Universitat Oberta de Catalunya

uoc.edu

Análisis de alteraciones del número de copias y variantes de un único nucleótido en la evolución de adenoma a adenocarcinoma colorrectal mediante secuenciación completa del genoma

Manuel Lozano García
Máster universitario de Bioinformática y Bioestadística UOC-UB
Área 3: estudio genómico del cáncer y otras enfermedades genéticas

Consultores/as: Laia Bassaganyas Bars (UOC), Jordi Camps Polo (IDIBAPS)
Profesor/a responsable de la asignatura: Ferran Prados Carrasco

5 de enero de 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de alteraciones del número de copias y variantes de un único nucleótido en la evolución de adenoma a adenocarcinoma colorrectal mediante secuenciación completa del genoma*
Nombre del autor:	*Manuel Lozano García*
Nombre del consultor/a:	*Laia Bassaganyas Bars, Jordi Camps Polo*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	*01/2021*
Titulación:	*Máster universitario de Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Área 3: estudio genómico del cáncer y otras enfermedades genéticas*
Idioma del trabajo:	*Castellano*
Palabras clave	*cáncer colorrectal, copy number alteration, single nucleotide variant, genómica humana*

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

Durante la progresión de adenoma (AD) a adenocarcinoma (ADK) en el cáncer colorrectal (CCR), las células adquieren alteraciones genómicas, tanto en el número de copia (CNA) como variantes de un solo nucleótido (SNV), que favorecen la proliferación tumoral. El análisis integral de las CNA y las SNV es necesario para comprender la heterogeneidad intra-tumoral y la progresión de los tumores y para mejorar los tratamientos. La secuenciación completa del genoma (WGS) ha revolucionado el estudio genómico del CCR, pero su alto coste limita la accesibilidad a esta técnica. En este sentido, el WGS de baja cobertura (LP-WGS) ofrece las principales ventajas del WGS pero a un coste más asequible.

El objetivo de este estudio era analizar las diferencias entre regiones de AD y de ADK de 23 muestras de AD avanzados, principalmente en el perfil de CNA, a partir de datos LP-WGS, y adicionalmente en las SNV identificadas mediante secuenciación dirigida.

El análisis de las CNA ha permitido identificar correctamente las principales CNA amplias (a nivel de brazo cromosómico) asociadas a la progresión del CCR. Los datos LP-WGS también han permitido identificar CNA focales (alteraciones localizadas) que pueden estar relacionadas con la evolución AD-ADK. Además,

se ha logrado identificar las SNV, y los genes correspondientes, que tienen un mayor impacto en la carcinogénesis del CCR.

Por lo tanto, el LP-WGS es una técnica coste-efectiva para la detección de CNA que, en combinación con otras técnicas de secuenciación, puede contribuir firmemente a una mejor comprensión de la genética del CCR.

Abstract (in English, 250 words or less):

During the progression from adenoma (AD) to adenocarcinoma (ADK) in colorectal cancer (CRC), cells acquire genomic alterations, both in copy number (CNAs) and single nucleotide variants (SNVs), which favor tumor proliferation. Comprehensive analysis of CNAs and SNVs is necessary to understand intra-tumour heterogeneity and tumour progression and to improve treatments. Whole genome sequencing (WGS) has revolutionised the genomic study of CRC, but its high cost limits the accessibility of this technique. In this sense, low-coverage WGS (LP-WGS) offers the main advantages of WGS but at a more affordable cost.

The aim of this study was to analyse the differences between AD and ADK regions of 23 advanced AD samples, mainly in the CNA profiles, from LP-WGS data, and additionally in the SNVs identified by means of targeted sequencing.

The analysis of the CNAs has allowed the correct identification of the main broad CNAs (at chromosomal arm level) associated with CRC progression. LP-WGS data have also allowed to identify focal CNAs (localised alterations) that may be related to AD-ADK evolution. Furthermore, it has been possible to identify the SNVs, and the corresponding genes, which have a greater impact on the carcinogenesis of CRC.

Therefore, LP-WGS is a cost-effective technique for the detection of CNAs that, in combination with other sequencing techniques, can strongly contribute to a better understanding of CRC genetics.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1. Contexto y justificación del trabajo

El cáncer colorrectal (CCR) fue el tercer tipo de cáncer con mayor incidencia (10,2% del total) y el segundo con mayor mortalidad (9,2% del total) a nivel mundial en 2018, según datos del *Global Cancer Observatory* (Globocan 2018) para hombres y mujeres de todas las edades [1]. En ese mismo año, el CCR fue el tipo de cáncer que tuvo mayor incidencia (14% del total) y el segundo con mayor mortalidad (14% del total) en España, según datos del Observatorio del Cáncer de la Asociación Española Contra el Cáncer (AECC) para hombres y mujeres de entre 50 y 69 años [2].

Los factores de riesgo del CCR se pueden dividir en factores modificables y factores no modificables. De entre los primeros destacan: la obesidad, el sedentarismo, el tabaquismo y el consumo elevado de carnes rojas o procesadas y de alcohol. Factores de riesgo no modificables son: la edad, los antecedentes familiares de CCR, la enfermedad intestinal inflamatoria crónica, la diabetes y algunos trastornos genéticos hereditarios, como el síndrome de Lynch [3].

Según la *American Cancer Society*, la tasa relativa de supervivencia a 5 años del CCR es del 64%. Sin embargo, el pronóstico de la enfermedad varía en función del grado de propagación del cáncer, siendo la tasa relativa de supervivencia a 5 años de más del 90% para cánceres en etapas localizadas, en las que no hay propagación fuera del colon o del recto, y del 14% para cánceres en etapas distantes, en las que el cáncer se ha propagado a distintas partes del cuerpo [3].

Los adenomas (AD) colorrectales, cuya prevalencia aumenta con la edad hasta alcanzar valores superiores al 30% [4]–[6], son el tipo de lesión premaligna más común en el CCR [7]. Sin embargo, solamente alrededor de un 5% de los AD evolucionan hasta formar un CCR [8]. La principal vía en la carcinogénesis colorrectal se conoce como secuencia adenoma-carcinoma (Figura 1) y empieza con la transformación del epitelio colorrectal normal en AD, para después formar el CCR. Este proceso, que puede producirse a lo largo de varios años, está condicionado por la acumulación progresiva de alteraciones genéticas que tienen lugar durante la transformación del AD en adenocarcinoma (ADK) [9].

La predicción de la evolución de los AD a ADK es, por lo tanto, un reto de gran relevancia para la mejora del cribado y la consiguiente reducción de la incidencia del CCR.

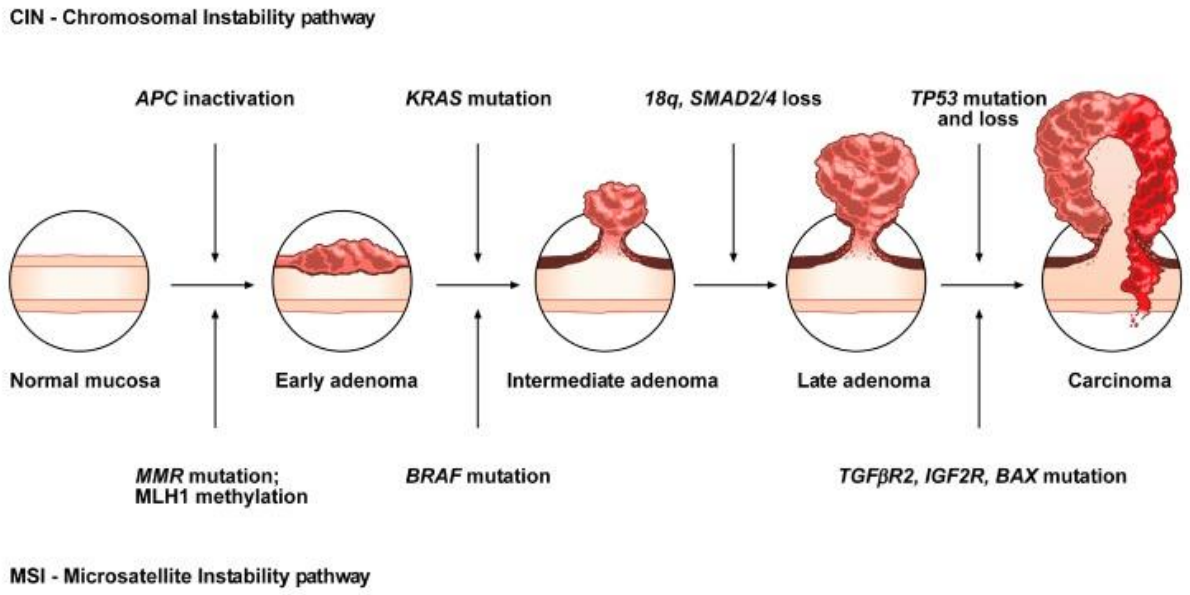

Figura 1. *Modelo genético de la secuencia adenoma-carcinoma. Extraída de [10].*

La mayoría de CCR se desarrollan como consecuencia de alteraciones a nivel cromosómico, lo que se conoce como vía de inestabilidad cromosómica (CIN, del inglés *chromosomal instability*), de acuerdo con el modelo clásico de progresión adenoma-carcinoma (Figura 1). Durante el proceso tumoral, la CIN adquirida por las células tumorales hace que éstas adquieran alteraciones genómicas, conocidas como mutaciones conductoras (*driver*, en inglés), que les proporcionan ventaja selectiva sobre el resto de células, favoreciendo así su proliferación y el proceso de carcinogénesis [11].

Una de las consecuencias de la CIN son las alteraciones del número de copias (CNA, del inglés *copy number alterations*). Estas alteraciones se pueden clasificar en alteraciones amplias (*broad* en inglés), que derivan en aneuploidía numérica (ganancia o pérdida de cromosomas enteros o brazos cromosómicos) y alteraciones focales (*focal* en inglés) (aneuploidía segmentaria), que afectan a regiones específicas y localizadas del genoma de las células cancerosas y consisten en amplificaciones o deleciones locales [12]. La aneuploidía está presente en alrededor del 90% de los tumores sólidos y es uno de los procesos principales que impulsan la progresión tumoral [13], [14]. Las CNA afectan directamente a los genes ubicados en las regiones genómicas alteradas, provocando una desregulación génica característica de cada tipo de tumor. De hecho, se ha demostrado la existencia de una correlación positiva entre las CNA y los niveles de expresión de los genes implicados [15].

En el caso del CCR, son conocidas las ganancias específicas de los cromosomas 7, 8q, 13 y 20q, y las pérdidas de los cromosomas 8p, 17p y 18 [16]. La caracterización de las CNA implicadas en el proceso de carcinogénesis ha permitido relacionar estas alteraciones con los genes afectados y las vías moleculares implicadas [17], [18]. Estudios recientes señalan que el 25% de los AD avanzados presentan CNA asociadas a la progresión del CCR y un 2-4% de los AD no avanzados también presentan este tipo de alteraciones [18]. Por lo tanto, las CNA asociadas a la carcinogénesis podrían utilizarse para

identificar AD de alto riesgo, proporcionando indicadores para el cribado del CCR.

Otro tipo de alteraciones genéticas que, junto con las CNA, tienen un papel importante en el proceso de carcinogénesis son las mutaciones puntuales o variantes de un único nucleótido (SNV, del inglés *single nucleotide variants*). Entre las mutaciones más relevantes en el proceso de carcinogénesis del CCR se encuentran las que afectan a los genes supresores de tumores *TP53*, *APC* y *SMAD4*, o las que afectan al gen *KRAS*, un oncogén que interviene en el crecimiento y la diferenciación celular [19]–[21].

Uno de los principales retos en el tratamiento del CCR, y del cáncer en general, es la heterogeneidad intra-tumoral [20], [22]. La heterogeneidad tumoral se debe a las alteraciones subclonales, es decir, a los distintos patrones que pueden presentar las alteraciones genéticas que favorecen el proceso de carcinogénesis, incluso entre distintas regiones de un mismo tumor, dando lugar a distintas subpoblaciones celulares. Entender la heterogeneidad del CCR, tanto a nivel de CNA como a nivel de SNV, es, por lo tanto, de gran importancia para lograr un diagnóstico y tratamiento exitosos. Sin embargo, el rol de las CNA y las SNV en la transición AD-ADK todavía no se ha determinado completamente [21] y se necesitan estudios más exhaustivos, que integren el análisis de estos dos tipos de alteraciones y permitan conocer mejor cuántos y qué genes están implicados en el proceso de carcinogénesis y cómo varían los patrones de estas alteraciones durante la evolución del tumor.

Tradicionalmente, los estudios de genética del cáncer han utilizado técnicas de hibridación, como la hibridación fluorescente *in situ* (FISH, del inglés *fluorescence in situ hybridization*), o tecnologías basadas en arrays, como la hibridación genómica comparativa (CGH, del inglés *comparative genomic hybridization*) o los arrays de SNP (del inglés *single nucleotide polimorfism*). No obstante, el desarrollo y la evolución durante los últimos años de las tecnologías de secuenciación masiva (NGS, del inglés *next generation sequencing*) han permitido estudiar la genética del cáncer de una forma más completa, fácil y rápida que con las técnicas anteriores. Estas tecnologías permiten detectar tanto mutaciones puntuales como CNA con una elevada resolución y precisión. Sin embargo, el principal inconveniente de estas técnicas es su elevado coste, especialmente en el caso de la secuenciación completa del genoma (WGS, del inglés *whole genome sequencing*).

Recientemente, se ha empezado a utilizar la WGS de baja cobertura (LP-WGS, del inglés *low-pass whole genome sequencing*) [23]–[25] como una alternativa más económica que la WGS original. La técnica LP-WGS permite igualmente realizar estudios genéticos con una buena precisión y abarcando todo el genoma pero a un coste más asequible, por lo que el uso de esta técnica para la detección de CNA es cada vez más frecuente. No obstante, debido a la baja cobertura empleada, la LP-WGS no puede utilizarse para la detección de SNV, para lo cual suele utilizarse la secuenciación de paneles de genes (*targeted sequencing* en inglés). La combinación de diferentes técnicas de secuenciación permite la detección de distintos tipos de alteraciones genéticas y, por

consiguiente, aportaría una visión más completa y una mejor comprensión de la genética del CCR.

1.2. Objetivos del trabajo

1.2.1. Objetivo principal

El objetivo principal de este estudio es analizar las alteraciones genéticas, especialmente CNA y SNV, que aparecen en la transición de AD a ADK colorrectal, a partir de datos de secuenciación masiva.

1.2.2. Objetivos específicos

- Analizar las posibles diferencias en el número y tipo de CNA y sus perfiles genómicos entre regiones de AD y regiones de ADK en muestras de AD avanzados, a partir de datos genómicos obtenidos mediante LP-WGS.
- Comparar los resultados del análisis de CNA obtenidos con los datos LP-WGS con los obtenidos previamente en las mismas muestras mediante FISH.
- Analizar las posibles diferencias en el número y tipo de SNV entre las regiones de AD y las regiones de ADK de los AD avanzados estudiados, a partir de datos genómicos obtenidos mediante *targeted sequencing*.
- Analizar las posibles diferencias en la clonalidad de las CNA encontradas entre las regiones de AD y las regiones de ADK de los AD avanzados estudiados.

1.3. Enfoque y método seguido

1.3.1. Análisis de CNA

El análisis de CNA se ha realizado a partir de datos LP-WGS obtenidos de regiones de AD y regiones de ADK en muestras de AD avanzados.

El WGS es una de las principales estrategias de secuenciación NGS utilizadas en el diagnóstico del cáncer. La evolución de la secuenciación NGS durante los últimos años ha permitido entender mejor, y de forma más sencilla y rápida, la genética del cáncer. Las plataformas de secuenciación de *Illumina* son las más utilizadas hoy en día. Estas plataformas están basadas en la secuenciación por síntesis, cuya secuencia de pasos se puede observar en la Figura 2. De manera resumida, el proceso consta de los siguientes pasos:

- Preparación de la **librería** de la muestra, fragmentando la muestra de ácido desoxirribonucleico (ADN) y uniendo unos ligandos específicos a ambos extremos de los fragmentos.
- **Hibridación** de la librería en una celda de flujo.
- **Amplificación** de los fragmentos mediante PCR (del inglés *polymerase chain reaction*) de puente, formando clústeres clonales.
- **Secuenciación** mediante reactivos que incluyen nucleótidos marcados con fluorescencia. Tras incorporar la primera base, se captura una imagen de la celda y se registra la intensidad y el color de emisión de cada clúster, que sirven

para identificar la base. Se repiten *n* ciclos de secuenciación para obtener *reads* de *n* bases.

- **Alineación** de los *reads* al genoma de referencia.

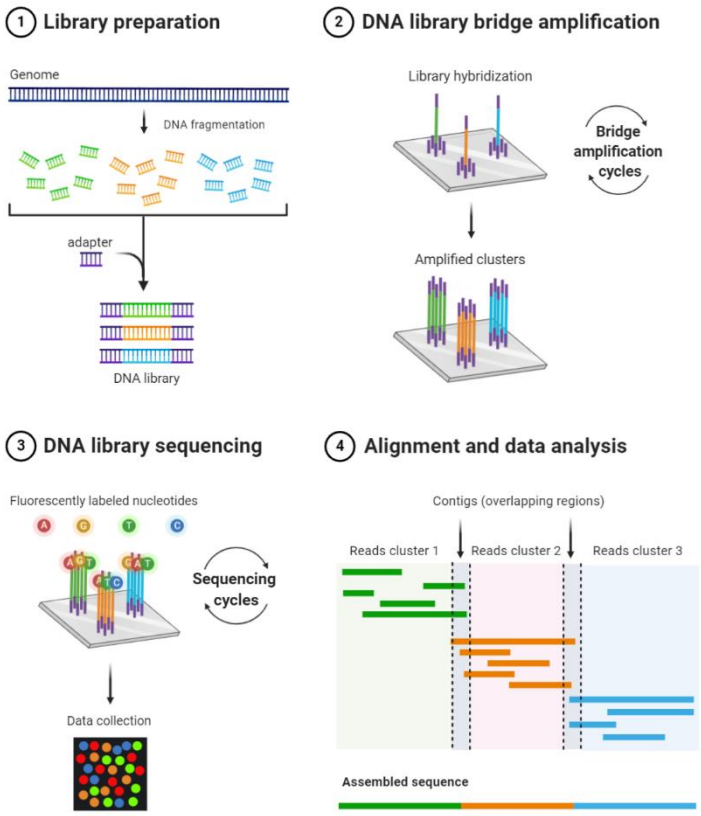

Figura 2. *Secuencia de pasos en la secuenciación NGS. Created with BioRender.*

La secuenciación NGS presenta varias ventajas respecto a las técnicas basadas en microarrays, las cuáles han sido comúnmente utilizadas para el análisis de CNA. En primer lugar, la secuenciación NGS proporciona una alta resolución (a nivel de nucleótido), lo cual permite detectar con más precisión y sensibilidad las alteraciones genéticas, tanto mutaciones puntuales, como inserciones, deleciones o CNA. Además, esta técnica permite la identificación de poblaciones celulares subclonales presentes en distintas proporciones en la muestra tumoral. En segundo lugar, los microarrays están basados en la hibridación de fragmentos de ADN de una muestra con un conjunto de secuencias diana previamente seleccionadas. La secuenciación NGS, sin embargo, consiste en determinar de manera simultánea la secuencia de nucleótidos de millones de fragmentos de ADN de una muestra, sin necesidad de tener conocimiento de la secuencia *a priori*. Por lo tanto, la secuenciación NGS, a diferencia de los microarrays, es una técnica mucho más flexible, pues no requiere de un rediseño cada vez que se quiere estudiar una nueva alteración, a la vez que permite descubrir nuevas alteraciones genéticas.

De entre las técnicas NGS, el WGS ofrece la ventaja añadida de abarcar todo el genoma, con lo que permite identificar un amplio abanico de posibles

variantes genéticas, tanto en regiones codificantes como no codificantes. Sin embargo, el WGS es una técnica generalmente poco accesible debido a su elevado coste, a pesar de que éste ha bajado drásticamente en la última década [26]. Uno de los factores más determinantes en el coste del WGS es la profundidad de cobertura, aparte del tamaño de la muestra y de la tecnología utilizada. La profundidad de cobertura se suele calcular en ventanas genómicas no solapadas de longitud fija y se define como el número promedio de *reads* alineados dentro de cada ventana. Los requisitos en cuanto a profundidad de cobertura dependen de cada aplicación y se debe elegir teniendo en cuenta su influencia no solamente en el coste sino también en la precisión de la secuenciación, de manera que cuanto mayor es la profundidad de cobertura, mayor es la fiabilidad de los resultados biológicos derivados de los datos de secuenciación. Por lo tanto, es importante elegir la profundidad de cobertura que proporcione un equilibrio entre precisión y coste.

El LP-WGS consiste en aplicar el WGS pero con una profundidad de cobertura muy baja, con lo que el coste de esta técnica es menor. A pesar de que la baja profundidad de cobertura hace que el LP-WGS no sea adecuado para la detección de mutaciones puntuales, se ha demostrado que esta técnica supone una alternativa coste-efectiva y con mayor resolución que los arrays en la detección de CNA, permitiendo identificar cambios en el número de copias con precisión [27]–[31].

Las regiones de AD y las regiones de ADK de las mismas muestras de AD avanzados que se han analizado mediante datos LP-WGS en este estudio, se habían analizado previamente a partir de datos FISH [32], lo cual ha permitido llevar a cabo la comparación entre ambas técnicas.

1.3.2. Análisis de SNV y clonalidad

El análisis de mutaciones puntuales, como las SNV, requiere de una secuenciación con una profundidad de cobertura mucho mayor que en el caso de las CNA para evitar la detección excesiva de falsos positivos. Es por ello que para analizar SNV se han utilizado datos de *targeted sequencing* obtenidos de regiones de AD y regiones de ADK en las mismas muestras de AD avanzados.

El *targeted sequencing*, secuenciación de paneles de genes o secuenciación dirigida es una técnica de secuenciación NGS que consiste en aislar y secuenciar un subconjunto de genes o regiones genómicas de interés con una profundidad de cobertura muy alta. Esta técnica se utiliza para dirigir la secuenciación a una determinada patología y permite detectar, además de variantes puntuales, aquellas que son raras o de muy baja frecuencia (variantes subclonales), que son difíciles y caras de detectar mediante otras técnicas. Sin embargo, a diferencia del WGS, el *targeted sequencing* no permite identificar nuevas alteraciones ya que la secuenciación está dirigida a regiones conocidas.

1.3.3. Herramientas de análisis

Para llevar a cabo el análisis de CNA, se propone el uso de CNApp. Se trata de una aplicación web desarrollada por un equipo de investigadores del *Institut d'Investigacions Biomèdiques August Pi i Sunyer* (IDIBAPS) de Barcelona [33]:

https://tools.idibaps.org/CNApp/cnapp_tool.html

El uso de esta herramienta viene dado por el acceso fácil y rápido a este recurso y por el asesoramiento que pueden proporcionar los consultores de este estudio, que contribuyeron a su desarrollo. CNApp permite llevar a cabo un análisis amplio e integrativo de las CNA a través de una interfaz de usuario sencilla y de fácil manejo.

De manera alternativa, se propone explorar el uso de ichorCNA, una herramienta desarrollada por investigadores del *Broad Institute* de Cambridge, Massachusetts, para la predicción y estimación de CNA y su estado clonal en el estudio de tumores [28]:

https://github.com/broadinstitute/ichorCNA

Esta herramienta, además de estar desarrollada por un centro de referencia a nivel mundial en el uso de datos genómicos para el estudio de distintas enfermedades humanas, entre ellas el cáncer, está desarrollada especialmente para analizar CNA en muestras con porcentajes bajos de células tumorales y/o secuenciadas mediante la técnica LP-WGS, con lo que es una herramienta idónea para el tipo de análisis que se plantea en este trabajo. Además, con ichorCNA se puede realizar un análisis más completo ya que, partiendo de los datos en crudo (archivos BAM) y siguiendo la *pipeline* de análisis que proporciona el programa, esta herramienta permite llevar a cabo tanto el análisis de CNA como el análisis de clonalidad.

De forma complementaria, se propone el uso de *scripts* en el lenguaje de programación R para completar el análisis de CNA y de SNV.

1.4. Planificación del trabajo

En la Figura 3 se muestra el diagrama de Gantt de este trabajo.

