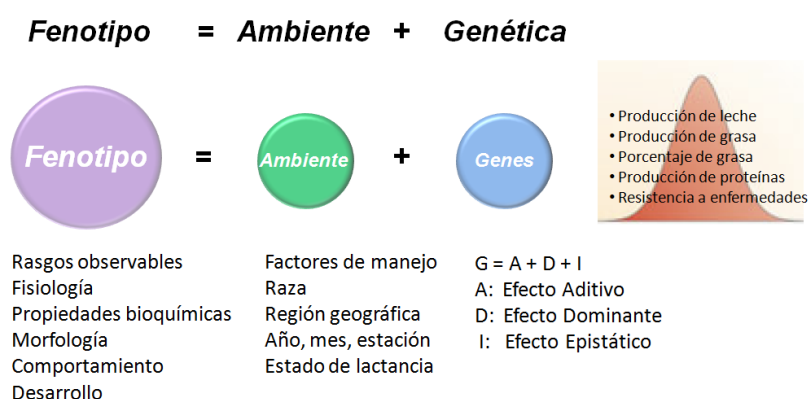


Ilustración 2. La mayor parte de los rasgos de importancia económica en el ganado bovino son epigenéticos. La mayor parte de los fenotipos son regulados por la interacción de los genes con el medio ambiente.

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 [6]. 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 Valores de Crianza (EBV, “Estimated Breeding Values”, por sus siglas en Inglés).

Sin embargo, aun cuando la idea de usar marcadores de ADN para incrementar la tasa de ganancia genética en el ganado lechero, ha existido por décadas(ej. [7]), la adopción de la selección asistida por marcadores (MAS, “Marker Asisted Selection”, por sus siglas en inglés) en la industria lechera había estado limitada, hasta hace muy poco tiempo. Existían diferentes razones para esto; 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. 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 crianza incluyendo la información de los marcadores era una barrera más para la aplicación de la selección asistida por marcadores [7].

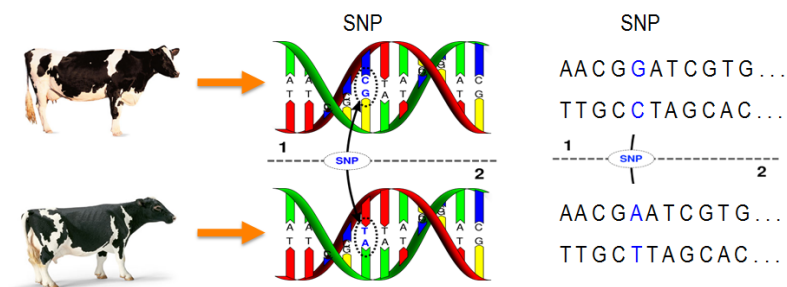
En el año 2001, investigadores de Holanda y Australia demostraron a través de simulaciones por computadora, que usando marcadores de ADN distribuidos a lo largo de todo el genoma permitía capturar de forma simultánea los efectos de todos los genes que afectan y regulan las características de importancia económica, aumentando con esto de forma substancial la precisión de la selección asistida por marcadores [7]. Este trabajo inicial sobre los beneficios de utilizar información de todo el genoma sirvió para establecer las necesidades tanto de información como de tecnología para lograr los avances demostrados en las simulaciones. Estos requerimientos incluían; 1) Contar con poblaciones de animales con mediciones fenotípicas precisas y sus muestras de ADN genómico, 2) Marcadores de polimorfismo de ADN posicionados en todo el genoma, 3) Tecnología de laboratorio

para realizar análisis de una alta densidad de marcadores y 4) Algoritmos y métodos para el análisis de datos genotípicos y fenotípicos. Estos requerimientos, aunados a otros beneficios de gran importancia, impulsaron a varios países, a través de sus centros de investigación y universidades a generar dos consorcios con el fin de secuenciar el genoma y crear un mapa de haplotipos de la evolución *Bos Taurus*.

1.2.4 Avances de la tecnología de genotipificación de salida masiva para bovinos.

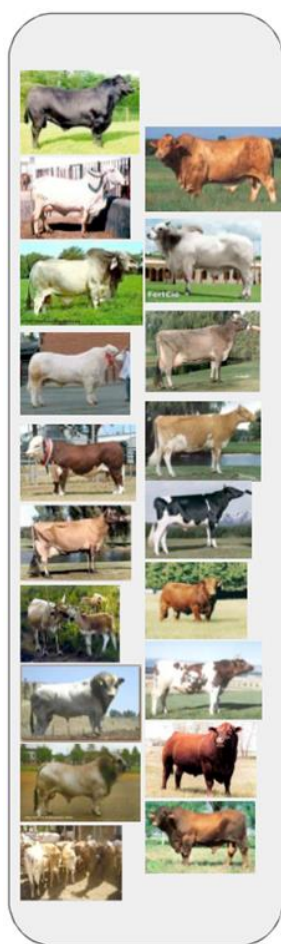
Los proyectos de secuenciación de genomas completos, en especial los genomas humano y bovino, han sido detonadores del desarrollo de nuevas tecnologías para la captura de información genética proveniente de muestras de ADN. Un ejemplo de ellas son las tecnologías de salida masiva para extracción de genotipos (“HTGT High-Throughput Genotyping Technologies”, por sus siglas en inglés) las cuales han jugado un rol invaluable en la captura de diversidad genética y variaciones hereditarias entre individuos. La disponibilidad de las HTGTs ha permitido la extracción de genotipos de especies adicionales, que han servido como organismos modelos para resolver la complejidad de la evolución humana y para extrapolar de una forma efectiva, información genética por medicina comparativa (veterinaria) a medicina humana.

En especial en la especie bovina, con el proyecto de secuenciación del genoma fue posible descubrir miles de marcadores de ADN, en forma de Polimorfismo de Nucleótido Simple (“SNP Single Nucleotide Polymorphism” por sus siglas en Ingles) [8]. Para hacer esto se realizó un estudio de descubrimiento de SNPs obteniendo como resultado más de 6 millones de SNPs putativos.



- Los SNPs son la forma más común de variación genética entre individuos
- Se estima que el genoma humano contiene más de 10 millones de SNPs, lo que equivale a un SNP cada 300 bases

Ilustración 3. La forma más común de polimorfismos en los genomas son los SNPs.



BREED	SAMPLES	POLYMORPHIC LOCI*	MEAN MAF†	MEDIAN MAF†
Angus	60	41,491	0.21	0.21
Beefmaster	24	42,925	0.22	0.21
Bos indicus Gir	24	23,971	0.11	0.02
Bos indicus Nelore	21	25,814	0.11	0.02
Brahman	25	30,284	0.13	0.08
Brown Swiss	24	36,347	0.19	0.17
Charolais	26	42,589	0.22	0.21
Guernsey	21	38,632	0.19	0.17
Hereford	32	42,992	0.20	0.23
Holstein	64	42,730	0.22	0.22
Jersey	28	35,976	0.18	0.14
Limousin	45	42,821	0.22	0.22
N'Dama	25	29,049	0.14	0.08
Norwegian Red	21	42,782	0.22	0.21
Piedmontese	24	42,185	0.22	0.21
Red Angus	15	40,188	0.21	0.20
Romagnola	24	38,830	0.20	0.19
Santa Gertrudis	24	42,064	0.22	0.21
Sheko	20	35,726	0.17	0.12
Outgroup‡	18	11,206	0.05	0.00
Overall	565	47,545	0.25	0.24

Ilustración 4. Razas utilizadas para la validación de los 54,609 SNPs. Se tomaron muestras de 565 animales pertenecientes a 21 razas de ganado bovino y se realizaron los primeros estudios de estructura genética de genoma completo (imagen tomada de Illumina.com).

Inicialmente de validaron 54,609 SNPs (50K) muestreando 565 animales de 19 razas distintas ubicadas alrededor del mundo, y se realizaron los primeros estudios de caracterización de la estructura genética de la especie [8].

La validación de los SNPs facilitó el diseño e implementación del primer Chip de Genotipificación de Salida Masiva para ganado bovino (Bovine SNPchip50 comercializado por Illumina Inc. desde 2009). El SNPchip50 [9], en solo unos meses se convirtió en el estándar para realizar estudios genéticos de las razas de ganado pertenecientes a la evolución *Bos Taurus*.

- Herramienta estándar para realizar genotipado en el genoma-completo para investigación y aplicaciones en la industria
- 54,609 SNPs distribuidos en los 30 cromosomas del genoma bovino
- 60,800 Beads

Parámetro	Rendimiento*	Especificación del producto
Tasa de lectura	99.50 %	> 99 %
Inconsistencia Mendeliana	0.04 %	> 0.1 %
Reproducibilidad	100 %	< 99 %
Frecuencia de alelo menor	0.19 media / 0.2 mediana	x
Locus polimórficos	37,758 media / 40,188 mediana	x

* Basado en 565 animales de 19 razas de ganado bovino, 46 trios, y dos replicas

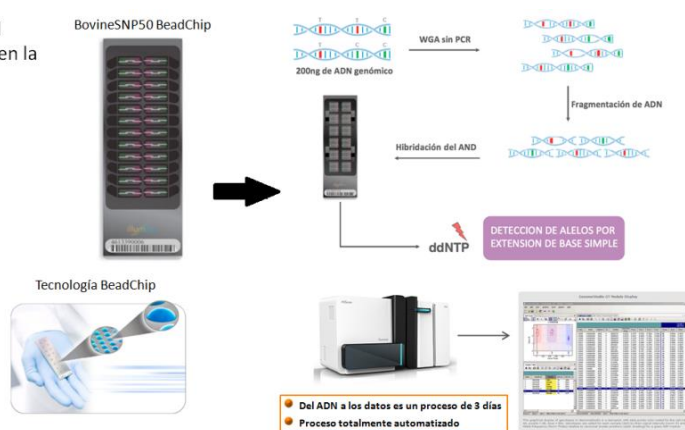


Ilustración 5. Características del SNPchip50, y la tecnología de genotipificación de salida masiva. Estándares para realizar Genotipificación masiva en programas de mejora genéticas asistida por marcadores.

Actualmente existen en el mercado seis SNPchips distintos para ganado bovino, cuya capacidad va desde baja hasta alta densidad. Cinco de estos son de la compañía Illumina: Bovine3K [10], BovineLD [11], BovineSNP50v.1 [9], BovineSNP50v.2 [9], y el BovineHD [12], los cuales tienen capacidad de 2900, 6909, 54001, 54609 y 777962 SNPs respectivamente. Uno es de la empresa Affimetrix: Axiom Bos 1 [13] el cual tiene una capacidad de 648875 SNPs. Los propósitos específicos de estos chips van

desde cálculo del mérito neto para características productivas, hasta cálculo de valores genómicos y estudios de asociación de SNPs con enfermedades y distintos fenotipos. La tabla 1 presenta un resumen de los chips incluyendo el propósito específico de cada uno de ellos.

Tabla 1. Resumen de SNPchips comerciales y el tipo de análisis que puede ser realizado con ellos.

SNPchip	Total de SNPs	Compañía	Propósito principal
Bovine3K	2,900	Illumina	Merito genético, verificación de porcentaje, trazabilidad de ganado
BovineLD	6,909	Illumina	Valor genómico de crianza (GEBV), variación genética
BovineSNP50v.1	54,001	Illumina	Selección genómica, Identificación de QTLs, Evaluación de mérito genético, comparación genética
BovineSNP50v.2	54,609	Illumina	Selección genómica, Identificación de QTLs, Evaluación de mérito genético, comparación genética
BovineHD	777,962	Illumina	Variación genética dentro de poblaciones, selección genómica, identificación de QTLs, evaluación de mérito genético, mapeo entre razas, desequilibrio de ligamiento, comparación genética, caracterización de razas para estudios de biodiversidad
Axiom BOS 1	648,875	Affimetrix	Merito genético, estudios de asociación, evaluación genómica, estudios de respuesta a fármacos, desequilibrio de ligamiento

1.2.5 Selección Genómica.

La revolución de la selección genómica empezó con dos desarrollos claves; el primero fue la reciente secuenciación del genoma bovino [14], y el segundo el desarrollo del Bovine SNPchip50 [9]. La Selección genómica está revolucionando a la industria del ganado y se refiere a la toma de decisiones de selección basada en el valor genómico de crianza de los animales (GEBV, “Genomic Estimated Breeding Value” por sus siglas en inglés). Los GEBVs son calculados como la suma de los efectos de una cantidad grande de marcadores moleculares (SNPs), distribuidos a lo largo de todo el genoma. Los genotipos y/o haplotipos de los marcadores involucrados capturan los efectos de todos los Locus de Características Cuantitativas (QTL, “Quantitative Trait Loci” por sus siglas en inglés) que contribuyen a la variación de un rasgo de valor económico. El efecto de los QTLs, inferidos ya sea de los haplotipos o genotipos de marcadores de tipo polimorfismo de nucleótido simple (SNPs), es primeramente estimado en una población grande de referencia con información fenotípica. En las generaciones subsecuentes solo es necesaria la información de los marcadores para calcular los GEBVs.

La tecnología de selección genómica involucra tres pasos principales; 1) captura de las muestras y obtención del ADN genómico, 2) aplicación de los SNPchips para obtención de los genotipos y 3) análisis bioinformático de la información para obtener los GEBVs. La captura de las muestras puede ser hechas de esperma, sangre, saliva, segregación nasal o una pinchada en la oreja, entre otras. Estas muestras deben ser transportadas al laboratorio para su procesamiento químico de extracción de ADN. El ADN debe ser ADN genómico completo y purificado. El ADN es amplificado y depositado en los SNPchip. 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 los polimorfismos y se genera una base de datos con la información genética de cada uno de los animales de la muestra.

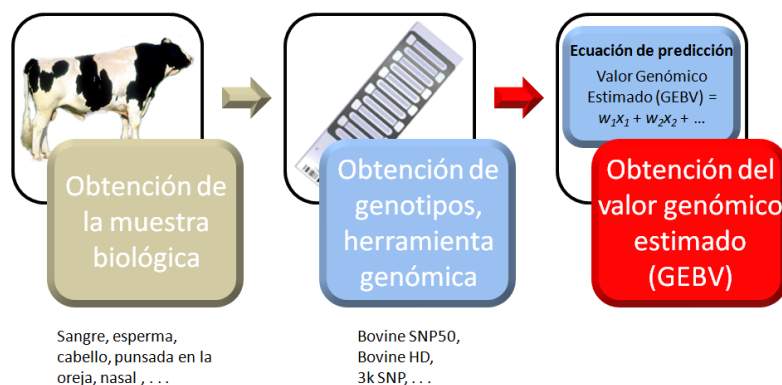


Ilustración 6. Pasos de la Selección Genómica. La selección genómica involucra tres pasos principales: 1) Captura de la muestras y obtención de ADN genómico, 2) Extracción de genotipos y 3) Análisis de la información.

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.

La selección genómica ha sido adoptada por la mayor parte de los países productores de leche de Europa y por Estados Unidos y Canadá en América. La figura tabla 2 muestra los resultados en la efectividad de hacer selección basado en datos genómicos.

