CENTRO DE INVESTIGACIÓN Y ESTUDIOS AVANZADOS DEL
INSTITUTO POLITÉCNICO NACIONAL

UNIDAD DE GENÓMICA AVANZADA
UNIDAD IRAPUATO

"FIRMAS DE SELECCIÓN EN EL GENOMA HUMANO Y DESARROLLO DE UN MÉTODO DE EXPRESIÓN HETERÓLOGA PARA SU ANÁLISIS FUNCIONAL"

Tesis que presenta:
César Mauricio Campa Álvarez

Para Obtener el Grado de
Maestro en Ciencias

En la Especialidad de
Biotecnología de Plantas

Codirectores de la tesis:

Dr. Alexander de Luna Fors Dr. Rafael Montiel Duarte

Irapuato, Gto. Diciembre, 2018

El presente trabajo fue realizado en el grupo de Interacción Núcleo Mitocondrial y Paleogenómica y el grupo de Sistemas Genéticos del Laboratorio Nacional de Genómica para la Biodiversidad, Unidad de Genómica Avanzada (LANGEBIO-UGA) del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, bajo la codirección del Dr. Rafael Montiel Duarte y del Dr. Alexander de Luna Fors, así como la asesoría del Dr. Andrés Moreno Estrada y el Dr. Eugenio Mancera Ramos.

Agradecimientos

Quiero agradecer Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca 615694 otorgada. Así mismo, quiero agradecer a mis asesores, Dr. Rafael Montiel y el Dr. Alexander de Luna, por darme la oportunidad para realizar mi tesis de maestría en ambos laboratorios, es muy agradable y enriquecedor al aprender de ambas líneas de investigación. A mis asesores, Dr. Eugenio Mancera y Dr. Andrés Moreno, por su retroalimentación, observaciones y aportaciones para que este proyecto saliera adelante.

Agradecer a los integrantes del Laboratorio de Interacción Núcleo-Mitocondrial y Paleogenómica y también a cada uno de los integrantes del laboratorio de sistemas genéticos, por acogerme como un compañero, por su ayuda y por tener esas agradables experiencias con cada uno de ellos.

Quiero agradecer a mis amigos por siempre estar conmigo y apoyarme incondicionalmente. Además, quisiera agradecer a mis padres y hermanos que me han enseñado a seguir adelante, nunca rendirme, a enseñarme a hacer siempre las cosas bien, apoyándome en todo momento, alentándome a seguir adelante y siempre estar conmigo en mis momentos de felicidad, tristeza y superación, muchas gracias por su cariño y apoyo.

Gracias

Índice General

Índice de figuras

Índice de tablas

Resumen

La evolución de las poblaciones de humanos modernos ha sido acompañada por cambios drásticos en el ambiente y el estilo de vida. Cada uno de estos cambios probablemente resultó en poderosas presiones selectivas que generaron nuevos genotipos que se adaptaron mejor a los entornos novedosos. El paso de cazadores-recolectores a las sociedades agrícolas, que comenzó hace unos 10,000-12,000 años, es uno de los cambios más radicales que presenció el humano, y cada cambio sociocultural ocurrido durante esta transición pudo dar como resultado presiones selectivas que repercutirían en la selección de variantes alélicas del genoma a lo largo del tiempo.

Hasta la fecha, los mejores ejemplos de selección reciente en humanos se han descubierto en estudios de genes candidatos en los que existía una hipótesis previa de selección. Al mismo tiempo, se han realizado mapas de variación genética, solo de poblaciones humanas actuales para detectar firmas de selección en el genoma. Por lo tanto, se puede hipotetizar que, con el estudio genómico global comparativo de poblaciones ancestrales contra poblaciones actuales, se podrán identificar firmas de selección en el genoma humano y con ello identificar genes bajo el contexto de adaptación evolutiva. Para el presente trabajo se establecieron dos poblaciones, la primera constituida de 8 individuos antiguos cazadores-recolectores de entre 24,000 a 3,000 años antes del presente (A.P.) encontrados en América, Europa y Asia, y la segunda población, que consta de 5 grupos continentales de humanos actuales del proyecto de los 1,000 genomas; para realizar el estudio comparativo de la diversidad nucleotídica y el cambio de las frecuencias alélicas en el tiempo mediante el mapeo de los genomas de cada una de las poblaciones. De este modo, se logrará detectar firmas de selección en el genoma humano y con ello identificar genes candidatos sujetos a cambios genéticos en respuesta a condiciones ambientales.

Dentro de este estudio se encontró que la mayoría de las regiones con variaciones alélicas entre las poblaciones comparadas están dentro de las regiones no codificantes, sugiriendo que los cambios alélicos se presentan con mayor frecuencia en regiones regulatorias. Además, se identificaron polimorfismos de un solo nucleótido (SNP por sus siglas en ingles) en 28 genes, posibles candidatos de haber sido seleccionados por adaptación al ambiente agrícola.

Para contribuir a elucidar si estos cambios fueron producidos por selección y no por deriva, se propone la implementación del modelo de expresión heteróloga en levadura. Este modelo eucariota ha demostrado ser útil para detectar diferencias funcionales entre alelos humanos, mismo que se aplica para la investigación de un gen candidato (Adenilato quinasa 2) que reúne las condiciones adecuadas para ser probado en este modelo.

Abstract

The evolution of modern human populations has been accompanied by drastic changes in the environment and lifestyle. Each of these changes probably resulted in powerful selective pressures that generated new genotypes that were better adapted to novel environments. The passage from hunter-gatherers to agricultural societies, which began about 10,000-12,000 years ago, is one of the most radical changes that humankind witnessed, and every sociocultural change that occurred during this transition could result in selective pressures that would affect the selection of allelic variants of the genome over time.

The best examples of recent selection in humans have been discovered in candidate gene studies in which there was a prior selection hypothesis. At the same time, genetic variation maps have been made, only of current human populations to detect selection signatures in the genome. Therefore, it can be hypothesized that, with the global comparative genomic study of ancestral populations against current populations, it will be possible to identify selection signatures in the human genome and thereby identify genes under the context of evolutionary adaptation. For the present work, two populations were established. The first consisting of 8 hunter-gatherer ancient individuals from 24,000 to 3,000 years before present found in America, Europe and Asia. The second population, consisting of 5 continental groups of current humans from the 1,000 genome project; to perform the comparative study of nucleotide diversity and the change of allelic frequencies over time by mapping the genomes of each of the populations. In this way, it will be possible to detect selection signatures in the human genome and thereby identify candidate genes subject to genetic changes in response to environmental conditions.

Within this study it was found that most of the regions with allelic variations between the compared populations are within the non-coding regions, suggesting that allelic changes occur more frequently in regulatory regions. In addition, Single Nucleotide Polymorphisms (SNP) were identified genes, potential candidates having been selected for adaptation to the agricultural environment.

To contribute to elucidate if these changes were produced by selection and not by genetic drift, the implementation of the model of heterologous expression in *S. cerevisiae* is proposed. This eukaryotic model has been shown to be useful for detecting functional differences between human alleles, which is applied to the investigation of a candidate gene Adenylate kinase 2 (AK2) that meets the appropriate conditions to be tested in this model.

Introducción

El *Homo sapiens*, como todas las especies, ha surgido por selección natural positiva (Sabeti et al., 2006), ya que indudablemente ha jugado un rol crucial en su evolución (Vallender & Lahn, 2004). Los rasgos beneficiosos formados por su acción, probablemente incluyan el bipedismo, el habla, la resistencia a las enfermedades infecciosas y otras adaptaciones a entornos nuevos y diversos (Sabeti et al., 2006). Por ejemplo, en los últimos 100,000 años, los humanos anatómicamente modernos se han extendido desde África para colonizar la mayor parte del mundo. Dentro de esta expansión uno de los cambios más dramáticos ocurridos fue la transición de las sociedades de cazadores-recolectores a las agrícolas, comenzando hace unos 10,000-12,000 años en el Creciente Fértil, y un poco más tarde en otros lugares (Voight, Kudaravalli, Wen, & Pritchard, 2006).

La identificación de blancos a selección positiva en humanos ha sido, hasta hace poco, frustrantemente lenta, confiando en el análisis de genes candidatos. Estos análisis genéticos únicos han ilustrado profundos conocimientos sobre la historia evolutiva humana reciente. Sin embargo, la genómica ha proporcionado los recursos necesarios para investigar sistemáticamente todo el genoma en busca de firmas de selección natural (Akey, 2009). Con la cantidad creciente de información sobre la variación genética, junto con nuevos métodos analíticos y avances en genómica comparativa para definir las regiones de codificación y regulación, y el seguimiento biológico de candidatos prometedores, está haciendo posible explorar más la historia evolutiva reciente de la población humana (Sabeti et al., 2007).

La búsqueda para identificar los rasgos seleccionados no solo se debe a la curiosidad sobre el pasado, sino también a la preocupación por la salud humana (Sabeti et al., 2006). La identificación, en este trabajo, de firmas de selección por adaptación al entorno agrícola en el genoma humano puede proporcionar información sobre la capacidad de adaptación que tienen los humanos a ambientes cambiantes a lo largo del tiempo y sobre las modificaciones funcionales específicas que conlleva dicha adaptación.

Capítulo I. Antecedentes

1.1 Historia humana

El origen y la diseminación de nuestra especie, *Homo sapiens*, ha sido y seguirá siendo un tema de sumo interés. Como tema de investigación, combina los dos grandes aspectos del proceso evolutivo, el cambio temporal y la variación espacial. En todo el mundo, los investigadores están explorando la evolución de nuestra especie realizando desde excavaciones en los desiertos más remotos hasta análisis de laboratorio que hubieran sido imposibles hace solo unos pocos años (GROUCUTT, 2018).

La morfología y el comportamiento de los humanos modernos se fueron formando a través de procesos evolutivos acumulativos durante la época del pleistoceno, hace aproximadamente 2.5 millones años. La evidencia fósil y genética ubica el origen del *Homo sapiens* hace unos 200,000 años (Veile, 2018). A su vez, la evidencia genética y paleoantropológica está de acuerdo con que la población humana de hoy es el resultado de una gran expansión demográfica y geográfica, que comenzó hace unos 45,000 a 60,000 años en África y dio lugar rápidamente a la ocupación humana de casi todas las regiones habitables de la Tierra (Henn, et al., 2012). Sin embargo, este modelo de dispersión es controversial, ya que recientes descubrimientos de arqueología, paleontología, geocronología, genética y estudios paleoambientales sugieren múltiples dispersiones anteriores a los 60,000 años en regiones como el sur y el este de Asia (Bae, et al., 2017).

1.1.1 Cazadores recolectores

La definición de cazador-recolector se basa completamente en el modo de subsistencia y se refiere a sociedades que obtienen sus alimentos y otros requisitos directamente de fuentes naturales silvestres, como son las plantas y animales de caza (Crittenden & Schnorr, 2017). En la búsqueda de explicar la cultura humana, los antropólogos han prestado una gran atención a las recientes sociedades de cazadores-recolectores o recolectores. Una de las principales razones de este enfoque ha sido la creencia generalizada de que el conocimiento de las sociedades de cazadores-recolectores podría abrir una ventana hacia la comprensión de las primeras culturas humanas (Veile, 2018).

Durante la vasta extensión de la historia humana, la pérdida de variación genética resultó de la forma en que el mundo fue colonizado por grupos de cazadores-recolectores que, después de colonizar un nuevo hábitat y expandirse allí, arrojaron pequeños grupos que fundaron nuevas colonias cercanas. Este proceso condujo a la reducción sucesiva de la variación en las colonias recién fundadas (efecto fundador), siendo la reducción proporcional al número de fundadores (Henn et al., 2012). No fue sino hasta el inicio de la época del Holoceno, hace 11 mil años, que las sociedades en el sudoeste de Asia (en el Creciente Fértil) comenzaron a cultivar y domesticar plantas y animales, adoptando la agricultura (Veile, 2018). Se ha concluido que la mayoría de las poblaciones africanas parecen haber permanecido relativamente constantes en tamaño hasta la invención de la agricultura en África, el crecimiento asociado y los cuellos de botella que ocurrieron después de 12,000 años (Henn et al., 2012).

Los datos limitados disponibles de cazadores-recolectores ancestrales impiden una comprensión de los procesos selectivos asociados con esta transición crucial a la agricultura en la evolución humana reciente (Olalde et al., 2014). Con los avances tecnológicos se ha aumentado la capacidad de aislar DNA antiguo (Shapiro & Hofreiter, 2014). Estos avances han proporcionado un medio único para estudios genéticos evolutivos y de población para reconstruir el pasado (Willerslev & Cooper, 2005). Con ello haciendo posible acceder directamente a la información genética de individuos cazadores-recolectores.

1.1.2 La agricultura como presión de selección

El estudio de datos de individuos contemporáneos ha revelado varios ejemplos de adaptación humana. La posibilidad de secuenciar muestras antiguas, sin embargo, proporciona un nivel de resolución sin precedentes para delinear un proceso más detallado de cómo la selección positiva actuó sobre la variación genética a través del tiempo. Se han encontrado genes que han estado sujetos a cambios genéticos en respuesta a condiciones ambientales versátiles, como lo es el grosor del cabello (EDAR), pigmentación de la piel (SLC24A5), color de ojos (HERC2/OCA2), hipoxia (EPAS1, EGLN1, BHLHE1), inmunidad (HLA-DQA1) y metabolismo (TBX15 y FADS) (Antelope, Marnetto, Casey, & Huerta-Sanchez, 2017), pero además existen genes que han estado sujetos a cambios socioculturales.

Un ejemplo que representa una adaptación a la domesticación de los animales lecheros y el posterior consumo de su leche es la persistencia de la lactasa (capacidad de producir lactasa en etapa adulta para digerir la lactosa de la leche). La frecuencia de la persistencia a la lactosa es paralela a los hábitos ancestrales de consumo regular de leche de las poblaciones, con altas frecuencias encontradas en poblaciones pastorales o agropastorales que tradicionalmente incorporaron grandes cantidades de leche en sus dietas. Esto condujo a un aumento en la frecuencia de variantes genéticas que mantienen la expresión de lactasa en la adultez (hipótesis histórico-cultural). Estas variantes permitieron a los portadores ampliar su repertorio dietético, lo que a su vez permitió a las poblaciones incorporar más leche en sus dietas. Esta adaptación es un excelente ejemplo de coevolución gen-cultura y de evolución convergente (diferentes variantes genéticas que causan el mismo cambio fenotípico que surgieron en poblaciones múltiples que adoptaron prácticas de ordeño en diferentes áreas geográficas) (Ségurel & Bon, 2017).

La agricultura es la palabra utilizada para denotar las muchas formas en que las plantas cultivadas y los animales domésticos sostienen a la población humana, al proporcionar alimentos y otros productos (Harris & Fuller, 2014). Con la aparición de la agricultura en los últimos ~10,000-11,000 años (Veile, 2018), los recursos alimenticios se volvieron más abundantes y constantes. El consumo de cereales y otras plantas aumentó drásticamente. La domesticación de plantas y animales provocó cambios en numerosos aspectos de la vida, incluidos los alimentos disponibles, las actividades físicas, la reproducción, las relaciones psicosociales, las interacciones microbianas, la exposición a toxinas/alérgenos y el sedentarismo (Münster et al., 2018).

La producción de alimentos confirió enormes ventajas a los agricultores en comparación con los cazadores. La producción de alimentos pudo soportar densidades de población mucho más altas, ya que se pudo almacenar excedentes de alimentos, que fue un requisito previo para el desarrollo de tecnología compleja, estratificación social, estados centralizados y ejércitos profesionales (Diamond & Bellwood, 2003). Incluso, con la agricultura vinieron los asentamientos permanentes y las primeras ciudades, con lo cual ya se necesitaban menos personas para cultivar el alimento de grupo, y muchas de ellas pudieron especializarse en otras actividades. La agricultura, por tanto, abrió el desarrollo de la industria, la siguiente gran ola cultural, que continúa hasta nuestros días (Campbell, 2001). Sin embargo, la agricultura también ha sido referida como "El peor error en la

historia de la raza humana" debido a la evidencia arqueológica que la asocia a la disminución de la salud y el aumento de la desigualdad social en varios sitios neolíticos (Veile, 2018).

Crosby (1999) sugiere que los orígenes de la agricultura resultaron en una presión demográfica importante. Los hombres dejaron de depender de manadas de animales de gran tamaño, pasando a la explotación de animales pequeños y de algunas plantas; estos y cada uno de los diferentes cambios ocurridos durante esta transición, pudieron dar como resultado las presiones selectivas que repercutirían en la selección de variantes alélicas del genoma a lo largo del tiempo. Luca et al (2010), propone que las respuestas adaptativas a nuevas presiones selectivas pueden haber involucrado nuevos alelos (variaciones genéticas) generados por mutación, introducidos de una población cercana o alelos existentes que no eran ventajosos (ya sea neutrales o levemente deletéreos) antes del cambio ambiental. El último escenario proporciona una respuesta de adaptación más rápida y, por lo tanto, puede ser común en especies, como los humanos, que han sufrido cambios ambientales y dietéticos dramáticos.

1.2 Evolución molecular

El estudio de la historia evolutiva de los organismos y la elucidación de las fuerzas evolutivas que moldean la biodiversidad y las adaptaciones que producen, han sido dos campos de estudio dentro de la biología (Eguiarte, 2007). Los incrementos en la accesibilidad y resolución del análisis genómico, así como otras tecnologías de alto rendimiento que pueden ahora medir el impacto funcional de la variación genética, pueden ayudar a abordar cuestiones antropológicas importantes (Gokcumen, 2018).

El campo de la evolución molecular implica el estudio de secuencias de DNA, RNA y proteínas con el objetivo de dilucidar los procesos que causan tanto el cambio como la constancia entre las secuencias a lo largo del tiempo. A menudo se busca distinguir si un patrón de variación en una muestra de secuencias de DNA es consistente con la deriva genética o con ciertas formas de selección natural. La característica común de todas las pruebas de hipótesis en estudios de evolución molecular es el uso de hipótesis nulas y alternativas para los patrones y tasas de cambio de secuencia (Hamilton, 2009). La revolución en genética molecular humana ha producido una gran cantidad de información, no solo sobre la estructura y función de nuestros genes, sino también sobre la expresión génica, la mutación y la variación polimórfica. Los estudios a nivel del genoma

completo ya han dado abundantes pruebas de las firmas de selección ocurridas en el pasado y la evolución adaptativa en secuencias de genes humanos. Ahora se pueden identificar muchos de los eventos moleculares que se han producido durante la evolución (Cooper & Kehrer-Sawatzki, 2008).

El estudio de la adaptación a nivel molecular se ha abordado sólo de manera reciente como un intento de evaluar cuál es el papel de la selección a nivel molecular. El estudio de la adaptación molecular se ha realizado a partir de dos aproximaciones estadísticas diferentes, la primera es la distribución del polimorfismo en las secuencias de DNA y la segunda es la determinación de las sustituciones sinónimas y no-sinónimas en secuencias codificantes de DNA, utilizando el codón como mínima unidad evolutiva (Eguiarte, 2007).

1.2.1 Selección natural y deriva genética

Comprender las fuerzas evolutivas a las que las poblaciones humanas han estado sujetas en el pasado ha sido un tema central en la biología evolutiva. En particular, determinar cómo dos mecanismos principales, la deriva genética y la selección natural, han dado forma a la variación genética es un tema que sigue siendo muy estudiado (Antelope et al., 2017). Tanto la selección natural como la deriva genética pueden conducir a la eliminación o fijación de un alelo particular (Cavalli-Sforza & Feldman, 2003). Además hay muchas formas de selección natural y, a veces, varios tipos de selección conducen a patrones genéticos similares (Antelope et al., 2017), lo que dificulta distinguir los cambios de frecuencia debidos a selección, de los cambios debidos a deriva.

La teoría neutral, que propone que la mayoría de las variaciones genéticas observadas, tanto dentro como entre especies, son neutrales (es decir, no tienen ningún efecto sobre la capacidad reproductiva de un individuo), por lo que su prevalencia poblacional cambia con el tiempo por casualidad, proceso llamado deriva genética (Sabeti et al., 2006). Además, los procesos neutros también pueden generar complejidad. Cuando las variantes generadas son efectivamente neutrales se acumulan y se encuentran en relación con el tamaño de la población. Así mismo la "Evolución Neutral Constructiva" denota la retención efectivamente irreversible de las interacciones moleculares que inicialmente surgen de manera neutral. La selección está involucrada, pero solo contra la pérdida de complejidad, pero no por su origen (Brunet & Doolittle, 2018).

Los cambios aleatorios en la frecuencia de los alelos de una generación a la siguiente, en poblaciones biológicas, debido a las muestras finitas de individuos, gametos y, en última instancia, alelos que contribuyen a la siguiente generación, se le conoce como deriva (Hamilton, 2009). La deriva genética (aunque es aleatoria y no direccional) puede dar como resultado la aparición de características neutrales o incluso desfavorables en el fenotipo (Campbell, 2017). La deriva genética aumenta a medida que disminuye el tamaño de la muestra utilizada (Hamilton, 2009). Por ejemplo, cuando una pequeña población está aislada en un ambiente relativamente rico, bajo estas condiciones, las características novedosas, pero aleatorias, pueden sobrevivir por un tiempo debido a que las presiones de selección son bajas (Campbell, 2017). Los efectos de la deriva genética son más obvios durante periodos de tiempo más largos. Las frecuencias de los alelos a lo largo de unas pocas generaciones cambian al azar, tanto aumentando como disminuyendo (Figura 1), a veces cambiando muy poco en una generación y otras veces cambiando más sustancialmente (Hamilton, 2009).

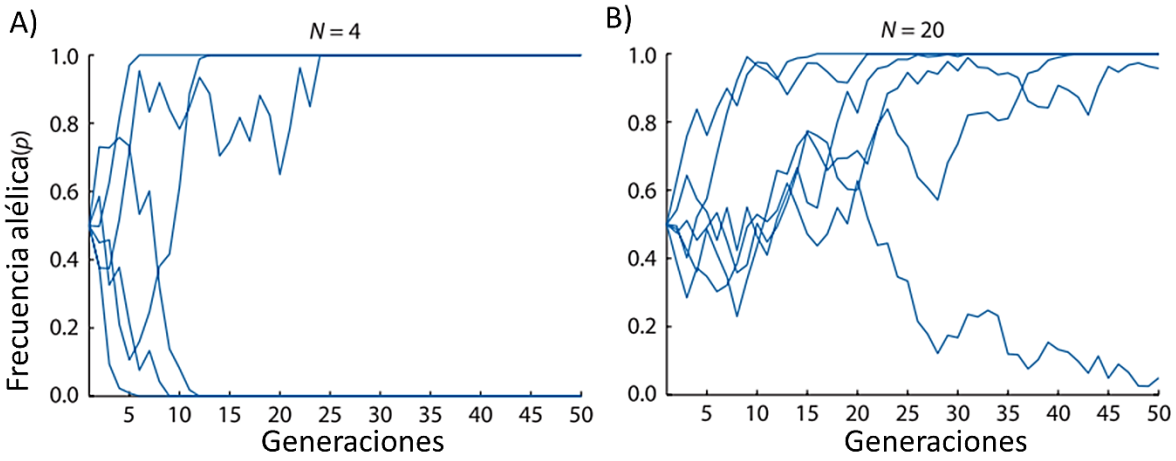


Figura 1.- Ejemplo de Comportamiento de la deriva genética por generación en poblaciones de N = 4 (A) y N = 20 (B). (Modificado de Hamilton, 2009).

En la Figura 1 se ejemplifica el comportamiento de la deriva en dos poblaciones de 4 y 20 individuos, las seis líneas representan poblaciones independientes que experimentan deriva genética a partir de la misma frecuencia inicial de alelos (p = 0.5). La naturaleza aleatoria de la deriva genética puede verse por los cambios en zigzag en la frecuencia de los alelos que no tienen una dirección aparente. Las frecuencias de alelos que alcanzan los ejes superior o inferior representan casos de fijación (p=1) o pérdida (p=0).

La biología de los organismos está conformada por la selección natural para maximizar la adecuación biológica o el éxito reproductivo (Veile, 2018). Los rasgos beneficiosos que son hereditarios aumentan en frecuencia con el tiempo, mientras que los rasgos heredables desfavorables se vuelven menos comunes. Este proceso puede llevarse a cabo de acuerdo con una variedad de modelos, que generan diferentes firmas en los patrones de variación genética (Hamilton, 2009). La selección direccional ocurre cuando uno de los dos alelos en un sitio polimórfico es favorecido sobre el otro de modo que aumentará rápidamente en frecuencia; si la selección es estable durante un período suficientemente largo de tiempo, el alelo favorecido alcanzará la fijación (Luca et al., 2010).

Una de varias explicaciones para variaciones en el genoma inusualmente antiguas es el efecto de la selección balanceadora: la fuerza adaptativa que favorece no solo a uno, sino a varios alelos en una región genómica y se pueden mantener a lo largo de millones de años (Gokcumen, 2018), a una frecuencia de equilibrio estable durante el tiempo que la presión selectiva esté presente; bajo un modelo particular de selección de equilibrio, los individuos heterocigotos tienen una adecuación mayor que la de ambos homocigotos, y se espera que un nuevo alelo aumente rápidamente en frecuencia hasta que alcance el equilibrio (Luca et al., 2010). Cuando los alelos bajo selección positiva aumentan en la prevalencia en una población, dejan "firmas" distintivas o patrones de variación genética en la secuencia de DNA. Estas firmas se pueden identificar en comparación con la distribución de fondo de la variación genética en humanos, que generalmente se argumenta que evolucionan en gran parte bajo neutralidad (Sabeti et al., 2006).

1.2.2 Identificación de selección

Los patrones de variación genética humana actuales no solo resultan de procesos selectivos, sino también de la historia de la población, y desentrañar los efectos de estos procesos es un desafío. Para lidiar con este problema, muchos investigadores han tomado una ruta empírica que no hace suposiciones sobre la historia de las poblaciones. De modo a apuntar la identificación de los loci con los patrones más inusuales, en comparación con los conjuntos de datos a gran escala de variación genética. Por ejemplo, los patrones de variación de todo el genoma se pueden clasificar por medio de una o más pruebas estadísticas; aquellos loci que caen por encima de un límite arbitrario se identifican como inusuales y a menudo se los denomina valores atípicos. Bajo el supuesto de que la mayoría de los loci en el genoma humano evolucionan de forma neutral, estos

loci con valores atípicos representan posibles candidatos de fuertes presiones selectivas (Luca et al., 2010).

Es por ello que se espera que mediante el análisis de grandes conjuntos de datos genómicos comparativos y grandes conjuntos de datos de SNP podremos determinar cómo y dónde, tanto la selección positiva como la negativa, han afectado la variación en los humanos (Nielsen, 2005). Existen estimadores que permiten identificar regiones determinadas en las que actuó la selección, y que se basan en la distribución de los polimorfismos de las poblaciones en estudio (medidas de polimorfismos de DNA), como lo son la diversidad nucleotídica y theta.

1.2.2.1 Diversidad nucleotídica

Una medida o estimador de polimorfismos de DNA es el número de sitios segregantes (S). Un sitio de segregación es cualquier sitio nucleotídico (L), que mantiene dos o más variantes dentro de la población. El número de sitios segregantes por sitio nucleotídico (P_s), se obtiene dividiendo el número de sitios segregantes *S* por el número total de sitios *L* (Hamilton, 2009), como se muestra en la fórmula 1.1.

$$P_s = \frac{S}{L} \qquad (1.1)$$

Otra medida es el índice de diversidad nucleotídica para una muestra de secuencias de DNA, simbolizada por π, y también conocida como la diferencia media entre pares de secuencias de una muestra de secuencias. La diversidad nucleotídica es calculada por la sumatoria de los sitios nucleotídicos que difieren entre cada par único de las secuencias de DNA (Hamilton, 2009). Se deduce que π se ve afectada mayormente por los alelos que poseen mayor frecuencia y es independiente del tamaño de la muestra, y con ello determinar si las secuencias se encuentran bajo el modelo neutral o se desvían del mismo (Eguiarte, 2007). De acuerdo a la distribución de los polimorfismos, si el valor de un locus se encuentra en la media de la distribución, se sugiere que este locus tiene una substitución aleatoria por lo que se considera que se encuentra en equilibrio neutral. Si existe un valor diferente al de la media de la distribución, se sugiere que se desvían de la neutralidad.

1.2.2.2 Theta (θ)

A la probabilidad de que dos alelos muestreados al azar de una población en equilibrio mutación-deriva sean alocigotos (genotipo en el que dos alelos en un locus provienen de fuentes completamente diferentes, como en la mayoría de los apareamientos normales y aleatorios), se le conoce como θ (Hamilton, 2009). θ sí se ve afectada por el tamaño de la muestra y por la deriva génica, es decir, por los alelos poco frecuentes (Eguiarte, 2007).

La relación que existe entre π y θ permite determinar si las secuencias se encuentran bajo el modelo neutral o se desvían del mismo. Si ambos estimadores dan el mismo resultado en cuanto a variación genética, quiere decir que el polimorfismo observado es neutro y se encuentra distribuido aleatoriamente. En cambio, si existen diferencias entre ambos quiere decir que la selección está afectando alguno de ellos promoviendo su incremento o decremento: si existe selección positiva, ésta incrementará las frecuencias alélicas y eso se reflejará en el incremento de π; si existe un mayor número de alelos deletéreos en la muestra, θ se verá incrementada (Eguiarte, 2007). De esta forma, si determinamos estadísticamente las diferencias entre ambos podremos detectar de manera indirecta la participación de la selección direccional en el mantenimiento del polimorfismo en las poblaciones.

1.2.2.3 Prueba de Tajima

Los primeros trabajos desarrollados para el estudio de la selección a nivel molecular fueron los basados en la distribución del polimorfismo como la prueba de Tajima, que se basada en las diferencias entre los estimadores π y θ. La *D* de Tajima es una prueba del modelo coalescente estándar (todas las mutaciones son selectivamente neutrales y la población permanece constante en el tiempo) que se aplica comúnmente a los datos de polimorfismo de DNA muestreados de una sola especie. La prueba utiliza la diversidad nucleotídica y el número de sitios segregantes observados en una muestra de DNA, en donde la hipótesis nula de la prueba es que la muestra de secuencias de DNA se tomó de una población con un tamaño de población efectivo constante y neutralidad selectiva de todas las mutaciones. La selección natural que opera en secuencias de DNA y los cambios en el tamaño efectivo de la población a lo largo del tiempo llevan al rechazo de esta hipótesis nula (Hamilton, 2009). Si D resulta negativa quiere decir que θ posee un valor mayor que π, lo que indica la presencia de mutaciones deletéreas. En cambio, si D resulta positiva quiere decir

que π tiene un mayor valor que θ, indicación de que algunos alelos se encuentran bajo selección positiva (por ejemplo, selección balanceadora) incrementando sus frecuencias. Si D es igual a cero quiere decir que no existe diferencia alguna entre ambos estimadores y nos encontramos bajo equilibrio neutral (Eguiarte, 2007).

1.2.3 Complementación heteróloga

A pesar de que se pueden identificar firmas se selección, se desconoce el alcance real que pudiera tener en el linaje humano. La evidencia genética molecular para adaptaciones a menudo viene sin una comprensión completa de las consecuencias funcionales y fenotípicas de las variantes o cambios genéticos implicados (Luca et al., 2010). Por ello, la demostración funcional sustenta y mejora considerablemente dicha evidencia; desde la correlación del alelo seleccionado con la variación fenotípica humana, usando un sistema modelo o estudios de laboratorio *in vitro* del alelo seleccionado (Sabeti et al., 2006).

En este caso, el análisis funcional de los genes ortólogos en varios modelos genéticos es posible y puede, por ejemplo, dar lugar a desarrollos rápidos en la comprensión de mecanismos de las enfermedades humanas (Culetto & Sattelle, 2000). Para este trabajo se analiza el modelo de clonación, potencial para la complementación heteróloga. El modelo de *Saccharomyces cerevisiae* ha sido muy bien estudiado, presenta genes ortólogos respecto al humano y se han usado para entender enfermedades humanas (Sun et al., 2016). Este organismo modelo puede ayudar a entender y ver diferencias funcionales y fenotípicas entre alelos, que a continuación se describen.

1.2.3.1 Saccharomyces cerevisiae

Saccharomyces cerevisiae es un modelo eucariota probado para estudios de biología molecular y celular (Mager & Winderickx, 2005). *S. cerevisiae* se han convertido en el genoma eucariota mejor caracterizado, proporcionando un conocimiento detallado de las vías genéticas y metabólicas, además del conocimiento de la estructura y función de un genoma eucariótico completo. Las ventajas de este modelo incluyen: tiempos de generación cortos, la facilidad de propagación, historias de vida simples, control preciso sobre las variables ambientales, la capacidad de preservar y la facilidad de replicar a cualquier nivel o etapa de un experimento (Zeyl, 2000). Además, es un organismo genéticamente tratable, susceptible de modificaciones tales como la alteración de genes,

el marcado de genes, la mutación o los efectos de dosificación de genes (Mager & Winderickx, 2005).

Debido a estas características ventajosas, la levadura se ha convertido en el organismo modelo elegido para la investigación relacionada con la medicina (Mager & Winderickx, 2005). Así mismo, usando la complementación entre especies, el organismo modelo *S. cerevisiae*, puede utilizarse como plataforma para probar variantes genéticas humanas (Hamza et al., 2015). Se sabe que casi la mitad de los genes esenciales de levadura que presentan un ortólogo en humano, pueden ser complementados con éxito (47% de ellos complementa la cepa *knockout*). Como los genes humanos desempeñan un papel muy similar en ambos organismos, el DNA codificante de proteínas de un gen humano puede sustituir al de la levadura, aumentando la posibilidad de humanizar procesos celulares completos en la levadura. Tales cepas simplificarían el descubrimiento de fármacos contra las proteínas humanas, permitirían estudiar las consecuencias de los polimorfismos genéticos humanos y potenciar los estudios funcionales de procesos celulares humanos completos en un organismo simplificado (Kachroo et al., 2015).

1.3 Paleogenómica

La rama de la biología que estudia cómo interactúan los diferentes componentes del genoma y los diferentes productos genéticos, utilizando un conjunto de herramientas en combinación con las metodologías de mapeo y secuenciación, se conoce como genómica (Primrose & Twyman, 2009), siendo un objetivo principal el inferir directamente la función de un genoma (Shapiro & Hofreiter, 2014). La genómica ofrece la oportunidad de identificar los genes que son responsables de la evolución de los caracteres clave compartidos de los organismos, o sinapomorfías, que en última instancia se utilizan para reconstruir el árbol de la vida (Bottjer et al., 2006). La paleogenómica, que es el estudio de los genomas de organismos antiguos (Lalueza-Fox & Gilbert, 2011), permite tanto de registros fósiles geológicos como genéticos, elucidar sobre el origen y la posterior evolución a través del tiempo, de genes y sinapomorfías clave. En otras palabras, la paleogenómica es la adición del componente tiempo al campo de la genómica (Bottjer et al., 2006).

La paleogenómica desempeña un papel cada vez más importante en mejorar la comprensión de los procesos evolutivos a corto y mediano plazo. La anotación mejorada de los genomas modernos facilita enormemente el análisis e interpretación de los paleogenomas con una mejor integración

con otros campos de investigación, incluida la biología y la bioquímica de desarrollo y sintética, que sin duda facilitará el logro de estos objetivos (Shapiro & Hofreiter, 2014).

La investigación en genómica antigua avanza nuestros conocimientos sobre la variación genética humana, cimentando nuestra visión de la variación humana como una mezcla en constante cambio de interacciones complejas. Por ejemplo, los nuevos datos genómicos permiten, por primera vez, una cuantificación de los grados de mezcla y los patrones no aleatorios con respecto a la geografía y el tiempo (Gokcumen, 2018). La mayoría de los estudios han utilizado DNA antiguo (aDNA) para proporcionar nuevos conocimientos sobre loci seleccionados positivamente, identificados en las poblaciones humanas actuales e inferir dónde se originaron las variantes beneficiosas, cuándo se produjo la adaptación y la magnitud de la selección (Antelope et al., 2017).

1.3.1 Alcances y limitaciones del DNA antiguo

Durante mucho tiempo, el análisis del aDNA humano representó una de las disciplinas más controvertidas en un campo de investigación ya controvertido. Experimentos comprometidos por la contaminación de DNA humano moderno generaron la duradera controversia sobre la autenticidad del aDNA. Sin embargo, el desarrollo de la secuenciación de próxima o segunda generación (SGS) en 2005 y los avances tecnológicos asociados con ella, han generado una nueva confianza en el estudio genético de restos humanos antiguos. La capacidad de secuenciar fragmentos de DNA más cortos, junto con un rendimiento de secuenciación muy alto, han reducido el riesgo de secuenciar la contaminación moderna y han proporcionado herramientas para evaluar la autenticidad de los datos de secuencia de DNA (Knapp et al., 2015).

Además de la contaminación, la degradación del DNA también ha comprometido la posibilidad de obtener una alta profundidad de secuencia y ningún genoma nuclear antiguo ha sido secuenciado a un nivel suficiente para el genotipado y la exclusión de errores debido a la secuenciación o daño post mortem (Rasmussen et al., 2010). Sin embargo, el DNA que se ha recuperado de restos arqueológicos y paleontológicos permite retroceder en el tiempo y estudiar las relaciones genéticas de organismos extintos con sus parientes contemporáneos. Esto proporciona una nueva perspectiva sobre la evolución de los organismos y las secuencias de DNA (Hofreiter et al., 2001). Por ejemplo con el aDNA se ha proporcionado información sobre la relación entre los humanos anatómicamente

modernos, que se extendieron desde África al resto del mundo comenzando hace unos 100,000 años, y sus precursores en Europa, los Neandertales (Pääbo et al., 2004).