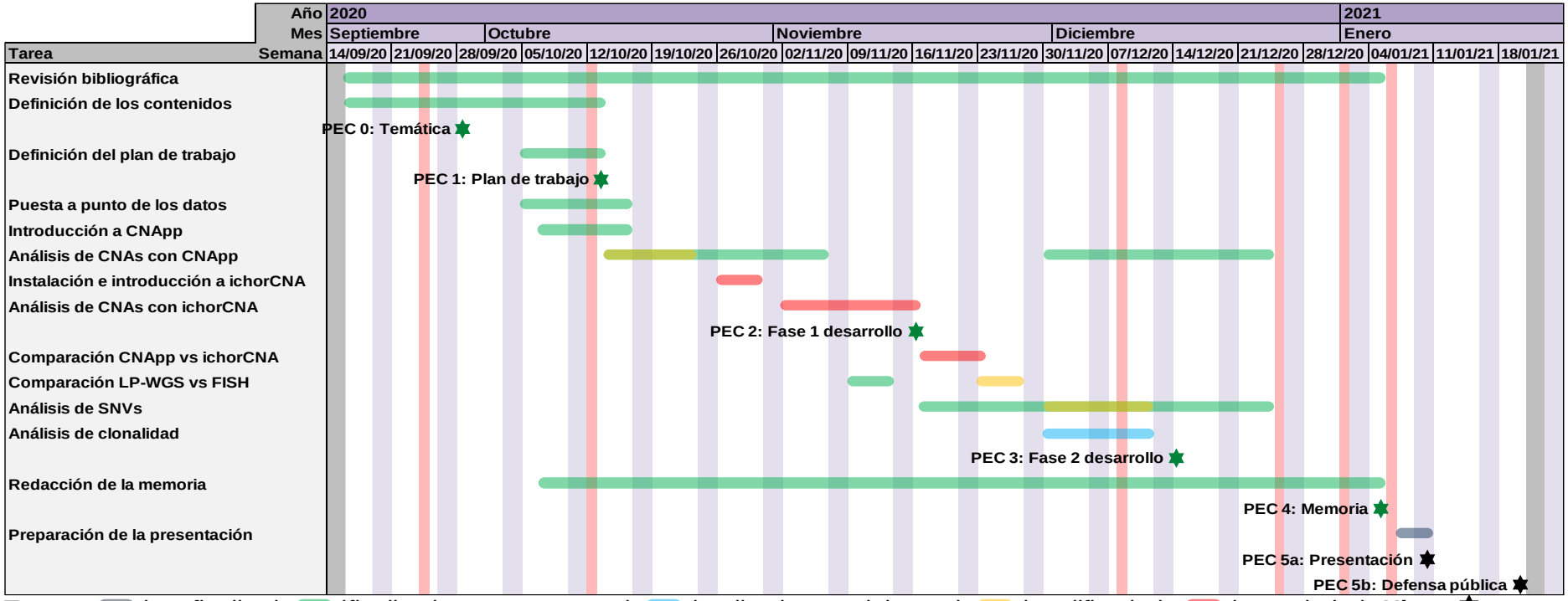

Figura 3. *Diagrama de Gantt del trabajo.*

Al inicio de este trabajo, se hizo un análisis de los principales riesgos del estudio que podían llevar a desviaciones tanto en las tareas programadas como en la temporización de las mismas. Los riesgos detectados así como las correspondientes acciones de mitigación propuestas pueden observarse en la siguiente tabla.

Tabla 1. *Análisis de riesgos*

Riesgo	Atenuación
Carencia de experiencia del estudiante en el tema de estudio	Se tomará una actitud proactiva, haciendo especial hincapié en la revisión bibliográfica y recibiendo asesoramiento frecuente por parte de los consultores y otros investigadores con experiencia, mediante reuniones periódicas
No disponibilidad de los archivos de datos en formato BAM	Se priorizará el análisis de CNA con la herramienta CNApp y se dedicará más tiempo al análisis de SNV
Problemas con la herramienta ichorCNA	Se priorizará el análisis de CNA con la herramienta CNApp y se dedicará más tiempo al análisis de SNV
Reducción del tiempo de trabajo por factores externos (pandemia, salud, otros contratiempos de índole personal)	Se priorizará el análisis de CNA y SNV, definiendo los perfiles genómicos de las muestras, y se empezará, si es posible, el análisis de clonalidad hasta donde dé tiempo
Conclusiones erróneas en el análisis de los datos	Se utilizarán como resultados de referencia los obtenidos anteriormente en la misma muestra utilizando la técnica FISH. El análisis de clonalidad se realizará, como mínimo, a partir de los resultados de referencia

A continuación se detallan las tareas realizadas y agrupadas según los hitos de este estudio:

- ***Hito*: Entrega de la PEC 0**
  - ***Tareas previstas***: propuesta de título y palabras clave, descripción de la temática escogida y de la problemática a resolver, definición orientativa de los objetivos y revisión bibliográfica.
  - ***Estado***: todas las tareas se finalizaron a tiempo y de manera satisfactoria.
  - ***Entregables***: documento de la PEC 0.
- ***Hito*: Entrega de la PEC 1**
  - ***Tareas previstas***: definición clara de los objetivos generales y específicos del trabajo, definición y temporización de las tareas e hitos del trabajo, análisis de riesgos, descripción general del enfoque y método a seguir.
  - ***Estado***: todas las tareas se finalizaron a tiempo y de manera satisfactoria.
  - ***Entregables***: documento de la PEC 1.
- ***Hito*: Entrega de la PEC 2**
  - ***Tareas previstas***: puesta a punto de los datos LP-WGS, familiarización con la herramienta CNApp, análisis de CNA con CNApp, instalación y familiarización con la herramienta ichorCNA y análisis de CNA con ichorCNA.

- ***Tareas no previstas***: comparación de los resultados LP-WGS y FISH.
- ***Cambios realizados***: ha habido dificultades para conseguir los archivos BAM necesarios para realizar el análisis de CNA con ichorCNA. Siguiendo las acciones de mitigación propuestas en el análisis de riesgos, se ha priorizado y extendido el análisis de CNA con CNApp, se ha cancelado el análisis de CNA con ichorCNA y se ha avanzado la comparación de los resultados LP-WGS y FISH.
- ***Estado***: se está llevando a cabo el análisis de CNA con CNApp y R. Falta por completar el análisis de enriquecimiento de conjuntos de genes. Se han comparado los resultados LP-WGS y FISH.
- ***Entregables***: documento de la PEC 2.

- ***Hito*: Entrega de la PEC 3**
  - ***Tareas previstas***: análisis de SNV y de clonalidad a partir de los datos de *targeted sequencing*.
  - ***Tareas no previstas***: análisis de CNA.
  - ***Cambios realizados***: se ha extendido el tiempo dedicado al análisis de CNA a partir de los datos LP-WGS, con el objetivo de completar el análisis de enriquecimiento de conjuntos de genes, analizar posibles coexistencias de CNA en las muestras de estudio y refinar el análisis de las *focal* CNA. Se ha extendido el tiempo dedicado al análisis de SNV a partir de los datos de *targeted sequencing*.
  - ***Estado***: se está completando el análisis de CNA. También se está trabajando en el análisis de SNV y se ha empezado el análisis de clonalidad (se han mirado los perfiles de las frecuencias alélicas de las SNV). Se ha decidido no proseguir con esta última parte (clonalidad), debido a los contratiempos en la obtención de los archivos BAM y a la falta de fluidez en el asesoramiento requerido para esta línea de trabajo.
  - ***Entregables***: documento de la PEC 3.

- ***Hito*: Entrega de la PEC 4**
  - ***Tareas previstas***: redacción de la memoria.
  - ***Tareas no previstas***: análisis de CNA y análisis de SNV.
  - ***Cambios realizados***: se ha extendido el tiempo dedicado tanto al análisis de CNA como al análisis de SNV.
  - ***Estado***: se ha finalizado correctamente tanto el análisis de CNA como de SNV. También se ha completado la memoria del proyecto.
  - ***Entregables***: memoria del trabajo junto con los códigos R y los archivos de datos necesarios para la replicación de los resultados de este estudio.

- ***Hito*: Entrega de la PEC 5a y 5b**
  - ***Tareas previstas***: preparación de la presentación y defensa del proyecto.
  - ***Estado***: en curso.
  - ***Entregables***: presentación en formato PowerPoint para la defensa del proyecto.

1.5. Breve sumario de productos obtenidos

Los productos obtenidos de la realización de este trabajo son los siguientes:

- **Memoria del trabajo**: incluye el plan de trabajo y una descripción detallada de la metodología seguida y los resultados obtenidos del análisis de *broad* CNA y *focal* CNA y del análisis de SNV en regiones de AD y regiones de ADK de muestras de AD avanzados.
- **Códigos R** y **archivos de datos**: incluye todo el material necesario para replicar los resultados de este estudio. En el anexo A se puede consultar un listado detallado de los archivos que se adjuntan a esta memoria.

1.6. Breve descripción de los otros capítulos de la memoria

El **capítulo 2** está dedicado al análisis de CNA a partir de datos LP-WGS. El capítulo está dividido en cuatro partes. La primera parte contiene la descripción de los datos utilizados para el análisis de CNA. La segunda parte contiene una explicación detallada de la metodología seguida. La tercera parte contiene las figuras y la explicación detallada de los resultados del análisis. Esta tercera parte se divide, a su vez, en dos subsecciones, una para las *broad* CNA y otra para las *focal* CNA. Por último, la cuarta parte corresponde a la discusión de los resultados.

El **capítulo 3** está dedicado al análisis de SNV a partir de datos de *targeted sequencing*. La organización de este capítulo es similar a la del capítulo 2: descripción de los datos, metodología, resultados y discusión.

Por último, el **capítulo 4** corresponde a las conclusiones del estudio, en las que se destacan los principales resultados, las dificultades encontradas a lo largo del estudio, las posibles mejoras y líneas de trabajo futuras, así como una opinión personal del autor sobre la realización de este trabajo.

2. Análisis de CNA

2.1. Descripción de los datos

Se dispone de los datos LP-WGS de un total de 23 AD avanzados (tabla 2). De éstos, se han obtenido resultados de las regiones de AD en 22 muestras y de las regiones de ADK en 19 muestras. Así, en un total de 18 muestras se tienen datos tanto de la región AD cómo de la región ADK de la misma lesión tumoral.

Tabla 2. *Muestras de AD avanzados utilizadas en el análisis de CNA*

ID	Muestra AD	Muestra ADK
1	●	●
2	●	
3	●	
5	●	●
6	●	●
7	●	●
8	●	
9	●	●
10	●	●
12	●	●
13	●	●
14		●
15	●	●
16	●	●
19	●	●
20	●	●
21	●	●
22	●	●
23	●	●
24	●	●
25	●	●
26	●	●
27	●	

2.2. Metodología

Una de las maneras de identificar CNA a partir de datos LP-WGS, aplicable en general a datos WGS, es midiendo el número de *reads* alineados a un genoma de referencia [18], [34], [35]. Suponiendo que la secuenciación es uniforme, el número de *reads* mapeados en una cierta región se espera que sea proporcional al número de veces que esa región aparece en la muestra analizada. Por lo tanto, el número de copia de una región genómica se puede estimar contando el número de *reads* (*read counts, RC*) alineados a esa región.

La figura 4 muestra el *pipeline* de análisis utilizado para el estudio de CNA en este trabajo.

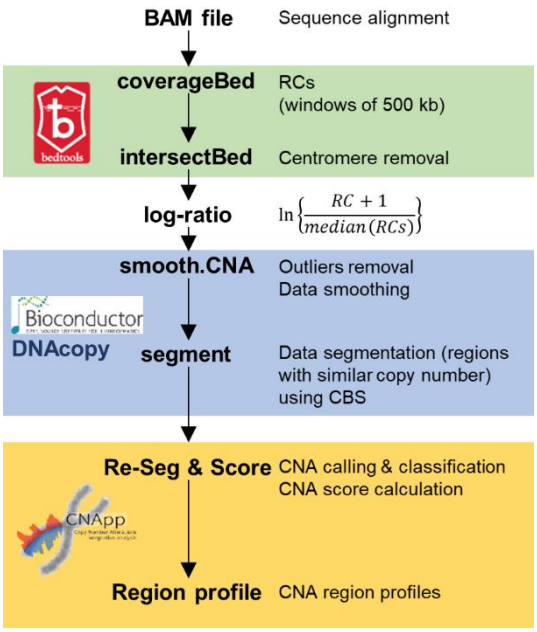

Figura 4. *Pipeline de análisis de CNA. BAM: Binary Alignment Map; CBS: Circular Binary Segmentation; CNA: Copy Number Alteration.*

Los datos LP-WGS disponibles inicialmente en este trabajo habían sido previamente procesados con los dos primeros bloques de la *pipeline* (bloques verde y azul), de manera que se encontraban en un formato apto para poder analizar fácilmente las CNA. No obstante, a continuación se da una explicación más detallada de las distintas etapas de la *pipeline*.

El análisis de CNA parte de los archivos BAM obtenidos de la secuenciación de las muestras de adenomas avanzados. Los archivos BAM contienen los *reads* alineados a un genoma de referencia, en este caso el genoma humano GRCh38/hg38. El primer paso del análisis consistió en calcular los RC en ventanas genómicas no solapadas de longitud fija. En el análisis de datos LP-WGS se utiliza un tamaño de ventana mayor que el utilizado en datos WGS, con lo que se pierde resolución en la detección de las CNA. Sin embargo, es necesario utilizar un tamaño de ventana mayor para compensar la baja cobertura que proporciona la técnica LP-WGS [28]. En este trabajo se utilizó una ventana de 500 kilobases. Una vez calculados los RC, se eliminaron las ventanas que solapaban con regiones centroméricas, ya que éstas acumulan una profundidad de cobertura sistemáticamente distinta al resto de regiones. Estos dos pasos se habían realizado utilizando las funciones *coverageBed* e *intersectBed*, respectivamente, del conjunto de herramientas *bedtools*:

https://bedtools.readthedocs.io/en/latest/#

Los RC de cada muestra se centraron respecto a la mediana de todas las ventanas (todo el genoma) y se les aplicó una transformación logarítmica, obteniendo así los valores de log-ratio. De esta manera, una ganancia en el número de copia (*gain*) corresponde a un valor de log-ratio positivo y una pérdida en el número de copia (*loss*) corresponde a un valor de log-ratio negativo.

El siguiente paso del análisis consistió en aplicar un algoritmo de segmentación para agrupar regiones comunes, es decir, ventanas contiguas que tienen un número de copia muy similar, en segmentos. En este estudio se utilizó la segmentación binaria circular (CBS, del inglés *circular binary segmentation*), que es uno de los algoritmos de referencia para identificar cambios en el número de copia a partir de datos de secuenciación genómica. En particular, se utilizó el algoritmo implementado en el paquete *DNAcopy* de R:

https://bioconductor.org/packages/release/bioc/html/DNAcopy.html

Como resultado de la segmentación se obtuvo los archivos de los cuáles partió el análisis de CNA en este estudio (ver anexo A). Se disponía de un archivo delimitado por tabuladores por cada muestra. Cada fila de estos archivos correspondía a un segmento para el cual se indicaba: el ID de la muestra a la que pertenece el segmento (*ID*), el número de cromosoma donde se encuentra el segmento (*chrom*), las posiciones inicial (*loc.start*) y final (*loc.end*) del segmento, el número de regiones (ventanas) incluidas en el segmento (*num.mark*) y el valor promedio de log-ratio en el segmento (*seg.mean*).

A partir de los valores *seg.mean* se debe decide qué segmentos corresponden a CNA y cuáles no, para luego poder definir perfiles regionales de las CNA. En este trabajo se ha utilizado CNApp para realizar parte del análisis de las CNA.

Para poder analizar los datos con CNApp, se juntaron todos los archivos resultantes de la segmentación de los datos con DNAcopy en un único archivo llamado ***data_all.txt***, cuyas primeras cinco líneas se pueden ver a continuación:

ID	*chr*	*loc.start*	*loc.end*	*seg.mean*	*type_s*
10AD	*1*	*0*	*121000000*	*0.0322*	*AD*
10AD	*1*	*125000000*	*142000000*	*-7.9725*	*AD*
10AD	*1*	*142500000*	*249000000*	*0.0657*	*AD*
10AD	*10*	*0*	*2500000*	*-0.0856*	*AD*
10AD	*10*	*3000000*	*5500000*	*0.5033*	*AD*

Para adaptar el formato de los datos al requerido por CNApp, se eliminó la variable *num.mark*, se cambió el nombre de la variable *chrom* a *chr* y se añadió la variable *type_s*, que indica si un segmento dado corresponde a una muestra AD o a una ADK.

CNApp permite hacer una re-segmentación de los datos antes de su clasificación. Sin embargo, tras comprobar que había pocas variaciones respecto a la segmentación hecha previamente con DNAcopy, se optó por no utilizar esta opción. Por lo tanto, el primer paso que se hizo con CNApp fue identificar los segmentos que correspondían a CNA. Para ello, CNApp utiliza, por defecto, los umbrales indicados en la siguiente tabla para los valores *seg.mean*.

Tabla 3. *Valores por defecto de los umbrales de seg.mean y de número de copia para la detección de CNA en CNApp*

Nivel de CNA	*seg.mean*	Nº de copias
High-level gain	1	≥ 4 copias
Medium-level gain	0,58	[3 – 4) copias
Low-level gain	0,2	[2,3 – 3) copias
Low-level loss	-0,2	(1 – 1,7] copias
Medium-level loss	-1	(0,6 – 1] copias
High-level loss	-1,74	≤ 0,6 copias

Para comprobar que los umbrales por defecto se ajustaban bien a los datos de este estudio, se representó la distribución de los valores *seg.mean* por muestra y para todas las muestras juntas (Figura 5).

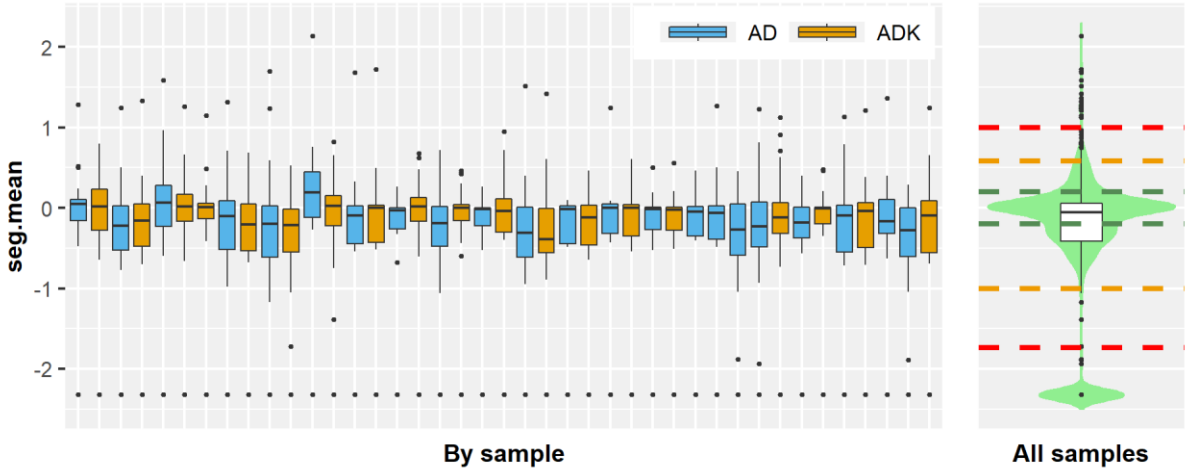

Figura 5. *Diagramas de caja de los valores seg.mean por muestra y para todas las muestras. Las líneas discontinuas indican los umbrales de seg.mean utilizados para la detección de CNA: 0,2/-0,2 (low gain/loss, verde), 0,58/-1 (normal gain/loss, naranja), 1/-1,74 (high gain/loss, rojo).*

En la Figura 5 se observa que los valores *seg.mean* están centrados aproximadamente en 0 y su distribución tiene una forma acampanada con largas colas tanto en la parte superior como inferior. Como es de esperar, el grueso de los segmentos son segmentos normales, con valores *seg.mean* comprendidos entre -0,2 y 0,2, por lo que parece razonable suponer que los segmentos cuyo valor *seg.mean* está fuera de ese intervalo son segmentos con algún tipo de alteración en el número de copia. Por lo tanto, se decidió mantener los umbrales por defecto para la detección de CNA.

Además de identificar las CNA, CNApp las clasifica en *broad* CNA, las cuales incluyen las que afectan a un cromosoma entero (*chromosomal*) o a un brazo cromosómico (*arm*), y *focal* CNA. Por defecto, CNApp considera como *broad* CNA aquellas que afectan, como mínimo, al 50% de un brazo cromosómico, y como *focal* CNA las que afectan a menos del 50% de un brazo cromosómico. El listado completo de segmentos clasificados se exportó al archivo ***Re-segmented_samples.xlsx***.

Una vez identificadas y clasificadas las CNA, se utilizó también CNApp para obtener los CNA *scores*. Se trata de dos parámetros, *broad* (BCS) y *focal* (FCS) CNA *scores*, que correlacionan altamente con la fracción de genoma alterada por las CNA *broad* y *focal*, respectivamente [33]. Por lo tanto, estos parámetros pueden ser útiles para analizar posibles diferencias entre las muestras AD y las muestras ADK. Los CNA *scores* de cada una de las muestras se exportaron al archivo ***CNA_Scores.tsv***.

Por último, se utilizó CNApp para calcular los perfiles regionales de las CNA. Para ello, CNApp divide el genoma en regiones o ventanas de tamaño definido por el usuario y luego comprueba qué segmentos de cada muestra solapan con cada región. De esta manera, para una muestra dada, el programa calcula un valor de log-ratio promedio (*W*) para cada región (*i*), como la suma ponderada de los valores *seg.mean* (*S*) de los *n* segmentos que solapan con esa región:

$$W(i) = \sum_{t=1}^{n} S_t \cdot \frac{l_t}{L(i)}$$

donde l_t corresponde a la longitud de cada segmento y $L(i)$ al tamaño de la región *i*. El resultado es una matriz con los valores *W* de cada región (filas) y muestra (columnas). En este estudio, los perfiles regionales de las *broad* CNA se calcularon a nivel de brazo cromosómico y se exportaron al archivo ***cna_profile_arms.tsv***. Los perfiles regionales de las *focal* CNA se calcularon en regiones de 1 megabase (Mb) y se exportaron al archivo ***cna_profile_1Mb.tsv***. En este segundo caso, se eligió el tamaño de región mínimo permitido por CNApp, para tener la máxima resolución posible.

Una vez calculados los perfiles regionales, se trabajó en R con los resultados exportados desde CNApp para realizar distintos tipos de cálculos y generar diferentes gráficos. En particular, se calcularon las frecuencias relativas, en porcentaje de muestras, tanto de las *broad* CNA como de las *focal* CNA, para cada tipo de muestra (AD o ADK), cada tipo de CNA (*gain* o *loss*) y cada región cromosómica (brazo cromosómico en las *broad* CNA o región de 1 Mb en las *focal* CNA). En las *focal* CNA, antes de calcular las frecuencias relativas, se eliminaron las regiones que solapaban con los telómeros y centrómeros. En el caso de los telómeros, simplemente se eliminó la primera y la última región de 1 Mb de cada cromosoma. En el caso de los centrómeros, se extendió el intervalo correspondiente al centrómero de cada cromosoma en 1 Mb por cada extremo y se eliminaron todas las regiones de 1 Mb que solapaban con el nuevo intervalo.

Las diferencias entre tipos de muestras se evaluaron mediante un test exacto de Fisher para cada región genómica, calculando previamente la correspondiente tabla de contingencia entre el tipo de CNA y el tipo de muestra. Debido al bajo número de muestras disponibles en este estudio, el uso de *p* valores ajustados y del nivel de significancia convencional de 0,05 no permitía obtener ningún resultado significativo. Con el fin de poder profundizar más en el análisis de posibles diferencias entre muestras AD y muestras ADK, se

decidió analizar la significancia estadística utilizando los *p* valores sin ajustar y seleccionando como remarcables las regiones con un *p* valor inferior a 0,2.

La interpretación biológica de las diferencias encontradas entre muestras AD y muestras ADK se llevó a cabo mediante un análisis de enriquecimiento de genes (GSEA, del inglés *gene set enrichment analysis*):

http://www.gsea-msigdb.org/gsea/index.jsp

En particular se utilizó la herramienta *compute overlaps*, que permite identificar procesos y vías biológicas enriquecidas en un conjunto de genes (*gene set* en inglés) proporcionado por el usuario, en este caso los localizados en las regiones en las que se encontró una diferencia remarcable entre muestras AD y muestras ADK, mediante la comparación con una colección de *gene sets* anotados.

La lista de genes para el GSEA se obtuvo a través de CNApp, que permite exportar a un archivo los genes comprendidos en una cierta región genómica. En el caso de las *broad* CNA, se analizaron por separado los genes de cada uno de los brazos cromosómicos que presentaron una diferencia remarcable entre las muestras AD y las muestras ADK, ya que el número de genes en regiones tan grandes es muy elevado. En el caso de las *focal* CNA, se analizaron conjuntamente los genes de todas las regiones de 1 Mb que presentaron una diferencia remarcable entre las muestras AD y las muestras ADK.