Tabla 2. Resultados de la efectividad de la selección genómica.

País	Número de toros	Número de SNPs	Características de producción	Método estadístico de cálculo	Certeza de los GEBV
Nueva Zelanda (Harris et al, 2008)	4,500	44,146	Leche, peso al nacimiento, fertilidad, longevidad, Células Somáticas	BLUP, Bayes A, Bayes B	50-60% (34% PABV)
USA y Canadá (VanRaden, 2009)	3,576	38,416	Merito neto, producción de leche, producción de grasa, longevidad	Similar al BLUP	50% (27% PABV%)
Holanda (De Roos, 2009)	1,583	57,660	Producción de grasa, producción de proteínas, patas y piernas, profundidad de la ubre, células somáticas, fertilidad	-	Aumento de (20% PABV)
Australia (Hayes et al, 2009)	798	38,259	Producción de proteínas, fertilidad, Índice de selección Australiano, Ranquin de ganancia Australiano	BLUP, Bayes A	14-55% (41% PABV)

La selección genómica es la tecnología que ha generado el incremento más grande en la tasa de mejora genética para la industria lechera en los últimos 20 años. Y, aun cuando se había trabajado de forma simulada para proyectar sus beneficios, es hasta ahora que se está probando su verdadero potencial. El incremento en la efectividad de los GEBVs (Valor genético basado en información genómica) sobre los EBV (valor genético basado en información de fenotipos y pedigrí) para toros jóvenes sin record de sus hijas, es impresionante, va del 2% al 20%. El incremento en la efectividad de los GEBVs está siendo usado de dos formas por las compañías de crianza de

sementales. En algunos casos grupos grandes de toros jóvenes están siendo muestreados para seleccionar un número reducido de toros y llevarlos a programas de prole. Esto reduce el costo de los programas de crianza y resulta en una ganancia genética extra. Otras compañías están ofertando grupos de toros jóvenes basados únicamente en su GEBV, tan rápido como estos están listos para reproducción. Esto resulta en una ganancia genética mayor, como resultado de reducir la longitud del intervalo generacional.

1.3 Métodos Estadísticos utilizados para estimar valor genómico de crianza.

Un aspecto fundamental del cual depende, en gran parte, el éxito de un programa de mejora genética es la integración y procesamiento adecuados de una gran cantidad de datos de genotipos, de pedigrí, de salud, y del medio ambiente en el cual ha crecido cada animal. Para este efecto se han desarrollado distintos métodos estadísticos y matemáticos [15,16,17]. Los métodos para estimar valores genómicos de crianza podemos dividirlos en dos grupos; los que utilizan Modelos Bayesianos [16, 17], y los que utilizan Modelos de Regresión Aleatoria [15]. Los métodos que utilizan modelos Bayesianos están basados en hacer asociación multi-locus, y son resueltos con métodos matemáticos como Modelos de Markov entrenados con el método Monte Carlo, o Máxima Verosimilitud; mientras que los métodos que utilizan regresión aleatoria están basados en modelar sistemas de variables aleatorias utilizando Modelos Lineales de Mezclas, para calcular efectos aleatorios (genéticos y medio ambiente), partiendo de efectos fijos (pedigrí). Estos métodos son poco entendidos, razón por la cual es limitada la comunidad de científicos que

están desarrollando sistemas de software bioinformático para programas de mejora genética utilizando selección genómica.

1.4 Planteamiento del Problema.

Un problema histórico con el ganado lechero de las regiones áridas de Baja California es la baja eficiencia productiva en los meses de verano, debido al estrés calórico provocado por las drásticas condiciones climáticas. Este problema provoca una reducción en la producción de leche hasta en 25% y permite tan sólo 10% de concepciones, y ha tenido un impacto tanto económico como social incalculable. En adición, en los últimos 5 años, la situación de la producción lechera en el estado se ha agudizado provocando la mayor crisis productiva en los últimos 50 años. El factor detonador tal crisis es la escasez de alimentos para el ganado, porque aun cuando el cultivo de alfalfa y otros granos en el estado de Baja California son buenos, el problema de desabasto debido a la preferencia de los productores de exportar a Estados Unidos las cosechas, no ha cesado, y se proyecta que siga en aumento. En los últimos tres años, el 80% de las cosechas de alfalfa no han sido utilizadas para alimentación del ganado local. Esto ha provocado la desaparición de más del 50% de los hatos lecheros del estado, con el sacrificio de más de 15,000 animales, mismos que salieron del inventario habitual de producción.

Ante esta realidad, el productor lechero debe procurar en hacer más eficientes sus recursos, buscando tener cada vez mejores animales, capaces de responder a las nuevas exigencias que implica la producción de leche en las condiciones actuales, dotados de una genética superior, aparejados con nuevas estrategias de

alimentación, administración, manejo zootécnico, medicina preventiva, crianza de reemplazos y biotecnología de la reproducción. 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 los rasgos de interés económico, así como la resistencia a efectos de estrés calórico y/o una baja susceptibilidad a enfermedades de predisposición genética, aunada a un programa de salud animal, de bioseguridad y de registro productivo adecuado.

Como respuesta a esta importante necesidad, empresarios productores ganaderos del estado, en coordinación con grupos de investigación de la Universidad Autónoma de Baja California, emprendieron una iniciativa para generar programas de mejora genética basados en el genoma para contrarrestar los efectos de estrés calórico, aumentar la resistencia a enfermedades típicas de la región y generar linajes de ganado más adaptado y productivo en Baja California. La iniciativa se generó con el propósito de caracterizar a nivel de diferencias genómicas el ganado susceptible a sufrir estrés calórico, localizar los genes involucrados con el estrés calórico y otras patologías que afectan la producción de leche, diseñar e implementar estrategias de reproducción de ganado óptimo genéticamente para la producción de leche bajo las condiciones ambientales de las regiones áridas del estado de Baja California y desarrollar la tecnología de instrumentación necesaria para experimentos de exploración y análisis de ganado [18].

A la fecha (Mayo 2019), se realizó un muestreo de 150 animales en distintas regiones del estado de Baja California, se midió una serie de fenotipos de interés a cada animal, se obtuvo sangre, se extrajo ADN de las muestras y se realizó la

genotipificación con el SNPchip LD de la empresa Illumina, que captura 6,909 SNPs a lo largo de los 30 cromosomas del genoma bovino. Dentro de los avances que se han logrado están: la caracterización de variaciones genómicas en alta densidad (CNVs; Copy Number Variations, por sus siglas en inglés), hecha en 12 vacas lecheras la ciudad de Mexicali con el SNPchip Axiom Bos 1 [19], y la caracterización de variables de manejo, de salud y de producción del Modelo Experimental de Producción Lechera del Instituto de Investigaciones en Ciencias Veterinarias [20], donde se implementará el programa de Selección Genómica. Parte del trabajo actualmente en proceso es la caracterización la estructura genética, los estudios de asociación de los marcadores SNP con los fenotipos medidos, y la estimación de los valores genómicos de crianza para implementar el primer programa de mejora genética basado en Selección Genómica, en el estado de Baja California.

Para continuar con la implementación del programa de Selección Genómica es necesario desarrollar las herramientas bioinformáticas para administrar toda la información de salud, de manejo y producción de los animales, para la estimación de los Valores Genómicos de Crianza, para realizar las proyecciones de cruzamiento vaca-semental, para medir el progreso genético y para evaluar la eficacia del programa de mejora. En este trabajo de tesis se propone estudiar e implementar métodos estadísticos para estimar Valores Genómicos de Crianza y desarrollar una herramienta bioinformática para ser utilizada como base en la implementación de un software administrador que apoye el desarrollo del primer programa de Selección Genómica en ganado bovino en el Estado de Baja California, y en México.

De lo anterior que, en este trabajo doctoral se proponen los siguientes objetivos y metas:

1.5 Objetivo y metas.

Para el desarrollo de este trabajo se propone un objetivo general ocho metas específicas; las cuales presentan a continuación:

1.5.1 Objetivo.

Crear una Herramientas bioinformática para la administración de hatos ganaderos, incluyendo mejora genética basada en selección genómica, lo que permitirá a los productores de la región optimizar los recursos dentro del hato, así como seleccionar ganado genéticamente óptimo para la producción de leche, lo que garantizaría una transmisión genética adecuada de acuerdo a las características de medio ambiente y recursos de la región, propiciando así, que el ganado sea mejor adaptado a las condiciones ambientales de la región, disminución en los problemas típicos de salud y aumento de la producción.

Metas.

Para lograr el objetivo se plantean las siguientes metas:

1. Realizar un estudio de los esquemas de mejora genética basados en selección genómica.
2. Hacer el análisis y filtrado de datos de animales genotipificados.
 - a. Filtrado de Genotipos.
 - b. Análisis de los registros de producción, salud, etc. (Fenotipos).
 - c. Análisis de los registros de pedigrí.

3. Implementar la estimación de los valores genómicos de crianza de los animales en estudio, utilizando los genotipos datos obtenidos por nuestro muestreo de ganado.
4. Diseño de un software bioinformático que atienda las necesidades básicas de la administración de hatos ganaderos, como producción, salud, reproducción, incluyendo una herramienta bioinformática para cálculo de valores genómicos.
 - a. Hacer un estudio de las necesidades de administración en un hato ganadero.
 - b. Generar un diagrama de flujo del software.
 - c. Definir el lenguaje de programación y/o estudio de desarrollo de aplicaciones, así como el motor de bases de datos a utilizar.
5. Generar el software bioinformático diseñado.
6. Implementar el software bioinformático en un hato ganadero de la región.
7. Escritura del documento de tesis.
8. Defensa del trabajo para obtener el grado.

2. MODELO ESTADÍSTICO PARA ESTIMACIÓN DE VALORES GENÓMICOS DE CRIANZA.

El nombre Modelos Lineales de Mezclas viene del hecho de que estos modelos son lineales en los parámetros, y que las covariables, o variables independientes, pueden implicar una mezcla de efectos fijos y aleatorios [21]. Estos modelos son útiles en ajustes donde las mismas unidades estadísticas son medidas repetidamente (estudios longitudinales). Debido a estas particularidades, basaremos nuestro estudio en un método que utilice los modelos lineales de mezclas.

Los modelos lineales de mezclas son asumidos en muchas de las aplicaciones genéticas y pueden ser representados matricialmente de la siguiente forma:

$$y = Xb + Zu + e \quad (1)$$

O

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ \vdots & \vdots \\ Z_{n1} & Z_{n2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Donde y es un vector de n observaciones (fenotipos, mediciones repetidas a través del tiempo), X y Z son matrices conocidas, b es un vector de efectos fijos, u es un vector de efectos aleatorios y e vectores de valores residuales aleatorios [22] con:

$$\text{Var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \sigma^2$$

Donde σ^2 es un escalar, posiblemente desconocido, $G = \text{Var}(u)$ y $R = \text{Var}(e)$.

2.1 Método BLUP (Best Linear Unbiased Prediction).

En la crianza animal, BLUP o Mejor Predicción Lineal Insesgada, es una técnica para la estimación del mérito genético. En general, este es un método de estimación de efectos aleatorios. El contexto del BLUP es el modelo lineal de mezclas.

Por insesgado se refiere a que $E(\text{predictor}) = E(k'b + m'u) = k'b$. Henderson [15] mostró que la Mejor Predicción Lineal Insesgada (BLUP, por sus siglas en ingles) de $k'b + m'u$ es

$$k'\hat{b} + m'GZ'V^{-1}(y - X\hat{b}) \quad (2)$$

donde $V = R + ZGZ'$, \hat{b} es cualquier solución a (3), la ecuación generalizada de mínimos cuadrados

$$XV^{-1}\hat{b} = X'V^{-1}y \quad (3)$$

La dificultad con esto es que V frecuentemente es una matriz tan grande que la inversión de esta puede no ser viable. Un método alternativo fue sugerido por

Henderson [15]. La predicción de $k'b+m'u$ es $k'\hat{b} + m'\hat{u}$, donde \hat{b} y \hat{u} son cualquier solución para (4);

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad (4)$$

Esta solución es conocida como las Ecuaciones del Modelo Mixto o MME (Mixed Model Equations, por sus siglas en ingles).

Henderson comprobó que \hat{b} de (4) es una solución a (3), y que \hat{u} de (4) es igual a $GZ'V^{-1}(y - X\hat{b})$ de (2) [15].

La ventaja computacional de (4) sobre (2) es que no se requiere V ni su inversa. R tiene las mismas dimensiones que V , pero R es usualmente una matriz identidad [15].

El método BLUP tiene algunas variantes, en las que podemos encontrar el ABLUP (Traditional BLUP) y GBLUP (Genomic BLUP). La diferencia entre estas dos variantes del método BLUP recae en que para el ABLUP o BLUP tradicional, la Matriz de Relaciones Genéticas, llamada matriz A, utilizada para la solución, está basada solo en los datos de pedigrí de los animales en el grupo de estudio, mientras que en el GBLUP o BLUP Genómico, dicha matriz es definida como Matriz de Relaciones Genómicas, llamada matriz G, y se calcula a partir de la matriz de incidencia de marcadores genéticos (SNPs), los cuales son obtenidos a través de la genotipificación del ADN de cada animal, utilizando SNPchips.

2.2 Método GBLUP (Genomic Best Linear Unbiased Prediction).

GBLUP es un método que utiliza relaciones genómicas para estimar el mérito genético de un individuo. Para este propósito, se necesita una matriz de relaciones genómicas, estimada de la información de los marcadores de ADN. La matriz define la covarianza entre individuos basado en similitud observada a nivel genómico, en lugar de la similitud esperada basada en el pedigrí, de modo que las predicciones de mérito genético estimadas pueden ser más exactas.

2.2.1 Matriz de Relaciones Genómicas.

Para obtener la matriz de relaciones genómicas, o matriz **G**, partimos de la Matriz **MAF** (Minor Allele Frequency matrix). Para genotipos, los elementos en esta matriz **MAF** son valores de 0, 1 o 2, 0 para homocigotos alelo mayor, 1 para heterocigotos y 2 para homocigotos alelo menor, las dimensiones de la matriz **MAF** son número de individuos (**n**) por número de marcadores (**m**)[23] .

El siguiente paso es obtener la matriz **M** del resultado de **MAF -1**, con lo que los valores de los elementos en la matriz **M** serían de -1 para homocigotos alelo mayor, 0 para heterocigotos y 1 para homocigotos alelo menor. También se necesita calcular una matriz **P**, las columnas de la matriz **P** son las frecuencias de alelos expresadas como $P_i = 2(p_i - 0.5)$, donde p_i es la frecuencia de alelo menor del marcador *i*. Obteniendo las matrices **M** y **P** ya podemos calcular la matriz **Z**, para esto sustraemos la matriz **P** de **M** ($Z = M - P$). Sustraer **P** de **M** da mayor crédito a los alelos raros que a los alelos comunes cuando calculamos relaciones genómicas[23].

