

**UNIVERSIDAD DE CHILE  
FACULTAD DE MEDICINA  
ESCUELA DE POSTGRADO**



**ESTUDIO GENÓMICO DE RIESGO PARA CÁNCER  
GÁSTRICO Y SU ASOCIACIÓN CON ANCESTRÍA EN LA  
POBLACIÓN CHILENA**

**ROCÍO MARIANA RETAMALES ORTEGA**

**TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS BIOMÉDICAS**

**Director de Tesis: Prof. Dr. Ricardo Verdugo Salgado  
Co-Director de Tesis: Dr. Patricio González Hormazábal**

**2020**

## DEDICATORIA

Para mi familia, Sylvia Mariana, José Miguel, Francisco José y mi Abueli Silvia y  
Nonna Edda.

## **AGRADECIMIENTOS**

Cada uno de los pasos dados para llegar a este momento han sido un gran desafío, y todas las situaciones que he vivido han servido para convertirme en lo que hoy soy. No ha sido fácil y el camino se puso cuesta arriba en muchos momentos, pero pude llegar hasta el final gracias al apoyo de un grupo grande de personas, que incluye a mis familiares, amigos, compañeros de laboratorio, apoyo médico y psicológico, y también a la comisión y mis tutores. Es difícil poder decir en una sola palabra o frase lo agradecida y bendecida que me encuentro de haber tenido de su ayuda, porque sin Uds. terminar esto no hubiese sido posible.

En primer lugar, quiero agradecer a mi familia. Soy la primera persona que se dedica a la ciencia entre un gran grupo de ingenieros, por lo que a veces ha sido complicado que entiendan la vocación que siento que me lleva a entregarlo todo y a seguir adelante. La verdad es que no me hallo haciendo otra cosa que esto. Sé que andar con el computador a todos lados durante casi dos años fue complicado para la vida familiar, así que les agradezco infinitamente a mis padres, José y Sylvia, que me hayan ayudado y apoyado incondicionalmente para llegar a ser Doctora. Quiero también agradecerle a mi hermano Francisco, a quien lo llamo mi alter ego, siempre ha sido un apoyo y un gran amigo. A Karina y primos Alberto, Benito y Javiera, que me levantaron, escucharon e instaron a seguir adelante.

Otras personas que han sido importantes en mi vida, son mis amigas Fernanda, Carla y Camila, que muchas veces me sacaron de la inercia en la que me encontré durante un periodo oscuro y me dieron su apoyo y fuerza para seguir adelante. A Elizabeth, Lorena, Sandra y Norma, que me escucharon y acompañaron durante este proceso. Todas me han ayudado en distintos episodios y crisis que tuve, y sin ese apoyo no podría haberme levantado.

A mis compañeros de laboratorio que participaron de alguna manera para el desarrollo de esta tesis, Eduardo, Mara y Lourdes. A aquellos que me ayudaron con todas esas cosas que no sabía realizar en el “laboratorio seco”, Patricio, Evelin y Cristian, todos los análisis informáticos que pude realizar (que al final es toda la tesis) es gracias a ellos. Pablo, juntos sacamos adelante nuestra tesis, tal como lo prometimos.

Agradecer a CONICYT (Beca Doctorado Nacional 21160297) que me financió durante el desarrollo del doctorado, también a mis tutores que me apoyaron desde el primer día en este proceso y me han guiado durante el desarrollo de esta tesis. Intentaba avanzar lo más rápido que podía, pero en aquellos momentos en los que se ponía más complicada la tesis, me ayudaron para salir adelante. Finalmente, a la comisión evaluadora de mi tesis que participaron en su desarrollo y me ayudaron a mejorar e instaron a trabajar más duro para hacer que mi tesis fuese más clara y “redondita”.

Estoy muy agradecida del apoyo de todos, y a todos los que no nombré pero estuvieron ahí, gracias, gracias, gracias.

## **TABLA DE CONTENIDOS**

<b>DEDICATORIA</b>	<b>2</b>
<b>AGRADECIMIENTOS</b>	<b>3</b>
<b>TABLA DE CONTENIDOS</b>	<b>5</b>
<b>ÍNDICE DE TABLAS Y FIGURAS</b>	<b>8</b>
Índice de Tablas	8
Índice de Figuras	9
Índice de Tablas Suplementarias	10
Índice de Figuras Suplementarias	11
<b>LISTA DE ABREVIATURAS</b>	<b>13</b>
<b>RESUMEN</b>	<b>15</b>
<b>ABSTRACT</b>	<b>17</b>
<b>I. INTRODUCCIÓN</b>	<b>19</b>
I.1. Epidemiología	19
I.2. Aspectos histológicos	20
I.3. Factores de Riesgo	21
I.3.1. Ambientales	21
I.3.2. Moleculares y genéticos	22
I.4. Genética poblacional del Cáncer Gástrico	23
I.4.1 Aspectos Generales	23
I.4.2 América Latina y Chile	24
I.4.3 Relación de la ancestría con el cáncer	25
I.5. Riesgo Genético	26
I.6. Búsqueda de marcadores genéticos de riesgo	28
<b>II. HIPÓTESIS</b>	<b>30</b>
<b>III. OBJETIVO GENERAL</b>	<b>30</b>
III.1 Objetivo específico 1	30
III.2 Objetivo específico 2	30
III.3 Objetivo específico 3	30
<b>IV. MATERIALES Y MÉTODOS</b>	<b>31</b>
IV.1. Selección de SNPs asociados a cáncer gástrico en distintas poblaciones	31

IV.2. Genotipificación de las muestras	32
IV.2.1 Objetivo 1	32
IV.2.2 Objetivo 2	33
IV.3. Control de Calidad	34
IV.3.1 Objetivo 1	34
IV.3.2 Objetivo 2	34
IV.4. Estimación de ancestrías globales	35
IV.4.1 Objetivo 1	35
IV.4.2 Objetivo 2	35
IV.5. Puntaje de Riesgo Poligénico	36
IV.6. Estimación y análisis de ancestría local	36
IV.7. Análisis de Asociación	38
IV.8. Análisis de Componentes Principales	39
IV.9. Evaluación de Riesgo Genético	39
IV.9.1 Selección de modelos genéticos de riesgo	39
IV.9.2 Puntaje de Riesgo Poligénico	40
IV.10. Admixture Mapping	41
IV.10.1 Cálculo de poder de la muestra	41
IV.10.2 Análisis de Admixture mapping en regiones asociadas a Cáncer Gástrico por su ancestría	42
IV.11. Identificación de enriquecimiento de funciones génicas en regiones genómicas	43
IV.12. Estadística	44
<b>V. RESULTADOS</b>	<b>45</b>
V.1. Objetivo Específico 1	45
V.1.A Selección SNPs asociados a cáncer gástrico en distintas poblaciones.	45
V.1.B Estimaciones del porcentaje de las ancestrías Aymara, Mapuche y Europea para individuos de diferentes regiones de Chile.	46
V.1.C Estimación del riesgo genético para CG mediante un Puntaje de Riesgo Poligénico.	49
V.1.D Relación del puntaje de riesgo poligénico con la ancestría global Europea, Aymara y Mapuche.	51
V.1.E Evaluación de la asociación de alelos de riesgo de los SNPs seleccionados con su ancestría local.	53
V.2. Objetivo Específico 2	58
V.2.A Evaluación de la necesidad de controlar por estructuración genética y/o ancestría entre casos y controles chilenos de cáncer gástrico	58

V.2.B Genotipificación y control de calidad de las variantes genéticas seleccionadas para el objetivo específico 1, en casos de CG y controles de la población chilena.	60
V.2.C Selección de variantes genéticas de riesgo para desarrollar Cáncer Gástrico en individuos de la población chilena.	66
V.2.D Comparación de modelos logísticos de puntaje de riesgo y ancestría para el Cáncer Gástrico en individuos de la población chilena.	70
V.3. Objetivo Específico 3	76
V.3.A Faseo de los cromosomas en los individuos del set de datos de casos de GC para determinar su haplotipo parental	76
V.3.B Estimación de la ancestría local en regiones del genoma de individuos con fenotipo de caso de cáncer gástrico.	76
V.3.C Admixture Mapping para identificar regiones asociadas a Cáncer Gástrico por su ancestría.	79
V.3.D Evaluación enriquecimiento de funciones génicas en regiones genómicas identificadas por Admixture Mapping.	81
<b>VI. DISCUSIÓN</b>	<b>84</b>
<b>VII. CONCLUSIONES</b>	<b>90</b>
<b>VIII. REFERENCIAS</b>	<b>92</b>
<b>IX. PUBLICACIONES GENERADAS</b>	<b>102</b>
<b>X. ANEXOS</b>	<b>104</b>
X.1 Objetivo Específico 1	106
X.2 Objetivo Específico 2	114
X.3 Objetivo Específico 3	121

## ÍNDICE DE TABLAS Y FIGURAS

### Índice de Tablas

Tabla 1. Variantes genéticas de riesgo para el Cáncer Gástrico genotipificados con la técnica de GT-Seq.	45
Tabla 2. Proporción de ancestrías Africana (AFR), Europea (EUR), Aymara (AYM), Mapuche (MAP) y Amerindia (AMR) en 1691 individuos del proyecto ChileGenómico.	48
Tabla 3. Puntaje de Riesgo Poligénico por macrozonas de Chile en una muestra de 1691 participantes del proyecto de ChileGenómico.	50
Tabla 4. Poder del estudio para un análisis de asociación con 300 casos y 300 controles con distintas frecuencias alélicas y valor de Odds Ratios (razón de momios).	61
Tabla 5. SNPs evaluados en las muestras caso-control genotipificados mediante el arreglo Global Screening (GSA).	62
Tabla 6. Análisis de asociación de las variantes genéticas de riesgo para el desarrollo del Cáncer Gástrico en la muestra caso-control de 317 controles y 301 casos de Cáncer Gástrico.	62
Tabla 7. Análisis de asociación de las variantes genéticas de riesgos para el desarrollo del Cáncer Gástrico en la muestra caso-control de 317 controles y 143 casos de Cáncer Gástrico tipo Difuso y 156 casos de Cáncer Gástrico tipo Intestinal.	63
Tabla 8. Modelos logísticos de riesgo basal y de ancestría con sus efectos de ajuste para la muestra de 308 controles y 294 casos de Cáncer Gástrico.	67
Tabla 9. Modelos logísticos que incluyen variables genéticas de riesgo para la muestra caso-control de 308 controles y 294 casos de Cáncer Gástrico.	68
Tabla 10. Comparación de los modelos logísticos analizados mediante un análisis anidado realizado con la prueba de la razón de verosimilitudes (LRT).	69
Tabla 11. Modelos logísticos de riesgo que incluyen el Puntaje de Riesgo Poligénico (PRS) y la ancestría global Mapuche (MAP) para la muestra caso-control de 308 controles y 294 casos de Cáncer Gástrico.	72



Tabla 12. Comparación de los modelos logísticos analizados mediante un análisis anidado realizado con la prueba de la razón de verosimilitudes (LRT).	73
Tabla 13. Genes presentes en las regiones cromosómicas asociadas a ancestría Amerindia identificadas con el análisis de solo casos de Admixture Mapping.	82
Tabla 14. Genes que se encuentran en regiones cromosómicas asociadas a la ancestría Amerindia, por Admixture Mapping, relacionados al cáncer.	83

## Índice de Figuras

Figura 1. Proporción de ancestrías Europea, Aymara y Mapuche por macrozonas del país.	48
Figura 2. Puntaje de Riesgo Poligénico en las poblaciones ancestrales Europea (EUR), Asiática del Este (EAS), Africana (AFR), Aymara (AYM), Mapuche (MAP) y población mestiza chilena del proyecto de ChileGenómico (CLG).	50
Figura 3. Relación entre el Puntaje de Riesgo Poligénico y las Ancestrías Europea, Amerindia, Aymara o Mapuche.	52
Figura 4. Proporción de ancestrías Europea (EUR), Amerindia (AMR) y Africana (AFR) para los alelos de riesgo de seis SNPs.	54
Figura 5. Proporción de cromosomas que portan un alelo de riesgo con ancestría local Europea en la muestra de ChileGenómico (CLG) y su muestra de referencia Europea (IBS) para seis SNPs de riesgo de Cáncer Gástrico.	55
Figura 6. Proporción de cromosomas que portan un alelo de riesgo con ancestría local Amerindia en la muestra de ChileGenómico (CLG) y su muestra de referencia Amerindia (AMR) para seis SNPs de riesgo de Cáncer Gástrico.	56
Figura 7. Proporción de cromosomas que portan un alelo de riesgo con ancestría local Africana en la muestra de ChileGenómico (CLG) y su muestra de referencia Africana (YRI) para seis SNPs de riesgo de Cáncer Gástrico.	57
Figura 8. Estimaciones de las Ancestrías Globales por individuo en la muestra caso-control de Cáncer Gástrico.	59

Figura 9. Relación entre los ORs extraídos desde la literatura para los SNPs seleccionados en el Objetivo 1 y los ORs calculados en la muestra caso-control (OR Chileno).	64
Figura 10. SNPs con diferencias significativas en la frecuencia de sus alelos de riesgo entre controles y casos de Cáncer Gástrico.	65
Figura 11. Predicción lineal de la interacción entre terciles de ancestría con cuartiles de Puntaje de Riesgo Poligénico.	74
Figura 12. Relación entre el Puntaje de Riesgo Poligénico y la ancestría Mapuche en 317 controles (izquierda) y 301 casos de Cáncer Gástrico (derecha).	75
Figura 13. Proporción de ancestrías Europea (EUR), Amerindia (AMR) y Africana (AFR) para los alelos de riesgo de cinco SNPs en 316 controles y 301 casos de Cáncer Gástrico.	78
Figura 14. Admixture Mapping de regiones cromosómicas de ancestría Amerindia con un análisis de Solo-Casos para el Cáncer Gástrico.	80
Figura 15. Predicciones de efectos de 14550 variantes genéticas anotadas en ENSEMBL para las regiones cromosómicas asociadas a ancestría Amerindia identificadas con el análisis de solo casos de Admixture Mapping.	82

## Índice de Tablas Suplementarias

Tabla Suplementaria 1. Resultados del meta-análisis realizado por el grupo de Mocellin et al (2015): Variantes genéticas asociadas significativamente con el riesgo de cáncer gástrico con un nivel de evidencia intermedio o alto.	105
Tabla Suplementaria 2. Región de nacimiento de los 1691 individuos analizados que pertenecen al proyecto ChileGenómico.	106
Tabla Suplementaria 3. Resumen edad y sexo de los 1691 individuos analizados que pertenecen al proyecto ChileGenómico.	106
Tabla Suplementaria 4. Proceso de selección de SNPs de riesgo de Cáncer Gástrico para su genotipificación por medio de la técnica GT-Seq.	109
Tabla Suplementaria 5. Características de edad y sexo de 317 controles y 301 casos de Cáncer Gástrico.	114
Tabla Suplementaria 6. Proporción del genoma que pertenece a las	118

ancestrías Mapuche, Aymara, Europea, Africana y Asiática de 317 controles y 301 casos de Cáncer Gástrico.

Tabla Suplementaria 7. Frecuencias de ancestrías locales Europea (EUR) o Amerindia (AMR) de los alelos de riesgo de cinco SNPs. 121

Tabla Suplementaria 8. Nombre de los genes encontrados por la herramienta web PANTHER dentro de las regiones genómicas asociadas a la ancestría Amerindia identificadas por AM 125

Tabla Suplementaria 9. Nombre de las vías de señalización encontradas por la herramienta web PANTHER dentro de las regiones genómicas asociadas a la ancestría Amerindia identificadas por AM 125

## **Índice de Figuras Suplementarias**

Figura Suplementaria 1. Primeros 10 países con la mayor tasa de mortalidad cruda para el Cáncer Gástrico para ambos sexos, todas las edades en el 2018. 104

Figura Suplementaria 2. Regresiones de las ancestrías globales Europea, Africana, Aymara, Mapuche y Amerindia estimadas en 1465 individuos del proyecto de ChileGenómico con 137 Marcadores Informativos de Ancestría (AIMs) genotipificados con KASPar con 150 AIMs genotipificados por GT-Seq. 107

Figura Suplementaria 3. Regresiones entre la ancestría global estimada con el programa ADMIXTURE y ancestría global estimada desde el total de la ancestría local inferida con RFMix para las 420 muestras del proyecto de ChileGenómico. 108

Figura Suplementaria 4. Frecuencia de las proporciones de las proporciones de ancestrías globales Europea, Aymara y Mapuche de 1691 individuos del proyecto de ChileGenómico. 110

Figura Suplementaria 5. División de Chile por macrozonas 111

Figura Suplementaria 6. Puntaje de Riesgo Poligénico (PRS) en las distintas macrozonas del País. 112

Figura Suplementaria 7. Representación de cromosomas mestizos con sus regiones asociadas a las ancestrías Africana: AFR, Ancestría Europea: EUR (azul), Ancestría Amerindia: AMR (verde), y Ancestría Desconocida: UNK (negro). 113

Figura Suplementaria 8. Flujo de trabajo del control de calidad realizado a los datos de la muestra caso-control obtenidos con el microarreglo Global Screening (GSA).	114
Figura Suplementaria 9. Estimaciones de las Ancestrías Globales Europea (CEU e IBS), Africana (YRI), Asiática (CHB), Aymara (PEP) y Mapuche (MAP) en la muestra caso-control de Cáncer Gástrico (GSA).	115
Figura Suplementaria 10. Error de Validación Cruzada para las estimaciones de ancestrías globales de 329 controles y 316 casos de Cáncer Gástrico.	116
Figura Suplementaria 11. Error de Validación Cruzada para las estimaciones de ancestría de 317 controles y 301 casos de Cáncer Gástrico.	117
Figura Suplementaria 12. Análisis de Componentes Principales (PCA) en la muestra de 317 controles y 301 casos de Cáncer Gástrico.	119
Figura Suplementaria 13. Frecuencia de los alelos de nueve SNPs que no tienen diferencias significativas entre controles y casos de Cáncer Gástrico.	120
Figura Suplementaria 14. Número de Individuos (N) necesarios para identificar una variante significativa, en un estudio con un poder del 80% y un valor de p de significancia $< 0.001$ , por el riesgo (odd ratio) de desarrollar la enfermedad en un análisis de solo casos de admixture mapping.	122
Figura Suplementaria 15. Número de Individuos (N) necesarios para identificar una variante significativa con un poder del 80% y un $p < 0.001$ de acuerdo al riesgo en un análisis de admixture mapping con un estudio caso-control	123
Figura Suplementaria 16. Gráfico cuantil-cuantil (Q-Q) de los valores de p observados en el análisis de solo casos de admixture mapping para la ancestría Amerindia y los valores de p esperados.	124

## LISTA DE ABREVIATURAS

Cáncer Gástrico	CG
Puntaje de Riesgo Poligénico	PRS
Odd Ratio	OR
Polimorfismos de nucleótido único	SNP
Marcadores Informativos de Ancestría	AIMs
Admixture Mapping	AM
Estudios de Asociación de Genoma Completo	GWAS
Genotyping in thousand sequencing	GT-Seq
Global Screening Array	GSA
Número de poblaciones ancestrales	K
Análisis de Componente Principales	PCA
Desequilibrio de Ligamiento	LD
Criterio de Información Akaike	AIC
ChileGenomico	CLG
Europea	EUR
Europea Ibérica	IBS
Africana	AFR/YRI
Asiática del Este	EAS
Amerindia	AMR
Aymara	AYM
Mapuche	MAP
Prueba de razón de verosimilitud	LRT

Valor de p	p
Coefficiente de Determinación	R <sup>2</sup>
Análisis de Varianza	ANOVA
Intervalo de confianza	IC
Puntaje Z (Z-score)	Statistic
Genomic Regions Enrichment of Annotations Tools	GREAT
Protein ANalysis THrough Evolutionary Relationships	PANTHER
	R

## RESUMEN

El cáncer gástrico (CG) es la principal causa de muerte oncológica en Chile, no tiene síntomas específicos y tiene una tasa de supervivencia a 5 años de hasta un 20% debido a su diagnóstico en estadios tardíos. La tasa de mortalidad de CG en Chile tiene una distribución heterogénea, pudiendo ser relacionada a factores genéticos como la ancestría. Los chilenos tienen principalmente tres componentes ancestrales, Africano, Europeo y Amerindio, los cuales pueden tener distinta susceptibilidad a algunas enfermedades. Dado que la asociación de SNPs con riesgo a CG puede ser diferente en chilenos, evaluamos si las asociaciones reportadas en la literatura generalizan a la población chilena y si su asociación con el fenotipo depende de la ancestría.

Para determinar si el riesgo predicho de CG se asocia a las diferencias de ancestría en la población chilena, se genotipificaron 150 Marcadores Informativos de Ancestría y 13 SNPs de riesgo de CG en 1691 individuos del proyecto ChileGenómico mediante secuenciación dirigida. A cada individuo se les calculó un Puntaje de Riesgo Poligénico (PRS) en base al número de alelos de riesgo y se les estimó las proporciones de las ancestrías Europea, Africana, Aymara y Mapuche. Los resultados revelaron que el PRS está asociado positivamente con la ancestría Mapuche ( $\beta_1$  ( $\beta_1$ ) = 0.344 y  $p < 0.01$ ) y negativamente con la Europea ( $\beta_1$  = -0.392 y  $p < 0.01$ ).

Para evaluar si las asociaciones genéticas publicadas para CG de los SNPs seleccionados son reproducibles en la población chilena, se estudió una muestra chilena de 301 casos y 317 controles genotipificada con el arreglo Global Screening Array. Se analizaron 11 SNPs mediante modelos logísticos que incluyeron la edad, sexo y cualquier posibilidad de estructuración poblacional en la muestra. Se observó

que tres de los 11 SNPs seleccionados son predictores en los modelos logísticos de ajuste, pero solamente rs2294008 es significativo (OR: 1.36, CI: 1.05-1.78, p: 0.023). Además se confirmó una relación positiva entre el PRS y la ancestría Mapuche ( $\beta_1 = 0.6811$ ,  $p < 0.01$ ).

Para determinar si existen loci de riesgo en regiones genómicas con origen Amerindio que no hayan sido previamente asociadas a CG, se realizó un análisis de Admixture Mapping en los 301 pacientes chilenos de CG. Se identificaron tres regiones cromosómicas sugerentes de asociación: 8p23.1, 13q33 y 20q13.32 ( $p=0.00489$  para los tres loci), que contienen en conjunto 26 genes. De estos, solo cuatro genes han sido relacionados al CG (*GNAS*, *STX16-NPEPL1*, *MIR296* y *MIR298*) y otros cuatro han sido relacionados a otros tipos de cáncer (*LINKC0028*, *BLK*, *MTMR9* y *TDH*).

Nuestros resultados confirman la asociación de tres loci de riesgo para CG en pacientes chilenos y sugieren la existencia de loci en tres regiones cromosómicas de origen amerindio que no han sido previamente asociadas al riesgo de padecer CG en pacientes chilenos. Por lo tanto, proponemos realizar estudios con mayor número de casos de CG chilenos para confirmar y mapear a mayor resolución la asociación de estas regiones cromosómicas, de manera de identificar los genes candidatos de modificar este fenotipo de riesgo asociado a ancestría.



## ABSTRACT

Gastric cancer (GC) is the main cause of oncological death in Chile, it does not have specific symptoms and has a 5-year survival rate of up to 20% due to its diagnosis in late stages. Mortality rate of GC in Chile has a heterogeneous distribution, and can be related to genetic factors such as ancestry. Chileans have mainly three ancestral components, African, European and Amerindian, which may have different susceptibility to some diseases. Given that the association of SNPs with GC risk may be different in Chileans, we evaluate whether the associations reported in the literature generalize to the Chilean population and whether their association with the phenotype depends on ancestry.

To determine whether the predicted GC risk is associated with ancestry differences in the Chilean population, 150 Informative Ancestry Markers and 13 GC risk SNPs were genotyped in 1691 individuals from the ChileGenomico project using directed sequencing. A Polygenic Risk Score (PRS) was calculated for each individual based on the number of risk alleles and the proportions of European, African, Aymara and Mapuche ancestry were estimated. The results revealed that PRS is positively associated with the Mapuche ancestry ( $\beta_1$  ( $\beta_1$ ) = 0.344 and  $p < 0.01$ ) and negatively with the European ( $\beta_1$  = -0.392 and  $p < 0.01$ ).

To evaluate if the genetic association with CG can reproduce the published results in the Chilean population, a Chilean sample of 301 cases and 317 controls were genotyped with the Global Screening Array. Eleven SNPs were analyzed using logistic models that accounted for age, sex, and any possibility of population structuring in the sample. It was observed that three of the 11 selected SNPs are predictors in the logistic adjustment models, but only rs2294008 has significance (OR: 1.36, CI: 1.05-1.78,  $p$ : 0.023). Besides it was confirmed a positive relationship

between PRS and Mapuche ancestry was confirmed ( $\beta_1 = 0.6811$ ,  $p < 0.01$ ).

To determine if there are risk loci in genomic regions of Amerindian origin that have not previously been associated with GC, an Admixture Mapping analysis was performed in the 301 Chilean GC patients whose genomes were typed by DNA microarrays. Three chromosomal regions suggestive of association ( $p = 0.00489$ ) were identified: 8p23.1, 13q33 and 20q13.32, which contain 26 genes. Of these, only four genes have been related to GC (*GNAS*, *STX16-NPEPL1*, *MIR296* and *MIR298*) and another four have been related to other types of cancer (*LINKC0028*, *BLK*, *MTMR9* and *TDH*).

Our results confirm the association of three risk loci for GC in Chilean patients and suggest the existence of loci in chromosomal regions of Amerindian origin that have not previously been associated with the risk of suffering from GC in Chilean patients. Therefore, we propose to carry out studies with a greater number of Chilean GC cases to confirm and map the association of these chromosomal regions at a higher resolution, in order to identify the candidate genes that modify this risk phenotype association with ancestry.

## I. INTRODUCCIÓN

### I.1. Epidemiología

Durante el año 2018, el Cáncer Gástrico (CG) sumó más de un millón de casos nuevos y fue la causa de 783.000 muertes en el mundo, ubicándose en el sexto lugar para la tasa de incidencia y cuarto para la tasa de mortalidad<sup>1</sup>. Se presenta mayormente en hombres<sup>2</sup> y tiene sintomatología inespecífica, por ejemplo dispepsia y pérdida de peso en estadios tempranos<sup>3,4</sup>, por lo que este cáncer es diagnosticado mayormente en estadios tardíos reflejándose en una tasa de supervivencia a cinco años cercana a un 20%<sup>5,6</sup>. Su tratamiento depende de la clasificación TNM del informe histopatológico. Si es identificado en los estadios Ib-III el tratamiento principal es la gastrectomía, en cambio, si se encuentra en el estadio IV, se realiza un tratamiento paliativo de quimioterapia<sup>3,7,8</sup>.

Se ha descrito una asociación entre incidencia de este cáncer y la etnicidad, ya que Europa central y del este, Este de Asia y América del Sur son los continentes que tienen las tasas de incidencia y mortalidad más altas a nivel mundial<sup>9</sup>, del mismo modo que tienen las mayores tasas de mortalidad<sup>10</sup>. Globalmente, Japón tiene las mayores tasas crudas de incidencia (90.8/100.000 habitantes) y mortalidad (38.2/100.000 habitantes), seguido por China y Lituania (Figura Suplementaria 1)<sup>1</sup>. En Latinoamérica, los países que tienen altas tasas de mortalidad para el CG se encuentran principalmente en la región de Los Andes, donde se ubica Colombia, Costa Rica, Perú, Chile y Ecuador<sup>11</sup>. Chile ocupa el octavo lugar a nivel mundial, para las tasas de incidencia y mortalidad, tiene la mayor tasa de mortalidad en Sudamérica, con 19.1 cada 100.000 habitantes, 26.0 cada 100.000 habitantes en hombres y 12.4 cada 100.000 habitantes en mujeres<sup>1,12,13</sup>. La tasa de mortalidad es heterogénea a lo largo del país, siendo más alta en las regiones del Maule, del

Bío-bío, La Araucanía y Los Ríos, con 26,47, 23.4, 25.4 y 24.06 cada 100.000 habitantes, respectivamente; mientras que las tasas más bajas están en la región de Arica y Parinacota y de Antofagasta, con 9.87 y 9.02 cada 100.000 habitantes<sup>13</sup>.

## **I.2. Aspectos histológicos**

El principal tipo de cáncer gástrico es el adenocarcinoma, mientras que otras formas como los tumores linfoproliferativos y neuroendocrinos son menos frecuentes<sup>3</sup>. Desde un punto de vista histológico, la clasificación de Lauren divide al adenocarcinoma gástrico en Difuso e Intestinal<sup>14</sup>. El primero, ha disminuido su incidencia a nivel mundial, pero sigue siendo el más común en lugares con alta incidencia de CG. Está precedido por una secuencia de lesiones histológicas que llevan al desarrollo de este tipo de adenocarcinoma; tiene un tejido bien diferenciado y posee estructuras glandulares que asemejan a la mucosa intestinal. El segundo, es más agresivo y tiene un peor pronóstico que el intestinal. Es diagnosticado en pacientes más jóvenes (40-60 años), su tejido es pobremente diferenciado, compuesto de células tumorales solitarias o muy poco cohesivas y no posee formaciones glandulares<sup>15</sup>.

Otra clasificación es diferenciando los tumores entre cuatro tipos moleculares de acuerdo a resultados de secuenciación masiva de tumores: i) Positivo para el virus Epstein-Barr, ii) Inestabilidad de microsatélites, iii) Genómicamente estable, y iv) Inestabilidad cromosómica, los cuales corresponden aproximadamente al 9%, 22%, 20% y 50% de los casos respectivamente<sup>16</sup>. La mayoría de los tumores positivos al virus Epstein-Barr están localizados en el fondo o el cuerpo del estómago, se presentan mayormente en hombres, tienen hipermetilación en el promotor de *CDKN2A* y también pueden presentarla en los genes *JAK2* y *PD-L1/2*. La inestabilidad de microsatélite se caracteriza por un sistema deficiente en la

reparación del DNA, el cual es causado frecuentemente por un evento epigenético, como la hipermetilación en la región promotora del gen *MLH1*, y además se relaciona al CG tipo intestinal. El tipo genómicamente estable se presenta en un alto porcentaje en el CG de tipo difuso, y presenta mutaciones de *RHOA* o fusiones relacionadas a la familia de *RHOA* que activan proteínas GTPasa. Finalmente, la inestabilidad cromosómica está relacionada en los adenocarcinomas de la unión gastroesofágica, y está caracterizada por aneuploidías, cambios estructurales de cromosomas y mutaciones que involucran genes promotores y supresores de tumor<sup>10,17</sup>.

