



UNIVERSIDAD PERUANA  
**CAYETANO HEREDIA**

**BÚSQUEDA DE GENES ASOCIADOS A  
RASGOS SELECCIONADOS PARA LA  
PRODUCCIÓN DE FIBRA DE ALPACA**

**TESIS PARA OPTAR EL  
GRADO DE MAESTRO EN  
BIOQUÍMICA Y BIOLOGÍA MOLECULAR**

**DIEGO MARTIN VÉLIZ OTANI**

**LIMA – PERÚ**

**2017**



**Dra. Rosa Patricia Herrera Velit**

**ASESORA**

## **AGRADECIMIENTOS**

A la Dra. Patricia Herrera y al Dr. José Espinoza por el apoyo brindado

A Jorge Rodríguez e Irene Delgado, por las sugerencias y consejos durante la ejecución de este proyecto.

A mis padres y a Keren por su paciencia y apoyo constante.

Al Programa Nacional de Innovación para la Competitividad y Productividad (Innovate Perú), de acuerdo al contrato 119-FINCYT-IA-2013: “Una aproximación de genómica de poblaciones aplicada a la filogenia y mapeo de rasgos productivos en alpaca”.

Diego Veliz Otani es becario del Programa de Bioquímica y Biología Molecular, CONV-00031-2015-FONDECYT-DE

## Tabla de contenidos

I.	INTRODUCCIÓN Y PLANTEAMIENTO DEL PROBLEMA .....	2
II.	MARCO TEÓRICO.....	13
1.	Genética de poblaciones .....	13
1.1.	Desviaciones del equilibrio de Hardy-Weinberg .....	13
1.2.	Evolución neutral .....	14
1.3.	Efectos de la selección y variación del tamaño poblacional .....	21
1.4.	Pruebas de hipótesis de equilibrio neutral .....	26
2.	Principios de genética cuantitativa y modelos de asociación genética.....	29
III.	OBJETIVOS.....	34
1.	Objetivo general .....	34
2.	Objetivos específicos.....	34
IV.	METODOLOGÍA .....	35
1.	Diseño .....	35
2.	Muestreo .....	35
2.1.	Población de interés: .....	37
2.2.	Población elegible: .....	37
2.3.	Tamaño muestral: .....	37
2.4.	Criterios de inclusión: .....	38
2.5.	Criterios de exclusión: .....	38
3.	Selección de genes candidatos: .....	39
4.	Identificación de variantes genéticas en los genes candidatos: .....	40
4.1.	Preprocesamiento de genomas secuenciados .....	40
4.2.	Identificación de variantes genéticas en los genes candidatos .....	42
4.3.	Predicción de los efectos de las variantes genéticas .....	43
5.	Estandarización de la amplificación del gen BMP4 .....	44
6.	Amplificación de los genes KAP1-1 y KAP13-1 .....	47
7.	Electroforesis en agarosa.....	47
8.	Secuenciamiento mediante método Sanger .....	48
9.	Estudio exploratorio del gen BMP4 en una fracción de la muestra representativa de todo el Perú .....	51
10.	Análisis genético-poblacional.....	52

11.	Pruebas de hipótesis de equilibrio evolutivo neutral en la muestra Rural Alianza	56
12.	Análisis de asociación genética .....	56
13.	Análisis de potencia estadística .....	58
V.	RESULTADOS .....	61
1.	Descripción de la calidad de la fibra.....	61
2.	Selección de genes candidatos y descripción de las variantes genéticas identificadas en la fase bioinformática .....	62
3.	Estandarización de la amplificación del gen BMP4 .....	74
4.	Amplificación del gen BMP4 en la muestra representativa de todo el Perú .....	75
5.	Amplificación de BMP4, KAP1-1 y KAP13-1 en la muestra Rural Alianza para el análisis de asociación y pruebas de evolución neutral .....	75
6.	Resultados del secuenciamiento Sanger .....	76
6.1.	Secuenciamiento del gen BMP4 en la muestra Nacional y en la muestra Rural Alianza .....	76
6.2.	Secuenciamiento del gen KAP13-1 en la muestra Rural Alianza .....	84
6.3.	Secuenciamiento del gen KAP1-1 en la muestra Rural Alianza .....	90
7.	Análisis genético-poblacional .....	90
7.1.	Análisis de endogamia y estructuración poblacional en la muestra Rural Alianza empleando diez microsatélites .....	91
7.2.	Descripción de la variabilidad genética de BMP4, KAP1-1 y KAP13-1 .....	92
8.	Pruebas de equilibrio evolutivo neutral .....	99
9.	Estudio de asociación genética .....	104
9.1.	Análisis de potencia estadística .....	104
9.2.	Análisis de asociación de los genes BMP4, KAP1-1 y KAP13-1 .....	107
VI.	DISCUSIÓN .....	111
VII.	CONCLUSIÓN .....	122
VIII.	RECOMENDACIONES .....	123
IX.	BIBLIOGRAFÍA.....	124

## ABREVIATURAS

LD	Desequilibrio de ligamiento ( <i>linkage disequilibrium</i> )
MAF	Frecuencia del Alelo Menor ( <i>Minor allele frequency</i> )
MCMC	Markov-Chain Monte Carlo
Min	Minutos.
MLM	Modelos Lineales Mixtos
pb	Pares de bases
PCR	Reacción en cadena de la polimerasa
SNP	Polimorfismo de un nucleótido ( <i>Single nucleotide polymorphism</i> )
STR	Repeticiones cortas en tándem ( <i>Short Tandem Repeats</i> ).
$\mu\text{M}$	Micromolar
V	Voltios

## LISTA DE FIGURAS

**Figura 1.** Relación entre el diámetro y el tipo de fibra y su precio.

**Figura 2.** Señales que intervienen en la secuencia de eventos durante el desarrollo del pelo.

**Figura 3.** Representación gráfica de los valores genotípicos  $-a$ ,  $d$  y  $a$  para un modelo de un locus.

**Figura 4.** Resumen gráfico de las actividades que componen este trabajo de investigación.

Figura 5. Diámetro de la fibra de alpaca en la muestra completa de 151 alpacas.

**Figura 6.** Resultados de la electroforesis de los amplicones de BMP4 en un gel de agarosa al 2.5%.

**Figura 7.** Alineamiento de las secuencias del gen BMP4 obtenidas a partir de las 10 mezclas de ADN escogidas aleatoriamente de la muestra Nacional.

**Figura 8.** Consecuencias del microsatélite GGC a nivel del secuenciamiento en el gen BMP4 en la muestra Nacional.

**Figura 9.** Cromatogramas de una región amplificada del gen KAP13-1 en la muestra Rural Alianza.

**Figura 10.** Se muestra el alineamiento de un grupo de secuencias del gen KAP13-1 en la muestra Rural Alianza

**Figura 11.** Se muestra el alineamiento de un grupo de secuencias del gen KAP1-1 en la muestra Rural Alianza.

**Figura 12.** Análisis de estructuración poblacional.

**Figura 13.** Desequilibrio de ligamiento de los SNPs con frecuencia del alelo mayor por encima del 5%.

**Figura 14.** Desequilibrio de ligamiento en los tres genes candidatos. a)

**Figura 15.** Distribución del número de diferencias entre pares de secuencias (*Mismatch Distribution*).

**Figura 16.** Gráficas de potencia estadística alcanzada según la proporción de la varianza fenotípica explicada por el efecto los efectos genéticos aditivos

**Figura 17.** Gráficos de cajas del diámetro de fibra para diferentes categorías genotípicas.

## LISTA DE TABLAS

**Tabla 1.** Valores estimados de heredabilidad y promedio del diámetro de fibra de alpaca.

**Tabla 2.** Condiciones para la estandarización de la amplificación del gen BMP4

**Tabla 3.** Condiciones de ciclaje de la estandarización de la amplificación del gen BMP4.

**Tabla 4.** Condiciones de la amplificación de los genes KAP1-1 y KAP13-1.

**Tabla 5.** Condiciones de ciclaje de la amplificación del gen KAP1-1.

**Tabla 6.** Condiciones de ciclaje de la amplificación del gen KAP13-1.

**Tabla 7.** Ubicación de los transcritos codificados por los genes reguladores seleccionados.

**Tabla 8.** Variantes en regiones promotoras de los genes candidatos seleccionados.

**Tabla 9.** Variantes sinónimas en genes candidatos.

**Tabla 10.** Variantes no sinónimas en genes candidatos.

**Tabla 11.** Variantes del gen BMP4 observadas en los 3 genomas secuenciados y en las 10 mezclas (*pools*) de ADN escogidas aleatoriamente de la muestra representativa de todo el Perú.

**Tabla 12.** Variantes del gen BMP4 descubiertas en el análisis poblacional (muestra Nacional).

**Tabla 13.** SNPs del gen KAP 13-1 detectadas en el análisis bioinformático.

**Tabla 14.** Análisis de heterocigocidad de 10 microsatélites.

**Tabla 15.** Correlación entre el desequilibrio de ligamiento y la distancia.

**Tabla 16.** Valores de los estimadores de  $\theta$  por sitio ( $4N_e\mu$ ).

**Tabla 17.** Resultado de las pruebas de equilibrio entre deriva y mutación bajo evolución neutral.

**Tabla 18.** Resultados del análisis de asociación – solo efectos aditivos

**Tabla 19.** Descripción del par de cebadores empleados para amplificar el gen BMP4

**Tabla 20.** Descripción de los cebadores empleados para la amplificación de los genes KAP1-1 y KAP13-1

**Tabla 21.** Frecuencias Alélicas del gen BMP4 en la muestra Nacional

**Tabla 22.** Frecuencias Alélicas de los tres genes candidatos en la muestra Rural Alianza.

**Tabla 23.** Análisis de heterocigocidad en la muestra Nacional.

**Tabla 24.** Análisis de heterocigocidad en la muestra Rural Alianza.

**Tabla 25.** Resultados del análisis de asociación – Modelo de un locus con dominancia.

**Tabla 26.** Selección del mejor modelo mediante Análisis de Varianza.

## RESUMEN

La alpaca (*Vicugna pacos*) es un animal de gran valor económico para la región andina del Perú debido a su carne y fibra. El precio de su fibra es altamente variable y está determinado principalmente por su diámetro y por la raza de la cual es obtenida. A la fecha, no se han identificado genes que determinen o afecten ningún rasgo morfológico de la fibra de alpaca.

En este estudio se evaluó si el gen BMP4 (regulador del desarrollo del pelo) o los genes KAP1-1 y KAP13-1 (componentes estructurales del pelo) presentan huellas de selección, y si alguno de estos genes posee polimorfismos asociados al diámetro de la fibra en una población de alpacas destinada a su producción. Se encontró evidencia de un barrido selectivo en o cerca del gen KAP13-1, lo que sugiere que este gen o uno muy cercano a él influencia alguna característica monogénica u oligogénica de posible relevancia para la producción de fibra. No se encontró asociación entre los polimorfismos en los genes y el diámetro de la fibra. Dado que la potencia estadística del análisis de asociación fue suficiente para detectar un gen con efecto mayor, se concluye que la huella de selección en KAP13-1 se debería a algún rasgo diferente del diámetro de la fibra de alpaca.

## **ABSTRACT**

The alpaca is a South American camelid of great economic importance in the Andean region because of its meat and fiber. The price of alpaca fiber is highly variable and is largely determined by its diameter and the breed from which it proceeds. So far, no genes influencing morphological traits of alpaca fiber have been identified.

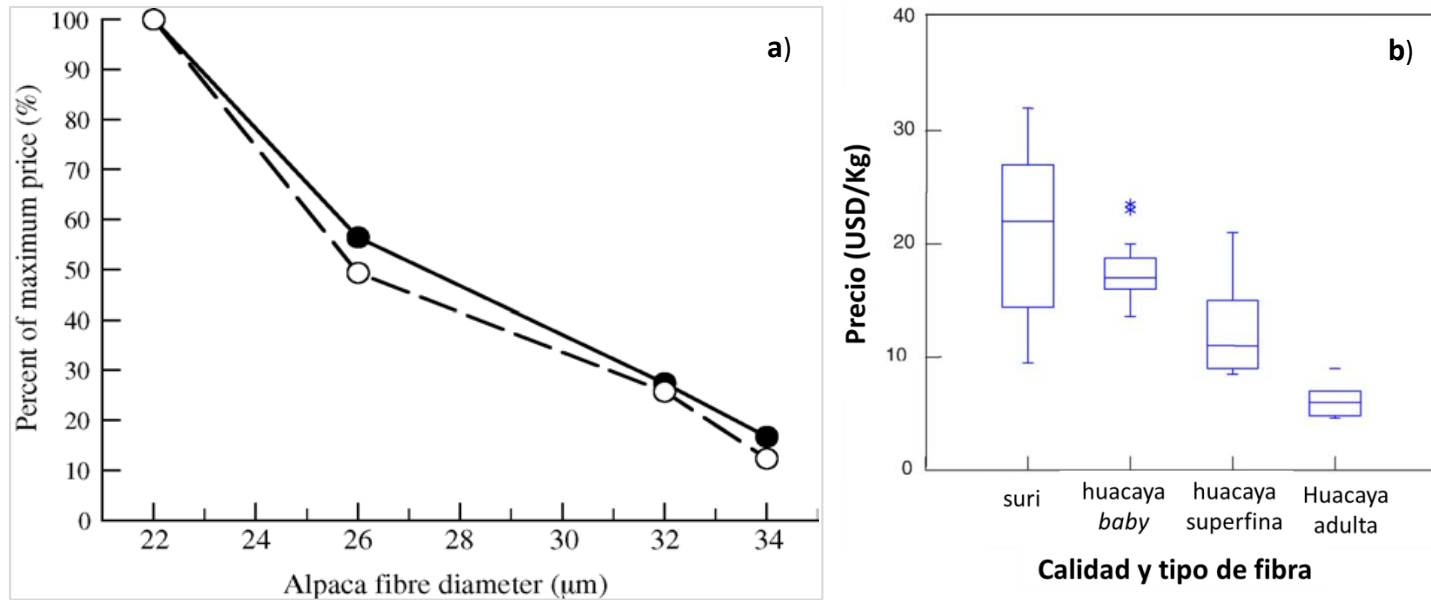
This study assessed whether the gene BMP4, which regulates hair development, or the genes KAP1-1 and KAP13-1, which are structural components of hair, present selection signatures. Additionally, a genetic association study between these genes and fiber diameter was performed. KAP13-1 showed several signatures of a selective sweep, suggesting that this gene or another very close to it affects a fiber trait of monogenic or oligogenic inheritance. No association was found between fiber diameter and polymorphisms in these genes. Given that the association study reached enough power to detect a major gene, it is concluded that the selective sweep on KAP13-1 was produced by selection on a trait other than fiber diameter.

## **I. INTRODUCCIÓN Y PLANTEAMIENTO DEL PROBLEMA**

La alpaca (*Vicugna pacos*) es un animal de gran valor económico para la región andina del Perú debido a su carne y fibra, siendo el Perú el principal productor de fibra de Alpaca a nivel mundial (1).

El precio de su fibra es altamente variable y está determinado principalmente por su diámetro y el tipo de fibra (fig. 1)(2,3). El tipo de fibra es empleado para definir las dos razas de alpaca. La raza Huacaya representa aproximadamente el 88% de la población del Perú y se caracteriza por tener una fibra rizada de apariencia esponjosa (4). La raza suri presenta una fibra menos rizada y más larga y lustrosa (5,6), por lo que suele compararse al casimir. Debido a estas características, la fibra de raza suri es mejor cotizada (fig. 1b). Otras variables de importancia económica son el peso total por esquila y el color.

El tipo de fibra está determinado genéticamente. Se ha predicho, mediante técnicas de genética clásica, que la fibra huacaya está determinada por los alelos recesivos de dos genes ligados con una frecuencia de recombinación de 9.9% (intervalo de confianza al 95% = [3-20%]) (7). Debido a que el fenotipo suri es dominante sobre el fenotipo huacaya, es posible obtener crías huacaya del cruce de parentales suri(7). Por lo tanto, la identificación de los genes que determinan el tipo de fibra puede acelerar el establecimiento de rebaños de alpacas suri homocigotas.



**Figura 1.** Relación entre el diámetro y el tipo de fibra y su precio. a) Relación entre el diámetro de la fibra de alpaca y el porcentaje del máximo valor alcanzado. Esferas negras: datos colectados entre 1985 y 1994. Esferas blancas: precios durante los años de depreciación (1986, 1991, 1992). Figura obtenida de referencia (2). b) Comparación entre el precio de la fibra procesada (ovillos) suri y fibra huacaya de tres calidades diferentes durante el periodo 1999-2004. El diámetro máximo de las 4 categorías (de izquierda a derecha) es 27; 22; 25 y 27.5 μm. Imagen adaptada de referencia (3).

Un estudio reciente reporta haber encontrado dos genes que determinan los fenotipos suri y huacaya; sin embargo, no detalla cuáles son (8). Uno de estos se encuentra en un clúster de Proteínas Asociadas a Keratina (KAP). Ambos genes se ubican en diferentes *scaffolds* en el genoma de alpaca, por lo que no se pudo determinar la distancia entre estos. Sin embargo, estos genes se encuentran a una distancia de 7Mb en el genoma de vaca.

Se ha estimado que la *heredabilidad* en sentido restringido (proporción de la variabilidad fenotípica explicada por efectos genéticos transmisibles de una generación a la siguiente) del diámetro de la fibra se encuentra entre 0.30 y 0.70 (tabla 1), y se presume que constituye un rasgo cuantitativo con herencia poligénica (9–12). Sin embargo, los genes involucrados no han sido identificados, principalmente debido a que el genoma de la alpaca aún no se encuentra ensamblado a nivel de cromosomas, lo que ha dificultado la identificación y selección de marcadores genéticos.

Se ha propuesto que existe un gen *mayor* (un gen con un gran efecto sobre un rasgo poligénico) que explica más del 50% de la heredabilidad del diámetro de la fibra; sin embargo, este gen no ha sido identificado (10). Su identificación permitiría la implementación de programas de selección asistida por marcadores. La ventaja de la selección asistida por marcadores —en caso de que se comprobara que existe un gen con efecto mayor— es que se podría considerar los efectos genéticos aditivos en la selección de individuos reproductores. Cabe destacar que solo los efectos genéticos aditivos son transmitidos de una generación a la siguiente (13).

**Tabla 1.** Valores estimados de heredabilidad y promedio del diámetro de fibra de alpaca.

Raza	Heredabilidad (h <sup>2</sup> )*	Diámetro ( $\bar{x} \pm DE$ )	Localidad	Método‡	Referencia
Huacaya	0.63(0.16)	26.4	Nueva Zelanda	Modelo animal	(14)
Huacaya	0.36[0.31-0.41]	22.97±4.18	Pacomarca (Arequipa)	MCMC-MLM – sin gen con efecto mayor	(10)
Huacaya	0.63[0.57-0.70]	22.97±4.18	Pacomarca	MCMC- MLM –gen con efecto mayor	(10)
Huacaya	0.67	22.2±2.97	Australia	MLM-Modelo animal	(11)
Huacaya	0.369(0.012)	23.07±4.26	Pacomarca	<i>Multitrait</i> -MLM	(12)
Suri	0.46[0.36-0.55]	24.71±4.88	Pacomarca	MCMC-MLM – sin gen con efecto mayor	(10)
Suri	0.69[0.61-0.76]	24.71±4.88	Pacomarca	MCMC-MLM –gen con efecto mayor	(10)
Suri	0.699(0.018)	24.73±5.01	Pacomarca	<i>Multitrait</i> -MLM	(12)

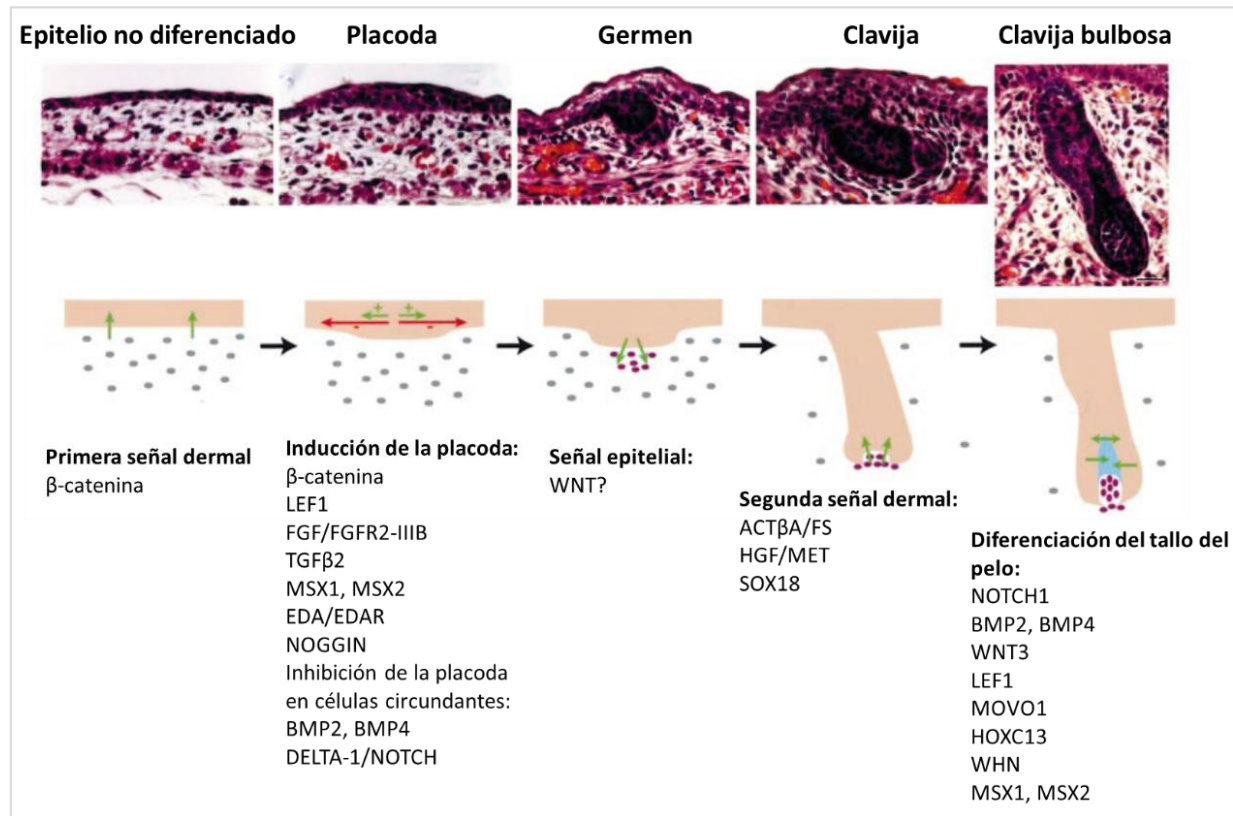
‡: MLM = Modelo Lineal Mixto. MCMC = *Markov Chain* Monte Carlo (método bayesiano). Los métodos no bayesianos emplearon el algoritmo REML (máxima verosimilitud restringida). El método “*Multitrait*” ajusta simultáneamente varios rasgos fenotípicos.\*: Los valores entre paréntesis corresponden al error estándar en los métodos REML. Los valores entre corchetes son la región de alta probabilidad posterior al95% para los métodos bayesianos

Dos estudios han encontrado microsatélites asociados al diámetro de la fibra(15,16). Sin embargo, no se ha estimado la proporción de la heredabilidad explicada por estos marcadores. Además, no se conoce la ubicación de estos microsatélites en el genoma de alpaca. Asimismo, estos estudios se basaron en una asociación indirecta y no se exploró la posible ubicación de las variantes genéticas funcionales. El inconveniente con las asociaciones indirectas (basadas en desequilibrio de ligamiento entre el marcador genético y la variante causal desconocida), es que no se pueden extrapolar fácilmente de una población a otra debido a que el perfil de desequilibrio de ligamiento de una población depende de su historia demográfica en particular (17).

La mayoría de estudios de asociación de genes candidatos en diversos tipos de ganado se han enfocado en los genes que codifican proteínas estructurales. Por ejemplo, los estudios que buscan genes asociados al diámetro de la lana de oveja emplearon como genes candidatos aquellos que codifican keratinas y proteínas asociadas a keratina (18–20). Destaca un estudio que encontró evidencia de ligamiento entre el diámetro de la lana y el gen que codifica la proteína estructural HGT-KAP6-1 (proteína asociada a Keratina 6-1 alta en glicina y tirosina) (21). Otro estudio encontró ligamiento entre el diámetro de la lana de oveja y la región 200cM del cromosoma 3 de oveja (20). Los autores de ese trabajo propusieron que el gen KRT2, que se localiza en esa zona, sería responsable del ligamiento; sin embargo, no realizaron un análisis formal de su hipótesis. Cabe destacar que en esta región genómica se han identificado dos genes regulador asociados al rizado de la lana (22), uno de los cuales (GPRC5A) regularía la diferenciación epitelial durante la etapa embrionaria (23). Asimismo un trabajo que buscó genes asociados al diámetro

de la fibra de alpaca se enfocó en proteínas asociadas a keratina (24), y un estudio que buscó identificar el locus responsable del fenotipo suri eligió el gen TCHH como candidato (25). Este gen codifica la proteína estructural tricohialina, asociada al rizado del pelo en humanos (26).

Las proteínas regulatorias también pueden influir sobre la morfología del pelo de ratón y de humano, cuyo desarrollo es muy similar al de la fibra de alpaca a nivel celular y molecular (Fig. 2) (27). El desarrollo del pelo, la fibra e incluso las plumas empieza con la inducción de un engrosamiento de la epidermis (conocido como placoda) por la dermis subyacente (primera señal dermal), seguida por la inducción del condensado dermal por la placoda recién formada (primera señal epitelial), seguida de la proliferación y diferenciación del epitelio de la placoda debido a señales provenientes de la *papila dermal* (*segunda señal dermal*), que deriva del condensado dermal. Durante este último paso, de proliferación y diferenciación, se forman los diferentes apéndices, que pueden ser pelo, lana, plumas, etc.



**Figura 2.** Señales que intervienen en la secuencia de eventos durante el desarrollo del pelo. Imagen adaptada de ref. (27).

La alteración experimental de proteínas que regulan el desarrollo del pelo puede afectar la morfología y la apariencia de este. Por ejemplo, los ratones que expresan un mutante dominante negativo del gen *Fgfr2-IIIb* (*dnFgfr2-IIIb*) presentan un pelo anormalmente delgado, con apariencia sedosa y brillante y con menor número de columnas celulares en la médula del pelo (28). Estos resultados son reminiscentes de la correlación positiva ( $r = 0.75$ ) entre la proporción de fibras con médula y el diámetro promedio de la fibra de alpaca (2). Además, la expresión ectópica de IGFBP5 (una proteína regulatoria que se sobreexpresa en ratones *dnFgfr2-IIIb*) produce un fenotipo similar a los ratones *dnFgfr2-IIIb* (28).

Se han identificado SNP en genes reguladores asociados a las características morfológicas del pelo y la lana. Por ejemplo, en un estudio de asociación de genoma completo (GWAS, *Genome-Wide Association Study*) en ovejas, se identificaron 15 SNP asociados al rizado y 9 SNP asociados al diámetro de la lana (22). Ninguno de estos SNP se ubica en genes que codifican proteínas estructurales. En contraste, los 10 SNP intragénicos (3 asociados al diámetro y 7 al rizado de la lana) se ubican en genes que codifican proteínas con funciones regulatorias (unión a ADN, transporte de bases nitrogenadas fosforiladas, transducción de señales de receptores asociados a proteínas G, división del ciclo celular, etc.). Asimismo, la sustitución p.V370A (valina por alanina en la posición 370 de la cadena de aminoácidos), causada por el polimorfismo c.1540T>C (timina por citosina en la posición 1540 de la región codificante) del gen *EDAR* humano está asociado a un incremento de 4.2-4.5  $\mu\text{m}$  en el diámetro del pelo (29,30). Además, cada copia del alelo T del SNP rs4752566 del gen *FGFR2* (Receptor tipo 2 del factor de crecimiento de fibroblastos) humano está asociado a un incremento de 1.4-3.4  $\mu\text{m}$  en el diámetro del pelo. Este efecto

fenotípico estaría mediado por diferencias en la expresión genética del mismo gen (31).

Este estudio empleó dos enfoques diferentes, uno “directo” y uno “indirecto”, para identificar genes relacionados a características morfológicas de la fibra de alpaca de probable importancia productiva. El enfoque directo consistió de un análisis de asociación genética cuyo propósito fue encontrar polimorfismos asociados específicamente al diámetro de la fibra, por ser este uno de los principales factores determinantes de su costo. Este enfoque tiene la ventaja de que si se identificaran polimorfismos asociados al diámetro, estos podrían incluirse en un programa de *selección asistida por marcadores*. De otro lado, este enfoque tiene la desventaja de que no permitiría detectar ningún posible efecto de estos genes sobre algún otro rasgo productivo no medido. Se escogió el diámetro de la fibra y no el tipo de fibra para el análisis de asociación debido a que las alpacas de raza suri son escasas y su muestreo hubiera sido complicado.