Para calcular la matriz **G** o matriz de relaciones genómicas utilizando marcadores SNPs existen diferentes métodos de los cuales vamos a exponer tres de estos métodos [23]. El **primero de los métodos** para obtener **G** usa la siguiente fórmula,

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2 \sum p_i(1-p_i)} \quad (5)$$

La división entre $2 \sum p_i(1-p_i)$ escala a **G** para ser análoga a la matriz de relaciones genéticas (matriz **A** del método ABLUP). Este método requiere el cálculo de la frecuencia de los alelos en la matriz de incidencia de los marcadores [23].

El **segundo método** para obtener **G** esta dado por la siguiente fórmula,

$$\mathbf{G} = \mathbf{ZDZ}' \quad (6)$$

Donde **D** es una diagonal con $D_{ii} = \frac{1}{m[2 p_i(1-p_i)]}$. Al igual que el primer método requiere el cálculo de la frecuencia de los alelos en la matriz de incidencia de los marcadores [23]. Esta fórmula fue propuesta para estudios genéticos en humanos.

El **tercer método** para obtener **G** no requiere la frecuencia de los alelos y en cambio se ajusta a la homocigocidad media por regresión **MM'** en **A**(matriz de relaciones Genéticas) para obtener **G** [23] usando el modelo

$$\mathbf{MM}' = g_0 \mathbf{11}' + g_1 \mathbf{A} + \mathbf{E} \quad (7)$$

Donde g_0 es la intersección y g_1 es la pendiente. Y para obtener **G** se usa la formula

$$\mathbf{G} = \frac{\mathbf{MM}' - g_0 \mathbf{11}'}{g_1} \quad (8)$$

La matriz **G** es positiva semidefinida para los 2 primeros métodos, pero puede ser singular si el número de loci es limitado o si 2 individuos tienen genotipos idénticos; **G** debe ser singular si el número de marcadores genéticos (SNPs) es menor al

número de individuos. Gemelos idénticos o clones pueden causar singularidad aun en \mathbf{A} . Una matriz \mathbf{Gw} mejorada y no singular se puede obtener como el ponderado (\mathbf{w}), $\mathbf{wG} + (1 - \mathbf{w})\mathbf{A}$ si el numero de marcadores es limitado y \mathbf{A} no es singular.

Donde $\mathbf{w} = \frac{0.05^2}{(0.05^2 + \frac{0.125}{m})}$, m es el número de marcadores. \mathbf{G} debería obtener más peso que \mathbf{A} si $m > 50$ y casi todo el peso (>0.99) si $m > 5000$ [23].

2.2.2 Estimado del Valor Genómico de Crianza (GEBV).

La solución propuesta por Henderson para calcular el valor genómico de crianza implica resolver (4) el sistema de Ecuaciones del Modelo Mixto (MME), el cual, adecuado para datos genómicos quedaría como

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (9)$$

Donde $\lambda = \sigma_e^2/\sigma_u^2$, el factor de índice de reducción en GBLUP, σ_e^2 es la varianza del vector de residuales y σ_u^2 es la varianza aditiva genética [24].

También, la ecuación de índice de selección es una solución al GBLUP para predecir el valor genómico \mathbf{u} directamente y evitar todos los pasos que implicaría resolver las Ecuaciones del Modelo Mixto. Esta ecuación nos permite obtener los Estimados del Valor Genómico de Crianza (**GEBV**, Genomic Estimated Breeding Value) en un solo paso.

Cuando derivamos la matriz \mathbf{G} utilizando los primeros 2 métodos expuestos en el capítulo 2, utilizando las frecuencias de los alelos, la ecuación de índice de selección para obtener los **GEBV**($\hat{\mathbf{u}}$) que se sugiere es,

$$\mathbf{GEBV}(\hat{\mathbf{u}}) = \mathbf{G} [\mathbf{G} + \mathbf{R}\lambda]^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad (10)$$

Donde, $\mathbf{X}\hat{\mathbf{b}}$ (efectos fijos) es la media de las observaciones (\mathbf{y}) [24].

Otra solución, si utilizamos la matriz G derivada el tercer método descrito (de regresión), para obtener los Estimados del Valor Genómico de Crianza se sugiere utilizar una solución diferente, esta es representada con la siguiente ecuación,

$$GEBV(\hat{u}) = [R^{-1} + G^{-1}\lambda]^{-1} R^{-1}(y - X\hat{b}) \quad (11)$$

Esta solución resulta más eficiente, ya que G se invierte solo una vez y entonces rasgos adicionales con heredabilidad diferente o R pueden procesarse con iteraciones [24].

2.3 Interpretación práctica del Estimado Valor Genómico de Crianza (GEBV).

El BLUP con su naturaleza de modelo de mezclas integra datos de distinta naturaleza para estimar el efecto que estos tienen, de forma simultánea y/o correlacionada, sobre un evento o medición. En el contexto de la mejora genética integra por un lado datos del pedigrí, en los cuales viene implícita la aptitud heredada de su linaje, incluyendo el grado de adaptabilidad que se ha logrado a las condiciones ambientales, o el grado de afectación de la misma; la genética propia del individuo, la cual es mérito individual y actúa respondiendo a las condiciones de crianza, de manejo, de salud, de medio ambiente y de presión de producción, sumado a factores aleatorios impredecibles, para lograr los valores fenotípicos que lo caracterizan.

De lo anterior que, el Valor Genómico de Crianza, medido como efectos aleatorios en el modelo de mezclas, indica como la genética propia del individuo influye en las observaciones de su rendimiento.

2.4 Cálculo de Heredabilidad.

Además de la estimación de los Valores Genómicos de Crianza, en este trabajo de tesis se calculan otros parámetros necesarios en la mejora genética, uno de ellos es la heredabilidad, la cual es un concepto que resume que tanto de la variación de un rasgo es debido a la variación de factores genéticos. A menudo, este término es usado en referencia a la semejanza entre padres e hijos. En este contexto una alta heredabilidad implica una semejanza fuerte entre padres e hijos con respecto a un rasgo específico, mientras que una baja heredabilidad implica un nivel bajo de semejanza.

En nuestro estudio, es necesario medir la proporción de la varianza fenotípica total que está determinada por la varianza aditiva genética, lo cual es llamado Heredabilidad en Sentido Estricto (*“Narrow Sense Heritability”*) y representada como h^2 ; y se calcula de la siguiente de la siguiente forma:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \quad (12)$$

Donde σ_a^2 es la varianza aditiva genética y σ_e^2 es la varianza de los residuales. La heredabilidad puede variar de 0 a 1, donde una heredabilidad de 0 representa un rasgo que no tiene una contribución genética a la herencia y una heredabilidad de 1 representa un rasgo que no tiene una contribución ambiental a la herencia.

2.5 Cálculo del coeficiente de Endogamia.

La endogamia es simplemente el cruce entre dos individuos emparentados, lo que significa que tienen uno o más antepasados en común. Cuanto más cercano sea el

parentesco entre dos animales, mayor es el porcentaje de consanguinidad en la progenie resultante.

Es así como la endogamia constituye una práctica común en el desarrollo y la propagación de las razas de animales para intentar concentrar características de interés económico en la progenie. Para muchas especies pecuarias, los caracteres de importancia económica son aquellos que afectan la productividad, longevidad, salud y capacidad reproductiva de los animales.

El coeficiente de endogamia es la reducción promedio de la heterocigocidad en un individuo debido al apareamiento no aleatorio dentro de una subpoblación. Esta se calcula de la siguiente forma:

$$F_{IS} = \frac{H_S - H_I}{H_S} \quad (13)$$

Donde F_{IS} es una medida del grado de endogamia (consanguinidad) dentro de una subpoblación, H_S es la heterocigocidad esperada entre subpoblaciones y H_I es la heterocigocidad promedio observada por individuo dentro de una subpoblación. Este coeficiente varía desde -1 donde todos los individuos son heterocigotos a 1 donde no hay heterocigotos observados.

2.6 Cálculo de la Precisión de los Valores Genómicos de Crianza.

La precisión de los Valores Genómicos de Crianza indica que tan bien esta estimación representa el verdadero Valor de Crianza. En otras palabras, representa la correlación entre el Valor Genómico de Crianza Estimado y el verdadero. Esta correlación puede tener valores de 0 que significa que la estimación es totalmente inexacta, a 1 que significa que el Valor Genómico e Crianza Estimado es el predictor

ideal del valor de crianza real. La precisión del Valor Genómico de Crianza Estimado se calcula de la siguiente forma:

$$\sqrt{\frac{1-PEV}{\sigma_a^2}} \quad (13)$$

Donde ***PEV*** es la varianza del vector de residuales, y σ_a^2 es la varianza aditiva genética.

3. IMPLEMENTACIÓN.

En este trabajo de tesis se implementó una Herramienta bioinformática para la administración de hatos ganaderos, incluyendo mejora genética basada en selección genómica. Para esto se programó una aplicación para escritorio de Windows, el cual incluye parte de la administración que requiere un hato ganadero; tal como registro de los animales del hato, registro de nacimiento de animales dentro del hato, registro de los datos genómicos en caso de tenerlos, registros de producción, gráficas de producción, y otras características importantes para el cumplimiento de los objetivos del proyecto; tales como el cálculo de valores genómicos, cálculo de la heredabilidad de algún rasgo en específico, cálculo del coeficiente de consanguinidad así como el cálculo de la precisión de los valores genómicos de crianza estimados.

A continuación, se describe el funcionamiento de las ventanas del ambiente gráfico, y funcionamiento del software implementado. Para la descripción de la aplicación, hacemos la dividiremos en 2 partes, la primera como las herramientas para la administración del hato ganadero y la segunda parte como la parte de la sección de

uso y manejo de datos de genotipos de genoma completo para el cálculo de los Valores Genómicos.

3.1 Herramienta bioinformática para la administración de hato Ganaderos.

Dentro de la aplicación desarrollada se creó una parte enfocada a la administración de un hato ganadero, el cual se describirá a continuación.

La venta principal de dicha aplicación se muestra en la siguiente imagen:

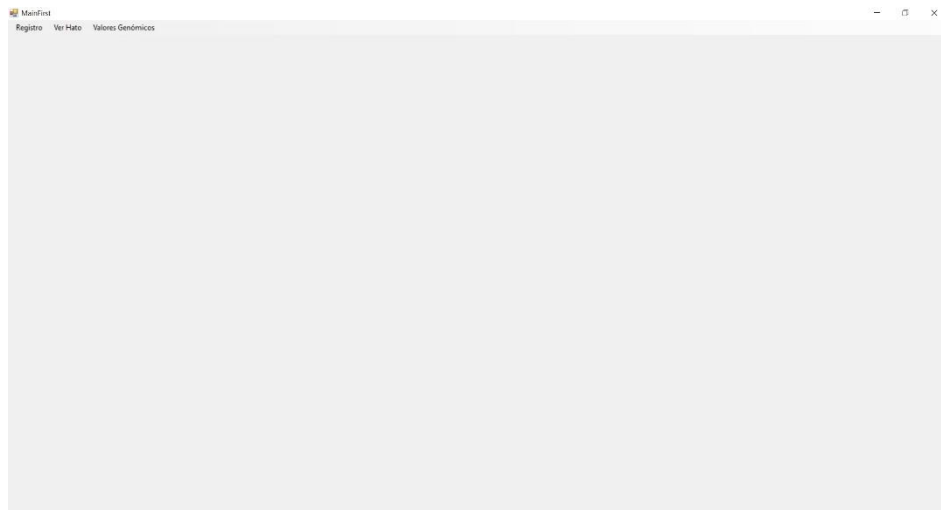


Ilustración 7. Ventana principal del software implementado.

En esta ventana encontramos como menú principal tres opciones, *Registro*, *Ver Hato* y *Valores genómicos*. Como se muestra en las figuras 8 y 9.

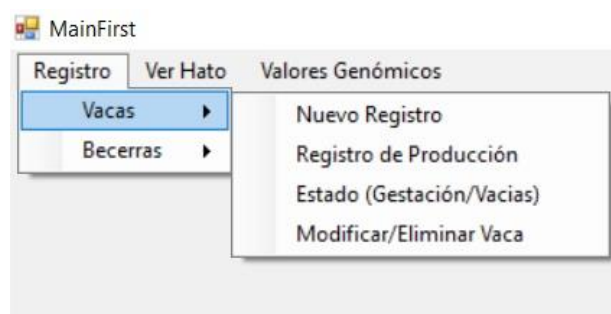


Ilustración 8. Submenú de la opción de Vacas.



Ilustración 9. Submenú de la opción Becerras

En la figura 8 podemos observar las opciones que ofrece el submenú de registro en su opción de Vacas. Como podemos ver, es posible hacer un nuevo registro (estas pueden ser vacas nuevas, incorporadas al hato, ya en producción), al momento de que una nueva vaca sea añadida al hato, o al registro. Igualmente podemos ver los registros de producción de cada vaca. Esto incluye el historial de producción de leche, de proteína en la leche, de células somáticas, etc., incluyendo todas las variables de interés del productor. De igual forma podemos ver el estado actual de cada vaca, incluyendo si está en gestación y/o produciendo leche o no. Finalmente, es posible editar los datos de cada registro y modificar la información.

En la figura 9 se observa la subsección de *Becerras*, la cual es usada para dar de alta becerras nuevas al registro. Igualmente, puede editarse el registro de las becerras existentes en el hato para realizar modificaciones a la información existente.

A continuación se presenta de forma gráfica la información contenida en ambas opciones del menú de *Registro*.

En la figura 10 se observa la venta de *Nuevos Registros de Vacas*. En esta hacemos registros de nuevos animales en el hato: se registran datos básicos como nombre, padre, madre, raza, fecha de nacimiento, la condición en que llegó al hato, así como opción de guardar los datos del genotipo del animal, siempre y cuando se cuenten con ellos.

The screenshot shows a software window titled "Nuevos Registros Vacas". It is divided into two main sections: "Datos Generales" and "Datos Genómicos".

Datos Generales:

- No. de Vaca:** A text box containing "199".
- Fecha de Nacimiento:** A date picker showing "6/20/2019".
- Condition:** Three checkboxes: "Nacida en el establo" (checked), "Comprada Seca", and "Comprada en Leche".
- Registration Table:** A table with three columns: "Nombre", "Registro", and "Raza".

	Nombre	Registro	Raza
Vaca	199	xxxxxx	H
Padre	Toro1	xxxxxxx	H
Madre	Vaca1	xxxxxxxx	H

Datos Genómicos:

- Genotipos:** A checkbox that is checked.
- Cargar:** A button.
- Fecha de Registro:** A date picker showing "6/20/2019".

At the bottom center is a button labeled "Agregar Registro".

Ilustración 10. Ventana de Registro de vacas nuevas.

En la figura 11 se observa la ventana de *Registro de Producción*. En esta se muestra como se registramos los pesajes de la producción de cada vaca individualmente. Se selecciona el número de vaca, y por último se selecciona y se registra el peso de su producción en cada una de las ordeñas que se tuvieron en la fecha de la medición.

Ilustración 11. Ventana de Registro de Producción.