El resultado del GSEA es un listado con los *gene sets* anotados (o vías biológicas predefinidas) que mejor solapan con el *gene set* analizado. Para cada *gene set* encontrado se obtiene: el número de genes que contiene, una descripción, el número de genes que solapan con el *gene set* analizado, un *p* valor y un q valor FDR. Para limitar los resultados, sobre todo en el caso de las *broad* CNA, se seleccionaron en cada análisis los 50 *gene sets* con un q valor FDR más pequeño y menor que 0,01.

2.3. Resultados

2.3.1. Resultados globales

En primer lugar, se analizan las CNA de manera global, incluyendo tanto *broad* CNA como *focal* CNA. La Figura 6 muestra el número medio de eventos CNA por muestra, de forma separada para las muestras AD y las muestras ADK, así como por cromosoma. En general, se detectan más *focal* CNA que *broad* CNA por muestra, lo cual es de esperar por el menor tamaño de las regiones genómicas analizadas en el caso de las *focal* CNA. También se observa que en los cromosomas 1, 9 y 16 hay un número claramente más elevado de *focal losses* (en torno a 2 por muestra) que en el resto de cromosomas, tanto en muestras AD como ADK, lo cual hace pensar que puede deberse a errores sistemáticos de secuenciación. Algo similar ocurre con las *broad* CNA en los cromosomas 13, 14, 15, 21 y 22, los cuales presentan una *arm loss* en todas las muestras. En este caso, se trata de cromosomas acrocéntricos, los cuales

tienen uno de los brazos muy corto. Este hecho, junto con la baja cobertura del LP-WGS, provoca la detección de una falsa *arm loss* en esas regiones genómicas en el 100% de las muestras.

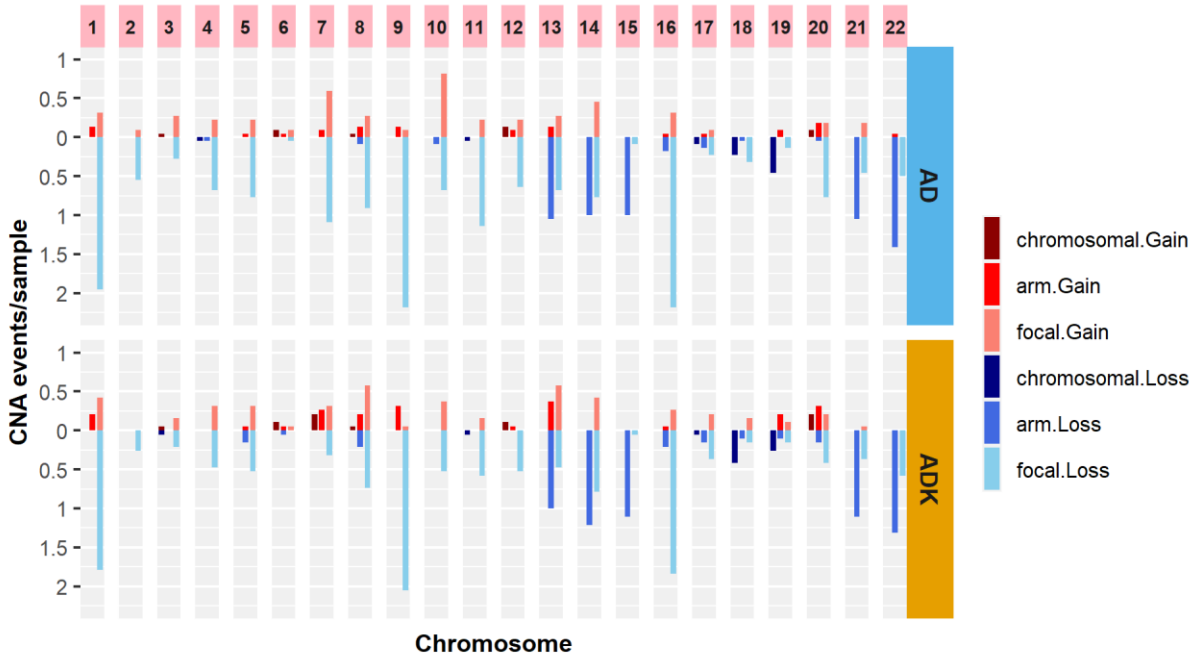

Figura 6. *Número medio por muestra de los distintos tipos de CNA, por tipo de muestra y cromosoma.*

La Figura 7 muestra el número medio de eventos CNA por muestra y los valores *seg.mean* de las CNA, de forma separada para las muestras AD y las muestras ADK.

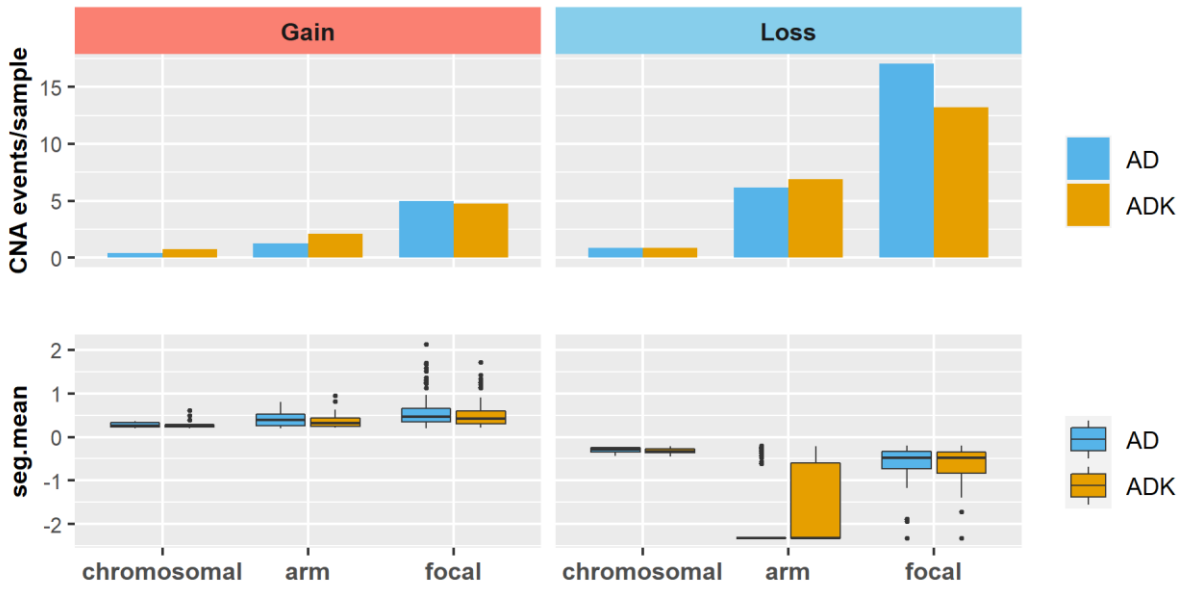

Figura 7. *Número medio por muestra y valores seg.mean de los distintos tipos de CNA, por tipo de muestra.*

Como resultado destacable, se observa que las muestras ADK tienen un número de *broad* CNA por muestra ligeramente superior a las muestras AD. Sin embargo, ocurre lo contrario con el número de *focal* CNA por muestra, que es superior en las muestras AD.

En cuanto a los valores *seg.mean*, no se aprecian diferencias relevantes entre los dos tipos de muestras. Cabe destacar, que los valores *seg.mean* anormalmente bajos en el caso de las *arm losses* están relacionados con las falsas CNA detectadas en los cromosomas 13, 14, 15, 21 y 22. En la Figura 20 (anexo B) se puede observar este hecho con mayor claridad, ya que se muestra la relación entre valores *seg.mean* y tipos de CNA, muestra a muestra y para cada cromosoma.

Para terminar con el análisis global de las CNA, en la Figura 8 se muestran los *broad* y *focal* CNA *scores* para las muestras AD y las muestras ADK por separado. La significancia de las diferencias entre los dos tipos de muestras se analizó mediante pruebas de suma de rangos de Wilcoxon. No se obtuvo diferencias significativas en ninguno de los *scores*. Sin embargo, en el caso de los BCS, el *p* valor obtenido (0,074) está cerca de ser significativo. Esto se puede interpretar como que, a pesar de no ser una diferencia significativa, las muestras ADK tienen una mayor fracción del genoma alterada por *broad* CNA que las muestras AD.

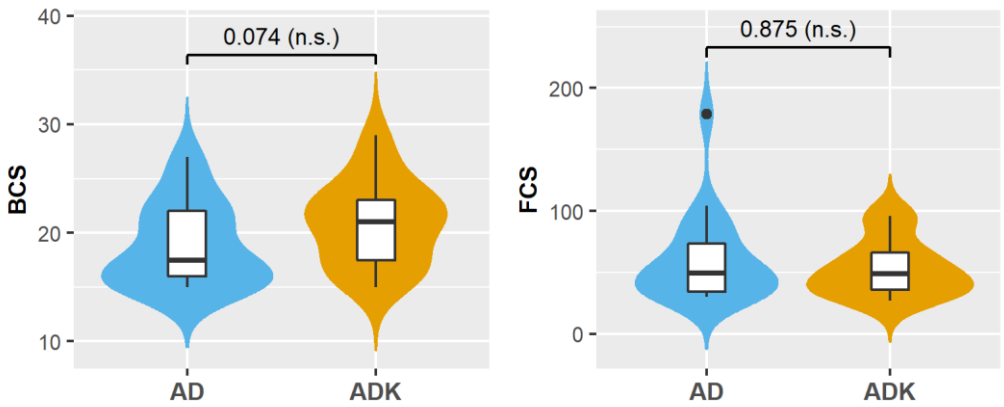

Figura 8. *Distribución de los broad (izquierda) y focal (derecha) CNA scores por tipo de muestra.*

2.3.2. *Broad* CNA

En este apartado se analizan en detalle las *broad* CNA. En la Figura 9 se muestra la frecuencia relativa de las *broad* CNA, de forma separada para las muestras AD y las muestras ADK, así como para cada brazo cromosómico. Cabe destacar que los brazos *p* de los cromosomas 13, 14, 15, 21 y 22 se han eliminado del análisis ya que se identificó un alto número de falsos positivos (*arm losses*) en esas regiones, tal y como se explica en el apartado anterior.

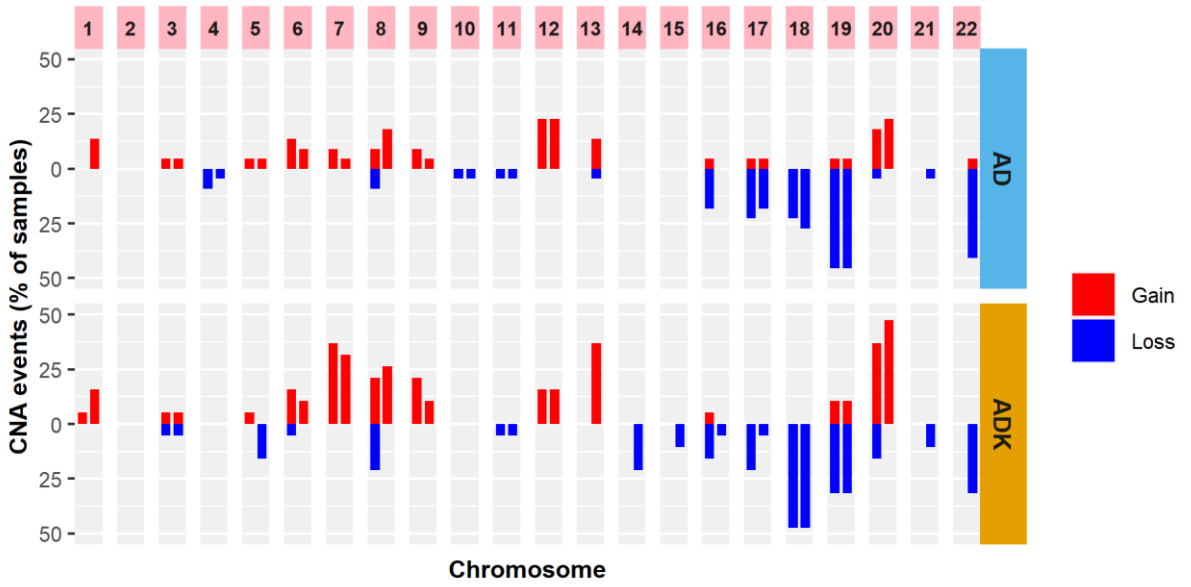

Figura 9. *Frecuencia relativa de las broad CNA por tipo de muestra y brazo cromosómico. En cada cromosoma, las barras de la izquierda corresponden al brazo p y las barras de la derecha al brazo q.*

En la Figura 9 destacan las siguientes *broad* CNA (con frecuencias > 25% en las muestras ADK):

- ***Gains* en 7p y 7q:** 9,1% (AD) vs 36,8% (ADK) y 4,5% (AD) vs 31,6% (ADK), respectivamente.
- ***Gain* en 8q:** 18,2% (AD) vs 26,3% (ADK).
- ***Gain* en 13q:** 13,6% (AD) vs 36,8% (ADK).
- ***Losses* en 18p y 18q:** 22,7% (AD) vs 47,4% (ADK) y 27,3% (AD) vs 47,4% (ADK), respectivamente.
- ***Gains* en 20p y 20q:** 18,2% (AD) vs 36,8% (ADK) y 22,7% (AD) vs 47,4% (ADK), respectivamente.

Además de las anteriores, otras *broad* CNA destacables son las de los brazos 5q (*loss*), 8p (*loss*), 9p (*gain*), 14q (*loss*) y 17p (*loss*). De entre las CNA anteriores, las más recurrentes son las pérdidas de los brazos 18p y 18q, así como la ganancia del brazo 20q.

Para conocer qué CNA son más relevantes en la transición AD-ADK, se comprobó la significancia de las diferencias en la frecuencia de las *broad* CNA entre tipos de muestras mediante tests exactos de Fisher. La Figura 10 muestra los resultados más destacables del análisis estadístico.

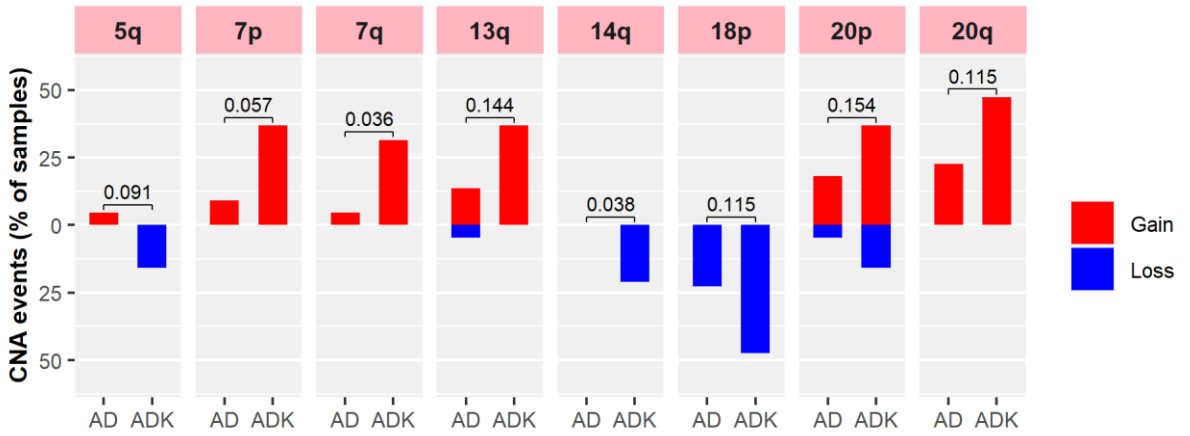

Figura 10. *Broad CNA cuyas diferencias entre las muestras AD y las muestras ADK tienen un p valor < 0,2.*

Las diferencias más significativas (p < 0,1) están asociadas a la ganancia del brazo 7q (p = 0,036) y a la pérdida del brazo 14q (p = 0,038), seguidas de la ganancia del brazo 7p (p = 0,057) y la pérdida del brazo 5q (p = 0,091). Aunque menos significativas, son también destacables las ganancias de los brazos 13q (p = 0,144), 20p (p = 0,154) y 20q (0,115), así como la pérdida del brazo 18p (p = 0,115). Estos resultados son muy similares a los obtenidos previamente en las mismas muestras de AD avanzados a partir de datos FISH [32] (Figura 11).

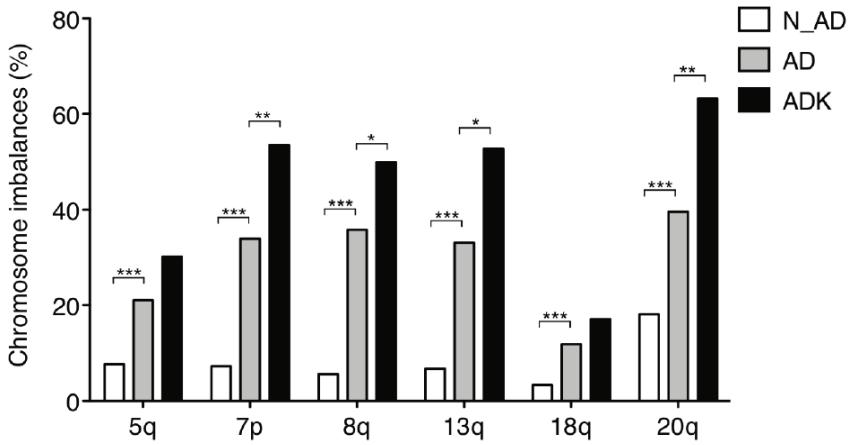

Figura 11. *Frecuencia media de las alteraciones cromosómicas observadas en AD no avanzados (N_AD) y en regiones AD y regiones ADK de AD avanzados mediante la técnica FISH. * p valor < 0.05, ** p valor < 0.005. Extraída de* [32].

En los resultados de los datos FISH, destacan las diferencias significativas entre las regiones de AD y las regiones de ADK en los brazos cromosómicos 7p, 8q, 13q y 20q. También, aunque de forma no significativa, se encontraron diferencias en la frecuencia de las alteraciones de los brazos 5q y 18q. En todos los casos, la frecuencia de las alteraciones cromosómicas fue mayor en las muestras ADK que en las muestras AD.

Es especialmente destacable la pérdida del brazo 14q encontrada con los datos LP-WGS, con una frecuencia del 21,1% en las muestras ADK frente al 0% en

las muestras AD. A pesar de no ser de las CNA más recurrentes, es de las diferencias más significativas encontradas con los datos LP-WGS. Sin embargo, no se reportó esta alteración en el análisis de los datos FISH.

Además de la frecuencia de las *broad* CNA, también es interesante comprobar si existe algún tipo de relación entre ellas, es decir, si hay algún tipo de coexistencia de *broad* CNA. En ese caso, podrían definirse grupos o clústeres de *broad* CNA característicos del CCR. Para comprobar este hecho, se han creado *heatmaps* con la distribución muestra a muestra de las *broad* CNA más relevantes (Figura 12).

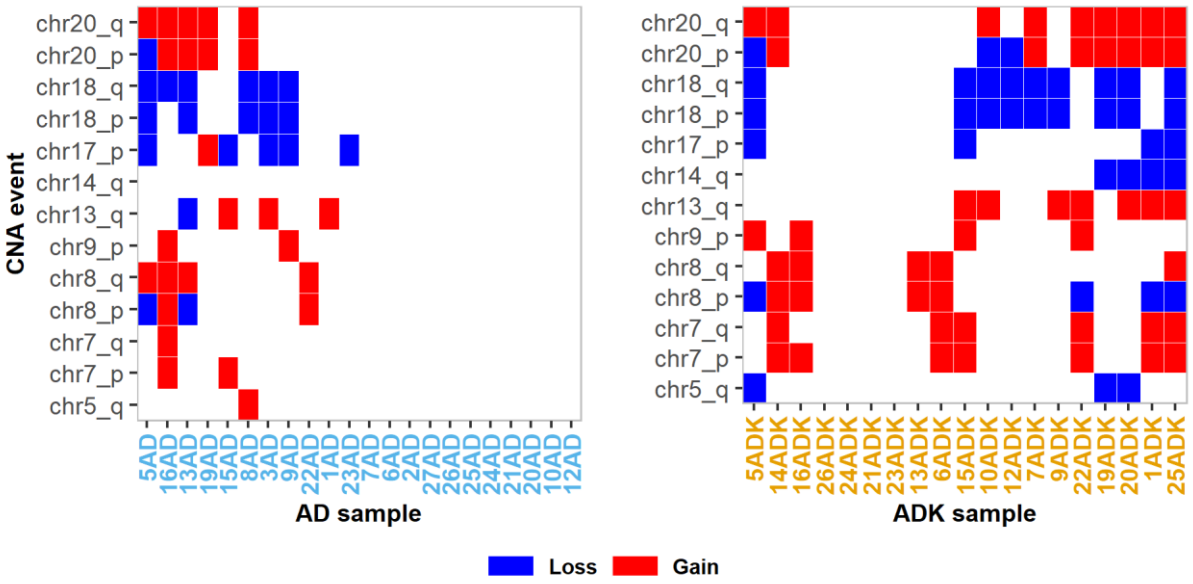

Figura 12. *Mapa de distribución de las principales broad CNA en las muestras AD y las muestras ADK.*

Debido al reducido número de muestras disponibles, es difícil identificar un patrón firme de coexistencia de *broad* CNA a partir de los *heatmaps* de la Figura 12, especialmente en las muestras AD, en las que la frecuencia de las CNA es baja. En cualquier caso, es evidente la elevada coexistencia entre *broad* CNA de brazos de un mismo cromosoma, especialmente en los cromosomas 7, 18 y 20. Para valorar el resto de casos, se han utilizado los siguientes criterios:

- Se tienen en cuenta las *broad* CNA que aparecen en cuatro o más muestras.
- Se considera que una *broad* CNA *X* coexiste con una *broad* CNA *Y* si *Y* aparece en más del 75% de las muestras en las que aparece *X*.

Siguiendo los criterios anteriores, destacan las siguientes coexistencias:

- Muestras AD: ganancia en los brazos 20p y 20q y pérdida en el brazo 18q.
- Muestras ADK:
  - Ganancia en los brazos 7p y 7q y pérdida en el brazo 8p.
  - Ganancia en los brazos 7p y 8q.
  - Pérdida en el brazo 14q y ganancias en los brazos 20p y 20q.

Para terminar el análisis de *broad* CNA, se llevó a cabo el GSEA en los brazos cromosómicos en los que se encontraron diferencias entre las muestras AD y las muestras ADK con un *p* valor < 0,2 (ver Figura 10).

Al tratarse de regiones genómicas muy grandes, se encontró un número de *gene sets* significativos muy elevado. En el anexo C se pueden consultar las tablas con el listado completo de *gene sets* encontrados para los distintos brazos cromosómicos. Para sintetizar los resultados, se agruparon los *gene sets* teniendo en cuenta su función biológica, tal y como se puede observar en la Figura 13.

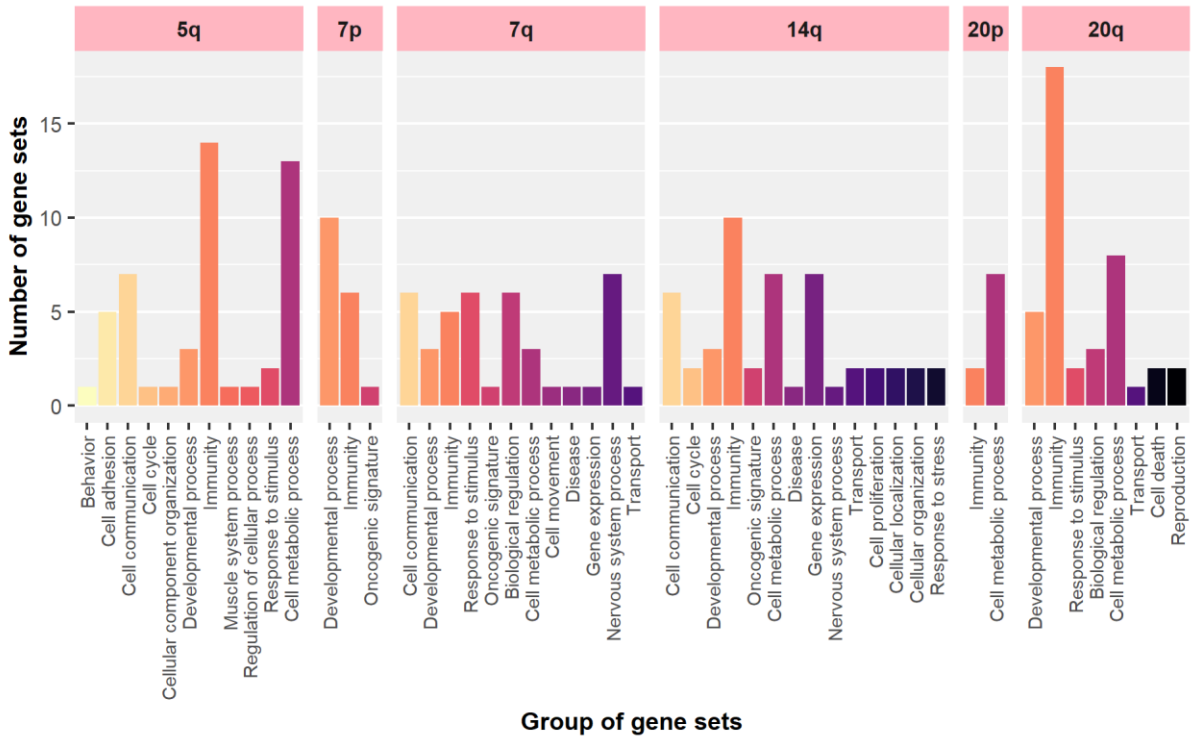

Figura 13. *Distribución de los gene sets más significativos asociados a las broad CNA según su función biológica.*

De entre los grupos de *gene sets* definidos, los tres más frecuentes son los que engloban genes relacionados con el sistema inmunitario, así como con el metabolismo y desarrollo celular. Se trata, en los tres casos, de funciones biológicas que, generalmente, se ven alteradas en el proceso de carcinogénesis.

