Expresión diferencial de genes del cáncer gástrico en diferentes poblaciones: EAST vs. WEST: ¿posible efecto de la dieta?

Alba Salietti Rodríguez
Máster en Bioinformática y Bioestadística UOC-UB
Área 4

Nombre Consultor/a: Helena Brunel Montaner
Nombre Profesor/a responsable de la asignatura: Antoni Pérez Navarro
Enero 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Expresión diferencial de genes del cáncer gástrico en diferentes poblaciones: EAST vs. WEST: ¿posible efecto de la dieta?*
Nombre del autor:	*Alba Salietti Rodríguez*
Nombre del consultor/a:	*Helena Brunel Montaner*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	*01/2021*
Titulación:	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Bioinformática y Bioestadística Área 4*
Idioma del trabajo:	*Español*
Palabras clave	*Transcriptómica, expresión génica, factores medioambientales*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Objetivo: Se pretende realizar un estudio comparativo de expresión génica entre distintas poblaciones con cáncer gástrico e intentar establecer relaciones entre la expresión diferencial de genes y posibles factores de presión selectiva medioambiental y dietética.

Materiales y métodos: Se estudia la expresión diferencial de genes en muestras de cáncer gástrico de distintas poblaciones, mediante el paquete edgeR del software R. Se estudian las diferencias clínicas y biológicas entre las poblaciones, y se intenta describir una relación plausible entre la expresión diferencial y los factores medioambientales y dietéticos mediante la búsqueda bibliográfica para establecer vías fisiológicas interrelacionadas.

Resultados: En el estudio de expresión diferencial de genes, se incluyen un total de 71 muestras de 4 poblaciones distintas (latina, blanca, asiática y negra). Se describen diferencias significativas entre el tipo de tumor, el estadio en el momento del diagnóstico y la presencia de metástasis. Se describe la expresión diferencial de genes en las comparaciones 2 a 2 de las distintas poblaciones, destacando una expresión diferencial en los genes implicados en el metabolismo de las hormonas tiroideas y en el fenotipo de moléculas de histocompatibilidad, que podrían tener relación con factores de presión selectiva como la obesidad y la infección por Helicobacter pylori, respectivamente.

Conclusiones: Se confirman diferencias significativas en el tipo histológico y el estadio en el momento del diagnóstico en distintas poblaciones, que presentan una expresión diferencial en genes que podrían estar modulados por efectos medioambientales y dietéticos como la obesidad y la infección por Helicobacter pylori.

Abstract (in English, 250 words or less):

Objective: to compare gene expression between different populations with gastric cancer and to try to establish relationships between gene differential expression and possible environmental and dietary selective pressure factors.

Materials and methods: The differential expression of genes in gastric cancer samples from different populations is studied using the edgeR package of the R software. Clinical and biological differences between populations are studied, and an attempt is made to establish a plausible relationship between differential expression and environmental and dietary factors by establishing interrelated physiological pathways.

Results: The gene differential expression study includes a total of 71 samples from 4 distinct populations (latin, white, asian and black). Significant differences between tumor type, stage at diagnosis, and presence of metastasis are described. Gene differential expression is described in comparisons 2 to 2 of the different populations, highlighting a differential expression in genes involved in thyroid hormone metabolism and histocompatibility molecule phenotype, which could be related to selective pressure factors such as obesity and Helicobacter pylori infection, respectively.

Conclusions: Significant differences in histological type and stage at diagnosis are confirmed in different populations, which have differential expression in genes that could be modulated by environmental and dietary effects such as obesity and Helicobacter pylori infection.

Índice

Lista de figuras

Lista de tablas

1. INTRODUCCIÓN

1.1 CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO

En la actualidad existen evidencias que sugieren un papel clave de factores medioambientales y dietéticos que actúan como señales de presiones selectivas que inducen al desarrollo de cáncer gástrico. Estos factores podrían explicar las diferencias en la distribución geográfica de la incidencia global y del tipo de cáncer gástrico, pero se necesitan más estudios en este campo para confirmar dicha hipótesis. Estudios poblacionales de expresión génica del cáncer gástrico que apoyen la hipótesis de la presión selectiva medioambiental y dietética en el desarrollo tumoral, podrían dar apoyo a la instauración de medidas preventivas y a la detección precoz, ambos básicos para disminuir la mortalidad por este tipo de cáncer. Actualmente se carece de estudios de expresión génica comparativos entre distintas poblaciones con cáncer gástrico.

1.2 OBJETIVOS DEL TRABAJO

OBJETIVOS GENERALES:

- Investigar la expresión diferencial de genes de pacientes con cáncer gástrico de diferentes poblaciones (latina vs. blanca vs. asiática vs. negra).
- Estudiar diferentes características clínicas y biológicas de los tumores, para detectar posibles factores pronósticos, en los grupos previamente descritos; e intentar relacionar los resultados obtenidos con los diferentes factores de presión selectiva medioambientales y dietéticos.

OBJETIVOS ESPECÍFICOS:

- Investigar la expresión diferencial de genes de pacientes con cáncer gástrico de diferentes poblaciones (latina vs. blanca vs. asiática vs. negra).
- Estudiar diferentes características clínicas de los tumores, para detectar posibles factores pronósticos, en los grupos previamente descritos.
- Estudiar diferentes características biológicas de los tumores, para detectar posibles factores pronósticos, en los grupos previamente descritos.
- Relacionar los factores pronósticos detectados con los diferentes factores de presión selectiva medioambientales.
- Relacionar los factores pronósticos detectados con los diferentes factores de presión selectiva dietéticos.

1.3 ENFOQUE Y MÉTODO SEGUIDO

Se plantea un proyecto de análisis bioinformático de la expresión diferencial de genes comparativo entre distintas poblaciones. Para dicho fin, se emplearán paquetes estadísticos de "bioconductor" del programa R dado que se trata de un programa gratuito y muy potente para este tipo de análisis de datos ómicos. El enfoque a seguir o *pipeline* intenta estandarizar el análisis de los datos para disminuir posibles sesgos y ruidos que podrían interferir en los resultados. Consiste en:

1) Definir las poblaciones sujetas a estudio.
2) Control de calidad de los datos crudos mediante *clustering* / PCA.
3) Normalización de los datos.
4) Identificación de los genes diferencialmente expresados.
5) Anotación de los resultados.
6) Comparación entre distintas comparaciones (cuando existan más de dos subpoblaciones).
7) Análisis de significación biológica ("*Gene Enrichment Analysis*").

Una vez llevado a término el estudio de la expresión diferencial de genes propiamente dicho, se intentará explorar también variables clínicas y biológicas que puedan tener un valor pronóstico, y se investigará si se pueden relacionar con factores de selección medioambientales y dietéticos.

1.4 PLANIFICACIÓN DEL TRABAJO

	TAREAS		FECHA	DÍAS
PEC1	**1. Plan de trabajo.**		**29/9-13/10/20**	**10**
	1.1	Contextualizar y justificar el trabajo.	29-30/09/20	2
	1.2	Exponer los objetivos generales y específicos.	1-2/10/20	2
	1.3	Explicar el enfoque y método a seguir.	5-6/10/20	2
	1.4	Planificar con hitos y temporización.	7-8/10/20	2
	1.5	Redactar y entregar la PEC1.	9-13/10/20	2
PEC2	**2. Investigar la expresión diferencial de genes de pacientes con cáncer gástrico de diferentes poblaciones (latina, blanca, asiática y negra).**		**14-29/10/20**	**12**
	2.1	Abrir los datos sobre cáncer gástrico de la plataforma http://www.cbioportal.org/ e importar los datos en el software R.	14/10/20	1
	2.2	Estudiar cómo aplicar los paquetes de "bioconductor" del programa R para el análisis de la expresión diferencial de genes.	15/16/10/20	2
	2.3	Analizar la expresión diferencial de genes de pacientes con cáncer gástrico de diferentes poblaciones mediante paquetes de "bioconductor" del programa R, aplicando el *pipeline* previamente descrito.	19-21/10/20	3
	2.4	Identificar los genes con expresión diferencial entre diferentes poblaciones con cáncer gástrico y análisis de significación biológica.	22-26/10/20	3
	2.5	Representar los resultados de forma gráfica.	27/10/20	1
	2.6	Describir los resultados de forma escrita.	28-29/10/20	2

	3. Estudiar diferentes características clínicas de los tumores, para detectar posibles factores pronósticos, en los grupos previamente descritos.		30/10-16/11/20	12
	3.1	Abrir los datos sobre cáncer gástrico de la plataforma http://www.cbioportal.org/ e importar los datos en el software R.	30/10/20	1
	3.2	Identificar las características clínicas de los tumores que se van a estudiar.	2-5/11/20	3
	3.3	Aplicar los test estadísticos correspondientes según el tipo de variable para analizar posibles diferencias entre las características clínicas de los tumores en las diferentes poblaciones.	6-9/11/20	2
	3.4	Representar los resultados de forma gráfica.	10/11/20	2
	3.5	Describir los resultados de forma escrita.	11-12/11/20	2
	3.6	Redactar y entregar la PEC2.	13-16/11/20	2
	4. Estudiar diferentes características biológicas de los tumores, para detectar posibles factores pronósticos, en los grupos previamente descritos.		**17-26/11/20**	**8**
	4.1	Abrir los datos sobre cáncer gástrico de la plataforma http://www.cbioportal.org/ e importar los datos en el software R.	17/11/20	1
	4.2	Identificar las características biológicas de los tumores que se van a estudiar.	18-19/11/20	2
	4.3	Aplicar los test estadísticos correspondientes según el tipo de variable para analizar posibles diferencias entre las características biológicas de los tumores en las diferentes poblaciones.	20-23/11/20	2
	4.4	Representar los resultados de forma gráfica.	24/11/20	1
	4.5	Describir los resultados de forma escrita.	25-26/11/20	2
	5. Relacionar los factores pronósticos detectados con los diferentes factores de presión selectiva medioambientales.		**27-30/11/20**	**3**
	5.1	Búsqueda bibliográfica sobre la posible relación entre los factores pronósticos detectados y posibles factores de presión selectiva medioambientales.	27/11/20	1
	5.2	Describir los resultados de forma escrita.	30/11-1/12/20	2
PEC3	**6. Relacionar los factores pronósticos detectados con los diferentes factores de presión selectiva dietéticos.**		**2-14/12/20**	**8**
	6.1	Búsqueda bibliográfica sobre la posible relación entre los factores pronósticos detectados y posibles factores de presión selectiva dietéticos.	2/12/20	1
	6.2	Identificar las características clínicas de los tumores que se van a estudiar.	3-4/12/20	2
	6.3	Aplicar los test estadísticos correspondientes según el tipo de variable para analizar posibles diferencias entre las características clínicas de los tumores en las diferentes poblaciones.	7/12/20	1
	6.4	Describir los resultados de forma escrita.	9-10/12/20	2
	6.5	Redactar y entregar la PEC3	11-14/12/20	2

Tabla 1. *Tareas*

Se adjunta el cronograma de Gantt en el Anexo 1.

1.5 BREVE SUMARIO DE PRODUCTOS OBTENIDOS

Al final del proyecto se habrán entregado:

- **Plan de trabajo**
- **Memoria**
- **Presentación virtual**
- **Autoevaluación del proyecto**

1.6 BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA

2. ANTEDECENTES: Incluye la introducción, la epidemiología del cáncer gástrico, sus factores de riesgo y los factores medioambientales.

3. MATERIALES Y MÉTODOS: Se expone la metodología para el estudio de expresión diferencial de genes y la metodología para el análisis descriptivo de la muestra.

4. RESULTADOS: Se expone el resultado del análisis de expresión diferencial de genes, con el resultado de las comparaciones 2 a 2 de las 4 poblaciones. El segundo apartado es el resultado del análisis descriptivo de las muestras donde se analizan variables basales generales, variables basales entre poblaciones, variables biológicas-pronósticas generales y variables biológicas-pronósticas entre poblaciones.

5. DISCUSIÓN: Se comentan de forma razonada y referenciada con bibliografía publicada en la literatura los resultados del estudio, haciendo énfasis en la posible interrelación entre las funciones biológicas en los que están implicados los genes diferencialmente expresados y las posibles relaciones con factores de presión selectiva medioambientales y dietéticos que pueden actuar de forma distinta en las diferentes poblaciones. Se exponen las limitaciones del trabajo.

6. CONCLUSIONES: Se resumen los resultados y conceptos más importantes del proyecto. Se apuntan futuras líneas de investigación que podrían aportar más consistencia a los resultados del proyecto.

7. GLOSARIO: Se describen las abreviaturas y términos más usados durante la memoria del proyecto.

8. BIBLIOGRAFÍA: Se incluyen las citas bibliográficas a las que se hace referencia durante la memoria del proyecto, siguiendo el estilo Vancouver

9. ANEXOS

Anexo 1. Cronograma de Gantt. Se representa la planificación del proyecto.

Anexo 2. Ejemplo del código R. Se incluye un ejemplo del código usado para el análisis bioinformático mediante el programa RStudio.

2. ANTECEDENTES

2.1 INTRODUCCIÓN

Prevenir y curar enfermedades crónicas, como el cáncer gástrico, representa un gran reto para el campo médico moderno. Existen esfuerzos desde las ciencias básicas y se han realizado grandes avances en campos como el epidemiológico, pero los aspectos sociales y culturales asociados a esta enfermedad han sido poco estudiados y son presentados como elementos de segundo orden al momento de construir programas de salud pública, ya sea para su detección temprana y/o para su control [1].

2.2 EPIDEMIOLOGÍA

El cáncer gástrico se diagnostica en casi 1 millón de pacientes nuevos cada año a nivel mundial, representa la quinta causa de enfermedad neoplásica en el mundo y la tercera causa de muerte relacionada con el cáncer [2-4]. Su incidencia varía ampliamente entre las regiones del mundo, con la incidencia más alta en Asia oriental, seguido de Europa oriental y partes de América Latina (siendo la primera causa de muerte en hombres y mujeres en Colombia)[1,5]. Las ratios de incidencia estandarizadas por edad son 6 veces superiores en Asia oriental comparadas con las de América del norte para ambos sexos [2]. Así pues, aproximadamente el 70% del cáncer gástrico en todo el mundo ocurre en países en vías de desarrollo, incluidos Asia oriental, Europa central y oriental y América del Sur, en comparación con los países desarrollados. Los países desarrollados tienen una incidencia de 173.000 casos en hombres y 102.000 en mujeres, en comparación con 467.000 y 247.000, respectivamente, en los países en desarrollo [6].

Además de las diferencias geográficas de la incidencia global del cáncer gástrico a nivel mundial, destaca que la incidencia de los diferentes tipos de cáncer gástrico también varían entre las regiones del mundo [5]. Los tumores proximales localizados en la unión gastroesofágica y el cardias gástrico son más comunes en Occidente y en raza blanca, en la que supone el doble de los casos respecto a los tumores no localizados en el cardias [3,5,7]. Se asocian con peores resultados debido a un estadio más avanzado en la presentación, mayor tamaño del tumor y asociación con histología pobremente diferenciada [5,7]. A pesar de eso, los pacientes con tumores gástricos proximales en Oriente (Asia) parecen tener una mejor supervivencia.

Y no sólo eso, sino que también es de interés poblacional que en los países occidentales, la incidencia de cáncer gástrico va en aumento; aproximadamente 25.000 casos nuevos se diagnostican cada año en los Estados Unidos. Este incremento se acompaña de un cambio en la incidencia relativa de los diferentes tipos de cáncer gástrico. Los tumores localizados en la unión gastroesofágica y el cardias gástrico se están volviendo más comunes, probablemente relacionados con la epidemia de obesidad y la prevalencia de la enfermedad por reflujo gastroesofágico. Si bien la incidencia de cáncer en cuerpo gástrico ha ido disminuyendo en casi

todos los grupos étnicos y de edad, la incidencia de cáncer gástrico distal ha aumentado en un 70% en pacientes caucásicos en el grupo de edad de 25 a 39 años (5).

Se han descrito disparidades raciales pronunciadas en el momento de la cirugía curativa que disminuyen después de años de supervivencia. Este hecho sugiere que la raza tiene menos influencia sobre los resultados cuanto más tiempo sobreviven los pacientes, pero se necesitan más estudios al respeto [8].

Dentro de los factores de riesgo se consideran tanto factores biológicos como socioculturales y medioambientales [1].

2.3 FACTORES DE RIESGO BIOLÓGICOS

Los **factores genéticos** plantean un riesgo 2 a 3 veces mayor de desarrollar cáncer de estómago en familiares de primer grado de pacientes con cáncer gástrico [1,9]. Dentro de los factores genéticos, existen varios estudios que identifican mutaciones de riesgo como la de CDH1 [10] y PALB2 [11] en poblaciones específicas, y se han identificado también algunos polimorfismos de IL-17 como factores de riesgo para el cáncer gástrico en población asiática [12]. A pesar de ello, se carece de estudios de expresión génica comparativos entre distintas poblaciones con cáncer gástrico.

Otro factor biológico importante es la **edad**, ya que más de la mitad de los casos se diagnostican en personas de más de 65 años [3,13], y el **sexo masculino** [3]. Se ha descrito también que los individuos del **grupo sanguíneo A** tienen un 20% más de probabilidad de presentar cáncer gástrico que las personas de los grupos sanguíneos O, B o AB (Muñoz Francesani, 1997), aunque existen opiniones contradictorias [1].

Existen además algunas **lesiones precursoras propias del estómago** que están asociadas con un mayor riesgo de cáncer, como la gastritis crónica atrófica, el reflujo gastroesofágico, los adenomas gástricos, la enfermedad de Menetrier y la anemia perniciosa [1,3]. Otro de los factores de riesgo, y quizás uno de los más importantes, es la **infección por Helicobacter pylori (H.pylori)**, que es una bacteria que coloniza la mucosa gástrica debido a la contaminación alimentaria y se ha asociado con el desarrollo de una inflamación local, la gastritis crónica atrófica, una conocida lesión precursora del cáncer gástrico [1,3].

2.4 FACTORES DE RIESGO MEDIOAMBIENTALES

El estudio realizado por Cairns (1986)(14) en el que comparó los índices de mortalidad por cáncer gástrico de japoneses residentes en Japón, japoneses inmigrantes a EE.UU., hijos de inmigrantes japoneses residentes en EE.UU. (Nisseis) y blancos estadounidenses, demostró que el índice de mortalidad por cáncer gástrico en los japoneses que migraron a EE.UU.

disminuyó en relación con el de los japoneses residentes en Japón. A su vez los hijos de los inmigrantes japoneses a EE.UU. presentaron una incidencia de mortalidad menor a la de sus padres y cercana a la de los blancos estadounidenses. Estas diferencias se podrían explicar por factores medioambientales y dietéticos más que por los factores genéticos [1,3].

En particular, la **dieta** ha sido objeto de numerosos estudios epidemiológicos y antropológicos. La dieta como factor medioambiental, social y cultural, podría explicar en parte esta divergencia geográfica descrita en la incidencia del adenocarcinoma gástrico [1]. Existen diferentes estudios basados más específicamente en los factores dietéticos del cáncer gástrico [1]. Un estudio desarrollado por Zhang (1994) evidenció que un consumo alto de frutas y verduras ricas en vitamina C estaba asociado con la disminución en el riesgo de desarrollar cáncer gástrico; así mismo, se estableció que una dieta rica en cereales de grano entero, carotenoides y té verde, disminuye el riesgo de desarrollar la enfermedad. A esta misma conclusión llegaron Buiatti et al. (1990) en un estudio de casos y controles de cáncer gástrico y dieta realizado en Italia, Graham et al. (1972) en un estudio sobre factores alimentarios en la epidemiología del cáncer gástrico, y Haenszel et al. (1972) al realizar una investigación sobre cáncer gástrico en japoneses residentes en Hawai [1,3]. Por otro lado, una ingesta habitual de alimentos ahumados y asados, ricos en hidrocarburos policíclicos aromáticos, está asociada con el riesgo de cáncer gástrico (WCRF, 1997). De acuerdo con los estudios de los investigadores de la división de epidemiología y genética del Instituto Nacional de Cáncer de Estados Unidos sobre la dieta y los hábitos culinarios se demostró que quienes comían carne bien asada tenían un riesgo tres veces mayor que quienes comían la carne a término medio. También se encontró que las personas que comían carne cuatro o más veces a la semana tenían el doble del riesgo de desarrollar ese tipo de cáncer que aquellos que consumían carne con menor frecuencia. También está bien establecido que una dieta con alta ingesta de sal está asociada a este tipo de cáncer [1,2,3,15]. El efecto carcinogénico de la sal está posiblemente ligado a un daño de la mucosa del estómago, como lo demostró Kodama y col. (1984) en un estudio sobre el efecto de la sal en el estómago de los ratones [1].

Algunos investigadores consideran que las personas de **escasos recursos económicos** tienen un mayor riesgo de presentar la enfermedad, pues obtienen el agua para consumo de pozos artesanales o aljibes que presentan altos niveles de nitratos y contaminación por H. Pylori. Adicionalmente estos grupos de población tienen una dieta rica en carbohidratos y baja en frutas y verduras frescas, así como un menor acceso y uso de los servicios de salud [1,3,13].

Como otro factor medioambiental y sociocultural para el desarrollo de cáncer gástrico se ha descrito el **tabaquismo** [2,3,15].

3. MATERIALES Y MÉTODOS

3.1 METODOLOGÍA PARA EL ESTUDIO DE EXPRESIÓN DIFERENCIAL DE GENES

IDENTIFICACIÓN DE LAS MUESTRAS

En este estudio se dispone de 5 muestras de cáncer gástrico de la población latina y de 6 muestras de cáncer gástrico de población negra, las cuales se analizan en su totalidad respectivamente. De la población blanca y asiática se analizan 30 muestras de cada población del total del que se dispone.

Posteriormente se analiza la distribución de las muestras conjuntamente y se realiza el análisis de expresión diferencial de genes comparando las poblaciones 2 a 2.

Los datos se descargan de la plataforma http://www.cbioportal.org/.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGESLIST

Una vez cargado el archivo que contiene la *count table* con los *reads counts* de las muestras a procesar mediante el software R, se recodifican las variables para obtener una tabla inicial con todas las muestras correctamente identificadas.

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los grupos de comparación (por ejemplo latina vs. blanca), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

- Transformación de los datos: la transformación de los *counts* a *pseudoCounts* en escala logarítmica en base 2 (log2) nos permite reducir la variabilidad y aproximar la distribución de los contajes a la normalidad, de forma que facilita la visualización de los datos. (Ver ejemplo en el Anexo 2. Código R).
- Distribuciones de los *pseudoCounts* entre muestras: la visualización de la distribución de *pseudoCounts* entre las diferentes muestras es útil para comparar la expresión génica entre ellas.
  1) **BARPLOT:** el gráfico de barras muestra el tamaño de las bibliotecas (conjunto de *pseudoCounts*) para cada una de las muestras, observando cuantas lecturas tenemos por muestra.
  2) **BOXPLOT:** el gráfico Boxplot proporciona una manera sencilla de visualizar la distribución de los *pseudoCounts* de cada muestra, observando la dispersión de las muestras y puntos alejados que permiten la detección de posibles *outliers*.

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

El análisis multivariante se lleva a cabo mediante el *multidimensional scaling* como representación del análisis de componentes principales, y el análisis de *clústers* y *heatmap*:

1) ***MULTIDIMENSIONAL SCALING PLOT*:** un MDSplot es una visualización de un análisis de componentes principales. Si el experimento está bien controlado y ha funcionado bien, lo que esperamos ver es que las mayores fuentes de variación en los datos son los grupos que nos interesan. También es una herramienta increíblemente útil para el control de calidad y la verificación de valores atípicos. Podemos usar la función plotMDS para crear el gráfico MDS.

2) ***CLUSTERING* I *HEATMAP*:** para explorar las similitudes y diferencias entre las muestras, a menudo es instructivo buscar un *clustering image map* (CIM) o mapa de calor de la matriz de distancia de muestra a muestra.
   Un mapa de calor es una cuadrícula bidimensional, rectangular y coloreada. Muestra datos que a su vez vienen en forma de matriz rectangular:
   - El color de cada rectángulo está determinado por el valor de la entrada correspondiente en la matriz
   - Las filas y columnas de la matriz se reorganizan de forma independiente de acuerdo con algún método de agrupamiento jerárquico, de modo que filas y columnas similares se coloquen una al lado de la otra, respectivamente.

FILTRADO

Muchos de los genes representados en la *count table* presentarán una expresión muy baja que no contribuirá en el resultado y dificultará el análisis. Por tanto, aplicamos un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en cierto número de muestras (según el número de muestras en la comparación).

NORMALIZACIÓN

Nos basaremos en la aproximación de normalización a través de la función calcNormFactors() del paquete de edgeR. Al normalizar los datos obtenemos el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

1) **MDS PLOT:** permite la comparación de la distribución de las muestras y las distancias entre ellas respecto al MDS Plot de los datos crudos.

2) **ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS:** se calcula la dispersión comuna, que estima la dispersión del global del conjunto de datos promediada para todos los genes; y la dispersión específica para los genes.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

Aplicando la función exacTest() i topTag() del paquete edgeR, se obtiene la tabla con los genes diferencialmente expresados con el p-valor correspondiente. La función decideTestDGE permite ver los genes sobre o infraexpresados en la comparación entre muestras 2 a 2.

ANOTACIÓN DE LOS RESULTADOS

Una vez tenemos seleccionados los genes diferencialmente expresados en la comparación entre las poblaciones (2 a 2) se tienen que identificar estos genes con sus anotaciones equivalentes en base a *Gene Ontology* (*Entrez Gee identifier*, *RefSeq*, *Ensembl*, *Gene Symbol*), permitiendo obtener más información. Este procedimiento se llevará a cabo a través de la descarga de la base de datos "org.Hs.eg.db" del paquete BiocManager que contiene las diferentes anotaciones posibles de los genes humanos.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Para interpretar la significación biológica de los genes diferencialmente expresados detectados, se utiliza la aproximación de goana(), especificando que se quiere estudiar los "biologycal process" que permite identificar en qué vías biológicas participan los genes en base a su anotación de *Gene Ontology*.

3.2 METODOLOGÍA PARA EL ANÁLISIS DESCRIPTIVO DE LA MUESTRA

Las variables cualitativas se expresarán en número total y porcentaje. Las variables cuantitativas serán evaluadas mediante un test de normalidad, como *Shapiro-Wilks* (para muestras n < 50) o *Kolmogorov-Smirnov* (para muestras n > 50).

En el análisis de contraste, se aplicarán test paramétricos en caso de haber comprobado la distribución normal de las variables, mientras que si no se sigue la normalidad aplicarán test no paramétricos. En caso de requerir análisis bivariante de dos variables cualitativas se aplicará el test de Chi-cuadrado, excepto si alguna de las frecuencias es <5 donde se aplicaría el exacto de *Fisher*. En caso de una variable cuantitativa y una variable cualitativa dicotómica, el contraste se realizará mediante el test paramétrico *T-Student* o el test no paramétrico *U-Mann-Whitney*. En caso de que la variable cualitativa sea politómica, se aplicará el test paramétrico *ANOVA* o el test no paramétrico *Kruskal-Wallis*. Si hay que contrastar dos variables cuantitativas se utilizará la correlación de *Pearson* (paramétrica), o la correlación de *Spearman* (no paramétrica).

Se utilizará un valor de p <0.05 bilateral como estadísticamente significativo. Los cálculos se realizarán con el paquete informático R software.

Se estudiarán las diferencias clínicas y biológicas entre las poblaciones, y se intentará describir una relación plausible entre la expresión diferencial y los factores medioambientales y dietéticos mediante la búsqueda bibliográfica para establecer vías fisiológicas interrelacionadas.

4. RESULTADOS

4.1 RESULTADOS DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL DE GENES

Inicialmente se explora la distribución de las muestras en las 4 poblaciones conjuntamente.