En la figura 12 se observa la ventana de registro de *Estado de la Vaca*, en esta se actualiza su estado actual, ya sea que esté preñada o está vacía.

Ilustración 12. Registro de la Vaca.

En la figura 13 se observa la ventana de la opción *Modificar Registro*, que es utilizada para editar cualquier registro existente. En esta se, podemos modificar los datos generales de cada una de las vacas en nuestro hato, también presentamos la opción de eliminar el registro de nuestro hato. Es importante tener cuidado que datos y por qué vamos a modificar y/o eliminarlos, al eliminarlos ya no podremos recuperarlos.

Modificar_Registro

Datos Generales

No. de Animal: 65

Fecha de Nacimiento: 6/20/2019

Edad: 0

	Nombre	Registro	Raza
Vaca	Vaca65	xxxxxx1	H
Padre	Toro1	xxxxxx0	H
Madre	Vaca2	xxxxxx	H

Estado: Preñada

Desde: 6/20/2019

Genotipos: ☒

Cargar

Modificar Eliminar

Ilustración 13. Ventana de Modificación de Registros o su Eliminación.

En la figura 14 se observa la ventana de *Nuevos Registros Becerras*. En esta podemos registrar los nuevos nacimientos y que tipo de nacimiento fue, así como su padre, madre, raza, fecha de nacimiento, y sus datos genómicos si se cuentan con ellos.

Nuevos Registros Becerras

Datos Generales

No. de Becerra: 256

Madre: 65

Fecha de Nacimiento: 6/20/2019

Tipo de Cria: ☒ Normal ☐ Gemelar

Datos Genómicos

☒ Genotipos

Cargar

	Nombre	Registro	Raza
Vaca	256	xxxxxx256	H
Padre	Toro2	xxxxxxx1	H
Madre	65	xxxxxxx2	H

Fecha de Registro: 6/20/2019

Agregar Registro

Ilustración 14. Ventana para Nuevos Registros de Becerras.

En figura 15 se observa la ventana de la opción *Ver Hato*. En esta se presenta un submenú en el que encontramos las opciones: *Datos de Vacas*, *Datos de Becerras* y *Análisis de Producción*.

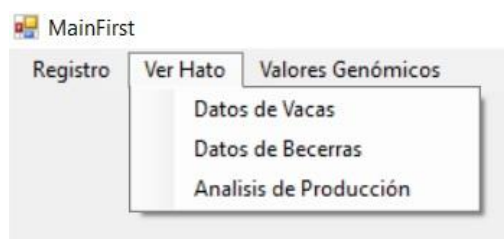


Ilustración 15. Submenú de la opción Ver Hato.

En la figura 16 observamos la ventana de *Datos de las Vacas*. En esta ventana mostramos los datos más importantes de manera individual de cada vaca en nuestros registros, además se muestra una tabla de los registros de producción por fecha, los cuales también se muestran en una gráfica de línea donde X son las fechas de registro de la producción y Y es el registro en Kilogramos.

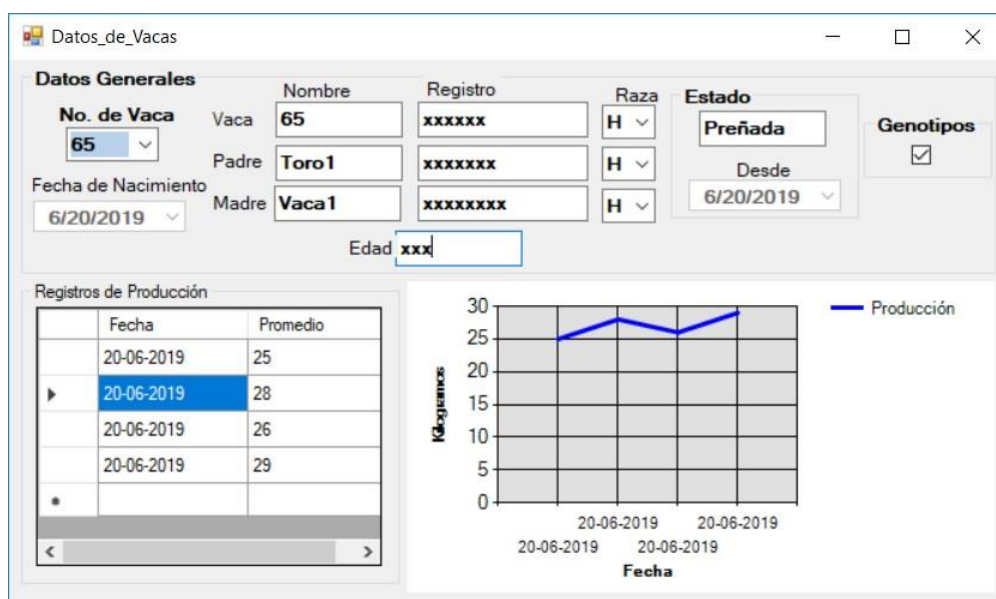


Ilustración 16. Ventana de Datos de Vacas.

En la figura 17 observamos la venta de *Datos de Becerra*. En esta podemos consultar los datos generales de las becerras registradas en nuestro hato.

Datos de Becerras

Datos Generales

No. de Becerra: 256

Fecha de Nacimiento: 6/20/2019

Nombre: Vaca

Padre: Toro2

Madre: 65

Edad: 0

Registro: xxxxxx

Raza: H

Estado: xxxx

Desde: 6/20/2019

Genotipos: ☒

Ilustración 17. Ventana Datos de Becerras.

En la figura 18 observamos la ventana de *Análisis de Producción*. En esta podemos consultar la producción registrada por número de lactancia, por fecha, o por un periodo de tiempo. En la parte de abajo se muestra la tabla de registros de producción encontrados y se representan en la gráfica de línea en la parte de arriba.

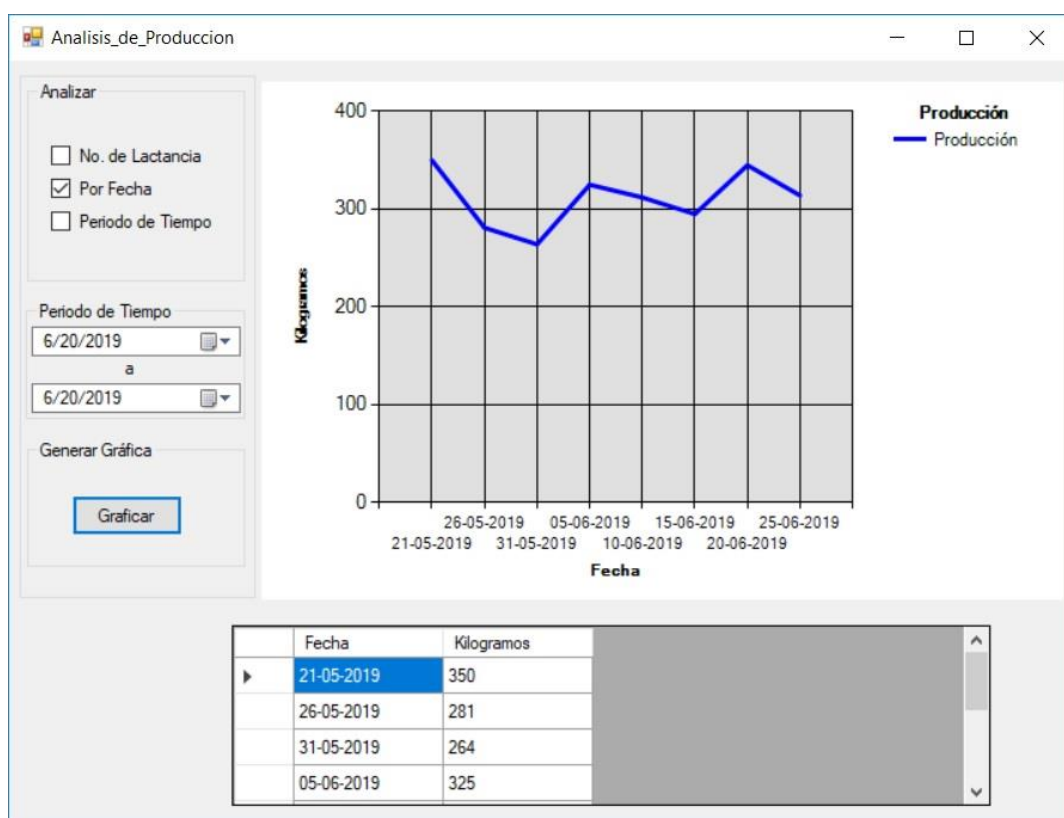


Ilustración 18. Ventana de Análisis de Producción.

En la figura 19 se observa el submenú de la opción de *Valores Genómicos*. En esta encontramos dos opciones, *Generación de Archivos* y *Herramienta para el Cálculo*, a continuación, describiremos ambas opciones.

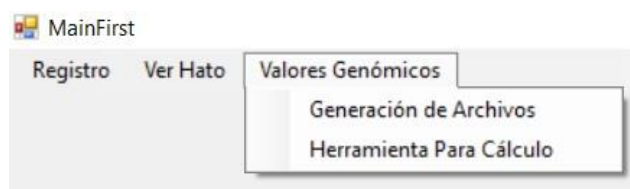


Ilustración 19. Submenú de la opción Valores Genómicos.

En la opción *Generación de Archivos* son desplegadas 3 opciones: Genotipos, Pedigrí y Fenotipos, como se observa en la figura 20. Estos nos sirven para generar los archivos necesarios para el cálculo de los Valores Genómicos de Crianza.

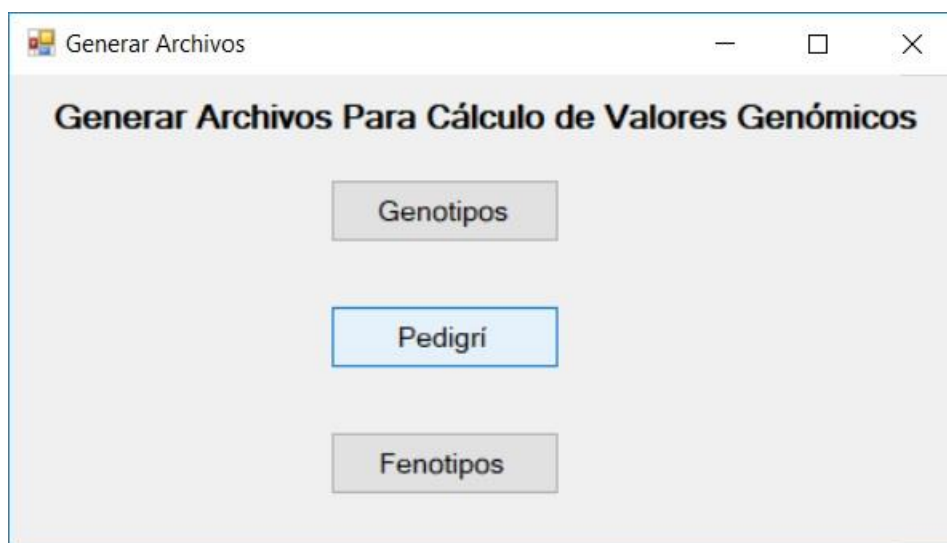


Ilustración 20. Ventana Generación de Archivos.

La opción *Herramienta Para Cálculo*, dentro del submenú *Valores Genómicos* abre la Herramienta desarrollada para el cálculo de Valores Genómicos de Crianza, la cual se describirá en la siguiente sección de este capítulo.

3.2 Datos necesarios para el cálculo de Los Valores Genómicos de Crianza.

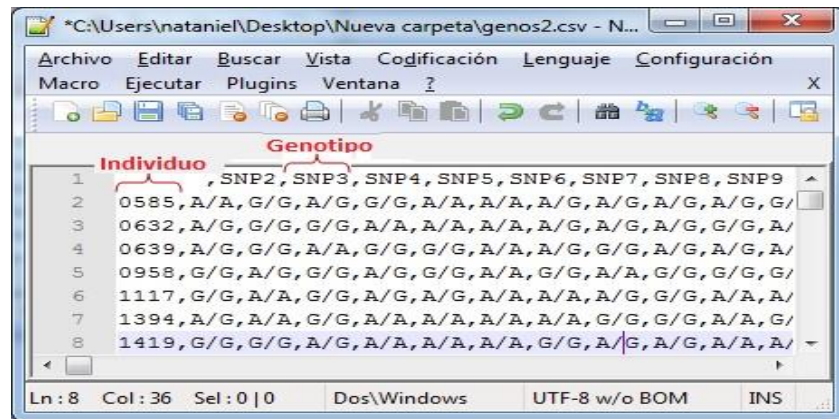
En esta sección se describen los datos a utilizar, el manejo de los archivos que contienen estos datos, y como a partir de éstos se van derivando las matrices necesarias para el cálculo de los valores genómicos, bajo ambiente R; además de la descripción de la aplicación desarrollada para el cálculo de los Valores Genómicos de Crianza.

Para el cálculo de los Valores Genómicos (GEBVs), utilizando el método GBLUP es preciso contar con los genotipos (matriz de incidencia de los marcadores de ADN, SNPs), fenotipos (estudio longitudinal del rasgo que queremos mejorar) y la información de pedigrí de los individuos en el grupo de estudio. Los genotipos son obtenidos a partir de la genotipificación del ADN de cada individuo utilizando SNPchips tales como los de la empresa Illumina [9][11][12], expuestos anteriormente. A las lecturas resultantes (marcadores SNPs) se debe de aplicar ciertos filtros para elevar la calidad de estos datos; tales como eliminar SNPs cuyos datos faltantes sean mayores a 10% (estos datos faltantes pueden deberse a errores de lectura en el SNPchip), eliminar SNPs que tengan una frecuencia de alelo menor <0.05 , eliminar SNPs monomórficos (SNPs que solo tienen el mismo alelo para todos los individuos), entre otros. Los fenotipos y pedigrí es información la cual de manera estándar los administradores de hatos llevan registros y actualizan constantemente.

3.2.1 Matriz de incidencia de Genotipos de los marcadores (SNPs).

La matriz de genotipos es la matriz que contiene el valor en pares de los SNPs para cada individuo en el grupo de estudio. En pares se refiere a que el genotipo de un SNP específico consta de dos alelos, de los cuales, uno fue heredado por el padre y el

otro por la madre del individuo. Por lo general, esta información se maneja en archivos donde los datos en un renglón son separados por comas, y con extensión .CSV.



The screenshot shows a text editor window titled '*C:\Users\natanie\\Desktop\Nueva carpeta\genos2.csv - N...'. The menu bar includes Archivo, Editar, Buscar, Vista, Codificación, Lenguaje, Configuración, Macro, Ejecutar, Plugins, and Ventana. The toolbar contains various icons for file operations. The text area displays a table with two main headers: 'Individuo' and 'Genotipo'. The 'Individuo' column lists IDs: 0585, 0632, 0639, 0958, 1117, 1394, and 1419. The 'Genotipo' column lists genotypes for SNPs 2 through 9, separated by commas. The status bar at the bottom shows 'Ln : 8 Col : 36 Sel : 0 | 0 Dos\Windows UTF-8 w/o BOM INS'.