El enfoque indirecto consistió en la realización de varios análisis de evolución neutral. Estos análisis se basaron en que la selección, ya sea natural o artificial, deja marcas reconocibles (“*selection signatures*”) sobre el perfil de diversidad de los genes sujetos a presión de selección (32). La ventaja de este método es que no requiere información fenotípica. Por lo tanto, si alguno de los genes candidatos afectara algún rasgo productivo no medido (es decir, diferente del diámetro), este fenómeno podría ser detectado mediante las pruebas de evolución neutral, mas no mediante el análisis de asociación. La desventaja de este método es que puede ser

difícil deducir qué rasgo fenotípico es responsable de una posible huella de selección detectada.

El enfoque indirecto empleado en este estudio se fundamenta en que los mecanismos de la selección natural o artificial presentan dos características potencialmente útiles para identificar genes asociados a rasgos productivos. Primero, solo los rasgos monogénicos u oligogénicos producirían huellas de selección lo suficientemente fuertes como para ser detectadas mediante las pruebas de evolución neutral empleadas en este estudio. Segundo, las huellas de selección (“*selection signatures*”) tienen una extensión limitada, usualmente de pocas decenas de megabases *o menos*. Por lo tanto, la detección de una huella de selección sugeriría que algún gen cercano ejerce una fuerte influencia sobre un rasgo fenotípico de probable relevancia productiva.

En la primera fase de este estudio se identificaron las variantes en un grupo de 22 genes reguladores. Dada la semejanza entre los procesos celulares y moleculares que gobiernan el desarrollo del pelo en diferentes mamíferos, la selección de genes reguladores se basó en 1) la identificación de los genes que regulan pasos claves en el desarrollo del pelo en *Mus musculus* y humano; 2) genes que contienen polimorfismos asociados al diámetro del pelo en otras especies, específicamente humano y ratón por ser las especies mejor estudiadas. Se seleccionó el gen *BMP4* por poseer tres variantes no sinónimas cercanas entre sí que producen grandes cambios en las características físico-químicas de los residuos aminoácidos.

Cabe destacar que la selección del gen regulador candidato priorizó el análisis de asociación. Ello debido a que un resultado significativo en la prueba de asociación otorgaría información puntual sobre variantes genéticas que podrían incluirse en programas de selección asistida por marcadores. En contraste, si se encontrase un gen bajo presión de selección sería difícil inferir qué rasgo específico es el responsable de las huellas de selección en dicho gen, y su posible utilidad en programas de mejoramiento genético requeriría ser confirmada mediante estudios adicionales.

En la segunda fase de este estudio se describió la variabilidad del gen regulador BMP4 a nivel nacional. También se describió la variabilidad genética de BMP4 y los estructurales KAP1-1 y KAP13-1, además de 10 microsatélites, en una muestra de la Empresa de Propiedad Social (EPS) Rural Alianza de Puno. La elección de los genes BMP4, KAP1-1 y KAP13-1 se fundamenta en la metodología. El propósito de los microsatélites fue establecer un punto de referencia que permita discernir fenómenos demográficos (estructuración poblacional y consanguinidad) de fenómenos específicos a los genes candidatos (p.ej. selección). En la tercera fase, se emplearon varias pruebas de evolución neutral para determinar si estos genes han estado sujetos a presión de selección reciente. Finalmente, los tres genes fueron incluidos en un análisis de asociación genética con el diámetro de alpaca.

Este es el primer estudio de asociación genética realizado en alpacas que empieza por la identificación de variantes genéticas en una extensa lista de genes candidatos y prioriza los polimorfismos incluidos en el análisis de asociación de acuerdo a su impacto funcional esperado (cambios del marco de lectura,

terminación prematura de la traducción, sustituciones puntuales sinónimas, no sinónimas, en regiones UTR, promotoras, etc.). Este es, también, el primer estudio que busca evidencia de huellas de selección como método indirecto para identificar regiones genómicas asociadas a rasgos de posible relevancia productiva.

## II. MARCO TEÓRICO

### 1. Genética de poblaciones

#### 1.1. Desviaciones del equilibrio de Hardy-Weinberg

La estratificación poblacional y la endogamia (“*inbreeding*”) incrementan la probabilidad de cometer errores estadísticos tipo I y tipo II en los estudios de asociación genética (33). Por lo tanto, estos estudios deben empezar por evaluar si estos fenómenos están presentes en la población estudiada.

La estratificación poblacional y la endogamia reducen la el número de heterocigotos con respecto al esperado bajo equilibrio de Hardy-Weinberg. La relación entre estos dos fenómenos y la deficiencia total de heterocigotos se describe por las siguientes ecuaciones (34):

$$\text{Fit} = 1 - (\text{Heterocigosidad observada} / \text{Heterocigosidad esperada})$$

$$1 - \text{Fit} = (1 - F_{st}) (1 - F_{is})$$

Donde el índice Fit refleja la deficiencia total de heterocigotos debido a la endogamia ( $F_{is}$ ) y estratificación poblacional ( $F_{st}$ ). La heterocigosidad esperada se calcula mediante la ecuación de Hardy-Weinberg.

En una población en la que no existe estratificación poblacional ( $F_{st} = 0$ ), se puede atribuir la deficiencia de heterocigotos a lo largo de todo el genoma a la endogamia. Sin embargo, si solo algunos marcadores presentan deficiencia de heterocigotos, altos valores de  $F_{is}$  pueden deberse a fenómenos con un efecto de

menor rango, tales como selección natural o artificial, o apareamiento preferencial entre individuos fenotípicamente similares entre sí (“*assortative mating*”) (35).

## **1.2. Evolución neutral**

La selección, sea esta natural o artificial, deja marcas distintivas en el genoma y su detección puede servir para identificar loci que influyen el *fitness* biológico (36). En especies domésticas, encontrar estas huellas puede dar indicios sobre la ubicación de genes que afectan rasgos fenotípicos de relevancia agropecuaria.

Existen varias pruebas de hipótesis que evalúan si un gen se encuentra sujeto a presión de selección (37). La característica común que tienen muchas de estas pruebas es que consideran como hipótesis nula que el gen evoluciona neutralmente y que el tamaño de la población se ha mantenido constante durante un periodo prolongado; de tal modo que la diversidad genética ha alcanzado un equilibrio neutral.

A continuación se detalla el fundamento de algunas pruebas de selección que evalúan si un gen se encuentra en equilibrio neutral. Para ello, se partirá de un sistema simple sin mutación ni recombinación y bajo las condiciones del modelo de Wright-Fisher, el cual asume generaciones discretas, individuos hermafroditas, tamaño poblacional constante, producción de un número infinito de gametos (de tal modo que el muestreo de gametos ocurre con reemplazo, según la distribución binomial) y la misma capacidad reproductiva para todos los individuos de la población (ausencia de selección natural o artificial). Estas restricciones se irán eliminando una por una.

La deriva génica es la variación aleatoria de las frecuencias alélicas de una población debido a que en cada generación se transmite un número finito de gametos. Una de las consecuencias a largo plazo es la disminución de diversidad genética debido a la pérdida o fijación de alelos (35).

La fuerza de la deriva génica puede medirse por la probabilidad de que un individuo posea dos alelos idénticos por descendencia (IBD). Una consecuencia del tamaño finito de las poblaciones reales es que conforme transcurre el tiempo, incrementa la probabilidad de que un individuo porte dos alelos que provienen de un ancestro común varias generaciones en el pasado (alelos IBD, *identical by descent*). La intensidad de la deriva génica también puede medirse por la varianza de las frecuencias alélicas entre dos generaciones consecutivas o entre réplicas de una misma población obtenidas, por ejemplo, por subdivisión poblacional o experimentalmente (38).

La fuerza de la deriva génica depende del tamaño poblacional. Tanto la varianza de las frecuencias alélicas como la probabilidad de que un individuo porte dos alelos IBD son inversamente proporcionales al tamaño muestral. En una población ideal bajo el modelo de Wright-Fisher(35):

$$F_t \equiv \Pr(\text{IBD}) = 1 - (1 - F_0) \left(1 - \frac{1}{2N}\right)^t$$

Ecuación 1

$$\sigma_{p(t)}^2 = p_0(1 - p_0) \left[1 - \left(1 - \frac{1}{2N}\right)^t\right]$$

Donde:

- $F_t$  = Coeficiente de endogamia (*Inbreeding coefficient*) en la generación “t”. Es la probabilidad de que un individuo muestreado aleatoriamente de la población posea dos alelos IBD.
- $F_0$  = Coeficiente de endogamia en la generación inicial.
- $\sigma_{p(t)}^2$  = Varianza de las frecuencias alélicas luego de “t” generaciones
- $p_0$  = Frecuencia alélica en la generación inicial.
- $N$  = Tamaño poblacional (se asume constante).

El *tamaño poblacional efectivo* ( $N_e$ ) mide la fuerza de la deriva génica en poblaciones reales (39). Las ecuaciones anteriores fueron desarrolladas para describir la evolución de poblaciones ideales según el modelo de Wright-Fisher. El tamaño efectivo es el tamaño que debería tener una población ideal para que la deriva génica actúe con la misma fuerza con que actúa sobre una población real.

Bajo evolución neutral, la diversidad genética en un locus alcanza un punto de equilibrio. La pérdida de heterocigocidad debida a la deriva génica es contrarrestada por el surgimiento de mutaciones *de novo*. Definiendo la heterocigocidad ( $H_t$ ) como la probabilidad de que un individuo *no* porte dos alelos idénticos por descendencia (es decir,  $H_t = 1 - F_t$ ), entonces de la ecuación 1 se deduce:

$$H_{t+1} = \lambda H_t$$

Ecuación 2

Donde  $\lambda = 1-1/(2N)$  y se asume que  $F_0=0$  (es decir, se considera la población inicial como punto de referencia). Cabe destacar que en este contexto la

homocigocidad y heterocigocidad están definidos en términos de identidad por descendencia (IBD), no por identidad por estado (IBS, *identical by state*). Ello significa que un individuo puede portar dos alelos IBS (es decir, con secuencia idéntica), pero puede considerarse heterocigoto si ambos alelos no son rastreados a un ancestro común en la población de referencia. La heterocigocidad en la siguiente generación dependerá de la tasa de disminución de la heterocigocidad debido a la deriva génica ( $1/2N$ ), y la tasa de mutación por generación por alelo ( $\mu$ ):

$$H_{t+1} = \lambda H_t \left[ \underbrace{(1 - 2\mu)}_{\text{Heterocigocidad en ausencia de mutación}} + \underbrace{2\mu \left( \frac{k-2}{k-1} \right)}_{\text{Probabilidad de que ninguno de los alelos de un individuo heterocigoto mute}} + \underbrace{2\mu(1 - \lambda H_t)}_{\text{Probabilidad de que un heterocigoto mute a un alelo que no porta}} + \underbrace{2\mu(1 - \lambda H_t)}_{\text{Probabilidad de que un homocigoto mute a heterocigoto}} \right]$$

Ecuación 3

Donde “k” es el número de alelos diferentes, y el término  $2\mu(k-2)/(k-1)$  es la probabilidad de que un heterocigoto siga siendo heterocigoto luego de mutar.

Si la tasa de mutación y el tamaño poblacional se mantienen constante durante un tiempo prolongado, la heterocigocidad alcanza un punto de **equilibrio neutral entre deriva y mutación**; es decir,  $H_{t+1} = H_t$ . En esta condición, la heterocigocidad esperada es:

$$E[H] = \frac{\theta}{1 + \left( \frac{\theta k}{k-1} \right)}$$

Ecuación 4

Donde “k” es el número de alelos diferentes;  $\theta = 4N\mu$ . Para una población no ideal, se reemplaza “N” por “Ne”. El parámetro  $\theta$  puede ser interpretado como la razón entre de la tasa de creación de heterocigotos a partir de homocigotos ( $2\mu$ ) y la tasa de pérdida de heterocigocidad por deriva génica ( $1/(2N)$ ) (40).

Si la unidad de análisis es el gen completo, k toma valores mucho mayores que la unidad, y la ecuación 4 se reduce al modelo de alelos infinitos propuesto por Kimura y Crow (41):

$$E[H] \cong \frac{\theta}{1 + \theta}$$

Usualmente, el valor de  $\theta$  es mucho menor que la unidad (40), por lo tanto, la heterocigocidad esperada es aproximada por el valor de este parámetro, el cual depende del tamaño poblacional efectivo:

$$E[H] \cong \theta = 4N_e\mu$$

Ecuación 5

Bajo evolución neutral, varios parámetros poblacionales que miden la diversidad genética dependen del estimador  $\theta$ (39). Estos parámetros son el número de sitios segregantes (S), la diversidad nucleotídica ( $\pi$ ), la distribución de frecuencias alélicas (SFS, *Site Frequency Spectrum*), etc. Asimismo, estos estadísticos pueden ser empleados para estimar  $\theta$  a partir de un grupo de secuencias. Bajo equilibrio neutral, todos los estimadores tienen valores esperados similares; sin embargo, cuando no se cumple esta condición, la discrepancia entre estos estimadores brinda indicios sobre las fuerzas evolutivas y fenómenos demográficos que han afectado a la población estudiada.

A continuación se muestran los estimadores basados en el número de sitios segregantes ( $\hat{\theta}_S$ ) y en el número de diferencias promedio entre pares de secuencias ( $\hat{\theta}_\pi$ ), los cuales son empleados en las pruebas de selección más usadas, tales como las pruebas D de Tajima (42), Fs de Fu (43), y R2 de Ramos-Onsins y Rozas (44).

### 1.2.1. Estimador de $\theta$ basado en el número de sitios segregantes (S)

El estimador de  $\theta$  basado en el número de sitios segregantes (posiciones de una secuencia en las que existen al menos dos alelos) fue planteado inicialmente por Watterson(45), pero a continuación se presenta en líneas generales el fundamento de una demostración alternativa basada en teoría de coalescencia (39).

Bajo el modelo de sitios infinitos cada mutación nueva genera un sitio segregante. Por lo tanto, el número de sitios segregantes esperado en un número determinado de muestras puede ser estimado por el número de mutaciones esperadas bajo evolución neutral mediante teoría de coalescencia (convergencia de los linajes actuales en un solo linaje ancestral).

Si la tasa de mutación por generación en un linaje (un haplotipo) es  $\mu$ , entonces el número esperado de mutaciones de “k” linajes durante “t” generaciones es “ $k\mu t$ ”. El número total de mutaciones durante la evolución de un grupo de muestras se obtiene por:

$$E[S] = E \left[ \sum_{k=2}^n \mu k T_k \right]$$

Ecuación 6

La probabilidad de que “k” linajes coalescan en una generación dada es:

$$\text{Pr}(\text{coalescencia}) = \frac{k(k-1)}{2} \frac{1}{2N_e}$$

El número de generaciones para que los “k” linajes coalescan sigue una distribución geométrica. Por lo tanto, el tiempo esperado es la inversa de la probabilidad:

$$E[T_k] = \frac{4N_e}{k(k-1)}$$

Ecuación 7

Reemplazando la ecuación 7 en la 6:

$$E[S] = 4N_e\mu \sum_{k=1}^{n-1} \frac{1}{k}$$

Reemplazando  $\theta$  por  $4N_e\mu$ :

$$\theta = \frac{E[S]}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

El estimador de Watterson ( $\hat{\theta}_W$  o  $\hat{\theta}_S$ ) emplea el número observado de sitios segregantes para estimar el valor de  $\theta$ . Este estimador puede presentarse con respecto a la secuencia completa ( $\hat{\theta}_{LW}$ ) o con respecto a cada nucleótido ( $\hat{\theta}_W$ ). Es decir:  $\hat{\theta}_W = \hat{\theta}_{LW}/L$ , donde  $L$  es la longitud de la secuencia analizada. Estos valores son estimados por:

$$\hat{\theta}_{LW} = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

$$\hat{\theta}_W = \frac{p_S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

Donde  $p_S = S/L$  es la proporción de sitios segregantes.

### 1.2.2. Estimador de $\theta$ basado en la diversidad nucleotídica ( $\pi$ )

La diversidad nucleotídica ( $\pi$ ) es el número promedio de diferencias entre cada par de secuencias de una muestra (46). Este valor equivale a la heterocigocidad esperada ( $\theta$ ) en un grupo de secuencias con evolución neutral bajo el modelo de sitios infinitos. La diversidad nucleotídica de la secuencia ( $\hat{\theta}_{\pi_L}$ ) es calculada mediante la siguiente ecuación:

$$\hat{\theta}_{\pi_L} = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^n \sum_{j>i}^n d_{ij}$$

Donde  $d_{ij}$  es el número de diferencias entre las secuencias “i” y “j”, y  $n$  es el número de secuencias. La diversidad nucleotídica por sitio es:  $\hat{\theta}_{\pi} = \hat{\theta}_{\pi_L}/L$ .

### 1.3. Efectos de la selección y variación del tamaño poblacional

La selección actúa mediante varios mecanismos (32). La selección negativa consiste en la eliminación de mutaciones que disminuyen el *fitness* biológico, mientras que la selección positiva promueve la fijación de mutaciones beneficiosas. La mayoría de las mutaciones no neutrales son deletéreas (47), es por ello que la selección negativa actúa constantemente. Las mutaciones beneficiosas son muy raras; es por ello que se suele considerar que la selección positiva actúa de manera periódica (48). Finalmente, la selección balanceante se da cuando el heterocigoto posee mayor *fitness* que los homocigotos. A diferencia de la selección positiva y

negativa, la selección balanceante incrementa la diversidad genética, ya que promueve la retención de ambos alelos(49).

La varianza del número de descendientes que procrea cada individuo afecta el tamaño poblacional efectivo. Bajo el modelo de Wright-Fisher, la relación entre el tamaño poblacional real, el tamaño poblacional efectivo, y la varianza del número de descendientes está dada por la siguiente ecuación (40):

$$Ne = \frac{2N_{t-1}}{\frac{\sigma_k^2}{\mu_k} + \mu_k - 1}$$

Ecuación 8

Donde:

- Ne: Tamaño efectivo poblacional en la generación actual.
- $N_{t-1}$ : Tamaño real de una población ideal en la generación anterior.
- k = Número de gametos transmitidos por cada individuo de la generación anterior.
- $\mu_k$  = Número promedio de gametos transmitidos por cada individuo de la generación anterior. Toma un valor de dos para poblaciones con tamaño constante
- $\sigma_k^2$  = Varianza del número de gametos transmitidos por cada individuo de la generación anterior.

La selección modifica el tamaño poblacional efectivo en todo el genoma. El modelo de Wright-Fisher asume que todos los individuos tienen la misma capacidad reproductiva, por lo tanto, la varianza en el número de descendientes se debe únicamente al azar. Sin embargo; en presencia de selección, la varianza en el fitness

incrementa la varianza del número de gametos que transmite cada individuo a la siguiente generación; por lo tanto, se reduce el tamaño efectivo poblacional (50). Este efecto es global; es decir, afecta el  $N_e$  de todos los loci en el genoma.

La selección disminuye aun más la diversidad genética en las regiones genómicas ligadas a las mutaciones bajo presión de selección. Cuando surge una mutación deletérea, la selección negativa disminuye la frecuencia de los alelos que conforman el haplotipo sobre el que surgió la mutación. Ello se debe a que el ligamiento no permite que los loci circundantes evolucionen independientemente. A esta disminución de la diversidad genética en loci neutrales ligados a loci con mutaciones deletéreas se le denomina selección de fondo (*background selection*) (51). Asimismo, cuando surge una mutación beneficiosa, el haplotipo en el que esta surgió tiende a fijarse en la población(52). A este fenómeno se le conoce como “barrido selectivo fuerte” (*hard selective sweep*). También es posible que un alelo neutral o casi neutral aumente su ventaja selectiva debido a cambios ambientales. Este alelo puede encontrarse en varios haplotipos, los cuales persistirán luego de que el alelo beneficioso se fije en la población por acción de la selección positiva. A este fenómeno se le denomina “barrido selectivo débil” (*soft selective sweep*)(53). El incremento de la frecuencia de los alelos neutrales ligados al locus bajo selección positiva se denomina “*hitch-hiking* genético”(52).

La pérdida de diversidad genética local debido a la selección implica una disminución en la heterocigocidad y el tamaño poblacional efectivo cerca al locus bajo presión de selección. La magnitud de la disminución de la heterocigocidad y el  $N_e$  depende de la varianza del fitness debido a la selección (54), la frecuencia de

recombinación entre el locus neutral y el locus bajo selección(55), la reducción de la varianza genética aditiva por generación debido a los efectos de la selección (50), la tasa de mutación, el coeficiente de selección (40,55), y el tiempo esperado entre dos barridos selectivos en el caso de *hitch-hiking* genético (56).

La selección y las variaciones del tamaño de una población afectan la topología de la genealogía. La selección balanceante promueve la retención de ambos alelos debido a que el heterocigoto posee mayor fitness(49). Ello ocasiona que en las regiones circundantes se observen alelos con frecuencia intermedia y un elevado desequilibrio de ligamiento. De otro lado, la selección positiva fuerte y la selección negativa usualmente incrementan el número de sitios con frecuencias alélicas bajas debido a que las mutaciones *de novo* perjudiciales tienen menos probabilidades de incrementar su frecuencia por azar y las mutaciones beneficiosas se fijan rápidamente y desplazan a los alelos ancestrales (barrido selectivo fuerte)(32).

En un evento de selección positiva debido a cambios en el ambiente (barrido selectivo débil), el alelo beneficioso suele encontrarse en más de un haplotipo. Ello reduce el número de haplotipos y sitios segregantes, e incrementa el perfil de desequilibrio de ligamiento (57). Cuando finaliza un barrido selectivo fuerte, se pierde toda la diversidad genética en las regiones circundantes. La disminución de diversidad genética en los loci cercanos es inversamente proporcional a la distancia entre estos y la mutación *de novo* seleccionada positivamente (40).

Las expansiones poblacionales incrementan el número de sitios segregantes con frecuencias alélicas bajas debido a que disminuye la probabilidad de que las

mutaciones *de novo* se pierdan por deriva génica. Considérese una población diploide bajo el modelo de Wright-Fisher con tamaño inicial  $N$  y tamaño final  $N_f$ . La frecuencia de una mutación *de novo* en la población inicial es  $1/(2N)$ , valor que también representa la probabilidad de que un gameto muestreado al azar porte la mutación. El número de gametos que originan la generación final es  $2N_f$ . Se puede modelar el número de gametos que portan el alelo mutante mediante una distribución de Poisson con parámetro  $\lambda = 2N_f \times \frac{1}{2N} = \frac{N_f}{N}$ . Por lo tanto, la probabilidad de que existan cero copias del alelo mutante en la generación final es  $e^{-\frac{N_f}{N}}$  (resultado que también ha sido obtenido mediante otro método (35)). Por lo tanto, cuando existen expansiones poblacionales, las mutaciones *de novo* tienen mayor probabilidad de sobrevivir a los efectos de la deriva génica e incrementa el número de sitios segregantes detectados en un grupo de muestras.

Las reducciones poblacionales disminuyen el número de sitios segregantes, además, los sitios segregantes detectados tienen frecuencias alélicas intermedias. Ello se debe a que los alelos con frecuencias bajas son eliminados rápidamente por acción de la deriva génica, y solo sobreviven los alelos que tenían frecuencias intermedias. Los cuellos de botella siguen una dinámica más compleja que depende de si el evento sigue en marcha o si ya acabó, así como de la magnitud de la reducción poblacional y de la expansión subsiguiente (37).

La selección y los eventos demográficos pueden tener efectos similares en los niveles de diversidad genética. Sin embargo; *la selección deja marcas focalizadas alrededor de los loci sujetos a presión de selección. En contraste, los fenómenos demográficos tienen efectos uniformes en todo el genoma* (39).

## 1.4. Pruebas de hipótesis de equilibrio neutral

### 1.4.1. Prueba D de Tajima:

Bajo equilibrio neutral entre deriva y mutación, se espera que los estimadores  $\hat{\theta}_W$  y  $\hat{\theta}_\pi$  tengan valores similares. Sin embargo, cuando el equilibrio no se cumple, estos dos estimadores toman diferentes valores. Ello se debe a que el estimador  $\hat{\theta}_\pi$  es más sensible a las frecuencias alélicas que  $\hat{\theta}_W$  debido a que los alelos más comunes son contabilizados más veces en las comparaciones entre pares de secuencias.

El test de Tajima evalúa si el número promedio de diferencias entre pares de secuencias concuerda con el valor esperado a partir del número de sitios segregantes observados si la población se encontrara en equilibrio neutral (42). El estadístico empleado en esta prueba es el siguiente:

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\text{Var}(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

Valores negativos de D indican selección negativa o positiva o expansión poblacional; mientras que los valores mayores a cero indican cuellos de botella en marcha, selección balanceante o estructuración poblacional. La significancia de la prueba estadística se calcula comparando el valor observado del estadístico D con la distribución esperada bajo la hipótesis nula de equilibrio neutral. Esta hipótesis nula se puede aproximar mediante una distribución Beta o mediante simulación de genealogías condicionadas en el número de sitios segregantes (S) o un estimador de  $\theta$ (42).

### 1.4.2. Prueba $F_s$ de $F_u$

Este test se basa en la Fórmula de Muestreo de Ewens (58), que predice la distribución de frecuencias alélicas bajo equilibrio neutral. A partir de esta fórmula,  $F_u$  planteó el estadístico  $F_s$  que evalúa si existe un exceso de alelos raros con respecto a la distribución esperada según el estimador  $\hat{\theta}_\pi$ (43). Esta prueba es bastante sensible a la selección direccional y expansiones poblacionales, situaciones en las cuales el estadístico  $F_s$  toma valores negativos. La significancia se evalúa mediante simulaciones bajo equilibrio neutral empleando un estimador de  $\theta$ .

### 1.4.3. Prueba $ZnS$ de Kelly

Esta prueba se basa en el perfil de desequilibrio de ligamiento esperado bajo equilibrio neutral. El estadístico  $ZnS$  es el desequilibrio de ligamiento promedio (medido mediante el coeficiente de correlación  $r^2$  de Hill y Robertson(59)) en una región con  $S$  sitios segregantes:

$$ZnS = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^S r_{ij}^2$$

La prueba  $ZnS$  puede emplearse como una prueba de dos colas. Cuando existe un barrido selectivo fuerte en ausencia de recombinación  $ZnS$  toma valores cercanos a cero (60). Ello se debe a que los barridos fuertes eliminan los alelos que existían antes de que empiece el barrido y los sitios segregantes observados, creados por mutaciones recientes, presentan valores bajos de  $r^2$  debido a su baja frecuencia poblacional. El mismo comportamiento se observa cuando existe selección negativa

o expansión poblacional bajo neutralidad (61). Sin embargo, bajo un barrido selectivo fuerte en presencia de recombinación (por ejemplo, en un locus neutral a una distancia intermedia al sitio bajo selección) o un barrido selectivo débil, varios haplotipos permanecen en la población con frecuencia intermedia, y los sitios segregantes en estos alcanzan valores de  $r^2$  altos (57,60). En una población de tamaño constante, un locus que evoluciona neutralmente pero en presencia de recombinación también mostrará valores bajos de  $ZnS$  debido a que el desequilibrio de ligamiento es destruido por acción de la recombinación (61).

La significancia del estadístico  $ZnS$  se obtiene comparando el valor observado con la distribución bajo la hipótesis nula de equilibrio neutral. Esta distribución se obtiene simulando genealogías condicionadas en el número de sitios segregantes. También se puede incluir el parámetro poblacional de recombinación  $R = 4N_eC$  (donde  $C$  es la frecuencia de recombinación por generación en toda la secuencia). Para ello, se debe estimar  $R$ , por ejemplo, mediante el método de Hudson (62).

#### 1.4.4. Prueba $R2$ de Ramos-Onsins y Rozas

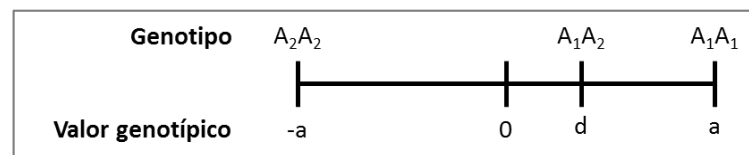
El estadístico  $R2$  se basa en que el número esperado de *singletons* luego de una expansión poblacional es igual a  $k/2$ , donde  $k$  es el número promedio de diferencias entre pares de secuencias (44). El estadístico  $R2$  se calcula mediante la siguiente ecuación:

$$R2 = \frac{\left( \sum_{i=1}^n \left( U_i - \frac{k}{2} \right)^2 / n \right)^{1/2}}{S}$$

El estadístico  $R^2$  adopta valores cercanos a cero cuando ha ocurrido una expansión poblacional. La significancia estadística se calcula simulando genealogías bajo la hipótesis nula.

## 2. Principios de genética cuantitativa y modelos de asociación genética

Asumiendo que un rasgo cuantitativo está influenciado por un polimorfismo bialélico, el fenotipo promedio de una población puede describirse a partir del fenotipo promedio de los portadores de cada genotipo y por la frecuencia de cada genotipo (63) (fig. 3). El promedio fenotípico de cada genotipo recibe el nombre de “*valor genotípico medio*”. En la figura 3, el fenotipo ha sido centrado en el punto equidistante a los valores genotípicos medios de los homocigotos  $A_1A_1$  y  $A_2A_2$ . En esta escala, el valor genotípico medio de  $A_2A_2$  es “ $-a$ ” y el valor genotípico medio de  $A_1A_1$  es “ $a$ ”. El valor genotípico medio del heterocigoto  $A_1A_2$  es “ $d$ ”, y equivale a la distancia entre el fenotipo promedio de los heterocigotos y el punto equidistante (el punto 0) entre los valores genotípicos de  $A_1A_1$  y  $A_2A_2$ . (13)



**Figura 3.** Representación gráfica de los valores genotípicos  $-a$ ,  $d$  y  $a$  para un modelo de un locus (Figura adaptada de (13)).