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 5 muestras de sujetos de población latina, 30 muestras de sujetos de población blanca, 30 muestras de sujetos de población asiática y 6 muestras de población negra.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene <chr>	latino1 <dbl>	latino2 <dbl>	latino3 <dbl>	latino4 <dbl>	latino5 <dbl>	blanco1 <dbl>	blanco2 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041	470.6122	342.4043
5T4	933.7998	796.9726	189.8969	489.3010	345.8573	351.4286	1018.0377
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578	2365.7143	1357.1056
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432	1.2245	0.4171
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937	175.1020	99.6768
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491	26.5837	17.6624
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000	105.7143	0.4171
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957	26388.3633	8842.9528
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607	84.2898	22.4419
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811	0.8163	1.6682

Figura 1. *Datos crudos de las 4 poblaciones a analizar.*

Por tanto, de los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 71 muestras (5 muestras de sujetos de población latina, 30 muestras de sujetos de población blanca, 30 muestras de sujetos de población asiática y 6 muestras de sujetos de población negra).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 4 grupos de comparación (latino vs. blanca vs. asiática, vs. negra), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

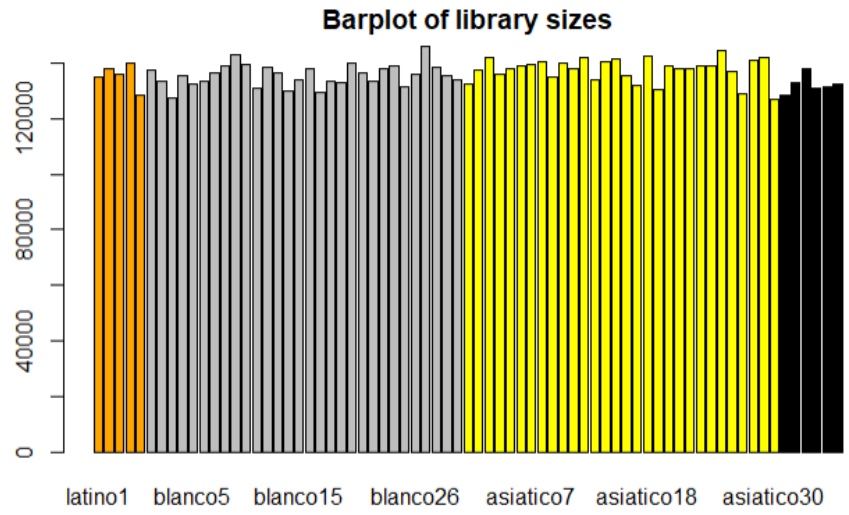


Figura 2. *Barplot de las 4 poblaciones. Población latina de color naranja, población blanca de color gris, población asiática de color amarillo y población negra de color negro.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aunque parece que las muestras de la población asiática, tienen recuentos un poco más elevados que el resto. En todo caso, el tamaño de la biblioteca se aproxima a los 14.000 recuentos para la mayoría de las muestras.

BOXPLOT

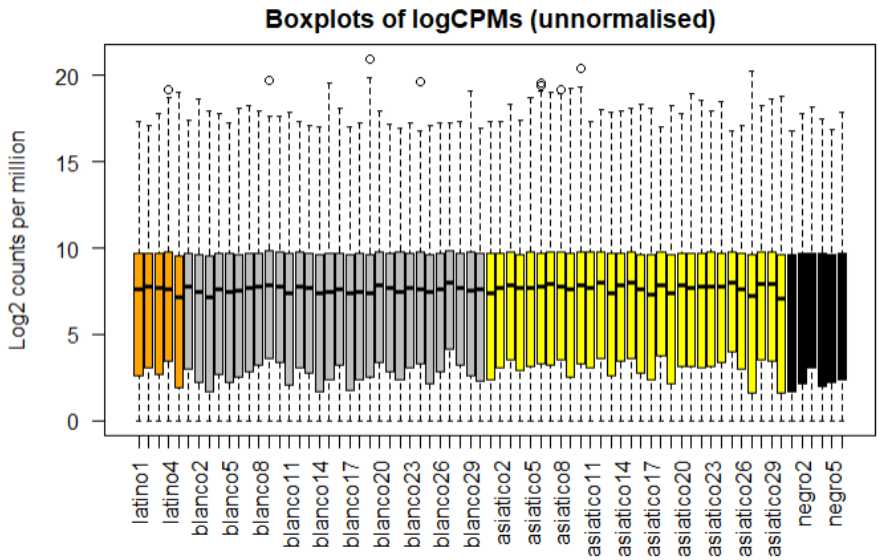


Figura 3. *Boxplot de las 4 poblaciones. Población latina de color naranja, población blanca de color gris, población "asiático" de color amarillo y población negra de color negro.*

Se observa que la distribución de los recuentos (*pseudoCounts*) es bastante homogénea entre las 71 muestras, con algún valor atípico o *outlier*.

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT (MDS PLOT)

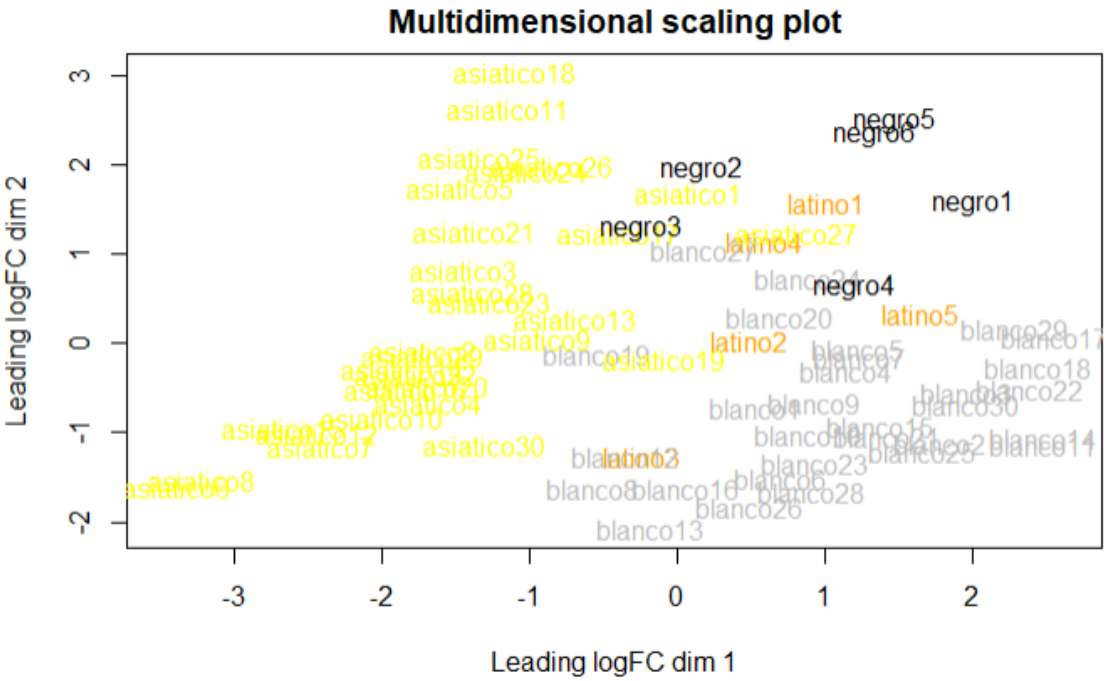


Figura 4. *MDS Plot de las 4 poblaciones.*

Se observa cierta agrupación por población. La población asiática tiene tendencia a distribuirse de forma agrupada a la izquierda, la población negra tiene tendencia a distribuirse por arriba, y las poblaciones blanca y latina más por el centro y derecha de forma mezclada entre las 2 poblaciones.

Esta agrupación menos definida de las poblaciones latina y negra puede ser consecuencia de la poca cantidad de muestras de que se dispone.

CLUSTERING I *HEATMAP*

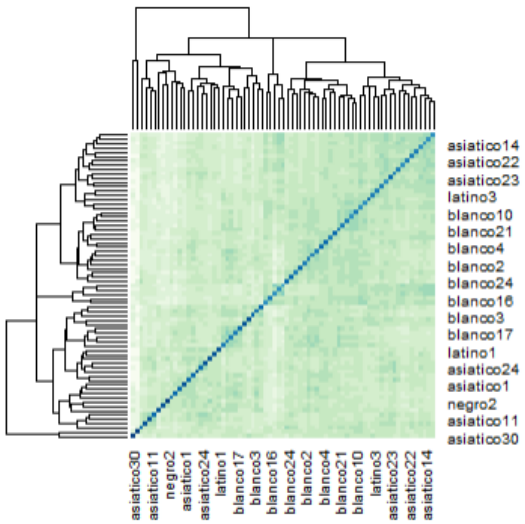


Figura 5. *Clustering i heatmap de las 4 poblaciones.*

Se puede observar cierta agrupación por población, sobre todo entre las muestras de la población asiática entre sí, y las muestras de la población blanca entre sí. Las muestras de las poblaciones latina y negra se solapan con las otras 2 poblaciones, debido probablemente a la poca cantidad de muestras de estas 2 poblaciones.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 4 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 17101 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18834670	1.0316325
latino2	latino	18874693	1.0325143
latino3	latino	19134901	1.0368160
latino4	latino	21024677	0.9743941
latino5	latino	20708185	0.8401281
blanco1	blanco	18324595	1.0632195
blanco2	blanco	21894258	0.8382132
blanco3	blanco	19546320	0.9003439
blanco4	blanco	18804940	1.0312474
blanco5	blanco	17617651	1.0615084

1-10 of 71 rows Previous 1 2 3 4 5 6 ... 8 Next

Figura 6. *Datos normalizados de las 4 poblaciones.*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDS PLOT

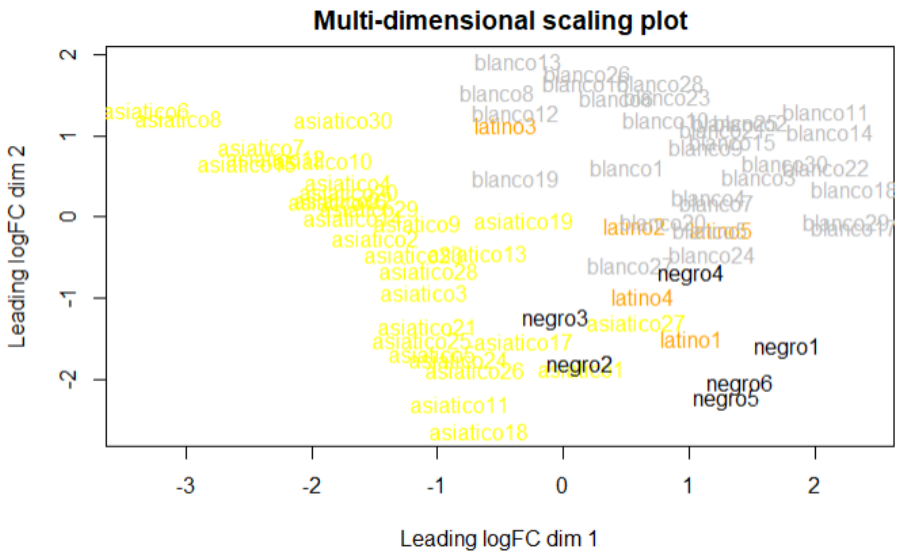
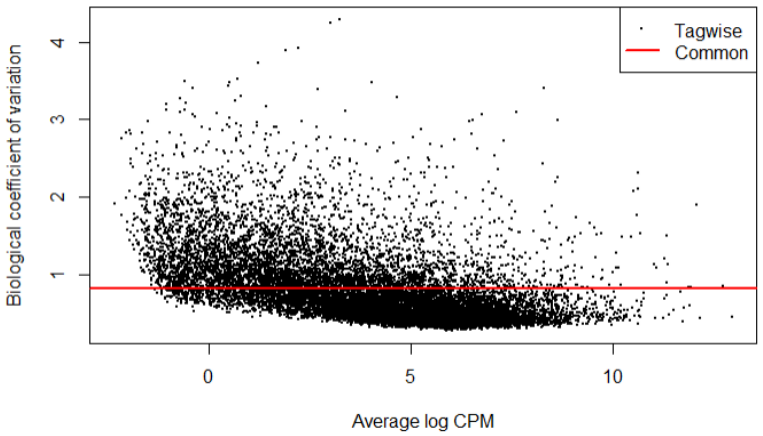


Figura 7. *MDS con los datos normalizados de las 4 poblaciones.*

Observamos que las posiciones de las muestras han variado respecto el MDS Plot de datos no normalizados, pero que las distancias entre las muestras se mantienen similares.
La población asiática tiene tendencia a distribuirse de forma agrupada a la izquierda, la población negra tiene tendencia a distribuirse por abajo a la derecha, y las poblaciones blanca y latina más por el centro y derecha de forma mezclada entre las 2 poblaciones.

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 66%
Dispersión específica: 66%

Figura 8. *Estimación de la dispersión de los datos de las 4 poblaciones.*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.8-0.9 aunque hay una dispersión de 66% con el coeficiente de variación biológica de algún gen superior a 4.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

Para el estudio de la expresión diferencial de genes se realizan comparaciones de las poblaciones 2 a 2.
Tal y como se ha comentado en el apartado 3.1, la identificación de genes diferencialmente expresados se ha realizado aplicando la función exacTest() i topTag() del paquete edgeR obteniendo la tabla con los genes diferencialmente expresados con el p-valor correspondiente.
La función decideTestDGE permite ver los genes sobre o infraexpresados en la comparación entre muestras 2 a 2, por lo que se mostrarán los resultados en las diferentes comparaciones que se detallan en apartados posteriores donde se muestran las comparaciones 2 a 2 entre las poblaciones.
La anotación de los resultados y el análisis de significación biológica también son producto de la comparación 2 a 2 entre las poblaciones, por lo que los resultados también se muestran en sus comparaciones correspondientes.

4.1.1 LATINA – BLANCA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 5 muestras de sujetos de población latina y 30 muestras de sujetos de población blanca.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene <chr>	latino1 <dbl>	latino2 <dbl>	latino3 <dbl>	latino4 <dbl>	latino5 <dbl>	blanco1 <dbl>	blanco2 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041	470.6122	342.4043
5T4	933.7998	796.9726	189.8969	489.3010	345.8573	351.4286	1018.0377
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578	2365.7143	1357.1056
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432	1.2245	0.4171
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937	175.1020	99.6768
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491	26.5837	17.6624
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000	105.7143	0.4171
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957	26388.3633	8842.9528
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607	84.2898	22.4419
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811	0.8163	1.6682

Figura 9. *Datos crudos de las 2 poblaciones a analizar (latina y blanca).*

Por tanto, de los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 35 muestras (5 muestras de sujetos de población latina y 30 muestras de sujetos de población blanca).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (latino vs blanco), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

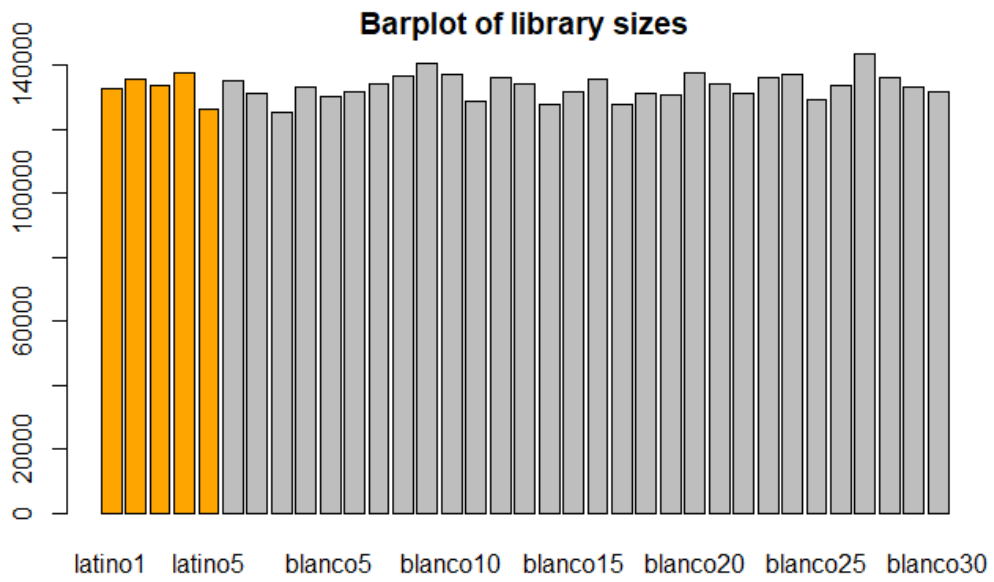


Figura 10. *Barplot de las 2 poblaciones. Población latina de color naranja y población blanca de color gris.*

Observamos que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras de las 2 poblaciones. En todo caso, el tamaño de la biblioteca se aproxima a los 14.000 recuentos para la mayoría de las muestras. Algunas muestras están entre los 12.000 y los 13.000 recuentos.

BOXPLOT

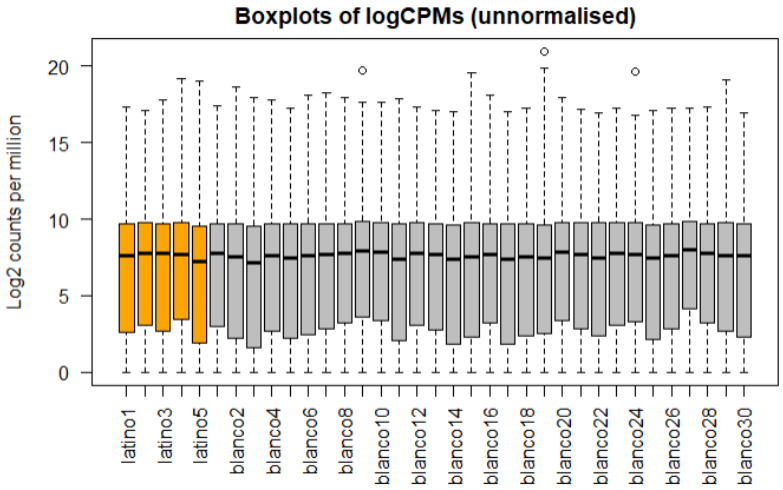


Figura 11. *Boxplot de las 2 poblaciones. Población latina de color naranja y población blanca de color gris.*

Observamos como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre las 35 muestras, con algún valor atípico o *outlier* ("blanco9", "blanco19" y "blanco24").

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

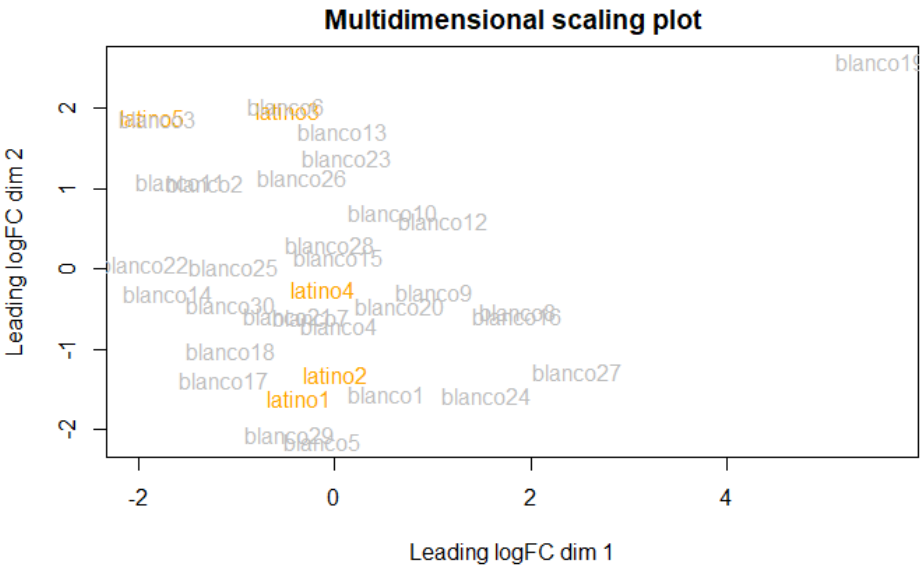


Figura 12. *MDS Plot de las 2 poblaciones (latina y blanca).*

En este caso se observa solapamiento en la distribución de las muestras latina y blanca. El hecho de disponer de solamente 5 muestras para la población latina representa una limitación en este análisis. Parece que hay una muestra "blanco19" con una distribución muy diferente al resto, pudiendo suponer un outlier tal y como ya se ha observado en la figura 11.

CLUSTERING I *HEATMAP*

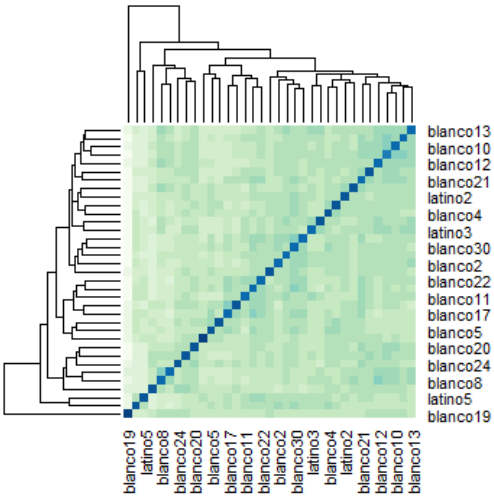


Figura 13. *Clustering i heatmap de las 2 poblaciones (latina y blanca).*

Se observa solapamiento entre las poblaciones latina y blanca, debido probablemente a la poca cantidad de muestras de la población latina.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 4 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 15965 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18619487	1.0397589
latino2	latino	18634649	1.0551192
latino3	latino	18889167	1.0749278
latino4	latino	20768991	0.9660329
latino5	latino	20540121	0.8240983
blanco1	blanco	18106215	1.1272630
blanco2	blanco	21660251	0.8331195
blanco3	blanco	19324523	0.8400748
blanco4	blanco	18585301	1.0383282
blanco5	blanco	17315282	1.0349502

1-10 of 35 rows Previous 1 2 3 4 Next

Figura 14. *Datos normalizados de las 2 poblaciones (latina y blanca).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDS PLOT

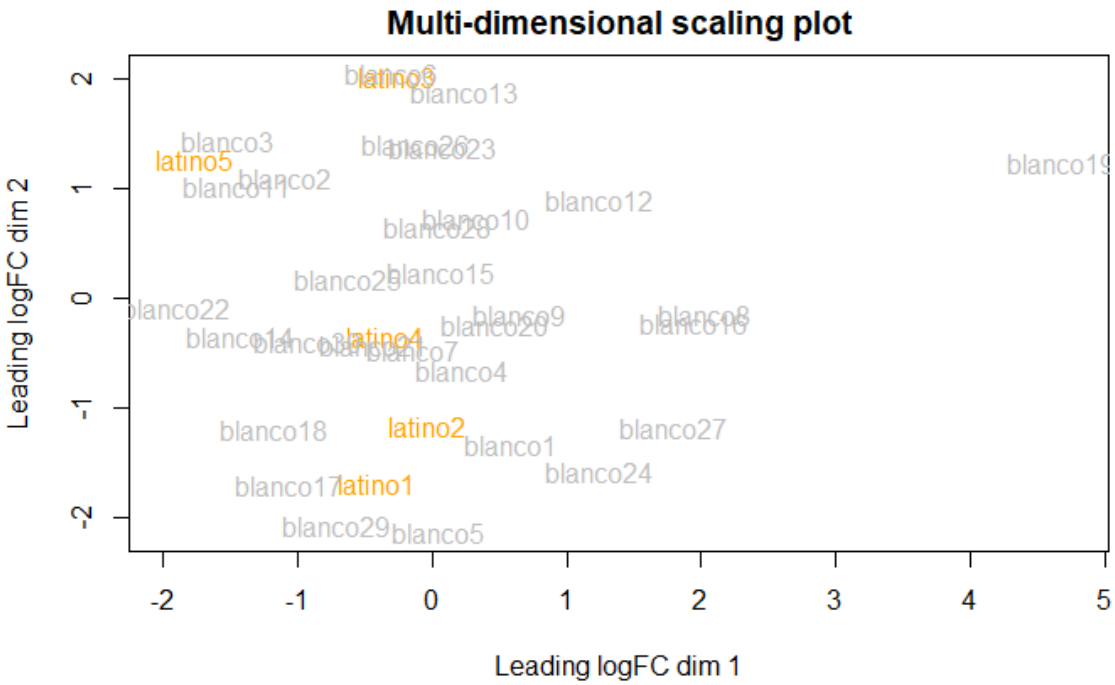
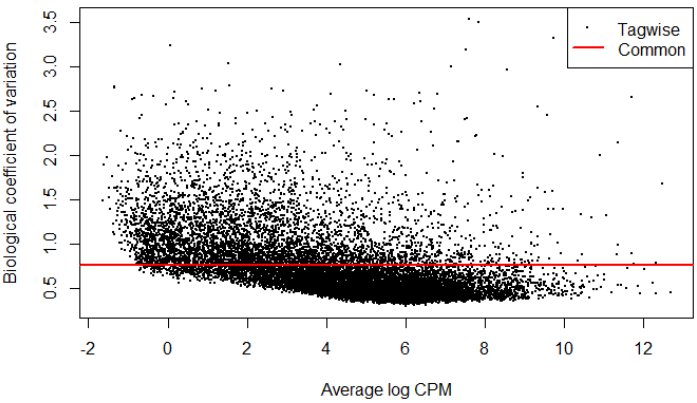


Figura 15. *MDS Plot con los datos normalizados de las 2 poblaciones (latina y blanca).*

Se observa que las posiciones de las muestras han variado respecto el MDS Plot de datos no normalizados, pero que las distancias entre las muestras se mantienen similares.En este caso, el MDS Plot con los datos normalizados es muy similar al de los datos crudos, con solapamiento de las muestras de la población latina con las muestras de la población blanca y un posible *outlier* en la muestra "blanco19" (ya observado en la figura 11 y en la figura 12).

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 58%
Dispersión específica: 58%

Figura 16. *Estimación de la dispersión de los datos de las 2 poblaciones (latina y blanca).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.8 aunque hay una dispersión de 58% con el coeficiente de variación biológica de algún gen superior a 3.5.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

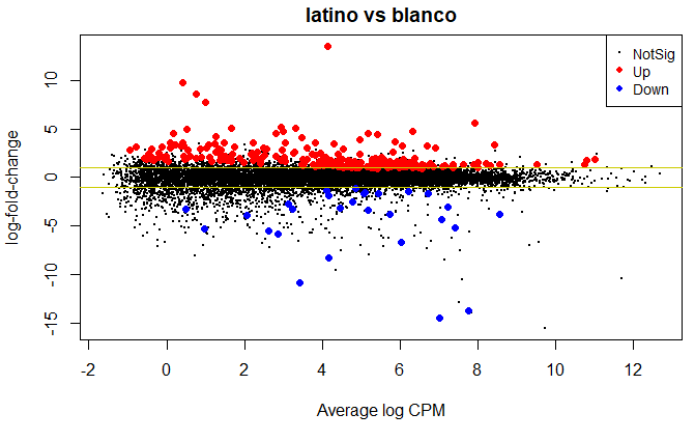
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
6881	HBM	13.523587	4.1403284	1.264001e-192	2.017977e-188
16262	SERP1	-13.753728	7.7644347	7.535138e-70	6.014924e-66
1781	C1orf84	5.081704	3.3004210	3.258914e-67	1.734285e-63
6641	GPX6	9.758006	0.4064400	4.699920e-57	1.875856e-53
2576	CDC23	3.251215	6.7036750	1.748273e-50	5.582236e-47
2123	CACNG5	8.582110	0.7611524	3.295214e-44	8.768015e-41
6882	HBM	-14.521512	7.0222703	2.005191e-43	4.573267e-40
2625	CDH24	7.678858	1.0054749	3.498143e-33	6.980982e-30
2577	CDC23	-3.787847	8.5466426	4.525090e-24	8.027007e-21
8883	LINC00875	4.481251	0.1675795	8.474915e-23	1.353020e-19

1-10 of 10 rows

Figura 17. *10 genes diferencialmente expresados entre las 2 poblaciones (latina y blanca).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	latino-blanco
Down	28
NotSig	15705
Up	232

Figura 18. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (latina y blanca).*

Se observa que 28 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 232 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 15705 genes no muestran una expresión diferencial significativa.

ANOTACIÓN DE LOS RESULTADOS

Una vez se han seleccionados los genes diferencialmente expresados en la comparación entre las poblaciones latina vs. blanca, se tiene que identificar estos genes con sus anotaciones equivalentes en base a *Gene Ontology* (*Entrez Gee identifier*, *RefSeq*, *Ensembl*, *Gene Symbol*), permitiendo obtener más información. Este procedimiento se llevará a cabo a través de la descarga de la base de datos "org.Hs.eg.db" del paquete BiocManager que contiene las diferentes anotaciones posibles de los genes humanos.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Se observan las 20 anotaciones GO enriquecidas:

	Term <chr>	Ont <chr>	N <dbl>	Up <dbl>	Do... <dbl>	PUp <dbl>
GO:0048822	enucleate erythrocyte development	BP	2	2	0	0.0001999502
GO:2001016	positive regulation of skeletal muscle cell dif...	BP	2	2	0	0.0001999502
GO:0006590	thyroid hormone generation	BP	10	3	0	0.0003108243
GO:0042403	thyroid hormone metabolic process	BP	11	3	0	0.0004229814
GO:0031572	G2 DNA damage checkpoint	BP	12	3	0	0.0005581698
GO:0071824	protein-DNA complex subunit organization	BP	142	8	1	0.0008791710
GO:0044818	mitotic G2/M transition checkpoint	BP	14	3	0	0.0009046192
GO:0048485	sympathetic nervous system development	BP	15	3	0	0.0011191552
GO:0060751	branch elongation involved in mammary gland duc...	BP	4	2	0	0.0011775333
GO:0060750	epithelial cell proliferation involved in mamma...	BP	4	2	0	0.0011775333
GO:0035162	embryonic hemopoiesis	BP	16	3	0	0.0013632772
GO:1902749	regulation of cell cycle G2/M phase transition	BP	119	7	0	0.0014481447
GO:0043970	histone H3-K9 acetylation	BP	5	2	0	0.0019443684
GO:0060745	mammary gland branching involved in pregnancy	BP	5	2	0	0.0019443684
GO:2000615	regulation of histone H3-K9 acetylation	BP	5	2	0	0.0019443684
GO:1902750	negative regulation of cell cycle G2/M phase tr...	BP	64	5	0	0.0020585739
GO:0008209	androgen metabolic process	BP	19	3	0	0.0022871160
GO:0043966	histone H3 acetylation	BP	19	3	0	0.0022871160
GO:0022603	regulation of anatomical structure morphogenesi...	BP	567	17	1	0.0024081726
GO:0043353	enucleate erythrocyte differentiation	BP	6	2	0	0.0028895537

Figura 19. *Gene Enrichment Analysis entre las 2 poblaciones (latina y blanca).*

Se observa que los principales genes diferencialmente expresados entre las muestras latina y blanca son proteinas implicadas en los *checkpoint* de la mitosis celular y de la reparación del daño del DNA. También llama la atención la diferencia de expresión de genes implicados en el metabolismo de las hormonas tiroidales así como en el proceso de la hematopoesis.