2.3.3. *Focal* CNA

En este apartado se analizan en detalle las *focal* CNA. En la Figura 14 se muestra la frecuencia relativa de las *focal* CNA, de forma separada para las muestras AD y las muestras ADK, así como para cada cromosoma. Cabe destacar que se eliminaron previamente todas las regiones de 1 Mb en las que aparecían de manera sistemática pérdidas o ganancias en más del 50% tanto de las muestras AD como de las muestras ADK.

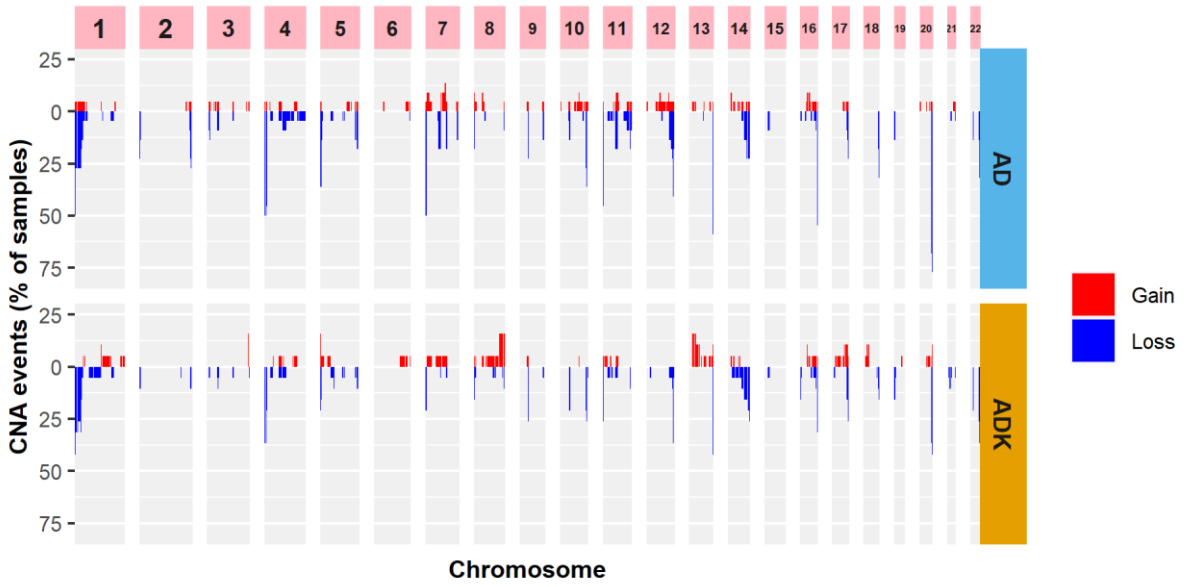

Figura 14. *Frecuencia relativa de las focal CNA por tipo de muestra y región genómica de 1 Mb.*

Un aspecto destacable es que la mayor parte de los picos, que corresponden a las regiones con mayor frecuencia de *focal* CNA, se encuentran en los extremos de los cromosomas, ya sea en las posiciones iniciales o finales.

Con el objetivo de identificar *focal* CNA que puedan tener un papel relevante en la transición AD-ADK, se comprobó la significancia de las diferencias en la frecuencia de las *focal* CNA entre tipos de muestras mediante un test exacto de Fisher en cada una de las regiones de 1 Mb. Tras el análisis estadístico, se agruparon las regiones contiguas de un mismo cromosoma, que tuvieran la misma frecuencia y el mismo *p* valor. La Figura 15 muestra los resultados más destacables del análisis estadístico.

Las diferencias más significativas ($p < 0,1$) aparecen en regiones de los cromosomas 7, 8, 11, 12, 13 y 20. De manera menos significativa ($p < 0,2$), también aparecen diferencias en regiones de los cromosomas 2, 4 y 7. De estas regiones, la más extensa es la que va de 116 a 135 Mb (19 Mb) en el cromosoma 8, seguida por las que van de 63 a 72 Mb (9 Mb) y de 1 a 6 Mb (5 Mb) en el cromosoma 7. Un aspecto destacable es que, excepto en las regiones de los cromosomas 8 y 13, las cuales presentan ganancias en el 15,8% de las muestras ADK versus el 0% de las muestras AD, en el resto de regiones aparecen pérdidas focales con una frecuencia mayor en las muestras AD que en las muestras ADK.

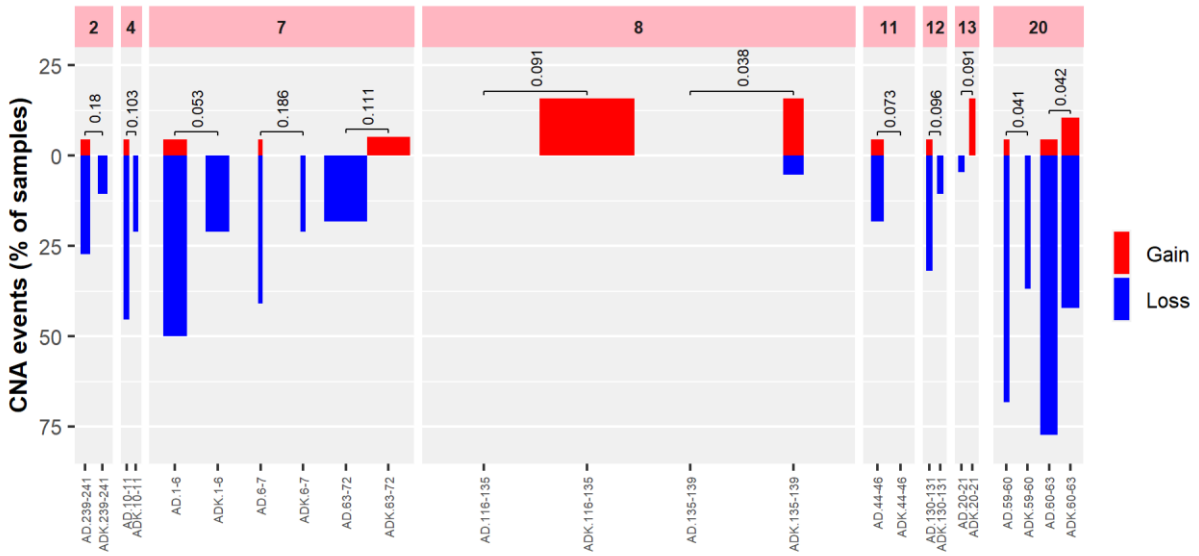

Figura 15. *Focal CNA cuyas diferencias entre las muestras AD y las muestras ADK tienen un p valor < 0,2. Las etiquetas del eje horizontal indican el tipo de muestra (AD o ADK) seguido de las posiciones inicial y final de la focal CNA correspondiente, en millones de bases. El ancho de las barras es proporcional a la longitud de la focal CNA.*

En la Figura 16 se pueden ver los *heatmaps* con la distribución muestra a muestra de las *focal* CNA más relevantes.

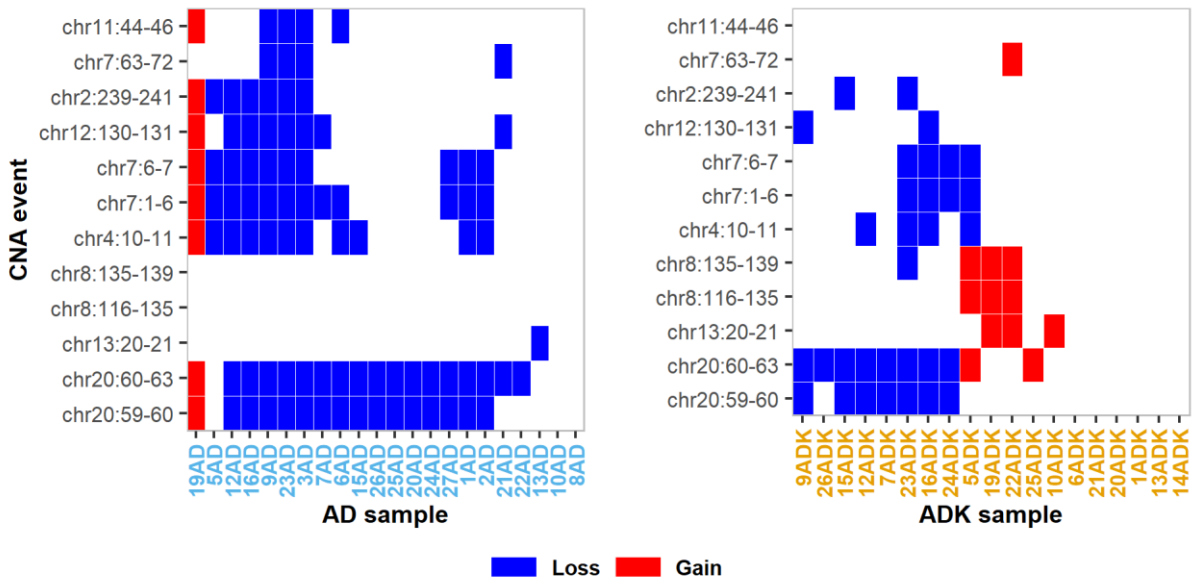

Figura 16. *Mapa de distribución de las focal CNA cuyas diferencias entre las muestras AD y las muestras ADK tienen un p valor < 0,2.*

En las muestras AD, las *focal* CNA localizadas en el cromosoma 20 afectan a más del 70% de las muestras. Por el contrario, no hay muestras que contengan *focal* CNA en el cromosoma 8 y solamente una muestra (13AD) tiene una pérdida focal en el cromosoma 13. El resto de las *focal* CNA tienen un alto grado de coexistencia, especialmente las del cromosoma 4 y las localizadas en la región de 1 a 7 Mb del cromosoma 7.

En las muestras ADK, se observa también una mayor prevalencia de las *focal* CNA ubicadas en el cromosoma 20, aunque con menor frecuencia que en las muestras AD. También con menor frecuencia, sigue habiendo cierta coexistencia de las *focal* CNA del cromosoma 4 y de la región inicial del cromosoma 7. A diferencia de las muestras AD, aparecen ganancias focales en regiones de los cromosomas 8 y 13.

En cuanto al GSEA, se realizó de manera conjunta para todos los genes ubicados en las 12 regiones con un *p* valor < 0,2 entre muestras AD y muestras ADK (Figura 15). Se encontraron los siguientes *gene sets* significativos:

- **GSE45365_CTRL_VS_MCMV_INFECTION_NK_CELL_DN**
  - **Nº de genes**: 194 (20 coincidencias)
  - **Descripción**: *genes down-regulated in NK cells: control versus acute primary viral infection*.
  - **q valor FDR**: 2,59e-11
- **GSE45365_CD8A_DC_VS_CD11B_DC_IFNAR_KO_DN**
  - **Nº de genes**: 196 (16 coincidencias)
  - **Descripción**: *genes down-regulated in dendritic cells with IFNAR1 [GeneID=3454] knockout: CD8A [GeneID=925] versus ITGAM+ [GeneID=3684]*.
  - **q valor FDR**: 3,13e-7
- **GSE18804_BRAIN_VS_COLON_TUMORAL_MACROPHAGE_DN**
  - **Nº de genes**: 194 (15 coincidencias)
  - **Descripción**: *genes down-regulated in tumor associated macrophages conditioned by: glioblastoma versus colorectal adenocarcinoma*.
  - **q valor FDR**: 1,82e-6

Los tres *gene sets* encontrados están relacionados con el sistema inmunitario. En la tabla 4 se muestra el listado completo de genes coincidentes con estos tres *gene sets*:

Tabla 4. *Genes asociados a las focal CNA e incluidos en los conjuntos de genes detectados en el GSEA*

Gene symbol	Chr	Start	End	GSE45365 Ctrl VS Mcmv Infection NK Cell DN	GSE45365 CD8A DC VS CD11B DC IFNAR KO DN	GSE18804 Brain VS Colon Tumoral Macrophage DN
RNF139	8	124.474.766	124.488.618	●	●	●
NSMCE2	8	125.091.822	125.182.572	●	●	●
MRPL13	8	120.395.843	120.445.407	●	●	●
PHF20L1	8	132.775.357	132.813.194	●	●	●
EIF3H	8	116.644.815	116.755.823	●	●	●
RAD21	8	116.845.933	116.874.866	●	●	●
NDUFB9	8	124.539.101	124.549.986	●	●	
CYRIB	8	129.839.593	130.017.129	●	●	
DSCC1	8	119.833.940	119.855.930	●		●
ATAD2	8	123.319.850	123.396.465	●		●
TRMT12	8	124.450.806	124.453.025	●		●
C8orf76	8	123.219.955	123.241.398	●		●
DERL1	8	123.013.163	123.042.423	●		

FBXO32	8	123.497.886	123.532.143	●			
UTP23	8	116.766.502	116.774.682	●			
NTAQ1	8	123.416.701	123.467.527	●			
FAM91A1	8	123.768.438	123.810.661			●	●
WASHC5	8	125.024.260	125.091.819			●	●
HAS2	8	121.613.030	121.641.390			●	
TAF2	8	119.730.773	119.832.834			●	
TBC1D31	8	123.072.618	123.152.152			●	
TATDN1	8	124.488.494	124.539.088			●	
HHLA1	8	132.061.485	132.105.265			●	
ZFAT	8	134.477.787	134.696.558			●	
TRIB1	8	125.430.320	125.438.405				●
SNTB1	8	120.535.744	120.812.069				●
ASAP1	8	130.052.104	130.443.660				●
ZNF736	7	64.313.872	64.337.362	●			
EIF2AK1	7	6.022.246	6.059.229	●			
ZNF92	7	65.373.798	65.401.135	●			
RAN	12	130.871.993	130.877.678	●			

Como se puede ver en la tabla anterior, la mayor parte de los genes están ubicados en el cromosoma 8, en particular en la región que va de 116 a 135 Mb.

2.4. Discusión

En este capítulo se llevado a cabo un análisis de las CNA detectadas a partir de datos LP-WGS obtenidos de regiones de AD y regiones de ADK de AD avanzados. El principal objetivo de este análisis era estudiar el potencial de la técnica LP-WGS para detectar tanto las CNA *driver* conocidas en el CCR, tradicionalmente estudiadas mediante técnicas de hibridación, como nuevas CNA, especialmente focales, que puedan estar involucradas en la transición AD-ADK.

Varios estudios han reportado el uso del LP-WGS para el análisis de CNA [18], [27], [28], [30], [31], [36], [37], así como las ventajas del LP-WGS frente a las técnicas basadas en arrays [29], [38]. El LP-WGS ofrece una alta resolución en la detección de alteraciones genómicas y permite la posibilidad de detectar nuevas alteraciones no conocidas. Además, gracias a la baja profundidad de cobertura, el LP-WGS tiene un precio más asequible que el WGS.

En primer lugar, se han analizado de manera global las *broad* CNA y las *focal* CNA. Los CNA *scores* [33] mostrados en la Figura 8, ofrecen una visión general de cuán afectados están los genomas de las muestras AD y las muestras ADK por los dos tipos de CNA. En el caso de las *broad* CNA, a pesar de ser menos numerosas que las *focal* CNA (ver Figura 7), abarcan regiones genómicas más grandes (brazos cromosómicos o cromosomas enteros), con lo que afectan a una fracción mayor del genoma. Por lo tanto, y dado que las muestras ADK presentan un mayor número de *broad* CNA que las muestras AD (ver Figura 7), la diferencia en el BCS entre tipos de muestras está cerca de ser significativa (p = 0,074), siendo el BCS mayor en las muestras ADK. Con respecto a las *focal* CNA, se ha identificado un gran número de ellas, especialmente de pérdidas focales, tanto en las muestras AD como en las muestras ADK. Dado que estas alteraciones abarcan regiones genómicas muy pequeñas, el hecho

de que las muestras AD tengan más *focal* CNA de no hace que el FCS sea diferente entre tipos de muestras (p = 0,875).

Con respecto a las *broad* CNA identificadas, las más recurrentes han sido las pérdidas de los brazos 18p y 18q y las ganancias en los brazos 7p, 7q, 8q, 13q, 20p y 20q (ver Figura 9). Otras *broad* CNA destacables que se han identificado han sido las pérdidas de los brazos 5q, 8p, 14q y 17p, así como la ganancia en el brazo 9p. En todos los casos, las *broad* CNA son más frecuentes en las muestras ADK que en las muestras AD. Estos resultados concuerdan con los reportados en la literatura, tal y como se puede ver en la tabla 5, que recoge los principales resultados de algunos estudios de genómica del CCR realizados en los últimos 20 años.

Tabla 5. *Estudios previos sobre las broad CNA asociadas al CCR*

Primer autor	Año	Técnica	Resultados principales
Hermsen M [16]	2002	CGH	**Pérdidas**: 8p, 15q, 17p y 18q **Ganancias**: 8q, 13q y 20q Asociadas a la progresión AD-ADK
Diep CB [17]	2006	CGH	**Pérdidas**: 17p, 18q, 4p, 8p y 14q **Ganancias**: 8q, 13q y 20, 7p y 17q Asociadas a la transición AD-ADK Aparecen pronto en carcinomas primarios Asociada a la transición a metástasis de hígado Alteración tardía
Leslie A [39]	2006	CGH	**Pérdidas**: 18q, 11q y 17p **Ganancias**: 12p, 20q, 13q y 20p Correlaciona con la mutación *KRAS* Correlacionan con la mutación *TP53* Correlacionan con la presencia de AD sincrónicos Asociadas a la aparición de displasia de alto grado
Camps J [15]	2009	CGH	**Pérdidas**: 1p, 4q, 5q, 8p, 17p, 18 y 21 **Ganancias**: 7, 8q, 11p, 13 y 20q Asociadas a carcinomas colorrectales primarios
Muzny DM [36]	2012	Arrays SNP/LP-WGS	**Pérdidas**: 1p, 4q, 5q, 8p, 14q, 15q, 17p, 17q, 18p, 18q, 20p y 22q **Ganancias**: 1q, 7p, 7q, 8p, 8q, 12q, 13q, 19q, 20p y 20q Asociadas a tumores colorrectales
Xie T [22]	2014	WGS	**Pérdidas**: 18q **Ganancias**: 8, 10, 13q, 14, 20q y 21q Aparecen solamente en el tumor primario Aparece solamente en el tumor metastásico Aparecen en los dos tumores
Mamlouk S [20]	2017	*Targeted sequencing*	**Pérdidas**: 17p entre otras **Ganancias**: 20q entre otras Alta heterogeneidad en las CNA tanto en tumores primarios como en tumores metastásicos
Cross W [21]	2018	WGS/WES	**Pérdidas**: 5q, 8p, 17p y 18q **Ganancias**: 1q, 7, 13q y 20 Asociadas al CCR
Carvalho B [18]	2018	LP-WGS	**Pérdidas**: 17p y 18q **Ganancias**: 13q y 20q Más frecuentes en AD avanzados que en AD no avanzados y en AD con displasia de alto grado
Molparia B [37]	2018	LP-WGS	**Pérdidas**: 8p, 18 y 9p **Ganancias**: 6p y 10p Asociadas al CCR en ADN tumoral circulante

De las *broad* CNA identificadas, se encontró diferencias remarcables (p < 0,2) entre muestras AD y muestras ADK en los brazos 5q, 7p, 7q, 13q, 14q, 18p, 20p y 20q. Estas *broad* CNA podrían, por lo tanto, estar más asociadas a la

progresión AD-ADK. Sin embargo, hay que tener en cuenta que la identificación de diferencias en las *broad* CNA entre muestras AD y muestras ADK ha estado condicionada por el número reducido de muestras disponibles en este estudio. El tamaño de la muestra dificulta la posibilidad de encontrar diferencias significativas entre tipos de muestras, motivo por el cual se decidió subir el umbral de significancia hasta 0,2 y utilizar los *p* valores sin ajustar. Por lo tanto, conviene interpretar con cautela los resultados estadísticos.

En cuanto a la comparación de los resultados obtenidos en este estudio con los obtenidos previamente mediante la técnica FISH en los mismos AD avanzados, cabe decir que ambos son muy parecidos. Al margen de las significancias, todas las *broad* CNA detectadas con FISH (Figura 11) se han detectado también con LP-WGS. Sin embargo, las ventajas del LP-WGS, y en general del WGS, sobre FISH son claras [40], [41]. En primer lugar, con FISH el análisis se centra en una serie de secuencias correspondientes a regiones genómicas previamente seleccionadas, mientras que con LP-WGS se analiza el genoma al completo, pudiéndose identificar nuevas alteraciones no conocidas. Por ejemplo, con FISH no se detectaron las *broad* CNA de los brazos 7q, 8p, 9p, 14q, 17p, 18p y 20p. Además, el LP-WGS permite identificar las CNA con más resolución, pues la secuenciación es a nivel de nucleótido.

Con respecto a la coexistencia de *broad* CNA, la encontrada en las muestras AD entre la ganancia en los brazos 20p y 20q y la pérdida del brazo 18q (Figura 12) había sido previamente reportada [42]. No se han observado otras coexistencias reportadas en la literatura del CCR como las de los cromosomas 13 y 14, las ganancias en 8q y 13q, las ganancias en 8q y 20q o la pérdida en 8p con la ganancia en 20q y las pérdidas en 17p y 18 [16], [42]. No obstante, cabe destacar la dificultad para encontrar coexistencia de CNA con el reducido número de muestras disponibles en este estudio.

El análisis GSEA de las *broad* CNA ha dado como resultado un predominio de *gene sets* relacionados con la respuesta inmune y el metabolismo celular (Figura 13), lo cual concuerda con el rol principal que tiene la alteración de estos dos procesos biológicos en el desarrollo del cáncer [43]–[46].

En relación con el análisis de *focal* CNA, es destacable que la mayor parte de las regiones en las que había una diferencia remarcable entre las muestras AD y las muestras ADK ($p < 0,2$) contienen pérdidas focales con una frecuencia mayor en las muestras AD que en las muestras ADK (Figura 15). Este resultado es opuesto a lo que se esperaría encontrar, que es una mayor frecuencia de *focal* CNA relacionadas con una mayor inestabilidad cromosómica y heterogeneidad en las muestras ADK. Este hecho, junto con los frecuentes picos de pérdidas focales que se observan en la Figura 14, generalmente localizados en los extremos de los cromosomas, hace pensar que puede haber algún tipo de sesgo en los datos que esté afectando al análisis de *focal* CNA.

Uno de los pasos más importantes en el análisis de CNA a partir de datos LP-WGS, y en general de datos WGS, es la normalización de los datos, un paso necesario para corregir posibles sesgos en los datos. Por ejemplo, son conocidos los sesgos introducidos por las variaciones en el porcentaje GC o las

regiones repetitivas del genoma. Otro paso importante en la normalización de los datos es la eliminación de regiones mal secuenciadas, con una cobertura sistemáticamente mayor o menor que la del resto de regiones. En el caso del LP-WGS, además, la combinación de los efectos mencionados con la baja profundidad de cobertura empleada, puede llevar a la identificación de falsas CNA. En el caso de las *broad* CNA, al analizarse regiones genómicas grandes, el efecto de la baja cobertura del LP-WGS y de los posibles sesgos o errores sistemáticos no es tan evidente como en el caso de las *focal* CNA. En el análisis de *focal* CNA, la falta de cobertura en una región concreta del genoma o las variaciones en la cobertura debidas al contenido GC de las distintas regiones pueden provocar la detección de falsas pérdidas focales.

