Estudi del paper funcional de la duplicació completa del genoma en càncer a partir de dades de seqüenciació d'ARN del programa The Cancer Genome Atlas (TCGA)

Nom Estudiant: Alba Mas Malavila
Pla d'estudis de l'estudiant: Màster en Bioinformàtica i Bioestadística
Àrea de treball final: Estadística i Bioinformàtica 3

Nom Consultor/a UOC: Laia Bassaganyas Bars

Nom Tutor extern: Jordi Camps Polo
Institució externa: Grup de recerca en Oncologia Gastrointestinal i pancreatica, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

Nom Professor/a responsable de l'assignatura: Ferran Prados Carrasco

Data Lliurament: 24 de juny de 2020

FITXA DEL TREBALL FINAL

Títol del treball:	Estudi del paper funcional de la duplicació completa del genoma en càncer a partir de dades de seqüenciació d'ARN del programa The Cancer Genome Atlas (TCGA).
Nom de l'autor:	Alba Mas Malavila
Nom del consultor/a:	Laia Bassaganyas Bars
Nom del PRA:	Ferran Prados Carrasco
Data de lliurament (mm/aaaa):	*06/2020*
Titulació o programa:	Màster en Bioinformàtica i Bioestadística
Àrea del Treball Final:	Estadística i Bioinformàtica
Idioma del treball:	Català
Paraules clau	*Genòmica del càncer, duplicació completa del genoma, expressió gènica*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

La duplicació completa del genoma (WGD) en càncer està associada amb inestabilitat genòmica, alteracions recurrents en gens de control del cicle cel·lular, nivells alts d'alteracions en el nombre de còpies i mal pronòstic clínic, però es desconeix l'efecte global que té sobre els programes transcripcionals.

El treball consisteix en un anàlisi d'expressió gènica diferencial entre tumors de càncer colorectal i adenocarcinoma de pulmó amb i sense WGD a partir de dades públiques de seqüenciació d'ARN, a partir del qual s'han identificat mecanismes moleculars i processos cel·lulars associats, posant especial èmfasi en mecanismes d'evasió del sistema immunitari.

En els dos tipus de càncer s'ha observat una correlació negativa entre WGD i resposta immunitària antitumoral. A més, en càncer colorectal s'han observat diferències metabòliques i en càncer de pulmó diferències en vies de divisió cel·lular i replicació. En conjunt, aquests resultats podrien explicar l'agressivitat i mal pronòstic associats als tumors que han patit WGD.

Abstract (in English, 250 words or less):

Whole genome doubling (WGD) in cancer is associated to genomic instability, recurrent alterations in cell cycle control genes, high levels of copy number alterations and a poor prognosis, but its global effect on the transcriptional programs is unknown.

This work consists of a differential gene expression analysis between colorectal and lung adenocarcinoma tumors with and without WGD using RNA-sequencing public data. We have identified the molecular mechanisms and biological processes associated, putting special attention to immune evasion strategies.

We show a negative correlation between WGD and antitumor immune response in both colorectal and lung adenocarcinoma tumors. In addition, we find metabolic changes in colorectal tumors and changes in cell division pathways and DNA replication in lung tumors. Overall, these results potentially explain the aggressiveness and poor prognosis associated with tumors that have undergone WGD.

Índex

Llista de figures

Llista de taules suplementàries

1. Introducció

A continuació s'introdueix la temàtica del treball, els objectius i la metodologia seguida. A més, s'explica com s'ha fet la planificació i es descriu com s'ha estructurat la memòria.

1.1. Context i justificació del treball

Un dels trets característics de molts càncers és el fet que presenten inestabilitat genòmica i contenen un nombre anormal de cromosomes (aneuploïdia). Una de les causes de l'aneuploïdia és l'acumulació d'alteracions cromosòmiques degut a errors en la segregació d'un o diversos cromosomes. Tot i així, també pot aparèixer durant la proliferació de cèl·lules tetraploides inestables generades per una duplicació completa del genoma a partir de cèl·lules diploides [1].

La duplicació completa del genoma (WGD, de l'anglès, *Whole Genome Doubling*) és una de les alteracions genètiques més comunes en càncer (afecta aproximadament un terç dels càncers humans[2]) i generalment està associada amb un mal pronòstic [3,4,5], de manera que el seu estudi té una rellevància clínica important. La taxa de WGD varia substancialment entre els diferents tipus de càncer i en funció de certes característiques histològiques i moleculars[3] . S'associa amb alteracions recurrents que afecten gens de control del cicle cel·lular i amb nivells alts d'alteracions en el nombre de còpies (CNA, de l'anglès, *Copy Number Alteration*) [2,3]. Es creu que té lloc a causa d'errors en la divisió cel·lular generalment en estadis inicials de la progressió tumoral, precedida per mutacions puntuals en gens *driver* i seguida de múltiples esdeveniments de pèrdua de cromosomes que acaben donant lloc a cèl·lules subtetraploides [4,6]. Se li otorga un paper important en l'evolució tumoral ja que les cèl·lules tetraploides originades per WGD, més tolerants a alteracions genètiques que les cèl·lules diploides, estableixen el context ideal per a la diversificació subclonal i l'augment de l'heterogeneïtat genètica intratumoral, la qual cosa fa augmentar el potencial adaptatiu d'aquestes cèl·lules i promou la progressió tumoral.

Recentment l'aneuploïdia (o nivells alts de CNA) s'ha associat amb nivells baixos d'infiltració immunitària i resistència a immunoteràpia en molts tipus de càncer[7,8]. El microambient tumoral està format per diversos tipus cel·lulars entre els quals s'inclouen les cèl·lules tumorals, diferents tipus de cèl·lules immunitàries i altres tipus de cèl·lules estromals. El paper de les cèl·lules immunitàries en la formació i progressió tumorals s'ha estudiat en diferents fases i s'han identificat diferents subgrups de cèl·lules amb accions complementàries o oposades. La immunitat antitumoral, activa sobretot durant les primeres fases de la formació del tumor, està mediada principalment per limfòcits T efectors, però també per cèl·lules NK (*natural killer*), cèl·lules dendrítiques i macròfags de tipus M1. Els limfòcits T són, després dels macròfags, el tipus cel·lular trobat amb més freqüència en el

microambient tumoral, i nivells alts d'infiltració estan associats a un pronòstic favorable. Es distingeixen dos tipus de limfòcits T efectors: els limfòcits T CD8+, els quals un cop activats es diferencien en limfòcits citotòxics i destrueixen les cèl·lules diana mitjançant l'exocitosi de grànuls que contenen perforina i granzimM; i els limfòcits T col·laboradors (*helper*) CD4+ (Th-1), els quals secreten citocines proinflamatòries que promouen l'activació i l'acció dels limfòcits T citotòxics i l'activitat anti-tumoral de macròfags i cèl·lules NK. Les cèl·lules dendrítiques també contribueixen a la inhibició tumoral com a cèl·lules presentadores d'antígens professionals; activen la resposta dels limfòcits T efectors per mitjà de la presentació de neoantígens units al complex major d'histocompatibilitat (MHC). Els macròfags de tipus I tenen un paper important en l'eliminació de cèl·lules cancerígenes immunogèniques mitjançant la fagocitosi. Així mateix, les cèl·lules NK exerceixen un efecte principalment antitumoral mitjançant l'alliberament de molècules citotòxiques i induint l'apoptosi de les cèl·lules tumorals. Tot i així, a mesura que el tumor progressa, el microambient tumoral canvia i apareixen tipus cel·lulars amb activitat immunosupressora, com MDSC (*myeloid derived suppressor cells*), macròfags de tipus M2 i limfòcits T reguladors, els quals suprimeixen l'activació i l'acció de les cèl·lules efectores mitjançant l'expressió de receptors i molècules immunosupressores. Altres mecanismes d'evasió immunitària que adquireixen les cèl·lules tumorals són la pèrdua d'expressió d'antígens tumorals i del MHC, la qual cosa les fa resistents a l'acció de les cèl·lules efectores; i l'expressió de molècules immunomoduladores *checkpoint* com CTLA-4 i PD-1, expressades també durant els processos inflamatoris normals com a mecanisme de protecció contra el dany tissular [9]. Tots aquests mecanismes afavoreixen la progressió tumoral i disminueixen l'eficàcia de la immunoteràpia.

Fins al moment, els estudis publicats sobre WGD en càncer correlacionen aquest esdeveniment amb inestabilitat genòmica [4], amb certs perfils mutacionals com la pèrdua de gens supressors de tumors i l'amplificació d'oncogens, i amb pitjor pronòstic clínic [3], però es desconeix l'efecte global que provoquen els guanys i les pèrdues de grans regions cromosòmiques, de centenars a milers de gens, sobre el funcionament cel·lular, incloent els programes transcripcionals. Tampoc s'ha estudiat el seu paper en la immunitat tumoral, que podria ser rellevant donada l'associació entre nivells alts de CNA i marcadors d'evasió immunitària, i que explicaria l'evolució genòmica accelerada i el mal pronòstic associat als tumors amb WGD.

Per tal de tenir una idea global de l'impacte funcional del WGD en la progressió tumoral s'han estudiat els perfils transcripcionals característics de tumors que han patit WGD i s'han identificat mecanismes moleculars i processos cel·lulars associats, posant especial èmfasi en mecanismes d'evasió del sistema immunitari.

1.2. Objectius del treball

1- Determinar les diferències significatives a nivell transcripcional entre tumors amb i sense WGD.

1.1- Identificar gens diferencialment expressats en tumors que hagin patit WGD.

1.2- Identificar programes d'expressió gènica característics dels tumors amb WGD.

2- Correlacionar l'estat de WGD dels tumors amb les seves característiques immunològiques.

2.1- Estimar els nivells d'infiltració de cèl·lules immunitàries presents en els tumors i correlacionar-ho amb l'estat WGD.

2.2- Identificar els immunofenotips dels tumors i correlacionar-ho amb l'estat WGD.

1.3. Enfocament i mètode seguit

S'han utilitzat dades públiques de seqüenciació d'ARN (*RNA-seq*) de múltiples tumors per tal de poder fer inferències robustes. Les dades s'han obtingut a partir de *The Cancer Genome Atlas* (TCGA), projecte impulsat pel *National Cancer Institute* que recopila dades genòmiques, epigenòmiques, transcriptòmiques i proteòmiques procedents de projectes d'investigació en càncer.

Inicialment es va plantejar una ànalisi *pan-cancer* de la cohort de TCGA. De totes maneres, es va començar amb les dades d'adenocarcinoma de còlon (COAD, *Colon Adenocarcinoma*) per ser uns dels tipus de càncer més estudiat pel grup d'investigació on es desenvolupa el treball, seguit de l'anàlisi de dades de càncer de recte (READ, *Rectum Adencarcinoma*), molt sovint estudiat juntament amb COAD com a càncer colorectal (CRC, *Colorectal Cancer*). A més a més, s'ha pogut estendre l'anàlisi a la cohort d'adenocarcinoma de pulmó (LUAD, *Lung Adenocarcinoma*).

Les dades de recomptes obtingudes de TCGA s'han processat i analitzat principalment mitjançant el programa estadístic R.

Els tumors s'han dividit en dos grups en funció de si han patit o no WGD. Per inferir l'estat de WGD s'han seguit dues estratègies: o bé utilitzant dades de MCN (*Major Copy Number*) o bé a partir de dades de ploïdia obtingudes mitjançant ASCAT (*Allele Specific Copy Number Analysis of Tumors*) [10]

S'ha realitzat un anàlisi d'expressió diferencial utilitzant els paquets `edgeR` i `limma` de Bioconductor, un dels mètodes més utilitzats per l'anàlisi de dades de *RNA-seq*. S'ha fet servir `edgeR` per organitzar, filtrar i normalitzar les dades pel mètode TMM (T*rimmed Mean of M values*) i `limma+voom` per determinar l'expressió diferencial mitjançant models lineals [11].

Per identificar perfils d'expressió característics dels tumors que han patit WGD s'ha fet una anàlisi d'enriquiment mitjançant GSEA (*Gene Set Enrichment Analysis*)[12]. Es tracta d'un programa d'escriptori de lliure accés que permet utilitzar diverses col·leccions que agrupen múltiples conjunts de gens obtinguts a partir d'anotacions conegudes o a partir de dades experimentals d'expressió. A diferència d'altres mètodes d'anàlisi d'enriquiment, utilitza la llista completa de gens ordenada, sense filtrar per un llindar de significància específic. L'anàlisi té en compte el nombre de gens associats a una via i la seva posició a la llista, així com el nombre total de gens que pertanyen a la via. La significància estadística es calcula utilitzant una distribució nul·la creada mitjançant testos de permutació.

Com a base de l'estratificació clínica, els tumors de càncer colorectal es poden classificar en subtipus moleculars consens (CMS, *Consensus Molecular Subtypes*), caracteritzats per diferents alteracions a nivell genòmic i transcriptòmic. S'ha utilitzat el paquet d'R `CMSclassifier` [13] per classificar els tumors en els diferents subtipus a partir de les dades d'expressió i correlacionar-ho amb el seu estat WGD.

Per obtenir les característiques immunològiques dels tumors s'han utilitzat diferents eines, entre les quals ESTIMATE (*Estimation of Stromal and Immune cells in Malignant Tumours using Expression data*) [14], un mètode que estima els nivells d'infiltració de cèl·lules immunitàries i estromals a partir de dades d'expressió gènica. L'algoritme està implementat en llenguatge R i disponible a través del paquet `estimate`. Per altra banda, s'ha utilitzat també una versió adaptada de l'IPS (*Immunophenoscore*) [15], una eina que a partir de les dades d'expressió permet obtenir puntuacions relacionades amb característiques representatives de diferents tipus d'activitats immunitàries (resposta antigènica, activitat immunomoduladora, funcions efectores i de memòria de les cèl.lules T, i activitat immuno-supressiva). Així mateix, s'han analitzat de forma similar els nivells d'un conjunt de 12 quimiocines que representen una signatura gènica lligada a la presència d'estructures limfoides terciàries (TLS, T*ertiary Lymphoid Structures*) [16] associades a la resposta antitumoral. S'han correlacionat totes aquestes característiques immunològiques amb l'estat WGD dels tumors.

1.4. Planificació del treball

La planificació del treball s'ha fet seguint com a guia les proves d'avaluació contínua (PACs) de l'assignatura. A continuació es llisten totes les tasques incloses dins de cada PAC d'acord amb la planificació inicial. Les tasques de la PAC 2 i la PAC 3 es corresponen amb aquelles necessàries per complir els objectius definits anteriorment. La PAC 2 o fase 1 inclou totes les tasques per a l'anàlisi de la cohort de COAD; la PAC 3 o fase 2 inclou les mateixes tasques per estendre l'anàlisi a la cohort de LUAD i la integració de tots els resultats.

PAC 0

- Redacció de la proposta de TFM.

PAC 1

- Definició dels objectius generals i específics.
- Definició de la metodologia utilitzada.
- Definició de les tasques.
- Temporització de les tasques (Diagrama de Gantt).
- Redacció del pla de treball.

PAC 2

Objectiu 1:

- Descàrrega de les dades.
- Determinació de l'estat WGD.
- Anàlisi d'expressió diferencial.
- Anàlisi d'enriquiment.
- Classificació CMS i correlació amb l'estat WGD.

Objectiu 2:

- ESTIMATE i correlació amb l'estat WGD.
- Immunophenoscore i correlació amb l'estat WGD.

PAC 3

Objectiu 1:

- Descàrrega de les dades.
- Determinació de l'estat WGD.
- Anàlisi d'expressió diferencial.
- Anàlisi d'enriquiment.

Objectiu 2:

- ESTIMATE i correlació amb l'estat WGD.

- Immunophenoscore i correlació amb l'estat WGD.

- Integració dels resultats obtinguts.

PAC4

- Redacció de la memòria.

PAC5a

- Elaboració de la presentació.

PAC5b

- Preparació de la defensa pública.

S'inclou com a annex el diagrama de Gantt que defineix la temporització del projecte (Annex 1). El diagrama s'ha creat a partir d'una plantilla d'Open Office obtinguda del lloc web Vertex42.com.

Al llarg del projecte s'ha modificat la planificació pel que fa a la temporització i l'ordre d'algunes tasques. També s'han afegit algunes tasques no previstes inicialment i s'ha limitat l'extensió de l'anàlisi a només un tipus de càncer addicional enlloc de ser *pan-cancer*. L'annex 2 conté el diagrama actualitzat, incloent les modificacions en la durada real de cada tasca.

1.5. Breu sumari de productes obtinguts

El producte obtingut d'aquest treball és una memòria que introdueix la temàtica escollida, els objectius, els mètodes d'anàlisi utilitzats, els resultats i la discussió d'aquests resultats.

1.6. Breu descripció dels altres capítols de la memòria

En primer lloc s'inclou el capítol "Materials i mètodes", on es descriuen les dades i eines informàtiques utilitzades per realitzar les anàlisis. Seguidament, en el capítol "Resultats", s'expliquen detalladament els resultats obtinguts en cada una de les anàlisis. A continuació, en el capítol "Discussió", es resumeixen els resultats més destacables i es comparen amb dades ja publicades. Finalment, a l'apartat de "Conclusions", s'avalua l'assoliment dels objectius i el seguiment de la planificació proposada inicialment.

2. Materials i mètodes

2.1. Dades

Les dades utilitzades provenen del consorci internacional *The Cancer Genome Atlas* (TCGA) (PMID: 2962505). Les mostres estan identificades amb codis de barres constituïts per diversos identificadors per cada un dels següents elements: projecte, origen del teixit, participant, tipus de mostra, vial, porció, anàlit, placa i centre de seqüenciació (veure figura 1, extreta de https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode).

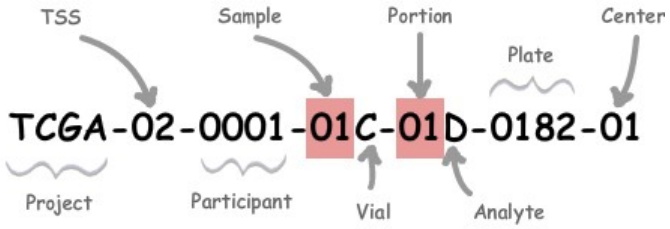

Figura 1: Codificació de les mostres de TCGA.

Hem utilitzat tres tipus d'informació sobre les mostres: dades d'expressió gènica obtingudes per seqüenciació d'ARN, dades de ploidia obtingudes mitjançant l'algoritme ASCAT i dades de l'estat WGD obtingudes a partir d'informació de *Major Copy Number* (MCN).

2.1.1. Dades d'expressió gènica

Les dades de TCGA harmonitzades es poden descarregar des del portal *Genomic Data Commons* (GDC) (https://gdc.cancer.gov/). Des d'aquest portal s'han descarregat tots els fitxers de recomptes de lectures d'*RNA-Seq* de les cohorts d'interès (projecte COAD, projecte READ i projecte LUAD). S'ha fet mitjançant una cerca en el repositori i filtrant per tipus de dades (*Gene expression quantification*), estratègia experimental (*RNA-seq*), tipus de *workflow* (*HTSeq-Counts*) i identificador de projecte (p.e.: *TCGA-COAD*). La descàrrega consta d'una carpeta comprimida que conté una carpeta per cada fitxer descarregat i un fitxer amb el nom MANIFEST.txt. Cada una de les carpetes conté el fitxer de recomptes comprimit i en alguns casos un fitxer amb el nom annotations.txt.

A més dels fitxers de recomptes, s'ha descarregat un fitxer en format TSV amb la informació sobre els fitxers de recomptes descarregats (*Sample Sheet*). Aquest fitxer conté, per cada fitxer descarregat, 8 variables corresponents a l'identificador del fitxer, el nom del fitxer, la categoria de dades, el tipus de dades, l'identificador de projecte, l'identificador de cas, l'identificador de mostra i el tipus de mostra. El tipus de mostra inclou les quatre categories

següents: *Metastatic, Primary Tumor, Recurrent Tumor* i *Solid Tissue Normal*. En aquesta anàlisi només s'han utilitzat mostres de tumors primaris. Com que els identificadors de mostra contenen informació del vial (alíquota), s'ha creat un identificador a nivell de mostra i s'han eliminat aquelles mostres repetides, deixant només una alíquota de cada una d'elles. A més, s'han eliminat les variables innecessàries, deixant només l'identificador de mostra i el nom del fitxer.