4.1.2 LATINA – ASIÁTICA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 5 muestras de sujetos de población latina y 30 muestras de sujetos de población asiática.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene <chr>	latino1 <dbl>	latino2 <dbl>	latino3 <dbl>	latino4 <dbl>	latino5 <dbl>	asiatico1 <dbl>	asiatico2 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041	154.7218	407.6782
5T4	933.7998	796.9726	189.8969	489.3010	345.8573	621.4165	334.5521
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578	948.9882	3186.0146
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432	0.4216	1.8282
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937	5.9022	155.3931
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491	9.2580	25.8090
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000	16.4418	806.6728
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957	1771.9098	45917.6920
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607	11.3954	35.6901
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811	1.6863	0.4570

Figura 20. *Datos crudos de las 2 poblaciones a analizar (latina y asiática).*

De los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 35 muestras (5 muestras de sujetos población latina y 30 muestras de sujetos de población asiática).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (latino vs. asiático), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

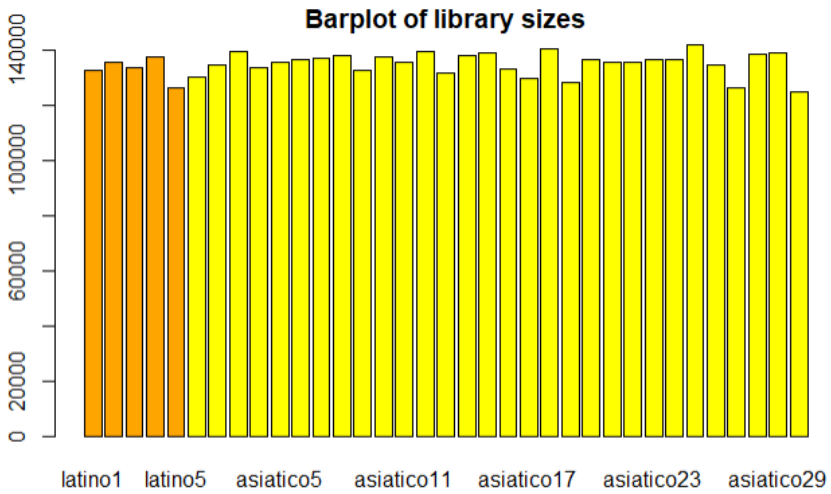


Figura 21. *Barplot de las 2 poblaciones. Población latina de color naranja y población asiática de color amarillo.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aproximándose a los 14.000 recuentos para la mayoría de las muestras. Algunas muestras están entre los 12.000 y los 13.000 recuentos.

BOXPLOT

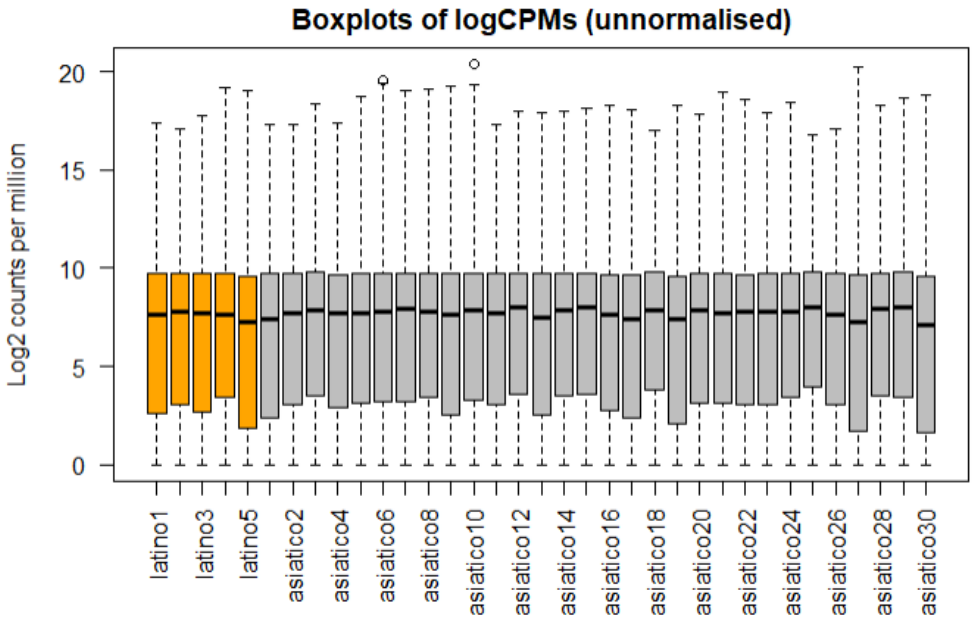


Figura 22. *Boxplot de las 2 poblaciones. Población latina de color naranja y población asiática de color amarillo.*

Se observa como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre les 35 muestras, con algún *outlier* ("asiatico6" y "asiatico10").

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

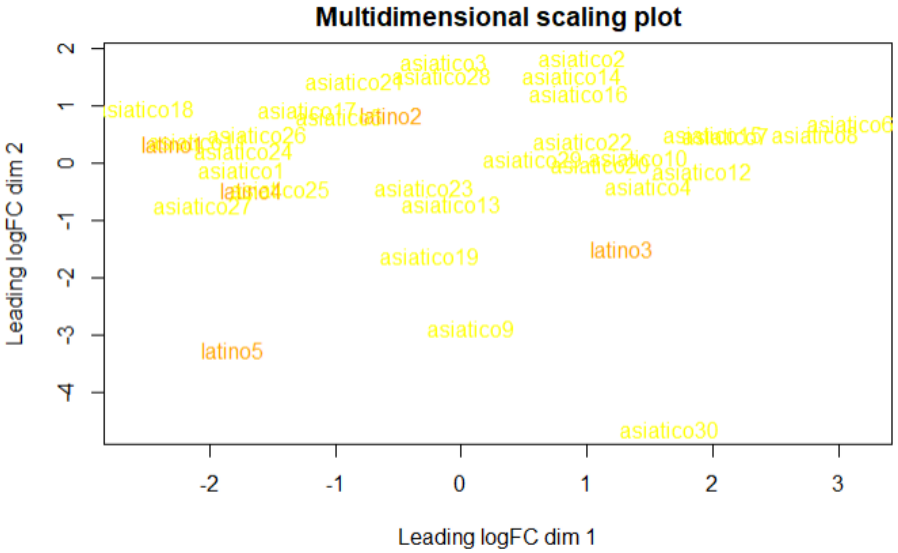


Figura 23. *MDS Plot de las 2 poblaciones (latina y asiática).*

En este caso se observa solapamiento en la distribución de las muestras latina y asiática. El hecho de disponer de solamente 5 muestras para la población latina representa una limitación en éste análisis. Parece que hay una muestra "asiático30" con una distribución muy diferente al resto, pudiendo suponer un outlier.

CLUSTERING I *HEATMAP*

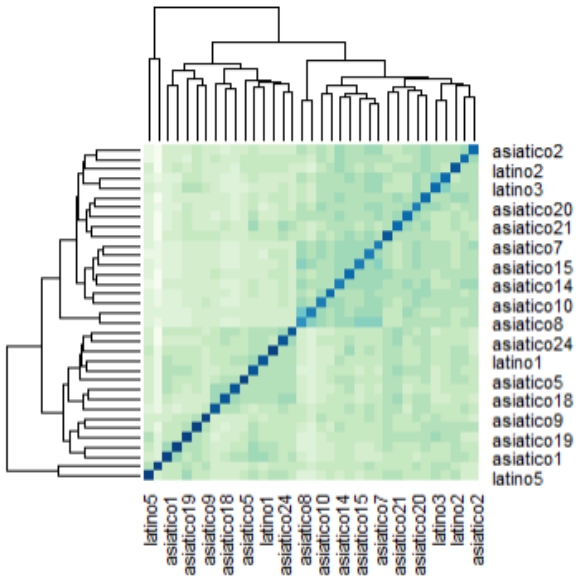


Figura 24. *Clustering i heatmap de las 2 poblaciones (latina y asiática).*

Se observa solapamiento entre las poblaciones latina y asiática, debido probablemente a la poca cantidad de muestras de la población latina.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 4 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 16234 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18620556	1.0064011
latino2	latino	18631953	1.0540321
latino3	latino	18890329	1.0507484
latino4	latino	20773795	0.9739578
latino5	latino	20541564	0.8483977
asiatico1	asiatico	18620363	1.0027432
asiatico2	asiatico	17369919	1.1226419
asiatico3	asiatico	19445800	1.1232011
asiatico4	asiatico	17090138	1.1652362
asiatico5	asiatico	18483524	1.0739849

1-10 of 35 rows Previous 1 2 3 4 Next

Figura 25. *Datos normalizados de las 2 poblaciones (latina y asiática).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDS PLOT

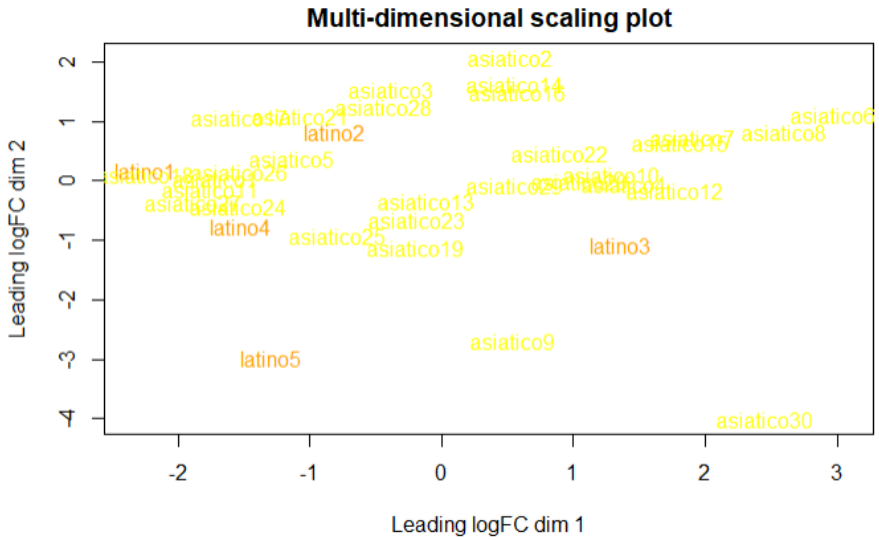
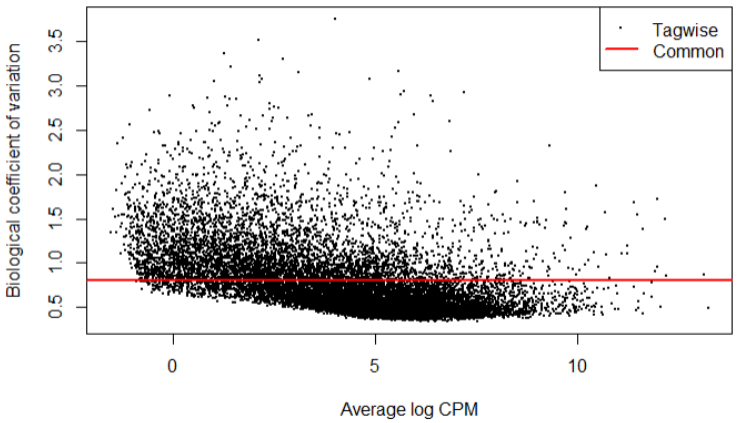


Figura 26. *MDS Plot con los datos normalizados de las 2 poblaciones (latina y asiática).*

Se observa que las posiciones de las muestras han variado respecto el MDS Plot de datos no normalizados, pero que las distancias entre las muestras se mantienen similares. En este caso, el MDS Plot con los datos normalizados es muy similar al de los datos crudos, con solapamiento de las muestras de la población latina con las de la población asiática y un posible *outlier* que corresponde a la muestra "asiatico30".

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 65%
Dispersión específica: 65%

Figura 27. *Estimación de la dispersión de los datos de las 2 poblaciones (latina y asiática).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.8 aunque hay una dispersión de 65% con el coeficiente de variación biológica de algún gen superior a 3.5.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

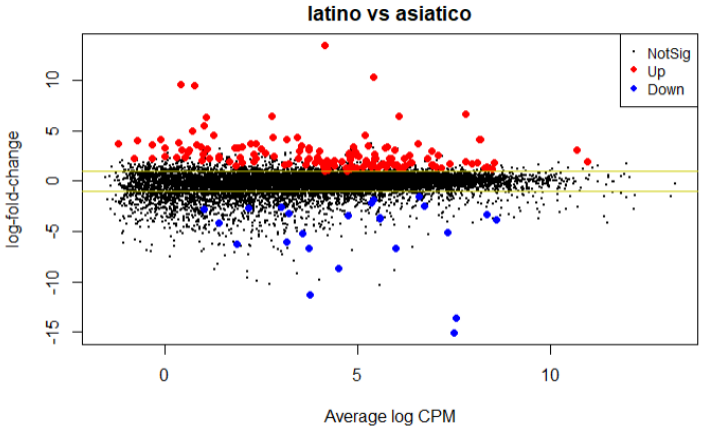
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
6881	HBM	13.473010	4.1482787	2.030509e-177	3.296328e-173
16262	SERP1	-13.563032	7.5645738	7.222338e-78	5.862371e-74
16263	SERP1	10.289655	5.4029635	4.374274e-60	2.367065e-56
2576	CDC23	3.671667	6.5662711	6.891487e-59	2.796910e-55
2123	CACNG5	9.497384	0.7597795	4.963308e-53	1.611487e-49
6641	GPX6	9.601543	0.4073036	2.802309e-52	7.582114e-49
1781	C1orf84	4.327485	3.4479448	5.192417e-45	1.204196e-41
6882	HBM	-15.040521	7.4978943	3.178721e-42	6.450420e-39
1780	C1orf84	-3.783787	5.5748022	4.975332e-26	8.974392e-23
8883	LINC00875	3.744956	0.3493408	2.485846e-22	4.035523e-19

1-10 of 10 rows

Figura 28. *10 genes diferencialmente expresados entre las 2 poblaciones (latina y asiática).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	latino-asiatico
Down	24
NotSig	16064
Up	146

Figura 29. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (latina y asiática).*

Se observa que 24 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 146 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 16064 genes no muestran una expresión diferencial significativa.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Observamos las 20 anotaciones GO enriquecidas:

	Term	Ont	N	Up	Do...	PUp
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0044818	mitotic G2/M transition checkpoint	BP	14	4	0	1.202380e-05
GO:0016573	histone acetylation	BP	59	6	0	3.924684e-05
GO:0018393	internal peptidyl-lysine acetylation	BP	62	6	0	5.217763e-05
GO:0035065	regulation of histone acetylation	BP	20	4	0	5.537492e-05
GO:1902275	regulation of chromatin organization	BP	64	6	0	6.254671e-05
GO:0007095	mitotic G2 DNA damage checkpoint	BP	8	3	0	6.703004e-05
GO:0006475	internal protein amino acid acetylation	BP	66	6	0	7.450101e-05
GO:0018394	peptidyl-lysine acetylation	BP	66	6	0	7.450101e-05
GO:2000756	regulation of peptidyl-lysine acetylation	BP	23	4	0	9.872463e-05
GO:0006473	protein acetylation	BP	75	6	0	1.527741e-04
GO:1901983	regulation of protein acetylation	BP	26	4	0	1.625937e-04
GO:0031056	regulation of histone modification	BP	50	5	0	1.949826e-04
GO:0010389	regulation of G2/M transition of mitotic cell c...	BP	114	7	0	2.246389e-04
GO:0031572	G2 DNA damage checkpoint	BP	12	3	0	2.551707e-04
GO:0043543	protein acylation	BP	84	6	0	2.853825e-04
GO:0035850	epithelial cell differentiation involved in kid...	BP	30	4	0	2.883628e-04
GO:1902749	regulation of cell cycle G2/M phase transition	BP	120	7	0	3.082585e-04
GO:0006325	chromatin organization	BP	294	11	1	3.100584e-04
GO:0016570	histone modification	BP	160	8	1	3.250721e-04
GO:0035066	positive regulation of histone acetylation	BP	13	3	0	3.291239e-04

Figura 30. *Gene Enrichment Analysis entre las 2 poblaciones (latina y asiática).*

Se observa que los principales genes diferencialmente expresados entre las muestras de la población latina y asiática son proteinas implicadas en los *checkpoint* de la mitosis celular y de la regulación de la acetilación.

4.1.3 LATINA – NEGRA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 5 muestras de sujetos de población latina y 6 muestras de sujetos de población negra.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene	latino1	latino2	latino3	latino4	latino5	negro1	negro2
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041	179.3048	180.0302
5T4	933.7998	796.9726	189.8969	489.3010	345.8573	97.1634	208.3866
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578	1899.1610	1153.3801
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432	46.3444	0.0000
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937	15.3416	19.7835
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491	9.0515	4.7942
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000	233.9592	12.1998
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957	2985.1506	3249.2907
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607	17.0068	20.5980
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811	0.0000	0.0000

Figura 31. *Datos crudos de las 2 poblaciones a analizar (latina y negra).*

De los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 11 muestras (5 muestras de sujetos de población latina y 6 muestras de sujetos de población negra).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (latina vs. negra), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

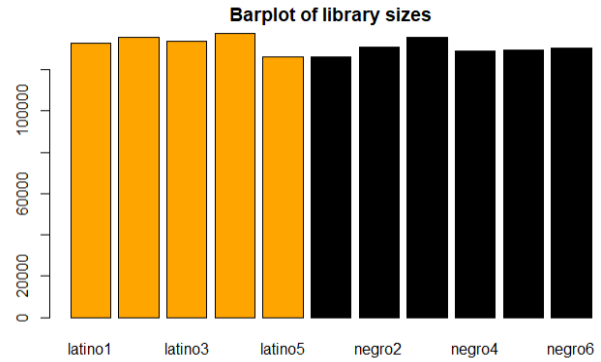


Figura 32. *Barplot de las 2 poblaciones. Población latina de color naranja y población negra de color negro.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aproximándose a los 14.000 recuentos para la mayoría de las muestras.

BOXPLOT

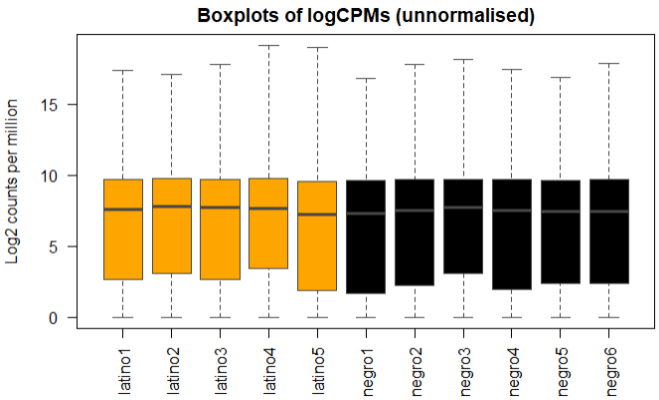


Figura 33. *Boxplot de las 2 poblaciones. Población latina de color naranja y población negra de color negro.*

Se observa como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre les 11 muestras, sin valores alejados que sugieran la presencia de *outliers*.

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

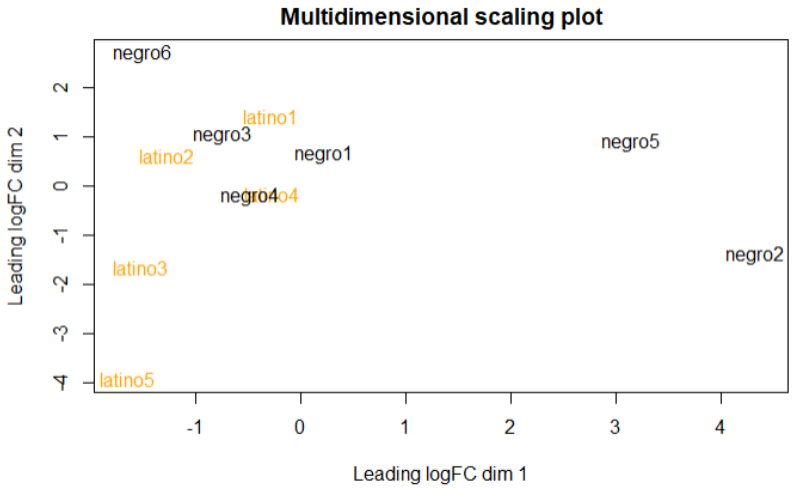


Figura 34. *MDS Plot de las 2 poblaciones (latina y negra).*

En este caso se observa solapamiento en la distribución de las muestras “latino1”, “latino2”, “latino4”, “negro1”, “negro3” y “negro4”. El hecho de disponer de solamente 5 muestras para la población latina y 6 muestras para la población negra representa una limitación en este análisis. Las muestras restantes de las 2 poblaciones se distribuyen por la zona inferior izquierda (población latina) y centro derecha (población negra).

CLUSTERING I *HEATMAP*

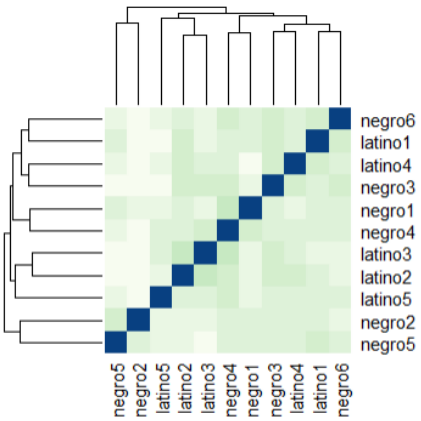


Figura 35. *Clustering i heatmap de las 2 poblaciones (latina y negra).*

Se observa solapamiento entre las 2 poblaciones, tal y como se observa en la figura 34.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 4 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 14645 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18630051	1.0527938
latino2	latino	18645375	1.0658185
latino3	latino	18896201	1.0460372
latino4	latino	20788823	0.9945032
latino5	latino	20549562	0.8383894
negro1	negro	16898409	1.0197157
negro2	negro	17616396	1.0482930
negro3	negro	20654738	0.9975418
negro4	negro	17569187	1.0629037
negro5	negro	17602053	1.0174362

1-10 of 11 rows Previous 1 2 Next

Figura 36. *Datos normalizados de las 2 poblaciones (latina y negra).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDS PLOT

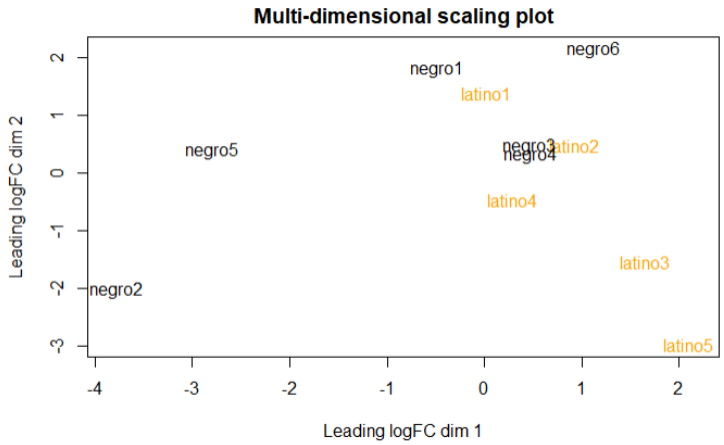
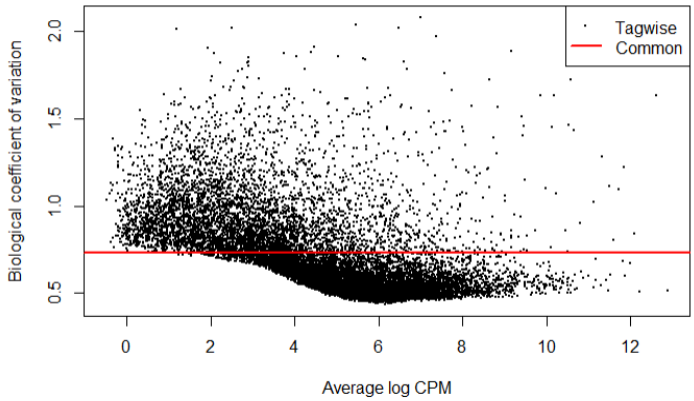


Figura 37. *MDS Plot con los datos normalizados de las 2 poblaciones (latina y negra).*

Se observa que las posiciones de las muestras han variado respecto el MDS Plot de datos no normalizados, y ha variado la distancia entre las muestras que antes formaban un clúster: se mantiene e incluso se ha acortado la distancia entre “latino2”, “negro3” y “negro4”, pero se han distanciado “latino1”, “latino4” y “negro1”. En este caso el resto de muestras se distribuyen por la zona centro izquierda (población negra) y inferior derecha (población latina).

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 53%
Dispersión específica: 53%

Figura 38. *Estimación de la dispersión de los datos de las 2 poblaciones (latina y negra).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.7-0.8 aunque hay una dispersión de 53% con el coeficiente de variación biológica de algún gen superior a 2.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

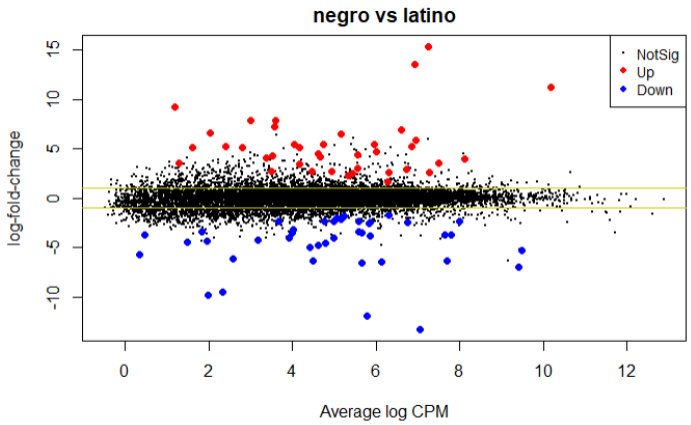
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
16263	SERP1	-13.271092	7.045378	1.693247e-61	2.479760e-57
6881	HBM	-11.896387	5.793760	1.556575e-55	1.139802e-51
16262	SERP1	13.556840	6.920802	1.493133e-45	7.288976e-42
6882	HBM	15.355558	7.237744	5.847657e-45	2.140974e-41
2123	CACNG5	-9.536060	2.344795	7.984572e-18	2.338681e-14
6641	GPX6	-9.792711	1.985593	1.015163e-17	2.477845e-14
1781	C1orf84	-4.544254	4.790158	1.437221e-16	3.006873e-13
2577	CDC23	3.964652	8.115425	2.585076e-15	4.732304e-12
16397	SH3D20	4.654733	5.999586	2.944782e-14	4.791814e-11
2576	CDC23	-3.761453	7.788096	9.600302e-14	1.405964e-10

1-10 of 10 rows

Figura 39. *10 genes diferencialmente expresados entre las 2 poblaciones (latina y negra).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	negro-latino
Down	41
Notsig	14565
Up	39

Figura 40. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (latina y negra).*

Se observa que 41 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 39 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 14565 genes no muestran una expresión diferencial significativa.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Observamos las 20 anotaciones GO enriquecidas:

	Term	Ont	N	Up	Do...	P.Up
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0006814	sodium ion transport	BP	90	3	0	0.002993449
GO:0003290	atrial septum secundum morphogenesis	BP	1	1	0	0.003287414
GO:0015878	biotin transport	BP	1	1	0	0.003287414
GO:0061026	cardiac muscle tissue regeneration	BP	1	1	0	0.003287414
GO:0015939	pantothenate metabolic process	BP	1	1	0	0.003287414
GO:0015887	pantothenate transmembrane transport	BP	1	1	0	0.003287414
GO:0051891	positive regulation of cardioblast differentiat...	BP	1	1	0	0.003287414
GO:0051890	regulation of cardioblast differentiation	BP	1	1	0	0.003287414
GO:0003285	septum secundum development	BP	1	1	0	0.003287414
GO:0048247	lymphocyte chemotaxis	BP	34	2	0	0.005420188
GO:0002548	monocyte chemotaxis	BP	35	2	0	0.005737331
GO:0070374	positive regulation of ERK1 and ERK2 cascade	BP	116	3	0	0.006113230
GO:0035054	embryonic heart tube anterior/posterior pattern...	BP	2	1	0	0.006564534
GO:1902083	negative regulation of peptidyl-cysteine S-nitr...	BP	2	1	0	0.006564534
GO:0086003	cardiac muscle cell contraction	BP	39	2	0	0.007088887
GO:0003289	atrial septum primum morphogenesis	BP	3	1	0	0.009831390
GO:0006768	biotin metabolic process	BP	3	1	0	0.009831390
GO:0003284	septum primum development	BP	3	1	0	0.009831390
GO:0072679	thymocyte migration	BP	3	1	0	0.009831390
GO:0071674	mononuclear cell migration	BP	49	2	0	0.011030440

Figura 41. *Gene Enrichment Analysis entre las 2 poblaciones (latina y negra).*

Se observa que los principales genes diferencialmente expresados entre las muestras de población latina y negra son proteinas implicadas en el transporte de moléculas (sodio, biotina, vitamina B_5 (*pantothenate*)).

