SEÑALES DE SELECCIÓN NATURAL EN POBLACIONES ANDINAS EXPUESTAS HISTÓRICAMENTE AL ARSÉNICO EN EL DESIERTO DE ATACAMA

Tesis

Entregada A La
Universidad De Chile
En Cumplimiento Parcial De Los Requisitos
Para Optar Al Grado De

Magíster en Ciencias Biológicas

Facultad De Ciencias

Por

Mario Andrés Apata Mamani

Agosto, 2018

Director de Tesis Dr.: Mauricio Moraga V.

Co-Director de Tesis Dr.: Ricardo Verdugo S.

FACULTAD DE CIENCIAS

UNIVERSIDAD DE CHILE

INFORME DE APROBACIÓN

TESIS DE MAGÍSTER

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Magíster presentada por el candidato.

Mario Andrés Apata Mamani

Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Magíster en Ciencias Biológicas, en el examen de Defensa Privada de Tesis rendido el día 01 de Agosto de 2018.

Director de Tesis

Dr. Mauricio Moraga .....

Co-director de Tesis

Dr. Ricardo Verdugo .....

Comisión de Evaluación de la Tesis

Dr. Elie Poulin .....

Dr. Rodrigo Medel .....

BIOGRAFÍA

Desde muy joven he estado interesado en entender las diferencias históricas y biológicas que caracterizan a las poblaciones humanas. Tal interés me condujo a iniciar mis estudios de pregrado en la carrera de Antropología Física de la Universidad de Chile (2009-2015), donde pude adquirir un mayor conocimiento sobre nuestra diversidad biológica y cultural. En mi memoria de titulo, y gracias a la tutela del Dr. Mauricio Moraga del Programa de Genética Humana de la Facultad de Medicina, pude aplicar los conocimientos de genética y evolución huamana, proponiendo la adaptación humana a los altos niveles de arsénico en la Quebrada Camarones (Región de Arica y Parinacota). Dicho trabajo me permitió realizar dos pasantías de investigación para aplicar métodos genómicos de detección de señales de selección natural en humanos, bajo la tutela de la Dra. Anna Di Rienzo del Departamento de Genética Humana de la Universidad de Chicago. Dichas experiencias me motivaron a continuar mis estudios de Magíster en Ciencias Biológicas de la Universidad de Chile (2015-2018). Con el apoyo del Dr. Moraga y el Dr. Ricardo Verdugo, continue mi investigación sobre los efectos de la selección natural en humanos, a través de un enfoque genómico en la presente tesis de Magíster, buscando contribuir a entender la compleja historia microevolutiva y posibles procesos adaptativos locales en las poblaciones andinas del norte de Chile.

AGRADECIMIENTOS

A la beca Conicyt para estudios de Magíster en Chile y al proyecto Fondecyt 1140544 por contribuir al financiamiento de esta tesis. Además, a la Vise-rectoría de Asuntos Académicos, Departamento de Postgrado y Postítulo de la Universidad de Chile por financiar una estadía de investigación en la Universidad de Chicago durante Enero-Marzo de 2017.

Al proyecto FONDEF D10I1007 de ChileGenomico por facilitar el uso de muestras de la región de estudio. De igual forma al Proyecto Conicyt USA2013-0015 que facilitó el uso de datos genotípicos de poblaciones del Sur de Chile, y al Dr. Andrés Moreno-Estrada (LANGEBIO, México) por el acceso a datos genotípicos de la población de Puno (Perú).

A la Dra. Anna Di Rienzo, David Witonsky y John Lindo de la Universidad de Chicago, por su apoyo técnico y teórico, y además por facilitar el uso de datos de secuenciación genómica de individuos de etnia Ayamara para imputar mis datos.

A los miembros del laboratorio de Genética de Poblaciones y Evolución Humana y GENOMED del Programa de Genética Humana, ICBM, Facultad de Medicina de la Universidad de Chile, por su permanente apoyo, crítica y los agradables momentos compartidos durante estos años. Especialmente a Michael Orellana y a Patricio Pezo por su crítica constructiva con el avance de tesis.

Al Dr. Ricardo Verdugo co-tutor de esta tesis quien me ha entregado una valiosa crítica y apoyo, especialmente en los métodos genómicos utilizados.

Al Dr. Mauricio Moraga por ser un pilar fundamental desde el inicio al fin de esta tesis, por su crítica y aportes en genética y evolución humana, por las agradables conversaciones en la Universidad, por su humor, por su apoyo a mis distintos intereses científicos y docentes, y por su incalculable comprensión y paciencia.

INDICE

Resume	…….....	1
Abstract	…….....	2
Antecedentes	…….....	3
	Selección natural y adaptación humana……..	3
	Ambientes extremos y genes bajo selección en las Américas……..	5
	Histórica exposición al arsénico en el Desierto de Atacama…….	8
	Signos de adaptación al arsénico en poblaciones andinas……..	11
	Estudios genómicos en selección positiva……..	13
Planteamiento		18
	Hipótesis……	18
	Objetivos……	18
Material y Método	…….....	19
	Poblaciones, muestras y bases de datos…….....	19
	Estructura genética poblacional…….....	24
	Métodos de selección positiva…….....	27
	Identificación y comparación de regiones bajo selección.....……..	30
Resultados	…….....	34
	Imputación, Filtros y Marcadores ……..…	34
	Estructura poblacional…….....	35
	Identificación de selección positiva …….	38
	Enriquecimiento de procesos biológicos …….	49
	Genes candidatos a selección positiva en Camarones …….....	53
Discusión	……...	56
Conclusión	…….....	64
Bibliografía	…….....	66

LISTA DE TABLAS

Tabla 1	Niveles de exposición y riesgo tóxico por arsénico	19
Tabla 2	Poblaciones, Muestras y Base de Datos Genotípicos	22
Tabla 3	Descripción y criterios de los métodos genómicos para identificar selección positiva	30
Tabla 4	Resultados de la imputación	34
Tabla 5	Frecuencias relativas de las ancestrías global y local inferidas en las poblaciones del estudio.	37
Tabla 6	Marcadores utilizados en PBS	39
Tabla 7	Top 10 de SNPs en el 1% de valores más altos de PBS para Camarones (Escenarios A y B)	40
Tabla 8	Marcadores utilizados en iHS	45
Tabla 9	Top 10 de SNPs y ventanas (200kb) en el 1% de valores más altos de iHS en Camarones	45
Tabla 10	Marcadores utilizados en XP-EHH	48
Tabla 11	Top 10 de SNPs en el 1% de valores más altos de XP-EHH (Escenarios A y B)	48
Tabla 12	Enriquecimiento de procesos biológicos en Camarones a partir de los genes identificados en PBS	49
Tabla 13	Enriquecimiento de procesos biológicos en Camarones a partir de los genes identificados en iHS	52

LISTA DE FIGURAS

Figura 1	Genes identificados bajo selección natural en poblaciones humanas	4
Figura 2	Barrido Selectivo (*Selective sweep*)	5
Figura 3	Área de estudio y los niveles de exposición a arsénico	10
Figura 4	Frecuencia del haplotipo protector del gen AS3MT en Chile	13
Figura 5	Métodos genómicos para identificar señales de selección positiva.	17
Figura 6	Escenarios de comparación utilizados para el cálculo de PBS	28
Figura 7	Mapa de flujo con objetivos del estudio y métodos genómicos	33
Figura 8	Gráfico de PCA con poblaciones amerindias	35
Figura 9	Inferencia de ancestría genética con Admixture	37
Figura 10	Manhhattan Plot de valores de PBS en Camarones	38
Figura 11	Scatter plot del intersecto entre los valores de PBS para Camarones respecto a Puno y Sur de Chile	41
Figura 12	Regiones genómicas bajo selección identificadas con el análisis de PBS en Camarones	43
Figura 13	Manhattan plots con valores de iHS en Camarones	44
Figura 14	Intersecto del 1% de valores más altos de iHS entre las poblaciones de estudio	46
Figura 15	Manhattan plot de los valores de XP-EHH para Camarones (versus Puno y Sur de Chile)	47
Figura 16	Enriquecimiento de procesos biológicos en el 1% de valores de PBS	50
Figura 17	Enriquecimiento de procesos biológicos en el 1% de valores de iHS	52
Figura 18	Regiones bajo selección en el cromosoma 6 con PBS en Camarones	53
Figura 19	Regiones bajo selección en el cromosoma 6 con iHS en Camarones	54
Figura 20	Región bajo selección en el cromosoma 10 en Camarones	55

RESUMEN

El Desierto de Atacama es un ambiente extremo para los seres humanos por su alta radiación UV, zonas de hipoxia y altos niveles de arsénico en esenciales recursos de agua. El arsénico es reconocido por ser carcinogénico y por aumentar la mortalidad infantil al provocar abortos espontáneos. La población de Quebrada Camarones tiene la mayor exposición al arsénico en América (1000 µg/L), superando la norma segura de la OMS (10 µg/L). Sin embargo, esta población ha subsistido bajo tales condiciones los últimos 7.000 años, sin evidencias de los efectos tóxicos antes mencionados en tiempos recientes.

Nuestro objetivo fue identificar procesos adaptativos locales en la población de Camarones mediante la búsqueda de señales genómicas de selección positiva. Analizamos muestras de Camarones (n=26), Puno (n=30) y Sur de Chile (n=39), siendo las dos últimas los controles. Las muestras se genotipificaron con dos Arrays de SNPs (Axiom LAT1 y Mega), y posteriormente, imputadas usando dos paneles de referencia (población Aymara y de 1000Genomas). Luego, caracterizamos su estructura genética e identificamos las señales de selección a través de los métodos de PBS y los basados en LD.

Hemos encontrado diferentes señales de selección en Camarones, mayoría en genes cercanos a la región de HLA o de genes relacionados con el arsénico y el estrés de hipoxia. Respecto a la adaptación al arsénico, nuestros resultados sugieren que Camarones y Puno comparten una misma señal de selección, sobre el gen CNNM2 en el cromosoma 10, el cual esta muy cerca al gen AS3MT. Por otra parte, procesos biológicos sobrerepresentados en Camarones están relacionados con la regulación muscular y presión arterial con genes involucrados en la homeostasis del oxigeno (ACE y NOX4). En conclusión, nuestros resultados sugieren un complejo escenario adaptativo entre las poblaciones del altiplano y el Desierto de Atacama, en que algunas señales son únicas en Camarones, pero otras como el arsénico han mostrado estar compartidas con otras poblaciones andinas.

ABSTRACT

The Atacama Desert is an extreme environment for humans due to high UV radiation, hypoxia and high arsenic levels in essential water sources. Arsenic is widely recognized to be carcinogenic and to increase child mortality rates by causing spontaneous abortions. People from Quebrada Camarones have the highest exposure to arsenic in the Americas (1000 µg/L), surpassing the WHO's safe norm (10 µg/L). However, these people have subsisted under these conditions over the last 7,000 years, and without showing the aformentioned toxicological effects in recent times.

The aim of this study was to identify unique local adaptations in Camarones people by searching for genomic signatures of positive selection. We used samples from Camarones (n=26), Puno (n=30) and South of Chile (n=39), the last two populations being used as controls. Those samples were genotyped by two SNPs arrays (Axiom LAT1 and Mega) and subsequently imputed by using two reference panels (Ayamra population and 1000Genomes). Then, we characterized their genetic structure, and searched for selection signatures by using PBS and LD-based methods.

We have found different positive selection signatures in Camarones, most of which are genes over HLA region or genes related to arsenic and hypoxic stress. Regard to arsenic adaptation, our results suggest that Camarones and Puno have similar selection signals, especially over the CNNM2 gene in chromosome 10, which is closely to the AS3MT gene. On the other hand, the biological processes overrepresented in Camarones were found in muscle and arterial pressure regulation with genes involved in oxygen homeostasis (ACE and NOX4). Therefore, our results suggest a more complex adaptative scenario between highlands and the Atacama Dessert, where some signals are unique in Camarones people, but others like arsenic have shown to be shared with other Andean populations.

ANTECEDENTES

I. Selección natural y adaptación humana

Desde su salida de África hace unos 120.000-60.000 años, las poblaciones humanas fueron colonizando diversas regiones geográficas del mundo, debiendo adaptarse biológica y culturalmente a diversas presiones selectivas de su entorno (Jeong & Di Rienzo, 2014). Tal expansión ocurrió en un relativamente corto periodo de tiempo, que ha sido suficiente para que la selección natural actuara sobre diferentes fenotipos ventajosos, posibilitando la adaptación humana a condiciones ambientales adversas, tales como temperaturas extremas, diferentes niveles de radiación UV, hipoxia de altura, dietas ricas en lactosa, grasas saturadas y trazas de selenio, y a enfermedades como la malaria y el cólera, e incluso a tóxicos naturales como el arsénico (Fan, Hansen, Lo &Tishkoff, 2016) (Figura 1).

Las fuerzas evolutivas como la mutación, el flujo génico, la deriva genética y la selección natural han provocado cambios en las frecuencias alélicas a lo largo del tiempo en las poblaciones humanas, configurando diferentes historias micro-evolutivas durante su expansión mundial (Jobling et al., 2013). Particularmente, es la selección natural la cual nos permite entender cómo algunas poblaciones humanas han logrado sobrevivir en ambientes extremos, a través de su principal consecuencia conocida como adaptación biológica (Herron & Freeman, 2015). La adaptación se caracteriza por el aumento en frecuencia de fenotipos con capacidad diferencial de sobrevivida y reproducción en sus portadores (*fitness darwiniano*), y como consecuencia de determinadas condiciones del entorno (presión selectiva). Lo que, en términos genéticos, se traduce en el aumento de la frecuencia de los alelos que determinan dichos fenotipos hasta alcanzar su fijación en la población (100%) (Vitti, Grossman and Sabeti, 2013).

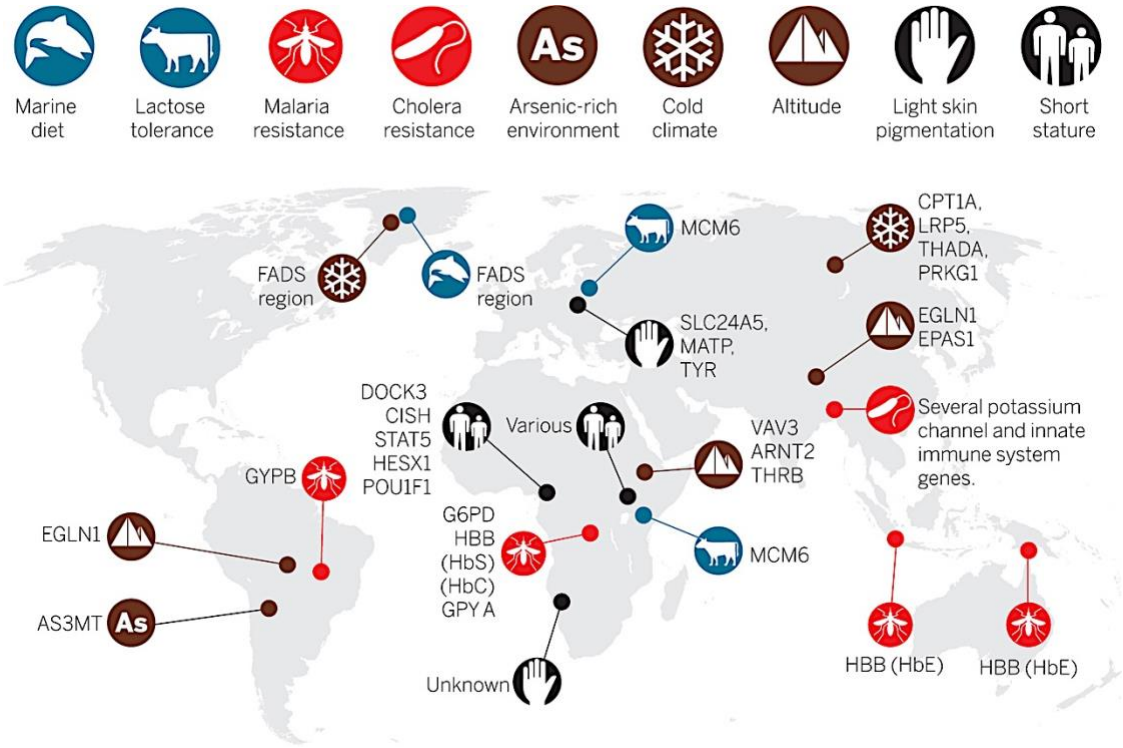


Figura 1. Genes identificados bajo selección natural en poblaciones humanas (Tishkoff S, 2015)

Existen al menos tres formas en que puede operar la selección natural a nivel genético. En primer lugar, la ***selección negativa o purificadora***, que consiste en la eliminación de alelos deletéreos que van apareciendo en el genoma producto de mutaciones o errores en la replicación del ADN. En segundo lugar, la ***selección balanceadora*** permite que se conserven varios alelos de un locus de manera equilibrada, favoreciendo la presencia de fenotipos intermedios (estabilizadora) o extremos (disruptiva) en una población. En tercer lugar, la ***selección positiva*** actúa incrementando rápidamente la frecuencia de los alelos que determinan los fenotipos con mayor *fitness* dentro de una población (Wollstein & Stephan, 2015; Vitti et al., 2013). Un efecto de este tipo de selección es la reducción de la variación tanto a nivel del locus seleccionado como de los *loci* cercanos físicamente, produciendo un efecto de

barrido selectivo (“*selective sweep”*). Los barridos selectivos se caracterizan por regiones del genoma con un alto desequilibrio de ligamiento (LD, “Linkage desequilibrium”), conformando largos haplotipos que tienen una menor tasa de recombinación que bajo condiciones de neutralidad (Figura 2). (Pritchard, Pickrell and Coop, 2010).

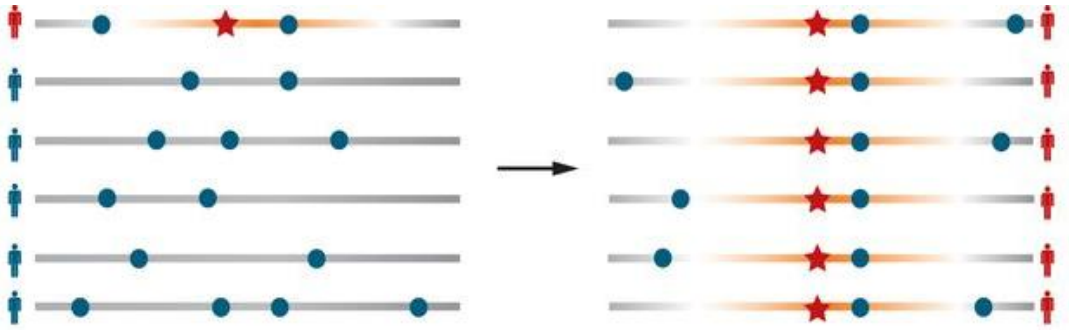


Figura 2. Barrido Selectivo (Selective sweep). Las barras representan los diferentes haplotipos de cada individuo antes y después del tiempo de acción de la selección natural (flecha). Los individuos en rojo son portadores de la variante con mayor fitness (estrella roja). Después de la acción de la selección, los portadores de la variante ventajosa también poseen las variantes que la acompañaban conformando largos haplotipos con alto desequilibrio de ligamiento (color naranja en las barras).

II. Ambientes extremos y genes bajo selección en las Américas

Gran parte de las adaptaciones locales en las poblaciones humanas ocurrieron al emigrar desde el continente africano y colonizar diferentes regiones del mundo. Hace unos 20.000-15.000 años, el ser humano colonizó por primera vez las Américas, cruzando el estrecho de Bering luego del último máximo glacial (LMG, Last Maximun Glacial) y alcanzando rápidamente el cono sur del continente hace unos 14.000-13.000 años (Goebel, Waters y O’Rouge, 2008). Sin embargo, aún no existe consenso sobre diferentes interrogantes como el tiempo de divergencia entre las primeras poblaciones amerindias y las

del noreste asiático, los posibles efectos de la deriva en la baja diversidad genética observada en los lineajes uniparentales de las poblaciones amerindias, el número de oleadas migratorias y si éstas pueden explicar las diferencias genéticas entre los macro-grupos amerindios, o si dichas diferencias son consecuencia de la diversificación *in situ* a lo largo del continente (Raghavan et al., 2015; Reich et al., 2012).

En los últimos años, a través de herramientas genómicas y utilizando tanto ADN antiguo como moderno, se ha demostrado diferencias a nivel de la estructura genética de las poblaciones amerindias, distinguiendo macro-grupos a nivel genético a lo largo del continente (Raghavan et al., 2016; Hamburger et al., 2015; Reich et al., 2012). Sugiriendo, que han ocurrido procesos micro-evolutivos distintos a nivel de sus historias demográficas, efecto de migraciones recientes y adaptaciones locales que han contribuido a explicar la diversidad biológica y cultural de las actuales poblaciones amerindias.

Las adaptaciones locales estudiadas en poblaciones amerindias tradicionalmente se han focalizado en el contraste de ambientes extremos del continente, como por ejemplo, los climas polares del ártico en norte América y en la Patagonia en el sur del continente, la hipoxia provocada en el altiplano de los Andes centro-sur, y los ambientes selváticos-tropicales y enfermedades asociadas en la Amazonía (Salzano F, 2016). La ocupación del altiplano en los Andes centro-sur se estima al menos hace unos 11.500 años atrás (Rademaker et al., 2013). Los ambientes de altura tienen como principal caracteristica generar condiciones de hipoxia (i.e. bajos niveles de oxígeno) en los organismos dependientes de oxígeno, lo cual ha permitido postular diferentes procesos adaptativos para las poblaciones humanas que allí habitan. Por ejemplo, recientes estudios han identificado genes bajo selección, en diferentes poblaciones de orígen étnico andino (e.g. Aymaras, Quechuas y Collas), relacionados con la vía inducida por hipoxia (HIF pathways) como el gen

EGLN1 (Bigham & Lee, 2014; Bigham et al., 2013; Bigham et al. 2010), con el control muscular de la presión sanguínea y procesos de óxido-reducción (NO pathways) como el PRKG1 (Eichstaedt et al., 2015a), y el gen codificante de la proteina del surfactante D o SFTPD1 que cumple una función protectora en los alveólos (Valverde et al., 2015).