2.1.2. Dades de ploidia

Mitjançant l'algoritme ASCAT (*Allele-Specific Copy number Analysis of Tumours*) [10] es pot determinar la ploïdia dels tumors a partir de dades de seqüenciació d'ADN. S'ha utilitzat un fitxer de dades de ploïdia de tumors de TCGA obtingut amb ASCAT per determinar el seu estat WGD. Aquest fitxer conté informació de ploïdia i puresa per 8524 mostres de TCGA de diferents tipus de càncer. S'han extret els casos de les cohorts d'interès i, igual que anteriorment, s'han creat identificadors a nivell de mostra i s'han eliminat les mostres repetides. S'ha considerat que una mostra ha patit WGD si té una ploidia igual o superior a 3,5; aquestes mostres s'han classificat com a 4N. Posteriorment, per a l'anàlisi d'expressió diferencial, s'han comparat les mostres 4N amb mostres amb una ploïdia inferior a 2,5; aquestes mostres s'han classificat com a 2N. Les mostres amb ploïdia entre 2,5 i 3,5 no s'han inclòs a l'anàlisi.

2.1.3. Dades de WGD

Una altra forma d'inferir si un tumor ha patit o no WGD és utilitzant informació de *Major Copy Number* (MCN), que fa referència al nombre de còpies de l'al·lel més freqüent. Es quantifica la fracció del genoma autosòmic amb MCN igual o superior a 2 i aquells tumors amb un 50% o més es classifiquen com a WGD [3].S'han utilitzat dades inferides utilitzant aquesta aproximació per classificar els tumors. El fitxer de dades utilitzat conté informació de l'estat WGD per 6184 mostres de TCGA de diferents tipus de càncer. Igual que abans, s'han extret els casos de les cohorts d'interès, s'han creat identificadors de mostra i s'han eliminat les mostres repetides.

2.2. Anàlisi d'expressió gènica diferencial entre tumors amb i sense WGD

L'anàlisi d'expressió diferencial s'ha fet seguint un protocol que utilitza els paquets `edgeR` i `limma` del programa estadístic R [11,17].

Per preparar les dades per a l'anàlisi s'ha creat una taula d'informació de mostres que relaciona l'identificador de cada mostra amb el fitxer de recomptes corresponent i amb l'estat WGD assignat. S'ha utilitzat la funció `readDGE` del paquet `edgeR` per llegir els fitxers de

recomptes inclosos en l'anàlisi i unir-los en una sola taula. L'objecte `DGEList` creat conté tota la informació relativa a les mostres i és utilitzat en els passos posteriors.

El preprocessament de les dades inclou un procés de filtració i un procés de normalització. En general, en qualsevol estudi d'expressió hi ha un nombre considerable de gens amb baixa expressió que cal filtrar abans de fer l'anàlisi d'expressió diferencial ja que no aporten informació biològica i contribueixen a disminuir la potència estadística de l'anàlisi en fer augmentar el nombre de testos simultanis. Per filtrar els gens amb baixa expressió s'ha utilitzat la funció `filterByExpr` del paquet `edgeR`, la qual calcula els CPM (*Counts Per Million*), normalitzats per la mida de la llibreria, i utilitza un llindar de filtració que es correspon amb un recompte aproximat de 10. Es mantenen aquells gens amb CPM superiors a aquest llindar en un nombre mínim de mostres que es correspon amb la mida del grup més petit.

Per altra banda, per poder comparar les mostres en l'anàlisi d'expressió diferencial posterior, cal normalitzar. Ho fem amb el mètode TMM (*Trimmed Mean of M-values*), el qual, assumint que la majoria de gens no estan diferencialment expressats, retalla els valors M (*log-ratios*) i els valors A (*log-averages*) per estimar factors de normalització a partir dels gens expressats de forma més estable[18] .La funció `calcNormFactors` d'`edgeR` calcula un factor de normalització per cada mostra i l'afegeix a la informació de l'objecte `DGEList` per tal que pugui ser utilitzat més endavant.

Seguidament s'ha fet una exploració de les dades mitjançant gràfics de densitat, gràfic d'escalat multidimensional (MDS) i un heatmap dels gens més variables. Com que els recomptes no solen tenir una distribució normal, abans de visualitzar les dades, es transformen a log-CPM. Els gràfics de densitat dels log-CPM abans i després de filtrar permeten veure l'efecte de la filtració i permeten detectar possibles *outliers*. Per tenir una idea general de l'estructura de les dades `s'han generat gràfics` MDS mitjançant la funció `plotMDS` del paquet `limma`, la qual calcula les distàncies entre cada parell de mostres com el log fold change pels 500 gens més diferents entre elles i les representa en un gràfic de dues dimensions. Aquest gràfic permet detectar diferències globals entre les mostres i identificar possibles factors de confusió. Una altra forma de veure si hi ha diferències globals entre les mostres dels dos grups és representant en un heatmap la matriu de distàncies dels log-CPM dels gens més variables. S'ha utilitzat la funció `heatmap.2` del paquet `gplots` per representar un heatmap ordenant les files (gens) i les columnes (mostres) per similaritat.

L'anàlisi d'expressió diferencial s'ha fet mitjançant models lineals amb el paquet `limma`. Per tal de poder ajustar els models lineals cal que les dades tinguin una distribució normal i variància constant. S'ha fet servir la funció `voom` per transformar les dades per tal que puguin ser utilitzades per l'ajust dels models lineals[19]. Específicament el que fa és transformar els recomptes en log-CPM utilitzant les mides de les llibreries i els factors de normalització. A més, estima la relació mitjana-variància i utilitza aquesta predicció per assignar un pes a cada

observació, que serà utilitzat en el procés de modelat lineal. Per ajustar els models lineals s'han utilitzat les funcions `lmFit` i `contrasts.fit`. Els testos de significació dels contrastos s'han realitzat mitjançant la funció `eBayes`, que utilitza el mètode T-moderat, el qual té en compte la variabilitat de tots els gens per estimar l'error estàndar de cada gen (moderació de la variància). L'objecte creat per aquesta funció conté, per cada test, valors estadístics com l'estadístic t, l'estadístic F, el log fold-change o el p-valor.

S'ha obtingut el nombre de gens diferencialment expressats utilitzant la funció `decideTests`. Per defecte, es consideren significatius aquells contrastos amb p-valor ajustat inferior a 0.05. El mètode d'ajust dels p-valors és el de Benjamini-Hochberg (BH), el qual controla la taxa de *False Discovery Rate* (FDR). Amb la funció `topTable` s'ha generat una taula amb tots els gens ordenats en funció del p-valor. La taula conté per cada gen el log-FC, el log-CPM mitjà, l'estadístic t moderat, el p-valor, el p-valor ajustat i l'estadístic B (log-odds). Els símbols oficials associats a cada un dels identificadors Ensembl s'han obtingut mitjançant la funció `getBM` del paquet `biomaRt`, la qual permet obtenir diferents atributs de la base de dades BioMart. Els noms dels gens obtinguts de TCGA són identificadors Ensembl amb un sufix corresponent a la versió, que cal treure abans d'executar la funció `getBM`. Els resultats de l'anàlisi d'expressió diferencial també s'han representat gràficament mitjançant un gràfic mitjana-diferència, un gràfic de volcà i un heatmap dels gens més significatius. El gràfic mitjana- diferència representa el log2-fold-change de cada gen en funció del seu valor de log-CPM mitjà i permet tenir una visió global del nombre de gens sobreexpressats i subexpressats. També podem visualitzar els resultats mitjançant un gràfic de volcà, en el qual es representa la significació biològica (log2FC) a l'eix horitzontal i la significació estadística (-log10(p-valor)) a l'eix vertical. Permet marcar llindars de significància que mostren els gens diferencialment expressats. Per últim, els log-CPM dels 100 gens més significatius (amb p-valor ajustat més baix) s'han represetat en un heatmap agrupant els gens i les mostres en funció de la seva similitud. Aquest gràfic permet veure si les mostres d'un mateix grup tenen patrons globals d'expressió similars.

2.3. Anàlisi d'enriquiment amb GSEA

Per identificar perfils d'expressió enriquits en un dels grups respecte l'altre s'ha utilitzat el programa d'escriptori GSEA 4.0.3.[12]

El programa permet realitzar l'anàlisi d'enriquiment amb dues aproximacions diferents: Standard GSEA o PreRanked GSEA. En l'anàlisi amb Standard GSEA, el programa parteix de les dades d'expressió i calcula un rank per cada gen per crear una llista ordenada. Per defecte la mesura que s'utilitza per calcular el rank és *Signal2Noise* i es calcula de la manera següent:

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

Per realitzar una anàlisi d'enriquiment amb Standard GSEA cal carregar tres fitxers de dades en un format específic:

- Fitxer de dades d'expressió en format TXT que conté els valors de recomptes normalitzats (log-CPM).

- Fitxer de classes en format CLS que relaciona cada mostra amb el grup al qual pertany.

- Fitxer de *gene sets* en format GMT que defineix els *gene sets* d'una col·lecció amb el nom, la descripció i la llista de gens inclosos a cada *gene set*.

Els fitxers d'expressió i de classes s'han generat utilitzant R. El fitxer de dades d'expressió ha de contenir a la primera columna els símbols HGNC, ja que són els identificadors utilitzats en els fitxers de *gene sets*. S'han afegit els símbols a la taula de log-CPM creada amb `voom` i s'ha eliminat la columna d'identificadors ensembl. La taula resultant s'ha guardat en un fitxer en format TXT amb els camps separats per tabulacions.

El fitxer de classes ha de contenir a la primera línia el nombre de mostres i el nombre de classes, a la segona línia els noms de les classes en l'ordre d'aparició dins del fitxer d'expressió i a la tercera línia les etiquetes de classe per cada una de les mostres. S'ha guardat en format CLS amb els camps separats per tabulacions.

El programa permet utilitzar diferents col·leccions de *gene sets*. S'ha utilitzat principalment la col·lecció de processos biològics de Gene Ontology pel fet de contenir termes generals i fàcilment interpretables i ser una de les més utilitzades. Es pot descarregar de la pàgina web de GSEA per tenir-la localment o carregar-la des de la web de GSEA a l'hora de realitzar l'anàlisi.

L'anàlisi d'enriquiment es basa en el càlcul d'un ES (*Enrichment Score*) per cada *gene set* en funció del nombre de gens associats i la seva posició a la llista. A partir de l'ES es calcula un NES (*Normalized Enrichment Score*), que normalitza en funció de la mida (nombre de gens) del *gene set*. Es determina la significància estadística mitjançant el càlcul de p-valors i FDR basant-se en testos de permutació per crear la distribució nul·la. Els testos poden ser de dos tipus: permutació de *gene sets* o permutació de fenotips. El mètode per defecte i el recomanat per dades d'RNA-seq amb un nombre alt de rèpliques és la permutació de fenotips ja que té més sentit a nivell biològic i produeix menys falsos positius. Tot i així, la permutació de *gene sets* també s'utilitza, però en ser menys rigorós, es recomana que el llindar de FDR utilitzat per a la significació sigui de 0.05 en comptes de 0.25. Inicialment s'ha utilitzat la permutació de fenotips però com que no ha donat resultats significatius s'ha repetit l'anàlisi utilitzant la permutació de *gene sets*, que s'ha seguit utilitzant en els anàlisis posteriors. Inicialment s'han utilitzat tots els paràmetres per defecte i s'han provat alguns canvis per tal d'obtenir *gene sets* significatius utilitzant la permutació de fenotips, com la utilització del test t com a mesura per calcular els ranks, la qual té en compte la mida dels grups; i la utilització d'altres

bases de dades de *gene sets* com Reactome, Hallmarks (col·lecció H) o la Gene Ontology completa.

L'anàlisi d'enriquiment amb GSEA dona com a resultat un informe que detalla el nombre de *gene sets* enriquits en cada un dels grups amb diferents nivells de significació (FDR < 0.25, p-valor < 0.01 i p-valor < 0.05). Aquesta informació també es mostra en un gràfic que representa el valor de NES a l'eix de les x i el p-valor ajustat i sense ajustar a l'eix de les y. A més, genera una taula en format excel i format html per cada grup o fenotip amb els *gene sets* enriquits ordenats en funció del NES. La taula mostra, per cada *gene set*, el nombre de gens que el formen, l'ES, el NES, el p-valor sense ajustar, el p-valor ajustat per FDR i per FWER, el rank al valor d'ES màxim i algunes informacions sobre el *leading edge subset*, que és el grup de gens que contribueixen a l'enriquiment del *gene set*. Dels 20 primers *gene sets* es pot obtenir més informació, com el gràfic d'enriquiment; se'n mostra un exemple a la figura 2 (extret de https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html)

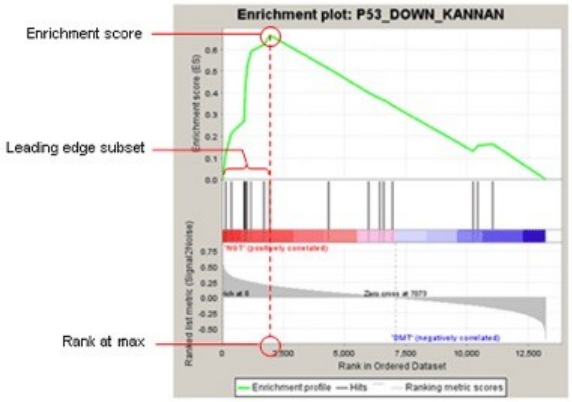

Figura 2: Exemple de gràfic d'enriquiment de la via "p53_down_kannan".

La part superior del gràfic mostra com varia l'ES al llarg de la llista ordenada de gens mentre corre l'anàlisi. El màxim representa l'ES pel *gene set*. La part intermèdia mostra la posició que ocupen els gens que pertanyen al *gene set* a la llista ordenada de gens. Els gens que apareixen abans o després (en funció de si el *gene set* està sobrerepresentat o infrarepresentat) del màxim ES formen part del *leading edge subset*. La part inferior del gràfic mostra el valor de la mesura de ranking, que correlaciona l'expressió de cada un dels gens amb el fenotip, al llarg de la llista ordenada de gens.

2.4. Classificació CMS i correlació amb l'estat WGD

En funció de l'expressió gènica les mostres de càncer colorectal es poden classificar en subtipus moleculars consens. S'ha vist que cada subtipus té unes característiques associades[13] :

- CMS1: nivells alts de cèl·lules immunitàries infiltrades, nivells alts de mutacions, inestabilitat de microsatèl·lits (MSI, *Microsatellite Instability*), nombre baix de CNA, hipermetilació.
- CMS2: diferenciació epitelial, nombre alt de CNA (inestabilitat cromosòmica), activació de vies WNT i MYC.
- CMS3: epitelial, desregulació metabòlica, nombre baix de CNA.
- CMS4: transició epiteli-mesènquima (EMT), activació de TGF-beta, angiogènesi, nivells alts de cèl·lules estromals, nombre alt de CNA.

S'ha utilitzat la funció `classifyCMS` del paquet `CMSclassifier` per classificar les mostres de COAD i READ en subtipus moleculars consens mitjançant un classificador *Random Forest*. La funció dona com a resultat les probabilitats de cada classe per cada mostra. A més, s'obté un vector (`predictedCMS`) amb la classe predita per cada tumor, sent la classe predita aquella amb probabilitat més alta superior a 0.5. En cas que cap de les classes superin 0.5, s'assigna `NA`. S'obté també un vector amb la classe més probable (`nearestCMS`). En aquest cas s'ha utilitzat el vector `predictedCMS`, assignant `NA` a les mostres poc clares.

Seguidament s'ha analitzat l'associació entre subtipus i WGD de forma gràfica i estadísticament amb un test d'independència de Fisher.

2.5. Anàlisis d'immunitat i correlació amb l'estat WGD

Per explorar les característiques immunològiques dels tumors s'han utilitzat tres aproximacions diferents, detallades a continuació.

2.5.1. ESTIMATE

ESTIMATE és un algoritme que estima els nivells d'infiltració de cèl·lules immunitàries i estromals a partir de dades d'expressió gènica mitjançant *single sample GSEA* (ssGSEA). Per cada mostra calcula un *immune score* i un *stromal score* a partir dels quals obté un *estimate score* que representa la puresa tumoral[14]. S'ha utilitzat l'algoritme implementat en el paquet `estimate` d'R.

Com a entrada s'utilitza la taula d'expressió en format GCT i amb una estructura determinada. Ha de contenir a la primera fila `#1.2`, a la segona el nombre de gens i el nombre de mostres i a la tercera els noms de les mostres. A partir de la quarta fila, la primera columna ha de contenir els símbols dels gens, la segona columna una descripció dels gens (pot estar buit) i la resta de columnes les dades, és a dir els valors d'expressió per cada una de les mostres. S'ha importat la taula d'expressió utilitzada per l'anàlisi amb GSEA i s'ha modificat tal com es demana. La funció `estimateScore` aplica l'algoritme i guarda el resultat

en un arxiu GCT que conté, per cada mostra, les tres puntuacions. Aquí només s'han utilitzat els *immune scores*. Per comparar les puntuacions entre els dos grups s'han representat les distribucions en diagrames de caixa i s'ha testat la diferència mitjançant un test de Mann-Whitney.

2.5.2. Immunophenoscore

L'*Immunophenoscore* (IPS) és una puntuació que representa la immunogenicitat tumoral i que es basa en l'expressió de 162 gens que han estat expressament identificats mitjançant R*andom forest*[15]. Aquests gens representen quatre categories específiques de resposta immunitària:

- Molècules MHC (processament d'antigens)
- Immunomoduladors / *Checkpoints*
- Cèl·lules efectores: limfòcits T citotòxics (CD8+) activats i de memòria; i limfòcits T *helper* (CD4+) activats i de memòria.
- Cèl·lules supressores: limfòcits T reguladors i MDSCs (*myeloid-derived suppressor cells*).

Les dues primeres categories inclouen gens expressats per les cèl·lules tumorals implicats en el processament i presentació d'antigens i en la modulació de la resposta immunitària; les altres dues inclouen gens expressats per diferents tipus de cèl·lules immunitàries infiltrades (veure figura 3[15]).

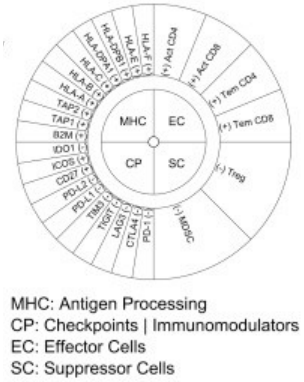

Figura 3: Immunofenograma.

L'script que s'ha utilitzat calcula per cada mostra un *z-score* pels valors d'expressió dels 162 gens i per 11 grups de gens que representen categories immunitàries. S'ha utilitzat la mateixa taula d'expressió creada per l'anàlisi amb GSEA, modificant-la per poder ser utilitzada per l'algoritme. També s'ha importat un fitxer de text amb la llista de gens inclosos en el càlcul de les puntuacions i s'ha aplicat l'algoritme. Per analitzar les diferències entre els grups s'ha fet servir el test de Mann-Whitney.

