



UNIVERSIDAD  
DE GRANADA



# **Characterization of Genetically Different Groups of Cancer Patients Using Unsupervised Methods.**

Lucía Almorox Antón

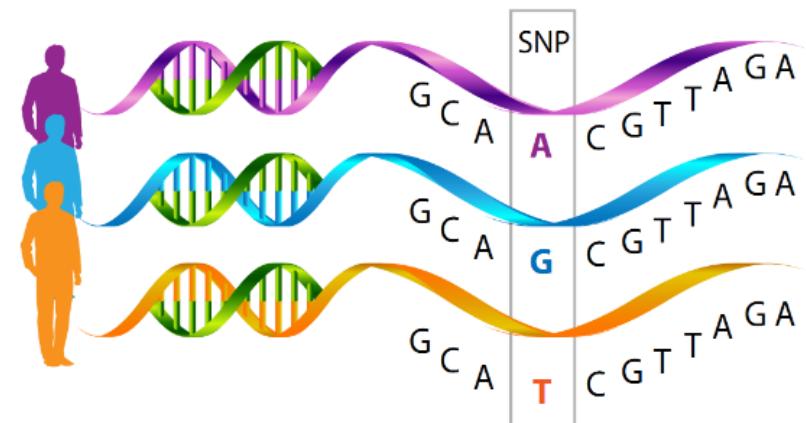
Tutora: María Coral del Val Muñoz

Mentora: Elisa Díaz de la Guardia Bolívar

# Índice

1. Introducción y objetivos.
2. Metodología.
3. Resultados y discusión.
4. Conclusiones y trabajo futuro.

- Los **genomas humanos no son idénticos** entre sí.
- Variaciones afectando a una sola base del genoma: **SNPs (Single Nucleotide Polymorphisms)**.
  - **Diversidad** entre individuos.
  - Predisposición a **enfermedades** [1].
- **Alelos de un SNP**: los distintos **nucleótidos** que pueden aparecer en la posición genómica del SNP.
  - **Alelo mayor**: el nucleótido más común para dicha posición, en una población de referencia.
  - **Alelo menor**: el nucleótido menos común.
    - Suele ser el “alelo de riesgo”.
    - MAF (Minor Allele Frequency) de un SNP en una población: proporción de veces que aparece el alelo menor en la posición genómica del SNP [2].



Concepto de SNP. Figura tomada de un recurso externo.

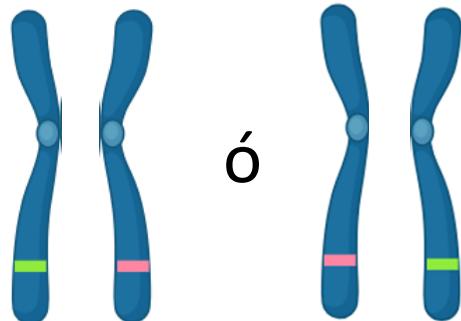
([www.nutrigeneticsspecialists.com](http://www.nutrigeneticsspecialists.com))

- Un SNP en un individuo **humano** en general tiene **dos alelos**. A la combinación se le llama “**genotipo**”.



**Homocigoto mayor (AA).**

Dos copias del alelo mayor.



**Heterocigoto (AB).**

Una copia del alelo mayor y otra del alelo menor.



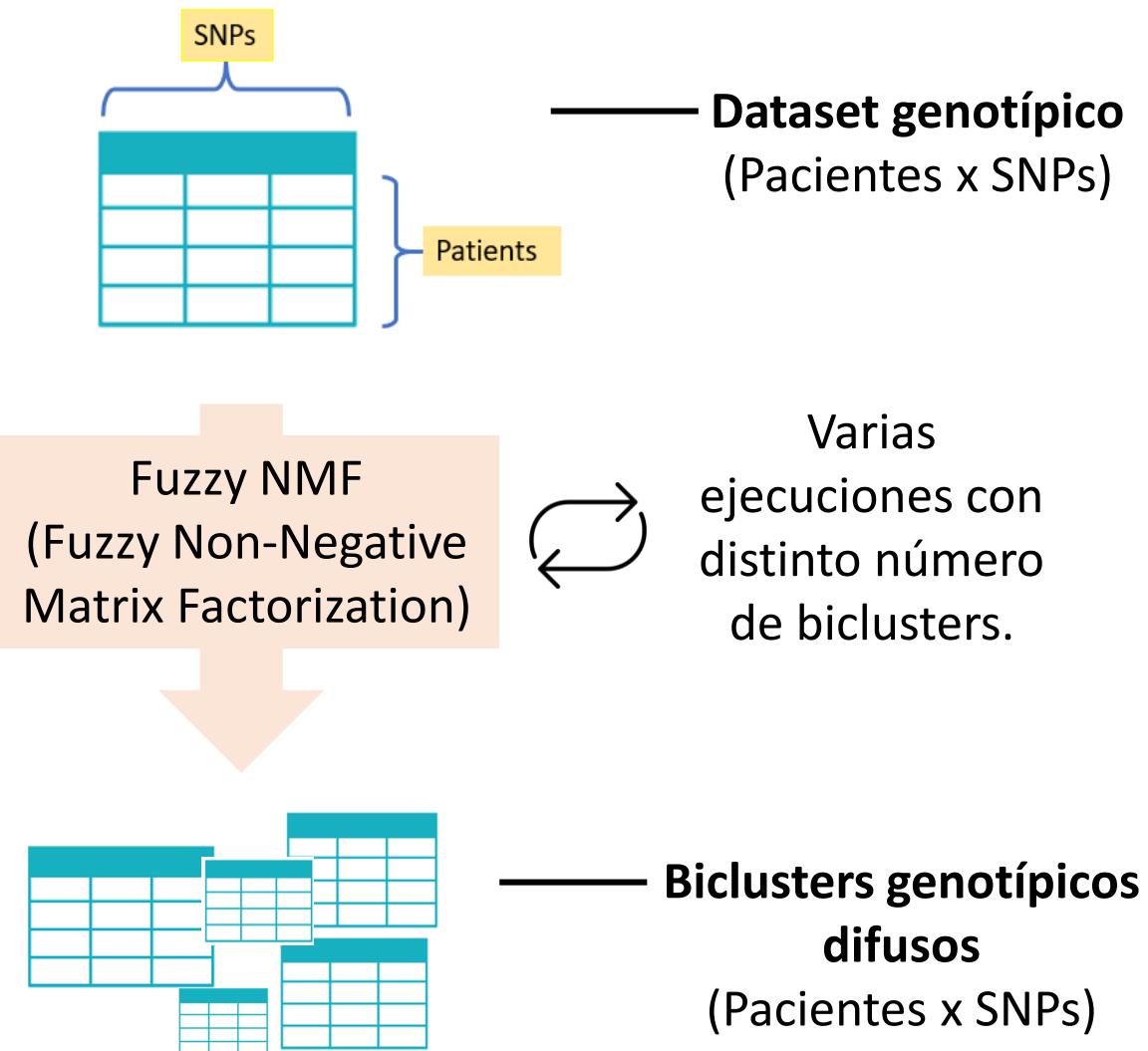
**Homocigoto menor (BB).**

Dos copias del alelo menor.

- El efecto de un SNP puede depender de su **genotipo** [2].
- También se dan **efectos de interacción** entre genotipos de distintos SNPs [3].

- **GWAS** (Genome-Wide Association Study) → búsqueda de **SNPs asociados con condiciones** particulares.
  - Se ha propuesto que los SNPs descubiertos por estos análisis explican una pequeña fracción de la variación genética de las enfermedades humanas [4].
- **PGMRA** (Phenotype x Genotype Many-to-Many Relation Analysis) → algoritmo no supervisado.
  - Supera varias limitaciones del análisis tradicional de GWAS [5].

**PGMRA:** obtención de biclusters difusos de calidad [5].



Esquema de la identificación de biclusters por PGMRA.

## Motivación del trabajo.

- **Cáncer colorectal (o cáncer de colon).**
  - Muy agresivo.
  - Prevalencia en aumento [6].
  - La variabilidad genética de los pacientes puede influenciar significativamente las manifestaciones del cáncer [7].
- **Biclusters de cáncer de colon de PGMRA.**
  - Identificación de variantes genéticas relevantes para distintos subgrupos de pacientes.
  - Diferentes orígenes o síntomas de la enfermedad.



## Motivación del trabajo.

- **Cáncer colorectal (o cáncer de colon).**
  - Muy agresivo.
  - Prevalencia en aumento [6].
  - La variabilidad genética de los pacientes puede influenciar significativamente las manifestaciones del cáncer [7].
  
- **Biclusters de cáncer de colon de PGMRA.**
  - Identificación de variantes genéticas relevantes para distintos subgrupos de pacientes.
  - Diferentes orígenes o síntomas de la enfermedad.



**Objetivo principal:**  
Evaluar la capacidad de PGMRA para, a partir de un tamaño de muestra pequeño, descubrir nuevo conocimiento en la caracterización de grupos genéticamente diferentes de cáncer de colon.

Introducción y objetivos

Metodología

Resultados y discusión

Conclusiones

## Proceso de generación de los datos de entrada.

- **Dataset genotípico** de individuos con **distintos diagnósticos**.
  - Cáncer de colon.
  - Cáncer de mama.
  - Cáncer de piel.
  - Controles sanos.
- Se excluyeron los pacientes con insuficiente calidad de muestra, así como los SNPs que se mantenían demasiado estables a nivel poblacional.
- **Codificación numérica de genotipos.**
  - Missing value  $\equiv 0$
  - Homocigoto mayor (AA)  $\equiv 1$
  - Heterocigoto (AB)  $\equiv 2$
  - Homocigoto menor (BB)  $\equiv 3$
- **Ejecución de PGMRA.**
- **Selección de biclusters que contenían exclusivamente pacientes de cáncer de colon.** Juntos recogían 77 de los 78 pacientes con este diagnóstico en la matriz de entrada.

## Proceso de generación de los datos de entrada.

- Detección y eliminación de 2 pares de muestras repetidas.
- **16 biclusters de cáncer de colon.** En total incluían:

- 74 pacientes.
- 2496 SNPs.



Información de todos los pacientes y SNPs.

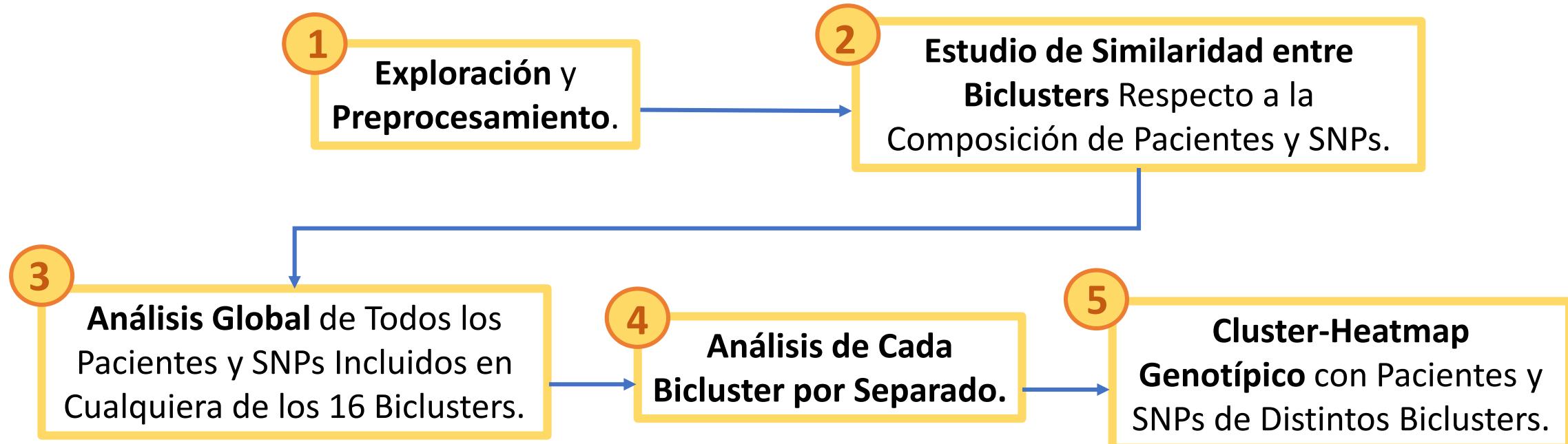
- Dataset genotípico (pacientes x SNPs).
- Dataset clínico (datos de pacientes).
- Dataset de Ensembl (datos de SNPs).