En cuanto a la adaptación a climas de extremo frío, se han estudiado poblaciones cercanas a las amerindias en Siberia, donde se han encontrado genes asociados con el control de energía y metabolismo mitocondrial como el CPT1A, LRP5, THADA y también el gen PRKG1 observado previamente en condiciones de hipoxia (Cardona et al., 2014). En Inuit de Groenlandia y el norte de América, Fumagalli et al. (2015) ha encontrado señales de selección sobre genes involucrados en el metabolismo de ácidos grasos, y específicamente con omega-3, proveniente de alimentos en su entorno (e.g. peces y mamíferos marinos). No obstante, la señal de selección sobre los genes de desaturasas de ácidos grasos (FADS) observada en los Inuit, también ha sido encontrada con menor intensidad en diferentes poblaciones amerindias a lo largo del continente, lo cual ha hecho hipotétizar que la capacidad de metabolizar ácidos grasos fue fuertemente seleccionada en una población ancestral a las amerindias durante el paso por Beringia o en tiempos cercanos a dicho proceso (Amorin et al., 2017).

En la literatura científica también se han encontrado adaptaciones locales a enfermedades, siendo la región genómica del sistema de antígenos leucocitarios humanos (HLA) una de las más estudiadas. Lindo et al. (2016), analizó restos humanos de la costa noroeste de América datados entre 6.260-1.036 años antes del presente (AP), e identificó una señal de selección sobre variantes del gen HLA-DQA1, sugiriendo una respuesta adaptativa en el sistema inmune previa a la colonización europea en el siglo XVI. A su vez, en los últimos años también se ha estudiado un evento adaptativo en la capacidad

de metabolizar tóxicos naturales como el arsénico en ambientes desérticos y de altura en poblaciones andinas en Sud América (Apata et al., 2017; Eichstaedt et al., 2015b; Schelebush et al., 2015).

III. Histórica exposición al arsénico en el Desierto de Atacama

El Desierto de Atacama ubicado entre el extremo norte de Chile y el sur del Perú, es el más árido del mundo y un ambiente extremo para la sobrevivencia de las poblaciones humanas, debido a su extrema aridez, abruptas oscilaciones térmicas en el día, alta radiación UV y zonas de hipoxia de altura, siendo los escasos recursos de agua dulce como ríos y quebradas, esenciales para la vida humana desde tiempos prehispánicos (Arriaza & Standen 2008). Sin embargo, muchos de esos recursos de agua se han encontrado naturalmente contaminados con altos niveles de arsénico durante milenios, debido a su origen geogénico y volcánico en la cordillera de los Andes (Figura 3) (Bunschuh et al., 2012; Mukherjee et al., 2014; López et al., 2012). Siendo Quebrada de Camarones, en la región de Arica y Parinacota, donde se encuentra los niveles de arsénico más altos en América (1000 µg/L), superando ampliamente el nivel seguro establecido por la Organización Mundial de la Salud (10 µg/L) (WHO 2011; Cornejo-Ponce et al., 2011; Arriaza et al., 2010; Bundschuh et al., 2012).

Los ambientes ricos en arsénico han demostrado ser muy nocivos para las poblaciones humanas, debido a sus efectos carcinogénicos, genotóxicos y principalmente por provocar abortos espontáneos (Marshall et al., 2007; Hopenhayn-Reich et al., 2000; Hopenhayn et al., 2003). Tales efectos han sido observados en poblaciones expuestas en tiempos relativamente recientes, como Antofagasta y Bangladesh durante la década de los sesenta y setenta, diferenciándose esta última por la prevalencia de lesiones a la piel (hiperqueratosis e hiper-pigmentación), signo diagnóstico de la intoxicación por arsénico en el sudeste asiático (Ashan et al., 2006; Ferrecio and Sancha, 2006;

Steinmaus et al., 2014). A diferencia de la exposición reciente de Antofagasta y Bangladesh, la población actual de Quebrada Camarones ha estado expuesta al arsénico de forma crónica sin presentar dichos efectos tóxicos en tiempos relativamente recientes. No obstante, algunos signos de envenenamiento por arsénico se han encontrado en las poblaciones prehispánicas de la zona durante el Periodo Arcaico (7000-3000 años AP), siendo los grupos humanos de la Cultura Chinchorro uno de los casos más estudiados (Figueroa 2001; Arriaza et al., 2010; Bartkus et al., 2011; Byrne et al., 2010; Swift et al., 2015).

Los Chinchorro fueron cazadores-recolectores y pescadores que habitaron el litoral y valles costeros del desierto de Atacama, siendo sus primeros asentamientos en Quebrada Camarones y en lo que hoy es la actual ciudad de Arica (Arriaza & Standen 2008). Arriaza (2005) planteó la hipótesis de que el envenenamiento por arsénico aumentó la tasa de abortos de infantes Chinchorro, provocando como respuesta emocional y cultural el inicio de la momificación artificial, principal práctica mortuoria de esta cultura. Diferentes estudios han medido las concentraciones de arsénico en muestras de cabello y hueso de restos momificados de Chinchorro, encontrando que la mayor exposición al arsénico ocurrió en el Periodo del Arcaico Temprano (7000-5000 AP) (Arriaza et al., 2010; Swift et al., 2015). Además, se ha observado una disminución de los valores promedio de arsénico en el cabello de individuos de diferentes periodos crono-culturales, a partir del periodo Arcaico con Chinchorro (5000-3000 años AP), el Formativo con poblaciones prehispánicas agro-alfareras (3000-500 años atrás), hasta las actuales poblaciones de Camarones (Yáñez et al., 2005; Bartkus et al., 2011; Byrne et al., 2010; Arriaza et al., 2010). Tal diferencia en los niveles de arsénico podría deberse a que las poblaciones locales de Camarones se han adaptado, durante los últimos 7000 años, a un ambiente tóxico mediante un incremento de la capacidad metabolizadora del arsénico, que es la principal forma de detoxificación en los seres humanos.

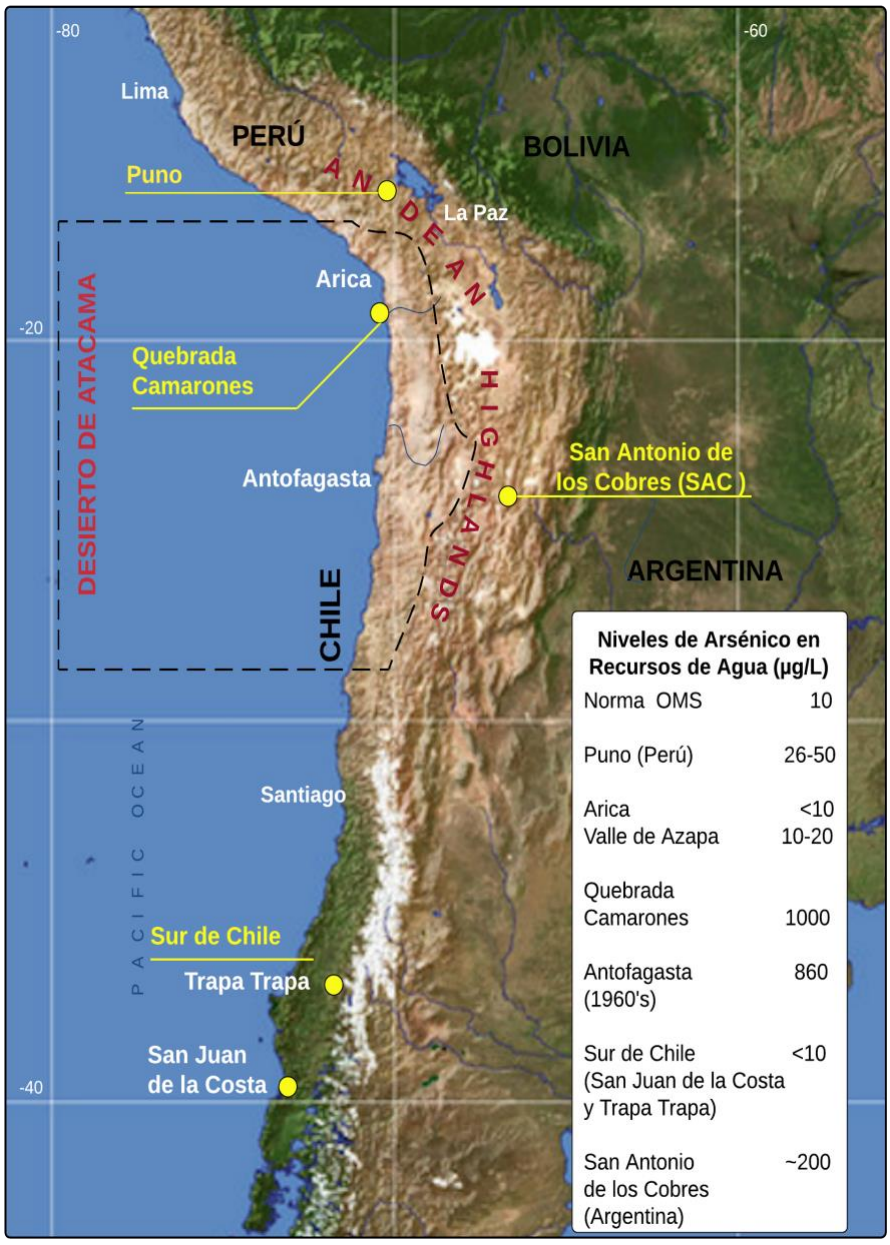


Figura 3. Área de estudio y niveles de exposición a arsénico. En el mapa se indica el lugar geográfico del Desierto de Atacama y de Quebrada Camarones. Así como los niveles de arsénico al que están expuestas las poblaciones humanas en la zona de los Andes Centro-Sur en Sud América.

IV. Signos de adaptación al arsénico en las poblaciones andinas

La eficiente metabolización del arsénico en humanos se caracteriza por una alta capacidad metiladora de la enzima Arsénico (+3) metiltransferasa (AS3MT) en la transformación de compuestos inorgánicos a mono-metilarsénico (MMA) y di-metilarsénico (DMA) (Lin et al., 2002; Thomas et al., 2007; Sumi & Himeno 2012). Siendo MMA, la especie más citotóxica y genotóxica del arsénico (Styblo et al., 2000; Mass et al., 2001). Una alta capacidad metiladora está asociada con mayores niveles de DMA y menores de MMA excretados en la orina (Gardner et al., 2010). La relación de especies metiladas del arsénico posee una amplia variabilidad inter-poblacional (Hughes et al., 2011; Agusa et al., 2011; Antonelli et al., 2014). Sin embargo, existen poblaciones indígenas de Sud América como la población de San Antonio de los Cobres (SAC) en Argentina, que presentan una alta capacidad metiladora (Engström et al., 2011). La población de SAC también ha estado expuesta históricamente a altos niveles de arsénico (200 µg/L), y ha mostrado una significativa asociación entre una alta capacidad metiladora con la mayor frecuencia de tres variantes en el gen codificante de la AS3MT (haplotipo CTA) (Schlawicke-Engström et al., 2007; Engström et al., 2011; Schlebush et al., 2013).

Las evidencias de variación fenotípica en la metilación del arsénico han llevado a proponer la hipótesis de adaptación al arsénico en poblaciones con una milenaria exposición a este metaloide en SAC (Schlebush et al., 2013; et al., 2015; Agusa et al., 2011). Dicha hipótesis también fue evaluada en la población de Camarones por Apata et al. (2017), donde encontraron que la frecuencia del haplotipo protector, compuesto por las tres variantes asociadas a una alta capacidad metiladora, tuvo una similar frecuencia a la observada en San Antonio de los Cobres (CTA=68%) (Figura 4). Además, en ese estudio se mostró que la frecuencia del alelo C asociado a mayor riesgo de toxicidad en el SNP 14458 (Met287Thr) del gen AS3MT, fue extremadamente baja en la población de Camarones (1%), lo que también ocurre en SAC (2%) (Antonelli et

al., 2014). Mientras que dicho alelo aumenta progresivamente hacia el norte de Camarones, con un 5% en el valle de Azapa, y hacia el sur de Chile alcanza el 16%(Apata et al., 2017). Dichas evidencias son congruentes con datos conocidos para las poblaciones del sudeste asiático y en Antofagasta donde las frecuencias también alcanzan el 15% (Hernández et al., 2008a; 2008b; Antonelli et al., 2014). Dicho alelo ha mostrado estar asociado con una baja eficiencia de la metabolización del arsénico (mayores niveles de MMA en la orina), efectos genotóxicos y con mayor riesgo de lesiones a la piel (Valenzuela et al. 2009; Hernández et al. 2014; Hsu et al., 2013). Por lo tanto, la baja frecuencia de ese alelo en Camarones y SAC, podría estar señalando que el proceso adaptativo en ambas poblaciones fue, por un lado, seleccionando positivamente los alelos de eficiente metabolización, y por otro, negativamente alelos de riesgo, disminuyendo los efectos tóxicos asociados al arsénico (Apata et al., 2017).

Los estudios de asociación a nivel genómico (GWAs) que se han desarrollado en poblaciones expuestas al arsénico, tanto en Bangladesh como en SAC, han mostrado una significativa asociación entre la región genómica cercana al gen AS3MT con la variación de los niveles de DMA y MMA (Pierce et al. 2012; Engström et al., 2013; Schlebush et al., 2015). Dichos estudios han identificado en la población de Bangladesh, la asociación de SNPs dentro del gen AS3MT y en otros genes de la región cromosómica 10q21 (CNNM2 y C10orf32-AS3MT), con una baja eficiencia metabolizadora del arsénico (altos niveles de MMA y bajos de DMA) (Pierce et al., 2012; 2013). Mientras que en la población de SAC, se ha evidenciado una única señal de asociación entre una eficiente metabolización (altos niveles de DMA y bajos de MMA) con la región genómica cercana al gen AS3MT (Schlebush et al., 2015). Schlebush et al. (2015) y Eichstaedt et al. (2015b) han identificado una señal de selección positiva en la región cercana al gen AS3MT para la población de SAC, dicha señal fue significativamente diferente de las poblaciones utilizadas como controles (peruanos, colombianos y otras poblaciones argentinas), confirmando

un proceso adaptativo local para el arsénico en la región altiplánica del noroeste argentino.

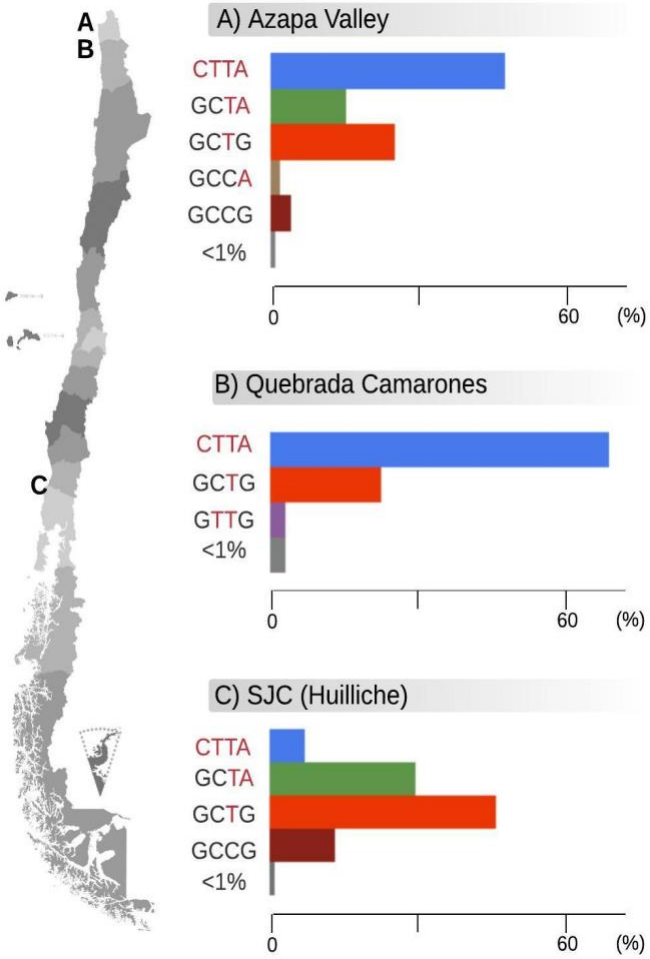


Figura 4. Frecuencias del haplotipo protector CTTA del gen AS3MT en Chile (Modificado de Apata et al., 2017).

V. Estudios genómicos en selección positiva

A partir de la secuenciación del genoma humano en el año 2001 y el desarrollo de nuevas tecnologías de genotipificación y secuenciación, la disciplina de la genética ha transitado progresivamente desde un enfoque clásico, focalizado en genes específicos, hacia el genómico, donde se analizan diferentes regiones genómicas de una especie. Con el uso de herramientas bioinformáticas, la genómica ha logrado manipular y analizar datos con una alta densidad de marcadores genéticos, como por ejemplo los Chips o Arrays

comerciales con cientos de miles de SNPs (por su sigla en inglés *Single Nucleotide Polymorphism*, o polimorfismos de un solo nucleótido) o por medio de secuenciación de segunda generación (NGS, *Next Generation Sequencing*).

En la era de la genómica, la selección es entendida como la propagación diferencial de un alelo como consecuencia de su efecto fenotípico y no del azar (Vitti et al., 2013). La genómica en estudios de selección se ha caracterizado por definir las hipótesis adaptativas *a posteriori*, diferenciándose del "enfoque clásico" o de "genes candidatos", donde las hipótesis adaptativas son definidas *a priori* sobre genes con fuerte evidencia de ser responsables de fenotipos bajo selección (Akey JM, 2009; Sabeti et al., 2006). En los últimos años, los estudios genómicos han buscado identificar mayoritariamente señales de selección positiva, mediante métodos basados en desequilibrio de ligamiento (LD de su nombre en inglés *Linkage Desequilibrium*) y por diferenciación de frecuencias alélicas a nivel inter-poblacional (Figura 5). Dichos métodos se han caracterizado por evaluar procesos micro-evolutivos (i.e. intra-especie), y han mostrado tener alcances temporales entre los 80.000 y 30.000 años, permitiendo conocer adaptaciones locales cercanas al *Out of Africa*, y a la ocupación temprana de nuevos entornos en el mundo por la especie humana (Oleksky, Smith and O'Brien, 2010).

Los métodos que utilizan LD son ampliamente usados para detectar barridos selectivos fuertes y suaves o incompletos (Sabeti et al. 2006). Tienen como alcance temporal los 30.000 años, ya que después de esta fecha un cromosoma humano típico habrá sufrido más de un cruce de recombinación por cada 100 kb, fragmentando los haplotipos y disminuyendo su frecuencia hasta hacerlos indistinguibles (Oleksky et al., 2010; Sabeti et al., 2005). Como principales limitantes metodológicas, estos métodos requieren conocer la fase gamética de los alelos para definir explícitamente los haplotipos, genomas sin eventos recientes de recombinación y muestras que no estén emparentadas.

Sin embargo, éstos han sido ampliamente utilizados para identificar eventos de selección recientes (9000-7000 años), tales como las variantes involucradas en la persistencia de la lactasa en poblaciones africanas y asiáticas. Los parámetros que utilizan son el número de SNPs en LD, la distancia y extensión física de los haplotipos, y la homocigosidad dentro de las diferentes regiones en las que es fragmentado el genoma (ventanas genómicas), que comúnmente son definidas entre 200 y 100Kb. Ejemplos de métodos basados en LD son el de homocigosidad del haplotipo extendido EHH (*Extended Haplotype Homozygosity*, Sabeti et al. 2002), haplotipos de amplio rango LRH (*Long-Range Haplotype*, Zhang et al. 2006; Sabeti et al. 2002), puntaje integrado de haplotipos iHS (*integrated Haplotype Score*, Voight, Kudaravalli, Wen y Pritchard, 2006), homocigosidad del haplotipo extendido inter-poblacional XP-EHH (*Cross-population extended haplotype homozygosity*, Sabeti et al., 2007), y el decaimiento de LD (*Linkage desequilibrium decay*, Wang, Kodama, Baldi y Moyzis, 2006).

El segundo grupo de métodos son los basados en diferencias alélicas a nivel inter-poblacional. Estos buscan detectar variantes genéticas fijadas como consecuencia de la selección positiva, bajo los supuestos que diferentes historias evolutivas y ambientes particulares entre dos o más poblaciones comparadas, se verán reflejadas en sus diferencias de frecuencias alélicas extremas y puntuales en ciertas regiones del genoma (Vitti et al., 2013). Este grupo utiliza como base metodológica el índice de diferenciación poblacional F_{ST}, considerando como alelos bajo selección a aquellos *loci* con valores extremos en comparación con el resto del genoma que podría estar más susceptible a eventos de deriva, por lo que se consideran convencionalmente regiones o genes bajo selección a aquellos que estén dentro del 1% de la distribución de valores de F_{ST} (Holsinger and Weir, 2009). El alcance temporal de los métodos que utilizan F_{ST} es de 80.000 años, ya que el barrido selectivo de escenarios adaptativos tempranos puede haber sido afectado tanto por la

recombinación y nuevas mutaciones, pero aún se mantendría la alta o moderada frecuencia de aquellos alelos bajo selección (Sabeti et al., 2006). Algunos métodos basados en F_{ST} a nivel genómico (F_{ST} de Weir and Cokehall), son el Locus con largo de rama especifica (LSBL, Locus Specific Branch Lengh, Shriver et al., 2004) y su versión mejorada al adherir la conversión de Cavalli-Sforza a distancias genéticas en cada rama con el estadístico de rama poblacional (PBS, Population Branch Statistic) (Yi et al., 2010).

