Anàlisi de l'expressió gènica de pacients amb Limfoma Plasmablàstic mitjançant dades d'RNA-seq.

Aleix Méndez López

Màster en Bioinformàtica i Bioestadística

Consultor:
Guillem Ylla Bou

Professor Responsable de l'assignatura:
Antoni Pérez Navarro Contactar

Tutors externs:
José Tomás Navarro Ferrando i Manuel Castro de Moura

8 de juny del 2021

FITXA DEL TREBALL FINAL

Títol del treball:	Anàlisi de l'expressió gènica de pacients amb Limfoma Plasmablàstic mitjançant dades d'RNA-seq.
Nom de l'autor:	*Aleix Méndez López*
Nom del consultor/a:	*Guillem Ylla Bou*
Nom del PRA:	*Antoni Pérez Navarro Contactar*
Data de lliurament (mm/aaaa):	6/2021
Titulació:	*Màster en Bioinformàtica i Bioestadística*
Àrea del Treball Final:	*TFM-Bioinformàtica i Bioestadística Àrea 4 aula 1*
Idioma del treball:	Català
Nombre de crèdits:	15
Paraules clau:	*RNA-seq, miRNA, Anàlisi d'expressió diferencial.*

Resum del Treball (màxim 250 paraules):

El limfoma plasmablàstic (LPB) és un subtipus de Limfoma B Difús de Cèl·lula Gran (LBDCG) que s'associa principalment a la infecció pel virus de la immunodeficiència humana (VIH), a altres tipus d'immunodepressió (com els receptors de trasplantament d'òrgans sòlids) i fins i tot en pacients immunocompetents. Tot i que és una malaltia rara en general, la incidència augmenta dràsticament en regions amb una prevalença elevada d'VIH. El limfoma plasmablàstic és un tumor poc estudiat i extremadament agressiu. L'aportació d'informació per definir l'expressió genòmica d'aquest limfoma és important per a la comprensió de la patogènesi d'aquesta malaltia. El present treball estudia els canvis en l'expresssió que es generen en el transcriptoma, mitjançant l'anàlisi bioinformàtic de les dades obtingudes a partir de tècniques d'RNA-seq realitzades sobre un conjunt de mostres de pacients amb LPB i Controls sense LPB. A més, gràcies a un estudi d'expressió de *microarrays* anterior on es van trobar un conjunt de miRNAs diferencialment expressats entre els mateixos dos grups, es determina quin és el paper dels miRNA diferencialment expressats en la modulació de l'expressió dels gens diana als que s'uneix. Els resultats d'aquest estudi inciten a crear línies de treball futures que permetin continuar estudiant aquest conjunt de dades.

Abstract (in English, 250 words or less):

Plasmablastic lymphoma (PBL) is a subtype of Diffuse B Large Cell Lymphoma (DBLCL) that is primarily associated with human immunodeficiency virus (HIV) infection, other types of immunosuppression (such as solid organ transplantation) and in immunocompetent patients. In general, PBL is a rare disease, but the incidence increases dramatically in regions with a high HIV prevalence. PBL is a poorly studied and extremely aggressive tumor. The provision of information to define the genomic expression of this lymphoma is important for understanding the pathogenesis of this disease. This study identify the transcriptome expression changes, using bioinformatics analysis of data from RNA-seq techniques performed on a set of samples from patients with PBL and Controls without PBL. In addition, using data of a previous *microarray* expression study where a set of differentially expressed miRNAs were found between the same two groups, the role of differentially expressed miRNAs in modulating gene expression is determined. The results of this study suggest the creation of future lines of work that will allow further study of this data set.

ÍNDEX

1. INTRODUCCIÓ

1.1. Descripció general

En el grup de recerca de Neoplàsies Limfoides de l'Institut Josep Carreras (IJC) s'han recopilat un conjunt de mostres de pacients amb limfoma plasmablàstic (LPB) a partir de les quals s'han obtingut dades d'RNA-seq per estudiar el transcriptoma d'aquests pacients, i dades de *microarrays* per estudiar l'expressió dels miRNA del mateix conjunt de pacients. En el present estudi s'analitzen les dades dels RNA-seq i es relacionen amb els resultats de l'anàlisi de les dades dels *microarrays* que es van obtenir en un estudi anterior.

1.2. Context i justificació del Treball

El limfoma plasmablàstic és una entitat clinicopatològica descrita per primera vegada el 1997 i reconeguda actualment com un subtipus de Limfoma B Difús de Cèl·lula Gran (LBDCG). S'associa principalment a la infecció pel virus de la immunodeficiència humana (VIH), en la qual representa fins a un 2% de tots els limfomes, encara que també s'observen casos en pacients amb altres tipus d'immunodepressió (com els receptors de trasplantament d'òrgans sòlids) i fins i tot en pacients immunocompetents. En la patogènia del limfoma plasmablàstic es troben implicats tant la infecció pel virus d'Epstein-Barr com reordenaments del gen MYC (ICO guia.2019). En una revisió del 2015 (Castillo et al. 2015) es van recopilar 590 casos, des de la primera publicació d'aquesta malaltia l'any 1997, dels quals 369 (63%) havien estat diagnosticats en pacients amb infecció per VIH, 164 (28%) en pacients sense infecció per VIH, 37 (6%) en pacients receptors de trasplantament i 20 (3%) com a limfoma plasmablàstic transformat. La presentació habitual és l'afecció extraganglionar, particularment a la cavitat oral en pràcticament la meitat dels casos, seguida del tracte gastrointestinal i la pell. El pronòstic del limfoma plasmablàstic és extremament dolent, amb mitjanes de supervivència de 8-15 mesos. El pitjor pronòstic sembla associar-se al limfoma plasmablàstic post-trasplantament, mentre que els pacients immunocompetents semblen tenir supervivències lleugerament més prolongades. La presència de reordenaments o guanys del gen MYC empitjora el pronòstic. No hi ha un tractament estàndard definit per als pacients amb limfoma plasmablàstic. Atesa la baixa freqüència d'aquest tipus de limfoma, l'evidència respecte a l'eficàcia dels tractaments es coneix d'estudis de casos i de sèries retrospectives. Sense tractament, la mitjana de supervivència és de 3-4 mesos. Actualment, el tractament consisteix en l'administració de quimioteràpia, amb o sense radioteràpia, mentre que la intensificació amb trasplantament autogènic de progenitors hematopoètics (TAPH) en primera línia és motiu de controvèrsia. Entre les futures vies de tractament es podrien trobar nous agents antivírics,

immunoteràpia cel·lular contra el VEB, utilització de CAR-T que tingui com a diana l'antigen CD30 (present en fins al 30% dels limfomes plasmablàstics) o fàrmacs que tinguin com a diana el gen MYC.

El limfoma plasmablàstic és un tumor poc estudiat i extremadament agressiu. L'aportació d'informació per definir l'expressió genòmica d'aquest limfoma és important per a la comprensió de la patogènesi d'aquesta malaltia i pel desenvolupament de tractaments personalitzats. El present estudi té la finalitat d'estudiar els canvis que es generen en el transcriptoma dels pacients amb LPB , estudiant l'expressió dels miRNAs mitjançant tècniques de *microarrays*, i dels RNA missatgers mitjançant tècniques de RNA-seq.

1.3. Objectius del Treball

Objectius generals:

- Descriure les diferències transcriptòmiques entre un conjunt de pacients amb LPB i un grup Control sense LPB.

- Revelar el paper dels miRNA en la modulació de l'expressió dels seus gens diana.

Objectius específics:

- Identificar els mRNAs diferencialment expressat per determinar les diferències del transcriptoma codificant entre els LPB i els Controls.

- Determinar si els nivells d'expressió de les dianes dels miRNAs prèviament identificats diferencialment expressats en pacients PBL mostren canvis d'expressió respecte el grup Control.

- Determinar si els nivells d'expressió de les dianes, dels miRNA prèviament identificats diferencialment sobreexpressats en els pacients LPB, es veuen disminuïts.

- Determinar si els nivells d'expressió de les dianes, dels miRNA prèviament identificats diferencialment inhibits en els pacients LPB, es veuen augmentats.

1.4. Enfocament

Durant l'assignatura de pràctiques en empresa es van analitzar dades de *microarrays* per estudiar els *miRNAs* de pacients amb LPB i de Controls sense LPB. Aquest estudi va permetre determinar un conjunt de *miRNAs* que estaven diferencialment expressats en el grup de pacients LPB en comparació amb el grup Control.

El present estudi analitzarà les dades d'RNA-seq, per estudiar el m-RNA total, de la mateixa cohort que en l'estudi de *microarrays*. Els resultats d'aquest estudi permetran descriure les diferències transcriptòmiques entre el grup de pacients amb LPB i el grup Control sense LPB. A més, es determinarà l'expressió dels gens diana dels miRNAs que es van trobar diferencialment expressats en l'estudi anterior.

1.5. Mètode seguit

En aquest projecte es va realitzar una estratègia ben clara i lineal que es va dividir en dos blocs:

1- Anàlisi de les dades d'RNA-seq

En primer lloc, l'estudi parteix dels arxius *fastq* generats a partir de la seqüenciació de les mostres. El processat d'aquestes dades es va desenvolupar utilitzant com a sistema operatiu, la distribució de Linux, Ubuntu. Es va considerar que aquest entorn facilitava la visualització i manipulació dels d'arxius que es volien estudiar i, a més, facilitava l'ús dels programes necessaris per al processament de les dades. El processat d'aquestes dades, va consistir en l'estudi de qualitat dels arxius, l'eliminació dels fragments de mala qualitat, l'alineament dels fragments amb el transcriptoma de referència, el control de qualitat d'aquest alineament, i finalment, l'obtenció dels recomptes de lectures alineades.

En segon lloc, es van modelar les dades del recompte mitjançant un model binomial negatiu i es van determinar els gens expressats diferencialment. A més, es va realitzar la visualització dels resultats amb *Heatmaps* i *volcanoplots,* es van identificar els gens més diferencialment expressats i es va fer un petit anàlisi de significat biològic dels gens seleccionats. Per realitzar aquesta part s'ha utilitzat el llenguatge R amb l'entorn de desenvolupament integrat R Studio per a l'anàlisi estadístic de les lectures obtingudes. En concret, s'ha utilitzat el paquet *DESeq2* que proporciona *Bioconductor*, per la normalització, visualització i l'anàlisi diferencial de les lectures. Fins aquest punt, s'ha considerat realitzat l'objectiu del projecte.

2- Expressió dels gens diana miRNA diferencialment expressats

La segona part del projecte consisteix en perllongar els resultat anteriors per acomplir els objectius restants. Aquesta etapa s'ha iniciat cercant els gens diana dels miRNA que es van determinar diferencialment expressats. En aquest cas, no s'han utilitzat algoritmes predictors per calcular les dianes dels miRNAs, sinó que s'ha decidit emprar la base de dades de la companyia Affymetrix amb la que es van efectuar els estudis de *microarrays*. Per aquest tipus de *microarray*, Affymetrix proporciona les dianes, que han estat validades per ells mateixos, de les seqüències complementaries de les sondes del seu *microchip*. El compliment de la cerca dels gens diana dels miRNAs, i per tant, del segon objectiu, ha permès estudiar i visualitzar els nivells d'expressió dels gens diana dels miRNA diferencialment expressats.

1.6. Planificació del treball

Anàlisi d'RNA-seq:

a. Estudiar el context clínic de les dades i preparar l'entorn de treball.
b. Plantejament del problema i anotació de les variables dels arxius (*target* file).
c. Control de qualitat de les dades.
d. Alineament amb el transcriptoma i quantificació de l'expressió de cada mostra.
e. Control de qualitat de l'alineament.
f. Alineament amb el genoma complert.
g. Anàlisi de l'origen dels fragments alineats.
h. Anàlisi d'exploració i visualització dels recomptes de cada mostra (PCA, Heatmaps).
i. Anàlisi d'Expressió Diferencial.
j. Gràfiques dels Resultats i anotació dels Resultats.

Anàlisi de l'expressió dels gens diana dels miRNA:

a. Recopilar els miRNA diferencialment expressats entre el grup LPB i el grup Control en l'estudi de *microarrays* .
b. Identificar els gens diana dels miRNAs llistats en el punt anterior.
c. Determinar els nivells d'expressió dels gens diana dels miRNAs .
d. Visualització de l'expressió dels gens diana.
e. Determinar si els nivells d'expressió dels gens diana dels miRNAs sobreexpressats en els pacients PBL es veuen disminuïts.
f. Determinar si els nivells d'expressió dels gens diana dels miRNAs inhibits en els pacients PBL es veuen augmentats.
g. Visualització dels resultats.

Aleix Méndez López

Calendari:

Mes	Lu.	Ma.	Mi.	Ju.	Vi.	Sá.	Do.
Febr. 2021	15 TRAMITS	16 TRAMITS	17 TRAMITS	18 TRAMITS	19 TRAMITS	20 TRAMITS	21 TRAMITS
	22 **Set up**	23 **Set up** Estudiar Contextclínic	24 **Set up** Estudiar Contextclínic	25 **Set up** Estudiar Contextclínic	26 **Set up** Estudiar Contextclínic	27 **Set up** Estudiar Contextclínic	28 **Set up** Estudiar Contextclínic
Mar. 2021	1 **Set up** Workstaton Planificar l'estudi	2 **Set up** Workstaton Planificar l'estudi	3 **Set up** Workstaton Planificar l'estudi	4 **Set up** Workstaton Planificar l'estudi	5 **Set up** Workstaton Planificar l'estudi	6 **Set up** Workstaton Planificar l'estudi	7 **Set up** Workstaton Planificar l'estudi
	8 **Set up** Workstaton Planificar l'estudi	9 **Set up** Configurar SO i softwares	10 **Set up** Configurar SO i softwares	11 **Set up** Configurar SO i softwares	12 **Set up** Configurar SO i softwares	13 **Set up** Configurar SO i softwares	14 **Set up**
	15 **RNA-seq**	16 QC	17 QC	18 QC	19 QC	20 QC	21 QC
	22 Alineament i Quantificació	23 Alineament i Quantificació	24 Alineament i Quantificació	25 Alineament i Quantificació	26 Alineament i Quantificació	27 Alineament i Quantificació	28 Alineament i Quantificació
	29 Visualització i Exploració dades	30 Visualització i Exploració dades	31 Visualització i Exploració dades	1 Visualització i Exploració dades	2 Visualització i Exploració dades	3 Visualització i Exploració dades	4 Visualització i Exploració dades
Abr. 2021	5 Anàlisi d'Expressió Diferencial	6 DE	7 DE	8 DE	9 DE	10 DE	11 DE
	12 Gràfics i Resultats	13 Gràfics i Resultats	14 Gràfics i Resultats	15 Gràfics i Resultats	16 Gràfics i Resultats	17 Gràfics i Resultats	18 **RNA-seq**
	19 **mRNA vs miRNA**	20 miRNA DE	21 miRNA DE	22 miRNA DE	23 miRNA DE	24 miRNA DE	25 miRNA DE
	26 *Targets* miRNA DE	27 *Targets* miRNA DE	28 *Targets* miRNA DE	29 *Targets* miRNA DE	30 *Targets* miRNA DE	1 *Targets* miRNA DE	2 *Targets* miRNA DE
May. 2021	3 *Targets* inhibits	4 *Targets* inhibits	5 *Targets* inhibits	6 *Targets* inhibits	7 *Targets* inhibits	8 *Targets* inhibits	9 *Targets* inhibits
	10 *Targets* sobreexpressats	11 *Targets* soberexpressats	12 *Targets* sobreexpressats	13 *Targets* sobreexpressats	14 *Targets* sobreexpressats	15 *Targets* sobreexpressats	16 *Targets* sobreexpressats
	17 Gràfics i Resultats	18 Gràfics i Resultats	19 Gràfics i Resultats	20 Gràfics i Resultats	21 Gràfics i Resultats	22 Gràfics i Resultats	23 **mRNA vs miRNA**
	24 **ESCRIURE LA MEMÒRIA**	25 **ESCRIURE LA MEMÒRIA**	26 **ESCRIURE LA MEMÒRIA**	27 **ESCRIURE LA MEMÒRIA**	28 **ESCRIURE LA MEMÒRIA**	29 **ESCRIURE LA MEMÒRIA**	30 **ESCRIURE LA MEMÒRIA**
	31 **ESCRIURE LA MEMÒRIA**	1 **ESCRIURE LA MEMÒRIA**	2 **ESCRIURE LA MEMÒRIA**	3 **ESCRIURE LA MEMÒRIA**	4 **ESCRIURE LA MEMÒRIA**	5 **ESCRIURE LA MEMÒRIA**	6 **ESCRIURE LA MEMÒRIA**
Jun. 2021	7 **ESCRIURE LA MEMÒRIA**	8 **ENTREGA DE LA MEMÒRIA**	9 **PRESENTACIÓ**	10 **PRESENTACIÓ**	11 **PRESENTACIÓ**	12 **PRESENTACIÓ**	13 **PRESENTACIÓ**

1.7. Breu sumari de contribucions i productes obtinguts:

Entregables realitzats durant el període de desenvolupament del projecte:

- Proposta del Treball Final de Màster.
- Document de definició del treball (PAC 0).
- Pla de treball (PAC 1).
- Desenvolupament del treball Fase I (PAC 2).
- Desenvolupament del treball Fase II (PAC 3).
- Memòria (PAC 4).
- Presentació Virtual en format de video (PAC 5).

Elements generats durant l'estudi:

- Resultats, taules i gràfics dels diferents resultats obtinguts.
- Codi R per realitzar l'anàlisi i visualització de l'expressió diferencial de les dades.
- Codi SBATCH per llençar diferents tasques al gestor de cues SLURM que va permetre processar totes les dades mitjançant un *cluster* de supercomputació del CSUC.
- Un major coneixement del transcriptoma de la patologia estudiada i del paper dels miRNA sobre l'expressió dels seus gens diana.

1.8. Breu descripció dels altres capítols de la memòria:

En els següents apartats s'explicarà breument la biologia del que s'analitzarà, les eines i els mètodes utilitzats, i els resultats i conclusions que s'han obtingut :

- **Context Biològic**: Introducció biològica dels elements transcriptòmics que s'estan estudiant i de la tecnologia de seqüenciació utilitzada.
- **Material i mètodes**: Descripció de les eines, els mètodes i els passos seguits en el processament de les dades i en l'anàlisi diferencial.
- **Resultats:** Explicació i representació dels gens necessaris per resoldre els objectius plantejats.

2. ESTAT DE L'ART.

El transcriptoma és el conjunt complet de transcripcions d'una cèl·lula durant una etapa de desenvolupament específica o durant una condició determinada. La comprensió del transcriptoma és essencial per interpretar els elements funcionals del genoma, revelar els components moleculars de les cèl·lules i els teixits, i també per comprendre la diferenciació cel·lular i l'origen de moltes malalties. La transcriptòmica té com a objectiu identificar tots els tipus de transcripcions, inclosos els RNA missatgers (mRNA), els RNA no codificants (ncRNA) com els microRNA (miRNA) o *small* RNAs (Srna); determinar l'estructura transcripcional dels gens com patrons d'splicing i altres modificacions post-transcripcionals; i tal i com es realitzarà en el present estudi, quantificar els canvis en la transcripció entre diferents processos o condicions. Les condicions típiques que es solen comparar són els controls sans enfront dels pacients amb malaltia (com en el present treball), teixits diferents, moments temporals diferents... El motiu d'aquestes comparacions es deu al fet que gairebé totes les cèl·lules mantenen el conjunt complert de gens, tot i que no tots s'expressen. A causa d'aquest repartiment del conjunt complet de gens entre cèl·lules, l'anàlisi de comparació entre transcriptomes és el mètode principal per investigar els factors que generen diferències funcionals entre les cèl·lules.

Com que la quantitat de transcripcions es representa, amb un nombre real, i com que els nombres reals no poden ser exactament iguals entre ells per casualitat, són necessaris models estadístics que validin que les quantitats entre dues transcripcions difereixen entre elles. Les dues tecnologies per quantificar les transcripcions són els *microarrays* i seqüenciació d'alt rendiment (RNA-seq). Tot i que aquestes dues tecnologies mesuren biològicament les mateixes variables, és a dir, la quantitat de transcripcions, els resultats difereixen entre si. Els *microarrays* mesuren la quantitat de transcripcions a partir de la quantitat de fluorescència emesa pels fragments dels nucleòtids quan s'uneixen a les sondes complementàries. D'altra banda, la quantitat de transcripció mesurada per l'RNA-seq és un nombre enter, ja que recompta el nombre de fragments que s'alineen a una determinada regió del mRNA. En el present estudi s'han recopilat els resultats de l'anàlisi bioinformàtic/bioestadístic efectuat entre dos condicions, els nivells de transcripció dels miRNA en pacients amb LPB i en Controls sense LPB a partir de dades de *microarrays*. I s'ha generat l'anàlisi bioinformàtic/bioestadístic dels nivells de transcripció del mRNA entre les mateixes dos condicions, a partir de dades d'RNA-seq. Sabent que els miRNA tenen una funció relacionada amb la regulació gènica, es vol estudiar si s'observa una correlació entre els miRNA diferencialmet expressats i els mRNA diferencialment expressats.

3. MATERIALS I MÈTODES.

En aquest capítol es descriuen les eines i els programes utilitzats seguint l'ordre en que s'han efectuat durant l'anàlisi. Tal com s'ha explicat en l'apartat Mètode seguit, els programes s'han executat utilitzant com a sistema operatiu, la distribució de Linux, Ubuntu (v. 18.04.5), però alguns processos s'han desenvolupat en un *cluster* de supercomputació degut a la demanda de requisits computacionals necessaris per realitzar-se. L'Institut Josep Carreras contracta un *cluster* de supercomputació (CSUC) que dóna servei als diferents grups de recerca de l'Institut. El *cluster* va ser utilitzat en aquest estudi durant el procés de *trimming* de les lectures que es va executar amb el *software* Trimmomatic, i durant el procés d'alineament, i de creació de l'Índex genòmic de referencia, que es va realitzar amb el programa STAR.

3.1. Dades analitzades i disseny experimental.

Les dades analitzades han estat proporcionades pel grup de recerca de Neoplàsies Limfoides de l'Institut Josep Carreras i pertanyen a un conjunt de mostres parafinades de teixit tumoral, obtingudes en el moment del diagnòstic entre els anys 2003 i 2018, de pacients amb limfoma plasmablàstic. El conjunt recopilat consta d'un total de 64 mostres, de les quals 41 corresponen a pacients amb LPB, 14 a Controls sense LPB, 3 a Desordres Limfoproliferatius Post-Transplant (PTLD), 3 a Sarcomes relacionats amb l'Herpes Virus (HHV8+), 2 a Síndromes de Richter i una corresponent a un pacient amb Plasmacitoma. Han estat processades totes les dades, però el present estudi té com objectiu analitzar les diferències d'expressió entre el grup LPB i el grup Control, per tant, tot i que s'ha realitzat el control de qualitat, l'alineament i la quantificació de les lectures en totes elles, només s'han tingut en compte les 55 mostres del conjunt LPB i Control. Per facilitar la comprensió del projecte, d'ara endavant es descriuran els processos referint-se només al conjunt d'interès.

Primerament, es va extreure el miRNA de 55 mostres, 41 mostres de pacients amb LPB i 14 de Controls sense LPB i es va dur a terme l'anàlisi de *microarrays* d'expressió per a totes elles. El xip utilitzat va ser el *GeneChip miRNA 4.0 Arrays* de la casa comercial *Affymetrix*. Els resultats dels *microarrays* van ser 66 arxius .CEL que contenien el nivell d'expressió dels miRNA de cada mostra. Durant l'assignatura de pràctiques en empresa es va realitzar l'anàlisi bioinformàtic d'aquests arxius, es va efectuar l'anàlisi diferencial entre l'expressió dels miRNA dels pacients amb PBL i el grup Control i es van guardar els resultats del miRNA diferencialment expressats.

D'altra banda, es va extreure l'RNA total per a seqüenciar l'RNA missatger del mateix conjunt de pacients. L'RNA total, a més de contenir l'RNA missatger que es vol estudiar, allotja fins a un 97% de molècules d'RNAs (ncRNA resultants de la transcripció) que no codifiquen per

proteïnes i, de les quals, la més abundant és l'RNA ribosòmic (Jhon et al. 2001). Per a una detecció eficient dels transcrits, generalment, s'eliminen els rRNAs de l'RNA total, ja sigui per selecció positiva de polyA+ o per *rRNA Depletion* (selecció negativa) abans de ser seqüenciats. El mètode de selecció per polyA+ supera al de *l'RNAr Depletion* en quant a la cobertura exònica dels fragments seqüenciats i en la precisió de la quantificació gènica, d'altra banda, el mètode d'*RNA depletion* proporciona més informació sobre altres transcripcions (ncRNA) i el seu rendiment és millor per als RNA que estan degradats (Zhao et al. 2018). En aquest cas, l'antiguitat de les mostres i el fet d'estar parafinades, incrementava la possibilitat de que l'mRNA estigués degradat, a més a més, no es volia descartar la possibilitat d'estudiar els fragments d'ncRNA d'aquests pacients. Per aquestes raons es va obtenir l'mRNA per selecció negativa realitzant un procés d'rRNA *Depletion*.

Finalment, es van seqüenciar 55 mostres, cada mostra es va dividir entre 4 i 12 rèpliques segons el nombre de cops que es va decidir seqüenciar cadascuna, la meitat de les rèpliques de cada mostra van ser seqüenciades en *flowcells* diferents i cada rèplica va ser carregada en una *lane* diferent de la *flowcell.* En total es va realitzar *multiplexPaired-End Seqüencing* de 380 rèpliques.

Multiplex sequencing permet agrupar biblioteques de múltiples experiments en una única reacció de seqüenciació. Per identificar de quin experiment prové cada seqüència determinada, cada biblioteca es prepara amb adaptadors que contenen etiquetes diferents (índex o codi de barres). Les etiquetes solen ser seqüències curtes (5-7 pb) que es llegeixen durant la seqüenciació. *Paired-End Seqüencing* permet seqüenciar els dos extrems d'un fragment i facilita la detecció de reordenaments genòmics i elements repetitius de la seqüència, així com fusions entre gens i transcripcions noves. A més de produir el doble de seqüències, les lectures complementàries permeten un alineament més precís i la possibilitat de detectar insercions i delecions, cosa que no és possible amb dades d'una sola lectura (Nakazato et al. 2013).

L'elaboració del present estudi comença amb l'anàlisi dels arxius *fastq* que contenen els resultats de les seqüenciacions. S'analitzen un total de 760 arxius, ja que per cada rèplica en resulten dos arxius *fastq,* un per cada direcció de les lectures.

3.2. FastQC.

Un cop organitzats els arxius en directoris, es troben els dos arxius *fastq* de cada rèplica en una mateixa carpeta identificada amb el codi de la rèplica que pertoca i s'efectua l'anàlisi de qualitat dels fitxer que es troben comprimits en format GNU ZIP (gzip).

L'anàlisi de qualitat de les dades s'ha realitzat amb el software FastQC (v. 0.11.9), que és un programa dissenyat per detectar possibles problemes en les dades de seqüenciació. Executa un conjunt d'anàlisis sobre els fitxers en format *fastq* o *bam* i produeix un informe en format HTML que resumeix els resultats. FastQC destaca totes les àrees que semblen tenir poca qualitat o que són inusuals, i per tant, permet detectar aquells paràmetres que caldrà analitzar amb més detall (Andrews S. 2010).

Els paràmetres que analitza el software són els següents:

- ✓ *Basic Statics.*
- ✓ *Per base sequence quality.*
- ✓ *Per tile sequence quality.*
- ✓ *Per sequence quality scores.*
- ✓ *Per base sequence content.*
- ✓ *Per sequence GC content*
- ✓ *Per base N content.*
- ✓ *Sequence Length Distribution.*
- ✓ *Overrepresented sequences.*
- ✓ *Adapter Content.*

Aquesta eina ha estat utilitzada per avaluar la qualitat dels resultats del processament de les dades durant les diferents etapes realitzades. D'entrada es va utilitzar sobre les dades crues de la seqüenciació, en segon lloc, sobre els arxius resultants del pre-processament efectuat per eliminar els adaptadors i, finalment, sobre els arxius resultants de l'alineament i la quantificació de les lectures.

3.3. Trimmomatic.

La presència de seqüències producte de la tècnica o de seqüències de mala qualitat en les lectures, com ara els adaptadors, pot donar lloc a errors en els recomptes d'aquestes lectures i per tant, en els anàlisis posteriors (Bolger et al. 2014).

El software utilitzat per pre-processar les dades i eliminar aquestes seqüències ha estat Trimmomatic (v.0.39). És una eina *multithread* de la línia de comandament que s'utilitza per retallar dades generades majoritàriament per Illumina. En aquest cas, trimmomatic s'ha executat amb el mode *paired end*, el qual té en compte la informació addicional continguda en les lectures emparellades per trobar més eficientment els fragments que corresponen a *primers* o adaptadors introduïts durant la preparació de les llibreries (Bolger et al. 2014).