Información de todos los pacientes y SNPs.

- Dataset genotípico (pacientes x SNPs).
- Dataset clínico (datos de pacientes).
- Dataset de Ensembl (datos de SNPs).

Distribución en 16  
biclusters de cáncer  
de colon.

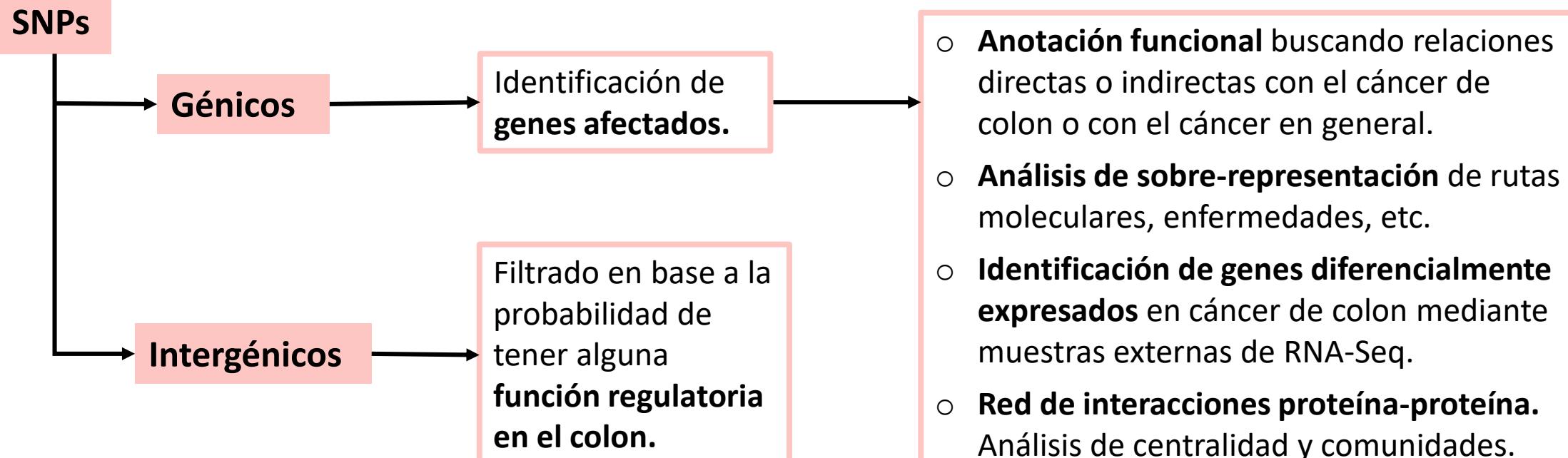
## Metodología



3

### Análisis Global de Todos los Pacientes y SNPs Incluidos en Cualquiera de los 16 Bioclusters.

Dataset clínico → Extracción de reglas de asociación.

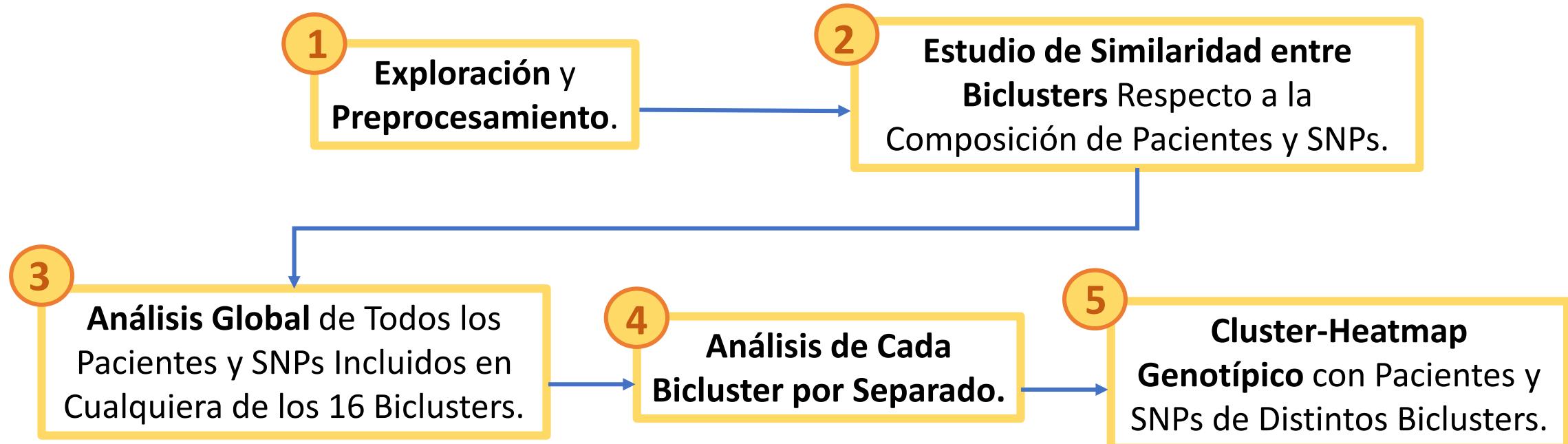


Información de todos los pacientes y SNPs.

- Dataset genotípico (pacientes x SNPs).
- Dataset clínico (datos de pacientes).
- Dataset de Ensembl (datos de SNPs).

Distribución en 16  
biclusters de cáncer  
de colon.

## Metodología



**4**

## Análisis de Cada Biocluster por Separado.

- 
- **Repetición de la mayoría de operaciones del análisis global** para los pacientes y SNPs de cada biocluster por separado → Comparación de resultados.
- **Mapas génicos** mostrando SNPs específicos del biocluster.
- **Cálculo de frecuencias genotípicas y MAFs** específicas del biocluster.

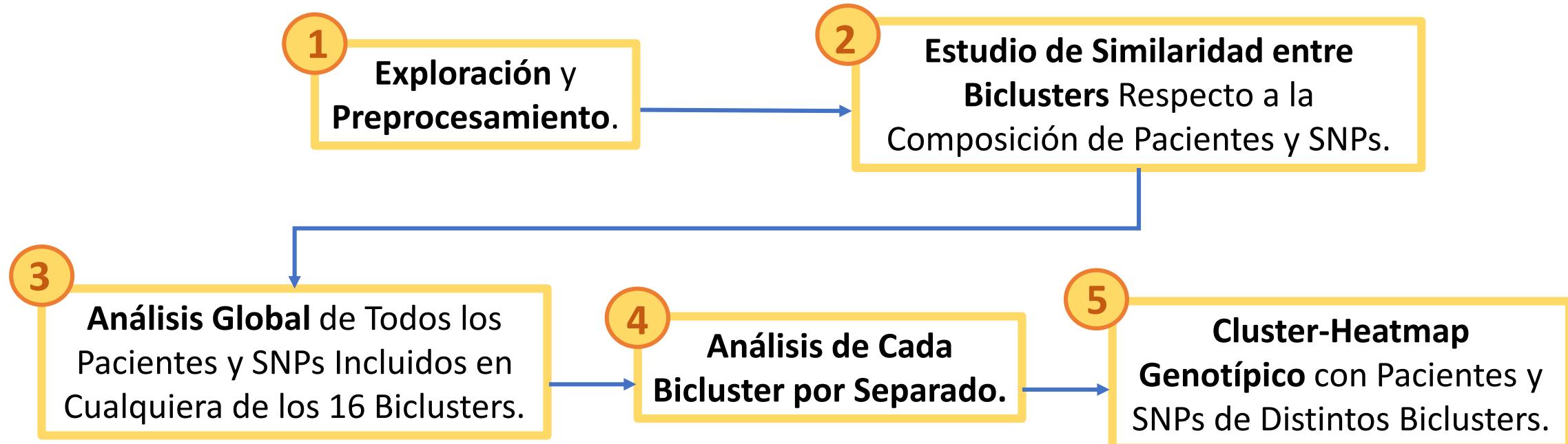
Todo ello permitió identificar biclusters y genes especialmente interesantes → Análisis extendido.

Información de todos los pacientes y SNPs.

- Dataset genotípico (pacientes x SNPs).
- Dataset clínico (datos de pacientes).
- Dataset de Ensembl (datos de SNPs).

Distribución en 16  
biclusters de cáncer  
de colon.

## Metodología



## Herramientas utilizadas (paquetes, servidores web, software, bases de datos).

### Exploración y Preprocesamiento

- Tidyverse (colección de paquetes de R)

### Estudio de Similaridad entre Biclusters

- Tidyverse (colección de paquetes de R)
- Gephi (Software)

### Análisis Global

#### Extracción de reglas de asociación.

- Arules (paquete de R)

#### Anotación de SNPs.

- Ensembl (base de datos)

#### Anotación de genes.

- VarElect (servidor web)
- GeneCards (servidor web)
- org.Hs.eg.db (paquete de R)
- Bio Entrez (paquete de Python)

#### Análisis de sobre-representación en conjuntos de genes.

- DisGeNET (base de datos)
- GO (ontología)
- KEGG (base de datos)
- DOSE (paquete de R)
- ClusterProfiler (paquete de R)

#### Análisis de expresión (RNA-seq)

- TCGA (base de datos)
- GDC (servidor web)
- KnowSeq (paquete de R)

#### Creación de redes de interacción proteína-proteína.

- NetworkAnalyst (servidor web)
- IMEx (base de datos)
- Cytoscape (software)
- Gephi (software)

#### Filtrado de SNPs intergénicos

- RegulomeDB (servidor web)

Herramientas adicionales a las del análisis global

### Análisis de cada Bicluster

#### Creación de mapas génicos mostrando SNPs.

- mapsnp v2 (paquete de R)
- UCSC (base de datos)
- BiomaRt (paquete de R)

#### Creación de tablas resumen

- gt (paquete de R)

### Heatmap Genotípico.

- ComplexHeatmap (paquete de R)

Introducción y objetivos

Metodología

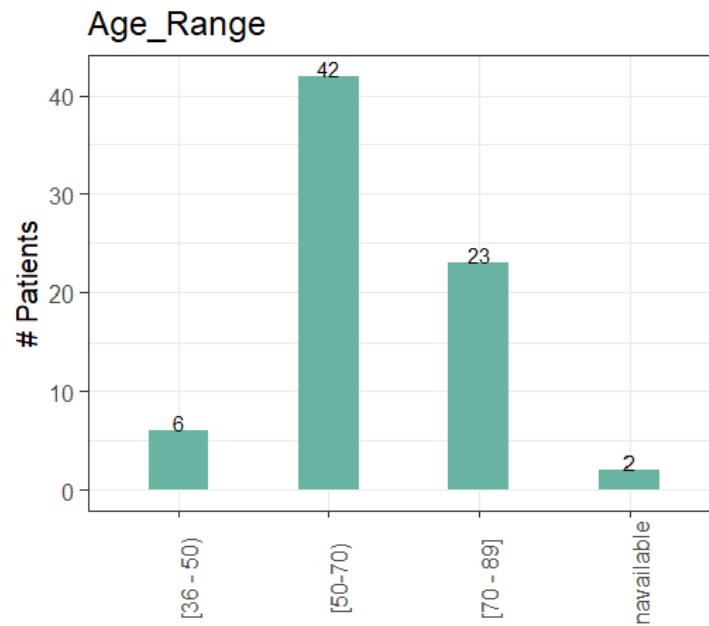
Resultados y discusión

Conclusiones

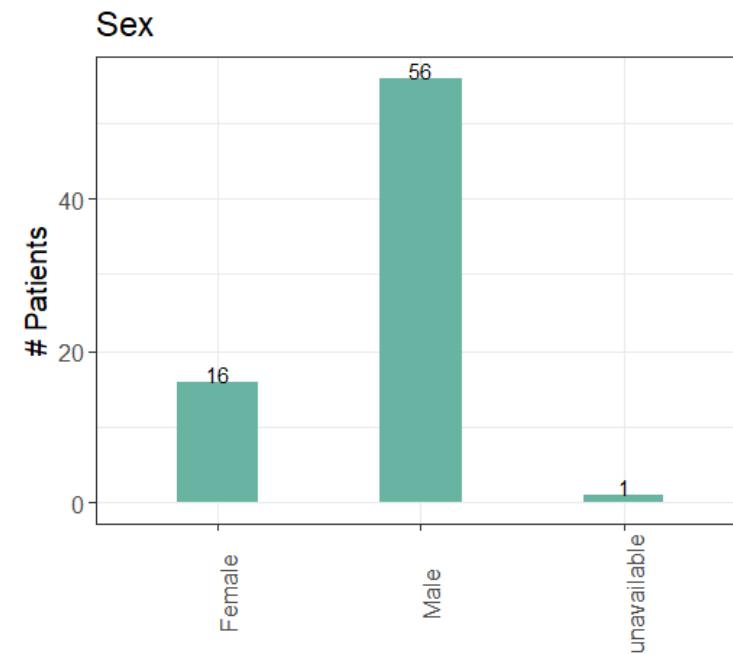
## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 73 pacientes.
  - Identificación de características clínicas generales. Ejemplos:

- Poca representación de pacientes menores de 50 años.
- Poca representación de mujeres.