Si “ $d$ ” tiene un valor igual a cero y “ $a$ ” es diferente de cero, significa que el valor genotípico medio está determinado únicamente por el número de copias de los alelos  $A_1$  y  $A_2$ , por lo que se considera que el locus  $A$  solo presenta efectos aditivos; es decir, los efectos genéticos contribuidos por este locus equivalen

exactamente a la suma de los efectos de los dos alelos portados. Si  $d$  y  $a$  son diferentes de cero, el locus presenta dominancia (parcial o total); es decir, los efectos genéticos contribuidos por este locus se pueden dividir en efectos aditivos (la suma de los efectos de los dos alelos) más los efectos de dominancia (la desviación observada respecto al valor genotípico medio que se esperaría observar si solo existieran efectos aditivos). Si  $a = d = 0$ , no existe asociación entre el fenotipo y el locus A (64).

El “*valor genotípico medio de la población*” ( $\mu$ ) está determinado por las frecuencias genotípicas y los valores genotípicos medios de cada genotipo (63). Bajo equilibrio de Hardy-Weinberg, la media poblacional de un rasgo cuantitativo es:

$$\mu = p^2(a) + 2pq(d) + q^2(-a)$$

$$\mu = a(p^2 - q^2) + 2pqd = a(p - q) + 2pqd$$

Ecuación 9

Donde:

- $p$  = frecuencia del alelo  $A_1$
- $q = 1-p$  = frecuencia del alelo  $A_2$

El parámetro  $\mu$  se encuentra en la misma escala que  $a$  y  $d$ , es decir, la medición se realiza con respecto al punto medio entre los valores genotípicos medios de  $A_2A_2$  y  $A_1A_1$  (el punto “0” en la figura 3).

A partir de la ecuación 9 se puede calcular el efecto promedio de cada alelo (*average allelic effect*), definido como la desviación respecto a la media poblacional

del valor fenotípico promedio de todos los individuos portadores de al menos una copia de dicho alelo (13). El valor genotípico promedio dado que al menos uno de los alelos es  $A_1$  estará definido por:

$$a \times Pr(A_1 \text{ se una con otro alelo } A_1) + d \times Pr(A_1 \text{ se una con un alelo } A_2) \\ = ap + dq$$

Por analogía, el valor genotípico promedio dado que al menos uno de los alelos es  $A_2$  será:

$$-aq + dp$$

Entonces:

$$\alpha_1 = (ap + dq) - \mu$$

$$\alpha_1 = q[a + d(q - p)]$$

Ecuación 10

$$\alpha_2 = (-aq + dp) - \mu$$

$$\alpha_2 = -p[a + d(q - p)]$$

Ecuación 11

Donde  $\alpha_1$  y  $\alpha_2$  representan el *efecto alélico promedio* de los alelos  $A_1$  y  $A_2$ , respectivamente.

Además:

$$\alpha = \alpha_1 - \alpha_2 = a + d(q - p)$$

Ecuación 12

Donde  $\alpha$  es el *efecto promedio de la sustitución alélica*, que representa el cambio esperado en el valor genotípico medio de un genotipo al reemplazar un alelo  $A_2$  por un alelo  $A_1$  (65). Reemplazando la ecuación 4 en las ecuaciones 2 y 3, se obtiene:

$$\begin{aligned}\alpha_1 &= q\alpha \\ \alpha_2 &= -p\alpha\end{aligned}$$

Por definición, los efectos alélicos promedio son aditivos. La suma de los efectos alélicos de un genotipo se denomina *valor genotípico aditivo*, *efecto genético aditivo* o *breeding value de un genotipo*. Si se suman, conceptualmente, los efectos alélicos a lo largo de todo el genoma, se obtiene el *breeding value de un individuo*. Un individuo transmite aleatoriamente un alelo por locus a cada descendiente; es decir, transmite la mitad de los efectos genéticos aditivos (13). Por lo tanto, el *breeding value* o valor genotípico aditivo de un individuo equivale al doble de la desviación del valor genotípico medio de su descendencia respecto al valor genotípico medio de la población (ya que la medición del *breeding value* se realiza con respecto a la media poblacional):

Desviación de la descendencia respecto a la media poblacional

$$\begin{aligned}&= \sum_k (\alpha_{ki} + \alpha_{kj}) \\ &= (\textit{breeding value del progenitor})/2\end{aligned}$$

Ecuación 13

Donde  $\alpha_{ki}$ ,  $\alpha_{kj}$  son los efectos alélicos de un individuo con alelos “i”, “j” en el locus “k”, y la suma se realiza sobre todos loci del genoma.

El *breeding value* de un genotipo suele diferir del valor genotípico medio. Es decir, los efectos genéticos no dependen únicamente de la suma de los efectos alélicos. A la diferencia entre estos dos valores se le denomina dominancia (desviación del modelo netamente aditivo), y estadísticamente representa la interacción entre los alelos de un genotipo. Por lo tanto el valor genotípico medio se puede representar mediante el siguiente modelo lineal general (64):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

Ecuación 14

Donde:

- $Y_i$ : Fenotipo del individuo “i”
- $X_i$ : *Score* genotípico. Número de copias del alelo  $A_1$  menos uno.
- $\beta_1$ : Coeficiente de regresión para los efectos alélicos aditivos (“a” en la fig. 3)
- $Z_i$ : *Score* de heterocigosidad. Cero para los homocigotos, y uno para los heterocigotos
- $\beta_2$ : Coeficiente de regresión para los efectos alélicos no aditivos. Para este modelo de un solo locus corresponde a la dominancia (“d” en la fig. 3).
- $\varepsilon_i$ : Residual del individuo i, porción del fenotipo no explicada por el modelo.

A partir de este modelo se pueden calcular  $\alpha_1$  y  $\alpha_2$ , y, por lo tanto, los efectos aditivos transmisibles de una generación a la siguiente.

### **III. OBJETIVOS**

#### **1. Objetivo general**

- Determinar la posible asociación de tres genes con rasgos seleccionados para la producción de fibra en una población de alpacas

#### **2. Objetivos específicos**

- Identificar un gen regulador de la ruta de señalización del desarrollo de la fibra de alpaca con polimorfismos de posible consecuencia funcional.
- Describir la variabilidad genética del gen regulador seleccionado a nivel nacional.
- Describir la variabilidad genética del gen regulador seleccionado y los genes estructurales KAP1-1 y KAP13-1 en una población de alpacas sometidas a selección artificial
- Evaluar si existen huellas de selección (*selection signatures*) en el gen regulador seleccionado y en los genes estructurales de la fibra de alpaca KAP1-1, KAP13-1 en una población de alpacas sometidas a selección artificial.
- Explorar la posible asociación de los genes antes mencionados con el grosor de la fibra de alpaca en la misma población.

## **IV. METODOLOGÍA**

En la figura 4 se presenta un resumen gráfico de la metodología. En este estudio se analizó el gen regulador BMP4 y los genes estructurales KAP1-1 y KAP13-1. Su elección se fundamenta más adelante.

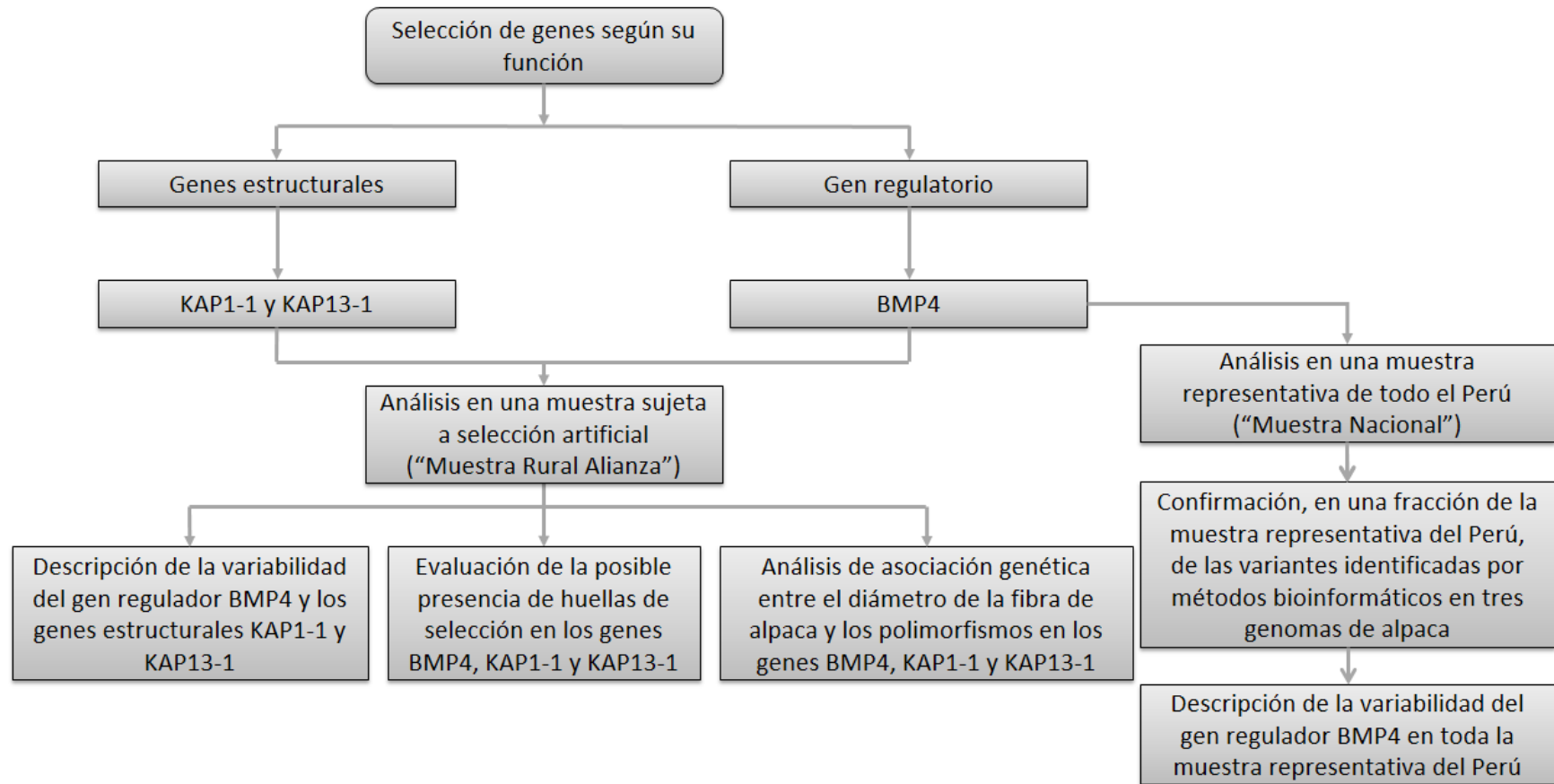
### **1. Diseño**

Estudio observacional descriptivo tipo transversal. Estudio de asociación genética basado en población de individuos no emparentados (*population-based genetic association study*).

### **2. Muestreo**

Se emplearon dos muestras diferentes para diferentes propósitos:

- Una muestra representativa de todo el Perú para describir la variabilidad genética de un gen regulador (BMP4) analizado por vez primera en alpaca.
- Una muestra para realizar el análisis de evolución neutral y de asociación genética con el gen regulador y los dos genes estructurales elegidos.



**Fig.4.** Resumen gráfico de las actividades que componen este trabajo de investigación.

### **2.1. Población de interés:**

- Para la descripción de la variabilidad del gen regulador (BMP4) a nivel nacional: Alpacas (*Vicugna pacos*) provenientes de Perú
- Para el análisis de asociación genética y de evolución neutral del gen regulador (BMP4) y los dos genes estructurales (KAP1-1 y KAP13-1): Alpacas del Perú sujetas a selección para características productivas de la fibra de alpaca.

### **2.2. Población elegible:**

- Para la descripción de la variabilidad del gen regulador (BMP4) a nivel nacional: Muestras de alpaca de todo el Perú cuyo ADN genómico se encuentra almacenado en la Unidad de Biotecnología Molecular-Señalización Celular (UBM-SC) de los Laboratorios de Investigación y Desarrollo de la Universidad Peruana Cayetano Heredia (LID-UPCH)
- Para el análisis de asociación genética y de evolución neutral del gen regulador (BMP4) y los dos genes estructurales KAP1-1 y KAP13-1: Alpacas de la Empresa de Propiedad Social (EPS) Rural Alianza (Puno, Perú) cuyo ADN genómico se encuentra almacenado en la UBM-SC.

### **2.3. Tamaño muestral:**

- Para la descripción de la variabilidad del gen regulador(BMP4) a nivel nacional:

Se utilizó el programa MINSAGE (66) para el cálculo del mínimo número de muestra necesario para detectar con un 99% de probabilidad un alelo con una frecuencia del 5% para marcadores bialélicos fuera de equilibrio de Hardy-Weinberg (HWE). Se consideró ausencia de HWE debido a que esta muestra es heterogénea (alpacas provienen de diferentes partes del Perú), por lo que se esperaba deficiencia de heterocigotos debido al efecto Wahlund (39).

- Para el análisis de asociación genética y de evolución neutral del gen regulador (BMP4) y los dos genes estructurales (KAP1-1 y KAP13-1):

Se analizaron todas las muestras provenientes de la EPS Rural Alianza disponibles en la UBM-SC que pasaron los criterios de inclusión y exclusión. Este número fue de 151 individuos .

#### **2.4. Criterios de inclusión:**

- Muestras no emparentadas

#### **2.5. Criterios de exclusión:**

- Para la descripción de la variabilidad del gen regulador a nivel nacional: No se empleó ningún criterio de exclusión.
- Para el análisis de asociación genética y de evolución neutral del gen regulador (BMP4) y los dos genes estructurales (KAP1-1 y KAP13-1): Muestras de individuos sin información fenotípica.

### 3. Selección de genes candidatos:

– Selección de dos genes estructurales:

En la UBM-SC se ha estandarizado la amplificación y se ha analizado el perfil de variabilidad genética a nivel nacional de 5 proteínas asociadas a keratina (KAP1-1, KAP6-1, KAP9-2, KAP11-1 y KAP13-1) (24). Además, se ha realizado un análisis de asociación del gen KAP11-1 con el diámetro, sin encontrarse ninguna asociación significativa. De otro lado, la amplificación y secuenciación del gen KAP6-1 fue dificultosa debido a la presencia de pseudogenes. Por lo tanto, en este estudio se eligieron los genes KAP1-1 y KAP13-1 para el análisis de asociación y para la búsqueda de huellas de selección en la muestra de alpacas de Rural Alianza. Estos genes no se analizaron en la muestra representativa de todo el Perú en este estudio, ya que dicho análisis ya fue realizado en anteriormente por Francesco Foppiano (24).

– Selección de un gen regulador:

El origen evolutivo y ontogénico del pelo de humano y de ratón es muy similar al de la fibra de alpaca(27). Por ello, la selección de los genes de alpaca candidatos para el análisis de asociación genética se basó en la identificación de genes ortólogos humanos y murinos que intervienen en el desarrollo del pelo en dichas especies. Se preseleccionaron genes reguladores con variantes genéticas asociadas al diámetro del pelo en otras especies (ser humano)y genes que al ser silenciados, eliminados o sobreexpresados provocaron variaciones en el fenotipo del pelo en la

especie *Mus musculus*. Finalmente, se escogió el gen BMP4 debido a que fue el que presentó mayor cantidad de variantes genéticas que podrían tener un importante impacto sobre la función de la estructura y función de las proteínas (ver más adelante).

#### **4. Identificación de variantes genéticas en los genes candidatos:**

(Nota: En este manuscrito se usa el término “variante genética” y “sitios segregantes” para hacer referencia tanto a mutaciones, cuya frecuencia poblacional es menor a 1%, como a polimorfismos de un nucleótido (SNP), que se distinguen por tener una frecuencia poblacional mayor a 1%.)

Esta etapa del proyecto consistió en el análisis bioinformático de 3 genomas de alpaca de raza Huacaya, obtenidos en un proyecto de la UBM-SC (Unidades de Biotecnología Molecular y Señalización Celular) financiado por FINCyT (119-FINCYT-IA-2013). Una de las alpacas fue de fibra fina (aprox. 18 micras), las otras dos de fibra intermedia-gruesa (24 -26 micras). Todo el análisis bioinformático se realizó en el servidor de la Unidad de Biotecnología Molecular (sistema operativo Ubuntu 14.04).

##### **4.1. Preprocesamiento de genomas secuenciados**

El formato VCF (*Variant Call Format*) es el formato estándar utilizado para representar las variantes genéticas identificadas en el genoma de una muestra de individuos con respecto a un genoma de referencia. La preparación de este archivo es el primer paso para identificar las variantes genéticas candidatas a utilizarse en el estudio de asociación genética.

Primero, se indizó el genoma de referencia (Vicugna\_pacos-2.0.2: obtenido de <http://www.ncbi.nlm.nih.gov/genome/905>) usando el programa *samtools faidx*. La indización permite acelerar el acceso a la información contenida en el archivo del genoma de referencia.

A continuación, se utilizó el programa *samtools mpileup* para crear un archivo en formato BCF (*Binary variant Call Format*), que contiene la verosimilitud (*likelihood*) de los datos (es decir, del alineamiento de los 3 genomas que obtuvimos contra el genoma de referencia) según cada genotipo posible observado a lo largo de todo el genoma. En términos simples, la verosimilitud de los alineamientos indica qué tan bien se han alineado los millones de fragmentos secuenciados al genoma de referencia. Estos valores de verosimilitud son utilizados por el programa en el siguiente paso para obtener los índices de calidad de los genotipos, que indican la probabilidad de que el genotipo asignado en cada posición sea correcto o incorrecto.

Luego, se utilizó el programa *bcftools call* para asignar los genotipos a partir de los valores de verosimilitud estimados en el paso anterior, obteniéndose un archivo en formato VCF (*Variant Call Format*). El archivo VCF describe los genotipos a lo largo de todo el genoma de los 3 individuos, y contiene un índice de calidad ("*Phred quality score*") para cada posición que difiere del genoma de referencia; es decir, por cada SNP (*Single Nucleotide Polymorphism*) o INDEL (*Insertion/deletion*). El *Phred Quality Score (Q)* se calcula del siguiente modo:

$$Q = - 10 \log_{10}(P)$$

Donde  $P$  es la probabilidad de que el genotipo sea incorrecto. Por ejemplo: Un valor  $Q = 30$  representa una probabilidad de 0.001 de que el genotipo sea incorrecto.

Debido al gran tamaño de los archivos VCF, estos fueron indizados para acelerar su exploración usando el programa tabix.

El formato VCF presenta gran flexibilidad para representar genotipos que difieren del genoma de referencia. Sin embargo, dicha flexibilidad ocasiona que el mismo genotipo pueda ser representado de diferentes formas en diferentes ejecuciones del programa, lo que resta reproducibilidad y dificulta el intercambio de información entre diferentes instituciones. Es por ello que los archivos VCF deben pasar por un proceso de normalización, y descomposición, especialmente cerca de INDELS (inserciones/delecciones) y polimorfismos multinucleótido. Los detalles de este procedimiento se pueden encontrar en el siguiente enlace de la Universidad de Michigan: [http://genome.sph.umich.edu/wiki/Variant\\_Normalization](http://genome.sph.umich.edu/wiki/Variant_Normalization). La normalización, descomposición y recorte de las variantes se realizó mediante los programas *vt normalize*, *vt decompose* y *vt decompose\_blocksub*.

#### **4.2. Identificación de variantes genéticas en los genes candidatos**

La identificación y selección de las variantes ubicadas en los genes candidatos se realizó utilizando los paquetes Rsamtools (67), VariantAnnotation (68), rtracklayer (69) y GenomicFeatures (70) del proyecto Bioconductor (71) para R mediante los pasos siguientes:

- Se obtuvo la anotación versión 2.02 del genoma de alpaca a partir de la base de datos de genomas del NCBI (<http://www.ncbi.nlm.nih.gov/genome/905>) y se indizó usando el programa tabix.
- Se utilizó la función `import.gff` del paquete `rtracklayer` para crear un objeto `TxDb` (*TranscriptDataBase*) utilizado por el paquete `GenomicFeatures` para acceder a la anotación del genoma.
- Se cargó el genoma de referencia y su índice mediante la función `FaFile` del paquete `Rsamtools`. Se cargó el archivo VCF junto con su índice (el cual contiene las variantes genéticas, su ubicación en el genoma, los genotipos y los índices de calidad) utilizando las funciones `readVcf` y `TabixFile` del paquete `VariantAnnotation`.
- Una vez cargados el genoma de referencia, las variantes genéticas (VCF) y la anotación del genoma (GFF), se seleccionaron solamente las variantes bialélicas con un índice de calidad  $\text{Phred} > 30$ ; es decir, variantes cuyos genotipos fueron asignados con una probabilidad mayor a 0.999 de ser correctos.

### **4.3. Predicción de los efectos de las variantes genéticas**

En esta etapa del estudio se buscó determinar qué variantes genéticas podrían tener un mayor impacto sobre la función de las proteínas que las contienen. Para ello, se predijeron los cambios a nivel de secuencia de aminoácidos, de sitios de splicing y en secuencias cercanas a los sitios de inicio de la transcripción, ya que podrían afectar la transcripción de los genes contiguos.

Se emplearon las funciones `predictCoding`, `locateVariants` y `subsetByOverlaps` del paquete `VariantAnnotation` para seleccionar las variantes en sitios codificantes y las cercanas a los sitios de inicio de transcripción (TSS) y sitios de *splicing* de los genes candidatos. Las únicas regiones no codificantes incluidas en la búsqueda de variantes fueron aquellas cercanas a los TSS.

Finalmente, se seleccionó el gen `BMP4` como gen candidato para el estudio de asociación genética por poseer 3 variantes no sinónimas que, al estar muy cerca entre sí, pueden ser genotipificadas mediante una única reacción de secuenciamiento Sanger. Además, tal como se describe más adelante, estas tres variantes genéticas provocan cambios bioquímicos severos en la secuencia de aminoácidos.

## **5. Estandarización de la amplificación del gen `BMP4`**

Se empleó la herramienta en línea primer-BLAST (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>) del NCBI para diseñar los cebadores para amplificar y secuenciar un fragmento del primer exón del gen `BMP4`, el cual contiene las 3 variantes identificadas en este gen.

Esta herramienta implementa el algoritmo del conocido programa `PRIMER3`, pero posee la ventaja adicional de realizar un alineamiento `BLAST` de los cebadores obtenidos contra el genoma de referencia; lo que en principio representa una PCR *in silico* que predice si los cebadores producirán productos de PCR inespecíficos.

Se amplificó el gen `BMP4` empleando un par de cebadores que flanquean una región de 394 pares de bases (anexo 3). Se realizó una PCR a 4 concentraciones de

MgCl<sub>2</sub> y 4 temperaturas de apareamiento (*annealing*) diferentes. Todas las reacciones se realizaron en un volumen total de 20 uL empleando 20ng de ADN (cuantificado en NanoDrop). Los detalles del protocolo se presentan en las tablas 2 y 3.

**Tabla 2. Condiciones para la estandarización de la amplificación del gen BMP4**

Reactivo	Concentración de la solución de trabajo	Volumen empleado (uL)	Concentración final
<i>Buffer Hot Start</i>	10X	2	1X
dNTPs	2mM	2	0.2mM
MgCl <sub>2</sub>	25mM	0.8 / 1.2 / 1.6 / 2	1 / 1.5 / 2 / 2.5 mM
<i>Forward primer</i>	10 uM	1	0.5 uM
<i>Reverse primer</i>	10 uM	1	0.5 uM
<i>HS Taq polymerase</i>	5 U/uL	0.1	0.5 U
Agua	-	11.1 / 10.7 / 10.3 / 9.9	-
ADN	10ng/uL	2	1 ng/uL

**Tabla 3. Condiciones de ciclaje de la estandarización de la amplificación del gen BMP4**

Temperatura	Tiempo	Ciclos
94 °C	4 m	1
94 °C	30 s	
56/58/60/62 °C	45 s	30
72°C	40 s	
72°C	4 m	1
4°C	∞	1

## 6. Amplificación de los genes KAP1-1 y KAP13-1

En un estudio previo conducido por Francesco Foppiano(24), se identificaron y validaron variantes genéticas en los genes que codifican las proteínas KAP1-1 y KAP13-1. Estos dos genes también se incluirán en el análisis de asociación genética, por lo que las condiciones de amplificación se detallan en las tablas 4-6, y las secuencias de los cebadores, en el anexo 3.

## 7. Electroforesis en agarosa

Se confirmó la presencia de amplicones mediante electroforesis en agarosa. Los genes KAP1-1 y KAP13-1, cuyos amplicones medían 800pb, aproximadamente, fueron detectados en geles de agarosa al 1% durante 50 min de

corrida a 90V. Los amplicones del gen BMP4, de 394 pb, fueron detectados en geles de agarosa al 2.5% durante 40 min de corrida a 90V.

## **8. Secuenciamiento mediante método Sanger**

Una vez que se confirmó la amplificación mediante electroforesis en agarosa, las muestras fueron enviadas a la empresa Macrogen (Corea) para ser secuenciadas mediante el método Sanger (*Dideoxy nucleotide chain termination*) en ambas direcciones utilizando un analizador genético 3730XL (Applied Biosystems).

Los archivos AB1 obtenidos fueron analizados con el software CodonCode. Primero se cortaron los extremos de las secuencias para maximizar las regiones con una probabilidad de error menor a 1%. Las secuencias recortadas con menos de 90% de bases con un Phred score < Q30 fueron descartadas por ser consideradas de baja calidad. Luego, las secuencias recortadas fueron alineadas y se realizó una edición manual en caso de ser necesario. Finalmente, se determinaron los genotipos mediante la herramienta de búsqueda de mutaciones incorporada en CodonCode y se realizó una inspección visual para descartar errores de genotipificación.

**Tabla 4.** Condiciones de la amplificación de los genes KAP1-1 y KAP13-1

Reactivo	Concentración de la solución de trabajo	Volumen empleado (uL)	Concentración final
Buffer Hot Start	10X	2	1X
dNTPs	2mM	2	0.2mM
MgCl <sub>2</sub>	25mM	1.6	2mM
Forward primer	10 uM	1	0.5 uM
Reverse primer	10 uM	1	0.5 uM
HS Taq Polymerase	5 U/uL	0.1	0.5 U
H <sub>2</sub> O	-	10.3	-
ADN	10ng/uL	2	1 ng/uL

**Tabla 5.** Condiciones de ciclaje de la amplificación del gen  
KAP1-1

Temperatura	Tiempo	Ciclos
94 °C	4 m	1
94 °C	30 s	
64 °C	45 s	30
72°C	1 m	
72°C	4 m	1
4°C	∞	1

**Tabla 6.** Condiciones de ciclaje de la amplificación del gen KAP13-1

Temperatura	Tiempo	Ciclos
94 °C	4 m	1
94 °C	30 s	
62 °C	45 s	30
72°C	55 s	
72°C	5 m	1
4°C	∞	1

### **9. Estudio exploratorio del gen BMP4 en una fracción de la muestra representativa de todo el Perú**

El poder estadístico de los estudios de asociación genética depende de la frecuencia poblacional de los polimorfismos analizados. Por ello, resulta necesario confirmar que las variantes genéticas del gen *BMP4* sean polimorfismos (frecuencia poblacional mayor a 1%) y no mutaciones (frecuencia poblacional menor a 1%). Este paso es de suma importancia ya que la identificación de variantes genéticas se realizó a partir de secuenciamiento de genomas completos, en los cuales se identifican todas las variantes de un individuo, muchas de las cuales pueden tener una baja frecuencia poblacional o incluso ser mutaciones privadas (mutaciones propias de un solo individuo).

Este análisis exploratorio empleó una submuestra aleatoria de 30 individuos provenientes de la muestra de 96 individuos que representan de manera

proporcional las poblaciones de alpaca de diferentes regiones del Perú. Los pasos para la validación fueron los siguientes:

- Se formaron 10 *pools* (mezclas) de 3 muestras de alpaca cada uno y se mandó a secuenciar a Macrogen (Corea del Sur) mediante el método Sanger. Este paso tuvo dos objetivos: a) confirmar que las variantes identificadas en la fase bioinformática no fuesen errores de genotipificación y b) confirmar que esta variante no es una mutación privada y que tiene una frecuencia poblacional suficientemente alta como para ser observada en al menos una de las 10 mezclas; es decir, en al menos una de las 30 muestras escogidas aleatoriamente.
- Luego de descartar en el paso anterior que las variantes observadas en la fase bioinformática fuesen errores de secuenciamiento, se secuenció la placa completa (n=96) de muestras representativas de todo el Perú para determinar las frecuencias alélicas, la heterocigosidad esperada, la heterocigosidad observada y el índice Fis.