2.5.3. Signatura 12-chemokines

Aquesta signatura inclou 12 quimiocines i és indicadora dels nivells d'inflamació i de la presència de *tertiary lymphoid structures* (TLS)[16]. Els TLSs són agregacions cel·lulars organitzades amb una estructura semblant al òrgans limfoides secundaris, els quals són bàsics per a la generació de la resposta immunitària adaptativa antitumoral. Es formen en teixits perifèrics en resposta a l'exposició perllongada a senyals inflamatoris, com ara en malalties autoimmunes i infeccioses, transplantament d'òrgans, desordres inflamatoris i

tumors. Es troben a l'estroma, a la part interna o a les zones externes invasives d'una part important de tumors de diferents tipus. Estan formats per una zona rica en limfòcits T, acompanyades de cèl·lules dendrítiques madures, i un fol·licle de limfòcits B rodejat de cèl·lules plasmàtiques. Juguen un paper important en la resposta antitumoral ja que constitueixen llocs privilegiats per a la presentació d'antígens tumorals i la generació de limfòcits T efectors de memòria, limfòcits B de mèmòria i anticossos, capaces d'induir una resposta llarga i controlar la recaiguda tumoral. La seva presència està associada a un pronòstic favorable en molts tipus de càncer i té potencial per predir la resposta a la immunoteràpia.

La signatura 12-chemokines ha estat derivada a partir de la correlació amb un patró d'expressió gènica relacionat amb inflamació i està associada amb l'augment de la supervivència en càncer colorectal, melanoma i càncer de pit.

De la mateixa manera que amb els gens de l'IPS, es calcula una puntuació per cada un dels 12 gens i una puntuació mitjana global. S'ha importat la taula d'expressió (la mateixa utilitzada per a l'IPS) i el fitxer que conté la llista de gens a analitzar i s'ha aplicat l'algoritme. S'han comparat les puntuacions de cada gen entre els dos grups mitjançant testos de Mann-Whitney.

2.6. Anàlisi del metabolisme

Per explorar les diferències metabòliques trobades s'ha fet un anàlisi d'enriquiment amb GSEA amb les vies metabòliques de la base de dades KEGG (*Kyoto Encyclopedia of Genes and Genomes*). Aquelles vies amb un major enriquiment s'han analitzat mitjançant l'eina Pathview i els resultats d'expressió diferencial del conjunt de gens metabòlics[20], la qual cosa ha permès representar gràficament les vies alterades i els gens implicats. (Col·laboració amb la Dra. Marta Cascante i Carles Foguet, Universitat de Barcelona).

3. Resultats

A continuació s'exposen els resultats obtinguts per cada un dels anàlisis realitzats.

3.1. Anàlisi de la cohort de COAD

Com que a l'inici del projecte encara no es diposava de les dades de l'estat WGD en funció del MCN, s'ha començat fent un anàlisi de la cohort de COAD separant els grups en funció del nivell de ploidia, tal com s'ha especificat a l'apartat de materials i mètodes. A continuació s'ha repetit l'anàlisi utilitzant les dades de WGD inferides a partir de MCN.

3.1.1. Separació de grups en funció de ploidia

Aquest anàlisi inclou un total de 338 mostres: 239 mostres 2N (N < 2,5) i 99 mostres 4N (N>3,5).

Exploració de les dades

Abans de l'anàlisi d'expressió diferencial s'han prepocessat les dades i s'han visualitzat gràficament. Mitjançant gràfics de densitat dels log-CPM abans i després de filtrar es pot veure l'efecte de la filtració (figura 4).

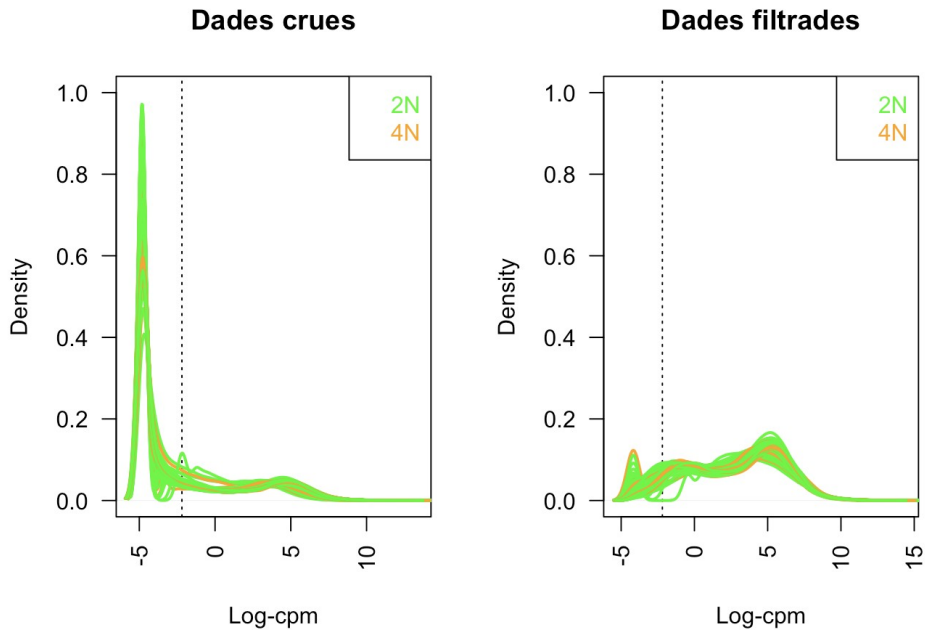

Figura 4: Gràfics de densitat dels log-CPM abans i després de filtrar. El llindar de filtració utilitzat es marca amb una línia discontínua.

Canviant l'escala i tornant a representar el gràfic de densitat dels log-CPM filtrats es pot veure bé la seva distribució i observar que la gran majoria de mostres tenen una distribució semblant (figura 5).

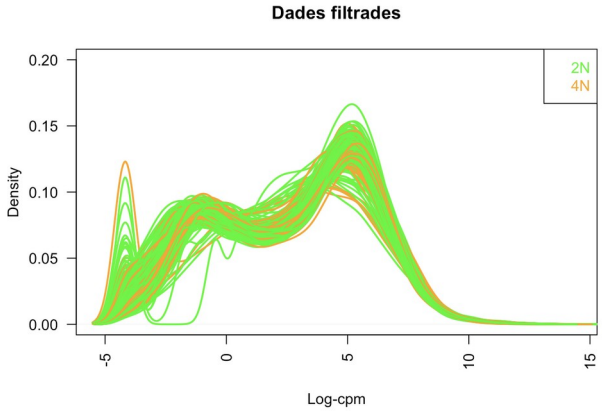

Figura 5: Gràfic de densitat dels log-CPM després de filtrar.

El gràfic d'escalat multidimensional mostra que les mostres 2N queden disperses de manera força homogènia ocupant tota l'àrea del gràfic, mentre que les mostres 4N s'agrupen en una regió del gràfic. Això indica que les mostres 4N són globalment més similars entre elles mentre que les mostres 2N tenen més variabilitat, amb una part de mostres similars a les 4N (figura 6). Això contrasta amb el fet que genòmicament els tumors 4N solen ser més heterogenis.

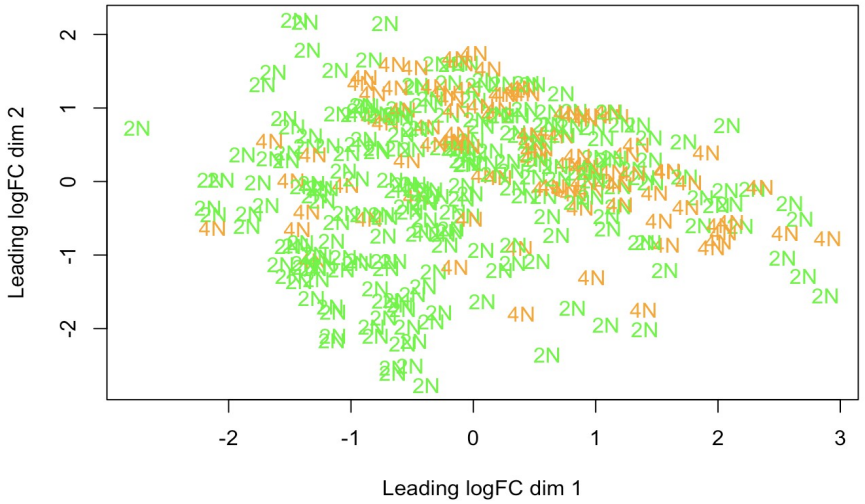

Figura 6: Gràfic d'escalat multidimensional.

Es representa també un heatmap de la matriu de distàncies dels log-CPM dels 500 gens més variables (figura 7). No s'observa cap patró particular.

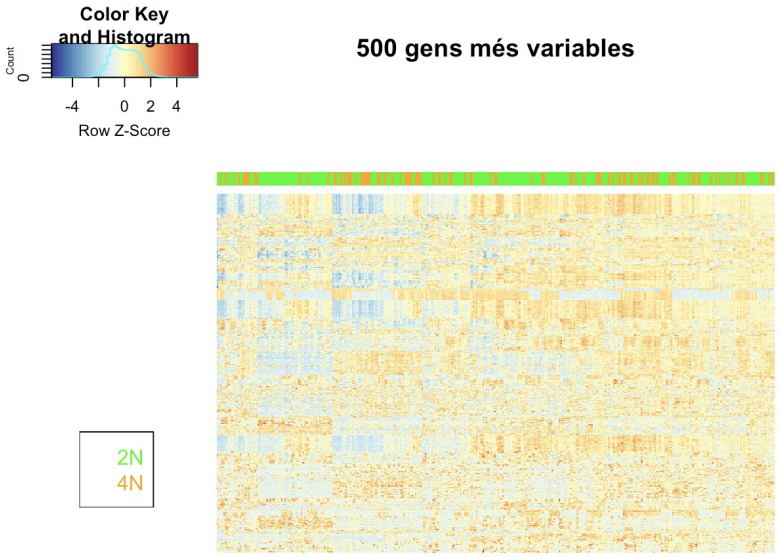

Figura 7: Heatmap dels 500 gens més variables.

Anàlisi d'expressió diferencial

L'anàlisi d'expressió diferencial identifica 1765 gens subexpressats i 1986 gens sobreexpressats utilitzant un FDR inferior a 0.05. Si només es consideren diferencialment expressats aquells gens que, a més de tenir un p-valor ajustat inferior a 0.05, tenen un log2-fold-change per sobre d'un cert llindar, el nombre de gens diferencialment expressats disminueix considerablement. Posant un llindar de log2-fold-change de 0.5 (valor absolut), aproximadament equivalent a un fold-change de 0.7 i 1.4, s'obtenen 438 gens downregulats i 754 gens sobreexpressats.

Es genera una llista de gens ordenats per significància. A la figura 8 es mostren els 10 primers gens amb el seu corresponent logFC i p-valor ajustat.

```
##      ensembl_gene_id hgnc_symbol     logFC    adj.P.Val
## 5010 ENSG00000124228       DDX27 0.6998349 1.328338e-10
## 5777 ENSG00000131043        AAR2 0.4885058 1.328338e-10
## 6012 ENSG00000132801      ZSWIM3 0.7663923 1.328338e-10
## 6010 ENSG00000132792     CTNNBL1 0.5147824 2.274391e-10
## 2352 ENSG00000101452       DHX35 0.4633117 3.907216e-10
## 2283 ENSG00000101158      NELFCD 0.5907585 5.862684e-10
## 389  ENSG00000022277        RTF2 0.4997010 5.932851e-10
## 2251 ENSG00000100982       PCIF1 0.4978491 5.932851e-10
## 2325 ENSG00000101337      TM9SF4 0.5070138 5.932851e-10
## 2357 ENSG00000101470       TNNC2 1.9949522 6.765073e-10
```

Figura 8: Primeres 10 línies de la llista de gens ordenats per significació.

La figura 9 mostra els resultats de l'anàlisi en un gràfic mitjana-diferència. Es marquen els gens sobreexpressats en vermell i els gens subexpressats en blau. S'observa que hi ha un nombre força elevat de gens diferencialment expressats.

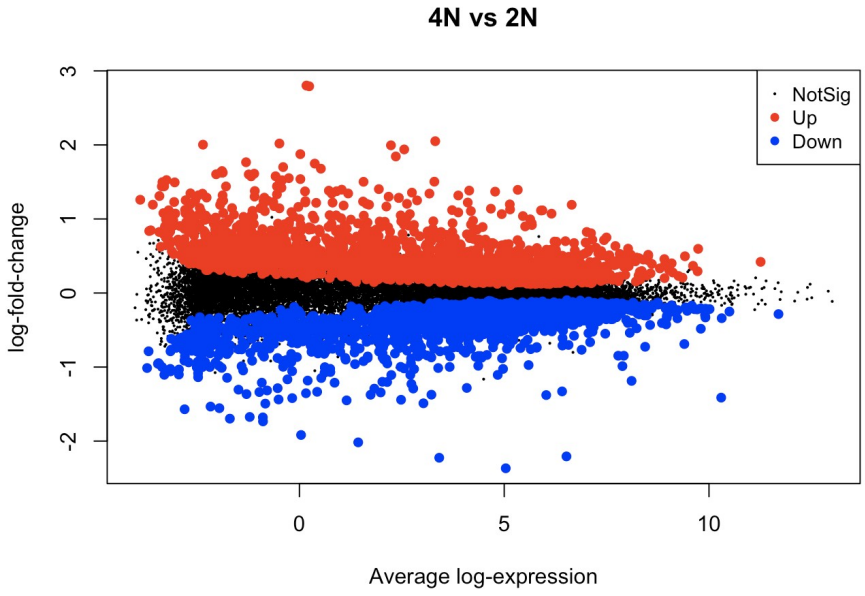

Figura 9: Gràfic mitjana-diferència.

El gràfic de volcà de la figura 10 dona una idea del nombre de gens diferencialment expressats tenint en compte també la significació biològica. La línia vermella marca un p-valor de 0.01 a l'eix vertical i les línies blaves marquen valors de fold change de 0.5 i 2 a l'eix horitzontal.

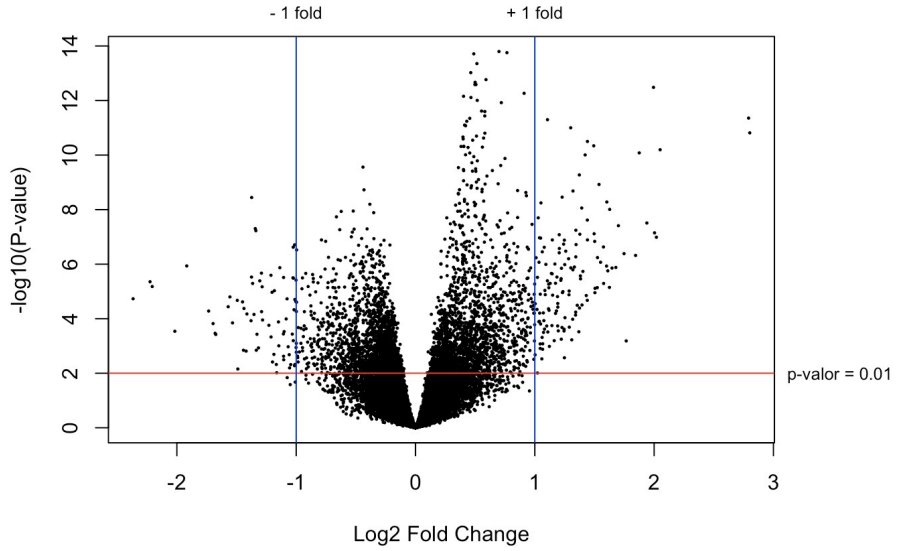

Figura 10: Gràfic de volcà.

A la figura 11 es representen en un heatmap els log-CPM dels 100 gens amb p-valor ajustat més baix, reordenant els gens i les mostres en funció de la seva similitud. S'observa que les mostres estan més o menys agrupades segons el grup al qual pertanyen, amb la majoria de les mostres 2N a l'esquerre del gràfic i la majoria de les mostres 4N a la dreta, tot i que s'intercalen algunes mostres del grup contrari. En general, la majoria de gens representats estan sobreexpressats en el grup 4N.

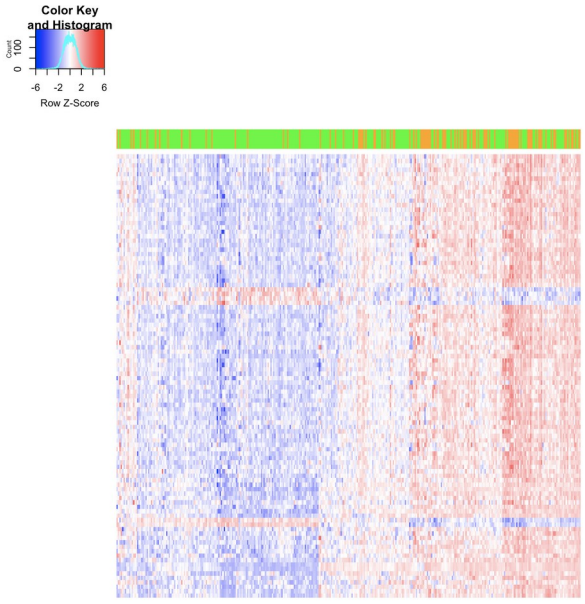

Figura 11: Heatmap dels 100 gens més significatius.

S'obté un resultat diferent si s'agafen els 100 gens amb p-valor ajustat més baix després d'haver filtrat per aquells gens amb un logFC en valor absolut superior a 0.5 (aproximadament equivalent a FC superior a 1.4 o inferior a 0.7) (figura 12).

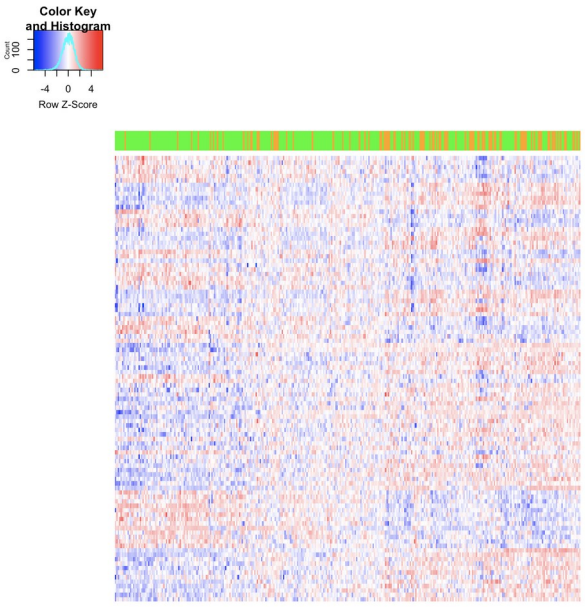

Figura 12: Heatmap dels gens més significatius havent filtrat per log-FC.

En aquest cas les mostres no queden tan ben agrupades i s'observen grups de gens sobreexpressats i grups de gens subexpressats en els dos costats del gràfic.

Anàlisi d'enriquiment

A partir de les dades d'expressió normalitzades generades per la funció `voom`, s'ha fet un anàlisi d'enriquiment amb GSEA per identificar perfils d'expressió enriquits en un dels grups respecte l'altre. Inicialment s'ha fet l'anàlisi utilitzant la permutació de fenotips i, tot i que sí que hi ha alguns *gene sets* amb p-valor inferior al 0,05 (9 *gene sets* enriquits a 4N i 111 a 2N),

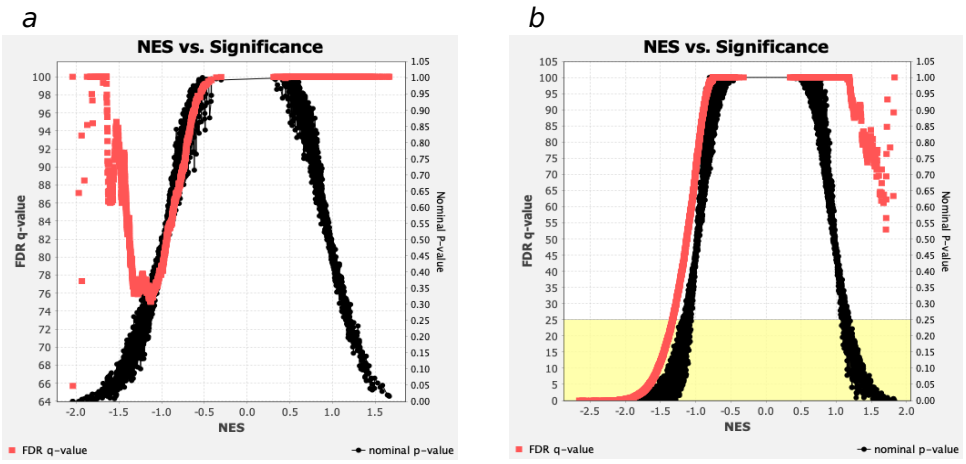

Figura 13: Gràfics p-valor vs NES de l'anàlisi amb permutació de fenotips (a) i amb permutació de gene sets (b)

cap d'ells té un p-valor ajustat (FDR) inferior al 0,25 (veure gràfic p-valors vs NES de la figura 13a).