Distribución del atributo "Age\_Range" en el conjunto total de pacientes.



Distribución del atributo "Sex" en el conjunto total de pacientes.

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 73 pacientes.
  - Identificación de tendencias de co-aparición de valores clínicos. Ejemplo:
    - Valores de CEA basales superiores a 10 se correlacionaban con metástasis.

Una de las reglas de asociación obtenidas con soporte mínimo de 0,05, confianza mínima de 0,7, longitud entre 2 y 4 ítems, y lift mínimo de 1,4.

rules	supp.	conf.	cover.	lift	count
{Basal_CEA=(10 - 12035)} ⇒ {Metastasic=Yes}	0.151	0.846	0.178	2.471	11

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.
  - **972 SNPs génicos.**
  - **1025 SNPs intergénicos.**
  - 499 no reconocidos por Ensembl → no pudimos analizarlos.

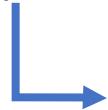
## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ **972 SNPs génicos.**

Valores más comunes en los atributos de Ensembl.

- Cromosoma: X (n= 184)
- Tipo de gen: codificador de proteína (n= 652).
- Tipo de consecuencia del transcripto de RNA: intrónica (n= 866).
- Impacto predicho sobre la función de la proteína: bajo (n= 939).



Se cree que la variación genética de numerosas enfermedades se debe principalmente a múltiples regiones genómicas, cada una contribuyendo con un efecto pequeño [8].

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. —————→ 695 genes afectados. Anotación génica funcional.

Affected Genes	Number of genes	Percentage
Colon Cancer Directly Related	146	21.01
Colon Cancer Indirectly Related	414	59.57
Cancer Directly Related	25	3.6
Cancer Indirectly Related	24	3.45
No relationship found	86	12.37

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. —————→ 695 genes afectados. Análisis de sobre-representación.

- En el conjunto total de genes: términos relacionados con desórdenes neurológicos.

*“Pervasive development disorder”, “Drug dependance”, “Narcolepsy”, “Neuron to neuron synapse”, etc.*

- En el conjunto de genes directamente relacionados con cáncer de colon: términos relacionados con cáncer, desórdenes neurológicos y regulación del calcio.

*“Gallbladder carcinoma”, “Impulsive Behaviour”, “Presynapse”, “Calcium reabsorption”, “Parathyroid hormone synthesis, secretion and action”, etc.*

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. —————→ 695 genes afectados. Análisis de sobre-representación.

- En el conjunto total de genes: términos relacionados con desórdenes neurológicos.

*“Pervasive development disorder”, “Drug dependance”, “Narcolepsy”, “Neuron to neuron synapse”, etc.*

- En el conjunto de genes directamente relacionados con cáncer de colon: términos relacionados con cáncer, desórdenes neurológicos y regulación del calcio.

*“Gallbladder carcinoma”, “Impulsive Behaviour”, “Presynapse”, “Calcium reabsorption”, “Parathyroid hormone synthesis, secretion and action”, etc.*

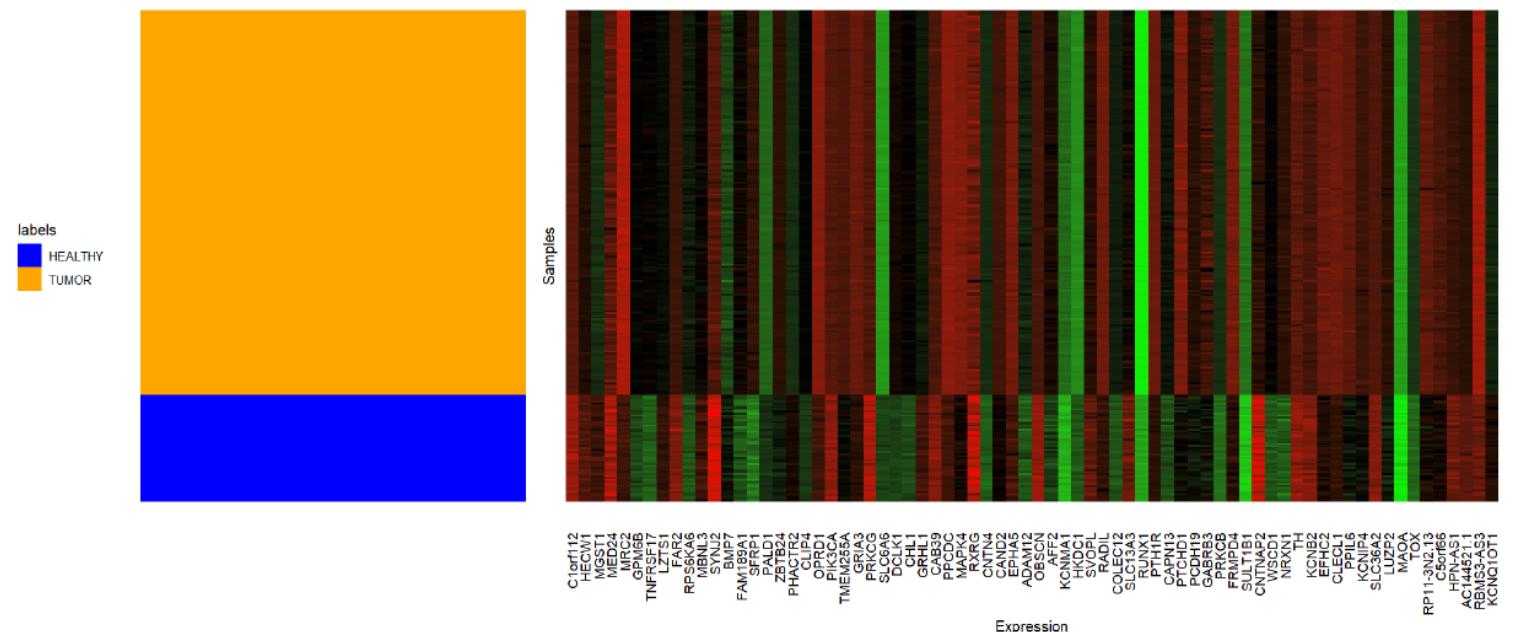
Los desórdenes neurológicos y la hipercalcemia son condiciones altamente prevalentes en pacientes de cáncer de colon [9,10]. Esta prevalencia podría estar originada por algunos de los SNPs génicos que PGMRA utilizó para agrupar pacientes con este diagnóstico.

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. → 695 genes afectados. Análisis de muestras externas RNA-Seq.

- Se identificó expresión diferencial en tejido colorectal canceroso para 72 de los 695 genes afectados. 43 estaban subexpresados y 29 sobreexpresados.



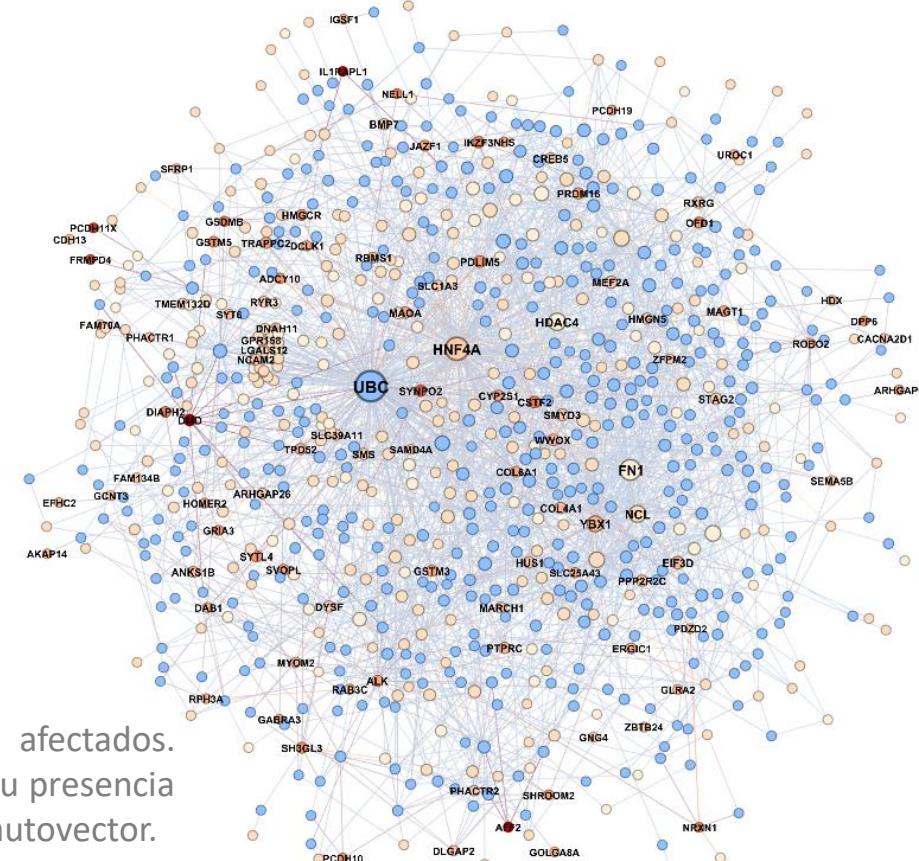
Estos genes podrían tener un papel clave en la enfermedad.

72 genes expresados diferencialmente en cáncer de colon: heatmap de su expresión en las dos clases de muestras de RNA-Seq de tejido colorectal: sanas y tumorosas.

**Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).**

- Análisis de los 2496 SNPs.
    - **972 SNPs génicos.** → **695 genes afectados.**
      - 398 nodos semilla y 452 nodos no-semilla.

## Red de interacciones proteína-proteína.



Red mínima de interacciones proteína-proteína derivada del conjunto total de genes afectados. Nodos no-semilla: azul. Nodos semilla: escala de rojos, con tonos más oscuros indicando su presencia en un mayor número de biclústeres. Mayor tamaño del nodo indica mayor centralidad de autovector.