Individuo	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
0585	A/A	G/G	A/G	G/G	A/A	A/A	A/G	A/G
0632	A/G	G/G	G/G	A/A	A/A	A/A	A/G	A/G
0639	A/G	G/G	A/A	G/G	G/G	A/A	A/G	A/G
0958	G/G	A/G	G/G	A/G	G/G	A/A	G/G	A/A
1117	G/G	A/A	G/G	A/G	A/A	A/A	A/G	G/G
1394	A/G	A/A	G/G	A/A	A/A	A/A	G/G	A/A
1419	G/G	G/G	A/G	A/A	A/A	A/A	G/G	A/A

Ilustración 21. Tabla de genotipos vista en un editor de texto.

El primer renglón corresponde al nombre o identificación de cada uno de los SNPs, a partir del segundo renglón, como podemos observar el primer dato de cada renglón corresponde al nombre del individuo, seguidos de los genotipos, separados por "," (coma).

3.2.1.1 Matriz de Frecuencia de Alelo Menor.

La matriz **MAF** (Minor Allele Frequency), es una matriz donde los genotipos son representados con un número, ya sea 0 para homocigoto alelo mayor, 1 para heterocigotos y 2 para homocigoto alelo menor, a esta codificación se le denomina contenido genético, esta información es usualmente manejada en un archivo con extensión .CSV.

Individuo	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
1	585	0.00	2.00	1.00	2.00	0.00	0.00	1.00
2	632	1.00	2.00	2.00	0.00	0.00	0.00	1.00
3	639	1.00	2.00	0.00	2.00	2.00	0.00	1.00
4	958	2.00	1.00	2.00	1.00	2.00	0.00	2.00
5	1117	2.00	0.00	2.00	1.00	1.00	0.00	1.00
6	1394	1.00	0.00	2.00	0.00	0.00	0.00	2.00
7	1419	2.00	2.00	1.00	0.00	0.00	0.00	2.00
8	1438	2.00	2.00	1.00	0.00	0.00	0.00	2.00

Ilustración 22. Tabla de genotipos vista en un editor de texto, codificada en 0, 1 y 2.

3.2.2 Fenotipos.

Como se mencionó, los fenotipos corresponden al estudio longitudinal de un rasgo en específico, o bien, mediciones continuas del rasgo en el que estamos interesados durante un periodo de tiempo, ya sea producción de leche, carne, etc.

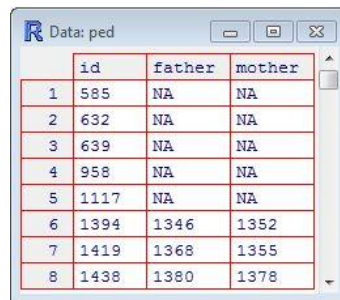
row.names	Fenotipos1
1	585
2	632
3	639
4	958
5	1117
6	1394
7	1419
8	1438

Ilustración 23. Tabla de fenotipos vista en ambiente R.

Como se observa en la tabla, la primera columna corresponde al nombre del individuo y la segunda columna al valor del fenotipo. En esta tabla también se pueden incluir datos de más de un fenotipo, solo incrementando las columnas a la derecha y nombrando la columna con el número de fenotipo y/o nombre del rasgo.

3.2.3 Pedigrí.

La información del pedigrí nos ayudará a establecer una relación de parentesco entre los individuos en el grupo de estudio, además, a partir de esta información podemos calcular la matriz de relaciones genéticas (matriz **A**).



	id	father	mother
1	585	NA	NA
2	632	NA	NA
3	639	NA	NA
4	958	NA	NA
5	1117	NA	NA
6	1394	1346	1352
7	1419	1368	1355
8	1438	1380	1378

Ilustración 24. Tabla de pedigrí vista en ambiente R.

Como se observa en la ilustración, esta tabla cuenta con 3 columnas, la primera corresponde al nombre del individuo, la segunda al nombre del su padre y la tercera al nombre de su madre. Los **NA** que se observan en la tabla de pedigrí de la Ilustración 5, se refiere a que el nombre del padre o de la madre, según corresponda a la columna, se desconoce.

La razón de por qué usar archivos donde los datos son separados por "," (coma), es que con este tipo de archivos se nos facilita la organización y manejo de los datos en ambiente R, además que este tipo de archivos son también manipulables en la hoja de cálculo de Excel.

3.3 Matriz de Relaciones Genómicas utilizando frecuencias de los alelos en ambiente R.

Para explicar la implementación en R del cálculo de la matriz G vamos a asumir que tenemos 4 individuos con 6 SNPs cada uno.

Matriz de Genotipos.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	GG	AA	AT	CC	AC	AG
Ind2	GA	GG	TT	CC	AA	GG
Ind3	AA	AG	AT	TT	CC	AA
Ind4	GG	AA	TT	CT	AA	GG

Esta matriz de genotipos es convertida a contenido genético contando los alelos de menor frecuencia, y otorgando el valor correspondiente a cada genotipo (0, 1 o 2). A esta matriz se le denomina Matriz **MAF** (Minor Allele Frequency) de tamaño n (número de individuos) x m (número de marcadores).

Matriz **MAF**.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	0	0	1	0	1	1
Ind2	1	2	0	0	0	0
Ind3	2	1	1	2	2	2
Ind4	0	0	0	1	0	0

Para iniciar nuestro cálculo en R debemos tener el archivo de datos con los genotipos codificados como se observa en la matriz **MAF**, abrir este archivo utilizando la función `read.csv()` (con los atributos necesarios, como el nombre y extensión del archivo), direccionando a un objeto llamado "**MAF**" el cual contendrá el contenido del archivo. La forma de ejecutarlo es la siguiente:

```
>MAF = read.table("MAF.csv",sep=",")
```

En R, se denomina objeto a las entidades que este crea y manipula, estos pueden ser variables, variables indexadas (arreglos), cadenas de caracteres, funciones, etc.

El siguiente paso es obtener la matriz **M**, para esto se resta 1 a todos los elementos de la matriz MAF generando valores de -1, 0 y 1. Para obtener la matriz **M** en entorno R utilizamos la siguiente instrucción,

```
>M = MAF - 1
```

Matriz **M**

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	-1	-1	0	-1	0	0
Ind2	0	1	-1	-1	-1	-1
Ind3	1	0	0	1	1	1
Ind4	-1	-1	-1	0	-1	-1

A partir de la matriz **M** ya estamos listos para calcular la matriz de relaciones genómicas entre cada par de individuos (matriz **G**).

Además, podemos obtener cierta información relevante a través del desarrollo de la matriz **M** en **MM'** y **M'M**, como a continuación describiremos. La Matriz **MM'**, que es el resultado de la multiplicación de la matriz **M** por su transpuesta **M'**.

```
>MMt = M %*% t(M)
```

Como se observa el operador “*” se refiere a multiplicación, y cuando está entre el símbolo “%” se indica que son matrices de dimensiones diferentes. La función **t()** obtiene la transpuesta de una matriz.

Matriz **MM'**

	SNP1	SNP2	SNP3	SNP4
Ind1	3	0	-2	2
Ind2	0	5	-3	2
Ind3	-2	-3	4	-3
Ind4	2	2	-3	5

Los elementos de la diagonal en esta matriz cuentan el número de SNPs homocigotos de cada individuo, y los elementos fuera de la diagonal miden el número de alelos compartidos por pariente. La matriz $M'M$, que es el resultado de la multiplicación de la transpuesta de M (M') por M .

```
>MtM= t(M)%*% M
```

Matriz $M'M$.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	3	2	1	2	2	2
Ind2	2	3	0	0	0	0
Ind3	1	0	2	1	2	2
Ind4	2	0	1	3	2	2
Ind4	2	0	2	2	3	3
Ind4	2	0	2	2	3	3

En esta matriz los elementos de la diagonal cuentan el número de individuos homocigotos para cada SNP, los elementos fuera de la diagonal miden el número de veces que alelos en diferente SNP fueron heredados por el mismo individuo.

El siguiente paso es calcular la matriz P , la matriz P contiene la frecuencia de los alelos expresada como $P_i = 2(p_i - 0.5)$, donde p_i es la frecuencia de alelo menor del **SNP** i . Partimos de la Matriz **MAF** para calcular las frecuencias de alelo menor de cada **SNP**. Las frecuencias de alelo menor según nuestra matriz **MAF**, serian para el **SNP1=0.375**, **SNP2=0.375**, **SNP3=0.25**, **SNP4=0.375**, **SNP5=0.375** y **SNP6=0.375**.

Entonces nuestra matriz P queda de la siguiente forma.

Matriz P .

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	-0.25	-0.25	-0.5	-0.25	-0.25	-0.25
Ind2	-0.25	-0.25	-0.5	-0.25	-0.25	-0.25

Ind3	-0.25	-0.25	-0.5	-0.25	-0.25	-0.25
Ind4	-0.25	-0.25	-0.5	-0.25	-0.25	-0.25

El siguiente paso es calcular la matriz **Z**, la cual es el resultado de sustraer **P** de **M**

$$Z = M - P$$

Matriz Z.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind1	-0.75	-0.75	0.5	-0.75	0.25	0.25
Ind2	0.25	1.25	-0.5	-0.75	-0.75	-0.75
Ind3	1.25	0.25	0.5	1.25	1.25	1.25
Ind4	-0.75	-0.75	-0.5	0.25	-0.75	-0.75

Para calcular la matriz **G** utilizando las frecuencias observadas de los alelos

empleamos la formula $G = \frac{ZZ'}{2 \sum p_i(1 - p_i)}$ descrita en el Capítulo 2. Donde $2 \sum p_i(1 - p_i) = 2.71875$. El producto de **Z** por su transpuesta **Z'** es igual a **ZZ'**.

$$ZZt = Z \%*\% t(Z)$$

Matriz ZZ'.

	Ind1	Ind2	Ind3	Ind4
Ind1	2.0625	-1.1875	-1.1875	0.3125
Ind2	-1.1875	3.5625	-2.4375	0.0625
Ind3	-1.1875	-2.4375	6.5625	-2.9375
Ind4	0.3125	0.0625	-2.9375	2.5625

Por lo tanto la matriz **G**, de tamaño **n x n**, sería

$$G = ZZt / 2 \sum p_i(1 - p_i)$$

	Ind1	Ind2	Ind3	Ind4
Ind1	0.758621	-0.43678	-0.43678	0.114943
Ind2	-0.43678	1.310345	-0.89655	0.022989
Ind3	-0.43678	-0.89655	2.413793	-1.08046
Ind4	0.114943	0.022989	-1.08046	0.942529

Obteniendo la matriz **G** utilizando el método de frecuencias observadas para calcular los Valores Genómicos se sugiere utilizar la ecuación de índice de selección (10) vista en el capítulo 2 de este trabajo, la cual es, $GEBV(\hat{u}) = G [G + R\lambda]^{-1} (y - X\hat{b})$, donde **R** es una matriz identidad de tamaño $n \times n$, y $\lambda = \sigma_e^2 / \sigma_u^2$. σ_e^2 y σ_u^2 son obtenidas utilizando el paquete regress() ya implementado en R.

Como ya vimos durante el Capítulo 2, también se puede usar el método de regresión de **MM'** en **A** como se muestra en (7) y utilizando (8). Si calculamos la matriz G por el método de regresión se recomienda usar (11) [24], también vista en el Capítulo 2.

Como observamos en el ejemplo de cómo calcular la matriz G, R nos facilita el manejo de datos indexados (matrices) al poder utilizar operadores aritméticos básicos como "*", "-", "/" para la manipulación de estos datos, entonces para realizar todos los cálculos se creó un Script en lenguaje R, el cual al ejecutarlo de manera secuencial desde la consola de comandos en R, calcula los valores genómicos utilizando dos diferentes métodos para calcular la matriz G (Frecuencias de los alelos y Regresión) y como solución final usando (9), (10) y (11) para calcular los GEBVs. También se desarrolló un ambiente gráfico que sea amigable al usuario, del que se creó un archivo ejecutable con extensión .EXE, el cual corre el Script de R en modo consola de comandos de MS-DOS evitando al usuario interactuar con la consola de comandos de R. Este ambiente grafico se le llamó GEBV.exe y cuenta con una serie de ventanas las cuales se explicará a continuación su funcionamiento.

3.4 Interfaz gráfica implementada.



Ilustración 25. Ventana Principal de GEBV.exe.

La ventana Principal es el cuerpo de este sistema de ventanas, desde esta se pueden operar todas las funcionalidades de la aplicación. En el centro de esta ventana se encuentran tres botones.

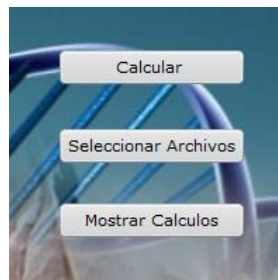


Ilustración 26. Botones en el centro de la Ventana Principal de GEBV.exe.

3.4.1 La opción Calcular.

El botón "**Calcular**" ejecuta el Script en R que calcula los valores genómicos, los resultados son guardados en varios archivos con extensión .CSV. Al dar clic en el botón "**Calcular**" se despliega la siguiente ventana como se aprecia en la imagen.



Ilustración 27. Ventana Calculando de GEBV.exe.

La barra de progreso nos indica el avance de los cálculos.

3.4.2 Seleccionar Archivos.

Este botón nos despliega la ventana que nos permite seleccionar los archivos necesarios para calcular los GEBVs. Estos archivos son, el de genotipos, fenotipos y pedigrí. Anteriormente se especificó el formato en el que deben de estar cada uno de ellos para poder utilizarlos en el Script en R. Para seleccionar los archivos se debe de dar clic en el ícono al final de la barra de selección (como se muestra en el círculo rojo), al seleccionar la ubicación de los 3 archivos se debe dar clic en el botón "**Guardar Archivos**", que se encuentra en la parte inferior de la ventana "**Seleccionar Archivos**".



Ilustración 28. Ventana Seleccionar Archivos de GEBV.exe.

Al dar clic en el botón "**Guardar Archivos**" se desplegará un aviso con la ubicación y nombre de los archivos seleccionados.

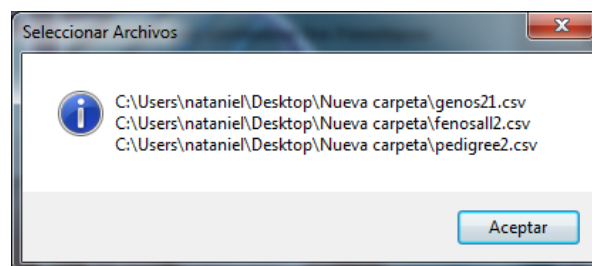


Ilustración 29. Subventana Seleccionar Archivos de GEBV.exe.

Al dar clic en aceptar se termina el proceso de seleccionar los archivos y automáticamente nos envía a la ventana "**Principal**".

3.4.3 Mostrar Cálculos.

Esta ventana muestra los cálculos obtenidos. Los cuales son desplegados en tablas como se muestra en la imagen.