Trimmomatic també permet introduir paràmetres de qualitat per retallar i eliminar fragments que no compleixen els llindars de qualitat que s'estipulen. Les seqüències dels arxius que s'estan analitzant tenen la qualitat suficient per utilitzar trimmomatic de la manera menys agressiva possible per tal d'evitar manipular els resultats de les lectures innecessàriament.

Per executar Trimmomatic és necessari introduir-li la seqüència de nucleòtids dels adaptadors presents en les lectures. Com hem vist anteriorment, FastQC és capaç de detectar la contaminació per adaptadors en les dades, però també és capaç d'identificar quin tipus d'adaptador s'ha utilitzat en la majoria de casos. El control de qualitat de les dades en cru mostrava que les lectures estaven contaminades per adaptadors universals d'Illumina. Es van cercar les seqüències nucleotídiques que constitueixen aquests tipus d'adaptadors i es va confirmar directament sobre els arxius *fastq* la presència real i abundant d'aquests tipus de seqüències. Un cop indicades les seqüències nucleotidíques dels adaptadors, es va procedir a executar Trimmomatic. Els arxius que es van obtenir van ser, per un costat, els *fastq* sense els adaptadors i per l'altre, arxius *fastq* de les lectures dels fragments retallats.

3.4. Softwares d'alineament i quantificació de les seqüències.

Un cop explorada la qualitat de les lectures en brut i un cop pre-processades les dades, s'inicia la quantificació de l'expressió a nivell transcripcional. L'objectiu d'aquest pas és identificar a partir de quina transcripció es va originar cada lectura i el nombre total de lectures associades a cada transcripció.

Hi ha molts enfocaments bioinformàtics existents sobre la quantificació de l'RNA-Seq. Tradicionalment, l'enfocament més popular consistia en alinear les lectures a un genoma de referència (o transcriptoma) mitjançant un alineador "*spliced*" com TopHat (Trapnell et al. 2009) o STAR (Dobin et al. 2013). Aquest pas és computacionalment pesat en el càlcul i cada mostra pot trigar varies hores, depenent de les prestacions que es disposin i dels paràmetres que s'utilitzin. A més a més, un cop generats els alineaments (arxius *BAM*) es necessiten eines addicionals per quantificar l'expressió gènica i per assignar les lectures als gens. Per aquesta

segona part, s'utilitzen des de tècniques simples de recompte contra una anotació coneguda, fins a estimacions basades en models d'expressió gènica (Robert et al. 2015).

Recentment, s'han desenvolupat mètodes computacionals capaços d'estimar l'abundància de les transcripcions molt ràpidament, en part, renunciant al pas d'alinear les lectures al genoma o al transcriptoma de referència. Aquests mètodes han guanyat popularitat a causa dels requisits computacionals marcadament més reduïts i la seva simplicitat d'ús en comparació amb els protocols de quantificació més tradicionals que requereixen l'alineació de les lectures amb el transcriptoma, seguides del processament posterior dels fitxers BAM resultants per obtenir estimacions de quantificació (Srivastava et al. 2020). Els nous mètodes computacionals es basen en alineaments lleugers (Patro et al. 2015) i pseudo-alineaments (Bray et al. 2016) com *Kallisto*, que comparteixen nocions similars però realment són conceptes diferents que implementen algoritmes distints. Un altre mètode és el "*quasi-mapping*", el quasi mapatge busca trobar les millors assignacions (objectius i posicions) per a cada lectura i està inspirada en els dos mètodes anteriors (Srivastava et al. 2016).

3.4.1. SALMON.

En el present estudi s'ha utilitzat *Salmon* (v. 1.4.0) per realitzar tant el mapatge com el recompte de lectures dels arxius resultants del tractament de les dades crues. *Salmon* és un quantificador transcriptòmic basat en el *quasi-mapping* que és capaç de corregir el biaix entre el contingut de GC dels fragments. Això permet millorar la precisió de les estimacions d'abundància i la fiabilitat de l'anàlisi d'expressions diferencials posteriors. *Salmon* combina un algorisme d'inferència paral·lela i models de biaix amb un procediment de mapatge de lectura ultra-ràpid (Patro et al. 2017). Per quantificar les lectures dels *fastq*, primer és necessària la creació d'un índex, que s'utilitza com a transcriptoma de referència.

Per a l'elaboració de l'índex, s'ha utilitzat l'arxiu (gencode.v38.transcripts.fa), proporcionat per la base de dades GENCODE, que conté les seqüències de nucleòtids de tots els transcrits dels cromosomes de referència humà GRCh38. A més a més, s'han concatenat les seqüències nucleotídiques del genoma primari (GRCh38.p13.genome.fa) juntament amb les seqüències del transcriptoma. D'aquesta manera, s'utilitzen les seqüències del genoma com a "parany" per descartar amb més precisió aquelles seqüències que són més afins a regions que no pertanyen al transcriptoma .

Els resultats de *Salmon* han generat 380 fitxers de quantificació. Aquests documents són de text pla, separat per tabulacions, amb una sola línia de capçalera que nomena totes les columnes. Els fitxers s'anomenen quant.sf i recullen la informació dels recomptes en les

columnes: *Name* (nom del transcrit diana), *Length* (llargada del tramscrit en nucleòtids), *EffectiveLength* (longitud efectiva calculada de la transcripció*)*, *TPM* (transcripcions per milió) i *NumReads* (estimació del nombre de lectures alineades a cada transcript). El TPM és la mesura d'abundància relativa que s'ha utilitzat per a l'anàlisi posterior.

3.4.2. STAR.

Els anàlisis diferencials dels recomptes de les lectures quantificades es va dur a terme amb els resultats obtinguts de l'alineament transcriptòmic amb el *softwareSalmon*, però també es va decidir alinear els fragments a tot el genoma amb STAR (v.2.7.3a) com a mètode comparatiu. Els resultats van permetre aportar més informació sobre l'origen dels fragments seqüenciats i contrastar la qualitat de l'alineament realitzat amb *Salmon*.

STAR (*software Spliced Transcripts Alignment to a Reference*) és un alineador tradicional que està basat en un algorisme d'alineació de tipus "*spliced*", que supera en velocitat els altres alineadors tradicionals, alhora que millora la sensibilitat i la precisió de l'alineació (Dobin et al. 2013). Per alinear les lectures amb STAR, primer és necessari generar un índex que actua com a genoma de referència.

Els dos passos desenvolupats durant la utilització d'STAR han estat la creació del genoma de referència i el mapatge dels fragments al genoma. La generació de l'Índex s'ha dut a terme, en primer lloc, introduint les seqüències de nucleòtids del genoma humà primari en format FASTA (GRCh38.primary_assembly.genome.fa) i en segon lloc, introduint les anotacions d'aquest genoma en format GTF (gencode.v37.primary_assembly.annotation.gtf). Tant el fitxer amb les seqüències com el fitxer amb les anotacions han estat proporcionats per la base de dades de referència GENCODE.

Els resultats d'STAR obtinguts per cada rèplica han estat tres fitxers log.out que contenen informació sobre l'estadística i el procés d'alineament, i un fitxer en format binari BAM que conté les seqüències alineades de les lectures.

3.5. Qualimap.

Qualimap (García-Alcalde et al. 2012) és una aplicació Java que té com a objectiu facilitar l'anàlisi de control de qualitat de les dades de mapatge. Qualimap pren els fitxers BAM com a entrada i explora les característiques de les lectures mapades i les seves propietats genòmiques. Ofereix una visió general de la qualitat de les dades en un format d'informe HTML. Els informes són complerts i contenen estadístiques, resums i gràfics de l'alineament. La informació generada per Qualimap es resumeix en els apartats:

- ✓ *Input data & parameters*
- ✓ *Summary*
- ✓ *Reads Genomic origin*
- ✓ *Transcript Coverage Profile*
- ✓ *Junction Analysis*

Es va executar Qualimap (v. 2.2.2d) sobre tots els fitxers BAM que contenien l'alineament genòmic de cada una de les rèpliques que es volia estudiar. Els resultats mostraven informació dels paràmetres anteriors. Per tal de comparar si els resultats de l'alineament amb *Salmon* eren semblants en quant a quantitat de fragments alineats, es va prestar especial atenció a l'origen genòmic de les lectures.

3.6. R, R Studio i Bioconductor.

Per a realitzar aquest treball, s'ha utilitzat R (v. 4.0.5) a través de l'entorn RStudio, un entorn de desenvolupament integrat de codi obert per a R que et permet treballar amb R dins d'una interfície que ofereix una millor disposició visual de les eines i les dades que s'utilitzen.

R és un llenguatge i entorn de programació destinat a l'anàlisi estadístic i la representació gràfica de dades. És un projecte GNU desenvolupat als Laboratoris Bell per John Chambers, que es basa en el llenguatge S, però té una implementació diferent. Està disponible com a programari lliure i complila i s'executa en una àmplia varietat de plataformes UNIX i sistemes similans (Linux), Windows i MacOS (Team RC. 2019). L'entorn d'R està dissenyat per integrar instal·lacions de programari amb un veritable llenguatge informàtic que permet als usuaris afegir funcions addicionals definint noves funcions. D'aquesta manera, R es pot ampliar mitjançant paquets. Hi ha una gran quantitat de paquets disponibles que cobreixen una àmplia gama d'estadístiques modernes en repostoris com CRAN (Team RC. 2019).

Els paquets estadístics més importants utilitzats per realitzar l'anàlisi estadístic del recompte de les lectures formen part de *Bioconductor* (v. 3.12.0). És un projecte de codi obert que es va iniciar al 2001 amb l'objectiu de desenvolupar amb R, les eines necessàries per a l'anàlisi estadístic de dades generades en laboratoris de biologia molecular, especialment, dades genòmiques d'alt rendiment. Inicialment, *Bioconductor* es va originar amb la finalitat d'aportar les eines necessàries per a l'anàlisi de dades generades pels *microarrays,* però actualment els avantatges que ofereix l'*RNA-seq* han produït l'adaptació d'aquestes eines i la creació de paquets, actualment consolidats, per l'anàlisi de les dades que genera aquesta tècnica.

Els paquets necessaris utilitzats en el present estudi es descriuen en els següents apartats:

3.6.1. TXIMETA

El paquet Tximeta és una evolució del paquet tximport (Soneson, Love i Robinson 2015) que s'utilitza per importar les dades generades durant el procés d'alineament i quantificació de les lectures a R. Tximeta ofereix les mateixes funcionalitats de Tximport, però a més permet afegir automàticament la *metadata* de les anotacions dels transcrits generades per Salmon, sempre que s'hagi utilitzat algun dels trancriptomes més comuns (GENCODE, Ensembl, RefSeq per humans o ratolí). El requisit per identificar la procedència de les transcripcions de referència és que tot el directori de sortida de Salmon estigui present i sense modificar (Love et al. 2020).

S'ha utilitzat Tximeta per importar en forma de matrius de recomptes les dades generades per Salmon i s'han guardat en un objecte de classe "*Sumarized Experiment*". Aquest objecte consta de tres parts: el bloc *assay* que conté la matriu de recomptes, el bloc anomenat *RowRange* que conté informació dels rangs genòmics i, finalment, el bloc *colData* que conté la informació sobre les variables de les mostres. Per tant, s'ha creat un objecte capaç de relacionar la informació anterior amb les tres matrius del bloc *assay*: La matriu amb els *counts*, la matriu amb l'estimació de l'abundància de cada transcrit i la matriu amb la mida dels gens.

3.6.2. DESeq2.

S'ha utilitzat el paquet *DESeq2* (Love et al. 2014) per a la normalització, visualització i anàlisi diferencial de les dades de seqüenciació. Aquest paquet es basa en estimar la variància mitjana dependent dels recomptes de la seqüenciació i calcular l'expressió diferencial basada en un model que utilitza una distribució binomial negativa.

Els valors de la matriu han de ser recomptes dels fragments seqüenciats i no normalitzats perquè el model estadístic *DESeq2* es mantingui i s'avaluï correctament la precisió de la mesura.

Bioconductor permet emmagatzemar les dades en objectes de diferents classes que faciliten i garanteixen que totes les variables de les dades es mantinguin estables mentre es treballa amb elles. A més, de proporcionar facilitats per manipular les dades (crear subconjunts, reordenar files i columnes), es poden utilitzar per moure les dades entre paquets.

DESeq2 utilitza una classe personalitzada que s'anomena *DESeqDataSet* i està basada en la classe d'objecte creat anteriorment *SummarizedExperiment,* per tant, és fàcil convertir d'una de les classes en l'altre. La diferència principal entre les dues classes s'objectes és que *DESeqDataSet* força els valors de la matriu a ser valors enters no negatius, i a més, té una

fórmula de disseny associada que indica, entre d'altres coses, quines columnes del colData (variables de les mostres) s'utilitzaran per al disseny experimental i com s'han d'utilitzar.
En el present estudi la formula ha estat dissenyada principalment per calcular l'expressió diferencial entre el grup LPB i el grup control. L'objecte que conté les dades, no només conté pacients LPB i Controls; per tant, s'ha creat un nou objecte *SummarizedExperiment* amb només les mostres LPB i Controls, reduint el nombre de columnes (rèpliques) que serà transformat a un objecte de classe *DESeqDataSet* utilitzant la fórmula de disseny apropiada amb la llibreria DESeq2.

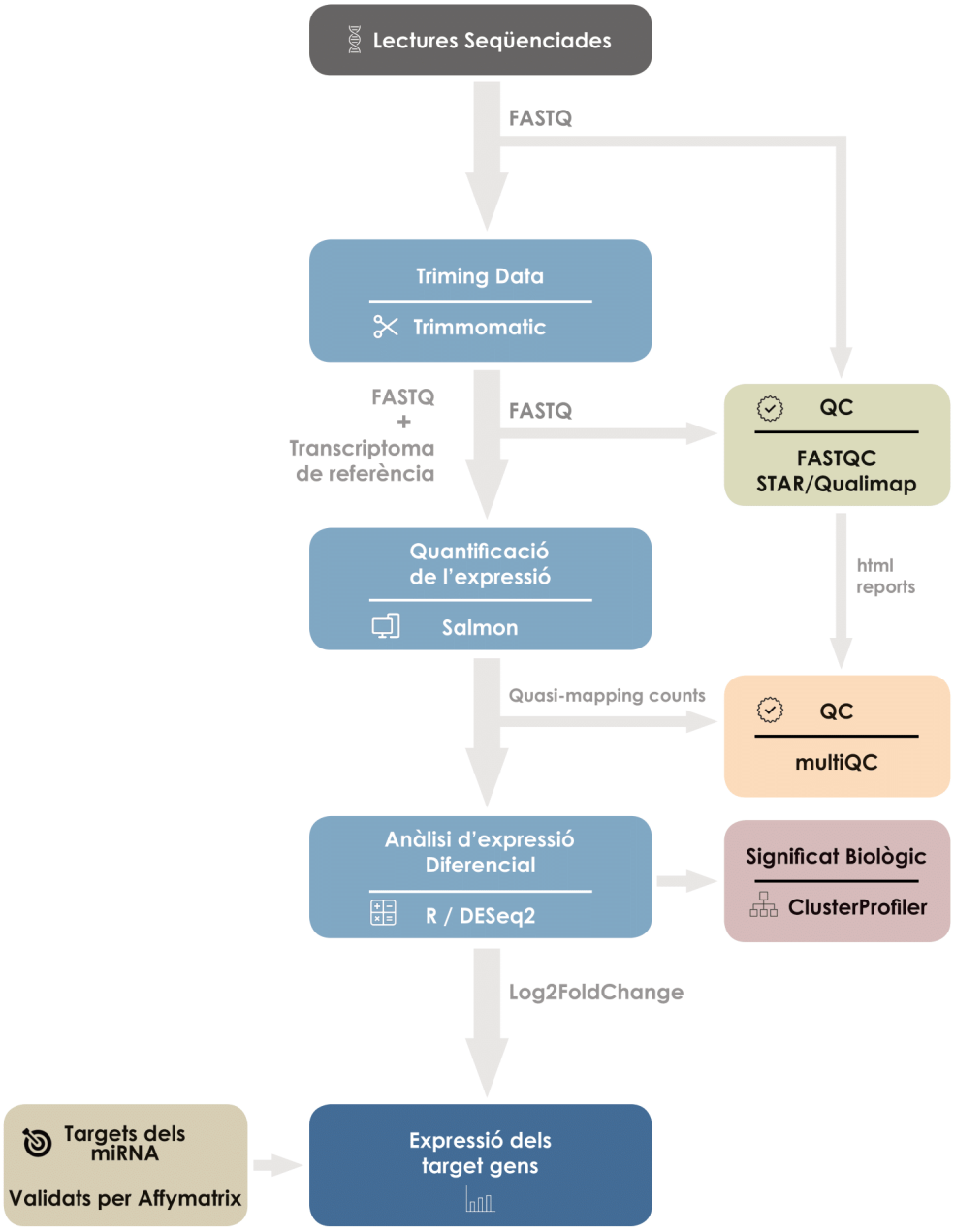
3.6.3. ClusterProfiler.

S'ha utilitzat el paquet *clusterProfiler* (v. 3.18.1) per implementar els diferents mètodes que han permès analitzar i visualitzar els perfils funcionals, els clústers dels gens i els gens resultants de l'anàlisi diferencial realitzat sobre els recomptes de les lectures seqüenciades (Yu G et al. 2012).

El paquet *clusterProfiler* depèn de les anotacions que ofereix *bioconductor* de les bases de dades GO i KEGG per obtenir els mapes d'informació genètic de tot el conjunt GO i KEGG. D'aquesta manera ofereix un mètode de classificació de gens, és a dir, groupGO, per classificar-los en funció de la seva projecció a un nivell específic del conjunt GO, proporciona funcions, enrichGO i enrichKEGG, per calcular l'enriquiment dels termes GO i les vies KEGG basades en la distribució hipergeomètrica i alhora facilita diversos mètodes de visualització (Yu G et al. 2012).

4. RESULTATS.

Els resultats del treball estan descrits seguint l'estratègia temporal que s'ha realitzat en conseqüència a la informació obtinguda en cada un dels actes efectuats. Aquesta estratègia s'esquematitza en la imatge 1:

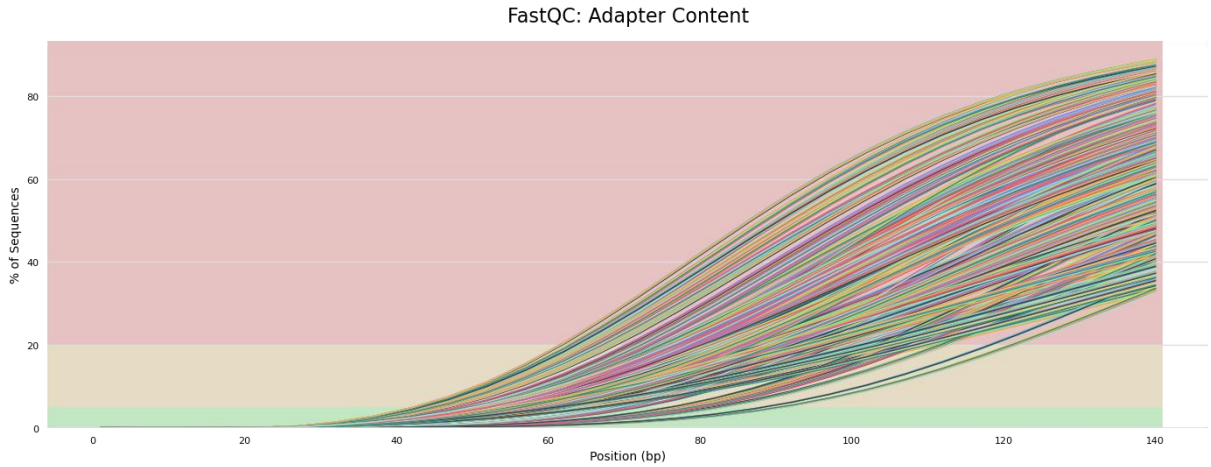


Imatge 1. Diagrama de flux de les etapes principals del projecte. Els processos s'identifiquen en la part superior del calaix i les eines principals utilitzades per realitzar-los en la part inferior del calaix.

4.1. Control de Qualitat dels arxius *fastq* crus.

Els informes generats amb FastQC mostraven que les lectures de totes les rèpliques seqüenciades complien els llindars de qualitat òptims establerts per poder desenvolupar un correcte anàlisi de les seqüències (multiQC suplementari), excepte pel paràmetre que indica el contingut d'adaptadors. Tot i que les lectures eren de bona qualitat, es va observar que el nombre de lectures seqüenciades en cada rèplica, en alguns casos no superava el milió de lectures. En aquest moment, tenint en compte que cada mostra està seqüenciada 4 o 8 vegades, es va considerar que la suma de les lectures de totes les rèpliques d'una mateixa mostra resultaria tenir els recomptes suficients per ser analitzats.

Com s'ha comentat anteriorment, els resultats del control de qualitat de les dades en cru mostrava que totes les lectures resultants de la seqüenciació presentaven contaminació per adaptadors. Tal com es mostra en la imatge 2, els diferents *fastq* presentaven entre un 40% i un 85% de fragments contaminats per adaptadors.



Imatge 2. Recompte percentual acumulat de la proporció d'adaptadors detectats a cada posició de cadascuna de les seqüències analitzades. Cada línia correspon a una rèplica.

Per seqüenciar els cDNA resultants de la transcripció reversa dels fragments d'RNA de les mostres, són necessàries seqüències adaptadores específiques als extrems dels fragments. Els rols i la composició dels adaptadors varien en funció de la plataforma de seqüenciació utilitzada, però principalment permeten inicialitzar la reacció primària de seqüenciació i l'amplificació dels fragments, així com l'opció d'agrupar múltiples experiments en una única reacció de seqüenciació (*multiplexing*), tal com s'ha realitzat en aquest experiment. En aquests casos, els adaptadors permeten la identificació de la mostra a la qual pertany cada fragment, oferint un índex únic (per cada mostra) integrat a l'adaptador. Tanmateix, la contaminació per

adaptadors en les lectures pot conduir a una incorrecta alineació dels fragments al transcriptoma de referència. Tal com es descriu en el següent capítol, s'ha efectuat el pre-processat necessari dels arxius per eliminar les lectures d'adaptadors dels fragments seqüenciats (Bolger et al. 2014).

L'altre paràmetre de qualitat a tractar degut als resultats inusuals detectats gràcies al programa FastQC, ha estat l'anormal distribució del contingut de GC que es troba en les 8 rèpliques d'una mateixa mostra, tal com s'observa en la imatge 3.

Imatge 3. Contingut mitjà de GC de les lectures de cada seqüenciació. Les lectures detectades han de tenir una distribució aproximadament normal del contingut de GC.

Aquesta mostra va ser descartada per no complir els criteris de qualitat suficients. Tot i que va ser inclosa en el pre-processat de les dades i en l'alineament del transcriptoma, la baixa qualitat que continuava presentant després d'eliminar els adaptadors, i el baix percentatge de lectures alineades en el transcriptoma, van acabar de decidir l'exclusió de totes les seves rèpliques de l'estudi.

4.2. Trimming Data.

Per garantir que la presència d'adaptadors no generés errors en els recomptes de les lectures es va haver de realitzar un pre-processat dels arxius *fastq.* Trimmomatic va ser l'eina utilitzada per abolir els fragments d'adaptadors. A més d'introduir la seqüència de nucleòtids dels adaptadors contaminants, es van especificar paràmetres de qualitat mínims que, tant els fragments com les bases de nucleòtids d'aquests fragments, han de tenir per no ser eliminats.

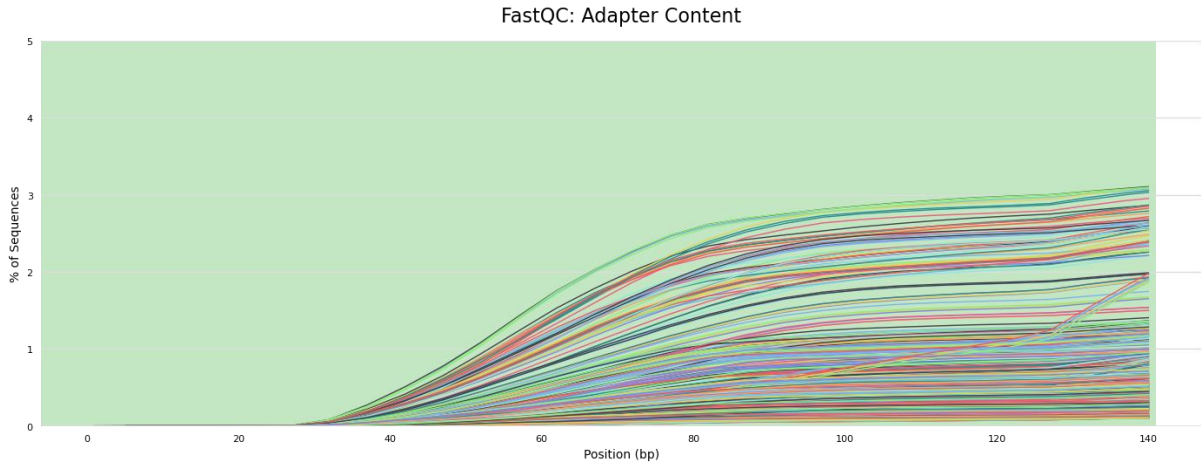
Existeix cert debat en quan a la retallada de les lectures seqüenciades, en dades d'RNA-seq aquest procés condueix a una disminució en el nombre de lectures, tot i que augmenta la proporció de les lectures alineades. No obstant això, efectuar una retallada agressiva pot

afectar negativament els anàlisis posteriors, com ara l'expressió diferencial i l'estimació de l'expressió (MacManes 2014). En conjunts de dades amb mitjanes de qualitat baixes o en protocols de preparació de llibreries susceptibles a la contaminació per adaptadors, aquest procés pot permetre la recuperació de lectures que d'una altra manera serien perjudicials per a l'estimació de les expressions. (Williams et al. 2016)

Tenint en compte que el nombre de fragments totals observats en el FastQC de les dades en cru era baix, els paràmetres introduïts a Trimmomàtic per retallar els fragments de baixa qualitat van ser poc restrictius.

Es va realitzar el control de qualitat dels *fastq* després de remoure el contingut dels adaptadors. L'informe de resultats del FastQC presentava tres punts importants a destacar:

- La contaminació per adaptadors dels fragments seqüenciats pràcticament ha desaparegut. El percentatge de lectures amb adaptadors és inferior al 3%, tal com s'observa en la següent imatge.



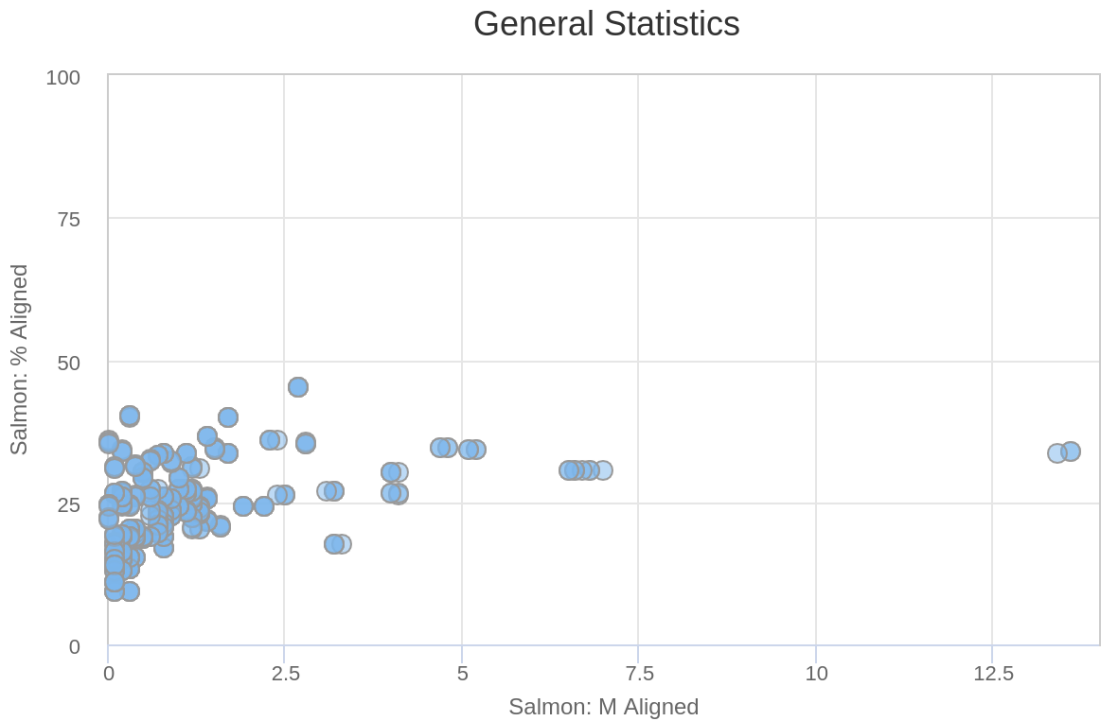
Imatge 4. Recompte percentual acumulat de la proporció d'adaptadors detectats a cada posició de cadascuna de les seqüències analitzades després del pre-processat. Cada línia correspon a una rèplica.