Con el objetivo de abreviar, a la muestra representativa de todo el Perú (n = 96) se denominará de ahora en adelante “muestra Nacional”.

## **10. Análisis genético-poblacional**

Los estudios de asociación genética se ven afectados por fenómenos demográficos como la estratificación poblacional y la endogamia (“*inbreeding*”)(33). Estos fenómenos también pueden afectar la diversidad genética y producir efectos que dificultarían la identificación de huellas de selección en la población estudiada (37). Por lo tanto, se empleó el software STRUCTURE (72,73) para evaluar si existe estratificación poblacional en las muestras de la EPS Rural

Alianza (Puno). Los resultados de este análisis contribuyeron a distinguir entre los efectos de la selección y de los fenómenos demográficos sobre los resultados de las pruebas de equilibrio neutral. Los resultados de STRUCTURE también ayudaron a definir si era necesario incluir una covariable adicional correspondiente a la subpoblación (latente) de origen en el análisis de asociación genética.

Este programa emplea un enfoque bayesiano para detectar de cuántas subpoblaciones latentes genéticamente diferenciadas provienen las muestras. El algoritmo evalúa si, dadas las frecuencias alélicas observadas, la probabilidad posterior de las frecuencias genotípicas observadas incrementaría si las muestras provinieran de diferentes subpoblaciones genéticamente diferenciadas. Es decir, encuentra el número de subpoblaciones latentes (cada una con diferentes frecuencias alélicas) que maximiza la probabilidad de los genotipos observados. Para este análisis, se utilizaron 10 microsatélites validados para análisis de parentesco (74) que fueron previamente genotipificados en la Unidad de Biotecnología Molecular.

Los parámetros empleados fueron los siguientes:

- *Burn in Period* = 100000. Corresponde al número de muestras aleatorias del espacio de parámetros durante los cuales se debe alcanzar la convergencia hacia una zona de alta credibilidad posterior.
- *Number of MCMC after burn in*: 200000. Luego de descartar las 100000 primeras muestras de parámetros, se generan 200000 muestras adicionales. Durante cada una de estas 200000 muestras, se calcula la probabilidad posterior de los valores muestreados.

- *Admixture model*: Se considera que los individuos de la muestra podrían presentar mestizaje entre diferentes poblaciones ancestrales. Si fuese el caso, STRUCTURE, trata de detectar las poblaciones ancestrales.
- *Allele frequencies correlated*: Se considera que algunos alelos pueden estar correlacionados entre sí.

La presencia de endogamia (“*inbreeding*”) fue evaluada mediante el índice Fis por el método de Weir & Cockerham (75) implementado en el software Genepop (76). Este método permite estimar el índice Fis cuando se trabaja con muestras de tamaño finito. Los valores de heterocigocidad esperada y observada fueron calculados con el software Cervus (77).

El desequilibrio de ligamiento fue calculado con el software DnaSP v5.1 (78). Los bloques de desequilibrio de ligamiento fueron graficados con el programa Haploview v4.2 (79).

El programa DnaSP requiere que las secuencias diploides estén separadas en haplotipos (2 secuencias por cada individuo diploide). Por lo tanto, se empleó el software PHASE v2.1 para reconstruir la fase haplotípica (80). Los parámetros empleados fueron los siguientes:

- Modelo con recombinación
- Posible existencia de un *hot spot* de recombinación (Modelo MR1)
- Estimación inicial (*prior*) del parámetro de recombinación poblacional ( $4N_e c$ ): 0.0004. Este valor equivale a un tamaño poblacional efectivo de

10000 y frecuencia de recombinación igual a 1cM/Mb. El tamaño efectivo poblacional fue obtenido de un estudio de filogenómica de camélidos (81).

- Probabilidad umbral de haplotipos: 0.95. Es el valor umbral de la probabilidad posterior de la fase estimada para considerar que la fase inferida es correcta.

## **11. Pruebas de hipótesis de equilibrio evolutivo neutral en la muestra Rural Alianza**

Se empleó el software DnaSP v.5.1 para calcular los estimadores de  $4N_e\mu$  y realizar todas las pruebas de neutralidad detalladas en el marco teórico: D de Tajima,  $F_s$  de Fu,  $Z_nS$  de Kelly,  $R_2$  de Ramos-Onsins y Rozas y la prueba de homocigocidad de Ewens-Watterson. Además, se graficó la distribución de diferencias entre pares de secuencias (MMD, *Mismatch Distribution*) y los intervalos de confianza al 95 y 99% usando el software Arlequin v3.5(82). Este programa no pudo calcular los intervalos de confianza para el gen KAP13-1 debido a “problemas de convergencia”, aparentemente debido a su escasa diversidad genética. Por lo tanto, la MMD de este gen fue graficada usando DnaSP y no incluye los intervalos de confianza. El *p-value* de todas las pruebas de neutralidad fue calculado comparando el valor observado del estadístico respectivo (D,  $F_s$ ,  $R_2$ ,  $Z_nS$ ) con la distribución observada en 1000 genealogías simuladas bajo la hipótesis neutral, tal como se detalla en el marco teórico.

## **12. Análisis de asociación genética**

Incluir todas las variantes en el análisis de asociación incrementaría la probabilidad de encontrar al menos una asociación falsa, ya que incrementa el *Family-wise error rate* (83). Este problema puede ser resuelto mediante correcciones para pruebas de hipótesis múltiples. Sin embargo, estos métodos vuelven las pruebas más conservadoras(83). Además, la potencia estadística para detectar asociación con variantes raras es muy baja, por lo que su inclusión en un estudio trae pocos beneficios a menos que se cuente con un tamaño muestral muy grande (84).

Por lo tanto, se descartaron los SNPs poco informativos, cuya inclusión en el análisis de asociación no tendría mayor impacto en la capacidad para detectar asociaciones reales, pero que sí incrementarían la probabilidad de cometer al menos un error tipo I. Solo se incluyeron los tag-SNPs en el análisis de asociación.

Los tag-SNPs son variantes que capturan, mediante desequilibrio de ligamiento, la mayor proporción posible de la variabilidad genética (85). Al incluir solo tag-SNPs en el análisis de asociación se reduce el número de pruebas de hipótesis realizadas. Además, este método no excluye la posibilidad de emplear correcciones para pruebas de hipótesis múltiples si se considera necesario. Una ventaja adicional de emplear tag-SNPs es que se evitan problemas de multicolinealidad en el análisis de asociación.

La selección de tag-SNPs se realizó con el método basado en el factor de inflación de la varianza (VIF) implementado en el software PLINK (86). Este método consiste en realizar una regresión lineal múltiple de un SNP a partir de todos los demás, y descartar todos los SNP que alcancen un coeficiente de determinación ( $R^2$ ) mayor a un valor umbral establecido. El valor de umbral empleado en este estudio fue 0.5.

El modelo de asociación se basó en los principios de genética cuantitativa descritos en el marco teórico. El modelo empleado fue el detallado en la ecuación 14:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

Ecuación 14

Donde:

- $Y_i$ : Fenotipo del individuo “i”
- $X_i$ : *Score* genotípico. Número de copias del alelo  $A_1$  menos uno.
- $\beta_1$ : Coeficiente de regresión para los efectos alélicos aditivos (“a” en la fig. 3)
- $Z_i$ : *Score* de heterocigosidad. Cero para los homocigotos, y uno para los heterocigotos
- $\beta_2$ : Coeficiente de regresión para los efectos alélicos no aditivos. Para este modelo de un solo locus corresponde a la dominancia (“d” en la fig. 3).
- $\varepsilon_i$ : Residual del individuo i, porción del fenotipo no explicada por el modelo.

El análisis estadístico se realizó en R (87).

### 13. Análisis de potencia estadística

La potencia estadística del análisis de asociación genética se calculó mediante dos métodos. El primer método empleado fue uno analítico, que se basa en la distribución esperada de los coeficientes de regresión cuando la hipótesis alternativa es correcta. Los detalles se encuentran en el anexo 2.

También se estimó la potencia estadística mediante simulaciones en R. La simulación se basa en el modelo presentado en la ecuación 14. En breve:

- Se simularon varios escenarios en los que un SNP en equilibrio de Hardy-Weinberg con diferentes frecuencias alélicas y diferentes valores genotípicos medios influyen sobre el fenotipo.

- Se consideró la diferencia observada en la muestra entre el diámetro de fibra de los machos y las hembras. Además, se empleó la proporción de machos y hembras observada en la muestra real.
- Se añadieron valores residuales (“error”) con distribución normal y media cero, de tal modo que la varianza total sea igual a la observada en la población real.
- Se empleó el siguiente modelo, que considera que no existe covarianza entre los términos a la derecha de la ecuación:

$$\sigma_{total}^2 = \sigma_{Aditiva}^2 + \sigma_{Dominancia}^2 + \sigma_{sexo}^2 + \sigma_{residual}^2$$

Las varianzas aditiva, de dominancia y debida al sexo fueron calculadas de acuerdo a las ecuaciones presentadas y demostradas en el anexo 1. La Varianza residual (comúnmente referida como “varianza ambiental”) es la varianza que se debe sumar a las varianzas anteriores para alcanzar la varianza observada en la muestra.

Para cada combinación de frecuencias alélicas y de valores genotípicos medios, se realizaron diez mil réplicas de la simulación. El poder estadístico se calculó como la proporción de las veces en que la hipótesis nula fue rechazada con un nivel de significancia de 0.05.

El error estándar de la potencia estadística estimada corresponde a la desviación estándar de la proporción de una distribución binomial(88), por lo tanto:

$$Error\ estándar(1 - \beta) = \sqrt{\frac{\beta(1 - \beta)}{N}}$$

Donde  $(1-\beta)$  es la potencia o poder estadístico predicho y N es el número de réplicas empleadas en la simulación. El máximo error estándar posible se alcanza cuando  $\beta$  es igual a 0.5. Por lo tanto:

$$Error\ estándar(1 - \beta) \leq \sqrt{\frac{0.5(1 - 0.5)}{10000}} = 0.005$$

La Potencia fue evaluada en un modelo que solo consideraba los efectos aditivos y un modelo que consideraba dominancia incompleta (efectos aditivos y de dominancia). En ambos casos se incluyó el sexo como covariable. También se evaluó si es que alguno de los dos modelos producía estimadores sesgados; es decir, se replicaron nuevamente diez mil sets de datos y se evaluó si la media de los diez mil estimadores correspondía a la media esperada, la cual es especificada al inicio de la simulación.

## V. RESULTADOS

### 1. Descripción de la calidad de la fibra

El número de alpacas hembra y macho fue 136 y 15, respectivamente. El diámetro de la fibra presentó una diferencia no significativa (t-test,  $p\text{-value} = 0.333$ ) de  $0.84\mu\text{m}$  entre ambos sexos (machos-hembras). Asimismo, el diámetro no presentó distribución normal (Shapiro-Wilk  $p\text{-value} = 0.043$ ) en la muestra total; sin embargo, sí presentó distribución normal en las muestras de alpacas macho y hembra por separado (valores de  $p\text{-value}$  de 0.848 y 0.052, respectivamente). La muestra completa (machos y hembras) presentó una fibra de muy alta calidad, con una media de  $20.93\mu\text{m}$  y una desviación estándar de  $3.06\mu\text{m}$  (fig.5). La proporción de fibra de calidad *Baby* (diámetro entre  $19.5$  y  $22.5 \mu\text{m}$ ) y *Royal* (diámetro menor a  $19.5\mu\text{m}$ ) alpaca fue de 39 y 32%, respectivamente.

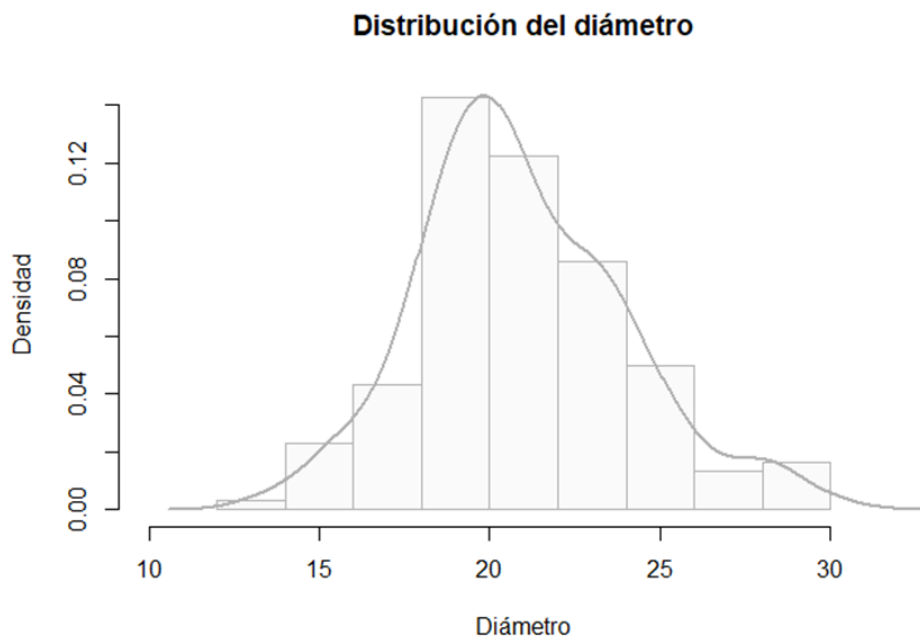


Figura 5. Diámetro de la fibra de alpaca en la muestra completa de 151 alpacas.

## **2. Selección de genes candidatos y descripción de las variantes genéticas identificadas en la fase bioinformática**

Se seleccionaron un total de 22 genes candidatos que intervienen en el desarrollo del pelo de ratón o humanos. Dos de estos genes codifican 4 transcritos diferentes, ya sea por variaciones en el sitio de inicio de la transcripción o por *splicing* alternativo. Asimismo, tres genes codifican dos transcritos diferentes, y los demás codifican un solo transcrito (tabla 7).

Se encontraron 64 variantes en los 22 genes, incluyendo variantes sinónimas, no sinónimas, de *splicing* y circundantes al sitio de inicio de transcripción (TSS). No se buscaron variantes en regiones no codificantes, excepto aquellas cercanas a los TSS. Se encontraron 13 variantes que afectan a 11 transcritos de 9 genes entre las posiciones -500 y +200 respecto al TSS (tabla 8). De las 13 variantes, dos son inserciones cortas; una, delección corta; y 10, polimorfismos de un único nucleótido (SNP). Asimismo, se encontraron 30 variantes (SNPs) sinónimas que afectan a 14 transcritos de 13 genes (tabla 9). En contraste, solo se encontraron 16 SNPs no sinónimos que afectan a 6 transcritos de 6 genes (tabla 10). Asimismo, se encontró una variante de *splicing* (alelo de referencia: A, alelo alternativo: G) en la posición 268 del transcrito XM\_015249781.1 del gen IGFBP3.

**Tabla 7.** Ubicación de los transcritos codificados por los genes reguladores seleccionados

<b>Gen</b>	<b>Scaffold (N° Accesoión)</b>	<b>Transcrito (N° Accesoión)</b>	<b>Inicio</b>	<b>Fin</b>	<b>Tamaño</b>	<b>Hebra</b>
BMP2	NW_005882770.1	XM_006207373.2	3962493	3973973	11481	+
BMP4	NW_005882719.1	XM_015235487.1	6120114	6137151	17038	+
BMP7	NW_005882888.1	XM_006213987.1	530240	617145	86906	-
	NW_005882888.1	XM_015247758.1	530240	617145	86906	-
DKK1	NW_005882763.1	XM_006206627.1	1203527	1206604	3078	-
DKK4	NW_005882908.1	XM_006215147.2	1664980	1668096	3117	+
EDA2R	NW_005882911.1	XM_006215240.2	2662606	2740096	77491	+
EDAR	NW_005882914.1	XM_015248843.1	2246405	2305162	58758	-
EDARADD	NW_005882956.1	XM_015249727.1	1451018	1505076	54059	+
FGF10	NW_005882775.1	XM_015242118.1	5576444	5651573	75130	-
		XM_006207704.2	5576444	5652163	75720	-
FGF7	NW_005882786.1	XM_006208358.2	6797676	6854679	57004	-

**Tabla 7.** Ubicación de los transcritos codificados por los genes reguladores seleccionados

<b>Gen</b>	<b>Scaffold (N° Accesoión)</b>	<b>Transcrito (N° Accesoión)</b>	<b>Inicio</b>	<b>Fin</b>	<b>Tamaño</b>	<b>Hebra</b>
FGFR2	NW_005882722.1	XM_015235881.1	886340	986490	100151	+
FOXN1	NW_005882847.1	XM_006211960.1	499956	525741	25786	-
HOXC13	NW_005882738.1	XM_015238026.1	11647938	11664806	16869	+
IGF1	NW_005882757.1	XM_006205895.1	9091631	9146433	54803	+
		XM_006205898.1	9091631	9159881	68251	+
		XM_006205897.2	9091710	9160402	68693	+
		XM_006205899.1	9093568	9159881	66314	+
IGFBP3	NW_005882964.1	XM_015249781.1	42808	112536	69729	-
IGFBP5	NW_005882769.1	XM_006207273.2	4332417	4351913	19497	+
		XM_015241762.1	4332456	4352366	19911	+
LEF1	NW_005882848.1	XM_015245895.1	4041024	4121659	80636	+
		XM_015245893.1	4041030	4145964	104935	+

**Tabla 7.** Ubicación de los transcritos codificados por los genes reguladores seleccionados

<b>Gen</b>	<b>Scaffold (N° Accesoión)</b>	<b>Transcrito (N° Accesoión)</b>	<b>Inicio</b>	<b>Fin</b>	<b>Tamaño</b>	<b>Hebra</b>
		XM_006212063.2	4041032	4146609	105578	+
		XM_015245894.1	4041032	4146609	105578	+
MSX1	NW_005882949.1	XM_006216368.1	1248130	1251733	3604	-
MSX2	NW_005882716.1	XM_006198966.2	3793921	3798701	4781	-
NOG	NW_005882953.1	XM_006216420.1	1389724	1390332	609	-
WNT10A	NW_005882769.1	XM_006207296.2	2589213	2603317	14105	-
WNT10B	NW_005882738.1	XM_006202929.2	7934852	7941195	6344	-

**Tabla 8.** Variantes en regiones promotoras de los genes candidatos seleccionados

<b>Gen</b>	<b>Scaffold</b>	<b>Inicio</b>	<b>Fin</b>	<b>Transcrito</b>	<b>Posición (TSS*)</b>	<b>Alelo ref.</b>	<b>Alelo alt.</b>
BMP2	NW_005882770. 1	3962549	3962549	XM_006207373.2	+56	T	C
EDAR	NW_005882914. 1	2305064	2305064	XM_015248843.1	+98	G	C
		2305084	2305084		+78	G	A
		2305099	2305099		+63	T	C
EDARA DD	NW_005882956. 1	1451103	1451103	XM_015249727.1	+85	C	G
FGF10	NW_005882775. 1	5652012	5652012	XM_015242118.1	-439	C	T
				XM_006207704.2	+151		
HOXC1 3	NW_005882738. 1	11647945	11647945	XM_015238026.1	+7	T	C

**Tabla 8.** Variantes en regiones promotoras de los genes candidatos seleccionados

<b>Gen</b>	<b>Scaffold</b>	<b>Inicio</b>	<b>Fin</b>	<b>Transcrito</b>	<b>Posición (TSS*)</b>	<b>Alelo ref.</b>	<b>Alelo alt.</b>		
IGF1	NW_005882757. 1	9093347	9093347	XM_006205899.1	-221	C	T		
IGFBP5	NW_005882769. 1	4332483	4332485	XM_006207273.2	+66	TCC	T		
				XM_015241762.1	+27				
				4332487	4332491	XM_006207273.2	+70	TTTTT	TTTTTATTTT
						XM_015241762.1	+31		
				4332493	4332493	XM_006207273.2	+76	T	TTTTTTC
						XM_015241762.1	+37		
				4332552	4332552	XM_006207273.2	+135	T	C
						XM_015241762.1	+96		
MSX2	NW_005882716. 1	3798654	3798654	XM_006198966.2	+47	C	T		

**Tabla 8.** Variantes en regiones promotoras de los genes candidatos seleccionados

Gen	Scaffold	Inicio	Fin	Transcrito	Posición (TSS*)	Alelo ref.	Alelo alt.
-----	----------	--------	-----	------------	--------------------	------------	------------

\*TSS: Posición con respecto a Sitio de inicio de transcripción. Alelo ref: Alelo de referencia. Alelo alt: Alelo alternativo.

**Tabla 9.** Variantes sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicación (CDS)	Ubicación (proteína)	Alelo Ref.	Alelo Alt.	Codón Ref.	Codón Alt.	AA
BMP2	XM_006207373	+	891	297	A	G	AGA	AGG	R
	.2	+	945	315	T	C	GCT	GCC	A
BMP4	XM_015235487	+	102	34	T	C	CTT	CTC	L
	.1	+	342	114	A	C	CCA	CCC	P
DKK4	XM_006215147	+	643	215	C	T	CTA	TTA	L
	.2								

**Tabla 9.** Variantes sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicación (CDS)	Ubicación n (proteína)	Alelo Ref.	Alelo Alt.	Codón Ref.	Codón Alt.	AA
EDAR	XM_015248843 .1	-	39	13	C	A	CTG	CTT	L
		-	537	179	C	T	TCG	TCA	S
		-	660	220	A	G	GCT	GCC	A
FGF7	XM_006208358 .2	-	46	16	T	G	AGA	CGA	R
		-	537	179	C	T	ACG	ACA	T
FGF10	XM_015242118 .1 XM_006207704 .2	-	333	111	T	C	CCA	CCG	P
		-	333	111	T	C	CCA	CCG	P
FGFR2	XM_015235881 .1	+	948	316	C	T	AAC	AAT	N
		+	540	180	G	A	AAG	AAA	K

**Tabla 9.** Variantes sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicación (CDS)	Ubicación n (proteína)	Alelo Ref.	Alelo Alt.	Codón Ref.	Codón Alt.	AA
		+	1731	577	A	G	CCA	CCG	P
		-	108	36	A	G	TCT	TCC	S
		-	294	98	A	G	CTT	CTC	L
FOXN1	XM_006211960	-	780	260	A	G	ACT	ACC	T
	.1	-	588	196	C	T	TCG	TCA	S
		-	1848	616	C	T	TCG	TCA	S
		-	195	65	C	T	GAG	GAA	E
IGFBP3	XM_015249781	-	282	94	A	G	GCT	GCC	A
	.1	-	492	164	G	A	AAC	AAT	N
		-	942	314	G	A	TAC	TAT	Y
MSX2		-	48	16	C	T	GAG	GAA	E

**Tabla 9.** Variantes sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicación (CDS)	Ubicación n (proteína)	Alelo Ref.	Alelo Alt.	Codón Ref.	Codón Alt.	AA
	XM_006198966	-	285	95	T	C	ACA	ACG	T
	.2	-	291	97	G	A	TTC	TTT	F
NOG	XM_006216420	-	300	100	G	A	TGC	TGT	C
	.1	-	348	116	T	A	TCA	TCT	S
WNT10A	XM_006207296	-	948	316	G	A	GGC	GGT	G
	.2								
WNT10B	XM_006202929	-	45	15	C	A	GCG	GCT	A
	.2								

CDS: Secuencia codificante; Ref.: Referencia; Alt.: Alternativo; AA: Aminoácido.

**Tabla 10.** Variantes no sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicaci ón (CDS)	Ubicació n (proteína)	Alelo Ref.	Alelo Alt.	Codó n Ref.	Codón Alt.	AA Ref.	AA Alt.
BMP4	XM_01523548 7.1	+	74	25	G	C	CGG	CCG	R	P
		+	1031	344	G	A	AGC	AAC	S	N
		+	25	9	G	A	GCC	ACC	A	T
		+	71	24	T	G	CTG	CGG	L	R
		+	98	33	A	G	TAC	TGC	Y	C
EDAR	XM_01524884 3.1	-	866	289	C	T	CGG	CAG	R	Q
		-	743	248	G	A	GCG	GTG	A	V
		-	1208	403	A	G	ATC	ACC	I	T
		-	755	252	C	G	CGA	CCA	R	P
		-	694	232	A	G	TGG	CGG	W	R

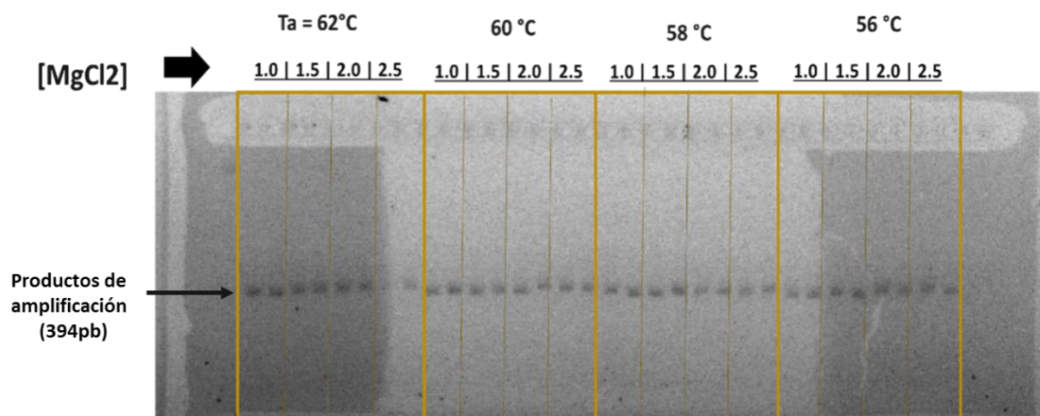
**Tabla 10.** Variantes no sinónimas en genes candidatos

Gen	Transcrito	Hebra	Ubicaci ón (CDS)	Ubicació n (proteína)	Alelo Ref.	Alelo Alt.	Codó n Ref.	Codón Alt.	AA Ref.	AA Alt.
EDAR	XM_01524972									
		+	301	101	C	G	CGG	GGG	R	G
ADD	7.1									
HOXC	XM_01523802									
		+	8	3	T	C	CTA	CCA	L	P
13	6.1									
IGFBP	XM_01524978									
		-	215	72	T	C	CAG	CGG	Q	R
		-	476	159	C	T	CGT	CAT	R	H
3	1.1									
		-	182	61	G	A	GCG	GTG	A	V
LEF1	XM_01524589									
		+	607	203	A	T	AAC	TAC	N	Y
	5.1									

### 3. Estandarización de la amplificación del gen BMP4

La herramienta en línea primer-BLAST produjo una lista de 10 pares de cebadores que podrían emplearse para amplificar la región deseada del gen BMP4. Se escogió el par de cebadores con mayor especificidad, cuyas características se detallan en el anexo 3.

El gen BMP4 fue amplificado en 4 condiciones de temperatura de *annealing* y 4 concentraciones de MgCl<sub>2</sub>. La amplificación fue exitosa en todas las condiciones probadas (Fig. 6). La especificidad de los productos fue confirmada por secuenciamiento Sanger.



**Figura 6.** Resultados de la electroforesis de los amplicones de BMP4 en un gel de agarosa al 2.5%. Se observa que el gen amplificó en todas las condiciones probadas. Las distancias migradas se observan ligeramente disperejas debido a que se utilizó un buffer de corrida con tinte fluorescente incorporado (“*Fluorescent Dye*”) que distorsiona ligeramente las distancias migradas. Se utilizó este reactivo para evitar el uso de bromuro de etidio.

Se escogió 62°C y [MgCl<sub>2</sub>] = 1.0mM como las condiciones de amplificación en las siguientes reacciones de PCR para disminuir la posibilidad de obtener productos inespecíficos.

#### **4. Amplificación del gen BMP4 en la muestra representativa de todo el Perú**

El gen BMP4 fue amplificado en 10 mezclas de 3 muestras cada uno para confirmar que era realmente polimórfico. Una vez que se confirmó mediante secuenciamiento Sanger que este gen era realmente polimórfico, se amplificó en una muestra representativa (muestra Nacional) de las poblaciones de alpaca de todo el Perú (n = 96). Su amplificación fue confirmada mediante electroforesis en geles de agarosa al 2.5%. Estas muestras fueron empleadas solo para la descripción de la variabilidad de BMP4 a nivel nacional mas no para el análisis de evolución neutral ni para el análisis de asociación genética.

#### **5. Amplificación de BMP4, KAP1-1 y KAP13-1 en la muestra Rural Alianza para el análisis de asociación y pruebas de evolución neutral**

Las 151 muestras empleadas en las pruebas de evolución neutral y el análisis de asociación provienen de la EPS Rural Alianza, Puno. Para abreviar, a estas muestras se les denomina “muestra Rural Alianza” en el resto del manuscrito. La amplificación de KAP1-1 y KAP13-1 fue confirmada por electroforesis en agarosa al 1%. La amplificación de BMP4 fue confirmada por electroforesis en agarosa al 2.5%. La especificidad y la calidad de los productos de amplificación fueron confirmadas en los resultados de secuenciamiento (ver más adelante).