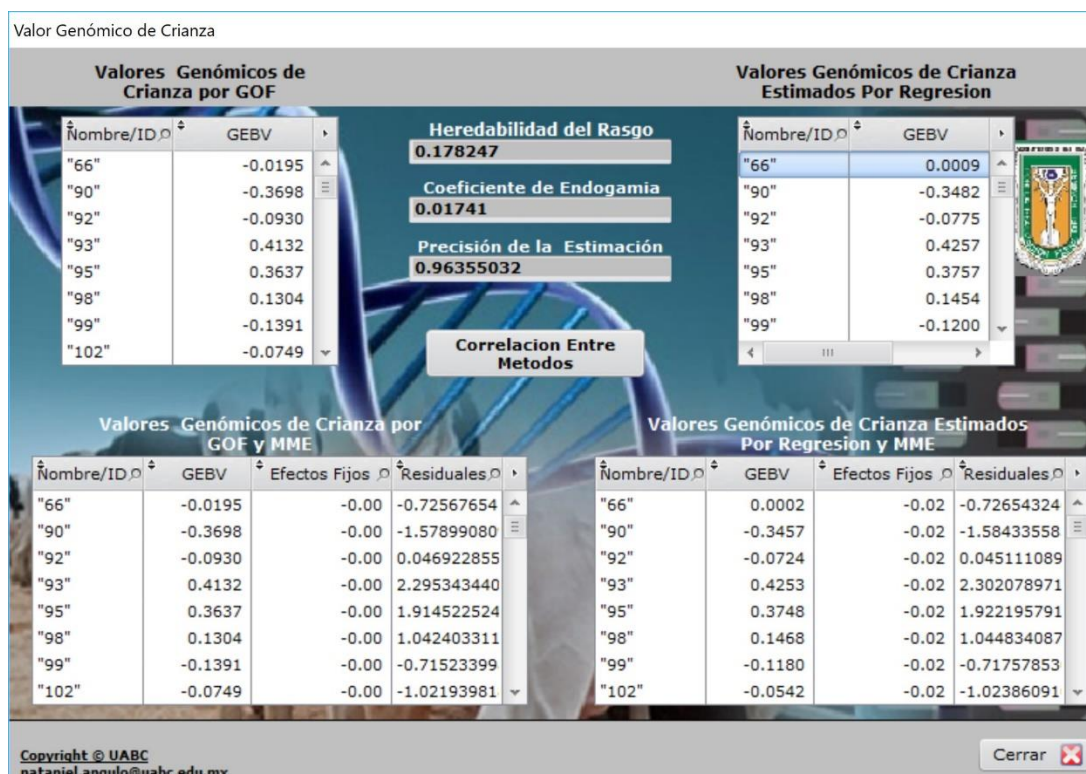


Ilustración 30. Ventana Valores Genómicos de Crianza de GEBV.exe.

En el encabezado de cada tabla se indica el método por el cual fueron calculados los valores genómicos. Cada una de las tablas cuenta con una columna de "**Nombre/ID**" seguida de otra llamada "**GEBV**", las cuales corresponden a la identificación y valor genómico respectivamente de cada animal. En las dos tablas que se encuentran en la parte inferior de esta ventana para calcular el valor genómico se utilizaron las ecuaciones del modelo mixto (MME), a través de este método podemos obtener conjuntamente el vector de efectos fijos (\hat{b}) y el vector de residuales (e). También encontramos, otros cálculos necesarios en la implementación de un programa de mejora genética, tales como la Heredabilidad del rasgo que queremos mejorar, el Coeficiente de Endogamia, por último, la precisión de la Estimación de los Valores Genómicos de Crianza.

Dentro de esta ventana podemos exportar los resultados de los cálculos, para ellos basta hacer un clic en la esquina superior derecha de la tabla que queremos exportar los datos como se muestra en la imagen, se despliega un menú Pop-up, el cual nos muestra diferentes opciones para este propósito.

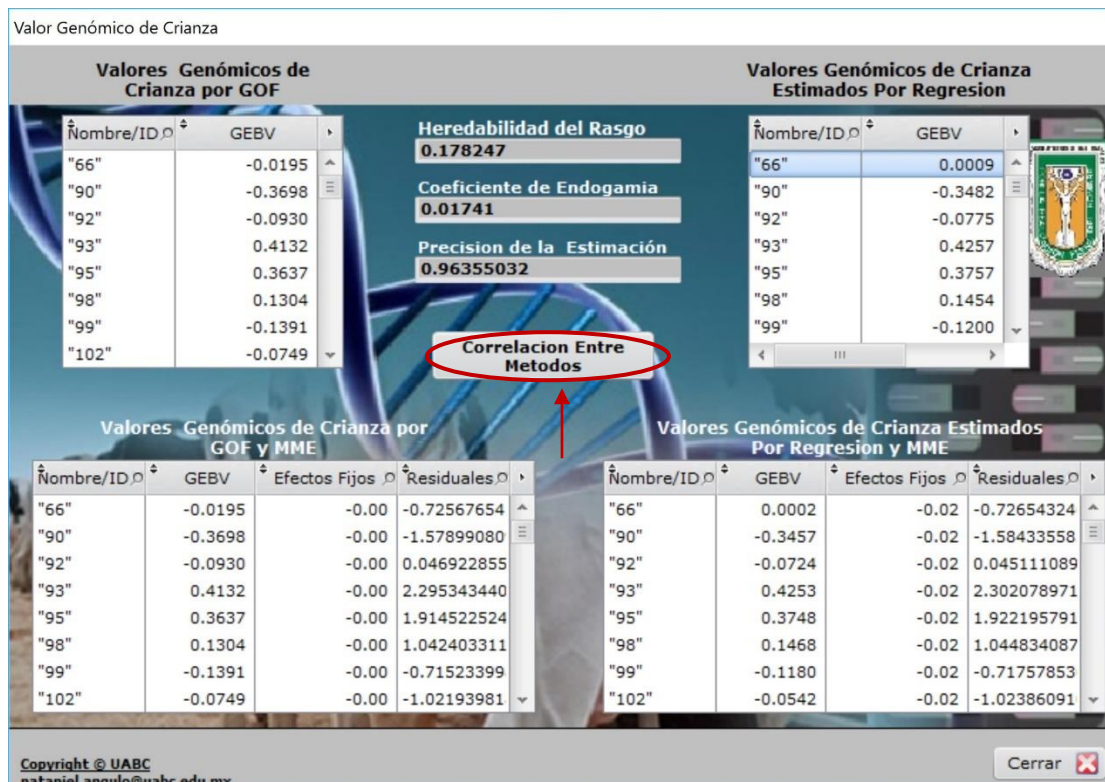


Ilustración 31. Descripción de ventana Valores Genómicos de Crianza de GEBV.exe.

En esta ventana se encuentra un botón llamado "**Correlación Entre Métodos**", la cual nos abre una subventana con las gráficas de correlación entre todos los métodos como se muestra a en la siguiente imagen.

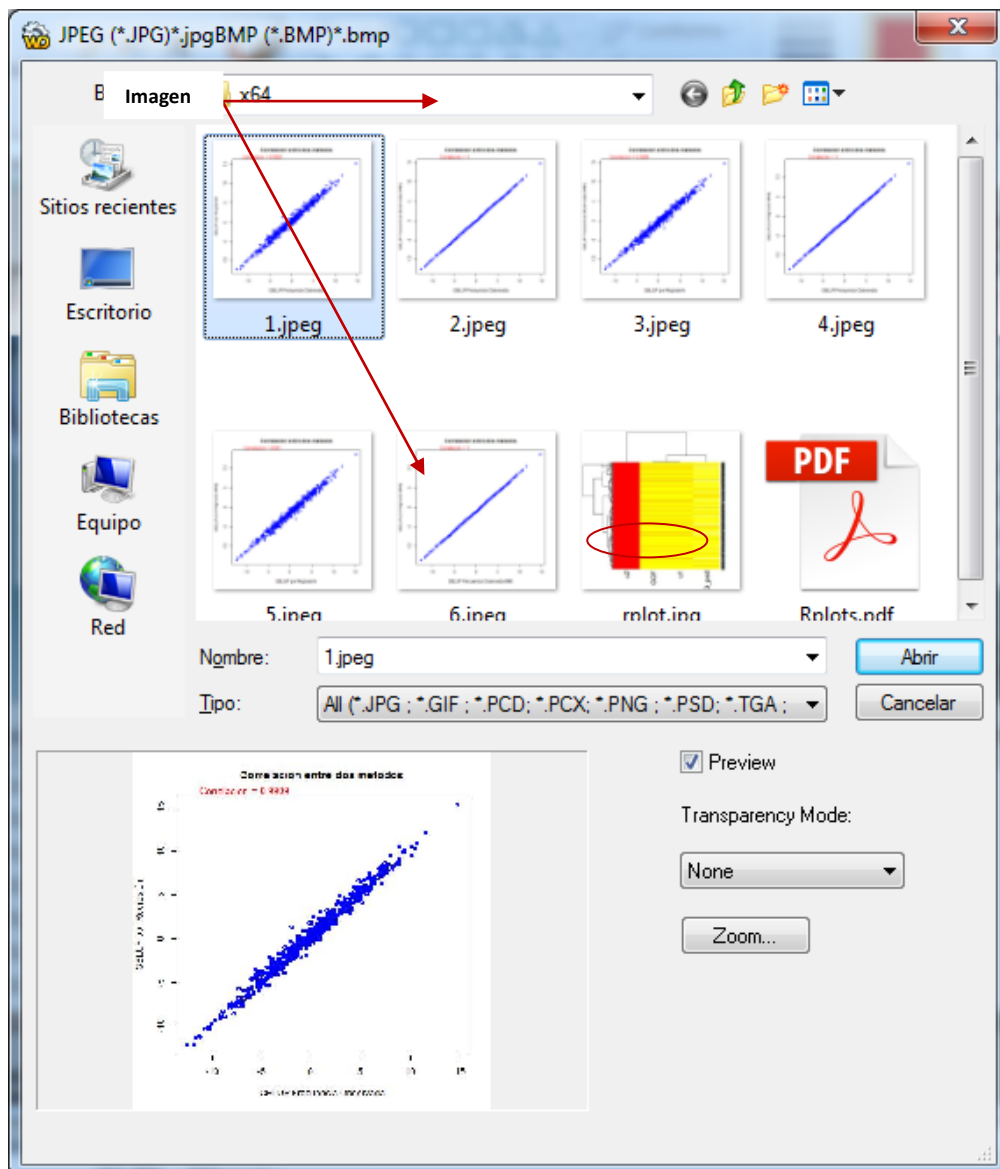


Ilustración 32. Ventana para mostrar Gráficas de GEBV.exe.

Al dar un clic sobre la imagen seleccionada la podremos ver en modo "**Preview**", para extender la imagen debemos dar clic en el botón "**Zoom**".

La ventana "**Principal**" cuenta con una barra de menú en su parte superior.

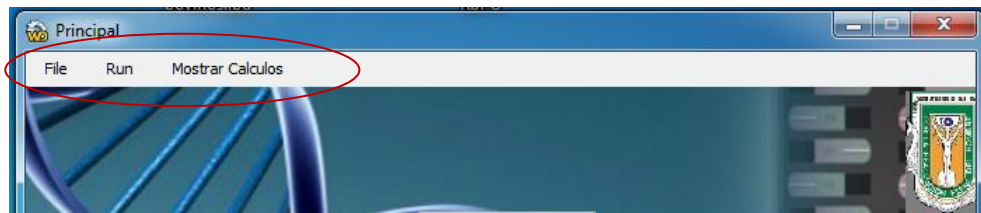


Ilustración 33. Barra de Menú en la Ventana Principal de Crianza de GEBV.exe.

En el submenú File podemos encontrar, dos opciones, "**Seleccionar Archivo**" y "**Cerrar**". "**Seleccionar Archivo**" nos lleva a la ventana de donde seleccionamos los tres archivos necesarios para realizar los cálculos de los valores genómicos, y la opción "**Cerrar**" sirve para cerrar por completo la aplicación.

En la barra de menú superior encontramos las opciones "**Run**" y "**Mostrar Cálculos**". Ambas opciones no cuentan con submenú. "**Run**" abre la ventana del cálculo de los valores genómicos y "**Mostrar Cálculos**" nos despliega el resultado de los cálculos una vez realizados.

4. PRUEBAS DEL MODELO Y ANÁLISIS DE RESULTADOS.

Para las pruebas del método implementado se utilizó un conjunto de datos que contiene genotipos, fenotipos y pedigrí de 500 toros. En este conjunto de datos cada individuo está etiquetado con una identificación única. Los Genotipos y el pedigrí fueron tomados de un conjunto de datos de ganado real, mientras que los fenotipos son hipotéticos. Los datos genotípicos consisten en 7250 marcadores SNPs para cada individuo en el grupo [26]. La información de entrada se manejó en tres archivos cuyo formato interno fue de datos separados por "," y con extensión **.CSV**, llamados "**genotipos.csv**", "**fenotipo.csv**" y "**pedigrí.csv**". Se utilizó el Script en R para hacer las diferentes corridas y calcular los valores genómicos de los 500 individuos para un fenotipo, utilizando matrices G derivadas de frecuencias alélicas de los marcadores SNPs y por método de regresión, y aplicando las diferentes soluciones expuestas en el Capítulo 2. Los resultados de los cálculos fueron llamados GEBVa (Matriz G por Frecuencias Alélicas y solución con 10) GEBVb (Matriz G por regresión y solución con 11), GEBVc (Matriz G por Frecuencias Alélicas y solución con 9), adicionalmente, se calcularon valores genómicos utilizando la matriz A (relaciones genéticas derivadas solo de pedigrí) a estos resultados se le llamó **A**. Se hicieron comparaciones entre los

resultados de diferentes métodos. La tabla 3 presenta una lista parcial de Valores Genómicos estimados para los datos de prueba.

Tabla 3. Lista parcial de Valores Genómicos calculados.

Nombre/ID	GEBVa	GEBVb	GEBVc	A
ID11430	-5.8358	-6.2155	-5.8358	-11.3695
ID11431	0.1005	-0.5880	0.1005	0.6604
ID11432	-5.0042	-5.1290	-5.0042	-6.5246
ID11433	-1.2909	-0.2433	-1.2909	-4.5475
ID11434	0.9323	1.7997	0.9323	-2.9489
ID11435	-0.7981	-0.0676	-0.7981	-0.5871
ID11436	-5.7081	-4.2169	-5.7081	-3.6692
ID11437	-3.9308	-3.9120	-3.9308	-7.0743
ID11438	-7.8363	-8.5829	-7.8363	-8.1255
ID11439	5.4315	5.9581	5.4315	5.8943
ID11440	-3.2646	-3.1041	-3.2646	-7.8390
ID11441	-8.2524	-7.5188	-8.2524	-10.7945

Con los vectores de los cuatro resultados de valores genómicos se creó una tabla de correlación obteniendo los siguientes resultados. La tabla 4 presenta las correlaciones resultantes.