- El nombre de fragments presents en els arxius *fastqc* resultants del processat de les dades en cru disminueix degut a l'eliminació de lectures que poden correspondre a *primers*, fragments curts producte de la tècnica i/o lectures de mala qualitat que no superen els llindars mínims de qualitat.
- Es manté la mateixa distribució inusual en el contingut de GC en les mateixes rèpliques que en el FastQC del *raw data*.

4.3. Alineament i Quantificació.

D'ara endavant, l'anàlisi de les dades s'efectua a partir dels arxius resultants del filtratge i la depuració realitzada dels *fastq* crus durant el *Trimming Data*. En primer lloc, s'ha realitzat l'alineament i la quantificació de les lectures sobre el transcriptoma humà de referència (gencode.v38.transcripts.fa) utilitzant el *software Salmon*. Els fitxers obtinguts contenen els recomptes dels fragments alineats a cada transcrit del transcriptoma de referència utilitzat.

L'anàlisi de qualitat de l'alineament (Taula 2 Suplementari) revela que el percentatge de fragments mapats per cada rèplica oscil·la entre el 10% i el 48% dels fragments, tal com es visualitza en la imatge 5. En altres termes, més de la meitat dels fragments seqüenciats no formen part del transcriptoma humà.



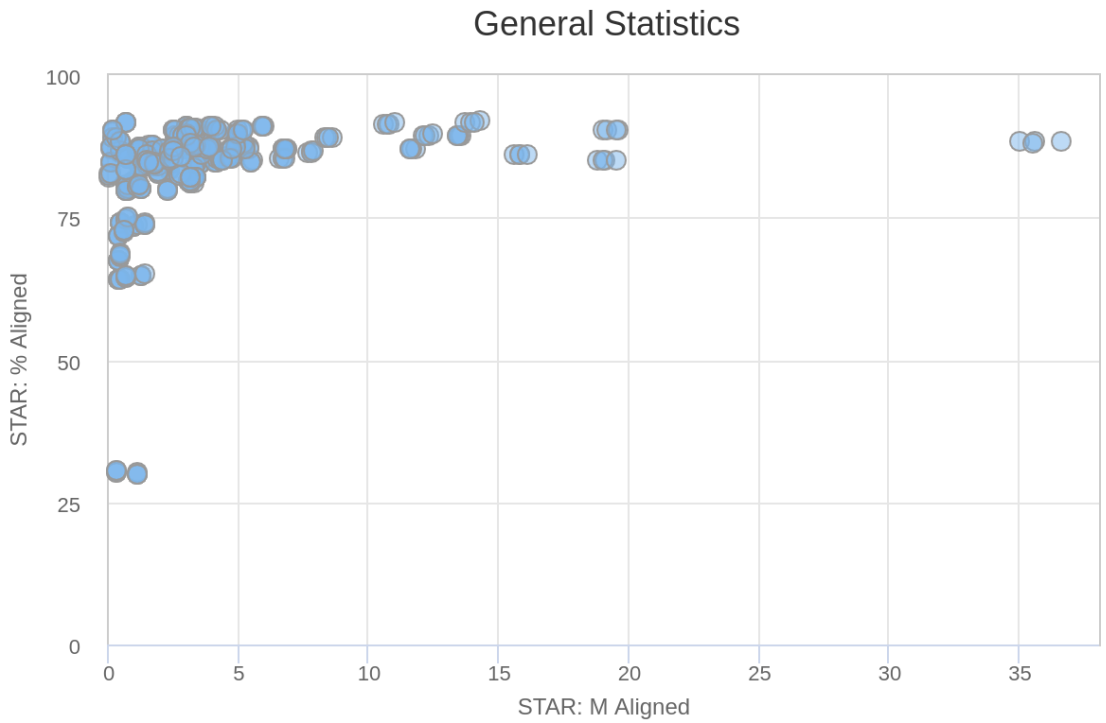
Imatge 5. Cada punt representa una rèplica seqüenciada. En l'eix de les Y es representa el percentatge dels fragments transcriptòmics respecte al total de fragments seqüenciats. En l'eix de les X es representen els milions de lectures alineades al transcriptoma de referència.

En aquest punt, va ser necessari cercar la procedència d'aquesta gran quantitat de fragments que no alineaven en l'anotació de gens codificant utilitzada. Un cop revisats i reanalitzats els controls de qualitat que confirmaven que els fragments seqüenciats, aparentment tenien bona qualitat, es va decidir alinear les seqüencies sobre tot el genoma humà (GRCh38.primary_assembly.genome). Cal recordar que *Salmon* no és capaç de quantificar

lectures que no formin part del transcriptoma, per aquest motiu es va decidir emprar el *software* STAR.

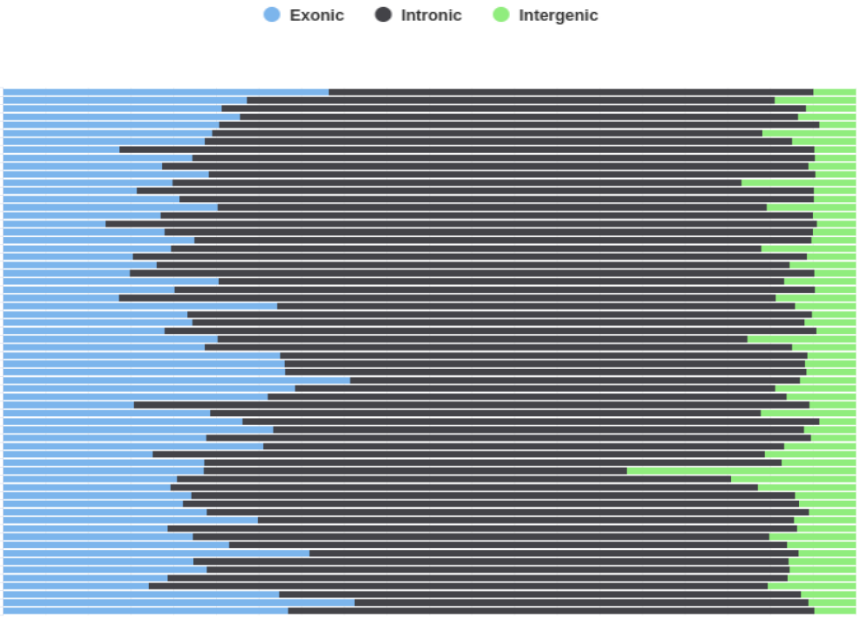
En aquest cas, els resultats generats per STAR són fitxers en format BAM. Aquests fitxers contenen els alineaments de les lectures, però no contenen els recomptes dels alineaments. No obstant, l'anàlisi de qualitat d'aquests alineaments és suficient per visualitzar l'origen i el percentatge dels fragments que s'han alineat en el genoma gràcies al programa Qualimap.

Representat en la imatge 6, el control de qualitat de l'alineament amb STAR, demostra que el percentatge d'alineament obtingut és superior al 65% en tots els casos i superior al 85% en la gran majoria de les rèpliques. En altres termes, els fragments seqüenciats formen part del genoma humà i es descarta que el baix percentatge d'alineament transcriptòmic fós causat per una possible contaminació o degradació de les mostres seqüenciades. Les rèpliques que només tenen un 30% de solapament pertanyen a la mostra que s'ha exclòs de l'estudi per criteris de qualitat.



Imatge 6. Cada punt representa una rèplica seqüenciada. En l'eix de les Y es representa el percentatge dels fragments alineats respecte al total de fragments seqüenciats. En l'eix de les X es representen els millons de lectures alineades al genoma.

D'altre banda, els resultats de l'informe generat per Qualimap demostren que més del 50% dels fragments de les seqüències pertanyen a regions intròniques, tal com s'observa en la següent figura:

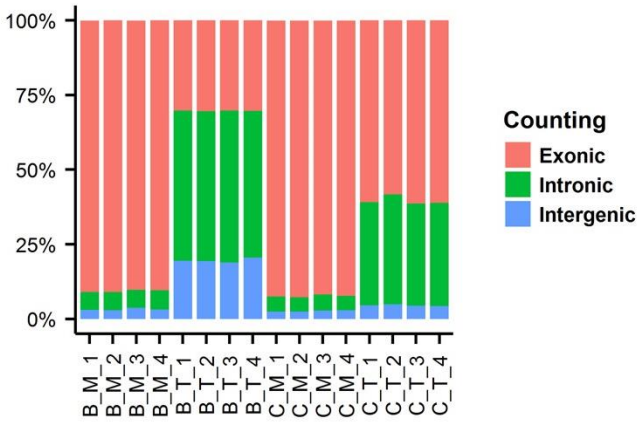


Imatge 7. Representació de la proporció de fragments segons la regió d'origen. Es representa la mitjana percentual de totes les rèpliques d'una mostra. Cada fila representa una mostra. Es representen en blau les regions exòniques, en gris les intròniques i en verd les intergèniques.

Anteriorment s'ha comentat que la seqüenciació es va realitzar sobre l'RNA total de les mostres. Cal recordar que aquest RNA, a més de contenir l'RNA missatger que es vol estudiar en el present estudi, allotja fins a un 97% de molècules d'RNAs (ncRNA resultants de la transcripció) que no codifiquen per proteïnes, de les quals la més abundant és l'rRNA. Pels motius indicats anteriorment, es va utilitzar un mètode de selecció negativa (rRNA *depletion)* per eliminar aquest rRNA i augmentar la concentració de l'mRNA d'interés. Aquest procés és efectiu a l'hora d'eliminar el material ribosòmic, però alhora permet conservar altres molècules d'RNA menys abundants, propiciant en alguns casos que un petit nombre de lncRNAs i *small* RNAs constitueixin una gran fracció de les lectures seqüenciades (Zhao et al. 2018).

Els resultats obtinguts en un estudi recent (Zhao et al. 2018) que compara els dos mètodes tradicionals d'eliminació de l' rRNA, coincideixen amb els del present estudi. Tal com s'observa en la figura següent, el percentatge de regions exòniques seqüenciades disminueix, mentre

que el percentatge de lectures intròniques augmenta quan es realitza rRNA *Depletion* en comptes de selecció positiva (PolyA+).



Imatge 8. Representació gràfica de l'origen dels fragments seqüenciats en l'estudi Zhao et al. 2018. En l'eix de la Y es quantifica el nombre percentual de fragment i l'eix de la X representa cada mostra seqüenciada. Les mostres identificades com B_M i C_M han estat depurades mitjançant PoliA+, en canvi les mostres identificades com B_T i B_T han estat tractades amb rRNA Depletion. Les mostres B_T tenen els resultats més semblants als obtinguts en el present estudi.

Les conclusions de l'alineament sobre el genoma humà s'exposen en els següents punts:

- Les mostres seqüenciades tenen un percentatge d'alineament mitjà superior al 85% sobre el genoma.
- El nombre i el percentatge dels fragments alineats amb *Salmon* sobre el transcriptoma són comparables als nombres i percentatges dels fragments alineats en regions exòniques del genoma (STAR).
- El nombre de lectures alineades sobre el transcriptoma es veu extremadament reduït degut a l'elevada presència de fragments intrònics.
- Més d'un 50% dels fragments seqüenciats són lectures intròniques que redueixen considerablement el nombre de lectures necessàries per estimar l'expressió gènica.
- S'hipotetitza que un petit nombre de lncRNAs i *small* RNAs constitueixen una gran fracció de les lectures degut al procés d'rRNA *depletion* realitzat durant la seqüenciació (Zhao et al. 2018).

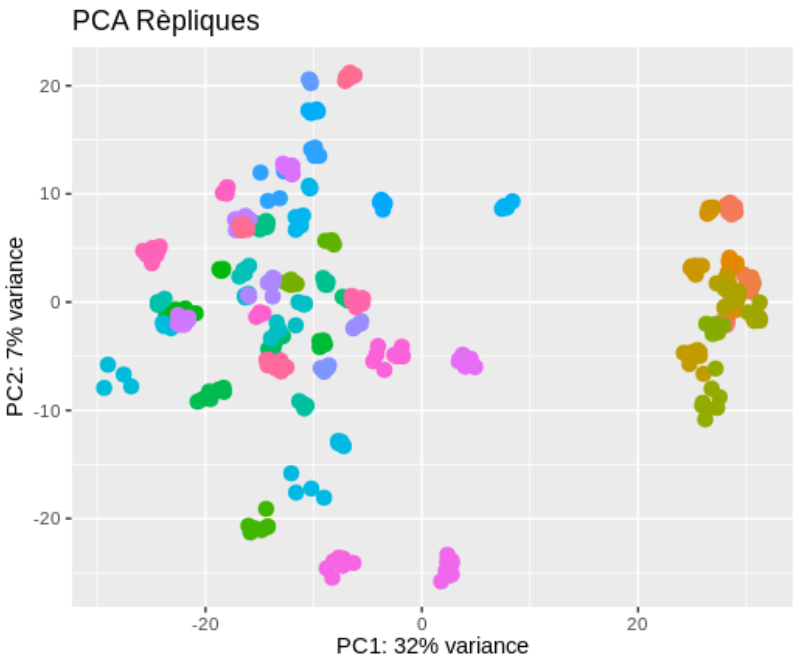
4.4. Exploració i visualització de les dades.

L'anàlisi estadístic s'ha realitzat a partir de la quantificació de l'alineament obtingut amb *Salmon.* En primer lloc, els recomptes estimats per cada transcrit, tenint en compte l'abundància i la llargada del transcrit, han estat importats a R amb el paquet *tximeta*. Aquesta informació, juntament amb les variables de cada rèplica, s'han emmagatzemat en un objecte R de classe *SumarizedExperiment* que té unes dimensions de 60232 files (transcrits) i 372 columnes (rèpliques). En segon lloc, s'ha transformat aquest objecte en un de tipus *DESeqDataSet.* Aquest nou objecte força la matriu de recomptes a que no contingui valors negatius i, a més, associa una fórmula de disseny que permet indicar les variables que es volen estudiar estadísticament.

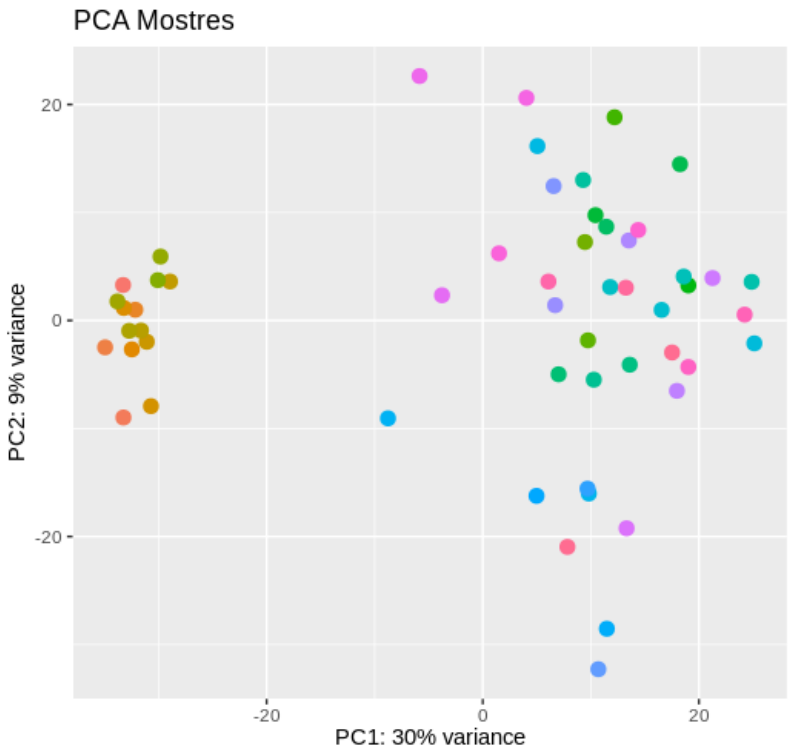
Els mètodes d'exploració de dades com *principal component anàlisi* (PCA) o *Heatmapclustering* funcionen molt bé per a dades que tenen el mateix rang de variància per als diferents intervals dels valors mitjans, és a dir, quan les dades són homoscedàstiques. En canvi, pels recomptes dels RNA-seq, la variància creix amb la mitjana i, per tant, els gens amb més recomptes esbiaixen els resultats. *DESeq2* permet utilitzar *the variance stabilizing transformation* (VST) amb la funció *vst* (Anders et al. 2010) per normalitzar els recomptes i reduir l'efecte dels gens amb més expressió, per tal d'explorar visualment la relació que hi ha entre les mostres. Aquesta funció crea un objecte de classe *DESeqTransform* amb les mateixes característiques que l'objecte *DESeqDataSet*, però amb els *counts* normalitzats.

Abans de realitzar el primer PCA, s'han eliminat totes aquelles files que tenien 0 recomptes en totes les rèpliques tècniques, quedant finalment un objecte de 45515 files i 372 columnes. El primer que s'ha observat és que totes rèpliques tècniques d'una mateixa mostra tenien la mateixa variabilitat i que s'agrupaven entre elles generant *clusters,* tal com s'observa en la imatge 9; per tant, els recomptes de les rèpliques es van concatenar obtenint com a resultat el recomptes totals de les lectures per cada mostra. En aquest punt, es va decidir continuar la resta de l'anàlisi amb aquest nou objecte de 45515 files i 54 columnes.

El següent PCA realitzat mostra la variabilitat genètica de les rèpliques unificades i confirma l'existència de dos grups clars, un que a més de separar-se de l'altre presenta una variabilitat més homogènia. I un altre que mostra una gran heterogeneïtat entre les mostres que el formen.

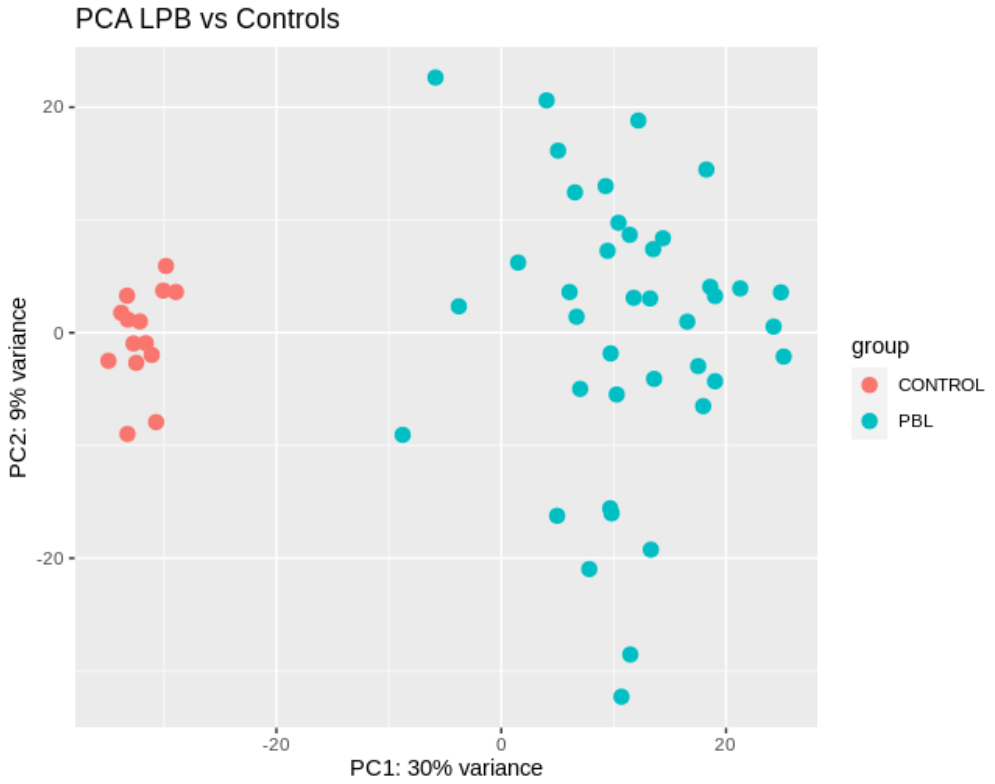


Imatge 9. PCA de les 372 rèpliques. Cada rèplica és un punt i totes les rèpliques d'una mateixa mostra presenta el mateix color. S'observa que les rèpliques d'una mateixa mostra *clusteritzen*.



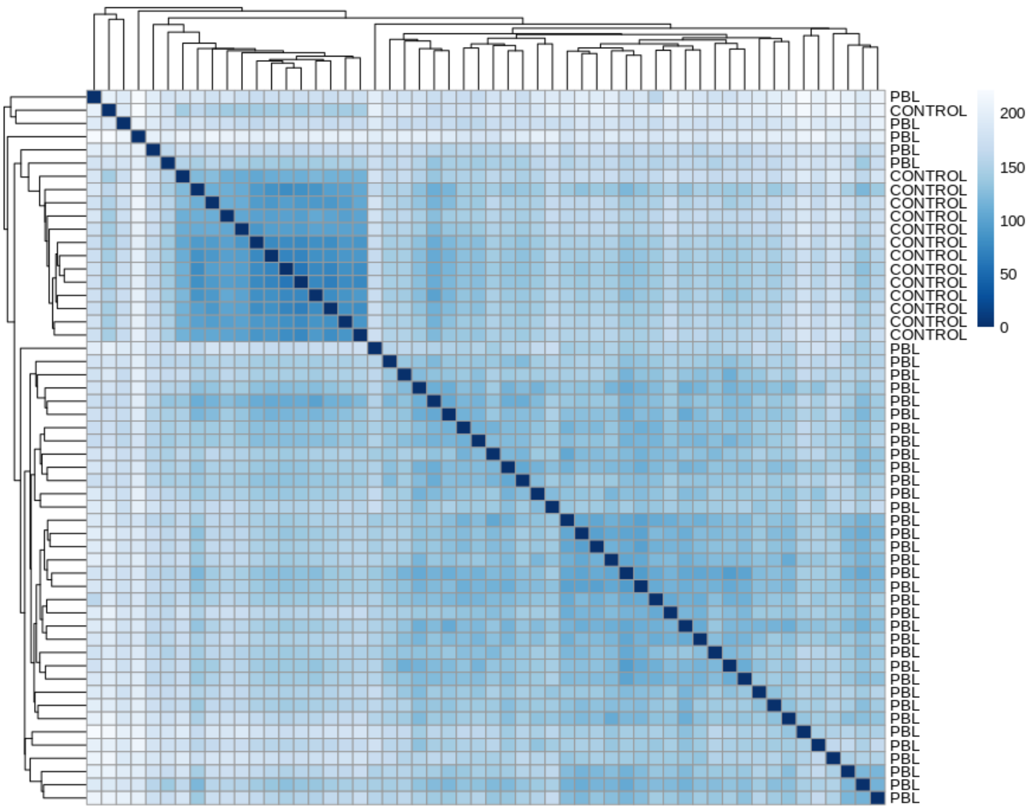
Imatge 10. PCA de les 54 mostres. Cada pacient és un punt i totes les mostres presenten colors diferents. S'observa que les mostres no *clusteritzen* entre elles excepte per les mostres de l'esquerra que corresponen als controls.

En la figura 11 s'observa que els dos grups marcadament separats corresponen al grup de pacients i al grup de controls. A més, s'identifica que la heterogeneïtat es troba present en el grup LPB i, en canvi, el grup control presenta una variabilitat més homogènia.



Imatge 11. PCA de les 54 mostres. Cada pacient és un punt. Les mostres de pacients amb LPB es representen en turquesa i en vermell les mostres Control sense LPB.

Una altra forma de visualitzar la semblança i els clústers entre les mostres ha estat mitjançant el càlcul de la distància Euclidiana de l'expressió de les dades normalitzades (VST). Amb els resultats de les distàncies s'ha generat una matriu que ha estat representada en forma de *heatmap* (Imatge 12). Tal com s'observa en els PCAs anteriors, es distingeixen dos grups que fan referència al grup LPB i el grup Control. En aquest cas, també s'identifica un petit grup de 4 pacients LPB, que segons el *hirerchical clustering* mostren expressions més properes al grup control que a la resta de pacients LPB. Seria interessant en propers estudis determinar quines diferències d'expressió tenen aquestes 4 mostres respecte la resta de LPB.



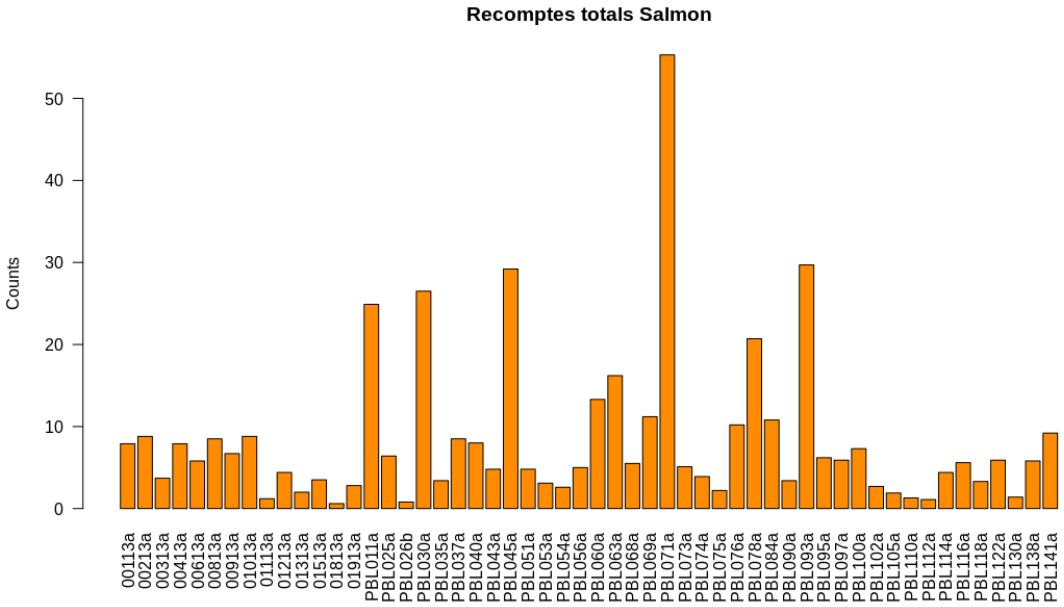
Imatge 12. Heatmap representatiu de les distancies en l'expressió de les 54 mostres (14 Controls i 41 LPB)

L'anàlisi d'exploració i visualització de les dades permet concloure els següents punts:

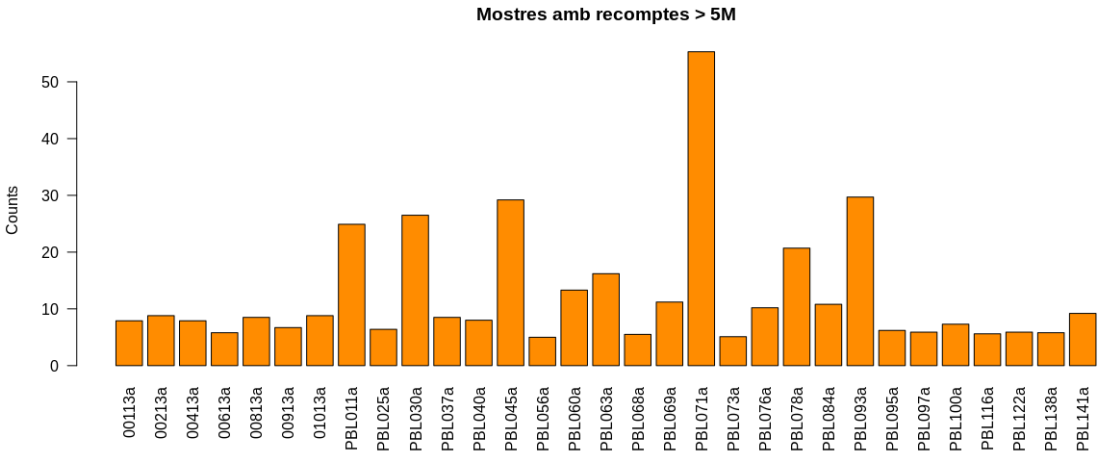
- Les rèpliques de cada mostra seqüenciada presenten la mateixa variabilitat d'expressió. Aquest fet és indicatiu de que no existeixen diferències tècniques entre les diferents seqüenciacions, que puguin haver generat lectures significativament diferents entre elles. Per tant, s'ha decidit concatenar els recomptes de les 372 rèpliques en les 54 mostres pertinents.
- Les mostres es separen en dos grups clars fruit de la variabilitat en l'expressió que presenten. S'ha observat que un dels grups correspon a pacients amb LPB i que l'altre grup està format per les mostres Control sense LPB.
- Les diferents mostres del grup LPB mostren una variabilitat molt heterogènia entre elles, mentre que les mostres del grup Control tenen una variabilitat més homogènia entre elles.
- S'ha identificat un petit grup de 4 mostres LPB que semblen tenir una variabilitat en l'expressió més allunyada a la resta dels LPB i es posiciona més pròxima al grup Control segons els resultats del *Hirerchical Clustering Heatmap*.