1.3.2 Daño molecular

Después de la muerte de un organismo, los compartimentos celulares que normalmente secuestran las enzimas catabólicas se descomponen. Como consecuencia, el DNA se degrada rápidamente por enzimas tales como las nucleasas lisosómicas. Además, la molécula de DNA se enfrenta a una avalancha de bacterias, hongos e insectos que se alimentan y degradan macromoléculas. El daño se acumula progresivamente hasta que el DNA pierde su integridad y se descompone, con una pérdida irreversible de información sobre la secuencia de nucleótidos (Pääbo et al., 2004). En la Tabla 1 se mencionan los diferentes tipos de daños que pueden ocurrir al DNA.

Tabla 1. Tipos de daño en el aDNA (Pääbo et al., 2004).

Tipo de daño	Proceso	Efectos en el DNA
Ruptura de hebras	-Degradación por microorganismos -Nucleasas en la célula Post mortem -Otros procesos químicos	-Reducción de las cantidades totales de DNA -Reducción de tamaño
Lesiones oxidativas	-Daño a las bases -Daño a los residuos de desoxirribosa	-Fragmentación de las bases -Fragmentación del azúcar -Modificación nucleotídica
Enlaces cruzados de DNA	-Reacciones entre DNA y con otras biomoléculas	-Por ejemplo, reacciones de Maillard
Lesiones hidrolíticas	-Perdida de grupos amino 1.-Adenina por hipoxantina 2.-Citocina por uracilo 3.- 5-metil-citocina por timina 4.-Guanina por xantina	-Cambio de potencial de codificación

A pesar de las limitaciones, el análisis del aDNA ofrece la posibilidad única de permitir que los individuos fallecidos y las especies extintas contribuyan a nuestra comprensión de la evolución genética molecular (Pääbo et al., 2004). El campo ahora se está desarrollando rápidamente, proporcionando información sin precedentes sobre la evolución de nuestra propia especie y la dinámica de la población humana en el pasado. Así mismo como la evolución y la historia de patógenos y epidemias humanas (Knapp et al., 2015). Por ejemplo, la secuenciación del aDNA de los homínidos arcaicos, como se muestra en la Figura 2, ha revelado una rica historia de mezcla entre los primeros humanos modernos, los neandertales y los denisovanos, y nos ha permitido desenmarañar procesos selectivos complejos (Slatkin & Racimo, 2016).

Figura 2.- Distribución geográfica de genomas completos de homínidos arcaicos existentes (Modificado de Morozova et al., 2016)

En la actualidad, no solo se pueden elucidar patrones de relación entre las poblaciones, sino que también pueden proporcionar respuestas detalladas a cuestiones históricas de relevancia para la arqueología y la paleoantropología (Slatkin & Racimo, 2016).

El análisis del genoma humano y sus procesos evolutivos llevará a entender de qué manera se dan las adaptaciones a cambios radicales, como lo es la transición de las sociedades de cazadores-recolectores a las sociedades agrícolas. Así mismo, el enfoque genómico ofrecerá información sobre los genes que le han permitido al humano adaptarse a dichos ambientes diferenciales. En este trabajo se pretende determinar que genes fueron sujetos a selección durante dicha transición mediante una aproximación bioinformática. Así mismo, generar una lista de cuales de ellos resultan idóneos para clonarlos en el modelo experimental de *S. cerevisiae* como una primera evaluación de posibles diferencias funcionales entre las variantes alélicas que pueda explicar y sustentar la selección direccional a la que habrían sido sujetos.

Objetivos

- Objetivo General:

Identificar y caracterizar funcionalmente regiones con firmas de selección en el genoma humano para la identificación de genes bajo selección, como candidatos a la adaptación al ambiente agrícola.

- Objetivos Específicos

- ✓ Análisis de la diversidad nucleotídica de los genomas humanos antiguos y actuales para la búsqueda de firmas de selección en el genoma humano.

- ✓ Determinar el subconjunto de regiones codificantes con firmas de selección para obtener una la lista de genes candidatos.

- ✓ Generar una lista de los genes ortólogos de cada gen candidato respecto al modelo de *Saccharomyces cerevisiae*.

- ✓ Establecer la metodología para analizar la adecuación de genes humanos en el modelo *Saccharomyces cerevisiae*.

Hipótesis

Mediante el estudio comparativo de la diversidad nucleotídica y la diferencia en frecuencias nucleotídicas del genoma de una población ancestral, constituida de cazadores-recolectores, respecto a la población actual, es posible identificar firmas de selección en el genoma humano.

Capítulo II. Materiales y Métodos.

2.1 Identificación de Regiones con firmas de selección

Este estudio se enfocó en la búsqueda de regiones con firmas de selección (ver esquema-resumen de la metodología en la Figura 3), para lo cual se establecieron dos poblaciones, la población antigua que consta de individuos cazadores-recolectores encontrados en disímiles partes del mundo y con diferente temporalidad (Figura 4.A), y la población actual del proyecto de los 1,000 genomas, que consta de distintos grupos continentales del mundo. Para la búsqueda de firmas de selección se utilizaron los estimadores de diversidad nucleotídica (π), θ y la D de Tajima.

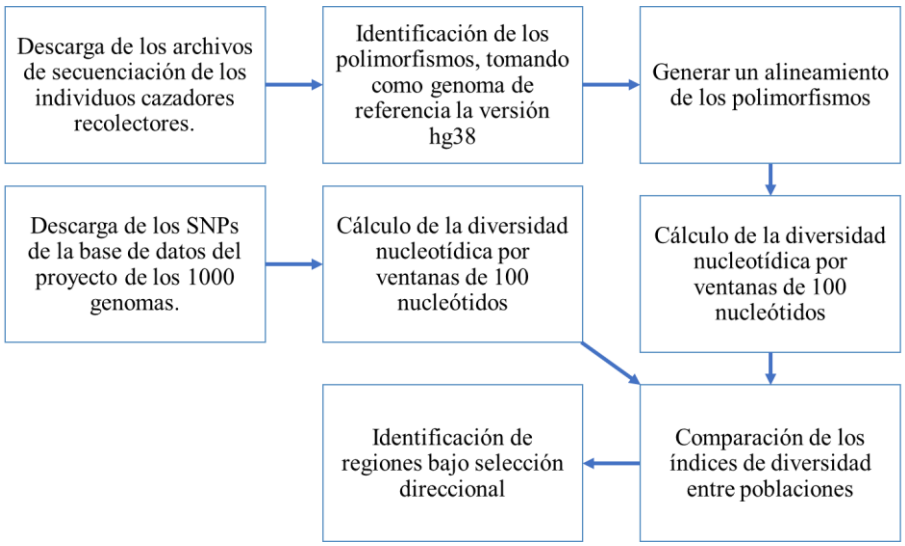


Figura 3.- Metodología para la identificación de regiones bajo selección direccional.

2.1.1 Población ancestral

Para la población antigua se seleccionaron 8 individuos antiguos, cada uno con indicios de ser cazador-recolector por la temporalidad, el análisis genómico y por las evidencias arqueológicas encontradas en su respectivo sitio arqueológico. Además, esta selección se debió a que se secuenció su genoma nuclear, que se encuentra disponible o se encuentran los archivos de secuenciación, necesarios para este estudio. En la Tabla 2 se muestra cada uno de los individuos antiguos seleccionados para el análisis, así mismo, se desglosa la calidad de mapeo tomada por cada autor, el genoma de referencia usado y el archivo que se descargó. La calidad de mapeo que se muestra

en la Tabla 2, es la escala de probabilidad "phred" de que una lectura o read sea alineado erróneamente (Li, Ruan, & Durbin, 2008), dada por la Fórmula 2.1.

$$Q_s = -10\ log_{10} Pr. \qquad (2.1)$$

Donde:
-***Qs*** es la calidad de mapeo.
-***Pr*** es el error de mapear un read.

Tabla 2. Resumen de los individuos antiguos usados para el análisis.
Se incluyen las características con las que se llevó a cabo el mapeo según sus autores.

Genoma	Calidad de mapeo (escala Phred)	Genoma de referencia	Archivo	Referencia
Anzick-1	30	hg18	BAM	Raghavan et al., 2014
Bichon	30	hg19	FASTQ	Jones et al., 2015
Kennewick	30	hg19	BAM	Rasmussen et al., 2015
Kotias	30	hg19	FASTQ	Jones et al., 2015
La Braña 1	25	hg19	FASTQ	Olalde et al., 2014
Loschbour	30	hg19	BAM	Lazaridis et al., 2014
Ma-1	30	hg18	BAM	Raghavan et al., 2014
Inuk	37	hg18	BAM	Rasmussen et al., 2010

2.1.1.1 MA-1

El individuo con mayor antigüedad (MA-1) es del Paleolítico superior (cazador-recolector), con una antigüedad de 24,157±266 años calibrados antes del presente (A.P.). Este es el genoma de humano anatómicamente moderno más antiguo reportado hasta la fecha. Se secuenció a partir de huesos encontrados en Mal'ta en el sur-centro de Siberia (Figura 4.B). Su genoma mitocondrial pertenece al haplogrupo U que también se ha encontrado en alta frecuencia en los antiguos cazadores-recolectores del Paleolítico superior y el Mesolítico europeo. El cromosoma Y de MA-1 es basal para los eurasiáticos occidentales de hoy en día y cerca de la raíz de la mayoría de los linajes nativo-americanos, lo que sugiere una conexión entre la Europa preagrícola y del Paleolítico superior de Siberia (Raghavan et al., 2014).

2.1.1.2 Bichon

Genoma del Paleolítico superior del oeste de Europa Occidental, secuenciado del hueso pétreo de un individuo masculino joven (de entre 20 y 23 años) de tipo Cro-magnon, con una edad de 13,665±105 años calibrados A.P., encontrado en "La cueva del Bichon", situada en las montañas Jura, Suiza, como se muestra en la Figura 4.B. Se encontraron en el sitio puntas de flecha de pedernal aparentemente proveniente de armas de cazador y además huesos de oso. A este espécimen se le asigno el haplogrupo U5b1h, la rama de los grupos U, especialmente el haplogrupo U5, que ha sido encontrado como haplogrupo dominante entre las comunidades cazadoras recolectoras (Jones et al., 2015).

2.1.1.3 Anzick-1

Anzick-1 es un infante varón con una antigüedad de 12,631±75 años calibrados A.P., cerca del final del período de Clovis. La secuenciación de su genoma fue a partir del hueso craneal pétreo encontrado en Montana, EUA. (Figura 4.B). Este individuo proviene de la cultura Clovis, ya que se asoció directamente con las herramientas encontradas de esta cultura, ya que son tecnológicamente consistentes con artefactos de este complejo. Estas personas, en última instancia, derivan de Asia y se relacionan directamente con los nativos americanos. Una hipótesis postula que los predecesores de los Clovis emigraron desde el suroeste de Europa durante el último máximo glacial. Su genoma mitocondrial es del haplogrupo D4h3a que es uno de los linajes de DNA mitocondrial raros específicos para los nativos americanos, se distribuye a lo largo de la costa del Pacífico en el norte y sur de América, entre las poblaciones contemporáneas y también está presente en muestras antiguas (Rasmussen et al., 2014).

2.1.1.4 Kotias

Genoma mesolítico, secuenciado de un diente molar de un adulto masculino joven encontrado en la cueva Kotias Klde en el oeste de Georgia, como se muestra en la Figura 4.B, con una edad de 9,712±183 años calibrados A.P. Este espécimen fue encontrado con evidencia de la industria mesolítica, además de huesos animales. Con el análisis del haplogrupo mitocondrial se le asignó el H13c, el más predominante y diverso encontrado en el oeste de Eurasia (Jones et al., 2015).

2.1.1.5 Kennewick

Kennewick es un esqueleto adulto masculino descubierto en el rio Columbia, en el estado de Washington, EUA. (Figura 4.B), con una edad de 8,770±430 años calibrados A.P., de la época del Holoceno temprano. Su genoma se secuenció a partir del hueso metacarpiano y con ello se ha relacionado su similitud genética con los nativos americanos. De acuerdo con el análisis mitocondrial, Kennewick es colocado en la raíz del haplogrupo X2a y no comparte ningún alelo derivado de las ramas que llevan a los haplogrupo X2a1 o X2a2, y el haplogrupo del cromosoma Y es Q-M3, siendo ambos linajes uniparentales encontrados casi exclusivamente entre nativos americanos contemporáneos (Rasmussen et al., 2015).

2.1.1.6 La Braña 1

Esqueleto masculino de un individuo europeo preagrícola del Mesolítico, descubierto en el sitio "La Braña-Arintero" en León, España (Figura 4.B), con una edad de 7,815±125 años calibrados A.P. Su genoma se secuenció a partir de la raíz del tercer molar superior izquierdo. Este espécimen presenta el haplogrupo mitocondrial U5b2c1y su haplogrupo del cromosoma Y es el C, relacionado con hombres del sureste de Europa, lo que sugiriere que puede ser un haplogrupo antiguo del clado europeo. Se cree que la distribución actual del haplogrupo C se debe a una sola salida migratoria de África a través del sur de Asia, siguiendo hacia el norte y que eventualmente alcanzo Siberia y las Américas (Olalde et al., 2014).

2.1.1.7 Loschbour

Esqueleto masculino del Mesolítico tardío, descubierto en el contexto de cazador recolector por los artefactos encontrados relacionados con este periodo, recuperado en *Loschbour rock shelter* en Heffingen, Luxemburgo (Figura 4.B), con una edad de 6,105±115 años calibrados A.P. La secuenciación de su genoma se llevó a cabo a partir de un diente. El haplogrupo de este individuo es el U5b1a que es típico de los genomas mesolíticos de cazadores recolectores (europeos pre-neolíticos), además su haplogrupo del cromosoma Y es el I, que es común en los europeos preagricultores (Lazaridis et al., 2014).

2.1.1.8 Inuk

El individuo más reciente (Inuk) es un Paleoesquimal de la cultura Saqqaq (cazador-recolector) con una antigüedad de 3,885±285 años calibrados A.P. Para la secuenciación de su genoma se tomó muestra del pelo encontrado en Qeqertasussuk en Groenlandia (Figura 4.B). Su haplogrupo del cromosoma Y es el Q1a, comúnmente encontrado entre siberianos y poblaciones de nativos americanos, mientras que el DNA mitocondrial muestra una cercana relación a Aleuts en las islas comandante (situado en el mar de Bering) y de Siberianos Sireniki Yupik (esquimales asiáticos) (Rasmussen et al., 2010).

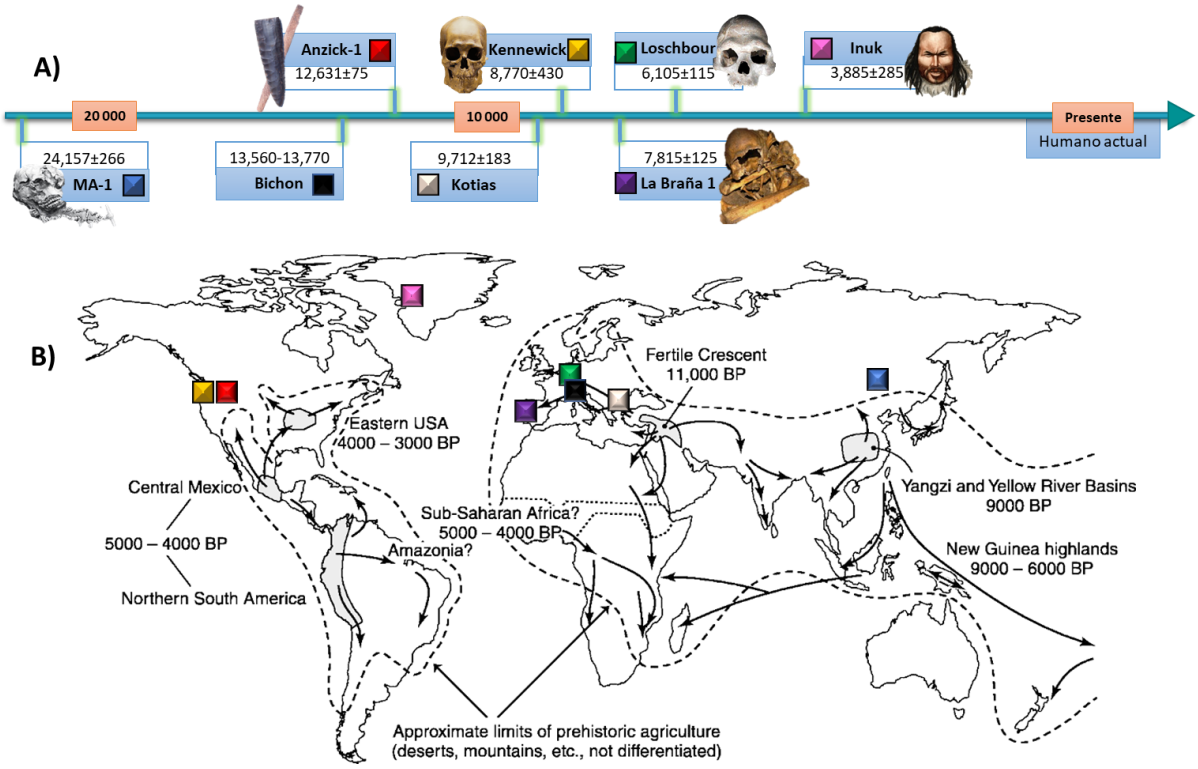


Figura 4. Ubicación y temporalidad de los humanos antiguos. A) Línea de tiempo en la que se ubica cada individuo. b) Mapa arqueológico de las tierras agrícolas, extensiones de las culturas neolíticas con fechas aproximadas de radiocarbono y ubicación aproximada de cada humano antiguo (Modificado de Diamond & Bellwood, 2003).

2.1.2 Población actual

2.1.2.1 Proyecto de los 1000 genomas

Para el estudio comparativo se descargaron los SNP reportados del proyecto de los 1,000 Genomas, que consta de 2,504 genomas actuales de 26 poblaciones, siendo un total de 5 grupos continentales, como se muestra en la Figura 5. Esta base de datos se seleccionó por ser la más completa al proporcionar un recurso integral sobre la variación genética humana, la asignación altamente

precisa de los genotipos en cada muestra, la detección eficiente de la mayoría de las variantes, el uso de secuenciación de nueva generación, los análisis del proyecto incorporan múltiples estrategias de análisis y además el mapeo de los genomas se llevó acabo usando el genoma de referencia más actual, hg38 (1000 Genomes Project Consortium, 2015; Sudmant et al., 2015).

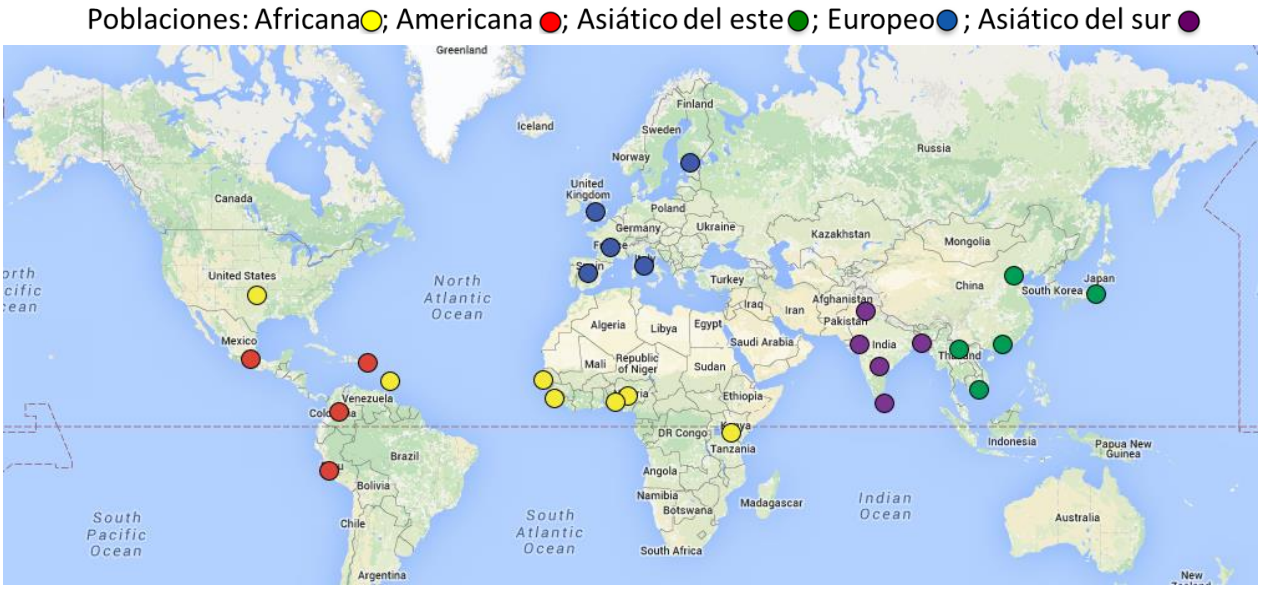


Figura 5. Ubicación de los grupos continentales del proyecto de los 1000 genomas. Cada color representa un grupo continental.

2.1.2.2 Genoma de referencia hg38/GRCh38

Para realizar la homogenización de los genomas antiguos y la comparación entre poblaciones se descargó el genoma de referencia hg38 (UCSC ID) o GRCh38 (ID de ensamblado), el cual es el genoma actual humano (*Homo sapiens*) ensamblado más reciente (diciembre de 2013), producido por el *Genome Reference Consortium* (conformado por NCBI, EMBL-EBI, *Sanger Institute*, y *Washington University*). Este genoma presenta un tamaño de secuencia total de 3,209,286,105 nucleótidos (Rosenbloom et al., 2014).

2.1.3 Mapeo de los genomas antiguos

En el caso de los individuos La Braña-1, Kotias y Bichon, el mapeo realizado por los respectivos autores no se encontró disponible, por ello se descargaron los archivos FASTQ (formato de archivo común para compartir datos de lecturas de secuencias que combinan tanto la secuencia como un puntaje de calidad por base [Cock et al., 2009]), para poder realizar el alineamiento con el genoma

de referencia hg38 y obtener el archivo BAM (Binary Alignment/Map). El archivo BAM es un formato binario y comprimido de un documento SAM, y el documento SAM es un formato de alineamiento genérico para el almacenamiento de reads alineados contra un genoma de referencia (Zhao et al., 2013).

Para esta metodología se utilizó el software BWA, el cual mapea secuencias de baja divergencia contra un genoma de referencia (Liu, Hsiao, & Dai, 2015). Para este estudio se estableció una calidad de mapeo mínima de 30 (escala phred), es decir que existe una probabilidad de error del 0.001.

2.1.4 Homogenización de los datos

Cada uno de los humanos antiguos se encuentra mapeado, ya sea con el genoma de referencia hg18, o con hg19 (como se muestra en la Tabla 2); para ello se tuvo que actualizar y homogenizar al genoma de referencia hg38. Para realizar este cambio de referencia, se utilizó la herramienta CrossMap, una herramienta para convertir coordenadas genómicas o archivos de anotación entre ensambles (Zhao et al., 2013). Este programa solo necesita el archivo BAM y el nuevo genoma de referencia.

2.1.5 SNP *calling*

Los genomas de humanos antiguos se analizaron con dos programas, SAMtools y BCFtools, con la finalidad de obtener los polimorfismos de un solo nucleótido (SNP) que difieren respecto al genoma de referencia hg38 (proceso conocido en inglés como *SNP calling*); todo esto utilizando un algoritmo que brinda evidencia estadística de la existencia del polimorfismo, como se muestra en la Figura 6. SAMtools es un paquete de librerías y software para el análisis y manipulación de alineamientos en los formatos SAM/BAM (Li et al., 2009). BCFtools es un set de utilidades que manipula variantes en VCF (*Variant Call Format*) y su contraparte binaria BCF (*Binary Call Format*) (Li et al., 2018).

El procedimiento consiste en:

1. Ordenado de los archivos bam (samtools sort file.bam file_sorted)
2. Generar el indexado del genoma de referencia (samtools faidx refernce.fasta)
   Donde faidx es para indexar/extraer el documento FASTA

3. Generar el archivo bcf (samtools mpileup – C 50 -g -f reference.fasta file_sorted > file.bcf)
   Donde el parámetro mpileup es el comando para calcular las probabilidades del genotipo, -C para reducir los efectos de reads con excesivos mismatches (el valor recomendado es de 50), -f es para indexar el archivo del genoma de referencia en formato FASTA y -g es para calcular las probabilidades de genotipo y dejarlas en formato BCF.
4. Generar el archivo vcf (bcftools call –mv file.bcf -o out.vcf)
   Donde "call" es para el llamado de los SNP o de inserciones/deleciones (indeles), -m es el modelo alternativo para llamado de variantes multialélicas y raras que se limitan con el comando -c, -v salida de solo sitios variantes y -o archivo salida
5. Filtrado y obtención de los SNP (vcftools –vcf file.vcf --minQ 20 --recode)
   Donde --minQ incluye solo los sitios con una calidad mínima de 20 y por encima de este y la opción –recode genera un nuevo archivo en formato ya sea VCF o BCF.

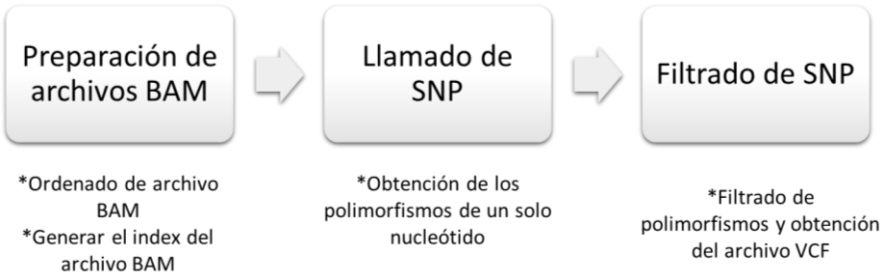


Figura 6. Metodología para la obtención y filtrado de polimorfismos.

Uno de los parámetros importantes para este análisis es la calidad del SNP (Q score), que está basado en la escala "phred" para cada nucleótido. Los puntajes de calidad de más de 20 son generalmente considerados como aceptables (Zeng et al., 2017), como se puede observar en la Figura 7, donde al llegar al valor de 20 de la escala phred, el error baja drásticamente, mientras que la exactitud aumenta.

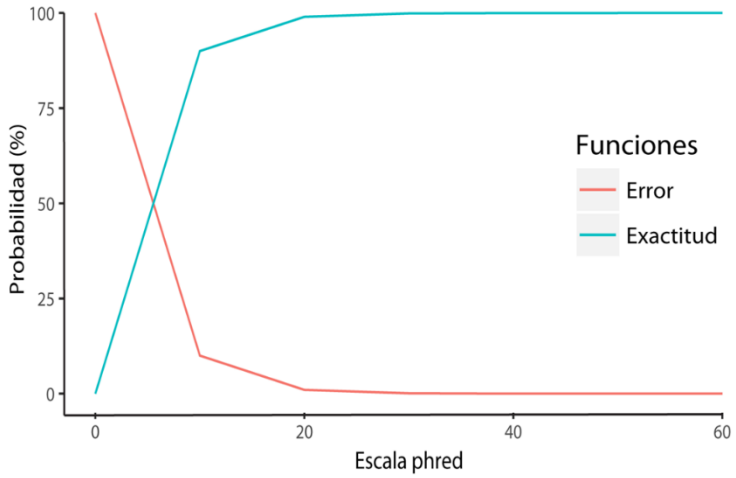


Figura 7. Conducta del error (color salmón) y la exactitud en la escala phred (color cian).

2.1.6 Alineamiento de los polimorfismos y cálculo de los estimadores

Con los pasos previos se obtuvieron ocho documentos VCF, que contienen los SNP de cada individuo antiguo. Previo al análisis es necesario realizar el alineamiento múltiple de secuencias por cada uno de los SNP obtenidos. En la práctica común, en el análisis se escoge un alelo (si es que existe más de uno), seleccionando así solo uno de los genotipos. Para este estudio se realizó el llamado de genotipos, pero el análisis no se realizó en base a ellos. Cada uno de los individuos se dividió en dos y se hizo la comparación por pares de los alelos, teniendo así 16 secuencias que correspondía a los 8 humanos antiguos.

Para esto se diseñó un programa basado en el lenguaje perl, para realizar dicho alineamiento por los sitios nucleotídicos homólogos. Teniendo este archivo se prosiguió, con otro programa diseñado también en base al lenguaje Perl, a calcular masivamente, por ventanas de 100 nucleótidos, la diversidad nucleotídica (π), θ y la *D* de Tajima.

2.2.6.1 Diversidad nucleotídica

La diversidad nucleotídica es calculada por la sumatoria de los sitios nucleotídicos que difieren entre cada par único de las secuencias de DNA, como se muestra en la fórmula 2.1.

$$\pi = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{n} \sum_{j>i}^{n} d_{ij} \qquad (2.1)$$

Donde:

- ***i*** y ***j*** son índices que hacen referencia a secuencias de DNA individuales.
- $\boldsymbol{d_{ij}}$ es el número de sitios nucleotídicos que difieren entre las secuencias de i y j.
- ***n*** es el número total de secuencias de DNA en la muestra.

El número de comparaciones por pares únicos en una muestra de *n* secuencias es (*n* (*n* - 1)) / 2 y así dividiendo la suma de d_{ij} por este número da el número promedio de las diferencias por cada par de secuencias (Hamilton, 2009).

2.2.6.2 Theta (θ)

Theta se calcula de la siguiente manera de:

$$\theta = 4N_e\mu \qquad (2.2)$$

Donde:

- $\boldsymbol{N_e}$ es el tamaño efectivo poblacional
- $\boldsymbol{\mu}$ es la tasa de mutación

Sin embargo, es difícil tener la determinación exacta de los parámetros N_e y μ, por lo que una manera indirecta de estimar θ es utilizando el número total de sitios segregativos en un grupo de secuencias (Eguiarte, 2007). Definiendo primero:

$$a_1 = \sum_{k=1}^{n-1} \frac{1}{k} \qquad (2.3)$$

Donde:

- $\boldsymbol{k}$ es igual al número de secuencias en la muestra.

Entonces θ, la cual se puede simbolizar como $\hat{\theta}_w$ (W por Watterson) o $\hat{\theta}_s$ (S por sitios segregantes) se calcula con la fórmula 2.4 (Hamilton, 2009).

$$\hat{\theta}_s = \frac{p_s}{a_1} \qquad (2.4)$$

La importancia de esta ecuación final (formula 2.4) es que $4N_e\mu$ puede ser estimada por el número de sitios segregantes (P_s) y el número de secuencias en una muestra.

2.2.6.3 Prueba de Tajima

La estadística de la *D* de Tajima se calcula a partir de la diferencia entre $\hat{\theta}_\pi$ y $\hat{\theta}_s$ dividida por la desviación estándar, como se muestra en la fórmula 1.6 (Hamilton, 2009).

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_s}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_s)}} = \frac{\hat{\theta}_\pi - \frac{p_s}{a_1}}{\sqrt{e_1 p_s + e_2 p_s(p_s - 1)}} \qquad (2.5)$$

Dónde:

- $\hat{\theta}_\pi$ es la estimación de θ basado en la diversidad nucleotídica
- $\hat{\theta}_s$ es la estimación de θ basado en los sitios segregativos

Las fórmulas usadas para calcular la varianza son:

$$e_1 = \frac{n+1}{3a_1(n-1)} - \frac{1}{a_1^2} \qquad (2.6)$$

Y

$$e_2 = \frac{c}{a_1^2 + a_2} \qquad (2.7)$$

Donde:

$$a_1 = \sum_{k=1}^{n-1} \frac{1}{k} \qquad (2.8)$$

$$a_2 = \sum_{k=1}^{n-1} \frac{1}{k^2} \qquad (2.9)$$

$$c = \frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2} \qquad (2.10)$$

En las que ***n*** es el número de secuencias muestreadas, asumiendo que no hay recombinación.

2.1.7 Identificación de ventanas con selección direccional

Para la identificación de las regiones bajo selección se realizaron dos metodologías, la primera es por la diferencia de la diversidad y la otra sobre las diferencias en las frecuencias nucleotídicas, que a continuación se describen.

2.1.7.1 Diferencias de la diversidad genética

Para esta estrategia, con ayuda de un programa en lenguaje perl, se obtuvo el valor absoluto de la diferencia entre el valor de cada estimador (previamente calculado por ventanas de 100 nucleótidos) de la población ancestral contra la actual. En la Formula 2.11 se muestra como ejemplo la manera para el calcular las diferencias de π, denotadas como $\Delta\pi$. Con este programa se genera un archivo de salida que presenta el cromosoma, la posición inicial de la venta, la posición final de la ventana, los SNP que se encontraron en dicha ventana y los valores de las diferencias absolutas de π, θ y D de Tajima.

$$\Delta\pi = | \pi_{antiguo} - \pi_{actual} | \quad (2.11)$$

Teniendo este archivo se seleccionaron las ventanas con los valores extremos de la distribución (*outliers*), siendo estas las regiones candidatas a estar bajo selección direccional. Con estas ventanas se procedió a buscar los SNP que generaban dicha variación en cada una de ellas, para clasificarlos posteriormente.

2.1.7.2 Diferencia entre frecuencias nucleotídicas

Con esta estrategia se buscan las regiones con nula o baja diversidad, pero con diferencias en las frecuencias de nucleótidos entre poblaciones. Para llevar a cabo esta metodología, se toma el archivo del alineamiento múltiple y se analiza por posición homóloga las frecuencias entre nucleótidos de cada población, como se ejemplifica en la Tabla 3.

Tabla 3. Ejemplo de regiones con diversidad nula o muy baja,
pero con diferencias en la frecuencia nucleotídica entre poblaciones.

Ejemplo de región con diversidad nula												
	Población Ancestral						**Población Actual**					
Humano 1	A	A	T	G	T		A	C	T	G	T	
Humano 2	A	A	T	G	T	**Diversidad =0**	A	C	T	G	T	**Diversidad=0**
Humano 3	A	A	T	G	T		A	C	T	G	T	
Humano 4	A	A	T	G	T	Frecuencia= 1	A	C	T	G	T	Frecuencia= 1
Posición	**1**	**2**	**3**	**4**	**5**	Respecto a A	**1**	**2**	**3**	**4**	**5**	Respecto a C
Ejemplo de región con baja diversidad												
Humano 1	A	A	G	G	T		A	C	G	G	T	
Humano 2	A	A	T	G	T	**Diversidad= 0.075**	A	C	G	G	T	**Diversidad=0**
Humano 3	A	A	T	G	T		A	C	G	G	T	
Humano 4	A	A	T	G	T	Frecuencia= 0.75	A	C	G	G	T	Frecuencia= 1
Posición	**1**	**2**	**3**	**4**	**5**	Respecto a T	**1**	**2**	**3**	**4**	**5**	Respecto a G

Para buscar estos casos se generó otro programa en lenguaje perl, que compara por posición homologa las frecuencias de cada nucleótido por población, seleccionado el nucleótido con mayor frecuencia para asignarle un valor a cada posición (siendo el valor igual a 1 cuando todas las muestras presentan un mismo nucleótido), para luego hacer la comparación entre poblaciones. Con esta metodología se pretende identificar las posiciones donde las poblaciones difieran respecto al nucleótido de mayor frecuencia y que la población actual conserve el nucleótido del genoma de referencia hg38. El límite de corte para este caso fue hasta el valor de frecuencia mínima de 0.55 (en humanos antiguos significa que, al menos, más de 4 individuos presentan el nucleótido de mayor frecuencia).

2.1.8 Anotación de las regiones

Con las metodologías antes descritas se identificaron diversos SNP, siendo el siguiente paso la anotación y clasificación de estos. Para poder hacer la anotación se usó *Variant Effect Predictor* (VEP), que es un conjunto de herramientas para el análisis, la anotación y la priorización de variantes genómicas en regiones codificantes y no codificantes, es de código abierto, de uso gratuito, y es compatible con la reproducibilidad completa de los resultados. VEP anota variantes de secuencia con cambios específicos y bien definidos, incluidos variantes únicas nucleotídicas, inserciones, deleciones, sustituciones de pares de bases múltiples, microsatélites y repeticiones en tándem (McLaren et al., 2016).

2.1.8.1 Documento de entrada para VEP

Se usó la interfaz web que necesita un formato específico para el documento de entrada y así poder hacer el análisis de los datos. Antes de su uso se generó dicho documento de entrada con los SNP encontrados por las metodologías anteriormente descritas. El formato del archivo es simple, separado por espacios (las columnas pueden estar separadas por espacios o caracteres de pestañas, como se muestra en la Tabla 4), que contiene cinco columnas requeridas más una columna de identificador opcional, como se muestra a continuación:

1.- Cromosoma: Solo el nombre o número, sin el prefijo 'chr'.
2.- Comienzo: Ubicación inicial de la región a analizar.
3.- Fin: Ubicación final de la región a analizar.
4.- Alelo: Por par, los alelos separados por una barra (/), con el alelo de referencia primero.
5.- Hebra: Definida como + (forward) o - (reverse).
6.- Identificador: Este identificador se utilizará en la salida de VEP. Si no se proporciona, VEP construirá un identificador a partir de las coordenadas y alelos proporcionados.

Tabla 4. Ejemplo del formato para la interfaz web de VEP (sin identificador expreso)

1	881907	881906	- / C	+
5	140532	140532	T / C	+
12	1017956	1017956	T / A	-

Teniendo el documento de entrada se procedió al uso de la interfaz gráfica de VEP (Disponible en https://www.ensembl.org/Tools/VEP). A continuación, se describen los pasos:

I. Selección del organismo: Human (*Homo sapiens*).
II. Elegir un nombre para los datos que se cargan: Identificación de trabajos y archivos cargados en VEP (opcional).
III. Carga de archivos: Elegir archivo en el sistema.
IV. Selección del símbolo genético: Proporcionar el gen en la salida, en este caso el identificador de HGNC para genes en humanos.
V. Selección de Predicciones de SIFT: SIFT predice si la sustitución de un aminoácido afecta la función de la proteína.
VI. Selección de predicciones de PolyPhen: PolyPhen predice el posible impacto de una sustitución de aminoácidos en la estructura y función.
VII. Ejecutar.