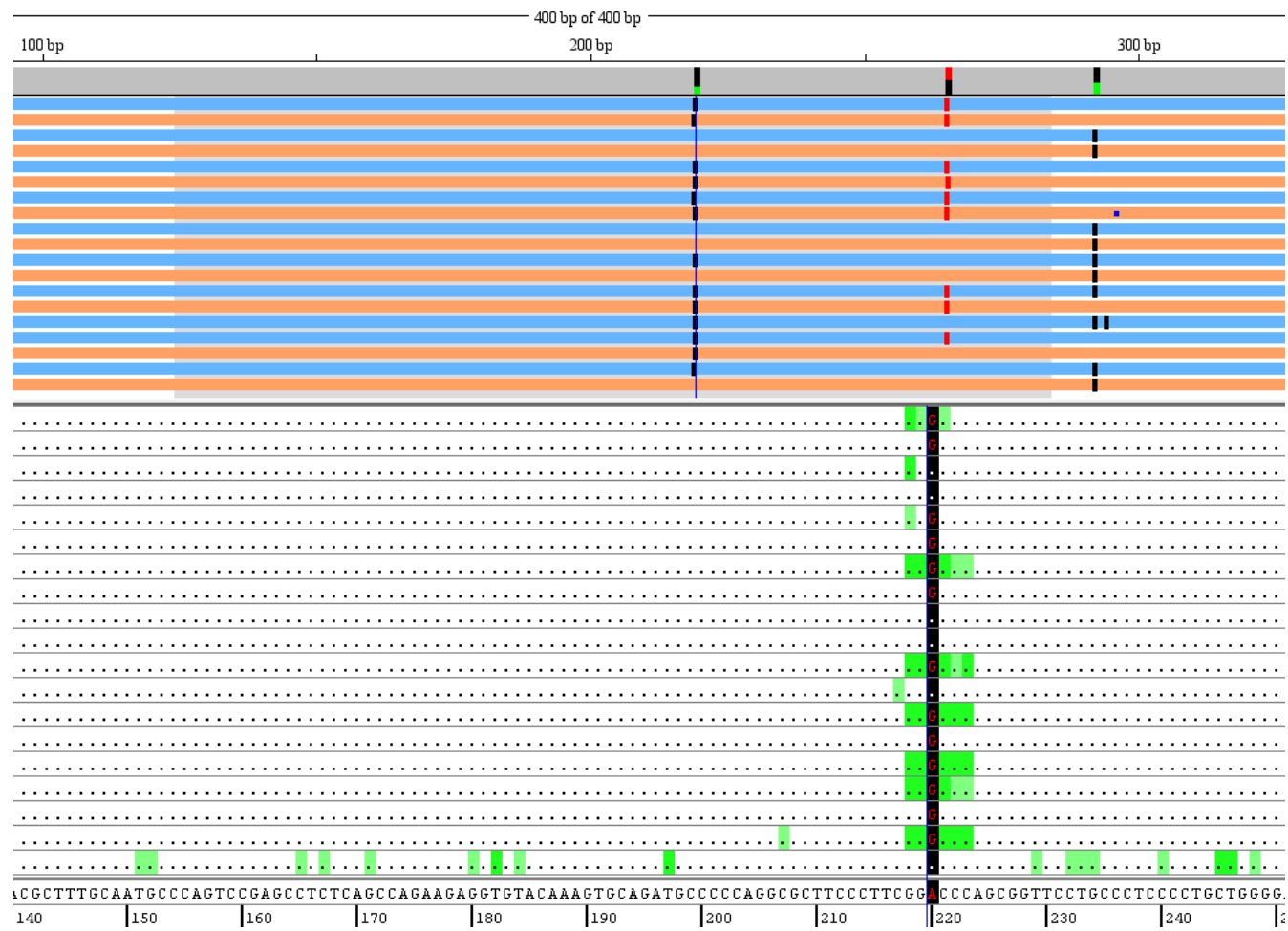
## **6. Resultados del secuenciamiento Sanger**

En el anexo 4 se detallan todas las variantes observadas en este estudio, así como el nombre oficial de los genes, el símbolo oficial, el alias empleado en este estudio (BMP4, KAP1-1 y KAP13-1) y el *scaffold* en el que se ubican.

### **6.1. Secuenciamiento del gen BMP4 en la muestra Nacional y en la muestra Rural Alianza**

En el análisis bioinformático inicial se identificaron 4 variantes genéticas. Tres de estas fueron encontradas con una alta frecuencia en las 10 mezclas (*pools*) de 3 muestras cada una (30 muestras en total). La variante restante fue observada en solo una de las mezclas. Los alelos alternativos en las 4 posiciones fueron los que se esperaban según los resultados obtenidos en la fase bioinformática (Fig. 7 y tabla 11). Por lo tanto, se consideró que ninguna de las 4 variantes genéticas eran errores de secuenciamiento y que tres de estas variantes podrían tener una frecuencia suficientemente alta como para ser incluidas en un estudio de asociación genética. Este segundo punto fue explorado con mayor detalle en el secuenciamiento de los 96 individuos de la muestra Nacional.

El análisis minucioso de esta región del gen BMP4, mostró que la variante en la posición 71 con respecto al sitio de inicio de transcripción predicho en la anotación del genoma, recae sobre un microsatélite con unidad repetitiva GGC (ver más adelante). Por lo tanto, el alelo de referencia pasaría a ser “G; y el alelo alternativo, “T”. Asimismo, el aminoácido de referencia sería R; y el alternativo, L.



**Figura 7.** Alineamiento de las secuencias del gen BMP4 obtenidas a partir de las 10 mezclas de ADN escogidas aleatoriamente de la muestra Nacional. En el panel superior se muestra el alineamiento de todas las secuencias. Las barras celestes horizontales representan las secuencias obtenidas con el primer Forward; y las barras anaranjadas, las obtenidas con el primer Reverse. Las barras verticales del panel superior representan las posiciones en las que la secuencia observada difiere de la secuencia consenso, y por lo tanto marcan los sitios polimórficos. En el panel inferior se observan los alineamientos de las secuencias. Los puntos representan las bases que NO difieren de la secuencia consenso. El “contig” mostrado en la parte inferior representa la secuencia consenso obtenida al alinear todas las secuencias. Las regiones coloreadas en verde en el panel inferior marcan las posiciones en las que la calidad de las secuencias disminuye. Esto es usualmente observado alrededor de las posiciones polimórficas.

**Tabla 11. Variantes del gen BMP4 observadas en los 3 genomas secuenciados y en las 10 mezclas (*pools*) de ADN escogidas aleatoriamente de la muestra representativa de todo el Perú**

	Identificadas en 3 genomas secuenciados(Fase bioinformática)				Observadas en 10 pools de ADN*		
<b>Alelo de referencia</b>	G	‡T	G	A	G	G	A
<b>Alelo alternativo</b>	A	‡G	C	G	A	T	G
<b>Ubicación en la secuencia codificante</b>	25	71	74	98	25	71	98
<b>Ubicación genómica</b>	61201	61201	612018	61202	61201	61201	612021
	38	84	7	11	38	84	1
<b>Codón de referencia</b>	GCC	CTG	CGG	TAC	GCC	CTG	TAC
<b>Codón alternativo</b>	ACC	CGG	CCG	TGC	ACC	CGG	TGC
<b>Aminoácido de referencia</b>	A	‡L	R	Y	A	R	Y
<b>Aminoácido alternativo</b>	T	‡R	P	C	T	L	C

<b>Posición del</b>							
<b>aminoácido</b>	9	24	25	33	9	24	33
<b>variante</b>							

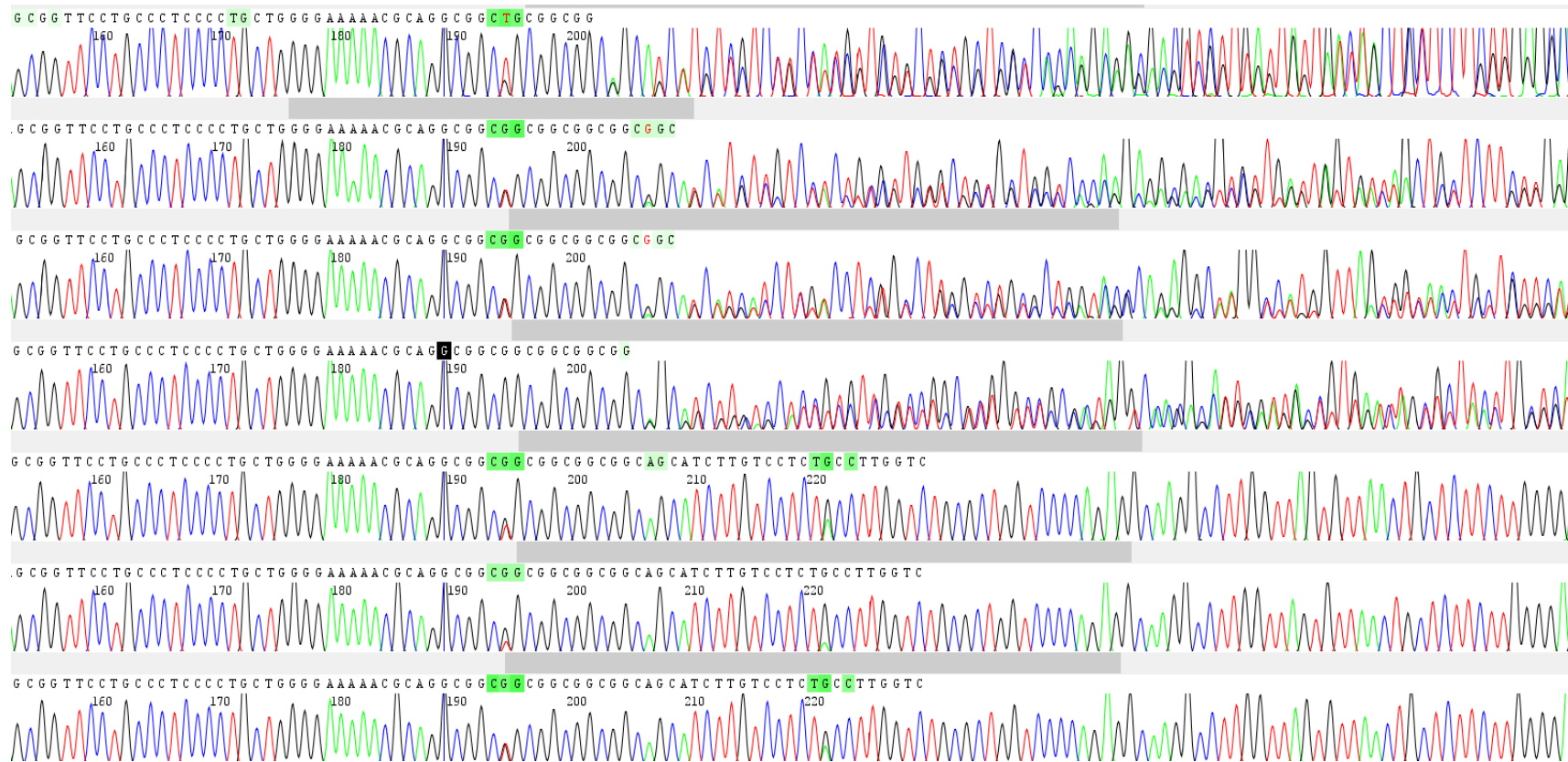
\*La variante en la posición 74 también fue encontrada en las mezclas de ADN, pero con una frecuencia aparentemente baja, lo cual se confirmó durante el análisis de una muestra representativa del Perú. ‡Luego se confirmó que la timina es el alelo de referencia, por lo tanto, el aminoácido de referencia es arginina (ver texto principal).

Luego de confirmar que el gen BMP4 fuese realmente polimórfico, se secuenciaron exitosamente 94 de las 96 muestras representativas de todo el Perú (muestra Nacional) y 147 de las 151 muestras de Rural Alianza. Durante esta fase, se detectaron en total 11 SNVs (Variantes de un único nucleótido. Este término abarca tanto a los SNPs como a las mutaciones puntuales, que conceptualmente difieren entre sí por su frecuencia poblacional) en las posiciones 6120071, 6120125, 6120138, 6120166, 6120184, 6120187, 6120190, 6120193, 6120211, 6120213, 6120215 del Scaffold17 del genoma de referencia VicPac2.02 de alpaca (N° de accesión: NW\_005882719.1). Las posiciones 6120184, 6120187, 6120190 y 6120193 corresponden al primer nucleótido de la unidad repetitiva “GGC” de un microsatélite polimórfico. La variante 6120166 tuvo una frecuencia poblacional muy baja y solo fue detectada en Rural Alianza (que presenta mayor tamaño muestral). Las variantes genéticas identificadas se listan en el anexo 4; y sus frecuencias alélicas en cada población, en el anexo 5.

Como se mencionó anteriormente, la presencia de un microsatélite en esta región afectó negativamente el secuenciamiento, ya que produjo un mosaico de secuencias que difieren —a partir de la posición del microsatélite— en unidades

completas del triplete GGC (figura 8). Este patrón observado podría deberse a que la secuencia en sí misma es polimórfica o debido al fenómeno de “*slippage*” en el que la polimerasa comete errores al sintetizar secuencias repetitivas, salteándose unidades repetitivas completas (tripletes GGC en este caso). Las muestras fallidas fueron secuenciadas hasta 3 veces con los mismos resultados, confirmándose que es realmente un microsatélite polimórfico y no un error de la polimerasa.

Debido a la presencia de este microsatélite, 12 de las 94 secuencias fueron genotipificadas parcialmente: entre uno de los extremos (posición 6120071) y el microsatélite (hasta la posición 6120190 o 6120193), cercano al cebador Reverse. Sin embargo, como se verá más adelante, esto no representó un mayor problema para el análisis de asociación debido al fuerte desequilibrio de ligamiento entre posiciones cercanas entre sí, que permite capturar la variabilidad genética de una región genómica mediante “tag-SNPs”.



**Figura 8.** Consecuencias del microsatélite GGC a nivel del secuenciamiento en el gen BMP4 en la muestra Nacional. En la parte inferior se presentan 3 secuencias correctas y sin ambigüedades. En la parte superior se muestran 3 muestras que presentan una mezcla de dos secuencias diferentes a partir de un punto cercano al microsatélite GGC, el cual empieza en la posición 178. Se

observa que la región ambigua contiene una superposición de las secuencias [ATC TTG...] y [GGC GGC...]. La secuencia [ATC TTG...] es la secuencia que se observa al terminar el microsatélite en las secuencias sin ambigüedades. Por lo tanto, resulta evidente que estas regiones ambiguas corresponden a una superposición de dos secuencias con diferente número de unidades repetitivas en el microsatélite GGC (microsatélites heterocigotos).

Además de las 4 variantes no sinónimas identificadas anteriormente, se identificaron otros 3 SNPs no sinónimos y 2 sinónimos con una frecuencia poblacional relativamente alta (tabla 12, no se muestra la SNV en 6120190 por ser idéntica a la SNV en 6120193). Uno de estos SNPs no sinónimos (100C>G) se ubica en un mismo codón que un SNP sinónimo identificado en la fase bioinformática (102T>C) y descrito en la tabla 10. El haplotipo con los alelos alternativos en estos dos SNPs (GUC) codifica el mismo aminoácido que el haplotipo que solo posee el alelo alternativo del segundo SNP (T102C). Ambos codifican valina.

El análisis de heterocigosidad, equilibrio de Hardy-Weinberg y desequilibrio de ligamiento de estos SNPs se detalla más adelante.

## **6.2. Secuenciamiento del gen KAP13-1 en la muestra Rural Alianza**

Se secuenciaron los 151 individuos seleccionados. Se obtuvieron secuencias de muy alta calidad para 147 muestras, con índices de calidad Phred score > Q30 en más del 95% de las bases. La calidad de las secuencias fue confirmada por inspección visual de los cromatogramas (Figura 15).

Luego de recortar los extremos de las secuencias, que siempre son de baja calidad, se obtuvo un fragmento de 498 pares de bases de longitud (figura 9). Se observaron cuatro polimorfismos con MAF (Frecuencia del Alelo Menor) > 0.05, y un polimorfismo con MAF < 0.05. Tres de estos polimorfismos son sustituciones no sinónimas, y uno de ellos se ubica en una región no codificante, 23pb *downstream* al codón de terminación (tabla 13).

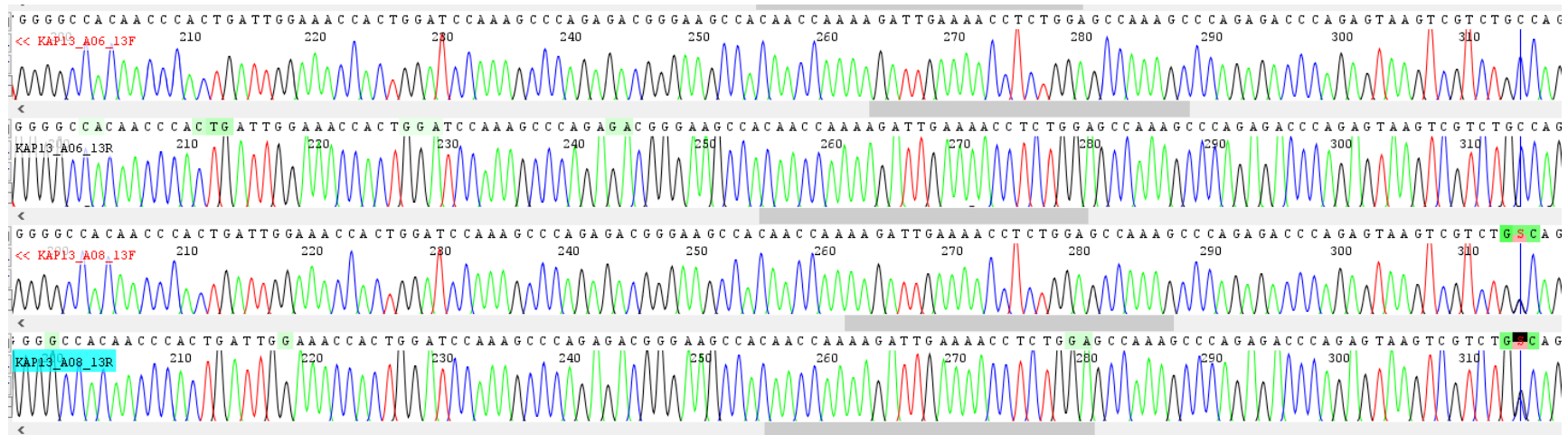
**Tabla 12.** Variantes del gen BMP4 descubiertas en el análisis poblacional (muestra Nacional)

	<b>Identificadas en 3 genomas secuenciados (Fase bioinformática)</b>				<b>Descubiertas en análisis poblacional</b>			
Alelo de referencia	G	T	G	A	C	G	C	T
Alelo alternativo	A	G	C	G	T	A	G	C
Ubicación en la secuencia codificante	25	71	74	98	12	80	100	102
Ubicación genómica	6120138	6120184	6120187	6120211	6120125	6120193	6120213	6120215
Codón de referencia	GCC	CTG	CGG	TAC	GGC	CGG	CTT	CTT
Codón alternativo	ACC	CGG	CCG	TGC	GGT	CAG	GTT	CTC
Aminoácido de referencia	A	L	R	Y	G	R	L	L
Aminoácido alternativo	T	R	P	C	G	Q	V	L
Posición del aminoácido variante	9	24	25	33	4	27	34	34

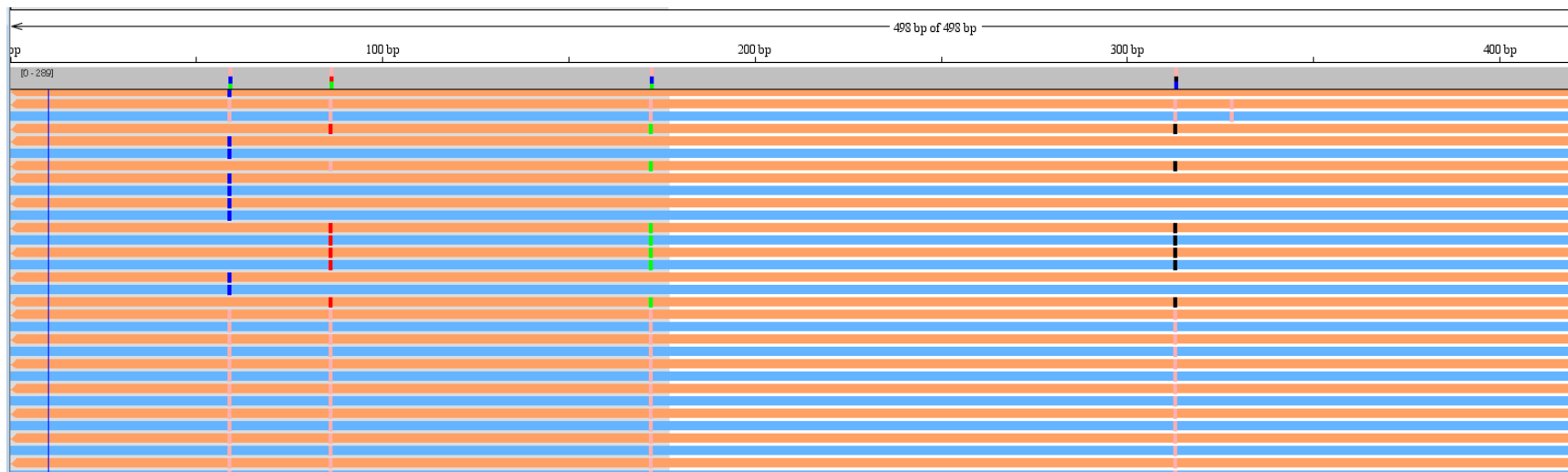
**Tabla 13.** SNPs del gen KAP 13-1 detectadas en el análisis bioinformático

	<b>SNPs observados</b>			
Alelo de referencia	G	G	T	G
Alelo alternativo	C	T	A	T
Ubicación respecto a CDS*	276	417	503	-
Ubicación respecto a TSS**	276	417	503	-
Ubicación genómica	151160	151301	151387	151414
Codón de referencia	TGG	TGG	TTC	-
Codón alternativo	TGC	TGT	TAC	-
Aminoácido de referencia	W	W	F	-
Aminoácido alternativo	C	C	Y	-
Posición del aminoácido variante	92	139	168	-

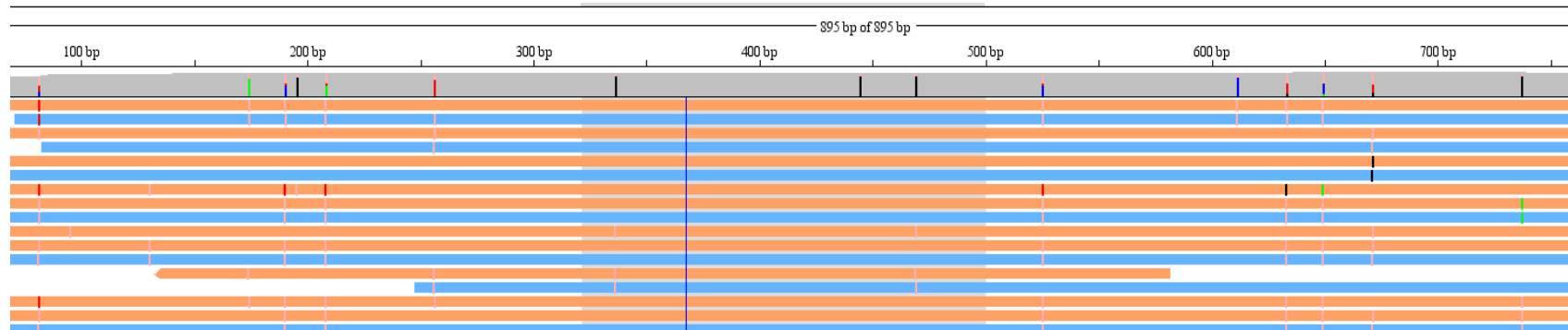
\*CDS: Secuencia que codifica la proteína (*coding sequence*). \*\*TSS: Sitio de inicio de la transcripción (*Transcription start site*). No se incluye el SNP con  $MAF < 0.05$ .



**Figura 9.** Cromatogramas de una región amplificada del gen KAP13-1 en la muestra Rural Alianza. La inspección visual confirmó que las secuencias obtenidas fueron de alta calidad. Se muestra una secuencia heterocigota (“S”, *strong*. Código IUPAC que representa ambigüedad “citosina o guanina”) en la posición 314 del alineamiento. Las terminaciones “\_13F” y “\_13R” en el nombre de las secuencias indican que estas fueron obtenidas con los cebadores *Forward* y *Reverse*, respectivamente.



**Figura 10.** Se muestra el alineamiento de un grupo de secuencias del gen KAP13-1 en la muestra Rural Alianza. Cada fila anaranjada corresponde a una secuencia obtenida con el cebador “Forward”; y cada fila celeste, a una obtenida con el cebador “Reverse”. Cada barra vertical marca las posiciones en que el genotipo observado difiere del homocigoto de referencia. Las barras rosadas corresponden a los heterocigotos; y las barras azules, rojas, verdes y negras, a los genotipos C/C, T/T, A/A y G/G, respectivamente. Las barras verticales de colores sobre la barra gris horizontal en la parte superior de la gráfica representan la frecuencia observada de cada genotipo en cada SNP.



**Figura 11.** Se muestra el alineamiento de un grupo de secuencias del gen KAP1-1 en la muestra Rural Alianza. Cada fila anaranjada corresponde a una secuencia obtenida con el cebador “Forward”; y cada fila celeste, a una obtenida con el cebador “Reverse”. Cada barra vertical marca las posiciones en que el genotipo observado difiere del homocigoto de referencia. Las barras rosadas corresponden a los heterocigotos; y las barras azules, rojas, verdes y negras, a los genotipos C/C, T/T, A/A y G/G, respectivamente. Las barras verticales de colores sobre la barra gris horizontal en la parte superior de la gráfica representan la frecuencia observada de cada genotipo en cada SNP

### **6.3. Secuenciamiento del gen KAP1-1 en la muestra Rural Alianza**

Este gen presentó 23 SNVs. En la figura 11 se presenta el alineamiento de las secuencias obtenidas.

## **7. Análisis genético-poblacional**

El análisis de genética de poblaciones tuvo dos propósitos: a) Describir la variabilidad genética del gen BMP4 a nivel nacional, debido a que, de encontrarse un SNP asociado al diámetro en BMP4, sería importante conocer si esta variante genética se presenta toda la población de alpacas del Perú. b) evaluar la posible estructuración poblacional y endogamia en la muestra Rural Alianza, ya que estos fenómenos pueden afectar negativamente el poder estadístico y la frecuencia de errores tipo I en los estudios de asociación genética, además de producir efectos que podrían dificultar la identificación de huellas de selección en las pruebas de evolución neutral.

La evaluación de presencia de endogamia (déficit de heterocigotos con respecto a lo esperado bajo equilibrio de Hardy-Weinberg) y de estructuración poblacional fue realizada empleando 10 microsatélites no ligados que se asume son neutrales y representan las características demográficas de la población. El análisis de deficiencia de heterocigotos fue repetido con los genes candidatos para evaluar si existía deficiencia de heterocigotos en algún locus en particular. Bajo equilibrio neutral se espera que todos los loci del genoma presenten aproximadamente los mismos valores de Fis. Finalmente, se realizaron pruebas de equilibrio entre deriva génica y mutación bajo evolución neutral en los tres genes candidatos empleados

en el análisis de asociación. Los análisis de estructuración poblacional y de equilibrio neutral solo se realizaron en la muestra Rural Alianza.

### 7.1. Análisis de endogamia y estructuración poblacional en la muestra Rural Alianza empleando diez microsatélites

Cuatro microsatélites mostraron déficit de heterocigotos (tabla 14). El índice Fis para los 10 microsatélites considerados en conjunto fue 0.034.

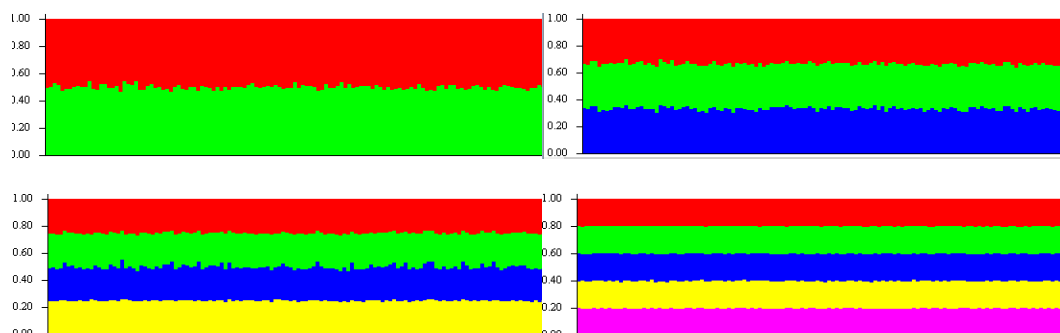
Se empleó el software STRUCTURE para evaluar la hipótesis de que las muestras provienen de  $K = 2, 3, 4$  o  $5$  poblaciones latentes. Para cada escenario, todos los individuos mostraron el mismo grado de ancestría respecto a cada población latente putativa (Figura 12). Por lo tanto, estos 10 microsatélites no mostraron evidencia de estructuración poblacional.