4.1.4 BLANCA – ASIÁTICA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 30 muestras de sujetos de población blanca y 30 muestras de sujetos de población asiática.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene	blanco1	blanco2	blanco3	blanco4	blanco5	blanco6	blanco7
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
133K02	470.6122	342.4043	224.6234	455.9140	179.0654	469.7354	229.1074
5T4	351.4286	1018.0377	115.2587	213.2616	815.7009	287.0605	887.5221
A-362G6.1	2365.7143	1357.1056	891.2901	1867.3835	2087.8505	1699.1664	1886.3689
A-C1	1.2245	0.4171	0.0000	2.5090	0.0000	0.0000	1.5376
a1/3GTP	175.1020	99.6768	90.3733	34.4086	35.5140	171.0765	35.6731
A1BG-AS1	26.5837	17.6624	29.6398	13.4301	9.9215	36.3791	17.2615
A1CF	105.7143	0.4171	0.0000	141.2186	349.1589	5.4368	18.1441
A2M	26388.3633	8842.9528	4547.7472	17226.2222	6961.6449	32182.7800	17942.2588
A2M-AS1	84.2898	22.4419	14.7937	40.0860	27.5140	58.2494	34.8305
A2ML1	0.8163	1.6682	52.3903	34.0502	0.3738	0.0000	56.8924

Figura 42. *Datos crudos de las 2 poblaciones a analizar (blanca y asiática).*

De los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 61 muestras (30 muestras de sujetos de población blanca y 30 muestras de sujetos de población asiática).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (blanca vs. asiática), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

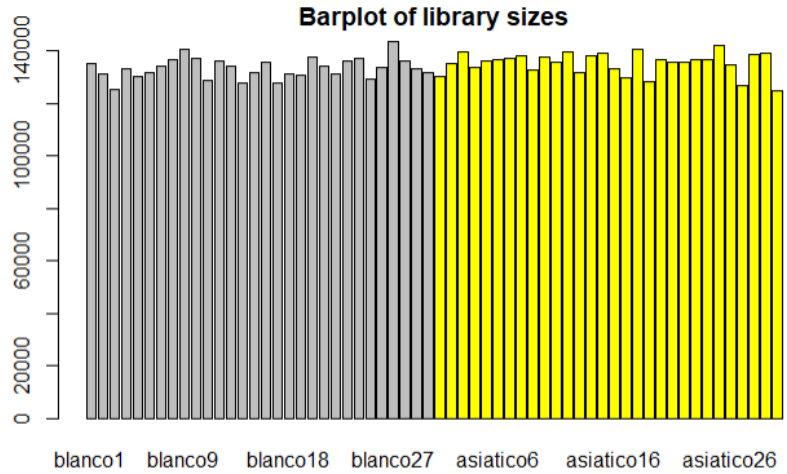


Figura 43. *Barplot de las 2 poblaciones. Población blanca de color gris y población asiática de color amarillo.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aproximándose a los 14.000 recuentos la gran mayoría las muestras. Algunas muestras están entre los 12.000 y los 13.000 recuentos.

BOXPLOT

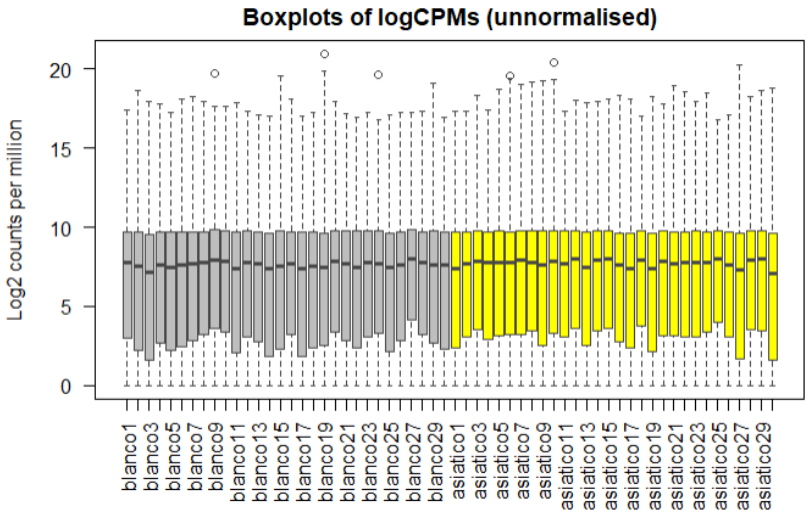


Figura 44. *Boxplot de las 2 poblaciones. Población blanca de color gris y población asiática de color amarillo.*

Se observa como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre les 60 muestras, con algún valor atípico ("blanco9", "blanco19", "blanco24", "asiatico6", "asiatico10").

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

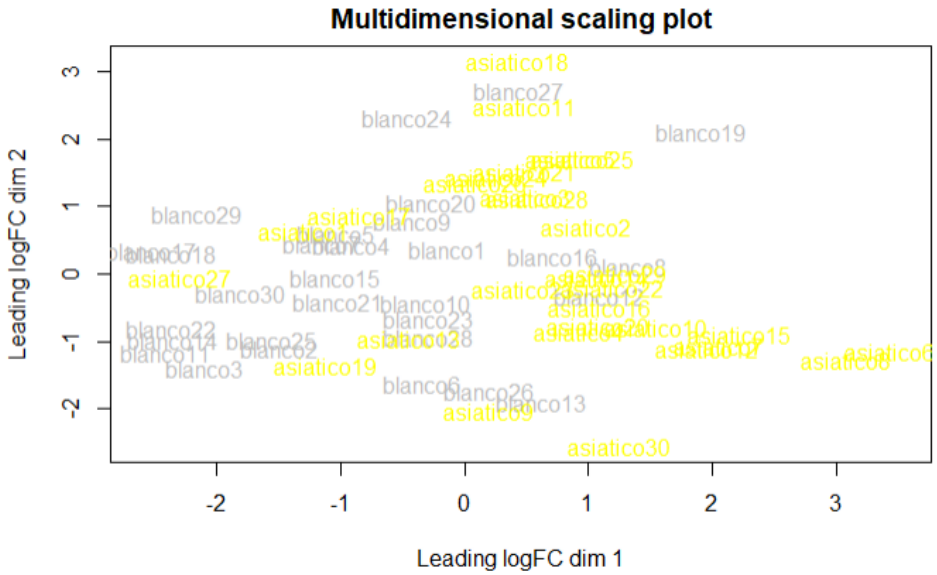


Figura 45. *MDS Plot de las 2 poblaciones (blanca y asiática).*

En este caso se observa solapamiento en la distribución de las muestras des las poblaciones blanca y asiática, aunque se puede observar cierta tendencia a agrupación de las muestras de población blanca a la izquierda y población asiática al centro-derecha. Se puede observar como la muestra "blanco19", "blanco24" y "blanco27" están más separadas del resto de muestras de la población blanca. Tanto "blanco19" como "blanco24" se han objectivado como posible *otulier* en la figura 44. Las muestras "asiatico6", "asiatico8", "asiatico11", "asiatico18" y "asiatico30" presentan separación respecto del resto de muestras de la población asiática. La muestra "asiatico10" ya se ha objectivado como posible *outlier* en la figura 44.

CLUSTERING I *HEATMAP*

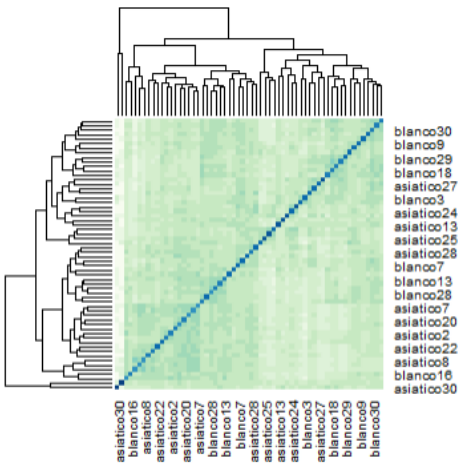


Figura 46. *Clustering i heatmap de las 2 poblaciones (blanca y asiática).*

Se observa solapamiento entre las poblaciones blanca y asiática, aunque se intuye cierta tendencia a agruparse por poblaciones.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 10 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 15735 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
blanco1	blanco	18116701	1.0651233
blanco2	blanco	21666671	0.8469262
blanco3	blanco	19336514	0.9058210
blanco4	blanco	18591124	1.0383091
blanco5	blanco	17321993	1.0745210
blanco6	blanco	20372930	0.8840328
blanco7	blanco	19386370	1.0046391
blanco8	blanco	20144072	1.0151853
blanco9	blanco	20657090	1.0374808
blanco10	blanco	19081573	1.0879939

1-10 of 60 rows Previous 1 2 3 4 5 6 Next

Figura 47. *Datos normalizados de las 2 poblaciones (blanca y asiática).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDSPLOT

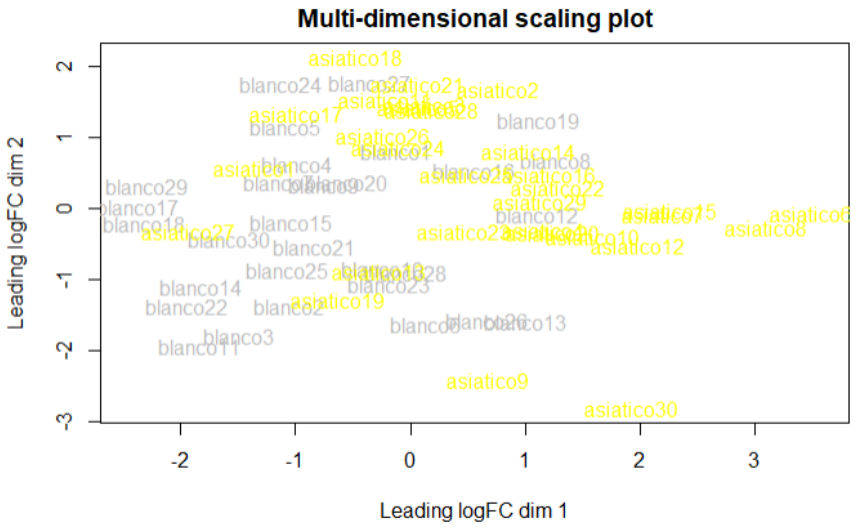
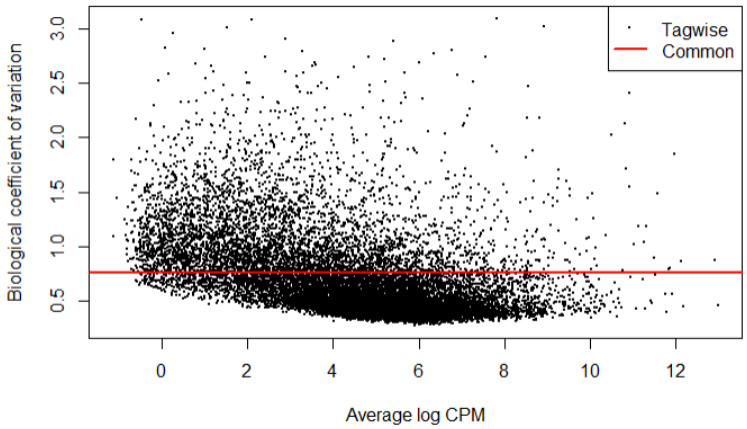


Figura 48. *MDS Plot con los datos normalizados de las 2 poblaciones (blanca y asiática).*

Se observa que las posiciones y las distancias entre las muestras han variado muy poco respecto el MDSplot de datos no normalizados, existiendo solapamiento entre las 2 poblaciones aunque sí cierta tendencia a agruparse por población.

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 58%
Dispersión específica: 58%

Figura 49. *Estimación de la dispersión de los datos de las 2 poblaciones (blanca y asiática).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.7-0.8 aunque hay una dispersión de 58% con el coeficiente de variación biológica de algún gen superior a 3.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

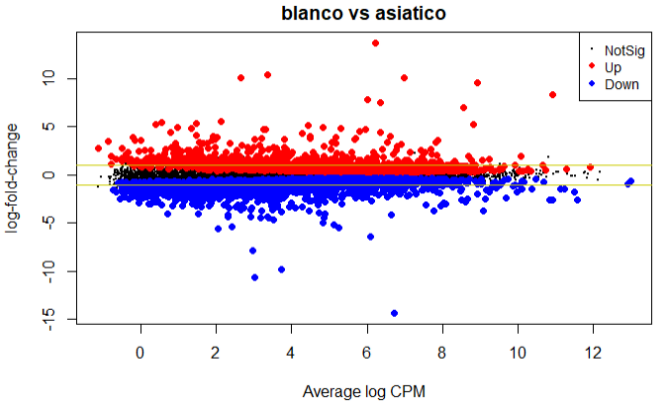
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
6881	HBM	13.745155	6.233778	1.661392e-257	2.614200e-253
6882	HBM	-14.315274	6.716651	7.831904e-244	6.161750e-240
1780	C1orf84	-5.007911	4.833406	1.726523e-178	9.055611e-175
16263	SERP1	10.054022	6.982697	6.674868e-136	2.625726e-132
2576	CDC23	3.609035	7.864344	3.020605e-133	9.505845e-130
6640	GPX6	-10.668170	3.010106	6.728902e-111	1.764655e-107
2577	CDC23	-3.254117	7.952363	2.670354e-102	6.002575e-99
6641	GPX6	10.126547	2.659420	6.321927e-102	1.243444e-98
1781	C1orf84	3.830561	4.481343	5.433903e-89	9.500273e-86
2123	CACNG5	10.374018	3.373530	5.839267e-86	9.188087e-83

1-10 of 10 rows

Figura 50. *10 genes diferencialmente expresados entre las 2 poblaciones (blanca y asiática).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	blanco-asiatico
Down	2807
NotSig	10728
Up	2200

Figura 51. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (blanca y asiática).*

Se observa que 2807 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 2200 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 10728 genes no muestran una expresión diferencial significativa.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Observamos las 20 anotaciones GO enriquecidas:

	Term	Ont	N	Up	Do...	PUp
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0010517	regulation of phospholipase activity	BP	50	19	7	1.887245e-05
GO:0060191	regulation of lipase activity	BP	67	22	10	6.013840e-05
GO:0043372	positive regulation of CD4-positive, alpha-beta...	BP	18	9	0	2.805666e-04
GO:0009067	aspartate family amino acid biosynthetic proces...	BP	12	7	1	4.126636e-04
GO:0010518	positive regulation of phospholipase activity	BP	43	15	7	4.250639e-04
GO:2000514	regulation of CD4-positive, alpha-beta T cell a...	BP	35	13	4	5.107696e-04
GO:2000516	positive regulation of CD4-positive, alpha-beta...	BP	23	10	1	5.301206e-04
GO:0002828	regulation of type 2 immune response	BP	20	9	2	7.463132e-04
GO:2000482	regulation of interleukin-8 secretion	BP	17	8	2	1.032235e-03
GO:0001819	positive regulation of cytokine production	BP	246	52	46	1.038897e-03
GO:0032729	positive regulation of interferon-gamma product...	BP	38	13	4	1.253435e-03
GO:0060193	positive regulation of lipase activity	BP	48	15	8	1.561033e-03
GO:1903555	regulation of tumor necrosis factor superfamily...	BP	88	23	14	1.646727e-03
GO:0031584	activation of phospholipase D activity	BP	5	4	1	1.656407e-03
GO:0030540	female genitalia development	BP	11	6	1	1.764527e-03
GO:0001817	regulation of cytokine production	BP	355	69	63	1.907212e-03
GO:0071706	tumor necrosis factor superfamily cytokine prod...	BP	89	23	15	1.935626e-03
GO:0071888	macrophage apoptotic process	BP	8	5	0	2.004969e-03
GO:0048806	genitalia development	BP	31	11	7	2.104599e-03
GO:0045624	positive regulation of T-helper cell differenti...	BP	15	7	0	2.289769e-03

Figura 52. *Gene Enrichment Analysis entre las 2 poblaciones (blanca y asiática).*

Se observa que los principales genes diferencialmente expresados entre las muestras blanca y asiática son proteinas implicadas en la respuesta inmune (regulación de CD4, citoquinas, respuesta inmunológica, TNF, apoptosis de macrófagos, regulación positiva de T-helper, …).

4.1.5 BLANCA – NEGRA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 30 muestras de sujetos de población blanca y 6 muestras de sujetos de población negra.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene	blanco1	blanco2	blanco3	blanco4	blanco5	blanco6	blanco7
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
133K02	470.6122	342.4043	224.6234	455.9140	179.0654	469.7354	229.1074
5T4	351.4286	1018.0377	115.2587	213.2616	815.7009	287.0605	887.5221
A-362G6.1	2365.7143	1357.1056	891.2901	1867.3835	2087.8505	1699.1664	1886.3689
A-C1	1.2245	0.4171	0.0000	2.5090	0.0000	0.0000	1.5376
a1/3GTP	175.1020	99.6768	90.3733	34.4086	35.5140	171.0765	35.6731
A1BG-AS1	26.5837	17.6624	29.6398	13.4301	9.9215	36.3791	17.2615
A1CF	105.7143	0.4171	0.0000	141.2186	349.1589	5.4368	18.1441
A2M	26388.3633	8842.9528	4547.7472	17226.2222	6961.6449	32182.7800	17942.2588
A2M-AS1	84.2898	22.4419	14.7937	40.0860	27.5140	58.2494	34.8305
A2ML1	0.8163	1.6682	52.3903	34.0502	0.3738	0.0000	56.8924

Figura 53. *Datos crudos de las 2 poblaciones a analizar (blanca y negra).*

De los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 36 muestras (30 muestras de sujetos de población blanca y 6 muestras de sujetos de población negra).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (blanco vs. negro), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

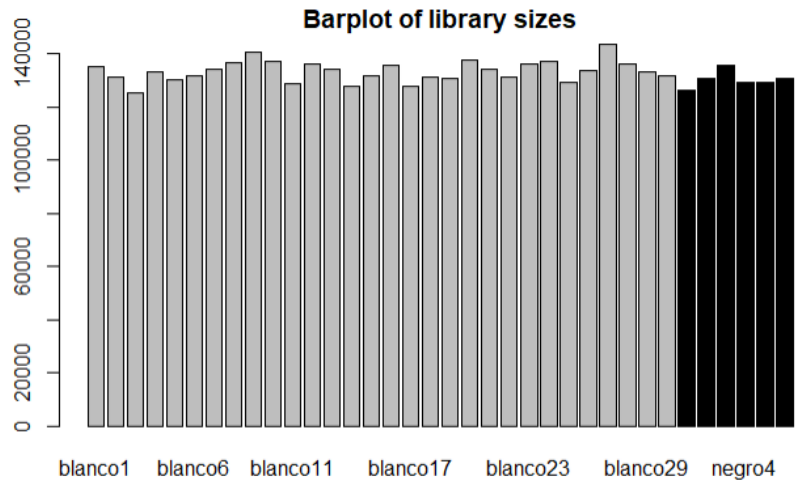


Figura 54. *Barplot de las 2 poblaciones. Población blanca de color gris y población negra de color negro.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aproximándose a los 14.000 recuentos para la gran mayoría de las muestras. Algunas muestras se acercan a los 13000 recuentos.

BOXPLOT

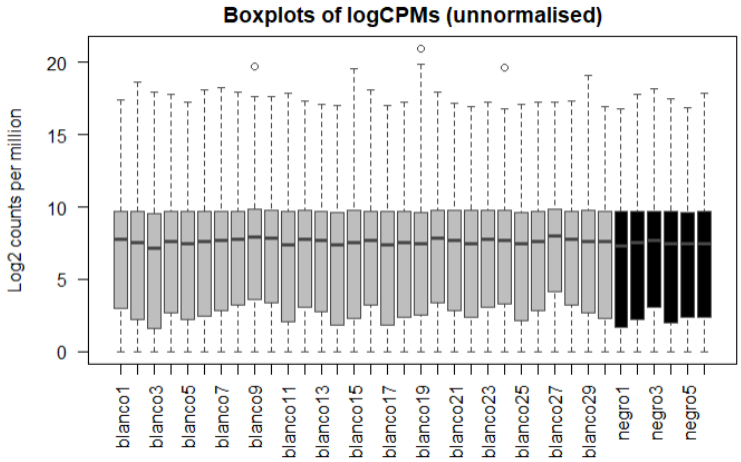


Figura 55. *Boxplot de las 2 poblaciones. Población blanca de color gris y población negra de color negro.*

Se observa como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre les 36 muestras, con algún valor atípico ("blanco9", "blanco19", "blanco24").

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

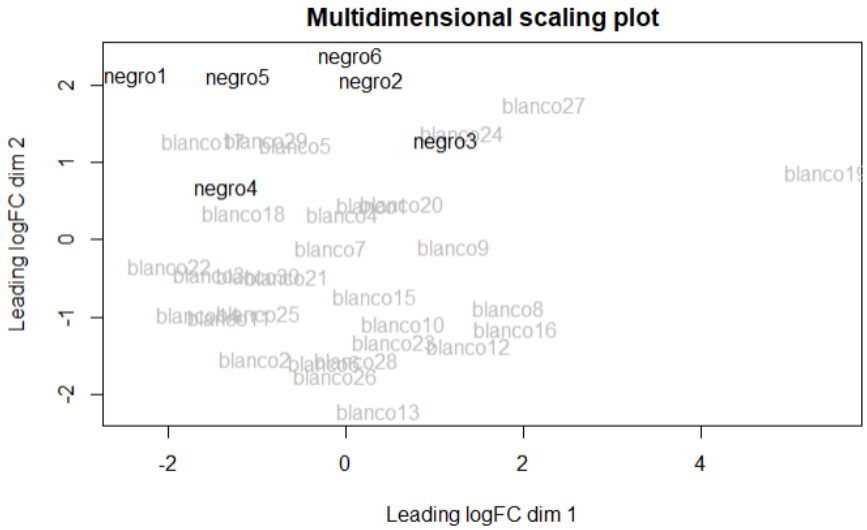


Figura 56. *MDS Plot de las 2 poblaciones (blanca y negra).*

En este caso se observa como la población negra se distribuye a la parte superior izquierda, mientras que la población blanca se distribuye por gran parte de la zona izquierda, aunque se puede observar solapamiento entre las muestras "negro3" y "negro4" con las muestras de la población blanca. La muestra "blanco19" se encuentra alejada de la distribución de la población, pudiendo ser un posible *outlier* (se observó también esta muestra como posible *outlier* en la comparación "latino-blanco").

CLUSTERING I *HEATMAP*

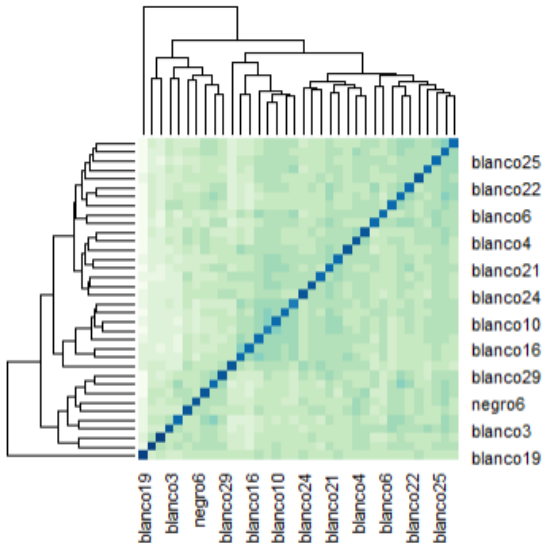


Figura 57. *Clustering i heatmap de las 2 poblaciones (blanca y negra).*

Se observa agrupación de las muestras de la población blanca con solapamiento de alguna muestra de la población negra. Como ya se ha explicado, la poca cantidad de muestras de la población negra puede ser el origen de este solapamiento y la falta de agrupación de la población negra.

FILTRADO

Dado que en este caso existe la limitación que la población latina solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 5 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 15785 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
blanco1	blanco	18116701	1.0859384
blanco2	blanco	21666671	0.8163791
blanco3	blanco	19336514	0.8183609
blanco4	blanco	18591124	1.0036928
blanco5	blanco	17321993	1.0175219
blanco6	blanco	20372930	0.9006501
blanco7	blanco	19386370	0.9766139
blanco8	blanco	20144072	1.0056215
blanco9	blanco	20657090	1.0524179
blanco10	blanco	19081573	1.0976538

1-10 of 36 rows Previous 1 2 3 4 Next

Figura 58. *Datos normalizados de las 2 poblaciones (blanca y negra).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDSPLOT

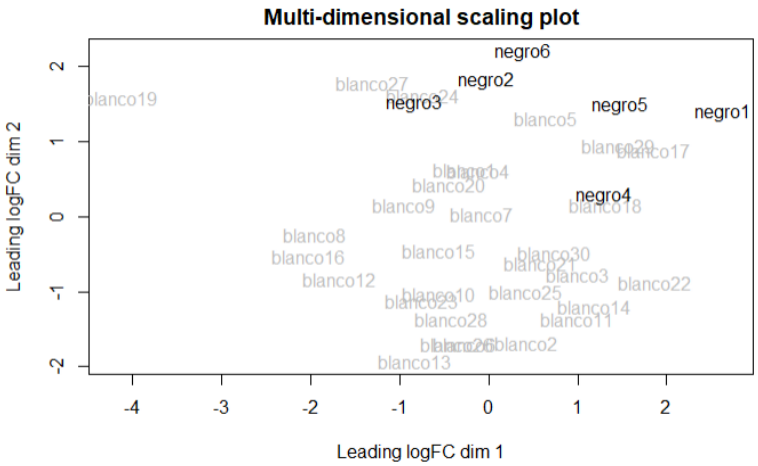
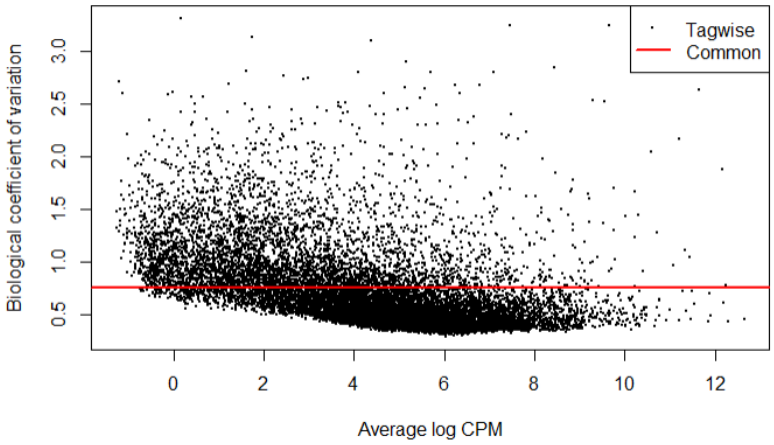


Figura 59. *MDS Plot con los datos normalizados de las 2 poblaciones (blanca y negra).*

Se observa que las posiciones han variado muy poco respecto el MDSplot de datos no normalizados. Las muestras de la población negra se sitúan en la parte superior (en esta ocasión en la parte superior derecha), pero algunas de ellas (“negro3”, “negro4” y “negro5”) se

acercan e incluso se solapan con las muestras de la población blanca. También se observa la muestra "blanco19" como posible *outlier*.