4.5. Valoració del recompte obtingut.

Abans de començar l'anàlisi d'expressió diferencial es va tenir en compte la diferència observada entre el nombre de lectures alineades de les diferents mostres. Tot i que en l'anàlisi diferencial es normalitzen les dades i es té en compte el nombre dels recomptes totals obtinguts, la gran diferència en el recompte de lectures entre les mostres va suscitar la possibilitat que els resultats es veiessin esbiaixats. En primer lloc, es van obtenir els recomptes totals de les lectures per cada mostra (imatge 12) i es va decidir establir un *treshold* que descartés aquelles mostres que no disposessin d'un mínim de 5 milions de recomptes totals (imatge 13).



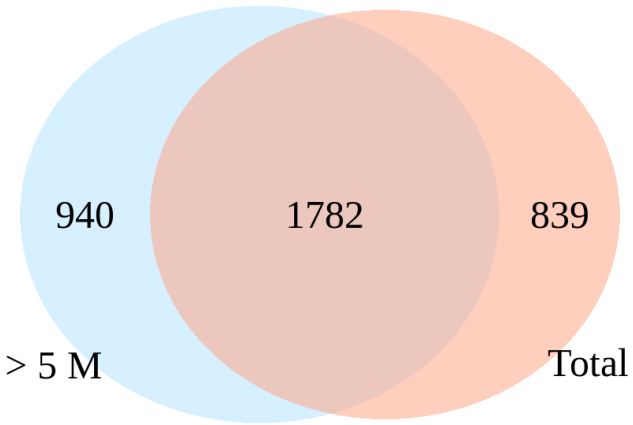
Imatge 12. Recomptes totals de l'alineament de les lectures per cada mostra. L'eix de les Y mostra en unitats de Milió el nombre de recomtes, i en l'eix de les X s'identifiquen les mostres.



Imatge 13. Mostres amb més de 5 milions de recomptes. L'eix de les Y mostra en unitats de Milió el nombre de recomptes, i en l'eix de les X s'identifiquen les mostres.

Es van obtenir un total de 31 mostres que superaven el llindar establert, 7 eren controls i 24 LPB.

En segon lloc, es va realitzar l'anàlisi diferencial entre controls i pacients LPB dels dos conjunts (el de 54 mostres i el de 31 mostres) amb l'objectiu de determinar si els resultats dels gens diferencialment expressats era diferent en cada conjunt. Es van recollir els transcrits amb un *Log2 Fold Change* (L2FC) major a 1 i inferior a -1 i un p valor ajustat (p.Adj) inferior a 0.001 dels dos conjunts i es van comparar. Aquesta comparació va determinar que un 68% dels gens diferencialment expressats (DE) en el conjunt format per les 54 mostres, també es trobaven diferencialment expressats en el conjunt amb 31 mostres tal i com es representa en el següent *Venn Diagram* (Imatge 14).



Imatge 14. Diagrama de Venn. El cercle de l'esquerra representa els gens DE en el grup de 31 mostres i el cercle de la dreta els gens DE en el grup de 54 mostres. Els nombres dels extrems són els gens DE només en aquells grups i el nombre de la intersecció els gens DE coincidents

Cal recordar que el nombre de seqüències duplicades detectades durant el control de qualitat dels *fastq* era molt baix (Taula suplementària 2). Un baix nombre de duplicacions és indicatiu de que el transcriptoma està ben cobert i representat. Tot i tenir un nombre de recomptes escàs, s'ha considerat que aquests recomptes representen adequadament els nivells d'expressió. Tenint en compte l'elevat nombre de gens diferencialment expressats en comú dels dos conjunts i valorant el baix nombre de lectures duplicades (Imatge 34 del suplementari), es va decidir realitzar l'anàlisi diferencial tenint en compte les 54 mostres.

4.6. Anàlisi d'expressió diferencial.

L'anàlisi d'expressió diferencial té com a objectiu determinar quins són els gens que es troben més significativament expressats i menys significativament expressats en un grup respecte l'altre, en el present estudi es determinarà quins són els gens que es troben significativament expressats i inhibits en el grup LPB respecte el grup de controls sense LPB.

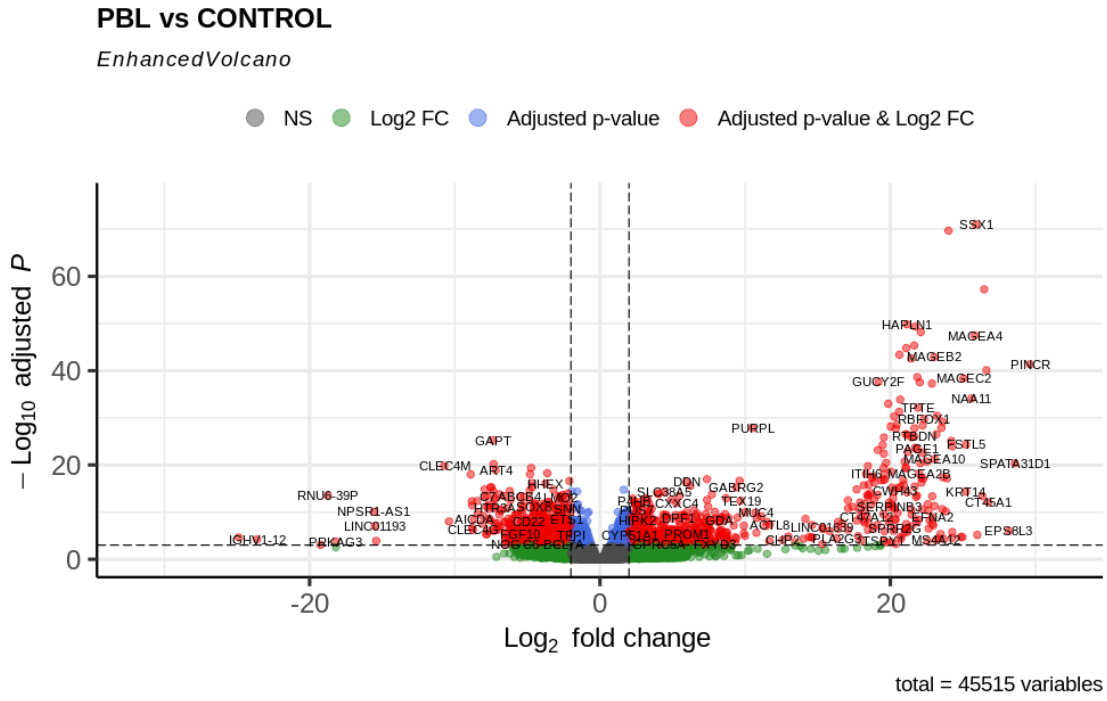
L'anàlisi s'inicia amb els recomptes sense normalitzar, és a dir, a partir de l'objecte de classe *DESeqDataSet* que conté les variables de les mostres i els recomptes emmagatzemats en matrius. *DESeq* és la funció encarregada d'estimar els factors de mida tenint en compte les diferències en la profunditat de la seqüència de cada mostra, a més realitza una estimació de la dispersió dels valors de cada gen, crea un model lineal generalitzat i, finalment, retorna un objecte que continua sent de la classe *DESeqDataSet.* Aquest objecte conté els resultats estadístics, per cada gen, entre d'altres el *log2 foldchange* i els p valors de les variables de la fórmula de disseny en forma d'una taula (*data frame*) que es visualitza amb R de la següent forma:

```
log2 fold change (MLE): Dx2 PBL vs CONTROL
Wald test p-value: Dx2 PBL vs CONTROL
DataFrame with 45535 rows and 6 columns
                    baseMean log2FoldChange     lfcSE      stat      pvalue        padj
                   <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
ENSG00000000003.15  1.9115383       0.221441 0.3156830  0.701467 4.83012e-01 6.38480e-01
ENSG00000000005.6   0.0290275      -0.538712 2.9655515 -0.181657 8.55852e-01          NA
ENSG00000000419.13  8.9641445       0.313169 0.1452528  2.156026 3.10817e-02 6.02995e-02
ENSG00000000457.14 13.7831039      -0.177890 0.0925369 -1.922368 5.45595e-02 1.00123e-01
ENSG00000000460.17 16.8385761       0.583021 0.1318036  4.423405 9.71573e-06 3.22885e-05
...                      ...            ...       ...       ...         ...         ...
ENSG00000288694.1   0.2748600       0.201176  0.714573  0.281534    0.778301    0.874457
ENSG00000288695.1   0.5670643       0.497869  0.457752  1.087639    0.276755    0.409775
ENSG00000288696.1   0.0109428      -0.433906  3.379646 -0.128388    0.897842          NA
ENSG00000288698.1   0.2783141       0.289167  2.294194  0.126043    0.899698    0.937070
ENSG00000288699.1   0.1875308      -0.595138  1.381276 -0.430861    0.666570    0.800875
```

Imatge 15. Objecte DESeqResults que conté les estadístiques resultants de l'anàlisi d'expressió entre el grup LPB i el grup Control sense LPB.

S'han trobat 1739 gens diferencialment expressats amb un valor absolut de log 2 foldchange (L2FC) superior a 2 i un p-valor ajustat per *multiple testing* (p.Adj) inferior a 0.001, dels quals 830 es troben sobreexpressats en pacients LPB i 909 es troben inhibits. La taula de resultats només conté els gens amb els identificadors d'*Ensmbl*, però per poder interpretar millor els resultats s'han utilitzat els paquets d'anotacions *AnnotationDbi* i *org.Hs.eg.db* i s'ha afegit una columna amb els identificadors *Gene Symbol* corresponents a cada codi *Ensmbl*. Finalment, s'han guardat els resultats en un arxius *csv*.

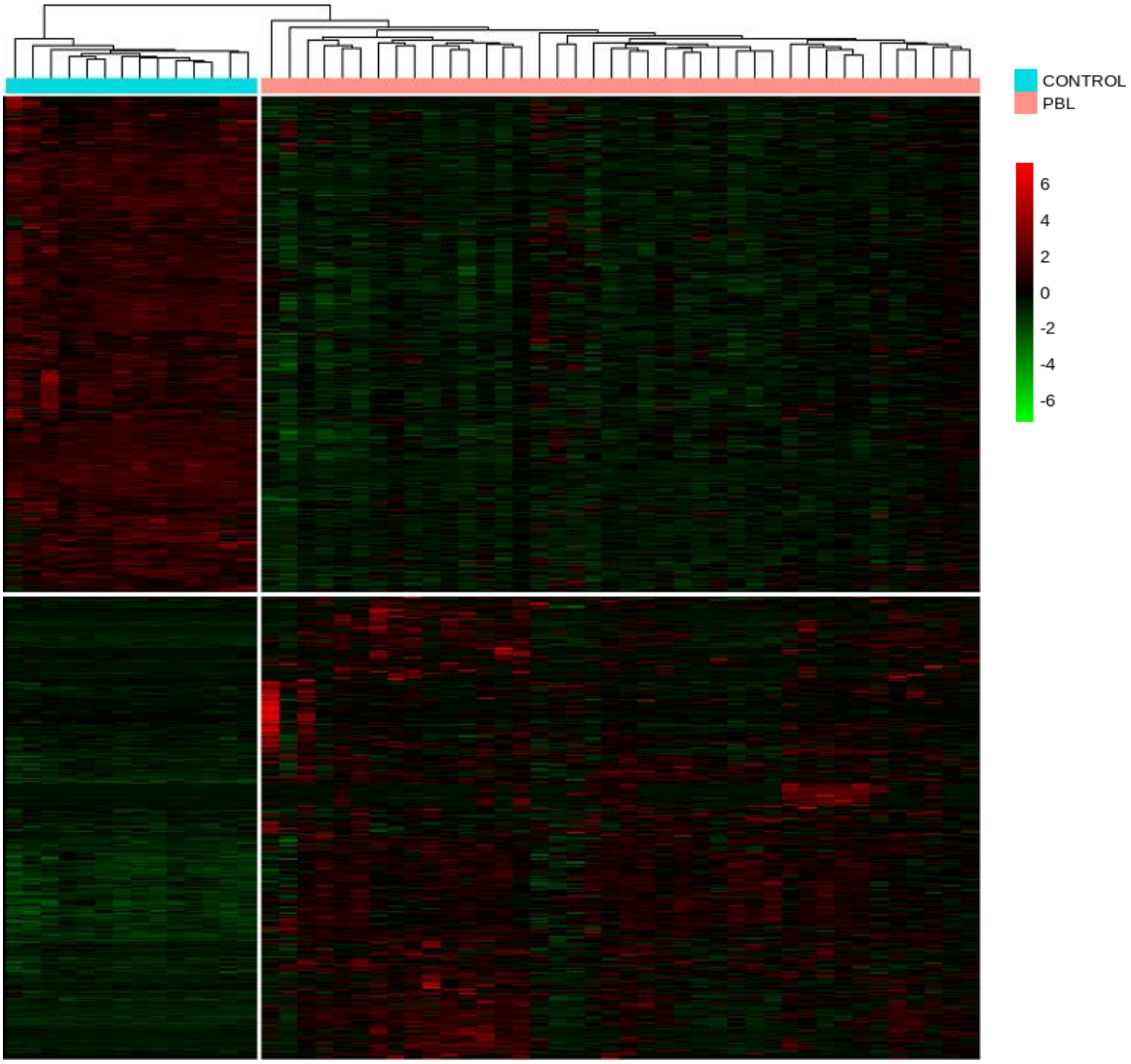
S'han utilitzat diferents estratègies per visualitzar els resultats d'expressió obtinguts. En primer lloc, s'ha realitzat un *volcanoplot* representant l'expressió diferencial dels 45515 gens obtinguts (Imatge 16), tenint en compte el valors estadístics de L2FC i de p.Adj per cada un d'ells.



Imatge 16. *Volcanoplot* de l'expressió diferencial on cada punt representa un gen. En vermell s'identifiquen els gens amb un valor absolut de L2FC superior a 2 i amb un p.Adj inferior a 0.001, en gris els gens amb un valor absolut de L2FC inferior a 2 i amb un p.Adj per sobre de 0.001, en verd els gens amb um L2FC superior a 2 però amb un p.Adj superior a 0.001, i en blau els gens amb un L2FC inferior a 2 però amb un p.Adj inferior a 0.001.

L'anàlisi s'ha realitzat de manera que un L2FC positiu indica una sobreexpressió del gen en el grup LPB i un L2FC negatiu determina una inhibició de l'expressió del gen en el grup LPB. En la imatge anterior s'observa que els gens amb un L2FC més extrem es troben sobreexpressats en el grup LPB.

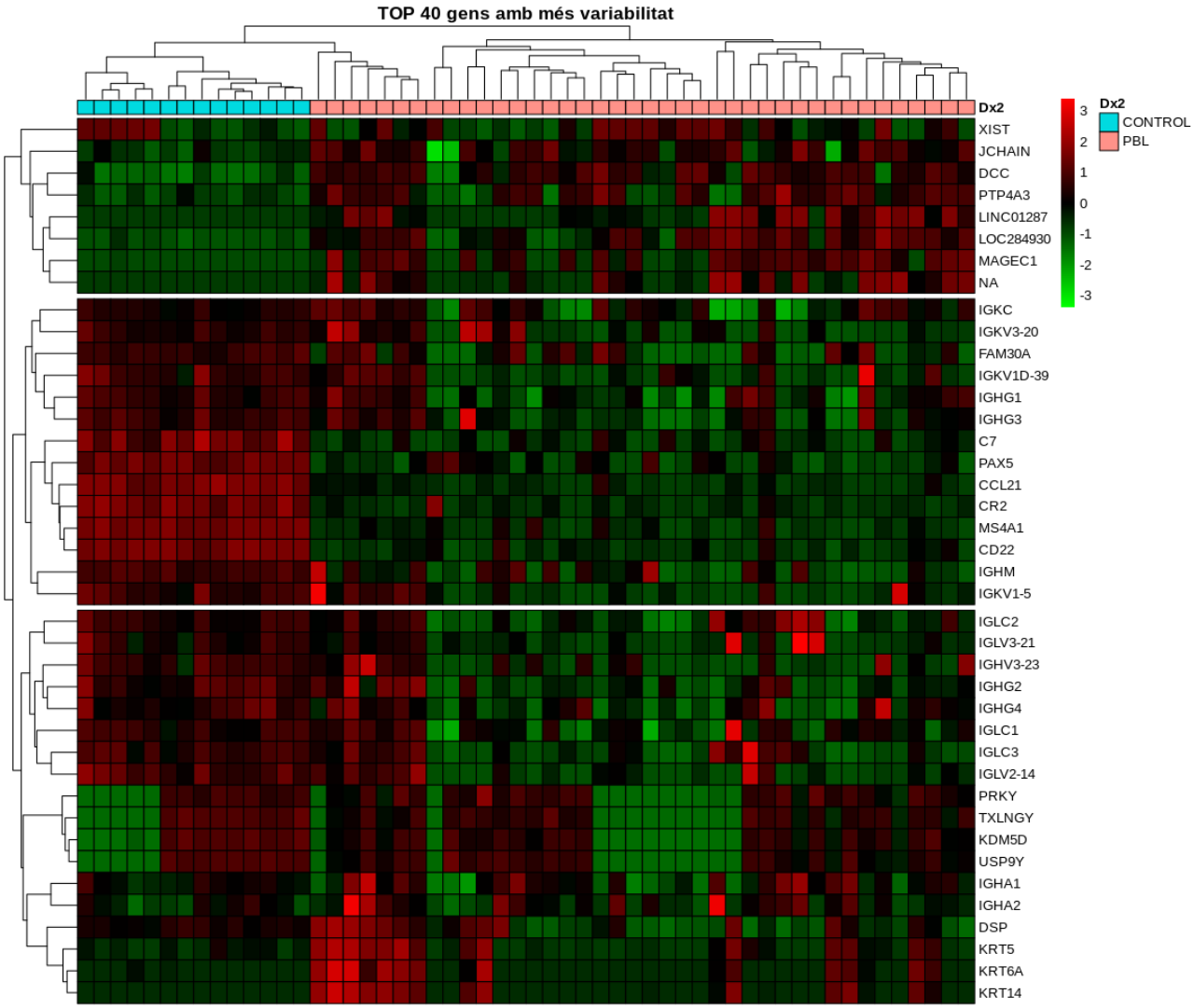
En segon lloc, s'ha representat la distribució de l'expressió diferencial de les mostres amb diferents *heatmap*. En la següent imatge es representen els nivells d'expressió normalitzats (VST) dels 1739 gens que s'han trobat diferencialment expressats.



Imatge 17. Heatmap de l'expressió normalitzada (VST) escalada (Z-*score)* dels 1739 gens més DE entre LPB i Controls. L'eix de la Y presenta els gens i l'eix de les X, les mostres. Les mostres marcades en turquesa són els pacients Controls. La resta, marcades amb color rosat, són els pacients amb LPB. L'expressió es representa amb la intensitat del color. Vermell pels gens més sobreexpressats, i verd pels gens més inhibits.

Els resultats de la imatge anterior demostren la clara diferència en l'expressió genètica entre el grup LPB i el grup Control d'aquest conjunt de gens. Tal com s'observava en el *volcanoplot*, s'identifica que els gens més sobreexpressats en LPB, formen part dels gens amb més nivell d'expressió (el grup LPB té les zones amb el color vermell més intens). A més a més, també es confirma el que s'ha vist en els PCAs (Imatge 11), és dir, s'observa un patró homogeni d'expressió en les mostres Controls, però en canvi, les mostres del grup LPB segueixen un patró d'expressió més heterogeni, especialment si ens fixem en el grup de gens sobreexpressat en LPB.

En tercer lloc, per tal d'aprofundir en l'estudi de l'heterogeneitat observada anteriorment, en l'expressió dels gens del grup LPB, s'ha representat l'expressió normalitzada (VST) dels 40 gens amb una variabilitat més elevada, és a dir, els gens que tenen la variança més elevada entre les mostres.



Imatge 18. Heatmap de l'expressió normalitzada (VST) escalada (Z-*score)* dels 40 gens amb més variabilitat. L'eix de la Y presenta els gens i l'eix de les X, les mostres. Les mostres marcades en turquesa són els pacients Controls. La resta, marcades amb color rosat, són els pacients amb LPB. L'expressió es representa amb la intensitat del color. Vermell pels gens més sobreexpressats, i verd pels gens més inhibits.

La interpretació d'aquest *heatmap* confirma que el grup LPB presenta una gran heterogeneïtat en l'expressió dels gens més variables, mentre que el grup de Controls manté una expressió més constant. Un exemple és el cas del gen KRT14 (l'últim gen representat en la imatge 18), mentre l'expressió d'aquest gen és baixa en totes les mostres del grup Controls, en el grup LPB es troba expressat amb una intensitat alta en 13 mostres, amb una intensitat baixa en 25 mostres i amb una intensitat mitjana en 3 mostres. També s'ha observat que una gran part dels gens amb més variabilitat estan relacionats amb la regulació de les Immunoglobulines.

La representació gràfica de l'anàlisi diferencial ha generat els següents punts a destacar:

- S'han trobat 1739 gens diferencialment expressats amb un L2FC superior a 2 i un p.Adj inferior a 0.001, dels quals 830 es troben sobreexpressats i 909 es troben inhibits en el grup LPB.
- Els gens amb un L2FC més extrem es troben sobreexpressats en els pacients LPB.
- El grup LPB mostra una variabilitat en l'expressió gènica heterogènia, mentre que el grup Control té una variabilitat més homogènia.
- Degut a la variabilitat genètica detectada en el grup LPB és important estudiar el possible efecte dels gens que presenten la variança més elevada entre les mostres. S'ha detectat que un elevat nombre dels gens amb més variabilitat estan relacionats en la regulació i el funcionament de les Immunoglobulines.

4.7. Significat Biològic dels resultats.

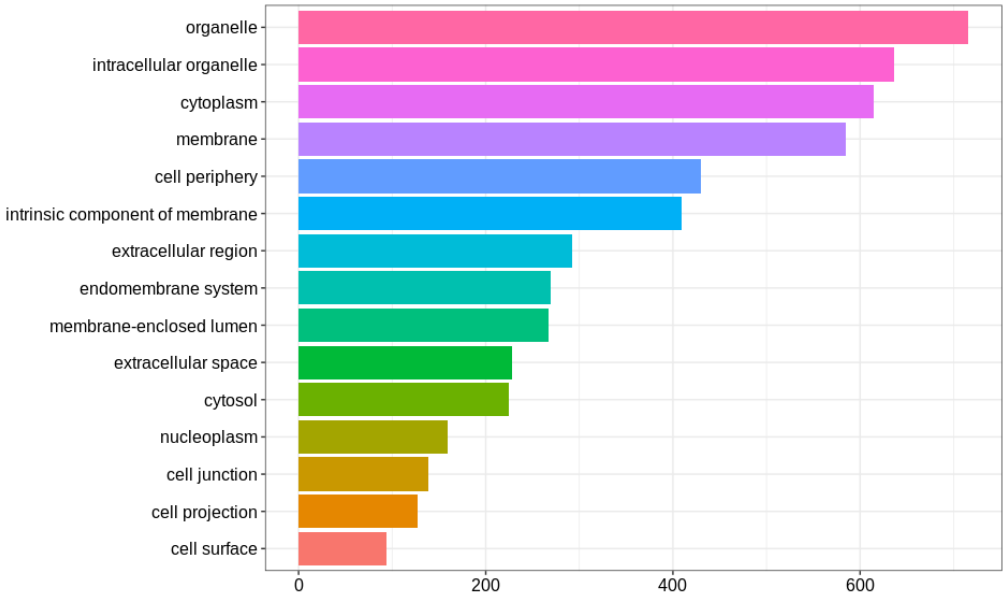
L'anàlisi de significat biològic dels resultats s'ha realitzat amb l'objectiu de trobar les funcions, els processos biològics i les vies moleculars que més apareixen en les llistes de gens generades en l'apartat anterior. Les anàlisis d'aquest tipus necessiten un nombre mínim de gens per ser fiables, de manera que és habitual realitzar una selecció menys restrictiva que en els passos anteriors, però en aquest cas s'ha utilitzat un p.Adj de 0.001 i un valor absolut de L2FC superior a 2 en l'anàlisi de classificació GO, en el GO *over-representationtest* i en el KEGG *over-representation test*. En canvi, ha estat necessari utilitzar una llista de gens més àmplia en el *GO Gene Set Enrichment Analysis* i el KEGG *Gene Set Enrichment Analysi*, per tant, s'han agafat aquells gens amb un valor absolut de L2FC superior a 2 i un p.Adj inferior a 0.05

El paquet de *Bioconductor* utilitzat ha estat *clusterProfiler.* Aquesta llibreria d'R implementa mètodes per analitzar i visualitzar perfils funcionals de coordenades genòmiques, grups de gens i gens individuals. És un paquet de R / Bioconductor que automatitza el procés de classificació de termes biològics i l'anàlisi d'enriquiment de grups de gens (Yu G. et al, 2012). Admet tres espècies, inclosos humans, ratolins i llevat, encara que els mètodes proporcionats poden estendre's fàcilment a altres espècies i ontologies. Incorpora el coneixement biològic (GO) i l'Enciclopèdia de Gens i Genomes (KEGG).

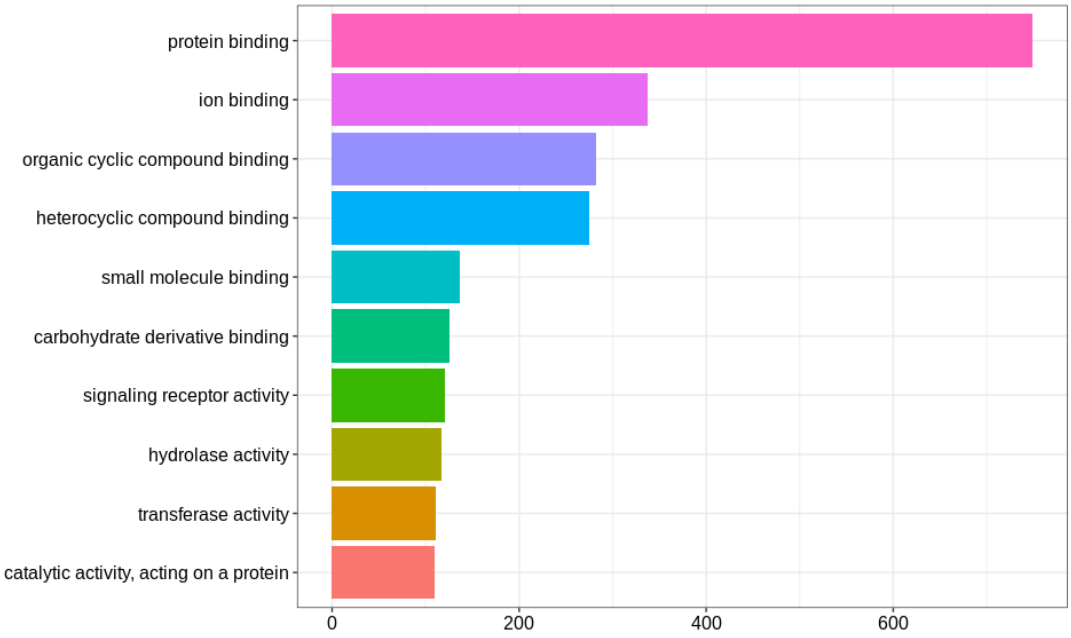
L'anàlisi GO s'ha realitzat dels gens (activats i inhibits de forma conjunta) diferencialment expressats sobre els termes: components cel·lulars (CC), procés biològic (BP) i de les funcions moleculars (MF).

4.7.1. Classificació GO

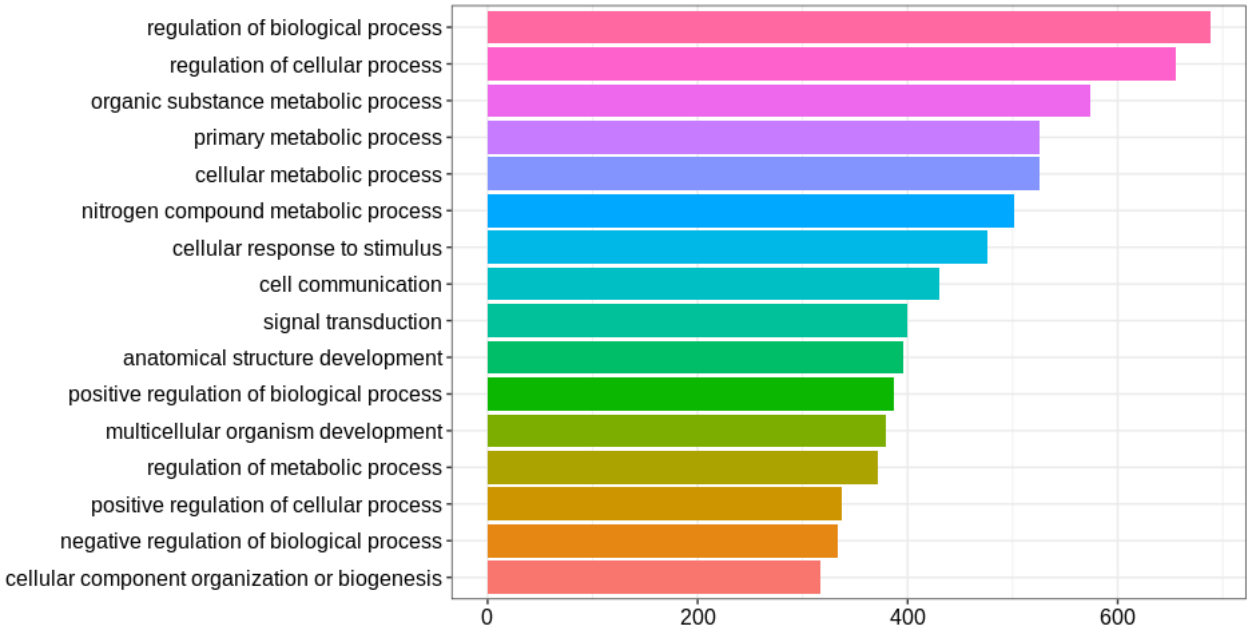
Es va cercar la classificació genètica basada en la base de dades GO per als gens diferencialment expressats que s'han obtingut. Les següents visualitzacions mostren els resultats del perfil funcional al nivell 3 de GO amb la funció *groupGO* per a les ontologies CC, BP i MF es representen amb els següents barplots :



Imatge 19. Nombre de gens DE classificats en els diferents Components Celulars de GO.



