Propuesta alternativa: "Análisis del genoma bovino en vacas de la raza Holando"

El conjunto de datos contiene información de un experimento realizado en Uruguay, que involucró el genotipado de una serie de vacas de la raza Holando. Para el genotipado se utilizó el chip "BovineHD BeadChip" de la empresa Illumina1 y el juego de datos de este ejercicio es un subconjunto del mismo, con marcadores elegidos al azar.

1- Describir la tecnología usada, incluyendo el chip utilizado (tipo de marcadores, cantidad de los mismos, distribución, criterios para elegirlos, especies que cubre, tipos de razas, etc.).

El "chip" utilizado para generar los datos que se analizan es el "BovineHD BeadChip", desarrollado por Illumina en colaboración con diversas instituciones del sector genética animal.

El "chip" contiene miles a millones de bolitas de silicona, localizadas en minúsculas cavidades en la superficie del chip. Cada bolilla está cubierta por sustancias sensitivas específicas para oligonucleótidos específicos en determinado locus. Conforme los fragmentos de DNA pasan sobre el chip, las sustancias reactivas se enlazan con la secuencia complementaria del DNA analizado. La especificidad de cada alelo es marcada por la incorporación de uno de cuatro nucleótidos etiquetados (ver Fig. 1). Al ser aplicado un láser, el nucleótido emite una señal. La intensidad de la misma devuelve información sobre la ratio de nucleótidos en ese locus.

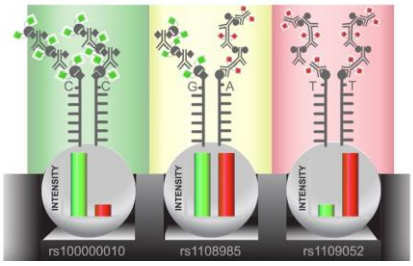

Fig. 1. Como funcionan los "microarrays" o chips"

Las barras de color verde y rojo indican las frecuencias de los alelos en el loci

Ver :(Illumina Microarray Technology)

El BovineHD BeadChip utilizado puede identificar marcadores, en este caso unos 777,000 SNPs ("single nucleotide polymorphisms"), que permiten, entre otros: análisis cuantitativos de loci, desequilibrios de ligamiento, merito genético, etc.

La técnica utilizada aquí cubre las especies bovinas productoras de leche o carne: 20 razas de *Bos taurus taurus* (Btt*) ,* 3 razas de *Bos taurus indicus* (Bti), y 4 cruzas Bti × Btt.

2 -¿Qué proporción de marcadores tiene el juego de datos en relación al total del chip? ·

La muestra analizada contiene información de 7,748 SNPs, lo que representa el 1% de la capacidad del chip.

3 - ¿Cuál es la distribución de marcadores por cromosoma en el juego de datos?

El máximo número de marcadores por cromosoma lo tiene el No 1 con 482. El mínimo el MT (cromosoma mitocondrial) con 2 y el promedio es de 250.