Un tercer enfoque metodológico son los que evalúan la asociación entre alguna región del genoma con algún fenotipo de interés. Métodos bajo este enfoque siguen la idea de los estudios de asociación genómica “Genome-Whole Association study o GWAs”, donde es posible relacionar estadísticamente regiones, genes o variantes genéticas con uno o más fenotipos, lo cual ha sido útil para identificar señales de selección positiva, balanceadora y sobre todo posibles eventos de selección poligénica de fenotipos complejos, donde existe la interacción de varios genes (p. Ej. genes relacionados con la variación de estatura, pigmentación de la piel, variación de temperaturas e hipoxia de altura) (Fan et al., 2016; Berg &Coop, 2014).

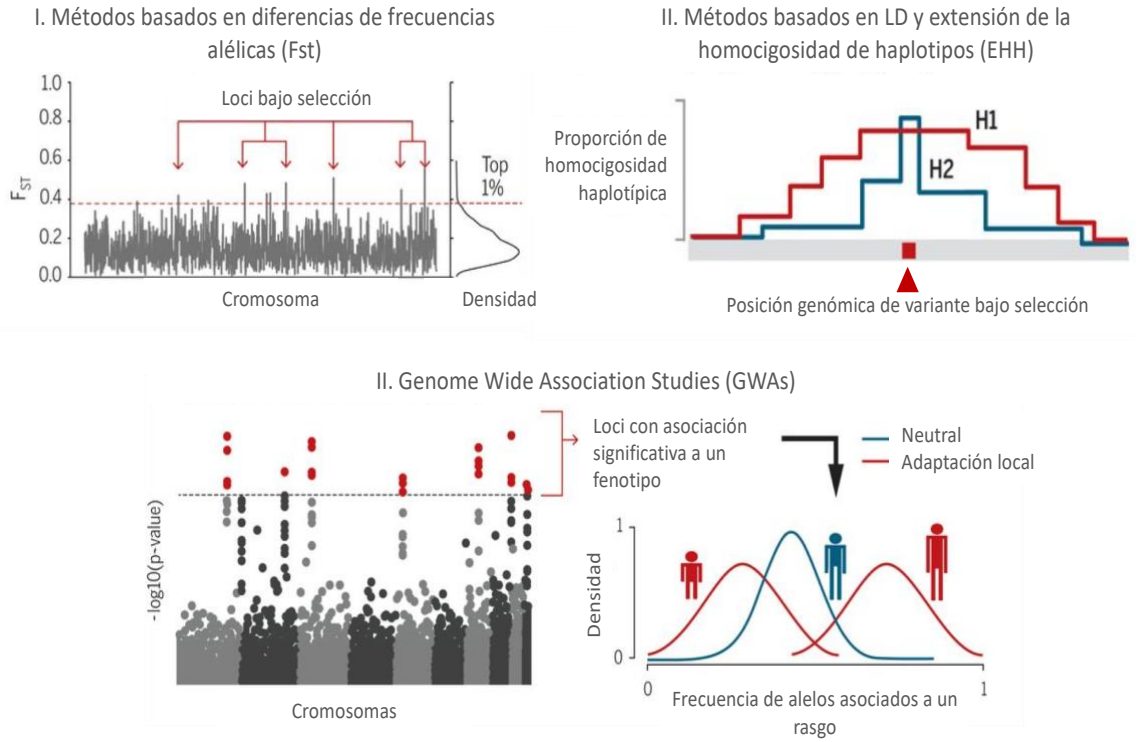


Figura 5. Métodos genómicos para identificar señales de selección positiva. Modificado de Fan et al. (2016). En rojo, se indican las variantes y haplotipo bajo selección.

Considerando las evidencias ecológicas, históricas y genéticas que apoyan un proceso adaptativo al arsénico en Quebrada Camarones, nos proponemos evaluar si la selección natural ha actuado sobre esta población, mediante la búsqueda de señales de selección positiva sobre regiones genómicas asociadas con el metabolismo del arsénico, o con otras regiones aún desconocidas, así como con otras presiones selectivas del entorno extremo en el Desierto de Atacama. Por lo que, nuestra pregunta de investigación es: *¿La población de Camarones posee procesos adaptativos asociadas con el metabolismo del arsénico o con otras presiones del entorno del Desierto de Atacama?*

PLANTEAMIENTO

Hipótesis

La población de Quebrada de Camarones posee adaptaciones biológicas en respuesta a una continua exposición a los altos niveles de arsénico, que la diferencian de otras poblaciones de similar ancestría amerindia sin exposición a este elemento.

Objetivo General

Evaluar posibles eventos adaptativos a partir de señales de selección postiva a nivel genómico en la población de Quebrada Camarones.

Objetivos Específicos

1. Caracterizar la estructura genética poblacional a nivel genómico de la población de Quebrada Camarones y de las poblaciones nativo americanas utilizadas como control (Puno y Sur de Chile).

2. Analizar e identificar regiones genómicas con señales de selección natural positiva en las poblaciones de estudio (Quebrada Camarones y controles).

3. Determinar si existen diferencias entre las señales de selección natural positiva identificadas en la población de Quebrada Camarones con las poblaciones utilizadas como control (Puno y Sur de Chile).

MATERIALES Y MÉTODOS

I. Poblaciones, muestras y bases de datos

Selección de poblaciones y muestras

Como diseño de estudio, se utilizó la población proveniente de la localidad de Quebrada Camarones, la cual posee antecedentes de una fuerte presión selectiva producto del arsénico y por lo tanto ha sido definida como población objetiva de selección para esa y otras posibles presiones selectivas propias de su historia evolutiva. Además, se incluyen dos poblaciones definidas como controles, siendo una proveniente de la localidad de Puno en Perú, y una población del sur de Chile (SCH) compuesta por las localidades de Trapa Trapa y San Juan de la Costa. Las muestras de Puno y SCH fueron elegidas debido a que son poblaciones con un fuerte componente genético amerindio, una ocupación temprana en sus respectivas zonas geográficas, así como poseer una histórica exposición a bajos niveles de arsénico y mayor disponibilidad de recursos hídricos (lluvias y lagos) (Tabla 1).

Tabla 1. Niveles de exposición y riesgo tóxico por arsénico

Población	Etnia	*N*	Niveles de arsénico (µg/L)	Riesgo tóxico en humanos [d]
Camarones	Aymara	26	1000-1300 [a]	Riesgo crítico
Puno	Aymara	30	26-50 [b]	Riesgo bajo
Sur de Chile	Huilliche-Pehuenche	39	<10 [c]	Sin riesgo

Las mediciones de arsénico son reportadas en los siguientes estudios [a]Arriaza et al. (2010), [b]George CM et al. (2014), [c]Informe del servicio sanitario SISS 2015/NCh409/1.Of.2005. [d]La OMS establece esta clasificación de riesgo tóxico al arsénico basado en los niveles de arsénico en agua de consumo y sus consecuencias para la salud humana, donde el límite seguro, sin riesgo tóxico, son los 10 µg/L; valores entre 10-50 µg/L son de riesgo bajo; 51-100 µg/L de riesgo medio; y >100 µg/L de riesgo crítico y es dónde hay efectos nocivos en la salud humana.

Quebrada de Camarones

Un total de 51 muestras fueron colectadas en el año 2013 en diferentes localidades de la Quebrada de Camarones, utilizando como criterio de muestreo individuos mayores de edad (≥18 años) y no emparentados (Apata et al., 2017). De los cuales se consideraron como "locales" de Camarones, aquellos con residencia permanente y que tuvieran una historia de ocupación de tres o más generaciones. A su vez, todos los individuos locales (n=36) declararon su pertenencia a la etnia Aymara. Con el objetivo de conocer la proporción de ancestría genética, se les realizó un set de 147 SNPs definidos como marcadores informativos de ancestría (AIMs), donde se estimó los componentes de ancestría mediante el algoritmo de Admixture, utilizando como poblaciones de referencia a individuos con sobre el 99% con ancestría europea (CEU), africana (YRI) y amerindia andina (Puno) y del sur de Chile (SCL) (Anexos 1). A partir del admixture se seleccionaron 19 individuos que poseían sobre el 90% de ancestría amerindia, los cuales se enviaron a genotipificar para 1.7 Millones de SNPs con el panel de Illumina Multi-ethnic Global Array (MEGA) en las instalaciones del laboratorio del Dr. Andrés Moreno-Estrada (LANGEBIO, México), financiado por el proyecto Fondecyt 1140544.

Por otra parte, se seleccionaron 15 individuos locales de Camarones obtenidos en el proyecto de ChileGenómico[1]. Estos individuos habían sido previamente genotipificados con un Array de 817.810 SNPs de Axiom Human Genome- Wide LAT 1 de Affymetrix (Hoffmann et al., 2011) en el Institute for Human Genetics, University of California, San Francisco, California, Estados Unidos (AXIOM). Estos individuos fueron elegidos bajo los criterios de residencia en la comuna de Camarones, y por poseer al menos un padre o madre que haya nacido en Camarones. Luego mediante un análisis de Admixture con el modelo de cinco poblaciones ancestrales antes mencionadas, se seleccionaron los individuos con más del 90% de ancestría amerindia, de los

[1] Proyecto Chile Genómico: http://www.chilegenomico.cl

cuales 7 individuos pasaron este filtro. Considerando los 19 y 15 individuos genotipificados con MEGA y AXIOM, la muestra total de Camarones se compone de 26 individuos (ver Tabla 2).

Puno

La población de Puno corresponde a individuos muestreados durante el año 2013 y genotipados por el grupo del Dr. Carlos Bustamante y Dr. Andrés Moreno-Estrada de la Universidad de Stanford, quienes autorizaron su uso en esta tesis. Los datos de Puno corresponden a 106 individuos, quienes residen en dicha localidad y que poseen al menos un abuelo perteneciente a la etnia Aymara. La etnia Aymara está ampliamente representada en la región de Puno, siendo la de mayor presencia, seguida por las etnias Quechua y Uros (Sandoval et al., 2016). Este set de datos fue genotipificado con el Array de 817.810 SNPs de Axiom Lat1 de Affimetrix. En base a análisis previos en el laboratorio se ha podido determinar que la proporción de ancestría genética amerindia de la muestra de Puno supera ampliamente el 90%, por lo que para este estudio seleccionamos 30 individuos con el 99% de ancestría genética amerindia (Tabla 2).

Sur de Chile (SLC)

La muestra del sur de Chile está compuesta por 22 individuos provenientes de las localidades de La Misión en la comuna de San Juan de la Costa (X Región), y 17 de Trapa-Trapa en la comuna del Alto Bío-Bío (VIII Región), ambas colectadas en el año 1991. Los individuos que componen la muestra del sur de Chile se adscriben a las etnias indígenas Huilliche y Pehuenche respectivamente. Esta muestra se encuentra disponible gracias al apoyo del proyecto PatagoniaDNA CONICYT DRI USA 2013-0015. También, se le realizó un análisis de Admixture para determinar el porcentaje de ancestría genética amerindia, seleccionándose 39 individuos con sobre el 90% de dicho componente (Tabla 2).

Poblaciones del Proyecto 1000 Genomas

Utilizamos los datos disponibles en 1000 Genome Project (1KG) para obtener poblaciones de referencia de ancestría continental, siendo los Yoruba de Ibadan (YRI) en Nigeria para África, Chinos Han en Beijin (CHB) para Asia y residentes de Utah con ancestría del norte y oeste de Europa (CEU) (The 1000 Genomes Project Consortium, 2015).

Control de calidad datos genotípicos pre-imputación

El número original de SNPs autosómicos de cada set de datos corresponde a 687.780 para los 7 individuos de Camarones genotipificados con AXIOM, mientras que para Sur de Chile corresponde a 598.826 SNPs dado que los 39 individuos incluidos provienen de dos Array de AXIOM de Affimetrix, y al hacer un solo set de datos para esta población se perdieron cerca de 112.242 SNPs. Por otro lado, el set de datos genotípicos para Puno corresponde al número de SNPs autosómicos (787.742 SNPs). Además, a cada set de datos se le aplico un filtro de tasa de llamada de variantes por SNP sobre el 0.1 (--geno) y con un mínimo de genotipos perdidos por individuo menor a 0.05 (--mind) en Plink1.9 (Chang et al. 2015).

Tabla 2. Poblaciones, Muestras y Base de Datos Genotípicos

Población	*N* con Ancestría Amerindia (>90%)	Array	Nº de SNPs	*N* Final	Proyecto/Acceso
Camarones	7/15 19/24	Axiom MEGA	687.780 1.265.900	26	ChileGenomico Fondecyt 1140544
Puno	30/106	Axiom	787.742	30	Dr. Andrés Moreno
Sur de Chile	39/57	Axiom	598.826	39	PatagoniaDNA CONICYT DRI USA

Determinación de fase gamética

Este proceso corresponde a la inferencia de las combinaciones alélicas (i.e. haplotipos) existentes a largo de los cromosomas a partir de los datos de genotipificados (SNPs). Conocer la fase de los datos genotípicos es un paso necesario para resolver ambigüedades tales como las diferentes combinaciones de haplotipos entre heterocigotos, lo cual es necesario determinar previamente a la aplicación de los análisis de selección. El programa SHAPEIT2 fue utilizado para este paso metodológico ya que ha mostrado ser altamente preciso cuando se usa numerosos paneles de referencia como 1KG (O'Connell et al., 2014). SHAPEIT2 fue utilizado incorporando un mapa genético del ensamble GRCh37 de NCBI, y un archivo con la información de todos de los individuos a analizar, así como el panel de 26 poblaciones de referencia de 1kG fase III, utilizando el formato de archivos binarios en el programa PLINK v1.9.

Imputación

La imputación es un paso metodológico que tiene por objetivo inferir genotipos perdidos en un set de individuos utilizando genomas de referencia con mayor densidad de datos, lo cual permite agregar aquellos *loci* (e.g. SNPs) perdidos al set de datos original. Aquellos loci son inferidos a partir de haplotipos que se comparten entre individuos y el panel de referencia con mayor densidad de SNPs, el cual puede ser proporcionado por International HapMap Project, 1000Genomes Project Fase III y por la secuenciación genómica de un subconjunto de los individuos del set original (Howie et al. 2012). El objetivo de este proceso es posibilitar el uso simultáneo de paneles de datos genotípicos diferentes (AXIOM y MEGA) en este estudio, aumentando la cobertura y el número de marcadores comunes entre cada muestra.

El método para realizar la imputación fue el algoritmo de IMPUT2 v2.3.2 (Howie et al., 2009; et al., 2012). El panel de referencia fue construido con el set

de datos poblacionales disponibles en 1KG fase 3[2], más 24 individuos de etnia Aymara con secuencias de genomas completos (4.5x y un individuo a 30x) en Illumina HiSeq 4000, PE125 de la University of Chicago Genomics Facility, y a los cuales se tuvo acceso gracias a la colaboración y pasantía de investigación en el laboratorio de la Dra. Anna Di Rienzo del Departamento de Genética Humana de la Universidad de Chicago. La imputación fue realizada independientemente para cada población de estudio por separado considerando que poseen diferentes Array de genotipificación. Los parámetros utilizados en la imputación corresponden al set de valores entregados por defecto en el programa. Luego, se hizo un control de calidad sobre los genotipos obtenidos, donde se seleccionaron aquellas muestras que poseían genotipos imputados con probabilidad posterior ≥ 0.9. Los genotipos fueron asumidos perdidos si ninguno de los tres posibles genotipos alcanzó dicho umbral (0.9). Además, dado el amplio número de *loci* alcanzados, también se hicieron controles de calidad adicionales, removiendo aquellos SNPs con tasas de valores perdidos mayores que 0.05 o p-valor menor que 10^{-6} en el equilibrio de HWE. El número de marcadores SNPs imputados que se obtuvo para cada población se muestra en la Tabla 3 de la sección de resultados.

II. Estructura genética poblacional

Análisis de componentes principales (PCA)

El método de componentes principales es frecuentemente usado en genética como un método para agrupar individuos e identificar a aquellos que pertenecen a grupos similares o conforman subpoblaciones, al igual que el análisis de mezcla (e.g. Admixture). Un número de componentes principales (PCs o eigenvectores) son estimados a partir de una matriz de datos genotípicos por individuo, donde cada PC resume alguna característica o

[2] Descargado desde: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html [Febrero, 2017].

variable de los datos que permite la discriminación entre ellos (Nielsen & Slatkin, 2013). A su vez, el PCA es un análisis multivariado que genera una representación gráfica con dos o tres dimensiones de la distancia entre poblaciones usando las frecuencias alélicas, más que distancias genéticas (Novembre et al., 2008). Los PCs son extraídos secuencialmente, siendo cada PC independiente y reflejando la mayor varianza posible de los datos. Usando el PCA es posible observar espacialmente si las muestras se agrupan como una población, así como sus diferencias con otras poblaciones a nivel génico (Novembre & Stephens, 2008).

Índice de diferenciación (F_{ST} genómico)

Se calcularon los valores de F_{ST} (Weir & Cockherman, 1984) a nivel genómico entre cada par de población con el programa Vcftools (Danecek et al., 2011) a partir de archivos en formato VCF ("variant call format"), que contienen la información de los genotipos por individuo. A partir de la matriz que posteriormente contiene las estimaciones de F_{ST} por SNPs, se procedió a hacer el calculo de PBS en la sección de métodos de inferencia de selección positiva.

Inferencia de ancestría global

Con el objetivo de conocer la proporción de ancestrías genéticas que comparten las muestras de cada población, utilizamos el programa ADMIXTURE (Alexander & Lange, 2011). Este software realiza una estimación de máxima verosimilitud de las ancestrías individuales a partir de una base de datos con multilocus y SNPs genotipados. El modelo de estimación asume un número hipotético de K poblaciones ancestrales que contribuyeron alelos a la muestra genotipificada. Por lo que es importante considerar que dichos componente genéticos ancestrales se encuentran incorporados en el panel evaluado. Como poblaciones de referencia se incorporaron datos del Proyecto 1000 Genomas con el propósito de tener ancestrías de referencia continentales de África, Europa y Asia (The 1000 Genomes Project Consortium, 2015). Dado

que ADMIXTURE asume independencia de individuos y de los SNPs, para este análisis se eliminaron los marcadores que presentaban una alta correlación por LD, medido por un r^2 mayor o igual a 0,1 en ventanas genómicas de 50 SNPs, dejando sólo un marcador por bloque de ligamiento. En este trabajo se realizaron las inferencias de ancestría global usando valores de K de 4 (i.e. ancestrías continentales) a 7 (i.e. ancestría única por población). Posteriormente, se calculó el error de validación (CV) cruzada para conocer qué valor de K tiene mayor capacidad predictiva sobre los datos observados.

Ancestría local

De forma de describir la variación de ancestría observada a nivel local en diferentes regiones genómicas (i.e. cromosomas) en las poblaciones analizadas, se utilizó el programa RFMix v1.5.4 para estimar la ancestría local (Maples, Grovel, Kenny and Bustamente, 2013). Especialmente, RFMix ha mostrado tener un alto poder de discriminación de ancestrías genéticas en poblaciones latino americanas mestizas. El análisis de ancestría local resulta informativo para conocer la proporción de ancestría amerindia en cada cromosoma y especialmente en aquellas regiones genómicas que podrían estar bajo presión selectiva como el arsénico (e.g. Cromosoma 10q21) o en otras regiones que podrían mostrar señales de selección a otras presiones del entorno. Para llevar a cabo este análisis, se requirió realizar la determinación de fase gamética de los cromosomas mediante SHAPEIT2. RFMix está basado en un análisis estadístico de *Random Forest.* El que permite hacer predicciones de ancestría sobre las diferentes ventanas definidas como tracts, a través de la generación de múltiples arboles de decisión, tomando el valor promedio mayor estimado para cada tract, reconstruyendo los dos pares de cromosomas con sus respectivas ancestrías. El panel de referencia para este análisis corresponde a Puno como población ancestral del componente andino

amerindio, además de las poblaciones CEU e YRI de 1kG para las ancestrías europea y africana respectivamente[3].

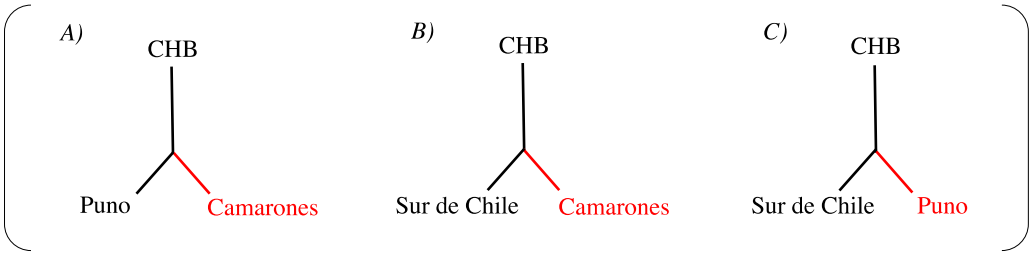
III. Métodos de inferencia de Selección Positiva

Population Branch Statistic (PBS)

El análisis de PBS permite identificar y comparar a nivel genómico los sitios (SNPs) que causan una mayor diferenciación entre tres poblaciones debido a cambios extremos producidos en las frecuencias alélicas (Yi et al. 2010). Bajo este modelo se consideran tres poblaciones filogenéticamente relacionadas: la primera es la población objetivo bajo selección; la segunda, es una filogenéticamente cercana, pero expuesta a presiones selectivas distintas; y la tercera, corresponde a la población filogenéticamente ancestral, que cumple la función de outgroup. En este análisis se consideró a los Han de China (CHB) disponibles en la base de datos de 1KG (N=49) como población filogenéticamente ancestral.