2.1.8.2 Documento de salida para VEP

El formato de salida predeterminado desde la interfaz web es un archivo delimitado por tabulaciones. Los valores vacíos se denotan con '-'. Las columnas de salida son varias, pero para este trabajo solo se seleccionaron las siguientes:

- ✓ Variación cargada: Cromosoma, posición y alelo.
- ✓ Alelo: Variante utilizada para calcular la consecuencia.
- ✓ Consecuencia: Tipo de consecuencia que genera esta variante (Tabla 5).
- ✓ Símbolo: El símbolo del gen.
- ✓ Gen: Identificador (ID) de Ensembl.
- ✓ Característica: Identificación estable (notación Ensembl).
- ✓ Exón: El número de exón (del número total).
- ✓ Posición en cDNA: Posición relativa del par de bases en la secuencia de cDNA.
- ✓ Posición en CDS: Posición relativa del par de bases en la secuencia de codificación.
- ✓ Posición en la proteína: Posición relativa del aminoácido en la proteína.
- ✓ Cambio de aminoácidos: Solo si la variante afecta a la secuencia de codificación de proteínas.
- ✓ Cambio de codón: Los codones alternativos con la base variante en mayúscula.
- ✓ Hebra: La cadena de DNA (1 o -1) en la que se basa la transcripción/característica.
- ✓ SIFT: Predicción y/o puntaje SIFT, con ambos dados como predicción (puntaje).
- ✓ PolyPhen: Predicción y/o puntaje de PolyPhen.

2.1.8.2.1 Anotación de las variantes por VEP

En el archivo de salida se muestra la consecuencia del alelo analizado, descritas en la Tabla 5. Las consecuencias son un conjunto de términos definidos por la *Sequence Ontology* (SO), que se puede asignar a cada combinación de un alelo y una transcripción, donde cada alelo puede tener un efecto

diferente en diferentes transcripciones y cada una de ellas una gravedad asignada (impacto) (Kumar et al., 2009; Li et al., 2009; Zerbino et al., 2017).

Tabla 5. Consecuencias que asigna VEP a cada variante analizada con su descripción e impacto.

Termino SO	Descripción SO	Variante	Impacto
3 prime UTR variant	Una variante del 3' UTR (del inglés *untranslated región*, es decir regiones no traducidas de los genes).	Variante principal 3 UTR	Modificador
5 prime UTR variant	Una variante del 5' UTR.	Variante principal 5 UTR	Modificador
Downstream gene variant	Variante ubicada en 3' de un gen.	Variante genética rio abajo	Modificador
Intergenic variant	Variante en la región intergénica (entre genes).	Variante intergénica	Modificador
Intron variant	Variante de transcripción que ocurre dentro de un intrón.	Variante de intrón	Modificador
Missense variant	Variante que cambia una o más bases, dando como resultado una secuencia de aminoácidos diferente, pero donde se preserva la longitud.	Variante sin sentido	Moderada
NMD transcript variant	Variante en una transcripción que es el objetivo de NMD (del inglés *nonsense- mediated mRNA decay*, que se encarga de la degradación de RNA mensajeros mediada por codones de este tipo).	Variante de transcripción NMD	Modificador
Non-coding transcript exon variant	Variante de secuencia que cambia la secuencia de exón no codificante en una transcripción no codificante.	Variante de exón de transcripción no codificante	Modificador
Non-coding transcript variant	Una variante de transcripción de un gen de ARN no codificante.	Variante de transcripción no codificada	Modificador
Splice donor variant	Variante de empalme que cambia la región de 2 bases en el extremo 5´ de un intrón.	Variante donante de empalme	Alto
Splice region variant	Una variante de secuencia en la que se ha producido un cambio dentro de la región del sitio de empalme ya sea dentro de 1-3 bases del exón o 3-8 bases del intrón.	Variante de la región de empalme	Bajo
Stop gained	Variante mediante la cual se cambia al menos una base de un codón, lo que da como resultado un codón de detención prematuro, lo que conduce a una transcripción acortada.	Codón de paro ganado	Alto
Synonymous variant	Variante de secuencia donde no hay cambio resultante en el aminoácido codificado.	Variante sinónima	Bajo
Upstream gene variant	Variante de transcripción de un gen de ARN no codificante.	Variante genética rio arriba	Modificador

La clasificación subjetiva de la gravedad de la variante presenta cuatro categorías, donde:

I. Alto: La variante tiene un alto impacto (disruptivo) en la proteína, lo que probablemente causa el truncamiento de la proteína, la pérdida de función o la activación mediada por la decadencia sin sentido.
II. Moderada: Una variante no disruptiva que podría cambiar la efectividad de la proteína.
III. Bajo: Se supone que es inofensivo o poco probable que cambie el comportamiento de las proteínas.
IV. Modificador: Usualmente variante o variantes no codificantes que afectan genes no codificadores, donde las predicciones son difíciles o no hay evidencia de impacto.

Para el fácil entendimiento del sitio génico en el que se encuentra cada una de las variantes se muestra la Figura 8, para una visualización relativa de la ubicación de cada una de ellas.

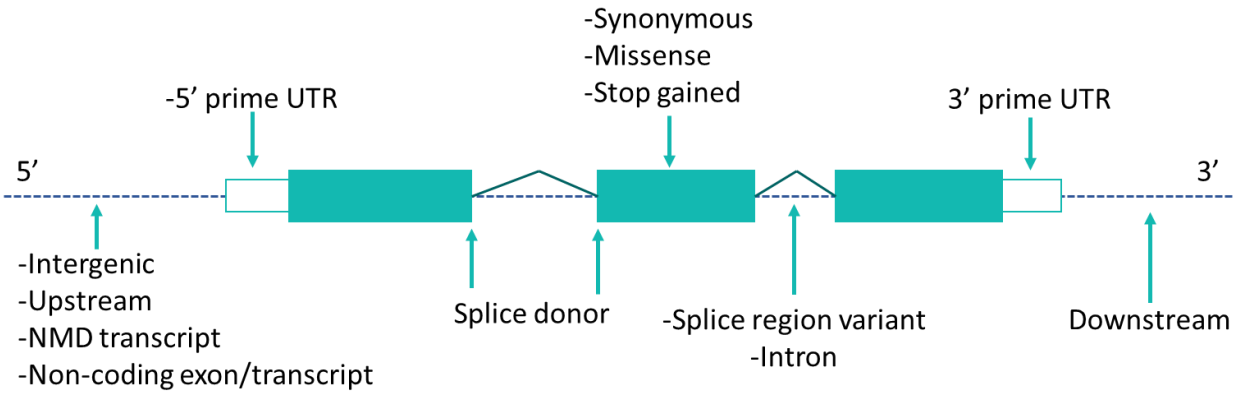


Figura 8. Diagrama de visualización relativa a la estructura de la transcripción de cada consecuencia (Disponible en https://www.ensembl.org/Tools/VEP).

2.1.8.2.2 Impacto de la sustitución de aminoácidos

Existen dos valores que reporta VEP, uno de ellos es **SIFT** el cual predice si es probable que la sustitución de un aminoácido afecte a la función de la proteína basándose en la homología de secuencia y la similitud fisicoquímica entre los aminoácidos alternativos. Proporciona para cada sustitución de aminoácidos una puntuación y una predicción cualitativa (ya sea 'tolerada' o 'perjudicial'). La puntuación es la probabilidad normalizada de que se tolere el cambio de aminoácidos, por lo que los puntajes más cercanos a cero tienen más probabilidades de ser nocivos. La predicción cualitativa se deriva de este puntaje de modo que las sustituciones con un puntaje <0.05 se denominan "deletéreas" y todas las demás se denominan "toleradas" (Kumar et al., 2009; McLaren et al., 2016).

El otro valor que genera es el PolyPhen que predice el efecto de una sustitución de aminoácido en la estructura y función de una proteína utilizando la homología de secuencia, las anotaciones Pfam, estructuras 3D de *Protein Data Bank*, y una serie de otras bases de datos y herramientas. Para cada sustitución de aminoácido proporciona tanto una predicción cualitativa (probablemente perjudicial, posiblemente perjudicial, benigna o desconocido) y una puntuación. La puntuación de PolyPhen representa la probabilidad de que una sustitución sea perjudicial, por lo que los valores más cercanos a 1 se predicen con mayor confianza como nocivos (Adzhubei et al., 2010; McLaren et al., 2016).

2.2 Regiones codificantes

Caracterizadas las regiones se analizó la frecuencia de cada una de ellas y solo se seleccionaron las variaciones en regiones codificantes (*missense* y *stop gained*) para identificar los genes a que corresponden, cuáles tienen un cambio diferencial en el aminoácido, cuáles un impacto en la proteína, cuántos de ellos presentan un gen ortólogo en *S. cerevisiae* y cuáles son fuertes candidatos para próximos análisis experimentales, como se describe en la Figura 9.

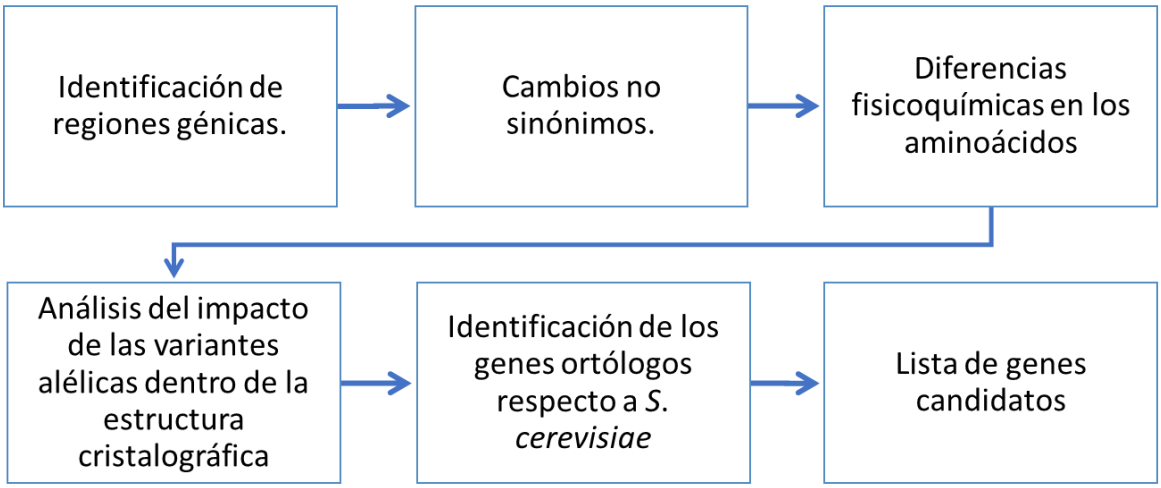


Figura 9.- Metodología para la selección de genes candidatos.

2.2.1 Clasificación de genes

Los genes asociados a un SNP de interés (genes candidatos), se clasificaron por los procesos biológicos en los que interactúan y así conocer su función a un nivel célula u organismo. Se optó por usar PANTHER (*Protein ANalysis THrough Evolutionary Relationships*), base de datos para clasificar proteínas por función, en la que se puede conocer la familia y subfamilia, la función molecular (la función de la proteína por sí misma o con proteínas que interactúan directamente a un nivel bioquímico), el proceso biológico (la función de la proteína en el contexto de una red más grande de proteínas que interactúan para lograr un proceso a nivel de la célula u organismo), la vía y la ubicación celular (Thomas et al., 2003).

2.2.2 Cambio en las propiedades fisicoquímicas del aminoácido

Teniendo los SNP que generan un cambio no sinónimo, se procedió a analizar las modificaciones respecto a las propiedades fisicoquímicas de los aminoácidos, para ello se tomó el código genético estándar de Budisa (2006), y las propiedades de cada aminoácido que reporta Masenko (2018), tomando solamente:

- Nombre del aminoácido.
- Estructura de la cadena lateral.
- Carga.
- Contenido de sulfuro.
- Peso molecular (masa molar g/mol).
- Formula estructural.

2.2.3 Estructura cristalográfica

Para ver las interacciones o impacto entre los polimorfismos encontrados con los sitios importantes para el funcionamiento de la proteína (sitios de unión, catalíticos, etc.), es de suma importancia que estén reportados dichos sitios (usando la base de datos UniProt (Consortium, 2016)) y que la proteína esté cristalizada. *The Protein Data Bank* es un banco de datos estructurales de macromoléculas biológicas (Berman et al., 2000), que fue consultada para la búsqueda de archivos PDB (archivo para almacenar estructuras de *Protein Data Bank*), de un gen o genes de interés.

En caso de no existir la proteína cristalizada, se pueden buscar proteínas cristalizadas con alto porcentaje de identidad de secuencia, incluyendo las de otros organismos, para tomarlas como plantilla o templado, y generar el modelado de la estructura con ayuda de SWISS-MODEL. SWISS-MODEL es un servidor automatizado para el modelado de proteínas, el cual depende de un complejo homólogo que usa como templado para el modelamiento de la proteína de interés (Waterhouse et al., 2018).

2.2.4 Genes ortólogos

Para conocer cuáles de los genes candidatos pueden ser clonados en el organismo modelo *S. cerevisiae*, se descargó la base de datos de ENSEMBL (Zerbino et al., 2017) tomando:

- ID Ensembl.
- Hebra.
- Nombre del gen.
- Descripción.
- Gen ortólogo del organismo modelo.
- Porcentaje de identidad del gen del modelo respecto al consultado.
- Porcentaje de identidad del gen consultado respecto al gen del modelo.

2.3 Complementación heteróloga en *Saccharomyces cerevisiae*

Para contribuir a elucidar si estos cambios fueron producidos por selección y no por deriva, se propone la implementación del modelo de expresión heteróloga en *S. cerevisiae.* Este modelo ha demostrado ser útil para detectar diferencias funcionales entre alelos humanos (Hamza et al., 2015; Kachroo et al., 2015; Sun et al., 2016). Por este motivo, se usará para identificar, si es que existen, diferencias entre el alelo humano antiguo y actual. En trabajos previos de búsqueda de selección en el genoma humano del grupo de Interacción Núcleo Mitocondrial y Paleogenómica, se encontró como candidato al gen adenilato quinasa 2 (AK2), siendo una enzima ubicua que cataliza la interconversión de tres nucleótidos de adenina en la célula: Mg_2 + ATP + AMP = Mg_2 + 2ADP, y se cree que contribuye a la homeostasis de la composición de nucleótidos de adenina en la célula (Noma et al., 1998). Además de contar con las características adecuadas para realizar el ensayo de complementación en *S. cerevisiae*, la ventana previamente identificada bajo selección de este gen se analizará en este nuevo estudio para corroborar su firma de selección con los tres estimadores propuestos.

Se propone el estudio de la proteína AK2 (ENSG00000004455) por ser un buen candidato para análisis de diferencias funcionales entre alelos, ya que presenta un tamaño de 239 aminoácidos (transcrito: ENST00000354858); se encontraron dos SNP con cambio no sinónimo, los cuales presentan diferencias en las propiedades fisicoquímicas de los aminoácidos. Los cambios encontrados fueron en el aminoácido 181, siendo el alelo ancestral el que presenta asparagina (neutro y polar cargado) y el alelo actual posee lisina (básico fuerte y polar cargado), y en el

aminoácido 191, presentado en el alelo ancestral ácido aspártico (polar cargado y ácido fuerte) mientras que el alelo actual tiene histidina (básico débil y polar cargado). Estos cambios fueron encontrados entre un sitio de unión a Adenosín Monofosfato (aminoácido 186), como se muestra en el Anexo 1, observándose una reducción aproximada de 5.65 Å entre la distancia de los SNP, al comparar las variantes ancestrales contra la actual.

Hipotetizando que estos cambios afectaron la actividad de la proteína, teniendo como base que *S. cerevisiae* es un sistema genético para identificar la variación funcional de genes humanos (Sun et al., 2016) y que este gen presenta su ortólogo en este organismo modelo, se realizó el análisis de las diferencias funcionales del gen AK2 actual en comparación con su variante alélica ancestral por medio de la complementación heteróloga en dicho organismo y con ello establecer esta metodología en el Langebio (Figura 10).

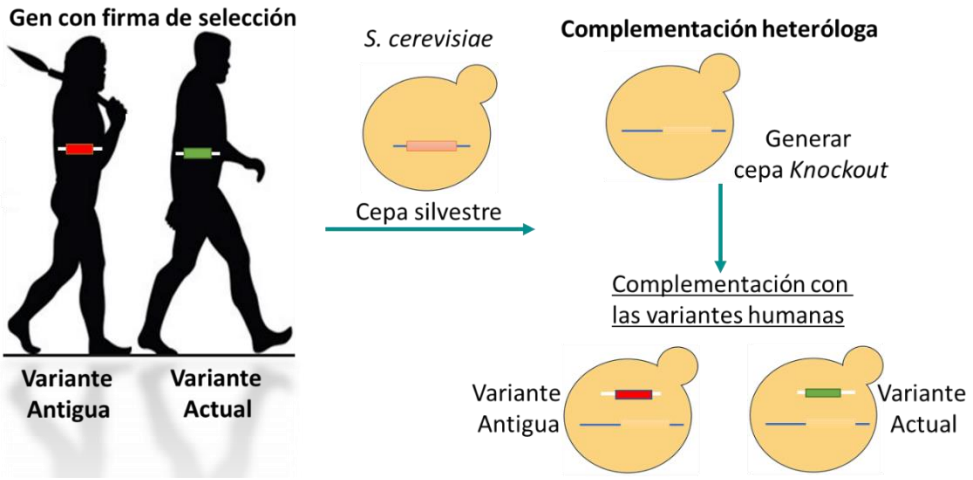


Figura 10.- Metodología propuesta para evaluar diferencias funcionales de variantes humanas en *S. cerevisiae*.

2.3.1 Cepas Knockout

AK2 presenta un gen ortólogo en levadura, *ADK1* (YDR226W), que presenta 54.9% de identidad en secuencia. *ADK1* a su vez, tiene un parálogo, *ADK2* (YER170W), que tiene un 35.1% de identidad en secuencia con *ADK1* y 36.8% respecto a AK2, según el análisis de matriz de identidad realizado con el software BioEdit (Hall et al., 2011). Se generaron las cepas *knockout* sencillas y la doble mutante de *S. cerevisiae* de estos genes como se describe a continuación.

2.3.1.1 Knockout sencillas

Para generar las cepas *knockout* se usó la cepa Y7092 (CFP-natR matα) y se utilizó la metodología de eliminación directa de genes según Baudin et al. (1993) y la metodología de transformación descrita por Gietz & Schiestl (2007), para deletar el gen de interés por medio de una recombinación homóloga dirigida por PCR (por las siglas en ingles de *polymerase chain reaction*). Se empleó el vector pFA6 como templado para el PCR, el cual presenta el cassette de resistencia al antibiótico geneticina (G418). Los oligos usados para eliminar los genes silvestres se muestran en la Tabla 6.

Tabla 6. Secuencias de deleción para los genes *ADK1* y *ADK2*.

Gen	Nombre	Secuencia 5' a 3'
YDR226W	ADK1_UP45	tttctctggtaaagtcaccacacagcatcaaatataacagtaatgccagctgaagcttcgtacgc
	ADK1_DN45	taaaaaaaagaaaagatatttagaagacattgcgcaaggtcattatcgatgaattcgagctcgtt
YER170W	ADK2_UP45	tgtaaatctttaaattttcaggaacagggttagcaagcatcaatgccagctgaagcttcgtacgc
	ADK2_DN45	gcgttaaatcaccctattagccgaatttgctggttataaggttcatcgatgaattcgagctcgtt

En la primera amplificación, las condiciones para generar los vectores (pFA6) con las regiones homólogas a los genes *ADK1* y *ADK2* fueron:

Tabla 7. Condiciones para la amplificación del vector.

	Temperatura °C	Tiempo	Ciclos
Primera desnaturalización	98	30 s	-
Desnaturalización	98	10 s	30
Alineamiento	60	30 s	
Extensión	72	45 s	
Extensión final	72	10 min	-

Para la selección de las cepas transformadas se usó el medio YPAD (el cual contiene 10g de extracto de levadura, 20g de peptona, 100ml de glucosa al 20%, 120mg de sulfato de adenina, 20g de agar y 900mL de agua), 1mL G418 y 1mL del antibiótico clonNAT. Para la confirmación de la transformación por medio de PCR, se usaron los oligos que se desglosan en la Tabla 8.

Tabla 8. Secuencias de confirmación de los genes *ADK1* y *ADK2*.

Gen	Nombre	SECUENCIA 5' A 3'
YDR226W	ADK1_primerA	gttgtgtcttcctgttttctctgtt
	ADK1_primerB	gaccattctaatggattctgagcta
	KanB	ctgcagcgaggagccgtaat
YDR226W	ADK2_primerA	cagcacttgaataagaaaataccgt
	ADK2_primerB	ttttaaaggctcatttgttttcttg
	KanB	ctgcagcgaggagccgtaat

Estos *primers* sirven para identificar las cepas que se transformaron exitosamente y las que aún mantienen el gen, como se muestra en la Figura 11. Con esta metodología se generaron las cepas knockout sencillas *Δadk1* y *Δadk2*.

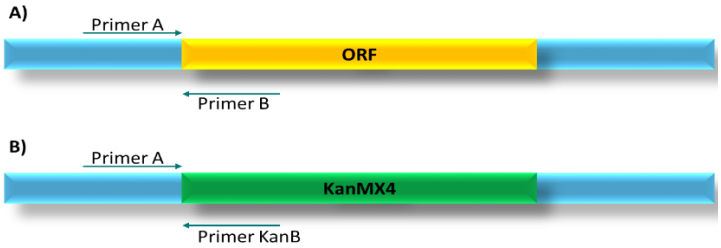


Figura 11.- Ubicación de los primers para la confirmación de la deleción. A) Sin deleción, el ORF (Open Reading Frame) sigue presente. B) Deleción exitosa, el cassette KanMx4 remplazo al ORF.

2.3.1.2 Doble knockout

Para generar la cepa doble knockout se usó la cepa *Δadk2*, el vector pAG32 (el cual contiene el cassette de resistencia al antibiótico higromicina), los *primers* para el gen YDR226W (Tabla 6) y se usó la misma metodología para generar las cepas knockout sencillas (eliminación por recombinación homologa dirigida por PCR). Para la selección de las cepas se empleó el medio YPAD con G418, clonNAT e higromicina. Para la confirmación se usaron los *primers* para el gen YDR226W, Tabla 8.

2.3.1.3 Fenotipificación

Teniendo las cepas knockout *Δadk1, Δadk2* y la doble knockout, se procedió a verificar su fenotipo y adecuación. Dentro de este ensayo se realizaron dos procedimientos que a continuación se describen.

2.3.1.3.1 Ensayo de Spots

Para este ensayo las colonias de interés se dejan crecer en medio liquido de YPAD a 30 ºC, en agitación constante por 48 hrs. Posteriormente se toma 1mL de medio saturado y se coloca en un tubo de 2 mL para centrifugar y remover el medio. Decantar el medio, agregar 1mL de agua estéril (para eliminar residuos del medio líquido), vortexear, centrifugar (*short*), eliminar el medio y repetir una vez más el lavado. Para medir densidad óptica a una onda de 600 (OD600), se agrega

de 1mL de agua estéril (aunque dependiendo de la saturación y viabilidad de las cepas, se puede agregar menos). Ajustar todas las cepas a OD600 de 1 (aproximadamente $1.89x10^7$ células), realizando los cálculos y las diluciones correspondientes. Teniendo los medios ajustados, se realizan diluciones de 10^{-1} a 10^{-5}. En este ensayo se usan dos medios sólidos, uno fermentable (YPAD) y un medio no fermentable (YPGE, el cual contiene extracto de levadura, peptona, etanol, agar, agua y glicerol, siendo este último la fuente de carbono). En las cajas Petri se agregaron 3µL de cada dilución, con ayuda de una pipeta multicanal, y se deja reposar hasta que se seque la gota inoculada. Incubar a 30 °C, para tomar las fotos a las 24 y 48 hrs. Este ensayo ayuda a contrastar la tasa de crecimiento celular en diferentes ambientes de crecimiento involucrando la dilución en serie.

2.3.1.3.2 Curvas de crecimiento

El otro procedimiento fue el análisis de curvas de crecimiento o adecuación, en diferentes medios como se muestra en la Tabla 9, donde se usaron tres replicas técnicas y tres replicas biológicas.

Tabla 9. Medios usados para el análisis de crecimiento.

Condición	Medio base	Concentraciones
Fuente de carbono: Ribosa	Sintético completo (SC)	2%
Fuente de carbono: Galactosa	SC	2%
Fuente de carbono: Glucosa	SC	2%
Estrés osmótico: Sorbitol	YPAD	1 M,1.2 M, 1.4 M y 1.6 M
Daño en la pared celular: SDS	YPAD	0.01%, 0.02%, 0.03% y 0.04%
Inhibidor de cAMP fosfodiesterasas: Cafeína	SC	0.1%, 0.15%, 0.2% y 0.25%
Estrés en el crecimiento celular: Etanol	YPAD	4%, 6%, 8% y 10%
Estrés Osmótico: CaCl2	YPAD	80mM, 100mM, 120mM y 140mM
Estrés oxidativo: Menadiona	YPAD	20mM, 30mM, 40mM y 50mM

Para este ensayo se inoculo medio liquido YPAD de cada una de las cepas de interés, para dejar a 30°C en agitación constante por 48 hrs y lograr saturación. Se utilizo una caja por condición, 2 líneas por concentración iniciando con la de menor a la de mayor concentración (como se muestra en la figura 12). Se usaron 150 µL de cada condición con 5 µL de medio saturado, inoculando con ayuda de la estación de trabajo robótica TECAN, ubicada en el LANGEBIO. Así mismo con ayuda de esta estación se midió OD por más de 48 hrs, para monitorear el crecimiento y obtención de datos para la generación de las curvas de crecimiento.

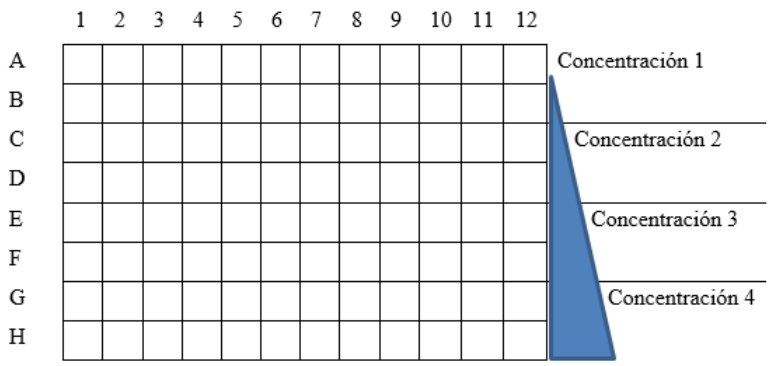


Figura 12.- Ejemplificación del ordenamiento una caja de 96 pozos para los ensayos de crecimiento.

2.3.2 Transformación y complementación

Para poder hacer los ensayos de complementación se usó el vector pCM189 (Gauthier et al., 2008). En los análisis hechos por Gauthier et al., (2008), con el gen humano adenilato quinasa 1 (parálogo al gen AK2) se observa que este gen humano puede complementar a la mutante de levadura *Δadk1*, a pesar de tener un 29% de identidad. Además, se menciona que se restauran los perfiles de HPLC (por sus siglas en inglés *high performance liquid chromatography*) cercanos a la de los niveles de tipo salvaje en los metabolitos estudiados, que fueron ATP, IMP, ADP, AMP, hipoxantina e inosina. Otra de las razones por las que se escogió el vector pcM189 fue por su sistema tetO7, el cual es un sistema reprimible de expresión génica con Doxiciclina y con ello poder ver diferentes niveles de expresión. Por esta razón este vector se usó para introducir las variantes alélicas de AK2 en las mutantes de levadura, como se describe más abajo. Como controles positivos se clonaron los genes *ADK1* y *ADK2* usando el mismo vector e introducirlos en las respectivas mutantes nulas. Estos genes fueron amplificados a partir de la extracción del DNA genómico de *S. cerevisiae* cepa S288C con ayuda de los *primers* que se reportan en la Tabla 10.

Tabla 10. *Primers* para la amplificación de los genes *ADK1* y *ADK2*.

Gen	Nombre	Secuencia 5' A 3'	Tamaño del amplicón (pb)
YDR226W	ADK1_FWD	cgcgcggccgcATGTCTAGCTCAGAATCCATT	669
	ADK1_REV	ccactgcagTTAATCCTTACCTAGCTTGTT	
YDR226W	ADK2_FWD	cgcgcggccgcATGAAAGCAGACGCGAAACAA	678
	ADK2_REV	cgcctgcagTCAATAATTTCGGAAGATAAT	

2.3.2.1 Genes humanos

Los genes AK2 humanos (la variante actual y la ancestral) fueron sintetizados y clonados en el vector pUC18 por GenScript (New Jersey, EUA). Las secuencias se muestran en el Anexo 2. Primeramente, se obtuvieron células competentes de *E. coli* (DH5α) con cloruro de rubidio (RbCl), para que se volvieran susceptibles a la incorporación de los vectores con los genes humanos y así poder tener células transformadas. El siguiente paso fue la obtención del material plasmídico usando el protocolo: *Plasmid DNA Purification* del *QIAprep Spin Miniprep Kit*. Teniendo cuantificados y purificados los plásmidos se procedió a obtener los genes por digestión enzimática usando las enzimas de restricción *Not*I y *Pst*I (Buffer 3.1 de NEB) por 37 °C por 3 horas y posteriormente a 65 °C para su inhibición. Se realizó la electroforesis en gel de agarosa al 0.7% para la identificación de las bandas de interés (bandas de 720 pb), para posteriormente cortarlas y purificarlas con el protocolo: *PureLink Quick Gel Extraction Kit*.

2.3.2.2 Vector pCM189

En la Figura 13 se muestra a grandes rasgos la metodología de la clonación de los genes de interés y los controles positivos en el vector pCM189.

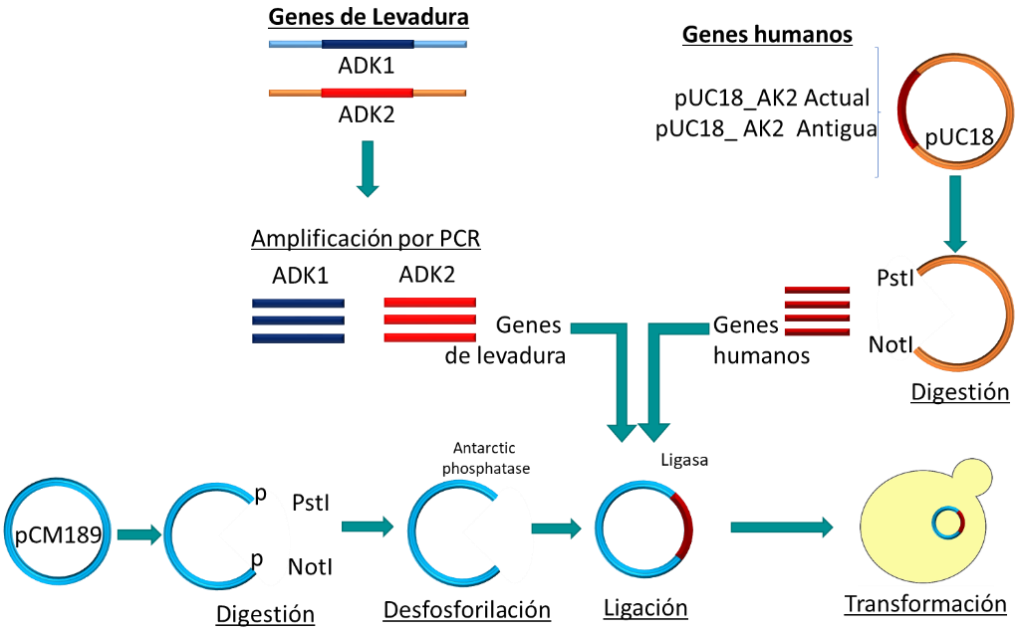


Figura 13.- Metodología para clonación de los genes humanos y de levadura en el vector pCM189.

Teniendo los genes aislados se procedió a realizar la digestión del vector pCM189 con las enzimas de restricción *Not*I y *Pst*I. Consecutivamente se eliminaron los grupos fosfato con ayuda de la enzima *Antarctic Phosphatase* para evitar la recircularización del vector y posteriormente se realizó la ligación (ligasa T4 DNA Ligase de *New England Biolabs*) con los genes (ligación a 16 °C toda la noche, con una inactivación a 65°c por 10 min), generando los vectores:

- ✓ pCM189 ADK1
- ✓ pCM189 ADK2
- ✓ pCM189 AK2 variante antigua
- ✓ pCM189 AK2 variante actual

Teniendo los vectores se procedió a transformar *E. coli* con cada uno de ellos, seleccionando las colonias con medio LB con ampicilina (pCM189 confiere dicha resistencia a *E. coli*) y se confirmó con los *primers* que se presentan en la Tabla 11.

Tabla 11. *Primers* para la confirmación del vector con los diferentes genes.

Confirmación de:	Nombre	Secuencia 5' A 3'	Tamaño del amplicón (pb)
pCM189 ADK1	pcM189_FWD	cctgtaggtcaggttgctttc	615
	ADK1_REV	ccactgcagTTAATCCTTACCTAGCTTGTT	
	pcM189_FWD	cctgtaggtcaggttgctttc	1323
	pcM189_REV	TCGGTTAGAGCGGATGTG	
pCM189 ADK2	pcM189_FWD	cctgtaggtcaggttgctttc	118
	ADK2_REV	cgcctgcagTCAATAATTTCGGAAGATAAT	
	pcM189_FWD	cctgtaggtcaggttgctttc	1332
	pcM189_REV	TCGGTTAGAGCGGATGTG	
pCM189 AK2 variante antigua	pcM189_FWD	cctgtaggtcaggttgctttc	724
	AK2_Human_REV	TGAGCCAGAAGCCACCATG	
	pcM189_FWD	cctgtaggtcaggttgctttc	1373
	pcM189_REV	TCGGTTAGAGCGGATGTG	
pCM189 AK2 variante actual	pcM189_FWD	cctgtaggtcaggttgctttc	724
	AK2_Human_REV	TGAGCCAGAAGCCACCATG	
	pcM189_FWD	cctgtaggtcaggttgctttc	1373
	pcM189_REV	TCGGTTAGAGCGGATGTG	

Para la confirmación se usaron las condiciones que reporta la Tabla 12, donde se aplicó el promedio de la temperatura de alineamiento de los oligos.

Tabla 12. Temperatura para la amplificación del vector.

	Temperatura °C	Tiempo	Ciclos
Primera desnaturalización	98	30 s	-
Desnaturalización	98	10 s	30
Alineamiento	65	30 s	
Extensión	72	45 s	
Extensión final	72	10 min	-

Teniendo las células transformadas y confirmadas por PCR se procedió a purificar los vectores (*QIAprep Spin Miniprep Kit*), para la transformación de *S. cerevisiae* (metodología de Gietz & Schiestl, 2007), con el medio de selección SC menos uracilo (SC -Ura), G418 y clonNat; con ello se generaron las cepas que se muestran en el Anexo 3 para los análisis posteriores. Teniendo las cepas de *S. cerevisiae* transformadas se procedió a verificar su complementación, fenotipo y adecuación con el ensayo de spots (en medio YPAD y YPGE, siguiendo la metodología previamente descrita, 2.3.1.3.1 Ensayo de spots) y el análisis de curvas de crecimiento (en medio SC-Ura con y sin doxiciclina, con sus respectivos controles), realizando mediciones de OD como se describió previamente (2.3.1.3.2 Curvas de crecimiento).

Capítulo III. Resultados

3.1 Genomas antiguos

3.1.1 Calidad de mapeo

Para este análisis se estableció una calidad mínima de mapeo de 30, lo que implica que exista 1 read incorrectamente mapeado de cada 1000 (99.9% de probabilidad de acierto); parámetro usado para el remapeo de los genomas de La Braña-1, Kotias y Bichon, ya que de estos individuos solo se obtuvo el archivo FASTQ. Los demás genomas presentaban el archivo BAM, que, de acuerdo a sus respectivos autores, presentaban una calidad de mapeo igual a 30, siendo homogéneos los parámetros usados en todos los humanos antiguos usados. En la Figura 14 se graficó las frecuencias de las calidades de mapeo de los reads de cada uno de los humanos antiguos y el corte de dicha calidad de mapeo para el análisis (línea roja), en donde se observa que la mayoría de los reads con mayor frecuencia se encuentran por encima de la calidad de mapeo de 30 (círculos de color azul).

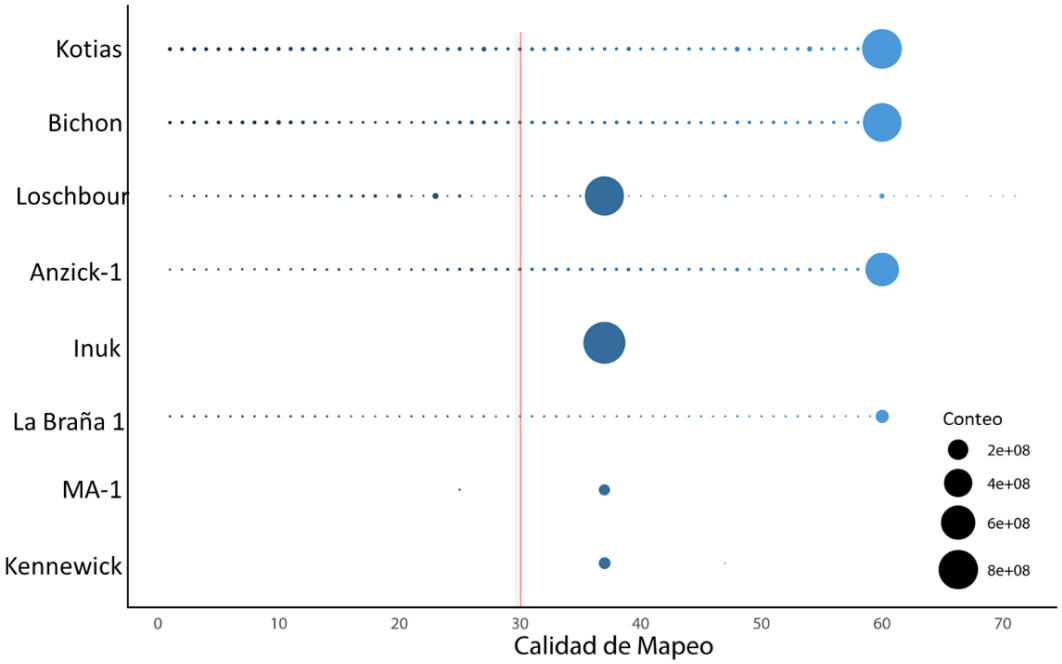


Figura 14.- Gráfica de las frecuencias de las calidades de mapeo de los humanos antiguos. La línea roja representa el corte de la calidad mínima de 30 que se tomó para el análisis. El tamaño de los círculos representa la frecuencia respecto a cada una de las calidades encontradas por individuo.