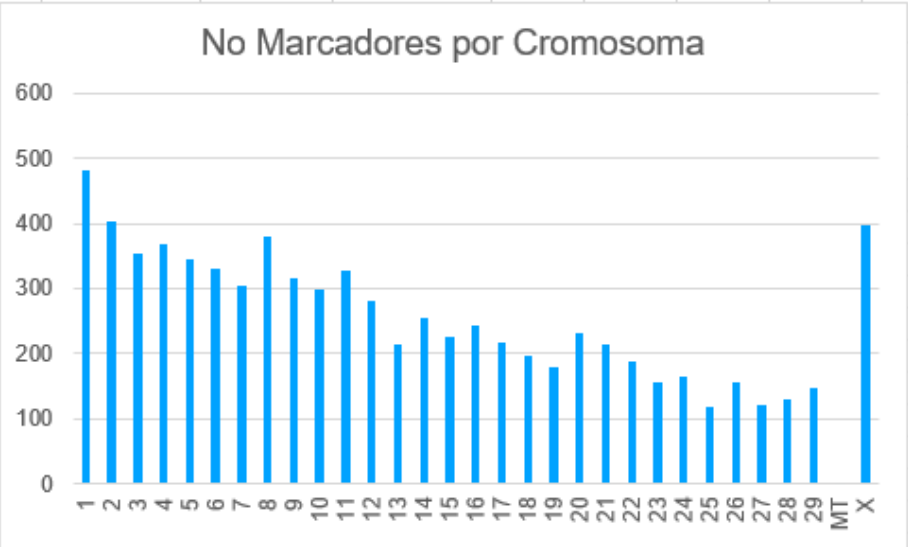

Fig. 2: Numero de marcadores (SNPs) por cromosoma.

4 -¿Es representativa del tamaño de los cromosomas bovinos?

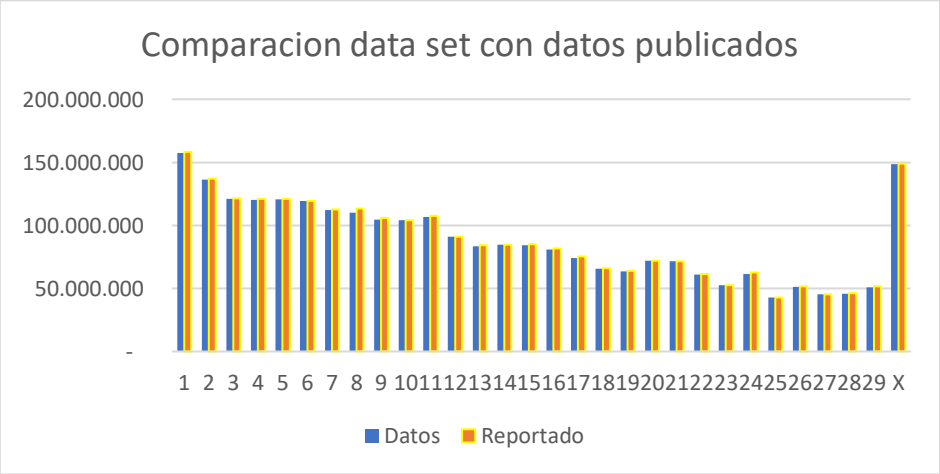

Fig. 3: Largo de cromosomas (1 al 30 más X) comparado con datos publicados

El largo de los cromosomas, según se deriva de los datos analizados, es similar a los reportados en trabajos científicos **(Bernt Guldbrandtsen & Joanna Szyda**. Patterns of DNA variation between the autosomes, the X chromosome and the Y chromosome in Bos taurus genome Bartosz Czech, 2020). La similitud arroja una correlación de 99.98% (exceptuando el cromosoma MT).

5 -¿Cuántos animales fueron genotipados en el juego de datos?

Fueron genotipados 24 animales para 7748 SNPs cada uno.

6 - ¿Qué representa cada una de las columnas del juego de datos?

Columna 1	Numero ordinal del 1 al 7748 (cantidad de SNPs analizados)
Columna 2	Identificador del SNP
Columna 3 - 26	Una columna por cada uno de los 24 animales genotipados
Columna 27	Nombre del marcador SNP (repite la columna 2)
Columna 28	Número del cromosoma (1 al 30 más X y MT)
Columna 29	Posición: Posición del marcador desde el inicio del cromosoma, en pares de bases
Columna 30	GenTraiScore: Es una métrica que mide la calidad del "calling" del marcador. Va de 0 a 1
Columna 31	NormID : es una clasificación de los diferentes tipos de sondas en el arreglo ("array")
Columna 32	alelo1: uno de los dos alelos del locus
Columna 33	alelo2: ídem

7 - ¿Cuántos datos faltantes hay en los genotipados?

Los datos faltantes en los resultados de genotipado aparecen como “-/-“ y suman 2,883.

8 - ¿Cómo es la distribución del score de calidad?

La distribución de 7,748 datos de calidad (“GenTrainScore”) se muestra en la fig. 4 abajo.