Para el cálculo de PBS se requiere como input la comparación pareada entre poblaciones, lo cual fue realizado mediante el cálculo del F_{ST} genómico de Weir & Cockerheim (1984) en VCFtools, considerando algunos filtros recomendados para el análisis, como la exclusión de SNPs con valores de alelos de menor frecuencia (MAF) menores a 0.01 (Danecek et al. 2011). A partir de los F_{ST} pareados, fue posible calcular los valores de PBS como se describe en Yi et al. (2011) mostrado en la Figura 6, donde además se indican los tres modelos de comparación utilizados en esta tesis para PBS.

[3] Utilizamos estos dos grupos (YRI y CEU), dado que son las poblaciones ancestrales que mayor componente genético han aportado a las poblaciones mestizas de América Latina, además del componente nativoamericano, según los recientes estudios de ancestría genética en poblaciones chilenas (Eyheramendy et al. 2015; Fuentes et al. 2014).



$$A)\ PBS_{Camarones} = \frac{T_{Camarones,Sur\ de\ Chile} + T_{Camarones,CHB} - T_{Sur\ de\ Chile,CHB}}{2}$$

$$B)\ PBS_{Camarones} = \frac{T_{Camarones,Sur\ de\ Chile} + T_{Camarones,CHB} - T_{Sur\ de\ Chile,CHB}}{2}$$

$$C)\ PBS_{Puno} = \frac{T_{Camarones,Sur\ de\ Chile} + T_{Camarones,CHB} - T_{Sur\ de\ Chile,CHB}}{2}$$

Figura 6. Escenarios de comparación utilizados para el cálculo de PBS (a, b y c). Donde T corresponde a la distancia genética estimada a partir de la clásica conversión de Fst utilizada en Cavalli-Sforza (1969): T=-log(1-F_{ST}).

Integrated Haplotype Score (iHS)

Este método es una medida estandarizada de la integral calculada sobre el área bajo la curva de cada haplotipo con homocigosidad extendida (EHH) perteneciente a las dos variantes de un SNP (i.e. alelo derivado y al ancestral) en una posición del genoma (i.e. SNP). Se calculó el puntaje de iHS por SNP y por ventanas de 200 kb, como recomiendan los autores de este método, para las poblaciones de Camarones, Puno y Sur de Chile de forma separada (Voight et al., 2006; Pickrell et al., 2009), utilizando el programa Selscan v3.2 (Szpiech & Hernandez, 2014). Dado que el método de iHS no confiere una prueba estadística, pues no se conoce la distribución de los datos bajo condiciones neutrales, se consideraron las recomendaciones de los autores descritas en Pickrell et al. (2009), respecto a considerar valores de iHS absoluto ≥ 2.5 y que

estén dentro del 1% de la distribución de valores a nivel genómico. Por otro lado, la definición del tamaño de las ventanas en 200 kb se debe a que dicho tamaño se ha estimado como suficiente para contener barridos selectivos descritos entre ~0.3-0.5 cM (Sabeti et al., 2006).

El cálculo de iHS en Selscan v3.2, fue realizado utilizando los archivos en formato PLINK (genotipos transpuestos) y las posiciones genéticas, extraídas desde un mapa genético usado en el ajuste de fase por cada cromosoma. Los valores de iHS fueron normalizados con la función –norm en Selscan (Szpiech & Hernandez, 2014), separando los valores obtenidos del análisis en 100 grupos de acuerdo con sus frecuencias alélicas, usando la desviación estándar y el promedio por grupo.

Cross-population Extended Homozygosity Haplotype (XP-EHH)

Este método es un análogo de iHS, el cual busca detector diferencias entre EHH para un mismo alelo entre dos poblaciones. Al igual que iHS, el objetivo es calcular un valor estadístico que describa las diferencias de EHH entre poblaciones por SNP. A diferencia de iHS, los autores de XP-EHH recomiendan utilizar un enfoque de análisis sobre marcadores para luego identificar aquellos que tengan un valor extremo, el cual posiblemente se encuentre bajo selección (Sabeti et al., 2007). Los valores de XP-EHH son calculados mediante el programa Selscan v3.2 (Szpiech & Hernandez, 2014). Las comparaciones utilizadas en este estudio corresponden a dos escenarios: (A). cuando se compara Camarones versus Puno, y (B) al comparar Camarones con Sur de Chile.

En la Tabla 3 se presenta un resumen de los métodos empleados en este trabajo, así como los criterios utilizados para identificar los *loci* bajo selección. Además, en la Figura 7 se muestra un mapa de flujo donde se asocia cada paso metodológico con los objetivos específicos de esta tesis.

Tabla 3. Descripción y criterios de los métodos genómicos para identificar selección positiva

Enfoque	Descripción	Método	Criterio de Selección	Límite temporal*	Referencias
Métodos basados en desequilibrio de ligamiento (LD)	Identifican barridos selectivos a nivel genómico. La asociación entre esas variantes define un haplotipo, que persiste en la población hasta que la recombinación desintegra dicha asociación.	iHS	1% de los valores más altos por SNPs y ventanas 200kb	30.000 años	Voight et al. (2006)
		XP-EHH	1% de los valores más altos por SNPs y ventanas 200kb		Sabeti et al. (2007)
Métodos basados en frecuencias alélicas	Identifican aquellos alelos con frecuencias sobre el promedio genómico, y que resultan ser diferentes entre poblaciones	PBS	1% de los valores más altos por SNPs y ventanas de 200kb**	80.000 años	Yi et al. (2010)

*Tiempo estimado de selección a partir de simulaciones en los respectivos estudios de cada método. **El método de PBS está diseñado originalmente por marcador, por lo que en este estudio se identificara la distribución de los marcadores de PBS bajo selección por ventanas genómicas de 200kb.

IV. Identificación y comparación de regiones bajo selección

Regiones y genes bajo selección

Los valores de PBS, iHS y XP-EHH que se encuentren en el 1% de su distribución se considerarán como un locus o región del genoma que está bajo selección positiva. Este criterio es definido como un umbral de discriminación, el cual diferentes estudios en selección positiva en humanos y los autores de los métodos empleados en esta tesis recomiendan. El uso del 1% es un valor de corte conservador en comparación con el utilizado en otros estudios (Bigham et al., 2009; Bigham et al., 2010). En el caso de los cálculos hechos por SNPs se seleccionaron los locus que estén en el 1% de la distribución de valores en sus respectivos métodos. Mientras que para las ventanas de 200kb estimadas, sólo

se considerarán aquellas que contengan un número ≥ 20 SNPs. Además, en el caso de iHS se ha enfatizado que se tiene mayor poder de detección de regiones bajo seleccionadas si se consideran aquellas que agrupen varios marcadores con valores altos de iHS, y no valores altos aislados por SNP individual (Voight et al., 2006). Para el test XPEHH, en cambio, se sugiere que es más conveniente utilizar aquellas regiones donde se encuentre un valor alto positivo, sin considerar los puntajes de los marcadores cercanos (Sabeti et al., 2007; Voight et al., 2006).

Comparación de señales de selección

Como una forma de conocer si los *loci* o regiones genómicas identificadas bajo selección por los análisis en iHS y PBS corresponden a señales locales o compartidas entre las poblaciones, se procedió a comparar dichos valores por *loci* comunes mediante las funciones de intersecto, unión y diferencia disponibles en el programa estadístico R. Particularmente, la función –*intersect* fue utilizada para comparar los *loci* con valores en el 1% de la distribución entre cada población tanto en PBS e iHS. Esta evaluación se graficó en R mediante un scatter plot donde cada eje corresponde a los valores de PBS o iHS por población.

Identificación de genes

A partir de la identificación de los loci y ventanas en el 1% de cada análisis se procedió a identificar sus respectivos genes asociados. Este paso se hizo mediante la utilización del programa de Annovar [4] (Wang, Li and Hakonarson, 2010). Dicho programa utiliza como input las posiciones físicas (bp) de cada *loci* o de inicio y final en el caso de las ventanas. Annovar hace una búsqueda por posición en la base de datos de “Known Canonical Genes” (21.680 genes a nivel genómico) publicada en el Genome Browser de UCSC

[4] Descargado de http://annovar.openbioinformatics.org/en/latest/ [Abril, 2017]

(Hsu et al., 2006)[5], obteniéndose como output los genes asociados, cercanos o contenidos en cada *loci* y ventana. Además, a través del análisis de intersectos en R, se identificaron los genes que son comunes entre las poblaciones para los análisis de PBS e iHS. Dicho análisis fue graficado mediante diagramas de Venn en R.

Enriquecimiento de procesos biológicos

El análisis de enriquecimiento corresponde a un método utilizado para identificar "clases de genes" que están sobrerrepresentados en un set de datos de genes mayor (e.g. Array), y podrían tener una asociación con fenotipos complejos, enfermedades u otros procesos biológicos. En este estudio, se llevó a cabo este análisis a través del software online GO term Enrichment Analysis (GOrilla). GOrilla utiliza las categorías de procesos biológicos definida para genes conocida como Gene Ontology (GO)[6]. GOrilla utiliza para el análisis de enriquecimiento una lista de genes de interés (e.g. top 1% de los genes en PBS) y una lista de todos los genes en el *background* (e.g. genes presentes en todo el Array imputado). El método identifica entonces, independientemente, para cada categoría de GO al que los genes en la lista objetivo se encuentran asociados, definiendo un valor de significancia para la sobrerepresentación de cada una de las categorías respecto al background. El valor de significancia es preciso y corregido por múltiples evaluaciones simuladas para cada categoría GO conocido como valor de p-FDR, que utiliza el método de corrección de Benjamini & Hochberg (1995) (Eden, Navon, Steinfeld, Lipson and Yakhini, 2009).

Fenotipos asociados

Los *loci* y ventanas que se encuentren en el 1% de cada análisis, y que hayan mostrado ser señales locales para cada población fueron utilizadas para

[5] https://genome.ucsc.edu/
[6] http://www.geneontology.org

identificar fenotipos previamente asociados en otros estudios de Genome Wide Association study (GWAs) a través del programa PheGenI disponible en NCBI (Phenotype-Genotype Integrator)[7]. Método que requiere ingresar una lista compuesta por los marcadores, ya sea *loci* o ventanas específicas, para luego conocer los fenotipos asociados a variantes determinados en otros estudios con sus respectivos valores de significancia.

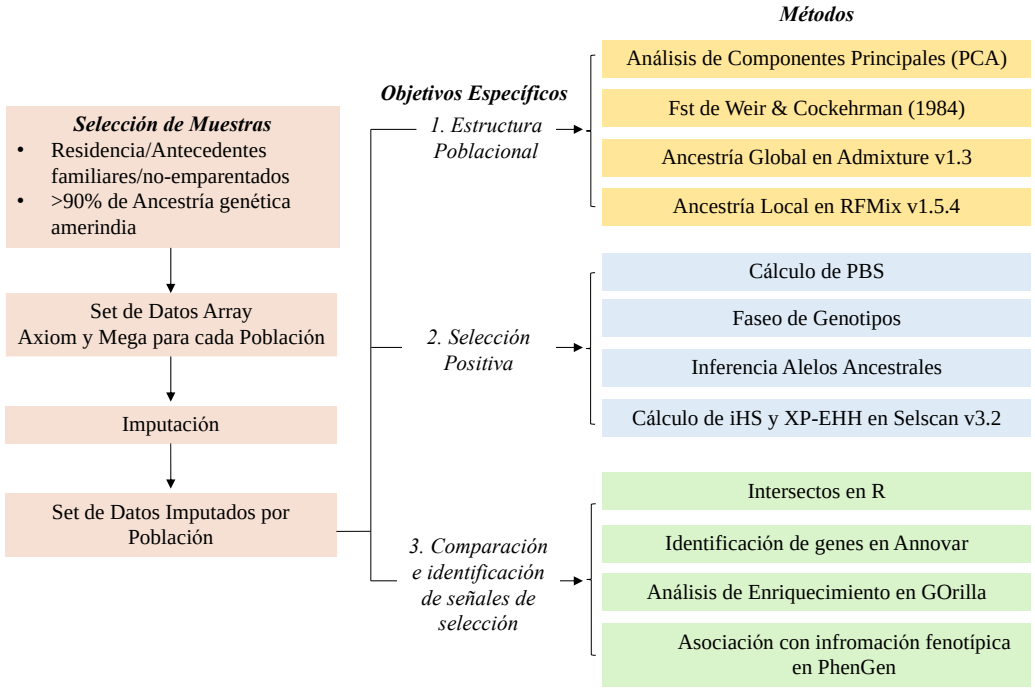


Figura 7. Mapa de Flujo que muestra los principales pasos metodológicos por cada objetivo específico, a partir de los datos imputados.

[7] https://www.ncbi.nlm.nih.gov/gap/phegeni

RESULTADOS

I. Imputación, Filtros y Marcadores

Es preciso señalar que de las bases de datos originales fueron removidos los marcadores uniparentales mitocondriales y cromosomas sexuales, dejando sólo marcadores autosómicos. Cada set de datos original por población (AXIOM y MEGA) fue imputado de manera independiente en Imput2. Durante el proceso de imputación en SHAPEIT2 se determinó la fase gamética de cada set de datos por población y luego se procedió con la imputación. El número de marcadores obtenidos para cada población fue ~33 Millones de SNPs. Como filtro se eliminaron los genotipos que tuvieron un valor de probabilidad posterior ≥ 0.9. Los genotipos fueron asumidos perdidos si ninguno de los tres posibles genotipos alcanzó dicho umbral de 0.9. Además, dado el amplio número de *loci* alcanzados, también se hicieron controles de calidad adicionales, removiendo aquellos SNPs con variantes perdidas mayores a 0.05 o p-valor menor que 10^{-6} en el equilibrio de HWE con funciones de VCFtools v0.1.13. Finalmente, sólo se conservaron los sitios con genotipos bialélicos dado que los posteriores análisis funcionan en base a dicho criterio. El número de marcadores SNPs imputados que se obtuvo para cada población se muestra en la Tabla 4.

Tabla 4. Resultados de la imputación

Población	N	Datos Arrays	Datos Imputados
Camarones	26	1.265.900 687.780	3.458.398*
Puno	30	787.742	4.592.409
Sur de Chile	39	598.826	4.567.254

*Este número corresponde a la combinación de SNPs entre ambos sets de datos imputados independientemente luego de aplicar los filtros de calidad post imputación. Camarones Axiom obtuvo un número de 4.585.367 SNPs, y el set de datos de MEGA, 4.596.768 SNPs.

II. Estructura poblacional

El análisis de PCA fue realizado considerando únicamente un set compuesto por las poblaciones amerindias correspondientes a Camarones, Puno y Sur de Chile (117.355 SNPs). En el caso de Sur de Chile se destacan las dos localidades desde donde provienen las muestras debido a estar compuesta por dos etnias (Huilliche de San Juan de la Costa y Pehuenche de Trapa Trapa). En la Figura 8, se puede observar que los componentes 1 y 2 son los que explican un mayor porcentaje de la varianza de los datos con un 4.41 y 2.54% respectivamente. Mientras que el componente 3 explica un 2.18% de la varianza de los datos. De estos resultados se puede destacar que, en la primera comparación, el componente 1 logra explicar las diferencias entre las poblaciones nortinas y las del sur de Chile, mientras que el componente 2 diferencia a algunos individuos de Camarones y también a las dos poblaciones del sur de Chile. Tales diferencias son mejor representadas en la comparación de los componentes 1 y 3, donde el PC3 podría estar diferenciando individuos costeros, como los provenientes de San Juan de la Costa (i.e. Huilliche), así como también algunos individuos de Camarones.

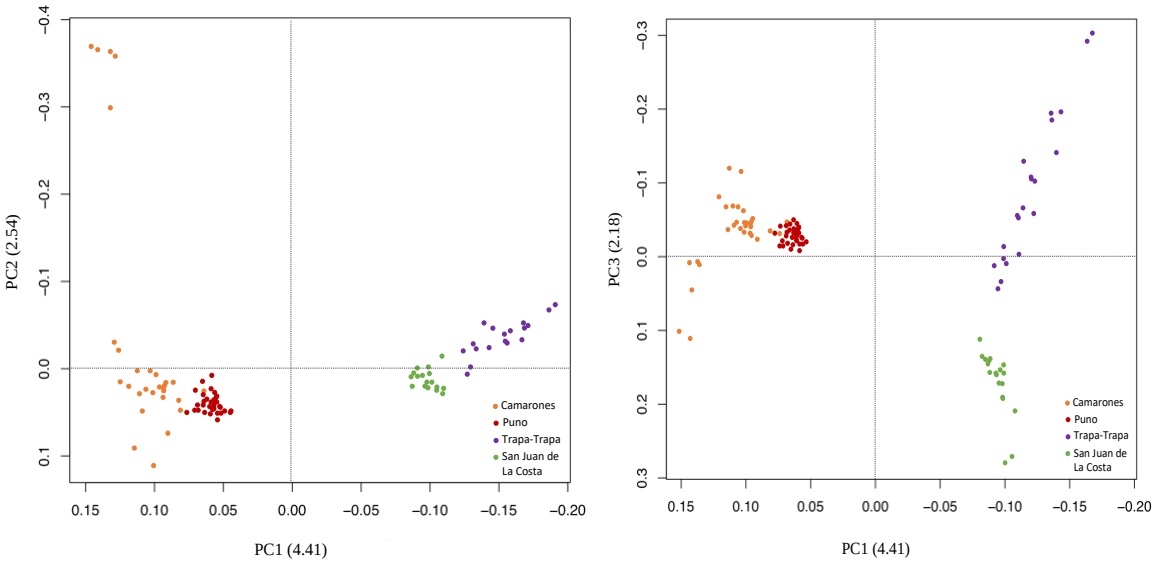


Figura 8. PCA con poblaciones amerindias

Ancestría Global y Local

Para la estimación de ancestría global se utilizó un set de datos compuesto por 6 poblaciones, incluyendo las de 1KG utilizadas como referencia continental (173.814 SNPs). El programa ADMIXTURE indica que al menos 10.000 marcadores permiten inferir ancestría a nivel continental (Alexander & Lange, 2011).

Respecto a los resultados del análisis de ancestría global, el modelo con un k=5 poblaciones fue el que presento el menor error de validación cruzada (0.60112). En dicho modelo, además de las ancestrías continentales de África, Europa y Este asiático, se observó que las poblaciones de Puno y Camarones difieren mayoritariamente del Sur de Chile a nivel de sus ancestrías, siendo estas una “amerindia andina” y la otra “amerindia del sur” (Figura 9, modelo k=5). El segundo modelo con menor error de validación cruzada fue el de k=4, donde se consideran cuatro poblaciones ancestrales, es decir, los cuatro continentes. En ese modelo, se observó que las poblaciones de Camarones, Puno y Sur de Chile poseen un alto porcentaje de ancestría amerindia (Tabla 5).

Con el objetivo de confirmar que las poblaciones estudiadas poseen un alto componente ancestral amerindio, factor clave en los análisis de selección positiva, se procedió a inferir la ancestría local del componente amerindio en Camarones mediante RFMIx. Los resultados de la ancestría local para Camarones muestran un 91% de ancestría amerindia a nivel poblacional (obtenida a partir de la suma de ancestría amerindia por cada cromosoma).

Tabla 5. Frecuencias relativas de las ancestrías global y local inferidas en las poblaciones del estudio.

Población	Admixture (k=4) AFR/EUR/EAS/AMR	Admixture (k=5) AFR/EUR/EAS/AMRA*/AMRS*	RFMix (k=3) ** AFR/EUR/AMRA
Camarones N=26	0.03/0.4/0.01/0.92	0.03/0.4/0.01/0.81/0.11	0.03/0.6/0.91
Puno N=30	0.03/0.3/0.00/0.94	0.03/0.3/0.00/0.92	-
Sur de Chile N=39	0.02/0.3/0.00/0.95	0.02/0.3/0.00/0.92	-

*Abreviaciones: ancestría amerindia (AMR), amerindia andina (AMRA) y amerindia sur (AMRS).