### **I.3. Factores de Riesgo**

#### **I.3.1. Ambientales**

Se ha reportado que el consumo de carnes rojas o procesadas son factores de riesgo de desarrollo de CG, mientras que las carnes blancas lo disminuyen<sup>18</sup>. El consumo de alcohol aumenta el riesgo en un 3% por cada 10 gramos de consumo diario en hombres, pero no se encontró esta relación en mujeres<sup>19</sup>. El bajo consumo de frutas y verduras, tiene un riesgo relativo de 5.5 comparado con aquellos que tienen un alto consumo<sup>20</sup>. Algunos factores protectores son una dieta alta en frutas y verduras, especialmente cítricos, el ejercicio (aeróbicos y de fuerza), suplementos antioxidantes tales como las vitaminas A, C y E, y el selenio<sup>21</sup>. Respecto al último, se ha encontrado que personas con una alta exposición a selenio tienen una menor incidencia y tasa de mortalidad<sup>22</sup>. Por último, otros factores son un IMC mayor a 25 kg/m<sup>2</sup> tiene un OR (Odd Ratio) de 1.13 de desarrollar este cáncer<sup>21</sup>; y el tabaquismo el cual es factor de riesgo para el GC y se ha estimado que un 22% de los casos de CG en el Reino Unido fueron relacionados a este hábito<sup>23</sup> (OR = 1.69)<sup>24</sup>.

Otros factores de riesgo para el CG son los infecciosos, que son dos: el Virus Epstein-Barr e *Helicobacter pylori*. El primero, se estima que está presente en cerca del 10% de los casos<sup>25</sup>, y se ha observado que los tejidos de CG que son positivos a este virus tienen menos niveles de apoptosis que aquellos que son negativos a este virus, y además tienen una mayor tasa de supervivencia<sup>26</sup>. Respecto al segundo, aunque tiene un riesgo atribuible al CG de un 74%, solo entre 1-3% de los pacientes infectados desarrollan CG<sup>27</sup>. El desarrollo del CG por medio de infección por esta bacteria está altamente descrito, iniciando la cascada de eventos con metaplasia, displasia y adenocarcinoma<sup>3,10</sup>.

### **I.3.2. Moleculares y genéticos**

Según su modo de herencia, el CG se divide en hereditario y esporádico. El primero corresponde al 1-3% de los casos y son enfermedades de herencia autosómica dominante. Estos incluyen a: CG difuso hereditario, Poliposis adenomatosa familiar y Síndrome Peutz-Jeghers, y están asociados a los genes de *E-cadherina*, *Poliposis Adenomatosa Coli* y *LKB1*, respectivamente<sup>6</sup>. El CG esporádico no se puede relacionar a un solo factor genético porque corresponde a una enfermedad de herencia compleja. Pocos estudios han calculado la heredabilidad del CG esporádico, sin embargo, en un estudio caso-control se encontró un valor de heredabilidad ( $h^2$ ) de 0.29 (Error Estándar: 0.157)<sup>28</sup>, lo cual es un nivel moderado de heredabilidad comparado con otros tipos de cáncer<sup>29</sup>.

Distintas variantes genéticas de riesgo han sido asociadas al CG. Un ejemplo para el CG hereditario son mutaciones en el gen de E-cadherina *CDH1*, el cual es un gen supresor de tumor, que se estima que cerca del 40% de los casos de HDCG presentan esta alteración germinal<sup>26</sup>. Por otro lado, para el CG esporádico se han identificado numerosos SNPs asociados al riesgo de desarrollar este cáncer y distintos genes. En el estudio realizado por Mocellin *et al*, recopilaron la información

de 2841 polimorfismos que se encontraban en 952 genes, de los cuales 156 variantes (101 genes) tenían una evidencia satisfactoria según el criterio de Venice (Tabla Suplementaria 1). Algunos de los polimorfismos identificados son rs2294008C>T, rs4072037C>T, rs2274223A>G<sup>30-32</sup> que están ubicados en los genes *PSCA*, *MUC1* y *PLCE1*, respectivamente. rs2294008C>T está asociado significativamente al CG difuso<sup>32-34</sup>; rs4072037C>T se ha descrito que tiene principalmente un efecto protector ante el CG<sup>31</sup>; y rs2274223A>G está asociado significativamente al aumento en el riesgo de desarrollar CG<sup>30</sup>. En Chile, se reportó que variantes que pertenecen a la vía de las MAP quinasas están asociadas con el CG, como rs3729932 (*RAF1*) y rs2283792 (*MAPK1*)<sup>35</sup>.

#### **I.4. Genética poblacional del Cáncer Gástrico**

##### **I.4.1 Aspectos Generales**

Las poblaciones se pueden agrupar según distintos factores culturales, fenotípicos y/o genéticos, por ejemplo: alimentación, estatura y ancestría, respectivamente. La ancestría está definida como la proporción del genoma de un individuo asociado a una o más poblaciones parentales y, según la clasificación de Linneaus, se divide principalmente en las poblaciones Europea, Africana, Asiática y Amerindia<sup>36</sup>. La ancestría se estima con Marcadores Informativos de Ancestría (AIMs), los cuales son polimorfismos con una gran diferencia de frecuencias entre las poblaciones ancestrales<sup>37</sup>. Para poder realizar las estimaciones de ancestría de un individuo se han realizado paneles que incluyen solo el ADN mitocondrial, el cromosoma Y o variantes autosómicas. El problema del primero y segundo es que tienen un sesgo porque representa solo una parte del linaje genético de un individuo, en cambio, el tercero provee mayor información genómica de la ancestría individual<sup>34</sup>.

#### I.4.2 América Latina y Chile

La población latinoamericana tiene ancestría mestiza, es decir, posee cromosomas mosaicos que derivan de más de una población ancestral<sup>37</sup>. Este proceso de mestizaje empezó a desarrollarse desde la época de la colonización Europea al continente americano, que trajo consigo familias Europeas y esclavos Africanos<sup>38</sup>. Además, la población latina tiene diferentes perfiles genéticos de ancestría Europea, Africana y Nativa Americana debido a que tienen una distinta historia de mestizaje<sup>39</sup>. Esto fue observado en un estudio de Homburger *et al* (2015) al realizar estimaciones de ancestría de individuos de distintos países de América del Sur. Reportaron que Perú es el país que tiene un mayor porcentaje de ancestría Nativa Americana y Argentina de Ancestría Europea<sup>40</sup>. Estos son ejemplos de la estructura subcontinental que existe entre los países de América Latina<sup>41</sup>. En Chile, se han realizado estudios para evaluar la estructura genética de la población. Uno de estos utilizó 40 AIMs y reportó que hay un promedio de 44.34% de ancestría Americana, 51.85% de ancestría Europea y 3.81% de ancestría Africana, además de mostrar que existe una subestructuración dentro del país<sup>42</sup>. En otro estudio realizado con individuos genotipados con un microarreglo por Eyheramendy *et al* (2015), se encontraron los mismos resultados para el promedio del país (ancestría Nativa Americana 44.74%, ancestría Europea 52.25%, ancestría Africana 3.01%). También estructuraron el país, en este caso en cinco zonas, debido a que observaron que en las zonas del centro (C1 y C2) hay un mayor porcentaje de ancestría Europea y en las zonas del sur (S1 y S2) un mayor porcentaje de ancestría nativa<sup>43</sup>.



### I.4.3 Relación de la ancestría con el cáncer

El meta-análisis de Mocellin *et al* (2015) resumió la evidencia de SNPs de riesgo para el CG esporádico con el fin de determinar si el riesgo SNPs pudiese estar relacionado a la etnicidad y características tumorales. Uno de los resultados más importantes que tuvieron fue que los SNPs de riesgo pueden dividirse entre aquellos que son de riesgo para Asiáticos, caucásicos y para ambos grupos<sup>32</sup>, lo que apoya la heterogeneidad tanto de la prevalencia como las tasas de incidencia y mortalidad observada en el mundo.

Por otro lado, se tienen antecedentes de distintos patrones de riesgo de cáncer entre los latinos, como los cáncer de mama, colorrectal y de pulmón que tienen una mayor incidencia que el cáncer de próstata<sup>44</sup>. Por ejemplo, en Chile se ha descrito que hay riesgo de cáncer de vesícula biliar asociado a la ancestría Mapuche, la cual es una subpoblación de la ancestría Amerindia en Chile<sup>45</sup>. Respecto al CG, en Perú ha sido asociada a la ancestría Amerindia<sup>46</sup>, y se ha encontrado una mayor incidencia y tasa de mortalidad en poblaciones indígenas como la Mapuche<sup>47</sup>. En otro estudio realizado en Perú se analizó la asociación de SNPs de riesgo previamente reportados para este cáncer en individuos que son casos de CG con un alto porcentaje de ancestría nativa americana y no se encontró una asociación<sup>48</sup>. Esto sugiere que es necesario realizar más estudios en poblaciones latinoamericanas para determinar si estos resultados son transversales o si varían por la heterogeneidad del mestizaje.

En Chile, Cannals *et al.* (2012) estudiaron la relación entre ancestría Mapuche y tasa de mortalidad en diferentes regiones de Chile. Observaron una correlación entre el aumento del porcentaje de ancestría Mapuche y el aumento de la tasa de mortalidad del CG ( $R^2$  de 0,558)<sup>49</sup>. De hecho, en la región de La Araucanía se observa un alto porcentaje de ancestría Mapuche (31,12%), al mismo tiempo que

existe una alta tasa de mortalidad para el CG (25,66 cada 100.000 habitantes). En conjunto, esto sugiere que podría existir una relación entre la ancestría Mapuche y la tasa de mortalidad del CG.

### **I.5. Riesgo Genético**

Se ha descrito que el riesgo de algunos SNPs podrían asociarse de manera dependiente con la población de estudio. Algunos ejemplos son: i) el haplotipo de *IL1B* -1464G/-511C/-31T que ha sido reportado como riesgo para el CG en asiáticos, mientras que el haplotipo *IL1B* -1464C/-511T/-31C es de riesgo para los caucásicos<sup>33</sup>; ii) *IL1B* rs419598-T aumenta el riesgo para CG tipo intestinal en caucásicos, mientras que en la asiática no tiene este efecto<sup>25</sup>; y iii) *IL-10* tiene un efecto que se conoce como “flip de alelos”, es decir, el riesgo que genera un alelo en una población, en otra población lo produce su otro alelo<sup>6,33</sup>. De hecho, se realizó un metanálisis donde agruparon SNPs según su riesgo para la población asiática o caucásica, dejando un precedente que demuestra de manera empírica que existen distintas variantes genéticas de riesgo para distintas poblaciones<sup>32</sup>. Otro ejemplo es el caso del gen *MUC1*, ya que se ha descrito que su alelo C (rs4072037) es protector en la población asiática, pero en la población caucásica no se observó este efecto<sup>49</sup>. Esto sugiere que las variantes genéticas pueden agruparse según su riesgo poblacional, lo que podría asociarse a la heterogeneidad en la distribución de la tasa de mortalidad CG tanto en América Latina como en el mundo.

Estimar la probabilidad de susceptibilidad de un individuo a una enfermedad (predicción de riesgo) es una ayuda para tomar decisiones clínicas, especialmente en el contexto de detección y prevención de condiciones que se desarrollan en la adultez. El cáncer esporádico es una enfermedad de herencia compleja la que se caracteriza por no seguir patrones clásicos de la herencia mendeliana, es decir, que

los fenotipos no dependen exclusivamente de la variación de un alelo, sino más bien por el efecto aditivo de múltiples variantes en el fenotipo más la interacción con factores ambientales. Hoy, la predicción se realiza principalmente con las variables de edad, sexo y factores de estilo de vida como el consumo de alcohol y la actividad física, y el perfil genético podría ser un componente útil para complementar los modelos con variables genéticas para mejorar el manejo de problemas de la salud. De hecho, para algunas enfermedades ya se han realizado perfiles poligénicos para calcular un puntaje de riesgo poligénico que provee información de uso clínico y personal<sup>50</sup>. Por ejemplo, un estudio con un puntaje de riesgo con 45 SNPs encontró que este valor se encuentra aumentado (en una desviación estándar) en pacientes con enfermedad coronaria arterial de inicio temprano (<40 años), respecto a individuos normales y con inicio tardío (hombres  $\geq 55$  años, y mujeres  $\geq 65$  años) de esta enfermedad<sup>51</sup>. Otro estudio, de cáncer de mama en europeos, se evaluó el efecto combinado de 10 SNPs, previamente asociados, observando que sólo 7 de 10 estaban fuertemente asociados, y que su efecto combinado aumenta fuertemente el riesgo de desarrollar este cáncer<sup>52</sup>. Dado estos antecedentes, se están utilizando modelos poligénicos para predecir el riesgo de desarrollar una enfermedad mediante estimaciones del puntaje de riesgo poligénico (PRS). El PRS consiste en un valor para cuyo cálculo debe incluir el efecto en conjunto de distintas variantes que tienen una asociación significativa con la enfermedad de interés. El PRS es principalmente calculado como la suma del peso del número de alelos de riesgo que lleva un individuo, donde los alelos de riesgo y sus pesos están definidos por el loci y su efecto está medido con el efecto detectado en GWAS<sup>50</sup>.

Respecto al CG, se observó que un PRS calculado con tres SNPs estimó un riesgo elevado de 1.8 veces en los pacientes comparado con los controles<sup>53</sup>. Dado que en el estudio de Mocellin *et al* (2015) se agruparon variantes genéticas según su riesgo con la población caucásica o asiática, una forma más exacta de predecir el

riesgo es calcular un puntaje de riesgo poligénico construido con variantes genéticas identificadas o que ya estén asociadas con la población de evaluada<sup>54</sup>.

## **I.6. Búsqueda de marcadores genéticos de riesgo**

La ancestría global es el promedio de un componente ancestral en el genoma de un individuo, y esta estimación es un buen indicador del mestizaje debido a que el genoma de individuos mestizos está compuesto por bloques de ADN de orígenes ancestrales diferentes. Sin embargo, la ancestría en un locus particular no está necesariamente determinada por la ancestría global, porque la distribución de esta última no es homogénea a lo largo de todo el genoma. Es decir, los alelos de un locus puede ser cualquiera de las ancestrías globales de un individuo. La inferencia de la ancestría de un locus permite realizar correlaciones entre su variación de ancestría con la variación de un fenotipo.

El mapeo de múltiples genes es una herramienta útil para asociar marcadores genéticos con un fenotipo, como el CG, en enfermedades que son de herencia compleja. Una herramienta para realizar esto son los estudios de asociación del genoma completo (Genome-wide association Study / GWAS) que permiten estudiar la asociación del genotipo de miles de SNPs con un fenotipo<sup>55</sup>. Muchos GWAS para el CG se han llevado a cabo en la población Europea<sup>56</sup> y Asiática<sup>57</sup>, sin embargo, en nuestro país no se han realizado. Es por esta razón, que en esta tesis se utilizará una aproximación de 'Admixture mapping' (AM) como herramienta de mapeo que permite identificar genes en poblaciones mestizas evaluando si existe una asociación del fenotipo con la ancestría del locus en vez del genotipo. Esta estrategia se utiliza en poblaciones mestizas y cuando hay una diferencia en la prevalencia del fenotipo en estudio entre las poblaciones ancestrales<sup>58</sup>. Es por esta razón que la población chilena es una población ideal para realizar este análisis dado a que es una

población mestiza (principalmente Europea y Amerindia), y por otro lado existe una diferencia entre las tasa de incidencia y mortalidad entre nuestra población y nuestra población de referencia española. España tiene una tasa de incidencia de 7,79/100.000 habitantes y de tasa de mortalidad de 4,92/100.000 habitantes, las que son bajas comparada a las que tiene Chile, con 15,59/100.000 habitantes y 13.76/100.000 habitantes respectivamente<sup>1</sup>.

La metodología de AM ya ha permitido la identificación de regiones cromosómicas asociadas significativamente a un fenotipo en poblaciones mestizas como la latina y la afroamericana. Un ejemplo es un estudio realizado con individuos de Puerto Rico y México, que incluye 529 casos de asma y 347 controles donde se encontró variaciones genéticas en los loci 6q15 y 8q12 que estarían asociados a las población Nativa Americana y a la susceptibilidad al asma<sup>59</sup>. Otro AM para identificar loci asociados a diabetes tipo 2 realizado en mujeres afroamericanas, identificó las regiones genómicas 3q26 y 12q23, que presentaron un enriquecimiento de ancestría Africana, fueron asociadas con un mayor riesgo de Diabetes tipo 2<sup>60</sup>.

Dado a que los marcadores pueden encontrarse en regiones con distinto origen ancestral, para realizar asociaciones de variantes genéticas con el riesgo es necesario considerar la ancestría del locus. Por lo tanto, dado a que se tiene el antecedente que de existe una tendencia entre el porcentaje de ancestría Mapuche tanto con los egresos hospitalarios como con la tasa de mortalidad, realizar un estudio de AM permitirá identificar si existen variantes genéticas, ya descritas o no, en regiones genómicas asociadas a regiones cromosómicas de ancestría Amerindia.

## **II. HIPÓTESIS**

Los cromosomas de origen amerindio en los chilenos aportan mayor riesgo de desarrollar cáncer gástrico que aquellos de origen europeo.

## **III.OBJETIVO GENERAL**

Evaluar si existe asociación entre alelos de riesgo y la ocurrencia de cáncer gástrico para loci previamente reportados, y si esto está asociado con el origen ancestral de las regiones cromosómicas portadoras.

### **III.1 Objetivo específico 1**

Evaluar la asociación del riesgo poligénico de loci descritos en literatura para cáncer gástrico con la ancestría en la población chilena.

### **III.2 Objetivo específico 2**

Validar el valor predictivo de cada locus de riesgo en la población chilena y determinar si depende de la ancestría.

### **III.3 Objetivo específico 3**

Identificar regiones genómicas que se asocian a cáncer gástrico en pacientes chilenos mediante un enfoque de Admixture Mapping.

## IV. MATERIALES Y MÉTODOS

### IV.1. Selección de SNPs asociados a cáncer gástrico en distintas poblaciones

El proceso de selección de SNPs de riesgo (Tabla Suplementaria 4) para el CG se realizó en los siguientes pasos:

- A. Recepción de una lista de SNPs inicial entregada por el Dr. P. González, que cumplió con los criterios de estar validados y reportados como factores de riesgo asociados a este cáncer.
- B. Revisión de SNPs de CG en GWAS catalog (<https://www.ebi.ac.uk/gwas/>)<sup>61</sup> para complementar la lista inicial. Se encontraron 45 asociaciones, pero se repitieron algunos SNPs.
- C. Aquellos SNPs cuya asociación fue encontrada en más de una publicación que tenían se seleccionó el estudio según:
  - a. Población de estudio: se prefirieron los estudios realizados en poblaciones latinas. Si la publicación se llevó a cabo en otros grupos de estudio se eligió en primer lugar el realizado en poblaciones europeas, si no en población asiática.
  - b. Robustez del estudio: Se prefirió los Estudios de Asociación de Genoma Completo (GWAS) con dos fases, de descubrimiento y replicación, sino se utilizaron GWAS de una sola fase.
- D. Se ordenó la lista de SNPs por OR y luego por valor de p para incluir aquellos que presenten  $OR \geq 1.2$  u  $OR \leq 0.8$  y valor de  $p \leq 3E^{-8}$ . Se seleccionaron 22 SNPs.
- E. Se revisaron los alelos de los SNPs, para determinar si se habían publicado correctamente o se analizaron sus reversos complementarios.

- F. Se eliminaron aquellos SNPs que no fue posible diseñar su partidador (materiales y métodos: IV.2.1).
- G. Se eliminaron aquellos SNPs que se encuentran en desequilibrio de ligamiento ( $r^2 > 0.8$ ).
- H. Finalmente, se utilizaron 13 SNPs.

La lista final de las variantes seleccionadas se encuentran en la Tabla 1.

## IV.2. Genotipificación de las muestras

### IV.2.1 *Objetivo 1*

Se utilizaron 1691 individuos del proyecto de ChileGenómico (*FONDEF D10E1007*) que fueron reclutados en las comunas de Arica, Iquique, La Serena, Coquimbo, Santiago, Chillán, Temuco y Puerto Montt. Este grupo contiene individuos nacidos en todas las regiones del país (desde la región de Arica y Parinacota hasta la de Magallanes y Antártica Chilena), son mayores de 18 años, tienen una dirección permanente en el país, y sus padres nacieron en Chile. El resumen de la descripción de la muestra se presenta en Tablas Suplementarias 2 y 3. Los protocolos y el consentimiento informado fueron aprobados por el Comité de Ética e Investigación en Humanos de la Facultad de Medicina de la Universidad de Chile.

El DNA se extrajo desde muestras de sangre y saliva utilizando la metodología descrita en la publicación de Verdugo *et al*<sup>62</sup>. La genotipificación fue mediante la técnica de “Genotyping in thousand sequencing” (GT-Seq)<sup>37</sup>, en una plataforma Illumina (NextSeq 550), donde se obtuvieron los genotipos de 150 AIMs y 13 SNPs de riesgo para el CG. Esta técnica utiliza dos procesos de PCR, el primero para crear los amplicones de interés y el segundo para agregar dos índices para realizar la identificación de las muestras, un índice para identificar el número de placa y otro el



pocillo. Luego, se normaliza la cantidad de ADN y se revisan los tamaños de los amplicones en un TapeStation. Finalmente, se secuencian en una plataforma Illumina NextSeq 550.

#### IV.2.2 *Objetivo 2*

La muestra caso-control está compuesta por 301 casos de CG (156 tipo Intestinal y 143 tipo Difuso) y 317 controles de la población chilena. Los individuos identificados como casos, fueron obtenidos en los siguientes centros: Hospital Clínico de la Universidad de Chile, Hospital del Salvador, Hospital Barros Luco Trudeau, Hospital San Juan de Dios y Biobanco de Tejidos y Fluidos de la Universidad de Chile; y diagnosticados histopatológicamente con adenocarcinoma gástrico de acuerdo al informe de anatomía patológica. Aquellos que pertenecen al grupo control, fueron sometidos a una endoscopia digestiva alta por orden de su médico tratante en el Hospital Clínico de la Universidad de Chile, siendo excluidos aquellos con lesiones premalignas o sospecha de CG. El resumen de Edad y Sexo se encuentra en la Tabla Suplementaria 5. Las muestras fueron genotipificadas con un arreglo de Illumina llamado Global Screening Array (GSA) v1 en Erasmus MC, Países Bajos (FONDECYT 1151015, IP Patricio González). El estudio cuenta con la aprobación del Comité de Ética de Investigación en Seres Humanos de la Facultad de Medicina de la Universidad de Chile.

### IV.3. Control de Calidad

#### IV.3.1 *Objetivo 1*

Para obtener los genotipos de las muestra de ChileGenómico se aplicó un protocolo bioinformático realizado por Campbell *et al*<sup>63</sup> que fue adaptado y automatizado para el proyecto de ChileGenómico que realiza: i) control de calidad a las lecturas secuenciadas, las que son identificadas de lecturas de los locus respectivos, se descartan aquellos que están sobrerrepresentados; ii) eliminar locus cuyos partidores generan dímeros y heterodímeros de partidior; iii) asignación de lecturas por individuos utilizando la combinación de índices (placa y pocillo); iv) llamado de genotipos por pareo con sondas *in-silico*, lo que fue realizado con secuencia de 15 nucleótidos que incluye el SNP (una para cada alelo) y sus nucleótidos aledaños, junto con la información del partidior forward; v) los genotipos se transforman a formato Plink; y, finalmente, vi) .Con el programa Plink1.9<sup>64</sup> filtro por tasa de genotipificación (mantienen  $\geq 80\%$ ) y tasa de datos perdidos de SNPs e individuos (mantienen  $<20\%$ ).

#### IV.3.2 *Objetivo 2*

Se utilizó el programa Plink1.9 para realizar los filtros de: Tasa de pérdida por SNP e individuo ( $< 5\%$ ), equilibrio de Hardy-Weinberg ( $<1E^{-6}$ ). Se eliminaron aquellos individuos cuyo el sexo fenotípico y genotípico no concordaban. Además, se eliminaron los SNPs ambiguos (aquellos que tienen alelos complementarios, es decir A y T o C y G) con el programa SNPFlip. Se extrajeron las variantes autosomales y se eliminaron individuos emparentados con el programa Relatedness Estimation in Admixture Populations (REAP)<sup>65</sup>. El proceso del control de calidad de las muestras se presenta en la Figura Suplementaria 2.

## IV.4. Estimación de ancestrías globales

### IV.4.1 *Objetivo 1*

Las estimaciones de ancestría de los individuos mestizos de ChileGenómico se realizó con el programa ADMIXTURE<sup>66</sup>, que entregó las proporciones del genoma de ascendencia Europea, Africana, Mapuche y Aymara. Para realizar estas estimaciones se necesitó un grupo de individuos de ancestría conocida (grupo de referencia), el cual estuvo compuesto por un subset de los datos genómicos de individuos que pertenecen al proyecto 1000 Genomas Fase 3<sup>67</sup>, el que incluyó 30 Europeos y 30 Africanos. Además, este grupo incluyó 110 Aymaras y 54 Mapuches, que fueron genotipificados con el arreglo Axiom LAT1.

### IV.4.2 *Objetivo 2*

Para realizar las estimaciones de los componentes ancestrales de cada individuo se utilizó el programa ADMIXTURE<sup>66</sup>, para el cual se necesita un grupo de individuos de ancestría conocida. Este grupo está compuesto con individuos que representan las poblaciones Europea, Africana y Asiática del Este, 40 individuos de cada uno, obtenidos del proyecto de 1000 Genomas Fase 3. Para la ancestría Amerindia se utilizaron 40 individuos que corresponden a la población Aymara de Perú y 54 individuos de la población Mapuche (Huilliche y Pehuenche)<sup>62</sup>. El grupo de ancestría desconocida son 618 individuos chilenos mestizos que pertenecen a la muestra caso-control. El programa realiza varias estimaciones de ancestrías globales por individuos con un distinto número de poblaciones ancestrales (K), y con el

análisis de validación cruzada se identifica el mejor K, eligiendo aquél que tiene el menor error de predicción<sup>66</sup>.

#### IV.5. Puntaje de Riesgo Poligénico

El puntaje de riesgo poligénico (PRS) considera un modelo aditivo para los alelos de riesgo. Para su cálculo se usan los valores de OR obtenidos de la literatura (Tabla 1), con la ecuación 1<sup>53</sup>:

$$PRS = \ln \left( \prod_{i=1}^n OR_i^x \right)$$

ecuación (1)

donde  $\Pi$  es la multiplicatoria de los  $OR_i$ ;  $OR_i$  es el efecto asociado al  $i$ -ésimo SNP de riesgo obtenido desde la literatura;  $n$  el número de SNPs seleccionados; y  $x$  el número de alelos de riesgo (0, 1, 2) de cada variante por individuo.

La obtención del PRS de las poblaciones de referencia se realizó con datos de individuos de poblaciones Europeas, Africanas y Asiáticas obtenidos desde el proyecto 1000 Genomas Fase 3<sup>67</sup>, 30 individuos de cada una, y 110 Aymaras y 54 Mapuches, que fueron genotipificados en arreglos Axiom LAT1.

#### IV.6. Estimación y análisis de ancestría local

Para evaluar las frecuencias de las ancestrías para los alelos de riesgo a nivel país se utilizaron 506 individuos que pertenecen al proyecto de ChileGenómico que fueron genotipificados por la plataforma Axiom LAT-1<sup>62</sup>, y para el estudio caso-control los individuos del proyecto FONDECYT genotipificados con el arreglo GSA. En

ambos casos, se infirieron los haplotipos (fasearon) con el programa SHAPEIT2, utilizando como grupo de referencia todos los individuos que pertenecen al proyecto 1000 Genomas Fase 3 y los mismos grupos de individuos de ancestría Mapuche y Aymara utilizados para la estimación de la ancestría global. Las inferencias de ancestrías Europea, Africana y Amerindia para cada locus se realizaron con el programa RFMix<sup>68</sup>. El control de calidad fue una correlación entre la ancestría global estimada con el software ADMIXTURE con la estimación de ancestría global calculada con el total de loci de cada población ancestral inferidos por ancestría local (Figuras Suplementarias 3 y 4). Se utilizó un programa “collapse\_ancestry.py”<sup>69</sup> para realizar el colapso de los SNPs adyacentes que tienen la misma ancestría, creando ventanas de ancestría dentro de un cromosoma. Luego, se obtuvo la ancestría local de los SNPs de riesgo con el programa BEDTools (<https://bedtools.readthedocs.io/en/latest/>), tanto para las muestras de ChileGenómico como para las muestras caso-control.

Los arreglos AXIOM LAT-1 y GSA no incluyen todos los SNPs seleccionados para su genotipificación, así que para el siguiente análisis se utilizó un subgrupo. Los SNPs encontrados fueron:

- a) Arreglo AXIOM LAT-1: rs4072037, rs2976392, rs2274223, rs1695, rs3765524 y rs3781264.
- b) Arreglo GSA: rs1695, rs2274223, rs4072037, rs6916921 y rs9820958.

Con estos datos, se comparó la frecuencia del alelo de riesgo del SNP ubicado en regiones cromosómicas de ancestría Europea o Africana o Amerindia en la muestra de ChileGenómico con su frecuencia en la población ancestral Ibérica, Yoruba (ambos obtenidos del proyecto 1000 Genomas) o Amerindia (obtenidos de proyectos del laboratorio) respectivamente. La prueba estadística utilizada fue un test de Chi-cuadrado.

La manera de estimar las frecuencias de la ancestría para un locus de riesgo está en la ecuación 2.

$$\% \text{ Cromosomas de Riesgo Población}_i = \frac{\text{Número de Cromosomas de Riesgo Población}_i}{\text{Total de Cromosomas Población}_i} \quad \text{Ecuación (2)}$$

donde  $i$  es el componente ancestral Europeo, Amerindio o Africano.

#### IV.7. Análisis de Asociación

Se calculó el poder de los análisis para un estudio que contiene 300 casos y 300 controles fue calculado con la herramienta GWAS Power Calculator<sup>70</sup>, suponiendo un modelo aditivo, prevalencia de CG de 0.0003 y un valor de  $p = 0.0045$  (corresponde a  $0.05/11$  (número de hipótesis independientes) = 0.0045). El poder de un estudio mayor a 80%, se obtiene para el OR de 1.8 desde una frecuencia del alelo de riesgo de 0.2, y el OR de 2, desde la frecuencia del alelo de riesgo de 0.1 (Tabla 4).