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

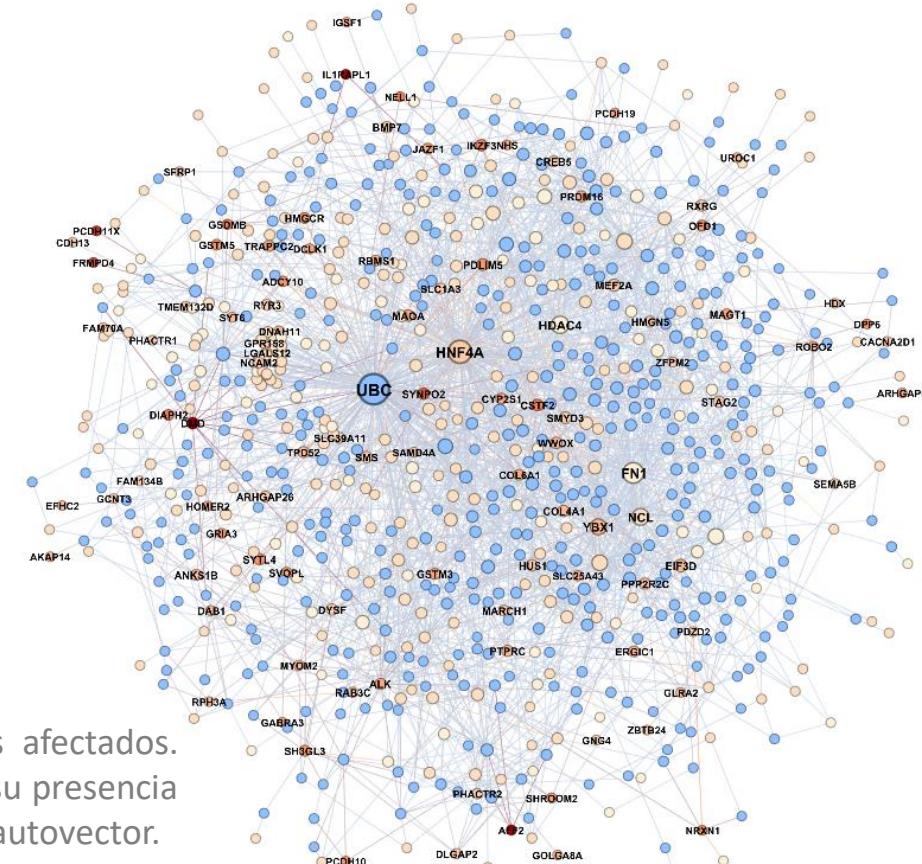
➤ 972 SNPs génicos. → 695 genes afectados.

- 398 nodos semilla y 452 nodos no-semilla.
- El nodo más central (mayor grado, centralidad de autovector y de intermediación): gen UBC (Ubiquitin C). Nodo no-semilla.

Grado medio: 6,18  
Grado UBC: 232

La proteína que codifica lleva a cabo sus funciones mediante la unión con gran variedad de proteínas [11].

Red de interacciones proteína-proteína.



Red mínima de interacciones proteína-proteína derivada del conjunto total de genes afectados. Nodos no-semilla: azul. Nodos semilla: escala de rojos, con tonos más oscuros indicando su presencia en un mayor número de biclústeres. Mayor tamaño del nodo indica mayor centralidad de autovector.

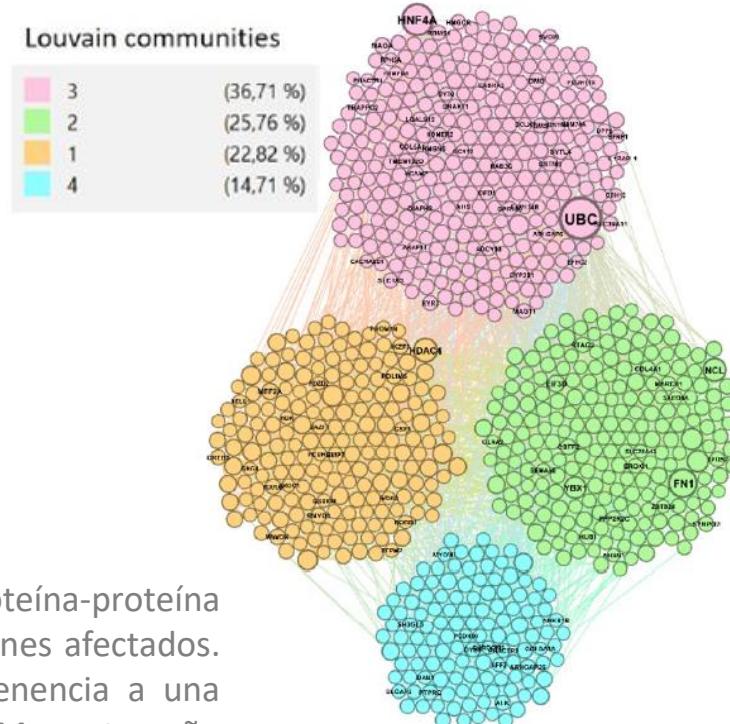
## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. —————→ 695 genes afectados.

Red de interacciones proteína-proteína.

- Louvain (algoritmo de detección de comunidades): partición de la red en 4 comunidades, con una modularidad de 0.388.



Red mínima de interacción proteína-proteína derivada del conjunto total de genes afectados. Cada color de nodo indica pertenencia a una comunidad de Louvain diferente. Mayor tamaño de nodo indica mayor centralidad de autovector.

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

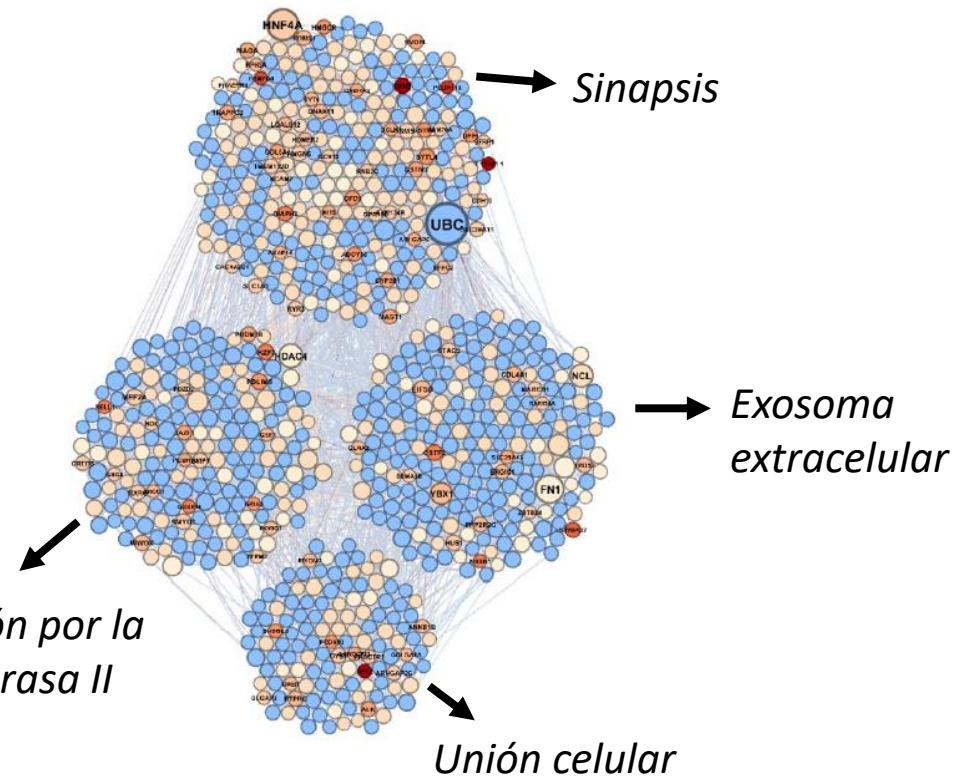
- Análisis de los 2496 SNPs.

➤ 972 SNPs génicos. —————→ 695 genes afectados.

- Louvain (algoritmo de detección de comunidades): partición de la red en 4 comunidades, con una modularidad de 0.388.

Red mínima de interacción proteína-proteína derivada del conjunto total de genes afectados. Nodos no-semilla: azul. Nodos semilla: escala de rojos, con tonos más oscuros indicando su presencia en un mayor número de biclústeres. Mayor tamaño de nodo indica mayor centralidad de autovector. Se muestran términos sobre-representados en cada comunidad.

Red de interacciones proteína-proteína.



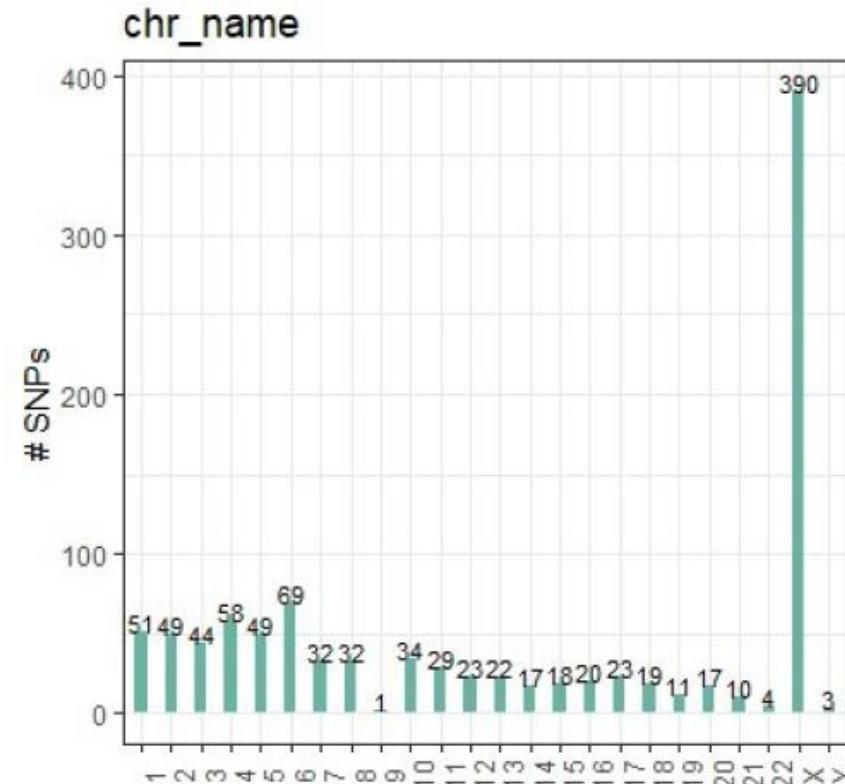
## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.

- **1025 SNPs intergénicos.**

Distribución en cromosomas (info. Ensembl).

- Al igual que para los SNPs génicos, el cromosoma donde se ubicaba mayor número de SNPs intergénicos era el X.



Distribución del atributo “chr\_name” en el conjunto total de SNPs intergénicos.

## Resultados del análisis global (de todos los pacientes y SNPs incluidos en cualquiera de los 16 biclusters).

- Análisis de los 2496 SNPs.
  - **1025 SNPs intergénicos.** Filtrado en base a la probabilidad de tener función regulatoria en el colon.
    - 47 SNPs intergénicos fueron filtrados por tener una probabilidad mínima de 0,25 de tener alguna función regulatoria específica en el colon (o en el intestino o en el intestino grueso), según RegulomeDB. Ejemplos:

9 de los 47 SNPs intergénicos filtrados usando RegulomeDB.

SNP	reg. prob	ranking	chrom	colon	intestine	large intestine
rs503832	0.969	2b	chr1	0.287	0.287	0.287
rs2482806	0.949	5	chr1	0.28	0.28	0.28
rs12023396	1	2a	chr1	0.296	0.296	0.296
rs1902719	1	3a	chr10	0.296	0.296	0.296
rs6584977	0.974	3a	chr10	0.288	0.288	0.288
rs6578958	0.857	2b	chr11	0.253	0.253	0.253
rs12420118	1	5	chr11	0.296	0.296	0.296
rs4405339	1	5	chr11	0.296	0.296	0.296
rs7933981	0.86	3a	chr11	0.254	0.254	0.254

## Resultados del bicluster 16.12.

Características generales en el total de pacientes/SNPs y en el bicluster 16.12.