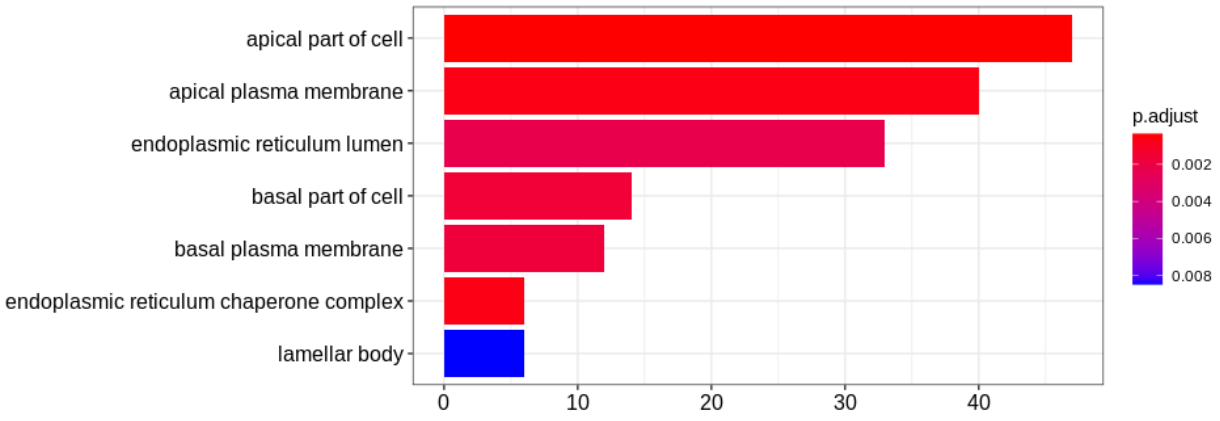
Imatge 20. Nombre de gens DE classificats en les diferents Funcions Moleculars de GO.



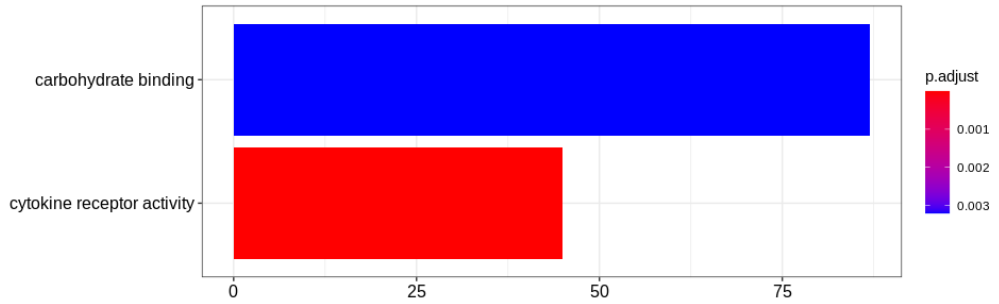
Imatge 21. Nombre de gens DE classificats en els diferents Processos Biològics de GO.

4.7.2. Prova de sobre representació GO

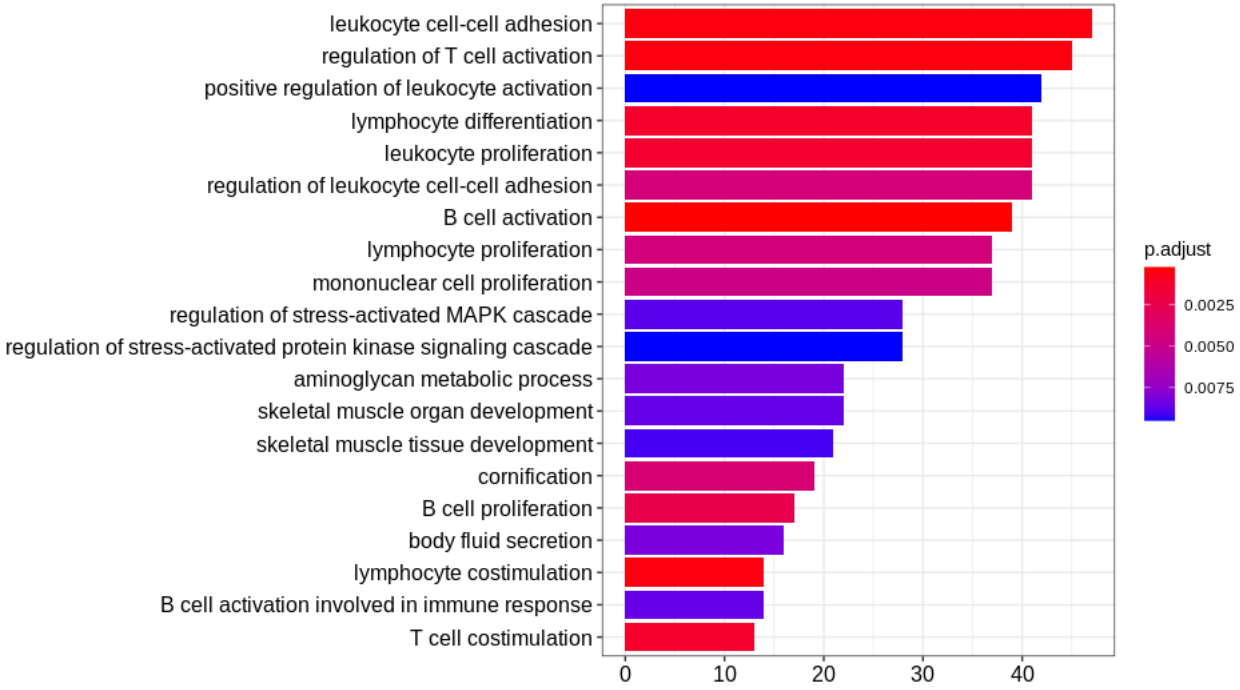
A diferencia de la Classificació GO, els análisis d'enriquiment implementen un model hipergeomètric per avaluar si el nombre de gens seleccionats associats a la classificació és més gran del que s'espera. Les visualitzacions del perfil funcional del conjunt de gens diferencialment expressats de GO amb la funció enrichGO retorna les categories GO després d'un control FDR, pValueCutoff = 0.01, per a les ontologies CC, BP i MF.



Imatge 22. Nombre de gens DE enriquits en els diferents Components Celulars de GO.



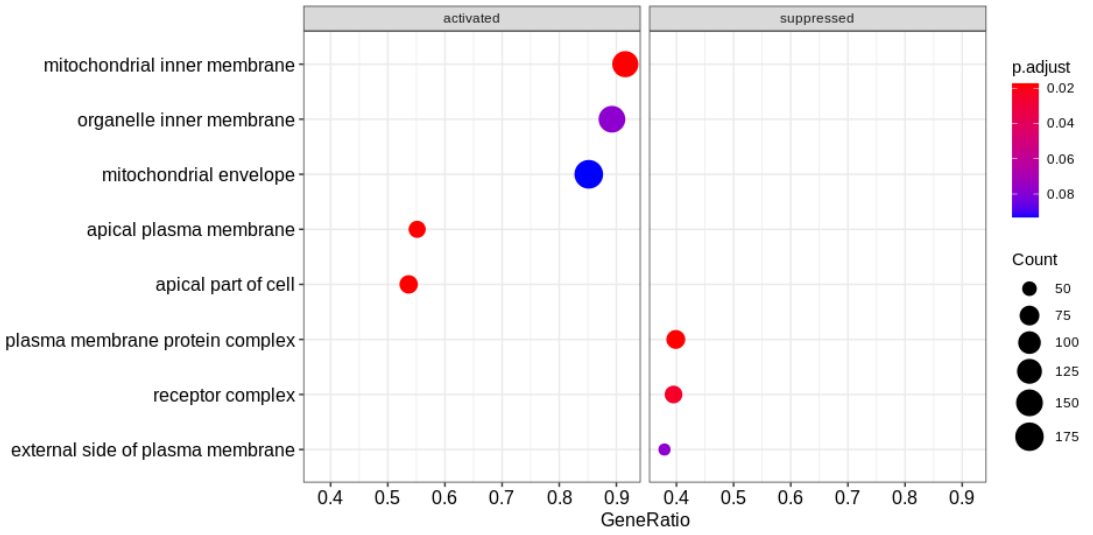
Imatge 23. Nombre de gens DE enriquits en les diferents Funcions Moleculars de GO. En aquest cas només s'ha obtingut resultat a partir dels gens amb un valor absolut de L2FC superior a 1 i un p.Adj inferior a 0.05



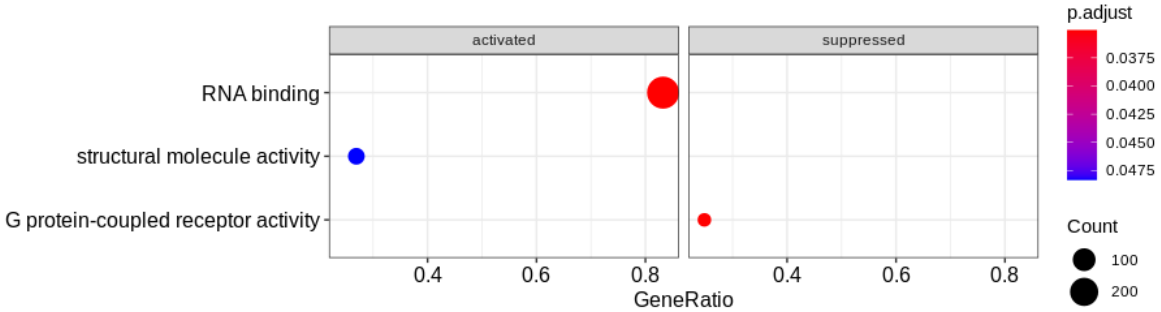
Imatge 24. Nombre de gens DE enriquits en els diferents Processos Biològics de GO.

4.7.3. Anàlisi d'enriquiment sobre el conjunt de gens GO

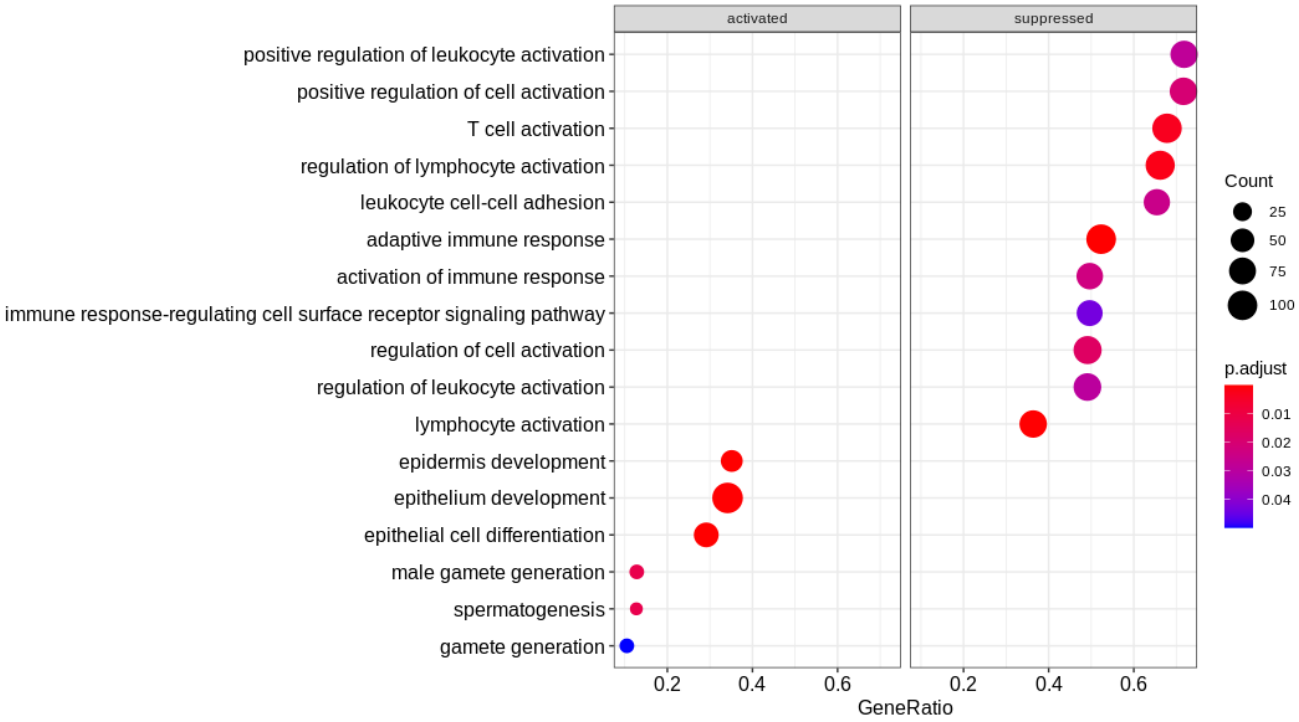
El test de sobre representació GO cerca els gens amb una diferència gran en l'expressió, però en canvi, no detecta una diferència petita en un conjunt coordinat de gens. L'objectiu de l'anàlisi d'enriquiment *GO* és suplir aquesta limitació. Les Visualitzacions del perfil funcional del conjunt de gens diferencialment expressats de GO amb la funció gseGO, per a les ontologies CC, BP només ha tingut resultat quan s'ha utilitzat una llista de gens amb un L2FC superior a 1 i un p.Adj inferior a 0.001.



Imatge 25. Nombre de gens DE conjuntament enriquits en els diferents Components Celulars de GO.



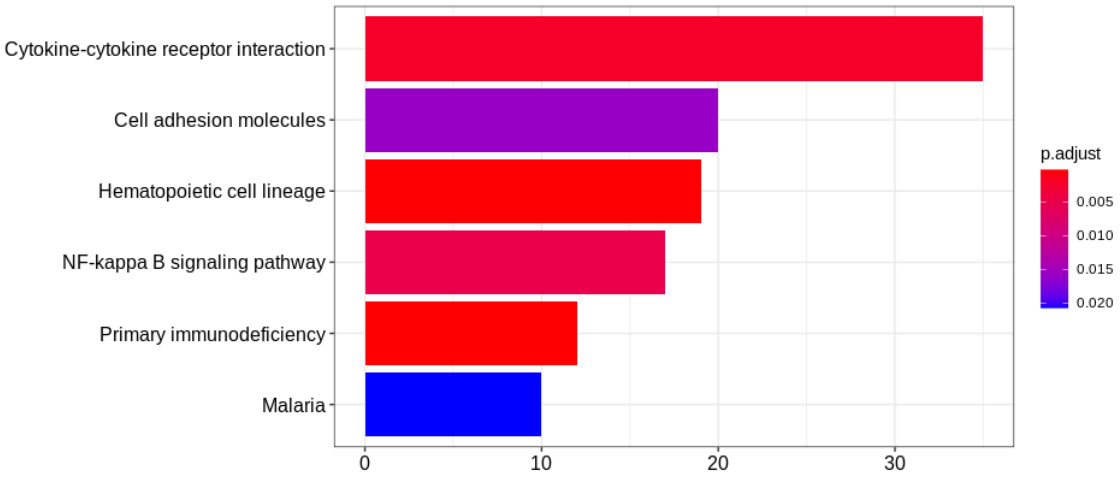
Imatge 26. Nombre de gens DE conjuntament enriquits en les diferents Funcions Moleculars de GO.



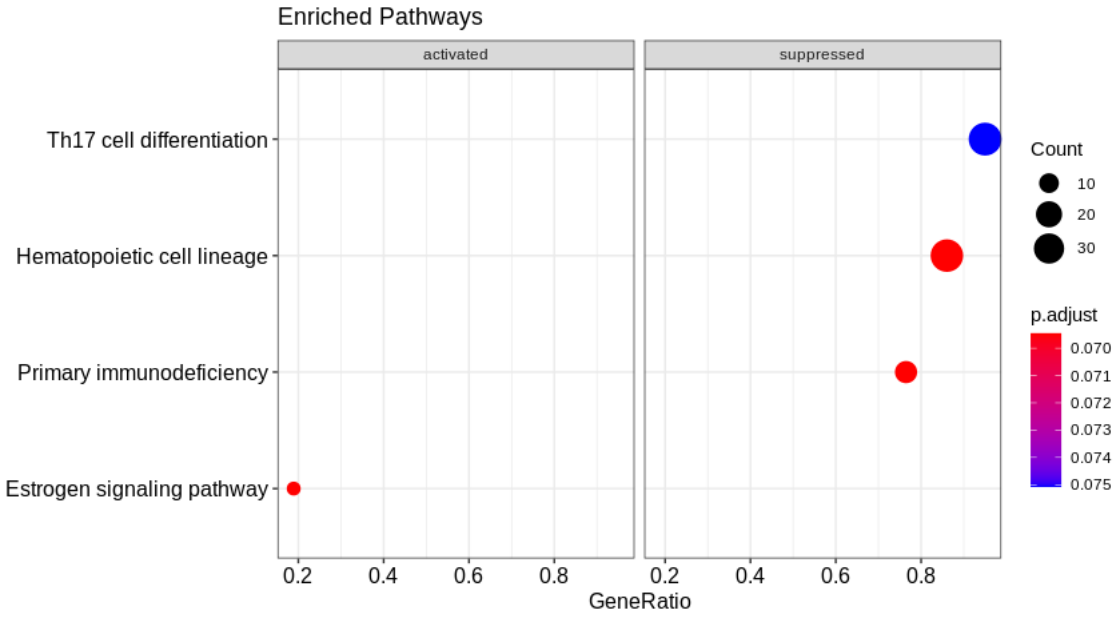
Imatge 27. Nombre de gens DE conjuntament enriquits en els diferents Processos Biològics de GO.

4.7.4. Prova de sobrerepresentació KEGG i anàlisi d'enriquiment sobre el conjunt de gens KEGG

Finalment s'ha realitzat l'anàlisi biològic utilitzant les annotacions de la base de dades KEGG. S'ha efectuat l'anàlisi d'enriquiment i el *gene set enrichment* KEGG dels gens diferencialment expressats (els sobreexpressats i dels infraexpressats de forma conjunta) amb la funció enrichKEGG i gseKEGG, aquestes funcions retornen les categories KEGG després d'un control FDR.



Imatge 28. Nombre de gens DE enriquits en la base de dades KEGG.



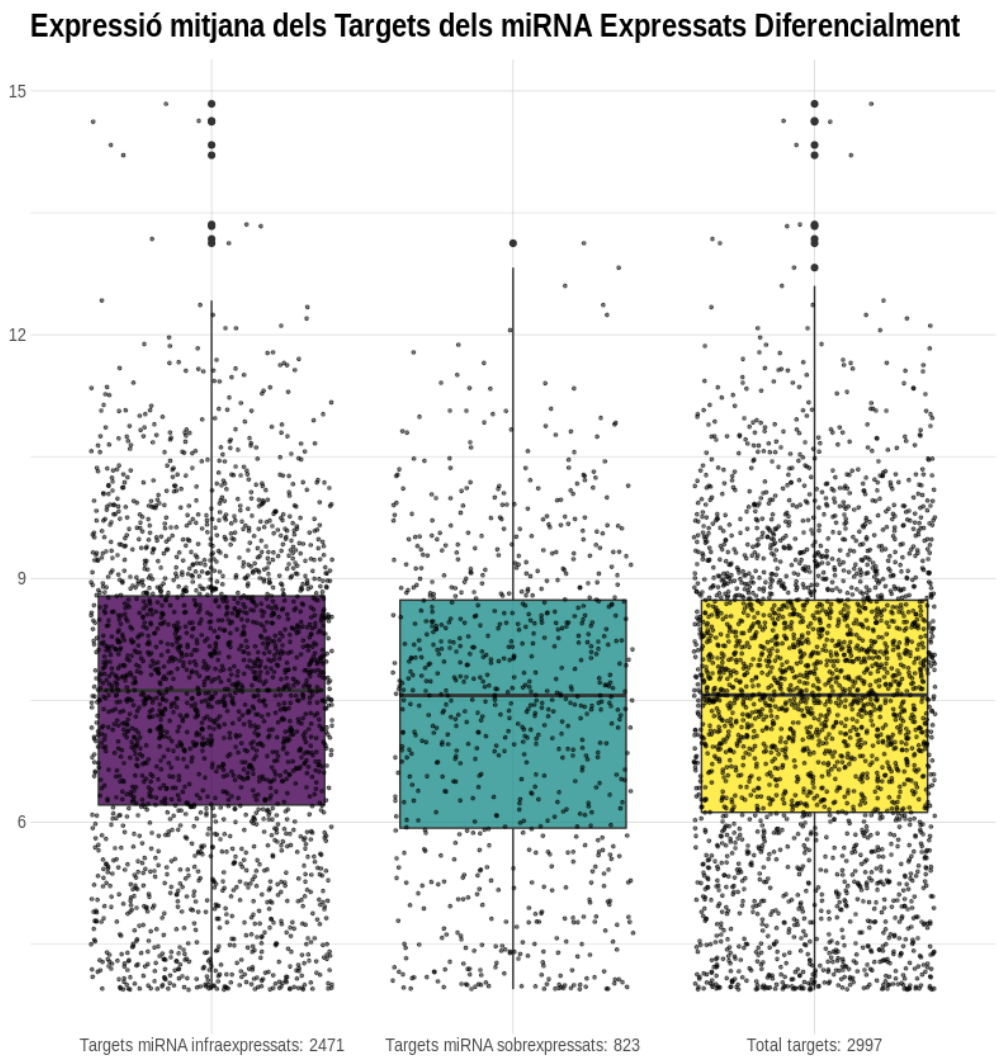
Imatge 29. Nombre de gens DE conjuntament enriquits en la base de dades KEGG.

4.8. Expressió de les dianes dels miRNA diferencialment expressats.

Per trobar els gens diana dels miRNAs diferencialment expressats en l'estudi de *microarrays* realitzat durant l'assignatura de pràctiques en empresa, s'ha utilitzat la base de dades proporcionada per *affymetrix.* Aquesta base de dades proporciona els gens diana (validats per ells mateixos en el 2016) dels miRNA coberts per les sondes que constitueixen el xip utilitzat (miRNA-4_0-st-v1). Només s'han utilitzat les dianes que han estat validades per *affymetrix* i s'han descartat les dianes anotades obtingudes mitjançant prediccions. Cal destacar que el microxip utilitzat en l'anàlisi de *microarrays* també conté sondes capaces de detectar altres molècules de tipus lncRNA, especialment *Small nucleolar* RNA (snoRNA). Els snoRNAs són RNAs no codificants que contribueixen a la biogènesi del ribosoma i a l'*splicing* de l'RNA modificant-los. A més, s'ha vist que s'expressen de manera específica durant les diferents fases del desenvolupament hematopoètic. En estudis recents, s'han identificat snoRNAs sobreexpressats en progenitors hematopoètics, i en canvi inhibits durant la diferenciació mieloide (Warnen et al. 2018). En l'anàlisi diferencial es van trobar 20 molècules que no eren miRNA, la majoria de les quals eren snoRNAs. Tot i que en el present estudi no les tindrem en compte, és interessant tenir en compte aquests resultats per futurs estudis.

D'entrada, van ser detectats un total de 59 molècules d'RNA diferencialment expressades entre el grup LPB i el grup Control sense LPB, de les quals només 39 corresponien a miRNA. Seguidament, es van cercar les dianes d'aquest miRNAs. Es van trobar un total de 4756 gens diana pels 39 miRNAs. Alguns miRNA presentaven dianes coincidents, per tant, es van trobar un total de 3162 gens diana únics. En aquest moment es va decidir realitzar diferents estratègies per visualitzar l'expressió d'aquests gens diana.

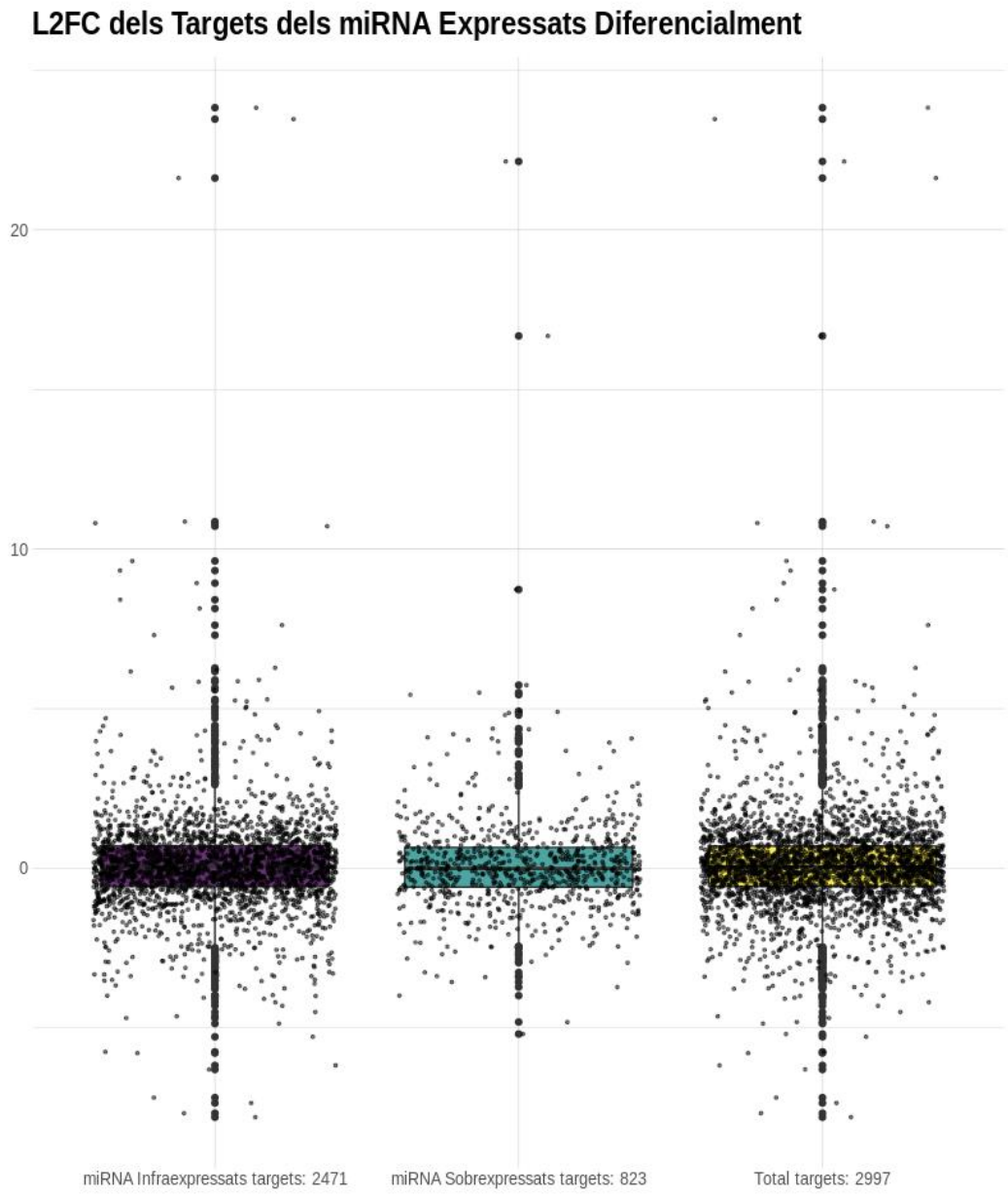
En primer lloc, es va representar l'expressió de totes les dianes de cada miRNA en les mostres LPB. Es va cercar l'expressió de les dades normalitzades (VST) de cada mostra LPB (excloent Controls) per tots els gens diana que l'anàlisi d'RNA-seq ha proporcionat un valor d'expressió. Es va trobar l'expressió de 2997 gens diana i es va representar la mitjana de l'expressió transformada de totes les mostres LPB (imatge 30). En el gràfic es representa l'expressió mitjana dels gens diana dels miRNA inhibits en LPB, l'expressió mitjana dels gens diana dels miRNA sobreexpressats en LPB i l'expressió mitjana dels gens diana de tots els miRNA.