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 58%
Dispersión específica: 58%

Figura 60. *Estimación de la dispersión de los datos de las 2 poblaciones (blanca y negra).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.7-0.8 aunque hay una dispersión de 58% con el coeficiente de variación biológica de algún gen aproximadamente a 3.5.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

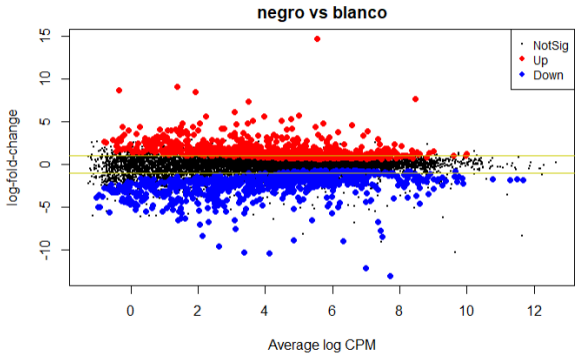
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
6882	HBM	14.704752	5.5425954	3.699611e-249	5.839835e-245
16263	SERP1	-13.058999	7.7222252	1.941545e-81	1.532365e-77
1780	C1orf84	5.625442	3.9622380	6.511601e-76	3.426188e-72
2577	CDC23	3.384899	6.9118874	2.209035e-67	8.717404e-64
6881	HBM	-12.202407	6.9847853	4.768518e-48	1.505421e-44
6640	GPX6	8.696852	-0.3460349	7.492516e-46	1.971156e-42
2626	CDH24	8.440175	1.9322206	4.601855e-44	1.037718e-40
2122	CACNG5	9.032549	1.3898141	2.829655e-37	5.583263e-34
19683	WASH5P	2.983787	4.1378730	9.292183e-30	1.629746e-26
16397	SH3D20	5.109478	4.4894837	3.791833e-27	5.985408e-24

1-10 of 10 rows

Figura 61. *10 genes diferencialmente expresados entre las 2 poblaciones (blanca y negra).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	negro-blanco
Down	770
NotSig	14002
Up	1013

Figura 62. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (blanca y negra).*

Se observa que 770 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 1013 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 14002 genes no muestran una expresión diferencial significativa.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Observamos las 20 anotaciones GO enriquecidas:

	Term <chr>	Ont <chr>	N <dbl>	Up <dbl>	Do... <dbl>	P.Up <dbl>
GO:0062009	secondary palate development	BP	14	6	0	0.0001311588
GO:0061408	positive regulation of transcription from RNA p...	BP	3	3	0	0.0002636328
GO:1902305	regulation of sodium ion transmembrane transpor...	BP	30	8	1	0.0004537230
GO:0034599	cellular response to oxidative stress	BP	169	23	7	0.0004651880
GO:0034447	very-low-density lipoprotein particle clearance	BP	7	4	0	0.0005037574
GO:0030510	regulation of BMP signaling pathway	BP	41	9	1	0.0009504470
GO:0018410	C-terminal protein amino acid modification	BP	4	3	0	0.0010040313
GO:0060022	hard palate development	BP	4	3	0	0.0010040313
GO:0086068	Purkinje myocyte to ventricular cardiac muscle ...	BP	4	3	0	0.0010040313
GO:0086029	Purkinje myocyte to ventricular cardiac muscle ...	BP	4	3	0	0.0010040313
GO:2000977	regulation of forebrain neuron differentiation	BP	4	3	0	0.0010040313
GO:1900407	regulation of cellular response to oxidative st...	BP	50	10	1	0.0010851566
GO:0030509	BMP signaling pathway	BP	77	13	2	0.0011129732
GO:1902882	regulation of response to oxidative stress	BP	52	10	1	0.0014890251
GO:1900408	negative regulation of cellular response to oxi...	BP	28	7	0	0.0015646003
GO:1903202	negative regulation of oxidative stress-induced...	BP	28	7	0	0.0015646003
GO:0000305	response to oxygen radical	BP	21	6	0	0.0016105104
GO:0034629	cellular protein-containing complex localizatio...	BP	9	4	1	0.0016346797
GO:1902883	negative regulation of response to oxidative st...	BP	29	7	0	0.0019503574
GO:0071773	cellular response to BMP stimulus	BP	83	13	3	0.0022536989

Figura 63. *Gene Enrichment Analysis entre las 2 poblaciones (blanca y negra).*

Se observa que los principales genes diferencialmente expresados entre las muestras blanca y negra son proteinas implicadas en la regulación (tanto positiva como negativa) de diferentes procesos como son la transcripción del RNA, la vía de señalización BMP, la respuesta oxidativa al estrés, la diferenciación neuronal.

4.1.6 ASIÁTICA – NEGRA

IDENTIFICACIÓN DE LAS MUESTRAS

Se incluyen 30 muestras de sujetos de población asiática y 6 muestras de sujetos de población negra.

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

gene <chr>	asiatico1 <dbl>	asiatico2 <dbl>	asiatico3 <dbl>	asiatico4 <dbl>	asiatico5 <dbl>	asiatico6 <dbl>	asiatico7 <dbl>
133K02	154.7218	407.6782	350.1565	385.2785	203.5119	461.1285	489.5939
5T4	621.4165	334.5521	1079.3549	196.9496	141.9599	401.5375	331.2473
A-362G6.1	948.9882	3186.0146	2910.0538	2973.4748	1680.7615	2618.8857	3368.6509
A-C1	0.4216	1.8282	9.3076	0.6631	0.2167	31.1995	9.8967
a1/3GTP	5.9022	155.3931	65.4738	259.9469	11.4868	304.1951	165.6236
A1BG-AS1	9.2580	25.8090	16.5321	30.1127	43.3747	17.6870	34.7315
A1CF	16.4418	806.6728	479.4993	69.6286	515.6069	6.8639	2.3286
A2M	1771.9098	45917.6920	19468.1088	26637.8846	8582.7073	22024.4011	33186.4881
A2M-AS1	11.3954	35.6901	23.5064	29.2241	28.7301	282.9264	181.2895
A2ML1	1.6863	0.4570	1.2838	0.0000	0.8669	4.3679	0.8732

Figura 64. *Datos crudos de las 2 poblaciones a analizar (asiática y negra).*

De los datos crudos obtenidos a partir de la *count table*, destaca que se ha evaluado la expresión de un total de 20.593 genes en 36 muestras (30 muestras de sujetos de población asiática y 6 muestras de sujetos de población negra).

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 2 grupos de comparación (asiático vs. negro), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

BARPLOT

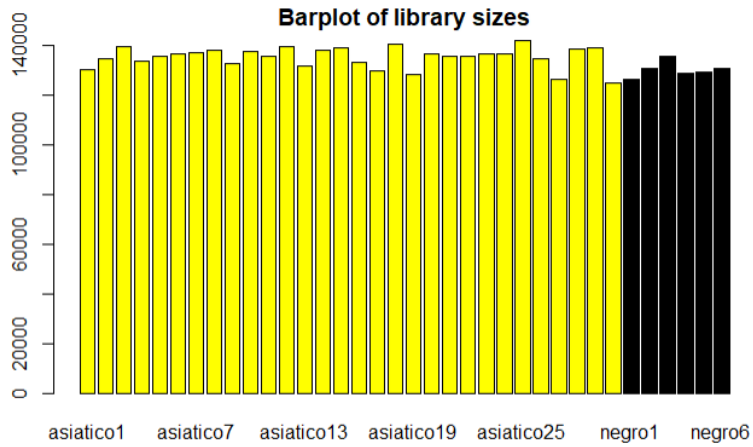


Figura 65. *Barplot de las 2 poblaciones. Población asiática de color amarillo y población negra de color negro.*

Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aproximándose a los 14.000 recuentos para la mayoría las muestras. Algunas muestras, sobre todo de la población negra se aproximan a 13.000 recuentos.

BOXPLOT

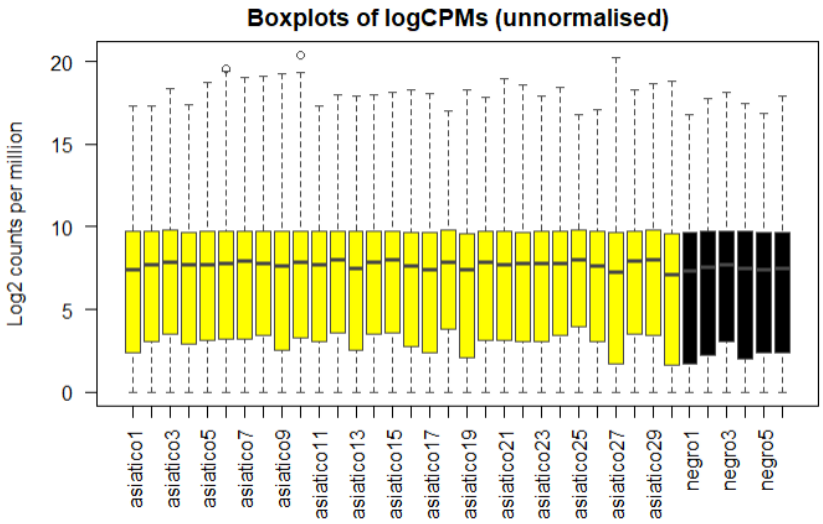


Figura 66. *Boxplot de las 2 poblaciones. Población asiática de color amarillo y población negra de color negro.*

Se observa como la distribución de los recuentos (*pseudoCounts*) es razonablemente homogénea entre les 36 muestras, con algún valor atípico ("asiatico6", "asiatico10").

APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

MULTIDIMENSIONAL SCALING PLOT

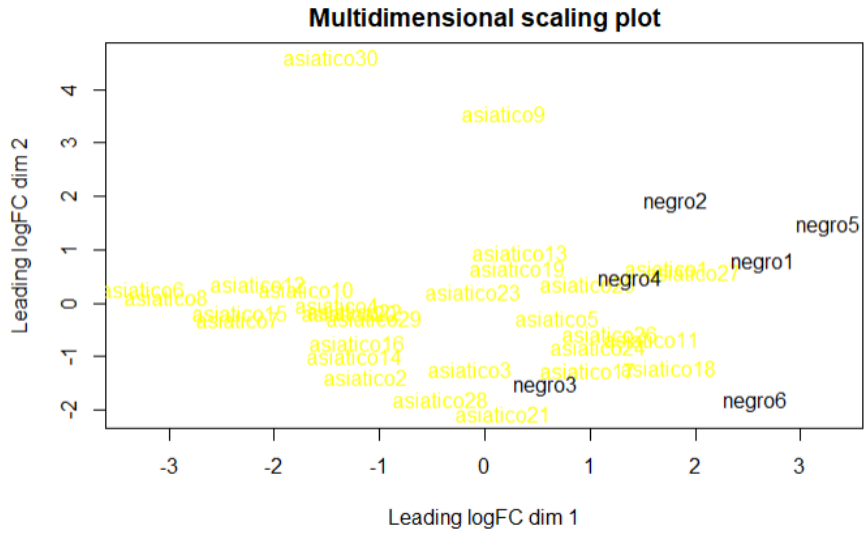


Figura 67. *MDS Plot de las 2 poblaciones (asiática y negra).*

En este caso se observa como la población negra se distribuye a la parte inferior derecha, mientras que la población asiática se distribuye por gran parte de la zona inferior, con cierto solapamiento de las muestras "negro3" y "negro4". Las muestras "asiatico9" y "asiatico30" se encuentran alejadas de la distribución de la población, pudiendo ser unos posibles *outlier* (se ha observado también la muestra "asiatico30" como posible *outlier* en la comparación "latino-asiatico").

CLUSTERING I *HEATMAP*

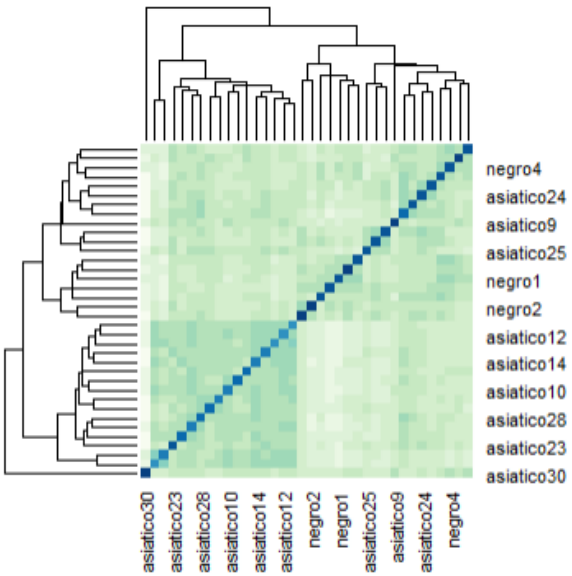


Figura 68. *Clustering i heatmap de las 2 poblaciones (asiática y negra).*

Se observa agrupación de las muestras de la población asiática con solapamiento de alguna muestra de la población negra. Como ya se ha explicado, la poca cantidad de muestras de la población negra puede ser el origen de este solapamiento y la falta de agrupación de la población negra.

FILTRADO

Dado que en este caso existe la limitación que la población negra solamente presenta 5 muestras, se aplica un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 5 muestras. Se observa que después de aplicar el filtrado especificado, se seleccionan 16044 genes, que serán los que se analizarán.

NORMALIZACIÓN

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
asiatico1	asiatico	18630652	1.0640806
asiatico2	asiatico	17378685	1.0920221
asiatico3	asiatico	19457064	1.0873860
asiatico4	asiatico	17098223	1.1071650
asiatico5	asiatico	18517188	1.1238638
asiatico6	asiatico	25636305	0.8013481
asiatico7	asiatico	20306512	1.0218869
asiatico8	asiatico	23989367	0.8823488
asiatico9	asiatico	20495522	0.9484006
asiatico10	asiatico	22753788	0.8989565

1-10 of 36 rows Previous 1 2 3 4 Next

Figura 69. *Datos normalizados de las 2 poblaciones (asiática y negra).*

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDSPLOT

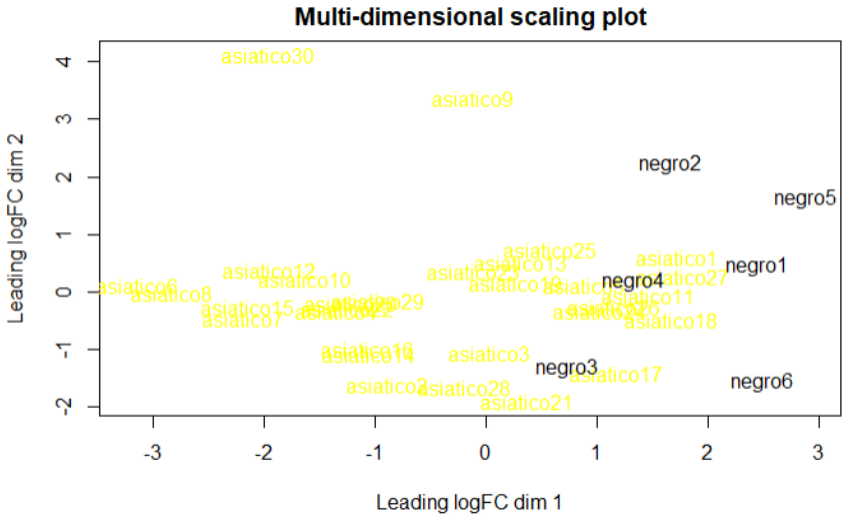
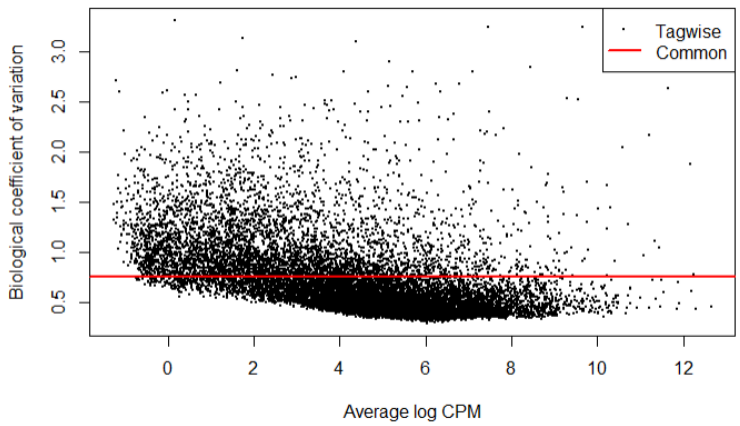


Figura 70. *MDS Plot con los datos normalizados de las 2 poblaciones (asiática y negra).*

Se observa que las posiciones han variado muy poco respecto el MDS Plot de datos no normalizados. La población negra se distribuye a la parte derecha (sobretodo en la zona centro e inferior), mientras que la población asiática se distribuye por gran parte de la zona inferior, con cierto solapamiento de las muestras "negro3" y "negro4". Las muestras "asiatico9" y "asiatico30" se encuentran alejadas de la distribución de la población, pudiendo ser unos posibles *outlier* (se ha observado también la muestra "asiatico30" como posible *outlier* en la comparación "latino-asiático").

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS



Dispersión comuna: 64%
Dispersión específica: 64%

Figura 71. *Estimación de la dispersión de los datos de las 2 poblaciones (asiática y negra).*

Se observa el mapa de dispersión de la media del log de cpm. La mayoría de genes están entorno a 0.8 aunque hay una dispersión de 64% con el coeficiente de variación biológica de algún gen superior a 3.

IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

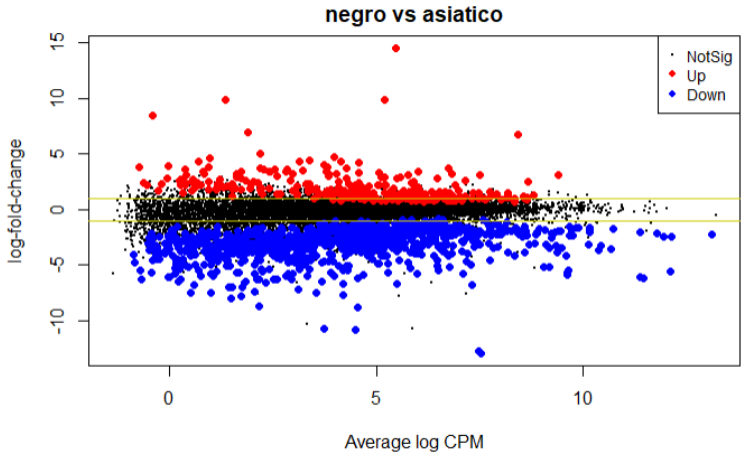
	genes <chr>	logFC <dbl>	logCPM <dbl>	PValue <dbl>	FDR <dbl>
6882	HBM	14.563016	5.4698418	1.437380e-234	2.306133e-230
16263	SERP1	-12.919150	7.5291799	2.931942e-94	2.352004e-90
2577	CDC23	3.716193	6.7480836	2.910522e-74	1.556547e-70
16262	SERP1	9.836563	5.1931239	1.742312e-57	6.988413e-54
1780	C1orf84	4.748371	3.9735713	1.112615e-51	3.570159e-48
6881	HBM	-12.731979	7.4630107	5.084705e-48	1.359650e-44
2122	CACNG5	9.879964	1.3444890	9.605983e-43	2.194010e-39
6640	GPX6	8.472576	-0.3944378	1.093996e-42	2.194010e-39
1781	C1orf84	-4.530036	5.5459541	1.309085e-34	2.333662e-31
16397	SH3D20	4.243086	4.5711462	1.772964e-29	2.844544e-26

1-10 of 10 rows

Figura 72. *10 genes diferencialmente expresados entre las 2 poblaciones (asiática y negra).*

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

A continuación se estudia cómo se expresan en las 2 poblaciones:



	negro-asiatico
Down	890
NotSig	14878
Up	276

Figura 73. *Expresión de los genes diferencialmente expresados en las 2 poblaciones (asiática y negra).*

Se observa que 890 de los genes se encuentran *down-regulated* (expresión disminuida, color azul) y 276 de los genes se encuentran *up-regulated* (expresión aumentada, color rojo) en la comparación génica entre las 2 poblaciones. El resto de 14878 genes no muestran una expresión diferencial significativa.

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

Observamos las 20 anotaciones GO enriquecidas:

	Term	Ont	N	Up	Do...	PUp
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0030032	lamellipodium assembly	BP	31	5	1	0.0003855968
GO:0033169	histone H3-K9 demethylation	BP	2	2	0	0.0004232285
GO:0022607	cellular component assembly	BP	1334	44	70	0.0006070820
GO:0031589	cell-substrate adhesion	BP	194	12	9	0.0006371808
GO:0032269	negative regulation of cellular protein metabol...	BP	482	21	23	0.0008367066
GO:0097581	lamellipodium organization	BP	39	5	3	0.0011444147
GO:1902083	negative regulation of peptidyl-cysteine S-nitr...	BP	3	2	0	0.0012524476
GO:0061408	positive regulation of transcription from RNA p...	BP	3	2	0	0.0012524476
GO:0044085	cellular component biogenesis	BP	1395	44	71	0.0015807731
GO:0043933	protein-containing complex subunit organization	BP	1046	35	54	0.0020557049
GO:0051248	negative regulation of protein metabolic proces...	BP	518	21	25	0.0020694615
GO:0065003	protein-containing complex assembly	BP	929	32	47	0.0020969609
GO:0002091	negative regulation of receptor internalization	BP	4	2	0	0.0024709423
GO:0071425	hematopoietic stem cell proliferation	BP	14	3	0	0.0026538362
GO:0033137	negative regulation of peptidyl-serine phosphor...	BP	14	3	0	0.0026538362
GO:0016043	cellular component organization	BP	2630	71	133	0.0028153208
GO:0010811	positive regulation of cell-substrate adhesion	BP	71	6	4	0.0033184017
GO:0035904	aorta development	BP	32	4	2	0.0039972682
GO:0016577	histone demethylation	BP	5	2	0	0.0040625065
GO:0070076	histone lysine demethylation	BP	5	2	0	0.0040625065

Figura 74. *Gene Enrichment Analysis entre las 2 poblaciones (asiática y negra).*

Se observa que los principales genes diferencialmente expresados entre las muestras de población blanca y negra son proteínas implicadas en la desmetilación de las histonas.

4.2 RESULTADOS DEL ANÁLISIS DESCRIPTIVO DE LAS MUESTRAS

4.2.1 VARIABLES BASALES GENERALES

Se estudian un total de 71 muestras de 4 poblaciones distintas: 5 latinos, 30 blancos, 30 asiáticos y 6 negros.

De las 71 muestras, 27 de los sujetos eran mujeres (38.03%) y 44 eran hombres (62.97%).

La edad media al diagnóstico fue de 64.29 años (SD +/- 13.34 años).

4.2.2 VARIABLES BASALES ENTRE POBLACIONES

En el estudio de las características basales entre poblaciones no existen evidencias estadísticas que apoyen una diferencia en la distribución del sexo entre poblaciones distintas (p- valor 0.6207). Tampoco existen evidencias estadísticas que apoyen una diferencia en la edad al diagnóstico entre poblaciones distintas (p- valor 0.0627).

4.2.3 VARIABLES BIOLÓGICAS - PRONÓSTICAS GENERALES

El tipo histológico de cáncer gástrico más frecuente con 42 casos (50.15%) fue el adenocarcinoma gástrico clásico, seguido del adenocarcinoma gástrico difuso con 11 casos (15.49%), 8 casos (11.26%) de adenocarcinoma mucinoso y de adenocarcinoma tubular, y tan sólo hubo 1 caso (1.40%) de adenocarcinoma papilar y 1 de carcinoma "ring cell" (1.40%).

Se confirma que en 25 casos (35%) no se identificó esófago de Barret como lesión pre-maligna, pero no hay datos en las muestras restantes.

La mediana del número de mutaciones identificadas en las muestras es de 114, con un rango intercuartílico de 65-583.

La mayoría de los cánceres fueron diagnosticados en un estadio avanzado T3 (21 casos, 29.58%) y T4a (18 casos, 25%), aunque la mayoría aún no presentaban metástasis en el momento del diagnóstico (58 casos, 81.7%).

Tan sólo se ha reportado que un 8.45% de los casos necesitaron terapia adyuvante postquirúrgica, mientras que un 18% no, pero tan sólo se dispone de datos de 19 de las muestras (26%).

4.2.4 VARIABLES BIOLÓGICAS - PRONÓSTICAS ENTRE POBLACIONES

TIPO HISTOLÓGICO

Se observan diferencias estadísticamente significativas en la distribución del tipo histológico de cáncer gástrico entre las poblaciones estudiadas, siendo la proporción:

	asiatico	blanco	latino	negro
Diffuse Type Stomach Adenocarcinoma	0.08450704	0.04225352	0.02816901	0.00000000
Mucinous Stomach Adenocarcinoma	0.09859155	0.00000000	0.01408451	0.00000000
Papillary Stomach Adenocarcinoma	0.01408451	0.00000000	0.00000000	0.00000000
Signet Ring Cell Carcinoma of the Stomach	0.01408451	0.00000000	0.00000000	0.00000000
Stomach Adenocarcinoma	0.15492958	0.36619718	0.01408451	0.05633803
Tubular Stomach Adenocarcinoma	0.05633803	0.01408451	0.01408451	0.02816901

Figura 75. *Proporción de los diferentes tipos histológicos de cáncer gástrico en las 4 poblaciones estudiadas.*

COMPARACIONES ENTRE POBLACIONES 2 A 2:

POBLACIONES	P-VALOR
Latina-blanca	0.007076
Latina-asiática	0.8005
Latina-negra	0.7056
Blanca-asiática	0.0003235
Blanca-negra	0.1295
Asiática-negra	0.4357

Tabla 2. *Comparaciones 2 a 2 entre poblaciones analizando el tipo histológico de cáncer gástrico.*

Sí existen diferencias significativas entre las poblaciones latina-blanca y entre blanca-asiática (p-valor<0.05). Por el contrario no existen diferencias significativas entre las poblaciones latina-asiática, latina-negra, blanca-negra y asiática-negra (p-valor>0.05).

NÚMERO DE MUTACIONES

No existen evidencias estadísticas que apoyen la presencia de diferencias significativas en el número de mutaciones de las muestras entre las distinas poblaciones (p-valor 0.333).

ESTADIO DIAGNÓSTICO

Existen diferencias estadísticamente significativas entre el estadio al diagnóstico y las poblaciones, siendo la proporción:

	asiatico	blanco	latino	negro
	0.00000000	0.00000000	0.00000000	0.00000000
T1	0.00000000	0.01408451	0.00000000	0.00000000
T1a	0.00000000	0.00000000	0.00000000	0.00000000
T1b	0.00000000	0.00000000	0.00000000	0.01408451
T2	0.02816901	0.02816901	0.00000000	0.02816901
T2a	0.00000000	0.05633803	0.00000000	0.00000000
T2b	0.00000000	0.01408451	0.01408451	0.00000000
T3	0.15492958	0.08450704	0.02816901	0.02816901
T4	0.02816901	0.05633803	0.01408451	0.00000000
T4a	0.18309859	0.04225352	0.01408451	0.01408451
T4b	0.02816901	0.00000000	0.00000000	0.00000000
TX	0.00000000	0.12676056	0.00000000	0.00000000

Figura 76. *Proporción de muestras según estadio de cáncer gástrico al diagnóstico en las 4 poblaciones estudiadas.*

COMPARACIONES ENTRE POBLACIONES 2 A 2:

POBLACIONES	P-VALOR
Latina-blanca	0.4121
Latina-asiática	0.2373
Latina-negra	0.7056
Blanca-asiática	8.811e-05
Blanca-negra	0.109
Asiática-negra	0.1276

Tabla 3. *Comparaciones 2 a 2 entre poblaciones analizando el estadio diagnóstico.*

Si se han encontrado diferencias estadísticamente significativas entre las poblaciones blanca-asiática (p valor < 0.05).
No existen diferencias entre el resto de poblaciones: latina-blanca, latina-asiática, latina-negra, blanca-negra y asiática-negra.

METÁSTASIS AL DIAGNÓSTICO

Existen diferencias estadísticamente significativas en la presentación de metástasis al diagnóstico y las poblaciones, siendo la proporción:

	asiatico	blanco	latino	negro
	0.00000000	0.00000000	0.00000000	0.00000000
M0	0.38028169	0.38028169	0.01408451	0.04225352
M1	0.04225352	0.00000000	0.02816901	0.01408451
MX	0.00000000	0.04225352	0.02816901	0.02816901

Figura 77. *Proporción de metástasis al diagnóstico en las 4 poblaciones estudiadas.*

COMPARACIONES ENTRE POBLACIONES 2 A 2:

POBLACIONES	P-VALOR
Latina-blanca	0.001358
Latina-asiática	0.001358
Latina-negra	0.7662
Blanca-asiática	0.05665
Blanca-negra	0.04513
Asiática-negra	0.02012

Tabla 4. *Comparaciones 2 a 2 entre poblaciones analizando metástasis al diagnóstico.*

Existen diferencias significativas entre las poblaciones latina-blanca, latina-asiática, blanca-negra y asiática-negra.
No existen diferencias significativas entre las poblaciones latina-negra y blanca-asiática.

ESÓFAGO DE BARRET

No se analiza la enfermedad de esófago de Barret por no disponer de datos suficientes (25 sujetos no presentaban esófago de Barret, no habiéndose estudiado en el resto).

TERAPIA ADYUVANTE POSTQUIRÚRGICA

Tampoco se analiza si los sujetos han recibido rerapia adyuvante postquirúrgica. En este caso se dispone de 13 sujetos que no la recibieron y de 6 que sí la recibieron. Del resto de sujetos (52 sujetos, 73% del total) no se ha recogido este dato.