**La población ancestral AMRA utilizada como referente fueron los 30 individuos de Puno.

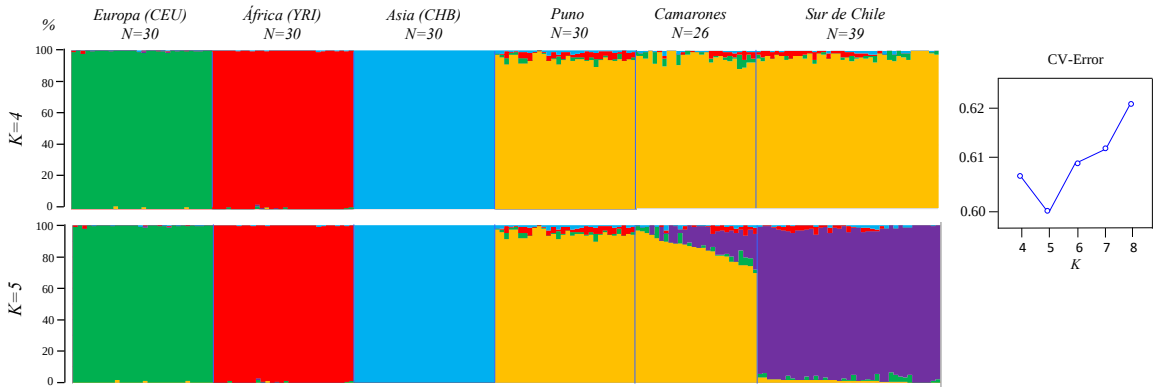
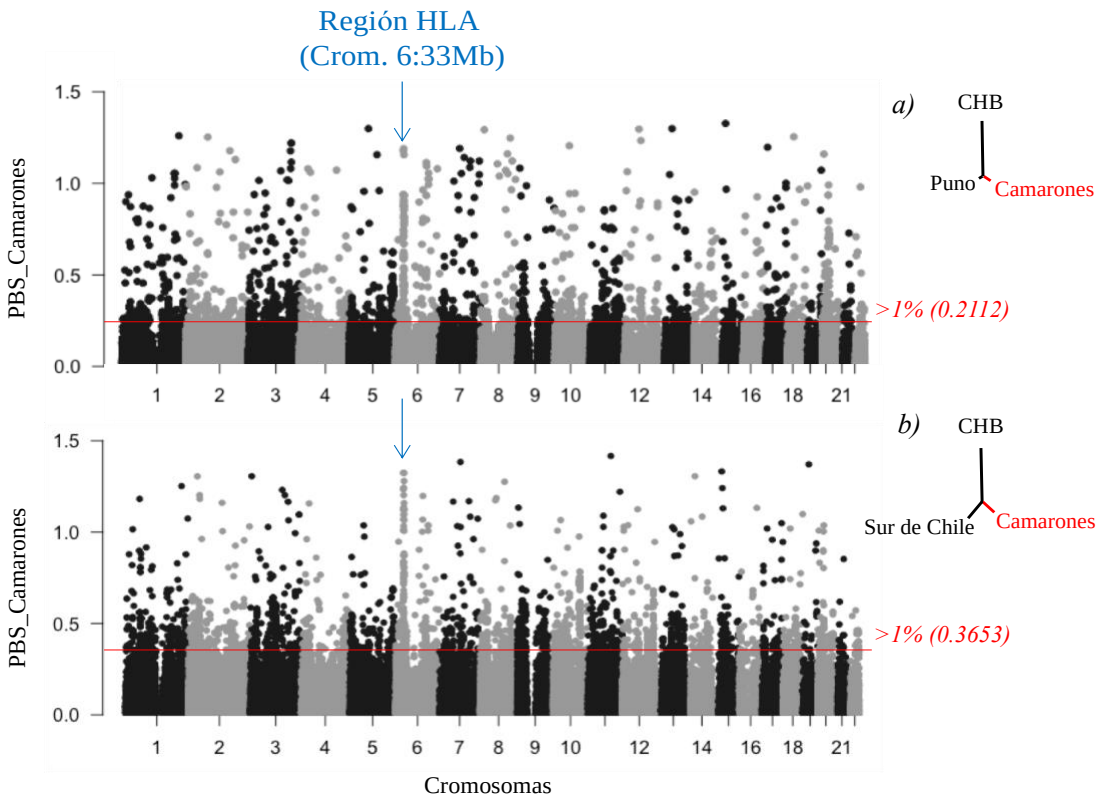


Figura 9. Inferencias de ancestría global en Admixture. Utilizando las poblaciones de referencia de 1KG de África (YRI), Europa(CEU), y Asia(CHB). Cada grupo está conformado por barras que representan a individuos y cuya proporción de ancestría corresponde a cada color.

III. Identificación de selección positiva

PBS

A partir de los tres escenarios calculados para PBS, mostrados en la Figura 10, se pudo identificar en el primer escenario de Camarones un total de 5.228 SNPs en el 1% de PBS, mientras que en el segundo escenario fueron 4.918 SNPs (Tabla 6). Los valores de PBS en el 1% se encuentran sobre el 0.211 y 0.365 en cada escenario graficado en los manhattan plot de la Figura 10. Los largos de bandas para cada población muestran que la mayor distancia genética de Camarones se da respecto a la población del sur de Chile (Fig. 10b), mientras que ésta es mucho menor respecto a Puno (Fig. 10a).



Figuera 10. Manhattan plot de valores de PBS calculados para los dos escenarios de Camarones como el *target* de selección. En la figura se indican los valores de corte en cada análisis (>1%). Además, en a y b se

muestran los largos de banda (i.e. distancias genéticas) estimadas a partir de los promedios de PBS para cada población.

Tabla 6. Marcadores utilizados en PBS

Escenarios de PBS	SNPs	SNPs Top1%	Ventanas (200kb) en el Top1% (>20 SNPs)	# Genes en el Top1%
A (Cam,Puno,CHB)	522.896	5228	20	887
B (Cam,SCL,CHB)	491.814	4918	54	962
C (Puno,SCL,CHB)	1.007.545	10075	37	1209

En la tabla 7 se describen los 10 primeros marcadores con los valores más altos de PBS en Camarones para los dos escenarios estudiados. En ambos modelos de PBS se destacan los SNPs con valores altos de PBS en la región cercana a los 31Mb del cromosoma 6 conocida por contener el Complejo Mayor de Histocompatibilidad (MHC) clase I (HLA-B, -C). lo que también puede observarse como un peak en ambos manhattan plot de la Figura 10. Otro marcador común que aparece entre los valores más altos es el presente en el gen codificante de la Caspasa inhibidora de apoptósis AVEN. Respecto al primer escenario, de Camarones y Puno, se destacan los SNPs ubicados en genes involucrados con procesos de regulación de óxido nítrico y vasodilatación como el gen DLC1, o el gen CAB39L (Calmudilina) asociado con características de los eritrocitos en el torrente sanguíneo (tamaño y concentración). Mientras que, en el escenario de Camarones y Sur de Chile, se pueden destacar los SNPs pertenecientes a los genes PCLO, MIR4510 y C15orf41 asociados con desórdenes mentales, vitamina D y colesterol respectivamente.

Tabla 7. Top 10 de SNPs en el 1% de valores más altos de PBS para Camarones (Escenarios A y B)

Ranking	SNP ID (Crom:pb)	Genes	Valores de PBS A (CAM:PUNO:CHB)
1	rs187848612(15:34223949)	AVEN	1.35:0.00:0.04
2	rs11617824(13:49982127)	CAB39L	1.32:0.00:0.08
3	rs1153188(12:55098996)	DCD, MUCL1	1.32:0.00:0.07
4	rs58807948(8:13220794)	DLC1	1.32:0.00:0.07
5	rs62514564(8:113951274)	CSMD3	1.27:0.00:0.12
6	rs6481066(10:55961210)	PCDH15	1.23:0.00:0.18
7	rs74744272(17:4892176)	CAMTA2, INCA1	1.22:0.00:0.04
8	rs10224522(7:76897499)	CCDC146	1.21:0.00:0.04
9	rs55891410(6:31219142)	HLA-C	1.21:0.00:0.13
10	rs13021344(2:170520306)	CCDC173	1.20:0.02:0.00
Ranking	**SNP ID (Crom:pb)**	**Genes**	**Valores de PBS B (CAM:SCL:CHB)**
1	rs74541541(11:90380750)	DISC1FP1	1.42:0.00:0.07
2	rs2877(7:82764425)	PCLO	1.38:0.00:0.01
3	rs75711841(19:24004984)	RPSAP58	1.37:0.00:0.02
4	rs187848612(15:34223949)	AVEN	1.33:0.00:0.06
5	rs3094055(6:30332146)	TRIM39-RPP21, HLA-E	1.32:0.00:0.02
6	rs9265671(6:31300894)	HLA-C, HLA-B	1.32:0.00:0.02
7	rs260547(6:32907278)	HLA-DMB	1.32:0.00:0.02
8	rs1169032(14:36249574)	RALGAPA1P1	1.31:0.00:0.78
9	rs1995206(3:5343748)	MIR4790, GRM7-AS3	1.31:0.04:0.00
10	rs35800906(2:34969243)	MIR4510, C15orf41	1.31:0.04:0.00

* Los PBS corresponden a los tres largos de bandas o distancias de cada población respecto a las otras. Abreviaciones SCL=Sur de Chile, CAM=Camarones, CHB=Han de China.

Con el análisis de los intersectos mostrado en la Figura 11, se pudo identificar que marcadores son únicos en Camarones o bien cuales son compartidos con otras poblaciones. Al comparar los valores de PBS entre Camarones y Puno por marcador, se encontraron 1.090 SNPs comunes que se pueden observar en el cuadrante superior derecho del gráfico de la Figura 11.

En dicho cuadrante, se destaca un grupo de marcadores en un círculo azul que muestran los 23 SNPs que caen específicamente en el gen de la Ciclina mediadora de transporte de iones de metales de membrana con dominio divalente 2 (CNNM2) ubicada en el cromosoma 10, la cual está a sólo ~50kb del gen AS3MT. A su vez, en el cuadrante superior izquierdo del mismo gráfico, se destaca en otro círculo azul un peak de marcadores que alcanzan valores altos sólo en Camarones, y no en Puno, los cuales corresponden a SNPs cercanos a los genes codificantes de la Caspasa inhibidora de la activación de apoptosis AVEN, un gen no codificante DISC1FP1 (Crom11:90.380.750), y un gran número de SNPs (>30) cercanos a la región del HLA clase I (-C, -B) y II (HLA-DRB1, -DQA1 y DQB1) del cromosoma 6 (30-33Mb).

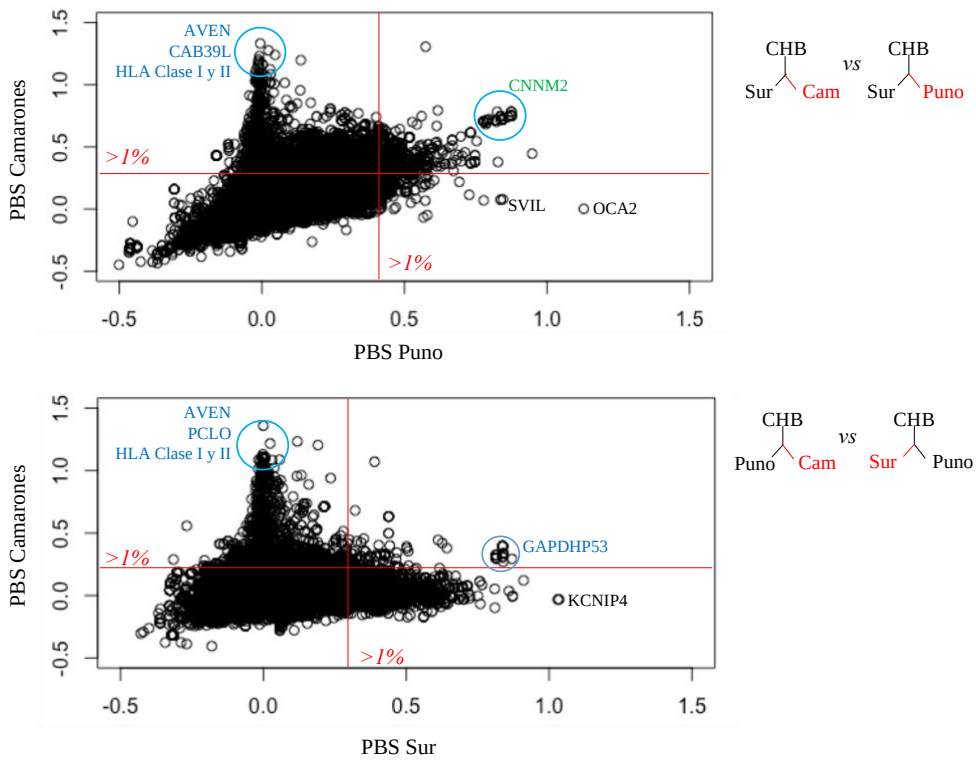


Figura 11. Scatter plot de los intersectos entre los valores de PBS para Camarones respecto a Puno y Sur de Chile. El límite del 1% se indica en una línea roja. A la derecha, se indican las poblaciones comparadas en rojo.

Por otro lado, al comparar los PBS de Camarones y Sur de Chile por marcador, se obtuvieron 155 SNPs comunes (Gráfico inferior de la Figura 11). A diferencia de la primera comparación, Camarones y Sur de Chile tiene una mayor diferenciación, observada en el menor número de marcadores bajo selección compartidos, donde sólo un grupo de 8 marcadores se agrupan en una parte del cuadrante superior derecho, los cuales pertenecen al cromosoma 20, en torno a 24Mb y cercanos a un pseudogen del gliceraldehído 3 fosfato deshidrogenasa (GAPDHP53). Otra diferencia respecto a la primera comparación es el pronunciado peak de marcadores que están en el cuadrante superior izquierdo correspondiente a marcadores bajo selección únicos en Camarones, mientras que, en el cuadrante inferior derecho, un gran número de SNPs con los valores más altos de PBS en Sur de Chile se encuentran en el gen codificante de la proteína de membrana voltaje dependiente de Potasio 4 o KCNIP4.

A partir del 1% de los marcadores calculados para los PBS de Camarones (escenario B) y Puno (escenario C), donde ambas poblaciones comparten los mismos outgroups de Han de China y Sur de Chile. Se procedió a la asignación de aquellos marcadores en ventanas genómicas de 200kb. A partir de dicha asignación, se identificaron aquellas regiones del genoma que acumulan un mayor número de marcadores en cada población. Además, como filtro adicional se seleccionaron aquellas ventanas que poseían más de 20 marcadores como una forma de despejar aquellas ventanas con valores altos aislados.

En la figura 12, se muestra la comparación entre las ventanas con el 1% de PBS para Camarones y Puno. Se presenta esta comparación, ya que es el escenario más parsimonioso, si se considera que ambas poblaciones comparten un mismo componente de ancestría genética y sólo difieren en los entornos que habitan. Al realizar un intersecto de ventanas entre ambas

poblaciones, se identificaron 18 ventanas compartidas (Figura 12b). Al conocer la distribución de valores del 1% de PBS en las ventanas de 200kb, se pudo identificar 54 ventanas en Camarones, de las cuales 18 son ventanas compartidas con la población de Puno y 36 son únicas en Camarones. En la Figura 12a, se muestra la regiones y genes asociados a nivel genómico que fueron determinadas como únicas en Camarones (en azul) o compartidas con Puno (en rojo).

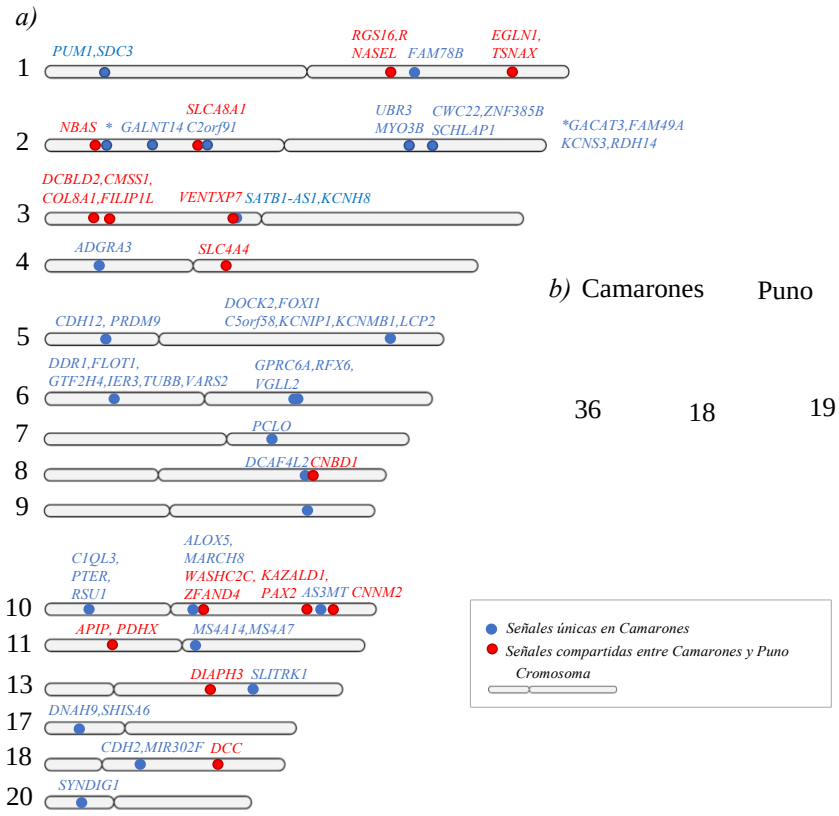


Figura 12. a) regiones bajo selección identificadas con el análisis de PBS en Camarones. En cada cromosoma, se destaca en rojo las regiones genómicas compartidas entre Camarones y Puno, y en azul, las que son únicas en Camarones. b) Diagrama de Venn que representa el intersecto de ventanas con el 1% de valores de PBS entre Camarones y Puno.

iHS

Los resultados de iHS calculados tanto por marcadores individuales (SNPs) y ventanas (200kb) para Camarones se muestran en la Figura 13. Se encontraron 10.650 SNPs en el 1% de valores más altos de iHS, donde los 10 marcadadores con mayor valor se muestran en la Tabla 9. Además, en la Figura 13b se muestra un manhattan plot con los valores de iHS promediados para cada ventana de 200kb, donde 121 ventanas se encontraron en el 1% de los valores promedios más altos (Tabla 8). Por otro lado, también se calculó el valor de iHS para las poblaciones de Puno y Sur de Chile de manera independiente, donde los marcadores en el 1% de iHS en Puno y Sur de Chile fueron 10.847 y 10.281 SNPs respectivamente. Mientras que el número de ventanas en el 1% de iHS, fueron 75 en Puno y 71 en Sur de Chile (Tabla 8).

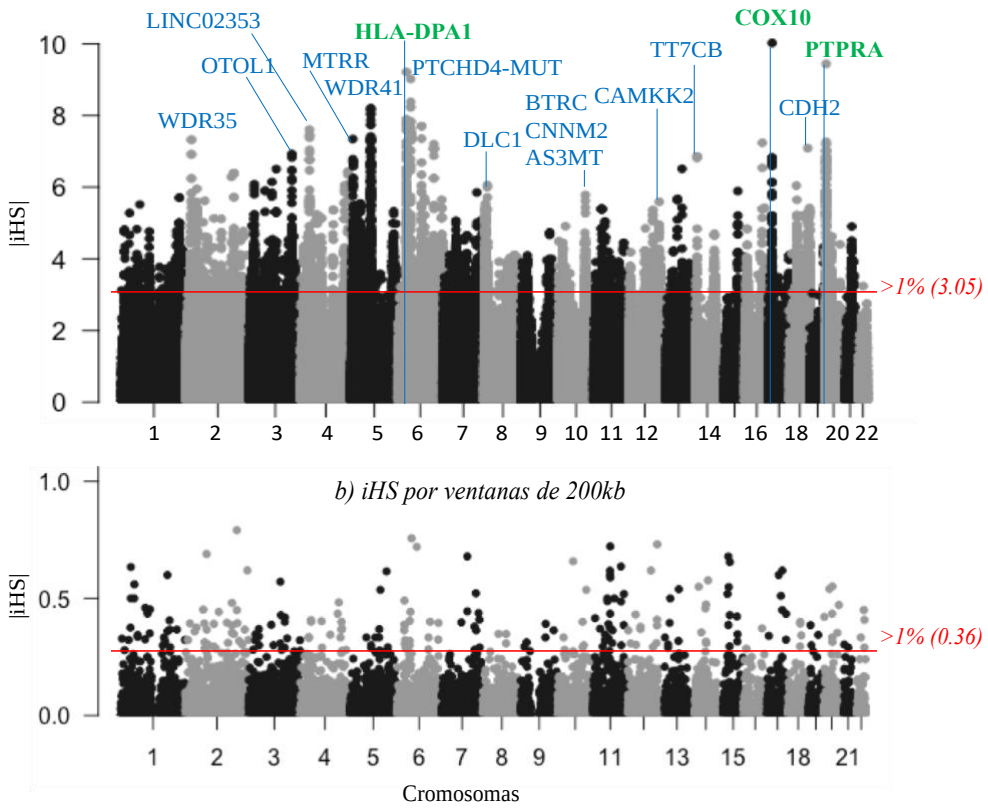


Figura 13. Manhattan plots con valores de iHS por SNP (a) y ventanas de 200kb (b) para la población de Camarones.

Tabla 8. Marcadores utilizados en iHS

Población	SNPs	SNPs Top1%	Ventanas 200kb (>20 SNPs)	Ventanas Top 1%	# Genes en el Top 1%
Camarones	1.065.061	10.650	12.121	121	716
Puno	1.084.711	10.847	7.475	75	681
Sur de Chile	1.028.191	10.282	7.187	71	697

Tabla 9. Top 10 de SNPs y ventanas (200kb) en el 1% de valores más altos de iHS en Camarones

Rank	**SNP ID (Crom:pb)**	**Genes**	**Valor iHS**
1	rs12941246(17:13975115)	COX10	10.02
2	rs605519(20:2893956)	PTPRA	9.44
3	rs10214910(6:33037675)	HLA-DPA1	9.21
4	rs12211733(6:48894648)	PTCHD4,MUT	9.01
5	rs10948453(6:48893852)	PTCHD4,MUT	8.37
6	rs7742488(6:48892734)	PTCHD4,MUT	8.23
7	rs7756563(6:48892960)	PTCHD4,MUT	8.23
8	rs1678775(5:76776893)	WDR41	8.20
9	rs437003(5:76769823)	WDR41	8.17
10	rs689657(5:76776692)	WDR41	8.16
Rank	**Ventana (200kb, Crom:bp)**	**Genes**	**Media iHS**
1	2:197200001-197400001	HECW2	0.79
2	6:56400001-56600001	DST	0.75
3	12:112600001-112800001	HECTD4	0.73
4	11:67000001-67200001	ANKRD13D, TBC1D10C	0.72
5	6:76200001-76400001	DLC1	0.72
6	15:43800001-44000001	DNAH6	0.68
7	10:59600001-59800001	CSMD3	0.67
8	6:33000001-33200001	HLA-DPA1, HLA-DPB1	0.67
9	11:108000001-108200001	PCDH15	0.65
10	1:39800001-40000001	CAMTA2, INCA1	0.65

A partir de los valores de iHS calculados por marcador (SNPs), se procedió a realizar el análisis de intersectos entre cada población como se muestra en la Figura 14a. En el caso de la comparación de Camarones con Puno, se encontraron 942 SNPs comunes, y al comparar con Sur de Chile se encontraron 422 SNPs (Figura 14b). Mientras que los marcadores comunes entre las tres poblaciones a nivel de marcadores fueron sólo 93 SNPs. Por otro lado, al comparar las ventanas de iHS entre Camarones y Puno, se pudo identificar sólo 4 ventanas compartidas, mientras que ninguna con Sur de Chile (Figura 14c).