3.1.2 Cobertura respecto al genoma de referencia hg38

Se puede observar en la Figura 15 las coberturas obtenidas de cada uno de los genomas antiguos, mostrando que cubren más del 50% del genoma de referencia (el cual consta de 3,209,286,105 nucleótidos), donde un nucleótido está siendo cubierto al menos una vez. Kennewick es el genoma con menor cobertura y Kotias con la mayor respecto al genoma hg38.

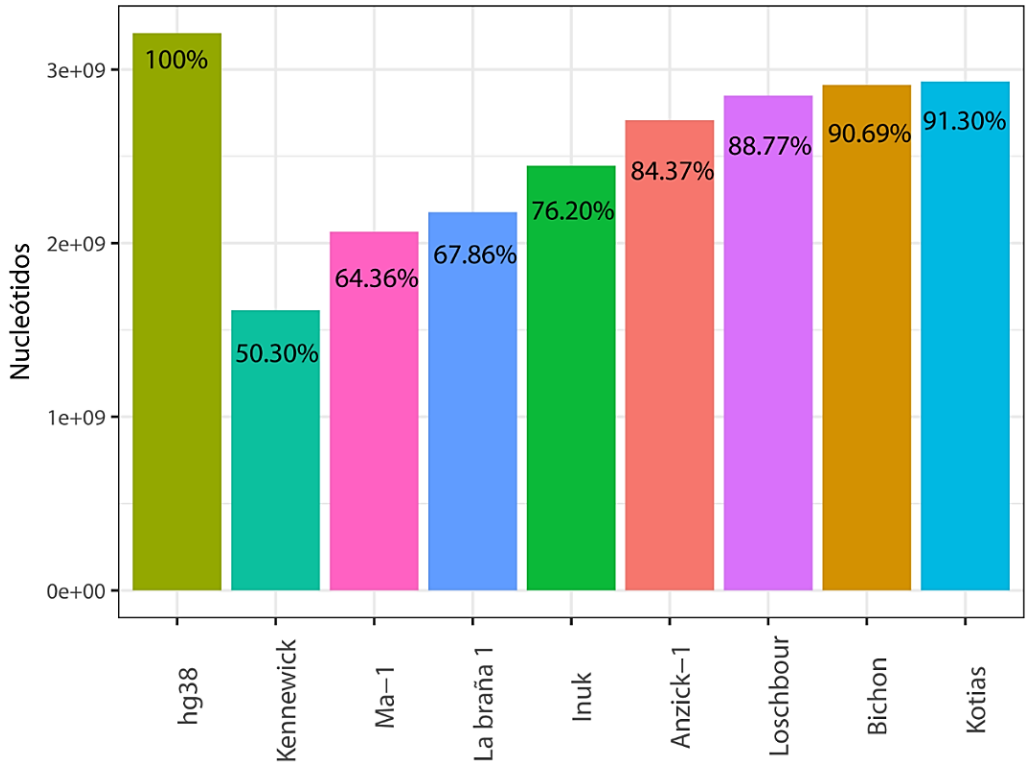


Figura 15.- Gráfica de la cobertura de los genomas antiguos respecto al genoma hg38.

3.1.3 SNP calling

En la Tabla 13 se puede observar la cantidad total de los SNP obtenidos por individuo y los SNP que fueron tomados para este análisis, que son los que muestran una calidad mínima de 20 de la escala phred (escala de probabilidad de que un SNP sea identificado erróneamente, en este caso, que existe una probabilidad de error del 0.01).

Tabla 13. Cantidad de los SNP encontrados por genoma.
Se presentan los SNP totales, los SNP con calidad mínima de 20 y el porcentaje de estos últimos.

Genoma	SNP	SNP con calidad mínima	Porcentaje
Anzick-1	6,627,890	3,573,504	53.92
Bichon	15,179,995	10,339,432	68.11
Kennewick	6,483,611	657,863	10.15
Kotias	6,241,201	3,672,818	58.85
La braña-1	3,667,132	1,027,678	28.02
Loschbour	3,310,571	3,133,032	94.64
Ma-1	3,499,205	878,456	25.10
Inuk	2,641,287	2,052,261	77.70

En la Figura 16 se puede observar de manera gráfica los SNP obtenidos (color cian) y los que tenían el parámetro mínimo establecido (color salmón). Se puede apreciar que Bichon es el genoma que presenta más cantidad de SNP con la calidad mínima de 20, mientras que Kennewick es el individuo que presenta menor cantidad de ellos.

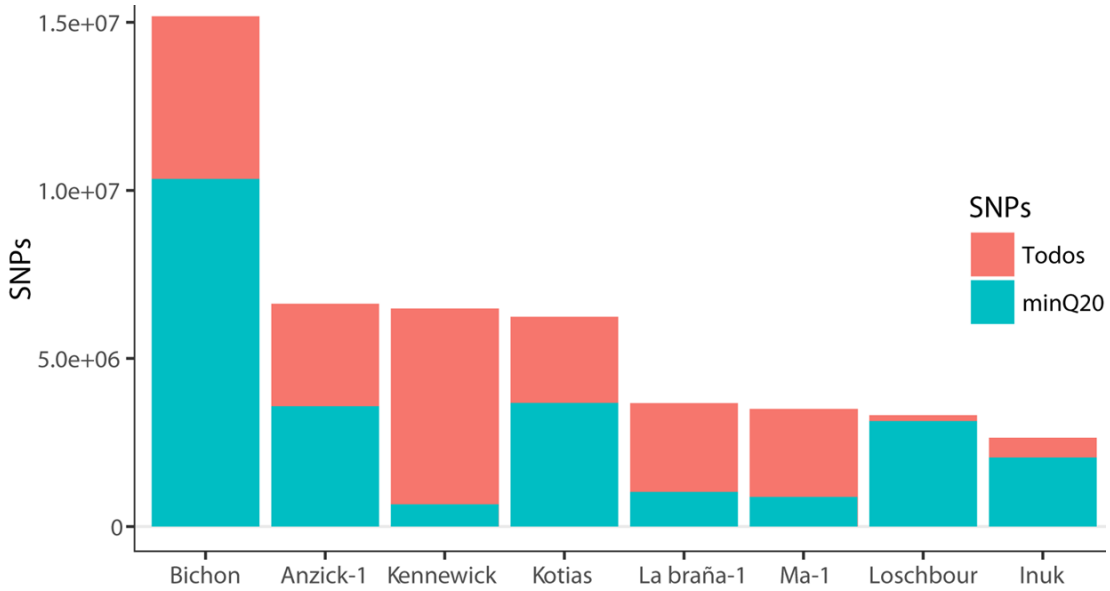


Figura 16.- Gráfica de barras de los SNP encontrados por humano antiguo analizado. De color salmón se presentan los SNP encontrados con calidad menor a 20 y de color turquesa se presentan los SNP con calidad mínima de 20.

En la Figura 17 se puede observar el boxplot de la distribución de solo aquellos SNP con calidad de al menos 20. Haciendo este corte, se observa que todas las distribuciones presentan una mediana

por arriba de la calidad mínima de 20 y que existen bastantes valores por encima de él, que sustentan que se tienen SNP con muy bajo porcentaje de error.

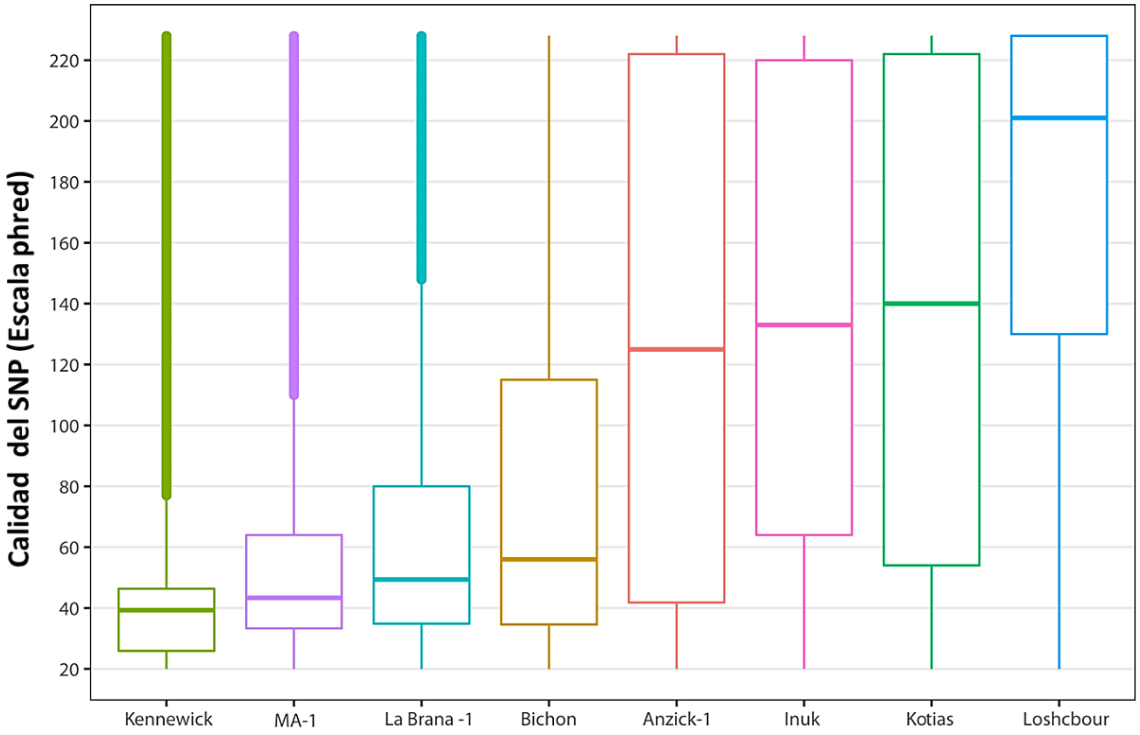


Figura 17.- Boxplot de las calidades de los SNP de los humanos antiguos analizados.

3.1.4 Comparación de los índices de diversidad entre poblaciones

Para ver el comportamiento de los índices de diversidad nucleotídica, se graficó dicho estimador de cada una de las poblaciones, como se muestra en la Figura 18, así como también θ (Figura 19) y la D de Tajima (Figura 20).

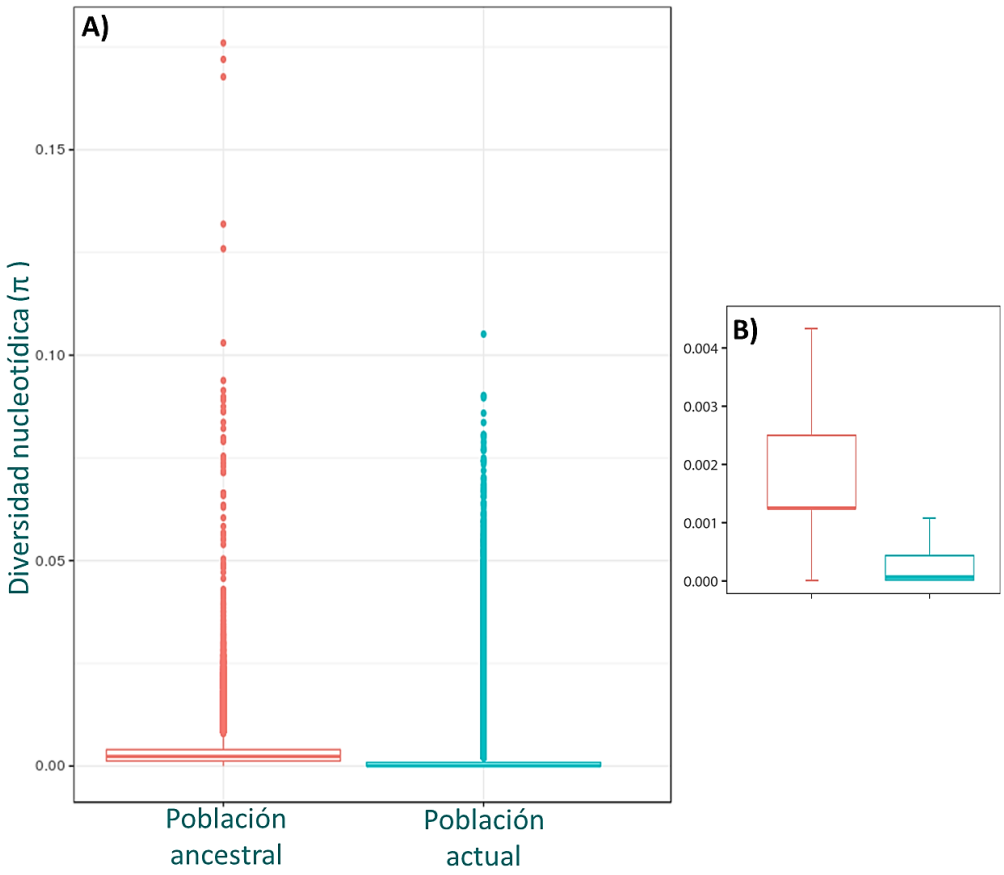


Figura 18.- Boxplot de la distribución de diversidad nucleotídica entre poblaciones. De color salmón se encuentra el boxplot de la población ancestral y de color cian la población actual. A) Comportamiento de todos los datos de ambas poblaciones. B) Comportamiento de los datos sin outliers.

La mediana de los datos de la población antigua se encuentra en el valor de π=0.002333 y la media en el valor de π=0.002333, mientras que en la población moderna se localiza en π=9.17E-05 y π=1.028e-03 respectivamente. Con la prueba de normalidad Shapiro-Wilk se sabe que el comportamiento de los datos en ambas poblaciones no es normal, con lo cual se usó la prueba estadística Wilcoxon para comparar las medias, al obtenerse un valor menor 0.05, se puede concluir que para nuestro intervalo de confianza (0.95) aceptamos la hipótesis alternativa, de que existen diferencias significativas entre poblaciones. Para conocer si ambas distribuciones son distintas se realizó una prueba de Kolmogórov-Smirnov, y se observó que ambas distribuciones son diferentes. Las pruebas estadísticas se incorporan en el Anexo 4. Se puede observar en la Figura 18 que existe una reducción de la diversidad, dicha disminución de la diversidad se puede deber a efectos demográficos, como cuellos de botella que ocurrieron durante la expansión geográfica (Henn et al., 2012).

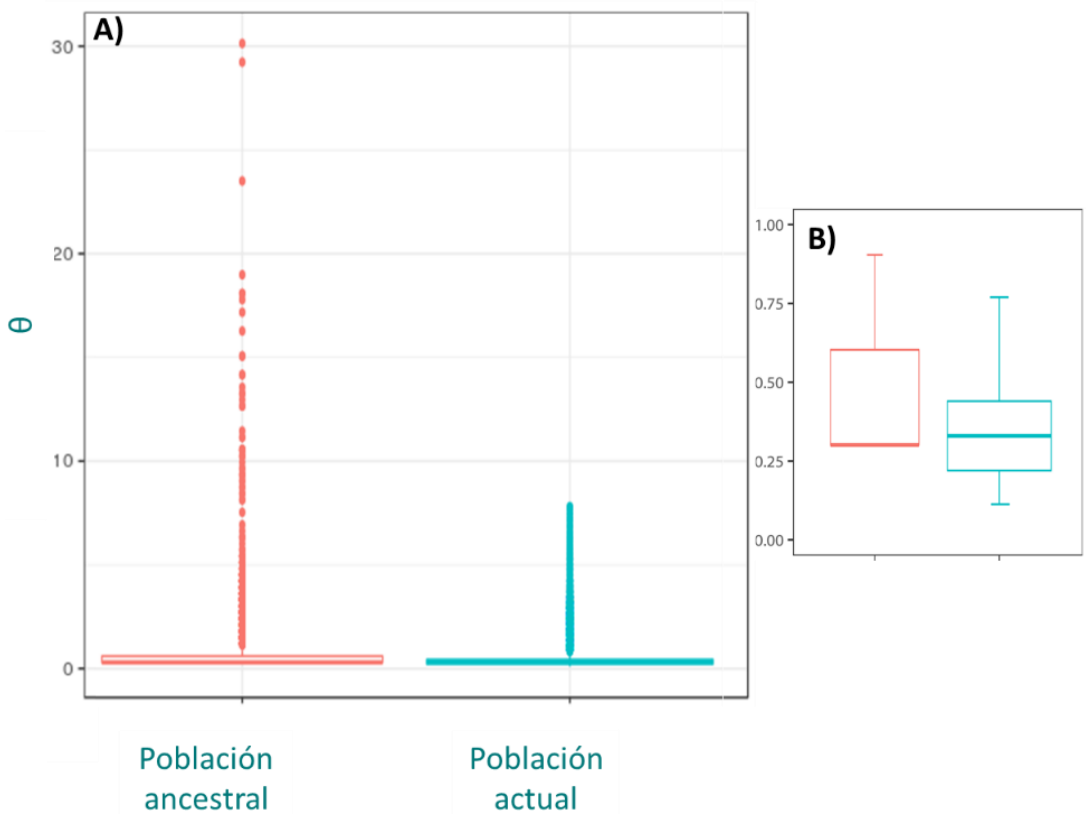


Figura 19.- Boxplot de la distribución de θ entre poblaciones. De color salmón se encuentra el boxplot de la población ancestral y de color cian la población actual. A) Comportamiento de todos los datos de ambas poblaciones. B) Comportamiento de los datos sin outliers.

La mediana de los datos de la población antigua se localiza en el valor de θ= 0.3014 y la media θ= 0.4162, mientras que en la población moderna se encuentran en θ= 0.3298 y θ=0.3417 respectivamente. Se sabe que el comportamiento de θ en ambas poblaciones no es normal (prueba de normalidad Shapiro-Wilk), las medias son significativamente diferentes (Prueba Wilcoxon), y que las distribuciones son distintas al compararse entre poblaciones (prueba de Kolmogórov-Smirnov). Las pruebas estadísticas se incorporan en el Anexo 4. Se puede apreciar que existe un mayor número de sitios segregativos en la población antigua. Además, hay una reducción de la diversidad en theta, de la población ancestral hacia la actual, que es consistente con π.

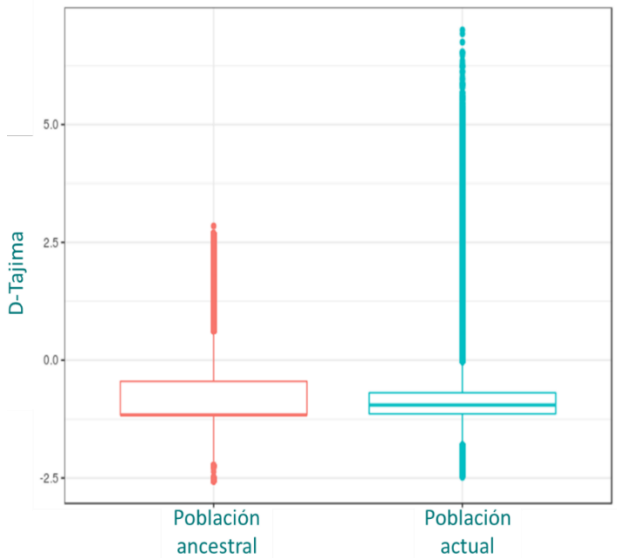


Figura 20.- Distribución de la D de Tajima entre poblaciones. De color salmón se encuentra el boxplot de la población ancestral y de color cian la población actual.

La mediana de los datos de la D de Tajima de la población antigua se encuentra en el valor de -1.1622 y su media en -0.6996, mientras que en la población actual se localizan en -0.9512 y -0.7744 respectivamente. Con las pruebas estadísticas (Anexo 4) se conoce que el comportamiento de la D de Tajima en ambas poblaciones no es normal (prueba de normalidad Shapiro-Wilk), las medias son significativamente diferentes (Prueba Wilcoxon) y que las distribuciones son distintas al compararse entre poblaciones (prueba de Kolmogórov-Smirnov). Con estas pruebas se afirma que existen diferencias significativas al compararse entre poblaciones.

3.2 Búsqueda de regiones bajo selección

3.2.1 Diferencias de la diversidad genética

Como se observó en las gráficas anteriores, existen diferencias significativas en el comportamiento de los datos en cada estimador (π, θ y D de Tajima) entre poblaciones. Así mismo existe una reducción de la diversidad de la población ancestral a la población actual, que puede deberse a selección, cambios aleatorios o procesos demográficos. En consecuencia, es difícil evaluar el significado que tienen las diferencias específicas del análisis global genómico en determinadas regiones del genoma, al no poderse distinguir que fuerza está actuando. Por lo tanto, se calculó la diferencia del valor ancestral de cada estimador (π, θ y D de Tajima) y su equivalente actual para analizar el cambio que existe en el valor ancestral respecto al actual. De este modo poder generar

la distribución de esta diferencia, y con ello, obtener una distribución más normal, y así, tratar de corregir el efecto demográfico al reescalar los cambios de la diversidad.

Si existiera un efecto demográfico, el valor de la diferencia estaría en la media o cercano a ella, ya que todo el genoma se comportó igual reduciendo su diversidad. Para esta estrategia se observó el comportamiento de la diferencia de los tres estimadores ($\Delta\pi$, $\Delta\theta$ y ΔD-Tajima), seleccionando los valores atípicos de diferencia en diversidad, seleccionando las ventanas que se encontraban en el 0.00001% de la distribución (ventanas candidatas a selección). Este valor se escogió arbitrariamente, además de ser el más alejado de la distribución, basándonos en lo que menciona Luca et al. (2010); Aquellos loci que caen por encima de un límite arbitrario se identifican como valores atípicos, bajo el supuesto de que la mayoría de los loci en el genoma humano evolucionan de forma neutral. Estos loci con valores atípicos representan posibles candidatos de fuertes presiones selectivas (Luca et al., 2010).

De manera que, si una ventana se encontraba en entre el límite establecido (0.00001%) en los tres estimadores se tomó como candidata para los análisis posteriores. Con ello se asegura a seleccionar los valores atípicos y, por ende, las regiones candidatas de fuertes presiones selectivas. Estas regiones difieren del comportamiento normal de la distribución de los polimorfismos (con los valores de π y θ) y de la desviación del modelo neutral (D de Tajima). En las Figuras 21, 22 y 23 se puede ver el comportamiento de los datos (de color cian) y el límite establecido (color rojo) para seleccionar las ventanas con diferencias atípicas entre las poblaciones. Se puede observar en cada una de las gráficas que los valores que no se tomaron en cuenta fueron los valores que se encuentra en las líneas rojas (límite establecido).

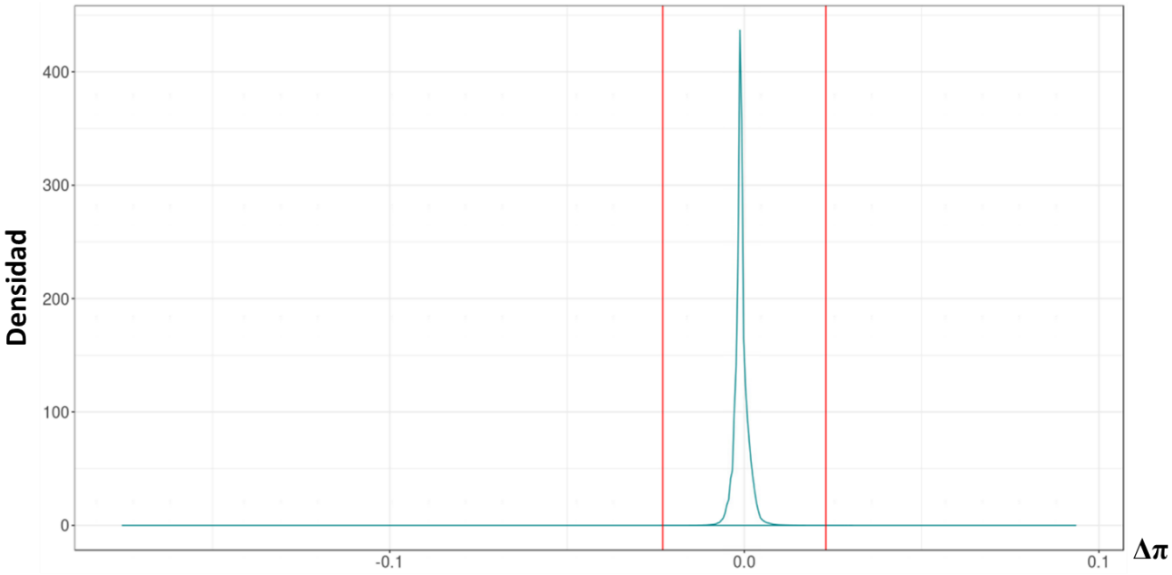


Figura 21.- Gráfica de la densidad de Δπ (Δπ= $\pi_{antiguo} - \pi_{actual}$).
Las líneas rojas representan los límites de confianza mínimo y máximo (0.00001% de los datos).

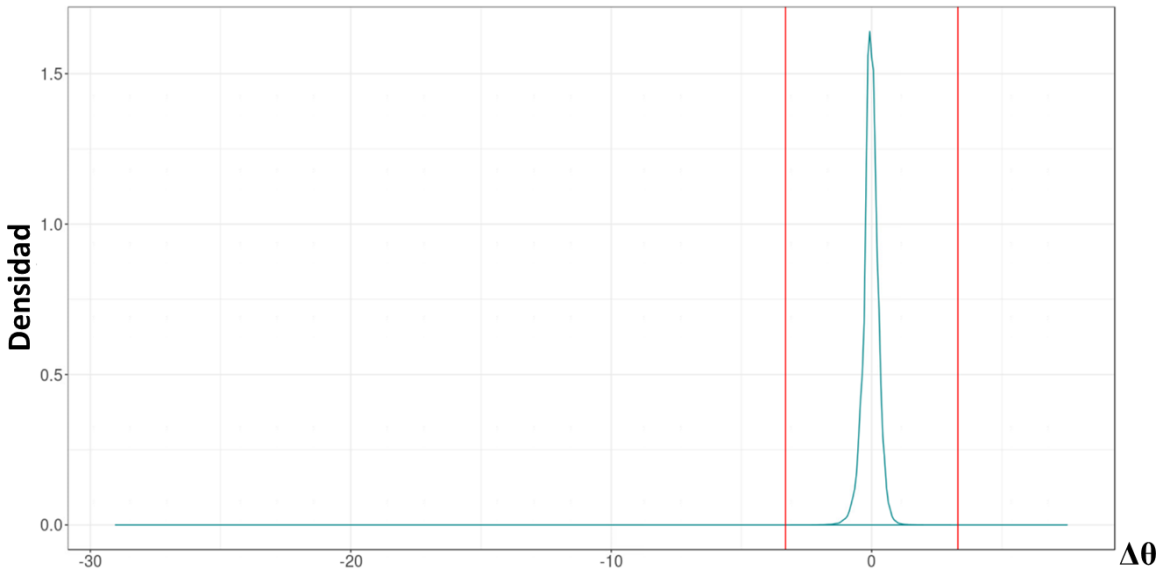


Figura 22.- Gráfica de la densidad de Δθ (Δθ= $\theta_{antiguo} - \theta_{actual}$).
Las líneas rojas representan los límites de confianza mínimo y máximo (0.00001% de los datos).

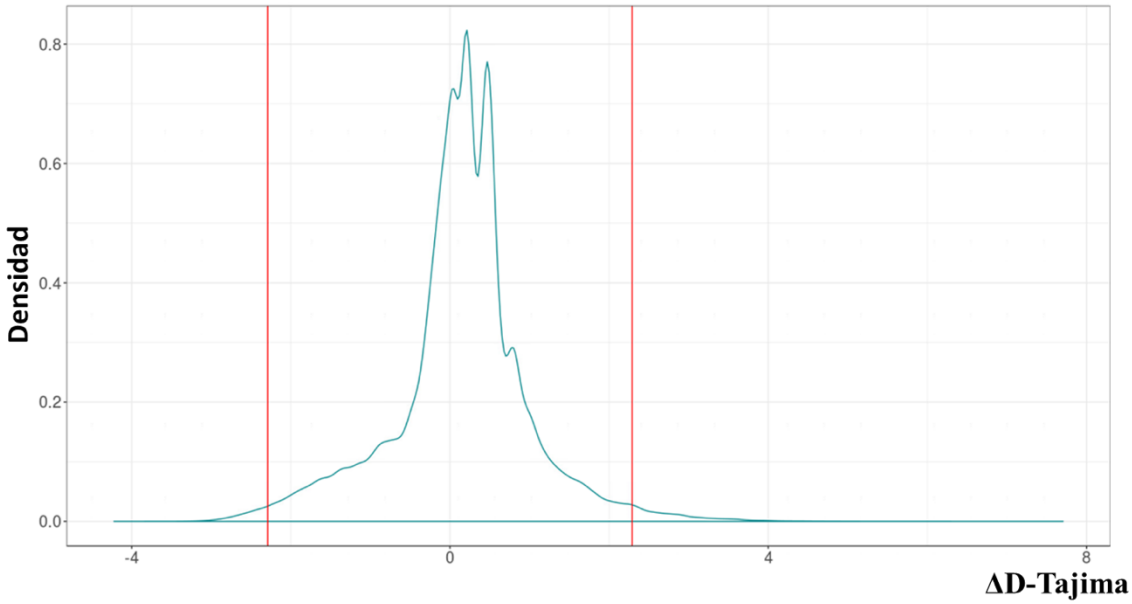


Figura 23.- Gráfica del logaritmo de la densidad de ΔD-Tajima (ΔD Tajima=D $\text{Tajima}_{\text{antiguo}}$ − D $\text{Tajima}_{\text{actual}}$). Las líneas rojas representan los límites de confianza mínimo y máximo (0.00001% de los datos).

3.2.1.1 Categorías de anotación por ventanas

Para conocer las regiones genómicas en las que se encontraban las ventanas atípicas, se clasificaron con ayuda de la base de datos Ensembl (Zerbino et al., 2017). Bajo la estrategia de seleccionar las ventanas con valores atípicos se encontraron 152 ventanas. Se pueden observar en la Figura 24 las categorías de anotación con Ensembl que generan las 152 ventanas seleccionadas. Cabe mencionar que una ventana puede encontrarse en una o varias regiones de anotación.

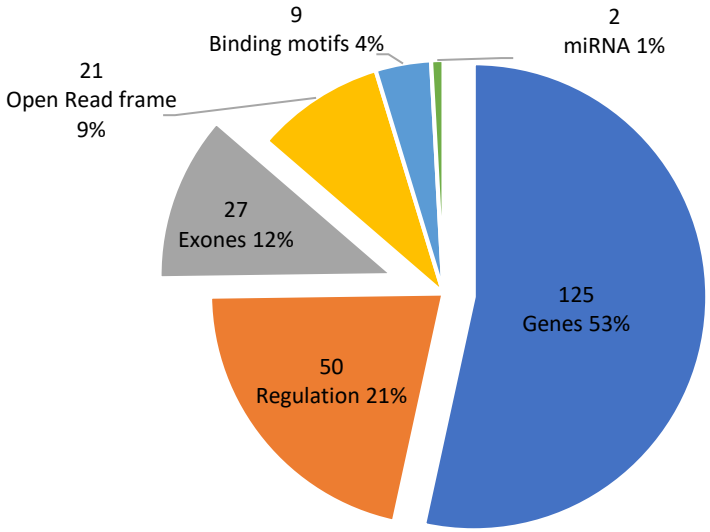


Figura 24.- Gráfica de pastel de la anotación de las ventanas por Ensembl.

3.2.1.2 Categorías de anotación por SNP

Del total de ventanas, solo se tomaron las que se relacionaban con regiones génicas (aquellas que estuvieran anotadas como genes y/o exones, sin repetir), para buscar los SNP que las diferenciaban. Teniendo la lista de los SNP ubicados en estas ventanas, se procedió a realizar la anotación con ayuda de VEP. Se puede observar, en la Figura 25, que hay más variación en regiones no codificantes (regiones intrónicas en este caso) que en otro tipo de regiones.

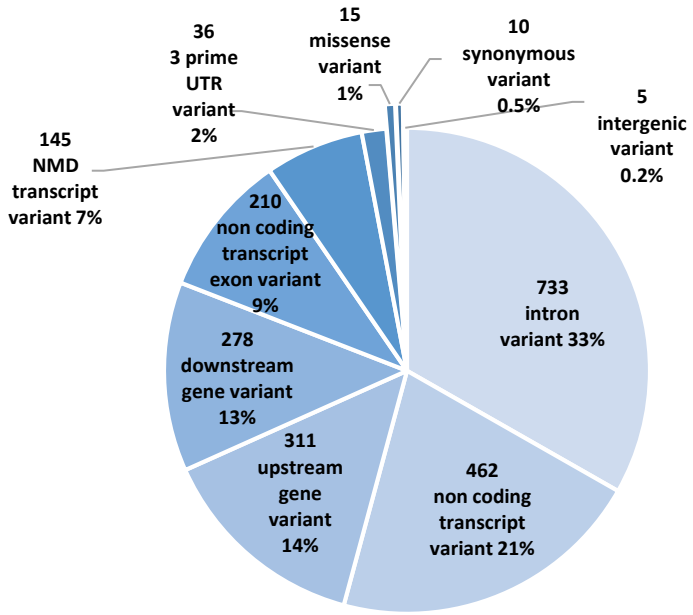


Figura 25.- Gráfica de pastel de la anotación por VEP de todos los SNP encontrados bajo esta metodología.

3.2.1.3 Regiones exónicas

Como este análisis se enfoca en buscar los SNP relacionados a regiones codificantes, solo se tomaron los cambios sinónimos y no sinónimos, los cuales se muestran en la Figura 26. Existen 10 SNP que generan un cambio sinónimo y 15 que generan un cambio no sinónimo.

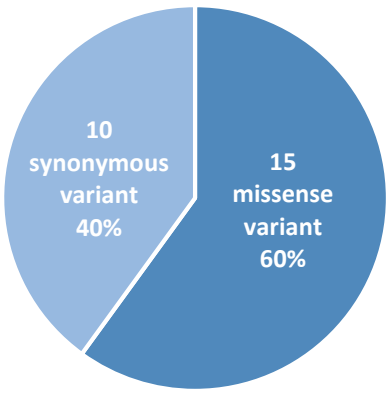


Figura 26.- Gráfica de pastel de la anotación por VEP de los SNP encontrados en regiones exónicas.

3.2.1.4 Genes candidatos

Analizando los 15 SNP que generan un cambio no sinónimo, se observó que están en dos genes, el gen PSMD13 (subunidad 26S del proteosoma), el cual presenta 9 SNP y el gen NDUFS7 (subunidad central de la ubiquinona oxidorreductasa) teniendo 6 SNP, que se describen a continuación:

3.2.1.4.1 PSMD13

El proteosoma 26S (degradación dependiente de ATP de proteínas ubiquitinadas) es un complejo de proteinasas multicatalítico de 2 complejos, un núcleo 20S y un regulador 19S. Este gen codifica una subunidad no ATPasa del regulador 19S, relacionado con:

-Homeostasis proteica.
-Progresión del ciclo celular.
-Apoptosis o la reparación del daño del DNA.

Los SNP que se encontraron se reportan en la Tabla 14 donde se colocan los cambios de aminoácido y los valores de SIFT y PolyPhen.

Tabla 14. Cantidad de SNP encontrados en el gen PSMD13.

Alelo		Aminoácidos	Codones	SIFT	PolyPhen
1	11_244106_A/G	E/G	GAA/GGA	Tolerada (0.15)	Benigna (0.001)
2	11_244108_A/G	N/D	AAT/GAT	Tolerada (0.15)	Benigna (0.03)
3	11_244115_G/A	C/Y	TGC/TAC	Tolerada (0.73)	Benigna (0)
4	11_244129_C/T	R/W	CGG/TGG	Tolerada (0.18)	Benigna (0)
5	11_244136_C/T	S/L	TCA/TTA	Deletérea (0.04)	Benigna (0)
6	11_244141_G/A	A/T	GCT/ACT	Tolerada (0.14)	Benigna (0)
7	11_244167_C/T	P/S	CCC/TCC	Tolerada baja confianza (0.25)	Benigna (0.089)
8	11_244171_T/C	F/S	TTT/TCT	Tolerada baja confianza (0.2)	Benigna (0.011)
9	11_244197_T/C	C/R	TGT/CGT	Deletérea baja confianza (0)	Posiblemente perjudicial (0.598)

De acuerdo con las puntaciones que arroja VEP de ambas probabilidades de sustitución, la mayoría no tiene un impacto grave al sustituir el alelo ancestral, aunque sí existen cambios en las propiedades de los aminoácidos, como se muestra en la Tabla 15. El SNP 9 es el único que podría tener un impacto en el funcionamiento de la proteína por los valores que arrojan SIFT y PolyPhen, además de que los cambios en las propiedades fisicoquímicas son notorios.

Tabla 15. Cambios de aminoácidos entre la población ancestral respecto a la actual de los SNP del gen PSMD13.