Primero, se revisó si las variantes genéticas seleccionadas (IV.1), asociadas a CG en distintas poblaciones, fueron genotipificadas en el arreglo GSA. Utilizando las herramientas web LD-proxy<sup>71</sup> y Haploreg v4.1<sup>72</sup> se buscaron variantes para reemplazar aquellas que no estaban presente en el arreglo GSA, con el criterio de disequilibrio de ligamiento de  $r^2 > 0.8$ . Una vez reemplazadas las variantes, se realizó el control de calidad a los genotipos obtenidos, cuyos pasos se muestran en la Figura Suplementaria 8. Finalmente, con el grupo final de SNPs, se realizó el análisis de asociación con el programa Plink1.9, módulo --assoc, el cual compara las frecuencias alélicas del alelo 1 (menor) entre casos y controles con un test de Chi-cuadrado, y se ajustan los datos con un post-test de Bonferroni. Finalmente, también entrega un valor de Odd Ratio respecto al alelo 1, que representa la chance de tener ese alelo y ser caso.

## IV.8. Análisis de Componentes Principales

Se realizó un Análisis de Componentes Principales (PCA) con el programa SMARTPCA<sup>73</sup> para determinar si existen grupos de individuos diferenciados genéticamente y si dichos grupos, de existir, pudieran estar diferencialmente representados entre casos y controles. De ser así, la estructuración genética podría causar falsas asociaciones entre el fenotipo caso/control y marcadores genéticos evaluados. El análisis de PCA permite representar los individuos según un número reducido de variables ortogonales, las cuales a su vez son combinaciones lineales de las variables presentes en el set de datos, en este caso, los genotipos del microarreglo. Luego del control de calidad y filtrado de SNPs por LD (se dejaron aquellas que tienen un  $r^2$  igual a 0.1) quedaron 58.000 variantes genéticas para este análisis. Entonces, luego de analizar las 58.000 variables independientes (SNPs) consideramos las 20 dimensiones principales, las cuales en conjunto explicaron un 8.6% de la varianza genética total, y se graficaron los primeros 10 componentes en el programa R con el paquete ggplot2.

## IV.9. Evaluación de Riesgo Genético

### IV.9.1 Selección de modelos genéticos de riesgo

Para seleccionar las variables genéticas que integran un modelo logístico que tiene un mejor ajuste a la varianza de los datos de la muestra caso-control, se realizaron búsquedas exhaustivas con el paquete glmulti de R, el cual realizó un ranking de modelos según el Criterio de Información Akaike (AIC) y la Devianza. Se agruparon las variables con distintos criterios, estos fueron ancestrías globales o SNPs de riesgo o ambos. Los grupos evaluados fueron:

Basal:  $Y \sim \text{Edad} + \text{Sexo}$

Ancestría:  $Y \sim \text{Edad} + \text{Sexo} + \text{MAP}$

SNPs:  $Y \sim \text{Edad} + \text{Sexo} + \text{SNPs}$

SNPs + Ancestría:  $Y \sim \text{Edad} + \text{Sexo} + \text{MAP} + \text{SNPs}$

SNPs x Ancestría:  $Y \sim \text{Edad} + \text{Sexo} + \text{MAP} \times (\text{selección de SNPs en el modelo "SNPs"})$

donde MAP es el porcentaje de ancestría Mapuche, "x" es el componente de interacción y SNPs los 11 SNPs de riesgo evaluados. Luego de realizar cinco búsquedas exhaustivas y tener los mejores modelos de riesgo para cada caso, se evaluó la significancia de cada variable en cada modelo mediante la función drop1 en R, usando la prueba de razón de verosimilitud (LRT).

#### IV.9.2 *Puntaje de Riesgo Poligénico*

Se evaluaron otros modelos de predicción de riesgo que incluyeron las variables de edad, sexo, PRS y/o el porcentaje de ancestría Mapuche del individuo (MAP). El PRS fue calculado con los 11 SNPs evaluados en el análisis de asociación. Los modelos evaluados fueron:

Basal:  $Y \sim \text{Edad} + \text{Sexo}$

Ancestría:  $Y \sim \text{Edad} + \text{Sexo} + \text{MAP}$

PRS:  $Y \sim \text{Edad} + \text{Sexo} + \text{PRS}$

PRS + MAP:  $Y \sim \text{Edad} + \text{Sexo} + \text{PRS} + \text{MAP}$



PRS x MAP: Y ~ Edad + Sexo + PRS x MAP

La significancia de las variables de cada modelo se evaluó mediante la función drop1 en R, usando LRT. Finalmente, se compararon los modelos con el modelo basal, que no incluye variables genéticas, con LRT para determinar si los modelos que tienen variables genéticas tienen una mejor predicción de ser caso de CG.

#### IV.10. Admixture Mapping

##### IV.10.1 Cálculo de poder de la muestra

El número de individuos necesarios para el análisis de Admixture mapping, con un estudio de solo casos, se calculó con las ecuaciones 3 y 4:

$$n \approx \left( \frac{Z_{1-\alpha} + Z_{1-\beta}}{\lambda} \right)^2 \frac{1}{V_0} \quad \text{ecuación (3)}$$

y

$$V_0 = \frac{1}{2} \theta(1 - \theta) \quad \text{ecuación (4)}$$

donde  $\lambda$  es el logaritmo natural del Odd Ratio del SNP y  $\theta$  la proporción de ancestría global, en este caso Amerindia<sup>74</sup>.

En el análisis de sólo casos (Figura Suplementaria 14) en una población con 40% de ancestría Amerindia, poder del 80% y para un valor de significancia de  $10^{-3}$  (valor de p utilizado en un estudio de de AM en latinos con tres poblaciones

ancestrales)<sup>75</sup> es posible identificar asociaciones con un OR mínimo de 2 usando 300 casos. En cambio, el análisis caso-control con una muestra de 300 casos y 300 controles necesita un OR mínimo de 3.7 (Figura Suplementaria 15). Por lo tanto, se decidió realizar un análisis de sólo casos.

#### IV.10.2 *Análisis de Admixture mapping en regiones asociadas a Cáncer Gástrico por su ancestría*

El método de Admixture mapping se realiza en poblaciones con cromosomas mestizos para correlacionar loci de una ancestría particular con un fenotipo. En este caso, se quiere evaluar si el CG está correlacionado con regiones cromosómicas de ancestría Amerindia.

Esto se realizó mediante un análisis de sólo casos<sup>74</sup>, que incluyó 301 individuos. Para comparar la frecuencia de ancestría local para un locus determinado respecto de la ancestría global se calculó el z-score<sup>76</sup> (ecuación 5):

$$Z_C(t) = \frac{\hat{\Pi}_d(t) - \hat{\Pi}_d(0.5)}{\sigma \hat{\Pi}_d(t)} \quad \text{ecuación (5)}$$

donde  $\hat{\Pi}_d(t)$  es la proporción estimada de la ancestría local Amerindia en el locus  $t$ ;  $\hat{\Pi}_d(0.5)$  representa la proporción esperada de ancestría global Amerindia de la muestra de casos, y  $\sigma \hat{\Pi}_d(t)$  el error estándar de la ancestría Amerindia del locus  $t$ .

Esto se traduce en la diferencia genética entre el porcentaje de ancestría global de la muestra menos la ancestría local en un locus, dividido por el error

estándar del locus. Un z-score positivo indica que a una alta ancestría Amerindia en ese locus particular está asociado con un mayor riesgo de CG; a su vez, un z-score negativo, indica que menor ancestría Amerindia está asociada al riesgo de CG. El valor de p se obtuvo con el Z-score, y para el análisis de solo-casos se utilizó un valor de p de  $\leq 0.001$ <sup>75</sup>.

#### IV.11. Identificación de enriquecimiento de funciones génicas en regiones genómicas

Para tener una mayor probabilidad de éxito en la anotación, se utilizó Haploreg v4.1 para identificar variantes genéticas presentes en las regiones genómicas identificadas. La herramienta web SNPnexus<sup>77</sup> se utilizó para anotar las variantes genéticas presente en las regiones cromosómicas asociadas a la ancestría Amerindia identificadas por AM. Esta herramienta utiliza distintas bases de datos, como Cosmic, ClinVar, GAD y 1000 Genomas.

Por otro lado, se utilizó un programa online que se llama Genomic Regions Enrichment of Annotations Tool (GREAT)<sup>78</sup> para asignar un significado biológico a las regiones no codificantes analizando la codificación de genes cercanos. Al mismo tiempo, se utilizó otra herramienta web Protein ANalysis THrough Evolutionary Relationships (PANTHER)<sup>79</sup> que clasifica proteínas y sus genes. Este análisis lo realiza de acuerdo a las siguientes clasificaciones: 1) Familia y subfamilias de proteínas; 2) Función molecular; 3) Procesos biológicos; y 4) Vías de señalización.

#### IV.12. Estadística

Para comparar los valores de PRS o estimaciones de ancestrías globales dentro del país en el Objetivo 1, se dividió el país en macrozonas. Las macrozonas incluyen un subgrupo de regiones políticas del país, y se dividen en 5: 1) Norte Grande: Arica y Parinacota (XV), Tarapacá (I) y Atacama (II); 2) Norte Chico: Copiapó (III) y Coquimbo (IV); 3) Centro: Valparaíso (V), Metropolitana (XIII o RM), Libertador General Bernardo O'Higgins (VI), Maule (VII), Biobío (VIII); 4) Sur: Araucanía (IX), Los Ríos (XIV) y Los Lagos (X); y finalmente, 5) Austral: Aysén del General Carlos Ibáñez del Campo (XI) y Magallanes y la Antártida Chilena (XII). La Macrozona Austral no fue utilizada para los análisis estadísticos debido al bajo número de muestras (10 individuos). Aquellos individuos que pertenecen a los grupos de Nacidos en el extranjero y No sabe/No responde (NS/NR) tampoco fueron incluidos. Para comparar las macrozonas se utilizó un ANOVA y un post-test de Duncan.

Las regresiones entre el PRS y la ancestría se realizaron con el programa R. Para determinar la significancia del modelo se calculó el valor de  $R^2$  y el valor de  $p$  ( $< 0.05$ ) del coeficiente  $\beta_1$  con el Estadístico F.

## V. RESULTADOS

### V.1. Objetivo Específico 1

Evaluar la asociación del riesgo poligénico de loci descritos en literatura para cáncer gástrico con la ancestría en población chilena.

#### V.1.A *Selección SNPs asociados a cáncer gástrico en distintas poblaciones.*

Seleccionamos 20 SNPs que están asociados al riesgo de desarrollar CG en otras poblaciones, que tienen valores de OR menor o igual a 0.8 o mayor o igual a 1.2, y valor de  $p \leq 3E^{-8}$ . Se diseñaron partidores para 18 SNPs y fueron genotificados por GT-Seq. Luego de filtrar los SNPs por desequilibrio de ligamiento quedaron 13 SNPs para el análisis (Tabla 1 y Tabla Suplementaria 4).

Cromosoma	rsID	Gen	Alelo de Riesgo	OR	p	Frecuencia Alelo de Riesgo	Porcentaje de Genotipado	PMID
1	rs4072037	MUC1	C	1.32	6.59E-08	0.68	75.51	26129866
1	rs140081212	GBAP1	G	1.27	7.90E-10	0.37	66.30	26098866
1	rs3762272	PKLR	T	1.41	1.42E-08	0.19	99.59	25731870
1	rs80142782	ASHL1	C	1.61	1.70E-19	1.00	100.00	26701879
2	rs3834129	CASP8	AGTAAG	1.37	3.48E-04	0.67	99.07	25731870
3	rs3087465	TGFBR2	G	1.64	1.54E-11	0.72	99.54	25731870
5	rs13361707	PRKAA1	T	1.41	7.60E-29	0.77	99.77	22037551
5	rs7712641	LOC105379076	T	1.19	1.00E-11	0.54	99.83	26701879
5	rs2910164	MIR146A	C	1.36	2.30E-02	0.26	99.13	27267319
8	rs2976392	PSCA	A	1.62	4.22E-15	0.48	93.63	25731870
10	rs2274223	PLCE1	G	1.57	3.78E-20	0.20	99.36	21642993
11	rs1695	GSTP1	G	1.19	7.88E-05	0.42	97.05	25731870
22	rs738722	CHEK2	C	1.3	1.41E-08	0.32	99.83	20729852

**Tabla 1.** Variantes genéticas de riesgo para el Cáncer Gástrico genotificados con la técnica de GT-Seq. Resumen de los datos más importantes para su selección y la fuente de los ORs y valores de p. OR: Odds Ratio (razón de momio), p: valor de p, PMID: identificador en la base de datos de PubMed.

### *V.1.B Estimaciones del porcentaje de las ancestrías Aymara, Mapuche y Europea para individuos de diferentes regiones de Chile.*

Para validar la técnica de GT-Seq, se realizaron correlaciones de las estimaciones de ancestría global de 1465 individuos calculadas con 131 marcadores informativos de ancestría (AIMs) genotipificados con GT-Seq con las estimaciones de ancestría calculadas con 137 AIMs genotipificados con la técnica KASPar. Esta última, fue considerada como estándar de comparación cuyos resultados fueron reportados anteriormente<sup>62</sup>.

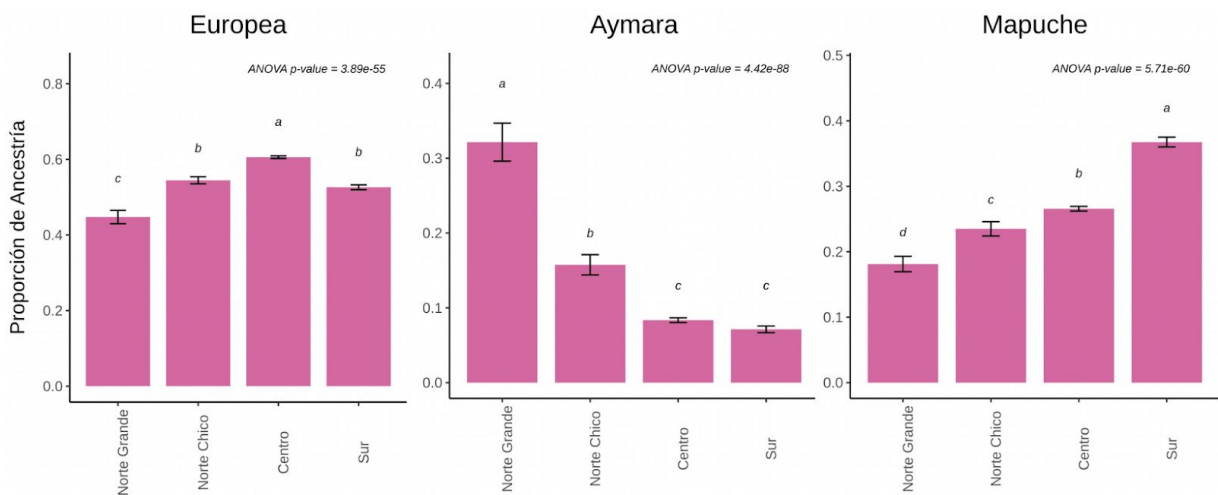
Las correlaciones entre las estimaciones de ancestrías<sup>62</sup> fueron de 0.73, 0.89, 0.60, 0.73 y 0.91, para las ancestrías Africana, Europea, Aymara, Mapuche y Amerindia, respectivamente (Figura Suplementaria 2). Aunque se espera que el coeficiente de correlación entre ambas plataformas fuese 1, debido a que se utilizaron casi los mismos SNPs para realizar la estimación de ancestría, esto probablemente no se observó debido a: i) la estocasticidad del algoritmo del programa ADMIXTURE que realiza los cálculos para obtener las estimaciones de ancestría global, y ii) se utilizó un el panel de individuos de ancestría conocida (referencia) distinto para estimar la ancestría de los individuos mestizos (muestra ChileGenómico)<sup>80</sup>. Respecto al segundo punto, el algoritmo de ADMIXTURE utiliza las frecuencias alélicas de las variantes presentes en los individuos de una población parental para inferirla en un individuo mestizo, por lo que si no se utilizaron los mismos individuos de referencia para realizar las estimaciones de ancestría global, un mismo individuo mestizo puede tener distintas estimaciones de ancestrías. Sin embargo, los valores de coeficiente de determinación ( $R^2$ ) indican que las ancestrías estimadas por ambas plataformas tienen una buena correlación. La tasa de concordancia para los genotipos obtenidos con ambas técnicas para los individuos en común es de 0.98, lo que valida la técnica de GT-Seq para realizar las estimaciones de ancestrías globales.

Las frecuencias de los porcentajes de ancestría globales estimados para la muestra de ChileGenómico se encuentran en la Figura Suplementaria 5. Los promedios fueron de las siguientes proporciones: 0.57 Europea, 0.38 Amerindia y 0.04 Africana. En Chile, el componente Amerindio se puede estratificar principalmente entre Aymara y Mapuche, representadas por las poblaciones indígenas que llevan ese nombre. Los promedios de las proporciones de ancestría para estos componentes son: 0.1 para Aymara y 0.28 para Mapuche (Tabla 2).

Para comparar si existen diferencias en las estimaciones de ancestría dentro del país, se dividió el país por macrozonas (Figura Suplementaria 6). En la Figura 1 se muestran las proporciones de las ancestrías Europea, Aymara y Mapuche estratificadas por Zonas. Se encontró un mayor porcentaje de: i) Ancestría Europea en la Zona Central ( $p < 0.001$ ); ii) Ancestría Aymara en la Zona del Norte Grande ( $p < 0.001$ ); y iii) Ancestría Mapuche en la Zona Sur ( $p < 0.001$ ).

	AFR	EUR	AYM	MAP	AMR
Promedio	0.043	0.572	0.102	0.282	0.384
Error Estándar	0.001	0.003	0.003	0.003	0.003

**Tabla 2.** Proporción de ancestrías Africana (AFR), Europea (EUR), Aymara (AYM), Mapuche (MAP) y Amerindia (AMR) en 1691 individuos del proyecto ChileGenómico.



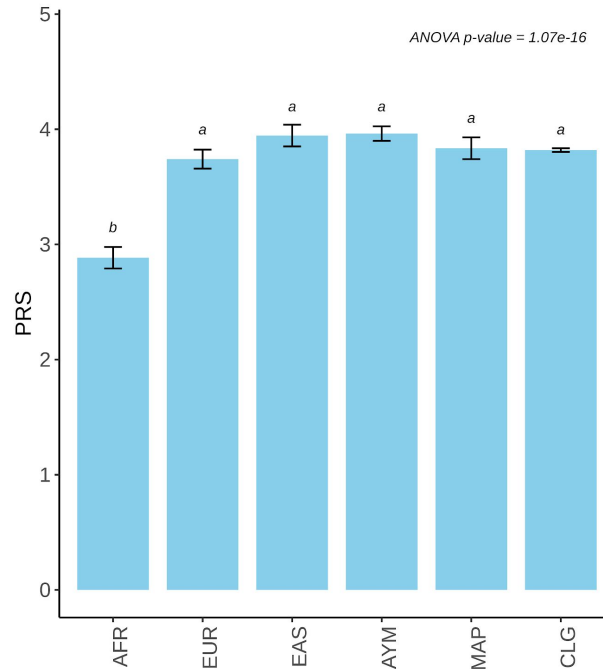
**Figura 1.** Proporción de ancestrías Europea, Aymara y Mapuche por macrozonas del país. Datos expresados en promedio  $\pm$  error estándar. Prueba de ANOVA y Post-hoc de Duncan. Barras etiquetadas con la misma letra identifican promedios sin diferencia significativa ( $p < 0.05$ ).



### *V.1.C Estimación del riesgo genético para CG mediante un Puntaje de Riesgo Poligénico.*

Para comparar el PRS entre las poblaciones Europea, Asiática, Africana, Aymara, Mapuche y de ChileGenómico, se utilizaron 10 de los 13 SNPs seleccionados (Figura 2) porque fueron los que se encontraron en común entre los datos de los genotipos de estas muestras. Se observó que el promedio más elevado de PRS se encuentra en la población Aymara y su puntaje no es significativamente distinto al de las poblaciones Mapuche, Europea, Asiática ni con la muestra de ChileGenómico. Todas las poblaciones nombradas anteriormente son significativamente distintas con la población Africana.

Para el análisis dentro del país, se calculó el PRS con 11 de los 13 SNPs (tasa de genotipificado > 80%). Se dividió el país en macrozonas y se calculó el promedio de los PRS de los individuos en cada macrozona. El promedio del PRS del país es de 3.98, y el de las zonas norte grande, norte chico, centro y sur es de 3.93, 4.04, 3.97 y 4.01, respectivamente. Para evaluar si existen diferencias en el PRS entre las zonas, se realizó una ANOVA. No hubo diferencias significativas (Tabla 3, Figura Suplementaria 7).



**Figura 2.** Puntaje de Riesgo Poligénico en las poblaciones ancestrales Europea (EUR), Asiática del Este (EAS), Africana (AFR), Aymara (AYM), Mapuche (MAP) y población mestiza chilena del proyecto de ChileGenómico (CLG). Datos expresados en promedio  $\pm$  error estándar. Prueba de ANOVA y Post-test de Duncan. Barras etiquetadas con la misma letra identifican promedios sin diferencia significativa ( $p < 0.05$ ).

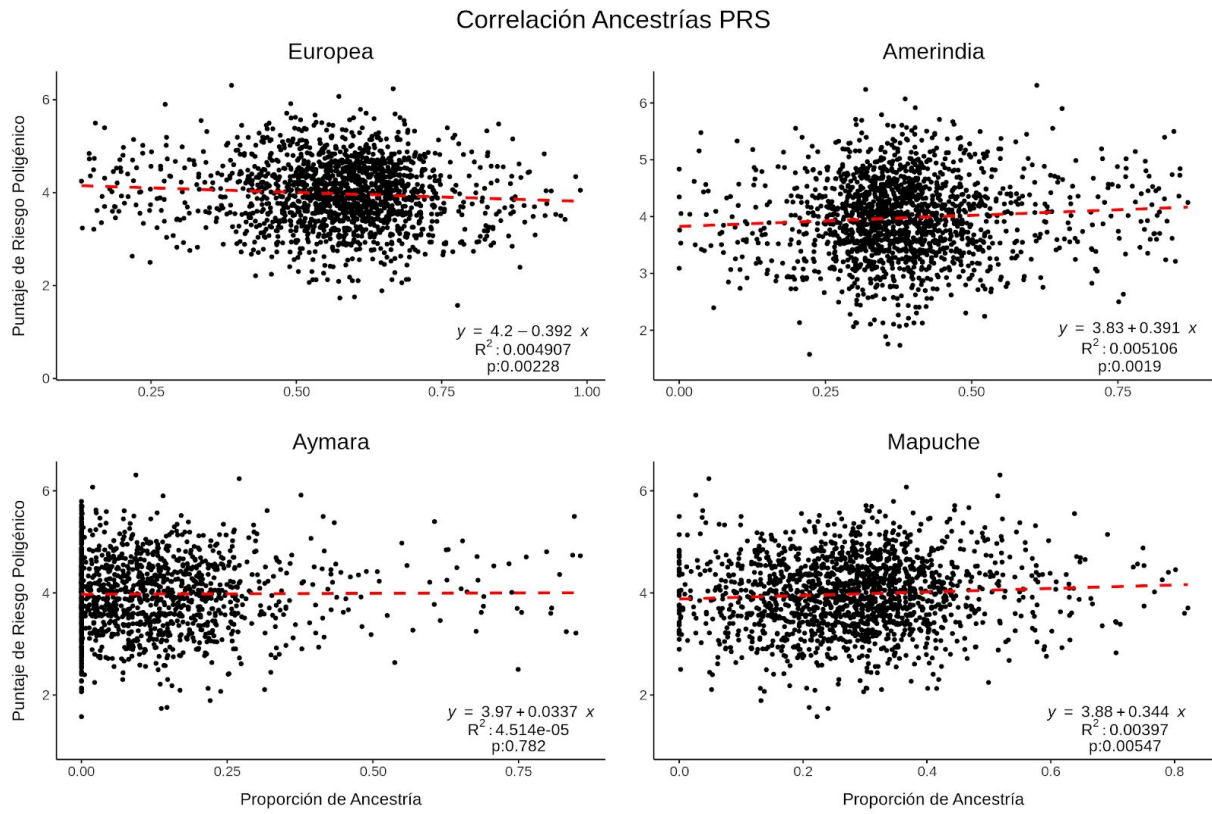
	Norte Grande	Norte Chico	Centro	Sur	País
Promedio PRS	3.93	4.04	3.97	4.01	3.98
Error Estándar	0.06	0.08	0.02	0.03	0.02

**Tabla 3.** Tabla 3. Puntaje de Riesgo Poligénico por macrozonas de Chile en una muestra de 1691 participantes del proyecto de ChileGenómico. Prueba de ANOVA ( $p=0.488$ ).

*V.1.D Relación del puntaje de riesgo poligénico con la ancestría global Europea, Aymara y Mapuche.*

La Figura 3 muestra las correlaciones entre las proporciones de ancestrías Europea, Aymara y Mapuche con el PRS. Se encontró que la ancestría Europea tiene una relación negativa con el PRS, ( $\beta_1$  ( $\beta_1$ ) = -0.392 y  $p < 0.01$ ); y una relación positiva con las ancestrías Amerindia ( $\beta_1$  = 0.391 y  $p < 0.01$ ) y Mapuche ( $\beta_1$  = 0.344 y  $p < 0.01$ ). Esto quiere decir, el OR del PRS para la ancestría Europea es de 0.676, para la ancestría Amerindia es de 1.478 y para la ancestría Mapuche 1.41.

Estos resultados indican que hay un aumento en riesgo de 1.411 entre una proporción de 0 a 1 de ancestría Mapuche, esto quiere decir que por cada tercil de esta ancestría el riesgo aumenta en 1.12. Para el caso de la ancestría Amerindia el aumento del riesgo entre la proporción de 0 a 1 es de 1.478, es decir, por cada tercil de ancestría el riesgo aumenta 1.38. Finalmente, cuando un individuo tiene una proporción de 1 de ancestría Europea, el riesgo se reduce en 0.676, implicando que el riesgo disminuye en 0.878 por cada tercil de ancestría.



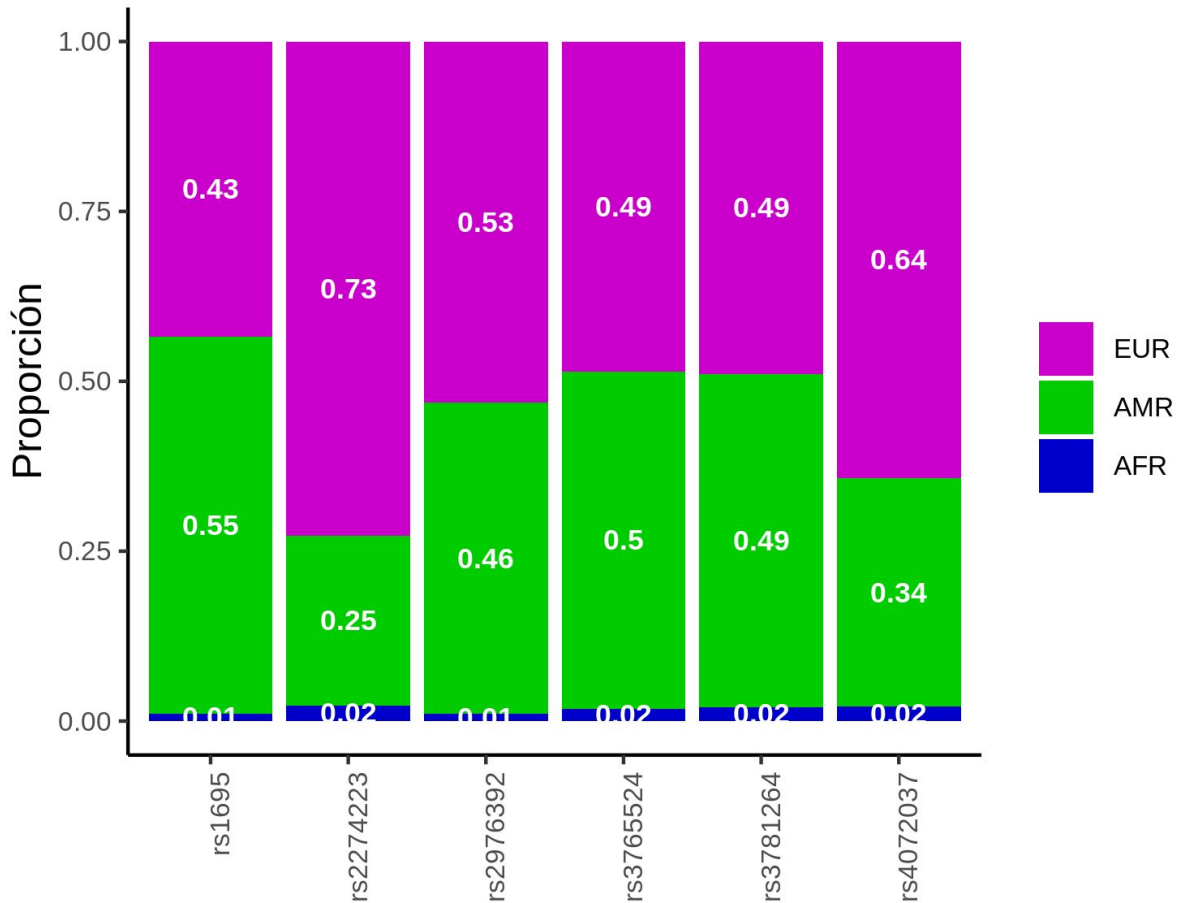
**Figura 3.** Relación entre el Puntaje de Riesgo Poligénico y las Ancestrías Europea, Amerindia, Aymara o Mapuche. Modelo de regresión lineal:  $y$ ; Coeficiente de determinación:  $R^2$  y valor de  $p$   $\beta_1$  (Estadístico F).