Imatge 30. El boxplot lila representa l'expressió mitjana dels gens diana dels miRNA inhibits en LPB, en verd la dels miRNA sobreexpressats en LPB i en groc la de tots els miRNA.

No es van observar diferències significatives en el nivell d'expressió normalitzat (VST) entre els gens diana dels miRNA inhibits i els sobreexpressats.

D'altra banda es va voler representar el mateix concepte, però amb el L2FC resultant de l'anàlisi diferencial (imatge 31).

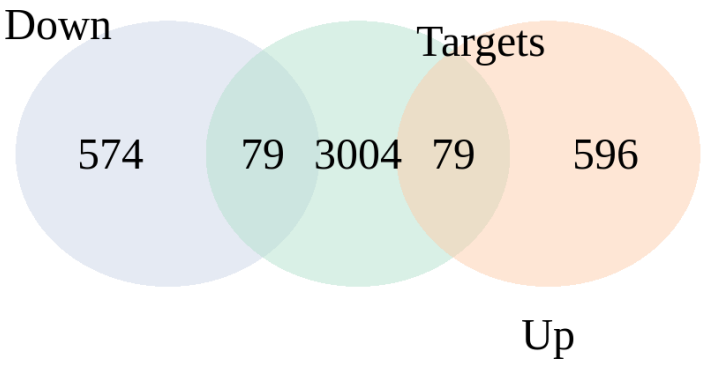


Imatge 31. El boxplot lila representa els valors de L2FC dels gens diana dels miRNA inhibits en LPB, en verd la dels miRNA sobreexpressats en LPB i en groc la de tots els miRNA.

No es van observar diferències significatives en l'expressió diferencial (L2FC resultant de la comparació LPB i Controls) entre els gens diana dels miRNA inhibits i els sobreexpressats.

En segon lloc, es va corelacionar l'expressió diferencial dels gens més diferencialment expressats amb l'expressió diferencial dels miRNA complementaris. Cal recordar que es van considerar miRNA DE els que tenien un valor absolut de L2FC superior a 1 i un p.Adj inferior a 0.001. I que s'han considerat gens DE expressats els que tenen un valor absolut de L2FC superior a 2 i un p.Adj inferior a 0.001.

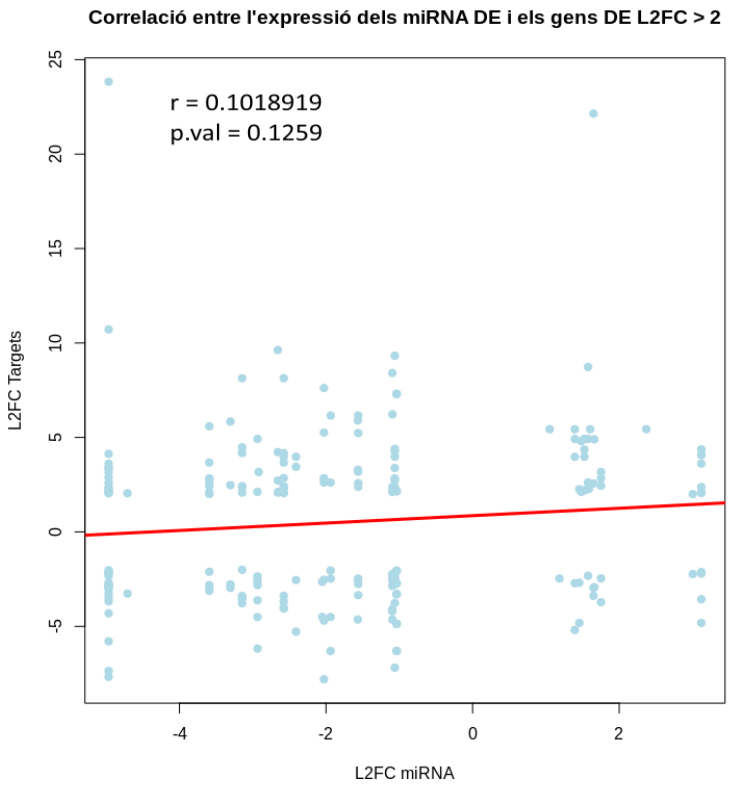
Es va trobar que 158 gens DE en l'anàlisi d'RNA-seq eren diana d'almenys un dels miRNA diferencialment expressat en l'anàlisi de *microarrays*, 79 d'aquests gens es troben sobreexpressats, mentre que els altres 79 es troben inhibits (imatge 32).



Imatge 32. El cercle morat etiquetat amb Down representa el nombre de gens DE inhibits en LPB, el cercle de la dreta etiquetat amb Up representa els gens DE sobreexpressats i el cercle verd situat al mig els gens diana dels miRNA DE.

El nombre de gens sobreexpressats i inhibits, tal i com s'observa en la imatge anterior, no coincideix amb el nombre obtingut en l'anàlisi diferencial degut a que alguns dels gens encara no tenen un *gene Symbol* associat. Les annotacions utilitzades per trobar els gens diana validats dels miRNA, només contenen l'etiqueta amb el gen Symbol,i per tant, no estan representats tots aquells gens que encara no existeix un gen *Symbol* associat.

A continuació, s'ha generat una taula (Taula 1)per estudiar si existeix una correlació entre els 39 miRNAs diferencialment expressats i les 158 dianes d'aquests miRNA obtinguts DE en l'anàlisi d'RNA-seq. La capacitat dels miRNA per adherir-se a més d'un gen ha generat una taula formada per 36 miRNA capaços d'unir-se a 227 gens DE, dels quals 158 són únics. La següent imatge mostra la correlació entre el L2FC dels miRNA i el L2FC de les dianes DE.



Imatge 33. Correlació entre el L2FC dels miRNA DE i el L2FC de les dianes DE.

Els resultats obtinguts mostren un coeficient de correlació de 0.10 i un p-valor de 0.12, per tant, s'observa una correlació positiva no significativa del 10% entre el L2FC dels miRNA DE i els gens diana DE. Tot i que aquests resultats no són significatius, no es descarta que si s'hagués trobat un nombre més elevat de coincidències entre dianes de miRNAs DE i gens DE (especialment entre les dianes de miRNAs sobreexpressats en LPB), s'hagués vist una correlació positiva més elevada.

5. Conclusions.

S'observen diferències transcriptòmiques significatives entre el grup de pacients LPB i el grup Controls sense LPB:

- Es detecten un total de 1739 gens diferencialment expressats amb un valor absolut de L2FC superior a 2 i un p.Adj inferior a 0.001, dels quals 830 es troben sobreexpressats i 909 es troben inhibits en el grup LPB.
- Els gens sobreexpressats en el grup LPB presenten els valors absoluts de L2FC més alts.
- El grup LPB mostra una variabilitat en l'expressió gènica molt heterogènia en comparació amb el grup Control.
- Entre els 40 gens amb més variabilitat, gairebé el 50% estan relacionats en la regulació i el funcionament de les Immunoglobulines.
- L'anàlisi de significat biològic demostra que els gens diferencialment expressats participen en vies importants relacionades amb l'activació de sistema immune, la diferenciació limfocítica i la proliferació cel·lular, entre d'altres.

L'expressió dels gens diana dels miRNA DE està condicionada per diversos factors que enmascaren el paper dels miRNA en la modulació de l'expressió:

- Es detecten 4756 gens dianes de 39 miRNAs diferencialment expressats.
- Es detecta una correlació possitiva no significativa i del 10% entre l'expressió dels miRNA DE i l'expressió dels gens diana DE.
- No es detecten diferències entre els gens diana dels miRNA sobreexpressats i els gens diana dels miRNA inhibits en la mitjana d'expressió del grup LPB.
- No es detecten diferències entre els gens diana dels miRNA sobreexpressats i els gens diana dels miRNA inhibits en el L2FC.

Existeixen diferents línies de treball futures que es podrien realitzar:

- Millorar la cerca dels gens diana, utilitzant bases de dades de miRNAs curades i de qualitat, per tal de comparar els resultats amb les anotacions utilitzades i valorar analitzar un nou estudi de correlació.
- S'ha vist que més del 50% de les lectures seqüenciades són regions d'origen intrònic que segurament pertanyen a lncRNAs i *small* RNAs. A més dels miRNA DE, també s'han trobat altres tipus de lncRNA en l'anàlisi de *microarrays*. Sería interessant utilitzar aquesta informació per estudiar l'expressió d'aquests altres tipus d'RNA en aquests pacients.

- Millorar l'anàlisi de Significat biològic realitzant una cerca detallada de les vies i els gens DE que hi participen. A més seria interessant no enfocar-se només en els gens DE, sinó també en el significat biològic dels gens que presenten la variància més elevada.

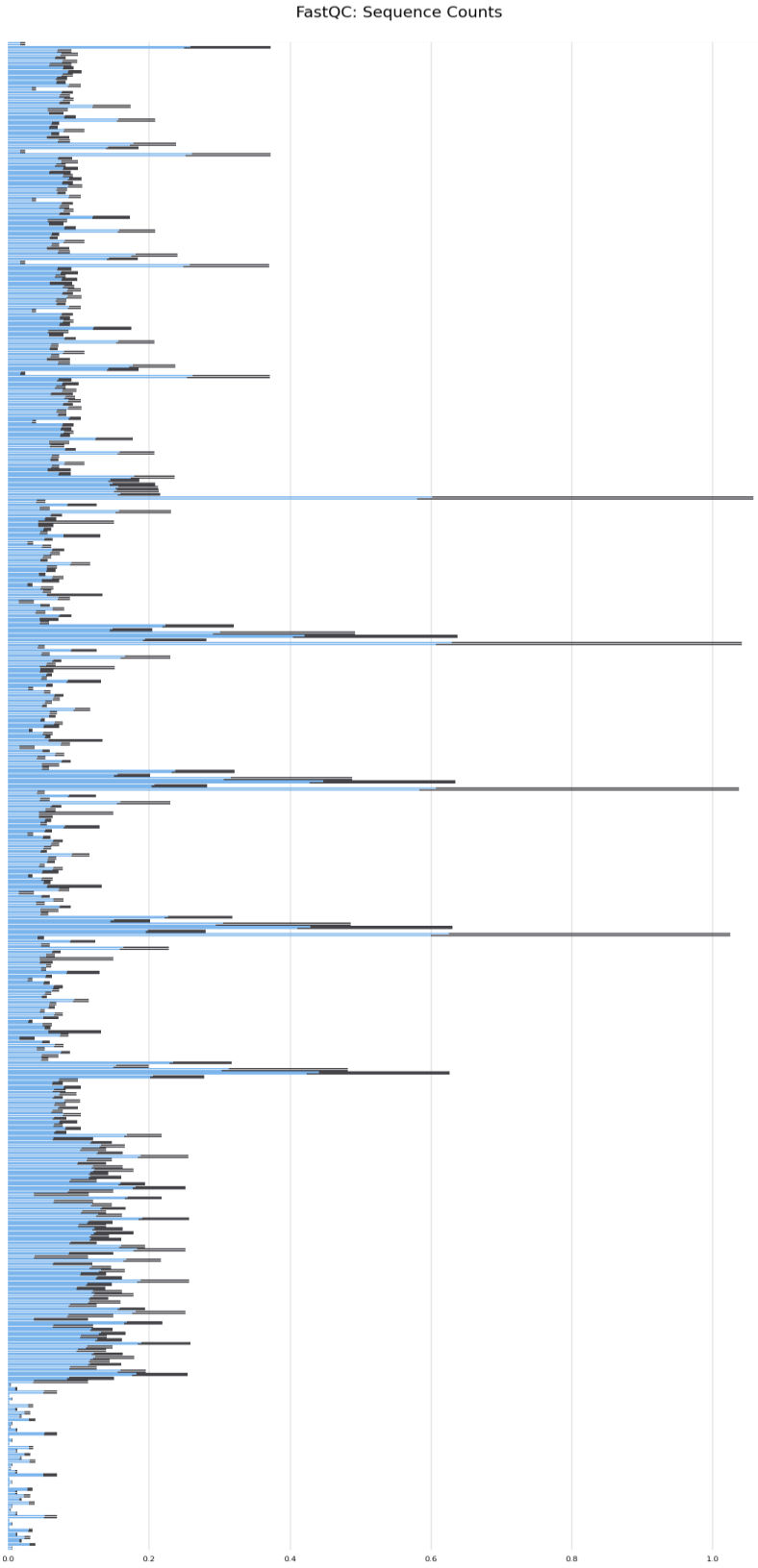
Sobre el seguiment de la planificació al llarg del projecte, cal destacar les adversitats que han anat sorgint durant l'acompliment de les fites i els terminis plantejats. El repte principal ha estat l'aplicació dels conceptes teòrics apresos a un estudi real que té algunes particularitats com una alta demanda de requisits computacionals. Aquests requisits alenteixen la manipulació de les dades i allarguen les diferents etapes dels processos. Un altre problema derivat de realitzar un estudi amb mostres reals ha estat la cerca dels motius i de les solucions en el moment en que es detectaven discordances o s'obtenien resultats no esperats duran l'execució de les diferents etapes. Aquesta última dificultat ha estat especialment present en l'avaluació dels fragments alineats sobre el transcriptoma. Per últim, cal destacar que aquests contratemps han permès la millorar d'aptituds i capacitats, tan a nivell tècnic i bioinformàtic gràcies al nombre d'eines i processos no previstos que s'han hagut de realitzar, com a nivell personal alhora de guanyar autonomia en la presa de decisions i en la cerca de les solucions als diferents obstacles que han anat apareixent.

6. Bibliografia

- Anders S, Hube, W. Differential expression analysis for sequence count data. Genome Biol 11, R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106.
- Andrews S (2010) FastQC: A quality control tool for high throughput sequence data.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
- Bray N.L. et al. . (2016) Near-optimal probabilistic RNA-seq quantification. Nature Biotech, 34 (5), 525 – 527.
- Castillo JJ, Bibas M, Miranda RN. The biology and treatment of plasmablastic lymphoma. Blood 2015;125:2323-2330.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
- García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012 Oct 15;28(20):2678-9. doi: 10.1093/bioinformatics/bts503. Epub 2012 Aug 22. PMID: 22914218.
- ICO guia Limfoma B Difús de Cèl·lula Gran, Març 2019.
- Mattick JS, Gagen MJ, The Evolution of Controlled Multitasked Gene Networks: The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms, Molecular Biology and Evolution, Volume 18, Issue 9, September 2001, Pages 1611–1630, https://doi.org/10.1093/oxfordjournals.molbev.a003951.
- Love MI, Soneson C, Hickey PF, Johnson LK, Pierce NT, Shepherd L, Morgan M, and Patro R. 2020. “Tximeta: Reference sequence checksums for provenance identification inRNA-seq”. Computational Biology. https://doi.org/10.1371/journal.pcbi.1007664.
- MacManes MD, On the optimal trimming of high-throughput mRNA sequence data. Front. Genet. (2014) 5:13. doi: 10.3389/fgene.2014.00013.
- Nakazato T, Hidemasa TO, Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read ArchiveOctober 22, 2013, https://doi.org/10.1371/journal.pone.0077910.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-419. doi:10.1038/nmeth.4197.
- Patro R, Duggal G, Kingsford C, Salmon: accurate, versatile and ultrafast quantification from RNA-Seq data using lightweight-alignment. (2015), bioRxiv, 9, 021592.

- Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol 16, 177 (2015). https://doi.org/10.1186/s13059-015-0734-x.
- Srivastava A, Malik L, Sarkar H. et al. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol 21, 239 (2020). https://doi.org/10.1186/s13059-020-02151-8
- Srivastava A, Sarkar H, Gupta N, Patro R, RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes, Bioinformatics, Volume 32, Issue 12, 15 June 2016, Pages i192–i200, https://doi.org/10.1093/bioinformatics/btw277.
- Soneson, Charlotte, Michael I. Love, and Mark Robinson. 2015. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Research 4 (1521). https://doi.org/10.12688/f1000research.7563.1.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
- Team RC, R: A Language and ENviroment for Statical Computing. R Foundation for Statistical Computing. R Foundation for Statical Computing Vienna, Austria 2019. https://www.r-project.org/.
- Warner WA, Spencer DH, Trissal M, et al. Expression profiling of snoRNAs in normal hematopoiesis and AML. Blood Adv. 2018, 2(2):151-163. doi:10.1182/bloodadvances.2017006668.
- Williams CR, Baccarella A, Parrish, JZ et al. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17, 103 (2016). https://doi.org/10.1186/s12859-016-0956-2.
- Yu G, Wang LG, Y Han, QY He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012, 16(5):284-287. doi:[10.1089/omi.2011.0118](http://dx.doi.org/10.1089/omi.2011.0118).
- Zhao, S, Zhang, Y, Gamini R. et al. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus depletion. Sci Rep 8, 4781 (2018). https://doi.org/10.1038/s41598-018-23226-4.

7. Material Suplementari.



Imatge 34. Proporció del Nombre de lectures duplicades. En blau es representen les lectures úniques i en negre les lectures duplicades per cada mostra.

miRNA	logFC	SYMBOL	log2FoldChange
hsa-miR-100-5p	-2,657578575	GRHL1	2,72054138
hsa-miR-100-5p	-2,657578575	LIN28B	9,627145855
hsa-miR-100-5p	-2,657578575	SMPDL3B	4,224468539
hsa-miR-100-5p	-2,657578575	RRM2	2,082935468
hsa-miR-100-5p	-2,657578575	COX2	2,113930778
hsa-miR-1273a	-2,05220552	CD209	-2,641693701
hsa-miR-1273a	-2,05220552	TNFRSF13C	-4,504296406
hsa-miR-140-3p	-2,576951768	TTC39C	-4,000919602
hsa-miR-140-3p	-2,576951768	CHRDL1	-3,3805302
hsa-miR-140-3p	-2,576951768	IL17REL	-4,056102128
hsa-miR-140-3p	-2,576951768	SLC23A1	2,853334529
hsa-miR-140-3p	-2,576951768	SLC7A5	3,976806223
hsa-miR-140-3p	-2,576951768	SLC30A3	-3,686537414
hsa-miR-140-3p	-2,576951768	PSAT1	3,670273257
hsa-miR-140-3p	-2,576951768	AHCY	2,284929151
hsa-miR-140-3p	-2,576951768	UCK2	2,367766505
hsa-miR-140-3p	-2,576951768	TTLL12	2,063237336
hsa-miR-140-3p	-2,576951768	LRPAP1	2,422661107
hsa-miR-140-3p	-2,576951768	TRIM31	8,136189525
hsa-miR-140-3p	-2,576951768	EBPL	2,058094425
hsa-miR-140-3p	-2,576951768	STC2	4,174829055
hsa-miR-148a-3p	3,120705613	GPRC5A	4,101993518
hsa-miR-148a-3p	3,120705613	SESN3	-3,564065338
hsa-miR-148a-3p	3,120705613	KIAA1549	3,614925637
hsa-miR-148a-3p	3,120705613	HSP90B1	2,066239744
hsa-miR-148a-3p	3,120705613	ARID3A	2,08213965
hsa-miR-148a-3p	3,120705613	CYCS	2,373514299
hsa-miR-148a-3p	3,120705613	GNB5	-2,123750721
hsa-miR-148a-3p	3,120705613	DNMT3B	4,374500133
hsa-miR-148a-3p	3,120705613	S1PR1	-2,199300004
hsa-miR-148a-3p	3,120705613	MYC	4,063836002
hsa-miR-148a-3p	3,120705613	NPTX1	-4,817148606
hsa-miR-150-5p	-4,966516091	A1CF	23,82722718
hsa-miR-150-5p	-4,966516091	KIAA1549	3,614925637
hsa-miR-150-5p	-4,966516091	ACSL6	-3,050859715
hsa-miR-150-5p	-4,966516091	SLC1A5	2,893386122
hsa-miR-150-5p	-4,966516091	BMP8B	4,134088382
hsa-miR-150-5p	-4,966516091	C11orf1	2,262329226
hsa-miR-150-5p	-4,966516091	REL	-2,766667854
hsa-miR-150-5p	-4,966516091	ANO7	3,350949749
hsa-miR-150-5p	-4,966516091	MUC4	10,71677283
hsa-miR-150-5p	-4,966516091	WDFY2	-2,077823624
hsa-miR-150-5p	-4,966516091	BCAS4	-3,491607505
hsa-miR-150-5p	-4,966516091	NPHS1	-7,673923855
hsa-miR-150-5p	-4,966516091	BCL11B	-2,773531091
hsa-miR-150-5p	-4,966516091	CCNE1	2,049385297
hsa-miR-150-5p	-4,966516091	QSOX1	2,18412302
hsa-miR-150-5p	-4,966516091	CCR6	-5,790202534
hsa-miR-150-5p	-4,966516091	SYNPO2	-2,326680459

miRNA	logFC	SYMBOL	log2FoldChange
hsa-miR-150-5p	-4,966516091	TMOD2	-2,686841047
hsa-miR-150-5p	-4,966516091	NR2F2	-2,02179202
hsa-miR-150-5p	-4,966516091	CAMK4	-2,907983494
hsa-miR-150-5p	-4,966516091	MANEAL	3,382372171
hsa-miR-150-5p	-4,966516091	CYCS	2,373514299
hsa-miR-150-5p	-4,966516091	FAM153B	-4,308525873
hsa-miR-150-5p	-4,966516091	GPR182	-7,355394259
hsa-miR-150-5p	-4,966516091	SLC7A11	2,615608367
hsa-miR-150-5p	-4,966516091	S1PR1	-2,199300004
hsa-miR-150-5p	-4,966516091	TTLL12	2,063237336
hsa-miR-150-5p	-4,966516091	HSPA4L	3,145028398
hsa-miR-150-5p	-4,966516091	S1PR3	-2,212479783
hsa-miR-150-5p	-4,966516091	FLT3	-2,858463277
hsa-miR-150-5p	-4,966516091	ZNF665	-2,906020704
hsa-miR-150-5p	-4,966516091	NKD1	-3,308758996
hsa-miR-150-5p	-4,966516091	PAK3	-3,671807736
hsa-miR-150-5p	-4,966516091	TXK	-2,899231994
hsa-miR-150-5p	-4,966516091	PDIA6	2,277046938
hsa-miR-150-5p	-4,966516091	GNB5	-2,123750721
hsa-miR-150-5p	-4,966516091	VEGFA	2,116028368
hsa-miR-151b	-1,936195436	BCL7A	-2,469141861
hsa-miR-151b	-1,936195436	KRT80	6,161916466
hsa-miR-1972	-3,304668308	IKZF2	-2,774396317
hsa-miR-1972	-3,304668308	DNAH3	5,841566352
hsa-miR-1972	-3,304668308	CPE	-2,960564161
hsa-miR-1972	-3,304668308	SLC16A1	2,47589307
hsa-miR-28-5p	-2,409608545	PIANP	-5,278701696
hsa-miR-28-5p	-2,409608545	SLC7A5	3,976806223
hsa-miR-28-5p	-2,409608545	CCND3	-2,545532099
hsa-miR-28-5p	-2,409608545	RAB36	3,447786762
hsa-miR-297	-1,095333409	SATB1	-2,228607164
hsa-miR-297	-1,095333409	TRAF5	-2,538722527
hsa-miR-297	-1,095333409	CLEC2D	-2,859349132
hsa-miR-297	-1,095333409	CERCAM	2,380247489
hsa-miR-297	-1,095333409	AKT3	-2,464345158
hsa-miR-297	-1,095333409	SIGMAR1	2,332371404
hsa-miR-297	-1,095333409	ENTHD1	-4,635788423
hsa-miR-297	-1,095333409	ZNF860	-4,19887784
hsa-miR-297	-1,095333409	PROX1	-4,089490479
hsa-miR-297	-1,095333409	VEGFA	2,116028368
hsa-miR-297	-1,095333409	GXYLT2	-2,321920561
hsa-miR-297	-1,095333409	HHIP	6,221388802
hsa-miR-297	-1,095333409	WDR72	8,412833833
hsa-miR-3178	1,587285165	SSR3	2,283870403
hsa-miR-3180-3p	2,369481918	DPF1	5,435856846
hsa-miR-3185	1,649502878	DYRK3	2,563878641
hsa-miR-3185	1,649502878	CHRDL1	-3,3805302
hsa-miR-3185	1,649502878	CPE	-2,960564161
hsa-miR-3185	1,649502878	SERPINB5	22,14247653
hsa-miR-3196	1,051430951	DPF1	5,435856846

Taula 1. Continuació de la Taula de Correlació entre els gens DE i les dianes DE

miRNA	logFC	SYMBOL	log2FoldChange
hsa-miR-342-3p	-3,592323177	TXNDC5	2,835806207
hsa-miR-342-3p	-3,592323177	SELPLG	2,594979428
hsa-miR-342-3p	-3,592323177	CD3D	-3,117932376
hsa-miR-342-3p	-3,592323177	PVRIG	-2,873298352
hsa-miR-342-3p	-3,592323177	RRM2	2,082935468
hsa-miR-342-3p	-3,592323177	C2orf72	5,590253235
hsa-miR-342-3p	-3,592323177	KLF8	-2,108334205
hsa-miR-342-3p	-3,592323177	PSAT1	3,670273257
hsa-miR-342-3p	-3,592323177	ZNF610	-2,815451474
hsa-miR-342-3p	-3,592323177	LRPAP1	2,422661107
hsa-miR-342-3p	-3,592323177	RRAGD	2,786943705
hsa-miR-342-3p	-3,592323177	SEPHS1	2,004121698
hsa-miR-342-5p	-4,709504986	VANGL2	-3,263085978
hsa-miR-342-5p	-4,709504986	CCNE1	2,049385297
hsa-miR-3609	-3,145132664	CERCAM	2,380247489
hsa-miR-3609	-3,145132664	LRPAP1	2,422661107
hsa-miR-3609	-3,145132664	PLS1	4,47982182
hsa-miR-3609	-3,145132664	KCNB1	-3,775946996
hsa-miR-3609	-3,145132664	RRM2	2,082935468
hsa-miR-3609	-3,145132664	RRAS2	-3,534708044
hsa-miR-3609	-3,145132664	WNK3	-3,495934346
hsa-miR-3609	-3,145132664	PXK	-2,005220621
hsa-miR-3609	-3,145132664	STC2	4,174829055
hsa-miR-3609	-3,145132664	TRIM31	8,136189525
hsa-miR-3609	-3,145132664	PRKCB	-3,385170382
hsa-miR-4417	3,004520616	SESN2	2,001857039
hsa-miR-4417	3,004520616	SATB1	-2,228607164
hsa-miR-4440	-2,918735209	NR6A1	3,171927961
hsa-miR-4488	1,18475644	BCL7A	-2,469141861
hsa-miR-4497	1,482444686	VEGFA	2,116028368
hsa-miR-4497	1,482444686	CAMK2N2	4,800150251
hsa-miR-4532	1,575565581	SELPLG	2,594979428
hsa-miR-4532	1,575565581	ZNF556	4,922588567
hsa-miR-4532	1,575565581	SLC7A11	2,615608367
hsa-miR-4532	1,575565581	LAX1	2,259255272
hsa-miR-4532	1,575565581	GXYLT2	-2,321920561
hsa-miR-4532	1,575565581	DSG3	8,73162189
hsa-miR-4634	1,661043643	SLC9A3	4,900882893
hsa-miR-4634	1,661043643	SNX22	-2,925957023
hsa-miR-4674	1,456280176	LAX1	2,259255272
hsa-miR-4674	1,456280176	NPTX1	-4,817148606
hsa-miR-4674	1,456280176	ZMAT1	-2,692065559
hsa-miR-4734	1,525186132	SLC7A5	3,976806223
hsa-miR-4734	1,525186132	ZNF556	4,922588567
hsa-miR-4734	1,525186132	ELL2	2,195128956
hsa-miR-4734	1,525186132	SCIN	4,357937725
hsa-miR-4793-3p	-2,935467698	ZNF556	4,922588567
hsa-miR-4793-3p	-2,935467698	SLC24A4	-2,352750691
hsa-miR-4793-3p	-2,935467698	KIF18B	2,125940565
hsa-miR-4793-3p	-2,935467698	CD209	-2,641693701

miRNA	logFC	SYMBOL	log2FoldChange
hsa-miR-4793-3p	-2,935467698	AFF3	-6,175758927
hsa-miR-4793-3p	-2,935467698	ZNF610	-2,815451474
hsa-miR-4793-3p	-2,935467698	TNFRSF13C	-4,504296406
hsa-miR-4793-3p	-2,935467698	SUSD5	-3,619144188
hsa-miR-4793-3p	-2,935467698	LPL	-2,464787881
hsa-miR-486-3p	-1,56128101	UST	-3,347822498
hsa-miR-486-3p	-1,56128101	BCL7A	-2,469141861
hsa-miR-486-3p	-1,56128101	BCL11A	-2,755495682
hsa-miR-486-3p	-1,56128101	ERBB2	2,581402898
hsa-miR-486-3p	-1,56128101	KRT80	6,161916466
hsa-miR-486-3p	-1,56128101	CASKIN1	5,228479632
hsa-miR-486-3p	-1,56128101	DNAJC6	3,19212167
hsa-miR-486-3p	-1,56128101	ZC3HAV1L	2,39071123
hsa-miR-500a-3p	-1,938477073	TNFRSF13C	-4,504296406
hsa-miR-500a-3p	-1,938477073	LRRC55	-6,303809056
hsa-miR-500a-3p	-1,938477073	SLC7A11	2,615608367
hsa-miR-500a-3p	-1,938477073	FYN	-2,045898676
hsa-miR-501-3p	-1,566150631	NPTXR	-2,547122714
hsa-miR-501-3p	-1,566150631	SLC2A12	3,292908723
hsa-miR-501-3p	-1,566150631	ENTHD1	-4,635788423
hsa-miR-501-3p	-1,566150631	CA12	5,901888075
hsa-miR-505-5p	-2,028594859	HOXB6	5,255161023
hsa-miR-505-5p	-2,028594859	TXNDC5	2,835806207
hsa-miR-505-5p	-2,028594859	SNX20	-2,527292572
hsa-miR-505-5p	-2,028594859	BEND4	-4,693087067
hsa-miR-505-5p	-2,028594859	PLD5	-7,798598254
hsa-miR-505-5p	-2,028594859	SLC7A11	2,615608367
hsa-miR-505-5p	-2,028594859	SATB2	2,83536587
hsa-miR-505-5p	-2,028594859	POU6F2	7,616358985
hsa-miR-5096	-1,061624346	POU3F2	9,325634353
hsa-miR-5096	-1,061624346	ANKRD44	-2,409224282
hsa-miR-5096	-1,061624346	C6orf132	4,286101252
hsa-miR-5096	-1,061624346	CYCS	2,373514299
hsa-miR-5096	-1,061624346	SLC7A5	3,976806223
hsa-miR-5096	-1,061624346	TEX9	-2,554994062
hsa-miR-5096	-1,061624346	QSOX1	2,18412302
hsa-miR-5096	-1,061624346	CDKL1	-2,458730122
hsa-miR-5096	-1,061624346	MANEAL	3,382372171
hsa-miR-5096	-1,061624346	SATB2	2,83536587
hsa-miR-5096	-1,061624346	KIR3DX1	4,394805079
hsa-miR-5096	-1,061624346	ZNF677	-2,555260427
hsa-miR-5096	-1,061624346	IKZF2	-2,774396317
hsa-miR-5096	-1,061624346	ABCA6	-2,220401425
hsa-miR-5096	-1,061624346	PTPRT	-3,76091789
hsa-miR-5096	-1,061624346	ART4	-7,184132891
hsa-miR-5096	-1,061624346	PDF	2,720419606
hsa-miR-548am-5p	-1,036711087	RAB8B	-2,057592986
hsa-miR-548am-5p	-1,036711087	SVOP	7,303500738
hsa-miR-548am-5p	-1,036711087	LRRC55	-6,303809056
hsa-miR-548am-5p	-1,036711087	ZNF154	-2,716317929