Bicluster	General and bicluster-specific information														
	Patients					SNPs					Affected Genes				
	# Patients	% Metast.	Mean age	% Females	% Surg.	# SNPs	% Interg.	% Genic	# Genes	% CRC dir.	% CRC ind.	% C. dir.	% C. ind.	% No rel.	% DEGs
All patients	73	34.72	63.38	21.92	72.60	1997	51.33	48.67	695	21.01	59.57	3.60	3.45	12.37	10.36
16.12	5	20.00	44.25	0.00	100.00	57	61.40	38.60	16	31.25	62.50	0.00	0.00	6.25	0.00

## Resultados del bicluster 16.12.

Características generales en el total de pacientes/SNPs y en el bicluster 16.12.

Bicluster	General and bicluster-specific information														
	Patients					SNPs					Affected Genes				
	# Patients	% Metast.	Mean age	% Females	% Surg.	# SNPs	% Interg.	% Genic	# Genes	% CRC dir.	% CRC ind.	% C. dir.	% C. ind.	% No rel.	% DEGs
All patients	73	34.72	63.38	21.92	72.60	1997	51.33	48.67	695	21.01	59.57	3.60	3.45	12.37	10.36
16.12	5	20.00	44.25	0.00	100.00	57	61.40	38.60	16	31.25	62.50	0.00	0.00	6.25	0.00

22 SNPs génicos

La mayoría con MAFs muy altas en el bicluster (todas sobrepasaban 0,5).

Tabla resumen de SNPs génicos en el biclúster 16.12.

Bicluster 16.12 - Genic SNPs											
rs	Gene	Rel	DEG	SNP - Ensembl info.		SNP - Bicluster 16.12 info.					
				chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	
rs10152417	RASGRF1	D.CRC	-	15	intron_variant	LOW	0	0	20	80	0.9
rs2859168	CSTF2	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs12009352	PCDH11X	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs6620925	SYTL4	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs10776699	GSTM5	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs1332018	GSTM3	D.CRC	-	1	5_prime_UTR_variant	LOW	0	0	20	80	0.9
rs12305014	PDE6H	IND.CRC	-	12	intron_variant	LOW	0	0	80	20	0.6
rs2475410	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503357	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503359	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4240130	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs4415478	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs5930459	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs5932901	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4458170	ALK	D.CRC	-	2	intron_variant	LOW	0	0	60	40	0.7
rs222364	MTM1	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4825330	NHS	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs1902957	DMD	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4828954	RP11-40F8.2	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs4970777	RP4-735C1.4	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs13108021	RP11-395F4.1	NO REL.	-	4	intron_variant	LOW	0	0	40	60	0.8
rs10454254	F11-AS1	IND.CRC	-	4	intron_variant	LOW	0	0	80	20	0.6

## Resultados del bicluster 16.12.

Características generales en el total de pacientes/SNPs y en el bicluster 16.12.

Bicluster	General and bicluster-specific information														
	Patients				SNPs				Affected Genes						
	# Patients	% Metast.	Mean age	% Females	% Surg.	# SNPs	% Interg.	% Genic	# Genes	% CRC dir.	% CRC ind.	% C. dir.	% C. ind.	% No rel.	% DEGs
All patients	73	34.72	63.38	21.92	72.60	1997	51.33	48.67	695	21.01	59.57	3.60	3.45	12.37	10.36
16.12	5	20.00	44.25	0.00	100.00	57	61.40	38.60	16	31.25	62.50	0.00	0.00	6.25	0.00

La edad media más baja de todos los biclusters.

22 SNPs génicos  
La mayoría con MAFs muy altas en el bicluster (todas sobrepasaban 0,5).

Tabla resumen de SNPs génicos en el biclúster 16.12.

Bicluster 16.12 - Genic SNPs											
rs	Gene	Rel	DEG	SNP - Ensembl info.		SNP - Bicluster 16.12 info.					
				chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	
rs10152417	RASGRF1	D.CRC	-	15	intron_variant	LOW	0	0	20	80	0.9
rs2859168	CSTF2	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs12009352	PCDH11X	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs6620925	SYTL4	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs10776699	GSTM5	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs1332018	GSTM3	D.CRC	-	1	5_prime_UTR_variant	LOW	0	0	20	80	0.9
rs12305014	PDE6H	IND.CRC	-	12	intron_variant	LOW	0	0	80	20	0.6
rs2475410	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503357	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503359	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4240130	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs4415478	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs5930459	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs5932901	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4458170	ALK	D.CRC	-	2	intron_variant	LOW	0	0	60	40	0.7
rs222364	MTM1	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4825330	NHS	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs1902957	DMD	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4828954	RP11-40F8.2	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs4970777	RP4-735C1.4	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs13108021	RP11-395F4.1	NO REL.	-	4	intron_variant	LOW	0	0	40	60	0.8
rs10454254	F11-AS1	IND.CRC	-	4	intron_variant	LOW	0	0	80	20	0.6

## Resultados del bicluster 16.12.

Hipótesis derivadas de las funciones de los genes afectados.

- Polaridad del epitelio: PCDH11X
- Reorganización de la actina: NHS, DMD, RASGRF1. Además, MTM1 y ALK se relacionan con el tejido muscular.
- Inflamación: GSTM3 y GSTM5.
- Possible relación con TGF-B: DMD y MTM1 están relacionados con el gigantismo, condición causada por grandes niveles de GH (Growth Hormone), hormona que induce TGF-B en los riñones.

**Funciones alteradas en el subtipo molecular de cáncer de colon 4 (CMS4 – Mesenquimal) [6]**

Los pacientes del bicluster 16.12 podrían ser casos de CMS4 y los SNPs del bicluster, orígenes genéticos del mismo.

## Resultados del bicluster 16.12.

Hipótesis derivadas de las funciones de los genes afectados.

Adicionalmente, los pacientes del bicluster 16.12 podrían ser casos de **cáncer de colon asociado a enfermedad inflamatoria intestinal** (Inflammatory Bowel Disease - IBD).

- La ruta del citoesqueleto de actina es la más importante en la progresión de colon sano a colon cancerígeno asociado a IBD [12].
- Los pacientes de cáncer de colon asociado a IBD llegan al diagnóstico aproximadamente 20 años antes que los de cáncer de colon esporádico [12].

## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

3 genes interesantes: número de biclusters en el que aparecen afectados y número de SNPs asociados.

Gen	Nº Biclusters (total: 16)	Nº SNPs
<b>ADAMTS9-AS1</b>	11	1
<b>PTCHD1-AS</b>	13 (1º)	14
<b>AFF2</b>	12	16 (1º)



Los tres estaban afectados en un gran número de biclusters.

Sus SNPs podrían contribuir a características clínicas comunes a una gran proporción de los pacientes de la muestra.

## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

Evidencias externas de implicación en el cáncer de colon.

- **ADAMTS9-AS1** → Ya ha sido descrito como un **marcador pronóstico en el cáncer de colon [13]**.
  - Validación externa de PGMRA.
  - PGMRA parece haber identificado un origen genómico de desregulación del gen.
- **PTCHD1-AS** → No hemos encontrado ninguna evidencia externa.
- **AFF2** → Solo un indicio de posible implicación [14].

Nuestros resultados sí sugieren dicha implicación.

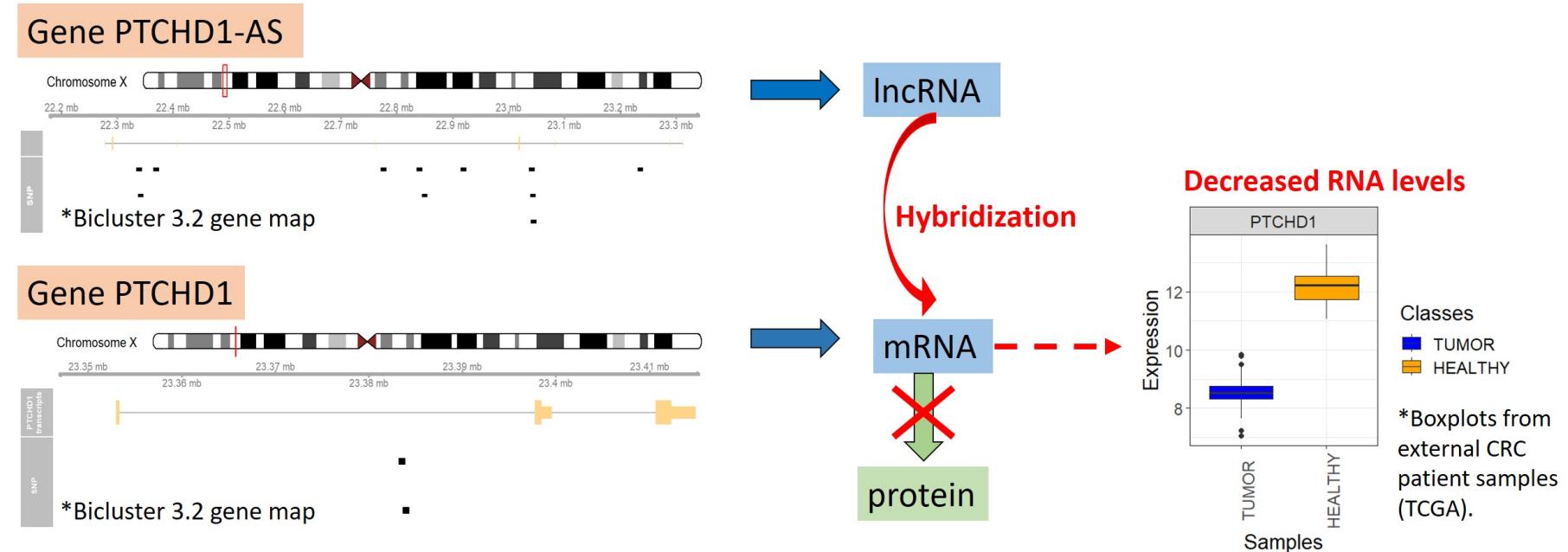
## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

### PTCHD1-AS

- **El gen diana de PTCHD1-AS (PTCHD1) estaba afectado en el bicluster 3.2.** Nuestro análisis de RNA-Seq lo identificó como **subexpresado** en cáncer de colon ( $\log FC = -3,66$ ).

La regulación que ejerce PTCHD1-AS sobre PTCHD1 parece estar intensificada en esta enfermedad. Esto podría estar influido por el impacto de los SNPs en ambos genes.

Regulación de la expresión génica ejercida por PTCHD1-AS sobre PTCHD1.

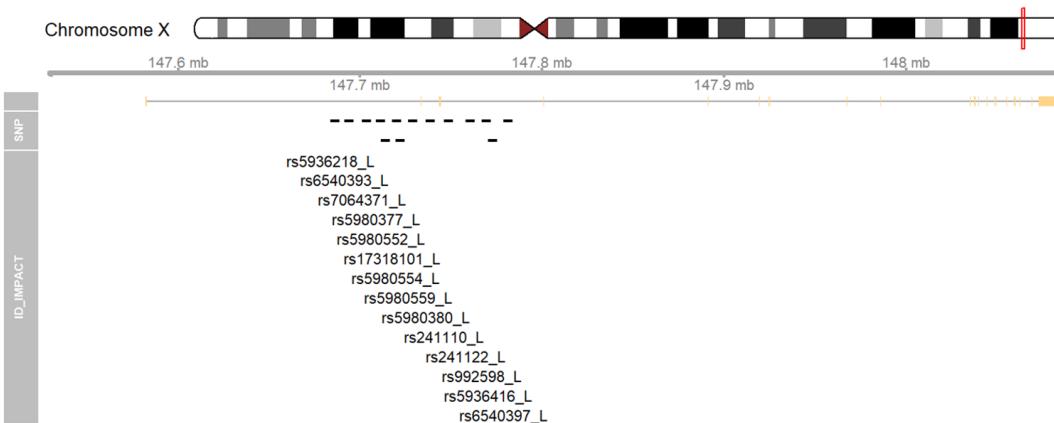


## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

### AFF2

- Sus SNPs presentaban MAFs muy altas en algunos biclusters.