5. DISCUSIÓN

5.1 VARIABLES CLÍNICAS Y PRONÓSTICAS

En primer lugar cabe destacar que se han observado diferencias estadísticamente significativas en la distribución del tipo histológico de cáncer gástrico entre las poblaciones latina-blanca y también entre las poblaciones blanca-asiática. Se han descrito 6 tipos histológicos diferentes en las 4 poblaciones: adenocarcinoma gástrico clásico (el más frecuente), adenocarcinoma gástrico difuso, adenocarcinoma mucinoso, adenocarcinoma tubular, adenocarcinoma papilar y carcinoma "ring cell". Según la literatura, la clasificación de los carcinomas gástricos parece reflejar diferencias biológicas importantes y posibles diferencias en el tipo celular que los originan, ya que los dos subtipos principales (intestinal y difuso), no se transforman en el otro y presentan una epidemiología diferente [4,16]. Los diferentes tipos de cáncer gástrico se localizan típicamente en una de las 3 capas mucosas. No sólo la célula madre, sino también la célula enterocromafin-like (ECL) puede proliferar y dar lugar a tumores. La célula madre probablemente da lugar al tipo intestinal, mientras que la célula ECL puede ser importante en el tipo difuso. Por tanto, es fundamental determinar el papel de la célula diana gastrina, la célula ECL, en la carcinogénesis gástrica, precisando más estudios al respecto [4]. No tan solo es importante y diferente entre poblaciones el tipo histológicos, sino la localización del tumor [7,16]. En este sentido, el cáncer gástrico a nivel de cardias suele ser de tipo adenocarcinoma difuso, y es el doble de frecuente en población blanca que en otras poblaciones, mientras que el cáncer no cardias lo es la mitad [7,16]. Contrariamente, el cáncer gástrico distal no cardias suele ser del tipo histológico intestinal, y es más frecuente en población negra [7]. La asociación entre la raza con la incidencia de los diferentes tipos de cáncer gástrico en las distintas poblaciones parece mediada más por factores de presión selectiva medioambientales que por variaciones genéticas [16]. En este sentido, cabe remarcar que Japón es uno de los países con mayor incidencia de cáncer gástrico a nivel mundial. Los migrantes japoneses a Estados Unidos, mantienen una elevada incidencia en su primera generación de descendientes, pero la tasa se asimila a los de los descendentes americanos tras dos generaciones [16]. Otro punto que pone en relevancia los factores ambientales en las diferencias poblaciones del tipo y localización del cáncer gástrico es que los factores de riesgo descritos presentan algunas diferencias y es que, a parte de los factores comunes (edad, sexo masculino, tabaquismo, raza, antecedentes familiares, sedentarismo, ingesta de fibra dietética, radiación), el cáncer no cardias se asocia más a la infección por H. pylori, un menor estatus socioeconómico, y factores dietéticos como el mayor consumo de sal y comida ahumada, y el menor consumo de frutas y vegetales [16]. Estas diferencias podrían explicar, al menos en parte, que el tipo histológico y la localización de los tumores gástricos sean significativamente distintos en población blanca comparativamente con el resto de poblaciones [16].

En este trabajo también se han encontrado diferencias estadísticamente significativas en el estadio en el momento del diagnóstico, sobre todo entre las poblaciones blanca y asiática, así como diferencias en la presencia de metástasis (menor presencia en la población asiática). Está descrito un alto porcentaje de enfermedad en estadio I en los asiáticos mientras que los

sujetos blancos son diagnosticados en estadio más avanzado (III) (8). Esta diferencia de estadio de la enfermedad en el momento del diagnóstico se puede relacionar con el programa nacional de vigilancia endoscópica para el cribado anual de cáncer gástrico con endoscopia anual para personas mayores de 40 años en Japón (población asiática) [7]. Con este cribado masivo aproximadamente la mitad de los tumores gástricos se detectan en una etapa temprana en individuos asintomáticos y la tasa de mortalidad por este cáncer se ha reducido a más de la mitad des de principios de la década de 1970. En China también se está llevando a cabo un estudio de intervención de cáncer gástrico que implica un enfoque integral para la prevención del cáncer gástrico incluida la erradicación de H.pylori, suplementos nutricionales y cribado masivo con radiografía de doble contraste y examen endoscópico [7].

Además, se han descrito diferencias de tratamiento en función de la población y la región geográfica. Por ejemplo los afroamericanos con adenocarcinoma gástrico resecable tienen más probabilidades de recibir una recomendación en contra de la cirugía que los individuos de otras poblaciones, presentando una supervivencia menor frente a otras poblaciones [17]. Esta variable no se ha podido analizar en este trabajo dado que nos faltan datos de tratamiento y supervivencia.

5.2 EXPRESIÓN DIFERENCIAL DE GENES E IMPLICACIÓN BIOLÓGICA

En el estudio de expresión diferencial de genes entre poblaciones se han descrito diversas posibles vías fisiopatológicas que podrían estar implicadas en el desarrollo del adenocarcinoma gástrico, viéndose influencias por posibles efectos externos medioambientales y dietéticos.

En primer lugar, en algunas de las comparaciones entre poblaciones (latina-blanca) se ha detectado una expresión diferencial de genes implicados en el metabolismo hematopoetico (GO:0043353). Está bien establecido que los sujetos con grupo sanguíneo A tienen un mayor riesgo de sufrir cáncer gástrico respecto el resto del grupos sanguíneos, aunque no parece existir una asociación en el pronóstico [18]. Sin embargo, debe tenerse en cuenta que la progresión del cáncer gástrico y la supervivencia relacionada con el grupo sanguíneo ABO también pueden estar asociados con otros factores, como la infección por H.pylori [18]. Los carbohidratos del grupo histo-sanguíneo ABO presentan una gran diversidad estructural que influye en la susceptibilidad humana a la infección por H. pylori [19], y también están bien establecido que determinados aspectos nutricionales tienen una influencia directa en la susceptibilidad a la infección por H. pylori [20].

Uno de los hallazgos interesantes es la expresión diferencial de genes implicados en el metabolismo de las hormonas tiroideas (GO:0006590, GO:0042403) entre las poblaciones estudiadas (en el caso de la comparación latina-blanca). En base a la literatura previa no existen evidencias de una posible relación entre las alteraciones del metabolismo de las hormones tiroideas y el cáncer gástrico. Sí que se ha establecido una relación con el adenocarcinoma gástrico y la obesidad [21], y se conoce que la obesidad se relaciona con alteraciones en el eje tiroideo. Por tanto, se considera este hallazgo un ámbito interesante

para futuros estudios. Evidentemente, los factores dietéticos y ambientales son uno de los puntos clave en el desarrollo de obesidad, por lo que en este caso también existiría un posible nexo de unión de los factores ambientales poblacionales y la expresión diferencial de genes.

En la comparación de las poblaciones latina-asiática y asiática-negra destaca la expresión diferencial de genes implicados en las vías de acetilación (GO:0016573, GO:0018393, GO:0018394, GO: 0031056) y metilación (GO:0033169, GO:0016577, GO:0070076). Este proceso es importante en relación a la epigenética. Varias líneas de estudio han corroborado la influencia de las variaciones epigenéticas dependientes de la dieta en modificaciones de las histonas en diferentes tipos de cáncer incluido el cáncer gástrico [22,23]. Nutrientes y compuestos bioactivos como el colecalciferol, cucrumina, resveratrol, qeurcetina, garcinol y butirato de sodio pueden regular las actividades HAT (histona acetiltransferasas) y HDAC (histona deacetilasas) y por lo tanto imponer firmas características en el epigenoma. Estos compuestos pueden favorecer la activación o represión de genes mediante la adición o eliminación de un grupo acetilo de los residuos de aminoácidos de las colas de histonas. Se han descrito avances en nuevas tecnologías "ómicas" que han ayudado a dilucidar el impacto y los efectos de la dieta en el genoma y el epigenoma para obtener resultados más prometedores para la prevención y el tratamiento del cáncer [22,23].

Otro hallazgo importante se encuentra en la comparación entre poblaciones latina-negra, donde destaca la expresión diferencial de proteínas implicadas en el transporte de moléculas (sodio, biotina, vitamina B_5 (*pantothenate*)) (GO:0015878, GO:0015939). Las vitaminas B (proporcionadas en la dieta a partir de alta cantidad de alimentos como cereales integrales, arros, nueces, leche, huevos, carne, pescado, frutas i verduras de hoja verde), mantienen y aumentan la tasa metabólica y promueven el crecimiento y la división celular. Funcionan como coenzimas en la síntesis de purinas y timidilato para la síntesis de DNA por lo que cuando sus niveles son insuficientes existe mayor incorporación de uracilo en el DNA con la consiguiente rotura de cromosomas, alteración de la reparación del DNA y transformación neoplásica. Se ha descrito asociación de forma signnificativa entre concentraciones séricas de vitaminas B y metabolitos relacionados y riesgo de adenocarcinoma gástrico, por lo que sugiere que las vitaminas B tienen un papel importante como agentes quimiopreventivos para el cáncer gástrico, pero que se debe investigar más a fondo su papel potencial [24].

En la comparación blanca-asiática destaca la expresión diferencial de genes implicados en la respuesta inmune (GO:2000514, GO:2000482, GO:0001819). En la literatura se encuentra la asociación entre los marcadores autoinmunes con la supervivencia en el cáncer gástrico. Los linfocitos infiltrantes son la principal célula inmune infiltrante, por lo que su densidad se considera una manifestación de la respuesta del sistema inmunitario del huésped contra las células tumorales (igual que sucede en otros tipos de cáncer). Se conoce de la evidencia de que las altas densidades de limfocitos infiltrantes intratumorales (CD8), son indicativas de mejor supervivencia [25]. Las alteraciones en la respuesta inmune de pacientes con enfermedades autoinmunes pueden predisponer a neoplasias malignas, y en muchos estudios se ha informado un vínculo entre la gastritis autoinmune crónica y el cáncer gástrico. La metaplasia

intestinal con displasia de la mucosa del cuerpo-fundus gástrico y la hiperplasia de las células cromafines, que son características típicas de la gastritis autoinmune en etapa tardía, se consideran lesiones precursoras [26]. La anemia perniciosa es la etapa terminal de la gastritis atrófica autoinmunitaria crónica en la que los anticuerpos contra las células parietales gástricas inhiben la secreción de factor intrínseco, que eventualmente conduce a anemia macrocítica (debido a la disminución de la absorción de vitamina B12). La absorción de vitamina B12 requiere una mucosa gástrica productora de ácido para permitir la escisión de la vitamina B12 de sus proteínas de unión en el estómago. Se ha descrito asociación entre niveles bajos de vitamina B12 y mayor riesgo de adenocarcinoma gástrico. La gastritis crónica atrófica disminuye la secreción de ácido gástrico, lo que disminuye la absorción de vitamina B12, por lo que representa otro vínculo potencial entre la malabsorción de vitamina B12 y un mayor riesgo de adenocarcinoma gástrico. Se deben realizar estudios adicionales donde se explora la utilidad de la vitamina B12 como marcador serológico de atrofia gástrica [26,27]. También la infección por H. pylori también se asocia con malabsorción de vitamina B12 probablemente debido a la inducción de gastritis atrófica y aclorhídia acompañante [27].

Destaca también en la comparación de las poblaciones blanca-negra la expresión diferencial de genes implicados en la regulación de diferentes procesos como el estrés oxidativo (GO:0034599, GO:1902882, GO:1903202, GO:0000305). El estrés oxidativo es el grado en que existe un desequilibrio entre la producción y la eliminación de metabolitos reactivos (por ejemplo ROS) que da como resultado daños en la estructura y función de las células, aumento de la proliferación celular, daño en el DNA e inestabilidad genómica, por lo que es un proceso biológico fundamental para el inicio y la progresión del cáncer y la respuesta al tratamiento, independientemente de los antecedentes poblacionales [28]. Es importante destacar que el desequilibrio entre la generación/eliminación/reparación de ROS conduce a un estado de inflamación crónica que también contribuye al desarrollo y progresión del tumor [28,29]. Tanto el estrés oxidativo como el funcionamiento inmunológico están influenciados por factores del huésped que incluyen comportamientos dietéticos, actividad física y variables psicosociales. Se ha descrito mayores niveles de estrés oxidativo en la población negra en comparación con la población blanca. También se ha demostrado que la población negra tiene una mayor exposición a los estresores psicosociales crónicos en comparación con la población blanca, como por ejemplo la discriminación racial, que se ha asociado específicamente con mayor estrés oxidativo de los glóbulos rojos [28].

5.3 LIMITACIONES

Este estudio tiene limitaciones relevantes.

En primer lugar cabe destacar que el número de muestras incluidas es bajo, con algunas poblaciones representadas por tan solo 5 (población latina) o 6 muestras (población negra), por lo que se tendrían que reproducir los cálculos en cohortes de mayor tamaño para

disminuir el efecto de la variabilidad interindividual y ganar potencia estadística para detectar diferencias poblacionales.

Otra limitación importante del estudio es el bajo número de variables fenotípicas disponibles en la base de datos utilizada, que dificulta la descripción clínica de la muestras, la descripción de las características biológicas y pronósticas de los tumores, y la correlación de los hallazgos en la expresión diferencial de genes con factores de presión selectiva dietéticos y medioambientales.

6. CONCLUSIONES

Se confirman diferencias significativas en el tipo histológico de cáncer gástrico, el estadio en el momento del diagnóstico y la presencia de metástasis. En las comparaciones binarias de expresión diferencial de genes entre población latina, blanca, asiática y negra, se describen posibles vías fisiopatológicas que pueden estar implicadas en el desarrollo del cáncer gástrico y, a la vez, verse influenciadas por factores medioambientales, destacando: el grupo sanguíneo ABO y la infección por H. pylori con factores dietéticos; el metabolismo de las hormonas tiroideas con la obesidad; y los niveles de vitaminas del complejo B, el estrés oxidativo, el estado inmunológico y los procesos de modificación epigenética con aspectos dietéticos.

Por lo que refiere al alcance de los objetivos del proyecto, se considera que se han resuelto de forma parcial, ya que se ha podido investigar la expresión diferencial de genes de pacientes con cáncer gástrico de diferentes poblaciones (latina vs. blanca vs. asiática vs. negra), pero ha faltado poder profundizar en el estudio de diferentes características clínicas y biológicas de los tumores, para detectar posibles factores pronósticos, por falta de datos disponibles. A pesar de ello, se ha intentado relacionar los resultados obtenidos en la expresión diferencial de genes con posibles vías fisiopatológicas del cáncer gástrico que pueden verse sometidas al efecto de diferentes factores de presión selectiva medioambientales y dietéticos.

Por lo anteriormente expuesto, se considera interesante como futuras líneas de investigación ampliar el número de muestras analizadas para el estudio comparativo de expresión diferencial de genes en población con cáncer gástrico, minimizando la variabilidad interindividual, así como que se acompañe de mayor cantidad de variables fenotípicas para poder correlacionar de forma más completa el binomio fenotipo-genotipo. También sería relevante estudiar las vías fisiopatológicas propuestas en este proyecto como precursoras del cáncer gástrico, a nivel molecular, para profundizar en el mecanismo carcinogénico de este tipo de tumor y así poder implementar nuevas medidas preventivas y/o terapéuticas.

7. GLOSARIO

Barplot: también conocido como diagrama de barras. Los niveles de factor se colocan en el eje x y las frecuencias (o proporciones) de varios niveles de factor se colocan en el eje y.

Boxplot: también conocido como diagrama de caja y bigote, *box plot*, *box-plot* o *boxplot*. Es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, el diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos, pudiendo también representar los valores atípicos de estos. Conviene recordar que se utilizan las bisagras de Tukey, y no los cuartiles a la hora de dibujar la caja del gráfico, aunque los resultados son semejantes en muestras grandes.

GO (*Gene Ontology*): es una iniciativa bioinformática con el objetivo de estandarizar la representación de los genes y los atributos de sus productos génicos de todas las especies: El proyecto proporciona un vocabulario controlado de términos y anotaciones de productos génicos.

Clustering* i *heatmap: técnica que muestra la magnitud de un fenómeno como el color en dos dimensiones. La variación de color puede ser por tono o intensidad., dando pistas visuales obvias al lector sobre cómo el fenómeno se agrupa o varía en el espacio. Hay dos categorías fundamentalmente diferentes de mapas de calor: el mapa de calor del clúster y el mapa de calor espacial.

Cronograma de Gantt: es una herramienta para planificar y programar tareas a lo largo de un período determinado. Desarrollado por Henry Laurence Gantt a inicios del siglo XX, el diagrama se muestra en un gráfico de barras horizontales ordenadas por actividades a realizar en secuencias de tiempo concretas.

***Gene Enrichment Analysis*:** es un método para identificar clases de genes o proteínas que están sobrerrepresentadas en un gran conjunto de genes o proteínas y pueden tener una asociación con fenotipos de enfermedades . El método utiliza enfoques estadísticos para identificar grupos de genes significativamente enriquecidos o empobrecidos. Las tecnologías de transcriptómica y los resultados de la proteómica a menudo identifican miles de genes que se utilizan para el análisis.

Multidimensional scaling plot: conjunto de técnicas de ordenación relacionadas que se utilizan en la visualización de información, en particular para mostrar la información contenida en una matriz de distancia. Es una forma de reducción de dimensionalidad no lineal .

***Outlier* o valor atípico:** observación anormal y extrema en una muestra estadística **o** serie temporal de datos que puede afectar potencialmente a la estimación de los parámetros del mismo.

RStudio: es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Permite un análisis y desarrollo para poder analitzar los datos con R.

8. BIBLIOGRAFÍA

1. Suárez R, Wiesner C, González C, Cortés C, Shinchi A. Antropología Del Cáncer. Investigación Aplicada En Salud Pública. Cancer Epidemiol Biomarkers Prev. 2014;23(5):700-13.

2. Suh YS, Yang HK. Screening and Early Detection of Gastric Cancer: East Versus West. Surg Clin North Am. 2015;95(5):1053–66.

3. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and Prevention. 1826;2(5):101–2.

4. Waldum HL, Fossmark R. Types of gastric carcinomas. Int J Mol Sci. 2018;19(12):1–12.

5. Russo AE, Strong VE. Gastric cancer etiology and management in Asia and the west. Annu Rev Med. 2019;70:353–67.

6. Guggenheim DE, Shah MA. Gastric cancer epidemiology and risk factors. J Surg Oncol. 2013;107(3):230–6.

7. Katherine D Crew AIN, Katherine. Epidemiology of gastric cancer. Gastric Cancer Princ Pract. 2006;12(3):354–62.

8. Luyimbazi D, Nelson RA, Choi AH, Li L, Chao J, Sun V, et al. Estimates of Conditional Survival in Gastric Cancer Reveal a Reduction of Racial Disparities with Long-Term Follow-Up. J Gastrointest Surg. 2015;19(2):251–7.

9. Song M, Camargo MC, Weinstein SJ, Best AF, Albanes D, Rabkin CS. gastric cancer and its precursors in a Western population. 2020;21(5):729–37.

10. Van der Post RS, Vogelaar IP, Carneiro F, Guilford P, Huntsman D, Hoogerbrugge N, et al. Hereditary diffuse gastric cancer: Updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. J Med Genet. 2015;52(6):361–74.

11. Fewings E, Larionov A, Redman J, Goldgraben MA, Scarth J, Richardson S, et al. Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. Lancet Gastroenterol Hepatol. 2018;3(7):489–98.

12. Long ZW, Yu HM, Wang YN, Liu D, Chen YZ, Zhao YX, et al. Association of IL-17 polymorphisms with gastric cancer risk in Asian populations. World J Gastroenterol. 2015;21(18):5707–18.

13. Nguyen DK, Maggard-Gibbons M. Age, poverty, acculturation, and gastric cancer. Surg (United States). 2013;154(3):444–52.

14. Cairns J. The cancer problem. Sci Am INC. 1975;66(6):364–71.

15. Ganfeng Luo, Yanting Zhang, Pi Guo, Li Wang, Yuanwei Huang KL. Global patterns and trends in stomach cancer incidence: Age, period and birth cohort analysis. Int J Cancer. 2017;141(7):1333–4.

16. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: Descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiol Biomarkers Prev. 2014;23(5):700–13.

17. Ulanja MB, Beutler BD, Rishi M, Konam KG, Zell SC, Patterson DR, et al. Influence of race and geographic setting on the management of gastric adenocarcinoma. J Surg Oncol [Internet]. 2019;120(2):270–9. Available from: http://dx.doi.org/10.1002/jso.25503.

18. Franchini M, Liumbruno GM, Lippi G. The prognostic value of ABO blood group in cancer patients. Blood Transfus. 2016;14(5):434–40.

19. Brandão de Mattos CC, de Mattos LC. Histo-blood group carbohydrates as facilitators for infection by Helicobacter pylori. Infect Genet Evol. 2017;53:167–74.

20. Aimasso U, D'Onofrio V, D'Eusebio C, Devecchi A, Pira C, Merlo FD, et al. Helicobacter pylori and nutrition: A bidirectional communication. Minerva Gastroenterol Dietol. 2019;65(2):116–29.

21. Avgerinos KI, Spyrou N, Mantzoros CS, Dalamaga M. Obesity and cancer risk: Emerging biological mechanisms and perspectives. Metabolism. 2019;92:121–35.

22. Calcagno DQ, Wisnieski F, Da Silva Mota ER, Maia De Sousa SB, Costa Da Silva JM, Leal MF, et al. Role of histone acetylation in gastric cancer: Implications of dietetic compounds and clinical perspectives. Epigenomics. 2019;11(3):349–62.

23. Fu DG. Epigenetic alterations in gastric cancer (review). Mol Med Rep. 2015;12(3):3223–30.

24. Ren J, Murphy G, Fan J, Dawsey SM, Taylor PR, Selhub J, et al. Prospective study of serum B vitamins levels and oesophageal and gastric cancers in China. Sci Rep. 2016;6(October):1–8.

25. Yu PC, Long D, Liao CC, Zhang S. Association between density of tumor-infiltrating lymphocytes and prognoses of patients with gastric cancer. Med (United States). 2018;97(27).

26. Bizzaro N, Antico A, Villalta D. Autoimmunity and gastric cancer. Int J Mol Sci. 2018;19(2):1–14.

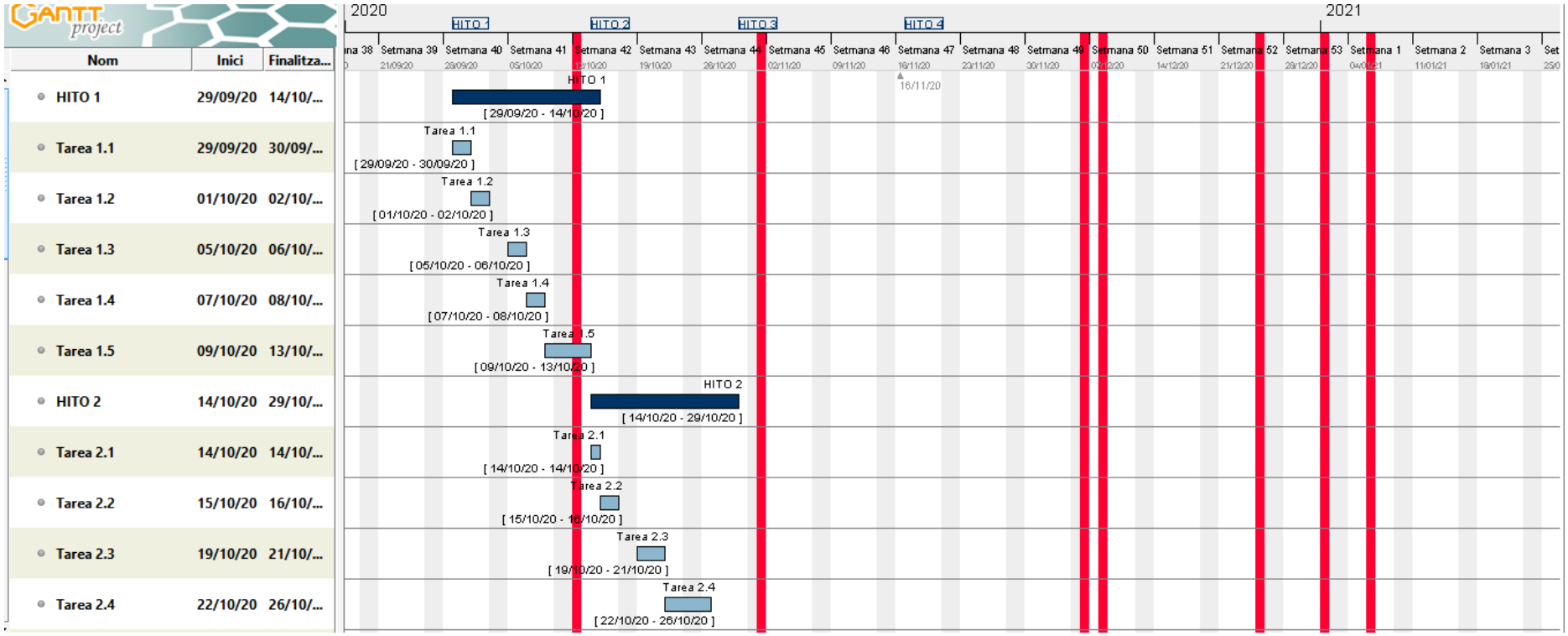
27. Eugenia H. Miranti, Rachael Stolzenberg-Solomon, Stephanie J. Weinstein, Jacob Selhub, Satu Männistö, Philip R. Taylor, Neal D. Freedman, Demetrius Albanes, Christian C. Abne GM. Low Vitamin B12 Increases Risk of Gastric Cancer: A Prospective Study of One-Carbon Metabolism Nutrients and Risk of Upper Gastrointestinal Tract Cancer. Int J Cancer. 2017;141(6):1120–1129.

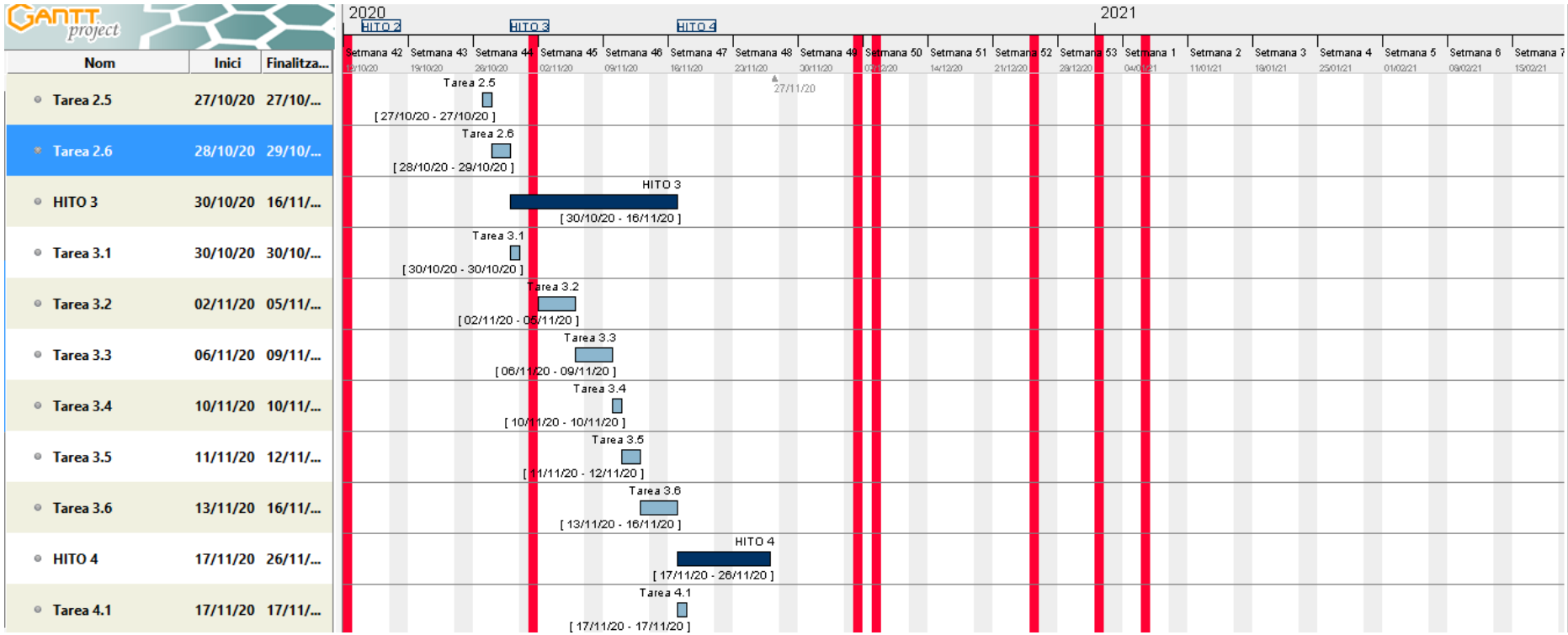
28. Zhang J, Ye Z, Townsend DM, Hughes-halbert C, Kenneth D, Therapeutics E, et al. Racial disparities, cancer and response to oxidative stress. Adv Cancer Res. 2019;144:343–83.