SNP	Alelo	Aminoácido	Estructura	Carga	Aminoácido con sulfuro	Masa molar	Formula
1	Actual	Ácido glutámico	Residuos cargados negativamente	ácido-polar	no	147.12926 g/mol	C5H9NO4
	Antiguo	Glicina	aminoácidos alifáticos	no polar	no	75.0666 g/mol	C2H5NO2
2	Actual	Asparagina	aminoácidos neutros	polar	no	132.11792 g/mol	C4H8N2O3
	Antiguo	Ácido aspártico	Residuos cargados negativamente	ácido-polar	no	133.10268 g/mol	C4H7NO4
3	Actual	Cisteína	aminoácidos neutros	polar	si	21.15818 g/mol	C3H7NO2S
	Antiguo	Tirosina	aminoácidos aromáticos	polar	no	181.18854 g/mol	C9H11NO3
4	Actual	Arginina	Residuos cargados positivamente	básico-polar	no	174.20096 g/mol	C6H14N4O2
	Antiguo	Triptófano	aminoácidos aromáticos	no polar	no	204.22518 g/mol	C11H12N2O2
5	Actual	Serina	aminoácidos neutros	polar	no	105.09258 g/mol	C3H7NO3
	Antiguo	Leucina	aminoácidos alifáticos	no polar	no	131.17292 g/mol	C6H13NO2
6	Actual	Alanina	aminoácidos alifáticos	no polar	no	89.09318 g/mol	C3H7NO2
	Antiguo	Treonina	aminoácidos neutros	polar	no	119.11916 g/mol	C4H9NO3
7	Actual	Prolina	aminoácidos alifáticos	no polar	no	115.13046 g/mol	C5H9NO2
	Antiguo	Serina	aminoácidos neutros	polar	no	105.09258 g/mol	C3H7NO3
8	Actual	Fenilalanina	aminoácidos aromáticos	no polar	no	165.18914 g/mol	C9H11NO2
	Antiguo	Serina	aminoácidos neutros	polar	no	105.09258 g/mol	C3H7NO3
9	Actual	Cisteína	aminoácidos neutros	polar	si	21.15818 g/mol	C3H7NO2S
	Antiguo	Arginina	Residuos cargados positivamente	básico-polar	no	174.20096 g/mol	C6H14N4O2

Estos SNP se encuentran repartidos en dos transcritos de *splicing* alternativo del gen, los SNP 1 al 6 se localizan en el transcrito ENST00000431206 y del 7 al 9 se encentran en el ENST00000382671. Para ver si estos SNP están en interacción con algún otro sitio de la proteína es necesario la estructura cristalográfica. Desafortunadamente para estos transcritos no se encontró la estructura cristalizada y, además, no existen reportes de los sitios de unión, catalíticos u otro sitio importante para el funcionamiento de la proteína.

Revisando la frecuencia de nucleótidos entre poblaciones, en la ventana que corresponde a este gen, como se muestra en la figura 27, se puede observar que todos los alelos se presentan en ambas poblaciones, excepto los del nucleótido 244136, en el que la población actual solo presenta cisteína, mientras que en la población antigua presenta cisteína y timina. Estos alelos solo se encuentran en el genoma antiguo de Bichon (heterocigoto), mientras que en los demás humanos antiguos tienen incompleta esta información por falta de cobertura.

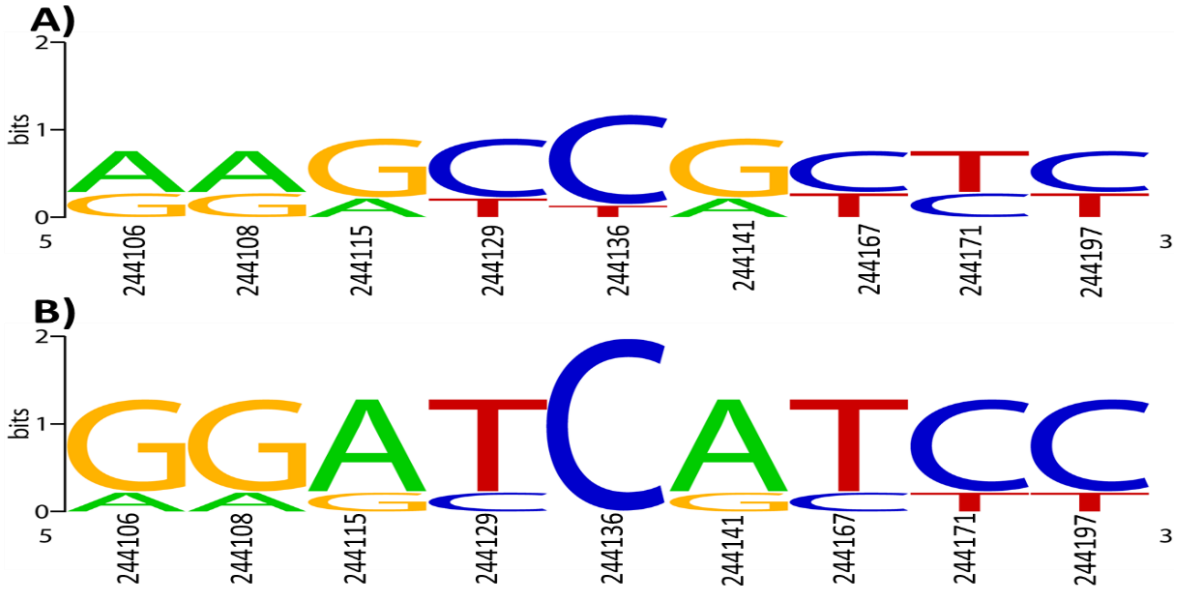


Figura 27.- Alineamiento de los SNP encontrados en la ventana 244100-244100 del gen PSMD13. A) Variantes de la población antigua. B) Variantes de la población actual.

3.2.1.4.2 NDFS7

Este gen codifica una proteína de la subunidad S7 del complejo I, del dinucleótido de nicotinamida adenina (NADH): ubiquinona oxidorreductasa, que forma parte de la cadena respiratoria mitocondrial. Este complejo funciona en la transferencia de electrones del NADH a la cadena respiratoria. Este gen lleva acabo la reacción: NADH + ubiquinona + $5H^+$ = NAD^+ + ubiquinol + $4H^+$.

Los SNP que se encontraron en este gen son 6, los cuales se desglosan en la Tabla 16. Cabe mencionar que los SNP caen en diferentes transcritos (*splicing* alternativo) por lo que cada uno genera un valor de SIFT y PolyPhen diferente. Son 7 transcritos a los que se relacionan estos SNP.

Tabla 16. SNP encontrados en el gen NDFS7. Transcritos a los que se relaciona cada uno de los SNP, así como los cambios no sinónimos que generan y los valores SIFT y PolyPhen.

Alelo		Transcrito	Aminoácido	Codón	SIFT	PolyPhen
1	19_1390913 G/A	ENST00000233627	V/M	GTG/ATG	deletérea baja confianza (0.01)	posiblemente dañino (0.879)
		ENST00000313408	V/M	GTG/ATG	deletérea baja confianza (0.01)	probablemente dañino (0.994)
		ENST00000414651	V/M	GTG/ATG	perjudicial baja confianza (0)	posiblemente dañino (0.843)
		ENST00000539480	V/M	GTG/ATG	deletérea baja confianza (0.01)	probablemente dañino (0.951)
		ENST00000546283	V/M	GTG/ATG	deletérea baja confianza (0.01)	probablemente dañino (0.994)
		ENST00000618074	V/M	GTG/ATG	nocivo (0.01)	probablemente dañino (1)
		ENST00000620479	V/M	GTG/ATG	nocivo (0.01)	probablemente dañino (0.93)
2	19_1390949 A/C	ENST00000233627	M/L	ATG/CTG	tolerada baja confianza (0.23)	benigno (0.05)
		ENST00000313408	M/L	ATG/CTG	tolerada baja confianza (0.22)	benigno (0.345)
		ENST00000414651	M/L	ATG/CTG	tolerada baja confianza (0.18)	benigno (0.091)
		ENST00000539480	M/L	ATG/CTG	tolerada baja confianza (0.25)	benigno (0.027)
		ENST00000546283	M/L	ATG/CTG	tolerada baja confianza (0.22)	benigno (0.345)
		ENST00000618074	M/L	ATG/CTG	tolerado (0.18)	probablemente dañino (0.995)
		ENST00000620479	M/L	ATG/CTG	tolerado (0.16)	benigno (0.091)
3	19_1390951 G/C	ENST00000233627	M/I	ATG/ATC	deletérea baja confianza (0.04)	benigno (0.135)
		ENST00000313408	M/I	ATG/ATC	deletérea baja confianza (0.03)	posiblemente dañino (0.809)
		ENST00000414651	M/I	ATG/ATC	deletérea baja confianza (0.02)	benigno (0.236)
		ENST00000539480	M/I	ATG/ATC	deletérea baja confianza (0.04)	benigno (0.33)
		ENST00000546283	M/I	ATG/ATC	deletérea baja confianza (0.03)	posiblemente dañino (0.809)
		ENST00000618074	M/I	ATG/ATC	nocivo (0.03)	probablemente dañino (0.997)
		ENST00000620479	M/I	ATG/ATC	nocivo (0.03)	benigno (0.236)
4	19_1390964 G/C	ENST00000233627	V/L	GTG/CTG	tolerada baja confianza (0.1)	posiblemente dañino (0.531)
		ENST00000313408	V/L	GTG/CTG	deletérea baja confianza (0.02)	posiblemente dañino (0.574)
		ENST00000414651	V/L	GTG/CTG	deletérea baja confianza (0.01)	posiblemente dañino (0.552)
		ENST00000539480	V/L	GTG/CTG	deletérea baja confianza (0.02)	benigno (0.188)
		ENST00000546283	V/L	GTG/CTG	deletérea baja confianza (0.02)	posiblemente dañino (0.574)
		ENST00000618074	V/L	GTG/CTG	tolerado (0.07)	probablemente dañino (0.999)
		ENST00000620479	V/L	GTG/CTG	tolerado (0.07)	posiblemente dañino (0.784)
5	19_1390967 G/C	ENST00000233627	V/L	GTC/CTC	tolerada baja confianza (0.12)	benigno (0.261)
		ENST00000313408	V/L	GTC/CTC	tolerada baja confianza (0.05)	posiblemente dañino (0.84)
		ENST00000414651	V/L	GTC/CTC	deletérea baja confianza (0.04)	benigno (0.389)
		ENST00000539480	V/L	GTC/CTC	tolerada baja confianza (0.06)	benigno (0.214)
		ENST00000546283	V/L	GTC/CTC	tolerada baja confianza (0.05)	posiblemente dañino (0.84)
		ENST00000618074	V/L	GTC/CTC	tolerado (0.07)	probablemente dañino (0.999)
		ENST00000620479	V/L	GTC/CTC	tolerado (0.08)	posiblemente dañino (0.498)
6	19_1390976 G/C	ENST00000233627	A/P	GCC/CCC	tolerada baja confianza (0.16)	posiblemente dañino (0.902)
		ENST00000313408	A/P	GCC/CCC	tolerada baja confianza (0.17)	probablemente dañino (0.913)
		ENST00000414651	A/P	GCC/CCC	tolerada baja confianza (0.14)	posiblemente dañino (0.631)
		ENST00000539480	A/P	GCC/CCC	tolerada baja confianza (0.18)	posiblemente dañino (0.626)
		ENST00000546283	A/P	GCC/CCC	tolerada baja confianza (0.17)	probablemente dañino (0.913)
		ENST00000618074	A/P	GCC/CCC	tolerado (0.14)	probablemente dañino (1)
		ENST00000620479	A/P	GCC/CCC	tolerado (0.13)	probablemente dañino (0.999)

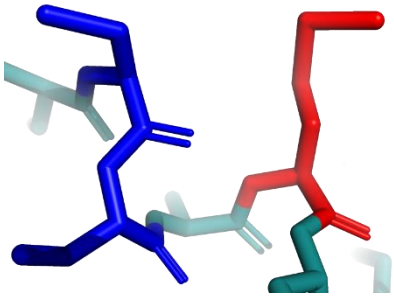
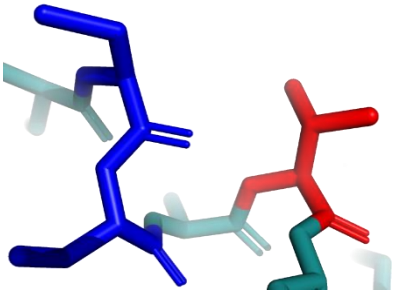
Se revisaron los cambios entre aminoácidos del alelo actual y el ancestral como se muestra en la Tabla 17.

Tabla 17. Cambios de aminoácidos entre la población ancestral respecto a la actual de los SNP del gen NDUFS7.

SNP	Alelo	aminoácido	Estructura	Carga	Aminoácido con sulfuro	Masa molar	Formula
1	Actual	Valina	Aminoácido alifático	no polar	no	117.14634 g/mol	C5H11NO2
	Antiguo	Metionina	Aminoácido neutral	no polar	si	149.21134 g/mol	C5H11NO2S
2	Actual	Metionina	Aminoácido neutral	no polar	si	149.21134 g/mol	C5H11NO2S
	Antiguo	Leucina	Aminoácido alifático	no polar	no	131.17292 g/mol	C6H13NO2
3	Actual	Metionina	Aminoácido neutral	no polar	si	149.21134 g/mol	C5H11NO2S
	Antiguo	Isoleucina	Aminoácido alifático	no polar	no	131.17292 g/mol	C6H13NO2
4	Actual	Valina	Aminoácido alifático	no polar	no	117.14634 g/mol	C5H11NO2
	Antiguo	Leucina	Aminoácido alifático	no polar	no	131.17292 g/mol	C6H13NO2
5	Actual	Valina	Aminoácido alifático	no polar	no	117.14634 g/mol	C5H11NO2
	Antiguo	Leucina	Aminoácido alifático	no polar	no	131.17292 g/mol	C6H13NO2
6	Actual	Alanina	Aminoácido alifático	no polar	no	89.09318 g/mol	C3H7NO2
	Antiguo	Prolina	Aminoácido alifático	no polar	no	115.13046 g/mol	C5H9NO2

Se realizó el modelado de la proteína (Anexo 5) para ver las interacciones con los sitios de unión a metales (4Fe-4S), sitios únicos que se reportan en la base de datos UniProt. Solo se realizó el modelado del transcrito ENST00000233627, ya que es el mismo que reporta UniProt. Dentro del análisis de la estructura se observó que solo el SNP 1, que se encuentra en el aminoácido 91 (cromosoma 19 posición 1390913 y alelos G/A), está junto a dos nucleótidos relacionados a unión a metales (aminoácido 88 y 89), como se muestra en la Tabla 18.

Tabla 18. Cambio de aminoácidos entre la población ancestral respecto a la actual del aminoácido 91 del gen NDUFS7. De color azul se muestran los aminoácidos relacionados a la unión a metales (aminoácidos 88 y 89) y en rojo el SNP (aminoácido 91).

Alelo Ancestral Aminoácido: Valina	Alelo Actual Aminoácido: Metionina
	

De acuerdo con los valores de SIFT y PolyPhen podría existir un impacto en la función de la proteína con este cambio de aminoácido, quizá relacionado con el hecho de que el cambio se encuentra cerca de un sitio importante para la función de la proteína.

Analizando los SNP de la ventana entre poblaciones, se puede observar que los humanos actuales no presentan variación a lo largo de dicha ventana, como se muestra en la Figura 28. Los alelos que generan un cambio no sinónimo solo se encuentran en Kotias (heterocigoto).

Figura 28.- Alineamiento de los SNP encontrados en la ventana 1390900-1391000 del gen NDFS7; de color gris claro se muestran los SNP que generan un cambio sinónimo. A) Variantes de la población antigua. B) Variantes de la población actual.

3.2.1.5 Genes ortólogos

Para los genes candidatos encontrados se procedió a buscar los genes ortólogos respecto al modelo de *S. cerevisiae.* En la Tabla 19 se puede observar que en el NDFS7 no presenta genes ortólogos, mientras que PSMD13 si presenta.

Tabla 19. Genes ortólogos de cada modelo respecto a los genes candidatos.

	Modelo	Gen del modelo	% id del gen del modelo respecto al consultado	% id del gen consultado respecto al gen del modelo
PSMD13	*S. cerevisiae*	RPN9	28.0423	26.972
NDFS7	*S. cerevisiae*	-	-	-

3.2.2 Diferencia entre frecuencias de los SNP

Con la metodología de las diferencias de la diversidad genética se obtuvieron pocas regiones. Por ese motivo se propone el análisis de las diferencias entre frecuencias de los SNP entre poblaciones para buscar SNP puntuales, posibles candidatos a selección. Bajo esta estrategia se analizaron dos tipos de datos: Humanos antiguos con frecuencia igual a 1 y Humanos antiguos con frecuencia igual a 1, que se describen más adelante.

3.2.2.1 Humanos antiguos con frecuencia igual a 1

En esta estrategia se examinaron, por posiciones nucleotídicas, las frecuencias entre poblaciones. Se analizaron las posiciones en las que los humanos antiguos presentaban una frecuencia igual a 1 (es decir, solo existe un tipo de nucleótido, lo que implica variación nula) y los humanos actuales presentaban variación en sus frecuencias nucleotídicas (frecuencia ≥ 0.5 pero < 1), siendo el nucleótido más frecuente diferente entre poblaciones. Con ello se encontraron 6 SNP, como se muestra en la siguiente Tabla:

Tabla 20. SNP encontrados cuando los humanos antiguos presentan una frecuencia igual a 1.

Cromosoma	Posición	Alelo antiguo	Frecuencia del alelo antiguo	Alelo actual	Frecuencia del alelo actual
10	116763432	A	1	T	0.576876996805112
3	53390849	G	1	C	0.662539936102236
3	195711572	C	1	T	0.773562300319489
6	46335020	G	1	A	0.581869009584665
6	46348830	G	1	A	0.58526357827476
6	46383784	A	1	G	0.618610223642173

Realizando la anotación con VEP de esta lista de SNP encontrados, se puede observar que ninguno de ellos se localiza en regiones codificantes, como se muestra en la Figura 29.

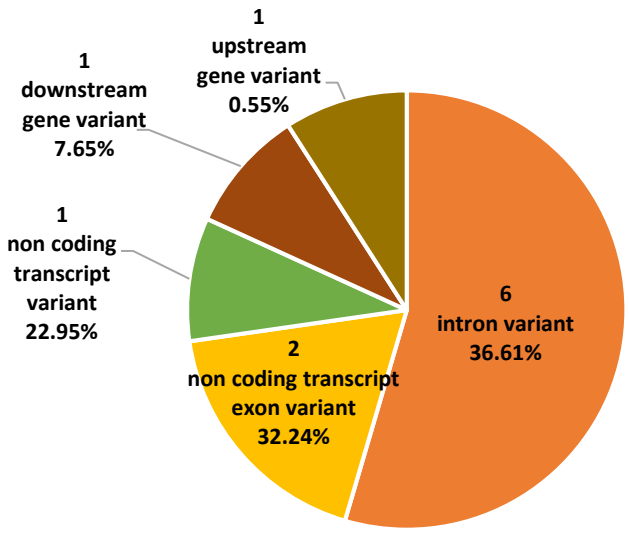


Figura 29.- Gráfica de pastel de la anotación de los SNP por VEP.

Cada una de estas regiones puede generar diversas variaciones, como se muestra en la Tabla 21, donde se muestran estas variaciones y regiones o genes a los que se relacionan.

Tabla 21. Regiones y genes relacionados a los SNP encontrados.

Cromosoma	Posición	Alelo	Región genómica	Gen
10	116763432	T/A	Intron variant	HSPA12A
3	195711572	T/C	Downstream gene variant	LINC00969
			Intron variant	LINC00969
			Non-coding transcript exon variant	AC233280.20
			Non-coding transcript exon variant	LINC00969
			Non-coding transcript variant	LINC00969
3	53390849	C/G	Intron variant	CACNA1D
			Upstream gene variant	SNORA26
6	46335020	A/G	Intron variant	RCAN2
6	46348830	A/G	Intron variant	RCAN2
6	46383784	G/A	Intron variant	RCAN2

3.2.2.2 Humanos actuales con frecuencia igual a 1

En este caso se analizaron las posiciones en las que los humanos actuales presentaban una frecuencia igual a 1 (es decir, solo existe un tipo de nucleótido, lo que implica variación nula) y los humanos antiguos presentaban variación en sus frecuencias nucleotídicas (frecuencia ≥ 0.5 pero ≤ 1), siendo el nucleótido más frecuente diferente entre poblaciones. Bajo esta estrategia se encontraron 20,186 SNP, los cuales generan 30,880 variaciones genéticas, como se muestra en la Figura 30.

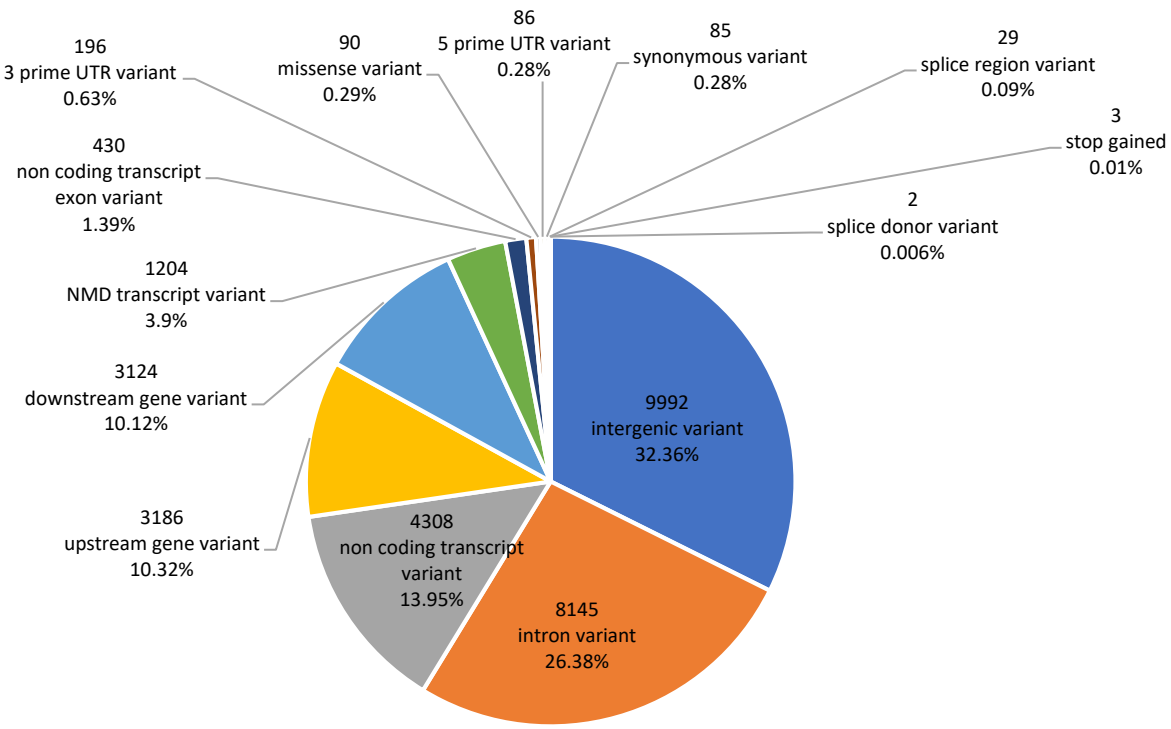


Figura 30.- Gráfica de pastel de la anotación de los SNP por VEP.

De los SNP que se encuentran en regiones codificantes son 90 con cambios no sinónimos, 85 con cambios sinónimos y 3 con cambios de ganancia de codón de paro (*stop gained*), como se muestra en la Figura 31.

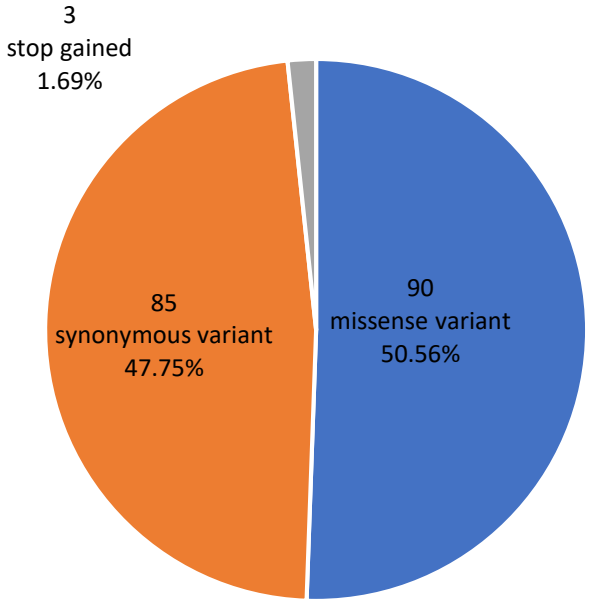


Figura 31.- Gráfica de pastel de la anotación de los SNP en regiones exónicas.

Estos 90 cambios no sinónimos y 3 de codón de paro están relacionados con 79 genes, los cuales se desglosan en el Anexo 6. En la Figura 32 se muestran los procesos biológicos (la función de la proteína en el contexto de una red más grande de proteínas que interactúan para lograr un proceso a nivel de la célula u organismo), a los que se encuentran relacionados los 79 genes encontrados bajo esta metodología. Los procesos biológicos se definen en el Anexo 8.

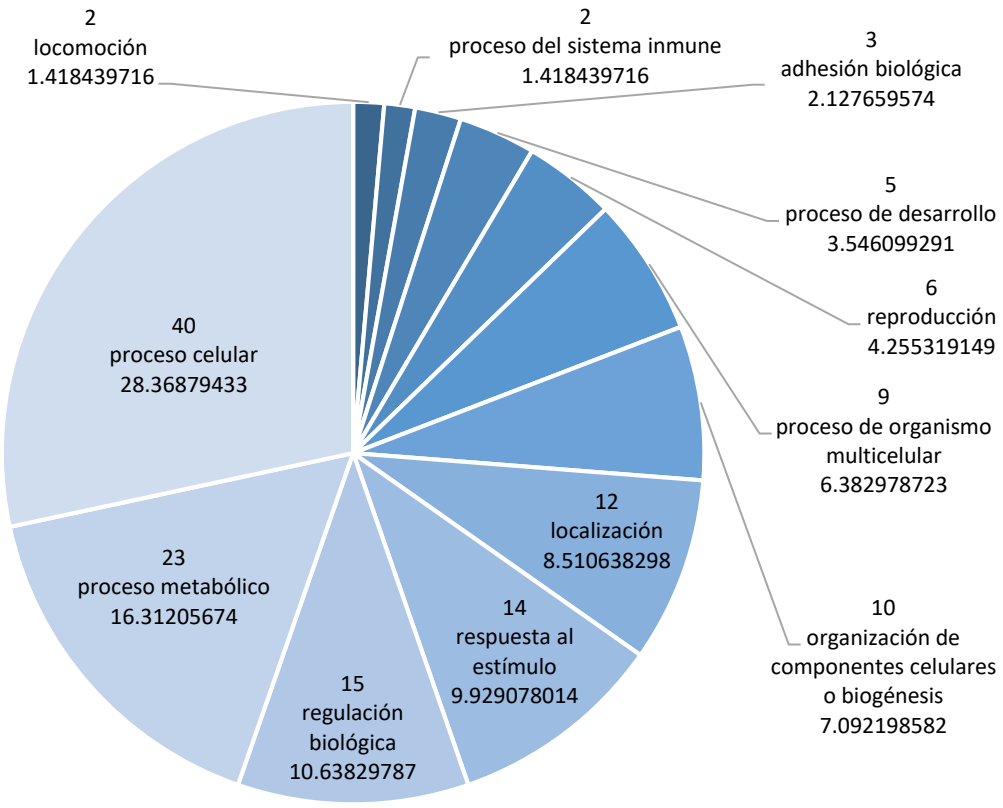


Figura 32.- Gráfica de pastel de la anotación por panther de los SNP con cambio sinónimo.

3.2.2.2.1 No sinónimas

En la figura 33, se puede observar la distribución de las frecuencias encontradas que corresponden a 90 cambios no sinónimos.

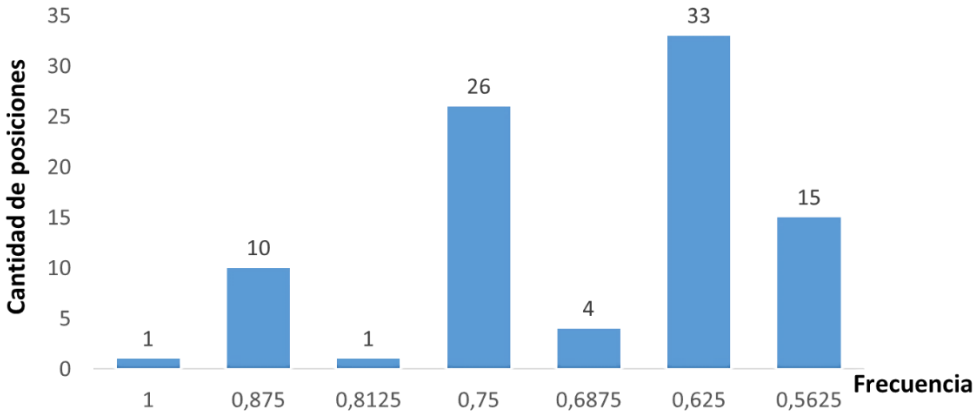


Figura 33.- Gráfica de barras de la distribución de las frecuencias por posición nucleotídica.

De las 90 posiciones encontradas con cambios no sinónimos (lista completa en Anexo 6), en la Tabla 22, se muestran los 12 mejores candidatos de acuerdo con su frecuencia alta o baja diversidad (en humanos antiguos). Cabe recordar que para la población actual la frecuencia es igual a 1 (no hay variación nucleotídica), mientras que en la población antigua sí existe. Solo se representa en la Tabla 22 aquellas posiciones donde la frecuencia va de 1 a 0.8, respecto al nucleótido de mayor frecuencia, que difiere con la población actual.

Tabla 22. SNP que se encuentran en regiones codificantes y generan un cambio no sinónimo.

Chr	Posición	Alelo antiguo	Frecuencia del alelo antiguo	Alelo actual	Frecuencia del alelo actual	Gen
3	195785236	A	1	T	1	MUC4
12	9095637	C	0.875	T	1	A2M
12	8852296	A	0.875	C	1	A2ML1
X	100988470	T	0.875	G	1	ARL13A
X	31478233	T	0.875	C	1	DMD
X	66602765	C	0.875	T	1	EDA2R
5	141101254	G	0.875	C	1	PCDHB3
X	151700795	G	0.875	A	1	PRRG3
5	149981314	C	0.875	T	1	SLC26A2
8	144414297	G	0.875	C	1	SLC39A4
14	22086905	T	0.875	C	1	TRAV23DV6
16	74391650	G	0.8125	A	1	NPIPB15

3.2.2.2.1.1 Genes candidatos

De esta lista de 12 genes que presentan baja diversidad nucleotídica en humanos antiguos y nula diversidad en humanos actuales, se tienen los valores que arrojan SIFT y PolyPhen de cada uno de los genes y de cada uno de los transcritos relacionados a los SNP encontrados, como se muestra en la Tabla 23. Todas las variantes y valores de SIFT y PolyPhen que arrojan se encuentran en el anexo 7.

Tabla 23. Valores de SIFT y PolyPhen de cada SNP.

Chr	Posición	Gen	Transcrito	Posición en proteína	Aminoácidos	SIFT	PolyPhen
3	195785236	MUC4	ENST00000462323	2115	D/V	tolerada baja confianza (1)	posiblemente dañino (0.908)
3	195785236	MUC4	ENST00000463781	2115	D/V	tolerada baja confianza (1)	posiblemente dañino (0.811)
3	195785236	MUC4	ENST00000466475	2115	D/V	tolerado (0.57)	desconocido (0)
3	195785236	MUC4	ENST00000470451	2115	D/V	tolerado (1)	benigno (0.018)
3	195785236	MUC4	ENST00000475231	2115	D/V	tolerada baja confianza (1)	posiblemente dañino (0.811)
3	195785236	MUC4	ENST00000477086	2115	D/V	tolerado (1)	posiblemente dañino (0.908)
3	195785236	MUC4	ENST00000477756	2115	D/V	tolerada baja confianza (0.38)	posiblemente dañino (0.811)
3	195785236	MUC4	ENST00000478156	2115	D/V	tolerada baja confianza (0.82)	posiblemente dañino (0.811)
3	195785236	MUC4	ENST00000479406	2115	D/V	tolerado (0.57)	desconocido (0)
3	195785236	MUC4	ENST00000480843	2115	D/V	tolerado (0.52)	benigno (0.003)
12	8852296	A2ML1	ENST00000299698	850	D/E	tolerado (1)	benigno (0)
12	8852296	A2ML1	ENST00000539547	359	D/E	tolerado (1)	benigno (0)
12	8852296	A2ML1	ENST00000541459	400	D/E	tolerado (1)	benigno (0)
12	9095637	A2M	ENST00000318602	639	N/D	tolerado (0.91)	benigno (0)
14	22086905	TRAV23DV6	ENST00000390451	103	S/L	tolerado (1)	benigno (0)
5	141101254	PCDHB3	ENST00000231130	202	P/R	tolerada baja confianza (1)	benigno (0)
5	149981314	SLC26A2	ENST00000286298	574	I/T	tolerado (0.84)	benigno (0)
8	144414297	SLC39A4	ENST00000276833	347	V/L	tolerado (1)	benigno (0)
8	144414297	SLC39A4	ENST00000301305	372	V/L	tolerado (1)	benigno (0)
X	31478233	DMD	ENST00000343523	208	R/Q	tolerado (0.58)	benigno (0)
X	31478233	DMD	ENST00000357033	2937	R/Q	tolerado (1)	benigno (0)
X	31478233	DMD	ENST00000358062	633	R/Q	tolerado (0.85)	benigno (0)
X	31478233	DMD	ENST00000359836	477	R/Q	tolerado (0.86)	benigno (0)
X	31478233	DMD	ENST00000378677	2933	R/Q	tolerado (1)	benigno (0)
X	31478233	DMD	ENST00000378707	477	R/Q	tolerado (0.89)	benigno (0)
X	31478233	DMD	ENST00000474231	477	R/Q	tolerado (0.83)	benigno (0)
X	31478233	DMD	ENST00000541735	477	R/Q	tolerado (0.87)	benigno (0)
X	31478233	DMD	ENST00000619831	2932	R/Q	tolerado (1)	benigno (0)
X	31478233	DMD	ENST00000620040	2936	R/Q	tolerado (1)	benigno (0)
X	66602765	EDA2R	ENST00000253392	129	T/A	tolerado (0.38)	benigno (0)

X	66602765	EDA2R	ENST00000374719	129	T/A	tolerado (0.48)	benigno (0)
X	66602765	EDA2R	ENST00000396050	129	T/A	tolerado (0.38)	benigno (0)
X	66602765	EDA2R	ENST00000451436	129	T/A	tolerado (0.48)	benigno (0)
X	100988470	ARL13A	ENST00000450457	279	M/I	tolerada baja confianza (0.12)	benigno (0)
X	100988470	ARL13A	ENST00000494863	113	M/I	tolerado (0.15)	benigno (0)
X	151700795	PRRG3	ENST00000370353	153	N/S	tolerado (1)	benigno (0)
X	151700795	PRRG3	ENST00000538575	153	N/S	tolerado (1)	benigno (0)
16	74391650	NPIPB15	ENST00000429990	301	Y/C	tolerada baja confianza (1)	benigno (0)

Existen diversos SNP que ya se encuentran publicados como variantes naturales o secuencia en conflicto. UniProt reporta variantes naturales de la secuencia de la proteína, que incluyen polimorfismos, variaciones entre cepas, aislados o cultivos, mutaciones asociadas a enfermedades y eventos de edición de ARN. Mientras que en la sección de secuencia en conflicto informa de las diferencias entre la secuencia canónica (secuencia de proteína más frecuente o la más conservada en especies ortólogas) y las diferentes presentaciones de secuencias fusionadas en la entrada. Estas diversas presentaciones pueden provenir de diferentes proyectos de secuenciación, diferentes tipos de experimentos o diferentes muestras biológicas. Los conflictos de secuencia generalmente son de origen desconocido. En la figura 34 se muestra la cantidad de variantes reportadas ya sea como variante natural o secuencia en conflicto, del total de los 90 cambios encontrados.

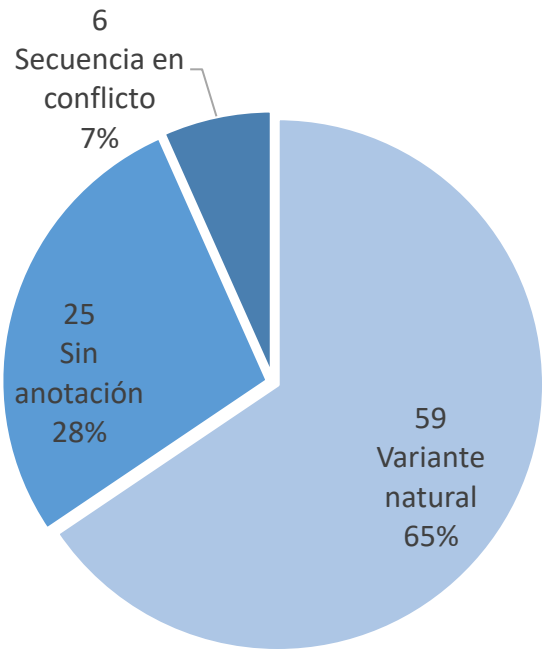


Figura 34.- Gráfica de pastel de la anotación encontrada en UniProt.

En la Tabla 24 se colocan algunos ejemplos de estas variantes reportadas.

Tabla 24. SNP relacionados a sitios de enlace disulfuro o reporte de variación natural.