Taula 2. Continuació de la Taula de Correlació entre els gens DE i les dianes DE

miRNA	logFC	SYMBOL	log2FoldChange
hsa-miR-548am-5p	-1,036711087	PARP15	-3,295671315
hsa-miR-548am-5p	-1,036711087	NPM3	2,159687772
hsa-miR-548am-5p	-1,036711087	CLDN16	-4,86662312
hsa-miR-548c-5p	-1,036711087	RAB8B	-2,057592986
hsa-miR-548c-5p	-1,036711087	ZNF154	-2,716317929
hsa-miR-548c-5p	-1,036711087	PARP15	-3,295671315
hsa-miR-548c-5p	-1,036711087	SVOP	7,303500738
hsa-miR-548c-5p	-1,036711087	NPM3	2,159687772
hsa-miR-548c-5p	-1,036711087	LRRC55	-6,303809056
hsa-miR-548c-5p	-1,036711087	CLDN16	-4,86662312
hsa-miR-548o-5p	-1,036711087	ZNF154	-2,716317929
hsa-miR-548o-5p	-1,036711087	LRRC55	-6,303809056
hsa-miR-548o-5p	-1,036711087	SVOP	7,303500738
hsa-miR-548o-5p	-1,036711087	CLDN16	-4,86662312
hsa-miR-548o-5p	-1,036711087	RAB8B	-2,057592986
hsa-miR-548o-5p	-1,036711087	PARP15	-3,295671315
hsa-miR-548o-5p	-1,036711087	NPM3	2,159687772
hsa-miR-6126	1,752262555	CDKL1	-2,458730122
hsa-miR-6126	1,752262555	SATB2	2,83536587
hsa-miR-6126	1,752262555	NRARP	2,446958118
hsa-miR-6126	1,752262555	LIFR	-3,721657726
hsa-miR-6126	1,752262555	NR6A1	3,171927961
hsa-miR-663a	1,394605945	ZNF556	4,922588567
hsa-miR-663a	1,394605945	SLC1A2	-5,1925612
hsa-miR-663a	1,394605945	DPF1	5,435856846
hsa-miR-663a	1,394605945	ZNF154	-2,716317929
hsa-miR-663a	1,394605945	SLC7A5	3,976806223
hsa-miR-6816-5p	1,603881369	DPF1	5,435856846

Taula 3. Continuació de la Taula de Correlació entre els gens DE i les dianes DE

Taula 2. Suplementària amb el recompte i el percentatge de lectures alineades. La tercera i la quarta columna és l'alineament sobre el transcriptoma (Salmon) i les dues útilmes l'alineament sobre el genoma (Star)

Sample Name	5'-3' bias	M Aligned	% Aligned	M Aligned	% Aligned	M Aligned
H2VW3DSXY_1_14UDI-idt-UMI	1.02	38.0	33.8%	14.0	88.3%	36.6
H2VW3DSXY_3_14UDI-idt-UMI	1.02	37.0	33.8%	13.6	88.3%	35.6
H2VW3DSXY_2_14UDI-idt-UMI	1.08	36.9	33.8%	13.6	88.2%	35.5
H2VW3DSXY_4_14UDI-idt-UMI	1.03	36.4	33.7%	13.4	88.3%	35.0
H2VW3DSXY_1_49UDI-idt-UMI	1.02	20.7	30.5%	7.0	85.2%	19.5
H2V3WDSXY_2_13UDI-idt-UMI	1.03	20.4	30.6%	6.6	90.4%	19.6
H2VW3DSXY_3_49UDI-idt-UMI	1.05	20.3	30.5%	6.8	85.1%	19.1
H2V3WDSXY_1_13UDI-idt-UMI	1.08	20.2	30.6%	6.6	90.3%	19.5
H2VW3DSXY_2_49UDI-idt-UMI	1.05	20.2	30.5%	6.8	85.1%	19.0
H2VW3DSXY_4_49UDI-idt-UMI	1.06	20.1	30.5%	6.7	85.1%	18.8
H2V3WDSXY_3_13UDI-idt-UMI	1.10	20.0	30.7%	6.5	90.2%	19.2
H2V3WDSXY_4_13UDI-idt-UMI	1.09	19.8	30.7%	6.5	90.2%	19.0
H2VW3DSXY_1_61UDI-idt-UMI	0.97	16.7	17.6%	3.3	86.2%	16.1
H2VW3DSXY_2_61UDI-idt-UMI	0.93	16.4	17.6%	3.2	86.2%	15.8
H2VW3DSXY_3_61UDI-idt-UMI	0.98	16.4	17.6%	3.2	86.2%	15.8
H2VW3DSXY_4_61UDI-idt-UMI	0.97	16.2	17.6%	3.2	86.1%	15.6
H2V3WDSXY_2_38UDI-idt-UMI-b	1.06	14.7	26.4%	4.1	91.9%	14.3
H2V3WDSXY_1_38UDI-idt-UMI-b	1.07	14.5	26.5%	4.1	91.8%	14.1
H2VWGDSXY_4_242UDI-idt-UMI	1.04	14.5	34.3%	5.2	89.4%	13.6
H2V3WDSXY_3_38UDI-idt-UMI-b	1.07	14.3	26.5%	4.0	91.7%	13.9
H2VWGDSXY_3_242UDI-idt-UMI	1.02	14.3	34.3%	5.2	89.3%	13.5
H2VWGDSXY_1_242UDI-idt-UMI	1.03	14.2	34.3%	5.1	89.3%	13.4
H2VWGDSXY_2_242UDI-idt-UMI	1.05	14.2	34.3%	5.1	89.3%	13.4
H2V3WDSXY_4_38UDI-idt-UMI-b	1.11	14.1	26.6%	4.0	91.8%	13.7
H2VW3DSXY_1_1UDI-idt-UMI	1.01	13.2	34.6%	4.8	89.6%	12.5
H2VW3DSXY_2_1UDI-idt-UMI	1.10	13.0	34.6%	4.7	89.5%	12.2
H2VW3DSXY_3_1UDI-idt-UMI	1.08	13.0	34.6%	4.8	89.5%	12.3
H2VW3DSXY_4_1UDI-idt-UMI	1.05	12.9	34.6%	4.7	89.5%	12.1
H2VWGDSXY_4_168UDI-idt-UMI	1.07	12.5	30.2%	4.1	87.2%	11.8
H2VWGDSXY_3_168UDI-idt-UMI	1.03	12.4	30.2%	4.0	87.2%	11.7
H2VWGDSXY_1_168UDI-idt-UMI	1.02	12.3	30.3%	4.0	87.2%	11.6
H2VWGDSXY_2_168UDI-idt-UMI	1.04	12.3	30.3%	4.0	87.2%	11.6

H2VW3DSXY_1_38UDI-idt-UMI-b	1.06	11.4	26.9%	3.2	91.6%	11.0
H2VW3DSXY_2_38UDI-idt-UMI-b	1.05	11.1	26.9%	3.2	91.5%	10.7
H2VW3DSXY_3_38UDI-idt-UMI-b	1.10	11.1	26.9%	3.2	91.5%	10.8
H2VW3DSXY_4_38UDI-idt-UMI-b	1.08	11.0	26.9%	3.1	91.5%	10.6
H2V3WDSXY_2_26UDI-idt-UMI	0.99	9.0	26.2%	2.5	89.2%	8.6
H2V3WDSXY_1_26UDI-idt-UMI	1.04	8.8	26.2%	2.5	89.1%	8.5
H2V3WDSXY_3_26UDI-idt-UMI	1.09	8.7	26.2%	2.5	89.1%	8.4
H2V3WDSXY_4_26UDI-idt-UMI	1.05	8.6	26.3%	2.4	89.0%	8.3
H2VW3DSXY_1_73UDI-idt-UMI	0.98	8.2	24.4%	2.2	86.6%	7.9
H2VW3DSXY_2_73UDI-idt-UMI	1.03	8.1	24.4%	2.2	86.5%	7.8
H2VW3DSXY_3_73UDI-idt-UMI	0.99	8.1	24.4%	2.2	86.6%	7.8
H2VW3DSXY_4_73UDI-idt-UMI	0.98	8.1	24.4%	2.2	86.5%	7.7
H2VTCDSXY_2_2UDI-idt-UMI	1.13	7.2	35.4%	2.8	87.2%	6.8
H2VTCDSXY_3_2UDI-idt-UMI	1.12	7.2	35.3%	2.8	87.1%	6.9
H2VTCDSXY_4_2UDI-idt-UMI	1.09	7.2	35.3%	2.8	87.1%	6.9
H2VTCDSXY_1_2UDI-idt-UMI	1.06	7.1	35.3%	2.8	87.2%	6.8
H2VWGDSXY_4_106UDI-idt-UMI	1.00	7.1	20.7%	1.6	87.2%	6.8
H2VW3DSXY_1_37UDI-idt-UMI	1.10	7.0	24.4%	1.9	85.5%	6.8
H2VWGDSXY_3_106UDI-idt-UMI	0.97	7.0	20.7%	1.6	87.2%	6.7
H2VW3DSXY_2_37UDI-idt-UMI	1.03	6.9	24.4%	1.9	85.4%	6.7
H2VW3DSXY_3_37UDI-idt-UMI	0.97	6.9	24.4%	1.9	85.4%	6.7
H2VWGDSXY_1_106UDI-idt-UMI	0.98	6.9	20.7%	1.6	87.2%	6.7
H2VWGDSXY_2_106UDI-idt-UMI	1.02	6.9	20.8%	1.6	87.2%	6.7
H2VW3DSXY_4_37UDI-idt-UMI	1.07	6.8	24.4%	1.9	85.4%	6.6
H2V3WDSXY_2_245UDI-idt-UMI	0.99	6.2	21.7%	1.4	91.1%	6.0
H2V3WDSXY_1_245UDI-idt-UMI	1.05	6.1	21.7%	1.4	91.1%	5.9
H2V3WDSXY_3_245UDI-idt-UMI	1.01	6.1	21.7%	1.4	91.0%	5.9
H2V3WDSXY_4_245UDI-idt-UMI	1.06	6.1	21.8%	1.4	91.0%	5.9
H2VWGDSXY_4_194UDI-idt-UMI	1.09	6.0	35.9%	2.4	84.9%	5.6
H2VWGDSXY_1_194UDI-idt-UMI	1.06	5.9	35.9%	2.3	84.8%	5.5
H2VWGDSXY_2_194UDI-idt-UMI	1.00	5.9	36.0%	2.3	84.8%	5.5
H2VWGDSXY_3_194UDI-idt-UMI	1.05	5.9	35.9%	2.3	84.8%	5.5
H2VWGDSXY_4_120UDI-idt-UMI	1.05	5.7	20.4%	1.3	87.3%	5.4
H2VWGDSXY_1_120UDI-idt-UMI	1.00	5.6	20.5%	1.2	87.2%	5.3
H2VWGDSXY_2_120UDI-idt-UMI	1.05	5.6	20.4%	1.2	87.3%	5.3

H2VWGDSXY_3_120UDI-idt-UMI	1.04	5.6	20.4%	1.3	87.2%	5.4
H2VWGDSXY_4_266UDI-idt-UMI	1.06	5.6	45.2%	2.7	87.8%	5.3
H2VWGDSXY_1_266UDI-idt-UMI	1.12	5.5	45.3%	2.7	87.9%	5.2
H2VWGDSXY_2_266UDI-idt-UMI	1.05	5.5	45.3%	2.7	87.8%	5.2
H2VWGDSXY_3_266UDI-idt-UMI	1.07	5.5	45.2%	2.7	87.8%	5.2
H2V3WDSXY_2_221UDI-idt-UMI	1.07	5.4	23.3%	1.3	90.5%	5.2
H2V3WDSXY_1_221UDI-idt-UMI	1.11	5.3	23.3%	1.3	90.5%	5.2
H2V3WDSXY_3_221UDI-idt-UMI	1.12	5.3	23.4%	1.3	90.4%	5.1
H2V3WDSXY_1_233UDI-idt-UMI	1.06	5.2	25.7%	1.4	89.7%	5.0
H2V3WDSXY_2_233UDI-idt-UMI	1.01	5.2	25.7%	1.4	89.7%	5.0
H2V3WDSXY_4_221UDI-idt-UMI	1.05	5.2	23.4%	1.3	90.4%	5.0
H2VWGDSXY_4_206UDI-idt-UMI	0.94	5.2	22.9%	1.3	88.5%	5.0
H2V3WDSXY_1_209UDI-idt-UMI	1.04	5.1	25.5%	1.4	87.2%	4.8
H2V3WDSXY_2_209UDI-idt-UMI	1.03	5.1	25.5%	1.4	87.3%	4.9
H2V3WDSXY_3_233UDI-idt-UMI	1.02	5.1	25.7%	1.4	89.6%	4.9
H2V3WDSXY_4_233UDI-idt-UMI	1.10	5.1	25.8%	1.4	89.6%	4.9
H2VWGDSXY_1_206UDI-idt-UMI	0.87	5.1	23.0%	1.3	88.5%	4.9
H2VWGDSXY_2_206UDI-idt-UMI	0.91	5.1	23.0%	1.3	88.5%	4.9
H2VWGDSXY_3_206UDI-idt-UMI	0.87	5.1	23.0%	1.3	88.4%	4.9
H2V3WDSXY_1_85UDI-idt-UMI	1.02	5.0	24.0%	1.3	85.4%	4.7
H2V3WDSXY_2_85UDI-idt-UMI	0.99	5.0	24.0%	1.3	85.5%	4.8
H2V3WDSXY_3_209UDI-idt-UMI	1.12	5.0	25.5%	1.4	87.1%	4.7
H2V3WDSXY_4_209UDI-idt-UMI	0.99	5.0	25.6%	1.4	87.2%	4.7
H2V3WDSXY_3_85UDI-idt-UMI	0.99	4.9	24.1%	1.3	85.3%	4.7
H2VWGDSXY_4_143UDI-idt-UMI	0.99	4.9	22.3%	1.2	85.6%	4.6
H2V3WDSXY_4_85UDI-idt-UMI	0.99	4.8	24.1%	1.3	85.3%	4.6
H2VWGDSXY_1_143UDI-idt-UMI	1.01	4.8	22.3%	1.2	85.6%	4.6
H2VWGDSXY_2_143UDI-idt-UMI	0.96	4.8	22.4%	1.2	85.6%	4.6
H2VWGDSXY_3_143UDI-idt-UMI	0.95	4.8	22.3%	1.2	85.6%	4.6
H2V3WDSXY_1_281UDI-idt-UMI	1.05	4.7	33.6%	1.7	85.1%	4.4
H2V3WDSXY_2_281UDI-idt-UMI	1.10	4.7	33.6%	1.7	85.2%	4.4
H2V3WDSXY_3_281UDI-idt-UMI	1.07	4.6	33.6%	1.7	85.0%	4.3
H2V3WDSXY_4_281UDI-idt-UMI	1.14	4.6	33.7%	1.7	85.1%	4.3
H2V3WDSXY_1_259UDI-idt-UMI	1.07	4.4	23.4%	1.1	90.4%	4.2
H2V3WDSXY_2_259UDI-idt-UMI	0.99	4.4	23.4%	1.1	90.5%	4.3

H2VWGDSXY_1_167UDI-idt-UMI	0.98	4.4	24.6%	1.2	85.5%	4.2
H2VWGDSXY_3_167UDI-idt-UMI	1.02	4.4	24.6%	1.2	85.6%	4.2
H2VWGDSXY_4_167UDI-idt-UMI	1.01	4.4	24.6%	1.2	85.6%	4.2
H2V3WDSXY_3_259UDI-idt-UMI	0.99	4.3	23.4%	1.1	90.3%	4.2
H2V3WDSXY_4_259UDI-idt-UMI	1.02	4.3	23.5%	1.1	90.3%	4.1
H2VWGDSXY_1_231UDI-idt-UMI	0.90	4.3	17.1%	0.8	84.8%	4.1
H2VWGDSXY_2_167UDI-idt-UMI	1.03	4.3	24.6%	1.2	85.6%	4.2
H2VWGDSXY_2_231UDI-idt-UMI	0.95	4.3	17.1%	0.8	84.8%	4.1
H2VWGDSXY_3_231UDI-idt-UMI	0.97	4.3	17.1%	0.8	84.7%	4.2
H2VWGDSXY_4_231UDI-idt-UMI	0.96	4.3	17.1%	0.8	84.8%	4.2
H2V3WDSXY_1_200UDI-idt-UMI	1.05	4.2	27.0%	1.2	91.1%	4.0
H2V3WDSXY_2_200UDI-idt-UMI	0.97	4.2	27.0%	1.2	91.1%	4.1
H2VW5DSXY_1_267UDI-idt-UMI	0.98	4.2	25.7%	1.2	87.8%	4.0
H2VW5DSXY_3_267UDI-idt-UMI	1.03	4.2	25.7%	1.2	87.6%	4.0
H2VW5DSXY_4_267UDI-idt-UMI	0.94	4.2	25.8%	1.2	87.7%	4.0
H2V3WDSXY_1_197UDI-idt-UMI	1.11	4.1	26.9%	1.2	87.3%	3.9
H2V3WDSXY_2_197UDI-idt-UMI	1.10	4.1	26.9%	1.2	87.3%	3.9
H2V3WDSXY_3_197UDI-idt-UMI	1.03	4.1	27.0%	1.2	87.1%	3.8
H2V3WDSXY_3_200UDI-idt-UMI	1.04	4.1	27.1%	1.2	90.9%	4.0
H2V3WDSXY_4_200UDI-idt-UMI	1.04	4.1	27.2%	1.2	91.0%	3.9
H2VW5DSXY_2_267UDI-idt-UMI	1.01	4.1	25.8%	1.2	87.7%	3.9
H2VY7DSXY_1_1UDI-idt-UMI	1.04	4.1	34.3%	1.5	89.6%	3.9
H2VY7DSXY_2_1UDI-idt-UMI	1.07	4.1	34.5%	1.5	89.6%	3.8
H2VY7DSXY_3_1UDI-idt-UMI	1.07	4.1	34.3%	1.5	89.5%	3.9
H2V3WDSXY_4_197UDI-idt-UMI	1.07	4.0	27.1%	1.2	87.1%	3.8
H2VY7DSXY_4_1UDI-idt-UMI	1.05	4.0	34.4%	1.5	89.6%	3.8
H2VWGDSXY_1_107UDI-idt-UMI	1.01	3.8	40.0%	1.7	86.2%	3.6
H2VWGDSXY_3_107UDI-idt-UMI	0.98	3.8	40.0%	1.7	86.2%	3.6
H2VWGDSXY_4_107UDI-idt-UMI	0.96	3.8	39.9%	1.7	86.2%	3.6
H2VW3DSXY_1_283UDI-idt-UMI	0.92	3.7	22.6%	0.9	90.2%	3.6
H2VW5DSXY_1_232UDI-idt-UMI	0.96	3.7	31.0%	1.2	87.5%	3.5
H2VW5DSXY_3_232UDI-idt-UMI	1.11	3.7	30.9%	1.3	87.3%	3.6
H2VWGDSXY_2_107UDI-idt-UMI	0.96	3.7	40.0%	1.7	86.2%	3.6
H2VW3DSXY_1_270UDI-idt-UMI	0.83	3.6	27.3%	1.1	82.0%	3.4
H2VW3DSXY_2_283UDI-idt-UMI	0.98	3.6	22.6%	0.9	90.1%	3.5

H2VW3DSXY_3_283UDI-idt-UMI	1.00	3.6	22.7%	0.9	90.2%	3.5
H2VW3DSXY_4_283UDI-idt-UMI	0.95	3.6	22.6%	0.9	90.2%	3.5
H2VW5DSXY_2_232UDI-idt-UMI	0.97	3.6	31.1%	1.2	87.5%	3.4
H2VW5DSXY_4_232UDI-idt-UMI	1.03	3.6	31.1%	1.2	87.5%	3.4
H2VWGDSXY_1_179UDI-idt-UMI	0.97	3.6	19.1%	0.8	84.4%	3.4
H2VWGDSXY_2_179UDI-idt-UMI	1.05	3.6	19.1%	0.8	84.5%	3.5
H2VWGDSXY_3_179UDI-idt-UMI	1.06	3.6	19.1%	0.8	84.4%	3.5
H2VWGDSXY_4_179UDI-idt-UMI	0.99	3.6	19.1%	0.8	84.4%	3.5
H2VWGDSXY_4_218UDI-idt-UMI	1.00	3.6	24.3%	1.0	82.2%	3.4
H2V3WDSXY_1_235UDI-idt-UMI	1.02	3.5	20.7%	0.8	87.5%	3.3
H2V3WDSXY_2_235UDI-idt-UMI	0.98	3.5	20.7%	0.8	87.6%	3.4
H2VW3DSXY_1_200UDI-idt-UMI	1.05	3.5	27.3%	1.0	90.8%	3.4
H2VW3DSXY_1_259UDI-idt-UMI	0.96	3.5	23.5%	0.9	90.1%	3.4
H2VW3DSXY_2_259UDI-idt-UMI	1.00	3.5	23.5%	0.9	90.1%	3.4
H2VW3DSXY_2_270UDI-idt-UMI	0.82	3.5	27.3%	1.1	82.0%	3.4
H2VW3DSXY_3_200UDI-idt-UMI	1.05	3.5	27.4%	1.0	90.7%	3.3
H2VW3DSXY_3_259UDI-idt-UMI	0.94	3.5	23.6%	0.9	90.1%	3.4
H2VW3DSXY_3_270UDI-idt-UMI	0.76	3.5	27.3%	1.1	81.9%	3.4
H2VW3DSXY_4_270UDI-idt-UMI	0.78	3.5	27.3%	1.1	82.0%	3.3
H2VWGDSXY_1_218UDI-idt-UMI	0.92	3.5	24.4%	1.0	82.2%	3.3
H2VWGDSXY_2_218UDI-idt-UMI	0.91	3.5	24.4%	1.0	82.2%	3.3
H2VWGDSXY_3_166UDI-idt-UMI	1.03	3.5	21.4%	0.8	83.5%	3.3
H2VWGDSXY_3_192UDI-idt-UMI	0.99	3.5	26.8%	1.1	81.2%	3.2
H2VWGDSXY_3_218UDI-idt-UMI	0.93	3.5	24.4%	1.0	82.2%	3.3
H2VWGDSXY_4_166UDI-idt-UMI	1.04	3.5	21.4%	0.8	83.6%	3.3
H2VWGDSXY_4_192UDI-idt-UMI	1.01	3.5	26.7%	1.1	81.2%	3.3
H2V3WDSXY_1_25UDI-idt-UMI	1.07	3.4	21.3%	0.8	82.1%	3.2
H2V3WDSXY_2_257UDI-idt-UMI	1.02	3.4	19.6%	0.7	88.0%	3.3
H2V3WDSXY_2_25UDI-idt-UMI	1.09	3.4	21.3%	0.8	82.2%	3.2
H2V3WDSXY_2_283UDI-idt-UMI	0.93	3.4	22.4%	0.8	90.8%	3.3
H2V3WDSXY_3_235UDI-idt-UMI	1.05	3.4	20.7%	0.8	87.5%	3.3
H2V3WDSXY_3_25UDI-idt-UMI	1.04	3.4	21.3%	0.8	82.0%	3.2
H2V3WDSXY_4_235UDI-idt-UMI	1.00	3.4	20.8%	0.8	87.5%	3.2
H2VW3DSXY_1_222UDI-idt-UMI	0.94	3.4	36.6%	1.4	81.5%	3.2
H2VW3DSXY_1_271UDI-idt-UMI	0.96	3.4	21.0%	0.8	89.2%	3.3

H2VW3DSXY_2_200UDI-idt-UMI	1.06	3.4	27.4%	1.0	90.7%	3.3
H2VW3DSXY_4_200UDI-idt-UMI	1.04	3.4	27.4%	1.0	90.7%	3.3
H2VW3DSXY_4_259UDI-idt-UMI	0.99	3.4	23.5%	0.9	90.0%	3.3
H2VWGDSXY_1_166UDI-idt-UMI	1.11	3.4	21.4%	0.8	83.6%	3.3
H2VWGDSXY_1_192UDI-idt-UMI	0.98	3.4	26.7%	1.1	81.1%	3.2
H2VWGDSXY_2_166UDI-idt-UMI	1.06	3.4	21.3%	0.8	83.6%	3.3
H2VWGDSXY_2_192UDI-idt-UMI	1.05	3.4	26.7%	1.1	81.2%	3.2
H2V3WDSXY_1_257UDI-idt-UMI	0.94	3.3	19.6%	0.7	87.9%	3.2
H2V3WDSXY_1_283UDI-idt-UMI	0.99	3.3	22.5%	0.8	90.7%	3.2
H2V3WDSXY_3_257UDI-idt-UMI	1.06	3.3	19.6%	0.7	87.9%	3.2
H2V3WDSXY_3_283UDI-idt-UMI	0.96	3.3	22.5%	0.8	90.6%	3.2
H2V3WDSXY_4_257UDI-idt-UMI	0.99	3.3	19.6%	0.7	87.8%	3.2
H2V3WDSXY_4_25UDI-idt-UMI	1.00	3.3	21.4%	0.8	81.9%	3.2
H2VW3DSXY_1_209UDI-idt-UMI	1.05	3.3	25.6%	0.9	87.1%	3.1
H2VW3DSXY_2_222UDI-idt-UMI	1.00	3.3	36.6%	1.4	81.3%	3.1
H2VW3DSXY_2_271UDI-idt-UMI	0.96	3.3	21.1%	0.8	89.1%	3.2
H2VW3DSXY_3_222UDI-idt-UMI	1.00	3.3	36.7%	1.4	81.4%	3.1
H2VW3DSXY_3_271UDI-idt-UMI	0.93	3.3	21.0%	0.8	89.1%	3.2
H2VW3DSXY_4_222UDI-idt-UMI	1.00	3.3	36.6%	1.4	81.3%	3.1
H2VW3DSXY_4_271UDI-idt-UMI	0.94	3.3	21.0%	0.8	89.1%	3.2
H2V3WDSXY_4_283UDI-idt-UMI	0.96	3.2	22.6%	0.8	90.6%	3.1
H2VW3DSXY_1_245UDI-idt-UMI	1.06	3.2	21.8%	0.7	91.0%	3.0
H2VW3DSXY_2_209UDI-idt-UMI	1.05	3.2	25.7%	0.9	87.1%	3.0
H2VW3DSXY_3_209UDI-idt-UMI	1.10	3.2	25.7%	0.9	87.1%	3.1
H2VW3DSXY_4_209UDI-idt-UMI	1.02	3.2	25.7%	0.9	87.1%	3.0
H2V3WDSXY_1_271UDI-idt-UMI	1.02	3.1	20.9%	0.7	89.5%	3.0
H2V3WDSXY_2_271UDI-idt-UMI	1.00	3.1	20.8%	0.7	89.6%	3.0
H2VW3DSXY_2_245UDI-idt-UMI	1.00	3.1	21.8%	0.7	90.9%	3.0
H2VW3DSXY_3_245UDI-idt-UMI	0.99	3.1	21.8%	0.7	91.0%	3.0
H2VW3DSXY_4_245UDI-idt-UMI	1.03	3.1	21.8%	0.7	90.9%	3.0
H2V3WDSXY_1_199UDI-idt-UMI	1.01	3.0	29.1%	1.0	85.7%	2.8
H2V3WDSXY_2_199UDI-idt-UMI	1.04	3.0	29.2%	1.0	85.8%	2.8
H2V3WDSXY_3_271UDI-idt-UMI	0.95	3.0	20.9%	0.7	89.4%	2.9
H2V3WDSXY_4_271UDI-idt-UMI	0.98	3.0	20.9%	0.7	89.4%	2.9
H2VW3DSXY_1_198UDI-idt-UMI	1.09	3.0	33.5%	1.1	82.6%	2.8