Tabla 4. Correlación entre los resultados de los diferentes métodos.

	GEBVa	GEBVb	GEBVc	A
GEBVa	1.0000	0.9909	1.0000	0.9110
GEBVb	0.9909	1.0000	0.9909	0.8978
GEBVc	1.0000	0.9909	1.0000	0.9110
A	0.9110	0.8978	0.9110	1.0000

Los resultados de la correlación entre los métodos utilizados nos indican que los cálculos de GEBVa, GEBVb y GEBVc son casi idénticos, ya que se obtuvieron correlaciones de entre 0.99 y 1.00, en estos tres métodos se utilizaron datos genómicos. Los resultados obtenidos en **A** se acercan a los obtenidos con los otros 3

métodos, solo que la matriz de relaciones utilizada en **A** se derivó solamente de los datos de parentesco (pedigrí), lo cual sería la razón de que **A** no logra capturar efectivamente la relación a nivel genómico entre todos los individuos del grupo. A continuación se presentan las gráficas de correlación resultantes.

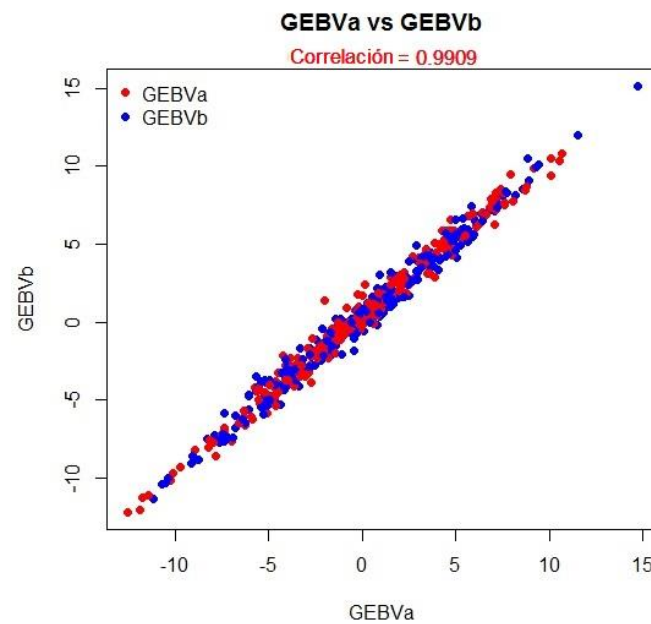


Ilustración 34. Correlación entre los GEBVs resultantes de los métodos GEBVa y GEBVb.

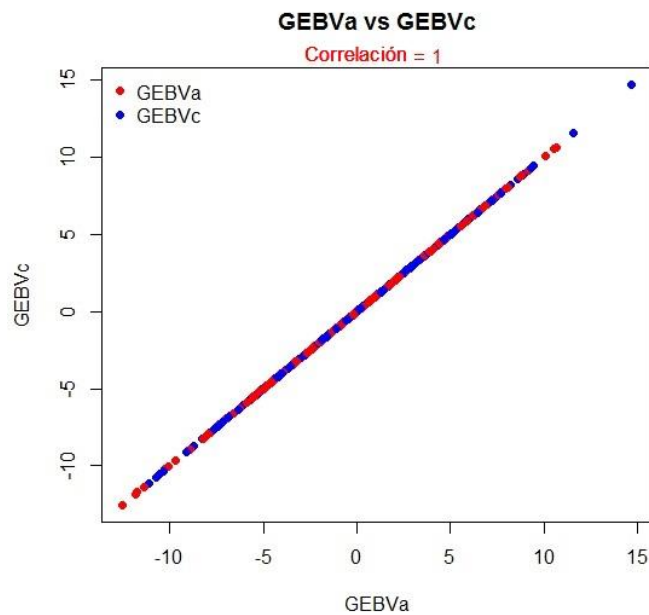


Ilustración 35. Correlación entre los GEBVs resultantes de los métodos GEBVa y GEBVc

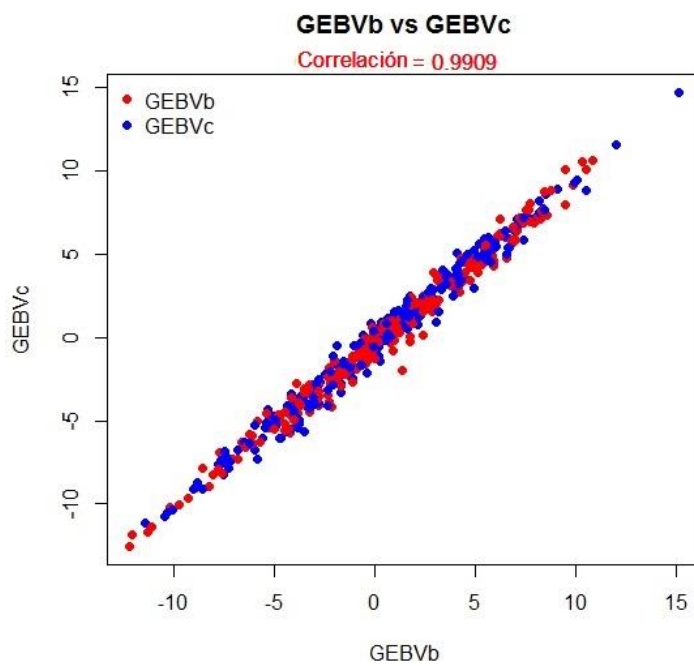


Ilustración 36. Correlación entre los GEBVs resultantes de los métodos GEBVb y GEBVc.

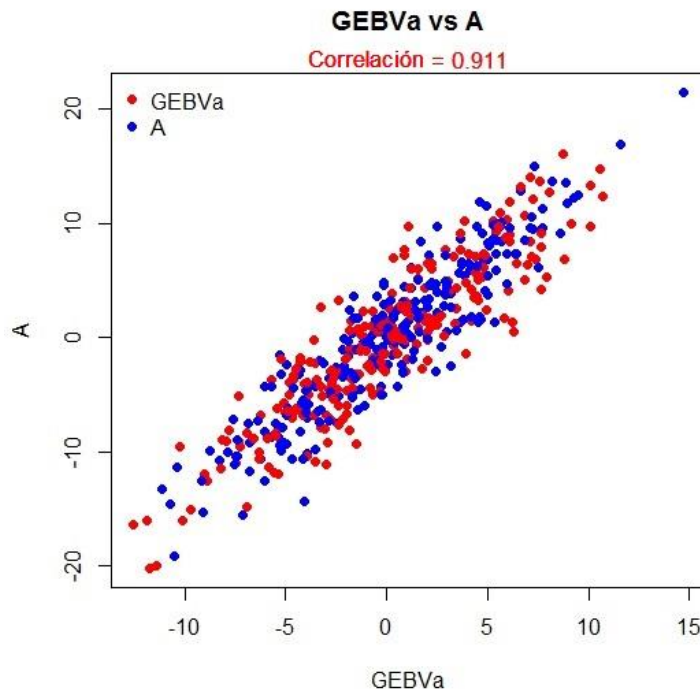


Ilustración 37. Correlación entre los GEBVs resultantes de los métodos GEBVa y A.

Las relaciones genómicas derivadas de datos de marcadores proporcionan una estimación más exacta de covarianzas genéticas entre parientes, que a su vez nos lleva a la obtención de predicciones más exactas. La figura 38 presenta encontramos las matrices de Relaciones Genómicas (G, por frecuencias de los alelos y regresión) y matriz de relaciones genéticas (A, derivada del pedigrí). Se puede observar que la matriz G es más precisa que la matriz A al encontrar relación entre los 500 individuos, esto debido a la utilización de datos genómicos, por tal motivo, con G podemos encontrar relación entre individuos aún sin tener datos de su parentesco.

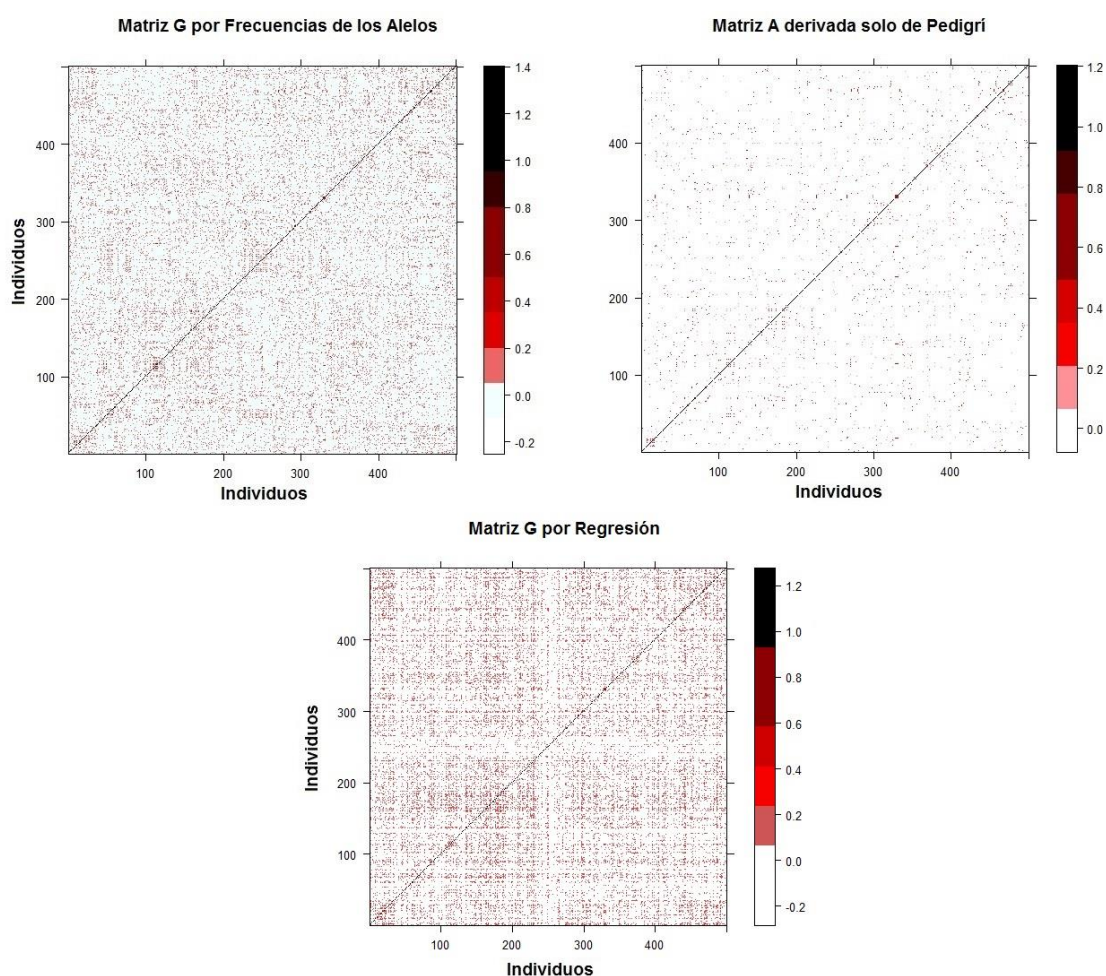


Ilustración 38. Precisión de las matrices de relaciones genómicas (G por frecuencias observadas, G por regresión, y A de pedigrí) para calcular la relación entre individuos.

En el modelo lineal de mezclas el $\text{GEBV}(\hat{\mathbf{u}})$ se refiere a los efectos aleatorios ocasionados por la incidencia de los marcadores genéticos en cada individuo. En la tabla 5, y figuras 39 y 40 se presenta la correlación entre GEBVa y GEBVb contra las observaciones del fenotipo utilizado en los cálculos.

Tabla 5. Correlación GEBVa vs Fenotipo y GEBVb vs Fenotipo.

	Fenotipo
GEBVa	0.8354
GEBVb	0.8226

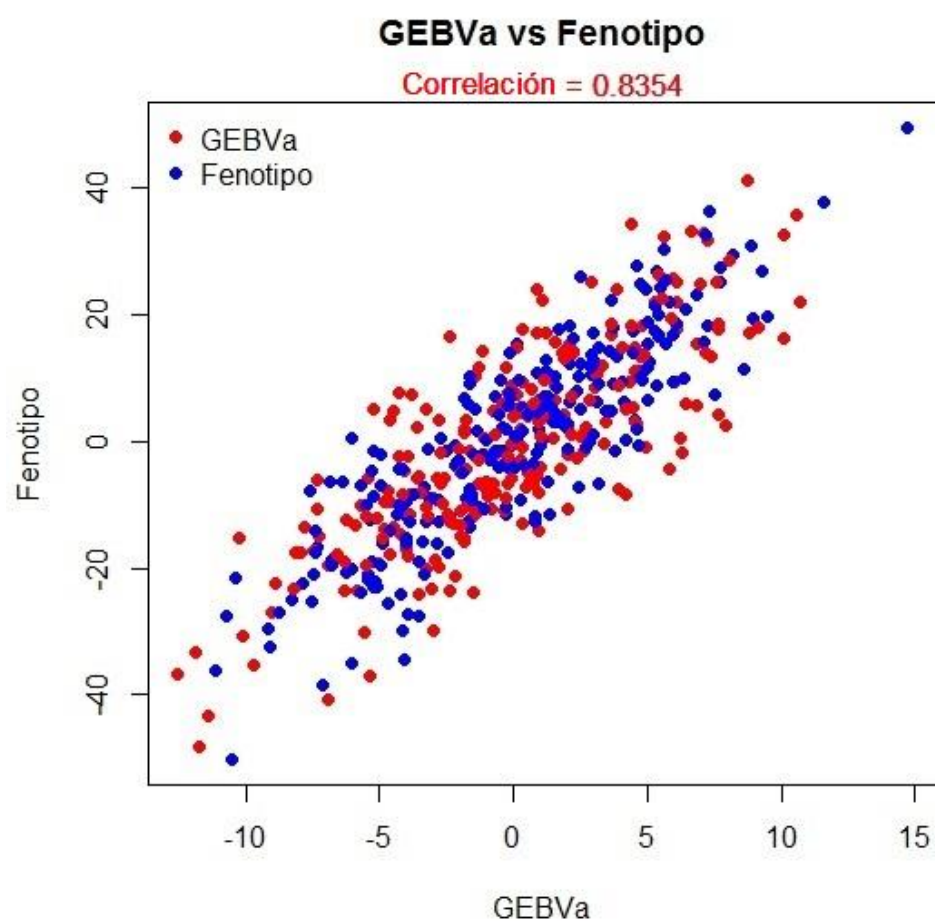


Ilustración 39. Correlación de los resultados de GEBVb y los valores del Fenotipo.