#### Bicluster 14.4 (19 pacientes)



Mapa génico de AFF2 mostrando SNPs del bicluster 14.4.

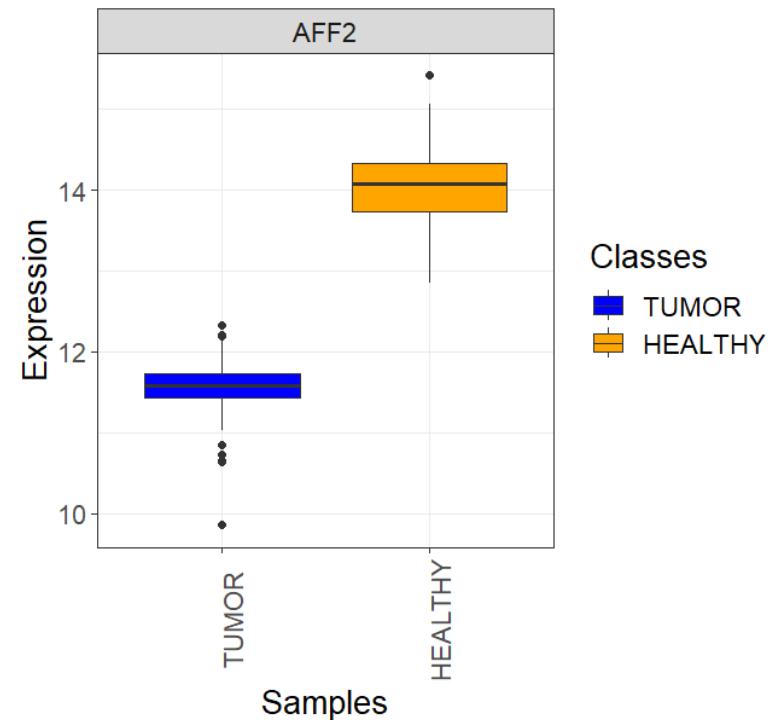
Segmento de la tabla resumen de los SNPs génicos del biclúster 14.4: SNPs del gen AFF2.

Bicluster 14.4 - Genic SNPs											
Gene info.			SNP - Ensembl info.			SNP - Bicluster 14.4 info.					
	Gene	Rel	DEG	chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	Calc_MAF
rs17318101	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63
rs241110	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63
rs241122	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs5936218	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs5936416	AFF2	IND.CRC	DOWN	X	3	LOW	0	31.58	0.00	68.42	0.68
rs5980377	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs5980380	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63
rs5980552	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63
rs5980554	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs5980559	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63
rs6540393	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs6540397	AFF2	IND.CRC	DOWN	X	3	LOW	0	42.11	0.00	57.89	0.58
rs7064371	AFF2	IND.CRC	DOWN	X	3	LOW	0	15.79	0.00	84.21	0.84
rs992598	AFF2	IND.CRC	DOWN	X	3	LOW	0	36.84	0.00	63.16	0.63

## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

### AFF2

- Nuestro análisis de RNA-Seq lo identificó como un gen muy subexpresado en cáncer de colon ( $\log FC = -2,49$ ).



Boxplots de la expresión de AFF2 en las dos clases de muestras de RNA-Seq de tejido colorectal: tumorosas y sanas.

## Genes cuyos SNPs parecen clave en la identificación de biclusters de cáncer de colon.

Pese a que los genes **ADAMTS9-AS1**, **PTCHD1-AS** y **AFF2** parecen ser relevantes para la mayoría de los pacientes, también encontramos diferencias entre los biclusters (respecto al número de SNPs, MAFs, etc.) que podrían contribuir a caracterizar distintos subtipos de cáncer de colon.

Introducción y objetivos

Metodología

Resultados y discusión

Conclusiones

Los resultados de PGMRA sobre tamaños de muestra pequeños mostraron capacidad de descubrir nuevo conocimiento en la caracterización de grupos genéticamente diferentes de pacientes de cáncer de colon.

- El análisis de los conjuntos completos de pacientes y SNPs apuntó hacia los orígenes genéticos de condiciones médicas que ya se sabe que son prevalentes en este cáncer.
- El análisis de biclusters separados reveló nuevos enfoques sobre la compleja interacción entre variantes genéticas, características clínicas y potenciales nuevas perspectivas sobre subtipos de cáncer de colon que deben de probarse experimentalmente en el laboratorio.

Futuras líneas de investigación derivadas de este trabajo supondrían explorar más a fondo las funciones regulatorias de SNPs intergénicos.

## Bibliografía.

1. Luca Del Giacco and Cristina Cattaneo. "Introduction to Genomics". In: Methods in Molecular Biology. Humana Press, Oct. 2011, pp. 79–88. doi: 10.1007/978-1-60327-216-2\_6.
2. Desiree J. Nava Cedeño, María Concepción Alonso Cerezo, Ancor Sanz García, Lorena Vega Zelaya, Juan Gordillo Perdomo, María de Toledo Heras, Jesús Pastor Gómez, Cristina V. Torres Díaz, Paloma Pulido Rivas and Rafael García de Sola. "Asociación entre los polimorfismos genéticos de nucleótido único en genes transportadores ABC con la epilepsia farmacorresistente en la población española". In: Revista de Neurología 75.09 (2022), p. 251. issn: 0210-0010. doi: 10.33588/rn.7509.2022133.
3. Felix Heinrich, Faisal Ramzan, Abirami Rajavel, Armin Otto Schmitt, and Mehmet Gultas. "MIDES: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotype". In: Biology 10.9 (Sept. 2021), p. 921. issn: 2079-7737. doi: 10.3390/biology10090921.
4. Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. "Common SNPs explain a large proportion of the heritability for human height". In: Nature Genetics 42.7 (June 2010), pp. 565–569. doi: 10.1038/ng.608.
5. J. Arnedo, C. del Val, G. A. de Erausquin, R. Romero-Zaliz, D. Svrankic, C. R. Cloninger, and I. Zwir. "PGMRA: a web server for (phenotype x genotype) many-to-many relation analysis in GWAS". In: Nucleic Acids Research 41.W1 (June 2013), W142–W149. doi: 10.1093/nar/gkt496.
6. Ahmed Malki, Rasha Abu ElRuz, Ishita Gupta, Asma Allouch, Semir Vranic, and Ala-Eddin Al Moustafa. "Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements". In: International Journal of Molecular Sciences 22.1 (Dec. 2020), p. 130. doi: 10.3390/ijms22010130.
7. Jialing Zhang, Stephan Stanislaw Späth, Sadie L Marjani, Wengeng Zhang, and Xinghua Pan. "Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment". In: Precision Clinical Medicine 1.1 (June 2018), pp. 29–48. issn: 2516-1571. doi: 10.1093/pcmedi/pby007.
8. John M Henshall, Rachel J Hawken, Sonja Dominik, and William Barendse. "Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples". In: Genetics Selection Evolution 44.1 (Apr. 2012). doi: 10.1186/1297-9686-44-12.
9. Shane Lloyd, David Baraghoshi, Randa Tao, Ignacio Garrido-Laguna, Glynn W. Gilcrease, Jonathan Whisenant, John R. Weis, Courtney Scaife, Thomas B. Pickron, Lyen C. Huang, Marcus M. Monroe, Sarah Abdelaziz, Alison M. Fraser, Ken R. Smith, Vikrant Deshmukh, Michael Newman, Kerry G. Rowe, John Snyder, Niloy J. Samadder, and Mia Hashibe. "Mental Health Disorders are More Common in Colorectal Cancer Survivors and Associated With Decreased Overall Survival". In: American Journal of Clinical Oncology 42.4 (Apr. 2019), pp. 355–362. doi: 10.1097/coc.0000000000000529.

## Bibliografía.

10. Wei Wang, Suyun Yu, Shuai Huang, Rui Deng, Yushi Ding, Yuanyuan Wu, Xiaoman Li, Aiyun Wang, Shijun Wang, Wenxing Chen, and Yin Lu. "A Complex Role for Calcium Signaling in Colorectal Cancer Development and Progression". In: Molecular Cancer Research 17.11 (Nov. 2019), pp. 2145–2153. doi: 10.1158/1541-7786.mcr-19-0429.
11. UBC- GeneCards. url: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=UBC>. [Accessed 05-09-2023].
12. Ziad Kanaan, Motaz Qadan, Maurice Robert Eichenberger, and Susan Galandiuk. "The Actin-Cytoskeleton Pathway and Its Potential Role in Inflammatory Bowel Disease-Associated Human Colorectal Cancer". In: Genetic Testing and Molecular Biomarkers 14.3 (June 2010), pp. 347–353. issn: 1945-0257. doi: 10.1089/gtmb.2009.0197.
13. Wanjing Chen, Qian Tu, Liang Yu, Yanyan Xu, Gang Yu, Benli Jia, Yunsheng Cheng, and Yong Wang. "LncRNA ADAMTS9-AS1, as prognostic marker, promotes cell proliferation and EMT in colorectal cancer". In: Human Cell 33.4 (Sept. 2020), pp. 1133–1141. issn: 1749-0774. doi: 10.1007/s13577-020-00388-w.
14. Kavitha Mukund, Natalia Syulyukina, Sonia Ramamoorthy, and Shankar Subramaniam. "Right and left-sided colon cancers - specificity of molecular mechanisms in tumorigenesis and progression". In: BMC Cancer 20.1 (Apr. 2020). issn: 1471-2407. doi: 10.1186/s12885-020-06784-7.

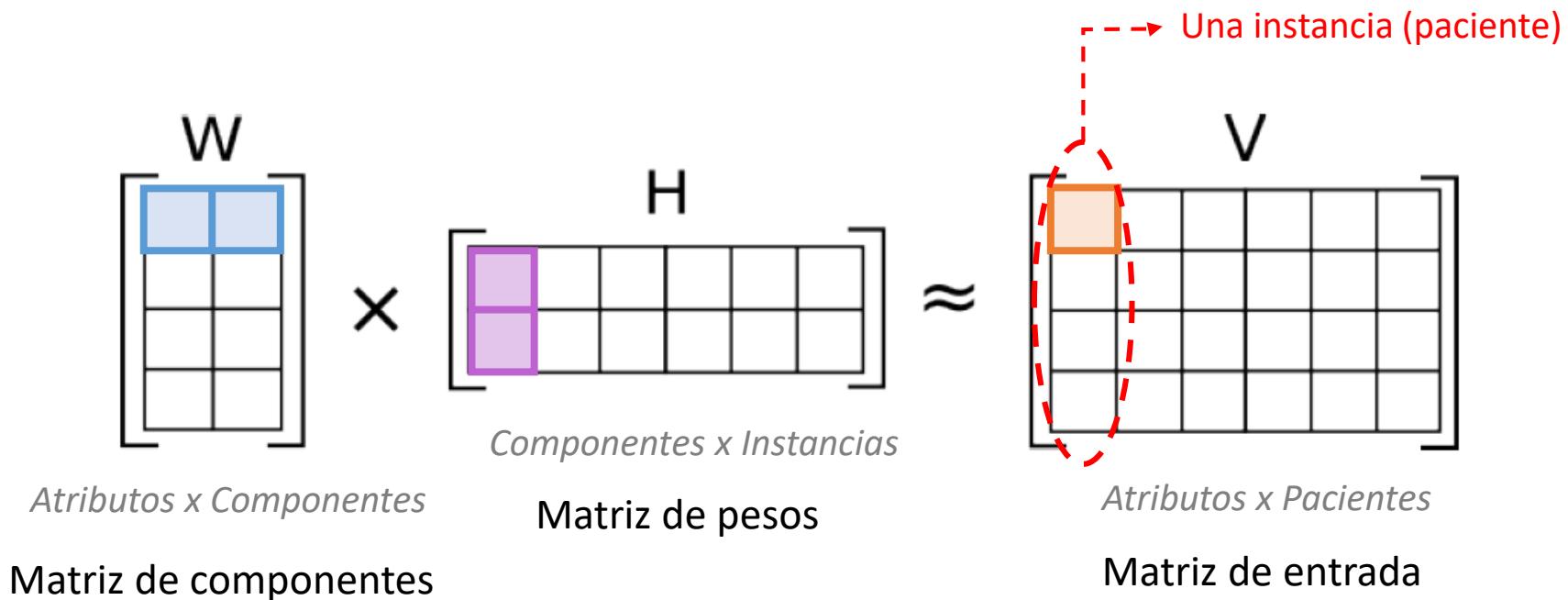


**MUCHAS  
GRACIAS**

## MATERIAL SUPLEMENTARIO

## NMF (factorización de matriz no negativa).

- NMF aproxima la matriz de entrada por el producto de dos matrices no negativas de menor rango. Para ello, encuentra componentes latentes en la matriz de entrada (instancias prototipo).
- NMF aproxima el **valor de una instancia (paciente) en un atributo (SNP)** por una combinación lineal. Esta combinación consta de los **valores prototipo del atributo (uno por componente)** ponderados por los **pesos asociados a cada componente en la instancia**.