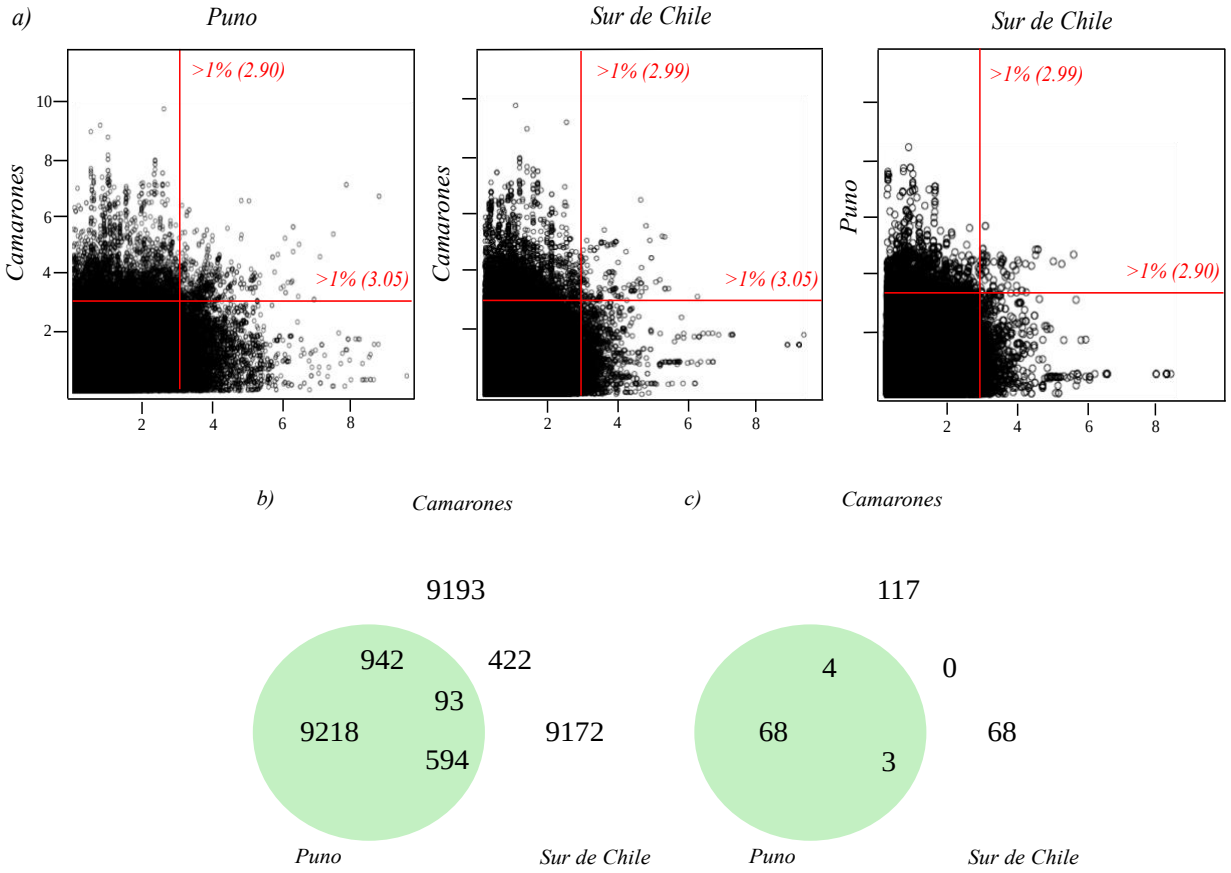


Figura 14. a) Intersectos del 1% de valores más altos de iHS. En rojo se indica el valor de corte >1% de iHS. b) y c) Diagramas de Venn que representan los intersectos entre Camarones, Puno y Sur de Chile, a nivel de marcadores SNPs y ventanas (200kb), respectivamente.

XP-EHH

Los valores de este método indican las diferencias entre los haplotipos extendidos entre los dos modelos de comparación utilizados (Tabla 10). En el gráfico superior de la Figura 15, se muestran los manhattan plot del análisis de XP-EHH por SNPs. A diferencia de los otros análisis, este método sólo indica valores altos entre mayor sea la diferenciación de haplotipos largos a nivel genómico (e.g. por SNPs y ventanas de 200kb), siendo los valores positivos correspondientes a la población objetivo bajo selección (Figura 15 para los marcadores y Tabla 11 para las ventanas).

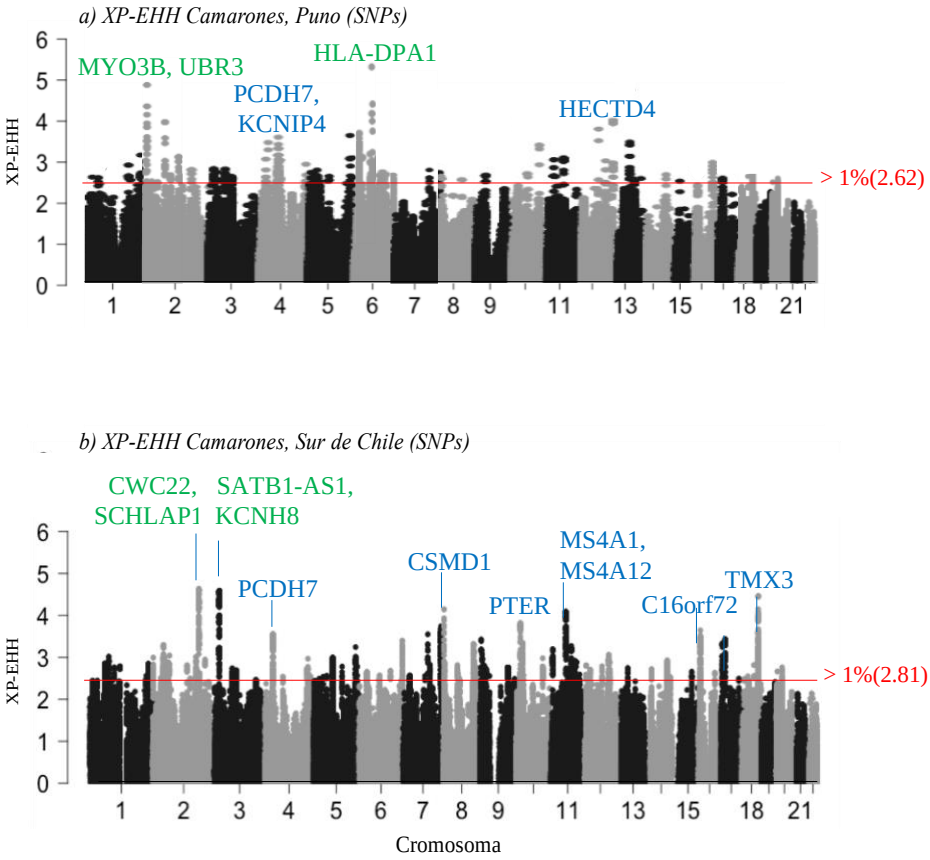


Figura 15. Manhattan plot de los valores de XP-EHH por marcadores SNPs entre Camarones y Sur de Chile y Camarones y Puno.

Tabla 10. Marcadores utilizados en XP-EHH

Escenarios de Comparación XP-EHH	SNPs	SNP Top1%	# Genes en el Top1%
Camarones-Puno	935.060	9.350	649
Camarones-SCL	1.169.775	11.697	706

Tabla 11. Top 10 ventanas en el 1% de XP-EHH (Escenarios A y B)

Rank	**Ventana (200kb, Crom:pb)**	**Genes**	**XP-EHH**
1	6:33000001-33200001	HLA-DPA1, HLA-DPB1	0.99
2	6:48800001-49000001	PTCHD4, MUT	0.98
3	9:108440911-108640910	PAX2	0.92
4	6:56400001-56600001	DST	0.75
5	10:99061114-99261113	BTRC	0.82
6	6:76200001-76400001	DLC1	0.78
7	5:10220584-10420583	WDR41	0.72
8	15:43800001-44000001	DNAH6	0.70
9	2:104672766-104872765	MYO3B, UBR3	0.69
10	1:34798511-34998510	PUM1, SCDC3	0.68
Rank	**Ventana (200kb, Crom:bp)**	**Genes**	**XP-EHH**
1	11:34798511-34998510	APIP,PDHX	0.98
2	8:88564985-88764984	CNBD1,DCAF4L2	0.90
3	9:108440911-108640910	TMEM38B	0.83
4	2:15412995-15612994	NBAS	0.80
5	3:99061114-99261113	DCBLD2,MIR548G	0.79
6	2:41212995-41412994	SLC8A1,LINC01913	0.78
7	5:10220584-10420583	CCT5,CMBL,FAM173B,MARCH6	0.78
8	1:231521291-231721290	EGLN1,TSNAX	0.77
9	10:104672766-104872765	CNNM2,NT5C2	0.65
10	11:34798511-34998510	APIP,PDHX	0.64

IV. Enriquecimiento de procesos biológicos

A partir de los genes identificados en Annovar para el 1% de los marcadores en Camarones por cada análisis, se procedió a determinar los procesos biológicos que se encuentran sobre-representados a nivel genómico. Considerando los genes en el 1% de los valores más altos de PBS por marcador (887 genes), se obtuvieron 10 procesos biológicos sobre-representados significativos (FDR<0.05) (Tabla 12). Mientras que en a partir de los genes obtenidos del análisis de iHS, se obtuvieron 4 procesos biológicos sobre-representados con significancia (FDR<0.05).

Tabla 12. Enriquecimiento de procesos biológicos en Camarones a partir de los genes identificados en PBS

	Proceso Biológico (Categorías GO)	C	O	E	R	P-Valor	FDR**
1.	Migración de células musculares	64	13	2.958	4.394	5.71E-06	0.004
2.	Señalización multicelular del organismo	170	21	7.857	2.672	3.70E-05	0.011
3.	Procesos del sistema muscular	342	33	15.807	2.087	5.01E-05	0.011
4.	Desarrollo del prosencéfalo	317	31	14.651	2.115	6.57E-05	0.011
5.	Desarrollo de dendritas	152	19	7.025	2.704	7.33E-05	0.011
6.	Desarrollo del sistema urogenital	289	28	13.357	2.096	0.000172053	0.018
7.	Desarrollo del tejido muscular	304	29	14.050	2.063	0.000173131	0.018
8.	Respuesta al estímulo mecánico	177	20	8.180	2.444	0.000195513	0.018
9.	Desarrollo del corazón	438	37	20.244	1.827	0.000283016	0.020
10.	Desarrollo de órgano sensorial	439	37	20.290	1.823	0.000295913	0.020

*GO=categorías de procesos biológicos del Gene Ontology. C=número de genes de referencia en la categoría GO. O=número de genes en el set de genes de referencia y categoría GO. E=el número esperado en la categoría. R=proporción de enriquecimiento. P-valor del enriquecimiento correspondiente a los genes sobre-representados en la categoría GO. **Valores significativos de p-valor con FDR<0.05 de Benjamini & Hochberg (1995).

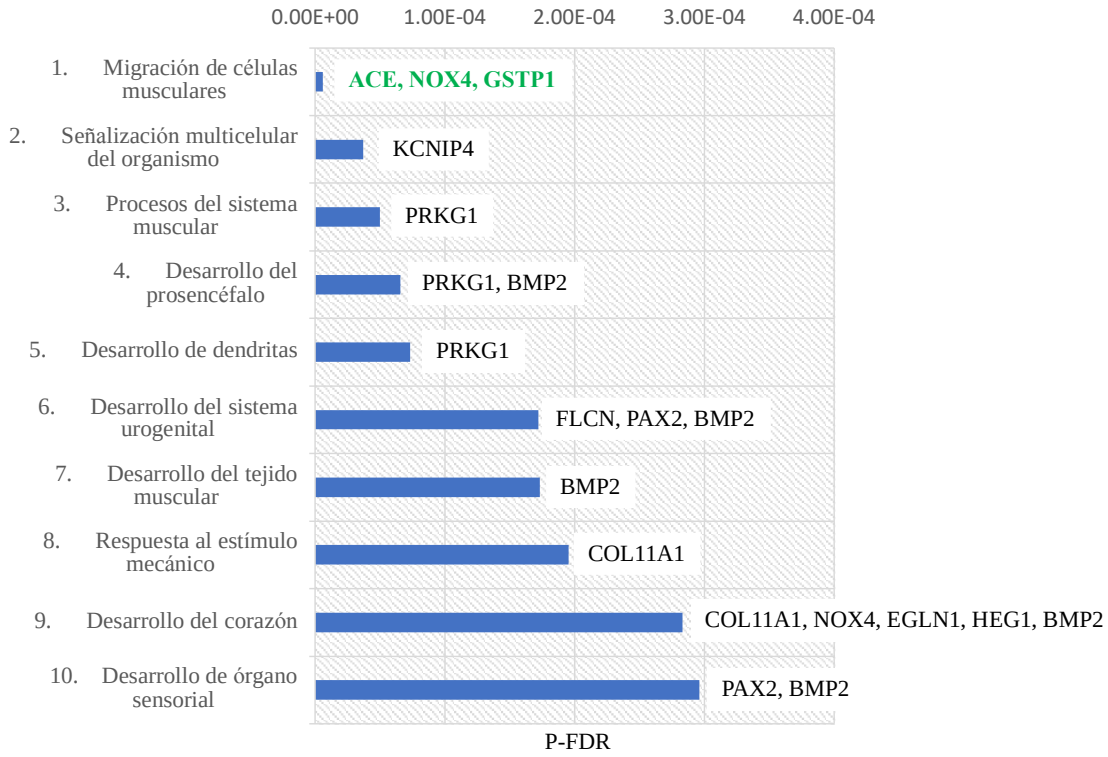


Figura 16. Procesos biológicos significativamente sobrerepresentados en el 1% de los genes bajo selección del análisis de PBS para Camarones. Se destacan los genes con función relevante en la categoría GO.

A partir de los valores en el 1% de PBS, el análisis de enriquecimiento muestra que a nivel genómico los procesos involucrados con la contracción muscular en zonas del torrente sanguíneo como son las categorías “migración de células musculares”, “proceso del sistema muscular”, “desarrollo del tejido muscular” y “desarrollo del corazón” que son procesos que han sido asociados previamente con respuestas e interacciones celulares en condiciones de estrés de hipoxia, regulación de oxígeno y respuesta térmica a condiciones de frío (Bigham, 2016: Bigham & Lee, 2014). Particularmente genes como el ACE (enzima convertidora de angiotensina I) tienen un rol importante en la saturación arterial (SaO2) en condiciones de hipoxia (Tomar et al., 2015) (Figura 16). Lo mismo suscede con NOX4 que es un sensor celular de Oxigeno

(Nisimoto et al., 2014). Además, entre los genes que participan de los procesos 8 y 9 (Figura 16), se encuentra COL11A1, que es el responsable de la síntesis de colágeno tipo XI, que cumple la función de regulación de la contracción muscular de cartílago. En este sentido, también aparece el gen codificante de la proteína responsable de la morfogénesis de hueso y cartílago (BMP2). Por otra parte, de forma particular, destacamos la presencia de Glutation S-transferasa (GSTP1) en la categoría 1 de migración de células musculares, pues GSTP1 es una reconocida en enzima que participa en los procesos de detoxificación en humanos. Mientras, que la categoría “desarrollo del sistema urogenital” podría estar involucrado con alguna interacción con el arsénico, dado que los efectos carcinogénicos asociados a arsénico son principalmente en el Sistema hepático y urogenital, así como su principal vía de detoxificación a través de la excreción del arsénico y otros tóxicos. Algunos genes asociados a tumores renales son FLCN (gen codificante de fuliculina) y PAX2 que es un supresor de genes tumorales WT1, y se a propuesto como un candidato de supresión de cáncer renal (Okumura et al., 2015).

Por otro lado, fueron 10 categorías de procesos biológicos identificados a partir de los genes en el 1% del análisis de iHS. Una categoría relevante dado el contexto de exposición en Camarones es la categoría sobre-representada de “Quimiotaxis negativa” (Tabla 13 y Figura 17). El proceso de quimiotaxis es definido como movimiento de células en respuesta a químicos, donde éstas pueden ser atraídas (quimiotaxis positiva) o ser repelidas por sustancias que muestran ciertas características químicas (quimiotaxis negativa). A su vez, la categoría de “vía señalización semaforina-plexina”, involucra genes en el control de la homeostasis y la morfogénesis de muchos tejidos, así como de la conectividad neuronal, cáncer, migración celular y respuestas inmunes a partir de la señalización intercelular de ligandos de semaforina y receptores de plexina. Respecto al “proceso metabolico flavonoide”, tiene relación con el

metabolismo de moléculas flavonoides que polifenoles natureles que provenien de la digestión de productos vegetales.

Tabla 13. Enriquecimiento de procesos biológicos en Camarones a partir de los genes identificados en iHS

	Proceso Biológico en Camarones	C	O	E	R	P-Valor	FDR
1.	Quimiotaxis negativa	33	9	1.161	7.749	1.38E-06	0.00105**
2.	Vía de señalización de semaforina-plexina	29	8	1.020	7.838	4.93E-06	0.00187**
3.	Proyección guía neuronal	191	19	6.721	2.826	4.35E-05	0.00888**
4.	Proceso metabólico flavonoide	20	6	0.703	8.524	4.67E-05	0.00888**
5.	Desarrollo axón	385	27	13.548	1.992	0.00051419	0.07836
6.	Desarrollo del procencéfalo	317	23	11.155	2.061	0.00083347	0.105851
7.	Transportadores de iones de Potasio	188	16	6.615	2.4184	0.00098584	0.10731
8.	Señalización celúla-célula via Wnt	401	26	14.111	1.842	0.00200943	0.19139
9.	Regulación de niveles hormonales	411	26	14.463	1.797	0.00281527	0.23835
10.	Transporte de amina	71	8	2.498	3.201	0.00336352	0.24737

*GO=categorías de procesos biológicos del Gene Ontology. C=número de genes de referencia en la categoría **GO.** O=número de genes en el set de genes de referencia y categoría GO. E=el número esperado en la categoría. R=proporción de enriquecimiento. P-valor del enriquecimiento correspondiente a los genes sobre-representados en la categoría GO. **Valores significativos de p-valor con FDR<0.05 de Benjamini & Hochberg (1995).

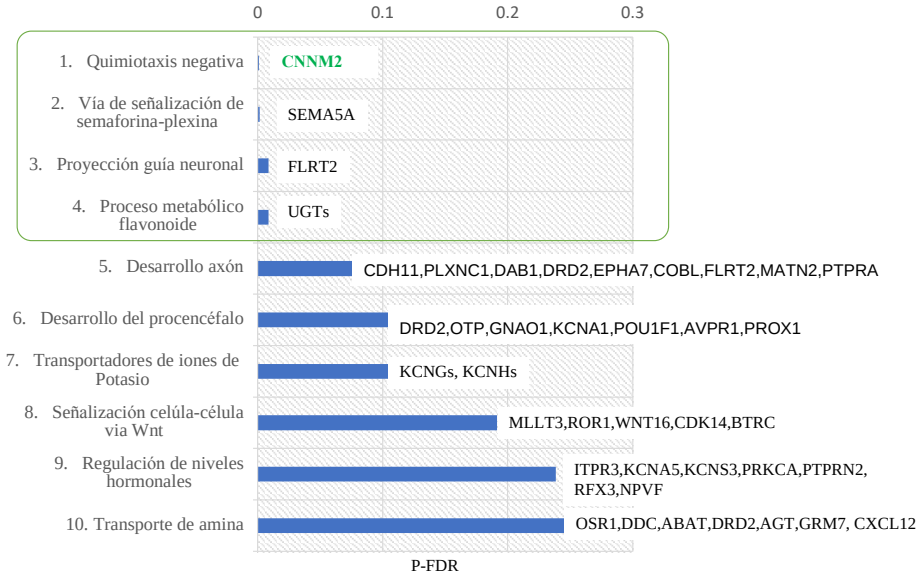


Figura 17. Procesos biológicos enriquecidos a partir de los genes bajo selección de iHS en Camarones (recuadro verde, P-FDR<0.05). Se destacan los genes con función relevante en cada categoría GO.

V. Genes candidatos a selección positiva en Camarones

Señales de selección en el sistema inmune HLA

La señal de selección más distintiva de Camarones corresponde a los genes ubicados entre los 33.02 y 33.04 Mb, en el cromosoma 6 donde se encuentra HLA -DPA1, -DPB1 y DPA2 (Figura 18 y 19). Sin embargo, sólo el gen HLA-DPA1 fue identificado con los tres métodos de selección sólo en Camarones (genes en el 1% de valores más altos de PBS, iHS y XP-EHH). El complejo de HLA ha sido identificado previamente bajo selección, siendo los genes de clase I y II mayoritariamente identificados. Además, el tipo de selección que afecta la región corresponde a selección balanceadora y positiva sobre genes de susceptibilidad y resistencia a enfermedades infecciosas (Meyer et al., 2018).