Gen	Posición en proteína	Aminoácidos	Tamaño de la proteína	Numero de Variante natural (UniProt)	The Single Nucleotide Polymorphism database
DMD	2937	R/Q	3685	VAR_005171	Q → R dbSNP:rs1800280
A2M	639	N/D	1,474	VAR_026820	N → D dbSNP:rs226405
EDA2R	129	T/A	299	VAR_044512	T → A dbSNP:rs1385698
PRRG3	153	N/S	231	VAR_046712	N → S dbSNP:rs4323608
SLC39A4	372	V/L	647	VAR_057483	V → L dbSNP:rs1871534
SLC26A2	574	I/T	739	VAR_058415	I → T dbSNP:rs30832
A2ML1	850	D/E	1,454	VAR_059083	D → E dbSNP:rs1860926

Esto dio pauta para verificar la base de datos usada de los 1000 genomas con hg38, encontrándose que no están reportados todos los SNP que reportan en el proyecto con el genoma hg19. De los sitios encontrados, se cambiaron sus coordenadas genómicas del genoma de referencia hg38 a la versión anterior hg19, con ayuda de la herramienta LiftOver herramientas de software para convertir coordenadas de un ensamblaje a otro (Kuhn, Haussler, & Kent, 2012). Se realizó una búsqueda de los SNP encontrados, en los 1000 genomas con hg19 para corroborar que no existiera variación en los humanos actuales. De los 90 cambios solo 28 no presentan variación en la población actual y estos se relacionan a 26 genes, los cuales se desglosan en la Tabla 25.

Tabla 25. SNP que se encuentran en regiones codificantes y generan un cambio no sinónimo.

chr	Coordenada hg38	Coordenada hg37	Alelo antiguo	Frecuencia	Alelo actual	Frecuencia	Variación	Gen
3	195785236	195512107	A	1	T	1	missense_variant	MUC4
5	141101254	140480838	G	0.875	C	1	missense_variant	PCDHB3
5	149981314	149360877	C	0.875	T	1	missense_variant	SLC26A2
14	22086905	22555182	T	0.875	C	1	missense_variant	TRAV23DV6
16	74391650	74425548	G	0.8125	A	1	missense_variant	NPIPB15
1	41512895	41978566	G	0.75	T	1	missense_variant	HIVEP3
5	35708993	35709095	T	0.75	C	1	missense_variant	SPEF2
11	130910330	130780225	A	0.75	C	1	missense_variant	SNX19
5	132813992	132149684	C	0.6875	G	1	missense_variant	SOWAHA
6	29555899	29523676	G	0.6875	C	1	missense_variant	UBD
3	75738575	75787726	A	0.625	G	1	missense_variant	ZNF717
3	195785171	195512042	C	0.625	T	1	missense_variant	MUC4
5	474989	475104	G	0.625	A	1	missense_variant	SLC9A3
7	142801067	142498751	G	0.625	A	1	missense_variant	TRBC2
9	128656549	131418828	C	0.625	A	1	missense_variant	WDR34
11	125961075	125830970	T	0.625	A	1	missense_variant	CDON

12	10434512	10587111	G	0.625	A	1	missense_variant	AC068775.2
12	10434512	10587111	G	0.625	A	1	missense_variant	KLRC2
15	23440196	23685343	C	0.625	T	1	missense_variant	GOLGA6L2
22	36028402	36424450	C	0.625	A	1	missense_variant	RBFOX2
1	16057247	16383742	G	0.5625	C	1	missense_variant	CLCNKB
2	95938843	96604591	G	0.5625	A	1	missense_variant	ANKRD36C
3	75738859	75788010	C	0.5625	A	1	missense_variant	ZNF717
5	78129204	77425028	T	0.5625	A	1	missense_variant	AP3B1
5	116005941	115341638	C	0.5625	G	1	missense_variant	LVRN
5	141956699	141336264	T	0.5625	G	1	missense_variant	PCDH12
12	103478394	103872172	G	0.5625	T	1	missense_variant	C12orf42
X	55146104	55172537	A	0.5625	G	1	stop_gained	FAM104B

3.2.2.2.1.2 Genes Ortólogos

En el Anexo 9 se presenta la lista completa de los genes ortólogos relacionados a los 26 genes encontrados bajo esta metodología. En la Tabla 26 se muestran los genes ortólogos de los 18 genes los cuales presentan ortólogo a S. cerevisiae.

Tabla 26. Genes ortólogos de cada gen.

Gen humano	Gen ortólogo	% id del gen del modelo respecto al consultado	% id del gen consultado respecto al gen del modelo
PHB2	PHB2	50.5017	48.7097
UBA1	UBA1	50	51.6602
ATP11C	DRS2	32.1555	26.8635
PUDP	YKL033W-A	30.2789	32.2034
LVRN	AAP1	22.7273	26.285
LVRN	APE2	22.3232	23.2143
TAF7L	TAF7	20.7792	16.2712
ARL13A	ARL3	20.7031	26.7677
PPP4R3C	PSY2	18.9904	18.4149
GYG2	GLG1	18.7625	15.2597
AP3B1	APL6	18.7386	25.3399
GYG2	GLG2	17.9641	23.6842
CLCNKB	GEF1	17.3217	15.276
MAP3K15	BCK1	17.0602	15.1556
SLC9A3	NHX1	15.7074	20.6951
SLC39A4	YKE4	11.2828	21.0983
MAP3K15	STE11	10.2818	18.8285
ATRX	RDH54	6.62119	17.2234
BRWD3	MDV1	5.88235	14.8459
BRWD3	CAF4	5.10544	14.3079

3.2.2.2.2 Stop gained

El SNP causante de variación *stop gained*, están relacionados al gen FAM104B como se muestra en la Tabla 27.

Tabla 27. SNP que se encuentran en región codificante y generan un cambio *stop gained*.

Chr	Posición	Alelo antiguo	Frecuencia	Alelo moderno	Frecuencia	Gen
X	55146104	A	0.5625	G	1	FAM104B

Estos cambios repercuten en diferentes transcritos del gen, como se muestra en la Tabla 28.

Tabla 28. SNP que se encuentran en regiones codificantes y generan un cambio *stop gained*.

Variación	Gen	Transcrito	Exón	posición cDNA	posición CDS	Posición en proteína	Aminoácidos	Codón	Hebra
X 55146104 G/A	FAM104B	ENST00000425133	3/3	370	331	111	R/*	CGA/TGA	-1
		ENST00000477847	3/3	453	319	107			
		ENST00000489298	3/3	546	325	109			

El gen FAM104B es un gen con una función actualmente desconocida, por lo que no se podría establecer un vínculo causal en el cambio sinónimo encontrado. El SNP encontrado se relaciona a 3 transcritos ENST00000425133, ENST00000477847 y ENST00000489298, que presentan 116, 112 y 114 aminoácidos, respectivamente. Con el codón de paro en cada uno de ellos se estarían perdiendo 6 aminoácidos, lo cual posiblemente no esté afectando la función de la proteína. Dentro del análisis de genes ortólogos, el gen FAM104B no presenta ortólogo respecto al organismo modelo propuesto.

3.3 Gen AK2

La ventana del gen AK2 (cromosoma 1 en la posición: 33013300-33013400), no se localizó en el análisis de las diferencias de la diversidad genética (en el corte del 0.00001%) y tampoco alguno de los SNP encontrados se identificó en el análisis de las diferencias entre frecuencias de alelos. Revisando los datos, dicha ventana presenta un valor de π del 0.012, θ con 1.507 y D de Tajima del 1.793, con lo cual esta ventana se encuentra entre el p valor del 0.001 para los tres estimadores.

3.3.1 Complementación heteróloga

3.3.1.1 Cepas knockout

Para verificar el fenotipo y adecuación de las cepas knockout generadas (*Δadk1*, *Δadk2* y la doble *knockout*), se utilizó el ensayo de spots en dos medios diferentes. En este ensayo se usó los medios YPAD e YPGE, en donde se puede observar en la Figura 35, que la cepa *Δadk2* presenta un fenotipo parecido al de la cepa silvestre, mientras que la cepa *Δadk1,* que de acuerdo a Gauthier et al., (2008) puede presentar una reducción de la actividad de la proteína de hasta un 15%, teniendo un fenotipo deletéreo en crecimiento. Así mismo, se puede inferir que el fenotipo de la doble mutante es similar al de *Δadk1*, de acuerdo con el ensayo de spots que se muestra en la figura 35, donde no se observan diferencias en el crecimiento. Al ser *ADK1* un gen esencial para la proliferación celular (Konrad, 1988), su deleción presenta un fenotipo deletéreo, mientas que el gen *ADK2* al no ser esencial (Gu, Gordon, Amutha, & Pain, 2005), no presenta fenotipo visible en el crecimiento de la cepa knockout correspondiente.

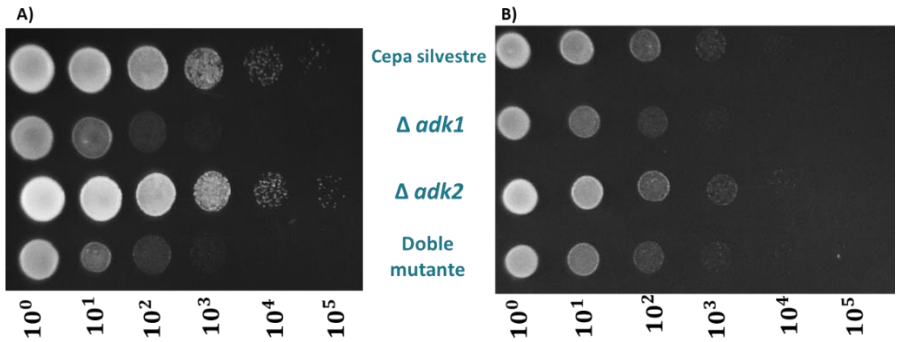


Figura 35.- Ensayo de spots. A) Medio fermentable YPAD. B) Medio no fermentable YPGE.

El siguiente procedimiento fue realizar un análisis más cuantificable del comportamiento de las cepas. Para ello se obtuvieron curvas de crecimiento en diferentes medios (Tabla 9), como se muestra en el Anexo 10, realizando mediciones de densidad óptica a cada hora por más de 2 días. Se observó que la cepa *Δadk1* presenta un crecimiento lento y también que en algunas condiciones su fenotipo es deletéreo, mientras que la cepa *Δadk2* presenta el mismo o muy similar comportamiento a la cepa silvestre, como se ejemplifica en la Figura 36, en la cual se muestran los efectos bajo diferentes fuentes de carbono.

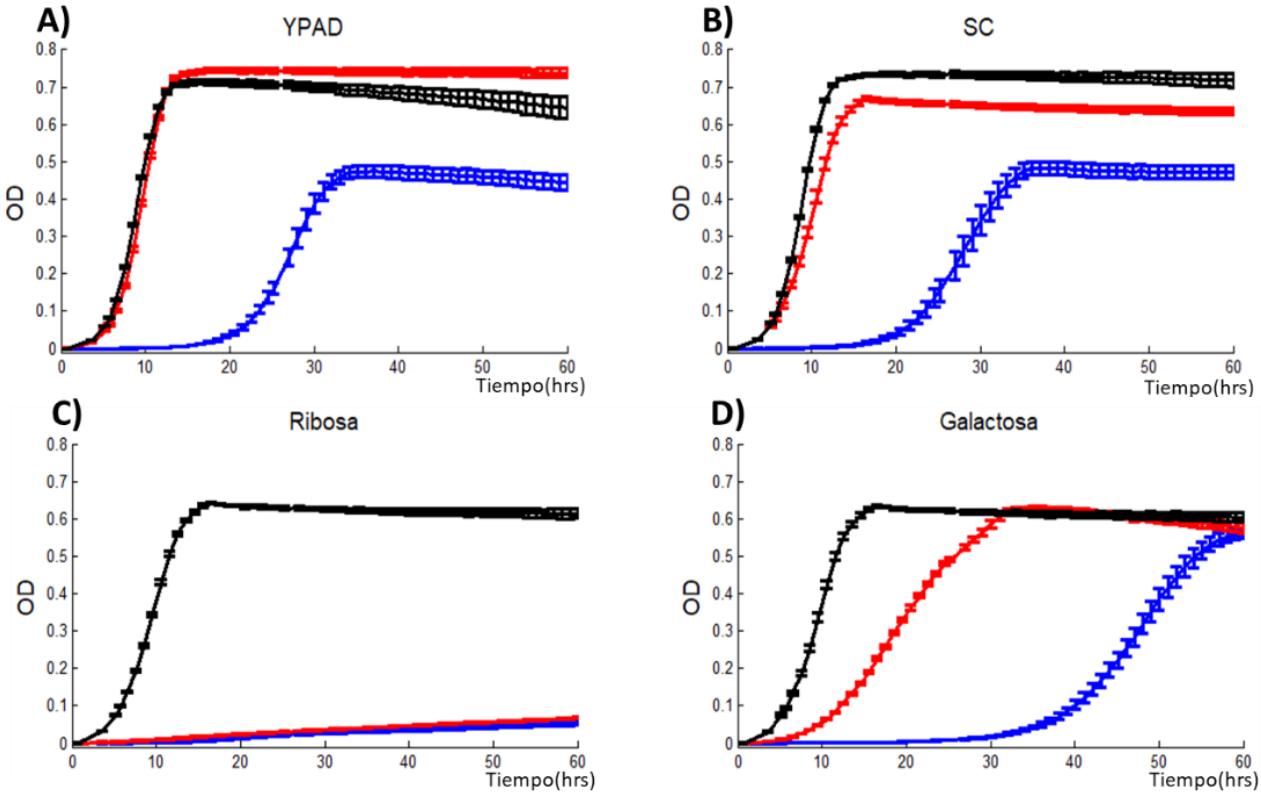


Figura 36.- Curvas de crecimiento en diferentes fuentes de carbono. Crecimiento de las cepas Δadk1 (línea azul), Δadk2 (línea roja) y silvestre (línea negra) en los medios YPAD, SC, Ribosa y Galactosa. Se puede apreciar que la cepa Δadk2 puede comportarse como la cepa silvestre (A) o tener una reducción de su tasa de crecimiento (B y D), mientras que la Δadk1 tiene una tasa crecimiento más lenta en comparación a las otras dos cepas (A, B y D). Solo se presentó un caso en donde ambas mutantes tienen un fenotipo letal ante una fuente de carbono específica (C).

3.3.1.2 Complementación

Las cepas knockout transformadas con los genes humanos se analizaron en medio SC-Ura con y sin doxiciclina, con sus respectivos controles como se muestra en las figuras 37 y 38. En la figura 37-A, se puede observar la restauración del fenotipo silvestre en la cepa *Δadk1* con el vector pCM189 que contiene la secuencia del gen *ADK1* (línea azul eléctrico). Además, se puede apreciar que las cepas *Δadk1* con el vector pCM189-AK2 variante antigua y la variante actual, no llevan a cabo la complementación (línea amarilla limón y azul celeste respectivamente), ya que se puede ver que estas se comportan igual que la cepa *knockout* con el vector vacío (línea gris). En la figura 37-B, se puede notar como la doxiciclina apaga la expresión del gen ADK1, al reducir su curva de crecimiento en comparación con la cepa silvestre (línea amarillo canario).

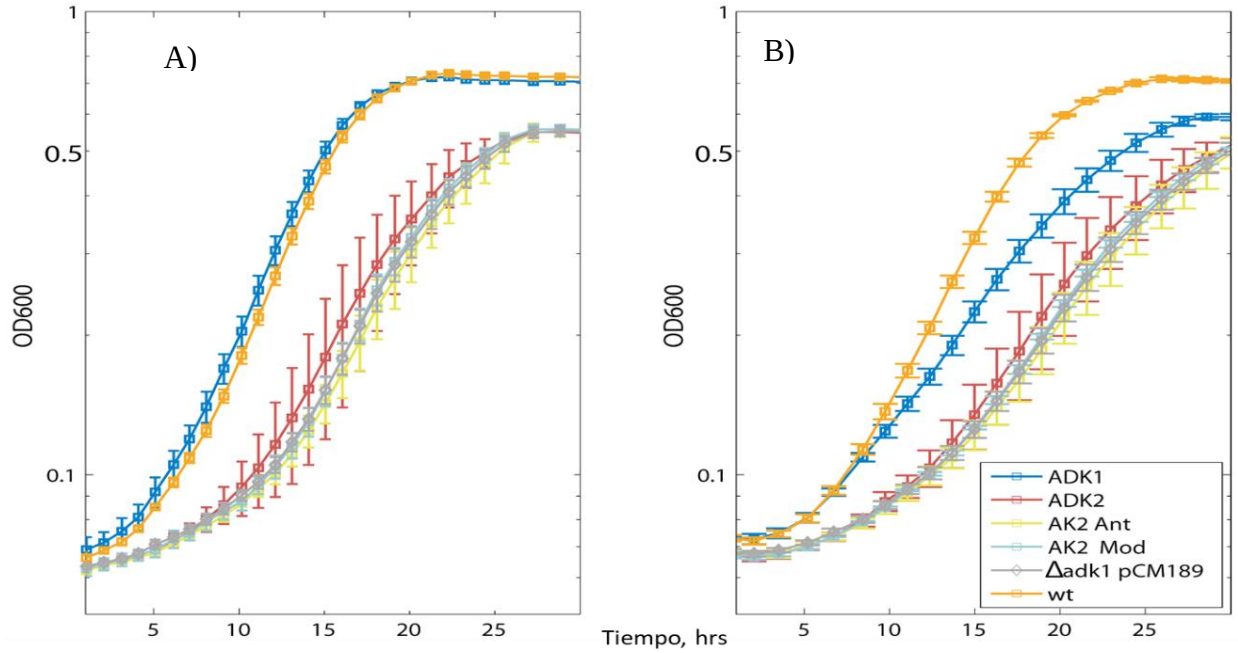


Figura 37.- Curvas de crecimiento de la cepa Δadk1. A) Medio SC-Ura. B) Medio SC-Ura con doxiciclina.

En la figura 38 se puede percatar en ambas graficas (A y B), que todas las cepas se comportan de manera similar. Al ser el gen *ADK2* un gen no esencial, no existe un fenotipo deletéreo (comportamiento igual que la cepa silvestre), además no hay un aumento o disminución del crecimiento de las cepas complementadas con los genes humanos. En otras palabras, no existen diferencias significativas entre las cepas.

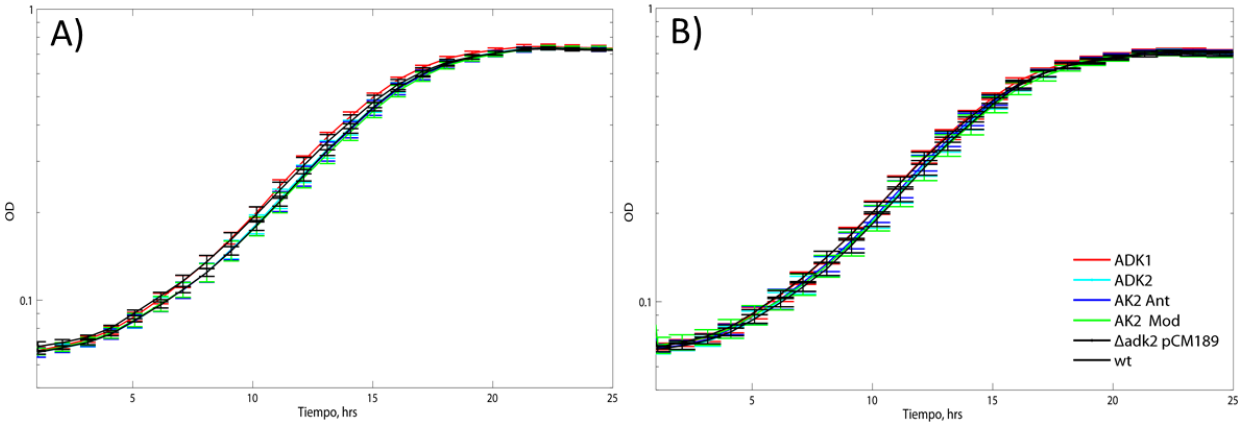


Figura 38.- Curvas de crecimiento de la cepa Δadk2. A) Medio SC-Ura. B) Medio SC-Ura con doxiciclina.

Capítulo IV. Discusión

4.1 Reducción de la diversidad

Los patrones de variación de todo el genoma se pueden resumir por medio de una o más pruebas estadísticas, que se usan para clasificar loci en el genoma. Para este estudio se utilizó los estimadores π, θ y D de Tajima, y realizando las pruebas estadísticas correspondientes se pudo observar que existen diferencias significativas de los estimadores entre la población antigua y la población actual. Además, existe una reducción de la diversidad que se puede deber a efectos demográficos, como cuellos de botella durante la invención de la agricultura, como se tiene reportado en poblaciones africanas (Henn et al., 2012). En los humanos antiguos el aumento de la diversidad no se le atribuye al llamado de los SNP, siendo uno de los filtros usados para el procesamiento de los datos al ser eliminado dicho factor. Los datos se filtraron con una calidad de mapeo phred mínima de 20, que de acuerdo con Zeng et al., (2017), los puntajes de calidad de más de 20 son generalmente considerados como aceptables. En el caso de una calidad de mapeo phred de 20 se reduce el error al 1% y aumentar la exactitud en esta calidad al 99%.

Para el análisis se estableció un valor arbitrario para la selección de ventanas, al ser el valor más alejado de la distribución y basándonos en lo que menciona Luca et al. (2010); Aquellos loci que caen por encima de un límite arbitrario se identifican como valores atípicos, bajo el supuesto de que la mayoría de los loci en el genoma humano evolucionan de forma neutral. Entonces aquellos valores que se encontraban por encina del pvalor del 0.00001% de la distribución de los datos se consideran valores inusuales y representan blancos candidatos de presiones selectivas. Se escogieron las ventanas que fueran valores atípicos dentro de los tres estimadores, esto para asegurar de que realmente estábamos seleccionando las ventanas que se encontraban bajo selección direccional, siendo un parámetro riguroso para asegurar dicha selección. Podría existir el caso en el que un SNP se localice en la lista de valor atípico de solo dos estimadores, de los tres usados, que pudiera ser un buen candidato, y se pudiera considerar para análisis posteriores.

Kimura (1983), menciona que el azar es lo que moldea las frecuencias alélicas, por ello la relación que existe entre π y θ permite determinar si las secuencias se encuentran bajo este modelo o se desvían del mismo. Además, con la prueba de D de Tajima se están seleccionado las regiones que se desvían del modelo del Kimura y se corrobora la relación entre π y θ (diferencias entre ambos

estimadores). Es decir, cuando los valores encontrados fuera de la distribución (outliers) de la prueba de D de Tajima, se está asegurando que los cambios no son azarosos y podrían tener un impacto en la adecuación.

4.2 Diferencias en la diversidad genética entre la población antigua y la actual

Con la metodología de las diferencias en la diversidad genética entre la población antigua y la actual, se observó mayores diferencias de diversidad en ventanas de regiones no codificantes (mayoritariamente en regiones intrónicas), en relación a las ventanas de las regiones codificantes. Por ello, se sugiriere que los cambios alélicos se presentan con mayor frecuencia en regiones regulatorias. Esto concuerda con el análisis a nivel genómico que se hizo en roedores, donde se demostró que en promedio los intrones evolucionan sustancialmente más rápido que las regiones intergénicas de DNA (Keightley & Gaffney, 2003) y también con el análisis de comparación de los genomas de humano y ratón, encontrando que existe una relajación en las regiones conservadas no codificantes mucho más profunda que en las regiones codificantes de proteínas (Kryukov et al., 2005). A pesar de que la mayor variación se encuentra en regiones no codificantes (98.5%), se observó variación en regiones codificantes (como lo fueron cambios no sinónimos y sinónimos, siendo el 1% y 0.5% respectivamente). Con ello se identificaron polimorfismos de un solo nucleótido en 2 genes, PSMD13 y NDUFS7, con 9 y 6 cambios no sinónimos respectivamente, posibles genes candidatos de haber sido seleccionados por un cambio diferencial. Las regiones no codificantes también son de interés para estudios posteriores, ya que posiblemente las fuerzas de selección actúan en mayor medida sobre la regulación génica y en menos en las regiones codificantes, ya que se pudo observar en la Figura 25, las ventanas con valores atípicos se encontraban en regiones 5’ o 3’, variantes de transcripción de genes de ARN no codificante, entre otros.

En el gen PSMD13, que está relacionado con el complejo que se encarga de la degradación dependiente de ATP de proteínas ubiquitinadas, la mayoría de las variaciones encontradas de acuerdo con SIFT y PolyPhen no tienen un impacto en la proteína. Estos SNP se encuentran repartidos en dos transcritos de *splicing* alternativo del gen. El SNP 9 (cromosoma 11 posición 244197 y alelos T/C) es el único que podría tener un impacto en el funcionamiento de la proteína por los valores que arrojan SIFT y PolyPhen, además de que los cambios en las propiedades fisicoquímicas son muy notorios. Además, no se puede determinar si estos sitios están afectando

la función de la proteína ya que no se tienen reportados los sitios con los cuales la proteína lleva acabo su función e igualmente no se tiene cristalizada la proteína. En este caso, la ventana de ambas poblaciones presenta gran diversidad lo que hace que sea un candidato al realizar la diferencia entre estimadores.

En el gen NDFS7, que forma parte de la cadena respiratoria mitocondrial, los SNP encontrados se encuentran en siete transcritos de la proteína y uno de ellos (cromosoma 19 posición 1390913 y alelos G/A), que puede tener un impacto en el funcionamiento de la proteína, ya que se encuentra cerca de dos nucleótidos de unión a metales, modificación que podría fundamentar los valores de SIFT y PolyPhen, que sugieren que la variante ancestral podría afectar dicha función. Analizando los SNP de la ventana entre poblaciones, se puede observar que los humanos actuales no presentan variación alguna, mientras que en la población antigua existen 14 sitios con variación nucleotídica, de los cuales solo 6 generan un cambio no sinónimo.

4.3 Diferencia entre frecuencias nucleotídicas

Esta metodología se realizó porque con la metodología de las diferencias de la diversidad genética se encontraron pocas regiones codificantes, además, existían regiones que no se estaban tomando en cuenta con la metodología anterior. Por ejemplo, si todas las secuencias muestreadas de la población A son idénticas, el estimador de diversidad da un valor de cero, ya que no existe diversidad (Tabla 29, población A), pero puede ocurrir que, en esa misma ventana en la población B tenga una secuencia distinta a la población A, pero idéntica en todas sus muestras, dando también un valor de 0 diversidad (Tabla 29, población B). Por lo tanto, desde el punto de vista de la diversidad no habría diferencia entre las poblaciones en esta ventana, a pesar de que a nivel de secuencia sí haya diferencias. Esta posibilidad se muestra en la Tabla 29, en la que se aprecia que existe un SNP en la posición 2, en la que la población A presenta el nucleótido adenina, mientras que la población B presenta citosina; ambas poblaciones con diversidad nula, pero existen diferencias entre la población A, respeto a la B.

Tabla 29. Ejemplo de una región sin diversidad entre muestras de la misma población, pero con diferencias en la frecuencia nucleotídica entre poblaciones.

	Población A						**Población B**					
Humano 1	A	A	T	G	T	**π=0**	A	C	T	G	T	**π=0**
Humano 2	A	A	T	G	T		A	C	T	G	T	
Humano 3	A	A	T	G	T	Frecuencia= 1 Respecto a A	A	C	T	G	T	Frecuencia= 1 Respecto a C
Humano 4	A	A	T	G	T		A	C	T	G	T	
Posición	1	2	3	4	5		1	2	3	4	5	

Con esta metodología se encontraron variaciones puntuales, y más SNP en regiones codificantes en comparación con la búsqueda de polimorfismos a partir de la diferencia en la diversidad genética. Bajo esta estrategia se encontraron 90 cambios no sinónimos, 87 cambios sinónimos y 3 cambios *stop gained.*

4.3.1 Variaciones missense

Revisando los SNP encontrados se halló que algunos de ellos se reportaban como variantes naturales y presentan variación en los humanos actuales en el proyecto de los 1000 genomas con hg19, inconsistencia con la base de datos usada para este análisis (1000 genomas, hg38). Realizando una verificación de ellos, solo 28 SNP no presentaban variación en la población actual (verificando en la base de datos de los 1000 genomas con hg19), que se relacionan a 25 genes. Esta inconsistencia se cree que pudo haber sido durante el cambio de coordenadas genómicas del genoma hg19 al hg38, y no se capturo todas las variantes que se encuentran reportadas en la base de datos de los 1,000 genomas.

4.3.2 Variación stop gained

El gen FAM104B tienen el codón de paro casi al final de la secuencia de codificación (solo se pierden 6 aminoácidos), de modo que posiblemente no esté afectando drásticamente la función de esta proteína, pero no se conoce la función del gen, información de la proteína, no se tiene cristalizada la proteína y no se tienen ortólogos en el modelo propuesto para ver si este cambio está perturbando a la proteína. Este gen también se verifico en la base de datos de los 1,000 genomas con hg19 para saber si no existía variación en la población actual.

4.4 Complementación heteróloga en *S. cerevisiae*

Se tiene una lista de 28 genes, posibles candidatos a la adaptación al entorno agrícola, encontrados por dos metodologías descritas previamente (Por las diferencias entre la diversidad genética y por las diferencias entre frecuencias nucleotídicas), por consiguiente, el próximo paso es la selección de alguno o algunos de ellos para poder hacer el análisis de las diferencias fenotípicas y funcionales, ver cómo es el impacto de esta variación en la eficiencia de la proteína y con ello proponer cómo es que la selección actuó en esta región. Por eso de cada uno de los genes candidatos se generó la lista de los genes ortólogos para *S. cerevisiae.* Se observó que se presentan pocos genes ortólogos de humano respecto al modelo propuesto.

4.4 Complementación heteróloga del gen AK2 en *S. cerevisiae*

En trabajos previos de búsqueda de selección en el genoma humano del grupo de Interacción Núcleo Mitocondrial y Paleogenómica, se encontró como candidato al gen *AK2*. Por contar con las características apropiadas para realizar el ensayo de complementación (tamaño, presentar ortólogo respecto a *S. cerevisiae* y función similar), se utilizó este gen para implementar la metodología de complementación heteróloga en *S. cerevisiae.* Esta metodología se pretende utilizar con los demás genes encontrados en este análisis, que tengan las características adecuadas, y así, elucidar como están afectando las variantes alélicas en el humano.

Para ello se realizó la fenotipificación de las cepas knockout sencillas en diferentes condiciones, observando que el gen *ADK1* es importante para el desarrollo óptimo de la cepa. Se pudo apreciar que la cepa *Δadk2* se comporta como la cepa silvestre en medio YPAD, tener una reducción de su tasa de crecimiento (como fue en el medio a base de galactosa). La cepa *Δadk1* tiene una tasa crecimiento más lenta en comparación a las cepas silvestre y *Δadk2,* que se puede apreciar en general en los medios probados (YPAD y SC, por ejemplo). Solo se presentó un caso en donde ambas mutantes tienen un fenotipo letal, el cual fue ante la fuente de carbono ribosa.

Se consiguió construir los plásmidos que presentaran las variantes alélicas humanas y sus ortólogos en levadura expresados bajo un promotor reprimible (tetO7). El sistema de represión por Doxiciclina es eficiente, ya que permitió silenciar de manera gradual el gen *ADK1* que

complementó a la cepa knockout *Δadk1* (Figura 37). No se observó la complementación de los genes humanos con respecto a las cepas knockout, ya que al tener transformada *S. cerevisiae* con los genes humanos, no se alcanzaron a ver diferencias en las tasas de crecimiento, es decir las transformantes no recuperaron el fenotipo silvestre (Figura 37). La falta de complementación podría deberse a que la localización subcelular del gen humano (la cual se encuentra en el espacio intermembranal de la mitocondria), respecto al de levadura (que se encuentra en la mitocondria y en el citosol), difiere como se muestra en la figura 39. Posiblemente se esté expresando la proteína y a pesar de ello, ésta no puede complementar a su ortólogo de levadura por esta diferencia de localización subcelular. También podría ser el caso en el que realmente el gen humano no puede llevar esta complementación. Para este gen se propone usar otro modelo de clonación para ver si existen variaciones funcionales entre alelos.

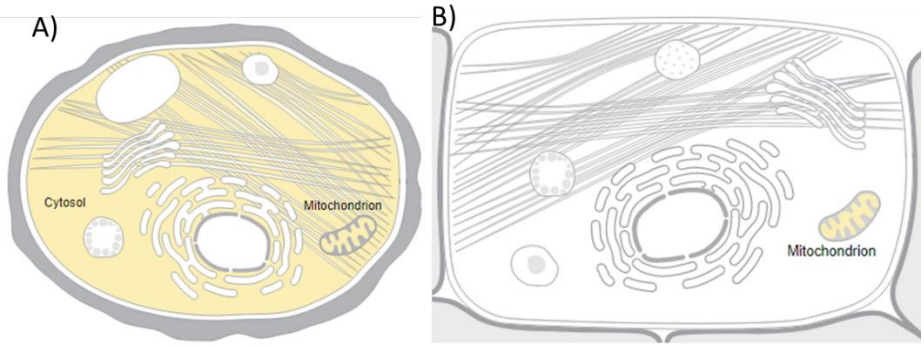


Figura 39. Localización subcelular. A) La proteína ADK1 se encuentra en la mitocondria y en el citosol. B) La proteína AK2 se encuentra expresada en el espacio intermembranal de la mitocondria

Se considera que se debe tener más de un análisis funcional, ya que existen genes que no se expresan de la misma manera *in vivo* que *in vitro*. De manera que, es importante tener ambos análisis, para poseer un argumento más sustentado de lo que ocurre cuando hay cambios alélicos dentro de la secuencia de un gen. De tal forma, generar evidencia funcional para tratar de tener una comprensión completa de las consecuencias funcionales de la selección y como ha afectado la variación genética en los humanos. Existen líneas celulares de eucariotas (de humanos, ratones y homínidos) que se pueden usar para el análisis de diferencias funcionales, si es que pudieran ser factibles, podrían ayudar a identificar qué impacto tienen los cambios alélicos en dichas líneas celulares.

Conclusiones

Las diferencias de diversidad observadas entre poblaciones son significativamente diferentes y se puede apreciar una reducción de esta al comparar la población antigua respecto a la actual, debida a efectos demográficos. La mayor diversidad encontrada en los humanos antiguos no se les atribuye a errores de selección de SNP.

Se realizaron dos métodos para la identificación de regiones bajo selección, por medio de las diferencias en la diversidad genética entre la población antigua y la actual, y la diferencia entre frecuencias nucleotídicas. Con la primera metodología (diferencias en la diversidad genética entre la población antigua y la actual) se encontraron dos ventanas bajo selección. Para su identificación se usó el criterio en el cual si la venta fuera un valor atípico en los tres estimadores (π, θ y D-Tajima), es una región candidata a selección. Este criterio es un poco riguroso, pero con ello se está asegurando que realmente se están escogiendo las ventanas bajo selección direccional. Sería interesante ver el comportamiento de las demás ventanas de acuerdo a su robustez, siendo los más robustos los que estén en las tres listas de los estimadores usados, seguidos de los que estén en dos y finalmente los que solo estén en una. También se podría disminuir el valor del corte para conocer las regiones que pudieran ser candidatas para análisis posteriores. Se hallaron pocos alelos por la metodología de las diferencias de la diversidad genética en regiones codificantes, y estos alelos se relacionaron con dos genes candidatos PSMD13 y NDUFS7, presentando 9 y 6 SNP respectivamente. Solamente del gen PSMD13 presenta un ortólogo en el modelo de *S.* cerevisiae.

Con la segunda metodología (diferencia entre frecuencias nucleotídicas) se identificaron variaciones puntuales, y más SNP en regiones codificantes, en comparación con la anterior, considerándose como una buena estrategia a nivel genoma, para la identificación de variación. Con la metodología de las diferencias entre frecuencias de SNP, se encontraron más regiones candidatas, considerando ésta ultima una estrategia bastante eficiente en la identificación puntual de SNP, ya que se encontraron 28 SNP relacionados a 26 genes.

En ambas metodologías (diferencias en la diversidad genética entre la población antigua y la actual, y la diferencia entre frecuencias nucleotídicas), se ha encontrado que las regiones con mayor índice de diversidad son las regiones no codificantes, con ello se puede inferir que las fuerzas de selección actúan en mayor medida sobre estas regiones y en menor proporción en las regiones codificantes,

por ello son consideradas también significativas para su posterior estudio, ya que pudieran tener un impacto dentro de las regiones codificantes como regiones regulatorias, que no se tomaron en consideración en el presente trabajo.

El estudio de la variación funcional, con el efecto fenotípico de las distintas variantes alélicas de del gen AK2 se llevó acabo en el modelo de *S. cerevisiae*, desafortunadamente no se estableció la complementación por el gen humano en las cepas *Δadk1* y *Δadk2*, debido a su localización subcelular o que el gen no puede llevar acabo la complementación del gen de levadura. El gen de levadura *ADK1* se encuentra en la mitocondria y en el citosol, mientras que el gen humano AK2 se encuentra únicamente en el espacio intermembranal de la mitocondria. A pesar de ello se logró implementar la metodología para posteriores análisis de diferencias funcionales entre alelos humanos en este modelo eucarionte. Montada la metodología, se puede seleccionar de la lista de genes candidatos algún gen de interés, que cumpla las características deseadas para realizar este ensayo, y ver si entre alelos existen variaciones funcionales.

Perspectivas

- ✓ Clasificar a los genes candidatos por su robustez, siendo los más robustos los que estén en las tres listas de los estimadores usados (π, θ y D de Tajima), seguidos de los que estén en dos y finalmente los que solo estén en una.
- ✓ Realizar el análisis de las regiones regulatorias que estén bajo selección direccional, por medio de aproximaciones bioinformáticas, con la finalidad de conocer si esta variación está impactando en la adecuación del humano.
- ✓ Analizar las variantes sinónimas ya que pudieran afectar la velocidad de traducción ya sea de manera positiva o negativa. Para elucidar esto se tomaría los reportes de optimización de codones para ver en qué manera afectan dichos cambios sinónimos.
- ✓ Realizar cinéticas enzimáticas de los distintos alelos de algún gen candidato para analizar si existen diferencias funcionales.
- ✓ De la lista de genes candidatos, realizar el análisis experimental de complementación y diferencias funcionales entre alelos, de uno o varios genes en líneas celulares humanas, en *M. musculus*, en *S. cerevisiae*, u otro organismo para complementar el análisis.