### *V.1.E Evaluación de la asociación de alelos de riesgo de los SNPs seleccionados con su ancestría local.*

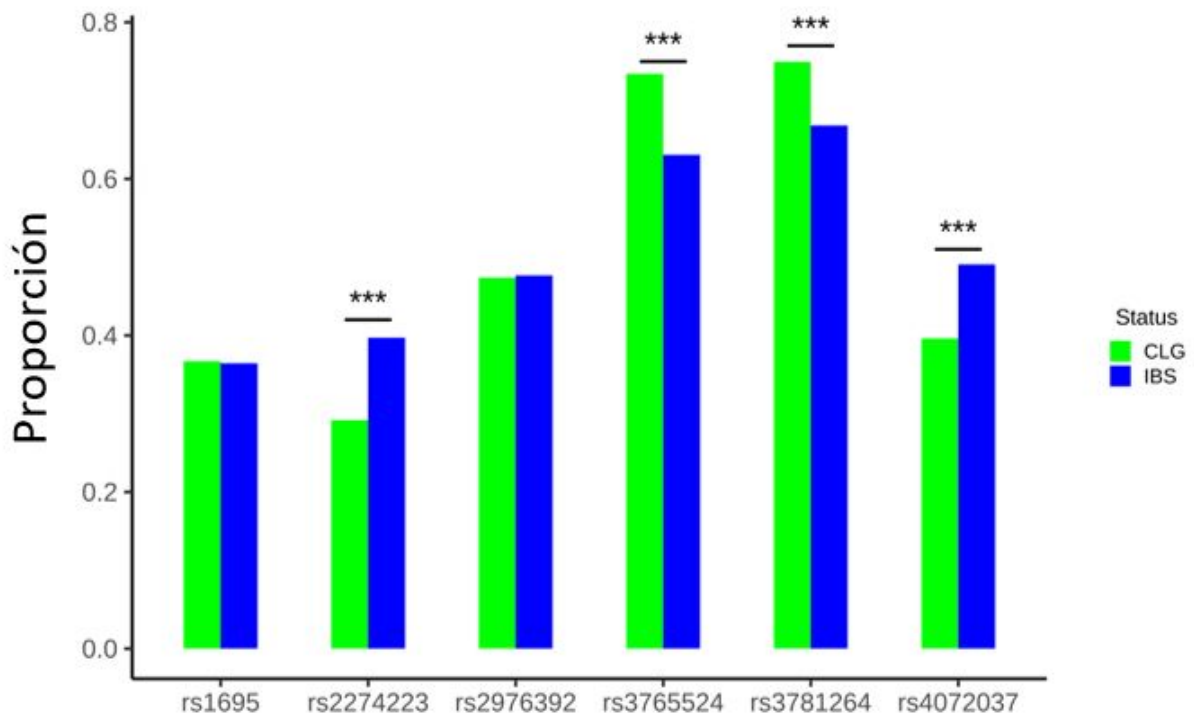
Como control de calidad, calculamos la correlación entre las estimaciones de ancestrías globales obtenidas en el punto 1.B, y el promedio de las ancestrías locales en todas las ventanas genómicas de cada individuo. La correlación observada fue  $R^2 > 0.85$  para cada ancestría (Figura Suplementaria 3).

Una manera de visualizar el análisis de ancestría local se presenta en las Figura Suplementaria 8, donde se observan cromosomas mestizos de dos individuos. Este análisis se realizó con individuos del proyecto de ChileGenómico que fueron genotificados con el arreglo Axiom LAT1, para poder inferir la ancestría de cada locus. Dentro de este arreglo, se encontraron seis de los 13 SNPs seleccionados. La proporción de ancestría local para los alelos de riesgo de los seis SNPs se muestra en la Figura 4, y se observa que los alelos de riesgo presentes en nuestra población provienen mayormente de las ancestrías Europea y Amerindia (Figura 4).

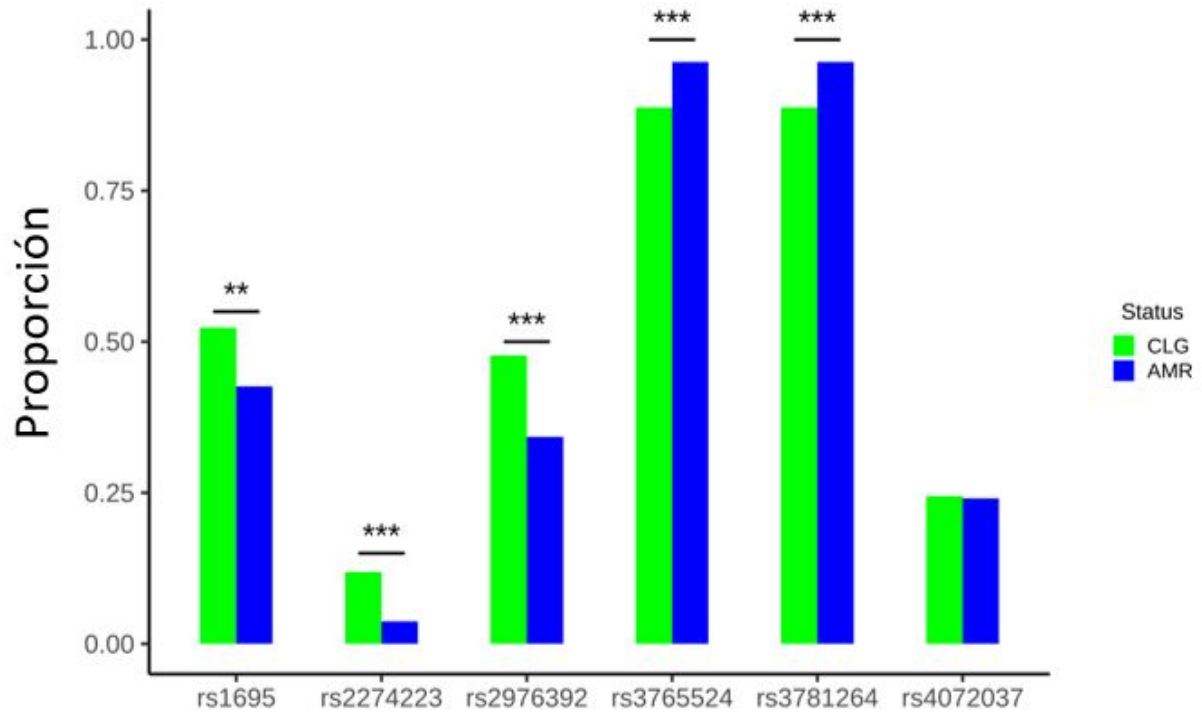
Para seis de los 13 SNPs seleccionados se comparó la frecuencia de su alelo de riesgo ubicado en regiones cromosómicas de ancestría Europea o Amerindia o Africana en la muestra de ChileGenómico con su frecuencia en la población ancestral Ibérica, Americana y Yoruba, obtenidos del proyecto 1000 Genomas, respectivamente. Se encontró que existen diferencias significativas de las frecuencias de los alelos de riesgo en las ancestrías: i) Europea: rs2274223, rs3765524, rs3781264 y rs4072037 (Figura 5); ii) Amerindia: rs1695, rs2274223, rs2976392, rs3765524 y rs3781264 (Figura 6); y iii) Africana: rs2274223, rs2976392, rs3765524 y rs4072037 (Figura 7). Esto indica que hubo procesos post-mestizaje colonial europeo-amerindio-africano que cambiaron las frecuencias alélicas asociadas a la ancestría de estos SNPs, posiblemente deriva génica u otros.



**Figura 4.** Proporción de ancestrías Europea (EUR), Amerindia (AMR) y Africana (AFR) para los alelos de riesgo de seis SNPs. La proporción fue calculada como el número de cromosomas portadores de alelos de riesgo con una ancestría local Europea o Amerindia o Africana dividido el número total de cromosomas con ancestría inferida en esta región.

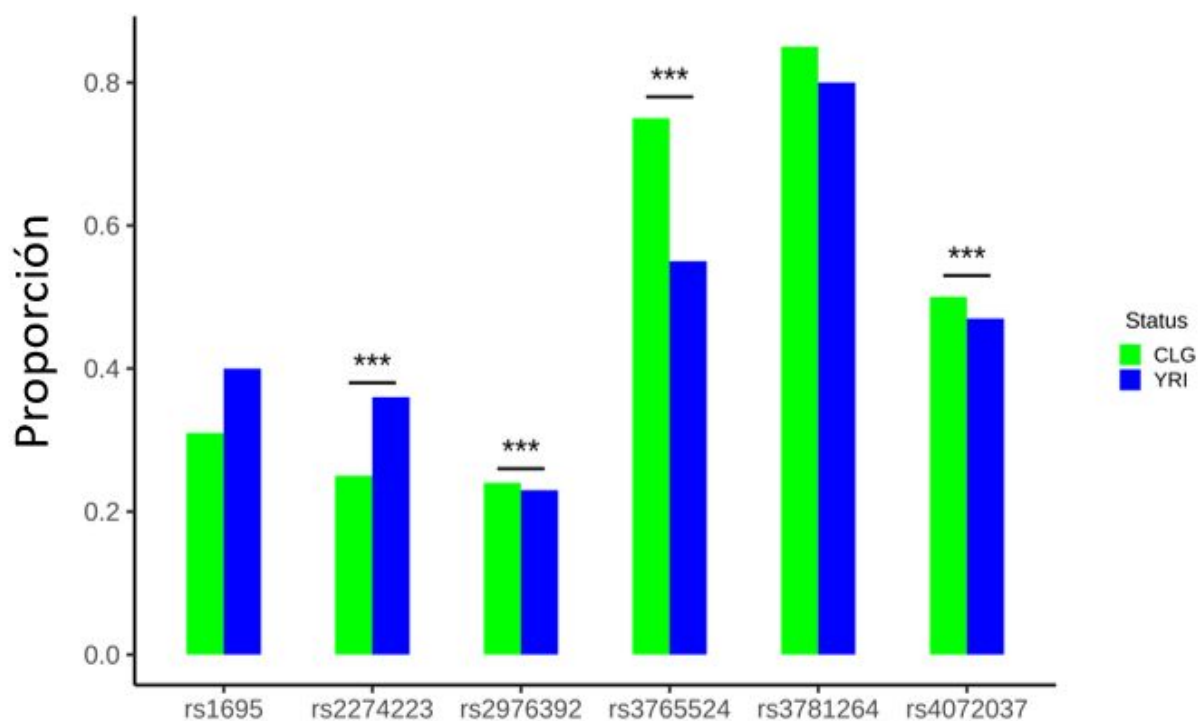


**Figura 5.** Proporción de cromosomas que portan un alelo de riesgo con ascendencia local Europea en la muestra de ChileGenómico (CLG) y su muestra de referencia Europea (IBS) para seis SNPs de riesgo de Cáncer Gástrico. La proporción fue calculada como el número de cromosomas portadores de alelos de riesgo con ascendencia local Europea dividido el número total de cromosomas con ascendencia inferida en esa región. Prueba de  $\chi^2$  (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ).



**Figura 6.** Proporción de cromosomas que portan un alelo de riesgo con ancestría local Amerindia en la muestra de ChileGenómico (CLG) y una muestra de referencia Amerindia (AMR) para seis SNPs de riesgo de Cáncer Gástrico. La proporción fue calculada como el número de cromosomas portadores de alelos de riesgo con ancestría local Amerindia dividido el número total de cromosomas con ancestría inferida en esa región. Prueba de  $\chi^2$  (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ).





**Figura 7.** Proporción de alelos de riesgo en regiones cromosómicas de ancestría Africana en la muestra de ChileGenómico (CLG) y la muestra de la población de referencia Africana (YRI) para seis SNPs de riesgo de Cáncer Gástrico. La proporción fue calculada como el número de cromosomas portadores de alelos de riesgo con ancestría local Africana dividido el número total de cromosomas con ancestría inferida en esa región. Prueba de  $X^2$  (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ).

## **V.2. Objetivo Específico 2**

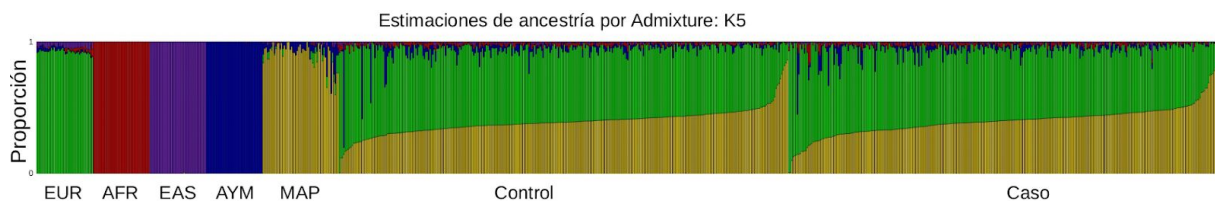
Validar el valor predictivo de cada locus de riesgo en la población chilena y determinar si depende de la ancestría.

### *V.2.A Evaluación de la necesidad de controlar por estructuración genética y/o ancestría entre casos y controles chilenos de cáncer gástrico*

El K6 de estimación de ancestrías globales mostró que algunos individuos tenían un 100% (Figura Suplementaria 9), por lo que se realizó un análisis con REAP<sup>65</sup> para identificar si se encontraban emparentados y determinar si era necesario eliminarlos de la muestra. Para realizar el análisis con REAP se utilizaron las estimaciones de ancestrías calculadas en un K5 por tener el menor error de validación cruzada (Figura Suplementaria 10). Se eliminaron 19 individuos, ya que se identificaron 38 individuos emparentados, de los cuales seis individuos eran controles, siete casos y seis individuos de referencia de ancestría Mapuche. Luego, se realizó un nuevo análisis de estimaciones de ancestría globales para 618 individuos (301 casos y 317 controles), para analizar las estimaciones de ancestría con K5 porque tienen el menor error de validación cruzada (Figura Suplementaria 11). Las estimaciones de ancestría se observan en la Figura 8. No se encontraron diferencias significativas en los porcentajes promedios de las ancestrías Europea, Aymara y Mapuche entre casos y controles (Tabla Suplementaria 6).

En el análisis de componentes principales no se observan agrupaciones espaciadas entre los grupos caso y control entre los primeros tres componentes principales de los 10 obtenidos (Figura Suplementaria 12). Se graficaron solo estos, debido a que el primer y segundo componente principal explican la mayor parte de la varianza de los datos, en este caso cerca del 0.21 y 0.13 respectivamente. Nuestros

resultados no evidencian diferencias en estos componentes principales de variación genética entre casos y controles de CG.



**Figura 8.** Estimaciones de las Ancestrías Globales por individuo en la muestra caso-control de Cáncer Gástrico. Se utilizó un K5, entonces se estimaron las ancestrías Europea (EUR), Africana (AFR), Asiática del Este (EAS), Aymara (AYM) y Mapuche (MAP). Se utilizó programa ADMIXTURE, seleccionando el K5 con una prueba de validación cruzada.

*V.2.B Genotipificación y control de calidad de las variantes genéticas seleccionadas para el objetivo específico 1, en casos de CG y controles de la población chilena.*

Se realizó un control de calidad a los datos cuyo flujo se encuentra en la Figura Suplementaria 2. De los 13 SNPs de riesgo para el CG seleccionados en el objetivo 1, sólo 9 están presentes en el arreglo y, con la búsqueda de SNPs que estuviesen en LD y en el arreglo, finalmente se pudo analizar 11 SNPs (Tabla 5).

Se realizaron tres tipos de análisis de asociación: a) Toda la muestra de casos (301 casos, Tabla 6), b) Casos solo de tipo intestinal (156 casos, Tabla 7) y c) Casos solo de tipo difuso (143 casos, Tabla 7). En los tres análisis el SNP rs2294008 está asociado, sin embargo, solo en a) y b) sigue siendo significativo luego de la corrección de Bonferroni, por lo tanto, sólo uno de los 11 SNPs evaluados está asociado en la muestra caso-control de CG. Sin embargo, la explicación más probable de nuestro resultado es una falta de poder (Tabla 4). De hecho, en la Figura 7 se observa que solamente los SNPs rs4072037 y rs2294008 tienen diferencias significativas en las frecuencias alélicas entre casos y controles. Sin embargo, tal como se evidenció en el análisis de asociación, luego de la corrección de Bonferroni rs4072037 dejó de ser significativo.

Para determinar si la falta de asociación para cada SNP es producto del tamaño muestral, se realizó un análisis de correlación para determinar, en conjunto, estos marcadores genéticos tienden a mostrar ORs similares a los publicados. Encontramos dos grupos de SNPs, rs4072037 y rs2294008 mostraron una alta correlación, con valores de OR casi idénticos a los reportados y fueron los que replicaron su asociación en la muestra chilena. El resto de SNP forma una nube que en conjunto muestra un sesgo hacia valores menores, sobre todo para los SNPs de mayor efecto. La pendiente de la asociación fue menor a 1 ( $b_1=0.419$ ,  $EE=0.166$ ) pero significativamente mayor a cero ( $p=0.039$ ). Por lo tanto, si bien nuestro estudio

tiene poder insuficiente para detectar asociaciones con  $OR < 1.8$ , nuestros resultados muestran que los SNPs replicaron de forma diferencial, y que otros factores pueden estar causando que el cluster de SNP que no replicaron su asociación (Figura 9).

		Frecuencias Alélicas					
		0.05	0.1	0.15	0.2	0.25	0.3
Odd Ratio	1.2	0.017	0.03	0.044	0.056	0.066	0.074
	1.4	0.067	0.149	0.227	0.289	0.335	0.365
	1.6	0.177	0.382	0.531	0.625	0.68	0.708
	1.8	0.342	0.64	0.789	0.858	0.89	0.902
	2	0.529	0.83	0.926	0.958	0.97	0.973

**Tabla 4.** Poder del estudio para un análisis de asociación con 300 casos y 300 controles con distintas frecuencias alélicas y valor de Odds Ratios (razón de momios). Resultados obtenidos en Genetic Association Study (GAS) Power Calculation ([http://csg.sph.umich.edu/abecasis/gas\\_power\\_calculator/](http://csg.sph.umich.edu/abecasis/gas_power_calculator/)).

Cromosoma	Inicial	Presente en GSA	Reemplazo	r	Alelo de Riesgo	OR
1	rs4072037	Sí			T	1.32
1	rs3762272		rs8179271	0.97	A	1.39
2	rs3834129		rs12693932	0.95	T	1.35
3	rs3087465		rs7648529	0.90	G	1.56
3	rs9841504		rs9820958	0.92	A	1.29
5	rs13361707	Sí			C	1.41
6	rs1799724		rs6916921	1.00	C	1.17
8	rs2294008	Sí			T	1.20
10	rs2274223	Sí			G	1.34
11	rs1695	Sí			G	1.19
22	rs738722	Sí			T	1.32

**Tabla 5.** SNPs evaluados en las muestras caso-control genotificados con el arreglo Global Screening (GSA). Para los SNPs presentes en GSA, se utilizó el OR utilizado en el Objetivo 1 (obtenido desde literatura). Para aquellas que no están presentes en GSA, se ajustó el OR del SNP (Objetivo 1) con la raíz cuadrada de la medida de desequilibrio de ligamiento  $r^2$  entre la variante inicial y la variante de reemplazo.  $r$ :  $\sqrt{r^2}$ . Fórmula utilizada para ajustar el OR:  $e^{\ln(OR) \times r}$ .

Crm	SNP	Riesgo	No Riesgo	OR Publicado	OR CG	IC	p	Bonferroni
1	rs4072037	T	C	1.32	1.29	[1.01-1.64]	0.04	0.42
1	rs8179271	A	G	1.39	1.07	[0.79-1.46]	0.65	1.00
2	rs12693932	T	C	1.35	1.05	[0.83-1.33]	0.67	1.00
3	rs7648529	G	A	1.56	1.21	[0.94-1.57]	0.14	1.00
3	rs9820958	A	G	1.29	1.09	[0.81-1.46]	0.57	1.00
5	rs13361707	C	T	1.41	1.20	[0.92-1.57]	0.17	1.00
6	rs6916921	C	T	1.17	1.03	[0.81-1.33]	0.79	1.00
8	rs2294008	T	C	1.20	1.49	[1.18-1.87]	0.00	0.01
10	rs2274223	G	A	1.34	1.17	[0.89-1.53]	0.25	1.00
11	rs1695	G	A	1.19	1.08	[0.86-1.35]	0.50	1.00
22	rs738722	T	C	1.32	1.03	[0.81-1.31]	0.83	1.00

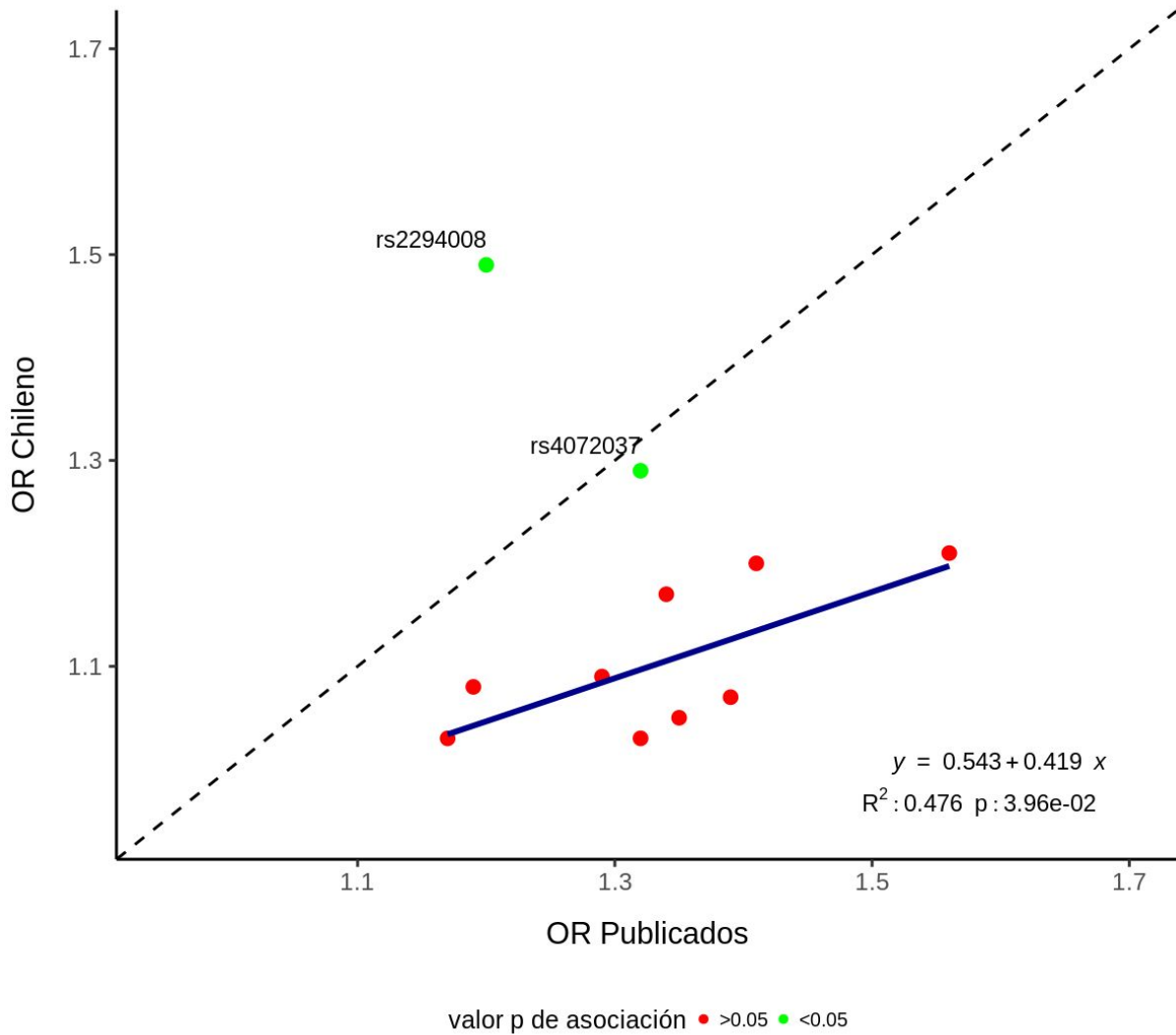
**Tabla 6.** Análisis de asociación de las variantes genéticas de riesgo para el desarrollo del Cáncer Gástrico en la muestra caso-control de 317 controles y 301 casos de Cáncer Gástrico. Crm: cromosoma; alelo de riesgo: Riesgo; alelo de no riesgo; No Riesgo; OR CG: odds ratio (razón de momios) para todas las muestras de

CG; IC: intervalo de confianza; p: valor  $p$ ; Bonferroni: valor  $p$  ajustado por comparaciones múltiples.

Variante	Riesgo	CG Difuso				CG Intestinal			
		OR	IC	valor de p	Bonferroni	OR	IC	valor de p	Bonferroni
rs4072037	T	1.34	[0.99-1.82]	0.06	0.66	1.23	[0.91-1.64]	0.17	1.00
rs8179271	A	0.92	[0.62-1.36]	0.68	1.00	1.23	[0.86-1.76]	0.26	1.00
rs12693932	T	1.03	[0.77-1.38]	0.84	1.00	1.07	[0.80-1.41]	0.65	1.00
rs7648529	G	1.11	[0.80-1.53]	0.53	1.00	1.32	[0.97-1.78]	0.08	0.85
rs9820958	A	1.16	[0.81-1.66]	0.42	1.00	1.00	[0.69-1.43]	0.99	1.00
rs13361707	C	1.03	[0.75-1.43]	0.84	1.00	1.41	[1.01-1.96]	0.04	0.49
rs6916921	C	1.06	[0.77-1.45]	0.72	1.00	1.04	[0.77-1.41]	0.79	1.00
rs2294008	T	1.44	[1.08-1.91]	0.01	0.14	1.57	[1.19-2.07]	0.00	0.02
rs2274223	G	1.38	[1.00-1.92]	0.05	0.55	0.99	[0.71-1.38]	0.94	1.00
rs1695	G	0.92	[0.69-1.22]	0.54	1.00	1.27	[0.96-1.66]	0.09	1.00
rs738722	T	1.19	[0.89-1.61]	0.24	1.00	0.89	[0.66-1.20]	0.45	1.00

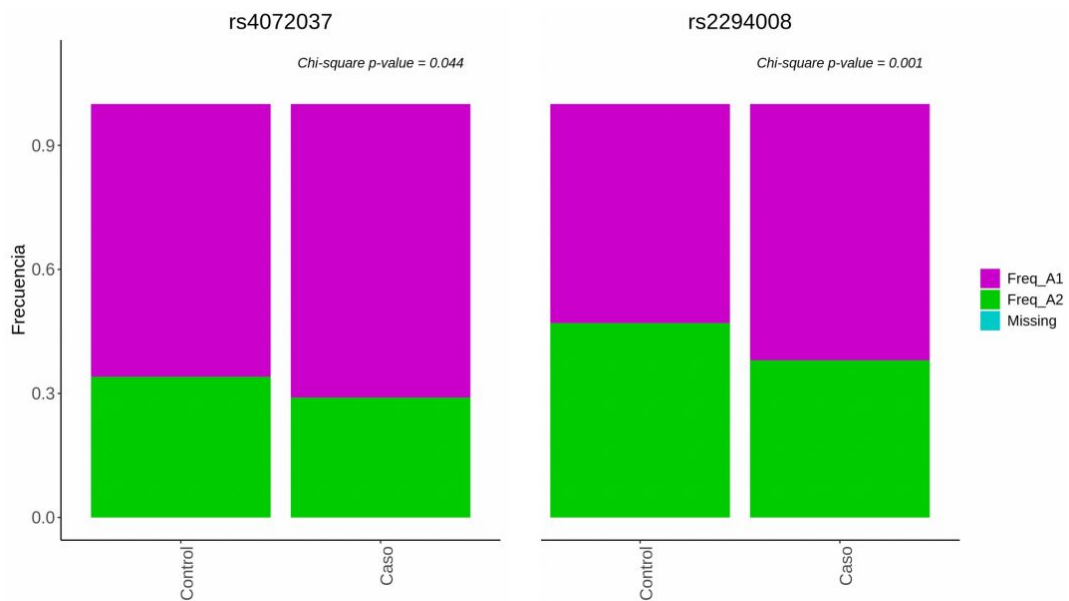
**Tabla 7.** Análisis de asociación de las variantes genéticas de riesgos para el desarrollo del Cáncer Gástrico en la muestra caso-control de 317 controles y 143 casos de Cáncer Gástrico tipo Difuso y 156 casos de Cáncer Gástrico tipo Intestinal. Riesgo: alelos de riesgo; OR: odds ratio (razón de momios); IC: intervalo de confianza; valor de  $p$ : valor de  $p$  crudo; Bonferroni: valor  $p$  ajustado por comparaciones múltiples.

## OR publicados y OR Chileno para Cáncer Gástrico



**Figura 9.** Relación entre los ORs extraídos desde la literatura para los SNPs seleccionados en el Objetivo 1 y los ORs calculados en la muestra caso-control (OR Chileno). Modelo lineal:  $y$ ; coeficiente de determinación:  $R^2$ ; y valor de  $p$   $\beta_1$ :  $p$  (Estadístico F).





**Figura 10.** SNPs con diferencias significativas en la frecuencia de sus alelos de riesgo entre controles y casos de Cáncer Gástrico. Freq\_A1: Frecuencia Alelo de Riesgo; Freq\_A2: Frecuencia de alelo no Riesgo; Missing: Datos perdidos. Valor de p para la prueba de Chi-cuadrado: Chi-square p-value.

### *V.2.C Selección de variantes genéticas de riesgo para desarrollar Cáncer Gástrico en individuos de la población chilena.*

Se realizó una selección de variables genéticas (SNPs y ancestría Mapuche) para crear modelos logísticos que se ajusten mejor a la muestra caso-control de CG. La selección de variables se realizó mediante una búsqueda exhaustiva para cada modelos logístico.

El modelo basal es aquél que incluye solamente las variables de edad y sexo (Tabla 8). Cabe recalcar que estas variables fueron significativas en todos los modelos logísticos seleccionados, y esto se condice con que son factores de riesgo para el CG. El modelo de ancestría más predictivo es aquél que presenta las variables de edad, sexo y la ancestría Mapuche (Tabla 8). Los modelos seleccionados que contenían los SNPs de riesgo están presentes en la Tabla 9. El mejor modelo de riesgo entre los que incluyen las variable de edad, sexo y tres de los 11 SNPs de riesgo (rs4072037, rs12693932 y rs2294008). El modelo que incluye la ancestría tiene las variables edad, sexo, ancestría Mapuche y dos de los 11 SNPs de riesgo (rs4072037 y rs2294008). Se evaluó también un modelo que incluye la interacción de tres SNPs con la ancestría Mapuche para detectar si la ancestría modifica el efecto que tiene el SNP. En este caso, se observa que los efectos de la interacción fueron positivos pero no significativos.