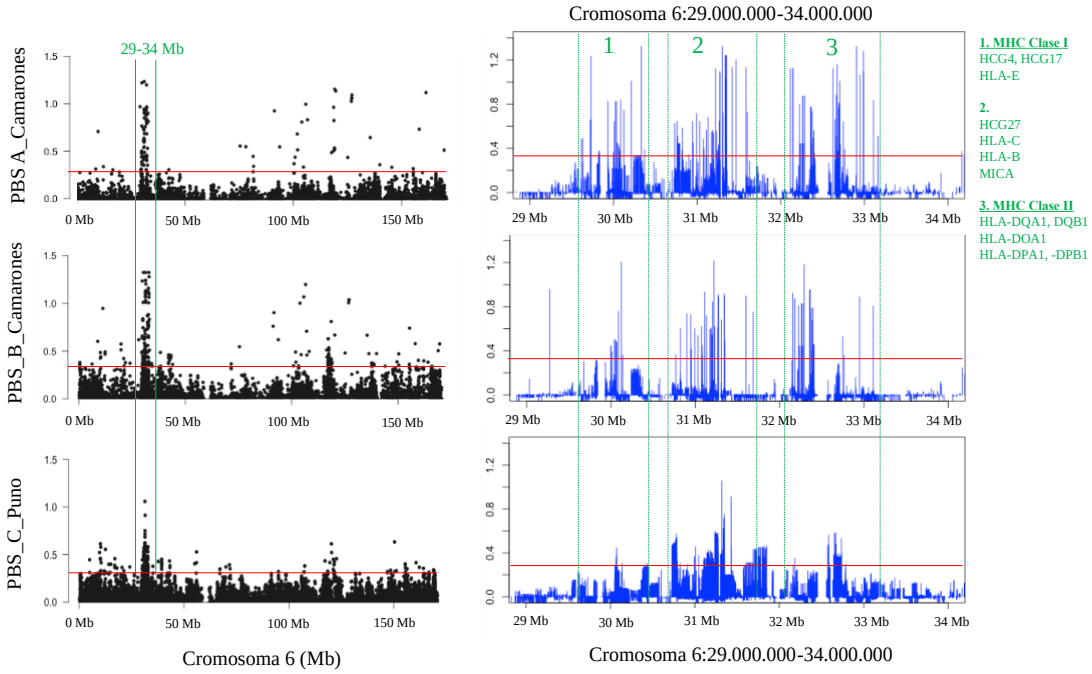


Figura 18. Regiones bajo selección en el cromosoma 6 con PBS. A la izquierda, se observa la señal de selección en el cromosoma 6 para los dos escenarios de PBS de Camarones y el de Puno. A la derecha, se observa dicha región y los genes presentes.

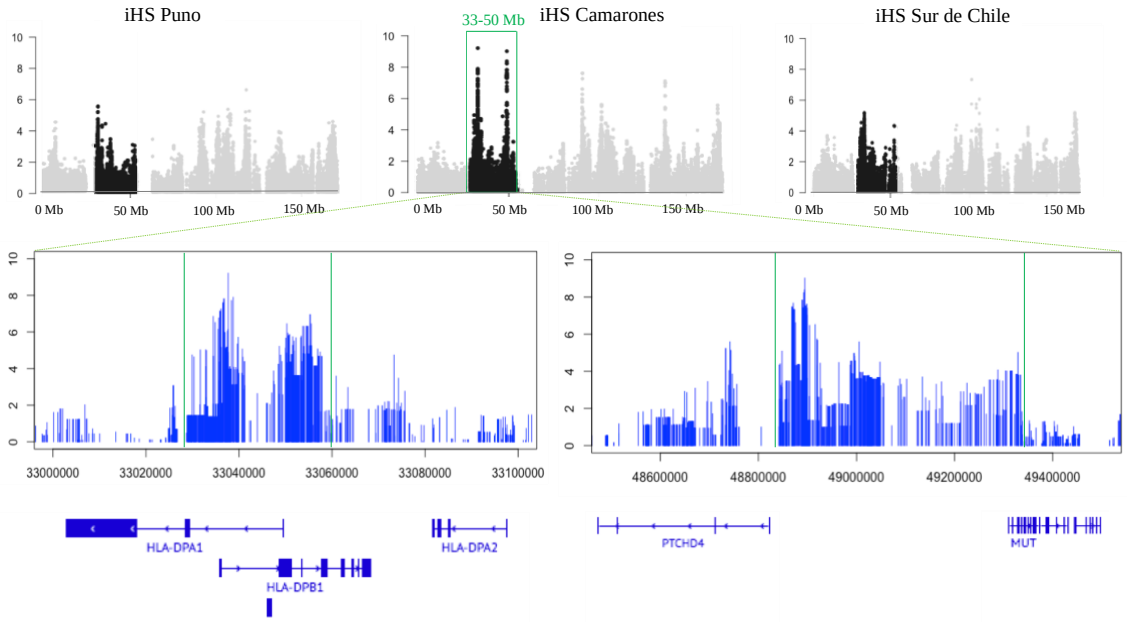


Figura 19. Regiones bajo selección en el cromosoma 6 a partir con iHS. A la izquierda se muestran los histogramas con los valores de PBS calculados para los dos escenarios de Camarones y para Puno a lo largo del cromosoma 6, donde se destaca una señal de selección sobre la región genómica entre 29-34Mb.

Otra región bajo selección que fue encontrada en los tres métodos, pero que no aparece en los top10, fue en el cromosoma 10 entre los 102-105Mb, y particularmente en el gen CNNM2 (proteína transmembrana transportadora de metales) (Figura 20). Sin embargo, esta región también se encuentra bajo selección en Puno. Por lo que es una señal compartida, y así lo muestran las comparaciones, especialmente en el intersecto de los PBS de Puno y Camarones en la Figura 11. En la región de 102-105 Mb, y particularmente la 104.3-104.9 Mb, también se encuentra el gen codificante de la enzima Arsénico

(+3) Metiltransferasa (AS3MT) el cual aparece bajo selección en Camarones (1% de PBS y XP-EHH) pero en un menor ranking que los top10.

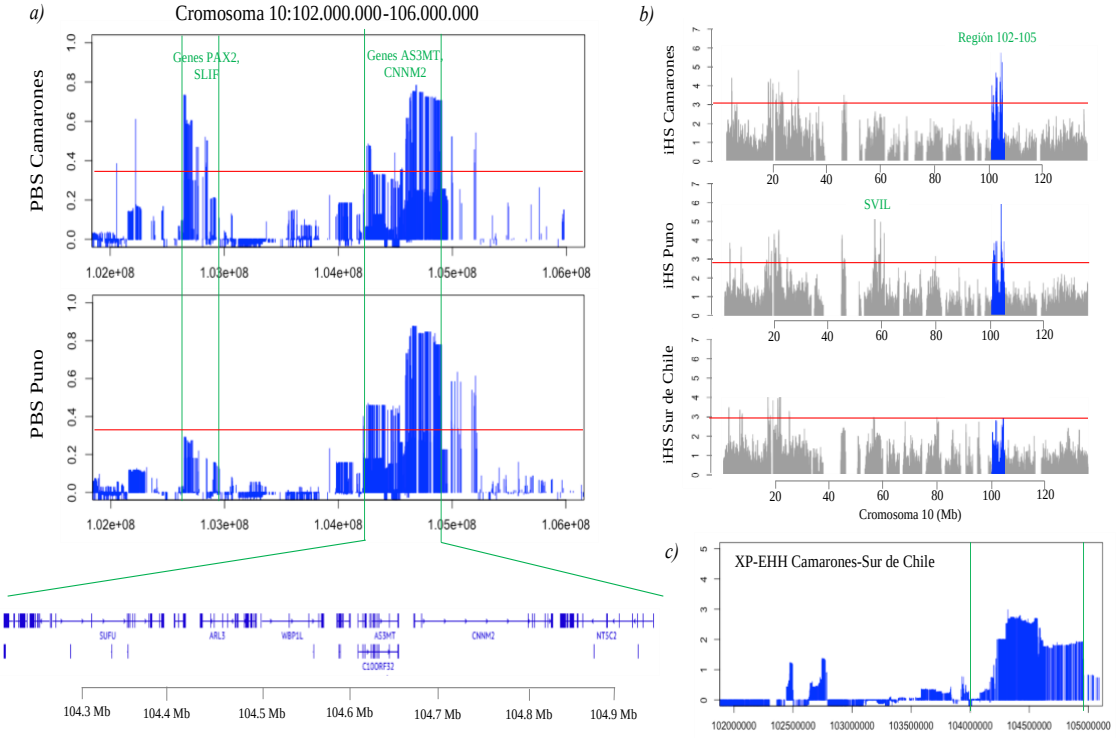


Figura 20. Región bajo selección en el cromosoma 10 de Camarones. A) valores de PBS en un histograma para Camarones y Puno y los genes presentes en la región 102-105Mb. B) valores de iHS en el cromosoma 10 para Camarones, Puno y Sur de Chile. C) Muestra los valores de XP-EHH para Camarones versus Sur de Chile.

DISCUSIÓN

Nuestros resultados han mostrado que la población de Quebrada Camarones posee diferentes señales de selección positiva a nivel genómico, donde algunas son distintivas de Camarones y otras compartidas con las poblaciones amerindias de Puno y Sur de Chile. A partir de la identificación de señales de selección y el análisis de enriquecimiento, proponemos tres posibles eventos adaptativos en Camarones, siendo estos posiblemente relacionados con la exposición al arsénico, el estrés de hipoxia y a patógenos locales. A su vez, también presentamos la discusión de la estructura genética de Camarones, así como de las limitaciones de los métodos empleados.

Estructura poblacional de Camarones

Los efectos de la deriva y el flujo génico contribuyen en gran medida a entender las similitudes o diferencias observadas entre los genomas de los individuos pertenecientes a una población. Junto a la recombinación, como consecuencia del flujo génico, genomas aislados geográficamente se pueden fragmentar en unas pocas generaciones, provocando la perdida de algunas regiones seleccionadas en el pasado (Vitti et al., 2013; Jobling et al., 2013; Novembre et al., 2009). Por lo tanto, nuestro primer objetivo fue caracterizar la estructura genética de Camarones, identificando las proporciones de ancestría genética a nivel global o local dado el contexto de miscegenación (i.e. mestizaje) entre las poblaciones amerindias y europeas en tiempos post-conquista española.

Los resultados muestran que la ancestría genética de la población de Camarones posee un alto componente genético amerindio que es compartido con Puno, pero distinto de Sur de Chile. En lo que respecta a esta ancestría, que aquí denominamos amerindia andina, puede estar reflejando un origen común entre las poblaciones andinas y/o del constante flujo biológico y cultural

en tiempos más tardíos dentro de la región de los andes centro-sur (e.g. entre zonas del lago Titicaca, costa sur del Perú, Norte de Chile y Noroeste argentino). Una ancestría genética común entre las poblaciones de Puno y Camarones, puede ser apoyada por su similar reconocimiento étnico Aymara, el contacto cultural macro-andino en tiempo prehispánicos y antecedentes biológicos como la alta frecuencia del sub-haplogrupo mitocondrial B2 como un linaje materno común entre las poblaciones amerindias de la zona andina (De Saint Pierre et al., 2012; Motti et al., 2013; Moraga et al., 2010).

Por otra parte, Camarones muestra un 11% de ancestría amerindia del sur de Chile, lo cual podría estar reflejando una migración histórica en los último 50 años o bien la persistencia aleatoria de un grupo de alelos comunes entre ambas poblaciones dado su común origen amerindio. En cuanto a un escenario de migración histórica desde el sur de Chile a Camarones, se puede considerar como antecedente, que la región del Desierto de Atacama fue anexada recientemente al territorio chileno a fines del siglo XIX, tras lo cual rápidamente comenzó a recibir oleadas de inmigrantes provenientes del centro-sur del país, debido a la alta demanda en el sector de la minería del salitre, y hoy del cobre.

Adaptación al arsénico

Respecto a nuestra hipótesis sobre un proceso adaptativo a los altos niveles de arsénico en la Quebrada de Camarones, distintivo de otras poblaciones amerindias con menor exposición, nuestros resultados no permiten aceptar del todo nuestra hipótesis. Si bien, sólo Camarones posee bajo selección la región cercana al gen AS3MT (Figura 12), así como otros genes posiblemente relacionados con el arsénico (e.g. GSTP1 en el cromosoma 11), también encontramos señales que son compartidas con la población de Puno. Algunas señales compartidas, son las presentes sobre el gen transportador de iones de metales CNNM2 en la región 10q21(i.e. región reconocida por contener además al gen AS3MT y ser asociada con el metabolismo del arsénico

en el cromosoma 10). Por lo tanto, en lo que respecta al arsénico, nuestras evidencias pueden estar reflejando un escenario micro-evolutivo más complejo, donde una hipótesis adaptativa podría considerar un proceso selectivo al arsénico temprano en una población ancestral a Camarones y Puno en alguna región de los Andes centro-sur, donde la exposición al arsénico hubiera sido más severa. Bajo ese escenario, la señal de selección sobre la región 10q21 (que contiene los genes AS3MT y CNNM2), se habría mantenido de forma más intensa sólo en la población de Camarones, dado que la presión selectiva en dicho valle ha sido constante. Mientras en Puno, dado que el contexto de exposición es mucho menor, la selección pudo haber sido menos intensiva, quedando solo como una “huella de selección” la señal encontrada sobre el gen CNNM2 como evento adaptativo común. Dicho escenario resulta interesante si consideramos que el análisis que mejor refleja este escenario es el de PBS (Figura 12), cuya característica es identificar señales de selección más antiguas y donde los barridos selectivos tienen una menor intensidad. Por otra parte, una segunda hipótesis no excluyente de la primera, puede ser el flujo génico entre ambas poblaciones andinas (Camarones y Puno) que podría explicar la presencia de CNNM2 bajo selección en Puno, donde la exposición actual es mucho menor.

Por otro lado, en ambos escenarios del análisis de PBS en Camarones, aparece otros genes bajo selección ausentes en Puno, como el gen codificante de glutatión (GSTP1) el cual participa en el transporte de las moléculas inorgánicas y orgánicas del arsénico (iAs, MMA y DMA) en los hepatocitos, siendo un reconocido participante en procesos de desintoxicación en humanos (Antonelli et al., 2014; Pierce et al., 2012). Además, también aparecen otros genes con un posible impacto en la supresión del efecto carcinogénico del arsénico como el gen supresor de apoptosis en vías tumorales (gen AVEN), el cual aparece únicamente en Camarones como candidato a estar bajo selección en el top10 de ambos modelos de PBS.

Adaptación a condiciones de hipoxia

A partir de los resultados del enriquecimiento en Camarones, se identificaron procesos biológicos sobre-representados como son la “migración de células musculares”, “procesos del sistema muscular”, “desarrollo del tejido muscular” y “desarrollo del corazón”. En los cuales encontramos genes relacionados con el control de la homeostasis del oxígeno y de la contracción muscular en arterias y cartílago del corazón como ACE, NOX4 y PRKG1. Por ejemplo, la proteína PRKG1 posee un rol central en la regulación de las funciones cardiovasculares y neuronales, además de relajar el tono del músculo liso, prevenir la agregación plaquetaria y modular el crecimiento celular, siendo previamente identificada bajo selección en poblaciones Andinas del noroeste argentino (Eichstaedt et al., 2015a). Mientras que ACE es una enzima implicada en la catálisis de la conversión de angiotensina I en un péptido fisiológicamente activo como angiotensina II. La angiotensina II es un potente péptido vasopresor y estimulante de la aldosterona que controla la presión arterial y el equilibrio hidroelectrolítico. En cuanto a NOX4, esta es una enzima que funciona como una subunidad catalítica del complejo NADPH oxidasa, actuando como un sensor de oxígeno y catalizando la reducción del oxígeno molecular a diversas especies reactivas de oxígeno (ROS). Sin embargo, tanto ACE como NOX4 no han sido previamente señaladas como genes bajo selección en ambientes de altura (Bigham et al., 2016), por lo cual representan nuevas evidencias que contribuyen con otras vías de acción distintas del modelo clásico de factores inducidos por hipoxia (HIF) obervado en Andinos (gen EGLN1) y en habitantes del Tibet (gen EPAS1). ACE, NOX4 y PRKG1 podrían estar mostrando una respuesta adaptativa a través del control de la homeostasis del oxígeno en diferentes tejidos y órganos en condiciones de hipoxia.

A partir de los métodos de PBS y XP-EHH se encontró que el gen EGLN1 tambien se encuentra bajo selección en Camarones y Puno, pero

ausente en Sur de Chile. La proteína codificada por este gen cataliza la formación postraduccional de 4-hidroxiprolina en proteínas alfa del factor inducible por hipoxia HIF (Hypoxia-inducible factor). EGLN1 como otros genes del complejo HIF, tiene un rol importante en el control de la homeostasis de oxígeno en mamíferos. Además, dicho gen ha sido previamente identificado a estar bajo selección en poblaciones andinas de Bolivia bajo condiciones de hipoxia de altura (~4.000 m.s.n.m.) (Bigham et al., 2011). La señal de selección sobre este gen en la población de Puno resulta congruente con una adaptación local a ambientes de altura (alrededor de los ~3.800 m.s.n.m.). Sin embargo, en Camarones tal señal, así como el de los otros genes presentes en el enriquecimiento, pueden estar reflejando una historia micro-evolutiva similar a la del arsénico. Donde la hipótesis de una población ancestral y/o un escenario tardío de flujo génico pueden contribuir a entender la presencia de genes asociados con hipoxia en una población de baja altura como Camarones (~1.000 m.s.n.m.).

Adaptación a patógenos

A partir de los tres métodos de detección de selección positiva, se identificó una región común bajo selección en Camarones, la cual se localiza sobre genes codificantes de las moléculas del sistema de antígenos leucocitario humano HLA Clase I y II, y especialmente sobre el gen HLA-DPA1. Diferentes genes en la región de HLA han sido identificadas bajo selección frente a presiones de patógenos locales, tales como los involucrados en la respuesta a anemia falciforme en contextos de malaria en África (Sánchez-Maza et al., 2017) o HLA-DQA frente a patógenos antiguos en individuos Tsimishian del norte de América (Lindo et al., 2016).

Se ha propuesto que la región de HLA ha estado evolucionando bajo los efectos de la selección positiva y balanceadora, entregando principalmente respuestas inmunes a patógenos locales a través de la historia humana (Meyer

et al., 2018). En cuanto a las moléculas de HLA-DPA1, estas presentan antígeno para las células T auxiliares CD4+, siendo la respuesta inmune de este tipo de células crucial en la eliminación algunos patógenos virales tales como el virus de la hepatitis B (VHB) (Katrinli et al., 2017). Estudios en poblaciones de Europa y Asia, han mostrados variantes en el antígeno leucocitario HLA-DPA1/DPB1 que influyen en la predisposición o el mayor riesgo al VHB (Katrinli et al., 2017; An et al., 2011). Mientras que, en el sudeste asiático, se han reportado algunas variantes que muestran disminuir el efecto del VHB (Wasityastuti et al., 2016). Si bien la hepatitis B es una enfermedad infecciosa que es endémica en América, especialmente con una distribución filogeográfica caracterizada por el predominio de la cepa VHB-F1b en regiones costeras del Perú, Chile y el sur de Argentina (Zehender et al., 2014), los efectos del VHB no dejan evidencias observables en los restos bioantropólógicos. Por lo cual, si bien VHB es una enfermedad mortal dado sus efectos carcinogénicos a nivel hepático (i.e. cirrosis), y especialmente a temprana edad con una deficiente respuesta inmune, una posible respuesta adaptativa a esta enfermedad necesitaría de mayores evidencias bioarqueológicas y epidemiológicas respecto a sus efectos sobre los tempranos y actuales habitantes de la región, respectivamente.

Las diferentes señales observadas sobre la región de HLA clase II y clase I mostradas en la Figura 18, podrían también estar indicando otras posibles respuestas adaptativas relacionadas con el sistema inmune en Camarones. En este sentido, la amplia evidencia bioarqueológica de afección de las poblaciones amerindias tempranas de Camarones, y de otros valles del Desierto de Atacama, a enfermedades infecciosas causadas por *Mycobacterium tuberculosis* para la tuberculosis (Tb) y parásitos *como Treponema cruzi* para Chagas, podrían constituir un posible escenario adaptativo, dado el largo periodo de interacción que han tenido las poblaciones amerindias con estos patógenos en los últimos ~9.000 años (Orellana-Halkyer

& Arriaza, 2010; Aufderheide et al., 2009; Arriaza et al., 1995). A diferencia de VHB, Tb y Chagas si dejan evidencias a nivel óseo y molecular en los restos bioantropológicos. Siendo las poblaciones costeras Chinchorro (7.000-3.000 AP) uno de los casos más estudiados por su alta prevalencia en el periodo del Arcaico Temprano (7.000-5000 AP) (Arriaza & Standen, 2008; Aufderheide et al., 2009). De hecho, en cuanto Chagas, se ha propuesto que algunas poblaciones de zonas tropicales de Bolivia, la presencia de un haplotipo protector en los genes HLA-DRB1 y HLA-B (del Puerto et al., 2012).

En este sentido, valles costeros como Camarones podrían haber constituido un nicho ecológico para los vectores de enfermedades infecciosas, como Tb y Chagas, actuando como ambientes aislados de microclima costero y semitropical, lo cual marca una diferencia con las zonas más áridas del desierto o el altiplano. A partir de nuestro análisis de XP-EHH, resulta interesante el hecho de que las señales observadas sobre la región de HLA marquen una fuerte diferencia entre Camarones y Puno, pero no así entre Camarones y Sur de Chile. Tal diferencia podría estar reflejando una señal que es propia de las condiciones costeras y/o de modos de vida cazador-recolector, donde la exposición a enfermedades infecciosas es mas frecuente, lo cual pudo haber sido un escenario similar entre los tempranos habitantes costeros de Camarones y los de la costa sur de Chile (e.g. Huilliches). Considerando dichos antecedentes, HLA pudo haber conferido una respuesta adaptativa a las condiciones de patógenos locales de los valles costeros del norte de Chile, lo cual resulta ser distintivo del altiplano andino.

Consideraciones de los métodos genómicos de detección de selección positiva

Es necesario mencionar que el enfoque genómico en estudios de selección positiva tiene limitaciones importantes de considerar. Un primer desafío es la interpretación coherente de los hallazgos con la información de los

posibles fenotipos bajo selección. En nuestro estudio, las señales de selección sobre la región 104Mb del cromosoma 10, previamente asociada con el metabolismo del arsénico, resulta coherente con la historia de exposición a arsénico. No obstante, en el caso de la región de HLA por ejemplo podría tener una diferente explicación biológica que los métodos de selección no están detectando. Al no considerar una prueba estadística que evalúe un escenario de evolución neutral versus selección, las señales de selección también podrían haber sido afectadas por los cambios demográficos de la población estudiada, las características del barrido selectivo o los efectos de otras fuerzas evolutivas (Wilson B, Petrov D y Messer P, 2014).