Bibliografía

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*(5), 711–722.

Antelope, C. X., Marnetto, D., Casey, F., & Huerta-Sanchez, E. (2017). Leveraging Multiple Populations across Time Helps Define Accurate Models of Human Evolution: A Reanalysis of the Lactase Persistence Adaptation. *Human Biology*, *89*(1), 81–97. Retrieved from http://www.bioone.org/doi/full/10.13110/humanbiology.89.1.05

Bae, C. J., Douka, K., & Petraglia, M. D. (2017). On the origin of modern humans: Asian perspectives. *Science*, *358*(6368), eaai9067. Retrieved from http://science.sciencemag.org/content/358/6368/eaai9067.full

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., & Cullin, C. (1993). A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Research*, *21*(14), 3329. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC309783/pdf/nar00063-0176.pdf

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. Retrieved from http://dx.doi.org/10.1093/nar/28.1.235

Bottjer, D. J., Davidson, E. H., Peterson, K. J., & Cameron, R. A. (2006). Paleogenomics of echinoderms. *Science*, *314*(5801), 956–960.

Brunet, T. D. P., & Doolittle, W. F. (2018). The generality of Constructive Neutral Evolution. *Biology & Philosophy*, *33*(1–2), 2.

Budisa, N. (2006). *Engineering the Genetic Code: Expanding the Amino Acid Repertoire for the Design of Novel Proteins*. Wiley. Retrieved from https://books.google.com.mx/books?id=U_PbUatJ0-sC

Campbell, B. (2017). *Human evolution: an introduction to man's adaptations*. Routledge.

Campbell, N. A. (2001). *Biología: Conceptos y relaciones*. Pearson Educación. Retrieved from https://books.google.com.mx/books?id=NI2qFwNNYX4C&pg=PA407&lpg=PA407&dq=La+agricultura,+por+tanto,+abrió+el+desarrollo+de+la+industria,+la+siguiente+gran+ola+cu

ltural,+que+continúa+hasta+nuestros+días&source=bl&ots=Xn1tGdsVtC&sig=2W0mY2nc0ZGVmk5rC9tZ3hUFx

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, *33*, 266.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. Retrieved from https://academic.oup.com/nar/article/38/6/1767/3112533

Consortium, 1000 Genomes Project. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68. Retrieved from https://www.nature.com/articles/nature15393

Consortium, U. (2016). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169.

Cooper, D. N., & Kehrer-Sawatzki, H. (2008). *Handbook of human molecular evolution*. John Wiley.

Crittenden, A. N., & Schnorr, S. L. (2017). Current views on hunter-gatherer nutrition and the evolution of the human diet. *American Journal of Physical Anthropology*, *162*(July 2016), 84–109. https://doi.org/10.1002/ajpa.23148

Crosby, A. W. (1999). *Visitando de nuevo Pangea. El neolítico reconsiderado, Imperialismo Ecológico*. (E. C. Barcelona, Ed.). Barcelona:

Culetto, E., & Sattelle, D. B. (2000). A role for Caenorhabditis elegans in understanding the function and interactions of human disease genes. *Human Molecular Genetics*, *9*(6), 869–877. Retrieved from https://academic.oup.com/hmg/article/9/6/869/618668

Diamond, J., & Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science*, *300*(5619), 597–603.

Gauthier, S., Coulpier, F., Jourdren, L., Merle, M., Beck, S., Konrad, M., … Pinson, B. (2008). Co-regulation of yeast purine and phosphate pathways in response to adenylic nucleotide variations. *Molecular Microbiology*, *68*(6), 1583–1594. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2008.06261.x/epdf

Gietz, R. D., & Schiestl, R. H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols*, *2*(1), 35. Retrieved from https://search.proquest.com/docview/1041216778?pq-origsite=gscholar

Gokcumen, O. (2018). The Year In Genetic Anthropology: New Lands, New Technologies, New

Questions. *American Anthropologist*, *120*(2), 266–277. Retrieved from https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1111/aman.13032

GROUCUTT, H. (2018). Modern Humans: Their African Origin and Global Dispersal. *PaleoAnthropology*, *3*, 6.

Gu, Y., Gordon, D. M., Amutha, B., & Pain, D. (2005). A GTP: AMP Phosphotransferase, Adk2p, in Saccharomyces cerevisiae ROLE OF THE C TERMINUS IN PROTEIN FOLDING/STABILIZATION, THERMAL TOLERANCE, AND ENZYMATIC ACTIVITY. *Journal of Biological Chemistry*, *280*(19), 18604–18609.

Hall, T., Biosciences, I., & Carlsbad, C. (2011). BioEdit: an important software for molecular biology. *GERF Bull Biosci*, *2*(1), 60–61.

Hamza, A., Tammpere, E., Kofoed, M., Keong, C., Chiang, J., Giaever, G., … Hieter, P. (2015). Complementation of yeast genes with human genes as an experimental platform for functional testing of human genetic variants. *Genetics*, *201*(3), 1263–1274. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4649650/pdf/1263.pdf

Harris, D. R., & Fuller, D. Q. (2014). Agriculture: definition and overview. In *Encyclopedia of global archaeology* (pp. 104–113). Springer.

Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. *Proceedings of the National Academy of Sciences*, *109*(44), 17758–17764. Retrieved from http://www.pnas.org/content/109/44/17758.short

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). ancient DNA. *Nature Reviews Genetics*, *2*(5), 353.

Jones, E. R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., … Gamba, C. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, *6*, 8912. Retrieved from https://www.nature.com/articles/ncomms9912.pdf

Kachroo, A. H., Laurent, J. M., Yellman, C. M., Meyer, A. G., Wilke, C. O., & Marcotte, E. M. (2015). Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, *348*(6237), 921–925.

Keightley, P. D., & Gaffney, D. J. (2003). Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proceedings of the National Academy of Sciences*, *100*(23), 13402–13406.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Knapp, M., Lalueza-Fox, C., & Hofreiter, M. (2015). Re-inventing ancient human DNA.

Investigative Genetics, *6*(1), 4.

Konrad, M. (1988). Analysis and in vivo disruption of the gene coding for adenylate kinase (ADK1) in the yeast Saccharomyces cerevisiae. *Journal of Biological Chemistry*, *263*(36), 19468–19474.

Kryukov, G. V, Schmidt, S., & Sunyaev, S. (2005). Small fitness effect of mutations in highly conserved non-coding regions. *Human Molecular Genetics*, *14*(15), 2221–2229.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2012). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073.

Lalueza-Fox, C., & Gilbert, M. T. P. (2011). Paleogenomics of archaic hominins. *Current Biology*, *21*(24), R1002–R1009.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., … Lipson, M. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409. Retrieved from https://media.nature.com/original/nature-assets/nature/journal/v513/n7518/extref/nature13673-s1.pdf

Li H, Handsaker B, Danecek P, McCarthy S, M. J. (2018). Bcftools-utilities for variant calling and manipulating VCFs and BCFs. Retrieved from https://samtools.github.io/bcftools/bcftools.html

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. Retrieved from https://watermark.silverchair.com/btp352.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAcowggHGBgkqhkiG9w0BBwagggG3MIIBswIBADCCAawGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMjFphRfDAQfdibMb8AgEQgIIBfUKb-AzScC2QZwoi9aKojS-Tx7gJuzgoFwl6wMH2fSxLe5P6

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. Retrieved from https://genome.cshlp.org/content/18/11/1851.full.pdf+html

Liu, W. Y., Hsiao, H.-I., & Dai, S. Y. (2015). Genomic analysis with MapReduce. In *Big Data (Big Data), 2015 IEEE International Conference on* (pp. 1330–1335). IEEE. Retrieved from https://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7363891

Luca, F., Perry, G. H., & Di Rienzo, A. (2010). Evolutionary adaptations to dietary changes. *Annual Review of Nutrition*, *30*, 291–314. Retrieved from https://www.annualreviews.org/doi/abs/10.1146/annurev-nutr-080508-141048

Luis E. Eguiarte, V. S. y X. A. (2007). *Ecología molecular*. México.

Mager, W. H., & Winderickx, J. (2005). Yeast as a model for medical and medicinal research. *Trends in Pharmacological Sciences*, *26*(5), 265–273.

Masenko, V. (2018). Amino Acids Guide.

Matthew B. Hamilton. (2009). *Population genetics. Population genetics.* UK: Wiley-Blackwell. https://doi.org/10.1111/j.1365-294X.1994.tb00109.x

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, *17*(1), 122. Retrieved from https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

Morozova, I., Flegontov, P., Mikheyev, A. S., Bruskin, S., Asgharian, H., Ponomarenko, P., … Gankin, Y. (2016). Toward high-resolution population genomics using archaeological samples. *DNA Research*, *23*(4), 295–310.

Münster, A., Knipper, C., Oelze, V. M., Nicklisch, N., Stecher, M., Schlenker, B., … Dresely, V. (2018). 4000 years of human dietary evolution in central Germany, from the first farmers to the first elites. *PloS One*, *13*(3), e0194862. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194862

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, *39*, 197–218. Retrieved from http://www.annualreviews.org/doi/pdf/10.1146/annurev.genet.39.073003.112420

Noma, T., Song, S., Yoon, Y.-S., Tanaka, S., & Nakazawa, A. (1998). cDNA cloning and tissue-specific expression of the gene encoding human adenylate kinase isozyme 21. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, *1395*(1), 34–39.

Olalde, I., Allentoft, M. E., Sánchez-Quinto, F., Santpere, G., Chiang, C. W. K., DeGiorgio, M., … Quilez, J. (2014). Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, *507*(7491), 225. Retrieved from https://www.nature.com/articles/nature12960.pdf

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., … Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.*, *38*, 645–679. Retrieved from https://www.annualreviews.org/doi/abs/10.1146/annurev.genet.37.110801.143214

Primrose, S. B., & Twyman, R. (2009). *Principles of genome analysis and genomics*. John Wiley & Sons.

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., … Metspalu, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, *505*(7481), 87–91. Retrieved from https://www.nature.com/articles/nature12736.pdf

Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., DeGiorgio, M., Stafford Jr, T. W., … Doyle, S. M. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, *506*(7487), 225. Retrieved from https://www.nature.com/articles/nature13025

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., … Gupta, R. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, *463*(7282), 757. Retrieved from https://www.nature.com/articles/nature08835.pdf

Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T. S., Moreno-Mayar, J. V., Poznik, G. D., … Moltke, I. (2015a). The ancestry and affiliations of Kennewick Man. *Nature*, *523*(7561), 455. Retrieved from https://www.nature.com/articles/nature13673.pdf

Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T. S., Moreno-Mayar, J. V., Poznik, G. D., … Moltke, I. (2015b). The ancestry and affiliations of Kennewick Man. *Nature*, *523*(7561), 455. Retrieved from https://www.nature.com/articles/nature14625.pdf

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., … Haeussler, M. (2014). The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, *43*(D1), D670–D681. Retrieved from https://academic.oup.com/nar/article/43/D1/D670/2439065

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., … Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Ségurel, L., & Bon, C. (2017). On the Evolution of Lactase Persistence in Humans. *Annual Review of Genomics and Human Genetics*, *18*. https://doi.org/10.1146/annurev-genom-091416-035340

Shapiro, áB, & Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function:

new insights from ancient DNA. *Science*, *343*(6169), 1236573. Retrieved from http://science.sciencemag.org/content/343/6169/1236573

Slatkin, M., & Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 201524306.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Fritz, M. H.-Y. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75. Retrieved from https://www.nature.com/articles/nature15394

Sun, S., Yang, F., Tan, G., Costanzo, M., Oughtred, R., Hirschman, J., … Yi, S. (2016). An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*, *26*(5), 670–680.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., … Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. Retrieved from https://genome.cshlp.org/content/13/9/2129.full

Vallender, E. J., & Lahn, B. T. (2004). Positive selection on the human genome. *Human Molecular Genetics*, *13*(suppl_2), R245–R254.

Veile, A. (2018). Hunter-gatherer diets and human behavioral evolution. *Physiology & Behavior*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031938418302506

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), e72. Retrieved from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040072

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., … Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, gky427-gky427. Retrieved from http://dx.doi.org/10.1093/nar/gky427

Willerslev, E., & Cooper, A. (2005). Ancient dna. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1558), 3–16. Retrieved from http://rspb.royalsocietypublishing.org/content/272/1558/3

Zeng, Q., Sun, L., Fu, Q., Liu, S., & Liu, Z. (2017). SNP Identification from Next-Generation Sequencing Datasets. *Bioinformatics in Aquaculture: Principles and Methods*, 288–307.

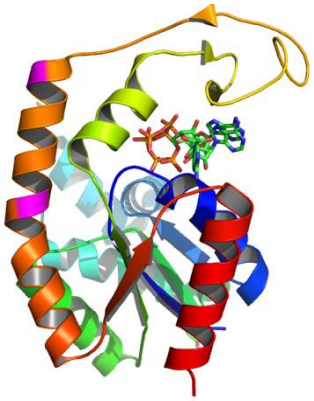
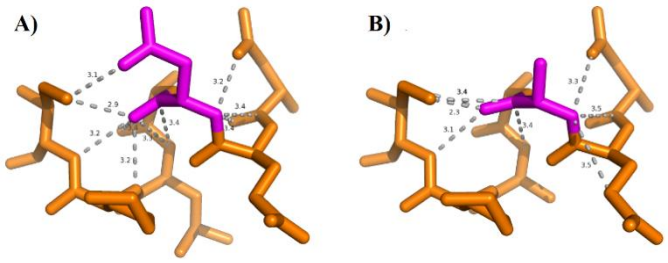
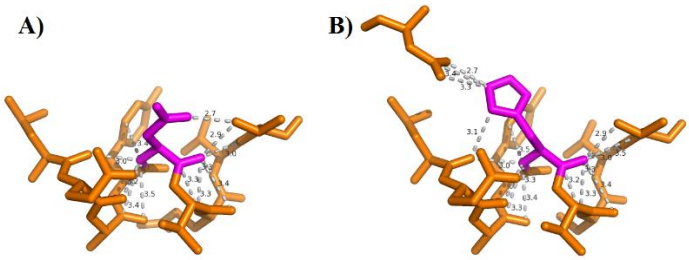
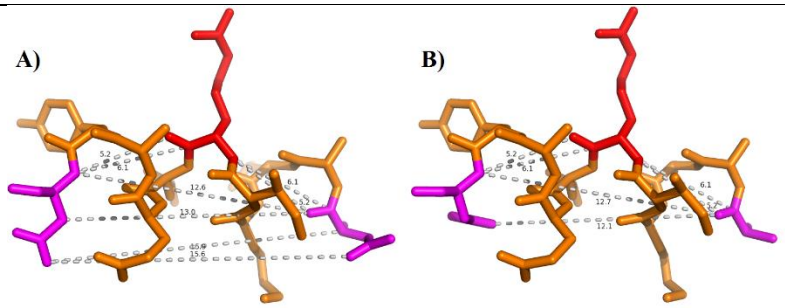
Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Girón, C. G. (2017). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. Retrieved from https://academic.oup.com/nar/article/46/D1/D754/4634002

Zeyl, C. (2000). Budding yeast as a model organism for population genetics. *Yeast*, *16*(8), 773–784.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2013). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. Retrieved from https://academic.oup.com/bioinformatics/article/30/7/1006/234947

Anexos

Anexo 1. Gen AK2 propuesto para el análisis de complementación y diferencias funcionales en *S. cerevisiae.*

Estructura modelada de la proteína de AK2 (PDB 2C9Y, con 99.16% de identidad en secuencia). En la hélice alfa de color anaranjado se indica la ubicación aproximada de los sitios variables en color magenta.	
Cambio alélico a nivel atómico del aminoácido 181. Se representa de color magenta el aminoácido con el cambio no sinónimo y las distancias que hay entre él y otros aminoácidos. A) Alelo ancestral que presenta asparagina B) alelo moderno que presenta lisina.	
Cambio alélico a nivel atómico del aminoácido 191. Se representa de color magenta el aminoácido con el cambio no sinónimo y las distancias que hay entre él y otros aminoácidos. A) Alelo ancestral que presenta ácido aspártico. B) Alelo moderno que presenta histidina.	
Distancias entre los alelos y el sitio de unión. De color rojo se representa el sitio de unión a AMP y de color magenta los alelos. A) Alelo ancestral. B) Alelo moderno.	

Anexo 2. Secuencia nucleotídica del gen AK2 actual y la variante ancestral.

Nombre	Secuencia
AK2_Modern **Transcript:** ENST00000354858 **-Length:** 734 bp **-Additional 5' sequence:** GCGGCCGC **-Additional 3' sequence:** CTGCAG **-Vector name:** pUC18	ATGGCTCCCAGCGTGCCAGCGGCAGAACCCGAGTATCCTAAAGGCATCCGGGCCGTGCTG CTGGGGCCTCCCGGGGCCGGTAAAGGGACCCAGGCACCCAGATTGGCTGAAAACTTCTGT GTCTGCCATTTAGCTACTGGGGACATGCTGAGGGCCATGGTGGCTTCTGGCTCAGAGCTAG GAAAAAAGCTGAAGGCAACTATGGATGCTGGGAAACTGGTGAGTGATGAAATGGTAGTG GAGCTCATTGAGAAGAATTTGGAGACCCCCTTGTGCAAAAATGGTTTTCTTCTGGATGGCT TCCCTCGGACTGTGAGGCAGGCAGAAATGCTCGATGACCTCATGGAGAAGAGGAAAGAG AAGCTTGATTCTGTGATTGAATTCAGCATCCCAGACTCTCTGCTGATCCGAAGAATCACAG GAAGGCTGATTCACCCCAAGAGTGGCCGTTCCTACCACGAGGAGTTCAACCCTCCAAAAG AGCCCATGAAAGATGACATCACCGGGGAACCCTTGATCCGTCGATCAGATGATAATGAAA AGGCCTTGAAAATCCGCCTGCAAGCCTACCACACTCAAACCACCCCACTCATAGAGTACT ACAGGAAACGGGGGATCCACTCCGCCATCGATGCATCCCAGACCCCCGATGTCGTGTTCG CAAGCATCCTAGCAGCCTTCTCCAAAGCCACATGTAAAGACTTGGTTATGTTTATCTAA
AK2_Ancient **-Length:** 734 bp **-Additional 5' sequence:** GCGGCCGC **-Additional 3' sequence:** CTGCAG **-Vector name:** pUC18	ATGGCTCCCAGCGTGCCAGCGGCAGAACCCGAGTATCCTAAAGGCATCCGGGCCGTGCTG CTGGGGCCTCCCGGGGCCGGTAAAGGGACCCAGGCACCCAGATTGGCTGAAAACTTCTGT GTCTGCCATTTAGCTACTGGGGACATGCTGAGGGCCATGGTGGCTTCTGGCTCAGAGCTAG GAAAAAAGCTGAAGGCAACTATGGATGCTGGGAAACTGGTGAGTGATGAAATGGTAGTG GAGCTCATTGAGAAGAATTTGGAGACCCCCTTGTGCAAAAATGGTTTTCTTCTGGATGGCT TCCCTCGGACTGTGAGGCAGGCAGAAATGCTCGATGACCTCATGGAGAAGAGGAAAGAG AAGCTTGATTCTGTGATTGAATTCAGCATCCCAGACTCTCTGCTGATCCGAAGAATCACAG GAAGGCTGATTCACCCCAAGAGTGGCCGTTCCTACCACGAGGAGTTCAACCCTCCAAAAG AGCCCATGAAAGATGACATCACCGGGGAACCCTTGATCCGTCGATCAGATGATAATGAAA ACGCCTTGAAAATCCGCCTGCAAGCCTACGACACTCAAACCACCCCACTCATAGAGTACT ACAGGAAACGGGGGATCCACTCCGCCATCGATGCATCCCAGACCCCCGATGTCGTGTTCG CAAGCATCCTAGCAGCCTTCTCCAAAGCCACATGTAAAGACTTGGTTATGTTTATCTAA

Anexo 3. Las 20 cepas generadas para el estudio de la variación funcional.

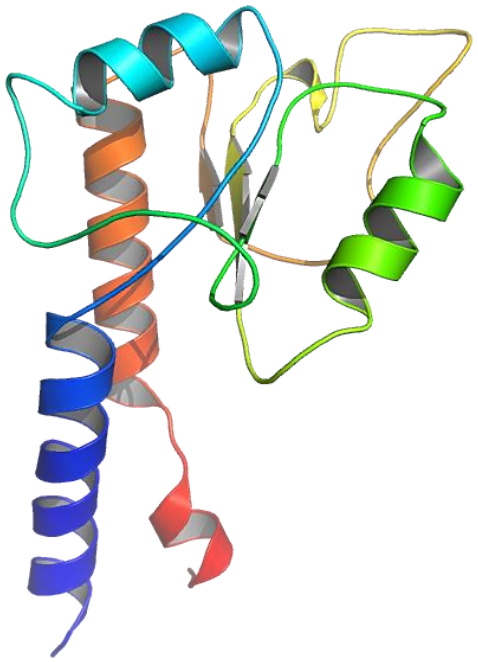
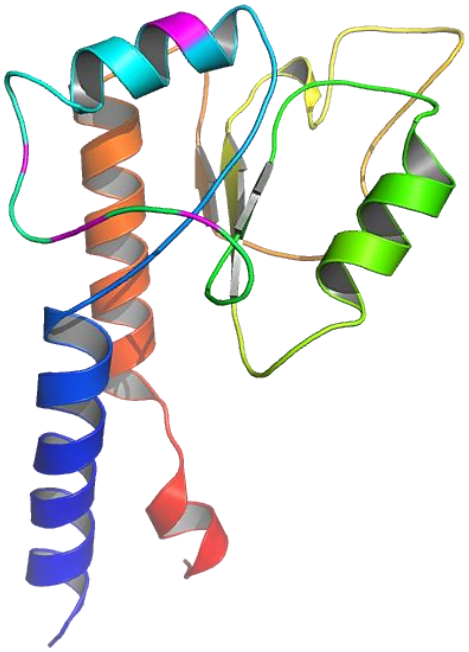
Cepa	Vector con	Variante
Silvestre	pCM189	Control
Δadk1		
Δadk2		
Δadk1 Δadk2		
Silvestre	pCM189_ADK1	Control
Δadk1		
Δadk2		
Δadk1 Δadk2		
Silvestre	pCM189_ADK2	Control
Δadk1		
Δadk2		
Δadk1 Δadk2		
Silvestre	pCM189_AK2	Antigua
Δadk1		
Δadk2		
Δadk1 Δadk2		
Silvestre	pCM189_AK2	Actual
Δadk1		
Δadk2		
Δadk1 Δadk2		

Anexo 4. Pruebas estadísticas.

Shapiro-Wilk normality test		
datos	pi	
Población Antigua	W=0.76455	p-value<2.20E-16
Población actual	W=0.57735	p-value<2.20E-16
Diferencia	W=0.80332	p-value<2.20E-16
datos	teta	
Población Antigua	W=0.59805	p-value<2.20E-16
Población actual	W=0.86953	p-value<2.20E-16
Diferencia	W=0.96344	p-value<2.20E-16
datos	Tajima	
Población Antigua	W=0.7748	p-value<2.20E-16
Población actual	W=0.78419	p-value<2.20E-16
Diferencia	W=0.94675	p-value<2.20E-16
Wilcoxon signed rank test with continuity correction		
datos	pi	
Población Antigua	V=6.0217e+13	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Población actual	V=3.2565e+14	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Diferencia	V=2.0389e+12	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
datos	teta	
Población Antigua	V=6.0263e+13	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Población actual	V=3.2565e+14	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Diferencia	V=3.9479e+12	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
datos	tajima	
Población Antigua	V=1.0169e+13	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Población actual	V=2.3016e+13	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		
Diferencia	V=5.9021e+12	p-value<2.20E-16
alternative hypothesis: true location is not equal to 0		

Wilcoxon: Comparing mean of two samples (population ancient VS current)		
datos	π	
W=2.35E+14	p-value<2.20E-16	
alternative hypothesis: true location shift is not equal to 0		
W=2.35E+14	p-value<2.20E-16	
alternative hypothesis: true location shift is greater than 0		
datos	teta	
W=1.60E+14	p-value<2.20E-16	
alternative hypothesis: true location shift is not equal to 0		
W=1.60E+14	p-value<2.20E-16	
alternative hypothesis: true location shift is greater than 0		
datos	Tajima	
W=1.18E+14	p-value<2.20E-16	
alternative hypothesis: true location shift is not equal to 0		
W=1.18E+14	p-value=1	
alternative hypothesis: true location shift is greater than 0		
Two-sample Kolmogorov-Smirnov test		
data: Anc$Piv and Mod$Piv		
alternative hypothesis: two-sided	D=0.77801	p-value<2.20E-16
In KS. Test (AncPiv, ModPiv) :		
p-value will be approximate in the presence of ties		
data: Anc$Teta and Mod$Teta		
alternative hypothesis: two-sided	D=0.4321	p-value<2.20E-16
In KS. Test (Anc$Teta, Mod$Teta):		
p-value will be approximate in the presence of ties		
data: Anc$Tajima and Mod$Tajima		
alternative hypothesis: two-sided	D=0.42629	p-value<2.20E-16
In KS. Test (Anc$Tajima, Mod$Tajima):		
p-value will be approximate in the presence of ties		

Anexo 5. Modelado de la proteína *NDUFS7*

***Splicing* alternativo:** ENST00000233627 **Aminoácidos:** 213 **Plantilla:** 5lnk.1.F **Identidad de secuencia:** 94.86	**A** 
Ubicación de los 6 SNPs **De color magenta se representan los SNP en la estructura.**	**B** 

Anexo 6. Cambios *missense* y *stop gained* encontrados con la metodología de baja diversidad, con diferencias en las frecuencias nucleotídicas entre poblaciones

Cromosoma	Posición	Alelo antiguo	Frecuencia	Alelo moderno	Frecuencia	Variación	Gen
3	195785236	A	1	T	1	missense_variant	MUC4
12	9095637	C	0.875	T	1	missense_variant	A2M
12	8852296	A	0.875	C	1	missense_variant	A2ML1
X	100988470	T	0.875	G	1	missense_variant	ARL13A
X	31478233	T	0.875	C	1	missense_variant	DMD
X	66602765	C	0.875	T	1	missense_variant	EDA2R
X	35803010	T	0.875	C	1	stop_gained	MAGEB16
5	141101254	G	0.875	C	1	missense_variant	PCDHB3
X	151700795	G	0.875	A	1	missense_variant	PRRG3
5	149981314	C	0.875	T	1	missense_variant	SLC26A2
8	144414297	G	0.875	C	1	missense_variant	SLC39A4
14	22086905	T	0.875	C	1	missense_variant	TRAV23DV6
16	74391650	G	0.8125	A	1	missense_variant	NPIPB15
12	7496888	C	0.75	T	1	missense_variant	CD163
12	7398426	T	0.75	A	1	missense_variant	CD163L1
12	9680226	T	0.75	C	1	missense_variant	CLEC2D
12	8222174	A	0.75	G	1	missense_variant	FAM90A1
8	144355665	T	0.75	C	1	missense_variant	FBXL6
X	151969707	C	0.75	A	1	missense_variant	GABRE
1	41512895	G	0.75	T	1	missense_variant	HIVEP3
X	35802308	T	0.75	C	1	missense_variant	MAGEB16
X	35802938	G	0.75	A	1	missense_variant	MAGEB16
X	35802939	A	0.75	T	1	missense_variant	MAGEB16
X	30218761	A	0.75	G	1	missense_variant	MAGEB2
X	19464358	T	0.75	C	1	missense_variant	MAP3K15
X	3317683	T	0.75	C	1	missense_variant	MXRA5
X	3321504	T	0.75	C	1	missense_variant	MXRA5
X	55263341	C	0.75	T	1	missense_variant	PAGE3
12	6970052	G	0.75	T	1	missense_variant	PHB2
X	154465991	C	0.75	G	1	missense_variant	PLXNA3
X	153397919	C	0.75	G	1	missense_variant	PNMA6E
12	9154777	G	0.75	T	1	missense_variant	PZP
X	17801257	C	0.75	T	1	missense_variant	RAI2
8	144415811	G	0.75	A	1	missense_variant	SLC39A4
11	130910330	A	0.75	C	1	missense_variant	SNX19
5	35708993	T	0.75	C	1	missense_variant	SPEF2
X	100686708	C	0.75	T	1	missense_variant	SYTL4

X	47198248	G	0.75	A	1	missense_variant	UBA1
X	8170039	C	0.75	G	1	missense_variant	VCX2
X	80688070	C	0.6875	T	1	missense_variant	BRWD3
X	35802678	G	0.6875	A	1	missense_variant	MAGEB16
5	132813992	C	0.6875	G	1	missense_variant	SOWAHA
6	29555899	G	0.6875	C	1	missense_variant	UBD
12	10434512	G	0.625	A	1	missense_variant	AC068775.2
X	139814971	C	0.625	A	1	missense_variant	ATP11C
X	77682471	G	0.625	C	1	missense_variant	ATRX
12	7097002	C	0.625	T	1	missense_variant	C1RL
11	125961075	T	0.625	A	1	missense_variant	CDON
X	66605144	T	0.625	C	1	missense_variant	EDA2R
15	23440196	C	0.625	T	1	missense_variant	GOLGA6L2
X	114731326	G	0.625	C	1	missense_variant	HTR2C
X	154013378	A	0.625	G	1	missense_variant	IRAK1
X	154018741	G	0.625	A	1	missense_variant	IRAK1
X	119147575	A	0.625	C	1	missense_variant	KIAA1210
12	10434512	G	0.625	A	1	missense_variant	KLRC2
X	152134921	T	0.625	C	1	missense_variant	MAGEA10
X	152135124	T	0.625	C	1	missense_variant	MAGEA10
X	26139103	T	0.625	C	1	missense_variant	MAGEB18
X	30236259	C	0.625	T	1	missense_variant	MAGEB3
X	26161415	G	0.625	A	1	missense_variant	MAGEB6P1
3	195785171	C	0.625	T	1	missense_variant	MUC4
X	55090033	G	0.625	C	1	missense_variant	PAGE2
X	72181757	A	0.625	G	1	missense_variant	PIN4
X	72181764	A	0.625	C	1	missense_variant	PIN4
X	85308129	A	0.625	T	1	missense_variant	POF1B
19	42935542	A	0.625	G	1	stop_gained	PSG7
X	7057741	G	0.625	C	1	missense_variant	PUDP
22	36028402	C	0.625	A	1	missense_variant	RBFOX2
X	115191608	T	0.625	C	1	missense_variant	RBMXL3
X	85108134	G	0.625	A	1	missense_variant	SATL1
5	474989	G	0.625	A	1	missense_variant	SLC9A3
X	48195304	G	0.625	C	1	missense_variant	SSX5
X	124333765	C	0.625	T	1	missense_variant	TEX13D
7	142801067	G	0.625	A	1	missense_variant	TRBC2
X	154485448	C	0.625	T	1	missense_variant	UBL4A
9	128656549	C	0.625	A	1	missense_variant	WDR34
3	75738575	A	0.625	G	1	missense_variant	ZNF717
12	8861159	T	0.5625	C	1	missense_variant	A2ML1
12	8863977	G	0.5625	A	1	missense_variant	A2ML1
2	95938843	G	0.5625	A	1	missense_variant	ANKRD36C

5	78129204	T	0.5625	A	1	missense_variant	AP3B1
12	103478394	G	0.5625	T	1	missense_variant	C12orf42
1	16057247	G	0.5625	C	1	missense_variant	CLCNKB
X	55146104	A	0.5625	G	1	stop_gained	FAM104B
X	103724558	C	0.5625	T	1	missense_variant	GLRA4
X	109465287	T	0.5625	C	1	missense_variant	GUCY2F
X	2859944	T	0.5625	C	1	missense_variant	GYG2
5	116005941	C	0.5625	G	1	missense_variant	LVRN
12	6830686	G	0.5625	A	1	missense_variant	P3H3
5	141956699	T	0.5625	G	1	missense_variant	PCDH12
X	27461222	G	0.5625	A	1	missense_variant	PPP4R3C
X	101292945	G	0.5625	A	1	missense_variant	TAF7L
3	75738859	C	0.5625	A	1	missense_variant	ZNF717

Anexo 7. Valores de SIFT y PolyPhen

chr	pos	SYMBOL	Feature	Protein position	Amino acids	Codons	SIFT	PolyPhen
3	195785236	MUC4	ENST00000462323	2115	D/V	gAc/gTc	tolerated_low_confidence(1)	possibly_damaging(0.908)
3	195785236	MUC4	ENST00000463781	2115	D/V	gAc/gTc	tolerated_low_confidence(1)	possibly_damaging(0.811)
3	195785236	MUC4	ENST00000466475	2115	D/V	gAc/gTc	tolerated(0.57)	unknown(0)
3	195785236	MUC4	ENST00000470451	2115	D/V	gAc/gTc	tolerated(1)	benign(0.018)
3	195785236	MUC4	ENST00000475231	2115	D/V	gAc/gTc	tolerated_low_confidence(1)	possibly_damaging(0.811)
3	195785236	MUC4	ENST00000477086	2115	D/V	gAc/gTc	tolerated(1)	possibly_damaging(0.908)
3	195785236	MUC4	ENST00000477756	2115	D/V	gAc/gTc	tolerated_low_confidence(0.38)	possibly_damaging(0.811)
3	195785236	MUC4	ENST00000478156	2115	D/V	gAc/gTc	tolerated_low_confidence(0.82)	possibly_damaging(0.811)
3	195785236	MUC4	ENST00000479406	2115	D/V	gAc/gTc	tolerated(0.57)	unknown(0)
3	195785236	MUC4	ENST00000480843	2115	D/V	gAc/gTc	tolerated(0.52)	benign(0.003)
12	8852296	A2ML1	ENST00000299698	850	D/E	gaC/gaA	tolerated(1)	benign(0)
12	8852296	A2ML1	ENST00000539547	359	D/E	gaC/gaA	tolerated(1)	benign(0)
12	8852296	A2ML1	ENST00000541459	400	D/E	gaC/gaA	tolerated(1)	benign(0)
12	9095637	A2M	ENST00000318602	639	N/D	Aat/Gat	tolerated(0.91)	benign(0)
14	22086905	TRAV23DV6	ENST00000390451	103	S/L	tCg/tTg	tolerated(1)	benign(0)
5	141101254	PCDHB3	ENST00000231130	202	P/R	cCg/cGg	tolerated_low_confidence(1)	benign(0)
5	149981314	SLC26A2	ENST00000286298	574	I/T	aTt/aCt	tolerated(0.84)	benign(0)
8	144414297	SLC39A4	ENST00000276833	347	V/L	Gtc/Ctc	tolerated(1)	benign(0)
8	144414297	SLC39A4	ENST00000301305	372	V/L	Gtc/Ctc	tolerated(1)	benign(0)
X	31478233	DMD	ENST00000343523	208	R/Q	cGg/cAg	tolerated(0.58)	benign(0)
X	31478233	DMD	ENST00000357033	2937	R/Q	cGg/cAg	tolerated(1)	benign(0)
X	31478233	DMD	ENST00000358062	633	R/Q	cGg/cAg	tolerated(0.85)	benign(0)
X	31478233	DMD	ENST00000359836	477	R/Q	cGg/cAg	tolerated(0.86)	benign(0)
X	31478233	DMD	ENST00000378677	2933	R/Q	cGg/cAg	tolerated(1)	benign(0)
X	31478233	DMD	ENST00000378707	477	R/Q	cGg/cAg	tolerated(0.89)	benign(0)
X	31478233	DMD	ENST00000474231	477	R/Q	cGg/cAg	tolerated(0.83)	benign(0)
X	31478233	DMD	ENST00000541735	477	R/Q	cGg/cAg	tolerated(0.87)	benign(0)
X	31478233	DMD	ENST00000619831	2932	R/Q	cGg/cAg	tolerated(1)	benign(0)
X	31478233	DMD	ENST00000620040	2936	R/Q	cGg/cAg	tolerated(1)	benign(0)
X	66602765	EDA2R	ENST00000253392	129	T/A	Aca/Gca	tolerated(0.38)	benign(0)
X	66602765	EDA2R	ENST00000374719	129	T/A	Aca/Gca	tolerated(0.48)	benign(0)
X	66602765	EDA2R	ENST00000396050	129	T/A	Aca/Gca	tolerated(0.38)	benign(0)