Los valores tienen una media de 0,8495 y una varianza de 0,0045, con un coeficiente de variación de 0,53%. Es decir que los datos se distribuyen con gran frecuencia hacia los valores de calidad altos. La distribución tiene un coeficiente de asimetría de -3,44, lo que indica una larga “cola” hacia la izquierda, por lo anteriormente explicado.

Fig. 4: Distribución de frecuencias de valores de calidad (GenTrainScore) de los datos analizados.

9 - Calcular para cada loci el contenido G+C observado (proporción de bases G o C del total de bases genotipadas. Utilizando esa información, describir para cada cromosoma el contenido G+C (por ejemplo, utilizando estadísticos resumen, histogramas, boxplots, etc.).

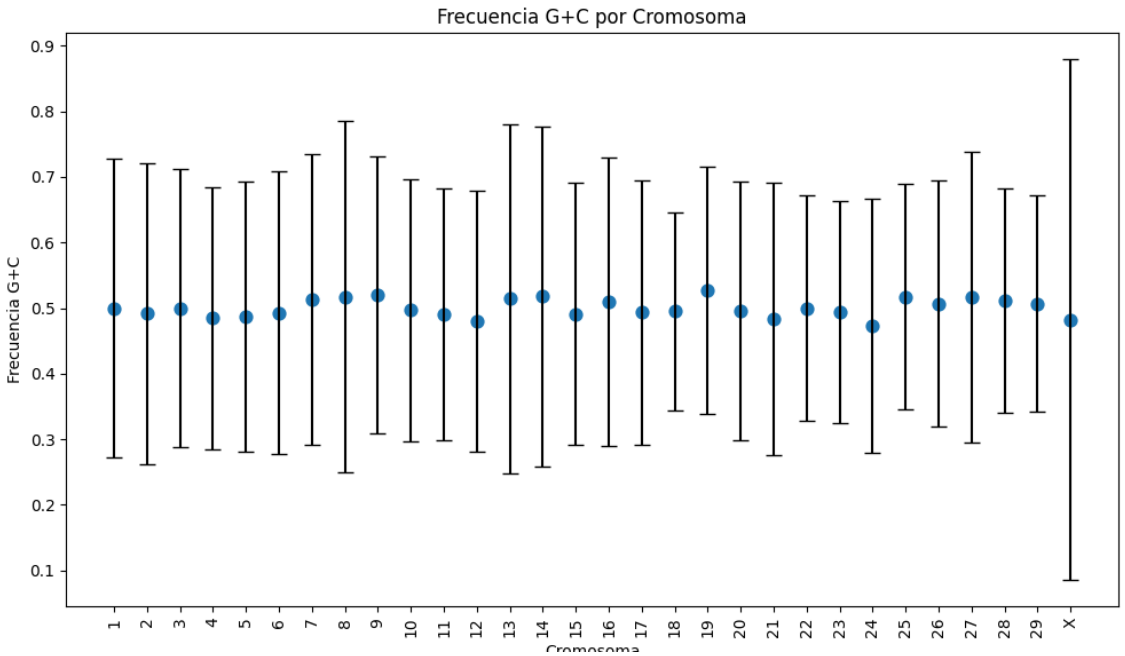

Fig. 5: Frecuencia de contenido de G+C por loci, para diferentes loci (se omite cromosoma MT)

10 - Calcular para cada animal la proporción de loci heterocigotas.

Se calculo el % de individuos heterocigotos, es decir los que tienen genotipo "A/G","A/C","T/G","T/C". El cálculo se realizó sin contar los “-/-“ (=sin datos).