Dentro de los modelos seleccionados hay variables que no tienen un valor de  $p$  significativo, esto significa que aunque esta variable ayuda a explicar parte de la varianza de los datos, predicción del modelo, no tiene una correlación con la variable dependiente (fenotipo) lo que puede deberse a una falta de poder de la muestra.

Para determinar si los modelos que incluyen variables genéticas mejoran la predicción de riesgo se compararon con el modelo basal. El modelo de ancestría

(MAP) no es más predictivo que el modelo basal, pero aquellos que incluyen SNPs (SNPs, MAP + SNPs y MAP\*SNPs) son significativamente más predictivos. Esto quiere decir que agregar los predictores genéticos (SNPs o MAP y SNPs) mejora significativamente el ajuste del modelo (Tabla 10).

Basal					Ancestría				
<i>Predictors</i>	<i>Odds Ratios</i>	<i>IC</i>	<i>Statistic</i>	<i>p</i>	<i>Predictors</i>	<i>Odds Ratios</i>	<i>IC</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	0.02	0.01 – 0.04	-9.25	< <b>0.001</b>	(Intercept)	0.01	0.00 – 0.03	-8.21	< <b>0.001</b>
Edad	1.08	1.06 – 1.09	9.72	< <b>0.001</b>	Edad	1.08	1.06 – 1.10	9.74	< <b>0.001</b>
Sexo: hombre	0.67	0.45 – 1.00	-1.97	<b>0.049</b>	Sexo: hombre	0.66	0.44 – 0.97	-2.08	<b>0.037</b>
MAP					MAP	4.11	0.81 – 22.18	1.68	0.093
Observations	602				Observations	602			
Deviance	702.357				Deviance	699.453			
AIC	708.357				AIC	707.453			
log-Likelihood	-351.179				log-Likelihood	-349.726			

**Tabla 8.** Modelos logísticos de riesgo basal y de ancestría con sus efectos de ajuste para la muestra de 308 controles y 294 casos de Cáncer Gástrico. Los modelos son: Basal:  $y = \text{edad} + \text{sexo}$ , Ancestría:  $\text{edad} + \text{sexo} + \text{Ancestría Mapuche (MAP)}$ . IC: Intervalo de Confianza al 95%, Statistic: Valor z, p: valor de p, Deviance: Devianza, AIC: Criterio de Información Akaike, log-Likelihood: logaritmo de la verosimilitud.

Predictors	SNPs				MAP + SNPs				MAP*SNPs			
	Odds Ratios	IC	Statistic	p	Odds Ratios	IC	Statistic	p	Odds Ratios	IC	Statistic	p
(Intercept)	0.01	0.00 – 0.02	-9.07	<0.001	0.01	0.00 – 0.02	-8.53	<0.001	0.01	0.00 – 0.02	-9.42	<0.001
Edad	1.08	1.06 – 1.09	9.57	<0.001	1.08	1.06 – 1.09	9.54	<0.001	1.08	1.06 – 1.10	9.63	<0.001
Sexo: hombre	0.64	0.43 – 0.96	-2.15	0.031	0.65	0.43 – 0.96	-2.14	0.032	0.63	0.42 – 0.94	-2.25	0.024
rs4072037	1.30	0.99 – 1.71	1.87	0.061	1.26	0.96 – 1.66	1.63	0.102				
rs12693932	1.23	0.94 – 1.61	1.51	0.131								
rs2294008	1.36	1.05 – 1.78	2.28	0.023	1.34	1.03 – 1.75	2.17	0.030				
MAP					3.21	0.62 – 17.82	1.36	0.172				
MAP:rs4072037									1.62	0.90 – 2.97	1.59	0.112
MAP:rs12693932									1.40	0.78 – 2.55	1.12	0.264
MAP:rs2294008									1.76	0.97 – 3.23	1.84	0.066
Observations	602				602				602			
Deviance	691.826				692.211				690.958			
AIC	703.826				704.211				702.958			
log-Likelihood	-345.913				-346.105				-345.479			

**Tabla 9.** Modelos logísticos que incluyen variables genéticas de riesgo para la muestra caso-control de 308 controles y 294 casos de Cáncer Gástrico. Se utilizaron aquellas muestras que se tenían los datos de edad, sexo y las variables genéticas de los 11 SNPs analizados y la ancestría global Mapuche (MAP). Se realizó una búsqueda exhaustiva para la selección de los predictores de los modelos logísticos SNPs y MAP + SNPs. Estos incluyeron las variables: a) SNPs:  $y = \text{edad} + \text{sexo} + \text{rs4072037} + \text{rs12693932} + \text{rs2294008}$ ; y b) MAP + SNPs:  $y = \text{edad} + \text{sexo} + \text{MAP} + \text{rs4072037} + \text{rs2294008}$ . Para el modelo MAP \* SNPs, se evaluó la interacción entre MAP y las variables genéticas seleccionadas en el modelo SNPs. El modelo MAP\*SNPs es:  $y = \text{edad} + \text{sexo} + \text{MAP} * \text{rs4072037} + \text{MAP} * \text{rs12693932} + \text{MAP} * \text{rs2294008}$ . La selección de los modelos se realizó con el Criterio de Información Akaike (AIC) y la Devianza (Deviance). IC: Intervalo de Confianza al 95%, Statistic: Estadístico Z-score, p: valor de p.

<b>Modelo 1</b>	<b>Modelo 2</b>	<b>LRT p value</b>
basal	MAP	0.09
basal	SNPs	0.01
basal	MAP + SNPs	0.02
basal	MAP*SNPs	9.75E-03

**Tabla 10.** Comparación de los modelos logísticos analizados mediante un análisis anidado realizado con la prueba de la razón de verosimilitudes (LRT). Los modelos comparados fueron: a) Basal:  $y = \text{edad} + \text{sexo}$ , b) MAP:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche}$ , c) SNPs:  $y = \text{edad} + \text{sexo} + \text{rs4072037} + \text{rs12693932} + \text{rs2294408}$ , d) MAP + SNPs:  $y = \text{edad} + \text{sexo} + \text{rs4072037} + \text{rs2294408} + \text{ancestría Mapuche}$ , y e) MAP \* SNPs:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche} * \text{rs4072037} + \text{ancestría Mapuche} * \text{rs12693932} + \text{ancestría Mapuche} * \text{rs2294408}$ . Valor de p de LRT: LRT p value, Observations: número de muestras.

#### *V.2.D Comparación de modelos logísticos de puntaje de riesgo y ancestría para el Cáncer Gástrico en individuos de la población chilena.*

Luego de dar cuenta de los efectos de edad y sexo, el PRS resultó significativamente asociado al riesgo de CG, con un OR de 1.76 (CI95=1.27-2.46,  $p=0.001$ ). Si bien la ancestría mapuche no estuvo asociada al riesgo ( $p=0.119$ ), el efecto de interacción entre ancestría y PRS sí resultó significativo, con un OR de 2.21 (CI95=1.37-3.76,  $p=0.002$ ). Este efecto se interpreta con el aumento de riesgo asociado a una unidad de aumento en PRS entre una persona con 0% de ancestría mapuche y una persona con 100% de esta ancestría (Tabla 11), es decir que el aumento del PRS está asociado a un mayor porcentaje de ancestría Mapuche. Para determinar si los modelos que incluyen variables genéticas mejoran la predicción de riesgo se compararon con el modelo basal. Los tres modelos que incluyen la variable PRS son más predictivos que el modelo basal, lo que quiere decir que agregar los predictores genéticos (PRS o MAP y PRS) mejora significativamente el ajuste del modelo (Tabla 12).

En el objetivo 1.D se evaluó si existe una relación entre el PRS de la población chilena y su proporción de ancestría Mapuche, cuyo resultado fue una relación positiva significativa. Aquí también se realizó lo mismo tanto en la muestra caso-control (como representación de un grupo heterogéneo de individuos) como en el grupo de casos de CG, evaluando los datos en cuartiles de PRS y terciles de ancestría Mapuche, y un análisis de correlación entre PRS y el porcentaje de ancestría Mapuche. En la Figura 11 se observa que la proporción de casos sobre los controles aumenta a medida que aumenta el tercil de ancestría, de un promedio de 47.7 en el tercil 1 de ancestría a un promedio de 54.63 en el tercil 3 (cuartil 4 de PRS, tercil 3 de Ancestría es 65.5%). Además, aumentan a medida que incrementa el cuartil de PRS, esto se observa más claramente en el tercil 3 de ancestría donde

el cuartil 1 de PRS tiene 48% de casos, el cuartil 2 49%, cuartil 3 56% y el cuartil 4 65.5%.

La Figura 12 muestra que existe una relación lineal positiva entre el PRS y la ancestría Mapuche en la muestra de solo casos ( $\beta_1 = 0.6811$ ,  $p < 0.01$ ), lo cual no fue observado en el grupo de solo controles ( $\beta_1 = 0.4916$ ,  $p > 0.05$ ). Esto indica que, tal como se observó en el Objetivo Específico 1, el grupo de solo casos tiene una relación positiva entre el PRS y la ancestría Mapuche cuyo beta es mayor a la observada en la muestra de ChileGenómico ( $\beta_1 = 0.344$ ,  $p < 0.01$ ). Esto puede deberse a un enriquecimiento de individuos con alto riesgo (solo casos), aumentando la varianza de PRS en la muestra.

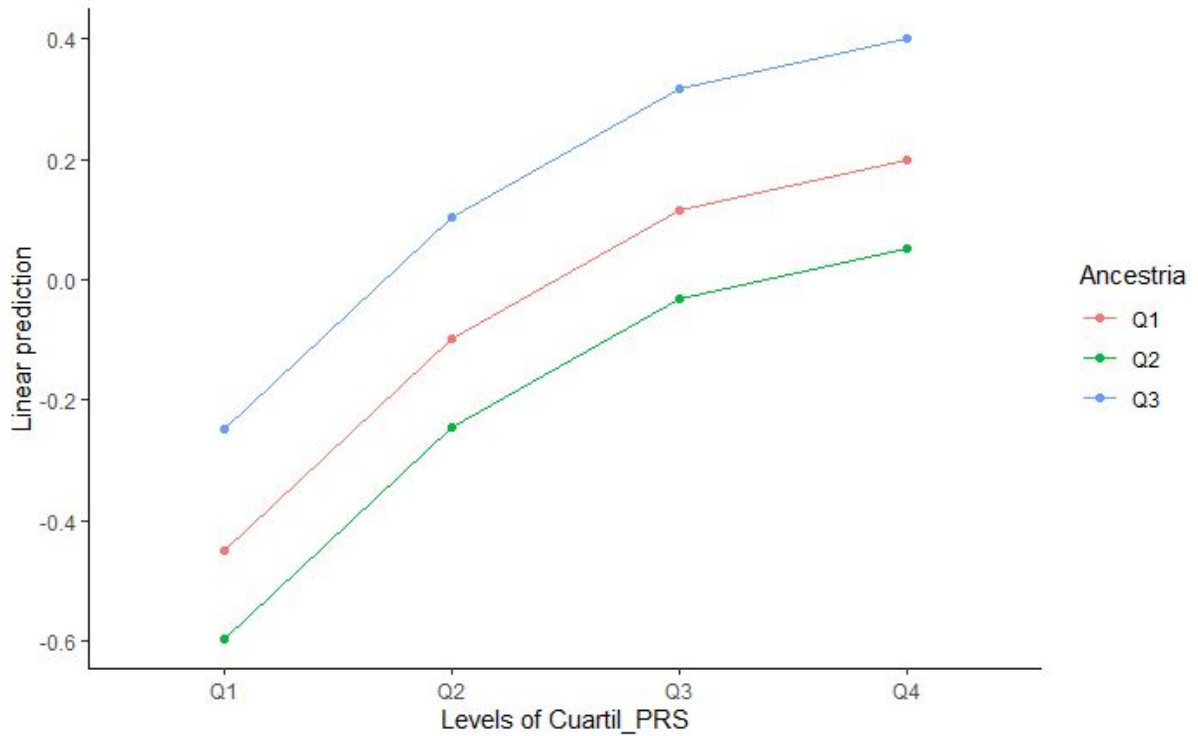
Predictors	PRS				MAP + PRS				MAP*PRS			
	Odds Ratios	IC	Statistic	p	Odds Ratios	IC	Statistic	p	Odds Ratios	IC	Statistic	p
(Intercept)	0.00	0.00 – 0.01	-8.66	<0.001	0.00	0.00 – 0.01	-8.30	<0.001	0.01	0.00 – 0.02	-9.27	<0.001
Edad	1.08	1.06 – 1.10	9.73	<0.001	1.08	1.06 – 1.10	9.74	<0.001	1.08	1.06 – 1.10	9.75	<0.001
Sexo: hombre	0.66	0.44 – 0.98	-2.06	0.040	0.65	0.43 – 0.96	-2.14	0.033	0.64	0.43 – 0.95	-2.18	0.029
PRS	1.76	1.27 – 2.46	3.34	0.001	1.72	1.23 – 2.41	3.16	0.002				
MAP					3.02	0.57 – 16.90	1.29	0.199				
MAP:PRS									2.21	1.37 – 3.66	3.16	0.002
Observations	602				602				602			
Deviance	690.821				689.134				691.534			
AIC	698.821				699.134				699.534			
log-Likelihood	-345.411				-344.567				-345.767			

**Tabla 11.** Modelos logísticos de riesgo que incluyen el Puntaje de Riesgo Poligénico (PRS) y la ancestría global Mapuche (MAP) para la muestra caso-control de 308 controles y 294 casos de Cáncer Gástrico. Se utilizaron aquellas muestras que se tenían los datos de edad, sexo y las variables genéticas de PRS y MAP. Los modelos evaluados son: a) PRS:  $y = \text{edad} + \text{sexo} + \text{PRS}$ , b) MAP + PRS:  $y = \text{edad} + \text{sexo} + \text{Ancestría Mapuche} + \text{PRS}$ , y c) MAP \* PRS:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche} * \text{PRS}$ . IC: Intervalo de Confianza al 95%, Statistic: Estadístico Z-score, p: valor de p, Observations: número de muestras.

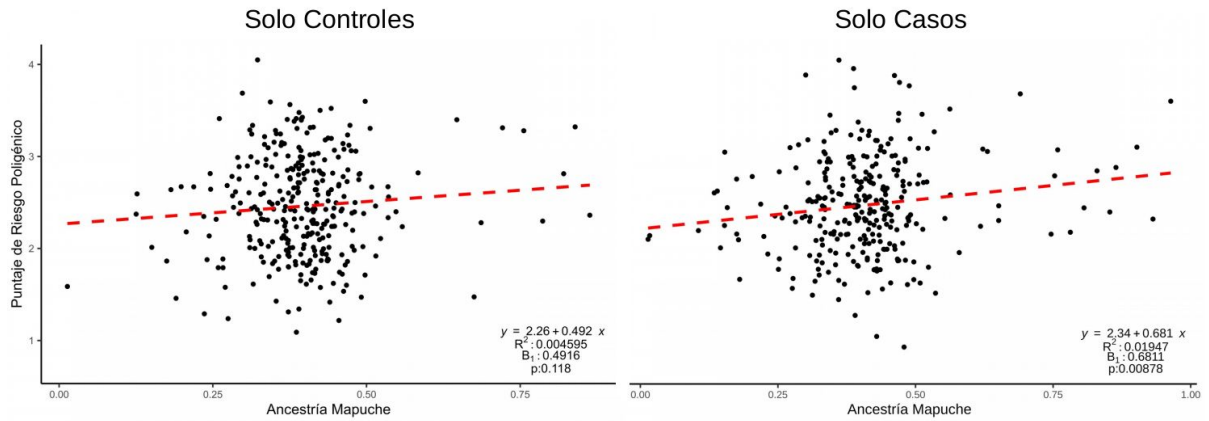


<b>Modelo 1</b>	<b>Modelo 2</b>	<b>LRT p value</b>
basal	MAP	0.09
basal	PRS	6.83E-04
basal	MAP + PRS	1.34E-03
basal	MAP*PRS	0.001

**Tabla 12.** Comparación de los modelos logísticos analizados mediante un análisis anidado realizado con la prueba de la razón de verosimilitudes (LRT). Los modelos comparados fueron: a) Basal:  $y = \text{edad} + \text{sexo}$ , b) MAP:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche}$ , c) PRS:  $y = \text{edad} + \text{sexo} + \text{PRS}$ , d) MAP + PRS:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche} + \text{PRS}$ , y e) MAP \* PRS:  $y = \text{edad} + \text{sexo} + \text{ancestría Mapuche} * \text{PRS}$ . Valor de p de LRT: LRT p value.



**Figura 11.** Predicción lineal de la interacción entre terciles de ancestría con cuartiles de Puntaje de Riesgo Poligénico. Los colores representan los distintos niveles de ancestría: a) rojo: tercil 1, b) verde: tercil 2, y c) azul: tercil 3. Los niveles de cuartiles de PRS (Levels of Cuartil\_PRS) están en el eje x: a) Q1: cuartil 1, b) Q2: cuartil 2, c) Q3: cuartil 3, y d) Q4: cuartil 4. Predicción lineal: Linear prediction.



**Figura 12.** Relación entre el Puntaje de Riesgo Poligénico y la ancestría Mapuche en 317 controles (izquierda) y 301 casos de Cáncer Gástrico (derecha). Modelo lineal:  $y$ , coeficiente de determinación:  $R^2$ , efecto de la ancestría Mapuche:  $\beta_1$ , y el valor  $p$  para  $\beta_1 \neq 0$ :  $p$ .

### **V.3. Objetivo Específico 3**

Identificar regiones genómicas que se asocian a cáncer gástrico en pacientes chilenos mediante un enfoque de Admixture Mapping.

#### *V.3.A Faseo de los cromosomas en los individuos del set de datos de casos de GC para determinar su haplotipo parental*

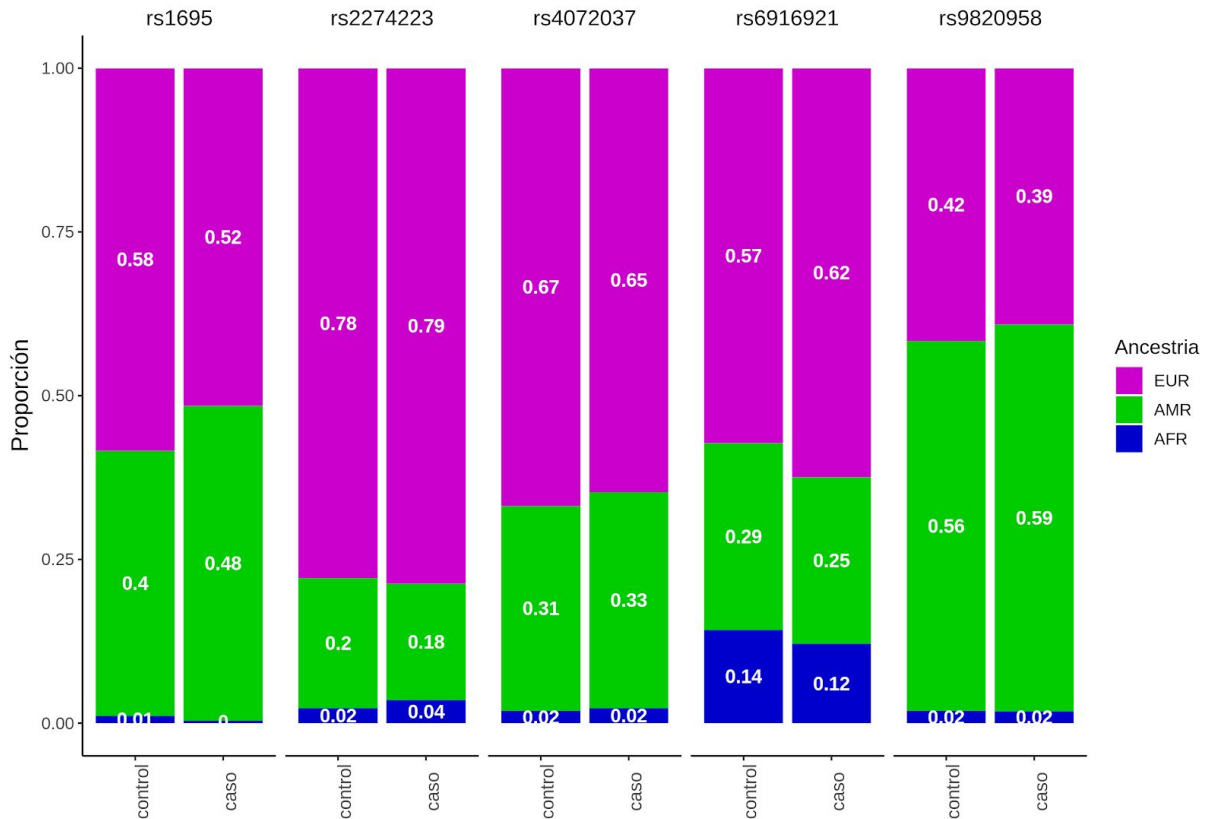
Se fasearon correctamente 792 individuos y 115.159 variantes, paso que se realiza para asignar alelos a los cromosomas paternos y maternos de cada individuo, dando un éxito del 100% en el procesamiento de los individuos y de 99.88% en las variantes. Sin embargo, se tuvo que eliminar el individuo EN219 que pertenece al grupo control, debido a que fue posible inferir correctamente solo un haplotipo parental del cromosoma 6.

#### *V.3.B Estimación de la ancestría local en regiones del genoma de individuos con fenotipo de caso de cáncer gástrico.*

Con el fin de evaluar si la asociación positiva entre PRS y la ancestría Mapuche puede explicarse por una asociación entre alelos de riesgo para SNPs conocidos y su ancestría local, se infirió la ancestría local de SNPs de cinco de los 11 SNPs de riesgo analizados (Figura 13). Se observa que los alelos de riesgo de cuatro SNPs están ubicados en regiones cromosómicas asociadas mayormente a ancestría europea (proporción > 0.5), exceptuando por rs9820958 que se encuentra mayormente en regiones cromosómicas de ancestría Amerindia, tanto para casos como para controles. Respecto a si estos los alelos de riesgo de estos SNPs están asociados al fenotipo, no se observan diferencias significativas entre casos y

controles para ninguno (Tabla Suplementaria 7), lo cual podría estar relacionado a la falta de asociación de cada SNP de manera individual en el punto 2.B.

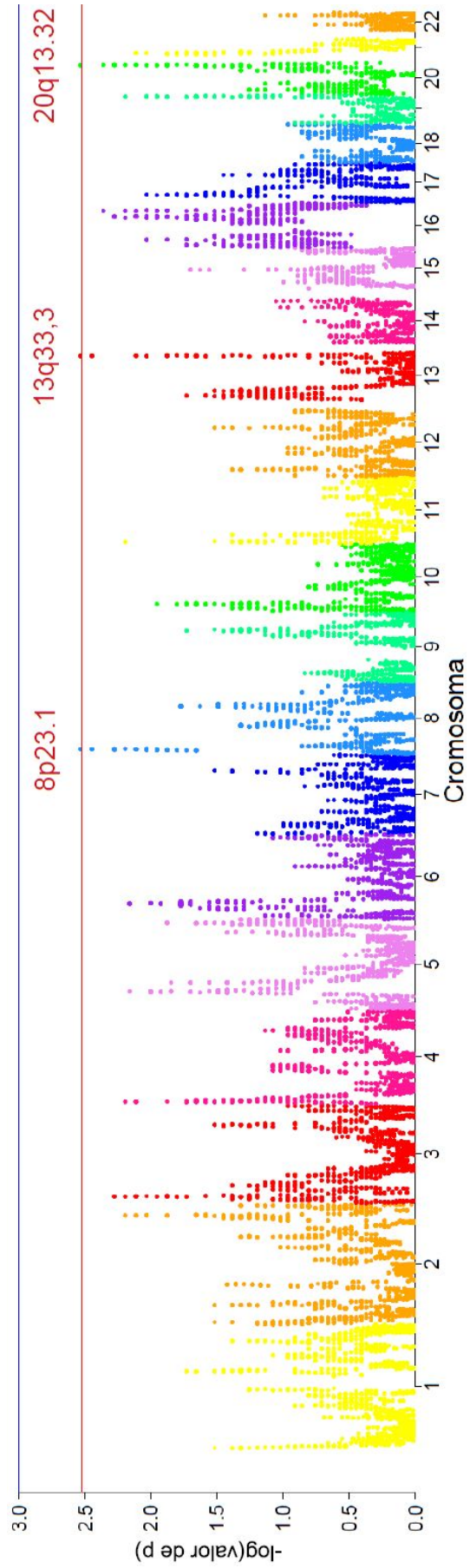
Esto quiere decir que la relación observada entre el PRS con la ancestría Mapuche, tanto en el análisis a nivel poblacional del objetivo 1 como en el de casos controles en este objetivo, se deba principalmente a la presencia de otros loci de riesgo en el genoma ubicados en regiones cromosómicas de ancestría Mapuche. Entonces, fue necesario evaluar si existen regiones cromosómicas enriquecidas de ancestría Mapuche asociadas al fenotipo caso de CG que pudiesen estar aportando susceptibilidad a este cáncer, lo que se realizó por medio de la metodología de Admixture Mapping.



**Figura 13.** Proporción de ancestrías Europea (EUR), Amerindia (AMR) y Africana (AFR) para los alelos de riesgo de cinco SNPs en 316 controles y 301 casos de Cáncer Gástrico. La proporción fue calculada como el número de cromosomas portadores de alelos de riesgo con una ancestría local Europea o Amerindia o Africana dividido el número total de cromosomas con ancestría inferida en esta región.

### *V.3.C Admixture Mapping para identificar regiones asociadas a Cáncer Gástrico por su ancestría.*

Se realizó un análisis de z-score, solo-casos, para determinar si existen loci enriquecidos en regiones genómicas de ancestría Amerindia o Europea en la muestra. En el análisis de z-score para regiones genómicas de ancestría Europea, no se encontraron variantes con valores de p significativos (datos no mostrados). En cambio, para el análisis de las regiones genómicas de ancestría Amerindia se observa que hay peaks en 3 regiones cromosómicas, 8p23.1, 13q33, y 20q13.32, que tienen un Z score positivo y un valor de  $p < 0.005$  (Figura 14). Esto quiere decir que existe un enriquecimiento de ancestría Amerindia. Sin embargo, estas regiones cromosómicas son sugerentes de asociación debido a que no alcanzan el nivel de significancia para establecer que están asociadas al riesgo de CG (valor de  $p < 0.001$ ). Esto implica que es necesario complementar este estudio con otros que tengan un mayor tamaño muestral para determinar si estas regiones cromosómicas están o no asociadas al CG en nuestra población. Cabe recalcar que los SNPs que fueron seleccionados desde la literatura, en el objetivo 1.1, no están incluidos dentro de las regiones genómicas identificadas por admixture mapping.



**Figura 14.** Admixture Mapping de regiones cromosómica de ancestría Amerindia con un análisis de Solo-Casos para 301 casos de el Cáncer Gástrico. Líneas valores umbrales: p de asociación  $< 1E^{-3}$  (azul) y p sugerente de asociación  $< 3E^{-3}$  (roja).

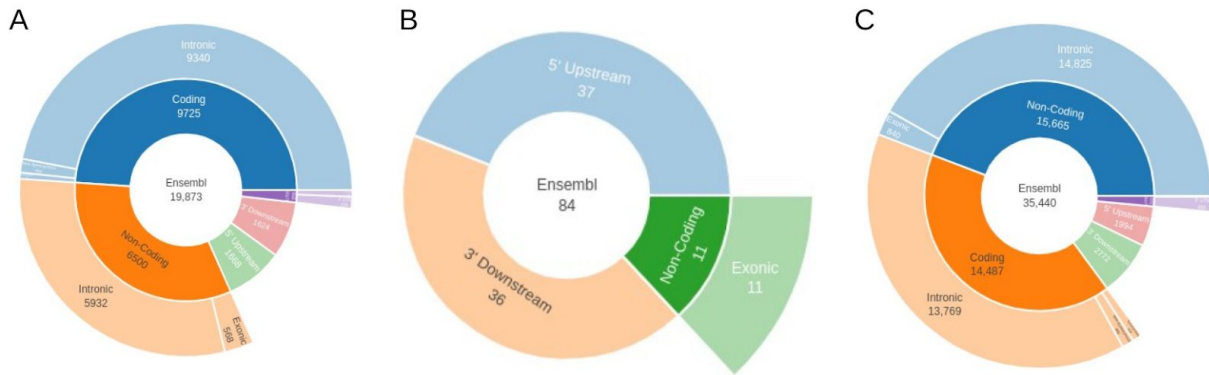


### V.3.D Evaluación enriquecimiento de funciones génicas en regiones genómicas identificadas por Admixture Mapping.

Las regiones genómicas que fueron identificadas en los cromosomas 8 (8p23.1), 13 (13q33) y 20 (20q13.32), inicialmente contenían 69 SNPs. Para tener una mayor probabilidad de éxito en la búsqueda de asociaciones para variantes presentes dentro de estas regiones cromosómicas. Esto se realizó en dos pasos: i) se complementó con 204 SNPs presente en las regiones genómicas presente en el arreglo GSA; y ii) se buscaron SNPs en desequilibrio de ligamiento con los 204 SNPs en la base de datos de Haploreg v4.1<sup>72</sup>. En total se utilizaron 14550 variantes genéticas.

En la base de datos de SNP Nexus<sup>77</sup> se anotó la información con 14550 variantes genéticas y las clasificaciones encontradas en Ensembl están en la Figura 15. En los cromosomas 8 y 20 se encontraron una mayor cantidad de predicciones de efectos y tienen un efecto en regiones codificantes de genes. De hecho, en el cromosoma 13 se encontró un solo gen, en cambio, el cromosoma 8 y 20 tienen 14 y 11 genes, respectivamente (Tabla 13). Dentro de las variantes genéticas patogénicas relacionadas al cáncer se encontró: 1) Cromosoma 8: los genes *BLK*, *LINC00208*, *MTMR9* y *TDH* se han relacionado al adenocarcinoma esofágico, y 2) Cromosoma 20: *GNAS-AS1* a cáncer de pulmón de células no pequeñas, *NPEPL1* al cáncer de mama, *GNAS*, *STX16-NPEPL1*, *MIR296* y *MIR298* que han sido relacionados a CG (Tabla 14).

Las herramientas web GREAT<sup>78</sup> o PANTHER no arrojaron resultados significativos con las 14450 variantes genéticas ingresadas. Un resumen de las vías de señalización y los genes encontrados en PANTHER están presente en las Tablas Suplementarias 8 y 9.