H2VW3DSXY_1_235UDI-idt-UMI	0.96	3.0	20.8%	0.7	87.2%	2.9
H2V3WDSXY_3_199UDI-idt-UMI	1.01	2.9	29.2%	1.0	85.6%	2.8
H2V3WDSXY_4_199UDI-idt-UMI	1.05	2.9	29.3%	1.0	85.6%	2.8
H2VW3DSXY_2_198UDI-idt-UMI	1.07	2.9	33.5%	1.1	82.5%	2.7
H2VW3DSXY_2_235UDI-idt-UMI	1.03	2.9	20.8%	0.7	87.2%	2.8
H2VW3DSXY_3_198UDI-idt-UMI	1.08	2.9	33.5%	1.1	82.5%	2.7
H2VW3DSXY_3_235UDI-idt-UMI	1.08	2.9	20.8%	0.7	87.2%	2.8
H2VW3DSXY_4_198UDI-idt-UMI	1.06	2.9	33.5%	1.1	82.5%	2.7
H2VW3DSXY_4_235UDI-idt-UMI	0.95	2.9	20.8%	0.7	87.1%	2.7
H2VWGDSXY_1_278UDI-idt-UMI	1.29	2.9	9.2%	0.3	30.0%	1.1
H2VWGDSXY_2_278UDI-idt-UMI	1.13	2.9	9.2%	0.3	30.0%	1.1
H2VWGDSXY_3_278UDI-idt-UMI	1.16	2.9	9.2%	0.3	30.0%	1.1
H2VWGDSXY_4_278UDI-idt-UMI	1.09	2.9	9.3%	0.3	30.1%	1.1
H2VW3DSXY_1_233UDI-idt-UMI	0.94	2.8	25.8%	0.8	89.4%	2.7
H2VW3DSXY_3_233UDI-idt-UMI	1.07	2.8	25.8%	0.8	89.4%	2.6
H2VW3DSXY_1_221UDI-idt-UMI	1.07	2.7	23.4%	0.7	90.2%	2.6
H2VW3DSXY_2_233UDI-idt-UMI	1.10	2.7	25.8%	0.8	89.4%	2.6
H2VW3DSXY_4_233UDI-idt-UMI	1.09	2.7	25.8%	0.8	89.4%	2.6
H2V3WDSXY_1_224UDI-idt-UMI	1.00	2.6	19.1%	0.5	86.7%	2.5
H2V3WDSXY_1_247UDI-idt-UMI	0.96	2.6	18.7%	0.5	87.3%	2.5
H2V3WDSXY_2_224UDI-idt-UMI	0.97	2.6	19.0%	0.5	86.7%	2.5
H2V3WDSXY_2_247UDI-idt-UMI	0.98	2.6	18.7%	0.5	87.4%	2.6
H2V3WDSXY_2_269UDI-idt-UMI	1.14	2.6	15.2%	0.4	85.3%	2.5
H2V3WDSXY_3_247UDI-idt-UMI	1.02	2.6	18.7%	0.5	87.2%	2.5
H2V3WDSXY_4_247UDI-idt-UMI	0.94	2.6	18.7%	0.5	87.2%	2.5
H2VW3DSXY_2_221UDI-idt-UMI	1.06	2.6	23.5%	0.7	90.2%	2.5
H2VW3DSXY_3_221UDI-idt-UMI	1.06	2.6	23.4%	0.7	90.2%	2.5
H2VW3DSXY_4_221UDI-idt-UMI	1.06	2.6	23.5%	0.6	90.2%	2.5
H2V3WDSXY_1_269UDI-idt-UMI	0.97	2.5	15.2%	0.4	85.3%	2.4
H2V3WDSXY_3_224UDI-idt-UMI	0.87	2.5	19.1%	0.5	86.6%	2.4
H2V3WDSXY_3_269UDI-idt-UMI	0.97	2.5	15.3%	0.4	85.2%	2.4
H2V3WDSXY_4_224UDI-idt-UMI	0.86	2.5	19.1%	0.5	86.5%	2.4
H2V3WDSXY_4_269UDI-idt-UMI	1.09	2.5	15.3%	0.4	85.2%	2.4
H2VW5DSXY_3_279UDI-idt-UMI	1.00	2.5	32.0%	0.9	84.5%	2.4
H2VW5DSXY_4_279UDI-idt-UMI	1.07	2.5	32.1%	0.9	84.7%	2.3

H2VWGDSXY_1_230UDI-idt-UMI	0.88	2.5	19.0%	0.5	83.4%	2.4
H2VWGDSXY_2_230UDI-idt-UMI	0.88	2.5	19.0%	0.5	83.4%	2.4
H2VWGDSXY_3_230UDI-idt-UMI	0.96	2.5	19.0%	0.6	83.4%	2.4
H2VWGDSXY_4_230UDI-idt-UMI	0.91	2.5	19.1%	0.6	83.5%	2.4
H2VW5DSXY_1_279UDI-idt-UMI	0.99	2.4	32.1%	0.9	84.7%	2.3
H2VW5DSXY_2_279UDI-idt-UMI	1.02	2.4	32.1%	0.9	84.6%	2.3
H2VWGDSXY_1_191UDI-idt-UMI	1.04	2.4	22.7%	0.6	79.9%	2.3
H2VWGDSXY_2_191UDI-idt-UMI	1.11	2.4	22.7%	0.7	80.0%	2.3
H2VWGDSXY_3_191UDI-idt-UMI	1.12	2.4	22.7%	0.7	79.9%	2.3
H2VWGDSXY_4_191UDI-idt-UMI	0.95	2.4	22.7%	0.7	80.0%	2.3
H2VW3DSXY_1_197UDI-idt-UMI	1.10	2.2	27.1%	0.7	87.0%	2.1
H2VW3DSXY_2_197UDI-idt-UMI	1.00	2.2	27.1%	0.6	87.0%	2.1
H2VW3DSXY_3_197UDI-idt-UMI	1.13	2.2	27.1%	0.6	87.0%	2.1
H2VW3DSXY_1_281UDI-idt-UMI	1.11	2.1	33.6%	0.8	85.0%	2.0
H2VW3DSXY_2_281UDI-idt-UMI	1.05	2.1	33.7%	0.8	85.0%	2.0
H2VW3DSXY_3_281UDI-idt-UMI	1.06	2.1	33.7%	0.8	85.0%	2.0
H2VW3DSXY_4_197UDI-idt-UMI	1.05	2.1	27.1%	0.6	87.0%	2.0
H2VW3DSXY_4_281UDI-idt-UMI	1.13	2.1	33.6%	0.8	84.9%	1.9
H2VW3DSXY_1_246UDI-idt-UMI	1.00	2.0	25.8%	0.6	82.9%	1.9
H2VW3DSXY_2_246UDI-idt-UMI	1.10	2.0	25.8%	0.6	82.8%	1.9
H2VW3DSXY_3_246UDI-idt-UMI	1.09	2.0	25.8%	0.6	82.9%	1.9
H2VW3DSXY_4_246UDI-idt-UMI	1.05	2.0	25.8%	0.6	82.8%	1.9
H2V3WDSXY_1_280UDI-idt-UMI	0.93	1.9	13.4%	0.3	84.5%	1.8
H2V3WDSXY_2_280UDI-idt-UMI	1.13	1.9	13.4%	0.3	84.6%	1.8
H2V3WDSXY_3_280UDI-idt-UMI	0.94	1.9	13.3%	0.3	84.4%	1.8
H2V3WDSXY_4_280UDI-idt-UMI	1.03	1.9	13.3%	0.3	84.4%	1.8
H2VW3DSXY_1_280UDI-idt-UMI	1.03	1.9	13.4%	0.3	84.1%	1.9
H2VW3DSXY_2_280UDI-idt-UMI	1.06	1.9	13.3%	0.3	84.0%	1.8
H2VW3DSXY_3_280UDI-idt-UMI	0.96	1.9	13.3%	0.3	84.0%	1.8
H2VW3DSXY_4_280UDI-idt-UMI	1.00	1.9	13.3%	0.3	84.1%	1.8
H2VW3DSXY_1_247UDI-idt-UMI	1.01	1.8	18.8%	0.4	86.6%	1.7
H2VW3DSXY_1_257UDI-idt-UMI	1.03	1.8	19.6%	0.4	87.7%	1.7
H2VW3DSXY_2_247UDI-idt-UMI	0.89	1.8	18.7%	0.4	86.5%	1.7
H2VW3DSXY_3_247UDI-idt-UMI	0.87	1.8	18.8%	0.4	86.6%	1.7
H2VW3DSXY_2_257UDI-idt-UMI	1.00	1.7	19.6%	0.4	87.6%	1.7

H2VW3DSXY_3_257UDI-idt-UMI	0.96	1.7	19.6%	0.4	87.7%	1.7
H2VW3DSXY_4_247UDI-idt-UMI	0.93	1.7	18.7%	0.4	86.5%	1.7
H2VW3DSXY_4_257UDI-idt-UMI	1.04	1.7	19.6%	0.4	87.6%	1.6
H2V3WDSXY_1_236UDI-idt-UMI	1.03	1.6	15.2%	0.3	84.7%	1.6
H2V3WDSXY_1_284UDI-idt-UMI	0.87	1.6	20.3%	0.4	73.8%	1.4
H2V3WDSXY_2_236UDI-idt-UMI	0.98	1.6	15.2%	0.3	84.8%	1.6
H2V3WDSXY_2_284UDI-idt-UMI	0.91	1.6	20.3%	0.4	74.0%	1.4
H2V3WDSXY_3_236UDI-idt-UMI	1.01	1.6	15.2%	0.3	84.7%	1.6
H2V3WDSXY_3_284UDI-idt-UMI	0.99	1.6	20.3%	0.4	73.7%	1.4
H2V3WDSXY_4_236UDI-idt-UMI	1.02	1.6	15.3%	0.3	84.6%	1.6
H2V3WDSXY_4_284UDI-idt-UMI	0.88	1.6	20.3%	0.4	73.6%	1.4
H2VW3DSXY_1_199UDI-idt-UMI	1.03	1.6	29.2%	0.5	85.2%	1.5
H2VW3DSXY_1_224UDI-idt-UMI	1.02	1.6	19.1%	0.3	86.1%	1.5
H2VW3DSXY_1_279UDI-idt-UMI	1.03	1.6	32.2%	0.6	84.3%	1.5
H2VW3DSXY_2_199UDI-idt-UMI	1.04	1.6	29.3%	0.5	85.1%	1.5
H2VW3DSXY_2_224UDI-idt-UMI	0.95	1.6	19.1%	0.3	86.0%	1.5
H2VW3DSXY_2_279UDI-idt-UMI	1.02	1.6	32.2%	0.6	84.2%	1.5
H2VW3DSXY_3_199UDI-idt-UMI	1.05	1.6	29.3%	0.5	85.1%	1.5
H2VW3DSXY_3_224UDI-idt-UMI	0.98	1.6	19.1%	0.3	86.1%	1.5
H2VW3DSXY_3_279UDI-idt-UMI	1.03	1.6	32.4%	0.6	84.3%	1.5
H2VW3DSXY_4_199UDI-idt-UMI	1.08	1.6	29.2%	0.5	85.1%	1.5
H2VW3DSXY_4_279UDI-idt-UMI	1.01	1.6	32.2%	0.6	84.2%	1.5
H2VW3DSXY_1_210UDI-idt-UMI	1.06	1.5	33.2%	0.7	65.0%	1.4
H2VW3DSXY_1_269UDI-idt-UMI	1.12	1.5	15.3%	0.3	85.1%	1.4
H2VW3DSXY_2_210UDI-idt-UMI	1.03	1.5	33.3%	0.7	64.9%	1.3
H2VW3DSXY_2_269UDI-idt-UMI	1.24	1.5	15.3%	0.3	85.1%	1.4
H2VW3DSXY_3_210UDI-idt-UMI	1.15	1.5	33.2%	0.7	64.9%	1.3
H2VW3DSXY_3_269UDI-idt-UMI	1.17	1.5	15.3%	0.3	85.1%	1.4
H2VW3DSXY_4_210UDI-idt-UMI	1.02	1.5	33.2%	0.7	64.9%	1.3
H2VW3DSXY_4_224UDI-idt-UMI	0.99	1.5	19.1%	0.3	86.0%	1.5
H2V3WDSXY_2_287UDI-idt-UMI	0.98	1.4	18.7%	0.3	80.2%	1.3
H2VW3DSXY_4_269UDI-idt-UMI	1.13	1.4	15.2%	0.2	85.1%	1.4
H2VY7DSXY_1_49UDI-idt-UMI	1.02	1.4	30.2%	0.5	85.3%	1.3
H2VY7DSXY_2_49UDI-idt-UMI	1.04	1.4	30.3%	0.5	85.3%	1.3
H2VY7DSXY_3_49UDI-idt-UMI	1.08	1.4	30.2%	0.5	85.3%	1.3

H2VY7DSXY_4_49UDI-idt-UMI	1.03	1.4	30.3%	0.5	85.3%	1.3
H2V3WDSXY_1_287UDI-idt-UMI	0.97	1.3	18.8%	0.3	80.2%	1.3
H2V3WDSXY_3_287UDI-idt-UMI	1.12	1.3	18.9%	0.3	80.1%	1.3
H2V3WDSXY_4_287UDI-idt-UMI	0.94	1.3	18.9%	0.3	80.0%	1.3
H2VW3DSXY_1_236UDI-idt-UMI	1.16	1.3	15.1%	0.2	83.9%	1.3
H2VW3DSXY_1_267UDI-idt-UMI	1.05	1.3	25.9%	0.4	87.4%	1.3
H2VW3DSXY_2_236UDI-idt-UMI	1.00	1.3	15.2%	0.2	83.8%	1.3
H2VW3DSXY_2_267UDI-idt-UMI	1.00	1.3	26.1%	0.4	87.3%	1.2
H2VW3DSXY_3_236UDI-idt-UMI	0.92	1.3	15.2%	0.2	83.9%	1.3
H2VW3DSXY_3_267UDI-idt-UMI	0.98	1.3	26.0%	0.4	87.3%	1.2
H2VW3DSXY_4_236UDI-idt-UMI	1.11	1.3	15.2%	0.2	83.9%	1.2
H2VW3DSXY_4_267UDI-idt-UMI	0.95	1.3	26.0%	0.4	87.3%	1.2
H2V3WDSXY_1_212UDI-idt-UMI	1.25	1.2	13.1%	0.2	80.6%	1.2
H2V3WDSXY_2_212UDI-idt-UMI	1.37	1.2	13.0%	0.2	80.7%	1.2
H2V3WDSXY_3_212UDI-idt-UMI	1.17	1.2	13.0%	0.2	80.5%	1.1
H2V3WDSXY_4_212UDI-idt-UMI	1.19	1.2	13.1%	0.2	80.5%	1.1
H2VW3DSXY_1_232UDI-idt-UMI	1.08	1.2	31.3%	0.4	87.0%	1.2
H2VW3DSXY_1_284UDI-idt-UMI	0.93	1.2	20.2%	0.3	73.6%	1.1
H2VW3DSXY_2_232UDI-idt-UMI	1.02	1.2	31.3%	0.4	86.9%	1.1
H2VW3DSXY_3_232UDI-idt-UMI	1.02	1.2	31.4%	0.4	87.0%	1.1
H2VW3DSXY_4_232UDI-idt-UMI	1.06	1.2	31.4%	0.4	86.9%	1.1
H2VW3DSXY_2_284UDI-idt-UMI	0.87	1.1	20.1%	0.3	73.5%	1.0
H2VW3DSXY_3_284UDI-idt-UMI	0.93	1.1	20.2%	0.3	73.5%	1.0
H2VW3DSXY_4_284UDI-idt-UMI	0.85	1.1	20.1%	0.3	73.5%	1.0
H2VW5DSXY_1_244UDI-idt-UMI	1.20	1.1	14.9%	0.2	81.6%	1.1
H2VW5DSXY_2_244UDI-idt-UMI	1.20	1.1	15.0%	0.2	81.5%	1.0
H2VW5DSXY_3_244UDI-idt-UMI	1.24	1.1	14.9%	0.2	81.4%	1.1
H2VW5DSXY_4_244UDI-idt-UMI	1.03	1.1	15.0%	0.2	81.5%	1.1
H2VY7DSXY_1_73UDI-idt-UMI	1.04	1.1	24.4%	0.3	86.6%	1.1
H2VY7DSXY_2_73UDI-idt-UMI	1.08	1.1	24.5%	0.3	86.7%	1.1
H2VY7DSXY_3_73UDI-idt-UMI	0.93	1.1	24.4%	0.3	86.6%	1.1
H2VY7DSXY_4_73UDI-idt-UMI	1.05	1.1	24.4%	0.3	86.6%	1.1
H2V3WDSXY_1_107UDI-idt-UMI	1.00	0.8	40.1%	0.3	86.2%	0.7
H2V3WDSXY_1_246UDI-idt-UMI	1.07	0.8	26.0%	0.2	83.5%	0.7
H2V3WDSXY_1_268UDI-idt-UMI	1.14	0.8	16.4%	0.2	75.0%	0.8

H2V3WDSXY_2_107UDI-idt-UMI	0.97	0.8	40.1%	0.3	86.2%	0.7
H2V3WDSXY_2_246UDI-idt-UMI	1.12	0.8	25.8%	0.2	83.5%	0.7
H2V3WDSXY_2_268UDI-idt-UMI	1.69	0.8	16.3%	0.2	75.0%	0.8
H2V3WDSXY_3_107UDI-idt-UMI	0.98	0.8	40.0%	0.3	86.2%	0.7
H2V3WDSXY_3_246UDI-idt-UMI	1.17	0.8	25.9%	0.2	83.3%	0.7
H2V3WDSXY_3_268UDI-idt-UMI	1.53	0.8	16.4%	0.2	74.8%	0.7
H2V3WDSXY_4_107UDI-idt-UMI	0.96	0.8	40.2%	0.3	86.0%	0.7
H2V3WDSXY_4_246UDI-idt-UMI	1.01	0.8	26.0%	0.2	83.4%	0.7
H2V3WDSXY_4_268UDI-idt-UMI	1.60	0.8	16.3%	0.2	74.8%	0.7
H2VW3DSXY_1_244UDI-idt-UMI	1.13	0.8	15.0%	0.1	80.6%	0.8
H2VW3DSXY_1_278UDI-idt-UMI	1.16	0.8	9.4%	0.1	30.4%	0.3
H2VW3DSXY_1_287UDI-idt-UMI	1.13	0.8	18.9%	0.2	80.0%	0.8
H2VW3DSXY_2_244UDI-idt-UMI	1.08	0.8	15.0%	0.1	80.5%	0.7
H2VW3DSXY_2_278UDI-idt-UMI	1.54	0.8	9.4%	0.1	30.4%	0.3
H2VW3DSXY_2_287UDI-idt-UMI	0.99	0.8	18.8%	0.2	79.9%	0.8
H2VW3DSXY_3_244UDI-idt-UMI	1.17	0.8	15.0%	0.1	80.6%	0.7
H2VW3DSXY_3_278UDI-idt-UMI	1.45	0.8	9.4%	0.1	30.3%	0.3
H2VW3DSXY_3_287UDI-idt-UMI	1.03	0.8	18.8%	0.2	79.9%	0.8
H2VW3DSXY_4_244UDI-idt-UMI	1.30	0.8	15.0%	0.1	80.4%	0.7
H2VW3DSXY_4_278UDI-idt-UMI	1.41	0.8	9.4%	0.1	30.2%	0.3
H2VW3DSXY_4_287UDI-idt-UMI	1.18	0.8	18.8%	0.2	80.0%	0.7
H2VY7DSXY_1_37UDI-idt-UMI	1.05	0.8	24.5%	0.2	85.5%	0.8
H2VY7DSXY_1_38UDI-idt-UMI-b	1.09	0.8	26.8%	0.2	91.7%	0.7
H2VY7DSXY_2_37UDI-idt-UMI	0.99	0.8	24.5%	0.2	85.6%	0.8
H2VY7DSXY_3_37UDI-idt-UMI	0.96	0.8	24.4%	0.2	85.5%	0.8
H2VY7DSXY_3_38UDI-idt-UMI-b	1.02	0.8	26.6%	0.2	91.7%	0.7
H2VY7DSXY_4_37UDI-idt-UMI	1.06	0.8	24.5%	0.2	85.6%	0.7
H2V3WDSXY_1_211UDI-idt-UMI	1.08	0.7	19.2%	0.2	64.7%	0.7
H2V3WDSXY_2_211UDI-idt-UMI	1.02	0.7	19.2%	0.2	64.8%	0.7
H2V3WDSXY_2_248UDI-idt-UMI	12.57	0.7	11.0%	0.1	72.9%	0.6
H2V3WDSXY_3_211UDI-idt-UMI	0.94	0.7	19.2%	0.2	64.5%	0.7
H2V3WDSXY_4_211UDI-idt-UMI	1.03	0.7	19.2%	0.2	64.3%	0.7
H2VW3DSXY_1_212UDI-idt-UMI	1.53	0.7	13.1%	0.1	79.8%	0.7
H2VW3DSXY_2_212UDI-idt-UMI	1.75	0.7	13.0%	0.1	79.7%	0.7
H2VW3DSXY_3_212UDI-idt-UMI	2.17	0.7	13.2%	0.1	79.8%	0.7

H2VW3DSXY_4_212UDI-idt-UMI	1.27	0.7	13.1%	0.1	79.8%	0.7
H2VY7DSXY_2_38UDI-idt-UMI-b	1.11	0.7	26.9%	0.2	91.7%	0.7
H2VY7DSXY_4_38UDI-idt-UMI-b	1.20	0.7	26.9%	0.2	91.8%	0.7
H2V3WDSXY_1_248UDI-idt-UMI	14.50	0.6	11.0%	0.1	72.8%	0.6
H2V3WDSXY_1_260UDI-idt-UMI	2.69	0.6	13.8%	0.1	68.5%	0.5
H2V3WDSXY_2_260UDI-idt-UMI	2.38	0.6	13.9%	0.1	68.7%	0.5
H2V3WDSXY_3_248UDI-idt-UMI		0.6	11.0%	0.1	72.7%	0.6
H2V3WDSXY_3_260UDI-idt-UMI	2.05	0.6	13.8%	0.1	68.3%	0.5
H2V3WDSXY_4_248UDI-idt-UMI		0.6	10.9%	0.1	72.4%	0.6
H2V3WDSXY_4_260UDI-idt-UMI	2.41	0.6	13.8%	0.1	68.0%	0.5
H2VW3DSXY_1_268UDI-idt-UMI	1.86	0.6	16.3%	0.1	74.2%	0.5
H2VW3DSXY_2_268UDI-idt-UMI	2.69	0.6	16.2%	0.1	74.1%	0.5
H2VW3DSXY_3_268UDI-idt-UMI	2.77	0.6	16.2%	0.1	74.2%	0.5
H2VW3DSXY_4_268UDI-idt-UMI	2.66	0.6	16.2%	0.1	74.1%	0.5
H2VY7DSXY_1_61UDI-idt-UMI	1.27	0.6	17.7%	0.1	86.7%	0.5
H2VY7DSXY_3_61UDI-idt-UMI	1.20	0.6	17.8%	0.1	86.5%	0.5
H2VW3DSXY_1_211UDI-idt-UMI	1.06	0.5	19.3%	0.1	64.1%	0.5
H2VW3DSXY_2_211UDI-idt-UMI	1.07	0.5	19.1%	0.1	64.1%	0.4
H2VW3DSXY_3_211UDI-idt-UMI	1.08	0.5	19.2%	0.1	64.0%	0.4
H2VW3DSXY_4_211UDI-idt-UMI	1.04	0.5	19.2%	0.1	64.1%	0.4
H2VY7DSXY_1_14UDI-idt-UMI	1.04	0.5	33.9%	0.2	88.3%	0.5
H2VY7DSXY_2_14UDI-idt-UMI	0.95	0.5	33.9%	0.2	88.3%	0.5
H2VY7DSXY_2_61UDI-idt-UMI	1.05	0.5	17.6%	0.1	86.7%	0.5
H2VY7DSXY_3_14UDI-idt-UMI	1.03	0.5	33.8%	0.2	88.3%	0.5
H2VY7DSXY_4_14UDI-idt-UMI	1.08	0.5	34.1%	0.2	88.3%	0.5
H2VY7DSXY_4_61UDI-idt-UMI	1.24	0.5	17.6%	0.1	86.5%	0.5
H2VW3DSXY_1_248UDI-idt-UMI		0.4	11.1%	0.1	71.8%	0.4
H2VW3DSXY_1_260UDI-idt-UMI	2.94	0.4	13.8%	0.1	67.4%	0.4
H2VW3DSXY_2_248UDI-idt-UMI		0.4	10.9%	0.1	71.6%	0.4
H2VW3DSXY_2_260UDI-idt-UMI	5.38	0.4	13.7%	0.1	67.4%	0.4
H2VW3DSXY_3_248UDI-idt-UMI		0.4	11.0%	0.1	71.8%	0.4
H2VW3DSXY_3_260UDI-idt-UMI	3.03	0.4	13.9%	0.1	67.4%	0.4
H2VW3DSXY_4_248UDI-idt-UMI		0.4	11.0%	0.1	71.6%	0.4
H2VW3DSXY_4_260UDI-idt-UMI	2.92	0.4	13.7%	0.1	67.4%	0.4
H2VY7DSXY_1_26UDI-idt-UMI	1.39	0.3	26.5%	0.1	89.0%	0.3

H2VY7DSXY_2_26UDI-idt-UMI	1.10	0.3	26.6%	0.1	88.9%	0.2
H2VY7DSXY_3_26UDI-idt-UMI	1.21	0.3	26.6%	0.1	88.9%	0.3
H2VY7DSXY_4_26UDI-idt-UMI	1.17	0.3	26.6%	0.1	89.2%	0.2
H2VY7DSXY_1_13UDI-idt-UMI	1.33	0.2	30.9%	0.1	90.4%	0.2
H2VY7DSXY_2_13UDI-idt-UMI	1.28	0.2	31.1%	0.1	90.2%	0.2
H2VY7DSXY_3_13UDI-idt-UMI	1.17	0.2	30.8%	0.1	90.1%	0.2
H2VY7DSXY_4_13UDI-idt-UMI	1.26	0.2	31.1%	0.1	90.3%	0.2
H2VY7DSXY_1_25UDI-idt-UMI		0.1	22.0%	0.0	82.7%	0.1
H2VY7DSXY_1_2UDI-idt-UMI	1.71	0.1	35.3%	0.0	87.4%	0.1
H2VY7DSXY_1_85UDI-idt-UMI	2.05	0.1	24.3%	0.0	84.8%	0.1
H2VY7DSXY_2_25UDI-idt-UMI	6.77	0.1	22.2%	0.0	82.3%	0.0
H2VY7DSXY_2_2UDI-idt-UMI	2.70	0.1	35.5%	0.0	87.4%	0.1
H2VY7DSXY_2_85UDI-idt-UMI	2.38	0.1	24.6%	0.0	84.6%	0.1
H2VY7DSXY_3_25UDI-idt-UMI	6.14	0.1	22.3%	0.0	82.8%	0.0
H2VY7DSXY_3_2UDI-idt-UMI	2.27	0.1	35.7%	0.0	87.2%	0.1
H2VY7DSXY_3_85UDI-idt-UMI	2.21	0.1	24.6%	0.0	84.7%	0.1
H2VY7DSXY_4_25UDI-idt-UMI	4.52	0.1	22.1%	0.0	82.2%	0.0
H2VY7DSXY_4_2UDI-idt-UMI	1.70	0.1	35.8%	0.0	87.3%	0.1
H2VY7DSXY_4_85UDI-idt-UMI	1.84	0.1	24.5%	0.0	84.7%	0.1