**Tabla 14.** Análisis de heterocigocidad de 10 microsatélites

Locus	Alelos	Heterocigosidad observada	Heterocigosidad esperada	HWE* ( <i>p-value</i> )	Fis‡ (W & C)
LCA19	14	0.662	0.676	0.4062	0.02
LCA37	16	0.799	0.850	0.0003	0.0608
LCA5	10	0.786	0.763	0.7915	-0.0303
LCA66	16	0.799	0.856	0.0252	0.0666
LCA8	11	0.857	0.853	0.4175	-0.005
YWLL29	9	0.649	0.728	0.0176	0.1082
YWLL36	12	0.87	0.872	0.3066	0.0023
YWLL40	7	0.623	0.665	0.1611	0.0633

YWLL44	14	0.851	0.896	0.0575	0.0503
YWLL8	25	0.922	0.937	0.1638	0.0164

\*HWE: Equilibrio de Hardy Weinberg. ‡W&C: Método de Weir y Cockerham (75).



**Figura 12.** Análisis de estructuración poblacional. Se muestran los resultados de STRUCTURE en 4 escenarios distintos: desde  $k = 2$  hasta  $k = 5$  poblaciones latentes. Cada columna representa un individuo (nótese que no se muestran las separaciones entre cada individuo). Todos los individuos tuvieron virtualmente la misma probabilidad de ser asignados a cada población. Por lo tanto, se concluye que todos los individuos provienen de una sola población genéticamente homogénea.

## 7.2. Descripción de la variabilidad genética de BMP4, KAP1-1 y KAP13-1

El gen KAP1 solo fue secuenciado en la muestra Rural Alianza y presentó 23 sitios segregantes, que comprenden tanto SNPs (frecuencia poblacional  $>1\%$ ) como mutaciones puntuales. Los 23 sitios fueron bialélicos; sin embargo, en 10 de ellos no se observó uno de los genotipos debido a la baja frecuencia poblacional. El gen BMP4 presentó 11 sitios segregantes, uno de los cuales presentó una frecuencia poblacional muy baja y fue detectado en la muestra Rural Alianza ( $N = 147$ ), pero no en la muestra Nacional ( $N = 94$ ), probablemente debido al menor tamaño muestral de esta. Finalmente, el gen KAP13-1, que solo fue secuenciado en Rural

Alianza, presentó 5 sitios segregantes. Las frecuencias alélicas de todos los sitios segregantes se detallan en el anexo 5.

El gen BMP4 no presentó deficiencia de heterocigotos en la muestra Nacional (Fis = -0.0318), ni en Rural Alianza (Fis = 0.0003). Además, las frecuencias alélicas y genotípicas de BMP4 fueron muy similares en ambas muestras (anexos 5 y 6), por lo tanto, los análisis posteriores solo se realizaron en la muestra Rural Alianza. El gen KAP1-1 presentó un moderado exceso de heterocigotos (Fis = -0.0595, muestra Rural Alianza). En contraste, el gen KAP13-1 mostró una marcada y significativa deficiencia de heterocigotos (Fis = 0.2182, muestra Rural Alianza). Los valores del índice Fis así como la significancia de las diferencias entre los valores de heterocigocidad esperada y observada de cada SNP se detallan en el anexo 6.

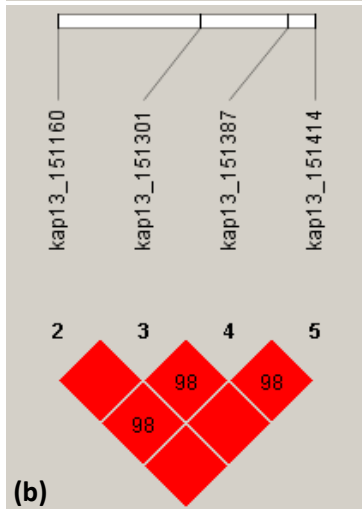
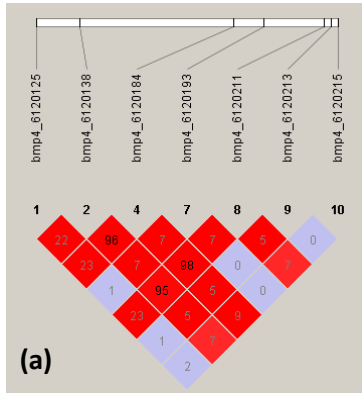
Se empleó el software PHASE v.2 para reconstruir las fases haplotípicas de los tres genes candidatos en la muestra Rural Alianza. La longitud de los fragmentos empleados para este análisis fue 177pb, 763pb, y 498pb para los genes BMP4, KAP1-1 y KAP13-1, respectivamente.

Se identificaron 12 haplotipos en el gen BMP4, 15 en el gen KAP1-1, y 4 en el gen KAP13-1. Este último presentó dos haplotipos comunes (54.8% y 42.5%) y dos con frecuencias muy bajas (2.4% y 0.3%).

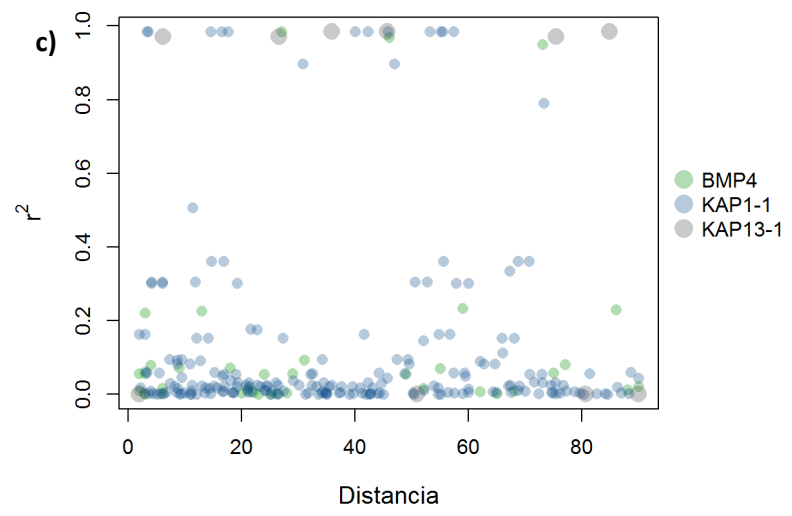
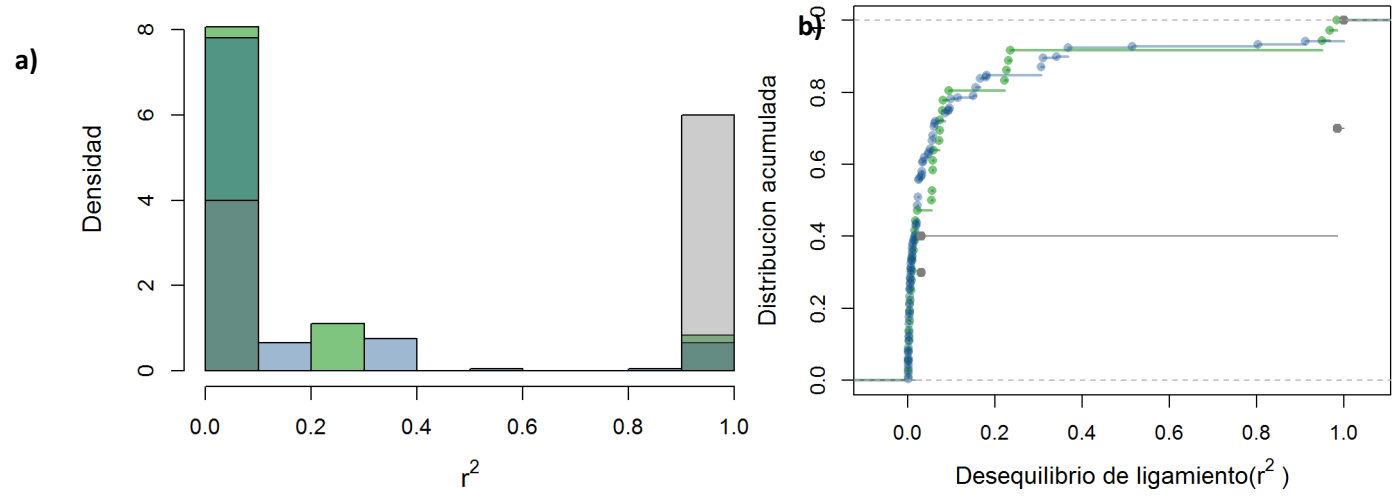
Los genes BMP4 y KAP1-1 presentaron 3y 4 pares de sitios segregantes con 4 tipos gaméticos, respectivamente. El número mínimo estimado de eventos de recombinación fue 2 en ambos genes. En contraste, el gen KAP13-1 no presentó

ningún par de sitios con 4 tipos gaméticos, y el mínimo número de eventos de recombinación fue cero.

Los perfiles de desequilibrio de ligamiento de los genes BMP4 y KAP1-1 fueron similares, pero difirieron del perfil de KAP13-1. El valor del estadístico  $|D'|$  fue 1 o muy cerca de 1 para todos los pares de sitios segregantes en los tres genes. El coeficiente de correlación al cuadrado ( $r^2$ ) presentó mayor variabilidad que el estadístico  $|D'|$ , especialmente en los genes BMP4 y KAP1-1 (figura 13). Este estadístico presentó una distribución bimodal en el gen KAP13-1, con valores de  $r$  entre 0.99 y 1 para todos los pares de sitios segregantes que no incluyen el SNP 151145 (figura 14), y valores de  $r^2$  igual a 0.03 entre este SNP y todos los demás. Este fenómeno se debió a la baja frecuencia de su alelo menor (2.4%). La prueba de Kolmogorov-Smirnov no detectó diferencias entre la distribución del desequilibrio de ligamiento (medido mediante el coeficiente de correlación al cuadrado) de BMP4 y KAP1-1 ( $p\text{-value} = 0.333$ ). En contraste, sí hubo diferencias entre la distribución de BMP4 y KAP13-1 ( $p\text{-value} = 0.007$ ) y entre KAP1-1 y KAP13-1 ( $p\text{-value} = 0.005$ ). Finalmente, no existió correlación entre la distancia y el desequilibrio de ligamiento en ninguno de los genes (tabla 15 y figura 14).



**Figura 13.** Desequilibrio de ligamiento de los SNPs con  $MAF > 0.05$  en los genes a) BMP4, b) KAP1-1 y c) KAP13-1 en las muestra Rural Alianza (el perfil de LD del gen BMP4 en la muestra Nacional fue indistinguible del observado en la muestra Nacional). Los valores ubicados en la intersección de cada par de SNP representan el valor de desequilibrio de ligamiento ( $R^2 \times 100$ ). Los colores se encuentran en una escala de colores desde blanco ( $|D'| = 0$ ) hasta rojo brillante ( $|D'| = 1$  y  $\log of the likelihood odds ratio [LOD] \geq 2$ ). El color lila representa valores de  $|D'| = 1$  pero con un valor de  $LOD < 2$ . El color blanco representa valores de  $|D'|$  menores a 1 y  $LOD < 2$ . Solo se muestran los SNPs con  $MAF > 0.05$ . El valor de LOD es un indicador del nivel de significancia, a mayor valor, mayor significancia.



**Figura 14.** Desequilibrio de ligamiento en los tres genes candidatos. a) Histograma de la densidad del coeficiente de correlación al cuadrado. b) Distribución acumulada empírica. c) Coeficiente de correlación al cuadrado versus la distancia medida en pares de bases.

**Tabla 15.** Correlación entre el desequilibrio de ligamiento y la distancia

Gen	Coeficiente de Correlación de Spearman	p-value
BMP4	0.157	0.360
KAP1-1	-0.047	0.501
KAP13-1	-0.031	0.931

El desequilibrio de ligamiento se midió mediante coeficiente de correlación ( $r^2$ ). La distancia se midió en pares de bases

## 8. Pruebas de equilibrio evolutivo neutral

Los estimadores  $\theta_\pi$  y  $\theta_s$  tuvieron valores similares en los genes BMP4 y KAP1-1, lo que concuerda con lo esperado bajo equilibrio neutral. En cambio,  $\theta_\pi$  fue 4 veces mayor que el estimador  $\theta_s$  para el gen KAP13-1 (Tabla 16).

**Tabla 16.** Valores de los estimadores de  $\theta$  por sitio ( $4N_e\mu$ )

Estimador	BMP4	KAP1-1	KAP13-1
$\theta_s$	0.0093	0.0047	0.0016
$\theta_\pi$	0.0129	0.0063	0.0041

Los genes BMP4 y KAP1-1 no mostraron evidencia de exceso o déficit de alelos raros mediante las pruebas D de Tajima o  $F_s$  de Fu (Tabla 18), lo que sugiere ausencia de expansión poblacional, selección negativa (*background selection*) o de barridos selectivos fuertes. En contraste, el gen KAP13-1 sí mostró evidencia de déficit de alelos raros mediante la prueba D de Tajima ( $p$ -value= 0.006), lo que sugiere un cuello de botella en curso, selección balanceante o barrido selectivo débil. La prueba de Ewens-Watterson (89) que evalúa la hipótesis alternativa de selección balanceante no fue significativa para ninguno de los genes (Tabla 17).

La distribución del número de diferencias entre pares de secuencias (MMD, *mismatch distribution*) fue multimodal para los tres genes analizados (figura 15); por lo tanto, no hay evidencia de expansión poblacional ni de selección negativa o barridos selectivos fuertes. Los mismos resultados fueron obtenidos con la prueba R2 de Ramos-Onsins y Rozas (tabla 17). Finalmente, la prueba ZnS de Kelly que

compara el perfil de desequilibrio de ligamiento observado con el perfil esperado bajo la hipótesis nula de equilibrio neutral, sugiere que KAP13-1 ha pasado recientemente por un barrido selectivo débil (tabla 17).

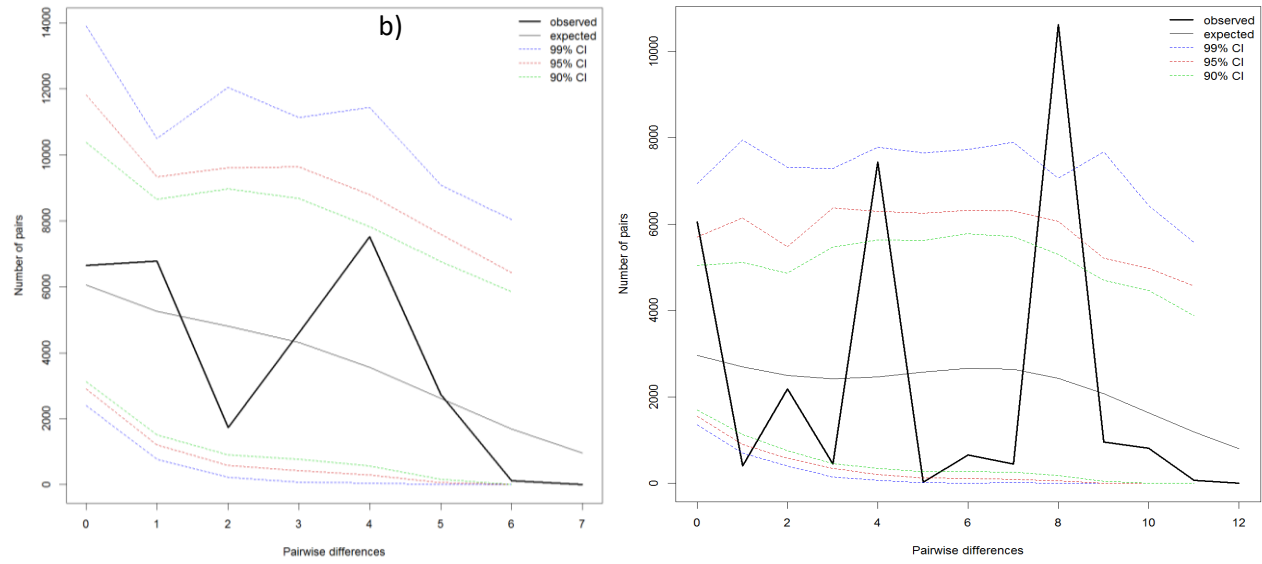
**Tabla 17.** Resultado de las pruebas de equilibrio entre deriva y mutación bajo evolución neutral

Test	Gen	Estadístico*	P-value‡
D de Tajima	BMP4	0.873	0.1287
	KAP1-1	0.894	0.102
	KAP13-1	2.822	0.006
Fs de Fu	BMP4	-0.234	0.891
	KAP1-1	1.834	0.99
	KAP13-1	6.325	0.995
Test de homocigidad de Ewens-Watterson	BMP4	0.224/0.276	0.358
	KAP1-1	0.204/0.22	0.532
	KAP13-1	0.481/0.627	0.287
R2 de Ramos-Onsins y Rozas	BMP4	0.114	0.86
	KAP1-1	0.110	0.872
	KAP13-1	0.204	0.998
Zns de Kelly	BMP4	0.106	0.137
	KAP1-1	0.116	0.069
	KAP13-1	0.608	0.007

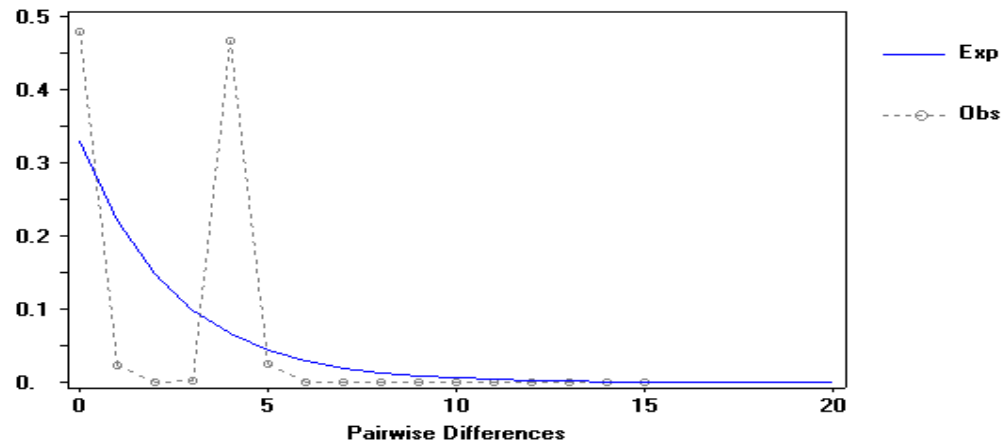
(\*)La prueba de Ewens-Watterson compara el estadístico F esperado y el observado. (‡) El p-value de las pruebas D de Tajima y ZnS de Kelly es

(Pr(estadístico simulado) > estadístico observado). En las tres pruebas restantes, el *p-value* es (Pr(estadístico simulado) < estadístico observado). El *p-value* de todas las pruebas se obtuvo simulando genealogías bajo la hipótesis nula condicionadas en el número de sitios segregantes observados en cada gen. Se obtuvieron valores muy similares en simulaciones condicionadas en el valor de  $\theta_\pi$  como estimador del parámetro  $4N_e\mu$ .

a)



c)



**Figura 15.** Distribución del número de diferencias entre pares de secuencias (*MismatchDistribution*). a) BMP4, b) KAP1-1, c)KAP13-1. Los intervalos de confianza fueron obtenidos con el software Arlequin 3.5, y corresponden a la hipótesis nula de ausencia de expansión poblacional. El algoritmo implementado en Arlequin fue incapaz de obtener los intervalos de confianza para el gen KAP13-1 debido a la baja variabilidad que este presenta.

## 9. Estudio de asociación genética

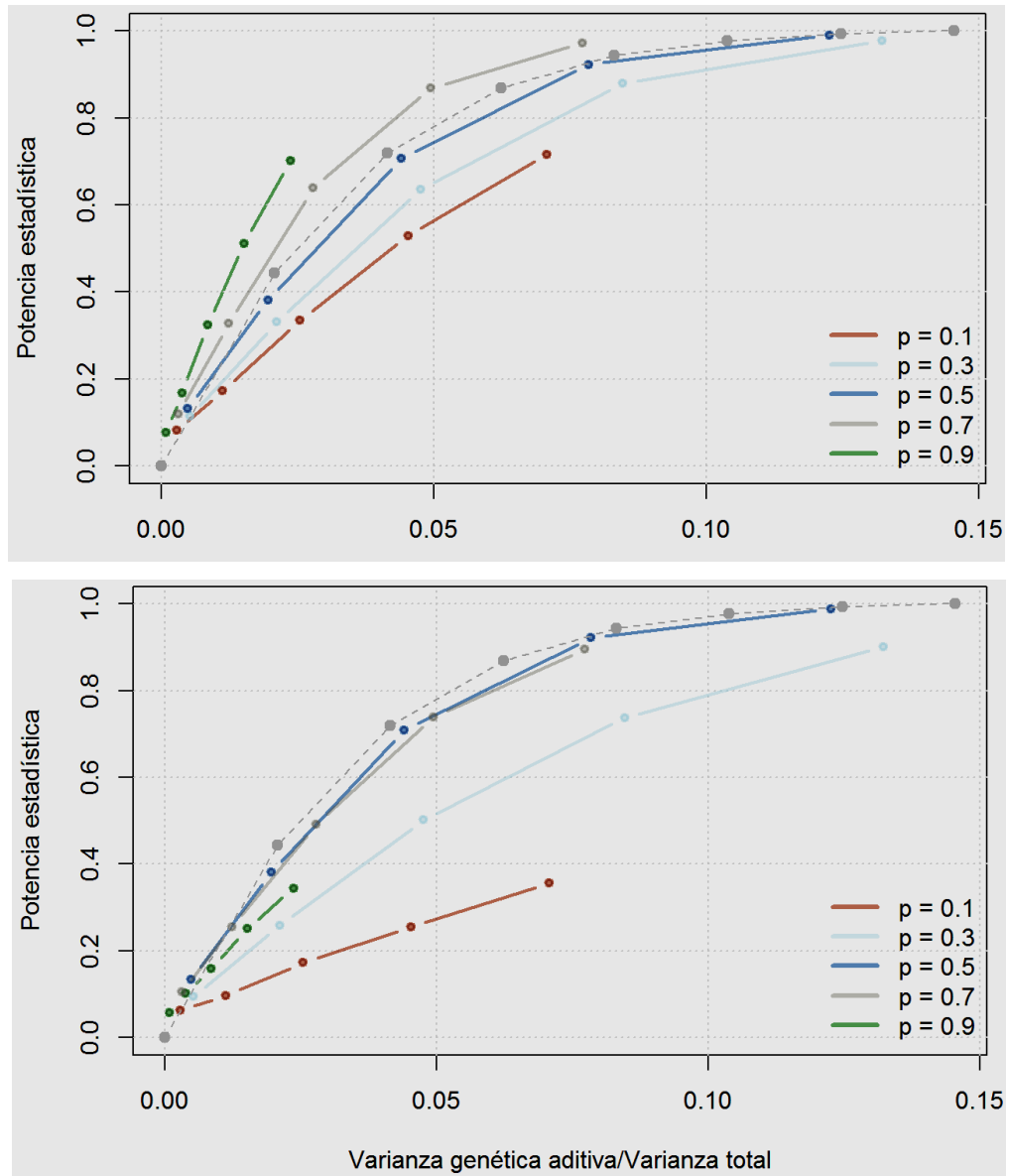
### 9.1. Análisis de potencia estadística

La potencia estadística se estimó mediante un método analítico y mediante simulaciones. El método analítico arroja resultados exactos (si se cumplen los supuestos del modelo), pero solo es aplicable para la situación poco realista en que el locus candidato tiene efectos puramente aditivos sobre el fenotipo. El método analítico se detalla en el anexo 2. En la simulación se consideró que el locus ejerce efectos de dominancia, con  $d = a/3$ , y se evaluó la potencia estadística cuando el modelo de asociación empleado incluye o excluye los efectos de dominancia.

En general, el modelo netamente aditivo alcanza mayor poder estadístico que el modelo que incluye dominancia. Las simulaciones mostraron que si el locus presenta dominancia, la potencia estadística no es simétrica con respecto a las frecuencias alélicas. Es decir, no se alcanza la misma potencia para frecuencias alélicas  $p$  y  $1-p$ . El modelo de asociación puramente aditivo alcanza mayor potencia cuando mayor sea la frecuencia del alelo cuyo efecto tiene el mismo signo que el efecto de dominancia. Por ejemplo, para un SNP con alelos  $B$  y  $b$ , y frecuencias  $f(B) = p$ ;  $f(b) = q$ ; y valores genotípicos medio “ $-a$ ”, “ $d$ ”, “ $a$ ” para los genotipos  $bb$ ,  $Bb$ ,  $BB$ , respectivamente, es posible detectar valores de “ $a$ ” más pequeños si los valores “ $a$ ” y “ $d$ ” tienen el mismo sentido y el efecto “ $a$ ” corresponde al homocigoto más frecuente (comparar las curvas para  $p=0.2$  y  $p=0.8$  en la figura 16). En un escenario pesimista, con  $f(B)=p=0.1$  (es decir, el efecto de dominancia tiene signo opuesto al valor medio del homocigoto más frecuente), se

alcanza una potencia estadística de 70% para un SNP que explica aproximadamente el 7% de la varianza genética aditiva.

La potencia estadística del modelo que incluye dominancia incompleta y efectos aditivos muestra una mayor dependencia de las frecuencias alélicas, alcanzando un máximo valor para frecuencias alélicas intermedias (figura 15). Este modelo también alcanza mayor potencia cuando el efecto de dominancia tiene el mismo sentido que el valor medio del homocigoto más frecuente.



**Figura 16.** Gráficas de poder estadístico alcanzado según la proporción de la varianza fenotípica explicada por el efecto los efectos genéticos aditivos. Arriba se muestran los resultados de un modelo en el que solo se incorporan los efectos genéticos aditivos en la regresión lineal. Abajo se muestran los resultados al incluir los efectos de dominancia (además de los efectos aditivos). La línea punteada en ambas gráficas representa la estimación mediante un método

analítico (detallado en el anexo 2) que asume que el rasgo cuantitativo es influenciado solo por efectos genéticos aditivos

## **9.2. Análisis de asociación de los genes BMP4, KAP1-1 y KAP13-1**

Se seleccionaron 5 tag-SNPs en el gen BMP4 y 9 en el gen KAP1-1. Las demás variantes fueron excluidas del análisis de asociación debido a que portaban escasa o nula información adicional. Por ejemplo, en la figura 17 se muestra la distribución del diámetro de la fibra según cada genotipo de los 4 SNPs de KAP13-1 con MAF (frecuencia del alelo menor)  $> 0.05$ . No se observan diferencias entre los genotipos en ninguno de los SNPs. Asimismo, el diámetro de la fibra se distribuye de manera idéntica en los 4 SNPs debido al fuerte desequilibrio de ligamiento que genera redundancia en la variabilidad genética que captura cada uno de estos.

El análisis de asociación se realizó de 3 modos diferentes. Primero se realizó el análisis empleando solo un gen a la vez, además del sexo. Los resultados se encuentran en la tabla 18. Luego, se repitió el análisis pero esta vez considerando los efectos de dominancia (anexo 7, tabla 25). Finalmente, se consideró un modelo completo con todos los tag-SNPs simultáneamente y se aplicó un análisis de varianza sobre el modelo global para seleccionar las variables significativas del modelo(90). El análisis de varianza compara el modelo completo con el modelo nulo, y evalúa si la reducción en la suma del cuadrado de los errores es significativa. Los resultados se detallan en la tabla 26 del anexo 7.

No se encontró asociación entre el diámetro de la fibra de alpaca y ninguno de los tag-SNP de los genes BMP4, KAP1-1 ni KAP13-1 ( $p\text{-value}>0.05$  para todos

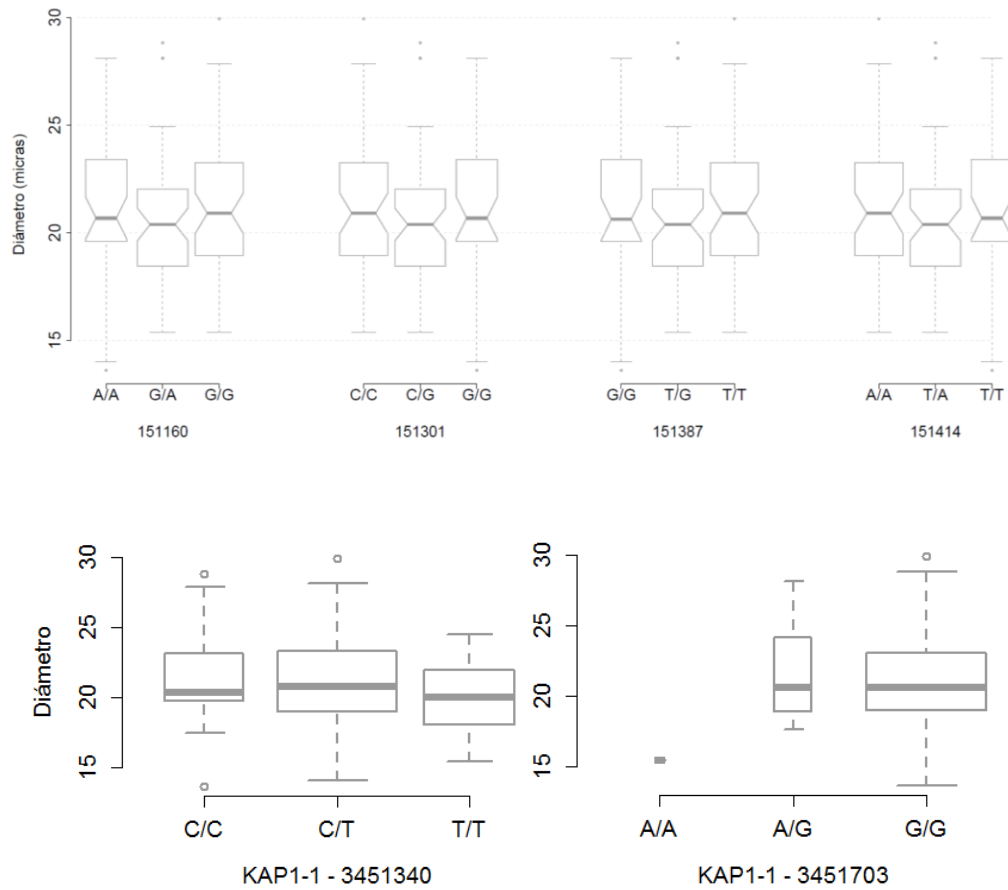
los SNPs) en ninguno de los modelos. El SNP 3451703 en el gen KAP1-1 fue el más próximo a alcanzar la significancia estadística. Sin embargo, este efecto se debió a un único valor extremo en uno de los genotipos (figura 17).