ID	% Heterocigotos	ID	% Heterocigotos
animal.1	29%	**animal.16**	30%
animal.3	29%	**animal.17**	30%
animal.4	29%	**animal.18**	32%
animal.5	30%	**animal.19**	29%
animal.6	30%	**animal.20**	30%
animal.8	29%	**animal.23**	29%
animal.9	29%	**animal.24**	28%
animal.10	29%	**animal.25**	30%
animal.11	29%	**animal.26**	31%
animal.12	32%	**animal.27**	30%
animal.14	30%	**animal.28**	32%
animal.15	29%	**animal.30**	32%

Fig. 5: Porcentaje de loci Heterocigotos para cada animal

11 - Para cada locus crear tres columnas que describan el conteo de cada uno de los tres genotipos posibles.

1) La que contiene "A/A","A/T","T/A" o "T/T" (que llamaremos A1A1, homocigotas)
2) La que contiene "A/G","A/C","T/G","T/C" (que llamaremos A1A2, heterocigotas)
3) La que contiene "G/G","G/C","C/G","C/C" (que llamaremos A2A2, homocigotas)

Los datos para cada locus pueden verse en archivo “vaca_out_final.xlsx” , columnas A1A1, A1A2 y A2A2, para cada genotipo. Abajo se muestra cuadro resumen para los 7748 loci y 24 individuos.

A1A1	A1A2	A2A2
genotipo	**cantidad**	**frecuencia**
A1A1	62.080	34%
A1A2	54.590	30%
A2A2	66.399	36%
Total	**183.069**	**100%**

Fig. 6: Conteo y frecuencia de cada uno de los tres genotipos hallados

12 - Calcular para cada locus la frecuencia relativa de genotipos heterocigotas observada.

Ver “vaca_out_final.xlsx” , Columna “Frec_He”, donde se muestra la frecuencia entre 0 y 1, del genotipo A1A2 para cada locus.

13 - Describir la heterocigosidad del conjunto de animales a partir de estas frecuencias relativas (por ejemplo, utilizando estadísticos resumen, histogramas, boxplots, etc.).

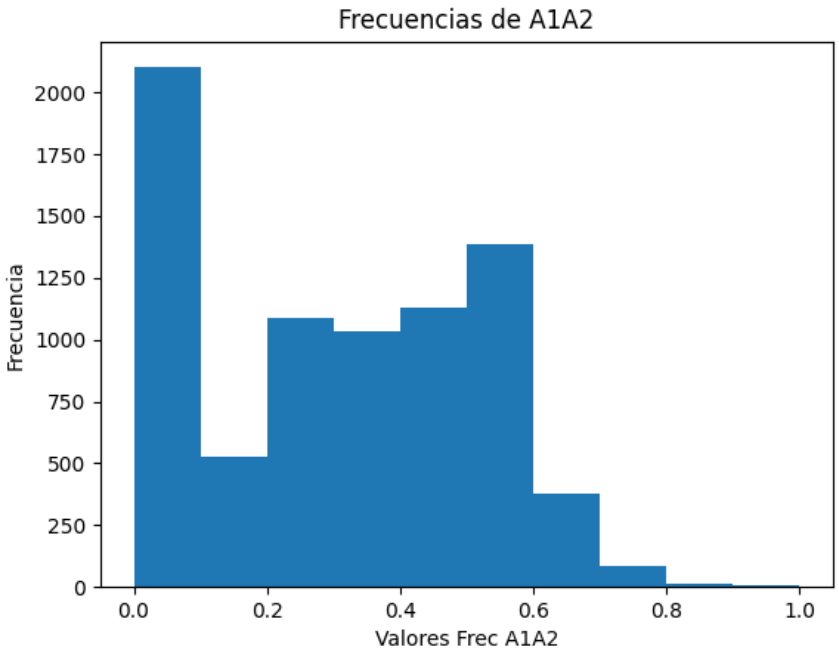

Fig. 7: Se observa la distribución de la frecuencia de heterocigotas A1A2 para los 24 animales.

14 - Calcular para cada locus las frecuencias alélicas y las frecuencias esperadas para los tres genotipos si el locus se encontrase en equilibrio de Hardy-Weinberg.