Tot i així, mirant la llista dels 20 primers *gene sets* enriquits en cada un dels grups s'observen algunes tendències. D'entre els *gene sets* enriquits al grup 4N destaquen termes relacionats amb metabolisme com *Regulation of glycogen metabolic process*, *Positive regulation of organic acid transport* o *Positive regulation of lipid storage*; i termes relacionats amb neurologia com *Regulation of postsynaptic density organization*. Per altra banda, dins el grup 2N també s'hi troben termes metabòlics com *Regulation of cellular respiration* o *Cell redox homeostasis* ; i termes relacionats amb immunitat com *Regulation of response to interferon gamma* o *Natural killer cell-mediated immunity*.

S'ha repetit l'anàlisi utilitzant permutació de *gene sets* en comptes de permutació de fenotips, menys rigurós pel que fa a significació dels testos. En aquest cas sí que s'han obtingut *gene sets* enriquits amb FDR inferior al 0,25, en concret 786 *gene sets* subregulats en el grup 4N. Posant com a llindar de significància un p-valor del 0,05, s'han obtingut 84 *gene sets* enriquits al grup 4N i 708 *gene sets* enriquits al grup 2N (veure gràfic p-valors vs NES de la figura 13b). Observant la llista dels 20 primers *gene sets* enriquits al grup 4N tornem a veure un enriquiment, aquest cop amb p-valors més baixos, en termes metabòlics relacionats amb el metabolisme de polisacàrids i el transport de ions; termes relacionats amb funcionament neuronal; i també termes associats al desenvolupament com *Embryonic digestive tract development*. Altrament, en la llista de gens subregulats en el grup 4N, aquest cop amb valors FDR propers a 0, tots els termes estan relacionats amb processos immunològics, confirmant i accentuant els resultats obtinguts anteriorment. A la figura 14 es mostren els gràfics

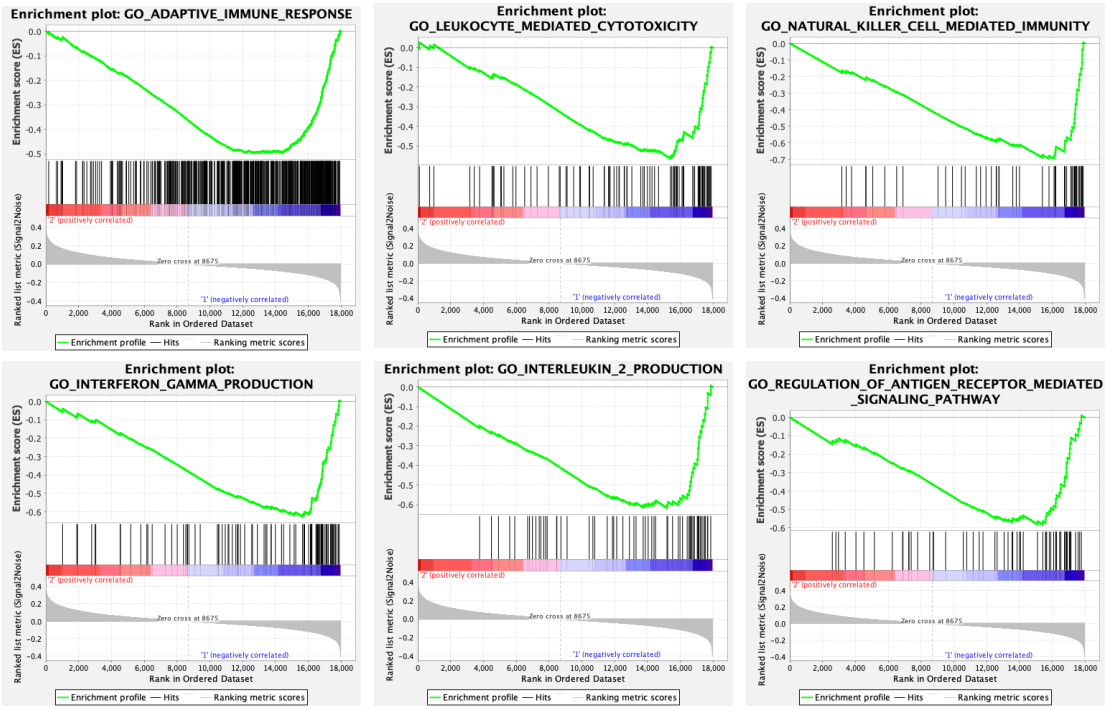

Figura 14: Gràfics d'enriquiment d'alguns dels gene sets relacionats amb la immunitat enriquits al grup 2N.

d'enriquiment d'alguns dels *gene sets* enriquits al grup 2N. Cal destacar que es tracta de vies relacionades amb l'acció de cèl·lules que inhibeixen la progressió tumoral, com ara cèl·lules NK i cèl·lules T citotòxiques i les seves molècules efectores (IFN gamma, IL-2).

Un informe més complet de l'anàlisi es pot trobar a l'annex 3.

Per optimitzar l'anàlisi i veure si surten *gene sets* significatius utiltizant la permutació de fenotips, s'han provat alguns canvis com la utilització del test t en comptes del *signal2noise* com a mesura per calcular els ranks. S'han obtingut resultats similars als del primer anàlisi, sense cap *gene set* significatiu. A més, s'han provat altres bases de dades de *gene sets*: Reactome, Hallmarks (col·lecció H) i Gene Ontology (completa). Utilitzant Reactome s'ha obtingut un sol *gene set* significatiu subregulat en el grup 4N: *Regulation signaling by CBL*, relacionat amb el sistema immunitari. Les altres bases de dades no han donat resultats significatius.

Canvi en el llindar de ploidia

Un altre factor que podria fer que no es detectessin diferències significatives és el que fet que els grups no siguin homogenis o hi hagi solapament. Com que en el gràfic MDS i en el heatmap de resultats de l'anàlisi d'expressió diferencial sembla que un grup gran de mostres 2N s'agrupa amb les mostres 4N, s'ha especulat que el grup 2N podria presentar una variabilitat gran degut a diferències de ploïdia. Per aquest motiu s'ha repetit l'anàlisi reduint el llindar de ploïdia del grup 2N, incloent només aquelles mostres amb ploïdia igual o inferior a 2 en comptes del llindar inicial de 2,5. Utilitzant aquest nou llindar de ploïdia, l'anàlisi comprèn 98 mostres en el grup 2N (ploïdia igual o inferior a 2) i 99 mostres en el grup 4N (ploïdia igual o superior a 3,5). A la figura 15 es mostra el gràfic MDS obtingut durant l'exploració de les dades.

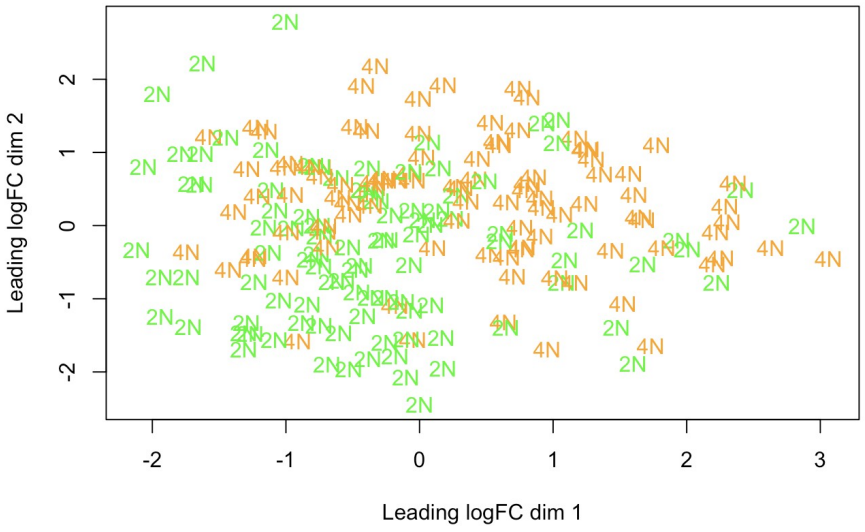

Figura 15: Gràfic d'escalat multidimensional de l'anàlisi utilitzant un llindar de ploidia diferent.

Tot i que també hi ha solapament entre els dos grups, en general sembla que queden més ben separats. Pel que fa als resultats de l'anàlisi d'expressió diferencial, utilitzant com a criteri un p-valor ajustat inferior a 0,05, s'obtenen 1784 gens subexpressats i 2094 gens sobreexpressats. Si, a més, es posa un llindar de log2-fold-change de 0,5, s'obtenen 594 gens subexpressats i 847 gens sobreexpressats. La figura 16 mostra el heatmap dels 100 gens amb p-valor ajustat més baix.

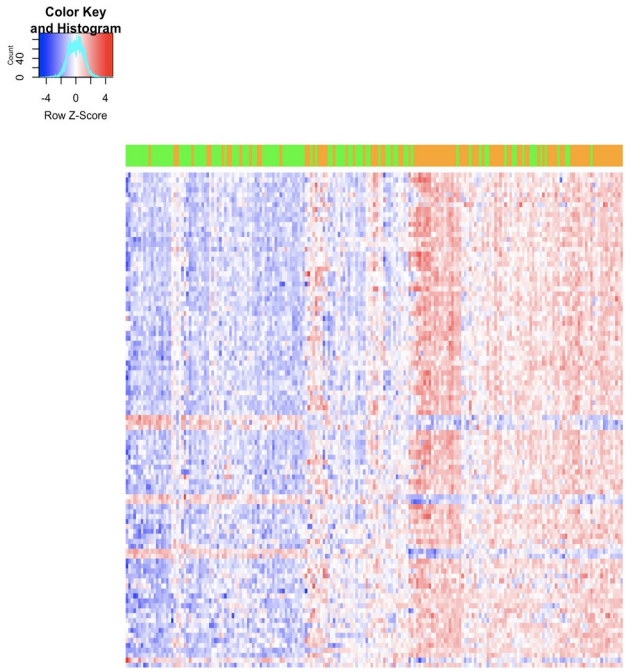

Figura 16: Heatmap dels 100 gens més significatius de l'anàlisi utilitzant un llindar de ploidia diferent.

L'agrupament de mostres és semblant a l'obtingut en l'anàlisi anterior, amb la majoria de mostres 4N agrupades a la dreta del gràfic i les mostres 2N a l'esquerre però també amb algunes mostres intercalades en el grup contrari. Els resultats de l'anàlisi d'enriquiment amb testos de permutació de fenotips són semblants als obtinguts en la primera anàlisi, sense cap *gene set* enriquit amb FDR superior a 0,25 i amb alguns termes immunològics a la part superior de la llista de *gene sets* enriquits en el grup 2N. Així doncs, el canvi en el llindar de ploidia pel grup 2N no ha fet variar els resultats de l'anàlisi d'enriquiment.

Anàlisi d'immunitat

Per explorar els canvis immunitaris trobats en l'anàlisi d'enriquiment, s'han utilitzat tres aproximacions diferents ESTIMATE, IPS i 12-chemokines, detallades a l'apartat de materials i mètodes.

Mitjançant ESTIMATE s'han calculat els nivells d'infiltració immunitària dels tumors inclosos a l'anàlisi i s'han observat diferències significatives entre els dos grups, sent lleugerament més baixos en el grup 4N (figura 17).

Aplicant l'*script* modificat de l'IPS s'obté una taula amb l'estadístic i el p-valor per cada un dels tests. Dels 168 tests, 93 són significatius amb p-valor inferior a 0,05. Es representen en diagrames de caixa les puntuacions per cada un dels gens de les categories MHC (processament d'antigens) i CP (immunomoduladors) i per cada un dels tipus cel·lulars (figures 18, 19 i 20).

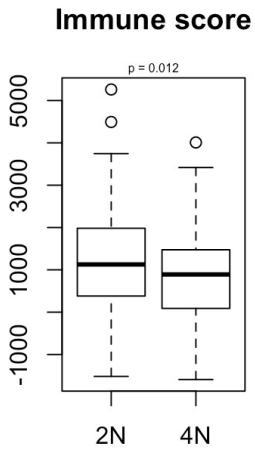

Figura 17: Resultats obtinguts amb l'aplicació de l'algoritme ESTIMATE.

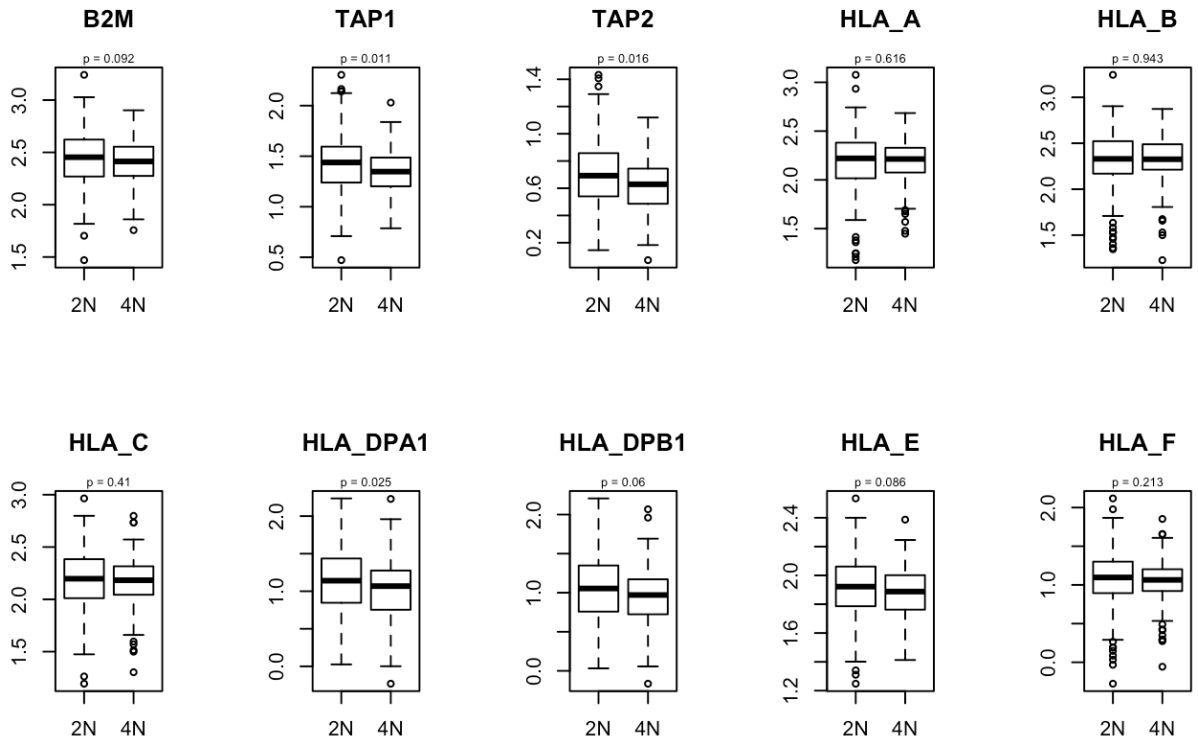

Figura 18: Puntuacions obtingudes pels gens de la categoria MHC, relacionada amb el processament d'antigens.

Els gens de la categoria MHC estan associats amb la presentació de neoantígens, importants per tal que les cèl·lules immunitàries puguin reconèixer i eliminar les cèl·lules tumorals. S'observa que estan majoritàriament subexpressats en el grup 4N; TAP1, TAP2 i HLA-DPA1 són significatius.

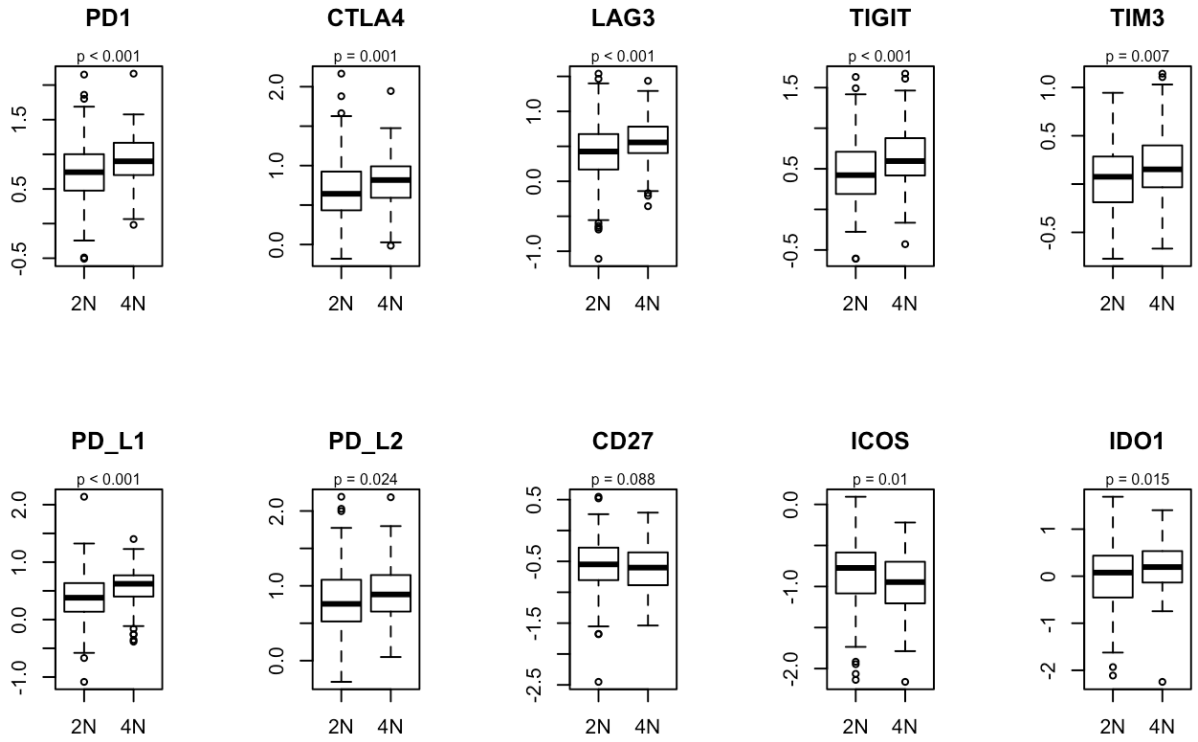

Figura 19: Puntuacions obtingudes pels gens de la categoria CP (immunomoduladors).