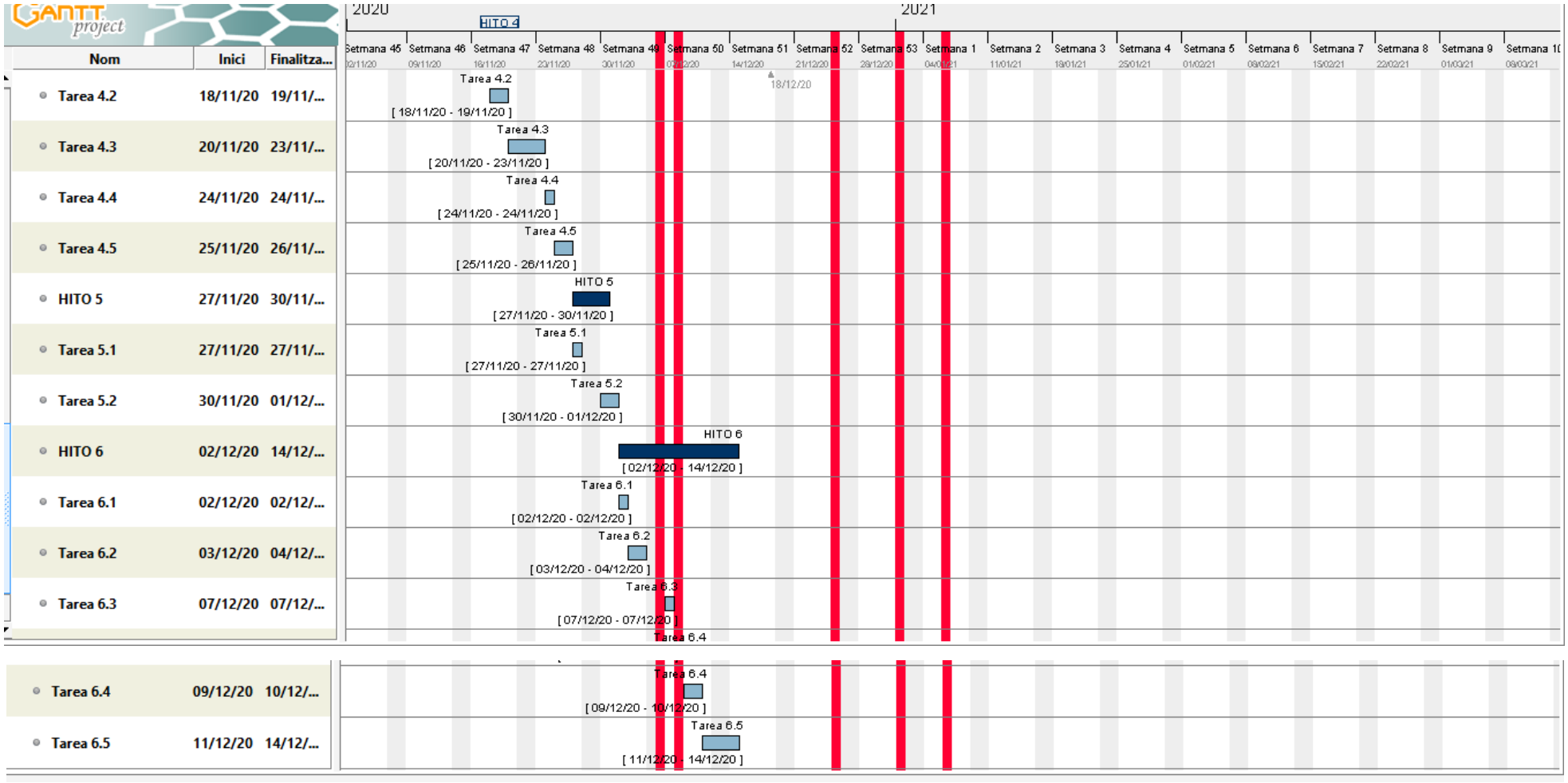
29. Murata M. Inflammation and cancer. Environ Health Prev Med. 2018;23(1):1–8.

9. ANEXOS

9.1 ANEXO 1. CRONOGRAMA DE GANTT




GANTT project
Nom
Inici
Finalitza...
Tarea 2.5 27/10/20 27/10/...
Tarea 2.6 28/10/20 29/10/...
HITO 3 30/10/20 16/11/...
Tarea 3.1 30/10/20 30/10/...
Tarea 3.2 02/11/20 05/11/...
Tarea 3.3 06/11/20 09/11/...
Tarea 3.4 10/11/20 10/11/...
Tarea 3.5 11/11/20 12/11/...
Tarea 3.6 13/11/20 16/11/...
HITO 4 17/11/20 26/11/...
Tarea 4.1 17/11/20 17/11/...
2020
2021
HITO 2
HITO 3
HITO 4
Setmana 42
Setmana 43
Setmana 44
Setmana 45
Setmana 46
Setmana 47
Setmana 48
Setmana 49
Setmana 50
Setmana 51
Setmana 52
Setmana 53
Setmana 1
Setmana 2
Setmana 3
Setmana 4
Setmana 5
Setmana 6
Setmana 7
27/11/20
Tarea 2.5
[27/10/20 - 27/10/20]
Tarea 2.6
[28/10/20 - 29/10/20]
HITO 3
[30/10/20 - 16/11/20]
Tarea 3.1
[30/10/20 - 30/10/20]
Tarea 3.2
[02/11/20 - 05/11/20]
Tarea 3.3
[06/11/20 - 09/11/20]
Tarea 3.4
[10/11/20 - 10/11/20]
Tarea 3.5
[11/11/20 - 12/11/20]
Tarea 3.6
[13/11/20 - 16/11/20]
HITO 4
[17/11/20 - 26/11/20]
Tarea 4.1
[17/11/20 - 17/11/20]


GANTT project
2020
2021
HITO 4
Setmana 45
Setmana 46
Setmana 47
Setmana 48
Setmana 49
Setmana 50
Setmana 51
Setmana 52
Setmana 53
Setmana 1
Setmana 2
Setmana 3
Setmana 4
Setmana 5
Setmana 6
Setmana 7
Setmana 8
Setmana 9
Setmana 10
18/12/20
Nom
Inici
Finalitza...
Tarea 4.2 18/11/20 19/11/...
Tarea 4.3 20/11/20 23/11/...
Tarea 4.4 24/11/20 24/11/...
Tarea 4.5 25/11/20 26/11/...
HITO 5 27/11/20 30/11/...
Tarea 5.1 27/11/20 27/11/...
Tarea 5.2 30/11/20 01/12/...
HITO 6 02/12/20 14/12/...
Tarea 6.1 02/12/20 02/12/...
Tarea 6.2 03/12/20 04/12/...
Tarea 6.3 07/12/20 07/12/...
Tarea 6.4 09/12/20 10/12/...
Tarea 6.5 11/12/20 14/12/...
Tarea 4.2 [18/11/20 - 19/11/20]
Tarea 4.3 [20/11/20 - 23/11/20]
Tarea 4.4 [24/11/20 - 24/11/20]
Tarea 4.5 [25/11/20 - 26/11/20]
HITO 5 [27/11/20 - 30/11/20]
Tarea 5.1 [27/11/20 - 27/11/20]
Tarea 5.2 [30/11/20 - 01/12/20]
HITO 6 [02/12/20 - 14/12/20]
Tarea 6.1 [02/12/20 - 02/12/20]
Tarea 6.2 [03/12/20 - 04/12/20]
Tarea 6.3 [07/12/20 - 07/12/20]
Tarea 6.4
Tarea 6.4 [09/12/20 - 10/12/20]
Tarea 6.5 [11/12/20 - 14/12/20]

9.2 ANEXO 2. EJEMPLO DE CÓDIGO R

9.2.1 ANÁLISIS DE EXPRESIÓN DIFERENCIAL DE GENES

IDENTIFICACIÓN DE LAS MUESTRAS

Se descarga la base de datos **de la plataforma** http://www.cbioportal.org/, que contiene las características clínicas de cada muestra de las diferentes poblaciones

```
taula_clinica_total <- read.delim ("C:/Users/a/Documents/MASTER/4T
TRIMESTRE/stad_tcga_clinical_data.tsv")
taula_clinica_total
```

Patient.ID <fctr>	Sample.ID <fctr>	Cancer.Type.Detailed <fctr>	Mutation.Count <int>
TCGA-3M-AB46	TCGA-3M-AB46-01	Stomach Adenocarcinoma	NA
TCGA-3M-AB47	TCGA-3M-AB47-01	Stomach Adenocarcinoma	NA
TCGA-B7-5816	TCGA-B7-5816-01	Diffuse Type Stomach Adenocarcinoma	1279
TCGA-B7-5818	TCGA-B7-5818-01	Diffuse Type Stomach Adenocarcinoma	326
TCGA-B7-A5TI	TCGA-B7-A5TI-01	Diffuse Type Stomach Adenocarcinoma	583
TCGA-B7-A5TJ	TCGA-B7-A5TJ-01	Stomach Adenocarcinoma	213
TCGA-B7-A5TK	TCGA-B7-A5TK-01	Stomach Adenocarcinoma	51
TCGA-B7-A5TN	TCGA-B7-A5TN-01	Stomach Adenocarcinoma	77
TCGA-BR-4183	TCGA-BR-4183-01	Stomach Adenocarcinoma	37
TCGA-BR-4184	TCGA-BR-4184-01	Stomach Adenocarcinoma	3530

```
summary(taula_clinica_total)
str(taula_clinica_total)
```

```
       Patient.ID            Sample.ID                                  Cancer.Type.Detailed
 TCGA-3M-AB46:  1   TCGA-3M-AB46-01:  1   Diffuse Type Stomach Adenocarcinoma      : 72
 TCGA-3M-AB47:  1   TCGA-3M-AB47-01:  1   Mucinous Stomach Adenocarcinoma          : 22
 TCGA-B7-5816:  1   TCGA-B7-5816-01:  1   Papillary Stomach Adenocarcinoma         :  8
 TCGA-B7-5818:  1   TCGA-B7-5818-01:  1   Signet Ring Cell Carcinoma of the Stomach: 14
 TCGA-B7-A5TI:  1   TCGA-B7-A5TI-01:  1   Stomach Adenocarcinoma                   :283
 TCGA-B7-A5TJ:  1   TCGA-B7-A5TJ-01:  1   Tubular Stomach Adenocarcinoma           : 79
 (Other)     :472   (Other)        :472
 Mutation.Count   Fraction.Genome.Altered Diagnosis.Age       Sex
 Min.    :   3.0   Min.    :0.0000        Min.    :30.00         : 35
 1st Qu.:  70.5   1st Qu.:0.0507          1st Qu.:58.00   Female:158
 Median : 108.0   Median :0.1764          Median :67.00   Male  :285
 Mean   : 355.0   Mean   :0.2290          Mean   :65.68
 3rd Qu.: 211.5   3rd Qu.:0.3690          3rd Qu.:73.00
 Max.   :6359.0   Max.    :0.9150         Max.    :90.00
 NA's   :83       NA's    :37             NA's    :40
             Ethnicity.Category                                    Race.Category
                       :155                                                : 97
 HISPANIC OR LATINO    :  5      ASIAN                                     : 89
 NOT HISPANIC OR LATINO:318      BLACK OR AFRICAN AMERICAN                 : 13
                                 NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER:  1
                                 WHITE                                     :278

 Adjuvant.Postoperative.Targeted.Therapy.Administered.Indicator
    :259
 NO :117
 YES:102
```

Se descarga la tabla que contiene la expresión de los genes en las diferentes muestras.

```
library(readr)
taula_expression_median_V2 <- read_delim("C:/Users/a/Documents/MASTER/4T
TRIMESTRE/stad_tcga(1)/data_RNA_Seq_v2_expression_median.txt", "\t", escape_double =
FALSE, trim_ws = TRUE)
taula_expression_median_V2
```

Hugo_Symbol <chr>	Entrez_Gene_Id <dbl>	TCGA-3M-AB46-01 <dbl>	TCGA-3M-AB47-01 <dbl>	TCGA-B7-5816-01 <dbl>	TCGA-B7-5818-01 <dbl>
LOC100130426	100130426	0.0000	0.0000	0.0000	0.0000
UBE2Q2P3	100133144	11.3383	14.2000	3.6576	4.7020
UBE2Q2P3	100134869	17.1023	7.8408	12.1906	41.1395
LOC149767	10357	40.5136	17.9592	0.4171	6.5488
TIMM23	10431	2163.3818	880.4082	750.2867	748.5265
MOXD2	136542	0.0000	0.0000	0.0000	0.0000
LOC155060	155060	181.5458	145.7143	37.9522	102.8160
RNU12-2P	26823	2.3701	0.8163	0.0000	0.0000
SSX9	280660	0.0000	0.0000	0.0000	0.0000
LOC317712	317712	0.0000	0.0000	0.0000	0.0000

Se seleccionan las muestras que pertenecen a la etnia “Hispanic or latino” y a la raza “White”, y se nombra a esta selección latina. La base de datos sólo cuenta con 5 muestras latina.

```
latino <- subset(taula_clinica_total, taula_clinica_total$Ethnicity.Category ==
"HISPANIC OR LATINO" & taula_clinica_total$Race.Category == "WHITE")
head (latino,5)
```

	Patient.ID <fctr>	Sample.ID <fctr>	Cancer.Type.Detailed <fctr>	Mutation.Count <int>
1	TCGA-3M-AB46	TCGA-3M-AB46-01	Stomach Adenocarcinoma	*NA*
303	TCGA-F1-A448	TCGA-F1-A448-01	Mucinous Stomach Adenocarcinoma	616
308	TCGA-FP-7998	TCGA-FP-7998-01	Diffuse Type Stomach Adenocarcinoma	62
387	TCGA-R5-A7O7	TCGA-R5-A7O7-01	Tubular Stomach Adenocarcinoma	111
390	TCGA-R5-A7ZI	TCGA-R5-A7ZI-01	Diffuse Type Stomach Adenocarcinoma	755

5 rows | 1-5 of 16 columns

Se seleccionan las 5 muestras latina, y se nombra a cada una de ellas “latino + número (del 1 al 5)”.

```
latino1 <- subset(taula_clinica_total, taula_clinica_total$Sample.ID == c("TCGA-3M-
AB46-01"))
latino2 <- subset(taula_clinica_total, taula_clinica_total$Sample.ID == c("TCGA-F1-
A448-01"))
latino3 <- subset(taula_clinica_total, taula_clinica_total$Sample.ID == c("TCGA-FP-
7998-01"))
latino4 <- subset(taula_clinica_total, taula_clinica_total$Sample.ID == c("TCGA-R5-
A7O7-01"))
latino5 <- subset(taula_clinica_total, taula_clinica_total$Sample.ID == c("TCGA-R5-
A7ZI-01"))
```

Se crea una tabla con las 5 muestras latina.

```
clinica_latino<-rbind(latino1, latino2, latino3, latino4, latino5)
clinica_latino
```

	Patient.ID <fctr>	Sample.ID <fctr>	Cancer.Type.Detailed <fctr>	Mutation.Count <int>
1	TCGA-3M-AB46	TCGA-3M-AB46-01	Stomach Adenocarcinoma	*NA*
303	TCGA-F1-A448	TCGA-F1-A448-01	Mucinous Stomach Adenocarcinoma	616
308	TCGA-FP-7998	TCGA-FP-7998-01	Diffuse Type Stomach Adenocarcinoma	62
387	TCGA-R5-A7O7	TCGA-R5-A7O7-01	Tubular Stomach Adenocarcinoma	111
390	TCGA-R5-A7ZI	TCGA-R5-A7ZI-01	Diffuse Type Stomach Adenocarcinoma	755

5 rows | 1-5 of 16 columns

Se seleccionan las muestras latina de la tabla de la expresión genética.

```
nom_gen<- taula_expression_median_V2["Hugo_Symbol"]
latino1_gen <- taula_expression_median_V2["TCGA-3M-AB46-01"]
latino2_gen <- taula_expression_median_V2["TCGA-F1-A448-01"]
latino3_gen <- taula_expression_median_V2["TCGA-FP-7998-01"]
latino4_gen <- taula_expression_median_V2["TCGA-R5-A7O7-01"]
latino5_gen <- taula_expression_median_V2["TCGA-R5-A7ZI-01"]
```

Se crea una tabla con las muestras seleccionadas.

```
gen_latino <- cbind(nom_gen, latino1_gen, latino2_gen, latino3_gen, latino4_gen,
latino5_gen)
gen_latino
```

Hugo_Symbol <chr>	TCGA-3M-AB46-01 <dbl>	TCGA-F1-A448-01 <dbl>	TCGA-FP-7998-01 <dbl>	TCGA-R5-A7O7-01 <dbl>	TCGA-R5-A7ZI-01 <dbl>
LOC100130426	0.0000	0.0000	0.3649	0.0000	0.0000
UBE2Q2P3	11.3383	35.1245	12.0966	20.8702	14.5046
UBE2Q2P3	17.1023	24.3668	11.2181	30.4850	16.3277
LOC149767	40.5136	14.7848	32.0933	57.0613	12.0648
TIMM23	2163.3818	871.9528	501.4361	1526.3909	1853.9557
MOXD2	0.0000	0.0000	0.0000	0.0000	0.0000
LOC155060	181.5458	120.3907	116.5737	99.8573	116.6263
RNU12-2P	2.3701	1.0561	1.0137	0.0000	0.0000
SSX9	0.0000	0.0000	0.0000	0.0000	0.0000
LOC317712	0.0000	0.0000	0.0000	0.0000	0.0000

Se realizan los mismos pasos para seleccionar las muestras:

- “blanco”: “not hispanic or latino” y “white”
- “asiatico”: “not hispanic or latino” y “asian”
- “negro”: “not hispanic or latino” y “black”

Se crea una tabla con las muestras de las 4 poblaciones.

```
agrup1 <- merge(gen_latino, gen_blanco, by="Hugo_Symbol")
agrup2 <- merge(gen_asiatico, gen_negro, by="Hugo_Symbol")
taula_gens_total <- merge(agrup1,agrup2, by="Hugo_Symbol")
taula_gens_total
```

Hugo_Symbol <chr>	TCGA-3M-AB46-01 <dbl>	TCGA-F1-A448-01 <dbl>	TCGA-FP-7998-01 <dbl>	TCGA-R5-A7O7-01 <dbl>	TCGA-R5-A7ZI-01 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041
5T4	933.7998	796.9726	189.8969	489.3010	345.8573
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811

Workflow - Pipeline. Análisis de la expresión génica de datos de ultasecuenciación de RNA

El análisis bioinformático de ultra-secuenciación de RNA se ha realizado siguiendo el procedimiento que se detalla a continuación:

INSTALACIÓN DE LOS PAQUETES BIOCONDUCTOR PARA EL PROCESAMIENTO DE LOS DATOS.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("edgeR")
```

CARGA DE DATOS CRUDOS Y CREACIÓN DEL OBJETO DGELIST

Una vez cargado el archivo que contiene la count table con los read counting de las muestras a procesar al Software R, se recodifican las variables del nombre de cada una de las 71 muestras para obtener una tabla inicial con cada muestra correctamente identificada:

```
raw_targets <- taula_gens_total
raw_targets
```

Hugo_Symbol <chr>	TCGA-3M-AB46-01 <dbl>	TCGA-F1-A448-01 <dbl>	TCGA-FP-7998-01 <dbl>	TCGA-R5-A7O7-01 <dbl>	TCGA-R5-A7ZI-01 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041
5T4	933.7998	796.9726	189.8969	489.3010	345.8573
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811

Se recodifican las variables:

```
colnames(raw_targets)[1] = paste0("gene")
colnames(raw_targets)[2:6] = paste0(latina, 1:5)
colnames(raw_targets)[7:36] = paste0(blanca, 1:30)
colnames(raw_targets)[37:66] = paste0(asiática, 1:30)
colnames(raw_targets)[67:72] = paste0(negra, 1:6)
raw_targets
```

gene <chr>	latino1 <dbl>	latino2 <dbl>	latino3 <dbl>	latino4 <dbl>	latino5 <dbl>	blanco1 <dbl>	blanco2 <dbl>
133K02	410.4927	327.0263	477.1076	216.8331	211.8041	470.6122	342.4043
5T4	933.7998	796.9726	189.8969	489.3010	345.8573	351.4286	1018.0377
A-362G6.1	1366.0969	1645.6922	3008.6163	1318.1170	1103.2578	2365.7143	1357.1056
A-C1	15.6423	1.0561	0.0000	89.8716	8.0432	1.2245	0.4171
a1/3GTP	21.8045	67.9398	294.9823	24.2511	135.3937	175.1020	99.6768
A1BG-AS1	9.3190	29.8372	41.9091	4.5934	36.4491	26.5837	17.6624
A1CF	7.1102	91.5251	8.1095	1.4265	0.0000	105.7143	0.4171
A2M	6558.4786	15557.6591	46655.4891	11145.9772	5076.7957	26388.3633	8842.9528
A2M-AS1	74.8177	53.0071	23.6797	49.4579	54.7607	84.2898	22.4419
A2ML1	0.0000	14.4328	0.3379	41.3695	2.6811	0.8163	1.6682

Estructura de la *count table*:

```
summary(raw_targets)
```

```
     gene              latino1            latino2            latino3        
 Length:20965       Min.   :     0.00   Min.   :     0.00   Min.   :     0.00  
 Class :character   1st Qu.:     5.21   1st Qu.:     7.55   1st Qu.:     5.41  
 Mode  :character   Median :   191.97   Median :   216.84   Median :   207.81  
                    Mean   :   898.39   Mean   :   900.30   Mean   :   912.71  
                    3rd Qu.:   827.15   3rd Qu.:   848.02   3rd Qu.:   838.99  
                    Max.   :167382.76   Max.   :140713.18   Max.   :225803.68  
    latino4            latino5            blanco1            blanco2        
 Min.   :     0.0   Min.   :     0.0   Min.   :     0.00   Min.   :     0.0  
 1st Qu.:    10.0   1st Qu.:     2.7   1st Qu.:     6.94   1st Qu.:     3.8  
 Median :   197.0   Median :   144.8   Median :   215.51   Median :   179.3  
 Mean   :  1002.8   Mean   :   987.8   Mean   :   874.06   Mean   :  1044.3  
 3rd Qu.:   851.3   3rd Qu.:   745.3   3rd Qu.:   832.45   3rd Qu.:   803.0  
 Max.   :590020.0   Max.   :530563.8   Max.   :175683.38   Max.   :409695.3  
    blanco3            blanco4            blanco5            blanco6        
 Min.   :     0.00   Min.   :     0.00   Min.   :     0.00   Min.   :     0.00  
 1st Qu.:     2.13   1st Qu.:     5.43   1st Qu.:     3.74   1st Qu.:     4.71  
 Median :   140.14   Median :   191.40   Median :   174.21   Median :   186.66  
 Mean   :   932.33   Mean   :   896.97   Mean   :   840.34   Mean   :   979.64  
 3rd Qu.:   753.65   3rd Qu.:   820.07   3rd Qu.:   813.08   3rd Qu.:   800.65  
 Max.   :249912.90   Max.   :228963.08   Max.   :156720.39   Max.   :282584.27  
    blanco7            blanco8            blanco9            blanco10       
 Min.   :     0.00   Min.   :     0.0   Min.   :     0.0   Min.   :     0.00  
 1st Qu.:     6.15   1st Qu.:     8.1   1st Qu.:    11.4   1st Qu.:     9.33  
 Median :   204.20   Median :   213.5   Median :   230.7   Median :   221.70  
 Mean   :   936.34   Mean   :   974.4   Mean   :   994.2   Mean   :   918.65  
 3rd Qu.:   832.46   3rd Qu.:   840.6   3rd Qu.:   903.4   3rd Qu.:   852.20  
 Max.   :315622.05   Max.   :246587.0   Max.   :848994.5   Max.   :203911.21  
    blanco11           blanco12           blanco13           blanco14       
 Min.   :     0.00   Min.   :     0.00   Min.   :     0.00   Min.   :     0.00  
 1st Qu.:     3.21   1st Qu.:     7.19   1st Qu.:     5.74   1st Qu.:     2.24  
 Median :   164.38   Median :   214.45   Median :   207.89   Median :   164.49  
 Mean   :   872.54   Mean   :   995.62   Mean   :   924.34   Mean   :   919.02  
 3rd Qu.:   810.63   3rd Qu.:   853.48   3rd Qu.:   838.28   3rd Qu.:   791.78  
 Max.   :245243.96   Max.   :161396.74   Max.   :143993.31   Max.   :134308.79  
    blanco15           blanco16           blanco17           blanco18       
 Min.   :     0.0   Min.   :     0.00   Min.   :     0.00   Min.   :     0.00  
 1st Qu.:     4.1   1st Qu.:     7.19   1st Qu.:     5.74   1st Qu.:     2.24  
 Median :   178.9   Median :   214.45   Median :   207.89   Median :   164.49  
 Mean   :  1017.2   Mean   :   995.62   Mean   :   924.34   Mean   :   919.02  
 3rd Qu.:   846.6   3rd Qu.:   853.48   3rd Qu.:   838.28   3rd Qu.:   791.78  
 Max.   :769302.8   Max.   :161396.74   Max.   :143993.31   Max.   :134308.79  
```

Por tanto, de los datos crudos obtenidos a partir de la *count table* destacar que se ha evaluado la expresión de un total de 20965 genes en 71 muestras (5 latina, 30 blanca, 30 asiática y 6 negra). La *count table* proporciona copias expresadas de cada uno de los genes analizados para cada una de las muestras.

A continuación, se crea el objeto en formato DGEList mediante la función DGEList() del paquete edgeR, definiendo los 4 grups de comparación (latina vs. blanca vs. asiática vs. negra), que será el objeto a partir del qual se procederá al análisis de datos de ultra-secuenciación de RNA:

```
library(edgeR)
poblacion <- c(latina, latina, latina,latina, latina,
blanca,blanca,blanca,blanca,blanca, blanca, blanca, blanca, blanca, blanca, blanca,
blanca, blanca, blanca, blanca, blanca, blanca, blanca, blanca, blanca, blanca,
blanca, blanca, blanca, blanca, blanca, blanca, blanca, blanca, blanca, asiática,
asiática, asiática, asiática, asiática, asiática, asiática, asiática, asiática,
asiática, asiática, asiática, asiática, asiática, asiática, asiática, asiática,
asiática, asiática, asiática, asiática, asiática, asiática, asiática, asiática,
asiática, asiática, asiática, asiática, asiática, negra, negra, negra, negra, negra,
negra)
col_poblacion <- ifelse(poblacion == latina, "orange",
                  ifelse(poblacion == blanca, "grey",
                  ifelse(poblacion == asiática, "yellow",
                        "black")))
d <- DGEList(counts=raw_targets[,2:72], genes = raw_targets[,1], group=poblacion)
d$samples
dim(d)
```

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18834670	1
latino2	latino	18874693	1
latino3	latino	19134901	1
latino4	latino	21024677	1
latino5	latino	20708185	1
blanco1	blanco	18324595	1
blanco2	blanco	21894258	1
blanco3	blanco	19546320	1
blanco4	blanco	18804940	1
blanco5	blanco	17617651	1

1-10 of 71 rows Previous 1 2 3 4 5 6 _ 8 Next

Y se comprueba que el objeto creado presenta las mismas dimensiones que la *count table* inicial, con información de 20965 genes en 71 muestras.

CONTROL DE CALIDAD DE LOS DATOS CRUDOS

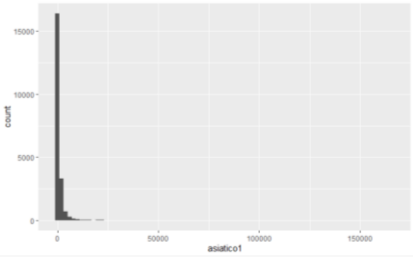
Transformación de los datos

La transformación de los *counts* a *pseudoCounts* en escala logarítmica en base 2 (log2) nos permite reducir la variabilidad y aproximar la distribución de los contajes a la normalidad, de forma que facilita la visualización de los datos.

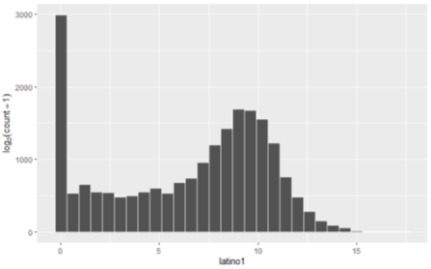
Se presenta la comparación de la representación de los *counts* y *pseudoCounts* de la muestra latino1:

```
library(ggplot2)
```

```
ggplot(raw_targets, aes(x = latino1)) + geom_histogram(fill = "#525252", binwidth =
2000)
```




```
library(ggplot2)
pseudoCount = log2(raw_targets[,2:21] + 1)
ggplot(pseudoCount, aes(x = latino1)) + ylab(expression(log[2](count + 1))) +
geom_histogram(colour = "white", fill = "#525252", binwidth = 0.6)
```



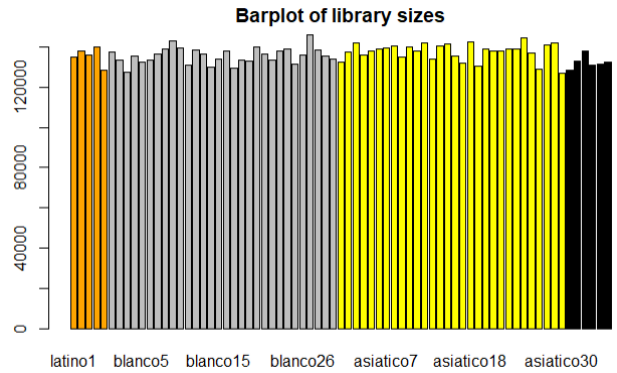
Se observa que cuando se trabaja con los datos transformados en escala log2 se gana precisión por la detección y representación de los resultados. En base a este hecho, se justifica que el análisis de calidad de los datos se base en la evaluación de los *pseudoCounts*.

Distribución de *pseudoCounts* entre muestras

La visualización de la distribución de *pseudoCounts* entre las diferentes muestras es útil para comparar la expresión génica entre ellas.

BARPLOT

```
librarySizes <- colSums(pseudoCount)
barplot(librarySizes,
        names=names(librarySizes),
        col = col_poblacion,
        main="Barplot of library sizes")
```

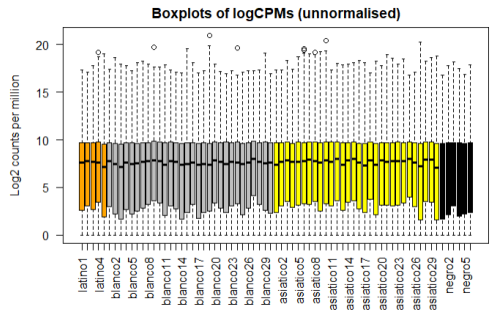


Se observa que antes de la normalización, a partir de los datos crudos transformados en escala log2, el tamaño de las bibliotecas es bastante homogéneo entre las muestras, aunque parece que las muestras de asiática, tienen recuentos un poco más elevados que el resto. En todo

caso, el tamaño de la biblioteca se aproxima a los 14.000 recuentos para la mayoría de las muestras.

BOXPLOT

```
library(ggplot2)
boxplot(pseudoCount, xlab="", ylab="Log2 counts per million",las=2)
title("Boxplots of logCPMs (unnormalised)")
```



Se observa que la distribución de los recuentos (*pseudoCounts*) es bastante homogénea entre las 71 muestras, con algún valor atípico o *outlier*.

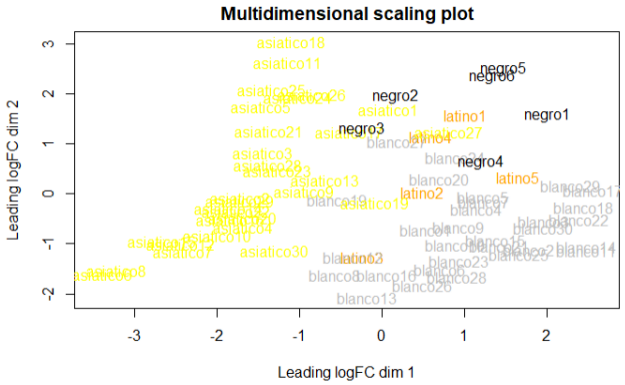
APROXIMACIÓN STAT WORLD (ANÁLISIS MULTIVARIANTE)

El análisis multivariante se lleva a cabo mediante el *multidimensional scaling plot* como representación del análisis de componentes principales, y el análisis de *clústers* y *heatmap*:

MULTIDIMENSIONAL SCALING PLOT (MDS PLOT)

Un MDS Plot es una visualización de un análisis de componentes principales. Si el experimento está bien controlado y ha funcionado bien, lo que se espera ver es que las mayores fuentes de variación en los datos son los tratamientos / grupos que nos interesan. También es una herramienta increíblemente útil para el control de calidad y la verificación de valores atípicos. Se puede usar la función plotMDS para crear el gráfico MDS.