Con el objetivo de entender mejor los resultados obtenidos, se ha representado el porcentaje GC de los cromosomas en los que se han identificado diferencias remarcables en las *focal* CNA entre muestras AD y muestras ADK (Figura 17).

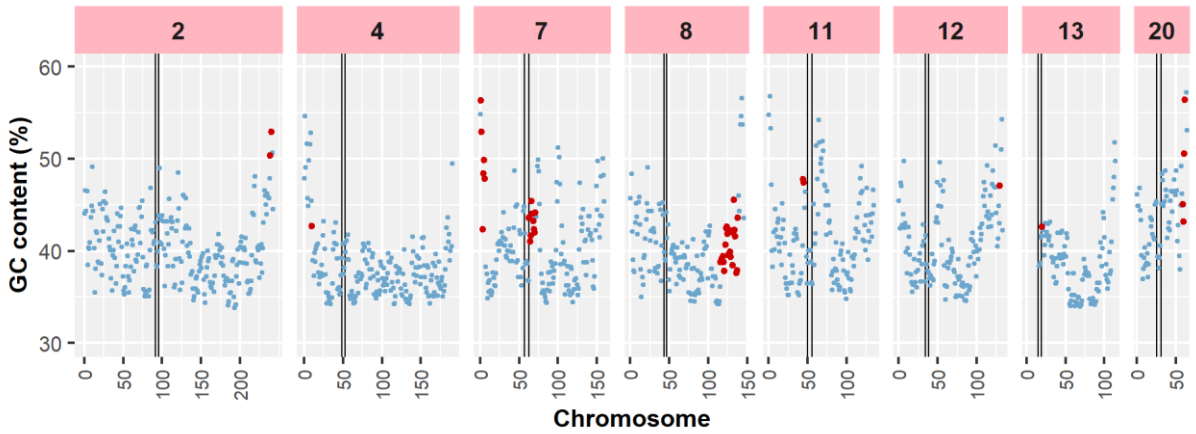

Figura 17. *Porcentaje GC calculado en ventanas de 1 Mb en los cromosomas con focal CNA cuyas diferencias entre las muestras AD y las muestras ADK tienen un p valor < 0,2. Los puntos rojos indican la localización de las focal CNA. Las líneas verticales delimitan las regiones centroméricas.*

En primer lugar, se observa que la mayor parte de las *focal* CNA se ubican cerca de los extremos de los cromosomas, es decir, de las regiones teloméricas, o cerca de las regiones centroméricas. A pesar de que tanto los centrómeros como los telómeros de todos los cromosomas se eliminaron para el análisis de las *focal* CNA, puede que las regiones próximas a ellos también tengan una cobertura sistemáticamente distinta al resto de regiones y, por lo tanto, hayan influido en la detección de las *focal* CNA. Además, excepto una parte de las *focal* CNA detectadas en el cromosoma 8, el resto de *focal* CNA se ubica en las regiones con mayor porcentaje GC de los respectivos cromosomas. Dado que en este estudio no se ha tenido en cuenta el porcentaje GC en la normalización de los datos, este hecho también puede haber influido en la detección de *focal* CNA.

Cabe destacar, que las *focal* CNA detectadas en la región del cromosoma 8 que va de 116 a 135 Mb, están más alejadas de regiones centroméricas y teloméricas y no corresponden a regiones con un porcentaje GC muy elevado. Por lo tanto, podría considerarse que la detección de esas *focal* CNA no se ha

visto afectada por los problemas anteriormente comentados. Además, coincide que el solapamiento con los *gene sets* más significativos del GSEA de las *focal* CNA se da principalmente con genes de esa región del cromosoma 8 (ver tabla 4).

En el caso de las *focal* CNA, los 3 *gene sets* identificados están relacionados con funciones del sistema inmunitario. En particular, contienen genes que aparecen desregulados en distintos tipos de células del sistema inmunitario, como las células NK, las células dendríticas con bloqueo del IFNAR1 y macrófagos asociados a tumores. Esa desregulación génica hace que fallen los distintos mecanismos que tiene el sistema inmunitario para combatir las células tumorales, favoreciendo así la proliferación tumoral.

3. Análisis de SNV

3.1. Descripción de los datos

Se dispone de los datos de *targeted sequencing* de un total de 24 AD avanzados (tabla 6). De éstos, se han obtenido resultados de las regiones de AD en 23 muestras y de las regiones de ADK en 23 muestras. Así, en un total de 22 muestras se tienen datos tanto de la región AD cómo de la región ADK de la misma lesión tumoral.

Tabla 6. *Muestras de AD avanzados utilizadas en el análisis de SNV*

ID	Muestra AD	Muestra ADK
1	●	●
2	●	●
3	●	●
4		●
5	●	●
6	●	●
7	●	●
8	●	
9	●	●
10	●	●
12	●	●
13	●	●
14	●	●
15	●	●
16	●	●
19	●	●
20	●	●
21	●	●
22	●	●
23	●	●
24	●	●
25	●	●
26	●	●
27	●	●

3.2. Metodología

Los datos de *targeted sequencing* disponibles inicialmente en este trabajo habían sido previamente procesados con MuTect, una herramienta desarrollada en el *Broad Institute* de Cambridge, Massachusetts, para la identificación de mutaciones puntuales somáticas a partir de datos NGS de genomas tumorales [47]:

https://software.broadinstitute.org/cancer/cga/mutect

La detección de SNV con MuTect, parte de los archivos BAM de dos muestras pareadas de ADN tumoral y normal y consta de los siguientes 4 pasos: eliminación de datos de baja calidad, detección de variantes en la muestra tumoral mediante un clasificador Bayesiano, filtrado para eliminar falsos positivos y clasificación de las variantes en somáticas o germinales mediante un segundo clasificador Bayesiano. El resultado es un archivo *vcf* (del inglés

variant call format) que contiene las variantes encontradas en las muestras analizadas.

En este estudio, las muestras tumorales corresponden a las regiones AD o a las regiones ADK de los AD avanzados, las cuales se compararon con muestras de tejidos normales, no tumorales, extraídos de los mismos pacientes. Las SNV encontradas en todas las muestras habían sido previamente anotadas y juntadas en un único archivo llamado ***combined.annotated.bed.xlsx***, el cual ha sido analizado, mediante un script en R, para llevar a cabo el análisis de SNV en este estudio. El archivo ***combined.annotated.bed.xlsx*** contiene tantas filas como SNV identificadas. En cuanto a las columnas, éstas se pueden dividir en 3 secciones:

- Las primeras 12 columnas contienen información sobre el cromosoma y la posición en la que se encuentra la variante, el identificador de la variante, el valor de los alelos de referencia (normal) y alternativo (tumoral), el *quality score*, el resultado del filtrado (*pass* o *fail*), así como información adicional sobre la variante.
- Las siguientes 67 columnas indican si la variante se encuentra o no en las distintas muestras de estudio (23 AD, 23 ADK y 21 normales).
- Las últimas 17 columnas contienen la anotación de las variantes. De estas últimas columnas, destacan las columnas *Consequence*, que describe la consecuencia de la variante en la secuencia proteica correspondiente, y *Extra*, que contiene los campos *SYMBOL* (símbolo del gen en el que se encuentra la variante) e *IMPACT* (grado de impacto de la consecuencia de la variante: *LOW*, *MODERATE*, *HIGH* o *MODIFIER*).

Los pasos que se siguieron para el análisis de SNV fueron los siguientes:

- Se eliminaron las variantes ubicada en los cromosomas X e Y.
- Se eliminaron las variantes con un valor en la columna *Consequence* igual a *synonymous_variant*. Las SNV sinónimas son aquellas que no modifican la secuencia de aminoácidos que producen y, generalmente, no tienen implicaciones funcionales.
- Se eliminaron las variantes con un valor en el campo *IMPACT* igual a *MODIFIER* o *LOW*, con lo que el análisis se centró en las variantes con un impacto moderado y alto.
- Se eliminaron las variantes que aparecían en muestras normales, manteniendo las variantes que solamente aparecen en las muestras AD, las muestras ADK o ambas.
- Se eliminaron variantes duplicadas. Se consideraron a tal efecto las variantes que tenían exactamente los mismos valores en las columnas 1 a 79 (descritas arriba).

Una vez realizados todos los pasos anteriores, se procedió a calcular la frecuencia relativa de cada SNV tanto en las muestras AD como en las muestras ADK.

3.3. Resultados

Inicialmente, se identificaron un total de 6643 SNV que afectaban, como mínimo, a una de las muestras AD o ADK y de las cuales 792 eran sinónimas y 5851 no sinónimas. De las SNV no sinónimas, 892 tenían un impacto moderado o alto, y de éstas se identificaron 250 que, además, aparecían solamente en muestras AD, muestras ADK o ambas, pero no aparecían en muestras normales. Finalmente, tras eliminar variantes duplicadas, se obtuvo un conjunto final de 229 SNV no sinónimas, con un impacto moderado o alto y asociadas únicamente a muestras AD y/o ADK. En la Figura 18 se puede observar la frecuencia relativa de las 229 SNV identificadas, de forma separada para las muestras AD y las muestras ADK, y por cromosoma.

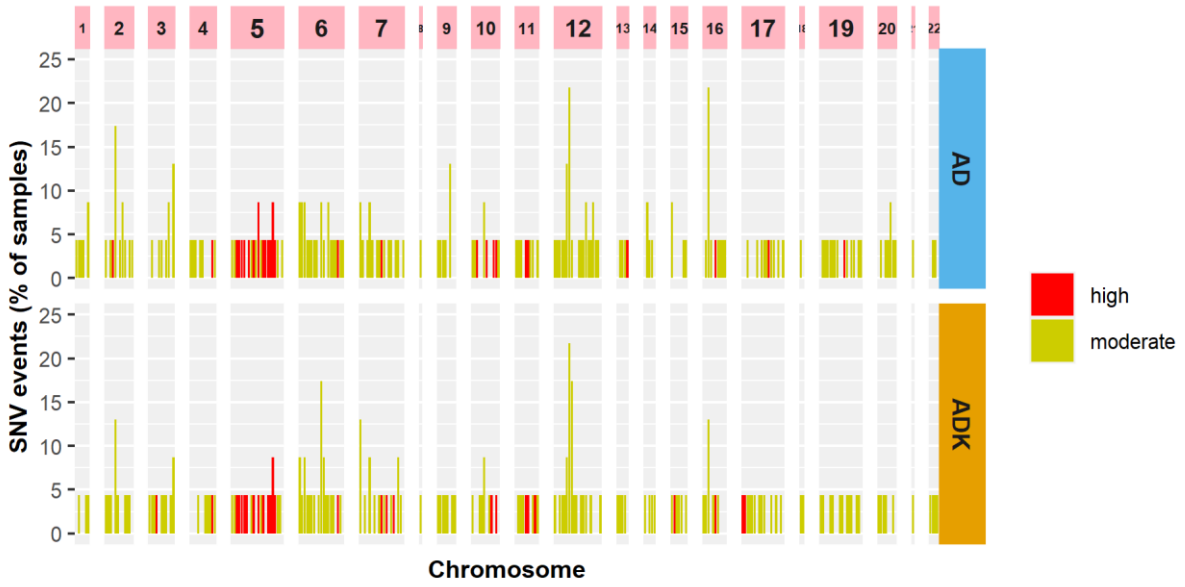

Figura 18. *Frecuencia relativa de las SNV por tipo de muestra y cromosoma.*

Se observa que, en general, la frecuencia de las SNV es baja, pues la mayoría solamente aparecen en una sola muestra. También se observa que se ha identificado un mayor número de SNV con un impacto moderado (193) que con un alto impacto (36). La tabla 7 muestra los genes correspondientes a las SNV con un alto impacto, el cromosoma en el que se encuentran y la frecuencia con la que aparecen las variantes tanto en las muestras AD como en las muestras ADK.

Tabla 7. *Genes asociados a SNV con un alto impacto*

Gen	Cromosoma	Frecuencia AD (%)	Frecuencia ADK (%)
ALK	2	4,3	
PBRM1	3		4,3
FBXW7	4	4,3	4,3
APC	5	4,3	4,3
APC	5	4,3	4,3
APC	5	4,3	4,3
APC	5	4,3	4,3
APC	5		4,3
APC	5	4,3	
APC	5	4,3	4,3

APC	5	8,7	4,3
APC	5	4,3	4,3
APC	5	4,3	
APC	5	4,3	4,3
APC	5	4,3	4,3
APC	5	8,7	8,7
APC	5	4,3	4,3
EPHA7	6	4,3	4,3
KMT2C	7	4,3	4,3
KMT2C	7		4,3
KMT2C	7		4,3
GATA3	10	4,3	
PTEN	10	4,3	
TCF7L2	10		4,3
TCF7L2	10	4,3	
TCF7L2	10	4,3	4,3
ATM	11	4,3	4,3
ATM	11	4,3	4,3
KMT2A	11		4,3
RB1	13	4,3	
MGA	15		4,3
CDH1	16	4,3	4,3
TP53	17		4,3
TP53	17		4,3
NCOR1	17	4,3	
NUMBL	19	4,3	

De entre las SNV con un alto impacto, destacan las 14 que afectan al gen *APC* ubicado en el cromosoma 5. De éstas, la mayoría aparecen tanto en las muestras AD como en las muestras ADK, y con la misma frecuencia. También cabe destacar las SNV con un alto impacto que aparecen solamente en muestras ADK, las cuales afectan a los genes *PBRM1*, *KMT2A*, *MGA* y *TP53*.

En cuanto a las SNV con un impacto moderado, las que aparecen en más de una muestra, tanto en las muestras AD como en las muestras ADK, son las que afectan a los siguientes genes: *PCBP1*, *PIK3CA*, *HLA-A*, *HLA-B*, *ZNF853*, *BRAF*, *NODAL*, *KRAS* y *ZNF764*. Otras SNV con un impacto moderado a destacar son las que aparecen solamente en muestras ADK, como las que afectan a los genes que se muestran en la tabla 8.

Tabla 8. *Genes asociados a SNV de impacto moderado que aparecen únicamente en muestras ADK*

Gen	Cromosoma	Gen	Cromosoma
GPR37L1	1	*FAT3*	11
DNMT3A	2	***KMT2D***	12
PCBP1	2	*C12orf49*	12
MYO7B	2	*FLT3*	13
TATDN2	3	*DHRS4L2*	14
RBM5	3	*PACS2*	14
ATR	3	***MGA***	**15**
PIK3CA	3	*TSC2*	16
PDGFRA	4	*CCDC102A*	16
KDR	4	***TP53***	**17**
BDH2	4	*CHD3*	17
FBXW7	4	*CDK12*	17
NSD1	5	*BRCA1*	17
HLA-B	6	***SMAD4***	18

MET	7	*ABHD17A*	19
OR10AC1P	7	*LGALS13*	19
KMT2C	7	*SERTAD1*	19
TAF1L	9	*ARHGAP35*	19
LRSAM1	9	*TMEM150B*	19
cds_end_NF	9	*CST9*	20
RET	10	*BPIFB3*	20
GPRIN2	10	*MN1*	22
FBXW4	10	*NCAPH2*	22

En la tabla anterior, se observa que hay SNV de impacto moderado que afectan a los genes *MGA* y *TP53*, los cuales también están afectados por SNV de alto impacto que solamente aparecen en muestras ADK.

La frecuencia alélica de las SNV (VAF, del inglés *variant allele frequency*) mide cuán frecuente es una variante en una determinada población de estudio. Este parámetro puede utilizarse para estudiar la heterogeneidad de un tumor, lo que se conoce como análisis de clonalidad. Analizar la clonalidad de un tumor consiste en determinar tanto la fracción de células cancerosas presentes en el tumor, como el número y la fracción de las distintas subpoblaciones de células tumorales, lo que se conoce en inglés como *cancer cell fraction* (CCF) [48], [49]. Conocer la heterogeneidad de un tumor a través del análisis de clonalidad es importante para entender mejor la posible progresión del tumor y establecer los tratamientos.

El cálculo de las VAF es solamente el primer paso para analizar clonalidad de un tumor. Luego, a partir de las VAF, puede estimarse la CCF. No obstante, la conversión de las VAF a CCF no es trivial ni directa, pues las VAF dependen tanto de la pureza del tumor (porcentaje de células tumorales en la muestra analizada) como del número de copias en la región en la que aparecen las variantes.

En este estudio, solamente se inició el análisis de clonalidad representando las VAF de una parte de las muestras disponibles (Figura 19). En este análisis no se incluyeron las muestras 6, 13 y 20 al no disponer de sus correspondientes archivos *vcf*, que contienen la información relativa a las VAF.

Como se puede observar, hay una gran variabilidad de VAF entre distintas muestras. Idealmente, las SNV tendrían una VAF de 0,5, ya que son mayoritariamente heterocigóticas. Sin embargo, eso equivaldría a que el tumor tuviera una pureza del 100%, hecho que es prácticamente imposible debido tanto a la contaminación de células normales que tienen todas las muestras de tumores como a la presencia de otros tipos celulares no tumorales (células inmunitarias, fibroblastos, etc.). De hecho, se puede ver que los picos máximos en muchas de las muestras están muy por debajo de 0,5, lo que significa que hay una baja pureza en algunas muestras y que la corrección por “pureza tumoral” sería imprescindible para poder hacer estimaciones de clonalidad. En algunos perfiles, también se observa que aparece más de un pico máximo, lo cual es indicativo de la presencia de distintos subclones.

Así, para seguir con el análisis de clonalidad, debería estimarse la pureza de las muestras para poder corregir los valores VAF (obviar la contaminación no

tumoral), para luego integrar esa información con la del número de copia y poder estimar la CCF. Sin embargo, como se comenta en las conclusiones de este trabajo, el análisis de clonalidad no se ha podido completar.

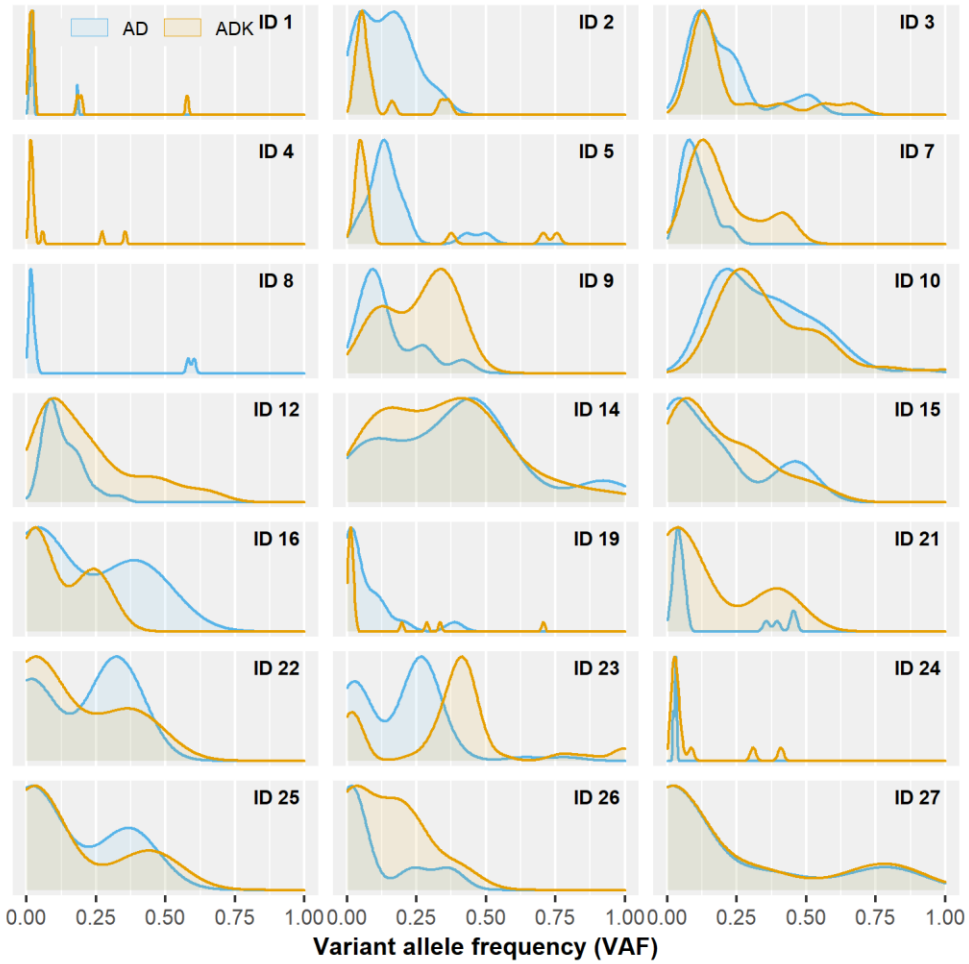

Figura 19. *VAF de las SNV identificadas en regiones AD y regiones ADK de AD avanzados.*

3.4. Discusión

En este capítulo se llevado a cabo un análisis de las SNV detectadas a partir de datos de *targeted sequencing* obtenidos de regiones de AD y regiones de ADK de AD avanzados, con el objetivo de identificar qué SNV están asociadas al CCR y, en especial, a la transición AD-ADK.

Cabe avanzar que los resultados obtenidos en este capítulo concuerdan con los reportado en la literatura sobre SNV en el CCR [19]–[21]. Algunas de las SNV identificadas con mayor frecuencia, tanto en las muestras AD como en las muestras ADK, han sido las que afectan al gen *APC*, ubicado en el cromosoma 5q. Esta mutación es bien conocida en el CCR y ya fue incluida en el primer modelo molecular de progresión AD-ADK (Figura 1) [9]. El *APC* es un gen supresor de tumores que se ve afectado en las primeras etapas del desarrollo del CCR, de ahí que aparezca tanto en las muestras AD como en las ADK.

Otras de las SNV ampliamente reportadas en la literatura y que se han identificado tanto en muestras AD como en muestras ADK son las que afectan a los genes *KRAS* (12p) y *BRAF* (7q). Las mutaciones de estos dos oncogenes, relacionados principalmente con la reproducción y la diferenciación celular, también aparecen pronto en la progresión AD-ADK, por lo que también se han identificado en los dos tipos de muestras.

A diferencia de los genes *APC*, *KRAS* o *BRAF*, los genes *TP53* (17p) y *SMAD4* (18q), suelen verse afectados en una etapa más avanzada en la progresión AD-ADK. Es por ello que las SNV asociadas a estos genes supresores de tumores solamente se han encontrado en muestras ADK y no en muestras AD.

Por último, son destacables también algunas SNV con alto impacto que solamente se han encontrado en las muestras ADK, como las que afectan a los genes *PBRM1* (3p) y *MGA* (15q). Se trata de dos genes supresores de tumores que, aunque de manera menos frecuente, también se han relacionado con el CCR [50]–[52].

4. Conclusiones

En este estudio se ha llevado a cabo un análisis de CNA y SNV a partir de datos LP-WGS y de *targeted sequencing*, respectivamente, obtenidos de regiones de AD y regiones de ADK de muestras de AD colorrectales avanzados. Las conclusiones derivadas de este estudio se han dividido en dos grupos. En primer lugar, se enumeran aquellas relacionadas con aspectos biológicos:

- Se ha encontrado una mayor frecuencia de *broad* CNA en las regiones ADK que en las regiones AD de los AD avanzados analizados. Las *broad* CNA identificadas de manera más recurrente han sido las pérdidas en 18p y 18q y las ganancias en 7p, 7q, 8q, 13q, 20p y 20q. Otras *broad* CNA que se han identificado con menor frecuencia han sido las pérdidas en 5q, 8p, 14q y 17p, así como la ganancia en 9p.
- La diferencia en el BCS entre regiones AD y regiones ADK de AD avanzados indica que las regiones ADK tienen una fracción mayor del genoma afectado por *broad* CNA, lo cual concuerda con el hecho de que las regiones ADK presentan un mayor número de *broad* CNA que las regiones AD.
- Se han encontrado diferencias remarcables entre las regiones AD y las regiones ADK de AD avanzados en los brazos 5q, 7p, 7q, 13q, 14q, 18p, 20p y 20q. Estas alteraciones podrían estar asociadas a la progresión AD-ADK. De hecho, en esos brazos cromosómicos se han identificado *gene sets* relacionados con la respuesta inmune y el metabolismo celular, dos procesos biológicos alterados en el desarrollo del cáncer.
- Se han encontrado diferencias significativas entre regiones AD y regiones ADK de AD avanzados en las *focal* CNA detectadas en la región del cromosoma 8 que va de 116 a 135 Mb, siendo mayor la frecuencia de las *focal* CNA en las regiones ADK. En esa región del cromosoma 8 se han identificado genes que aparecen normalmente desregulados en distintos tipos de células del sistema inmunitario: células NK, células dendríticas con bloqueo del IFNAR1 y macrófagos asociados a tumores, lo cual favorece la proliferación tumoral.
- Las SNV identificadas con mayor frecuencia, tanto en las regiones AD como en las regiones ADK de AD avanzados, han sido las que afectan a los genes *APC*, *KRAS* y *BRAF*. Estas mutaciones son bien conocidas en el CCR y aparecen en las primeras etapas de la progresión AD-ADK, de ahí que aparezcan tanto en las regiones AD como en las regiones ADK.
- Las SNV que afectan a los genes *TP53* y *SMAD4* solamente se han encontrado en las regiones ADK de AD avanzados, lo cual concuerda con el modelo de inestabilidad cromosómica de la progresión AD-ADK, en el que estas mutaciones se asocian a una etapa más avanzada del proceso de carcinogénesis.
- Otras SNV con un alto impacto que solamente se han encontrado en las regiones ADK de AD avanzados son las que afectan a los genes *PBRM1* y *MGA*, dos genes supresores de tumores que, aunque de manera menos frecuente, también se han relacionado con el CCR.