**Figura 15.** Predicciones de efectos de 14550 variantes genéticas anotadas en ENSEMBL para las regiones cromosómicas asociadas a ancestría Amerindia identificadas con el análisis de solo casos de Admixture Mapping. A) Cromosoma 8, B) Cromosoma 13 y C) Cromosoma 20.

Cromosoma	Gen	Cromosoma	Gen
8	AF131216	13	PPISP24
	BLK		
	C8orf12	20	GNAS
	CTC-493P15		GNAS-AS1
	FAM167A		MIR296
	LINC00208		MIR298
	LINC00529		NPEPL1
	MTMR9		PIEZO1P2
	RN7SL293P		RP11-261P9
	RNU6-1084P		RP1-309F20
	RP11-148O21		RP4-806M20
	RPL19P13		STX16
	SLC35G5		STX16-NPEPL1
	TDH		

**Tabla 13.** Genes presentes en las regiones cromosómicas asociadas a ancestría Amerindia identificadas con el análisis de solo casos de Admixture Mapping.

Cromosoma	Gen	Rol	Cáncer	DOI	PMID
8	<i>BLK</i>	Riesgo	Adenocarcinoma Esofágico y Esófago de Barrett	10.1016/S1470-2045(16)30240-6	27527254
8	<i>LINC00208</i>	Riesgo	Adenocarcinoma Esofágico y Esófago de Barrett	10.1016/S1470-2045(16)30240-6	27527254
8	<i>MTMR9</i>	Sobreexpresión	Adenocarcinoma Esofágico	10.1136/gut.2010.234179	21478220
8	<i>TDH</i>	Actividad antitumoral	Cáncer Esofágico	10.1155/2015/464937	26495015
20	<i>MIR296</i>	Sobreexpresión promueve crecimiento celular	Cáncer Gástrico	10.1016/j.prp.2018.09.026	30293950
20	<i>MIR298</i>	Podría tener como blanco BAX, el cual es usada como blanco molecular para terapia contra el cáncer	Cáncer Gástrico		26064232
20	<i>GNAS</i>	Asociación con el estadio tumoral y supervivencia	Cáncer Gástrico	10.2217/pgs-2018-0199	31124412
20	<i>GNAS-AS1</i>	Eje GNAS-AS1/miR-4319/NECAB3 promueve migración e invasión a través de la alteración de la polarización de macrófagos	Cáncer de Pulmón de células no-pequeñas	10.1007/s10142-019-00696-x	31267263
20	<i>NPEPL1</i>	Es blanco de miR-19a que es un miR oncogénico	Cáncer de Mama	10.1042/BSR20200626	32538432
20	<i>STX16</i>	Fusión con otro gen: <i>STX16-NPEPL1</i>	Tumores estromales Gastrointestinales	10.18632/oncotarget.3731	25987131

**Tabla 14.** Genes que se encuentran en regiones cromosómicas asociadas a la ancestría Amerindia, por Admixture Mapping, relacionados al cáncer. PMID: Identificador de la base de datos de PubMed.

## VI. DISCUSIÓN

Chile está dentro de los 10 países con la tasa de mortalidad más alta de CG a nivel mundial y no se tiene la información genética suficiente para poder realizar modelos de predicción de riesgo. Los principales estudios genéticos de asociación para el CG se han realizado en poblaciones europeas y asiáticas, que no son los mejores modelos para nuestra población mestiza porque posee un gran componente Amerindio. Sin embargo, se han realizado algunos estudios con individuos de América Latina, como el de Boqueti et al (2013) que reportó que IL-1RN\*2 aumenta el riesgo del CG en latinoamericanos<sup>81</sup>, y un estudio en Perú realizado por Zamudio et al (2016) donde no se encontró una asociación entre una selección de variantes ya descritas como riesgo para el CG y casos de CG de ese país<sup>48</sup>. En Chile, se reportaron asociaciones entre SNPs de genes que pertenecen a la vía de señalización RAS/RAF/MEK/ERK con el CG<sup>35</sup>. En América Latina existe una subestructura poblacional que se evidencia con distintas proporciones de mestizaje, lo que implica que haya una alta diferenciación fenotípica y genotípica<sup>41</sup>. Es más, la ancestría Amerindia está compuesta por distintas subpoblaciones indígenas entre los países latinoamericanos, lo que podría reflejarse en cambios en el riesgo que tiene un marcador genético. Por lo tanto, fue necesario evaluar si SNPs ya descritos están asociadas a nuestra población también, y si su ancestría particular es un factor confundente en estas asociaciones.

El PRS es un valor que permite combinar el efecto pequeño de varios SNPs que permite tener una mejor precisión del riesgo genético. Se esperaba correlacionar la distribución heterogénea de la tasa de mortalidad de CG descrita por el DEIS a lo largo del país<sup>13</sup> con un PRS, pero no fue posible con los SNPs de riesgo asociados en otras poblaciones. Sin embargo, encontramos una relación lineal positiva significativa entre la ancestría Amerindia y Mapuche con el PRS, y una relación lineal

negativa significativa con la Europea. Por lo tanto, existe una asociación entre el PRS con la ancestría que es heterogénea entre sus componentes ancestrales. Sin embargo, al comparar un PRS calculado con la selección de SNPs, en un grupo de individuos que representa los distintos componentes ancestrales, no se encontraron diferencias significativas entre las poblaciones Europea y Asiática con la Aymara, Mapuche y el grupo de individuos de ChileGenómico. Esto puede deberse a que los SNPs que están asociados en otras poblaciones no son predictivos en la nuestra, o los SNPs seleccionados no fueron suficientes para reproducir las diferencias en las tasas de incidencia y mortalidad observada a nivel mundial.

Luego, se evaluó si los SNPs de riesgo seleccionados estaban asociados de manera independiente en la muestra caso-control de CG de la población chilena. Encontramos que solamente rs2294008 está asociado a CG en nuestra población, el cual se ha descrito que aumenta el riesgo del CG en un 33%. Está se ubica en el gen *PSCA* (cromosoma 8q24.2), codifica para el antígeno prostático que es una proteína glicosilfosfatidilinositol de anclaje<sup>82</sup>. Esta proteína está relacionada a la proliferación celular y muerte celular, entre otros procesos, está presente en la próstata normal y sobreexpresado en el cáncer de próstata, y disminuido en los cáncer esofágico, vejiga y gástrico<sup>31,82</sup>. Cumple un rol importante en la adhesión celular que permite mantener la estructura del epitelio del estómago, que si es alterado permite la transición epitelial-mesenquimal que facultando la migración e invasión de las células tumorales a los vasos sanguíneos y linfáticos, que es el primer paso para la metástasis tumoral.

Al separar los pacientes por tipo de cáncer, difuso e intestinal, rs2294008 estuvo asociado solo con el tipo difuso. Sin embargo, el valor *p* nominal para el tipo intestinal fue menor a 0.05, mostrando una tendencia a tener efecto en ambos tipos, la diferencia sea sólo producto del bajo poder al dividir la muestra en dos. El hecho de que la mayoría de los SNPs estudiados no replicaron su asociación en nuestra

muestra es un resultado que podría darse ya que se ha reportado que hay variantes que no tienen la misma asociación en distintas poblaciones de estudio, e incluso pueden existir diferencias entre la histología difusa e intestinal de CG. Esta ya se ha reportado en algunos SNPs, como el rs4072037 que se ha descrito que se asociado en la población asiática, pero no en la caucásica<sup>49</sup>. La falta de asociación puede deberse a múltiples factores: i) factores ambientales no incluidos en los modelos ajustados, los cuales incluso podrían interactuar con los factores genéticos; ii) cambios en las frecuencias alélicas, como es el caso del rs2274223, y/o haplotípicas producto de la historia demográfica de la población chilena; iii) procesos de recombinación que hayan cambiado las asociaciones por ligamiento entre SNPs y variantes causales en la regiones estudiadas; iv) bajo poder para asociaciones con OR menor 1.8. Sin embargo, la comparación entre los valores de OR en la literatura y los estimados en el presente estudio, muestra una tendencia de asociación positiva en todos los SNPs, aunque con una tendencia a valores menores en Chile para un subconjunto de loci, sugiriendo una heterogeneidad de factores que afectan distintamente a los SNPs según su ubicación en el genoma.

Se evaluaron modelos lineales generalizados para determinar si el efecto en conjunto de distintas variables son más predictivas de riesgo que los SNPs de manera individual<sup>65</sup>. Se realizó una selección de los SNPs que fuesen más predictivos y se evaluó si las ancestrías podrían ser un factor que mejoren el puntaje de riesgo. La ancestría por sí sola no mejora el nivel predictivo del modelo basal. Esto concuerda con no haber encontrado diferencias significativas en las ancestrías globales entre casos y controles. Sin embargo, la selección de variables en el modelo que tiene solo las variables de ancestría, el primer modelo de predicción que incluye un componente ancestral es el que tiene la ancestría Mapuche. Este resultado podría relacionarse con la tendencia observada entre la tasa de mortalidad y la Zona Sur del país, donde se encuentra un mayor promedio de ancestría Mapuche.

Cuando se realizó la selección de los SNPs de riesgo que son más predictivos para nuestra población a través de una búsqueda exhaustiva de modelos de riesgo, se seleccionaron solamente rs2294008, rs4072037 y rs12693932, indicando que no son necesarios todos los SNPs seleccionados para tener un modelo de riesgo en esta muestra caso-control. El PRS fue una variable genética que mejora el ajuste de todos los modelos logísticos, y su OR más alto es en el modelo de interacción PRSxMAP (OR = 2.21). Es decir, que la proporción de ancestría Mapuche modifica el efecto del riesgo que confiere el PRS. Este resultado podría indicar que el PRS incluye variantes que tienen un efecto de epistasis con variantes que tienen distinta frecuencia en las poblaciones Europea y Mapuche. Esto es concordante con los resultados obtenidos en el análisis de correlación de PRS con ancestría Mapuche tanto en la muestra de ChileGenómico como en la de casos de CG. Este tipo de relaciones ya se ha descrito en otros tipos de cáncer, como el cáncer de mama, donde el estudio de Torres *et al* (2018) reportó que la proporción de ancestría Nativa Americana podría modificar el efecto que confiere variables de susceptibilidad comunes, y que esto podría predecir el riesgo del cáncer de mama tan preciso como factores de riesgo descritos como la historia familiar de éste cáncer<sup>83</sup>. Esto sustenta que la ancestría mejora los modelos de predicción de riesgo en poblaciones mestizas y, además, la importancia de evaluar marcadores genéticos en estas poblaciones.

La inferencia de la ancestría del locus de cinco variantes de riesgo en la muestra caso-control mostró que no hay diferencias significativas en la ancestría entre casos y controles, y dado que observamos que la variable de ancestría Mapuche modifica el efecto del PRS, fue importante determinar si existen otras variables presente en regiones de esta ancestría que puedan estar asociadas al CG. Para evaluar esto se utilizó Admixture Mapping (AM), el cual es un método utilizado en poblaciones mestizas para correlacionar la ancestría de un locus con un fenotipo<sup>84</sup>. Se buscó encontrar una sobrerrepresentación de la ancestría del locus de una población parental respecto al promedio muestral<sup>58</sup>. Esto fue evaluado con las

ancestría Europea y Amerindia, lo que no fue observado en regiones asociadas a la primera. En cambio, se encontraron tres regiones cromosómicas sugerentes de asociación ubicados en 8p23.1, 13q33 y 20q13.32 asociados a ancestría Amerindia. Anteriormente, se identificaron aberraciones en el número de copias en los mismos cromosomas en CG, junto que más de un 70% de los adenocarcinomas tienen una ganancia en los genes *STX16* y *GNAS*, los cuales también fueron identificados en el análisis de Admixture Mapping<sup>85</sup>. De los 26 genes identificados, han sido relacionados con el cáncer: *GNAS-AS1* con cáncer de mama<sup>86</sup>; *BLK* con leucemia linfocítica crónica<sup>87</sup>; *LINC00208* esofágico y esófago de Barrett<sup>88</sup>; *MTMR9* y *TDH* con el cáncer esofágico<sup>89,90</sup>. Aquellos que han sido relacionados específicamente con el CG son: *GNAS*, *STX16-NPEPL1*, *MIR296* y *MIR298*<sup>91-95</sup>. El gen *GNAS* codifica a la subunidad alfa s de la proteína G, que está involucrada en muchos procesos celulares, los que incluyen la proliferación celular y supervivencia celular, los cuales tienen un rol importante en el crecimiento tumoral, angiogénesis y la diseminación metastásica<sup>94</sup>. *MIR296* es un microRNA tiene resultados contradictorios en el CG, donde se ha reportado que puede tener un rol como supresor de tumor y de promover el crecimiento celular. Sin embargo, se ha observado que se encuentra elevado en pacientes de CG chinos, por lo que podría utilizarse como un biomarcador para este cáncer<sup>92</sup>. *MIR298* es otro microRNA cuyo blanco es el gen *BAX*, el cual regula la proliferación celular. Se ha descrito que este un tratamiento para el CG, bufalina, tiene como blanco a este microRNA, permitiendo que la proteína BAX cumpla su rol estimulando el proceso de apoptosis<sup>93</sup>. Finalmente, *STX16-NPEPL1* se transcribe a un lncRNA, RNA no codificante largo, cuyo rol tumoral no ha sido bien definido. Sin embargo, no hay estudios donde estos genes hayan sido asociados con el CG, es decir que tenga un valor de riesgo, ni un riesgo asociado a la ancestría. Esto implica que tanto las variantes genéticas que han identificadas en estudios de CG como las que aún no han sido relacionadas al CG, podrían estar asociadas al riesgo de CG en nuestra población, y no han sido evaluadas debido a que no se tiene la suficiente evidencia en otras poblaciones. Esto



demuestra que es necesario poder identificar variantes genéticas, descritas o no, enriquecidas en regiones cromosómicas asociadas a la ancestría Amerindia y que podrían ser factores de riesgo para el CG.

Si bien es cierto que el valor de  $p$  que se está utilizando como umbral para determinar si una variante genética identificada por AM es significativa fue obtenido desde un estudio realizado en latinoamericanos<sup>75</sup>, la muestra son individuos que residen en Estado Unidos, Puerto Rico y México. Éstos pertenecen a poblaciones que tienen marcadores de ancestría Amerindia parecidos a la nuestra, pero la diferencia reside en que tenemos distintas ascendencias indígenas y, por otro lado, la asociación encontrada entre ancestría y PRS fue significativa con la ancestría Mapuche, que es más difícil realizar la inferencia de esta ancestría en un locus. Esto puede estar generando un sesgo en los resultados encontrados por AM, lo que podría verse evidenciado en variantes genéticas que no están siendo identificadas que podrían estar asociadas a la ancestría Mapuche. Otra limitación, es que el valor de  $p$  esté siendo subestimado, por lo que no se están tomando en cuenta asociaciones que serían significativas en nuestra población, o que se esté sobreestimado y se necesita un valor de  $p$  menor para identificar SNPs asociados a esta ancestría.

El presente trabajo de tesis sugiere la necesidad de realizar estudios de asociación genética a nivel genómico en nuestra población mestiza, debido a que este estudio presenta limitaciones: 1) se analizaron solo algunas los SNPs que han sido asociados al CG en otras poblaciones, por lo que podrían existir variantes genéticas que están asociadas en nuestra población que no fueron analizadas; 2) es necesario un mayor tamaño muestral para replicar las asociaciones pequeñas de cada SNP.

## VII. CONCLUSIONES

Este es el primer estudio de riesgo genético que busca SNPs que estén enriquecidos en regiones genómicas de ancestría Amerindia en casos de CG. A partir de nuestros resultados, podemos concluir que: 1) el riesgo poligénico para CG de los chilenos, estimado con 13 variantes genéticas asociados en la literatura, está asociado positivamente con las ancestrías Amerindia y Mapuche; 2) el puntaje de riesgo poligénico calculado con estos marcadores también estuvo asociado con riesgo de CG en la muestra caso-control, y su grado de asociación fue mayor para individuos con más ancestría Mapuche; 3) tres de los once SNPs evaluados fueron seleccionados como predictores de riesgo en los modelos logísticos, y la interacción entre el PRS y la ancestría Mapuche indica que pueden existir variantes genéticas que tienen un efecto de epistasia con otras que tienen una diferencia en su frecuencia entre las poblaciones ancestrales Europea y Mapuche; y 4) los pacientes chilenos de CG presentan evidencia de un enriquecimiento de la ancestría amerindia en tres loci, en los cromosomas 8, 13 y 20, sugiriendo presencia de variación genética para riesgo de CG asociada a este componente ancestral.

Dentro de las variantes genéticas identificadas dentro de estas regiones cromosómicas no se han descrito asociaciones con el CG. Por lo tanto, nuestros resultados podrían guiar a un nuevo trabajo con más individuos para asociar marcadores genéticos para crear modelos predictivos de riesgo de CG en la población chilena. Estos modelos podrían usarse como parte de una medida preventiva que permita realizar un seguimiento de individuos que tienen un alto riesgo para desarrollar este cáncer. Esto también podría utilizarse para realizar un diagnóstico en estadios tempranos del CG, lo que mejoraría significativamente la supervivencia de estos pacientes. Es decir, el poder identificar de manera más exacta marcadores genéticos asociados a la ancestría Amerindia, y ojalá a la

ancestría Mapuche, podría crear un modelo de riesgo que podría incluirse en la guía clínica del Minsal para el CG, significando un ahorro considerable en tratamientos, al mismo tiempo que se evita que la calidad de vida de un paciente se vea mermada por el paso por esta enfermedad.

Futuros estudios con mayor tamaño muestral serán necesarios para replicar estas asociaciones en nuestra población y para realizar un análisis de mapeo fino de la variación genética en los loci identificados en este trabajo<sup>96</sup>.

## VIII. REFERENCIAS

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
2. Freedman, N. D. *et al.* Male predominance of upper gastrointestinal adenocarcinoma cannot be explained by differences in tobacco smoking in men versus women. *Eur. J. Cancer* **46**, 2473–2478 (2010).
3. Ajani, J. A. *et al.* Gastric adenocarcinoma. *Nat. Rev. Dis. Primer* **3**, 17036 (2017).
4. Correa, P. Gastric Cancer. *Gastroenterol. Clin. North Am.* **42**, 211–217 (2013).
5. Park, J. *et al.* Prognostic Factors for Advanced Gastric Cancer : Stage-stratified Analysis of Patients who Underwent Curative Resection. *Cancer Res Treat* **38**, 13–18 (2006).
6. Karimi, P., Islami, F., Anandasabapathy, S., Freedman, N. D. & Kamangar, F. Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and Prevention. *Cancer Epidemiol. Biomarkers Prev.* **23**, 700–713 (2014).
7. Cho, B. Y. Evaluation of the validity of current national health screening programs and plans to improve the system. *Univ. Seoul* 741–58 (2014).
8. Waddell, T. *et al.* Gastric cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **24**, (2013).
9. Guggenheim, D. E. & Shah, M. A. Gastric cancer epidemiology and risk factors: Gastric Cancer, Epidemiology. *J. Surg. Oncol.* **107**, 230–236 (2013).
10. Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K. & Prenen, H. Gastric cancer. *The Lancet* **388**, 2654–2664 (2016).
11. Torres, J. *et al.* Gastric cancer incidence and mortality is associated with altitude in

- the mountainous regions of Pacific Latin America. *Cancer Causes Control* **24**, 249–256 (2013).
12. Caglevic, C., Silva, S., Mahave, M., Rolfo, C. & Gallardo, J. The current situation for gastric cancer in Chile. *Ecancermedicalscience* **10**, 8–15 (2016).
  13. DEIS. DEIS (2015) Serie de Mortalidad Observada y Ajustada por Tumores Malignos, Según Región. (2015).
  14. Lauren, P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. *Acta Pathol Microbiol Scand* **64**, 31–49 (1965).
  15. Marqués-Lespier, J. M., González-Pons, M. & Cruz-Correa, M. Current Perspectives on Gastric Cancer. *Gastroenterol. Clin. North Am.* **45**, 413–428 (2016).
  16. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
  17. Liu, X. & Meltzer, S. J. Gastric Cancer in the Era of Precision Medicine. *Cmgh* **3**, 348–358 (2017).
  18. Kim *et al.* Effect of Red, Processed, and White Meat Consumption on the Risk of Gastric Cancer: An Overall and Dose–Response Meta-Analysis. *Nutrients* **11**, 826 (2019).
  19. Dong, J. & Thrift, A. P. Alcohol, smoking and risk of oesophago-gastric cancer. *Best Pract. Res. Clin. Gastroenterol.* **31**, 509–517 (2017).
  20. Terry, P., Nyren, O. & Yuen, J. Protective effect of fruits and vegetables on stomach cancer in a cohort of Swedish twins. *Int. J. Cancer* **76**, 35–37 (1998).
  21. Rawla, P. & Barsouk, A. Epidemiology of gastric cancer: global trends, risk factors and prevention. *Gastroenterol. Rev.* **14**, 26–38 (2019).
  22. Vinceti, M. *et al.* Selenium for preventing cancer. *Cochrane Database Syst. Rev.* (2018) doi:10.1002/14651858.CD005195.pub4.

23. Yusefi, A. R., Bagheri Lankarani, K., Bastani, P., Radinmanesh, M. & Kavosi, Z. Risk Factors for Gastric Cancer: A Systematic Review. *Asian Pac. J. Cancer Prev.* **19**, (2018).
24. Yoon, H. & Kim, N. Diagnosis and Management of High Risk Group for Gastric Cancer. *Gut Liver* **9**, 5–17 (2015).
25. Ruíz-García, E. *et al.* Gastric cancer in Latin America. *Scand. J. Gastroenterol.* **53**, 124–129 (2018).
26. Figueiredo, C., Costa, S., Karameris, A. & Machado, J. C. Pathogenesis of Gastric Cancer. *Helicobacter* **20**, 30–35 (2015).
27. Mommersteeg, M. C., Yu, J., Peppelenbosch, M. P. & Fuhler, G. M. Genetic host factors in Helicobacter pylori-induced carcinogenesis: Emerging new paradigms. *Biochim. Biophys. Acta - Rev. Cancer* **1869**, 42–52 (2018).
28. Sampson, J. N. *et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *J. Natl. Cancer Inst.* **107**, djv279 (2015).
29. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68 (2016).
30. Gigeek, C. O. *et al.* Genetic variants in gastric cancer: Risks and clinical implications. *Exp. Mol. Pathol.* **103**, 101–111 (2017).
31. Sun, H. *et al.* Associations of Genetic Variants in the PSCA, MUC1 and PLCE1 Genes with Stomach Cancer Susceptibility in a Chinese Population. *PLOS ONE* **10**, e0117576 (2015).
32. Mocellin, S., Verdi, D., Pooley, K. A. & Nitti, D. Genetic variation and gastric cancer risk: A field synopsis and meta-analysis. *Gut* **64**, 1209–1219 (2015).
33. Zabaleta, J. Multifactorial Etiology of Gastric Cancer. in *Cancer Epigenetics* (eds.

- Dumitrescu, R. G. & Verma, M.) vol. 863 411–435 (Humana Press, 2012).
34. Wang, T., Zhang, L., Li, H., Wang, B. & Chen, K. Prostate Stem Cell Antigen Polymorphisms and Susceptibility to Gastric Cancer: A Systematic Review and Meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **21**, 843–850 (2012).
  35. Gonzalez-Hormazabal, P. *et al.* Polymorphisms in RAS/RAF/MEK/ERK Pathway Are Associated with Gastric Cancer. *Genes* **10**, 20 (2018).
  36. Barnholtz-Sloan, J. S., McEvoy, B., Shriver, M. D. & Rebbeck, T. R. Ancestry Estimation and Correction for Population Stratification in Molecular Epidemiologic Association Studies. *Cancer Epidemiol. Biomarkers Prev.* **17**, 471–477 (2008).
  37. Ziyatdinov, A. *et al.* Mixed-model admixture mapping identifies smoking-dependent loci of lung function in African Americans. *Eur. J. Hum. Genet.* (2019)  
doi:10.1038/s41431-019-0545-8.
  38. Adhikari, K., Chacón-Duque, J. C., Mendoza-Revilla, J., Fuentes-Guajardo, M. & Ruiz-Linares, A. The Genetic Diversity of the Americas. *Annu. Rev. Genomics Hum. Genet.* **18**, 277–296 (2017).
  39. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of african americans, latinos, and european Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
  40. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet.* **11**, 1–26 (2015).
  41. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. *Curr. Opin. Genet. Dev.* **41**, 106–114 (2016).
  42. Fuentes, M. *et al.* Geografía génica de Chile. Distribución regional de los aportes genéticos Americanos, Europeos y Africanos. *Rev. Med. Chil.* **142**, 281–289 (2014).

43. Eyheramendy, S., Martinez, F. I., Manevy, F., Vial, C. & Repetto, G. M. Genetic structure characterization of Chileans reflects historical immigration patterns. *Nat. Commun.* **6**, 6472 (2015).
44. Stern, M. C. *et al.* Variability in Cancer Risk and Outcomes Within US Latinos by National Origin and Genetic Ancestry. *Curr. Epidemiol. Rep.* **3**, 181–190 (2016).
45. Lorenzo Bermejo, J. *et al.* Subtypes of Native American ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in Chile. *PLoS Genet.* **13**, e1006756 (2017).
46. Pereira, L. *et al.* Socioeconomic and nutritional factors account for the association of gastric cancer with amerindian ancestry in a latin american admixed population. *PLoS ONE* **7**, (2012).
47. Arnold, M. *et al.* The burden of stomach cancer in indigenous populations: a systematic review and global assessment. *Gut* **63**, 64–71 (2014).
48. Zamudio, R. *et al.* Population, Epidemiological, and Functional Genetics of Gastric Cancer Candidate Genes in Peruvians with Predominant Amerindian Ancestry. *Dig. Dis. Sci.* **61**, 107–116 (2016).
49. Liu, X. *et al.* MUC1 gene polymorphism rs4072037 and susceptibility to gastric cancer: a meta-analysis. *SpringerPlus* **3**, 599 (2014).
50. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
51. Christiansen, M. K. *et al.* A 45-SNP genetic risk score is increased in early-onset coronary artery disease but independent of familial disease clustering. *Atherosclerosis* **257**, 172–178 (2017).
52. Harlid, S. *et al.* Combined effect of low-penetrant SNPs on breast cancer risk. *Br. J.*



- Cancer* **106**, 389–396 (2012).
53. Choi, J., Jia, G., Wen, W., Long, J. & Zheng, W. Evaluating polygenic risk scores in assessing risk of nine solid and hematologic cancers in European descendants. *Int. J. Cancer* **ijc.33176** (2020) doi:10.1002/ijc.33176.
  54. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
  55. Marigorta, U. M., Rodríguez, J. A., Gibson, G. & Navarro, A. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
  56. Shriner, D., Adeyemo, A., Ramos, E., Chen, G. & Rotimi, C. N. Mapping of disease-associated variants in admixed populations. *Genome Biol.* **12**, 223 (2011).
  57. Yao, Y. *et al.* Genome-wide analysis of Epstein-Barr virus identifies variants and genes associated with gastric carcinoma and population structure. *Tumor Biol.* **39**, 101042831771419 (2017).
  58. Parra, E. J. *et al.* Admixture mapping in two Mexican samples identifies significant associations of locus ancestry with triglyceride levels in the BUD13/ZNF259/APOA5 region and fine mapping points to rs964184 as the main driver of the association signal. *PLOS ONE* **12**, e0172880 (2017).
  59. Galanter, J. M. *et al.* Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. *J. Allergy Clin. Immunol.* **134**, 295–305 (2014).
  60. Uribe-Salazar, J. M., Palmer, J. R., Haddad, S. A., Rosenberg, L. & Ruiz-Narváez, E. A. Admixture mapping and fine-mapping of type 2 diabetes susceptibility loci in African American women. *J. Hum. Genet.* **63**, 1109–1117 (2018).
  61. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide

- association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, (2019).
62. Verdugo, R. A. *et al.* Development of a small panel of SNPs to infer ancestry in Chileans that distinguishes Aymara and Mapuche components. *Biol. Res.* **53**, 15 (2020).
63. Campbell, N. R., Harmon, S. A. & Narum, S. R. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* **15**, 855–867 (2015).
64. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
65. Thornton, T. *et al.* Estimating Kinship in Admixed Populations. *Am. J. Hum. Genet.* **91**, 122–138 (2012).
66. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
67. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
68. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
69. Martin, A. R. & Maihofer, A. X. Ancestry pipeline. *armartin/ancestry\_pipeline* [https://github.com/armartin/ancestry\\_pipeline](https://github.com/armartin/ancestry_pipeline) (2017).
70. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
71. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring

- population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1. *Bioinformatics* **31**, 3555–3557 (2015).
72. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
  73. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
  74. Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. Design and Analysis of Admixture Mapping Studies. *Am. J. Hum. Genet.* **74**, 965–978 (2004).
  75. Torgerson, D. G. *et al.* Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J. Allergy Clin. Immunol.* **130**, 76-82.e12 (2012).
  76. Zhu, X., Tang, H. & Risch, N. Admixture Mapping and the Role of Population Structure for Localizing Disease Genes. in *Advances in Genetics* vol. 60 547–569 (Elsevier, 2008).
  77. Dayem Ullah, A. Z. *et al.* SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* **46**, W109–W113 (2018).
  78. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
  79. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **45**, D183–D189 (2017).
  80. Royal, C. D. *et al.* Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. *Am. J. Hum. Genet.* **86**, 661–673 (2010).
  81. Bonequi, P., Meneses-González, F., Correa, P., Rabkin, C. S. & Camargo, M. C. Risk

- factors for gastric cancer in Latin America: a meta-analysis. *Cancer Causes Control* **24**, 217–231 (2013).
82. Lott, P. C. & Carvajal-Carmona, L. G. Resolving gastric cancer aetiology: an update in genetic predisposition. *Lancet Gastroenterol. Hepatol.* **3**, 874–883 (2018).
83. Torres, D. *et al.* Interaction between genetic ancestry and common breast cancer susceptibility variants in Colombian women: Interaction between genetic ancestry and common breast cancer susceptibility variants. *Int. J. Cancer* **144**, 2181–2191 (2019).
84. Shriner, D. Overview of Admixture Mapping. *Curr. Protoc. Hum. Genet.* **76**, 1.23.1-1.23.8 (2013).
85. Buffart, T. E. *et al.* High resolution analysis of DNA copy-number aberrations of chromosomes 8, 13, and 20 in gastric cancers. *Virchows Arch.* **455**, 213–223 (2009).
86. Liu, S.-Q., Zhou, Z.-Y., Dong, X., Guo, L. & Zhang, K.-J. LncRNA GNAS-AS1 facilitates ER+ breast cancer cells progression by promoting M2 macrophage polarization via regulating miR-433-3p/GATA3 axis. *Biosci. Rep.* **40**, BSR20200626 (2020).
87. Lukas, M. *et al.* Survey of ex vivo drug combination effects in chronic lymphocytic leukemia reveals synergistic drug effects and genetic dependencies. *Leukemia* (2020) doi:10.1038/s41375-020-0846-5.
88. Gharahkhani, P. *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol.* **17**, 1363–1373 (2016).
89. Goh, X. Y. *et al.* Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut* **60**, 1317–1326 (2011).
90. Jia, Y., Hu, X., Gabriella, H., Qin, L. & Meggyeshazi, N. Antitumor Activity of

Tenacissoside H on Esophageal Cancer through Arresting Cell Cycle and Regulating PI3K/Akt-NF- $\kappa$ B Transduction Cascade. *Evid. Based Complement. Alternat. Med.* **2015**, 1–10 (2015).