Els gens de la categoria CP (Checkpoints) modulen l'activitat immunitària mitjançant la presentació d'antígens que poden activar o inhibir la resposta immunitària. ICOS i CD27 activen la resposta mentre que la resta la inhibeixen. S'observa que tots els inhibidors estan

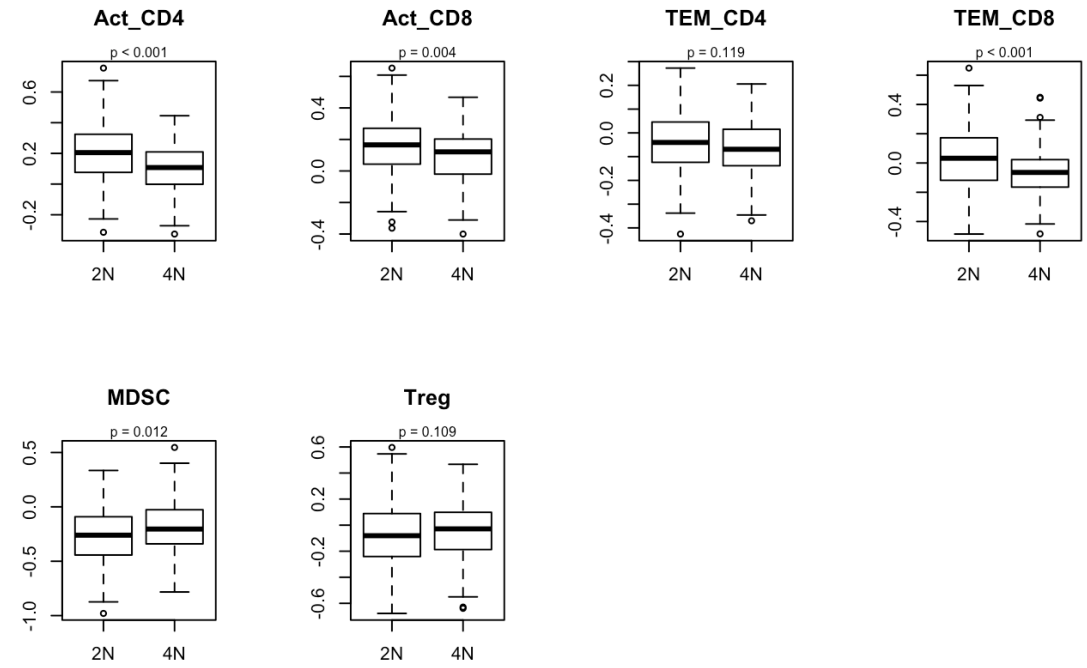

Figura 20: Puntuacions obtingudes pels diferents tipus cel·lulars efectors (a dalt) i supressors (a baix).

significativament sobreexpressats en el grup 4N mentre que els activadors estan subexpressats.

Els gens de tres tipus de cèl·lules efectores (limfòcits T citotòxics activats i de memòria i limfòcits T *helper* activats) estan significativament subexpressats en el grup 4N. No hi ha difèrencies significatives en l'expressió dels gens dels limfòcits T *helper* de memòria. Per altra banda, els gens de les cèl·lules supressores MDSC estan significativament sobreexpressats en el grup 4N. No hi ha diferències significatives en l'expressió dels gens dels limfòcits T reguladors tot i que els quartils de la distribució del grup 4N són lleugerament més alts.

Pel que fa a la signatura 12-chemokines, dels 13 testos realitzats, 10 són significatius amb p-valor inferior a 0,05, inclòs el test per la mitjana de les 12 quiomicines. Les puntuacions de cada gen individual es representen a la figura 21.

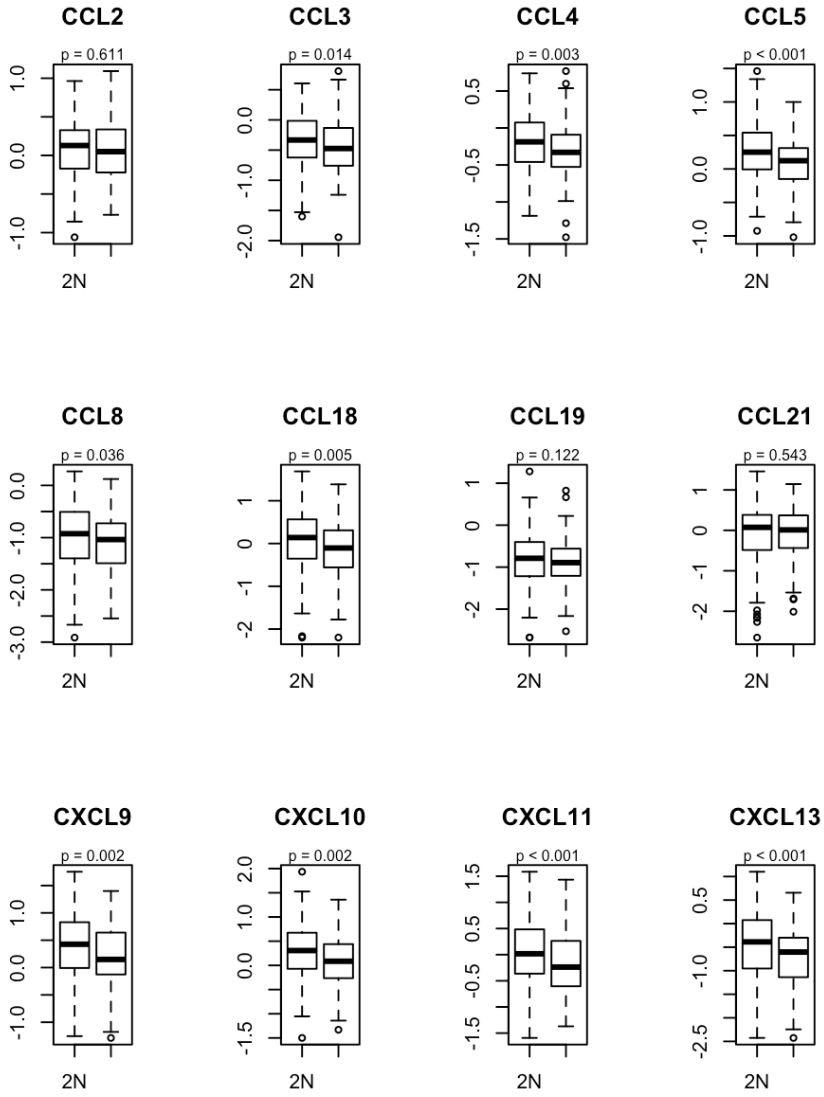

Figura 21: Puntuacions obtingudes per les quimioquines incloses a la signatura 12-chemokines.

Excepte CCL2, CCL19 i CCL21, la resta de quimiocines estan subexpressades en el grup 4N.

En conjunt, els resultats d'aquestes anàlisis mostren una correlació negativa entre WGD i resposta immunitària antitumoral.

Anàlisi del metabolisme

L'anàlisi d'enriquiment de vies metabòliques de KEGG no ha donat resultats significatius. Tot i així, s'han analitzat més detalladament aquelles vies més enriquides mitjançant la integració dels resultats de l'anàlisi d'expressió diferencial relatius a gens metabòlics. D'un total de 1724 gens associats al metabolisme s'han trobat 247 gens sobreexpressats i 410 gens subexpressats (p-valor ajustat inferior a 0,25). S'ha fet servir l'eina Pathview per identificar les vies metabòliques associades a aquests gens i s'ha observat una disminució al grup 4N en l'expressió de gens implicats en el metabolisme *one-carbon*, en la via de senyalització Hif-1, el metabolisme d'èters lípids i la biosíntesi d'arginina. Dins de la via de Hif-1 destaca la subexpressió de NOS pel paper que té en la resposta immunitària. A la figura 22 es mostra el mapa metabòlic d'aquesta última via.

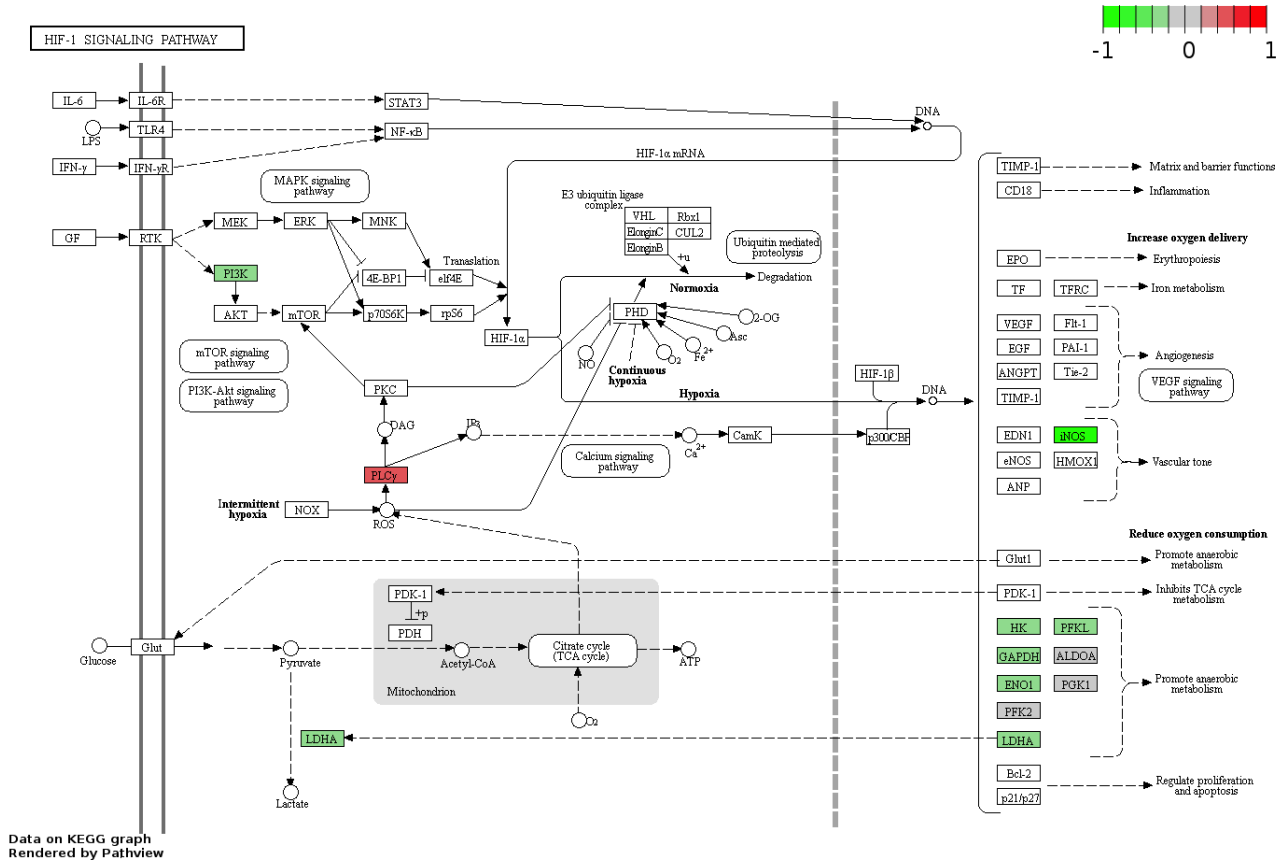

Figura 22: Mapa metabòlic de la via de senyalització Hif-1. Els gens sobreexpressats a 4N es marquen en vermell i els gens subexpressats es marquen en verd. La intensitat del color reflexa el valor del fold-change.

3.1.2. Separació de grups en funció de l'estat WGD inferit a partir de MCN

Aquesta anàlisi inclou 157 mostres: 107 mostres sense WGD i 50 mostres amb WGD. L'exploració de les dades prèvia a l'anàlisi no mostra diferències respecte la primera anàlisi.

Com a resultat de l'anàlisi d'expressió diferencial s'han obtingut 417 gens subexpressats i 460 gens sobreexpressats (FDR < 0,05). Comparat amb l'anàlisi utilitzant les dades de ploïdia, s'obté un nombre molt més baix de gens diferencialment expressats i amb p-valors més alts. Això s'explicaria pel fet que en aquesta anàlisi s'han utilitzat aproximadament la meitat de mostres respecte la primera anàlisi i que la separació de grups no és tan clara, ja que inclou totes les mostres de la cohort, mentre que en l'anàlisi amb dades de ploïdia s'han descartat aquelles mostres amb ploidies intermèdies.

En aquest cas, l'anàlisi d'enriquiment utilitzant permutació de *gene sets* ha donat resultats significatius (FDR < 0,25) en els dos sentits, amb 518 *gene sets* enriquits en el grup 4N i 545 *gene sets* enriquits en el grup 2N (figura 23).

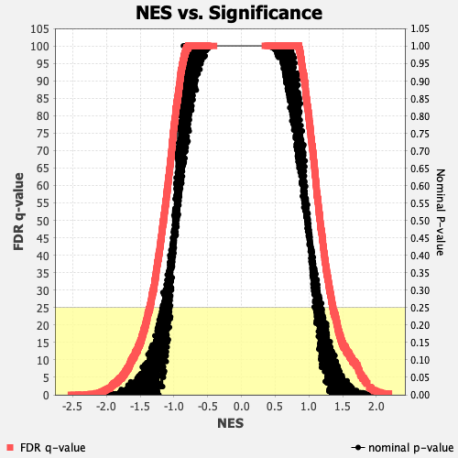

Figura 23: Gràfic p-valor vs NES obtingut en l'anàlisi d'enriquiment de la cohort de COAD (separació de grups per MCN).

Els resultats són similars als obtinguts en la primera anàlisi, amb un clar enriquiment de termes immunitaris en el grup sense WGD. Entre els 20 primers *gene sets* enriquits en el grup amb WGD també hi apareixen alguns termes metabòlics, com *Amine transport*, termes

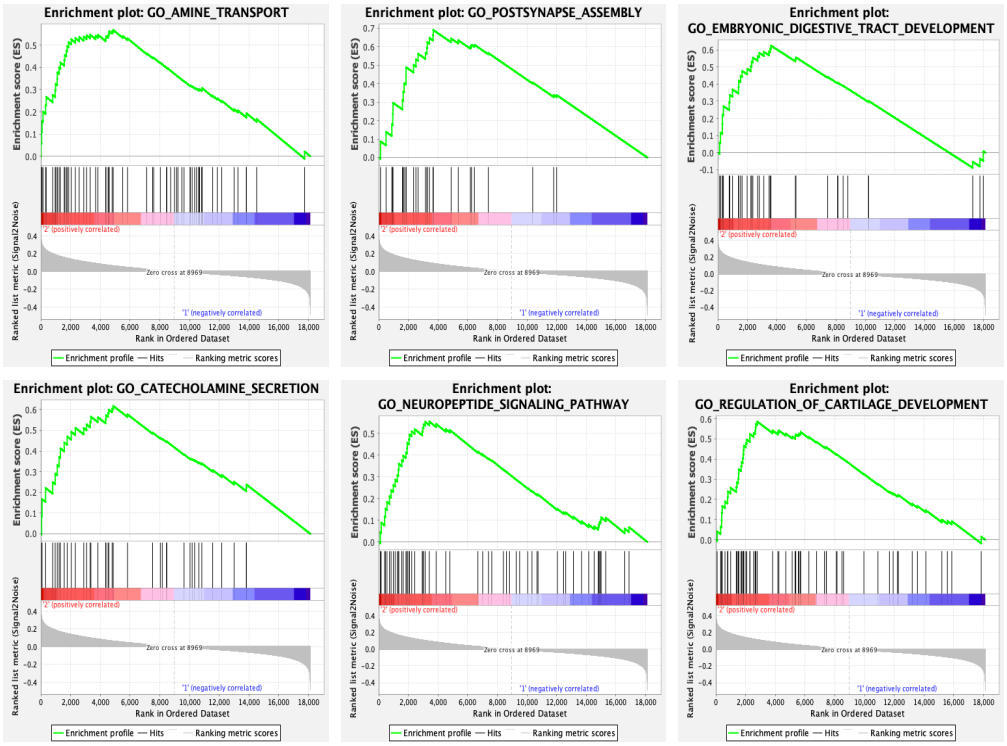

Figura 24: Gràfics d'enriquiment d'alguns dels gene sets enriquits al grup amb WGD.

relacionats amb funcionament neuronal i termes associats al desenvolupament i la morfogènesi. Es mostren els gràfics d'enriquiment d'alguns dels primers 20 *gene sets* enriquits (figura 24).

L'annex 4 conté un informe de resultats més complet de l'anàlisi.

En relació a les anàlisis d'immunitat, els resultats continuen mostrant una resposta immunitària antitumoral enriquida al grup sense WGD però no tant accentuada com en la primera anàlisi. En aquest cas no s'observa una disminució significativa dels nivells globals de cèl·lules immunitàries al grup WGD. A més, cap dels gens de processament d'antigens està diferencialment expressat i només tres dels immunomoduladors (LAG3, TIGIT i PD-L1) estan sobreexpressats significativament en el grup amb WGD. Pel que fa a les cèl·lules efectores, continuen significativament subexpressades excepte els T *helper* de memòria, igual que anteriorment; mentre que en aquest cas cap dels tipus cel·lulars amb funció supressora està sobreexpressat. Només dues (CCL4 i CCL5) de les dotze quimioquines estan downregualdes significativament. A l'annex 5 es troben els resultats complets de l'anàlisi d'immunitat.

Pel que fa a les anàlisis de gens metabòlics, s'han trobat 133 gens sobreexpressats i 197 gens subexpressats. Els canvis més importants es troben en les vies de fosforilació oxidativa, el metabolisme de sucres amino, el metabolisme d'esfingolípids, l'activitat d'hidrolases lisosomals i el metabolisme *1-carbon*, on l'expressió de diversos gens està subexpressada al grup amb WGD; i en el metabolisme d'acid araquidònic, on l'expressió està generalment

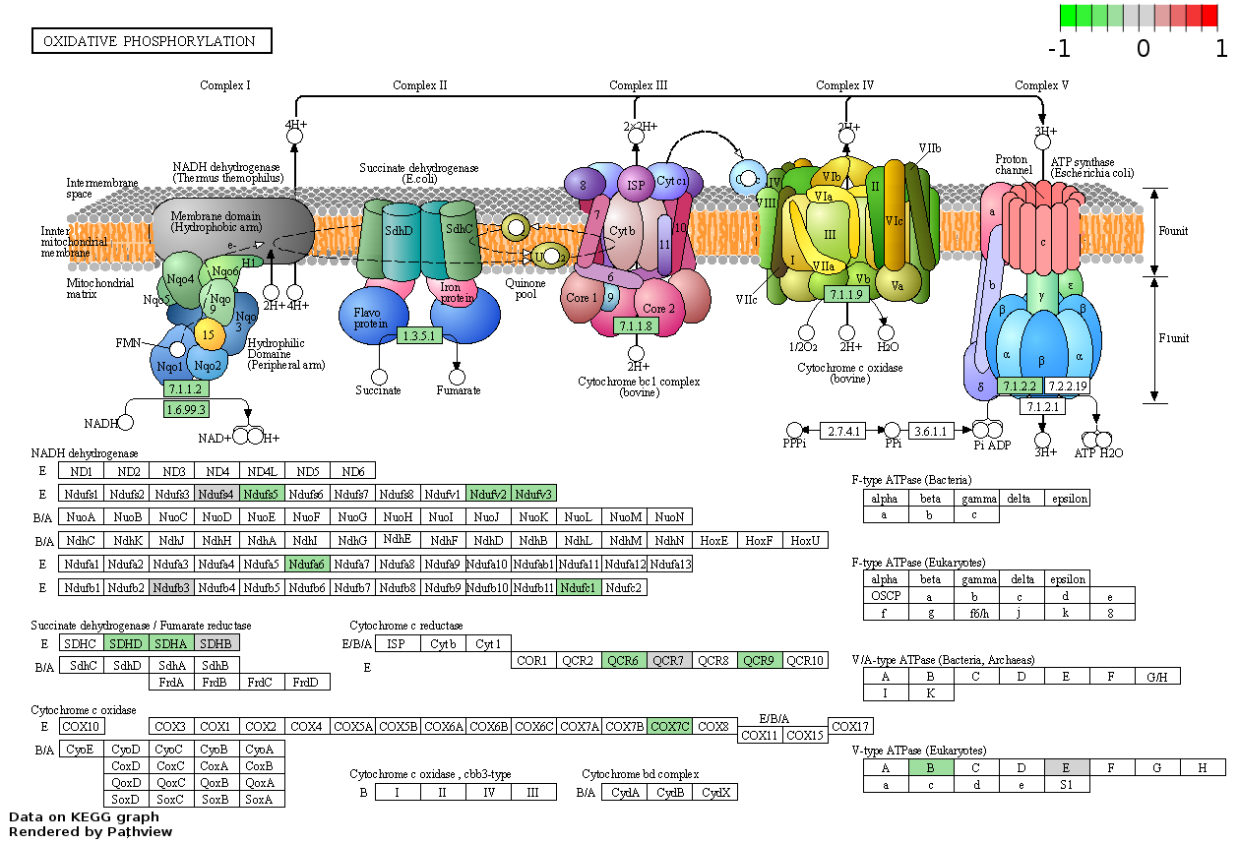

Figura 25: Mapa metabòlic de la via de foforilació oxidativa. Els gens sobreexpressats al grup amb WGD es marquen en vermell i els gens subexpressats es marquen en verd. La intensitat del color reflexa el valor del fold-change.

incrementada. També s'observa una tendència a la baixa en gens de la via de la glicòlisi, tot i que menys important. La representació d'algunes d'aquestes vies es mostra a la figura 25 i 26.