- La heterogeneidad inter-tumoral e intra-tumoral hace que el análisis genómico de los tumores sea una tarea compleja. Analizar la heterogeneidad del CCR implica integrar la información de las distintas alteraciones que se producen durante la progresión AD-ADK, tanto a nivel de CNA como a nivel de SNV, lo cual es fundamental para lograr entender la progresión del tumor y plantear posibles estrategias de tratamiento.

En segundo lugar, se enumeran las conclusiones relacionadas con aspectos técnicos y metodológicos:

- A partir de los datos LP-WGS se han podido identificar todas las *broad* CNA que se habían detectado previamente con FISH. No obstante, el LP-WGS ha permitido identificar otras *broad* CNA que no se detectaron con FISH, como las que afectan a los brazos 7q, 8p, 9p, 14q, 17p, 18p y 20p.
- El LP-WGS tiene las principales ventajas del WGS: una alta resolución en la detección de alteraciones genómicas y la posibilidad de detectar nuevas alteraciones no conocidas, pero a un precio más asequible que el WGS, por la baja profundidad de cobertura que emplea. Dado que en el caso de las *broad* CNA se analizan regiones genómicas grandes, el efecto de la baja cobertura del LP-WGS no es crítico y se ha demostrado que esta técnica permite identificarlas correctamente. Por lo tanto, el LP-WGS es una alternativa coste-efectiva a las técnicas basadas en arrays para la detección de CNA.
- El análisis de *focal* CNA, al centrarse en regiones genómicas pequeñas, es más sensible a los errores sistemáticos de secuenciación que pueden ocasionarse por la baja profundidad de cobertura del LP-WGS, pudiendo provocar la detección de falsas CNA.
- La detección de una mayoría de *focal* CNA cerca de las regiones centroméricas y teloméricas de los cromosomas, en las cuales hay un mayor porcentaje GC, indica que el porcentaje GC es un factor crítico en el análisis de CNA a partir de datos LP-WGS, especialmente en el caso de las *focal* CNA. Por lo tanto, es importante llevar a cabo una buena normalización de los datos para corregir el efecto del porcentaje GC y eliminar regiones con una cobertura sistemáticamente mayor o menor que la del resto de regiones.
- La combinación de diferentes técnicas de secuenciación permite la detección de distintos tipos de alteraciones genéticas y, por consiguiente, aporta una visión más completa y una mejor comprensión de la genética del CCR.

En cuanto al cumplimiento de los objetivos planteados inicialmente en este estudio, cabe destacar que se ha logrado cumplir la mayor parte de ellos. No obstante, debido a la falta de disponibilidad de los archivos BAM, no se ha podido analizar los datos con la herramienta ichorCNA. A nivel metodológico, este era un aspecto interesante, pues ichorCNA es una herramienta diseñada expresamente para analizar CNA a partir de datos LP-WGS, lo cual era el objetivo principal de este estudio. Además, ichorCNA también permite hacer de manera fácil y automática un análisis de clonalidad. Por lo tanto, la falta de acceso de los archivos BAM ha ocasionado también que no se haya podido realizar el análisis de clonalidad.

Con respecto a la planificación y la metodología propuestas al inicio de este trabajo, se considera que ambas han sido adecuadas. Con respecto a la planificación, se propuso una correcta temporización de las tareas a realizar, con tiempo suficiente para llevar a cabo cada una de ellas y dejando un margen suficiente para reaccionar correctamente frente a posibles contratiempos. También cabe destacar que la previsión de riesgos y las medidas de contingencia que se plantearon inicialmente han resultado ser efectivas, pues contemplaban, por ejemplo, la imposibilidad de disponer de los archivos BAM, tal y como se ha comentado anteriormente. Eso ha hecho que, frente a ese contratiempo, el trabajo no se haya visto afectado y haya habido una buena y rápida actuación, adaptando los objetivos y las tareas a realizar. En particular, siguiendo las medidas de contingencia que se propusieron, se decidió no continuar con el análisis de clonalidad, profundizar en el análisis de CNA y dedicar más tiempo al análisis de SNV. Si bien es cierto que el análisis de clonalidad podría haberse realizado, a pesar de no haber podido utilizar ichorCNA, esa opción no estaba contemplada en la planificación inicial y hubiera requerido de más tiempo para poder llevarse a cabo. En cuanto a la metodología propuesta, ésta también ha sido correcta. En concreto, no ha habido ningún problema en cuanto a las herramientas de análisis, pues tanto CNApp como R son dos herramientas de fácil acceso.

Algunas de las limitaciones de este estudio no suponen sino oportunidades de trabajo futuro. La primera y más importante es llevar a cabo el análisis de clonalidad que no se ha podido realizar en este estudio. Por otro lado, también sería interesante explorar el uso de la herramienta ichorCNA para analizar las CNA y realizar el análisis de clonalidad. En cuanto a aspectos metodológicos, sería conveniente mejorar la normalización de los datos, siguiendo los pasos que hay descritos en la literatura para analizar CNA a partir de datos de secuenciación masiva. También sería conveniente aumentar la muestra de estudio, pues el número de muestras disponibles en este estudio ha limitado el análisis estadístico de los datos. Por último, se podría profundizar en el análisis e interpretación biológica de los resultados encontrados en este estudio.

Para terminar, me gustaría aportar mi valoración personal sobre este proyecto. La elección de este proyecto se me planteó desde un principio, por decisión propia, como un reto personal y profesional. En el campo personal, la temática del proyecto era completamente nueva para mí. Esto, si bien ha sido una motivación a lo largo de todo el proyecto, también implicaba más tiempo de dedicación, sobre todo a nivel de revisión bibliográfica y asimilación de conceptos básicos. También hay que añadir que la complicada situación a nivel global en la que nos encontramos, debido a la pandemia, ha complicado en ciertos momentos el trabajo diario en este proyecto. A nivel profesional, el reto era aprender nuevas técnicas y herramientas de análisis que pudieran ser de utilidad en mi trabajo diario y/o me permitieran el acceso a nuevos perfiles profesionales. Tras la finalización del proyecto, he de decir que acabo con la sensación de tener una visión completa y global de la temática tratada, de haber podido profundizar en un tema que, desde un principio, me ha parecido apasionante y con la satisfacción de haber podido completar la mayor parte de los objetivos definidos.

5. Glosario

AD: adenoma

ADK: adenocarcinoma

ADN: ácido desoxirribonucleico

BAM: *binary alignment map*

CBS: *circular binary segmentation*

CGH: *comparative genomic hybridization*

CIN: *chromosomal instability*

CNA: *copy number alteration*

FISH: *fluorescence in situ hybridization*

GSEA: *gene set enrichment analysis*

LP-WGS: *low-pass whole genome sequencing*

NGS: *next generation sequencing*

PCR: *polymerase chain reaction*

RC: *read count*

SNP: *single nucleotide polimorfism*

SNV: *single nucleotide variant*

VAF: *variant allele frequency*

WGS: *whole genome sequencing*

6. Bibliografía

[1] "Globocan 2018," 2019. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf.

[2] "Incidencia y mortalidad de cáncer colorrectal en España en la población entre 50 y 69 años," 2019. [Online]. Available: https://www.aecc.es/sites/default/files/content-file/Incidencia-mortalidad-colon2018_CCAAyProvincias.pdf.

[3] "Cancer Facts & Figures 2020," 2020. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.

[4] J. C. Clark *et al.*, "Prevalence of polyps in an autopsy series from areas with varying incidence of large- bowel cancer," *Int. J. Cancer*, vol. 36, no. 2, pp. 179–186, Aug. 1985.

[5] J. DiSario, P. Foutch, H. Mai, K. Pardy, and R. Manne, "Prevalence and malignant potential of colorectal polyps in asymptomatic, average-risk men," *Am. J. Gastroenterol.*, vol. 86, no. 8, pp. 941–945, 1991.

[6] D. A. Lieberman, D. G. Weiss, J. H. Bond, D. J. Ahnen, H. Garewal, and G. Chejfec, "Use of colonoscopy to screen asymptomatic adults for colorectal cancer," *N. Engl. J. Med.*, vol. 343, no. 3, pp. 162–168, Jul. 2000.

[7] R. J. Davies, R. Miller, and N. Coleman, "Colorectal cancer screening: Prospects for molecular stool analysis," *Nat. Rev. Cancer*, vol. 5, no. 3, pp. 199–209, Mar. 2005.

[8] T. Muto, H. J. R. Bussey, and B. C. Morson, "The evolution of cancer of the colon and rectum," *Cancer*, vol. 36, no. 6, pp. 2251–2270, 1975.

[9] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–767, Jun. 1990.

[10] F. D. E. De Palma, V. D'argenio, J. Pol, G. Kroemer, M. C. Maiuri, and F. Salvatore, "The molecular hallmarks of the serrated pathway in colorectal cancer," *Cancers (Basel).*, vol. 11, no. 7, Jul. 2019.

[11] L. Sansregret, B. Vanhaesebroeck, and C. Swanton, "Determinants and clinical implications of chromosomal instability in cancer," *Nature Reviews Clinical Oncology*, vol. 15, no. 3. Nature Publishing Group, pp. 139–150, 01-Mar-2018.

[12] Y. C. Tang and A. Amon, "Gene copy-number alterations: A cost-benefit analysis," *Cell*, vol. 152, no. 3. Cell, pp. 394–405, 31-Jan-2013.

[13] B. A. Weaver and D. W. Cleveland, "Does aneuploidy cause cancer?," *Current Opinion in Cell Biology*, vol. 18, no. 6. Elsevier Ltd, pp. 658–667, 01-Dec-2006.

[14] R. Beroukhim *et al.*, "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, no. 7283, pp. 899–905, Feb. 2010.

[15] J. Camps *et al.*, "Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer," *Genes Chromosom. Cancer*, vol. 48, no. 11, pp. 1002–1017, Nov. 2009.

[16] M. Hermsen *et al.*, "Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability," *Gastroenterology*, vol. 123, no. 4, pp. 1109–1119, Oct. 2002.

[17] C. B. Diep, K. Kleivi, F. R. Ribeiro, M. R. Teixeira, O. C. Lindgjærde, and R. A. Lothe, "The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes," *Genes Chromosom. Cancer*, vol. 45, no. 1, pp. 31–41, 2006.

[18] B. Carvalho *et al.*, "Evaluation of cancer-associated DNA copy number events in colorectal (advanced) adenomas," *Cancer Prev. Res.*, vol. 11, no. 7, pp. 403–411, Jul. 2018.

[19] S. H. Zaidi *et al.*, "Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020.

[20] S. Mamlouk *et al.*, "DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer," *Nat. Commun.*, vol. 8, Jan. 2017.

[21] W. Cross *et al.*, “The evolutionary landscape of colorectal tumorigenesis,” *Nat. Ecol. Evol.*, vol. 2, no. 10, pp. 1661–1672, Oct. 2018.
[22] T. Xie *et al.*, “Patterns of somatic alterations between matched primary and metastatic colorectal tumors characterized by whole-genome sequencing,” *Genomics*, vol. 104, no. 4, pp. 234–241, Oct. 2014.
[23] Z. Dong *et al.*, “Low-pass whole-genome sequencing in clinical cytogenetics: A validated approach,” *Genet. Med.*, vol. 18, no. 9, pp. 940–948, Sep. 2016.
[24] A. Chaubey *et al.*, “Low-Pass Genome Sequencing: Validation and Diagnostic Utility from 409 Clinical Cases of Low-Pass Genome Sequencing for the Detection of Copy Number Variants to Replace Constitutional Microarray,” *J. Mol. Diagnostics*, vol. 22, no. 6, pp. 823–840, Jun. 2020.
[25] “Low-Coverage Whole Genome Sequencing - National Cancer Institute.” [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/blog/2019/low-coverage-seq. [Accessed: 24-Dec-2020].
[26] “The Cost of Sequencing a Human Genome.” [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost. [Accessed: 26-Dec-2020].
[27] E. Heitzer *et al.*, “Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing,” *Genome Med.*, vol. 5, no. 4, Apr. 2013.
[28] V. A. Adalsteinsson *et al.*, “Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors,” *Nat. Commun.*, vol. 8, no. 1, Dec. 2017.
[29] B. Zhou, S. S. Ho, X. Zhang, R. Pattni, R. R. Haraksingh, and A. E. Urban, “Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis,” *J. Med. Genet.*, vol. 55, no. 11, pp. 735–743, Nov. 2018.
[30] N. Van Roy *et al.*, “Shallow whole genome sequencing on circulating cell-free DNA allows reliable noninvasive copy-number profiling in neuroblastoma patients,” *Clin. Cancer Res.*, vol. 23, no. 20, pp. 6305–6315, Oct. 2017.
[31] L. C. Xia *et al.*, “Whole genome analysis identifies the association of TP53 genomic deletions with lower survival in Stage III colorectal cancer,” *Sci. Rep.*, vol. 10, no. 1, Dec. 2020.
[32] I. Quintanilla Leo, “New insights into the colorectal carcinogenesis: from early precursor lesions to the role of aneuploidy,” Universitat de Barcelona, 2017.
[33] S. Franch-Expósito *et al.*, “CNApp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications,” *Elife*, vol. 9, Jan. 2020.
[34] J. Kendall and A. Krasnitz, “Computational Methods for DNA Copy-Number Analysis of Tumors,” in *Methods in molecular biology (Clifton, N.J.)*, vol. 1176, NIH Public Access, 2014, pp. 243–259.
[35] A. Magi, L. Tattini, T. Pippucci, F. Torricelli, and M. Benelli, “Read count approach for DNA copy number variants detection,” *Bioinformatics*, vol. 28, no. 4, pp. 470–478, Feb. 2012.
[36] D. M. Muzny *et al.*, “Comprehensive molecular characterization of human colon and rectal cancer,” *Nature*, vol. 487, no. 7407, pp. 330–337, Jul. 2012.
[37] B. Molparia, G. Oliveira, J. L. Wagner, E. G. Spencer, and A. Torkamani, “A feasibility study of colorectal cancer diagnosis via circulating tumor DNA derived CNV detection,” *PLoS One*, vol. 13, no. 5, p. e0196826, May 2018.
[38] M. Kucharík, J. Budiš, M. Hýblová, 4, G. Minárik, and T. Szemes, “Copy number variant detection with low-coverage whole-genome sequencing is a viable replacement for the traditional array-CGH,” *medRxiv*, p. 2020.09.07.20183665, Sep. 2020.
[39] A. Leslie *et al.*, “Chromosomal changes in colorectal adenomas: Relationship to gene mutations and potential for clinical utility,” *Genes Chromosom. Cancer*, vol. 45, no. 2, pp. 126–135, Feb. 2006.
[40] G. A. Andriani *et al.*, “A direct comparison of interphase FISH versus low-coverage

single cell sequencing to detect aneuploidy reveals respective strengths and weaknesses," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019.
[41] D. Niu *et al.*, "Evaluation of Next Generation Sequencing for Detecting HER2 Copy Number in Breast and Gastric Cancers," *Pathol. Oncol. Res.*, vol. 26, no. 4, pp. 2577–2585, Oct. 2020.
[42] T. Xie *et al.*, "A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations," *PLoS One*, vol. 7, no. 7, p. 42001, Jul. 2012.
[43] J. L. Markman and S. L. Shiao, "Impact of the immune system and immunotherapy in colorectal cancer," *J. Gastrointest. Oncol.*, vol. 6, no. 2, pp. 208–223, 2015.
[44] H. Gonzalez, C. Hagerling, and Z. Werb, "Roles of the immune system in cancer: From tumor initiation to metastatic progression," *Genes and Development*, vol. 32, no. 19–20. Cold Spring Harbor Laboratory Press, pp. 1267–1284, 2018.
[45] R. E. Brown, S. P. Short, and C. S. Williams, "Colorectal Cancer and Metabolism," *Current Colorectal Cancer Reports*, vol. 14, no. 6. Current Medicine Group LLC 1, pp. 226–241, 01-Dec-2018.
[46] S. La Vecchia and C. Sebastián, "Metabolic pathways regulating colorectal cancer initiation and progression," *Seminars in Cell and Developmental Biology*, vol. 98. Elsevier Ltd, pp. 63–70, 01-Feb-2020.
[47] K. Cibulskis *et al.*, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nat. Biotechnol.*, vol. 31, no. 3, pp. 213–219, Mar. 2013.
[48] L. Oesper, G. Satas, and B. J. Raphael, "Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data," *Bioinformatics*, vol. 30, no. 24, pp. 3532–3540, Dec. 2014.
[49] M. Cmero *et al.*, "Inferring structural variant cancer cell fraction," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020.
[50] Y. S. Jo, M. S. Kim, N. J. Yoo, and S. H. Lee, "Somatic mutation of a candidate tumour suppressor MGA gene and its mutational heterogeneity in colorectal cancers," *Pathology*, vol. 48, no. 5. Elsevier B.V., pp. 525–527, 01-Aug-2016.
[51] H. Mathsyaraja *et al.*, "Loss of MGA mediated Polycomb repression promotes tumor progression and invasiveness," *bioRxiv*, p. 2020.10.16.334714, Oct. 2020.
[52] X. S. Shu *et al.*, "The epigenetic modifier PBRM1 restricts the basal activity of the innate immune system by repressing retinoic acid-inducible gene-I-like receptor signalling and is a potential prognostic biomarker for colon cancer," *J. Pathol.*, vol. 244, no. 1, pp. 36–48, Jan. 2018.

7. Anexos

7.1. Anexo A

- **Códigos R**:
  - **Archivo "cna_analysis.R"**: para realizar el análisis de CNA.
  - **Archivo "target_analysis.R"**: para realizar el análisis de SNV.
- **Archivos de datos**:
  - **Carpeta "DNAcopy_results"**: contiene los archivos con los valores *seg.mean* de las distintas muestras, necesarios para el análisis de CNA.
  - **Carpeta "target_seq"**: contiene los archivos *vcf* con las SNV identificas en cada muestra. Se utilizan para calcular las VAF.
  - **Carpeta "Res_no_seg"**
    - **Archivo "CNA_Scores.tsv"**: contiene los valores BCS y FCS de todas las muestras.
    - **Archivo "Re-segmented_samples.xlsx"**: contiene el listado completo de segmentos, con los correspondientes valores *seg.mean*, clasificados por CNApp para todas las muestras. Se utiliza para realizar el análisis global de CNA.
    - **Carpeta "Resultados broad"**
      - **Archivo "cna_profile_arms.tsv"**: contiene la matriz de valores *seg.mean* por brazo cromosómico y muestra, necesaria para el análisis de *broad* CNA.
      - **Archivos "genes_5q.txt", "genes_7p.txt", "genes_7q.txt", "genes_13q.txt", "genes_14q.txt", "genes_18p.txt", "genes_20p.txt" y "genes_20q.txt"**: contienen el listado de genes localizados en los correspondientes brazos cromosómicos. Se utiliza para el GSEA de las *broad* CNA.
      - **Archivo "GSEA_broad.xlsx"**: contiene el listado de *gene sets* identificados en el GSEA de las *broad* CNA.
    - **Carpeta "Resultados focal"**
      - **Archivo "cna_profile_1Mb.tsv"**: contiene la matriz de valores *seg.mean* por región genómica de 1 Mb y muestra, necesaria para el análisis de *focal* CNA.
      - **Archivo "genes.txt"**: contiene el listado de genes localizados en las regiones en las que se ha identificado diferencias remarcables en las *focal* CNA entre las regiones AD y las regiones ADK de AD avanzados. Se utiliza para el GSEA de las *focal* CNA.
  - **Archivo "data_all.txt"**: contiene el listado completo de segmentos, con los correspondientes valores *seg.mean*, de todas las muestras. Es el archivo de entrada a CNApp.
  - **Archivo "combined.annotated.bed.xlsx"**: contiene la información de las SNV identificadas en todas las muestras de estudio.
  - **Archivos "centromeres.txt" y "telomeres.txt"**: contienen las coordenadas de los centrómeros y telómeros de todos los cromosomas humanos.

- **Archivos “gc_chr2.txt”, “gc_chr4.txt”, “gc_chr7.txt”, “gc_chr8.txt”, “gc_chr11.txt”, “gc_chr12.txt”, “gc_chr13.txt” y “gc_chr20.txt”**: contienen los porcentajes GC, en ventanas de 1 Mb, de los cromosomas en los que se han identificado diferencias remarcables en las *focal* CNA entre las regiones AD y las regiones ADK de AD avanzados.