X	66602765	EDA2R	ENST00000451436	129	T/A	Aca/Gca	tolerated(0.48)	benign(0)
X	100988470	ARL13A	ENST00000450457	279	M/I	atG/atT	tolerated_low_confidence(0.12)	benign(0)
X	100988470	ARL13A	ENST00000494863	113	M/I	atG/atT	tolerated(0.15)	benign(0)
X	151700795	PRRG3	ENST00000370353	153	N/S	aAc/aGc	tolerated(1)	benign(0)
X	151700795	PRRG3	ENST00000538575	153	N/S	aAc/aGc	tolerated(1)	benign(0)
16	743916510	NPIPB15	ENST00000429990	301	Y/C	tAt/tGt	tolerated_low_confidence(1)	benign(0)
1	41512895	HIVEP3	ENST00000372583	2109	D/A	gAc/gCc	tolerated(1)	benign(0)
1	41512895	HIVEP3	ENST00000372584	2109	D/A	gAc/gCc	tolerated_low_confidence(1)	benign(0)
11	130910330	SNX19	ENST00000265909	618	L/F	ttG/ttT	tolerated(1)	benign(0)
11	130910330	SNX19	ENST00000533214	618	L/F	ttG/ttT	tolerated(1)	benign(0)
12	6970052	PHB2	ENST00000545167	75	E/A	gAg/gCg	tolerated_low_confidence(0.27)	unknown(0)
12	7398426	CD163L1	ENST00000313599	523	L/M	Ttg/Atg	tolerated(0.09)	possibly_damaging(0.589)
12	7398426	CD163L1	ENST00000416109	533	L/M	Ttg/Atg	tolerated(0.09)	possibly_damaging(0.484)
12	7496888	CD163	ENST00000359156	342	I/V	Atc/Gtc	tolerated(0.22)	benign(0.007)
12	7496888	CD163	ENST00000396620	342	I/V	Atc/Gtc	tolerated(0.31)	benign(0.001)
12	7496888	CD163	ENST00000432237	342	I/V	Atc/Gtc	tolerated(0.21)	benign(0)
12	7496888	CD163	ENST00000541972	330	I/V	Atc/Gtc	tolerated(0.22)	benign(0.003)
12	8222174	FAM90A1	ENST00000307435	348	T/I	aCa/aTa	tolerated(0.12)	benign(0)
12	8222174	FAM90A1	ENST00000538603	348	T/I	aCa/aTa	tolerated(0.12)	benign(0)
12	9154777	PZP	ENST00000261336	1205	T/P	Acc/Ccc	tolerated(1)	benign(0)
12	9680226	CLEC2D	ENST00000492359	33	T/M	aCg/aTg	deleterious_low_confidence(0.01)	unknown(0)
5	35708993	SPEF2	ENST00000356031	904	A/V	gCt/gTt	tolerated(0.56)	benign(0)
5	35708993	SPEF2	ENST00000440995	899	A/V	gCt/gTt	tolerated(0.49)	benign(0.001)
5	35708993	SPEF2	ENST00000509059	899	A/V	gCt/gTt	tolerated(0.56)	benign(0.001)
5	35708993	SPEF2	ENST00000637569	904	A/V	gCt/gTt	tolerated(0.39)	probably_damaging(0.98)
8	144355665	FBXL6	ENST00000331890	496	G/S	Ggc/Agc	tolerated(1)	benign(0.001)
8	144355665	FBXL6	ENST00000455319	490	G/S	Ggc/Agc	tolerated(1)	benign(0.001)
8	144415811	SLC39A4	ENST00000276833	133	M/T	aTg/aCg	tolerated(0.28)	benign(0)
8	144415811	SLC39A4	ENST00000301305	158	M/T	aTg/aCg	tolerated(0.17)	benign(0)
X	3317683	MXRA5	ENST00000217939	2000	G/S	Ggc/Agc	tolerated(0.4)	benign(0.164)
X	3321504	MXRA5	ENST00000217939	1394	G/D	gGc/gAc	tolerated(1)	benign(0)
X	8170039	VCX2	ENST00000317103	138	T/S	aCc/aGc	tolerated_low_confidence(1)	benign(0)
X	17801257	RAI2	ENST00000331511	252	M/V	Atg/Gtg	tolerated(1)	benign(0)
X	17801257	RAI2	ENST00000360011	252	M/V	Atg/Gtg	tolerated(1)	benign(0)
X	17801257	RAI2	ENST00000415486	202	M/V	Atg/Gtg	tolerated(1)	benign(0)

X	17801257	RAI2	ENST00000451717	252	M/V	Atg/Gtg	tolerated(1)	benign(0)
X	17801257	RAI2	ENST00000545871	252	M/V	Atg/Gtg	tolerated(1)	benign(0)
X	19464358	MAP3K15	ENST00000338883	192	A/T	Gct/Act	deleterious(0.02)	benign(0.356)
X	30218761	MAGEB2	ENST00000378988	61	E/K	Gag/Aag	tolerated(0.78)	benign(0.003)
X	35802308	MAGEB16	ENST00000399985	38	L/F	Ctc/Ttc	tolerated(0.73)	benign(0)
X	35802308	MAGEB16	ENST00000399987	38	L/F	Ctc/Ttc	tolerated(0.73)	benign(0)
X	35802308	MAGEB16	ENST00000399988	38	L/F	Ctc/Ttc	tolerated(0.73)	benign(0)
X	35802308	MAGEB16	ENST00000399989	38	L/F	Ctc/Ttc	tolerated(0.73)	benign(0)
X	35802308	MAGEB16	ENST00000399992	70	L/F	Ctc/Ttc	tolerated(0.74)	benign(0.007)
X	35802938	MAGEB16	ENST00000399985	248	M/V	Atg/Gtg	deleterious(0)	benign(0.005)
X	35802938	MAGEB16	ENST00000399987	248	M/V	Atg/Gtg	deleterious(0)	benign(0.005)
X	35802938	MAGEB16	ENST00000399988	248	M/V	Atg/Gtg	deleterious(0)	benign(0.005)
X	35802938	MAGEB16	ENST00000399989	248	M/V	Atg/Gtg	deleterious(0)	benign(0.005)
X	35802938	MAGEB16	ENST00000399992	280	M/V	Atg/Gtg	deleterious(0)	benign(0.003)
X	35802939	MAGEB16	ENST00000399985	248	M/K	aTg/aAg	tolerated(1)	benign(0)
X	35802939	MAGEB16	ENST00000399987	248	M/K	aTg/aAg	tolerated(1)	benign(0)
X	35802939	MAGEB16	ENST00000399988	248	M/K	aTg/aAg	tolerated(1)	benign(0)
X	35802939	MAGEB16	ENST00000399989	248	M/K	aTg/aAg	tolerated(1)	benign(0)
X	35802939	MAGEB16	ENST00000399992	280	M/K	aTg/aAg	tolerated(1)	benign(0)
X	47198248	UBA1	ENST00000451702	16	I/V	Atc/Gtc	tolerated_low_confidence(0.36)	unknown(0)
X	47198248	UBA1	ENST00000457753	16	I/V	Atc/Gtc	tolerated_low_confidence(0.53)	unknown(0)
X	55263341	PAGE3	ENST00000374951	35	N/D	Aat/Gat	tolerated(1)	benign(0)
X	55263341	PAGE3	ENST00000519203	35	N/D	Aat/Gat	tolerated(1)	benign(0)
X	100686708	SYTL4	ENST00000263033	420	I/V	Att/Gtt	tolerated(0.57)	benign(0.003)
X	100686708	SYTL4	ENST00000276141	420	I/V	Att/Gtt	tolerated(0.57)	benign(0.003)
X	100686708	SYTL4	ENST00000372989	420	I/V	Att/Gtt	tolerated(0.57)	benign(0.003)
X	151969707	GABRE	ENST00000370328	102	S/A	Tcc/Gcc	tolerated(0.15)	benign(0.415)
X	153397919	PNMA6E	ENST00000445091	311	P/A	Cct/Gct	tolerated(0.07)	benign(0.038)
X	154465991	PLXNA3	ENST00000369682	863	E/D	gaG/gaC	tolerated(0.45)	possibly_damaging(0.469)
5	132813992	SOWAHA	ENST00000378693	124	R/P	cGg/cCg	tolerated(0.33)	benign(0)
6	29555899	UBD	ENST00000377050	160	C/S	tGt/tCt	tolerated(0.43)	benign(0)
X	35802678	MAGEB16	ENST00000399985	161	H/R	cAc/cGc	tolerated(0.37)	benign(0.01)
X	35802678	MAGEB16	ENST00000399987	161	H/R	cAc/cGc	tolerated(0.37)	benign(0.01)
X	35802678	MAGEB16	ENST00000399988	161	H/R	cAc/cGc	tolerated(0.37)	benign(0.01)
X	35802678	MAGEB16	ENST00000399989	161	H/R	cAc/cGc	tolerated(0.37)	benign(0.01)

X	35802678	MAGEB16	ENST00000399992	193	H/R	cAc/cGc	tolerated(0.5)	benign(0)
X	80688070	BRWD3	ENST00000373275	1288	K/R	aAg/aGg	tolerated(1)	benign(0)
11	125961075	CDON	ENST00000263577	1221	I/N	aTc/aAc	tolerated(1)	benign(0)
11	125961075	CDON	ENST00000392693	1244	I/N	aTc/aAc	tolerated(1)	benign(0)
11	125961075	CDON	ENST00000531738	598	I/N	aTc/aAc	tolerated_low_confidence(1)	benign(0)
12	7097002	C1RL	ENST00000266542	285	I/V	Atc/Gtc	tolerated(0.26)	benign(0.264)
12	7097002	C1RL	ENST00000534950	118	I/V	Atc/Gtc	tolerated(0.87)	benign(0.06)
12	10434512	KLRC2	ENST00000381902	102	F/S	tTt/tCt	tolerated(1)	benign(0)
12	10434512	KLRC2	ENST00000535069	75	F/S	tTt/tCt	tolerated(1)	benign(0.009)
12	10434512	KLRC2	ENST00000536833	43	F/S	tTt/tCt	tolerated(1)	benign(0)
12	10434512	AC068775.2	ENST00000539033	102	F/S	tTt/tCt	tolerated(1)	benign(0)
15	23440196	GOLGA6L2	ENST00000567107	760	E/G	gAa/gGa	tolerated_low_confidence(0.12)	unknown(0)
22	36028402	RBFOX2	ENST00000438146	8	H/Q	caT/caG	tolerated_low_confidence(0.11)	benign(0)
3	75738575	ZNF717	ENST00000478296	300	R/C	Cgt/Tgt	deleterious(0.05)	possibly_damaging(0.862)
3	195785171	MUC4	ENST00000462323	2137	T/A	Aca/Gca	tolerated(0.75)	benign(0.398)
3	195785171	MUC4	ENST00000463781	2137	T/A	Aca/Gca	tolerated_low_confidence(0.36)	benign(0.224)
3	195785171	MUC4	ENST00000466475	2137	T/A	Aca/Gca	tolerated(0.45)	unknown(0)
3	195785171	MUC4	ENST00000470451	2137	T/A	Aca/Gca	tolerated(0.51)	benign(0.398)
3	195785171	MUC4	ENST00000475231	2137	T/A	Aca/Gca	tolerated_low_confidence(0.44)	benign(0.224)
3	195785171	MUC4	ENST00000477086	2137	T/A	Aca/Gca	tolerated(0.77)	benign(0.09)
3	195785171	MUC4	ENST00000477756	2137	T/A	Aca/Gca	tolerated_low_confidence(0.43)	benign(0.224)
3	195785171	MUC4	ENST00000478156	2137	T/A	Aca/Gca	tolerated_low_confidence(0.64)	benign(0.224)
3	195785171	MUC4	ENST00000479406	2137	T/A	Aca/Gca	tolerated(0.45)	unknown(0)
3	195785171	MUC4	ENST00000480843	2137	T/A	Aca/Gca	tolerated(0.76)	benign(0.224)
5	474989	SLC9A3	ENST00000264938	799	C/R	Tgc/Cgc	tolerated_low_confidence(1)	benign(0)
5	474989	SLC9A3	ENST00000514375	790	C/R	Tgc/Cgc	tolerated_low_confidence(1)	benign(0)
7	142801067	TRBC2	ENST00000466254	10	K/E	Aag/Gag	tolerated(0.21)	benign(0.011)
9	128656549	WDR34	ENST00000372715	60	W/G	Tgg/Ggg	tolerated(0.61)	benign(0)
9	128656549	WDR34	ENST00000419989	45	W/G	Tgg/Ggg	tolerated(0.66)	benign(0)
9	128656549	WDR34	ENST00000451652	51	W/G	Tgg/Ggg	tolerated(0.35)	benign(0)
X	7057741	PUDP	ENST00000540122	198	C/S	tGt/tCt	tolerated_low_confidence(0.51)	benign(0)
X	26139103	MAGEB18	ENST00000325250	40	P/S	Cct/Tct	tolerated(1)	benign(0)
X	26161415	MAGEB6P1	ENST00000416929	272	N/S	aAt/aGt	tolerated(1)	benign(0)
X	30236259	MAGEB3	ENST00000361644	112	I/T	aTc/aCc	tolerated(1)	benign(0)
X	30236259	MAGEB3	ENST00000620842	112	I/T	aTc/aCc	tolerated(1)	benign(0)

X	48195304	SSX5	ENST00000311798	19	E/Q	Gag/Cag	deleterious_low_confidence(0)	possibly_damaging(0.713)
X	48195304	SSX5	ENST00000347757	19	E/Q	Gag/Cag	deleterious(0.04)	possibly_damaging(0.721)
X	48195304	SSX5	ENST00000376923	19	E/Q	Gag/Cag	deleterious(0.04)	possibly_damaging(0.721)
X	55090033	PAGE2	ENST00000374965	5	L/V	Cta/Gta	tolerated(1)	benign(0)
X	55090033	PAGE2	ENST00000374968	5	L/V	Cta/Gta	tolerated(1)	benign(0)
X	55090033	PAGE2	ENST00000449097	5	L/V	Cta/Gta	tolerated_low_confidence(1)	benign(0)
X	66605144	EDA2R	ENST00000253392	57	R/K	aGa/aAa	tolerated(0.84)	benign(0.001)
X	66605144	EDA2R	ENST00000374719	57	R/K	aGa/aAa	tolerated(0.6)	benign(0.125)
X	66605144	EDA2R	ENST00000396050	57	R/K	aGa/aAa	tolerated(0.84)	benign(0.001)
X	66605144	EDA2R	ENST00000451436	57	R/K	aGa/aAa	tolerated(0.6)	benign(0.125)
X	72181757	PIN4	ENST00000218432	16	R/Q	cGg/cAg	tolerated_low_confidence(0.25)	benign(0)
X	72181757	PIN4	ENST00000373662	16	R/Q	cGg/cAg	tolerated_low_confidence(0.24)	benign(0)
X	72181757	PIN4	ENST00000373669	16	R/Q	cGg/cAg	tolerated_low_confidence(0.47)	benign(0)
X	72181757	PIN4	ENST00000423432	16	R/Q	cGg/cAg	tolerated_low_confidence(0.27)	benign(0)
X	72181764	PIN4	ENST00000218432	18	S/R	agC/agA	tolerated_low_confidence(0.32)	benign(0)
X	72181764	PIN4	ENST00000373662	18	S/R	agC/agA	tolerated_low_confidence(0.56)	benign(0)
X	72181764	PIN4	ENST00000373669	18	S/R	agC/agA	tolerated_low_confidence(0.61)	benign(0)
X	72181764	PIN4	ENST00000423432	18	S/R	agC/agA	tolerated_low_confidence(0.34)	benign(0)
X	77682471	ATRX	ENST00000373344	929	E/Q	Gag/Cag	deleterious_low_confidence(0.04)	benign(0.005)
X	77682471	ATRX	ENST00000395603	891	E/Q	Gag/Cag	deleterious_low_confidence(0.04)	benign(0.386)
X	77682471	ATRX	ENST00000624032	900	E/Q	Gag/Cag	deleterious_low_confidence(0.02)	possibly_damaging(0.743)
X	77682471	ATRX	ENST00000624166	861	E/Q	Gag/Cag	deleterious_low_confidence(0.01)	possibly_damaging(0.812)
X	85108134	SATL1	ENST00000395409	92	W/R	Tgg/Cgg	tolerated(0.56)	benign(0)
X	85108134	SATL1	ENST00000509231	279	W/R	Tgg/Cgg	tolerated(0.38)	benign(0.003)
X	85308129	POF1B	ENST00000262753	349	M/L	Atg/Ttg	tolerated(1)	benign(0)
X	85308129	POF1B	ENST00000373145	349	M/L	Atg/Ttg	tolerated(1)	benign(0)
X	114731326	HTR2C	ENST00000276198	23	S/C	tCt/tGt	tolerated_low_confidence(0.28)	benign(0)
X	114731326	HTR2C	ENST00000371950	23	S/C	tCt/tGt	tolerated_low_confidence(1)	benign(0)
X	114731326	HTR2C	ENST00000371951	23	S/C	tCt/tGt	tolerated_low_confidence(0.28)	benign(0)
X	115191608	RBMXL3	ENST00000424776	723	R/C	Cgc/Tgc	tolerated_low_confidence(0.24)	benign(0)
X	119147575	KIAA1210	ENST00000402510	103	G/V	gGc/gTc	tolerated(0.36)	benign(0)
X	124333765	TEX13D	ENST00000632372	283	L/P	cTg/cCg	tolerated(0.09)	benign(0.047)
X	139814971	ATP11C	ENST00000327569	114	C/W	tgT/tgG	tolerated(1)	benign(0)
X	139814971	ATP11C	ENST00000361648	114	C/W	tgT/tgG	tolerated(1)	benign(0)
X	139814971	ATP11C	ENST00000370557	111	C/W	tgT/tgG	tolerated(1)	benign(0)

X	152134921	MAGEA10	ENST00000244096	234	V/I	Gtc/Atc	tolerated(1)	benign(0.001)
X	152134921	MAGEA10	ENST00000370323	234	V/I	Gtc/Atc	tolerated(1)	benign(0.001)
X	152135124	MAGEA10	ENST00000244096	166	R/K	aGa/aAa	tolerated(1)	benign(0.006)
X	152135124	MAGEA10	ENST00000370323	166	R/K	aGa/aAa	tolerated(1)	benign(0.006)
X	152135124	MAGEA10	ENST00000444834	166	R/K	aGa/aAa	tolerated(1)	benign(0.006)
X	154013378	IRAK1	ENST00000369974	453	S/L	tCg/tTg	tolerated(0.36)	benign(0)
X	154013378	IRAK1	ENST00000369980	532	S/L	tCg/tTg	tolerated(0.33)	benign(0)
X	154013378	IRAK1	ENST00000443220	281	S/L	tCg/tTg	tolerated(0.43)	benign(0)
X	154013378	IRAK1	ENST00000444254	88	S/L	tCg/tTg	tolerated(0.79)	benign(0)
X	154018741	IRAK1	ENST00000369973	222	F/S	tTt/tCt	tolerated(0.55)	benign(0.001)
X	154018741	IRAK1	ENST00000369974	196	F/S	tTt/tCt	tolerated(1)	benign(0)
X	154018741	IRAK1	ENST00000369980	196	F/S	tTt/tCt	tolerated(1)	benign(0)
X	154018741	IRAK1	ENST00000393687	196	F/S	tTt/tCt	tolerated(1)	benign(0)
X	154018741	IRAK1	ENST00000429936	222	F/S	tTt/tCt	tolerated(0.75)	benign(0)
X	154485448	UBL4A	ENST00000369653	142	H/R	cAt/cGt	tolerated(0.12)	benign(0)
1	160572477	CLCNKB	ENST00000431772	135	P/A	Cct/Gct	tolerated_low_confidence(0.19)	unknown(0)
1	160572477	CLCNKB	ENST00000619181	452	P/A	Cct/Gct	deleterious_low_confidence(0)	possibly_damaging(0.737)
12	6830686	P3H3	ENST00000290510	301	T/A	Aca/Gca	deleterious(0.02)	benign(0.024)
12	8861159	A2ML1	ENST00000299698	1122	R/W	Cgg/Tgg	deleterious(0.02)	possibly_damaging(0.601)
12	8861159	A2ML1	ENST00000539547	631	R/W	Cgg/Tgg	deleterious(0.01)	possibly_damaging(0.599)
12	8861159	A2ML1	ENST00000541459	672	R/W	Cgg/Tgg	deleterious(0.01)	possibly_damaging(0.601)
12	8863977	A2ML1	ENST00000299698	1229	H/R	cAc/cGc	tolerated(0.54)	benign(0)
12	8863977	A2ML1	ENST00000539547	738	H/R	cAc/cGc	tolerated(0.51)	benign(0)
12	8863977	A2ML1	ENST00000541459	779	H/R	cAc/cGc	tolerated(0.44)	benign(0)
12	103478394	C12orf42	ENST00000378113	11	E/D	gaA/gaC	tolerated_low_confidence(0.35)	benign(0.072)
12	103478394	C12orf42	ENST00000547347	11	E/D	gaA/gaC	tolerated_low_confidence(0.19)	benign(0.072)
12	103478394	C12orf42	ENST00000548883	11	E/D	gaA/gaC	tolerated_low_confidence(0.35)	benign(0.072)
12	103478394	C12orf42	ENST00000551134	11	E/D	gaA/gaC	tolerated_low_confidence(0.26)	benign(0.072)
12	103478394	C12orf42	ENST00000552578	11	E/D	gaA/gaC	tolerated_low_confidence(0.31)	benign(0.072)
2	959388443	ANKRD36C	ENST00000456556	540	L/P	cTa/cCa	tolerated_low_confidence(0.17)	possibly_damaging(0.721)
3	757388559	ZNF717	ENST00000478296	205	V/G	gTg/gGg	tolerated(0.31)	benign(0)
5	781292044	AP3B1	ENST00000255194	585	V/E	gTa/gAa	tolerated(1)	benign(0)
5	781292044	AP3B1	ENST00000519295	536	V/E	gTa/gAa	tolerated(1)	benign(0)
5	116005941	LVRN	ENST00000357872	689	L/F	ttG/ttC	deleterious(0.02)	benign(0.148)
5	116005941	LVRN	ENST00000395528	677	L/F	ttG/ttC	deleterious(0.02)	benign(0.238)

5	116005941	LVRN	ENST00000504467	689	L/F	ttG/ttC	deleterious(0.02)	benign(0.133)
5	116005941	LVRN	ENST00000514509	206	L/F	ttG/ttC	deleterious(0.02)	possibly_damaging(0.522)
5	141956699	PCDH12	ENST00000231484	385	H/N	Cac/Aac	tolerated(1)	benign(0)
X	2859944	GYG2	ENST00000353656	270	A/V	gCg/gTg	tolerated(0.37)	benign(0.145)
X	2859944	GYG2	ENST00000381157	89	A/V	gCg/gTg	deleterious(0.02)	benign(0.033)
X	2859944	GYG2	ENST00000381163	270	A/V	gCg/gTg	tolerated(0.37)	benign(0.018)
X	2859944	GYG2	ENST00000398806	239	A/V	gCg/gTg	tolerated(0.37)	benign(0.04)
X	2859944	GYG2	ENST00000639373	270	A/V	gCg/gTg	tolerated(0.38)	benign(0.348)
X	27461222	PPP4R3C	ENST00000412172	692	V/A	gTa/gCa	tolerated(1)	benign(0)
X	101292945	TAF7L	ENST00000372907	34	L/P	cTt/cCt	tolerated(0.52)	benign(0)
X	103724558	GLRA4	ENST00000372617	85	I/V	Atc/Gtc	tolerated(0.4)	benign(0.015)
X	109465287	GUCY2F	ENST00000218006	296	R/Q	cGg/cAg	tolerated(0.62)	benign(0)

Anexo 8. Definición de los procesos biológico que reporta panther.

Nombre	Descripción
Locomoción	Movimiento autopropulsado de una célula u organismo de un lugar a otro.
Proceso del sistema inmune	Cualquier proceso involucrado en el desarrollo o funcionamiento del sistema inmune, un sistema orgánico para respuestas calibradas a posibles amenazas internas o invasivas.
Adhesión biológica	La unión de una célula u organismo a un sustrato, a otra célula u otro organismo. La adhesión biológica incluye la unión intracelular entre regiones de la membrana.
Proceso de desarrollo	Un proceso biológico cuyo resultado específico es la progresión de una unidad de vida integrada: una estructura anatómica (que puede ser una estructura subcelular, célula, tejido u órgano) u organismo a lo largo del tiempo desde una condición inicial a una condición posterior.
Reproducción	La producción de nuevos individuos que contienen una porción de material genético heredado de uno o más organismos parentales.
Proceso de organismo multicelular	Cualquier proceso biológico, que ocurre a nivel de un organismo multicelular, pertinente a su función.
Organización de componentes celulares o biogénesis	Un proceso que resulta en la biosíntesis de las macromoléculas constituyentes, el ensamblaje, la disposición de las partes constituyentes o el desmontaje de un componente celular.
Localización	Cualquier proceso en el que una célula, una sustancia o una entidad celular, como un complejo proteico u orgánulo, se transporta, se ata o se mantiene en una ubicación específica. En el caso de las sustancias, la localización también puede lograrse a través de la degradación selectiva.
Respuesta al estímulo	Cualquier proceso que resulte en un cambio en el estado o la actividad de una célula o un organismo (en términos de movimiento, secreción, producción de enzimas, expresión genética, etc.) como resultado de un estímulo. El proceso comienza con la detección del estímulo y finaliza con un cambio en el estado o actividad de la célula u organismo.
Regulación biológica	Cualquier proceso que modula un atributo medible de cualquier proceso, calidad o función biológica.
Proceso metabólico	Las reacciones químicas y las vías, incluido el anabolismo y el catabolismo, por el cual los organismos vivos transforman las sustancias químicas. Los procesos metabólicos típicamente transforman moléculas pequeñas, pero también incluyen procesos macromoleculares tales como la reparación y replicación del DNA y la síntesis y degradación de proteínas.
Proceso celular	Cualquier proceso que se lleva a cabo a nivel celular, pero no necesariamente restringido a una sola célula. Por ejemplo, la comunicación celular ocurre entre más de una célula, pero ocurre a nivel celular.

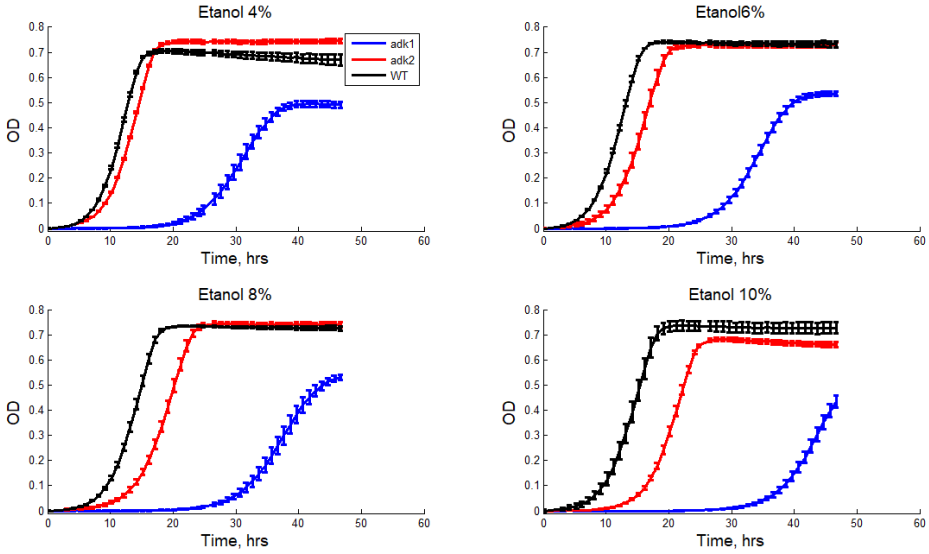
Anexo 9. Genes ortólogos a S. *cerevisiae* de los 79 genes

Human gen	Gene description	Model gene name	%id. target model gene identical to query gene	%id. query gene identical to target model gene
PHB2	prohibitin 2	PHB2	50.5017	48.7097
UBA1	ubiquitin like modifier activating enzyme 1	UBA1	50	51.6602
ATP11C	ATPase phospholipid transporting 11C	DRS2	32.1555	26.8635
PUDP	pseudouridine 5'-phosphatase	YKL033W-A	30.2789	32.2034
LVRN	laeverin	AAP1	22.7273	26.285
LVRN	laeverin	APE2	22.3232	23.2143
TAF7L	TATA-box binding protein associated factor 7 like	TAF7	20.7792	16.2712
ARL13A	ADP ribosylation factor like GTPase 13A	ARL3	20.7031	26.7677
PPP4R3C	protein phosphatase 4 regulatory subunit 3C	PSY2	18.9904	18.4149
GYG2	glycogenin 2	GLG1	18.7625	15.2597
AP3B1	adaptor related protein complex 3 beta 1 subunit	APL6	18.7386	25.3399
GYG2	glycogenin 2	GLG2	17.9641	23.6842
CLCNKB	chloride voltage-gated channel Kb	GEF1	17.3217	15.276
MAP3K15	mitogen-activated protein kinase kinase kinase 15	BCK1	17.0602	15.1556
SLC9A3	solute carrier family 9 member A3	NHX1	15.7074	20.6951
SLC39A4	solute carrier family 39 member 4	YKE4	11.2828	21.0983
MAP3K15	mitogen-activated protein kinase kinase kinase 15	STE11	10.2818	18.8285
ATRX	ATRX, chromatin remodeler	RDH54	6.62119	17.2234
BRWD3	bromodomain and WD repeat domain containing 3	MDV1	5.88235	14.8459
BRWD3	bromodomain and WD repeat domain containing 3	CAF4	5.10544	14.3079
A2M	alpha-2-macroglobulin			
A2ML1	alpha-2-macroglobulin like 1			
AC068775.2				
ANKRD36C	ankyrin repeat domain 36C			
C12orf42	chromosome 12 open reading frame 42			
C1RL	complement C1r subcomponent like			
CD163	CD163 molecule			
CD163L1	CD163 molecule like 1			
CDON	cell adhesion associated, oncogene regulated			
CLEC2D	C-type lectin domain family 2 member D			
DMD	dystrophin			
EDA2R	ectodysplasin A2 receptor			
FAM90A1	family with sequence similarity 90 member A1			
FBXL6	F-box and leucine rich repeat protein 6			
GABRE	gamma-aminobutyric acid type A receptor epsilon subunit			
GLRA4	glycine receptor alpha 4			

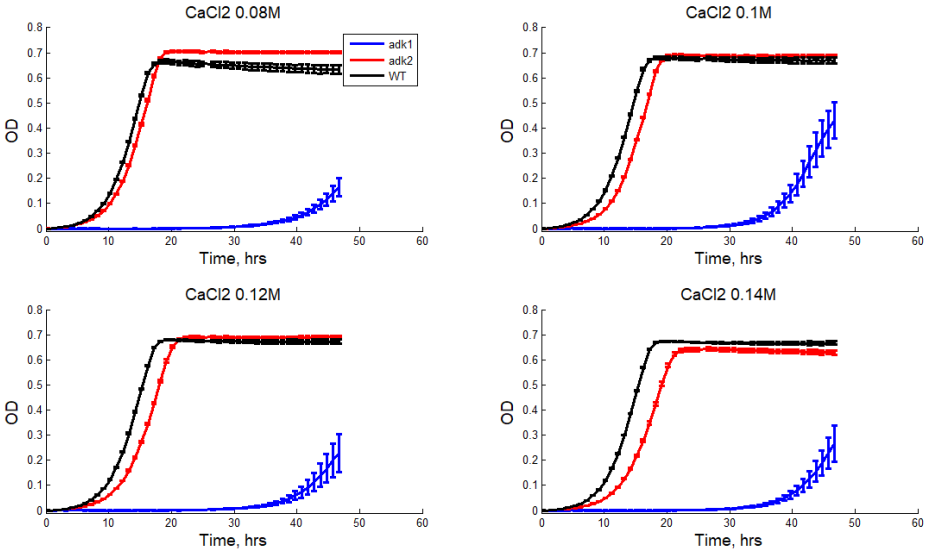
GOLGA6L2	golgin A6 family-like 2			
GUCY2F	guanylate cyclase 2F, retinal			
HIVEP3	human immunodeficiency virus type I enhancer binding protein 3			
HTR2C	5-hydroxytryptamine receptor 2C			
IRAK1	interleukin 1 receptor associated kinase 1			
KIAA1210	KIAA1210			
KLRC2	killer cell lectin like receptor C2			
MAGEA10	MAGE family member A10			
MAGEB16	MAGE family member B16			
MAGEB18	MAGE family member B18			
MAGEB2	MAGE family member B2			
MAGEB3	MAGE family member B3			
MAGEB6P1	MAGE family member B6 pseudogene 1			
MUC4	mucin 4, cell surface associated			
MXRA5	matrix remodeling associated 5			
P3H3	prolyl 3-hydroxylase 3			
PCDH12	protocadherin 12			
PCDHB3	protocadherin beta 3			
PIN4	peptidylprolyl cis/trans isomerase, NIMA-interacting 4			
PLXNA3	plexin A3			
PNMA6E	PNMA family member 6E			
POF1B	POF1B, actin binding protein			
PRRG3	proline rich and Gla domain 3			
PZP	PZP, alpha-2-macroglobulin like			
RAI2	retinoic acid induced 2			
RBFOX2	RNA binding fox-1 homolog 2			
RBMXL3	RNA binding motif protein, X-linked like 3			
SATL1	spermidine/spermine N1-acetyl transferase like 1			
SLC26A2	solute carrier family 26 member 2			
SNX19	sorting nexin 19			
SOWAHA	sosondowah ankyrin repeat domain family member A			
SPEF2	sperm flagellar 2			
SSX5	SSX family member 5			
SYTL4	synaptotagmin like 4			
TEX13D	TEX13 family member D			
TRAV23DV6	T-cell receptor alpha variable 23/delta variable 6			
TRBC2	T-cell receptor beta constant 2			
UBD	ubiquitin D			
UBL4A	ubiquitin like 4A			
WDR34	WD repeat domain 34			

Anexo 10. Curvas de crecimiento de las cepas knockout comparadas con la cepa silvestre.

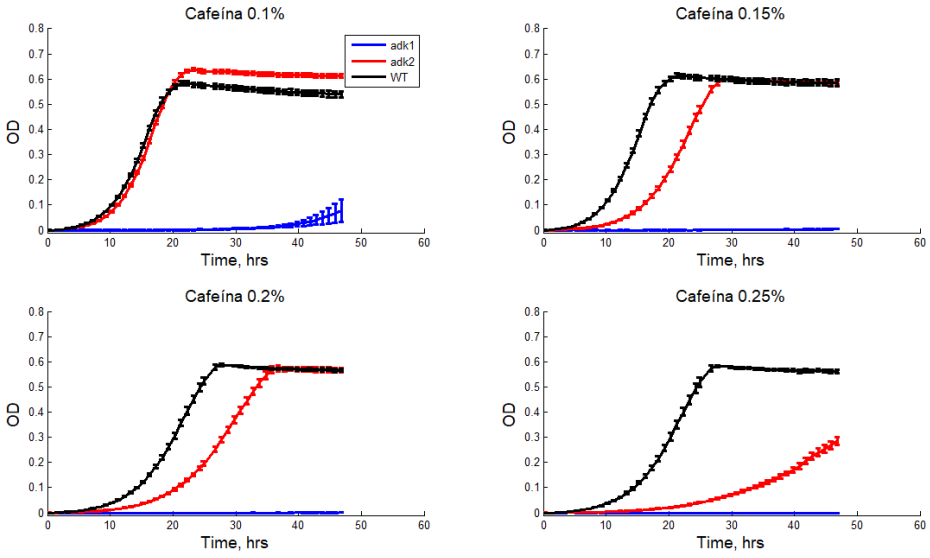
- Etanol



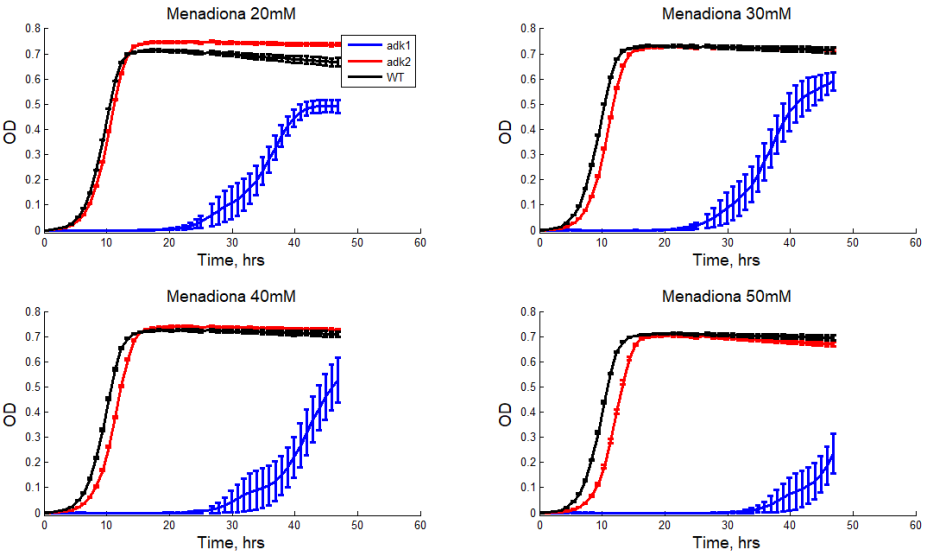
- CaCl2



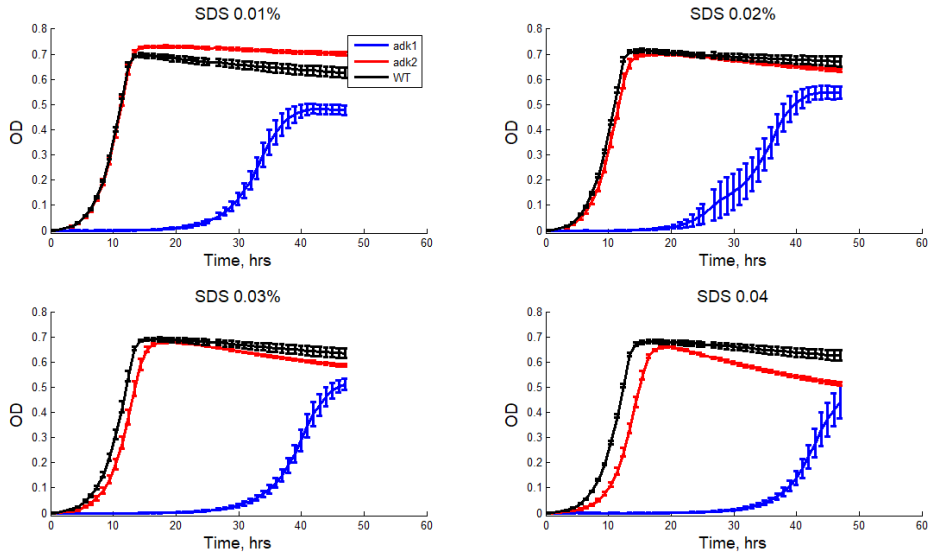
- Cafeína



- Menadiona



- SDS



- Sorbitol