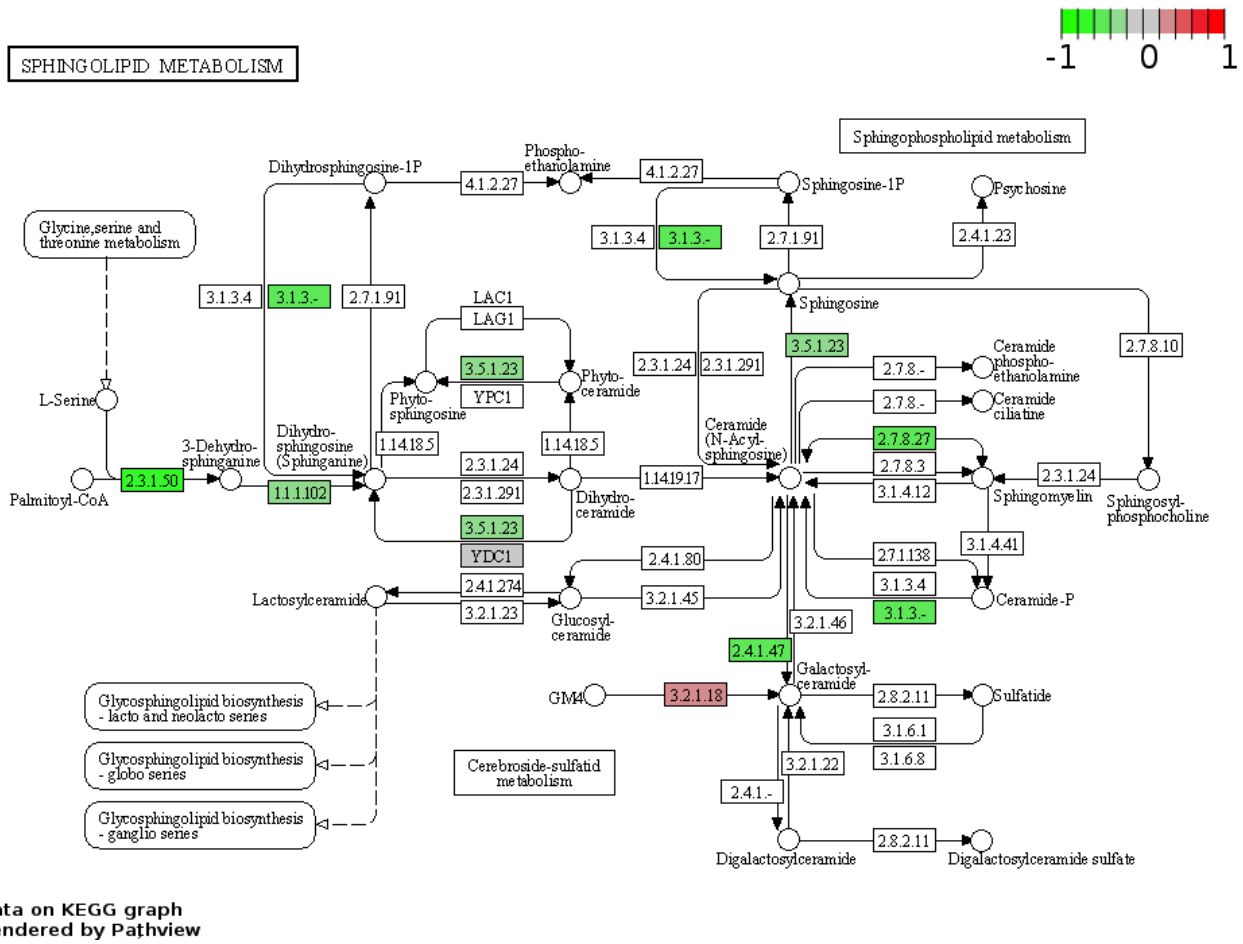

Figura 26:Mapa del metabolisme dels esfingolípids. Els gens sobreexpressats al grup amb WGD es marquen en vermell i els gens subexpressats es marquen en verd. La intensitat del color reflexa el valor del fold-change.

Cal destacar que les alteracions en el metabolisme de sucres amino, àcid araquidònic i esfingolípids podrien estar relacionades amb la baixada de la resposta immunitària al grup amb WGD.

(Col·laboració de Carles Foguet i Dra. Marta Cascante, Universitat de Barcelona)

3.2. Anàlisi de la cohort de READ

Els mateixos procediments d'anàlisi s'han repetit amb la cohort de READ ("Rectum Adenocarcinoma"), histològicament semblant a COAD, inferint l'estat WGD mitjançant les dues aproximacions. Cal destacar que en els dos casos es tenen moltes menys mostres que en les anàlisis de COAD (veure taula suplementària 1). En l'anàlisi separant els grups en funció de la ploïdia es tenen 110 mostres: 71 en el grup 4N i 39 en el grup 2N. En l'anàlisi separant

en funció de MCN encara es redueix més el nombre de mostres: 30 sense WGD i 19 amb WGD.

En l'exploració de les dades es pot veure que en tots dos casos el gràfic MDS no mostra agrupament de mostres en funció del grup, tal com es veia en COAD, i l'anàlisi d'expressió diferencial no dona resultats significatius, és a dir que no hi ha gens diferencialment expressats amb FDR inferior a 0,05.

En l'anàlisi d'enriquiment amb les dades de ploïdia, pel que fa a l'enriquiment en el grup 4N, s'obté només un *gene set* significatiu. Entre els primers 20 *gene sets* hi apareixen dos termes lligats al metabolisme, tot i que no en la mateixa línia que els observats en COAD, però sí que s'hi observen alguns termes de desenvolupament. Pel que fa a l'enriquiment al grup 2N, s'obtenen 354 *gene sets* significatius, tot i que dins dels primers 20 *gene sets* només apareixen dos termes relacionats amb la immunitat; en canvi, destaca l'enriquiment en termes relacionats amb el metabolisme bioenergètic o fosforilació oxidativa, els quals no es veien en l'anàlisi amb GSEA de la cohort de COAD però sí que es detectaven en els anàlisis metabòlics. Es mostren els gràfics d'enriquiment per alguns d'aquests termes (figura 27).

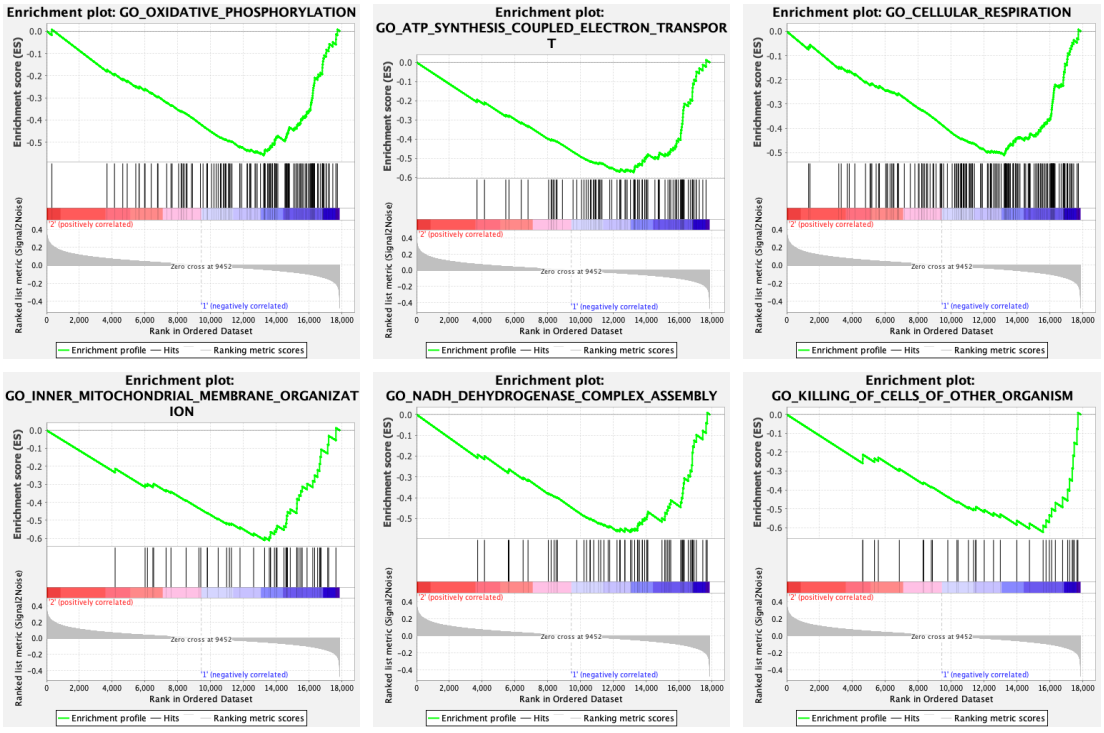

Figura 27: Gràfics d'enriquiment d'alguns dels gene sets enriquits al grup 2N de la cohort de READ.

L'annex 6 conté un informe més complet dels resultats d'aquestes anàlisis.

Utilitzant les dades de MCN només apareixen 5 *gene sets* enriquits significativament al grup amb WGD i un al grup sense WGD. Això podria ser degut a la disminució en el tamany mostral, que en aquest cas és considerable. Tot i així, seguint amb la tendència vista en COAD, dins de la llista d'enriquiment al grup amb WGD apareixen termes relacionats amb

funcionament neuronal i amb morfogènesi. També s'hi observen alguns termes metabòlics com *Oligosaccharide biosynthetic process* o *Arginine metabolic process*. Per altra banda, entre els termes enriquits en el grup sense WGD, apareixen alguns termes relacionats amb immunitat i altres associats al funcionament muscular. L'annex 7 conté un informe més complet dels resultats d'aquestes anàlisis.

3.3. Classificació CMS de les mostres de COAD i READ

S'han classificat els tumors de COAD i READ en subtipus moleculars consens (CMS) per identificar possibles factors de confusió. Es mostra l'anàlisi per les mostres agrupades segons la ploïdia. Els resultats amb les mostres agrupades en funció de MCN són semblants (annex 8).

En el cas de COAD s'han classificat 286 mostres amb una probabilitat superior a 0,5. Això representa un 84.6% de les mostres. La figura 28a mostra el gràfic MDS en el qual s'identifiquen les mostres en funció del grup (2N o 4N) al qual pertanyen i els subtipus moleculars assignats. Es pot observar que els subtipus CMS1 es concentren a la part superior del gràfic i es corresponen majoritàriament amb mostres del grup 2N. També els subtipus CMS3, localitzats a la part central, són majoritàriament del grup 2N. Les mostres del grup 4N, generalment localitzades a la part inferior del gràfic són majoritàriament de subtipus CMS2 i CMS4. Aquestes associacions es poden veure també mitjançant diagrames de barres de les freqüències relatives (figura 28b).

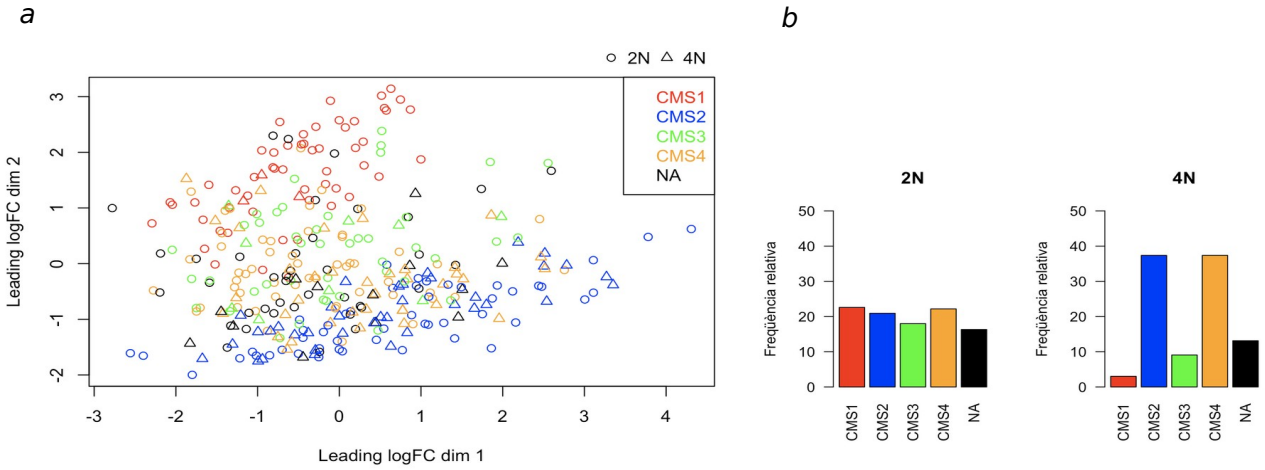

Figura 28: Classificació CMS de la cohort de COAD. a) Gràfic d'escalat multidimensional. b) Diagrames de barres de les freqüències relatives.

Es confirma així que les mostres del grup 4N són majoritàriament CMS2 i CMS4, caracteritzats per nivells alts de CNA, mentre que el grup 2N inclou els quatre subtipus en proporcions semblants. El test d'indepèndencia de Fisher (p-valor = 1.587e-07) conclou que hi

ha diferències significatives entre els grups 2N i 4N pel que fa a la distribució dels subtipus de tumors.

Els resultats de la classificació CMS de les mostres de READ es mostra a la figura 29. Per una banda s'observa que hi ha molt poques mostres CMS1; per altra banda, els altres subtipus queden més o menys agrupats però no sembla que hi hagi associació amb els grups.

Tant els diagrames de barres com el test de Fisher (p-value = 0.4023) demostren que els dos grups tenen distribucions semblants, contràriament al que es veia en COAD.

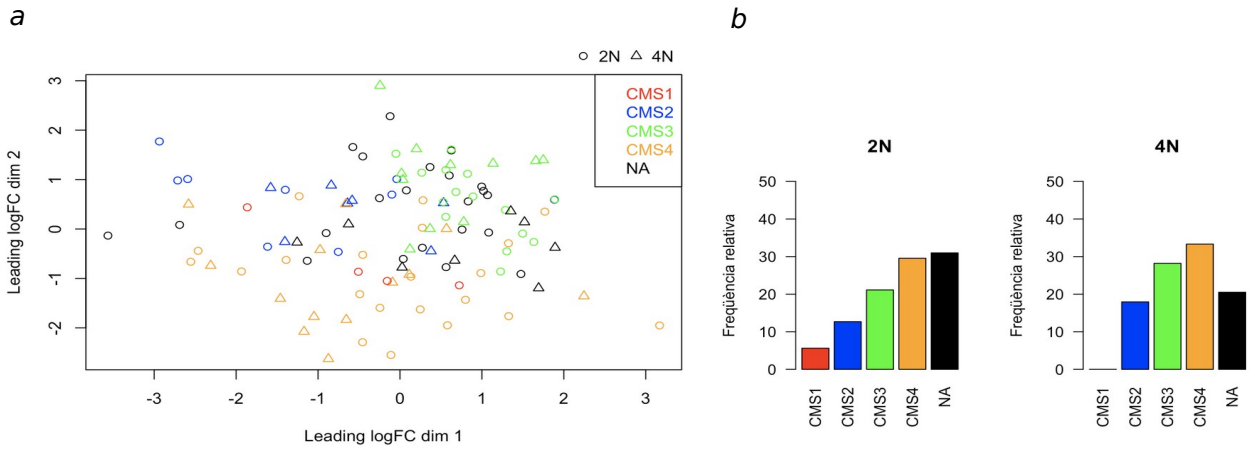

Figura 29: Classificació CMS de la cohort de READ. a) Gràfic d'escalat multidimensional. b) Diagrames de barres de les freqüències relatives.

Aquí s'ha de destacar que els subtipus CMS1 es caracteritzen perquè presenten MSI. En la cohort de READ gairebé no hi ha mostres tumorals d'aquest subtipus mentre que en la cohort de COAD sí que n'hi ha i estan majoritàriament en el grup 2N. Això ja es mostra en un estudi publicat (Bielski et al. 2018) en què es demostra que la taxa de WGD varia en funció del tipus i subtipus de càncer. Particularment, es troba que la taxa de WGD en els tumors de càncer colorectal és del 36% i que cap d'aquests tumors presenten MSI. Aquest patró també es troba en càncer endometrial i càncer d'estómac.

Els subtipus CMS1 també es caracteritzen per un enriquiment en la infiltració immunitària. Així doncs, les diferències pel que fa a la immunitat trobades en la cohort de COAD, molt més febles en la cohort de READ, podrien estar amplificades per l'enriquiment en subtipus CMS1 en el grup 2N.

Per validar l'associació entre WGD i canvis en la immunitat s'ha reptetit l'anàlisi en la cohort de COAD excloent les mostres CMS1. Els resultats es mostren a la secció següent.

3.4. Anàlisi de la cohort de COAD excloent les mostres CMS1

S'ha repetit l'anàlisi completa de la cohort de COAD excloent les mostres CMS1.

En primer lloc s'ha realitzat l'anàlisi separant les mostres en funció de ploïdia. S'han eliminat 57 mostres classificades com a CMS1 (un 20% del total), quedant 185 mostres del grup 2N i 96 mostres del grup 4N. En el gràfic MDS (figura 30) s'observa que les mostres dels dos grups queden repartides homogèniament, indicant que l'agrupació que es donava en l'anàlisi incloent tots els subtipus podria ser deguda a la presència de les mostres CMS1 en el grup 2N.

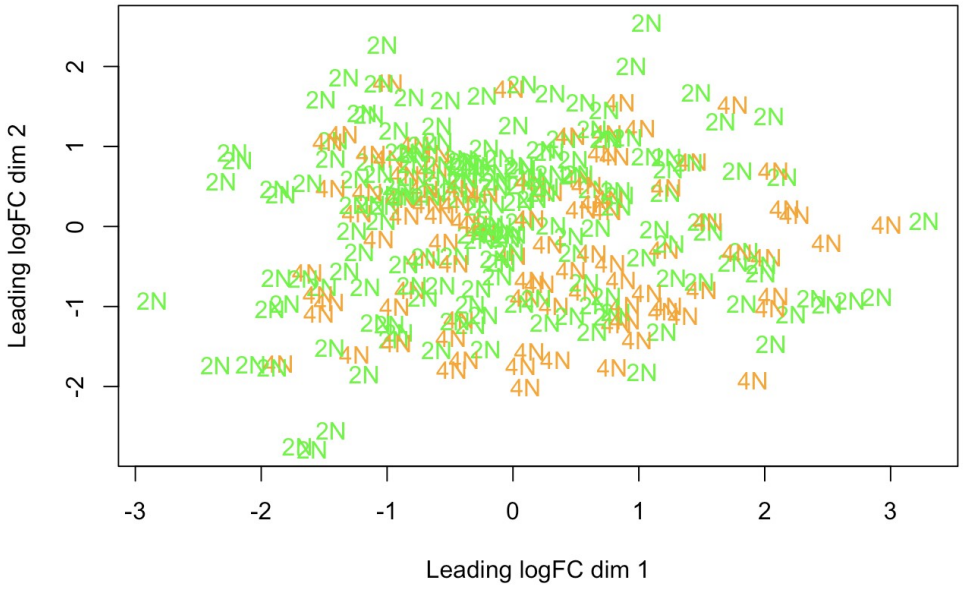

Figura 30: Gràfic d'escalat multidimensional de la cohort de COAD (grups separats per ploidia) excloent les mostres CMS1.