7.2. Anexo B

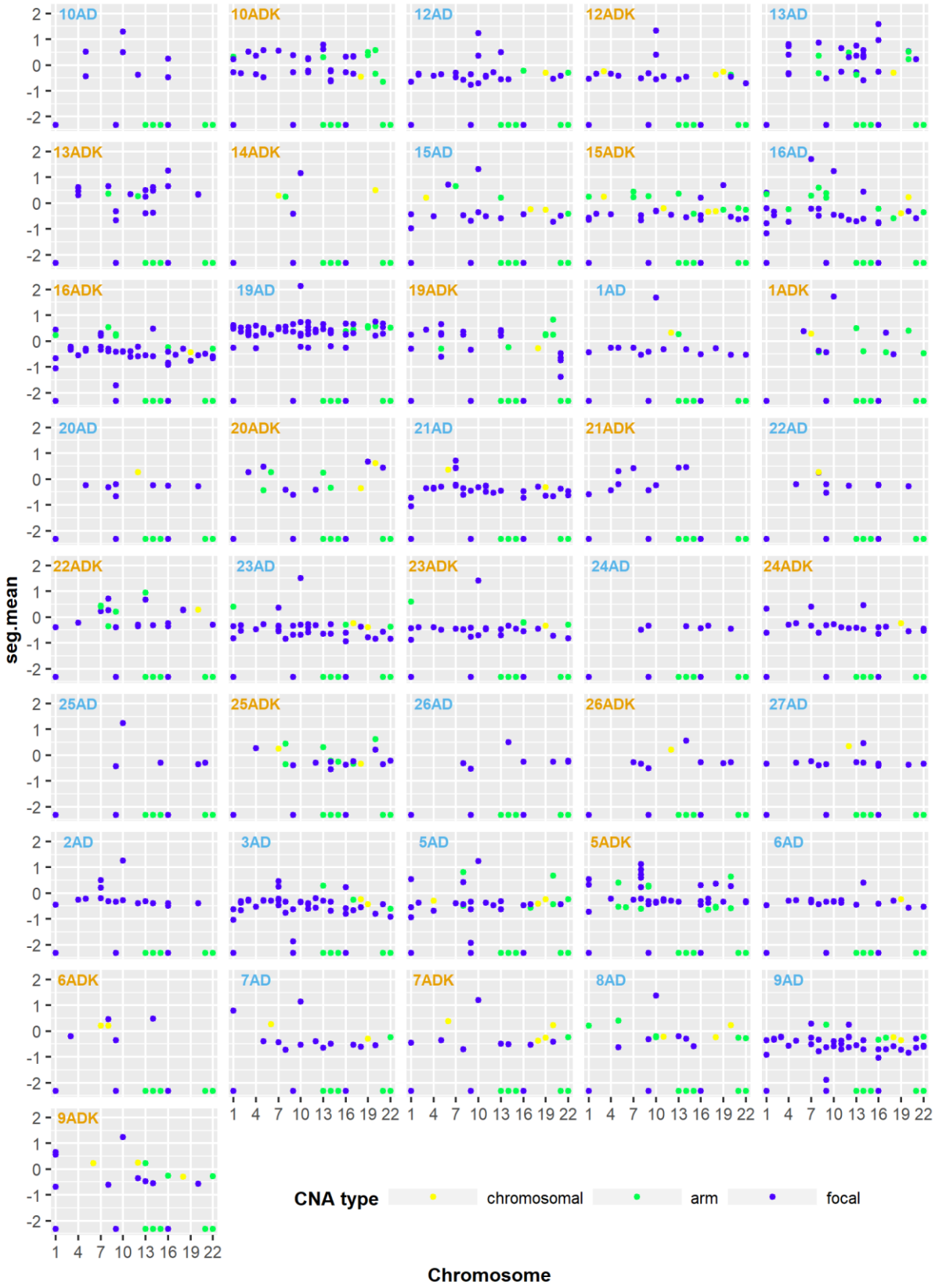

Figura 20. *Valores seg.mean de las CNA, por muestra, cromosoma y tipo de CNA.*

7.3. Anexo C

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
5q						
GO_HOMOPHILIC_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	168	The attachment of a plasma membrane adhesion molecule in one cell to an identical molecule in an adjacent cell.	60	0.3571	1.44E-48	Cell adhesion
GO_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	277	The attachment of one cell to another cell via adhesion molecules that are at least partially embedded in the plasma membrane.	62	0.2238	7.58E-37	Cell adhesion
GO_CELL_CELL_ADHESION	884	The attachment of one cell to another cell via adhesion molecules.	81	0.0916	1.85E-20	Cell adhesion
GO_BIOLOGICAL_ADHESION	1481	The attachment of a cell or organism to a substrate, another cell, or other organism. Biological adhesion includes intracellular attachment between membrane regions.	99	0.0668	1.18E-15	Cell adhesion
GO_CALCIUM_DEPENDENT_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_CELL_ADHESION_MOLECULES	43	The attachment of one cell to another cell via adhesion molecules that require the presence of calcium for the interaction.	14	0.3256	3.28E-9	Cell adhesion
GO_CELL_CELL_SIGNALING	1809	Any process that mediates the transfer of information from one cell to another. This process includes signal transduction in the receiving cell and, where applicable, release of a ligand and any processes that actively facilitate its transport and presentation to the receiving cell. Examples include signaling via soluble ligands, via cell adhesion molecules and via gap junctions.	92	0.0509	1.28E-7	Cell communication
GO_SYNAPTIC_SIGNALING	751	Cell-cell signaling to, from or within a synapse.	50	0.0666	6.83E-7	Cell communication
GO_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	1915	Any process that modulates the frequency, rate or extent of the covalent alteration of one or more amino acid residues within a protein.	91	0.0475	3.73E-6	Cellular metabolic process
GO_REGULATION_OF_PHOSPHORYLATION	1668	Any process that modulates the frequency, rate or extent of addition of phosphate groups into a molecule.	82	0.0492	4.84E-6	Cellular metabolic process
GO_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS	1868	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving phosphorus or compounds containing phosphorus.	88	0.0471	8.43E-6	Cellular metabolic process
GO_SYNAPSE_ASSEMBLY	184	The aggregation, arrangement and bonding together of a set of components to form a synapse. This process ends when the synapse is mature (functional).	20	0.1087	4.65E-5	Developmental process
GO_SYNAPSE_ORGANIZATION	433	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of a synapse, the junction between a neuron and a target (neuron, muscle, or secretory cell).	32	0.0739	5.82E-5	Developmental process
GO_TYROSINE_PHOSPHORYLATION_OF_STAT_PROTEIN	89	The process of introducing a phosphate group to a tyrosine residue of a STAT (Signal Transducer and Activator of Transcription) protein.	13	0.1461	3.21E-4	Cell communication

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_REGULATION_OF_ORGANELLE_ORGANIZATION	1310	Any process that modulates the frequency, rate or extent of a process involved in the formation, arrangement of constituent parts, or disassembly of an organelle.	63	0.0481	4.76E-4	Regulation of cellular process
GSE25123_CTRL_VS_ROSIGLITAZONE_STIM_PPARG_KO_MACROPHAGE_DN	199	Genes down-regulated in bone marrow-derived macrophages with PPARG [GeneID=5468] knockout: control versus rosiglitazone.	19	0.0955	5.69E-4	Immunity
GO_CELLULAR_AMIDE_METABOLIC_PROCESS	1192	The chemical reactions and pathways involving an amide, any derivative of an oxoacid in which an acidic hydroxy group has been replaced by an amino or substituted amino group, as carried out by individual cells.	58	0.0487	8.29E-4	Cellular metabolic process
GSE21063_WT_VS_NFATC1_KO_16H_ANTI_IGM_STIM_BCELL_DN	188	Genes down-regulated in B lymphocytes stimulated by anti-IgM for 16h: wildtype versus NFATC1 [GeneID=4772] knockout.	18	0.0957	9.08E-4	Immunity
GO_CELLULAR_RESPONSE_TO_ENDOGENOUS_STIMULUS	1406	Any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus arising within the organism.	65	0.0462	9.08E-4	Response to stimulus
GO_POSITIVE_REGULATION_OF_NUCLEOBASE_CONTAINING_COMPOUND_METABOLIC_PROCESS	1861	Any cellular process that activates or increases the frequency, rate or extent of the chemical reactions and pathways involving nucleobases, nucleosides, nucleotides and nucleic acids.	80	0.0430	9.08E-4	Cellular metabolic process
GO_RESPONSE_TO_ENDOGENOUS_STIMULUS	1662	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus arising within the organism.	73	0.0439	1.2E-3	Response to stimulus
GO_NEGATIVE_REGULATION_OF_RELAXATION_OF_MUSCLE	5	Any process that stops, prevents or reduces the frequency, rate or extent of relaxation of muscle.	4	0.8000	1.26E-3	Muscle system process
GSE14415_NATURAL_TREG_VS_TCONV_UP	157	Genes up-regulated in natural T reg versus T conv.	16	0.1019	1.26E-3	Immunity
GSE29618_PRE_VS_DAY7_FLU_VACCINE_BCELL_DN	196	Genes down-regulated in comparison of B cells from influenza vaccinee pre-vaccination versus those at day 7 post-vaccination.	18	0.0918	1.28E-3	Immunity
GSE11961_PLASMA_CELL_DAY7_VS_GERMINAL_CENTER_BCELL_DAY40_UP	200	Genes up-regulated in day 7 plasma cells versus day 40 germinal center B cells.	18	0.0900	1.45E-3	Immunity
GSE22886_TH1_VS_TH2_48H_ACT_DN	200	Genes down-regulated in comparison of stimulated CD4 [GeneID=920] Th1 cells at 48 h versus stimulated CD4 [GeneID=920] Th2 cells at 48 h.	18	0.0900	1.45E-3	Immunity
GSE23398_WT_VS_IL2_KO_CD4_TCELL_SCURFY_MOUSE_DN	200	Genes down-regulated in lymph node CD4 [GeneID=920] T cells: scurfy (non-functional form of FOXP3 [GeneID=50943]) versus scurfy and IL2 [GeneID=3558] knockout.	18	0.0900	1.45E-3	Immunity
GSE24972_MARGINAL_ZONE_BCELL_VS_FOLLICULAR_BCELL_IRF8_KO_UP	200	Genes up-regulated in spleen B lymphocytes with IRF8 [GeneID=3394] knockout: marginal zone versus follicular.	18	0.0900	1.45E-3	Immunity

GO_POSITIVE_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT	93	Any process that activates or increases the frequency, rate or extent of receptor signaling pathway via STAT.	12	0.1290	1.64E-3	Cell communication

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_CELL_JUNCTION_ORGANIZATION	713	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of a cell junction. A cell junction is a specialized region of connection between two cells or between a cell and the extracellular matrix.	39	0.0547	2.02E-3	Cellular component organization
GO_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT	151	Any process that modulates the frequency, rate or extent of receptor signaling via STAT.	15	0.0993	2.64E-3	Cell communication
GO_RESPONSE_TO_ORGANIC_CYCLIC_COMPOUND	946	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an organic cyclic compound stimulus.	47	0.0497	2.69E-3	Cellular metabolic process
GO_NEGATIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	1209	Any process that stops, prevents, or reduces the frequency, rate or extent of chemical reactions and pathways involving a protein.	56	0.0463	2.76E-3	Cellular metabolic process
GO_POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	1853	Any process that activates or increases the frequency, rate or extent of an organismal process, any of the processes pertinent to the function of an organism above the cellular level; includes the integrated processes of tissues and organs.	77	0.0416	2.78E-3	Developmental process
GO_CELL_CYCLE_PROCESS	1422	The cellular process that ensures successive accurate and complete genome replication and chromosome segregation.	63	0.0443	2.78E-3	Cell cycle
GO_POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	1732	Any process that activates or increases the frequency, rate or extent of the chemical reactions and pathways involving a protein.	73	0.0421	2.78E-3	Cellular metabolic process
GSE19888_ADENOSINE_A3R_INH_PRETREAT_AND_ACT_BY_A3R_VS_A3R_INH_AND_TCELL_MEMBRANES_ACT_MAST_CELL_UP	194	Genes up-regulated in HMC-1 (mast leukemia) cells incubated with the peptide ALL1 followed by treatment with: Cl-IB-MECA [PubChem=3035850] versus T cell membranes.	17	0.0876	2.78E-3	Immunity
GSE3400_UNTREATED_VS_IFNB_TREATED_MEF_UP	194	Genes up-regulated in mouse embryonic fibroblasts (MEF): untreated versus interferon beta.	17	0.0876	2.78E-3	Immunity
GO_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT	174	An intracellular signal transduction process in which STAT proteins (Signal Transducers and Activators of Transcription) convey a signal to trigger a change in the activity or state of a cell. The STAT cascade begins with receptor activation followed by activation of STAT proteins by kinases. It proceeds through STA dimerization and subsequent nuclear translocation of STAT proteins, and ends with regulation of target gene expression by STAT proteins.	16	0.0920	2.78E-3	Cell communication
GSE7568_IL4_TGFB_DEXAMETHASONE_VS_IL4_TGFB_TREATED_MACROPHAGE_DN	175	Genes down-regulated in macrophages: 5 days with IL4 [GeneID=3565] and dexamethasone [PubChem=5743] followed by TGFB1 [GeneID=7040] for 24h versus 5 days with IL4 [GeneID=3565] followed by TGFB1 [GeneID=7040] for 24h.	16	0.0914	2.84E-3	Immunity

GO_POSITIVE_REGULATION_OF_SIGNALING	1896	Any process that activates, maintains or increases the frequency, rate or extent of a signaling process.	78	0.0411	2.84E-3	Cell communication
GSE29618_PRE_VS_DAY7_POST_TIV_FLU_VACCINE_BCELL_DN	196	Genes down-regulated in comparison of B cells from TIV influenza vaccinee pre-vaccination versus those at day 7 post-vaccination.	17	0.0867	2.84E-3	Immunity

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_PEPTIDYL_TYROSINE_MODIFICATION	377	The modification of peptidyl-tyrosine.	25	0.0663	2.84E-3	Cellular metabolic process
GO_PEPTIDE_METABOLIC_PROCESS	907	The chemical reactions and pathways involving peptides, compounds of two or more amino acids where the alpha carboxyl group of one is bound to the alpha amino group of another.	45	0.0496	2.84E-3	Cellular metabolic process
GO_POSITIVE_REGULATION_OF_TYROSINE_PHOSPHORYLATION_OF_STAT_PROTEIN	71	Any process that activates or increases the frequency, rate or extent of the introduction of a phosphate group to a tyrosine residue of a STAT (Signal Transducer and Activator of Transcription) protein.	10	0.1408	2.84E-3	Cellular metabolic process
GSE17974_1.5H_VS_72H_IL4_AND_ANTI_IL12_ACT_CD4_TCELL_UP	198	Genes up-regulated in comparison of CD4 [GeneID=920] T cells treated with IL4 [GeneID=3565] and anti-IL12 at 1.5 h versus those at 72 h.	17	0.0859	2.84E-3	Immunity
GO_BEHAVIOR	606	The internally coordinated responses (actions or inactions) of animals (individuals or groups) to internal or external stimuli, via a mechanism that involves nervous system activity.	34	0.0561	2.84E-3	Behavior
GO_POSITIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	1260	Any process that activates or increases the frequency, rate or extent of the covalent alteration of one or more amino acid residues within a protein.	57	0.0452	2.84E-3	Cellular metabolic process
GO_REGULATION_OF_PEPTIDYL_TYROSINE_PHOSPHORYLATION	263	Any process that modulates the frequency, rate or extent of the phosphorylation of peptidyl-tyrosine.	20	0.0760	2.84E-3	Cellular metabolic process
GSE19941_LPS_VS_LPS_AND_IL10_STIM_IL10_KO_NFKBP50_KO_MACROPHAGE_DN	199	Genes down-regulated in NFKB1 and IL10 [GeneID=4790;3586] knockout macrophages stimulated by LPS versus those also stimulated by IL10 [GeneID=3586].	17	0.0854	2.84E-3	Immunity
7p						
GSE45365_CD8A_DC_VS_CD11B_DC_UP	197	Genes up-regulated in dendritic cells: CD8A [GeneID=925] versus ITGAM+ [GeneID=3684].	18	0.0914	6.03E-8	Immunity
GSE45365_HEALTHY_VS_MCMV_INFECTION_BCELL_IFNAR_KO_UP	200	Genes up-regulated during primary acute viral infection: B lymphocytes versus CD8 T cells.	15	0.0750	2.92E-5	Immunity
GO_EMBRYONIC_SKELETAL_SYSTEM_MORPHOGENESIS	97	The process in which the anatomical structures of the skeleton are generated and organized during the embryonic phase.	11	0.1134	3.2E-5	Developmental process
GO_REGIONALIZATION	355	The pattern specification process that results in the subdivision of an axis or axes in space to define an area or volume in which specific patterns of cell differentiation will take place or in which cells interpret a specific environment.	18	0.0507	1.93E-4	Developmental process
GO_PATTERN_SPECIFICATION_PROCESS	455	Any developmental process that results in the creation of defined areas or spaces within an organism to which cells respond and eventually are instructed to differentiate.	20	0.0440	2.91E-4	Developmental process

GO_EMBRYONIC_SKELETAL_SYSTEM_DEVELOPMENT	130	The process, occurring during the embryonic phase, whose specific outcome is the progression of the skeleton over time, from its formation to the mature structure.	11	0.0846	3.44E-4	Developmental process
GSE18804_SPLEEN_MACROPHAGE_VS_TUMORAL_MACROPHAGE_UP	199	Genes up-regulated in macrophages: control versus tumor associated.	13	0.0653	4.58E-4	Immunity

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GSE6259_33D1_POS_DC_VS_BCELL_DN	150	Genes down-regulated in splenic dendritic cells versus 33D1+ B lymphocytes.	11	0.0733	1.06E-3	Immunity
GO_ANTERIOR_POSTERIOR_PATTERN_SPECIFICATION	219	The regionalization process in which specific areas of cell differentiation are determined along the anterior-posterior axis. The anterior-posterior axis is defined by a line that runs from the head or mouth of an organism to the tail or opposite end of the organism.	13	0.0594	1.06E-3	Developmental process
GO_SKELETAL_SYSTEM_MORPHOGENESIS	227	The process in which the anatomical structures of the skeleton are generated and organized.	13	0.0573	1.43E-3	Developmental process
GO_EMBRYO_DEVELOPMENT_ENDING_IN_BIRTH_OR_EGG_HATCHING	663	The process whose specific outcome is the progression of an embryo over time, from zygote formation until the end of the embryonic life stage. The end of the embryonic life stage is organism-specific and may be somewhat arbitrary; for mammals it is usually considered to be birth, for insects the hatching of the first instar larva from the eggshell.	22	0.0332	3.5E-3	Developmental process
GO_TUBE_DEVELOPMENT	1148	The process whose specific outcome is the progression of a tube over time, from its initial formation to a mature structure. Epithelial and endothelial tubes transport gases, liquids and cells from one site to another and form the basic structure of many organs and tissues including lung and trachea, kidney, the mammary gland, the vascular system and the gastrointestinal and urinary-genital tracts.	30	0.0261	6.7E-3	Developmental process
KRAS.DF.V1_DN	193	Genes down-regulated in epithelial lung cancer cell lines over-expressing an oncogenic form of KRAS [Gene ID=3845] gene.	11	0.0570	7.86E-3	Oncogenic signature
GSE18804_SPLEEN_MACROPHAGE_VS_BRAIN_TUMORAL_MACROPHAGE_UP	196	Genes up-regulated in macrophages: control versus glioblastoma conditioned.	11	0.0561	8.45E-3	Immunity
GSE23568_ID3_KO_VS_WT_CD8_TCELL_UP	199	Genes up-regulated in CD8 T cells: ID3 [GeneID=3399] knockout versus wildtype.	11	0.0553	9.11E-3	Immunity
GO_EMBRYO_DEVELOPMENT	1066	The process whose specific outcome is the progression of an embryo from its formation until the end of its embryonic life stage. The end of the embryonic stage is organism-specific. For example, for mammals, the process would begin with zygote formation and end with birth. For insects, the process would begin at zygote formation and end with larval hatching. For plant zygotic embryos, this would be from zygote formation to the end of seed dormancy. For plant vegetative embryos, this would be from the initial determination of the cell or group of cells to form an embryo until the point when the embryo becomes independent of the parent plant.	28	0.0263	9.44E-3	Developmental process

GO_SKELETAL_SYSTEM_DEVELOPMENT	516	The process whose specific outcome is the progression of the skeleton over time, from its formation to the mature structure. The skeleton is the bony framework of the body in vertebrates (endoskeleton) or the hard outer envelope of insects (exoskeleton or dermoskeleton).	18	0.0349	9.44E-3	Developmental process

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
7q						
REACTOME_CLASS_C_3_METABOTROPIC_GLUTAMATE_PHEROMONE_RECEPTORS_	40	Class C/3 (Metabotropic glutamate/pheromone receptors).	11	0.2750	2.74E-6	Cell metabolic process
GO_SENSORY_PERCEPTION_OF_BITTER_TASTE	41	The series of events required to receive a bitter taste stimulus, convert it to a molecular signal, and recognize and characterize the signal. This is a neurological process.	11	0.2683	2.74E-6	Nervous system process
GO_SENSORY_PERCEPTION_OF_TASTE	64	The series of events required for an organism to receive a gustatory stimulus, convert it to a molecular signal, and recognize and characterize the signal. Gustation involves the direct detection of chemical composition, usually through contact with chemoreceptor cells. This is a neurological process.	13	0.2031	2.74E-6	Nervous system process
GO_DETECTION_OF_CHEMICAL_STIMULUS_INVOLVED_IN_SENSORY_PERCEPTION_OF_TASTE	42	The series of events involved in the perception of taste in which a gustatory chemical stimulus is received and converted into a molecular signal.	10	0.2381	3.96E-5	Nervous system process
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_SIZE	520	Any process that modulates the size of an anatomical structure.	32	0.0615	1.63E-4	Developmental process
REACTOME_SIGNALING_BY_GPCR	1184	Signaling by GPCR.	54	0.0456	1.63E-4	Cell communication
KEGG_TASTE_TRANSDUCTION	52	Taste transduction.	10	0.1923	2.02E-4	Nervous system process
GO_REGULATION_OF_CELLULAR_COMPONENT_SIZE	383	A process that modulates the size of a cellular component.	26	0.0679	2.8E-4	Developmental process
GO_CELLULAR_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	1220	Any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an oxygen-containing compound stimulus.	54	0.0443	2.82E-4	Cellular response to stimulus
GO_NERVOUS_SYSTEM_PROCESS	1462	A organ system process carried out by any of the organs or tissues of neurological system.	60	0.0410	6.7E-4	Nervous system process
REACTOME_GPCR_LIGAND_BINDING	463	GPCR ligand binding.	28	0.0605	7.24E-4	Cell communication
GO_SENSORY_PERCEPTION_OF_CHEMICAL_STIMULUS	523	The series of events required for an organism to receive a sensory chemical stimulus, convert it to a molecular signal, and recognize and characterize the signal. This is a neurological process.	30	0.0574	7.65E-4	Cellular response to stimulus
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_MOVEMENT	610	Any process that activates or increases the frequency, rate or extent of the movement of a cellular component.	33	0.0541	7.65E-4	Cell movement

GSE46242_TH1_VS_ANERGIC_TH1_CD4_TCELL_UP	196	Genes up-regulated in CD4 [GeneID=920] Th1 cells: control versus anergic.	17	0.0867	8.22E-4	Immunity
GO_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	1714	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an oxygen-containing compound stimulus.	66	0.0385	1.09E-3	Cellular response to stimulus

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
REACTOME_G_ALPHA_I_SIGNALLING_EVENTS	403	G alpha (i) signalling events.	25	0.0620	1.26E-3	Cell communication
GO_FORMATION_OF_QUADRUPLE_SL_U4_U5_U6_SNRNP	12	Formation of a quadruple snRNP complex composed of the spliced leader (SL) RNA along with the U4/U6-U5 tri-snRNP complex. Interactions that may facilitate this include a duplex between the SL and U6 RNAs and interactions between the U5 RNA and the exon sequence at the 5' splice site within the SL RNA.	5	0.4167	2.25E-3	Cellular metabolic process
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS_	94	Class B/2 (Secretin family receptors).	11	0.1170	3.09E-3	Cell communication
GO_OXIDATION_REDUCTION_PROCESS	1014	A metabolic process that results in the removal or addition of one or more electrons to or from a substance, with or without the concomitant removal or addition of a proton or protons.	44	0.0434	3.09E-3	Cellular metabolic process
GO_SENSORY_PERCEPTION	985	The series of events required for an organism to receive a sensory stimulus, convert it to a molecular signal, and recognize and characterize the signal. This is a neurological process.	43	0.0437	3.25E-3	Nervous system process
GSE11961_FOLLICULAR_BCELL_VS_PLASMA_CELL_DAY7_DN	200	Genes down-regulated in follicular B cells versus day 7 plasma cells.	16	0.0800	3.29E-3	Immunity
REACTOME_RNA_POLYMERASE_II_TRANSCRIPTION	1373	RNA Polymerase II Transcription.	54	0.0393	3.96E-3	Gene expression
GO_HOMEOSTATIC_PROCESS	1963	Any biological process involved in the maintenance of an internal steady state.	70	0.0357	4.54E-3	Biological regulation
REACTOME_SIGNALING_BY_BRAF_AND_RAF_FUSIONS	65	Signaling by BRAF and RAF fusions.	9	0.1385	4.54E-3	Oncogenic signature
GO_DETECTION_OF_STIMULUS_INVOLVED_IN_SENSORY_PERCEPTION	532	The series of events involved in sensory perception in which a sensory stimulus is received and converted into a molecular signal.	28	0.0526	4.57E-3	Nervous system process
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_ORGANIZATION	1217	Any process that activates or increases the frequency, rate or extent of a process involved in the formation, arrangement of constituent parts, or disassembly of cell structures, including the plasma membrane and any external encapsulating structures such as the cell wall and cell envelope.	49	0.0403	4.92E-3	Developmental process
REACTOME_DISEASE	1580	Disease.	59	0.0373	5.7E-3	Disease

GO_RESPONSE_TO_ENDOGENOUS_STIMULUS	1662	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus arising within the organism.	61	0.0367	6.39E-3	Cellular response to stimulus

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_G_PROTEIN_COUPLED_RECEPTOR_SIGNALING_PATHWAY	1375	A series of molecular signals that proceeds with an activated receptor promoting the exchange of GDP for GTP on the alpha-subunit of an associated heterotrimeric G-protein complex. The GTP-bound activated alpha-G-protein then dissociates from the beta- and gamma-subunits to further transmit the signal within the cell. The pathway begins with receptor-ligand interaction, or for basal GPCR signaling the pathway begins with the receptor activating its G protein in the absence of an agonist, and ends with regulation of a downstream cellular process, e.g. transcription. The pathway can start from the plasma membrane, Golgi or nuclear membrane.	53	0.0385	6.39E-3	Cell communication
GO_DETECTION_OF_STIMULUS	703	The series of events in which a stimulus is received by a cell or organism and converted into a molecular signal.	33	0.0469	6.63E-3	Cellular response to stimulus
GO_REGULATION_OF_ACTIN_FILAMENT_BASED_PROCESS	405	Any process that modulates the frequency, rate or extent of any cellular process that depends upon or alters the actin cytoskeleton.	23	0.0568	7.14E-3	Biological regulation
GO_CELLULAR_RESPONSE_TO_NITROGEN_COMPOUND	716	Any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a nitrogen compound stimulus.	33	0.0461	8.36E-3	Cellular response to stimulus
GSE17721_CTRL_VS_POLYIC_4H_BMDC_DN	200	Genes down-regulated in comparison of control dendritic cells (DC) at 4 h versus those stimulated with poly(I:C) (TLR3 agonist) at 4 h.	15	0.0750	8.36E-3	Immunity
GSE17721_PAM3CSK4_VS_GADIQUIMOD_0.5H_BMDC_UP	200	Genes up-regulated in comparison of dendritic cells (DC) stimulated with Pam3Csk4 (TLR1/2 agonist) at 0.5 h versus DC cells stimulated with Gardiquimod (TLR7 agonist) at 0.5 h.	15	0.0750	8.36E-3	Immunity
GSE21774_CD62L_POS_CD56_BRIGHT_VS_CD62L_NEG_CD56_DIM_NK_CELL_DN	200	Genes down-regulated in NCAM1+ SELL bright [GeneID=4684;6402] versus NCAM1- SELL dim [GeneID=4684;6402].	15	0.0750	8.36E-3	Immunity
GO_REGULATION_OF_ANION_CHANNEL_ACTIVITY	9	Any process that modulates the frequency, rate or extent of anion channel activity.	4	0.4444	8.54E-3	Biological regulation
GO_REGULATION_OF_ACTIN_FILAMENT_ORGANIZATION	278	Any process that modulates the frequency, rate or extent of actin filament organization.	18	0.0647	9.05E-3	Biological regulation
REACTOME_EXTRA_NUCLEAR_ESTROGEN_SIGNALING	75	Extra-nuclear estrogen signaling.	9	0.1200	9.37E-3	Cell communication
GO_SULFATE_TRANSPORT	18	The directed movement of sulfate into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.	5	0.2778	9.56E-3	Transport
GO_REGULATION_OF_CELL_SIZE	181	Any process that modulates the size of a cell.	14	0.0773	9.67E-3	Biological regulation