**Tabla 18. Resultados del análisis de asociación – solo efectos aditivos**

<b>Gen</b>	<b>SNP</b>	<b>Variable</b>	<b>Coef.</b>	<b>Error Est.</b>	<b>p-val</b>	<b>R<sup>2</sup></b>
<b>BMP4</b>		(Intercepto)	20.66	0.319	0.000	
	6120125	sexo	0.65	0.845	0.440	0.020
		aditivo	0.69	0.479	0.154	
		(Intercepto)	21.13	0.477	0.000	
	6120138	sexo	0.75	0.847	0.379	0.008
		aditivo	-0.21	0.367	0.560	
		(Intercepto)	21.26	0.291	0.000	
	6120213	sexo	-0.01	0.835	0.987	0.028
		aditivo	-1.28	0.681	0.062	
		(Intercepto)	20.78	1.189	0.000	
	6120215	sexo	0.41	0.923	0.655	0.002
		aditivo	0.18	0.641	0.779	
<b>KAP1-1</b>		(Intercepto)	21.69	0.495	0.000	
	3451340	sexo	0.84	0.877	0.339	0.033
		aditivo	-0.74	0.386	0.056	
		(Intercepto)	21.15	0.376	0.000	
	3451449	sexo	0.73	0.872	0.404	0.010
		aditivo	-0.39	0.435	0.369	

**Tabla 18. Resultados del análisis de asociación – solo efectos aditivos**

<b>Gen</b>	<b>SNP</b>	<b>Variable</b>	<b>Coef.</b>	<b>Error Est.</b>	<b>p-val</b>	<b>R<sup>2</sup></b>
		(Intercepto)	22.35	1.358	0.000	
	3451454	sexo	0.71	0.868	0.415	0.013
		aditivo	-0.77	0.718	0.285	
		(Intercepto)	19.67	0.907	0.000	
	3451515	sexo	0.60	0.863	0.486	0.020
		aditivo	0.76	0.522	0.150	
		(Intercepto)	20.53	1.561	0.000	
	3451703	sexo	0.66	0.870	0.451	0.005
		aditivo	0.21	0.808	0.800	
		(Intercepto)	20.55	1.040	0.000	
	3451870	sexo	0.67	0.872	0.441	0.005
		aditivo	0.21	0.569	0.707	
		(Intercepto)	20.92	0.414	0.000	
<b>KAP13-1</b>	151301	sexo	0.79	0.851	0.357	0.007
		aditivo	-0.04	0.332	0.893	



**Figura 17.** Panel superior: gráficos de cajas del diámetro observado según cada genotipo en los 4 SNPs del gen KAP13-1 con  $MAF > 0.05$ . Panel inferior: Gráfica de cajas de los dos SNPs que estuvieron más cerca de alcanzar la significancia estadística.

## VI. DISCUSIÓN

Este estudio tuvo como objetivo identificar genes que influyen rasgos productivos de la fibra de alpaca. Se emplearon varias pruebas de evolución neutral para evaluar si el gen regulador (BMP4) o los dos genes estructurales (KAP1-1 y KAP13-1) analizados presentaban huellas de selección en una población de alpacas destinada a la producción de fibra. Asimismo, se realizó en la misma población un análisis de asociación entre el diámetro de la fibra (una de las características morfológicas de la fibra de alpaca con mayor importancia económica) y las variantes genéticas en los tres genes mencionados.

No se encontró una diferencia significativa entre el diámetro promedio de la fibra de alpacas hembras y machos en la muestra Rural Alianza, población en la que se realizaron las pruebas de equilibrio evolutivo neutral y el análisis de asociación genética (no se pudo incluir la muestra Nacional por no contarse con información fenotípica de ella). Ello difiere de reportes anteriores en los que la fibra de alpacas hembras presentó un diámetro mayor por  $0.95\mu\text{m}$ (91). La carencia de significación estadística en este estudio podría deberse al escaso número de alpacas macho en la muestra empleada (15 individuos), lo que no habría permitido alcanzar una diferencia significativa.

El diámetro de la fibra no presentó distribución normal en la muestra completa (Shapiro-Wilk,  $p\text{-value} = 0.043$ ), pero sí en los machos y hembras por separado. La calidad de la fibra fue muy alta, con una media de  $20.95\mu\text{m}$  y una desviación estándar de  $3.04\mu\text{m}$ . La proporción de *Baby* (entre  $19.5$  y  $22.5 \mu\text{m}$ ) y *Royal*( $<19.5 \mu\text{m}$ ) alpaca fue de 39 y 32%, respectivamente. La alta calidad de la

fibra sugiere que esta población de alpacas se encuentra bajo selección artificial para el diámetro o que fue fundada a partir de una población con una fibra de alta calidad. Cabe destacar que el diámetro de la fibra no es la única característica que presenta evidencia de selección artificial: todas las muestras de alpaca fueron de color blanco y de raza huacaya, lo que implica intervención humana para mantener la población con estas características homogéneas.

En un esfuerzo por seleccionar y evaluar un gen regulatorio que pudiese jugar un papel importante en la determinación de los atributos de la fibra en alpaca, se identificaron y anotaron las variantes de 22 genes que regulan el desarrollo del pelo en mamíferos, explorando para ello el genoma completo de tres alpacas secuenciadas para un proyecto anterior. Se escogió el gen BMP4 debido a que fue el que presentó mayor cantidad de variantes genéticas que podrían tener un importante impacto sobre la función de la estructura y función de las proteínas (tabla 10).

Una variante sinónima, y las cuatro no sinónimas del gen BMP4 identificadas en la fase bioinformática, fueron confirmadas durante el secuenciamiento de una muestra de alpacas representativa del Perú (tabla 12). Por lo tanto, se confirma la alta calidad de las variantes identificadas en el secuenciamiento de genoma completo.

El análisis de 10 microsatélites en la población Rural Alianza encontró una ligera deficiencia de heterocigotos ( $F_{is} = 0.034$ ). Sin embargo, el análisis de STRUCTURE no mostró evidencia de estructuración poblacional (figura 12). Ello concuerda con un estudio anterior que no encontró estructuración poblacional

dentro de siete granjas de alpaca de Junín, Huancavelica y Puno, aunque sí encontró diferenciación entre la población de alpacas de Huancavelica y las otras regiones (92). Por lo tanto, se concluye que este ligero déficit de heterocigocidad en la muestra Rural Alianza se debe a la consanguinidad, posiblemente vinculada con la selección artificial a la que esta población estaría sujeta o al limitado tamaño poblacional. Cabe destacar que la selección reduce la variabilidad genética tanto a nivel global (es decir, en todo el genoma) como a nivel local (es decir, en el locus bajo selección y las regiones circundantes). Sin embargo, el efecto local es mucho más fuerte que el efecto global(40).

El análisis del gen KAP13-1 mostró un gran déficit de heterocigocidad para todos los SNPs (muestra Rural Alianza), con desequilibrio de Hardy-Weinberg altamente significativo ( $p\text{-value} < 0.01$ ), y con índices Fis entre 0.21 a 0.22 (anexo 6, tabla 24). Se descarta que el déficit de heterocigocidad de KAP13-1 se deba exclusivamente a la endogamia o a la estructuración poblacional debido a que los 10 microsatélites y los genes BMP4 y KAP1-1 presentaron valores bajos de Fis (anexo 6, tabla 24). Además, el análisis de los 10 microsatélites con STRUCTURE no encontró grupos latentes genéticamente diferenciados (figura 12). Cabe destacar que estos 10 microsatélites han permitido identificar 3 grupos de alpaca genéticamente diferenciados, dos de los cuales provienen de la misma región (Huancavelica) (92).

Los genes BMP4 y KAP1-1 presentaron un perfil de desequilibrio de ligamiento muy similar, pero diferente del de KAP13-1 (figuras 13 y 14). Este gen presentó un elevado desequilibrio de ligamiento, con valores de  $r^2$  entre 0.98 y 1

para todos los pares de sitios segregantes que excluyen al SNP 151145. Este SNP presentó una frecuencia alélica muy baja (2.4%) y un menor desequilibrio de ligamiento con los demás SNPs, lo que sugiere que esta mutación es mucho más reciente(93).

Valores de Fis tan altos como los observados en el gen KAP13-1 son muy inusuales en poblaciones naturales panmícticas, pero podrían observarse en un gen asociado a un rasgo que influencia el *fitness* en una población sujeta a una fuerte selección (94). La selección también podría explicar el elevado desequilibrio de ligamiento en el gen KAP13-1(60). Se aplicaron varias pruebas de neutralidad para evaluar la hipótesis de que los bajos niveles de diversidad genética en el gen KAP13-1 estarían causados por selección (tabla 17). Estas pruebas también fueron aplicadas a los genes BMP4 y KAP1-1, ya que fenómenos demográficos pueden dejar huellas genéticas similares a las generadas por la selección(37).

La distribución de diferencias entre pares de secuencias (MMD, *Mismatch Distribution*) del gen BMP4 fue bimodal y estuvo dentro del intervalo de confianza al 95% de los valores esperados bajo equilibrio neutral (figura 14); lo que indica que no hay evidencia de expansión poblacional, selección negativa ni barridos selectivos fuertes(95). La MMD del gen KAP1-1 también fue bimodal; sin embargo, la segunda moda excedió el intervalo de confianza al 99% de los valores esperados bajo equilibrio neutral. No obstante, ello no es indicativo de selección negativa o positiva ni de expansión poblacional, ya que bajo esas condiciones se hubiera esperado una distribución unimodal. KAP13-1 también mostró una distribución bimodal, siendo 0 la primera moda (figura 15). Ello se debe a que

debido a la escasa diversidad de este gen, muchos pares de haplotipos no presentan ninguna diferencia. La distribución bimodal de los tres genes sugiere que no ha habido una expansión poblacional reciente; además, ninguno de estos se encontraría bajo selección negativa ni habría pasado por un barrido selectivo fuerte(95).

La prueba R2 de Ramos-Onsins y Rozas no fue significativa para ninguno de los genes candidatos (tabla 18). Esta prueba también emplea el número de diferencias entre pares de secuencias, además del número de sitios segregantes y de *singletons* (mutaciones únicas), para evaluar si el gen se encuentra en equilibrio neutral (44). La prueba R2 corrobora los resultados del análisis MMD.

La prueba Fs de Fu que evalúa la distribución de las frecuencias alélicas no mostró evidencia de expansión poblacional, selección negativa ni barrido selectivo fuerte en ninguno de los genes. La prueba D de Tajima no fue significativa en los genes BMP4 ni KAP1-1 (tabla 17), lo que descarta expansión poblacional o selección en estos genes (42).

El estadístico D de Tajima del gen KAP13-1 fue muy alto y significativo ( $D = 2.822$ ), lo que sugiere déficit de alelos raros debido a reducción del tamaño poblacional o selección balanceante (37,42). La prueba D de Tajima también puede ser significativa debido a la ocurrencia de barridos selectivos débiles. En ese escenario, el valor esperado del estadístico D de Tajima es aproximadamente cero; sin embargo, su varianza incrementa mucho. Por ello, valores muy altos de D no descartan barridos selectivos débiles (57). En presencia de barridos selectivos fuertes, el estadístico D de Tajima tiende a adoptar valores muy negativos y significativos.

El gen KAP13-1 presentó un valor de ZnS más alto que el esperado bajo equilibrio neutral lo que sugiere la acción de un barrido selectivo débil reciente en o muy cerca de este gen (tabla 17) (57). Esta prueba está basada en el perfil de desequilibrio de ligamiento esperado bajo equilibrio neutral, y se diferencia de las demás por su capacidad de detectar barridos selectivos débiles (57,60). Esta prueba pierde potencia rápidamente tras finalizar el barrido débil debido a que el desequilibrio de ligamiento es destruido rápidamente por la recombinación; lo que sugiere que el barrido selectivo aún está en marcha o habría terminado durante las últimas  $0.05N_e$  generaciones, aproximadamente, ya que caso contrario sería difícilmente detectable (57).

Cuando existe un barrido selectivo fuerte, se espera que el estadístico ZnS tome valores cercanos a cero (60). Ello se debe a que el desequilibrio de ligamiento, medido por el coeficiente de correlación al cuadrado  $r^2$ , disminuye debido a que los alelos raros no logran alcanzar altos valores de  $r^2$ . El mismo comportamiento se observa cuando existe selección negativa o expansión poblacional bajo neutralidad (61). En una población de tamaño constante, una región genómica que evoluciona neutralmente pero con eventos de recombinación también mostrará valores bajos de ZnS debido que el desequilibrio de ligamiento es destruido por acción de la recombinación (61).

En este estudio se empleó el mínimo número de eventos de recombinación observados para estimar el parámetro poblacional de recombinación ( $4N_e c$ ) (62). Ello implica que este parámetro puede haber sido subestimado, pero es muy improbable que haya sido sobreestimado.

La recombinación tiene efectos opuestos sobre las pruebas de equilibrio neutral dependiendo de si la hipótesis nula es verdadera o falsa (56). Cuando la hipótesis nula es verdadera y el locus presenta recombinación, los estadísticos R2, Fs, D y ZnS toman valores muy bajos y las pruebas correspondientes son *liberales* (es decir, incrementa la frecuencia de errores estadísticos tipo I), favoreciendo la hipótesis alternativa de expansión poblacional, selección negativa, o barrido selectivo fuerte. En contraste, cuando ha ocurrido una expansión poblacional, las pruebas R2, ZnS y D pierden poder y favorecen la hipótesis nula. La prueba Fs tiene un comportamiento opuesto y se vuelve ligeramente *liberal*.

Las pruebas Fs y R2 fueron no significativas para los tres genes. Por lo tanto, la posible subestimación de la recombinación no indujo a rechazar incorrectamente la hipótesis nula y, al contrario, estas pruebas podrían haberse comportando de manera *conservadora*(61).

Los estadísticos ZnS y D presentaron valores positivos muy altos, en el gen KAP13-1, en contraposición a lo esperado bajo selección negativa, barridos fuertes o expansión poblacional(42,60). Asimismo, estos resultados no son explicados por una posible subestimación del parámetro de recombinación y respaldan la hipótesis de que KAP13-1 ha sido afectado por un barrido selectivo débil o por un barrido fuerte acompañado de recombinación (57,61,96).

El perfil de diversidad genética de KAP13-1 no corresponde al esperado bajo selección balanceante. Los valores elevados del estadístico D de Tajima y ZnS de Kelly, además de las frecuencias alélicas intermedias, concuerdan con una selección balanceante. Sin embargo, bajo este escenario se esperaría observar un

exceso de heterocigocidad, en contraposición a lo observado en KAP13-1 (39). Los resultados de la prueba de homocigocidad de Ewens-Watterson tampoco respaldan la hipótesis alternativa de selección balanceante en ninguno de los genes(89).

El apareamiento *selectivo* (“*assortative mating*”) puede generar un déficit de heterocigotos e incrementar el desequilibrio de ligamiento. Este fenómeno implica que los individuos eligen parejas reproductivas con fenotipos similares al de sí mismos (35). Sin embargo, KAP13-1 se encuentra rodeado de un clúster de genes que codifican queratinas y otras proteínas asociadas a queratinas. No existe evidencia de que alguno de estos genes determine alguna característica fenotípica relacionada a la elección de compañeros sexuales. Por lo tanto, resulta inverosímil que este fenómeno sea el responsable de la escasa diversidad genética en KAP13-1.

También se descarta que la baja diversidad genética en KAP13-1 se deba a una reducción del tamaño poblacional. Los eventos demográficos dejan señales a lo largo de todo el genoma(39). Por lo tanto, si hubiera ocurrido una reducción poblacional, los genes BMP4 y KAP1-1 también se habrían visto afectados.

Un barrido selectivo débil en KAP13-1 o en una región cercana es la explicación más probable para las características del gen KAP13-1. Este fenómeno explica satisfactoriamente el escaso número de haplotipos y sitios segregantes, las frecuencias alélicas intermedias, el elevado desequilibrio de ligamiento y el déficit de heterocigotos (57,96,97).

El análisis de potencia estadística mostró que el tamaño muestral empleado en el estudio de asociación genética permitió alcanzar un poder por encima de 0.8 para detectar SNPs que expliquen más del 10% de la varianza fenotípica (figura 16). Anteriormente se ha postulado que existiría un gen aún no conocido que explicaría hasta el 57% de la varianza del diámetro de la fibra de alpaca (10). El poder estadístico alcanzado en este estudio sería suficiente para detectar un gen con un efecto tan grande si este fuese o estuviera en LD con alguno de los SNPs evaluados en el análisis de asociación.

Se empleó un modelo de asociación que ha sido utilizado para identificar SNPs asociados a 8 rasgos cuantitativos en una población humana coreana (98), y para identificar marcadores asociados a los niveles de expresión de alrededor de más de 1000 genes en población humana europea (99). Además, se encuentra implementado en el software PLINK desarrollado para realizar estudios de asociación genética (86) y ha sido recomendado para estudios de asociación en individuos no emparentados en ausencia de estructuración poblacional (64).

El análisis de los genes KAP1-1, BMP4 y KAP13-1 no mostró asociación estadística con el diámetro de la fibra ( $p\text{-value} > 0.5$ ). Dado el poder estadístico alcanzado en este estudio, se puede concluir que es muy improbable que los polimorfismos con la mayor contribución sobre el diámetro de la fibra de alpaca se encuentren en la proximidad de alguno de los SNPs analizados. Cabe mencionar que un estudio anterior tampoco encontró asociación entre el diámetro de la fibra de alpaca y otra proteína asociada a Keratina (KAP11-1)(24), resaltando la

dificultad que suelen presentar los estudios de genes candidatos para encontrar asociaciones genotipo-fenotipo.

Es improbable que el barrido selectivo en KAP13-1 se deba a la selección del diámetro de la fibra por dos razones: a) Si KAP13-1 tuviese un efecto grande sobre el diámetro, el análisis de asociación hubiese resultado significativo (ver gráfica de potencia estadística, fig. 16); b) Si KAP13-1 tuviese un efecto pequeño sobre el diámetro de la fibra, no se hubiera detectado ninguna huella de selección, ya que la presión de selección no deja huellas de selección convencionales en genes con efectos pequeños. Por ejemplo, un estudio evaluó si existe coincidencia entre las regiones del genoma con evidencia de selección y las regiones del genoma que contienen polimorfismos que afectan 7 rasgos cuantitativos poligénicos que están sujetos a una fuerte selección en vacas lecheras (contenido de grasa, proteína, etc., en leche obtenida de vacas Holstein y Jersey) (100). De las 35 comparaciones realizadas (5 contrastes diferentes por 7 rasgos cuantitativos) solo dos fueron estadísticamente significativos. En cambio, sí se encontró evidencia de selección en genes que determinan rasgos simples (monogénicos u oligogénicos) empleados para definir las diferentes razas de ganado vacuno, tales como el gen POLLED, que determina la presencia o ausencia de cuernos, y 4 genes que determinan el color y el patrón de la pigmentación. También se detectó huellas de selección en genes *mayores* (genes con un efecto grande sobre un rasgo poligénico), aunque estas huellas de selección no fueron tan evidentes como en el caso de los rasgos monogénicos(100). Resultados similares se obtuvieron en un estudio independiente también en ganado bovino (101).

Es posible que KAP13-1 o un gen muy cercano afecte un rasgo monogénico u oligogénico, o que tenga un gran efecto sobre algún rasgo poligénico (el cual no sería el diámetro de la fibra). El tipo de fibra (suri/huacaya) es la única característica morfológica de la fibra de alpaca que se conoce sigue una herencia monogénica u oligogénica (7). Se ha postulado que existen dos genes que determinan el tipo de fibra, y uno de ellos se encontraría en un clúster de proteínas asociadas a keratina (8). El gen KAP13-1 se encuentra cerca del extremo del scaffold101 y está rodeado de varias proteínas parálogas (*Keratin-associated protein 13-1 like*) y otras proteínas asociadas a keratina. Es posible que KAP13-1 se encuentre cerca de o sea uno de los genes que determinan el fenotipo suri/Huacaya.

## **VII. CONCLUSIÓN**

Existen al menos 5 polimorfismos genéticos que podrían afectar significativamente la estructura y función del gen regulador BMP4.

La población de alpacas de la EPS Rural Alianza, de donde provienen las muestras empleadas en el análisis de asociación, presentan bajos niveles de endogamia y ausencia de estratificación poblacional.

Se encontró evidencia de a) selección positiva a través de un barrido selectivo en o cerca del gen KAP13-1. Este gen u algún otro muy cercano podría influenciar alguna característica morfológica de la fibra con herencia monogénica u oligogénica sujeta a selección. Los genes BMP4 y KAP1-1 no presentaron huellas de selección.

No se encontró una asociación entre el diámetro de la fibra y las variantes de los genes KAP1-1, KAP13-1 ni BMP4.

## **VIII. RECOMENDACIONES**

Se recomienda realizar un análisis de asociación genética entre el gen KAP13-1 y el tipo de fibra de alpaca.

Se recomienda sustituir el enfoque de genes candidatos por enfoques genómicos, tales como chips de genotipificación, genotipificación por secuenciamiento, secuenciamiento de exoma completo o secuenciamiento de genoma completo para acelerar la identificación de los genes que influyen los rasgos productivos de alpaca.

## IX. BIBLIOGRAFÍA

1. Organización de las naciones unidas para el desarrollo industrial subdivisión de promoción de inversión y tecnología. Producción textil de fibras de camélidos sudamericanos en el área altoandina de Bolivia, Ecuador y Perú [Internet]. 2006. Available from: [http://www.unido.org/fileadmin/import/58563\\_camelidos\\_final.pdf](http://www.unido.org/fileadmin/import/58563_camelidos_final.pdf)
2. McGregor BA. Production, attributes and relative value of alpaca fleeces in southern Australia and implications for industry development. *Small Rumin Res.* 2006 Feb 1;61(2):93–111.
3. Frank EN, Hick MVH, Gauna CD, Lamas HE, Renieri C, Antonini M. Phenotypic and genetic description of fibre traits in South American domestic camelids (llamas and alpacas). *Small Rumin Res.* 2006 Feb;61(2–3):113–29.
4. Instituto Nacional de Estadística e Informática. IV Censo Nacional Agropecuario [Internet]. 2012 [cited 2017 Feb 4]. Available from: <http://censos.inei.gob.pe/cenagro/tabulados/>
5. Instituto Nacional de Estadística e Informática. Perú: Anuario de Estadísticas ambientales 2014 [Internet]. Lima, Perú: Instituto Nacional de Estadística e Informática; 2015 Enero [cited 2017 Feb 24]. Available from: [https://www.inei.gob.pe/media/MenuRecursivo/publicaciones\\_digitales/Est/Lib1197/caratula.htm](https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1197/caratula.htm)
6. Ponzoni RW, Hubbard DJ, Kenyon RV, Tuckwell CD, McGregor BA, Howse A, et al. Phenotypes resulting from Huacaya by Huacaya, Suri by Huacaya and Suri by Suri alpaca crossings. In: IAI 1997: Shaping the future: Proceedings of the International Alpaca Industry 1997 seminar [Internet]. Australian Alpaca Association; 1997 [cited 2017 May 24]. p. 11–13. Available from: <http://dro.deakin.edu.au/eserv/DU:30066046/mcgregor-phenotypes-1997.pdf>
7. Presciuttini S, Valbonesi A, Apaza N, Antonini M, Huanca T, Renieri C. Fleece variation in alpaca (*Vicugna pacos*): a two-locus model for the Suri/Huacaya phenotype. *BMC Genet.* 2010 Jul 20;11:70.
8. Wong YT. P1007 Genetic and protein study of alpaca fiber. *J Anim Sci.* 2016 Sep 1;94(supplement4):18–18.
9. Wuliji null, Davis null, Dodds null, Turner null, Andrews null, Bruce null. Production performance, repeatability and heritability estimates for live weight, fleece weight and fiber characteristics of alpacas in New Zealand. *Small Rumin Res J Int Goat Assoc.* 2000 Aug 1;37(3):189–201.

10. Pérez-Cabal MA, Cervantes I, Morante R, Burgos A, Goyache F, Gutiérrez JP. Analysis of the existence of major genes affecting alpaca fiber traits. *J Anim Sci*. 2010 Dec;88(12):3783–8.
11. Ponzoni RW, Grimson RJ, Hill JA, Hubbard DJ, McGregor BA, Howse A, et al. The inheritance of and associations among some production traits in young Australian alpacas. <http://www.aaabg.org/livestocklibrary/1999/AB99111.pdf>. 1999;13:468–71.
12. Cervantes I, Pérez-Cabal MA, Morante R, Burgos A, Salgado C, Nieto B, et al. Genetic parameters and relationships between fibre and type traits in two breeds of Peruvian alpacas. *Small Rumin Res*. 2010 Jan;88(1):6–11.
13. Falconer PDS, Mackay PTFC. *Introduction to Quantitative Genetics*. 4 edition. Longman; 1995. 480 p.
14. Wuliji T, Davis GH, Dodds KG, Turner PR, Andrews RN, Bruce GD. Production performance, repeatability and heritability estimates for live weight, fleece weight and fiber characteristics of alpacas in New Zealand. *Small Rumin Res*. 2000 Aug 1;37(3):189–201.
15. Paredes MM, Membrillo A, Gutiérrez JP, Cervantes I, Azor PJ, Morante R, et al. Association of microsatellite markers with fiber diameter trait in Peruvian alpacas (*Vicugna pacos*). *Livest Sci*. 2014 Mar;161:6–16.
16. Paredes MM, Membrillo A, Azor PJ, Machaca JE, Torres D, Serrano AM. Genetic and phenotypic variation in five populations of Huacaya Alpacas (*Vicugna pacos*) from Peru. *Small Rumin Res*. 2013 Abril;111(1–3):31–40.
17. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008 Jun;9(6):477–85.
18. McLaren RJ, Rogers GR, Davies KP, Maddox JF, Montgomery GW. Linkage mapping of wool keratin and keratin-associated protein genes in sheep. *Mamm Genome*. 1997 Dec;8(12):938–40.
19. Parsons YM, Cooper DW, Piper LR. Evidence of linkage between high-glycine-tyrosine keratin gene loci and wool fibre diameter in a Merino half-sib family. *Anim Genet*. 1994 Apr 1;25(2):105–8.
20. Bidinost F, Roldan DL, Dodero AM, Cano EM, Taddeo HR, Mueller JP, et al. Wool quantitative trait loci in Merino sheep. *Small Rumin Res*. 2008 Jan;74(1–3):113–8.
21. Zhou H, Gong H, Li S, Luo Y, Hickford JGH. A 57-bp deletion in the ovine KAP6-1 gene affects wool fibre diameter. *J Anim Breed Genet Z Tierzucht Zuchtungsbiologie*. 2015 Aug;132(4):301–7.

22. Wang Z, Zhang H, Yang H, Wang S, Rong E, Pei W, et al. Genome-Wide Association Study for Wool Production Traits in a Chinese Merino Sheep Population. *PLOS ONE*. 2014 Sep 30;9(9):e107101.
23. Cheng Y, Lotan R. Molecular Cloning and Characterization of a Novel Retinoic Acid-inducible Gene That Encodes a Putative G Protein-coupled Receptor. *J Biol Chem*. 1998 Dec 25;273(52):35008–15.
24. Foppiano Florez Estrada FA. Caracterización de marcadores genéticos en genes que codifican a proteínas asociadas a queratina y evaluación de la asociación del gen KRTAP11-1 al diámetro de fibra en alpaca (*Vicugna pacos*) siguiendo una aproximación de gen candidato. 2016 [cited 2017 May 30]; Available from: <http://190.116.48.45/handle/upch/127>
25. Delgado de la Flor Montauban IRM. Caracterización de nuevos marcadores genéticos Microsatélites e identificación de SNP en el gen de Tricohialina en alpacas (*Vicugna pacos*). 2014 [cited 2017 May 25]; Available from: <http://repositorio.concytec.gob.pe/handle/CONCYTEC/74>
26. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, Zhu G, et al. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet*. 2009 Nov;85(5):750–5.
27. Millar SE. Molecular mechanisms regulating hair follicle development. *J Invest Dermatol*. 2002 Feb;118(2):216–25.
28. Schlake T. FGF signals specifically regulate the structure of hair shaft medulla via IGF-binding protein 5. *DevCamb Engl*. 2005 Jul;132(13):2981–90.
29. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, et al. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet*. 2008 Mar 15;17(6):835–43.
30. Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, et al. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *HumGenet*. 2008 Sep;124(2):179–85.
31. Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, et al. FGFR2 is associated with hair thickness in Asian populations. *J Hum Genet*. 2009 Jul 10;54(8):461–5.
32. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *CurrBiol CB*. 2010 Feb 23;20(4):R208-215.