```
plotMDS(pseudoCount,col=col_poblacion)
title("Multidimensional scaling plot")
```

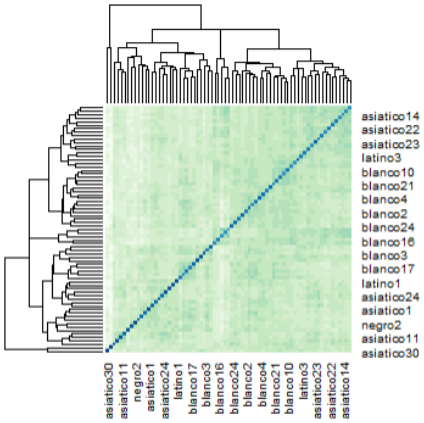


Se observa cierta agrupación por población. La población asiática tiene tendencia a distribuirse de forma agrupada a la izquierda, la población negra tiene tendencia a distribuirse por arriba, y las poblaciones blanca y latino" más por el centro y derecha de forma mezclada entre las 2 poblaciones.

Esta agrupación menos definida de las poblaciones latina y negra puede ser consecuencia de la poca cantidad de muestras que disponemos.

CLUSTERING I *HEATMAP*

```
library(stats)
library(RColorBrewer)
mat.dist = pseudoCount
colnames(mat.dist) = paste(colnames(mat.dist), sep = " : ")
mat.dist = as.matrix(dist(t(mat.dist)))
mat.dist = mat.dist/max(mat.dist)
hmcol = colorRampPalette(brewer.pal(9, "GnBu"))(16)
heatmap(mat.dist, col = rev(hmcol), symkey = FALSE, margins = c(9, 9))
```



FILTRADO

Muchos de estos genes presentarán una expresión muy baja que no contribuirá en el resultado y dificultará el análisis. Por tanto, aplicamos un filtro de manera que se conserven los genes que presenten como mínimo 1cmp (recuento por millón) en al menos 4 muestras (ya que en el diseño del estudio se dispone de una población (latina) con 5 muestras).

```
keep <- rowSums(cpm(d) > 1) >= 4
d_filtrada <- d[keep,]
dim(d_filtrada)
```

```
[1] 17101    71
```

Se observa que después de aplicar el filtrado especificado, en vez de 20965 genes se seleccionan 17101 genes, que serán los que se analizarán.

Otra forma de observar el efecto del filtrado sobre los datos es comparando las bibliotecas que contienen los recuentos totales de cada una de las muestras, antes y después de aplicar el filtro.

```
d$samples$lib.size <- colSums(d$counts)
d$samples$lib.size
```

```
 [1] 18834670 18874693 19134901 21024677 20708185 18324595 21894258 19546320 18804940 17617651
[11] 20538185 19630295 20428758 20843381 19259500 18292808 20873179 19378688 19267175 21325635
[21] 21408230 18046173 19672552 28424510 19081354 18249439 18433275 19121481 19628684 18033585
[31] 18585972 19838367 19008158 19617931 17782409 18850739 17705619 19827369 17359781 18742346
[41] 25932598 20584906 24261748 20671128 22964789 19227337 19261734 18986923 20389051 20201795
[51] 22830478 20798599 19591652 20001257 19569778 21678792 21045851 20794853 20326022 18585301
[61] 19298246 20431201 19392499 20032404 20598971 17175521 17857812 20912584 17787251 17795826
[71] 21022383
```

```
d_filtrada$samples$lib.size <- colSums(d_filtrada$counts)
d_filtrada$samples$lib.size
```

```
 [1] 18828654 18869231 19131475 21013888 20704310 18319033 21891533 19542815 18802405 17615274
[11] 20534325 19627329 20424033 20832886 19245350 18289742 20870016 19297619 19265666 21322517
[21] 21403180 18044036 19669496 27259170 19071015 18243889 18429801 19113074 19621611 18031594
[31] 18581648 19820947 19004034 19612925 17780365 18844824 17700995 19820966 17355695 18730156
[41] 25924978 20580137 24255657 20662410 22958353 19202133 19255560 18969165 20384198 20194911
[51] 22827405 20796240 19577415 19997624 19565639 21675183 21039315 20789707 20320496 18547457
[61] 19289894 20428052 19386735 20026347 20575195 17172162 17841813 20907378 17784397 17788012
[71] 21018554
```

Se observa como el número de lecturas (recuentos) de cada una de las bibliotecas disminuye después de aplicar el filtro, ya que se han eliminado aquellas lecturas correspondientes a los genes poco expresados.

NORMALIZACIÓN

Aproximación de normalización a través de la función calcNormFactors() del paquete de edgeR.

```
norm_targets <- calcNormFactors(d_filtrada)
norm_targets$samples
```

	group <fctr>	lib.size <dbl>	norm.factors <dbl>
latino1	latino	18828654	1.0309422
latino2	latino	18869231	1.0317924
latino3	latino	19131475	1.0359768
latino4	latino	21013888	0.9739310
latino5	latino	20704310	0.8394549
blanco1	blanco	18319033	1.0624912
blanco2	blanco	21891533	0.8374890
blanco3	blanco	19542815	0.8996154
blanco4	blanco	18802405	1.0303671
blanco5	blanco	17615274	1.0606024

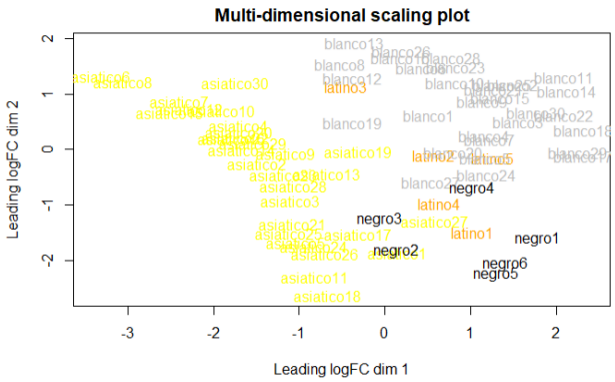
1-10 of 71 rows Previous 1 2 3 4 5 6 ... 8 Next

Al normalizar los datos se obtiene el tamaño de la librería de cada muestra con el factor de normalización aplicado, según la distancia con la línea de lecturas de referencia.

EXPLORACIÓN DE LOS DATOS NORMALIZADOS

MDS PLOT

```
plotMDS(norm_targets, col = col_poblacion, main = "Multi-dimensional scaling plot")
```



Se observa que las posiciones de las muestras han variado respecto el MDS Plot de datos no normalizados, pero que las distancias entre las muestras se mantienen similares.

ESTIMACIÓN DE LA DISPERSIÓN DE LOS DATOS

Se calcula la dispersión comuna, que estima la dispersión del global del conjunto de datos promediada para todos los genes; y la dispersión específica para los genes.

Dispersión comuna:

```
dispersio <- estimateCommonDisp(norm_targets, verbose=TRUE)
dispersio$common.dispersion
```

```
Disp = 0.66139 , BCV = 0.8133
[1] 0.6613888
```

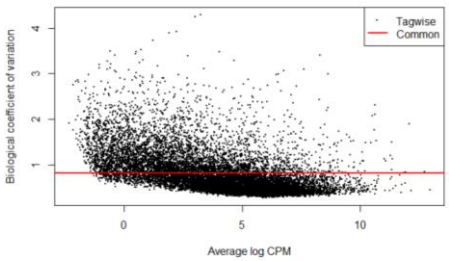
Dispersión específica:

```
dispersio_gens <- estimateTagwiseDisp(dispersio)
dispersio_gens$common.dispersion
```

```
[1] 0.6613888
```

En este caso se observa que la dispersión comuna y la dispersión específica coinciden, siendo del 66%.

Se representa la dispersión gráficamente:



IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

Para el estudio de la expresión diferencial de genes se realizan comparaciones de las poblaciones 2 a 2.
En este ejemplo se expone la comparación “latino-blanco”.
Aplicando la función exacTest() i topTag() del paquete edgeR, se obtiene la tabla con los genes diferencialmente expresados con el p-value correspondiente:

```
et_latino_blanco <- exactTest(dispersio_gens_latino_blanco)
top_latino_blanco <- topTags(et_latino_blanco)
top_latino_blanco
```

	genes	logFC	logCPM	PValue	FDR
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
6881	HBM	13.523587	4.1403284	1.264001e-192	2.017977e-188
16262	SERP1	-13.753728	7.7644347	7.535138e-70	6.014924e-66
1781	C1orf64	5.081704	3.3004210	3.258914e-67	1.734285e-63
6641	GPX6	9.758006	0.4064400	4.699920e-57	1.875856e-53
2576	CDC23	3.251215	6.7036750	1.748273e-50	5.582236e-47
2123	CACNG5	8.582110	0.7611524	3.295214e-44	8.768015e-41
6882	HBM	-14.521512	7.0222703	2.005191e-43	4.573267e-40
2625	CDH24	7.678858	1.0054749	3.498143e-33	6.980982e-30
2577	CDC23	-3.787847	8.5466426	4.525090e-24	8.027007e-21
8883	LINC00875	4.481251	0.1675795	8.474915e-23	1.353020e-19

1-10 of 10 rows

Se obtienen los 10 genes diferencialmente expresados (DE) entre las 2 poblaciones de forma estadísticamente significativa. Se observa también el logCPM (logaritmo del recuento de copias por millón) y el FDR (*false discovery rate*) que equivaldría a la tasa de falsos positivos.

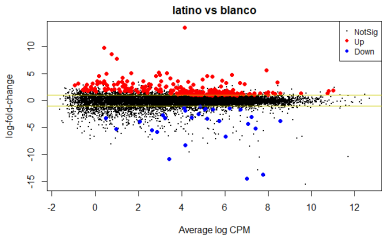
La función decideTestDGE permite ver los genes sobre o infraexpresados en la comparación blanco-asiático.

```
summary(de_latino_blanco <- decideTestsDGE(et_latino_blanco))
```

	latino-blanco
Down	28
NotSig	15705
Up	232

Y se representa gráficamente a través de la función plotMD():

```
plotMD(et_latino_blanco)
abline(h=c(-1,1), col="yellow3")
```



Se observa que 28 de los genes se encuentran *down-regulated* (expresión disminuida) y 232 de los genes se encuentran *up-regulated* (expresión aumentada) en la comparación génica entre las 2 poblaciones. El resto de 15705 genes no muestran una expresión diferencial significativa.

ANOTACIÓN DE LOS RESULTADOS

Una vez se han seleccionados los genes diferencialmente expresados en la comparación entre las poblaciones latina vs. blanca, se tiene que identificar estos genes con sus anotaciones equivalentes en base a Gene Ontology (Entrez Gee identifier, RefSeq, Ensembl, Gene Symbol), permitiendo obtener más información. Este procedimiento se llevará a cabo a través de la descarga de la base de datos "org.Hs.eg.db" del paquete BiocManager que contiene las diferentes anotaciones posibles de los genes humanos.

```
library(org.Hs.eg.db)
egENSEMBL <- toTable(org.Hs.egENSEMBL)
match_gens_ensembl_latino_blanco <- match(top_latino_blanco$table$genes,
egENSEMBL$ensembl_id)
top_latino_blanco$table$EntrezGene <-
egENSEMBL$gene_id[match_gens_ensembl_latino_blanco]

egREFSEQ <- toTable(org.Hs.egREFSEQ)
match_gens_refseq_latino_blanco <- match(top_latino_blanco$table$EntrezGene,
egREFSEQ$gene_id)
top_latino_blanco$table$RefSeq <- egREFSEQ$accession[match_gens_refseq_latino_blanco]

egSYMBOL <- toTable(org.Hs.egSYMBOL)
match_gens_symbol_latino_blanco <- match(top_latino_blanco$table$EntrezGene,
egSYMBOL$gene_id)
top_latino_blanco$table$SYMBOL <- egSYMBOL$symbol[match_gens_symbol_latino_blanco]

top_latino_blanco
```

ANÁLISIS DE SIGNIFICACIÓN BIOLÓGICA (*GENE ENRICHMENT ANALYSIS*)

```
go_latino_blanco <- goana(et_latino_blanco)
topGO(go_latino_blanco, ont="BP", sort="Up", n=20, truncate=50)
```

Se observan las 20 anotaciones GO enriquecidas:

	Term	Ont	N	Up	Do...	PUp
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0048822	enucleate erythrocyte development	BP	2	2	0	0.0001999502
GO:2001016	positive regulation of skeletal muscle cell dif...	BP	2	2	0	0.0001999502
GO:0006590	thyroid hormone generation	BP	10	3	0	0.0003108243
GO:0042403	thyroid hormone metabolic process	BP	11	3	0	0.0004229814
GO:0031572	G2 DNA damage checkpoint	BP	12	3	0	0.0005581698
GO:0071824	protein-DNA complex subunit organization	BP	142	8	1	0.0008791710
GO:0044818	mitotic G2/M transition checkpoint	BP	14	3	0	0.0009046192
GO:0048485	sympathetic nervous system development	BP	15	3	0	0.0011191552
GO:0060751	branch elongation involved in mammary gland duc...	BP	4	2	0	0.0011775333
GO:0060750	epithelial cell proliferation involved in mamma...	BP	4	2	0	0.0011775333
GO:0035162	embryonic hemopoiesis	BP	16	3	0	0.0013632772
GO:1902749	regulation of cell cycle G2/M phase transition	BP	119	7	0	0.0014481447
GO:0043970	histone H3-K9 acetylation	BP	5	2	0	0.0019443684
GO:0060745	mammary gland branching involved in pregnancy	BP	5	2	0	0.0019443684
GO:2000615	regulation of histone H3-K9 acetylation	BP	5	2	0	0.0019443684
GO:1902750	negative regulation of cell cycle G2/M phase tr...	BP	64	5	0	0.0020585739
GO:0008209	androgen metabolic process	BP	19	3	0	0.0022871160
GO:0043966	histone H3 acetylation	BP	19	3	0	0.0022871160
GO:0022603	regulation of anatomical structure morphogenesi...	BP	567	17	1	0.0024081726
GO:0043353	enucleate erythrocyte differentiation	BP	6	2	0	0.0028895537

Se observa que los principales genes diferencialmente expresados entre las muestras latina y blanca son proteinas implicadas en los *checkpoint* de la mitosis celular y de la reparación del daño del DNA. También llama la atención la diferencia de expresión de genes implicados en el metabolismo de las hormonas tiroidales así como en el proceso de la hematopoesis.

9.2.2 ANÁLISIS DESCRIPTIVO DE LAS MUESTRAS

VARIABLES BASALES GENERALES

POBLACIÓN

```
table(taula_clinica_total$Pobla)
```

```
asiatico   blanco   latino    negro
      30       30        5        6
```

Se estudian un total de 71 muestras de 4 poblaciones distintas: 5 latina, 30 blanca, 30 asiática y 6 negra.

SEXO

```
table(taula_clinica_total$Sex)
```

```
  Female   Male
0     27     44
```

De las 71 muestras, 27 de los sujetos eran mujeres (38,03%) y 44 eran hombres (62,97%).

EDAD AL DIAGNÓSTICO

```
library(nortest)
lillie.test(taula_clinica_total$Diagnosis.Age)
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  taula_clinica_total$Diagnosis.Age
D = 0.1032, p-value = 0.05858
```

Se puede asumir normalidad de la variable “edad al diagnóstico”.

```
mean(taula_clinica_total$Diagnosis.Age)
sd(taula_clinica_total$Diagnosis.Age)
```

```
[1] 64.29577
[1] 13.34851
```

La edad media al diagnóstico fue de 64,29 años (SD +/- 13,34 años).

VARIABLES BASALES ENTRE POBLACIONES

SEXO -POBLACIÓN

```
tabla_sexo_pobla <- table(taula_clinica_total$Sex, taula_clinica_total$Pobla)
fisher.test(tabla_sexo_pobla)
```

```
        Fisher's Exact Test for Count Data

data:  tabla_sexo_pobla
p-value = 0.6207
alternative hypothesis: two.sided
```

No existen evidencias estadísticas que apoyen una diferencia en la distribución del sexo entre poblaciones distintas (p- valor 0,6207).

EDAD AL DIAGÓSTICO - POBLACIÓN

```
summary(aov(Diagnosis.Age ~ Pobla, data = taula_clinica_total))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Pobla        3   1280   426.7   2.554 0.0627 .
Residuals   67  11193   167.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

No existen evidencias estadísitcas que apoyen una diferencia en la edad al diagnóstico entre poblaciones distintas (p- valor 0,0627).

VARIABLES BIOLÓGICAS - PRONÓSTICAS GENERALES

TIPO DE CÁNCER

```
table(taula_clinica_total$Cancer.Type.Detailed)
```

```
 Diffuse Type Stomach Adenocarcinoma           Mucinous Stomach Adenocarcinoma
                                  11                                         8
    Papillary Stomach Adenocarcinoma Signet Ring Cell Carcinoma of the Stomach
                                   1                                         1
             Stomach Adenocarcinoma            Tubular Stomach Adenocarcinoma
                                  42                                         8
```

El tipo histológico de cáncer gástrico más frecuente con 42 casos (50,15%) fue el adenocarcinoma gástrico clásico, seguido del adenocarcinoma gástrico difuso con 11 casos (15,49%), 8 casos (11,26%) de adenocarcinoma mucinoso y de adenocarcinoma tubular, y tan sólo hubo 1 caso (1,40%) de adenocarcinoma papilar y 1 de carcinoma "ring cell".

NÚMERO DE MUTACIONES

```
lillie.test(taula_clinica_total$Mutation.Count)
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  taula_clinica_total$Mutation.Count
D = 0.31915, p-value < 2.2e-16
```

Se puede observar que la variable mutaciones no sigue una distribución normal por lo que se utilizará un test no paramétrico.

```
summary(taula_clinica_total$Mutation.Count)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   21.0    65.0   114.0   516.7   583.5  6359.0       3
```

La mediana del número de mutaciones identificadas en las muestras es de 114, con un rango intercuartílico de 65-583.

ESTADIO DEL TUMOR

```
table(taula_clinica_total$American.Joint.Committee.on.Cancer.Tumor.Stage.Code)
```

```
   T1 T1a T1b  T2 T2a T2b  T3  T4 T4a T4b  TX
0   1   0   1   6   4   2  21   7  18   2   9
```

MESTÁSTASI

```
table(taula_clinica_total$American.Joint.Committee.on.Cancer.Metastasis.Stage.Code)
```

```
   M0 M1 MX
0 58  6  7
```

La mayoría de los cánceres fueron diagnosticados en un estadío avanzado T3 (21 casos, 29.58%) y T4a (18 casos, 25%), aunque la mayoría aún no presentaban metástasis en el momento del diagnóstico (58 casos, 81.7%).

BARRET

```
table(taula_clinica_total$Barretts.Esophagus)
```

```
    No Yes
46  25   0
```

Se confirma que en 25 casos (35%) no se identificó esófago de Barret como lesión pre-maligna, pero no hay datos en las muestras restantes.

TERAPIA ADJUVANTE POSTQUIRÚRGICA

```
table(taula_clinica_total$Adjuvant.Postoperative.Targeted.Therapy.Administered.Indicator)
```

```
    NO YES
52  13   6
```

Tan sólo se ha reportado que un 8,45% de los casos necesitaron terapia adyuvante postquirúrgica, mientras que un 18% no, pero tan sólo se dispone de datos de 19 de las muestras (26%).

VARIABLES BIOLÓGICAS - PRONÓSTICAS ENTRE POBLACIONES

TIPO DE CÁNCER

```
tabla_tipo_pobla <- table(taula_clinica_total$Cancer.Type.Detailed,
taula_clinica_total$Pobla)
chisq.test(tabla_tipo_pobla)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  tabla_tipo_pobla
X-squared = 27.327, df = 15, p-value = 0.02618
```

Se encuentran diferencias estadísticamente significativas entre el tipo de cáncer en al menos 2 de las poblaciones, por lo que se procede a realizar comparaciones múltiples 2 a 2.

```
prop.table(tabla_tipo_pobla)
```

```
                                          asiatico     blanco     latino      negro
Diffuse Type Stomach Adenocarcinoma       0.08450704 0.04225352 0.02816901 0.00000000
Mucinous Stomach Adenocarcinoma           0.09859155 0.00000000 0.01408451 0.00000000
Papillary Stomach Adenocarcinoma          0.01408451 0.00000000 0.00000000 0.00000000
Signet Ring Cell Carcinoma of the Stomach 0.01408451 0.00000000 0.00000000 0.00000000
Stomach Adenocarcinoma                    0.15492958 0.36619718 0.01408451 0.05633803
Tubular Stomach Adenocarcinoma            0.05633803 0.01408451 0.01408451 0.02816901
```

```
comparaciones_latino <- subset(taula_clinica_total, taula_clinica_total$Pobla ==
latina)
comparaciones_blanco <- subset(taula_clinica_total, taula_clinica_total$Pobla ==
blanca)
comparaciones_asiatico <- subset(taula_clinica_total, taula_clinica_total$Pobla ==
asiática)
comparaciones_negro <- subset(taula_clinica_total, taula_clinica_total$Pobla ==
negra)
```

Se realizan las comparaciones 2 a 2 entre las 4 poblaciones. Se expone un ejemplo:

Ejemplo comparación 2 a 2 entre latina y blanca:

```
com_latino_blanco <- rbind(comparaciones_latino, comparaciones_blanco)
```

```
tabla_comp_tipo_latino_blanco <- table(com_latino_blanco$Cancer.Type.Detailed,
com_latino_blanco$Pobla)
fisher.test(tabla_comp_tipo_latino_blanco)
```

```
        Fisher's Exact Test for Count Data

data:  tabla_comp_tipo_latino_blanco
p-value = 0.007076
alternative hypothesis: two.sided
```

En este caso se observan diferencias estadísticamente significativas (p-valor<0.05).

NÚMERO DE MUTACIONES

```
summary(aov(taula_clinica_total$Mutation.Count ~ taula_clinica_total$Pobla))
```

```
                         Df   Sum Sq Mean Sq F value Pr(>F)
taula_clinica_total$Pobla  3  3832041 1277347   1.157  0.333
Residuals                 64 70673720 1104277
3 observations deleted due to missingness
```

No existen evidencias estadísticas que apoyen la presencia de diferencias significativas en el número de mutaciones de las muestas entre las distinas poblaciones (p-valor>0.05).

ESTADIO DEL TUMOR

```
tabla_estadio_pobla <-
table(taula_clinica_total$American.Joint.Committee.on.Cancer.Tumor.Stage.Code,
taula_clinica_total$Pobla)
fisher.test(tabla_estadio_pobla, simulate.p.value = TRUE)
```

```
        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  tabla_estadio_pobla
p-value = 0.0009995
alternative hypothesis: two.sided
```

Existen diferencias estadísticamente significativas entre el estadío al diagnóstico y las poblaciones, siendo la proporción:

```
prop.table(tabla_estadio_pobla)
```

```
       asiatico     blanco     latino      negro
     0.00000000 0.00000000 0.00000000 0.00000000
T1   0.00000000 0.01408451 0.00000000 0.00000000
T1a  0.00000000 0.00000000 0.00000000 0.00000000
T1b  0.00000000 0.00000000 0.00000000 0.01408451
T2   0.02816901 0.02816901 0.00000000 0.02816901
T2a  0.00000000 0.05633803 0.00000000 0.00000000
T2b  0.00000000 0.01408451 0.01408451 0.00000000
T3   0.15492958 0.08450704 0.02816901 0.02816901
T4   0.02816901 0.05633803 0.01408451 0.00000000
T4a  0.18309859 0.04225352 0.01408451 0.01408451
T4b  0.02816901 0.00000000 0.00000000 0.00000000
TX   0.00000000 0.12676056 0.00000000 0.00000000
```

Se realizan las comparaciones 2 a 2 entre las 4 poblaciones. Se expone un ejemplo:

comparación 2 a 2 entre latina y blanca:

```
tabla_comp_estadio_latino_blanco <-
table(com_latino_blanco$American.Joint.Committee.on.Cancer.Tumor.Stage.Code,
com_latino_blanco$Pobla)
fisher.test(tabla_comp_estadio_latino_blanco)
```

```
        Fisher's Exact Test for Count Data

data:  tabla_comp_estadio_latino_blanco
p-value = 0.4121
alternative hypothesis: two.sided
```

En este caso no existen diferencias significativas entre las poblaciones latina y blanco” en la variable estadio al diagnóstico (p-valor >0.05).

MESTÁSTASIS

```
tabla_metastasis_pobla <-
table(taula_clinica_total$American.Joint.Committee.on.Cancer.Metastasis.Stage.Code,
taula_clinica_total$Pobla)
fisher.test(tabla_metastasis_pobla, simulate.p.value = TRUE)
```

```
        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  tabla_metastasis_pobla
p-value = 0.0004998
alternative hypothesis: two.sided
```

Existen diferencias estadísticamente significativas entre encontrar metástasis al diagnóstico y las poblaciones, siendo la proporción:

```
prop.table(tabla_metastasis_pobla)
```

```
      asiatico     blanco     latino      negro
   0.00000000 0.00000000 0.00000000 0.00000000
M0 0.38028169 0.38028169 0.01408451 0.04225352
M1 0.04225352 0.00000000 0.02816901 0.01408451
MX 0.00000000 0.04225352 0.02816901 0.02816901
```

Se realizan las comparaciones 2 a 2 entre las 4 poblaciones. Se expone un ejemplo:

comparación 2 a 2 entre latina y blanca:

```
tabla_comp_metastasis_latino_blanco <-
table(com_latino_blanco$American.Joint.Committee.on.Cancer.Metastasis.Stage.Code,
com_latino_blanco$Pobla)
fisher.test(tabla_comp_metastasis_latino_blanco)
```

```
        Fisher's Exact Test for Count Data

data:  tabla_comp_metastasis_latino_blanco
p-value = 0.001358
alternative hypothesis: two.sided
```

Se observa que existen diferencias significativas entre las poblaciones latina y blanca en relación a la presencia de metástasis al momento del diagnóstico.

No se estudian las variables Barret ni tratamiento adyuvante postquirúrgico dado que se dispone de pocos datos.