91. Kang, G. *et al.* Integrated genomic analyses identify frequent gene fusion events and VHL inactivation in gastrointestinal stromal tumors. *Oncotarget* **7**, 6538–6551 (2016).
92. Huang, Z. *et al.* Six Serum-Based miRNAs as Potential Diagnostic Biomarkers for Gastric Cancer. *Cancer Epidemiol. Biomarkers Prev.* **26**, 188–196 (2017).
93. Zhao, H. *et al.* Bufalin promotes apoptosis of gastric cancer by down-regulation of miR-298 targeting bax. *Int. J. Clin. Exp. Med.* **8**, 3420–3428 (2015).
94. Alakus, H. *et al.* Association of the GNAS1 T393C polymorphism with tumorstage and survival in gastric cancer. *World J. Gastroenterol.* **15**, 6061 (2009).
95. Matsubara, A. *et al.* Activating GNAS and KRAS mutations in gastric foveolar metaplasia, gastric heterotopia, and adenocarcinoma of the duodenum. *Br. J. Cancer* **112**, 1398–1404 (2015).
96. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* (2018)  
doi:10.1038/s41576-018-0016-z.

## IX. PUBLICACIONES GENERADAS

### Presentaciones en congresos:

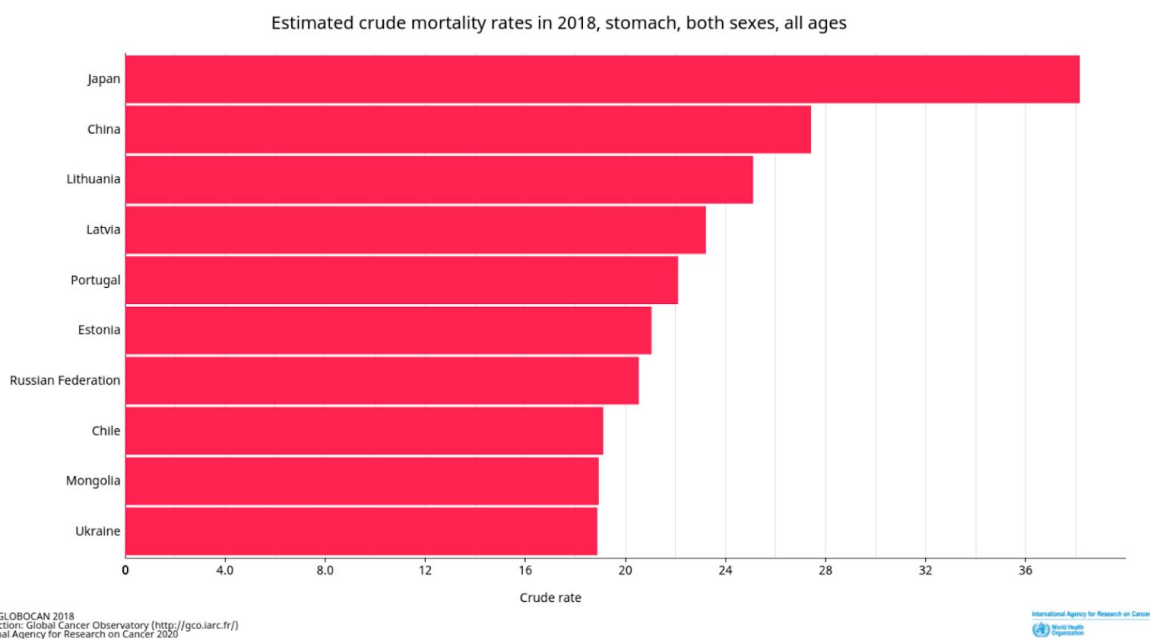
- Genotipificación por secuenciación de un panel de ancestría y riesgo genético en 2000 participantes del Proyecto ChileGenómico. Asociación Latinoamericana de Genética, Argentina, Octubre 2019.
- Puntaje de Riesgo Poligénico está relacionado a la Ancestría Mapuche en una muestra caso-control de Cáncer Gástrico en Chile. UGenoma. Julio 2019.
- Genetic Polymorphisms as risk factors for gastric cancer. Digestive Disease Week, Estados Unidos, Mayo 2019.
- Gastric Cancer's Polygenic Risk Score calculated with genetics variants described in other populations are not related with Mapuche Ancestry in the Chilean population. LI Congreso Sociedad Chilena de Genética, Chile 2018.

### Publicaciones:

- Polymorphisms PSCA rs2294008, IL-4 rs2243250 and MUC1 rs4072037 are associated with gastric cancer in a high risk population. Patricio Gonzalez-Hormazabal, *Rocío Retamales-Ortega*, Maher Musleh, Marco Bustamante, Juan Stambuk, Raul Pisano, Hector Valladares, Enrique Lanzarini, Hector Chiong, Jose Suazo, Luis A Quiñones, Nelson M Varela, V Gonzalo Castro, Lilian Jara, Ricardo A Verdugo, Zoltan Berger. Mol Biol Rep. 2020 Nov;47(11):9239-9243. doi: 10.1007/s11033-020-05943-0.

- Se está escribiendo un manuscrito para una publicación que incluye datos del objetivo 1 de esta tesis.

## X. ANEXOS



**Figura suplementaria 1.** Primeros 10 países con la mayor tasa de mortalidad cruda para el Cáncer Gástrico para ambos sexos, todas las edades en el 2018. Datos obtenidos de la Agencia Internacional para la Investigación del Cáncer (<https://gco.iarc.fr/today/home>).



**Table 1** Meta-analysis results: genetic variants significantly associated with gastric cancer risk with a high or intermediate level of summary evidence

Gene	Variant ID	Chr	Risk Allele	Datasets	Cases	Controls	Subgroup	sOR	LL	UL	p Value	Venice criteria	Level of evidence	FPRP (10E-3)	FPRP (10E-6)
MUC1	rs2070803	1	A	6	2376	4903	Diffuse	0.585	0.526	0.650	<1.00E-20	AAA	High	0.00	0.00
MTX1	rs2075570	1	G	5	2299	4603	Diffuse	0.587	0.525	0.655	<1.00E-20	AAA	High	0.00	0.00
PRKAA1	rs13361707	5	C	5	7587	7891	Non-cardia	1.406	1.334	1.482	<1.00E-20	AAA	High	0.00	0.00
PLCE1	rs2274223	10	G	5	4869	16 497	Cardia	1.565	1.486	1.648	<1.00E-20	AAA	High	0.00	0.00
PSCA	rs2294008	8	T	7	5690	6975	Non-cardia	1.329	1.254	1.408	<1.00E-20	AAA	High	0.00	0.00
TGFBR2	rs3087465	3	A	4	2323	2347	Asian	0.691	0.620	0.769	1.54E-11	AAA	High	0.00	0.53
PKLR	rs3762272	1	G	3	1362	3101	Diffuse	0.710	0.631	0.799	1.42E-08	AAA	High	0.03	0.97
PSCA	rs2976392	8	G	4	2602	4443	Intestinal	0.808	0.751	0.870	1.42E-08	AAA	High	0.00	0.36
GSTP1	rs1695	11	G	11	2425	4731	Asian	1.191	1.092	1.299	7.88E-05	AAA	High	0.12	0.99
CASP8	rs3834129	2	del	3	616	1085	Mixed	0.732	0.617	0.869	3.48E-04	AAA	High	0.93	1.00
TNF	rs1799724	6	T	10	2136	3817	Mixed	1.173	1.047	1.314	0.006	AAA	High	0.89	1.00
PSCA	rs2294008	8	T	13	7380	26 388	Intestinal	1.284	1.206	1.367	5.33E-15	ABA	Intermediate	0.00	0.00
TRIM46	rs3814316	1	C	3	1362	3101	Diffuse	0.582	0.506	0.670	3.71E-14	AAB	Intermediate	0.00	0.79
THBS3	rs2066981	1	C	3	1362	3101	Diffuse	0.585	0.507	0.674	1.57E-13	AAB	Intermediate	0.00	0.22
ABO	rs8176719	9	A (vs O)	15	8737	1 069 256	Caucasian	1.225	1.161	1.293	1.92E-13	ABA	Intermediate	0.00	0.00
MUC1	rs2070803	1	A	6	3886	4903	Asian	0.660	0.590	0.738	2.55E-13	ABA	Intermediate	0.00	0.01
KRTCAP2	rs4971088	1	T	3	1362	3101	Diffuse	0.608	0.526	0.703	1.60E-11	AAB	Intermediate	0.02	0.96
MTX1	rs2075570	1	G	5	3586	4603	Asian	0.672	0.596	0.759	1.44E-10	ABA	Intermediate	0.01	0.90
PPARG	rs1805192	3	G	5	546	903	Mixed	2.265	1.695	3.027	3.20E-08	BAA	Intermediate	0.78	1.00
PPARG	rs1805192	3	G	3	227	598	Caucasian	2.149	1.536	3.007	8.03E-06	BAA	Intermediate	0.96	1.00
PSCA	rs2294008	8	T	4	1187	3162	Caucasian	1.337	1.177	1.519	8.51E-06	ABA	Intermediate	0.45	0.99
PSCA	rs2976391	8	A	4	1921	2504	Asian	0.725	0.625	0.840	2.00E-05	ABA	Intermediate	0.71	1.00
FAM189B	rs2072647	1	G	3	1362	3101	Diffuse	0.727	0.626	0.844	2.82E-05	ABA	Intermediate	0.44	0.99
PTGS2	rs20417	1	C	6	1464	3000	Asian	1.773	1.356	2.318	2.85E-05	BBB	Intermediate	0.92	1.00
THBS3	rs2066981	1	C	3	2649	3101	Asian	0.721	0.615	0.845	5.63E-05	ABA	Intermediate	0.60	0.99
IL1B	rs1143634	2	T	4	434	450	HP positive	1.721	1.321	2.243	5.84E-05	BAB	Intermediate	0.93	1.00
IL17F	rs763780	6	G	4	1760	2362	Asian	1.288	1.138	1.458	6.28E-05	AAB	Intermediate	0.63	0.99
YY1AP1	rs3738590	1	T	3	1362	3101	Diffuse	0.750	0.651	0.864	6.78E-05	ABA	Intermediate	0.77	1.00
KRTCAP2	rs4971088	1	T	3	2649	3101	Asian	0.725	0.618	0.852	9.04E-05	ABA	Intermediate	0.87	1.00
TNF	rs1800629	6	A	38	7896	12 982	Mixed	1.166	1.074	1.267	2.67E-04	ABB	Intermediate	0.26	0.99
TLR4	rs4986790	9	G	3	774	1129	Non-cardia	2.022	1.351	3.026	6.17E-04	BBA	Intermediate	0.99	1.00
MDM2	rs2279744	12	G	4	778	2272	Cardia	1.384	1.133	1.690	0.001	ABB	Intermediate	0.97	1.00
TLR4	rs4986790	9	G	8	1859	3276	Caucasian	1.475	1.159	1.878	0.002	BBB	Intermediate	0.98	1.00
TNF	rs1800629	6	A	19	3516	6604	Caucasian	1.187	1.065	1.322	0.002	ABA	Intermediate	0.76	1.00
IL1B	rs16944	2	T	9	644	4867	Cardia	1.199	1.066	1.349	0.003	AAB	Intermediate	0.83	1.00
IL8	rs4073	4	A	8	1058	2620	Intestinal	1.240	1.077	1.428	0.003	ABA	Intermediate	0.95	1.00
ABO	rs8176719	9	A (vs AB)	13	4696	620 332	Caucasian	1.234	1.068	1.427	0.004	ABB	Intermediate	0.96	1.00
TNF	rs1799724	6	T	7	1503	2517	Asian	1.189	1.043	1.355	0.010	BAA	Intermediate	0.94	1.00
TLR4	rs4986790	9	G	9	1919	3438	Mixed	1.433	1.089	1.886	0.010	BBB	Intermediate	0.99	1.00
IL10	rs1800871	1	T	3	78	399	Diffuse	0.567	0.367	0.875	0.010	BAB	Intermediate	0.99	1.00
EPHX1	rs1051740	1	C	3	427	1401	Caucasian	1.236	1.049	1.456	0.011	BAB	Intermediate	0.98	1.00
IGFBP3	rs2854744	7	C	3	1100	1338	Mixed	1.181	1.026	1.360	0.020	BAB	Intermediate	0.98	1.00
TP53	rs17878362	17	ins	3	342	635	Mixed	1.373	1.051	1.793	0.020	BAB	Intermediate	0.99	1.00
PRKAA1	rs13361707	5	C	3	1113	5261	Cardia	1.177	1.025	1.350	0.020	AAB	Intermediate	0.97	1.00
TIMP2	rs2277698	17	T	3	692	685	Mixed	0.807	0.669	0.975	0.026	BAB	Intermediate	0.99	1.00
IL8	rs2227306	4	T	4	348	1040	Asian	1.225	1.017	1.476	0.032	BAB	Intermediate	0.99	1.00
TNF	rs1800629	6	A	7	1239	4520	Cardia	1.180	1.010	1.379	0.037	BAB	Intermediate	0.98	1.00
TNF	rs1800629	6	A	8	1338	3836	Diffuse	1.262	1.010	1.578	0.041	BBB	Intermediate	0.99	1.00
KIAA0907	rs625658	1	C	3	1362	3101	Diffuse	0.842	0.713	0.993	0.042	ABB	Intermediate	0.99	1.00

Venice criteria: A (high), B (moderate), C (weak) credibility for three parameters (amount of evidence, heterogeneity and bias; see text for more details); level of evidence: overall level of summary evidence according to the Venice criteria.  
Chr, chromosome; FPRP, false positive report probability (at two levels of prior probability, see text for more details); HP, *Helicobacter pylori*; LL, 95% lower level; sOR, summary OR; UL, 95% upper level.

**Tabla Suplementaria 1.** Resultados del meta-análisis realizado por el grupo de Mocellin et al (2015): Variantes genéticas asociadas significativamente con el riesgo de cáncer gástrico con un nivel de evidencia intermedio o alto.

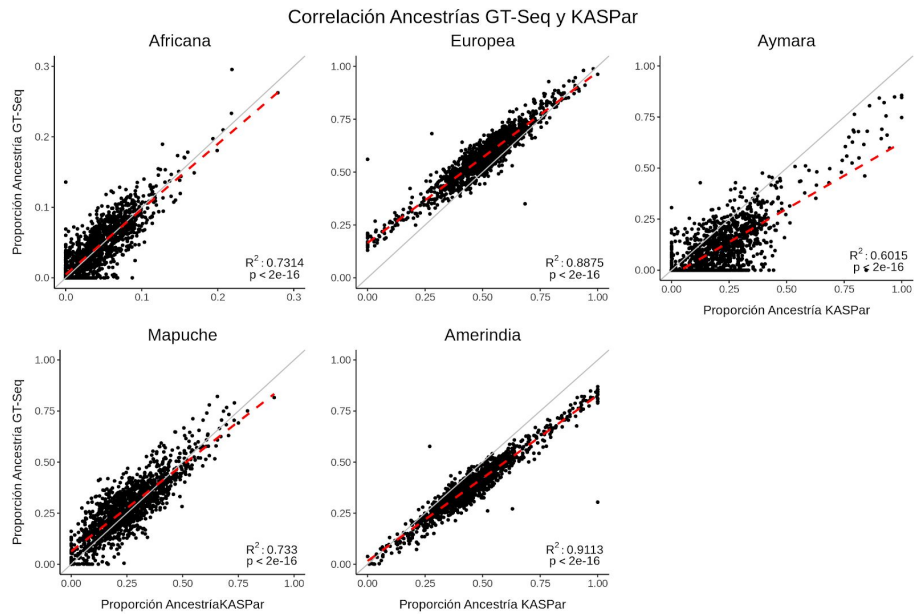
## X.1 Objetivo Específico 1

Región	Frecuencia
Nacidos en el extranjero	17
NS/NR	6
Region de Arica y Parinacota	68
Region de Tarapaca	30
Region de Antofagasta	22
Region de Atacama	14
Region de Coquimbo	76
Region de Valparaiso	26
Region Metropolitana	589
Region Lib. Gral. B. O.	34
Region del Maule	29
Region del Biobio	353
Region de la Araucania	291
Region de los Rios	24
Region de Los Lagos	102
Region de Aysen	3
Region de Magallanes	7
<b>Total</b>	<b>1691</b>

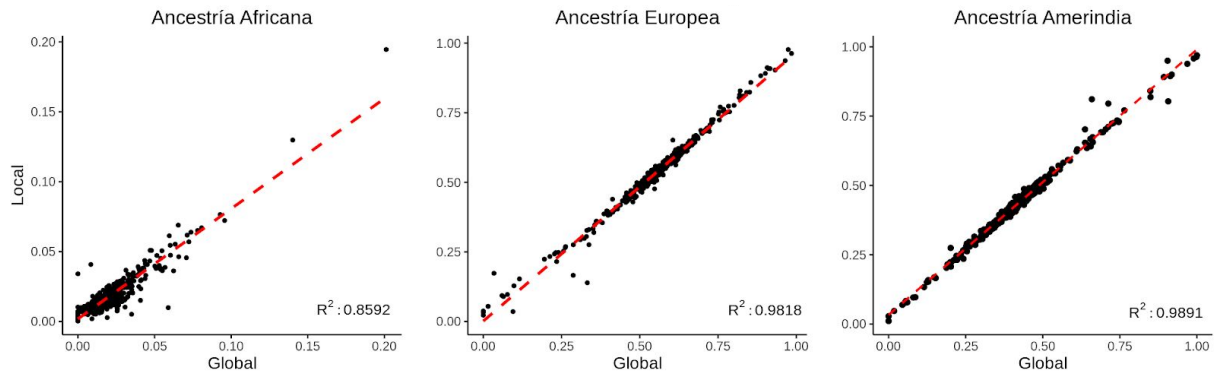
**Tabla Suplementaria 2.** Región de nacimiento de los 1691 individuos analizados que pertenecen al proyecto ChileGenómico.

Edad	
Promedio	30.9
Mínimo	17
Máximo	84
Sexo	
Hombre	802
Mujer	886

**Tabla Suplementaria 3.** Resumen edad y sexo de los 1691 individuos analizados que pertenecen al proyecto ChileGenómico.



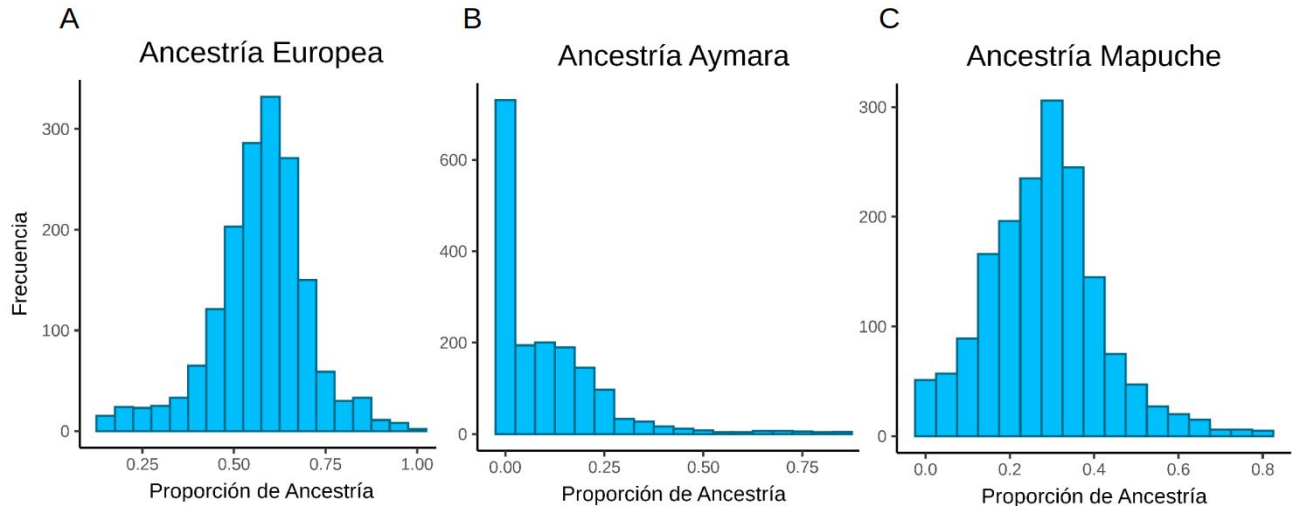
**Figura Suplementaria 2.** Regresiones de las ancestrías globales Europea, Africana, Aymara, Mapuche y Amerindia estimadas en 1465 individuos del proyecto de ChileGenómico con 137 Marcadores Informativos de Ancestría (AIMs) genotipificados con KASPar con 150 AIMs genotipificados por GT-Seq. Coeficiente de determinación:  $R^2$ ; y valor de  $p$  de  $\beta_1$  :  $p$ .



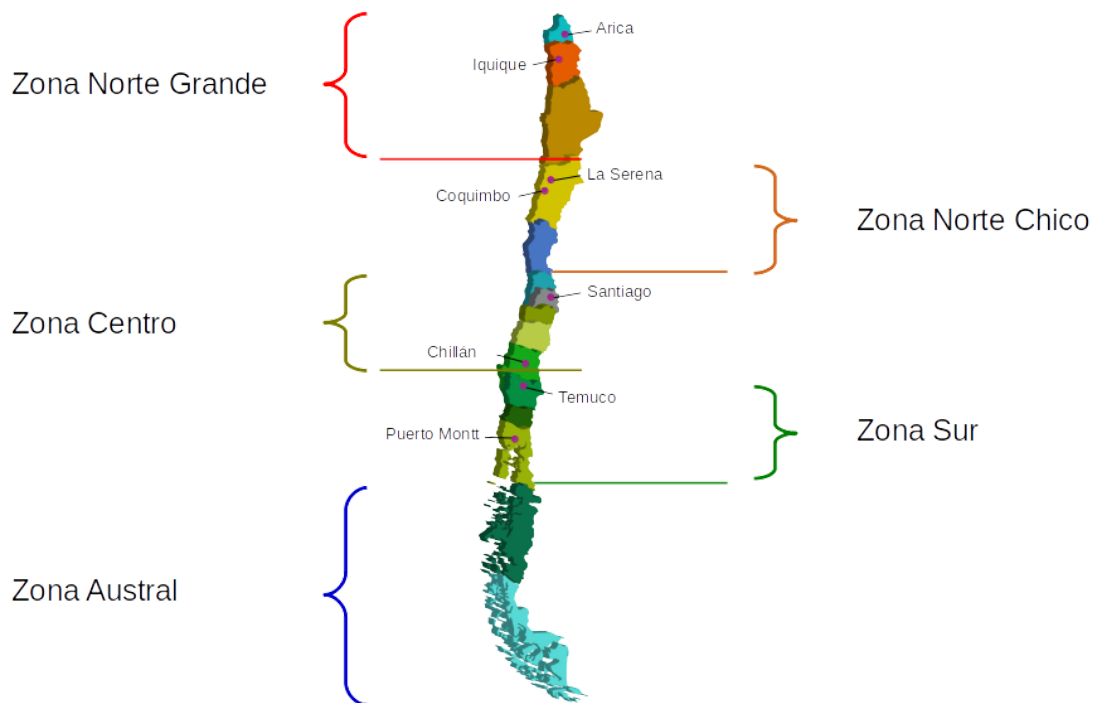
**Figura Suplementaria 3.** Regresiones entre la ancestría global estimada con el programa ADMIXTURE y ancestría global estimada desde el total de la ancestría local inferida con RFMix para las 420 muestras del proyecto de ChileGenómico. Modelos lineales con el estadístico  $R^2$ . Ancestría global (global): se estimó con el programa ADMIXTURE. Ancestría local (local): ancestría global calculada con el total de loci de cada población ancestral inferidos por ancestría local. Coeficiente de determinación:  $R^2$ .

Investigador	Selección $p < 3E^{-8}$ y $OR > 1.2$	No Diseñado	Corrida de Prueba	LD: $r^2 > 0.8$
rs4072037	rs10074991	rs760077	rs10074991	rs14007212
rs140081212	rs7712641	rs1799724	rs9841504	rs3762272
rs3762272	rs2976392			rs4072037
rs80142782	rs753724			rs70142782
rs3834129	rs11187842			rs3834129
rs3087465	rs3765524			rs3087465
rs13361707	rs2274223			rs13361707
rs2294008	rs760077			rs2910164
rs1695				rs7712641
rs1799724				rs2976392
rs2910164				rs2274223
rs3781264				rs1695
rs738722				rs738722
rs9841504				

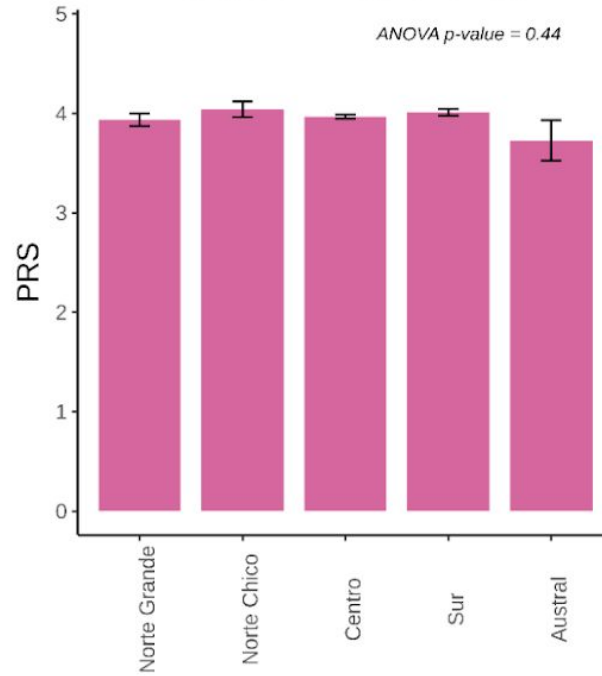
**Tabla Suplementaria 4.** Proceso de selección de SNPs de riesgo de Cáncer Gástrico para su genotipificación por medio de la técnica GT-Seq. Investigador: lista de SNPs entregadas por el investigador encargado del Cáncer Gástrico. Selección por  $p$  y OR: se agregaron SNPs que cumplieron con esos criterios. No Diseñado: SNPs cuyos partidores no fueron diseñados. Corrida de Prueba: SNPs que generaron homo o heterodímeros de partidores, por lo que fueron eliminados de la lista de SNPs seleccionados. Desequilibrio de Ligamiento (LD): Lista final de SNPs donde se eliminaron aquellos que se encontraban con un  $r^2$  mayor a 0.8.



**Figura Suplementaria 4.** Frecuencia de las proporciones de ancestrías globales Europea, Aymara y Mapuche de 1691 individuos del proyecto de ChileGenómico.

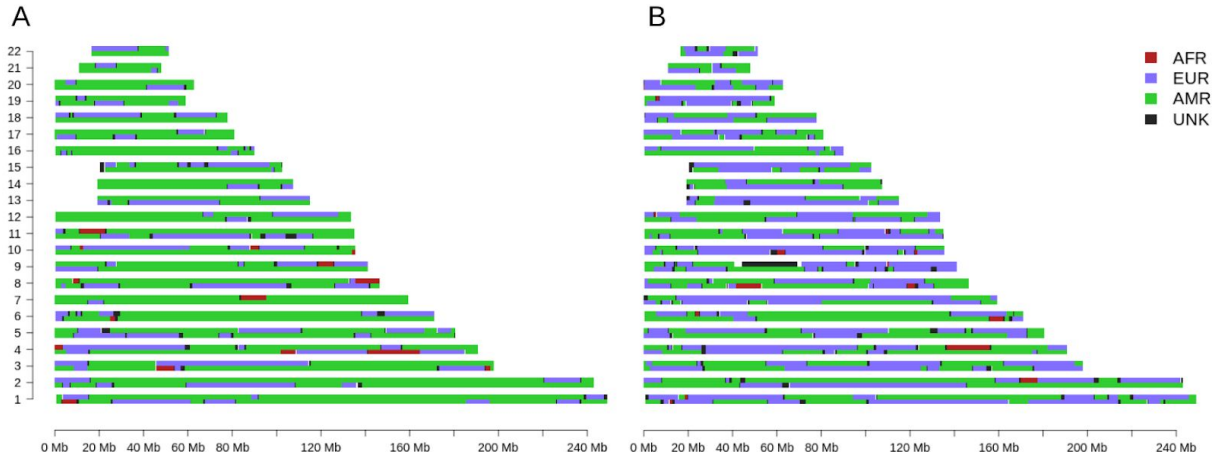


**Figura Suplementaria 5.** División de Chile por macrozonas. Las macrozonas incluyen un subgrupo de regiones políticas del país: a) Norte Grande: Arica y Parinacota (XV), Tarapacá (I) y Atacama (II); b) Norte Chico: Copiapó (III) y Coquimbo (IV); c) Centro: Valparaíso (V), Metropolitana (XIII o RM), Libertador General Bernardo O’Higgins (VI), Maule (VII), Biobío (VIII); d) Sur: Araucanía (IX), Los Ríos (XIV) y Los Lagos (X); y e) Austral: Aysén del General Carlos Ibáñez del Campo (XI) y Magallanes y la Antártida Chilena (XII).