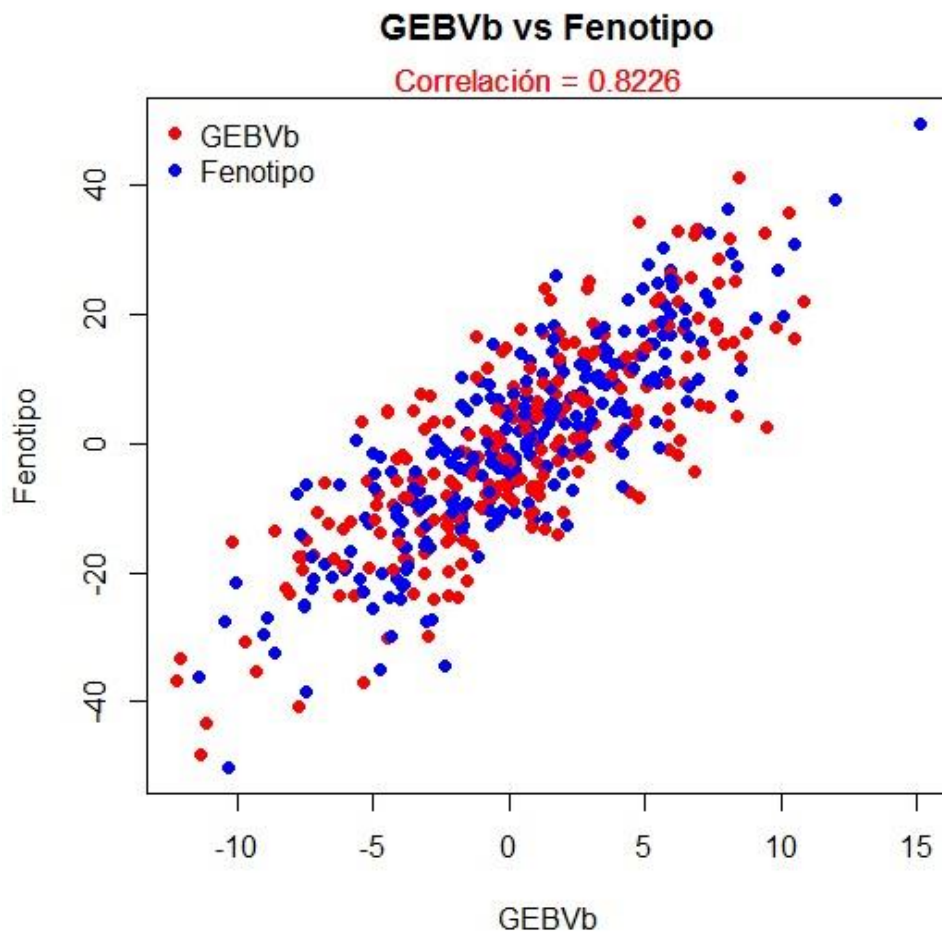


Ilustración 40. Correlación de los resultados de GEBVa y los valores del Fenotipo.

Se puede observar cómo los Valores genómicos (Efectos aleatorios) y los fenotipos observados (mediciones) presentan una correlación positiva de 0.8354 (GEBVa) y 0.8226 (GEBVb). Esto es un indicador de como los efectos aleatorios (genotipos) afectan a la expresión del fenotipo de interés en cada uno de los individuos.

Los cálculos obtenidos con la herramienta desarrollada en este trabajo de investigación también fueron comparados con resultados obtenidos con paquetes implementados en R, synbreed [26] y rrBLUP[27] obteniendo los mismos valores genómicos. Para comparar con synbreed se utilizó una función dentro de este

paquete llamada `gpMod()` y el cual a su vez calcula una matriz de relaciones genómicas basada en frecuencia de los alelos y el modelo GBLUP. En `rrBLUP` se encuentra la función `kin.blup()`, la cual calcula valores genómicos por el método GBLUP utilizando una matriz de relaciones genómicas basada en frecuencias de los alelos. Los resultados se muestran en las figuras 41 y 42.

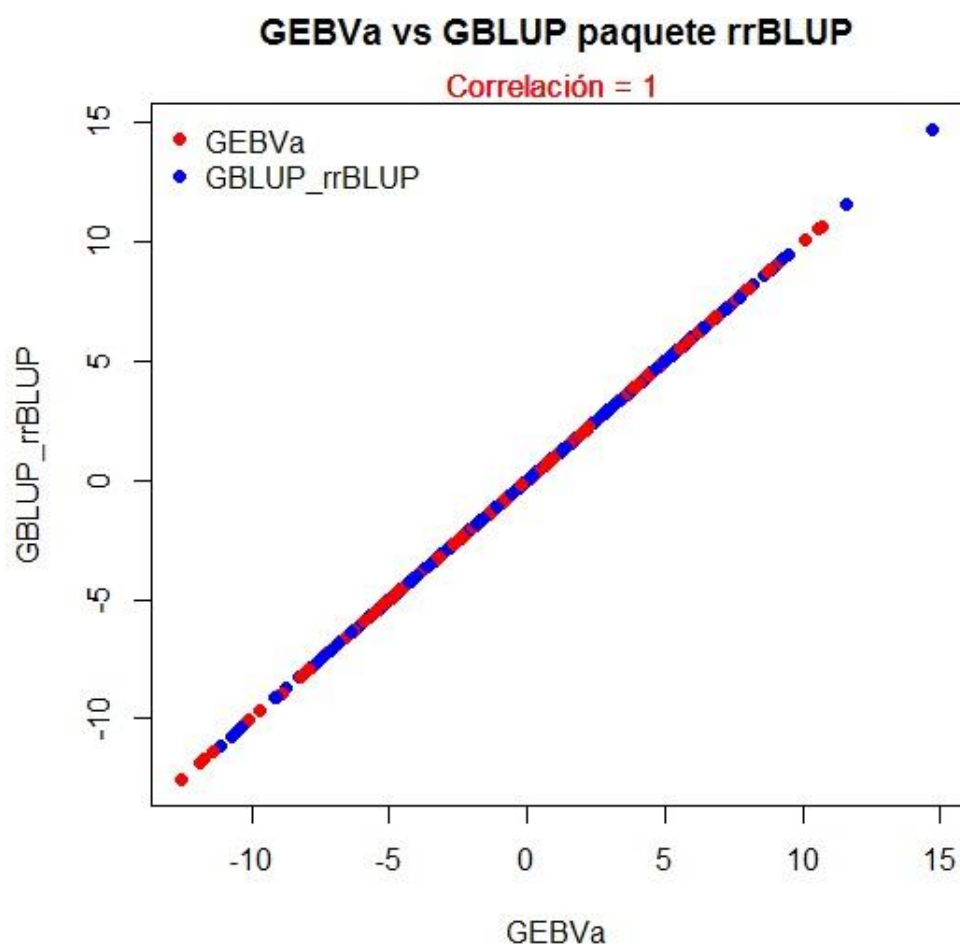


Ilustración 41. Grafica de correlación entre los resultados de GEBVa y el Paquete rrBLUP.

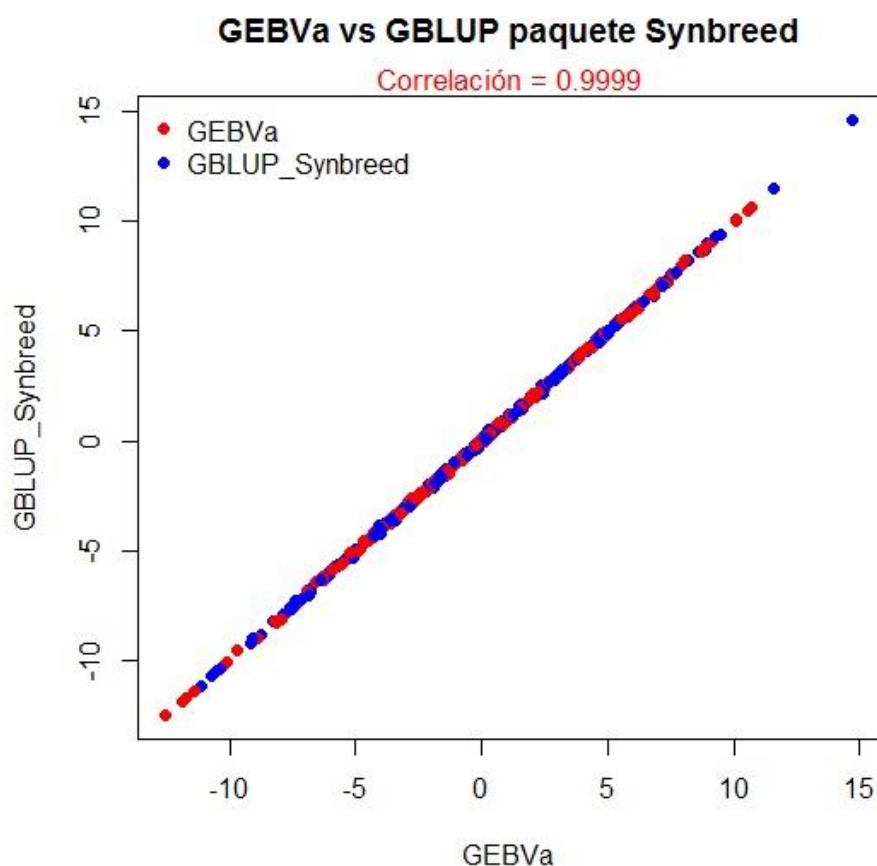


Ilustración 42. Gráfica de correlación entre los resultados de GEBVa vs Paquete Synbreed.

La ejecución del herramienta en R para el cálculo de los valores genómicos utilizando el set de datos presentado al inicio de este capítulo se hizo sobre una PC tipo laptop con sistema operativo Windows 7, procesador Intel® Core™ i3 2.5 Ghz, 8 Gb de memoria RAM, con un tiempo de ejecución de aproximadamente 180 segundos.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1 Conclusiones.

Calcular valores genómicos de crianza utilizando el método GBLUP es prácticamente eficaz y computacionalmente viable para un conjunto de datos como el utilizado (500 individuos x 7250 marcadores SNP), además de la facilidad que nos permite el integrar los datos necesarios al modelo. Otro punto para destacar es que los métodos para construir la matriz G a utilizar en GBLUP no son complicados, y podemos integrar datos genómicos y de pedigrí en ésta (Matriz G por Método de Regresión en A). De este modo, se desarrolló una herramienta estadística basada en el método GBLUP para estimar valores genómicos de ganado bovino utilizando marcadores SNP. Con esto, la integración de una herramienta para calcular Valores Genómicos de Crianza a una herramienta de administración de hatos ganaderos nos lleva a poder crear un panorama futuro basados en la situación actual de nuestro hato, ya que al ver el comportamiento de la producción de nuestro hato nos hace darnos cuentas de las oportunidades de mejora en nuestro grupo de animales, y así

basados en los valores genómicos calculados decidir las cruzas potenciales que en un futuro nos permitirán elevar la producción y mejorar la salud de nuestros animales dentro del hato. Todo lo anterior con el objetivo implementar un programa de mejora genética de ganado bovino en hatos del municipio de Mexicali, Baja California.

BIBLIOGRAFÍA

- [1] Ortega, J. y García, L. (2011). The bovine genome, methods and results of its analysis. *MVZ Córdoba*, 16(1), 2410-2424.
- [2] Hayes, B.J., Lewin, H.A., Goddard, M.E. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*, 29(4), 206-214.
- [3] Genómica. (s.f.). En Wikipedia. Recuperado el 18 de Mayo de 2015 de <http://es.wikipedia.org/wiki/Gen%C3%B3mica>.
- [4] Martínez, S.E. (12 de Junio de 2012). Tendrá Baja California ganado mejorado genéticamente. LA CRONICA.COM. Recuperado de <http://www.lacronica.com/EdicionEnLinea/Notas/Noticias/12062014/852281-Tendra-BC-ganado-mejorado-geneticamente.html>.
- [5] Michel A. Wattiaux. Dairy Essentials, Babcock Institute: for international dairy research and development Recuperado el 10 Octubre de 2014 de www.babcock.wisc.edu/node/121.
- [6] Genetics Home Reference. A service of the U.S. National Library of Medicine. (18 de Mayo 2015). Marker. Recuperado de <http://ghr.nlm.nih.gov/glossary=marker>.
- [7] Meuwissen, T. H. E., Hayes, B.J., y Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(4), 1819–1829.
- [8] Van Tassell C.P., Smith T.P., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C., y Sonstegard T.S., (2008), SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*, 5(3), 247-252.
- [9] Illumina (2012): BovineSNP50 Genotyping BeadChip. Illumina, Inc.
- [10] Illumina (2011). GoldenGate Bovine 3K Genotyping BeadChip. Illumina, Inc.
- [11] Illumina (2013): BovineLD v1.1 Genotyping BeadChip. Illumina, Inc.
- [12] Illumina (2012): BovineHD Genotyping BeadChip. Illumina, Inc.
- [13] Affymetrix (2011): Axiom Genome-Wide Bos 1 Array Plate.
- [14] The Bovine HapMap Consortium., (2009), Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, 324(5926), 528-532.

- [15] Henderson, I. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2), 423-447.
- [16] Kärkkäinen, H.P., Sillanpää, M.J. (2012). Back to basics for Bayesian model building in genomic selection. *Genetics*, 191(3), 969-87.
- [17] Takeshi, H, and Hiroyoshi, I. (2010). EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genetics*, 11:3.
- [18] Protocolo. (s.f.). Caracterización de variables de producción, de salud y genéticas en un hato lechero para la implementación de un programa de mejora genética basada en marcadores. Universidad Autónoma de Baja California, Mexicali, Baja California. México.
- [19] Salomón-Torres, R. (2014). Algoritmos Bioinformáticos para la Detección de Variaciones Estructurales en el Genoma Completo del Ganado Bovino, Utilizando SNPs de Alta Densidad. Universidad Autónoma de Baja California, Mexicali, Baja California, México.
- [20] Arango-Pérez, M.L. (2013). Caracterización del Modelo Experimental de Producción Lechera del Instituto de Investigaciones Veterinarias de la UABC para la implementación de un programa de mejora genética. Universidad Autónoma de Baja California, Mexicali, Baja California, México.
- [21] West, B., Welch, K., Galecki, A. (2007). Linear Mixed Models: A Practical Guide Using Statistical Software. Boca Raton, FL: Taylor & Francis Group, LLC.
- [22] Robinson, G.K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, 6(1), 15-51.
- [23] VanRaden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.*, 91(11), 4414-4423.
- [24] Zapata-Valenzuela, J., Whetten R. W., Neal D., Mckeand S. y Isik F. (2013). Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. *G3 Journal*, 3(5), 909-916.
- [25] R Project. (s.f.). What is R?. Recuperado el 4 de Mayo de 2015 de: <http://www.r-project.org/about.html>.
- [26] Wimmer, V., Albrecht, T., Auinger, H. J., Schon, C. C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, 28(15), 2086–2087.
- [27] Endelman, J.B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, 4(3), 250–255.
- [28] Taejeong, C., Dajeong, L., Byoungho, P., Aditi, S., Jong-Joo, K., Sidong, K. y Seung H. L. (2016). Accuracy of genomic breeding value prediction for

intramuscular fat using different genomic relationship matrices in Hanwoo (Korean cattle). *Asian-Australasian Journal of Animal Sciences*, 30:907-911.