Los datos se encuentran en la planilla "vaca_out_final.xlsx" y son:

f_A1	frecuencia de alelo A1
f_A2	frecuencia de alelo A2
E(A1A1)	Equilibrio H&W A1A1
E(A1A2)	Equilibrio H&W A1A2
E(A2A2)	Equilibrio H&W A2A2

15 - Realizar para cada locus una prueba de chi-cuadrado y determinar si se encuentra en equilibrio de Hardy-Weinberg considerando un alfa = 0,01 (1 %).

Realizada la prueba Chi cuadrado para 7,737 loci, comparando las frecuencias de los tres genotipos observadas : O A1A1, OA1A2 y O A2A2 versus E A1A1, E A1A2 y EA2A2, 6,181 (80%) de las comparaciones para los loci resulto NO significativa. Es decir que todos los loci se **encuentran en equilibrio Hardy – Weinberg. (ver "vaca_out_final.xlsx" chi 2).** Para 1,556 (20%) loci se encontró fijación de un alelo, lo que se aparta del equilibrio mencionado.

16 - Describir la frecuencia del alelo mayor en los datos. Opcional

Se muestra la frecuencia del alelo mayor (A1) en el gráfico, encontrándose una distribución de frecuencias bastante parecida por clase, (300 a 400 loci por clase), salvo para los extremos, donde aparece fijado el alelo A1 en un 14% y el A2 en 13% de los loci.

Fig. 8: Frecuencia del alelo mayor atraves de los 4778 loci

17 - ¿Hay alguna indicación de que las vacas pertenezcan a poblaciones diferentes desde el punto de vista genético? Por ejemplo, se pueden utilizar técnicas de análisis exploratorio multivariado.

Se transformaron los datos tipo “X/Y” a 0,1,2. Luego se corrió un análisis de componentes principales. Graficando los primeros dos componentes (que juntos explican un 58% de la varianza), se encuentra un grupo bastante homogéneo de animales (los animales se muestran como “b1 – b24”).

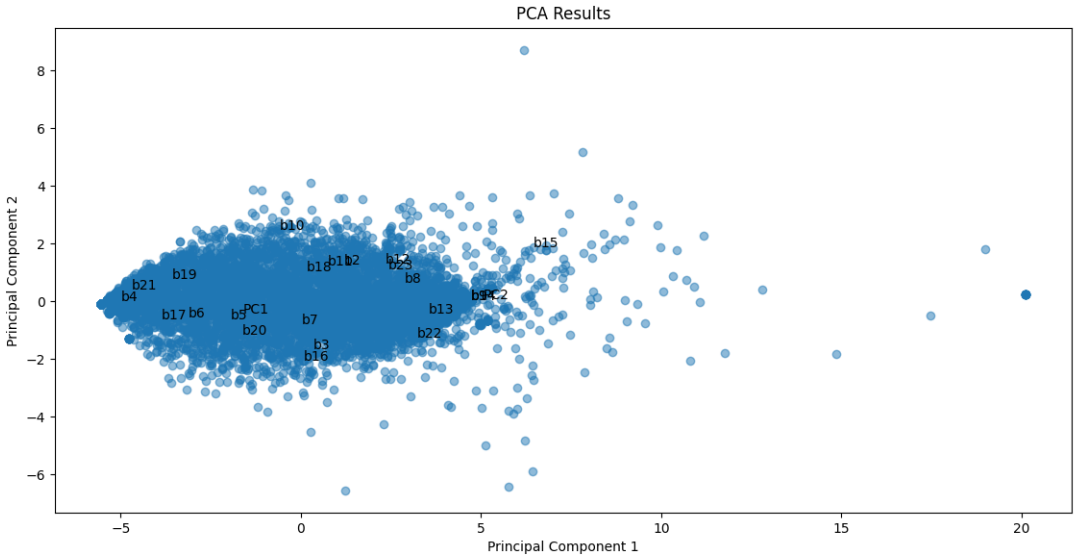

Fig. 9: Análisis grafico de los dos primeros componentes principales

Datos accesibles en : [Pabloau2/Genética-FAGRO: Contiene código Python y planillas Análisis datos Holando 2024 (github.com)]