D'acord amb això i amb la reducció de la mida de la mostra, els resultats de l'anàlisi d'expressió diferencial mostren molts menys gens diferencialment expressats. Tot i així, els resultats de l'anàlisi amb GSEA continuen mostrant clarament un enriquiment de la immunitat en el grup 2N i un enriquiment de metabolisme de polisacàrids en el grup 4N (poc significatiu, igual que en l'anàlisi inicial). Els anàlisis d'immunitat també continuen mostrant un enriquiment de la resposta immunitària en el grup 2N però molt menys accentuada. No hi ha diferències signficatives pel que fa a la infiltració immunitària; els gens implicats en el processament d'antigens tampoc mostren diferències significatives i només 4 (PD1, LAG3, TIGIT i PD-L1) dels 8 gens associats a la immunosupressió estan significativament sobreexpressats al grup 4N. D'entre les cèl·lules efectores només els limfòcits T *helper* activats i els T citotòxics de memòria estan subexpressats en el grup 4N i no hi ha diferències en els nivells de les cèl·lules supressores. Només dues (Cxcl11 i Cxcl13) de les quimiocines estan subexpressades al grup 4N. Aquesta baixada general en la significació es podria explicar en gran part per l'exclusió de les mostres CMS1, caracteritzades per nivells alts de cèl·lules immunitàries, però també perquè indirectament implica la reducció del nombre de mostres, fent disminuïr la potència estadística. Els annexos 9 i 10 contenen els informes complets d'aquestes anàlisis.

Paral·lelament s'ha realitzat l'anàlisi també separant les mostres en funció de MCN. En aquest cas s'han eliminat 25 mostres (20%), quedant 85 mostres sense WGD i 47 mostres amb WGD. També s'obtenen molts menys gens significativament expressats però l'anàlisi d'enriquiment continua mostrant un enriquiment significatiu de la immunitat en el grup 2N, tot i que en aquest cas més feble. Potser gràcies a això apareixen també, dins dels primers 20 *gene sets* enriquits, termes relacionats amb respiració cel·lular i metabolisme bioenergètic, tal com es veia a la cohort de READ. Els anàlisis d'immunitat gairebé no mostren diferències significatives entre els dos grups; igual que abans, això podria ser degut a l'exclusió de mostres CMS1, però també s'ha de tenir en compte que la mida de la mostra en aquest anàlisi és molt més petita. Els annexos 11 i 12 contenen els informes complets d'aquestes anàlisis.

3.5. Anàlisi de la cohort de LUAD

Finalment, per validar o contrastar els resultats obtinguts en COAD i READ, s'ha repetit l'anàlisi amb la cohort de LUAD ("Lung Adenocarcinoma"), una de les més estudiades i amb un nombre de mostres disponibles més alt. El càncer de pulmó és la principal causa de mort per càncer arreu del món. El tipus més comú és el *Non-Small-Cell Lung Carcinoma* (NSCLC), que comprèn principalment dos subtipus: *Lung Adenocarcinoma* (LUAD) i *Lung Squamous Cell Carcinoma* (LUSC). La taxa de WGD en NSCLC és relativament alta i recentment s'ha estudiat la seva relació amb l'acumulació d'alteracions deletèries (López et al. 2020). Per aquest motiu i pel fet que disposem d'informació transcriptòmica i de l'estat WGD d'un alt nombre de mostres d'aquest tipus de càncer, hem repetit l'anàlisi complet en la cohort de LUAD de TCGA.

L'anàlisi separant els grups en funció de la ploïdia inclou 337 mostres: 202 mostres 2N i 135 mostres 4N. En el gràfic MDS les mostres dels dos grups queden repartides homogèniament. En l'anàlisi d'expressió diferencial obtenim 2321 gens subexpressats i 1551 gens sobreexpressats (FDR < 0,05). El heatmap dels gens més significatius mostra que la majoria de gens estan downregulats en el grup 4N. L'anàlisi amb GSEA mostra en general moltes vies diferencialment expressades (570 al grup 4N i 1589 al grup 2N) (veure figura 31).

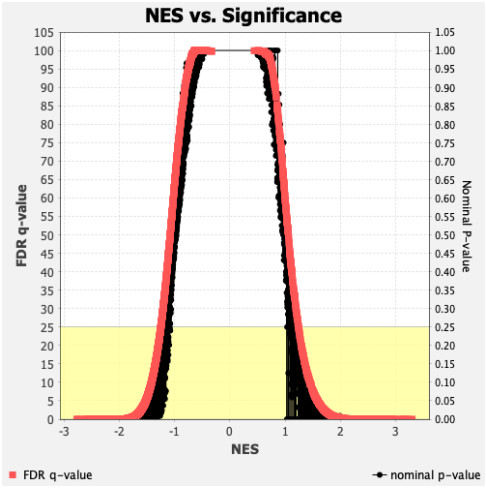

Figura 31: Gràfic p-valor-NES de l'anàlisi d'enriquiment de la cohort de LUAD (separació de grups per ploidia).

Això pot ser degut, en part, a que hi ha moltes vies enriquides que representen el mateix procés biològic. D'entre les vies enriquides en el grup 4N destaquen vies de replicació de l'ADN i de

segregació cromosòmica, la qual cosa no es veia en l'anàlisi de COAD. A la figura 32 es mostren els gràfics d'enriquiment d'alguns dels termes més enriquits.

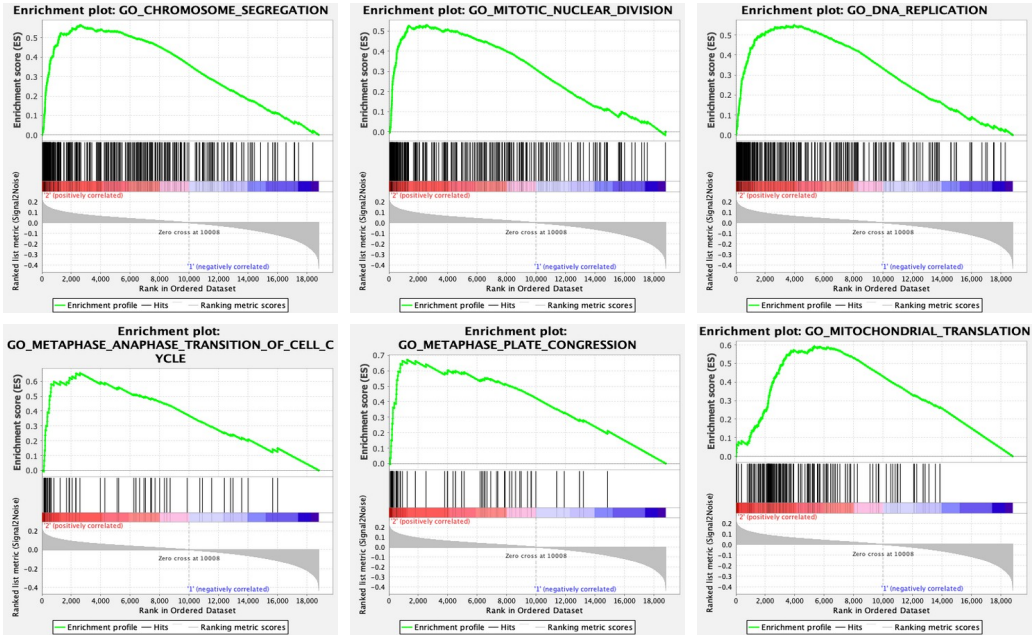

Figura 32: Gràfics d'enriquiment d'algunes de les vies enriquides al grup 4N de la cohort de LUAD.

En aquest cas no apareixen dins els 20 *gene sets* més enriquits termes relacionats amb metabolisme tal com es veia en COAD, tot i que no es pot descartar aquest enriquiment ja que hi ha molts *gene sets* significatius situats per sota dels 20 primers. En canvi, les vies més enriquides en el grup 2N corresponen, igual que en COAD, a vies del sistema immunitari,

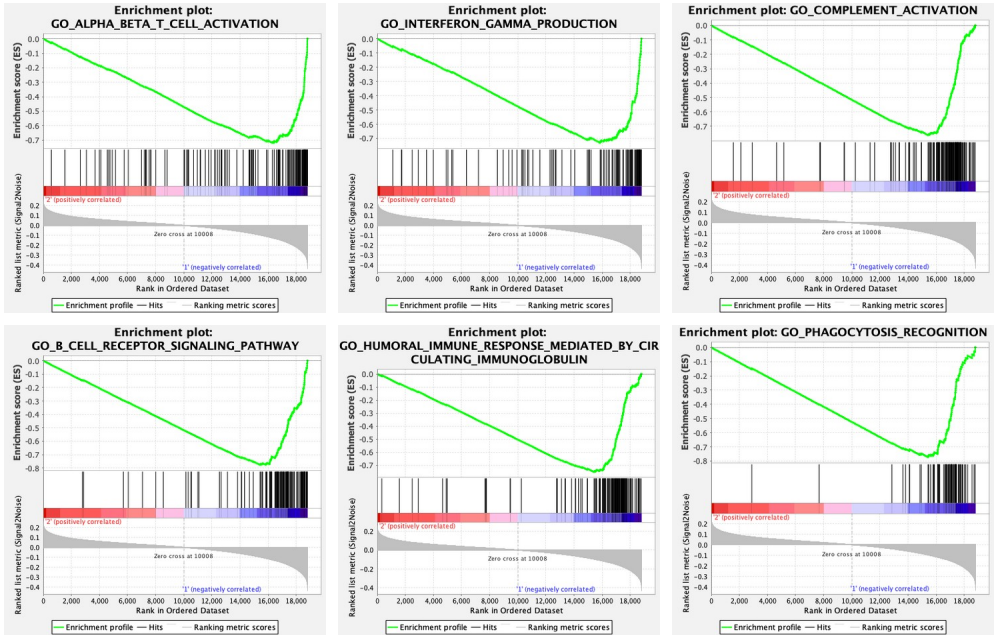

Figura 33: Gràfics d'enriquiment d'algunes de les vies immunitàries enriquides al grup 2N de la cohort de LUAD.

lligades a l'acció de cèl·lules T però també a l'activació del complement, resposta immunitària humoral, cèl·lules B i fagocitosi (figura 33).

Les anàlisis d'immunitat ho confirmen mostrant una resposta antitumoral enriquida en el grup 2N, generalment més significativa que en COAD. Excepte els nivells de limfòcits T activats CD4+, la resta de determinants immunitaris testats són significatius. Els annexos 13 i 14 contenen els informes complets d'aquestes anàlisis.

L'anàlisi separant els grups en funció de MCN inclou 388 mostres: 162 mostres sense WGD i 226 mostres amb WGD. En aquest cas, contràriament a l'anàlisi anterior, el gràfic MDS mostra agrupament en funció del grup, fent pensar que hi podria haver algun factor de confusió. S'obtenen 4122 gens subexpressats i 4108 gens sobreexpressats, un nombre molt elevat. Els resultats de GSEA conicideixen amb els de l'anàlisi anterior (veure figura 29b), amb enriquiment de vies relacionades amb divisió cel·lular en el grup amb WGD i enriquiment de termes immunitaris en el grup sense WGD, en aquest cas amb la presència de vies relacionades amb la producció de citocines. Els anàlisis d'immunitat mostren en general un enriquiment de la resposta antitumoral en el grup sense WGD, excepte els nivells de limfòcits T activats CD4+ i les quimiocines CCL8 i CXCL10, que estan enriquits en el grup amb WGD. Els annexos 15 i 16 contenen els informes complets d'aquestes anàlisis.

3.6. Correlació entre ploidia i estat WGD determinat per MCN

Tot i que els resultats utilitzant els dos tipus d'aproximacions per determinar l'estat WGD són similars, és interessant veure la correlació entre els nivells de ploïdia i l'estat WGD determinat per MCN per explicar les petites diferències entre un anàlisi i l'altre.

Per veure aquesta correlació s'ha analitzat el subconjunt de mostres de COAD i LUAD que disposen dels dos tipus d'informació. Cal destacar que en les dues cohorts moltes de les mostres utilitzades en l'anàlisi separant per ploïdia no disposen d'informació de l'estat WGD, però en canvi hi ha informació de ploïdia per gairebé totes les mostres incloses a l'anàlisi separant per estat WGD inferit per MCN.

Mitjançant diagrames de caixa es pot veure la distribució de ploïdia per cada un dels grups (figura 34a). En els dos casos la mediana de ploïdia del grup sense WGD està al voltant de 2, mentre que la mediana de ploïdia del grup amb WGD està al voltant de 3 (3,03 en COAD i 3,19 en LUAD). Per veure-ho més clarament s'han representat aquestes mostres en un gràfic de punts ordenant les mostres en funció de la ploïdia i diferenciant els dos estats amb colors (en verd sense WGD i en taronja amb WGD) (figura 34b).

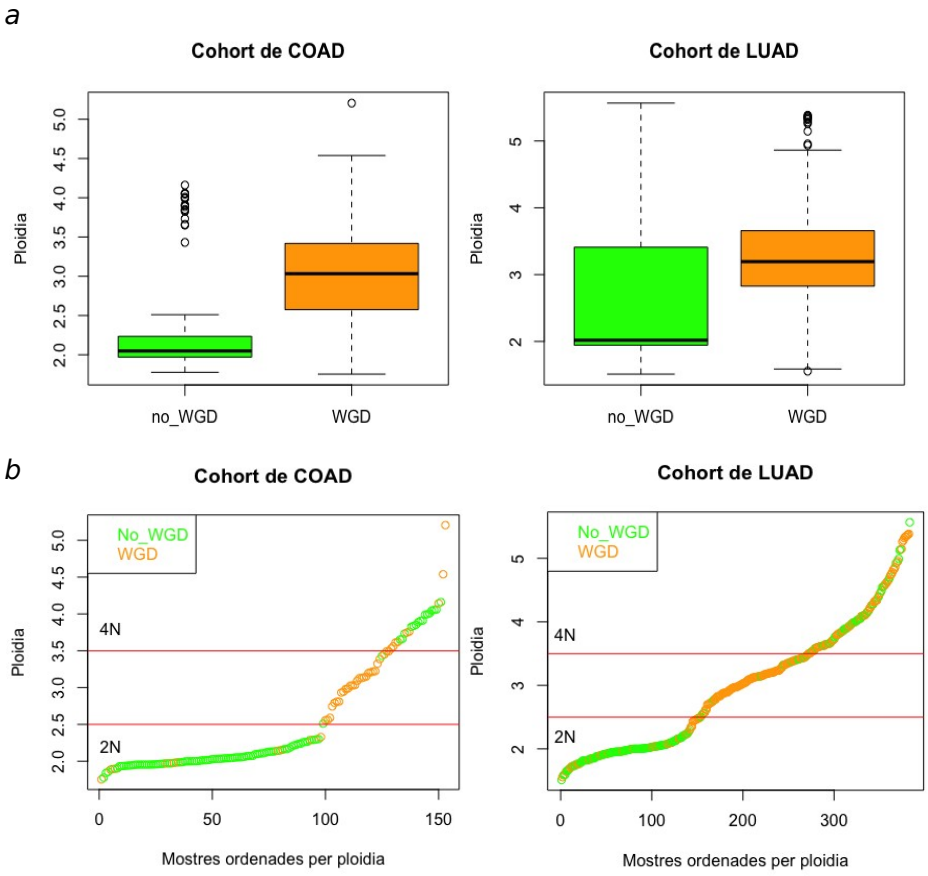

Figura 34: Correlació entre ploidia i estat WGD. S'han marcat amb línies vermelles els llindars de ploidia utilitzats per establir els grups 2N i 4N.

Tant en la cohort de COAD com en la de LUAD s'observa que la majoria de les mostres classificades com a 2N efectivament no han patit WGD. En canvi, el grup de mostres classificades com a 4N (amb ploïdia superior a 3.5) inclou mostres dels dos grups. S'observa, a més, que les mostres amb WGD es concentren a la regió de ploïdies intermèdies, no incloses dins de l'anàlisi separant per ploïdia.

Cal tenir en compte aquestes diferències a l'hora de comparar les anàlisis fetes amb un tipus o altre d'aproximació.

4. Discussió

S'han analitzat les diferències transcripcionals entre tumors que han patit WGD i tumors que no l'han patit. Tot i haver-hi certes diferències en la classificació de les mostres en funció de l'aproximació utilitzada per determinar l'estat WGD, s'obtenen resultats similars. El fet que les anàlisis en les que s'utilitza el MCN donin resultats generalment menys significatius es pot

explicar per la reducció de la mida de la mostra (veure taules suplementàries 1 i 2) i en una separació dels grups no tan clara.

En conjunt, tant en mostres de càncer colorectal com en mostres d'adenocarcinoma de pulmó s'observa una correlació negativa entre WGD i resposta immunitària antitumoral, fet que indica que aquesta associació podria tenir un caràcter més o menys universal. De fet, en un estudi previ similar, es va demostrar una correlació positiva entre nivells alts d'aneuploïdia, associats a WGD, i marcadors d'evasió immunitària en diversos tipus de tumors (Davoli et al. 2017).

S'ha demostrat que l'exclusió de mostres CMS1, caracteritzades per nivells d'infiltració immunitària alts, no fa desaparèixer del tot les diferències immunitàries. En concret, en l'anàlisi en què se separen els grups per ploïdia, continua havent-hi un augment d'expressió en el grup 4N d'algunes molècules immunosupressores com PD1 o PD-L1, la qual cosa podria explicar la relació inversa entre aneuploïdia i infiltració immunitària. El fet que les diferències immunitàries siguin menys accentuades en aquest grup, tal com passa en la cohort de recte, també es pot explicar, un cop més, per la reducció de la mida de la mostra. La mateixa idea serveix pels resultats observats en l'anàlisi de les mostres d'adenocarcinoma de pulmó, amb una mida mostral més gran i amb diferències en la infiltració immunitària més accentuades.

Comparant els resultats pel que fa la immunitat entre els dos tipus de càncer, s'observen algunes diferències. Mentre que l'anàlisi amb GSEA de la cohort de còlon mostra sobretot infraregulació de termes associats a l'acció de limfòcits T, en la cohort de pulmó apareixen també termes lligats a limfòcits B, dels quals no se'n coneix tan bé el paper en la progressió tumoral; tot i així s'estan trobant evidències que lliguen la seva presència amb la immunosupressió i la progressió tumoral[9]. Una altra diferència destacable es troba en els gens implicats en el processament d'antígens, molts d'ells diferencialment expressats en pulmó però no en còlon. La relació entre la baixa expressió de molècules HLA (*Human Leukocyte Antigen*), que formen part del complex major d'histocompatibilitat, i un pronòstic dolent s'ha descrit en diversos tipus de càncer. Concretament, s'ha identificat la pèrdua d'HLA en un 40% de tumors de NSCLS en estadis inicials, que se selecciona positivament al llarg de l'evolució tumoral i es creu que és un dels mecanismes existents d'evasió immunitària[21]. També ha estat descrita en diferents tipus de càncer (inclòs el càncer de pulmó i el de cólon) la pèrdua d'heterozigositat (LOH) de B2M, un factor essencial per l'assemblatge del complex HLA[21,22,23]. Aquesta alteració també es relaciona amb l'escapatòria de la resposta antitumoral i s'ha associat amb la resistència a immunoteràpia (bloqueig dels checkpoints). Tant B2M com les diferents molècules HLA estan subexpressades en els tumors amb WGD de la cohort de pulmó. Una altra diferència que s'observa entre les dues cohorts és l'alteració de la quantitat de limfòcits T helper (CD4+); aquesta és inferior en tumors amb WGD de còlon, seguint la mateixa tendència que la resta de cèl·lules efectores, però és superior o no hi ha diferències en el cas dels tumors amb WGD de pulmó. El paper dels limfòcits T *helper* i el complex major d'histocompatibilitat de classe II no es coneix tan bé,

però recentment s'ha demostrat que l'activitat dels limfòcits T, tant CD8+ com CD4+, és necessària per una resposta immunitària òptima[24].