33. Bouaziz M, Ambroise C, Guedj M. Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. *PLoS ONE*. 2011 Diciembre;6(12):e28845.
34. Weir BS, Hill WG. Estimating F-statistics. *Annu Rev Genet*. 2002;36:721–50.
35. Templeton AR. *Population Genetics and Microevolutionary Theory*. 1 edition. Hoboken, N.J: Wiley-Liss; 2006. 705 p.
36. Kohn MH, Pelz H-J, Wayne RK. Natural selection mapping of the warfarin-resistance gene. *ProcNatlAcadSci U S A*. 2000 Jul 5;97(14):7911–5.
37. Biswas S, Akey JM. Genomic insights into positive selection. *Trends Genet TIG*. 2006 Aug;22(8):437–46.
38. Buri P. Gene Frequency in Small Populations of Mutant *Drosophila*. *Evolution*. 1956;10(4):367–402.
39. Hamilton M. *Population Genetics*. 1 edition. Chichester, UK ; Hoboken, NJ: Wiley-Blackwell; 2009. 424 p.
40. Walsh B, Lynch M. *Evolution and Selection of Quantitative Traits: I. Foundations*. [Internet]. 1st ed. Unpublished manuscript [cited 2017 Mar 4]. Available from: [nitro.biosci.arizona.edu/zbook/NewVolume\\_2/newvol2.html](http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html)
41. Kimura M, Crow JF. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics*. 1964 Apr;49(4):725–38.
42. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989 Nov 1;123(3):585–95.
43. Fu Y-X. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics*. 1997 Oct 1;147(2):915–25.
44. Ramos-Onsins SE, Rozas J. Statistical Properties of New Neutrality Tests Against Population Growth. *MolBiolEvol*. 2000 Dec 1;19(12):2092–100.
45. Watterson GA. On the number of segregating sites in genetical models without recombination. *TheorPopul Biol*. 1975 Apr;7(2):256–76.
46. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *ProcNatlAcad Sci*. 1979 Oct 1;76(10):5269–73.
47. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 2007 Aug;8(8):610–8.
48. Nurminsky DI. *Selective Sweep*. 1 edition. Nurminsky DI, editor. Springer US; 2007. 121 p.

49. Kreitman M, Hudson RR. Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*. 1991 Mar;127(3):565–82.
50. Santiago E, Caballero A. Effective size of populations under selection. *Genetics*. 1995 Feb 1;139(2):1013–30.
51. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*. 1993 Aug;134(4):1289–303.
52. Maynard J, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res*. 2007 Dec;89(5–6):391–403.
53. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*. 2005 Apr;169(4):2335–52.
54. Robertson A. Inbreeding in artificial selection programmes. *Genet Res*. 1961;2(02):189–194.
55. Barton NH. Linkage and the limits to natural selection. *Genetics*. 1995 Jun 1;140(2):821–41.
56. Maruyama T, Birky CW. Effects of periodic selection on gene diversity in organelle genomes and other systems without recombination. *Genetics*. 1991 Feb 1;127(2):449–51.
57. Pennings PS, Hermisson J. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLOS Genet*. 2006 Dec 15;2(12):e186.
58. Ewens WJ. The sampling theory of selectively neutral alleles. *TheorPopul Biol*. 1972 Mar 1;3(1):87–112.
59. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *TheorAppl Genet*. 1968;38(6):226–231.
60. Kelly JK. A test of neutrality based on interlocus associations. *Genetics*. 1997 Jul;146(3):1197–206.
61. Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. Statistical Power Analysis of Neutrality Tests Under Demographic Expansions, Contractions and Bottlenecks With Recombination. *Genetics*. 2008 May;179(1):555–67.
62. Hudson RR. Estimating the recombination parameter of a finite population model without selection. *Genet Res*. 1987 Dec;50(3):245–50.

63. Fisher RA. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. 1918 [cited 2015 Sep 22]; Available from: <https://digital.library.adelaide.edu.au/dspace/handle/2440/15097>
64. Neale BM, Ferreira MAR, Medland SE, Posthuma D, editors. *Statistical Genetics: Gene Mapping Through Linkage and Association*. 1 edition. New York: Taylor & Francis; 2008. 608 p.
65. Fisher RA. Average excess and average effect of a gene substitution. *Ann Eugen*. 1941;11(1):53–63.
66. Gregorius HR. The probability of losing an allele when diploid genotypes are sampled. *Biometrics*. 1980 Dec;36(4):643–52.
67. Morgan M, Pages H, Obenchain V. Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import. R Package. 2010;
68. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*. 2014 Jul 15;30(14):2076–8.
69. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*. 2009 Jul 15;25(14):1841–2.
70. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*. 2013 Aug 8;9(8):e1003118.
71. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
72. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000 Jun;155(2):945–59.
73. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 2003 Aug;164(4):1567–87.
74. Agapito J, Rodríguez J, Herrera-Velit P, Timoteo O, Rojas P, Boettcher PJ, et al. Parentage testing in alpacas (*Vicugna pacos*) using semi-automated fluorescent multiplex PCRs with 10 microsatellite markers. *Anim Genet*. 2008 Apr;39(2):201–3.
75. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. 1984;38(6):1358–70.

76. Raymond M, Rousset F. GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *J Hered.* 1995 May 1;86(3):248–9.
77. Kalinowski ST, Taper ML, Marshall TC. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol Ecol.* 2007 Mar 1;16(5):1099–106.
78. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009;25(11):1451–1452.
79. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005;21(2):263–265.
80. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001 Apr;68(4):978–89.
81. Wu H, Guang X, Al-Fageeh MB, Cao J, Pan S, Zhou H, et al. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun.* 2014 Oct 21;5:5188.
82. Noble WS. How does multiple testing correction work? *Nat Biotechnol.* 2009 Dec;27(12):1135–7.
83. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: Designing rare variant association studies. *Proc Natl Acad Sci.* 2014 Jan 28;111(4):E455–64.
84. Barrett JC. Haploview: Visualization and Analysis of SNP Genotype Data. *Cold Spring Harb Protoc.* 2009 Oct 1;2009(10):pdb.ip71.
85. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007 Sep;81(3):559–75.
86. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. 2014.
87. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet.* 2014 May;15(5):335–46.
88. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 2010;10(3):564–567.
89. Watterson GA. The Homozygosity Test of Neutrality. *Genetics.* 1978 Feb 1;88(2):405–17.

90. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 1st ed. 2013, Corr. 6th printing 2016 edition. New York: Springer; 2013. 426 p.
91. Morante R, Goyache F, Burgos A, Cervantes I, Pérez-Cabal MA, Gutiérrez JP. Genetic improvement for alpaca fibre production in the Peruvian Altiplano: the Pacamarca experience. *AnimGenetResourGénétiquesAnim Génétiques Anim*. 2009 Oct;45:37–43.
92. Rodriguez J, Flor ID de la, Barreto T, Timoteo O, Herrera-Velit P, Espinoza JR, et al. Genetic diversity in alpaca (*Vicugna pacos*) populations using 10 microsatellite markers. In 2009 [cited 2017 May 23]. Available from: [http://inis.iaea.org/Search/search.aspx?orig\\_q=RN:41046519](http://inis.iaea.org/Search/search.aspx?orig_q=RN:41046519)
93. Kimura M, Ohta T. The age of a neutral mutant persisting in a finite population. *Genetics*. 1973 Sep;75(1):199–212.
94. Kimura M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics*. 1969 Apr;61(4):893–903.
95. Rogers AR, Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. *MolBiolEvol*. 1992 May;9(3):552–69.
96. Messer PW, Neher RA. Estimating the strength of selective sweeps from deep population diversity data. *Genetics*. 2012 Jun;191(2):593–605.
97. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLOS Genet*. 2015 Feb 23;11(2):e1005004.
98. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban H-J, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet*. 2009 May;41(5):527–34.
99. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*. 2005 Oct 27;437(7063):1365–9.
100. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics*. 2014;15:246.
101. Druet T, Pérez-Pardal L, Charlier C, Gautier M. Identification of large selective sweeps associated with major genes in cattle. *Anim Genet*. 2013 Dec 1;44(6):758–62.

102. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 2006 Feb;38(2):203–8.
103. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet.* 2011 Jul;19(7):807–12.

## ANEXOS

### Anexo 1. Derivación de las ecuaciones de la varianza genética

Sea “G” los efectos genéticos de un locus sobre un determinado rasgo cuantitativo.

Entonces:

$$G = A + D$$

Donde “A” es el efecto aditivo, y “D” el efecto de dominancia.

La varianza de la suma de dos variables aleatorias X, Y es:

$$\sigma_{(X+Y)}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y}$$

Por lo tanto, la varianza de G está definida por:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + 2\sigma_{A,D}$$

Por definición, no existe correlación entre los componentes aditivos y los no aditivos, por lo tanto, la covarianza es cero y la varianza de G se reduce a la siguiente expresión:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

... Ecuación a

La varianza de una variable aleatoria X con media  $\mu$  es:

$$\sigma_X^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Por lo tanto, se puede calcular la varianza de los efectos genéticos ponderando la diferencia entre cada valor genotípico medio y la media poblacional por las frecuencias genotípicas esperadas bajo equilibrio de Hardy-Weinberg:

$$\sigma_G^2 = p^2(a - \mu)^2 + 2pq(d - \mu)^2 + q^2(-a - \mu)^2$$

$$\sigma_G^2 = 2pq[a + d(q - p)]^2 + (2pqd)^2$$

$$\sigma_G^2 = 2pq\alpha^2 + (2pqd)^2$$

... Ecuación b

Además, por definición:

$$A_{ij} = \alpha_i + \alpha_j = \textit{breeding value}$$

Donde “A<sub>ij</sub>” es la suma de los efectos aditivos de los alelos “i”, “j” de un locus. No existe correlación entre los efectos alélicos  $\alpha_i$ ,  $\alpha_j$ . Por lo tanto, la varianza de los efectos genéticos aditivos de un genotipo A<sub>i</sub>A<sub>j</sub> equivale a la varianza de su *breeding value*:

$$\sigma_{A_{ij}}^2 = \sigma_{\alpha_i}^2 + \sigma_{\alpha_j}^2$$

Notar que la media poblacional de los *breeding values* es cero bajo equilibrio de Hardy-Weinberg:

$$\mu(\textit{breeding value}) = p^2(2q\alpha) + 2pq(q - p)\alpha + q^2(-2p\alpha)$$

$$\mu(\textit{breeding value}) = 2pq\alpha[p + (q - p) - q]$$

$$\mu(\textit{breeding value}) = 0$$

La varianza poblacional de los efectos genéticos aditivos se puede obtener ponderando la varianza de cada *breeding value* por su frecuencia poblacional esperada bajo equilibrio de Hardy-Weinberg (64). Por lo tanto, la varianza genética aditiva es:

$$\sigma_A^2 = p^2(2q\alpha - 0)^2 + 2pq[(q - p)\alpha - 0]^2 + q^2(-2p\alpha - 0)^2$$

$$\sigma_A^2 = 2pq\alpha^2$$

...Ecuación c

La varianza de los efectos de dominancia se halla reemplazando las ecuaciones b y c en la ecuación a:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

$$\sigma_D^2 = \sigma_G^2 - \sigma_A^2$$

$$\sigma_D^2 = 2pq\alpha^2 + (2pqd)^2 - 2pq\alpha^2$$

$$\sigma_D^2 = (2pqd)^2$$

## **Anexo 2. Cálculo de potencia estadística mediante método analítico**

A continuación se presenta una demostración de un método comúnmente empleado para calcular la potencia estadística de un estudio de asociación genética para rasgos cuantitativos en individuos no emparentados (102).

Sea  $Y_i$  el fenotipo de un individuo “i” influenciado por un locus B con dos alelos  $B_1$  y  $B_2$ . La desviación de  $Y_i$  con respecto a la media poblacional puede atribuirse a los efectos genéticos aditivos (A) y los efectos residuales (E), que incluyen dominancia, epistasis, efectos ambientales de todo tipo. Por lo tanto:

$$Y = \mu + A + E$$

Ecuación 1

Además,

$$A_i = x_{i1}\alpha_1 + x_{i2}\alpha_2$$

$$A = x_1\alpha_1 + x_2\alpha_2$$

Ecuación 2

Donde  $x_1, x_2$  representan el número de copias de los alelos  $B_1$  y  $B_2$ , respectivamente, que porta un individuo cualquiera, de tal modo que  $x_1 + x_2 = 2$  (especie diploide).

Reemplazando la ecuación 2 en 1:

$$Y_i = \mu + x_1\alpha_1 + x_2\alpha_2 + E$$

$$Y_i = \mu + x_1\alpha_1 + (2 - x_1)\alpha_2 + E$$

$$Y_i = \mu + 2\alpha_2 + x_1\alpha_1 - x_1\alpha_2 + E$$

$$Y_i = \mu + 2\alpha_2 + x_1(\alpha_1 - \alpha_2) + E$$

$$Y_i = \beta_0 + \beta_1 x_1 + E$$

Donde:

- $\beta_0 = \mu + 2\alpha_2$
- $\beta_1 = \alpha_1 - \alpha_2 = \text{Efecto promedio de la sustitución alélica}$

Sea el estadístico T:

$$T = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2(\widehat{\beta}_1)}} = \frac{\widehat{\beta}_1}{\widehat{\sigma}(\widehat{\beta}_1)}$$

$$E[T] = \frac{\beta_1}{E[\widehat{\sigma}(\widehat{\beta}_1)]}$$

Ecuación 3

Por teoría de regresión, T sigue aproximadamente una distribución normal con media E[T]:

$$T \sim N\left(\frac{\beta_1}{E[\widehat{\sigma}(\widehat{\beta}_1)]}, 1\right)$$

Ecuación 4

Además,  $T^2$  sigue una distribución chi-cuadrado no central con un grado de libertad (g.l):

$$T^2 = \frac{\widehat{\beta}_1^2}{\widehat{\sigma}^2(\widehat{\beta}_1)} \sim \chi^2_{(1,\lambda)}$$

Por teoría, la media de  $T^2$  es igual a  $\lambda + g.l$ . Por lo tanto, el valor del parámetro de no-centralidad es (102):

$$\lambda = E[T^2] - 1 = \frac{(N - 1)h_B^2}{1 - h_B^2}$$

$$\lambda = \frac{(N - 1)h_B^2}{1 - h_B^2}$$

Se empleó la distribución chi-cuadrado no central  $\chi^2(v,\lambda)$  para estimar la potencia estadística del siguiente modo:

1. Se determinó el valor crítico bajo la hipótesis nula, es decir para una distribución chi-cuadrado con un grado de libertad y  $\lambda=0$
2. Se empleó la ecuación 13 para calcular el valor de  $\lambda$  para loci con diferentes niveles de contribución a la varianza fenotípica.
3. Se empleó el parámetro  $\lambda$  del paso 2 para calcular la probabilidad de que el estadístico  $T^2$  con distribución  $\chi^2(v=1,\lambda)$  tome un valor menor al valor crítico (donde  $v=1$  es el g.l).

### Anexo 3. Cebadores utilizados para la amplificación de bmp4, kap1-1 y kap13-1

Las secuencias de los cebadores de BMP4 son las siguientes:

- *Forward*: 5'-CCG GGG GCT TAT GTA ACC TG-3'
- *Reverse*: 5'-AAG AAA CCG ACC TGT GGC TT-3'

**Tabla 19.** Descripción del par de cebadores empleados para amplificar el gen BMP4

Gen	Sentido	Hebra molde	longitud	Inicio*	Fin*	Tm	GC%	longitud del producto
BMP4	<i>Forward primer</i>	+	20	6119922	6119941	60.18	60	394
	<i>Reverse primer</i>	-	20	6120315	6120296	59.82	50	

\* Ubicación con respecto al scaffold17 (N. Accesoión: NW\_005882719.1)

**Tabla 20. Descripción de los cebadores empleados para la amplificación de los genes KAP1-1 y KAP13-1**

Gen	Cebador	Secuencia	Tm (°C)	Tamaño del amplicón	$\Delta G$ Hairpin (Kcal/mol)	$\Delta G$	
						Homodimero (Kcal/mol)	Heterodímero (Kcal/mol)
KAP1-1	F	5'-TGGCATTGGTGGTAGCATTG-3'	60	894	-0.52	-3.14	-6.21
	R	5'-AACTGGAGGCCTGTGTAAC-3'	58		-0.29	-3.07	
KAP13-1	F	5'-TGAGTCACGGGGATGACATC-3'	62	862	-2.2	-5	6.82
	R	5'-TGTGCTTCAGTGTCACATGTC-3'	61		-0.91	-7.07	

#### Anexo 4. Descripción de los genes empleados y las variantes genéticas observadas en este estudio

Gen	Símbolo	Alias	<i>Scaffold</i> (Número de accesión)	Sitios polimórficos	hebra
Bone Morphogenetic Protein 4	BMP4	BMP4	Scaffold17 (NW_005882719.1)	6120071	+
				6120125	+
				6120138	+
				6120166	+
				6120184	+
				6120187	+
				6120190	+
				6120193	+
				6120211	+
				6120213	+
				6120215	+
Keratin, high-sulfur matrix protein, B2C- like	LOC102528989	KAP1-1	Scaffold188 (NW_005882890.1)	3451340	-
				3451354	-
				3451389	-
				3451433	-
				3451449	-
				3451454	-

Gen	Símbolo	Alias	<i>Scaffold</i> (Número de accesión)	Sitios polimórficos	hebra
				3451467	-
				3451515	-
				3451595	-
				3451703	-
				3451716	-
				3451728	-
				3451784	-
				3451791	-
				3451870	-
				3451892	-
				3451907	-
				3451908	-
				3451930	-
				3451996	-
				3452036	-
				3452055	-
				3452067	-
	LOC102532866	KAP13-1	Scaffold101	151145	+

Gen	Símbolo	Alias	<i>Scaffold</i> (Número de accesión)	Sitios polimórficos	hebra
			(NW_005882803.1)	151160	+
Keratin-associated protein 13-1-like				151301	+
				151387	+
				151414	+

## Anexo 5. Frecuencias alélicas

**Tabla 21.** Frecuencias Alélicas del gen BMP4 en la muestra Nacional

Gen	SNP	Alelo	MAF*
	6120071	C>T	0.000
	6120125	C>T	0.213
	6120138	G>A	0.463
	6120166	G>A	<b><u>0.000</u></b>
	6120184	T>G	0.463
BMP4	6120187	G>C	0.027
	6120190	G>A	0.005
	6120193	G>A	0.076
	6120211	A>G	0.440
	6120213	C>G	0.078
	6120215	T>C	0.127

\*MAF: Frecuencia del Alelo menor (alelo con menor frecuencia)

**Tabla 22.** Frecuencias Alélicas de los tres genes candidatos en la muestra Rural Alianza

Gen	SNP	Alelo	MAF*	
<b>BMP4</b>	6120071	C>T	0.007	
	6120125	C>T	0.187	
	6120138	G>A	0.476	
	6120166	G>A	0.003	
	6120184	T>G	0.469	
	6120187	G>C	0.037	
	6120190	G>A	0.010	
	6120193	G>A	0.060	
	6120211	A>G	0.443	
	6120213	C>G	0.065	
	6120215	T>C	0.109	
	<b>KAP1-1</b>	3451340	T>C	0.469
		3451354	C>T	0.064
3451389		T>C	0.034	
3451433		A>T	0.114	
3451449		C>T	0.298	
3451454		G>A	0.070	
3451467		A>T	0.298	
3451515		T>C	0.169	
3451595		G>A	0.051	
3451703		G>A	0.051	
3451716		T>C	0.026	
3451728		G>C	0.051	
3451784		C>T	0.298	
3451791		C>G	0.007	
3451870		C>A	0.125	
3451892		T>G	0.298	
3451907		A>T	0.018	

**Tabla 22.** Frecuencias Alélicas de los tres genes candidatos en la muestra Rural Alianza

Gen	SNP	Alelo	MAF*
	3451908	C>A	0.289
	3451930	T>G	0.400
	3451996	G>A	0.057
	3452036	G>A	0.022
	3452055	G>C	0.066
	3452067	C>T	0.050
	151145	G>A	0.024
	151160	G>C	0.452
<b>KAP13-1</b>	151301	G>T	0.452
	151387	T>A	0.449
	151414	G>T	0.452

\*MAF: Frecuencia del Alelo menor (alelo con menor frecuencia)

## Anexo 6. Análisis de heterocigocidad

**Tabla 23.** Análisis de heterocigocidad en la muestra Nacional

<b>Gen</b>	<b>SNP (coordenada)</b>	<b>Heterocigosidad observada</b>	<b>Heterocigosidad esperada</b>	<b>HWE* (p-value)</b>	<b>Fis (W &amp; C)</b>
<b>BMP4</b>	6120125	0.340	0.337	0.6496	-0.01
	6120138	0.543	0.500	0.8514	-0.09
	6120184	0.543	0.500	0.8514	-0.09
	6120187	0.053	0.052	1.000	-0.02
	6120193	0.152	0.141	1.000	-0.08
	6120211	0.494	0.496	0.5752	0.00
	6120213	0.133	0.145	0.3987	0.09
	6120215	0.205	0.222	0.3733	0.08

\*HWE: Equilibrio de Hardy Weinberg. ‡ Método de Weir y Cockerham 1984.

**Tabla 24.** Análisis de heterocigocidad en la muestra Rural Alianza

Gen	SNP (coordenada)	Heterocigosidad observada	Heterocigosidad esperada	HWE* (p-value)	‡Fis (W & C)
BMP4	6120125	0.32	0.305	0.801	-0.05
	6120138	0.503	0.501	0.593	-0.01
	6120184	0.503	0.5	0.600	-0.01
	6120187	0.061	0.072	0.176	0.15
	6120193	0.12	0.113	1	-0.06
	6120211	0.492	0.495	0.542	0.01
	6120213	0.099	0.122	0.086	0.19
	6120215	0.202	0.195	0.796	-0.04
KAP13- 1	151160	0.388	0.497	0.006	0.22
	151301	0.388	0.497	0.006	0.22
	151387	0.395	0.496	0.010	0.21
	151414	0.388	0.497	0.006	0.22
KAP1-1	3451340	0.492	0.5	0.500	0.02
	3451354	0.114	0.121	0.423	0.06
	3451389	0.067	0.065	1	-0.03
	3451433	0.228	0.203	1	-0.13
	3451449	0.463	0.42	0.925	-0.10
	3451454	0.125	0.13	0.491	0.04
	3451467	0.463	0.42	0.925	-0.10
	3451515	0.309	0.282	0.933	-0.10

**Tabla 24.** Análisis de heterocigocidad en la muestra Rural Alianza

Gen	SNP (coordenada)	Heterocigosidad observada	Heterocigosidad esperada	HWE* (p-value)	‡Fis (W & C)
	3451595	0.103	0.098	1	-0.05
	3451703	0.088	0.098	0.297	0.10
	3451716	0.051	0.05	1	-0.02
	3451728	0.103	0.098	1	-0.05
	3451784	0.463	0.42	0.925	-0.10
	3451791	0.015	0.015	1	0.00
	3451870	0.221	0.22	0.648	0.00
	3451892	0.463	0.42	0.925	-0.10
	3451907	0.036	0.035	1	-0.01
	3451908	0.45	0.413	0.902	-0.09
	3451930	0.514	0.482	0.837	-0.07
	3451996	0.114	0.108	1	-0.06
	3452036	0.043	0.043	1	-0.02
	3452055	0.131	0.123	1	-0.07
	3452067	0.085	0.096	0.272	0.11

\*HWE: Equilibrio de Hardy Weinberg. ‡ Método de Weir y Cockerham 1984.

## Anexo 7. Resultados complementarios del análisis de asociación

Tabla 25. Resultados del análisis de asociación – Modelo de un locus con dominancia

Gen	SNP	Variable	Coefficiente	Error Estándar	P-value	R <sup>2</sup>
<b>BMP4</b>	6120125	(Intercepto)	20.64	0.323	0.000	0.021
		sexo	0.61	0.854	0.478	
		aditivo	0.42	0.789	0.596	
		dominante	0.39	0.918	0.672	
	6120138	(Intercepto)	21.20	0.556	0.000	0.009
		sexo	0.770	0.854	0.369	
		aditivo	-0.22	0.369	0.551	
		dominante	-0.12	0.521	0.811	
	6120213	(Intercepto)	21.25	0.294	0.000	0.028
		sexo	-0.01	0.839	0.992	
		aditivo	-1.41	1.050	0.182	
		dominante	0.21	1.330	0.875	
6120215	(Intercepto)	17.91	3.172	0.000	0.01	
	sexo	0.66	0.956	0.494		
	aditivo	1.59	1.581	0.317		
	dominante	1.66	1.699	0.332		
<b>KAP1-1</b>	3451340	(Intercepto)	21.32	0.585	0.000	0.044
		sexo	0.88	0.876	0.317	
		aditivo	-0.70	0.387	<b>0.073</b>	
		dominante	0.64	0.545	0.240	
	3451449	(Intercepto)	24.44	3.072	0.000	0.011
		sexo	0.71	0.869	0.415	
		aditivo	-1.81	1.543	0.244	
		dominante	-1.30	1.713	0.449	
	3451454	(Intercepto)	20.93	0.565	0.000	0.017
		sexo	0.71	0.876	0.417	
		aditivo	-0.27	0.553	0.631	
		dominante	-0.25	0.673	0.710	
3451515	(Intercepto)	20.94	2.166	0.000	0.023	

**Tabla 25. Resultados del análisis de asociación – Modelo de un locus con dominancia**

<b>Gen</b>	<b>SNP</b>	<b>Variable</b>	<b>Coefficiente</b>	<b>Error Estándar</b>	<b>P-value</b>	<b>R<sup>2</sup></b>
		sexo	0.62	0.866	0.476	
		aditivo	0.14	1.096	0.900	
		dominante	-0.76	1.193	0.523	
		(Intercepto)	15.38	3.047	0.000	
	3451703	sexo	0.56	0.862	0.516	0.033
		aditivo	2.77	1.530	<b>0.073</b>	
		dominante	3.47	1.768	<b>0.052</b>	
		(Intercepto)	18.70	2.178	0.000	
	3451870	sexo	0.60	0.875	0.493	0.012
		aditivo	1.12	1.100	0.309	
		dominante	1.20	1.240	0.336	
		(Intercepto)	21.13	0.499	0.000	
<b>KAP13-1</b>	151301	sexo	0.66	0.855	0.439	0.017
		aditivo	0.01	0.332	0.968	
		dominante	-0.66	0.524	0.212	

**Tabla 26. Selección del mejor modelo mediante Análisis de Varianza.**

Variable	SC	g.l	Estadístico F	P-value
add_bmp4_6120125	7.01	1	0.755	0.387
add_bmp4_6120138	0.19	1	0.021	0.886
add_bmp4_6120193	0.95	1	0.102	0.750
add_bmp4_6120213	10.71	1	1.154	0.286
add_bmp4_6120215	7.94	1	0.855	0.358
add_kap1_3451433	25.21	1	2.714	0.103
add_kap1_3451454	1.88	1	0.202	0.654
add_kap1_3451703	0.00	1	0.000	0.996
add_kap1_3451870	15.54	1	1.674	0.199
add_kap1_3451996	7.47	1	0.804	0.372
add_kap1_3452055	17.83	1	1.920	0.170
add_kap13_151301	1.08	1	0.116	0.734
dom_bmp4_6120125	0.85	1	0.091	0.763
dom_bmp4_6120138	3.99	1	0.430	0.514
dom_bmp4_6120193	ND	0	ND	ND
dom_bmp4_6120213	3.99	1	0.430	0.514
dom_bmp4_6120215	6.81	1	0.734	0.394
dom_kap1_3451433	ND	0	ND	ND
dom_kap1_3451454	ND	0	ND	ND
dom_kap1_3451703	ND	0	ND	ND

**Tabla 26. Selección del mejor modelo mediante Análisis de Varianza.**

Variable	SC	g.l	Estadístico F	P-value
dom_kap1_3451870	6.46	1	0.696	0.407
dom_kap1_3451996	ND	0	ND	ND
dom_kap1_3452055	ND	0	ND	ND
dom_kap13_151301	11.93	1	1.285	0.260
sexo	0.05	1	0.005	0.944
Residuals	822.34	88	-	-

SC: Suma del cuadrado de los errores; g.l: Grados de libertad.

El prefijo “add” se refiere al efecto aditivo; y “dom” al efecto de dominancia.

ND: No disponible (no fue posible calcular los coeficientes debido a la ausencia de una de las categorías genotípicas debido a la baja frecuencia del alelo menor).