A pesar de lo anterior, los métodos aquí utilizados poseen un amplio poder de identificación de selección, especialmente de haplotipos recientes con frecuencias intermedias o altas en el caso iHS, y de alta frecuencia o fijación en el caso de XPEHH. Mientras que PBS tiene mayor poder al detectar eventos selectivos de mayor antigüedad, lo cual se complementa de buena forma con XPEHH (Barreiro et al. 2008; Lappalainen et al. 2010), y lo cual podemos observar bastante bien en la mayor cantidad de señales compartidas entre PBS y XP-EHH en nuestros resultados. Bajo diferentes simulaciones, los tres métodos aquí utilizados poseen un alto poder de identificar señales de selección, y con amplios ejemplos de adaptaciones locales halladas mediante su uso en la historia microevolutiva de las poblaciones humanas post *Out of Africa* (Fan et al., 2015, Vitti et al., 2013; Yi et al., 2010; Pickrell et al., 2009; Sabeti et al., 2009).

CONCLUSIONES

A partir de nuestros resultados, podemos concluir que la población de Quebrada Camarones posee adaptaciones biológicas que la diferencian de otras poblaciones de similar ancestría amerindia. Respecto a nuestra hipótesis, esta no puede ser aceptada del todo, debido a que encontramos señales de selección que son compartidas con otras poblaciones amerindias, como es el caso de una señal común sobre el gen CNNM2 en la región cercana al gen AS3MT entre Camarones y Puno. En este sentido, un evento adaptativo al arsénico podría haber sido un proceso selectivo común en las poblaciones de ancestría amerindia andina de la región de los Andes Centro-Sur, reflejando un escenario micro-evolutivo más complejo, que otras fuerzas evolutivas como la migración, podrían contribuir a comprender.

De nuestro principal objetivo podemos señalar que la población de Camarones presenta señales de selección que permiten al menos proponer tres eventos adaptativos asociados con: respuestas inmunes a patógenos locales, estrés por hipoxia y el antes mencionado caso del arsénico. Respecto a nuestros objetivos específicos, es posible concluir que Camarones presenta una estructuración poblacional similar a Puno, compartiendo una ancestría genética común amerindia andina, que es distinta de la amerindia del Sur de Chile. No obstante, dicho componente genético sureño, también se encuentra en Camarones, pero en mucho menor intensidad. De la identificación de señales de selección, los tres métodos empleados permitieron identificar regiones del genoma bajo selección, siendo las principales cercanas a la región HLA en el cromosoma 6 y en genes asociados al metabolismo del arsénico (GSTP1, AS3MT y CNNM2). A partir de las comparaciones realizadas, podemos señalar que las señales de selección sobre la región HLA son distintivas de Camarones, especialmente aquellas variantes en el gen HLA-DPA1. A su vez, a partir del análisis de enriquecimiento, fue posible identificar procesos biológicos y genes

involucrados en la regulación de la homeostasis del oxígeno en la población de Camarones (e.g. genes ACE y NOX4), los que podrían tener un rol adaptativo a hipoxia. En este sentido, nuestro estudio devela eventos adaptativos locales en la población de Quebrada Camarones asociados con las diferentes condiciones naturales adversas del Desierto de Atacama, y con una posible historia común entre las poblaciones amerindias de la región de los Andes Centro-Sur.

Bibliografía

Agusa T, Fujihara J, Takeshita H, YIwata H. 2011. Individual Variation in Inorganic Arsenic Metabolism Associated with AS3MT Genetic Polymorphisms. Int. J. Mol. Sci.12: 2351-2382.

Antonelli R, Shao K, Thomas DJ, Sams II R, Cowden J. 2014. AS3MT, GSTO, and PNP polymorphisms: Impact on arsenic methylation and implications for disease susceptibility. Environmental Research 132: 156–167.

Apata M, Arriaza B, Llop E, Moraga M. 2017. Human adaptation to arsenic in Andean people of the Atacama Desert. Am J Phys Anthropol 163: 192–199.

Arriaza BT, Salo WL, Aufderheide AC, Holcomb TA 1995. Pre-Columbian tuberculosis in Northern Chile: molecular and skeletal evidence. Am J Phys Anthropolol 98: 37-45.

Arriaza BT. 2005. Arseniasis as an environmental hypothetical explanation for the origin of the oldest artificial mummification practice in the world. Chungara, Revista de Antropología Chilena 37(2): 255-260.

Arriaza B, Standen V. 2008. Bioarqueología. Historia Biocultural de los Antiguos Pobladores del Extremo Norte de Chile. Editorial Universitaria. Santiago de Chile.

Arriaza BT, Amarasiriwardena D, Cornejo L, Standen V, Byrne S, Bartkus L, et al. 2010. Exploring chronic arsenic poisoning in pre-Columbian Chilean mummies. Journal of Archaeological Sciences 37: 1274–1278.

Ahsan H, Chen Y, Parvez F, Zablotska L, Argos M, Hussain I, et al. 2006. Arsenic Exposure from Drinking Water and Risk of Premalignant Skin Lesions in Bangladesh: Baseline Results from the Health Effects of Arsenic Longitudinal Study. Am. J. Epidemiol. 163(12): 1138-1148.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5): 711-22.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics, 12(1): 246.

Amorim CE, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hünemeier T. 2017. Genetic signature of natural selection in first Americans. Proceedings of the National Academy of Sciences 114 (9): 2195-2199.

An P, Winkler C, Guan Li, O'Brien SJ, Zengcorresponding Z, HBV Study Consortium. 2011. A Common HLA–DPA1 Variant is a Major Determinant of Hepatitis B Virus Clearance in Han Chinese. J Infect Dis 203(7): 943-947.

Aufderheide AC, Salo W, Madden M, Streitz J, Buikstra J, Guhl F, et al. 2009. A 9,000-year record of Chagas' disease. Proc Natl Acad Sci 101(7): 2034–2039.

Bartkus L, Amarasiriwardena D, Arriaza B, Bellis D, Yáñez J. 2011. Exploring lead exposure in ancient Chilean mummies using a single strand of hair by laser ablation-inductively coupled plasma-mass spectrometry (LA-ICP-MS). Microchemical Journal 98: 267-274.

Bigham AW, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, ... Shriver, MD. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genetics, 6(9).

Bigham AW, Wilson MJ, Julian CG, Kiyamu M, Vargas E, Leon-Velarde F, et al. 2013. Andean and Tibetan patterns of adaptation to high altitude. Am. J. Hum. Biol. 25:190–197.

Bigham AW & Lee F. 2014. Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes & Development. Cold Spring Harbor Laboratory Press. 28: 2189-2204

Bundschuh J, Litter ML, Parvez F, Román-Ross G, Nicolli HB, Jean JS, et al. 2012. One century of arsenic in Latin America: A review of history and occurrence from 14 countries. Science of the Total Environment 429: 2–35.

Byrne S, Amarasiriwardena D, Bandak B, Bartkus L, Kane J, Jones J, et al. 2010. Where Chinchorro exposed to arsenic? Arsenic determination in Chinchorro mummies'hair by Laser Ablation Inductively Coupled Plasma-Mass Spectrometry (LA-ICP-MS). Microchemical Journal 94: 28-35.

Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, et al. 2014. Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. PLoS ONE 9(5): e98076.

Cavalli-Sforza LL. 1969. Human Diversity. Proc. 12th Int. Congr. Genet. 2, 405-416.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience Vol 4, Issue 1: 1-16.

Cornejo-Ponce L, Lienqueo H, Arriaza BT. 2011. Levels of total arsenic in edible fish and shellfish obtained from two coastal sectors of the Atacama Desert in north of Chile: Use of non-migratory marine species as bioindicators of sea environmental pollution. Journal of Environmental Sciences Health, Part A: Toxic/Hazardous Substances and Environmental Engineering 46: 1274-1282.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics 27(15): 2156-8.

del Puerto F, Nishizawa JE, Kikuchi M, Roca Y, Avilas C, et al. 2012. Protective Human Leucocyte Antigen Haplotype, HLA-DRB1*01-B*14, against Chronic Chagas Disease in Bolivia. PLoS Negl Trop Dis 6(3): e1587

Eden E, Navon R, Steinfeld, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:48.

Eichstaedt CA, Antão T, Cardona A, Pagani L, Kivisild T & Mormina M. 2015a. Genetic and phenotypic differentiation of an Andean intermediate altitude population. Physiol Rep 3 (5): e12376.

Eichstaedt CA, Antão T, Cardona A, Pagani L, Kivisild T, & Mormina M. 2015b. Positive selection of AS3MT to arsenic water in Andean populations. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis.

Engström K, Vahter M, Mlakar SJ, Concha G, Nermell B, Raqib R, et al. 2011. Polymorphisms in Arsenic (+III Oxidation State) Methyltransferase (AS3MT) Predict Gene Expression of AS3MT as Well as Arsenic Metabolism. Environ Health Perspect 119(2): 182–188.

Engström KS, Hossain MB, Lauss M, Ahmed S, Raqib R, Vahter M, et al. 2013. Efficient Arsenic Metabolism —The AS3MT Haplotype Is Associated with DNA Methylation and Expression of Multiple Genes Around AS3MT. PLoS ONE 8(1): e53732.

Fan S, Hansen MEB, Lo Y, Tishkoff SA. (2016). Going global by adapting local: A review of recent human adaptation. Science 354(6308), 54-59.

Ferrecio C, Sancha AM. 2006. Arsenic Exposure and its impact on health in Chile. J Health Popul Nutr 24(2): 164-175.

Figueroa L. 2001. Arica inserta en una region arsenical: El arsénico en el ambiente que la afecta y sus 45 siglos de arsenicismo crónico. Ediciones Universidad de Tarapacá. Arica, Chile.

Gardner RM, Nermell B, Kippler M, Grander M, Li L, Ekstrom EC, et al. 2010. Arsenic methylation efficiency increases during the first trimester of pregnancy independent of folate estatus. Reprod Toxicol 31:210–218.

George CM, Sima L, Arias MLJ, Mihalic J, Cabrera LZ, Danz D, Checkley W, Gilman RH. 2014. Bulletin of the World Health Organization 92: 565-572.

Goebel T, Waters MR, O'Rourke DH. 2008. The Late Pleistocene Dispersal of Modern Humans in the Americas. Science 319 : 1497-1502.

Hernández A, Ximena N, Surralles J, Sekaran C, Tokunaga H, Quinteros D, et al. 2008a. Role of the Met(287)Thr polymorphism in the AS3MT gene on the metabolic arsenic profile. Mutat Res 637(1-2): 80-92.

Hernández A, Sekaran C, Tokunaga H, Sampayo-Reyes A, Quinteros D, Creus A, et al. 2008b. High arsenic metabolic efficiency in AS3MT Met(287)Thr allele carriers. Pharmacogenet Genomics 18(4): 349-355.

Hernández A, Paiva L, Creus A, Quinteros D, Marcos R. 2014. Micronucleus frequency in cooper-mine workers exposed to arsenic is modulated by the AS3MT Met287Thr polymorphism. Mutat Res 759: 51-55.

Herron JC & Freeman S. 2015. Evolutionary Analysis (5 Ed.). Pearson.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet 11(12): e1005602.

Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, ... Risch N. 2011. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics, 98(6), 422–430.

Hopenhayn-Rich C, Browning S, Hertz-Picciotto I, Ferreccio C, Peralta C, Gibb H. 2000. Chronic arsenic exposure and risk of infant mortality in two areas of Chile. Environ Health Perspect 108(7): 667-73.

Hopenhayn C, Ferreccio C, Browning SR, Huang B, Peralta C, Gibb H, et al. 2003. Arsenic exposure from drinking water and birth weight. Epidemiology 14 (5): 593-602.

Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5(6): e1000529.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics volume 44: 955-959.

Hsu L, Chen GS, Lee CH, Yang TS, Chen YH, Wang YH, et al. 2013. Use of Arsenic-Induced Palmoplantar Hyperkeratosis and Skin Cancers to Predict Risk of Subsequent Internal Malignancy. Am J Epidemiol 177: 202-212.

Hsu F, Kent JW, Clawson H, Kuhn RM, Diekhans M, & Haussler D. 2006. The UCSC known genes. Bioinformatics, 22(9): 1036-1046.

Hughes MF, Beck BD, Chen Y, Lewis AS, Thomas DJ. 2011. Arsenic Exposure and Toxicology: A Historical Perspective. Toxicol Sci 132(2): 305-332.

Jeong C, Di Rienzo A. 2014. Adaptations to local environments in modern human populations. Curr Opin Genet Dev 29:1-8.

Katrinli S, Enc FY, Ozdil K, Ozturk O, Tuncer I, Doganay GD, Doganay L. Effect of HLA-DPA1 alleles on chronic hepatitis B prognosis and treatment response. North Clin Istanb 25(3):168-174.

Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS. 2016. A time transect of exomes from a Native American population before and after European contact. Nature Communications 7:13175.

Lin S, Shi Q, Nix FB, Styblo M, Beck MA, Herbin-Davis KM, et al. 2002. A novel S-adenosyl-l-methionine:arsenic(III) methyltransferase from rat liver cytosol. J Biol Chem 277(13): 10795–10803.

López D, Bundschuh J, Birkle P, Armienta MA, Cumbal L, Sracek O, et al. 2012. Arsenic in volcanic geothermal fluids of Latin America. Sci Total Environ 429: 57-75.

Mass MJ, Tennant A, Roop BC, Cullen WR, Styblo M, Thomas DJ, et al. 2001. Methylated trivalent arsenic species are genotoxic. Chem Res Toxicol 14(4): 355–361.

Marshall G, Ferrecio C, Yuan Y, Bates MN, Stenmaus C, Selvin S, et al. 2007. Fifty-Year Study of Lung and Bladder Cancer Mortality in Chile Related to Arsenic in Drinking Water. JNCI J Natl Cancer Inst 99(12): 920-928.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics 93 (2): 278-288.

Meyer D, Aguiar VRC, Bitarello BD, Brandt DYC, Nunes K. 2018. A genomic perspective on HLA evolution. Immunogenetics (2018) 70:5–27.

Moraga M, Santoro C, Standen V, Carvallo P, Rothhammer F. 2006. Microevolution in Prehistoric in Andean Populations: Chronologic mtDNA Variation in the Desert Valleys of Northern of Chile. Am. J. Phys. Anthropol. 127: 170-181.

Mukherjee A, Verma S, Gupta S, Henke K, Bhattacharva P. 2014. Influence of tectonics, sedimentation and aqueous flow cycle on the origin of global groundwater arsenic: Paradigms from three continents. Journal of Hidrology 518, Part C: 284-299.

Nielsen R & Slatkin MW. 2013. An introduction to population genetics: theory and applications. Sinauer Associates.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. Nature Vol 456: 98-101.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40(5): 646-9.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, ... Marchini J. 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genetics, 10(4).

Oleksyk TK, O'Brien SJ, Smith MW. 2010. Scanning genomes for the footprints of historic selection. Phil Trans R Soc B 365: 185-205.

Orellana-Halkyer N, Arriaza B. 2010. Enfermedad de Chagas en poblaciones prehistóricas del norte de Chile. Revista Chilena de Historia Natural 83: 531-541.

Pierce BL, Kibriya MG, Tong L, Jasmine F, Argos M, et al. 2012. Genome-Wide Association Study Identifies Chromosome 10q24.32 Variants Associated with Arsenic Metabolism and Toxicity Phenotypes in Bangladesh. PLoS Genet 8(2):1002522.

Pierce B, Tong L, Argos M, Gao J, Jasmine F, Roy S, et al. 2013. Arsenic metabolism efficiency has a causal role in arseni toxicity: Mendelian randomization and gene-environment interaction. International Journal of Epidemiology 42: 1862-1872.

Pickrell, J. K., Coop, G., & Novembre, J. 2009. Signals of recent positive selection in a worldwide sample of human populations, 826–837.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20(4): 208–15.

Rademaker K, Hodgins G, Moore K, Zarrillo S, Miller C, Bromley GRM, Leach P, Reid DA, Álvarez WY, Sandweiss DH. 2013. Paleoindian settlement of the high-altitude Peruvian Andes. Science Vol. 346, Issue 6208: 466-469.

Raghavan M. et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349: aab3884.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra M V., Rojas W, Duque C, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Rothammer FA, Allison M, Núñez L, Standen V, B Arriaza (1985) Chagas disease in pre-Columbian South America. American Journal of Physical Anthropology 68: 495-498.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909): 832–37.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. 2006. Positive natural selection in the human lineage. Science 312(5780):1614–20.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, ... Lander ES. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature, 449: 913–919.

Sandoval JR, Lacerda DR, Acosta O, Jota MSA, Robles-Ruiz P, Salazar-Granara A, et al. 2016. The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations. Annals of Human Genetics 80: 88-101.

Salzano F. 2016. The role of natural selection in human evolution – insights from Latin America. Genet Mol Biol. 39(3): 302-311.

Schlawicke Engström K, Broeberg K, Concha G, Vahter M, Malakar SJ, Vahter M. 2007. Genetic polymorphisms influencing arsenic metabolism: evidences from Argentina. Environ Health Perspect 115(4): 559-605.

Schlebusch CM, Lewis CM, Vahter M, Engström K, Tito RY, Huerta D, et al. 2013. Possible Positive Selection for an Arsenic-Protective Haplotype in Humans. Environ Health Perspect 121(1): 53-58.

Schlebusch CM, Gattepaille L, Engström K, Vahter M, Jakobsson M, Broberg K. 2015. Human Adaptation to Arsenic-Rich Environments. Mol Biol Evol 32(6): 1544-1555.

Shriver M, Kennedy G, Parra E, Lawson H, Sonpar V, Huang J, et al. 2004. The genomic distribution of population substructure in four populations using 8,525 automsomal SNPs. Human Genomics vol. 1 (4): 274-286.

Sokal R, Michener C. 1958. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28: 1409-1438.

Sumi D, Himeno S. 2012. Role of Arsenic (+3 Oxidation State) Methyltransferasa in Arsenic Metabolism and Toxicity. Biol Pharm Bull 35(11): 1870-1875.

Steinmaus C, Ferrecio C, Acevedo J, Yuan Y, Liaw J, Durán V, et al. 2014. Increased Lung and Bladder Cancer Incidence in Adults After In Utero and Early-Life Arsenic Exposure. Cancer Epidemiol Biomarkers Prev 23(8): 1529-38.

Styblo M, Del Razo L, Vega L, Germolec DR, LeGluyse EL, Hamilton GA, et al. 2000. Comparative toxicity of trivalent and pentavalent inorganic and methylated arsenicals in rat and human cells. Arch Toxicol 74(6): 289-299.

Swift J, Cupper ML, Greig A, Westaway MC, Carter C, Santoro CM, et al. 2015. Skeletal arsenic of the pre-columbian population of Caleta Vitor, northern Chile. Journal of Archaeological Science 58: 31-45.

Szpiech ZA, Hernandez RD. 2014. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol Biol Evol (3): 1–4.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30(12), 2725–2729.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. 2007. Convergent adaptation of human lactase persistance in Africa and Europe. Nat Genet 39: 31-40.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature volume 526: 68–74.

Thomas DJ, Li J, Waters SB, Xing W, Adair BM, Drobna Z, et al. 2007. Arsenic (+3 Oxidation State) Methyltransferasa and methylation of arsenicals. Exp Biol Med 232(1): 3-13.

Valverde G, Zhou H, Lippold S, de Filippo C, Tang K, López Herráez D, et al. 2015. A Novel Candidate Region for Genetic Adaptation to High Altitude in Andean Populations. PLoS ONE 10(5): e0125444.

Valenzuela OL, Drobná Z, Hernández-Castellano E, Sanchez-Peña LC, García-Vargas GC, et al. 2009. Association of AS3MT polymorphisms and the risk of premalignant arsenic skin lesions. Toxic Appl Pharmacol 239(2): 200-7.

Vitti J, Grossman S, Sabeti P. 2013. Detecting Natural Selection in Genome Data. Annu. Rev. Genet. 47: 97-120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biology, 4(3), e72.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16): e164.

Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc. Natl. Acad. Sci. USA 103(1): 135–40.

Wasityastuti W, Yano Y, Ratnasari N, Triyono T, Triwikatmani C, Indrarti F, Heriyanto DS, Yamani LN, Liang Y, Utsumi T, Hayashi Y. 2016. Protective effects of HLA-DPA1/DPB1 variants against Hepatitis B virus infection in an Indonesian population. Infect Genet Evol 41:177-184.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution Vol. 38, No. 6: 1358-1370.

Wilson B, Petrov D y Messer P. 2014. Soft selective sweeps in complex demographic scenarios. Genetics 198(2):669-84.

Wollstein A, Stephan W. 2015. Inferring positive selection in humans from genomic data. Investigative Genetics, 6(1), 1–8.

World Health Organization. 2011. Guidelines for drinking-water quality, fourth edition.

Yáñez J, Fierro V, Mansilla H, Figueroa L, Cornejo L, Barnes RM. 2005. Arsenic speciation in human hair: a new perspective for epidemiological assessment in chronic arsenicism. J Envon Monit 7: 1335-1341.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Ping Z, Pool J, et al. 2010. Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. Science 329(5987): 75-78.

Zehender G, Ebranati E, Gabanelli E, Sorrentino C, Lo Presti A, Tanzi E, et al. 2014. Enigmatic origin of hepatitis B virus: An ancient travelling companion or a recent encounter? World J Gastroenterol 20:7622-34.

Zhang C, Bailey DK, Awad T, Liu G, Xing G, et al. 2006. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22(17): 2122–28.