Un altre aspecte a tenir en compte sobre els resultats de les anàlisis d'immunitat són les diferències que s'observen en funció de si se separen les mostres per ploïdia o per MCN, especialment en l'expressió de la signatura de quimiocines. Mentre que en la cohort de cólon separant els grups per ploïdia 9 de les 12 quimiocines estan subexpressades en el grup 4N, en l'anàlisi en què se separen els grups per MCN només 2 quimiocines estan significativament subexpressades. Tot i que això podria ser degut a la reducció en el nombre de mostres que implica la reducció en la potència estadística, aquesta tendència també es dona en pulmó, on el nombre de mostres de les dues anàlisis és semblant. En aquest cas s'observa subexpressió de totes les quimiocines quan s'utilitza la separació per ploïdia i, en canvi, quan s'utilitza la separació per MCN només una quimiocina (CCL19) està significativament subexpressada. Una possible explicació d'aquestes diferències seria que la mediana de la ploïdia del grup amb WGD inferit per MCN és aproximadament de 3, mentre que en l'anàlisi en el qual s'usa la informació de ploïdia, s'han considerat WGD aquelles mostres amb ploïdies superiors a 3,5, descartant totes les mostres intermèdies. Així doncs, s'ha de tenir en compte que aquests resultats podrien estar lligats a nivells de ploïdia alts i potser no a haver patit WGD.

A part dels canvis immunitaris s'han observat també diferències metabòliques entre els tumors amb i sense WGD de càncer colorectal; donat que aquestes no s'observen en adenocarcinoma de pulmó, podrien ser específiques del càncer colorectal. Tot i així, caldria explorar-ho més específicament ja que en l'anàlisi GSEA de càncer de pulmó s'identifiquen una gran quantitat de *gene sets*, la qual cosa podria amagar diferències també en metabolisme. En un anàlisi preliminar de les vies metabòliques en la cohort de còlon s'ha trobat una disminució en l'expressió de gens implicats en la via de fosforilació oxidativa. Termes relacionats amb aquesta via també apareixen enriquits al grup sense WGD tant en la cohort de recte com en la cohort de còlon quan s'exclouen els subtipus CMS1. El fet que no apareguin en l'anàlisi de còlon amb tots els subtipus podria indicar que estan relacionats amb tumors MSS (*Microsatellite stable*) o podria ser simplement una qüestió tècnica donada pel fet que les diferències immunitàries n'emmascaren d'altres de menys significatives. En un article recent[25] es parla de la connexió entre dues de les principals capacitats que adquireixen les cèl·lules tumorals: la inestabilitat genòmica i l'alteració del metabolisme. Han descobert que l'acumulació de diversos metabòlits (2-hidroxi-glutarat, fumarat i succinat) en cèl·lules tumorals inhibeix la reparació de l'ADN i causa inestabilitat genòmica. L'enzim succinat deshidrogenasa (SDH), que provoca l'acumulació de succinat quan no és funcional, és un dels gens de la fosforilació oxidativa que s'ha trobat subexpressat en l'anàlisi metabòlica realitzada sobre la cohort de còlon.

De la mateixa manera que els canvis metabòlics només s'han observat en càncer colorectal, l'enriquiment en vies de segregació cromosòmica i replicació s'ha vist només en mostres amb WGD de càncer de pulmó, la qual cosa indica que podria estar lligat a aquest tipus de teixit.

Tot i així, aquestes diferències també s'han trobat en un estudi d'expressió gènica d'un model tetraploide de cèl·lules DLD-1, una línia cel·lular d'adenocarcinoma de cólon[26].En aquest estudi s'ha observat un enriquiment en la resposta a l'estrès replicatiu en les cèl·lules 4N, les quals, a més, presenten més capacitat de migració i d'invasió.

Altres diferències trobades en còlon i recte tenen a veure amb l'enriquiment, al grup amb WGD, de vies relacionades amb el desenvolupament i la morfogènesi, la qual cosa podria anar lligada al remodelat tissular, la plasticitat i la invasió, processos característics de tumors més agressius. Tot i això, s'hauria d'explorar més detalladament amb bases de dades més específiques.

Per poder validar tots aquests resultats s'haurien de repetir les anàlisis en altres cohorts de càncer colorectal i de càncer de pulmó i estendre-les als altres tipus de càncer per identificar alteracions específiques d'un tipus cel·lular o més generalitzades. Per altra banda, en relació als resultats associats a canvis en el metabolisme, s'està fent una anàlisi més específica per tal de determinar les vies metabòliques alterades.

5. Conclusions

Pel que fa a la realització del treball puc extreure algunes conclusions.

En primer lloc, la meva valoració sobre la planificació inicial és bona ja que generalment s'ha seguit el que s'havia establert a l'inici del treball, modificant lleugerament la temporització de les tasques i afegint-ne algunes que no s'havien previst per tal de completar els resultats obtinguts. Pel que fa al compliment dels objectius establerts inicialment també ho valoro satisfactòriament, ja que s'ha realitzat l'anàlisi completa de la cohort principal i s'ha estès a una cohort addicional. L'estudi *pan-cancer* planificat en un principi es queda pendent com a línia de treball futur.

En segon lloc, la metodologia usada m'ha permès consolidar els coneixements adquirits durant el màster i aprendre'n de nous. A més, el treball amb una quantitat gran de dades públiques m'ha permès trobar-me amb els problemes propis d'aquest tipus d'anàlisi i aprendre a solucionar-los.

Finalment, la valoració sobre els resultats obtinguts també és bona. La correlació negativa trobada entre la immunitat anti-tumoral i el WGD sembla consistent i dóna suport a altres estudis similars, mentre que les diferències metabòliques identificades entre tumors amb i sense WGD obren noves línies per explorar.

6. Glossari

Alteracions del nombre de còpia (CNA): guanys o pèrdues de fragments genòmics de més d'un 1 kb.

Complex major d'histocompatibilitat (MHC): locus polimòrfic que codifica per proteïnes de superfície cel·lular implicades en la presentació d'antigens al sistema immunitari.

Duplicació completa del genoma (WGD): procés mitjançant el qual es duplica el nombre de cromosomes d'una cèl·lula.

***False Discovery Rate* (FDR)**: mètode d'ajust del p-valor utilitzat quan es realitzen múltiples tests simultanis i que representa la proporció esperada de falsos positius.

Gene set: Conjunt de gens que representa un procés biològic, una via de senyalització o altres.

Immunomoduladors: molècules implicades en la regulació de la resposta immunitària.

***Log-fold-change* (Log-FC)**: Logaritme en base 2 del *fold-change*.

Nombre major de còpies (MCN): nombre de còpies de l'al·lel més freqüent.

***Normalized Enrichment Score* (NES)**: Puntuació d'enriquiment normalitzada pel nombre de gens que componen el *gene set*.

Ploïdia: nombre complet de conjunts cromosòmics en una cèl·lula.

Recomptes per milió (CPM): Nombre de recomptes d'un trànscrit dividit per la mida de la llibreria i multiplicat per un milió.

Subtipus moleculars consens (CMS): sistema de classificació dels tumors colorectals en funció de l'expressió gènica.

7. Bibliografia

1. Kuznetsova, Anastasia Y, Katarzyna Seget, Giuliana K Moeller, Mirjam S de Pagter, Jeroen ADM de Roos, Milena Dürrbaum, Christian Kuffer, et al. 2015. "Chromosomal Instability, Tolerance of Mitotic Errors and Multidrug Resistance Are Promoted by Tetraploidization in Human Cells." *Cell Cycle* 14 (17). Taylor & Francis: 2810–20.

2. Ben-David, Uri, and Angelika Amon. 2019. "Context Is Everything: Aneuploidy in Cancer." *Nature Reviews Genetics*. Nature Publishing Group, 1–19.

3. Bielski, Craig M., Ahmet Zehir, Alexander V. Penson, Mark T. A. Donoghue, Walid Chatila, Joshua Armenia, Matthew T. Chang, et al. 2018. “Genome Doubling Shapes the Evolution and Prognosis of Advanced Cancers.” *Nature Genetics* 50 (8): 1189–95.

4. Dewhurst, SM. 2015. “Whole Genome Doubling Propagates Chromosomal Instability and Accelerates Cancer Genome Evolution.” PhD thesis, UCL (University College London).

5. López, Saioa, Emilia L. Lim, Stuart Horswell, Kerstin Haase, Ariana Huebner, Michelle Dietzen, Thanos P. Mourikis, et al. 2020. “Interplay Between Whole-Genome Doubling and the Accumulation of Deleterious Alterations in Cancer Evolution.” *Nature Genetics* 52 (3): 283–93. https://doi.org/10.1038/s41588-020-0584-7.

6. Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, et al. 2020. “The Evolutionary History of 2,658 Cancers.” *Nature* 578 (7793). Nature Publishing Group: 122–28.

7. Davoli, Teresa, Hajime Uno, Eric C Wooten, and Stephen J Elledge. 2017. “Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy.” *Science* 355 (6322). American Association for the Advancement of Science: eaaf8399.

8. Taylor, Alison M, Juliann Shih, Gavin Ha, Galen F Gao, Xiaoyang Zhang, Ashton C Berger, Steven E Schumacher, et al. 2018. “Genomic and Functional Approaches to Understanding Cancer Aneuploidy.” *Cancer Cell* 33 (4). Elsevier: 676–89.

9. Gonzalez, Hugo, Catharina Hagerling, and Zena Werb. 2018. “Roles of the Immune System in Cancer: From Tumor Initiation to Metastatic Progression.” *Genes & Development* 32 (19-20). Cold Spring Harbor Lab: 1267–84.

10. Van Loo, Peter, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, et al. 2010. “Allele-Specific Copy Number Analysis of Tumors.” *Proceedings of the National Academy of Sciences* 107 (39). National Acad Sciences: 16910–5.

11. Law, Charity W, Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K Smyth, and Matthew E Ritchie. 2016. “RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and edgeR.” *F1000Research* 5. Faculty of 1000 Ltd.

12. Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences* 102 (43). National Acad Sciences: 15545–50.

13. Guinney, Justin, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, et al. 2015. “The Consensus Molecular Subtypes of Colorectal Cancer.” *Nature Medicine* 21 (11). Nature Publishing Group: 1350–6.

14. Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martı'nez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, et al. 2013. “Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data.” *Nature Communications* 4 (1). Nature Publishing Group: 1–11.

15. Charoentong, Pornpimol, Francesca Finotello, Mihaela Angelova, Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, and Zlatko Trajanoski. 2017. “Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.” *Cell Reports* 18 (1). Elsevier: 248–62.

16. Sautès-Fridman, Catherine, Florent Petitprez, Julien Calderaro, and Wolf Herman Fridman. 2019. “Tertiary Lymphoid Structures in the Era of Cancer Immunotherapy.” *Nature Reviews Cancer* 19 (6). Nature Publishing Group: 307–25.

17. González, Ignacio. 2014. *Statistical Analysis of Rna-Seq Data*. INRA Toulouse / IMT Université Toulouse.

18. Robinson, Mark D, and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of Rna-Seq Data.” *Genome Biology* 11 (3). BioMed Central: R25.

19. Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for Rna-Seq Read Counts.” *Genome Biology* 15 (2). Springer: R29.

20. Brunk, Elizabeth, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, et al. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3). Nature Publishing Group: 272.

21. McGranahan, Nicholas, Rachel Rosenthal, Crispin T Hiley, Andrew J Rowan, Thomas BK Watkins, Gareth A Wilson, Nicolai J Birkbak, et al. 2017. “Allele-Specific Hla Loss and Immune Escape in Lung Cancer Evolution.” *Cell* 171 (6). Elsevier: 1259–71.

22. Sade-Feldman, Moshe, Yunxin J Jiao, Jonathan H Chen, Michael S Rooney, Michal Barzily-Rokni, Jean-Pierre Eliane, Stacey L Bjorgaard, et al. 2017. “Resistance to Checkpoint Blockade Therapy Through Inactivation of Antigen Presentation.” *Nature Communications* 8 (1). Nature Publishing Group: 1–11.

23. Grasso, Catherine S, Marios Giannakis, Daniel K Wells, Tsuyoshi Hamada, Xinmeng Jasmine Mu, Michael Quist, Jonathan A Nowak, et al. 2018. “Genetic Mechanisms of Immune Evasion in Colorectal Cancer.” *Cancer Discovery* 8 (6). AACR: 730–49.

24. Alspach, Elise, Danielle M Lussier, Alexander P Miceli, Ilya Kizhvatov, Michel DuPage, Adrienne M Luoma, Wei Meng, et al. 2019. “MHC-II Neoantigens Shape Tumour Immunity and Response to Immunotherapy.” *Nature* 574 (7780). Nature Publishing Group: 696–701.

25. Chen, Lei-Lei, Xiong, Yue. 2020. “Tumour metabolites hinder DNA repair”. Nature **582**, 492-494.

26. Wangsa, Darawalee, Isabel Quintanilla, Keyvan Torabi, Maria Vila-Casadesús, Amaia Ercilla, Gregory Klus, Zeynep Yuce, et al. 2018. “Near-Tetraploid Cancer Cells Show Chromosome Instability Triggered by Replication Stress and Exhibit Enhanced Invasiveness.” *The FASEB Journal* 32 (7). Wiley Online Library: 3502–17.

27. Campbell, Joshua D, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, et al. 2016. “Distinct Patterns of Somatic Genome Alterations in Lung Adenocarcinomas and Squamous Cell Carcinomas.” *Nature Genetics* 48 (6). Nature Publishing Group: 607.

28. Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. 2017. “RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool.” *PloS One* 12 (12). Public Library of Science.

29. Fontana, E, K Eason, A Cervantes, R Salazar, and A Sadanandam. 2019. “Context Matters —Consensus Molecular Subtypes of Colorectal Cancer as Biomarkers for Clinical Trials.” *Annals of Oncology* 30 (4). Oxford University Press: 520–27.

30. Ho, Ping-Chih, and Pu-Ste Liu. 2016. “Metabolic Communication in Tumors: A New Layer of Immunoregulation for Immune Evasion.” *Journal for Immunotherapy of Cancer* 4 (1). BioMed Central: 4.

31. Jamal-Hanjani, Mariam, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, et al. 2017. “Tracking the Evolution of Non–Small-Cell Lung Cancer.” *New England Journal of Medicine* 376 (22). Massachusetts Medical Society: 2109–21. https://doi.org/10.1056/nejmoa1616288.

32. La Vecchia, Sofia, and Carlos Sebastián. 2020. “Metabolic Pathways Regulating Colorectal Cancer Initiation and Progression.” In *Seminars in Cell & Developmental Biology*, 98:63–70. Elsevier.

33. Pavlova, Natalya N, and Craig B Thompson. 2016. “The Emerging Hallmarks of Cancer Metabolism.” *Cell Metabolism* 23 (1). Elsevier: 27–47.

34. Reimand, Jüri, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, et al. 2019. “Pathway Enrichment Analysis and Visualization of Omics Data Using G: Profiler, Gsea, Cytoscape and Enrichmentmap.” *Nature Protocols* 14 (2). Nature Publishing Group: 482–517.

35. Seton-Rogers, Sarah. 2020. “Weighing up Effects of Extra Chromosomes.” *Nature Reviews Cancer*. Nature Publishing Group, 1–1.

36. Shukla, Ankit, Thu HM Nguyen, Sarat B Moka, Jonathan J Ellis, John P Grady, Harald Oey, Alexandre S Cristino, et al. 2020. “Chromosome Arm Aneuploidies Shape Tumour Evolution and Drug Response.” *Nature Communications* 11 (1). Nature Publishing Group: 1–14.

37. Tamborero, David, Carlota Rubio-Perez, Ferran Muiños, Radhakrishnan Sabarinathan, Josep M Piulats, Aura Muntasell, Rodrigo Dienstmann, Nuria Lopez-Bigas, and Abel Gonzalez-Perez. 2018. “A Pan-Cancer Landscape of Interactions Between Solid Tumors and Infiltrating Immune Cell Populations.” *Clinical Cancer Research* 24 (15). AACR: 3717–28.

38. Van de Peer, Yves, Eshchar Mizrachi, and Kathleen Marchal. 2017. “The Evolutionary Significance of Polyploidy.” *Nature Reviews Genetics* 18 (7). Nature Publishing Group: 411.

8. Taules suplementàries

Taula 1: Nombre de mostres incloses en cada un dels grups d'anàlisi.

	no WGD	WGD	Total
COAD ploidia	239	99	338
COAD MCN	107	50	157
COAD no CMS1 ploidia	185	96	281
COAD no CMS1 MCN	85	47	132
READ ploidia	71	39	110
READ MCN	30	19	49
LUAD ploidia	202	135	337
LUAD MCN	162	226	388

Taula 2: Nombre de gens diferencialment expressats (FDR < 0.05) en cada una de les anàlisis.

	Gens subexpressats	Gens sobreexpressats
COAD ploidia	1765	1986
COAD MCN	417	460
COAD no CMS1 ploidia	260	522
COAD no CMS1 MCN	28	35
READ ploidia	0	0
READ MCN	0	0
LUAD ploidia	2321	1551
LUAD MCN	4122	4108

Taula 3: Nombre de gene sets enriquits (FDR < 0.25) en les anàlisis GSEA amb permutació de gene sets.

	no WGD	WGD
COAD ploidia	786	0
COAD MCN	545	518
COAD no CMS1 ploidia	624	0
COAD no CMS1 MCN	144	711
READ ploidia	354	1
READ MCN	1	5
LUAD ploidia	1589	570
LUAD MCN	1369	574

Taula 4: Resum dels 20 primers processos biològics enriquits en l'anàlisi GSEA amb permutació de gene sets.

	Immunitat	Metabolisme	Desenvolupament	Funcions neuronals	Divisió cel·lular
COAD ploidia	no WGD	WGD	WGD	WGD	-
COAD MCN	no WGD	WGD	WGD	WGD	-
COAD no CMS1 ploidia	no WGD	WGD	-	WGD	-
COAD no CMS1 MCN	no WGD	WGD / no WGD	WGD	-	-
READ ploidia	no WGD	WGD / no WGD	WGD	-	-
READ MCN	no WGD	WGD	WGD	WGD	-
LUAD ploidia	no WGD	-	-	-	WGD
LUAD MCN	no WGD	-	-	-	WGD

9. Annexos

A continuació es descriuen els fitxers que s'adjunten com a annexos:

Annex 1: Diagrama de Gantt inicial

Annex 2: Diagrama de Gantt final

Annex 3: Informe de resultats de la cohort de COAD utilitzant informació de ploidia.

Annex 4: Informe de resultats de la cohort de COAD utilitzant informació de MCN.

Annex 5: Anàlisis d'immunitat de la cohort de COAD utilitzant informaicó de MCN.

Annex 6: Informe de resultats de la cohort de READ utilitzant informació de ploidia.

Annex 7: Informe de resultats de la cohort de READ utilitzant informació de MCN.

Annex 8: Resultats de la classificació CMS.

Annex 9: Informe de resultats de la cohort de COAD excloent mostres CMS1 i utilitzant informació de ploidia.

Annex 10: Anàlsis d'immunitat de la cohort de COAD excloent mostres CMS1 i utilitzant informació de ploidia.

Annex 11: Informe de resultats de la cohort de COAD excloent mostres CMS1 i utilitzant informació de MCN.

Annex 12: Anàlsis d'immunitat de la cohort de COAD excloent mostres CMS1 i utilitzant informació de MCN.

Annex 13: Informe de resultats de la cohort de LUAD utilitzant informació de ploidia.

Annex 14: Anàlsis d'immunitat de la cohort de LUAD utilitzant informació de ploidia.

Annex 15: Informe de resultats de la cohort de LUAD utilitzant informació de MCN.

Annex 16: Anàlsis d'immunitat de la cohort de LUAD utilitzant informació de MCN.

Annex 17: Codi R utilitzat per a l'anàlisi d'expressió diferencial i enriquiment.

Annex 18: Codi R utilitzat per els anàlisis d'immunitat.

Annex 19: Codi R utilitzat per a la classificació CMS.

Annex 20: Codi R utilitzat per a la correlació entre ploidia i estat WGD determinat per MCN.