GO_POSITIVE_REGULATION_OF_CYTOSKELETON_ORGANIZATION	230	Any process that activates or increases the frequency, rate or extent of the formation, arrangement of constituent parts, or disassembly of cytoskeletal structures.	16	0.0696	9.67E-3	Biological regulation

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
14q						
GSE45365_HEALTHY_VS_MCMV_INFECTION_BCELL_IFNAR_KO_UP	200	Genes up-regulated during primary acute viral infection: B lymphocytes versus CD8 T cells.	39	0.1950	6.75E-21	Immunity
GO_POSTTRANSCRIPTIONAL_REGULATION_OF_GENE_EXPRESSION	1193	Any process that modulates the frequency, rate or extent of gene expression after the production of an RNA transcript.	85	0.0712	9.37E-17	Gene Expression
GO_RNA_PROCESSING	1442	Any process involved in the conversion of one or more primary RNA transcripts into one or more mature RNA molecules.	91	0.0631	7.2E-15	Gene Expression
GO_GENE_SILENCING	838	Any process carried out at the cellular level that results in either long-term transcriptional repression via action on chromatin structure or RNA mediated, post-transcriptional repression of gene expression.	62	0.0740	1.66E-12	Gene Expression
GSE18804_SPLEEN_MACROPHAGE_VS_TUMORAL_MACROPHAGE_UP	199	Genes up-regulated in macrophages: control versus tumor associated.	28	0.1407	4.24E-11	Immunity
GO_RNA_PHOSPHODIESTER_BOND_HYDROLYSIS	160	The RNA metabolic process in which the phosphodiester bonds between ribonucleotides are cleaved by hydrolysis.	24	0.1500	6.43E-10	Cellular metabolic process
GSE18804_SPLEEN_MACROPHAGE_VS_BRAIN_TUMORAL_MACROPHAGE_UP	196	Genes up-regulated in macrophages: control versus glioblastoma conditioned.	22	0.1122	1.73E-6	Immunity
GO_NUCLEIC_ACID_PHOSPHODIESTER_BOND_HYDROLYSIS	308	The nucleic acid metabolic process in which the phosphodiester bonds between nucleotides are cleaved by hydrolysis.	27	0.0877	4.9E-6	Cellular metabolic process
GSE45365_CD8A_DC_VS_CD11B_DC_UP	197	Genes up-regulated in dendritic cells: CD8A [GeneID=925] versus ITGAM+ [GeneID=3684].	21	0.1066	8.24E-6	Immunity
GO_NEGATIVE_REGULATION_OF_ARTERY_MORPHOGENESIS	5	Any process that stops, prevents or reduces the frequency, rate or extent of artery morphogenesis.	5	10.000	8.55E-6	Developmental proccess
GSE29618_PRE_VS_DAY7_POST_LAIV_FLU_VACCINE_MONOCYTE_DN	200	Genes down-regulated in comparison of monocytes from LAIV influenza vaccinee pre-vaccination versus those at day 7 post-vaccination.	20	0.1000	4.6E-5	Immunity
LTE2_UP.V1_UP	188	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells (breast cancer) and long-term adapted for estrogen-independent growth.	19	0.1011	7.79E-5	Oncogenic signature
GO_RNA_PHOSPHODIESTER_BOND_HYDROLYSIS_ENDONUCLEOLYTIC	82	The chemical reactions and pathways involving the hydrolysis of internal 3',5'-phosphodiester bonds in one or two strands of ribonucleotides.	12	0.1463	3.61E-4	Cellular metabolic process
GO_HEMATOPOIETIC_PROGENITOR_CELL_DIFFERENTIATION	172	The process in which precursor cell type acquires the specialized features of a hematopoietic progenitor cell, a class of cell types including myeloid progenitor cells and lymphoid progenitor cells.	17	0.0988	4.48E-4	Developmental proccess

GO_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE_SIGNAL_SEQUENCE_RECOGNITION	9	The process in which SRP binds to the signal peptide in a nascent protein, causing protein elongation to pause, during cotranslational membrane targeting.	5	0.5556	6.66E-4	Cell communication

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GSE5589_LPS_AND_IL10_VS_LPS_AND_IL6_STIM_MACROPHAGE_45MIN_DN	200	Genes down-regulated in bone marrow-derived macrophages (45 min): IL10 [GeneID=3486] and LPS versus IL6 [GeneID=3469] and LPS.	18	0.0900	7.26E-4	Immunity
REACTOME_CELLULAR_RESPONSE_TO_HYPOXIA	75	Cellular response to hypoxia.	11	0.1467	8.34E-4	Response to stress
GO_INTRACELLULAR_TRANSPORT	1758	The directed movement of substances within a cell.	72	0.0410	8.46E-4	Transport
GO_PROTEIN_LOCALIZATION_TO_MEMBRANE	654	A process in which a protein is transported to, or maintained in, a specific location in a membrane.	36	0.0550	9.03E-4	Cellular localization
GO_MRNA_METABOLIC_PROCESS	887	The chemical reactions and pathways involving mRNA, messenger RNA, which is responsible for carrying the coded genetic 'message', transcribed from DNA, to sites of protein assembly at the ribosomes.	44	0.0496	9.59E-4	Cellular metabolic process
GO_CELLULAR_MACROMOLECULE_LOCALIZATION	1989	Any process in which a macromolecule is transported to, and/or maintained in, a specific location at the level of a cell. Localization at the cellular level encompasses movement within the cell, from within the cell to the cell surface, or from one location to another at the surface of a cell.	78	0.0392	1.3E-3	Cellular localization
GO_SENSORY_PERCEPTION_OF_SMELL	452	The series of events required for an organism to receive an olfactory stimulus, convert it to a molecular signal, and recognize and characterize the signal. Olfaction involves the detection of chemical composition of an organism's ambient medium by chemoreceptors. This is a neurological process.	28	0.0619	1.3E-3	Nervous system process
GO_RIBONUCLEOPROTEIN_COMPLEX_SUBUNIT_ORGANIZATION	238	Any process in which macromolecules aggregate, disaggregate, or are modified, resulting in the formation, disassembly, or alteration of a ribonucleoprotein complex.	19	0.0798	1.51E-3	Cellular organization
MEK_UP.V1_UP	195	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active MAP2K1 [Gene ID=5604] gene.	17	0.0872	1.51E-3	Oncogenic signature
REACTOME_CYCLIN_A_CDK2_ASSOCIATED_EVENTS_AT_S_PHASE_ENTRY	85	Cyclin A:Cdk2-associated events at S phase entry.	11	0.1294	2.01E-3	Cell cycle
GO_FORMATION_OF_QUADRUPLE_SL_U4_U5_U6_SNRNP	12	Formation of a quadruple snRNP complex composed of the spliced leader (SL) RNA along with the U4/U6-U5 tri-snRNP complex. Interactions that may facilitate this include a duplex between the SL and U6 RNAs and interactions between the U5 RNA and the exon sequence at the 5' splice site within the SL RNA.	5	0.4167	2.28E-3	Cellular metabolic process
REACTOME_DEGRADATION_OF_DVL	57	Degradation of DVL.	9	0.1579	2.73E-3	Cellular communication

GO_CELLULAR_PROTEIN_CONTAINING_COMPLEX_ASSEMBLY	1129	The aggregation, arrangement and bonding together of a set of components to form a protein complex, occurring at the level of an individual cell.	50	0.0443	3.25E-3	Cellular organization
GO_REGULATION_OF_ARTERY_MORPHOGENESIS	13	Any process that modulates the frequency, rate or extent of artery morphogenesis.	5	0.3846	3.26E-3	Developmental proccess

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_RNA_SPLICING	484	The process of removing sections of the primary RNA transcript to remove sequences not present in the mature form of the RNA and joining the remaining sections to form the mature form of the RNA.	28	0.0579	3.47E-3	Gene Expression
REACTOME_DEGRADATION_OF_GLI1_BY_THE_PROTEASOME	60	Degradation of GLI1 by the proteasome.	9	0.1500	3.57E-3	Cell communication
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	542	G alpha (s) signalling events.	30	0.0554	3.57E-3	Cell communication
KEGG_PROTEASOME	46	Proteasome.	8	0.1739	3.57E-3	Gene Expression
REACTOME_DEFECTIVE_CFTR_CAUSES_CYSTIC_FIBROSIS	61	Defective CFTR causes cystic fibrosis.	9	0.1475	3.76E-3	Disease
GO_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER_IN_RESPONSE_TO_HYPOXIA	77	Any process that modulates the frequency, rate or extent of transcription from an RNA polymerase II promoter as a result of a hypoxia stimulus.	10	0.1299	3.76E-3	Response to stress
REACTOME_ABC_TRANSPORTER_DISORDERS	77	ABC transporter disorders.	10	0.1299	3.76E-3	Transport
GO_POST_TRANSLATIONAL_PROTEIN_MODIFICATION	363	The process of covalently altering one or more amino acids in a protein after the protein has been completely translated and released from the ribosome.	23	0.0634	4.33E-3	Gene Expression
GO_REGULATION_OF_CELLULAR_AMINO_ACID_METABOLIC_PROCESS	63	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways involving amino acids.	9	0.1429	4.51E-3	Cellular metabolic process
GO_RNA_SPLICING_VIA_TRANSESTERIFICATION_REACTIONS	391	Splicing of RNA via a series of two transesterification reactions.	24	0.0614	4.51E-3	Gene Expression
GO_ENDOTHELIAL_CELL_PROLIFERATION	199	The multiplication or reproduction of endothelial cells, resulting in the expansion of a cell population. Endothelial cells are thin flattened cells which line the inside surfaces of body cavities, blood vessels, and lymph vessels, making up the endothelium.	16	0.0804	4.82E-3	Cell proliferation
GSE29618_LAIV_VS_TIV_FLU_VACCINE_DAY7_MONOCYTE_UP	199	Genes up-regulated in comparison of monocytes from LAIV influenza vaccinee at day 7 post-vaccination vesus those from TIV influenza vaccinee at day 7.	16	0.0804	4.82E-3	Immunity
GSE26912_TUMORICIDAL_VS_CTRL_MACROPHAGE_DN	200	Genes down-regulated in macrophages: tumoricidal versus control.	16	0.0800	4.86E-3	Immunity
REACTOME_DOWNSTREAM_SIGNALING_EVENTS_OF_B_CELL_RECEPTOR_BCR_	81	Downstream signaling events of B Cell Receptor (BCR).	10	0.1235	4.86E-3	Immunity

GO_REGULATION_OF_MRNA_METABOLIC_PROCESS	344	Any process that modulates the frequency, rate or extent of mRNA metabolic process.	22	0.0640	4.86E-3	Cellular metabolic process
REACTOME_HEDGEHOG_LIGAND_BIOGENESIS	65	Hedgehog ligand biogenesis.	9	0.1385	4.87E-3	Cell communication

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES_	50	Cross-presentation of soluble exogenous antigens (endosomes).	8	0.1600	4.87E-3	Immunity
REACTOME_SIGNALING_BY_NOTCH4	82	Signaling by NOTCH4.	10	0.1220	5.02E-3	Cell communication
GO_EPITHELIAL_CELL_PROLIFERATION	453	The multiplication or reproduction of epithelial cells, resulting in the expansion of a cell population. Epithelial cells make up the epithelium, the covering of internal and external surfaces of the body, including the lining of vessels and other small cavities. It consists of cells joined by small amounts of cementing substances.	26	0.0574	5.02E-3	Cell proliferation
GO_CELL_CYCLE_G2_M_PHASE_TRANSITION	273	The cell cycle process by which a cell in G2 phase commits to M phase.	19	0.0696	5.02E-3	Cell cycle
20p						
GO_NEGATIVE_REGULATION_OF_PEPTIDASE_ACTIVITY	270	Any process that stops or reduces the rate of peptidase activity, the hydrolysis of peptide bonds within proteins.	15	0.0556	3.2E-6	Cellular metabolic process
GO_NEGATIVE_REGULATION_OF_PROTEOLYSIS	366	Any process that stops, prevents, or reduces the frequency, rate or extent of the hydrolysis of a peptide bond or bonds within a protein.	16	0.0437	1.3E-5	Cellular metabolic process
GO_NEGATIVE_REGULATION_OF_HYDROLASE_ACTIVITY	471	Any process that stops or reduces the rate of hydrolase activity, the catalysis of the hydrolysis of various bonds.	17	0.0361	4.67E-5	Cellular metabolic process
GO_REGULATION_OF_PEPTIDASE_ACTIVITY	464	Any process that modulates the frequency, rate or extent of peptidase activity, the hydrolysis of peptide bonds within proteins.	16	0.0345	1.81E-4	Cellular metabolic process
REACTOME_BETA_DEFENSINS	42	Beta defensins.	6	0.1429	7.27E-4	Immunity
GO_NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY	837	Any process that stops or reduces the activity of an enzyme.	20	0.0239	9.96E-4	Cellular metabolic process
GO_REGULATION_OF_HYDROLASE_ACTIVITY	1302	Any process that modulates the frequency, rate or extent of hydrolase activity, the catalysis of the hydrolysis of various bonds, e.g. C-O, C-N, C-C, phosphoric anhydride bonds, etc. Hydrolase is the systematic name for any enzyme of EC class 3.	25	0.0192	1.67E-3	Cellular metabolic process
REACTOME_DEFENSINS	52	Defensins.	6	0.1154	1.67E-3	Immunity
GO_REGULATION_OF_PROTEOLYSIS	748	Any process that modulates the frequency, rate or extent of the hydrolysis of a peptide bond or bonds within a protein.	17	0.0227	9.72E-3	Cellular metabolic process
20q						

GO_ANTIMICROBIAL_HUMORAL_RESPONSE	137	An immune response against microbes mediated through a body fluid. Examples of this process are seen in the antimicrobial humoral response of Drosophila melanogaster and Mus musculus.	21	0.1533	7.11E-13	Immunity
REACTOME_ANTIMICROBIAL_PEPTIDES	97	Antimicrobial peptides.	18	0.1856	1.85E-12	Immunity

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_ANTIBACTERIAL_HUMORAL_RESPONSE	59	An immune response against bacteria mediated through a body fluid. Examples of this process are the antibacterial humoral responses in Mus musculus and Drosophila melanogaster.	14	0.2373	6.84E-11	Immunity
GO_DEFENSE_RESPONSE_TO_BACTERIUM	341	Reactions triggered in response to the presence of a bacterium that act to protect the cell or organism.	27	0.0792	1.37E-10	Immunity
GO_NEGATIVE_REGULATION_OF_MOLECULAR_FUNCTION	1195	Any process that stops or reduces the rate or extent of a molecular function, an elemental biological activity occurring at the molecular level, such as catalysis or binding.	45	0.0377	1.81E-7	Biological regulation
GO_NEGATIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	1209	Any process that stops, prevents, or reduces the frequency, rate or extent of chemical reactions and pathways involving a protein.	45	0.0372	2.2E-7	Biological regulation
GO_RESPONSE_TO_BACTERIUM	743	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus from a bacterium.	32	0.0431	3.12E-6	Immunity
GO_NEGATIVE_REGULATION_OF_PEPTIDASE_ACTIVITY	270	Any process that stops or reduces the rate of peptidase activity, the hydrolysis of peptide bonds within proteins.	19	0.0704	3.12E-6	Cellular metabolic process
GO_DEFENSE_RESPONSE_TO_OTHER_ORGANISM	1203	Reactions triggered in response to the presence of another organism that act to protect the cell or organism from damage caused by that organism.	41	0.0341	1.3E-5	Immunity
GO_HUMORAL_IMMUNE_RESPONSE	370	An immune response mediated through a body fluid.	21	0.0568	1.61E-5	Immunity
GO_NEGATIVE_REGULATION_OF_PROTEOLYSIS	366	Any process that stops, prevents, or reduces the frequency, rate or extent of the hydrolysis of a peptide bond or bonds within a protein.	20	0.0546	6.11E-5	Cellular metabolic process
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	1466	Any process that activates or increases the rate or extent of development, the biological process whose specific outcome is the progression of an organism over time from an initial condition (e.g. a zygote, or a young adult) to a later condition (e.g. a multicellular animal or an aged adult).	44	0.0300	1.08E-4	Developmental process
GO_REGULATION_OF_PEPTIDASE_ACTIVITY	464	Any process that modulates the frequency, rate or extent of peptidase activity, the hydrolysis of peptide bonds within proteins.	22	0.0474	1.26E-4	Cellular metabolic process
GO_INNATE_IMMUNE_RESPONSE	986	Innate immune responses are defense responses mediated by germline encoded components that directly recognize components of potential pathogens.	34	0.0345	1.26E-4	Immunity
REACTOME_INNATE_IMMUNE_SYSTEM	1114	Innate Immune System.	36	0.0323	2.38E-4	Immunity
GO_NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY	837	Any process that stops or reduces the activity of an enzyme.	30	0.0358	2.76E-4	Cellular metabolic process

GO_POSITIVE_REGULATION_OF_CELL_DIFFERENTIATION	1033	Any process that activates or increases the frequency, rate or extent of cell differentiation.	34	0.0329	3.02E-4	Developmental process
GO_DEVELOPMENTAL_PROCESS_INVOLVED_IN_REPRODUCTION	1001	A developmental process in which a progressive change in the state of some part of an organism specifically contributes to its ability to form offspring.	33	0.0330	4.11E-4	Developmental process

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GO_NEGATIVE_REGULATION_OF_HYDROLASE_ACTIVITY	471	Any process that stops or reduces the rate of hydrolase activity, the catalysis of the hydrolysis of various bonds.	21	0.0446	4.83E-4	Cellular metabolic process
REACTOME_BETA_DEFENSINS	42	Beta defensins.	7	0.1667	5.98E-4	Immunity
GO_REGULATION_OF_PROTEOLYSIS	748	Any process that modulates the frequency, rate or extent of the hydrolysis of a peptide bond or bonds within a protein.	27	0.0361	7.23E-4	Cellular metabolic process
GO_REGULATION_OF_CELL_DIFFERENTIATION	1945	Any process that modulates the frequency, rate or extent of cell differentiation, the process in which relatively unspecialized cells acquire specialized structural and functional features.	50	0.0257	7.71E-4	Developmental process
GO_RESPONSE_TO_BIOTIC_STIMULUS	1615	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a biotic stimulus, a stimulus caused or produced by a living organism.	44	0.0272	7.71E-4	Cellular response to stimulus
GO_POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	1853	Any process that activates or increases the frequency, rate or extent of an organismal process, any of the processes pertinent to the function of an organism above the cellular level; includes the integrated processes of tissues and organs.	48	0.0259	9.69E-4	Biological regulation
GO_DEFENSE_RESPONSE	1814	Reactions, triggered in response to the presence of a foreign body or the occurrence of an injury, which result in restriction of damage to the organism attacked or prevention/recovery from the infection caused by the attack.	47	0.0259	1.19E-3	Immunity
GO_REPRODUCTION	1502	The production of new individuals that contain some portion of genetic material inherited from one or more parent organisms.	41	0.0273	1.52E-3	Reproduction
REACTOME_DEFENSINS	52	Defensins.	7	0.1346	1.97E-3	Immunity
GO_REGULATION_OF_CELL_DEATH	1746	Any process that modulates the rate or frequency of cell death. Cell death is the specific activation or halting of processes within a cell so that its vital functions markedly cease, rather than simply deteriorating gradually over time, which culminates in cell death.	45	0.0258	2.04E-3	Cell death
GSE45365_HEALTHY_VS_MCMV_INFECTION_CD8_TCELL_IFNAR_KO_UP	191	Genes up-regulated during primary acute viral infection in dendritic cells with IFNAR1 [GeneID=3454] knockout: CD8A [GeneID=925] versus ITGAM+ [GeneID=3684].	12	0.0628	2.81E-3	Immunity
GO_POSITIVE_REGULATION_OF_CELLULAR_BIOSYNTHETIC_PROCESS	1967	Any process that activates or increases the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of substances, carried out by individual cells.	48	0.0244	3.71E-3	Cellular metabolic process
GSE17721_LPS_VS_CPG_12H_BMDC_UP	199	Genes up-regulated in comparison of dendritic cells (DC) stimulated with LPS (TLR4 agonist) at 12 h versus DC cells stimulated with CpG DNA (TLR9 agonist) at 12 h.	12	0.0603	3.71E-3	Immunity

GSE15735_2H_VS_12H_HDAC_INHIBITOR_TREATED_CD4_TCELL_DN	200	Genes down-regulated in CD4 [GeneID=920] T cells treated with HDAC inhibitors: 2h versus 12h.	12	0.0600	3.71E-3	Immunity

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	FDR q-value	Biological_Type
GSE19888_ADENOSINE_A3R_INH_VS_ACT_WITH_INHIBITOR_PRETREATMENT_IN_MAST_CELL_DN	200	Genes down-regulated in HMC-1 (mast leukemia) cells incubated the peptide ALL1 versus those followed by treatment with Cl-IB-MECA [PubChem=3035850].	12	0.0600	3.71E-3	Immunity
GSE22601_DOUBLE_NEGATIVE_VS_CD4_SINGLE_POSITIVE_THYMOCYTE_DN	200	Genes down-regulated in thymocytes: double negative versus CD4 [GeneID=920] single positive.	12	0.0600	3.71E-3	Immunity
GSE5589_LPS_VS_LPS_AND_IL10_STIM_IL10_KO_MACROPHAGE_180MIN_DN	200	Genes down-regulated in bone marrow-derived macrophages with IL10 [GeneID=3486] and 180 min stimulation of: LPS versus IL10 [GeneID=3486] and LPS.	12	0.0600	3.71E-3	Immunity
GO_PROTEOLYSIS	1809	The hydrolysis of proteins into smaller polypeptides and/or amino acids by cleavage of their peptide bonds.	45	0.0249	3.88E-3	Cellular metabolic process
GO_EMBRYO_DEVELOPMENT	1066	The process whose specific outcome is the progression of an embryo from its formation until the end of its embryonic life stage. The end of the embryonic stage is organism-specific. For example, for mammals, the process would begin with zygote formation and end with birth. For insects, the process would begin at zygote formation and end with larval hatching. For plant zygotic embryos, this would be from zygote formation to the end of seed dormancy. For plant vegetative embryos, this would be from the initial determination of the cell or group of cells to form an embryo until the point when the embryo becomes independent of the parent plant.	31	0.0291	5.48E-3	Developmental process
GO_NEGATIVE_REGULATION_OF_CALCIUM_ION_TRANSPORT	64	Any process that stops, prevents, or reduces the frequency, rate or extent of the directed movement of calcium ions into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.	7	0.1094	5.73E-3	Transport
GO_MULTICELLULAR_ORGANISM_REPRODUCTION	876	The biological process in which new individuals are produced by one or two multicellular organisms. The new individuals inherit some proportion of their genetic material from the parent or parents.	27	0.0308	6.88E-3	Reproduction
GO_APOPTOTIC_PROCESS	1994	A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathway phase) which trigger an execution phase. The execution phase is the last step of an apoptotic process, and is typically characterized by rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. When the execution phase is completed, the cell has died.	47	0.0236	8.6E-3	Cell death
GO_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	1714	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an oxygen-containing compound stimulus.	42	0.0245	9.65E-3	Cellular response to stimulus