**Fuzzy NMF utiliza los valores prototipo y los pesos para identificar los atributos y las instancias más relevantes para cada componente, creando un bicluster por componente.**

**1. Se selecciona la primera componente de W** (primera columna) y se calcula un **umbral**, usando la siguiente formula:

$$\text{Threshold} = \max(W_{i1}) \times (1 - \text{fuzziness})$$

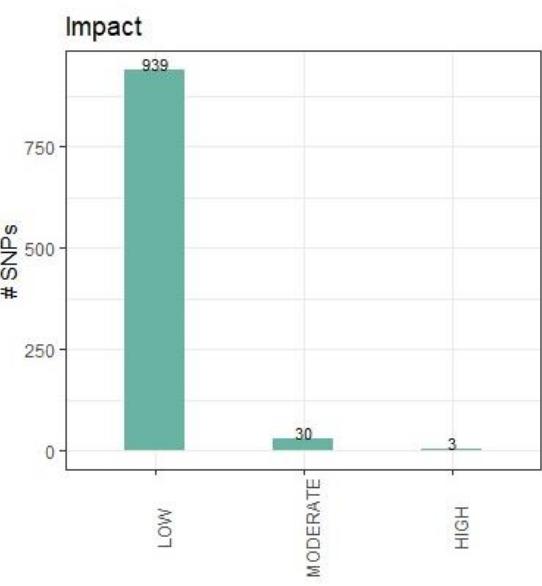
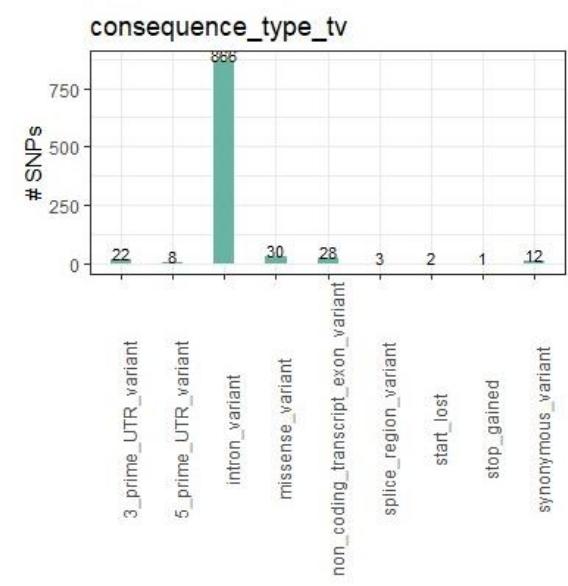
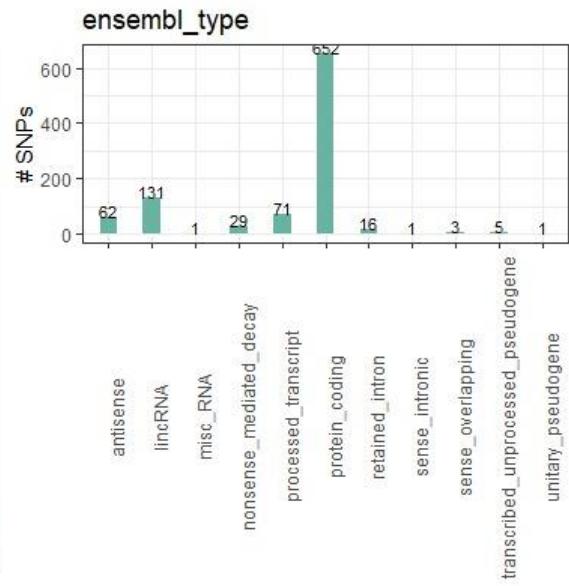
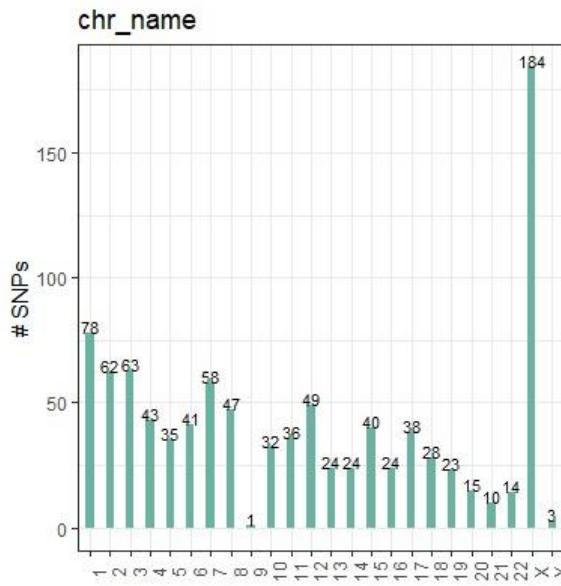
- Todas las filas de W (atributos-SNPs) con valores en la columna 1 superiores o iguales al umbral son seleccionadas.

**2. Se selecciona la primera componente de H** (primera fila) y se calcula un **umbral**, usando la siguiente formula:

$$\text{Threshold} = \max(H_{1i}) \times (1 - \text{fuzziness})$$

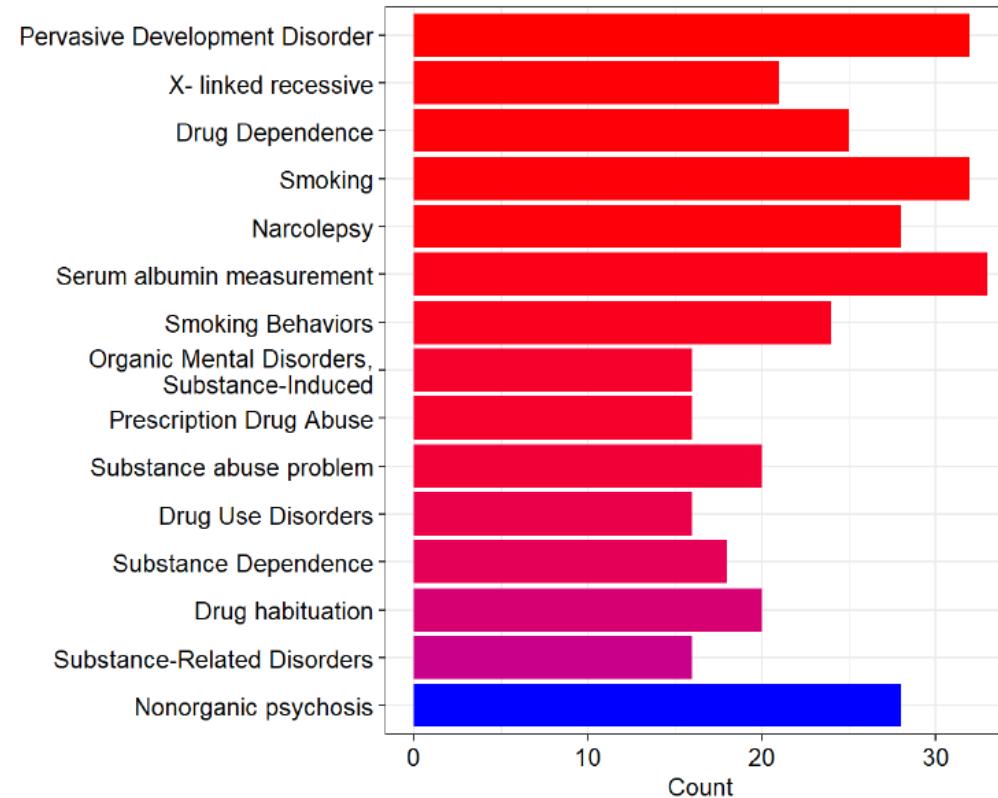
- Todas las columnas de H (instancias–pacientes) con valores en la fila 1 superiores o iguales al umbral son seleccionadas.
3. Las filas y columnas seleccionadas forman el bicluster de la primera componente. El mismo procedimiento se repite para el resto de componentes.

# Distribución de atributos de Ensembl en los SNPs génicos.



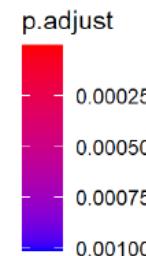
## Análisis de sobre-representación en el conjunto total de genes.

### Enfermedades



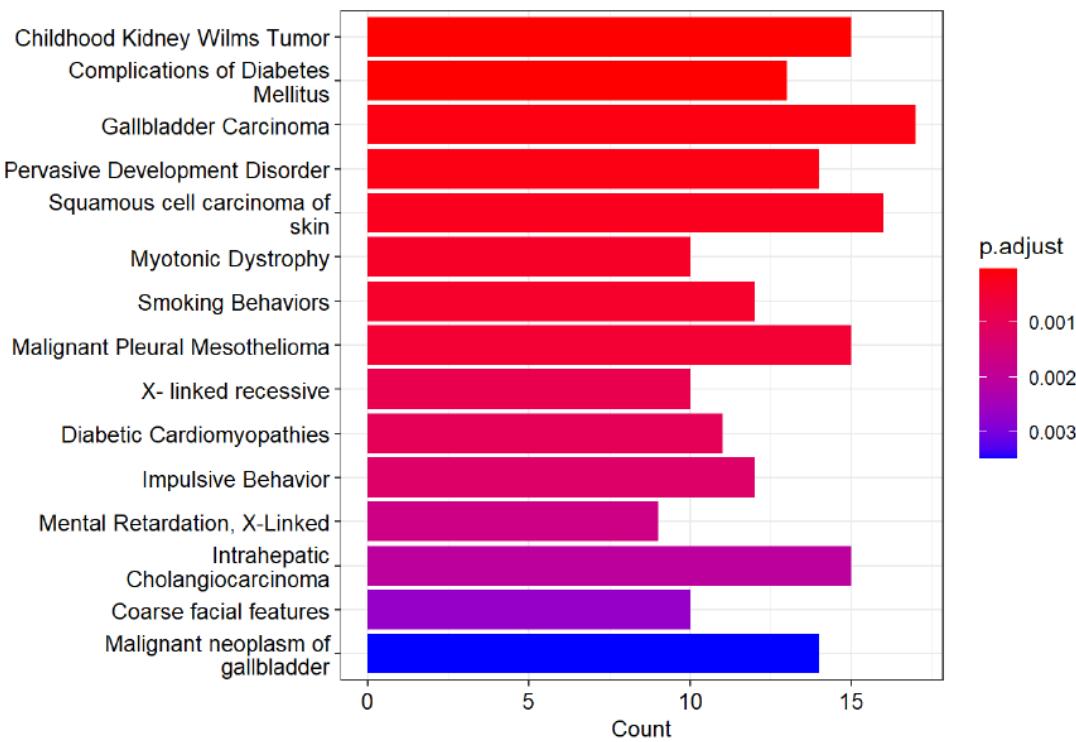
### Componentes celulares

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0098984	neuron to neuron synapse	28/486	7.44E-05	3.13E-02	2.60E-02