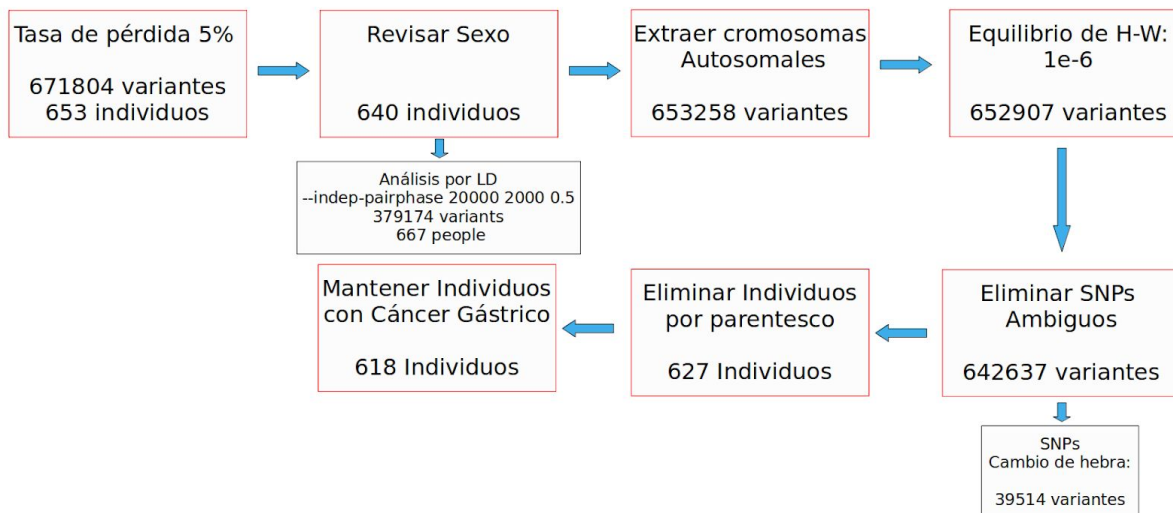
**Figura Suplementaria 6.** Puntaje de Riesgo Poligénico (PRS) en las distintas macrozonas del País. Datos representados por el promedio  $\pm$  error estándar. Prueba de ANOVA. Valor de p de la ANOVA: ANOVA p-value.





**Figura Suplementaria 7.** Representación de cromosomas mestizos con sus regiones asociadas a las ancestrías Africana: AFR, Ancestría Europea: EUR (azul), Ancestría Amerindia: AMR (verde), y Ancestría Desconocida: UNK (negro). La ancestrías locales de las regiones cromosómicas fueron inferidas con el programa RFMIX. Se presentan dos individuos: A) Individuo con 72% de ancestría Nativa y 24% de ancestría Europea; y B) Individuo con 49% de ancestría Europea y 47% de ancestría Nativa. Megabases: Mb.

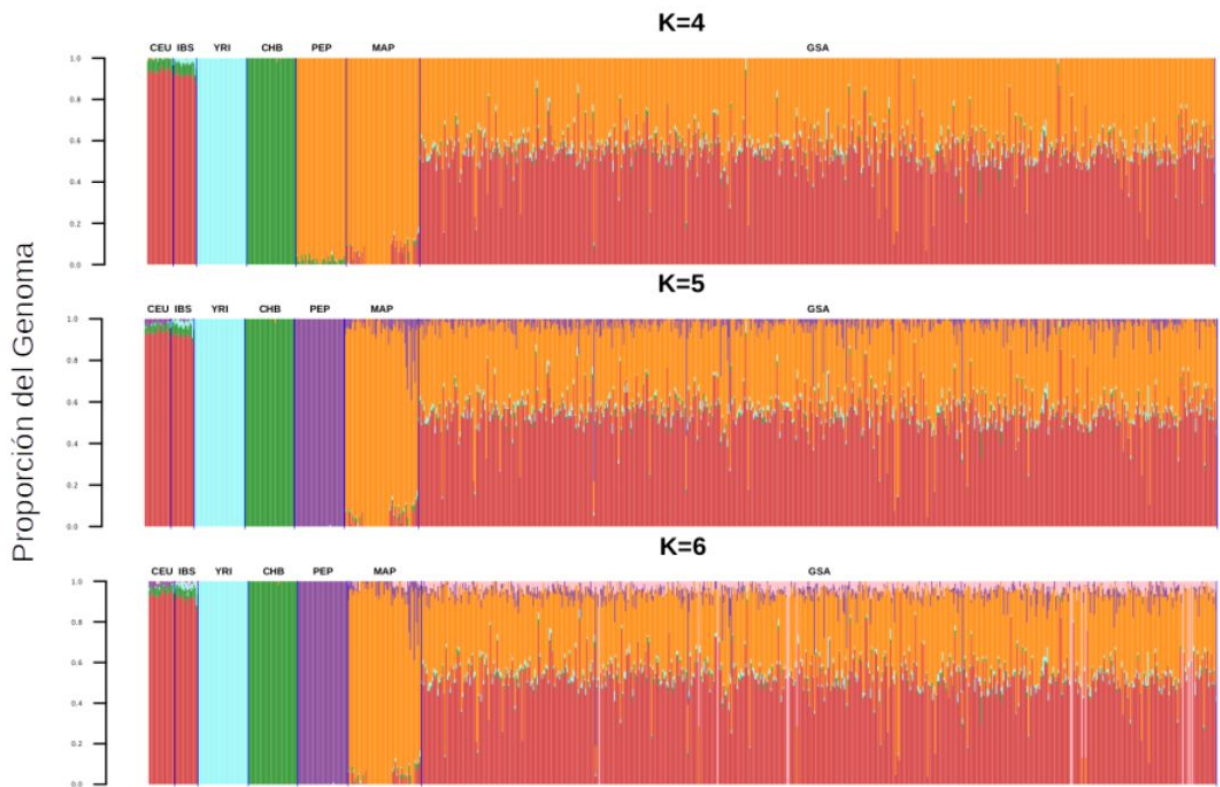
## X.2 Objetivo Específico 2



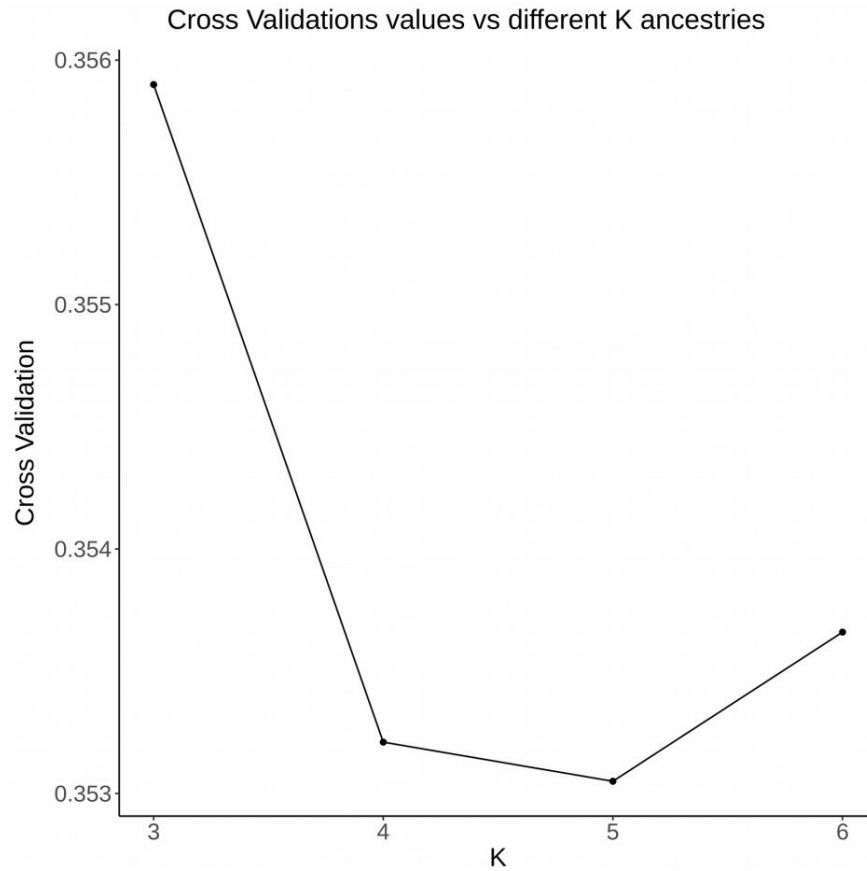
**Figura Suplementaria 8.** Flujo de trabajo del control de calidad realizado a los genotipos de la muestra caso-control obtenidos con el microarreglo Global Screening (GSA).

	Control	Caso	p
Edad			
Promedio	50.562	64.017	< 2.2e-16
EE	0.014	0.015	
Sexo			
Hombre	190 (61.7%)	193 (65.6%)	1
Mujer	118 (38.3%)	101 (34.4%)	

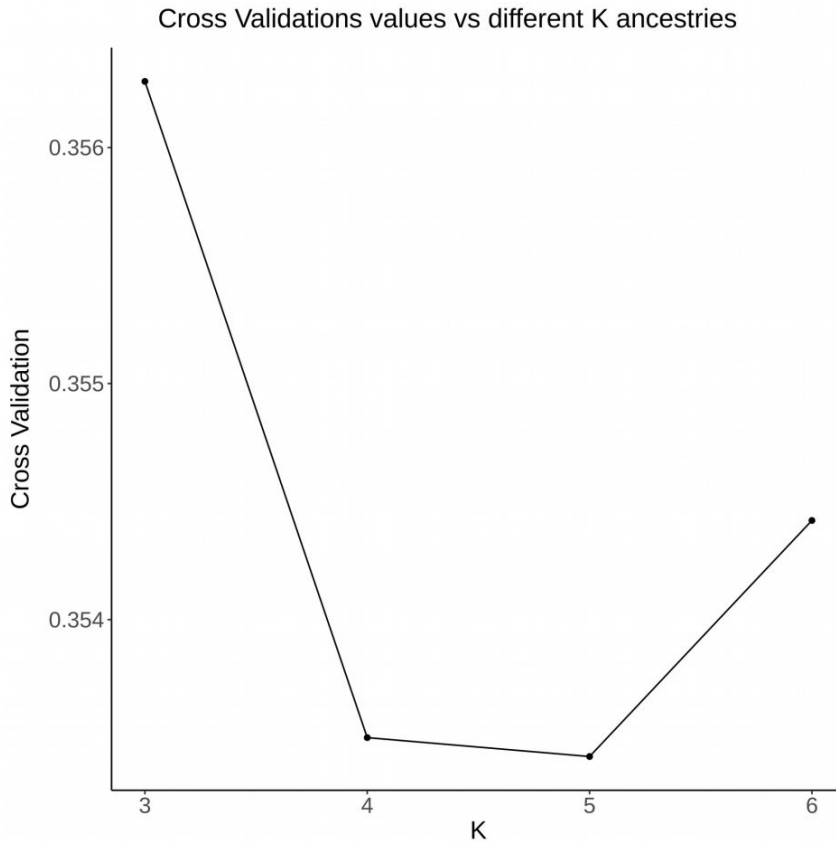
**Tabla Suplementaria 5.** Características de edad y sexo de 317 controles y 301 casos de Cáncer Gástrico. Datos representados con el número de muestras y el porcentaje del total de la muestra. Valor de p de la prueba de t-student: p.



**Figura Suplementaria 9.** Estimaciones de las Ancestrías Globales Europea (CEU e IBS), Africana (YRI), Asiática (CHB), Aymara (PEP) y Mapuche (MAP) en la muestra caso-control de Cáncer Gástrico (GSA). La muestra GSA incluye 326 controles y 316 casos de Cáncer Gástrico. Proporción del genoma: proporción de ancestría global asociada a una componente ancestral (CEU, IBS, YRI, CHB, PEP o MAP). K es el número de componentes ancestrales utilizados para estimar las ancestrías globales de cada individuo.



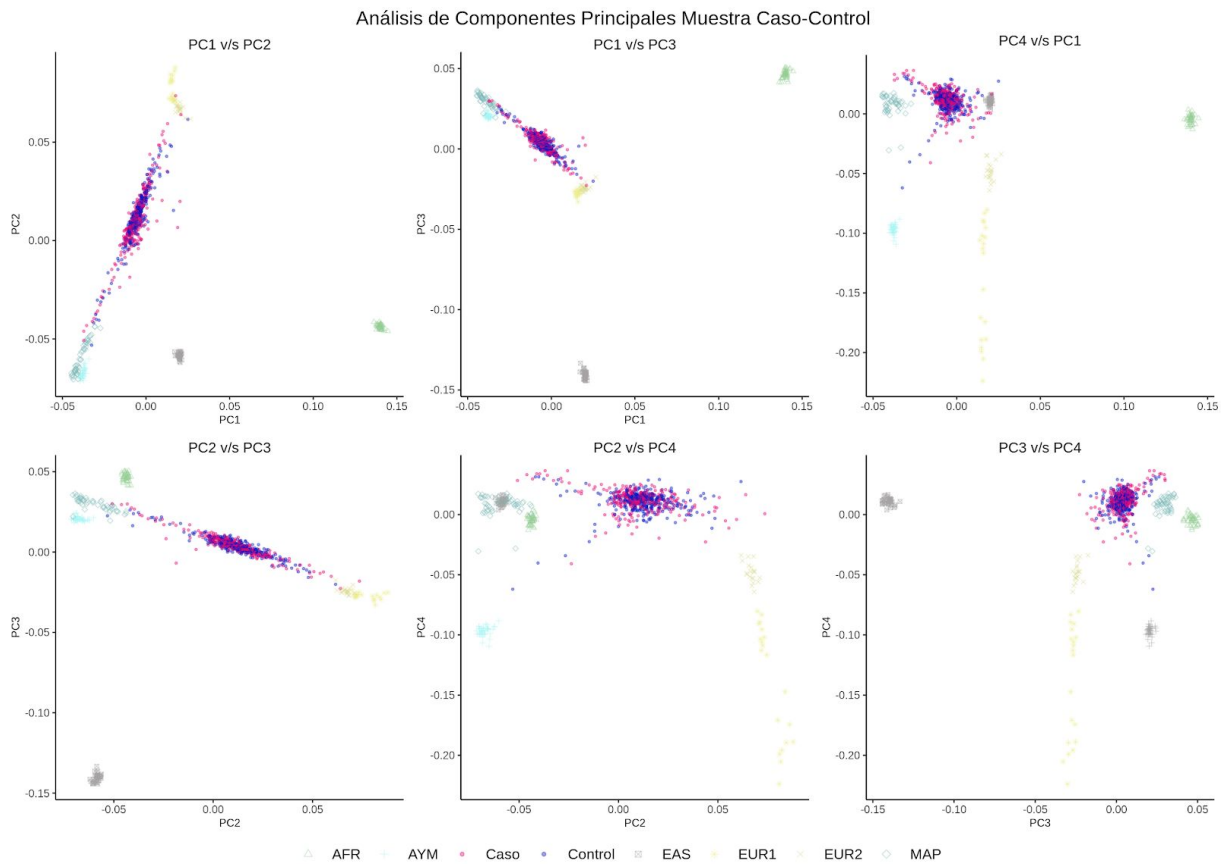
**Figura Suplementaria 10.** Error de Validación Cruzada para las estimaciones de ancestrías globales de 329 controles y 316 casos de Cáncer Gástrico. Estimación de ancestrías globales por individuo fue realizado con ADMIXTURE. K es el número de poblaciones que se utilizaron para estimar la ancestría global de cada individuo.



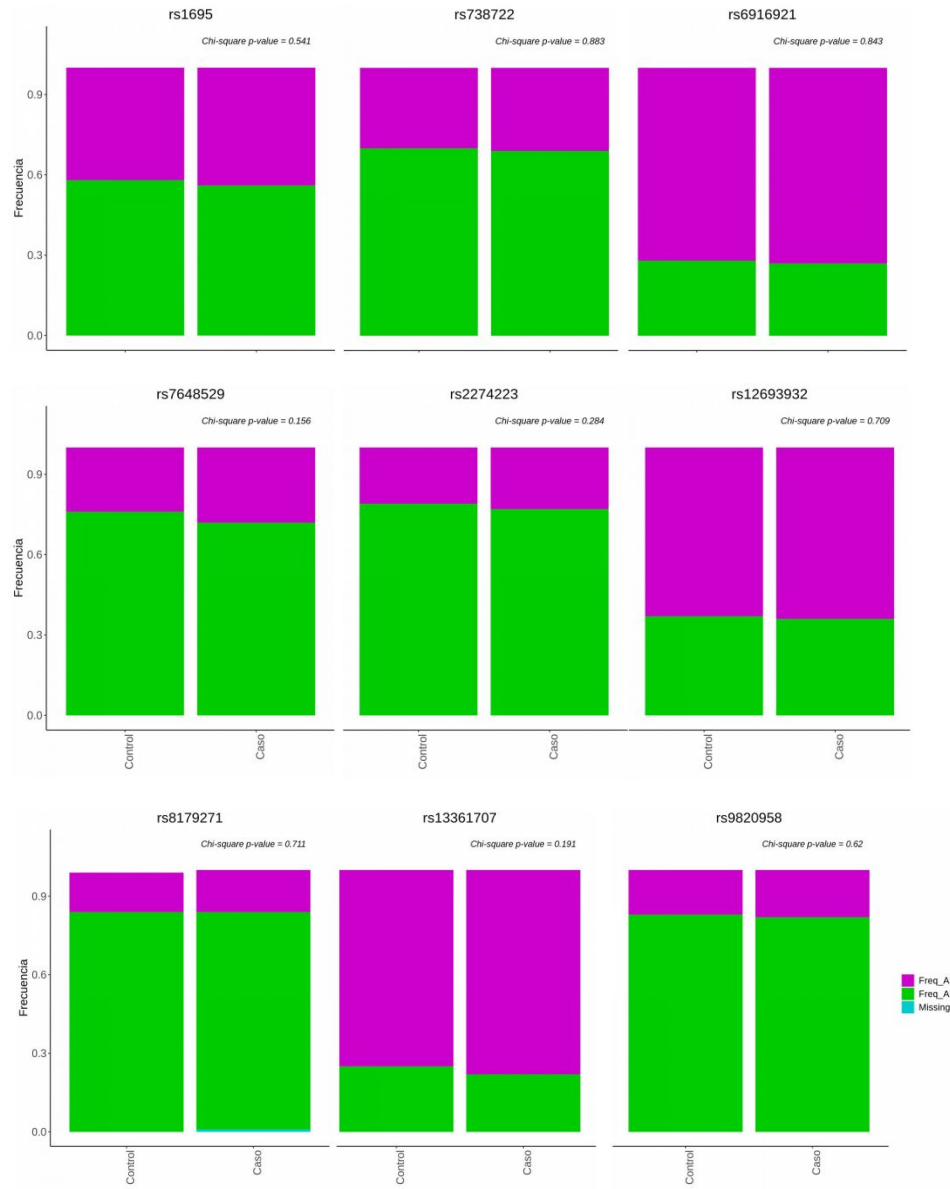
**Figura Suplementaria 11.** Error de Validación Cruzada para las estimaciones de ancestría de 317 controles y 301 casos de Cáncer Gástrico. Se eliminaron individuos por parentesco con el programa REAP. Estimaciones de ancestrías globales por individuo fue realizado con el programa ADMIXTURE. K es el número de poblaciones que se utilizaron para estimar la ancestría global de cada individuo.

Ancestría	Muestra	Promedio	EE	p
Mapuche	Control	0.396	0.005	0.15
	Caso	0.409	0.007	
Aymara	Control	0.033	0.004	0.44
	Caso	0.03	0.003	
Europea	Control	0.544	0.006	0.28
	Caso	0.535	0.007	
Africana	Control	0.017	0.001	0.51
	Caso	0.018	0.001	
Asiática	Control	0.009	0	0.7
	Caso	0.008	0.001	

**Tabla Suplementaria 6.** Proporción del genoma que pertenece a las ancestrías Mapuche, Aymara, Europea, Africana y Asiática de 317 controles y 301 casos de Cáncer Gástrico. Error Estándar: EE, valor de p para la prueba de t-student: p.



**Figura Suplementaria 12.** Análisis de Componentes Principales (PCA) en la muestra de 317 controles y 301 casos de Cáncer Gástrico. Se utilizaron cinco poblaciones de referencia; Europea (CEU), Africana (YRI), Este de Asia (EAS), Aymara (AYM) y Mapuche (MAP), y las muestras mestizas de casos y controles. Se utilizó el programa Plink1.9. Componente principal: PC.



**Figura Suplementaria 13.** Frecuencia de los alelos de nueve SNPs que no tienen diferencias significativas entre controles y casos de Cáncer Gástrico. Freq\_A1: Frecuencia Alelo de Riesgo; Freq\_A2: Frecuencia de alelo no Riesgo; Missing: Datos perdidos. Test: Chi-cuadrado.

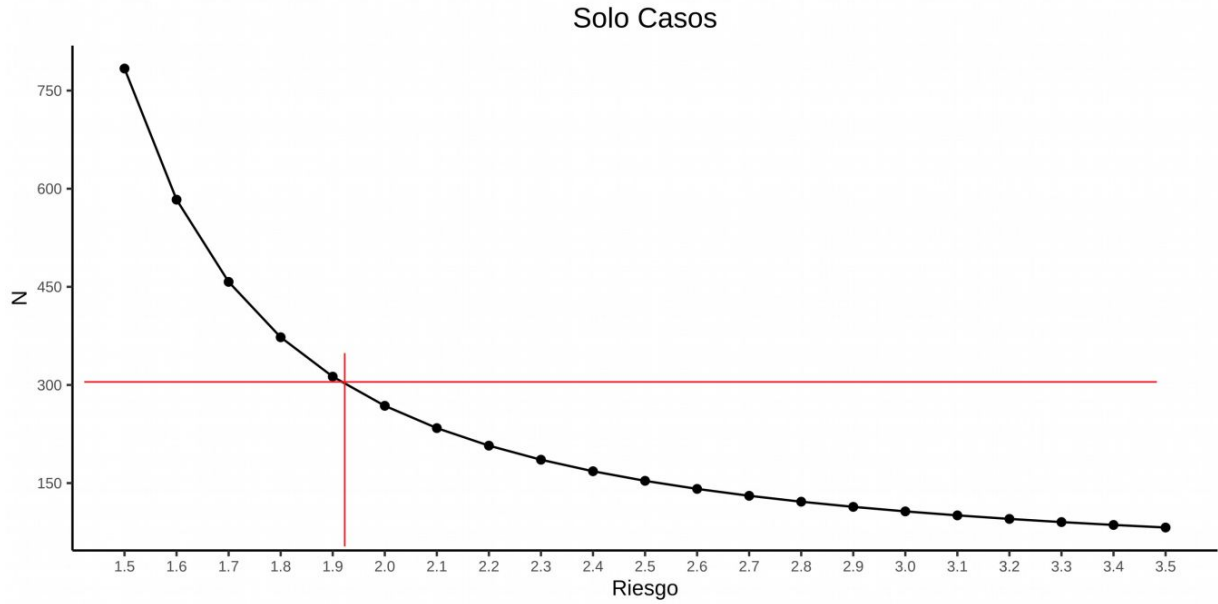


### **X.3 Objetivo Específico 3**

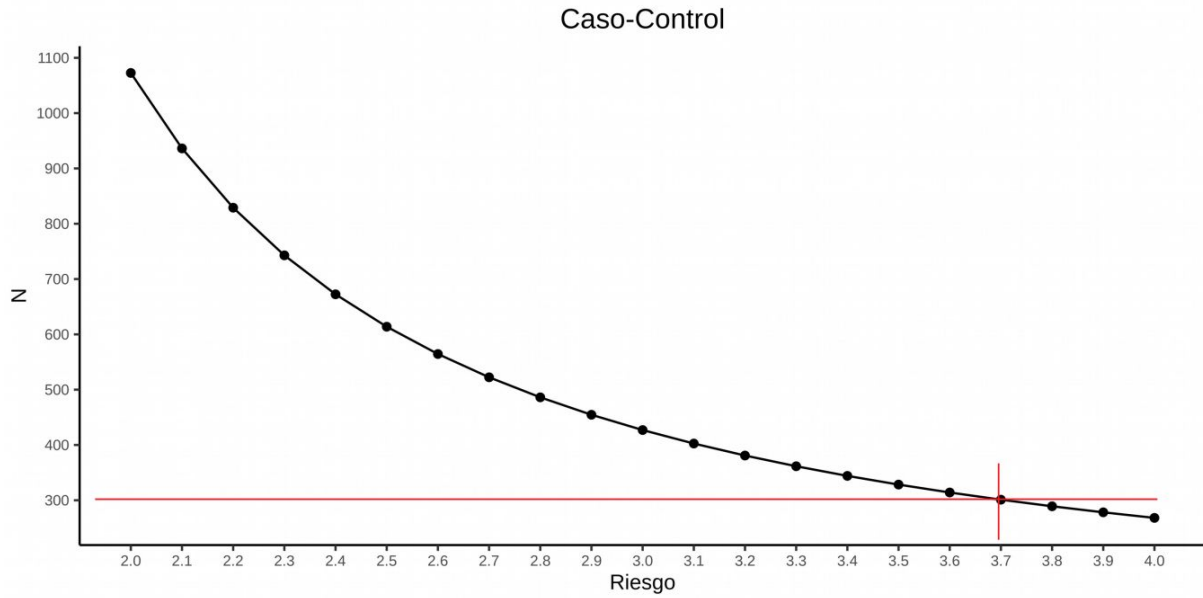
Control de calidad ancestría local vs global muestra caso-control

SNP	Fenotipo	EUR	AMR	valor de p
rs4072037	control	145	68	0.792
	caso	112	57	
rs2274223	control	102	26	0.81
	caso	111	25	
rs1695	control	153	106	0.11
	caso	135	126	
rs9820958	control	45	61	0.8
	caso	43	65	
rs6916921	control	73	91	0.4
	caso	60	93	

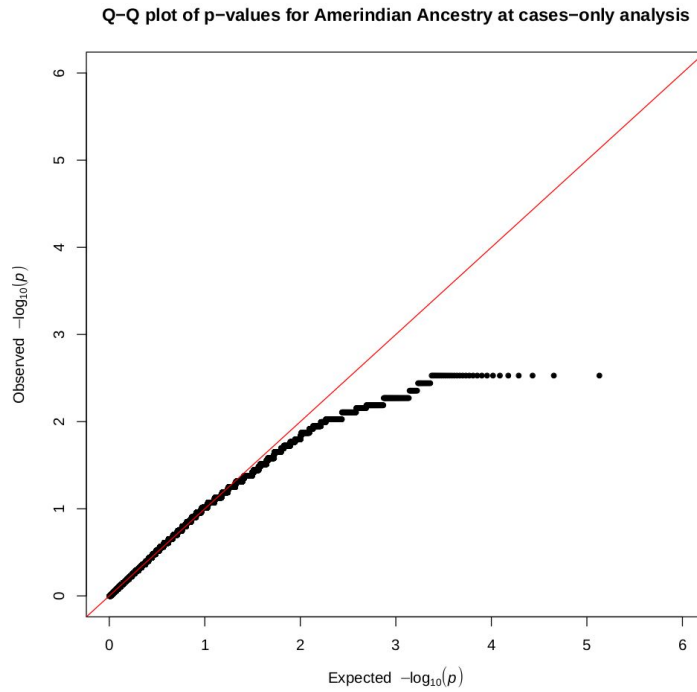
**Tabla Suplementaria 7.** Frecuencias de ancestrías locales Europea (EUR) o Amerindia (AMR) de los alelos de riesgo de cinco SNPs. Valor de p de la prueba de Chi-cuadrado: valor de p.



**Figura Suplementaria 14.** Número de Individuos (N) necesarios para identificar una variante significativa, en un estudio con un poder del 80% y un valor de p de significancia  $< 0.001$ , por el riesgo (odd ratio) de desarrollar la enfermedad en un análisis de solo casos de admixture mapping. N de la muestra: 294 casos de Cáncer Gástrico.



**Figura Suplementaria 15.** Número de Individuos (N) necesarios para identificar una variante significativa, en un estudio con un poder del 80% y un valor de p de significancia  $< 0.001$ , por el riesgo (odd ratio) de desarrollar la enfermedad en un análisis de caso-control de admixture mapping. N de la muestra: 308 controles y 294 casos de Cáncer Gástrico.



**Figura Suplementaria 16.** Gráfico cuantil-cuantil (Q-Q) de los valores de p observados en el análisis de solo casos de admixture mapping para la ancestría Amerindia y los valores de p esperados. Observed  $-\log_{10}(p)$ :  $-\log_{10}$  de los valores de p calculados con z-score. Expected  $-\log_{10}(p)$ :  $-\log_{10}$  de los valores de p esperados en una muestra que tiene una distribución normal.

ID mapeado	Nombre del Gen	Símbolo del Gen	Clase de la Proteína
LINC00208	-	<i>LINC00208</i>	
GNAS	Subunidad alpha de la proteína que se une al nucleótido guanina, isoforma Xlas	<i>GNAS</i>	Proteína G heterotrimétrica
FAM167A	FAM167A	<i>FAM167A</i>	
TDH	Deshidratasa serina-like	<i>SDSL</i>	
STX16	Sintaxina 16	<i>STX16</i>	Proteína SNARE
TDH	L-serina deshidratasa/L-treonina deaminasa	<i>SDS</i>	
SLC35G5	Miembro G5 de la familia transportadora de soluto	<i>SLC35G5</i>	Transportador
MTMR9,AF131216	Gen relacionada a miotubularina 9	<i>MTMR9</i>	Fosfatasa
NPEPL1	Probable aminopeptidasa	<i>NPEPL1</i>	Metaloproteasa
BLK	Tirosina quinasa	<i>BLK</i>	
TDH,AF131216	3-deshidrogenasa L-treonina inactiva, mitocondrial	<i>TDH</i>	Oxidoreductasa

**Tabla Suplementaria 8.** Nombre de los genes encontrados por la herramienta web PANTHER dentro de las regiones genómicas asociadas a la ancestría Amerindia identificadas por AM.

Vía de Señalización
Endotelina
Receptor del factor liberador de corticotropina
Receptor de histamina H2
Hormona liberadora de gonadotropina
Enfermedad de Parkinson
Proteína G heterotrimétrica: Gi alpha y Gs alpha
Liberación de encefalina
Receptor adrenérgico beta 1, beta 2 y beta 3
Receptor muscarínico de acetilcolina 2 y 4
Receptor tipo 5HT4
Receptor nicotínico de acetilcolina
Activación de células B

**Tabla Suplementaria 9.** Nombre de las vías de señalización encontradas por la herramienta web PANTHER dentro de las regiones genómicas asociadas a la ancestría Amerindia identificadas por AM.