# Análisis de sobre-representación en el conjunto de genes directamente relacionados con cáncer de colon.

## Enfermedades



## Componentes celulares

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0098793	presynapse	19/133	1.21E-06	3.53E-04	3.04E-04
GO:0044306	neuron projection terminus	9/133	3.73E-06	1.09E-03	4.68E-04
GO:0099522	cytosolic region	4/133	1.92E-05	5.60E-03	1.60E-03
GO:0043679	axon terminus	7/133	9.66E-05	2.81E-02	6.05E-03
GO:0099524	postsynaptic cytosol	3/133	1.59E-04	4.64E-02	7.98E-03

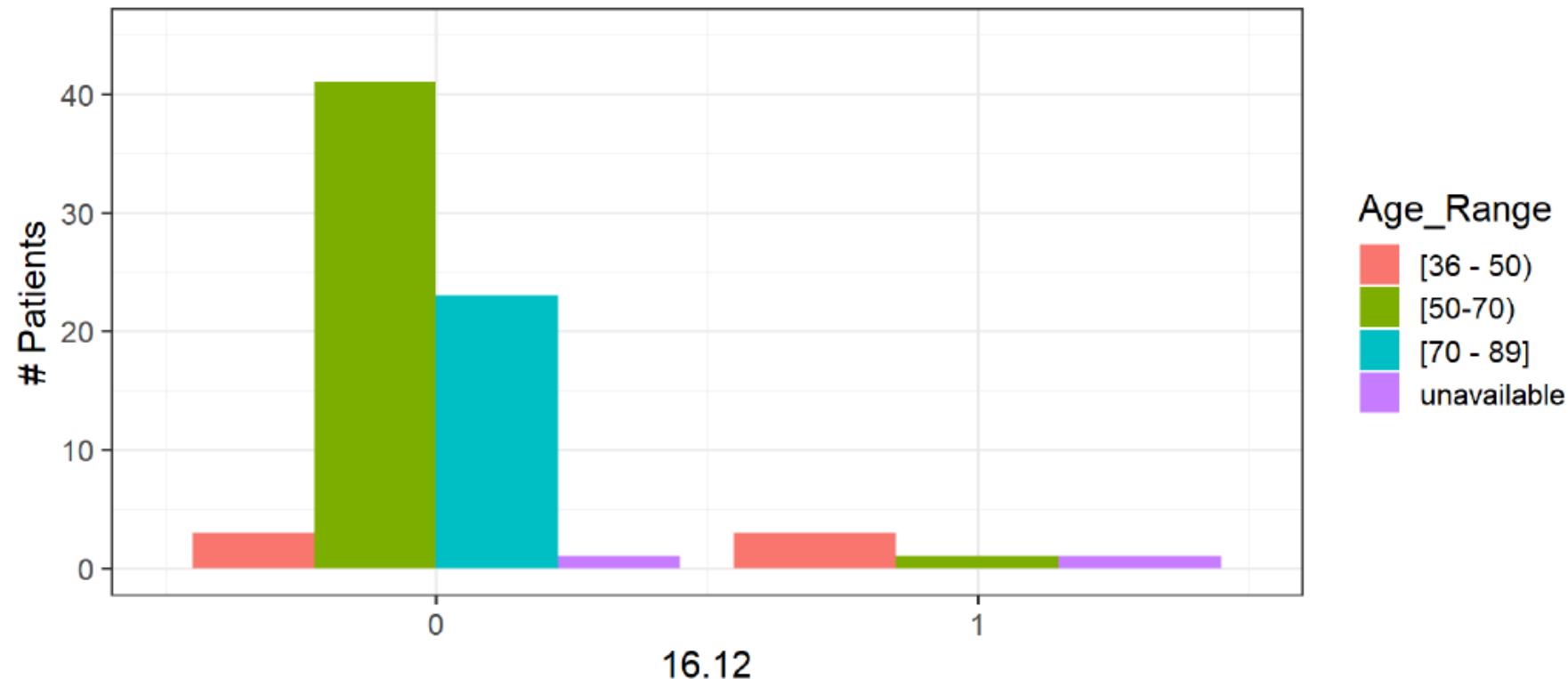
## Rutas moleculares

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
hsa04961	Endocrine and other factor-regulated calcium reabsorption	7/82	5.88E-07	1.29E-04	9.84E-05
hsa04070	Phosphatidylinositol signaling system	7/82	3.50E-05	7.71E-03	2.93E-03
hsa04928	Parathyroid hormone synthesis, secretion and action	7/82	6.20E-05	1.37E-02	3.46E-03
hsa04510	Focal adhesion	9/82	1.23E-04	2.71E-02	5.16E-03

## Tabla resumen de biclusters

General and bicluster-specific information																
Bicluster	Patients					SNPs				Affected Genes						
	# Patients	% Metast.	Mean age	% Females	% Surg.	# SNPs	% Interg.	% Genic	# Genes	% CRC dir.	% CRC ind.	% C. dir.	% C. ind.	% No rel.	% DEGs	
All patients	73	34.72	63.38	21.92	72.60	1997	51.33	48.67	695	21.01	59.57	3.60	3.45	12.37	10.36	
2.2	35	29.41	61.73	8.57	68.57	1471	49.29	50.71	546	21.43	58.79	2.93	4.03	12.82	11.17	
3.2	3	66.67	63.67	0.00	66.67	1439	49.97	50.03	532	21.24	61.47	2.63	3.20	11.47	11.47	
8.4	54	37.74	63.13	12.96	72.22	181	54.70	45.30	58	20.69	56.90	5.17	1.72	15.52	6.90	
10.1	34	29.41	64.26	11.76	76.47	171	49.71	50.29	55	20.00	56.36	7.27	3.64	12.73	12.73	
10.4	3	0.00	52.50	33.33	100.00	110	71.82	28.18	16	25.00	56.25	6.25	6.25	6.25	6.25	
11.5	35	41.18	63.97	2.86	71.43	341	50.73	49.27	133	16.54	63.16	4.51	3.01	12.78	7.52	
12.1	11	54.55	62.82	9.09	81.82	59	52.54	47.46	14	21.43	64.29	7.14	0.00	7.14	7.14	
12.9	8	37.50	61.62	0.00	62.50	64	51.56	48.44	19	10.53	63.16	10.53	0.00	15.79	10.53	
13.8	27	34.62	64.88	0.00	81.48	154	61.04	38.96	42	11.90	66.67	7.14	4.76	9.52	11.90	
13.10	7	42.86	62.71	0.00	71.43	126	50.00	50.00	39	23.08	66.67	2.56	2.56	5.13	17.95	
14.4	19	26.32	65.00	0.00	94.74	65	33.85	66.15	17	11.76	76.47	5.88	0.00	5.88	11.76	
15.6	11	45.45	64.82	9.09	63.64	229	55.46	44.54	75	21.33	58.67	5.33	1.33	13.33	12.00	
16.7	26	40.00	65.04	15.38	65.38	103	57.28	42.72	23	17.39	69.57	4.35	4.35	4.35	17.39	
16.12	5	20.00	44.25	0.00	100.00	57	61.40	38.60	16	31.25	62.50	0.00	0.00	6.25	0.00	
17.7	12	33.33	65.00	16.67	66.67	172	59.88	40.12	49	14.29	67.35	8.16	4.08	6.12	10.20	
17.8	8	62.50	58.14	0.00	75.00	69	82.61	17.39	7	14.29	71.43	14.29	0.00	0.00	14.29	

## Distribución de la edad por rangos en función de la pertenencia al bicluster 16.12.



## Subtipo molecular consenso de cáncer de colon 4 (CMS4 – Mesenquimal).

Subtype	Gene Expression
CMS4 (Mesenchymal)	<ul style="list-style-type: none"><li>Activation of Transforming growth factor-<math>\beta</math> (TGF-<math>\beta</math>)</li><li>Upregulated expression of EMT genes</li><li>Enhanced expression of genes regulating inflammation, matrix remodeling, stromal invasion and angiogenesis</li></ul>

Características biológicas del subtipo CMS4.  
Figura tomada de una fuente externa [6].

- **Transición epitelial-mesenquimal (EMT):** proceso mediante el cual se pierde la polaridad adhesiva de las células cancerosas epiteliales, que cambian a células mesenquimales. Asociado con:
  - Pérdida de cadherinas.
  - Ruptura de las uniones célula-célula.
  - Reorganización citoesqueleto de actina para formar pseudópodos (extensiones citoplasmáticas) invasivos.
- La transición mesenquimal permite a las células invasivas navegar a través de la matriz extracelular y hacia el sistema vascular.
- Otras características del subtipo incluyen la activación de TGF-B (Factor de Crecimiento Transformante Beta) y la sobreexpresión de genes que regulan la inflamación.

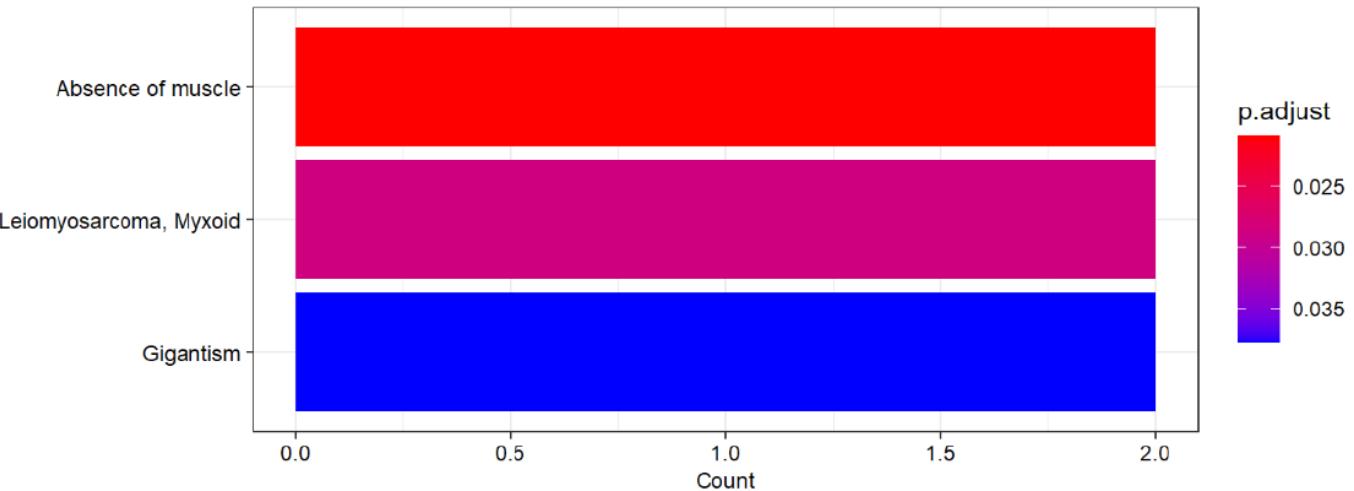
## **Subtipo molecular consenso de cáncer de colon 4 (CMS4 – Mesenquimal).**

### **Funciones génicas del bicluster 16.12 relacionadas con CMS4:**

- Polaridad del epitelio: PCDH11X
- Reorganización de la actina: NHS, DMD, RASGRF1. Además, MTM1 y ALK se relacionan con el tejido muscular.
- Inflamación: GSTM3 y GSTM5.
- Possible relación con TGF-B: DMD y MTM1 están relacionados con el gigantismo, condición causada por grandes niveles de GH (Growth Hormone), hormona que induce TGF-B en los riñones.

## Análisis de sobre-representación en el bicluster 16.12.

### Enfermedades



### Funciones moleculares

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0004364	glutathione transferase activity	2/11	1.66E-04	8.95E-03	5.20E-03
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	2/11	9.15E-04	4.94E-02	1.45E-02

# Cluster-heatmap genotípico de los pacientes y SNPs incluidos en los biclusters 2.2 y/o 8.4.

