UNIVERSITAT DE BARCELONA

Mètodes estadístics per al refinament del diagnòstic dels síndromes limfoproliferatius de cèl.lula B

Guillem Clot Razquin

Mètodes estadístics per al refinament del diagnòstic dels síndromes limfoproliferatius de cèl·lula B

Memòria presentada per **Guillem Clot Razquin** per a optar al grau de doctor per la Universitat de Barcelona

Programa: Estadística. Mètodes estadístics en bioinformàtica.

Departament: Genètica, Microbiologia i Estadística.

Centre de realització de la tesi: IDIBAPS.

Data: Juny 2017, Barcelona.

Guillem Clot Razquin (Doctorand)

Dra. Sílvia M. Beà Bobet (Directora) **Dr. Alexandre Sánchez Pla (Director)**

Resum

Els síndromes limfoproliferatius crònics de cèl·lula B (B-CLPD) engloben diverses entitats heterogènies de tumors hematològics. Aproximadament un 15% dels casos no es poden classificar correctament en cap entitat mitjançant els criteris diagnòstics convencionals. Aquests casos sense un diagnòstic definitiu es consideren B-CLPD, *not otherwise specified* (B-CLPD, NOS). Determinades informacions moleculars, com són l'expressió gènica i les alteracions en el material genètic, podrien contenir trets específics de cada entitat. La tecnologia dels *microarrays* permet mesurar l'expressió de milers de gens o la quantitat de material genètic en milers de regions de manera simultània, el qual facilita la identificació de biomarcadors candidats.

El primer objectiu d'aquesta tesi ha sigut el de construir un predictor, basat en informació de dues plataformes de *microarrays* diferents (expressió i *copy-number*), capaç de classificar les diferents entitats de B-CLPD. Per tal de construir-lo, s'ha utilitzat una cohort de 159 pacients leucèmics distribuïts en nou entitats de B-CLPD diferents. Degut al cost econòmic d'utilitzar dues plataformes, ha sigut molt important avaluar l'increment que suposa en la precisió del predictor utilitzar-ne dues en comptes d'una. La metodologia utilitzada ha permès concloure que l'expressió gènica, com a font única, obté precisions similars a les de combinar les dues fonts.

El segon objectiu ha sigut construir un predictor utilitzant només la font d'expressió. Per tal de minimitzar el risc d'*overfitting*, la metodologia proposada ha permès obtenir un predictor senzill i fàcil de contrastar amb coneixements biològics previs. Aquest predictor ha inclòs 55 gens i ha sigut capaç de distingir sis de les nou entitats. Les tres restants s'han categoritzat com *Miscellaneous*.

Els predictors basats en tecnologies d'alt rendiment, com són els *microarrays*, són

difícils d'implementar a la rutina clínica. Existeixen tècniques més simples, com la *quantitative* PCR (qPCR), però que tenen l'inconvenient de només poder mesurar uns pocs gens simultàniament. El primer pas per tal de construir un predictor basat en qPCR és seleccionar un petit subconjunt de gens a mesurar en aquesta tecnologia, on la selecció es fa en base a la informació obtinguda dels *microarrays*. Aquest canvi de tecnologia té dos inconvenients: *i*) el nombre de gens a traslladar a qPCR és limitat, i *ii*) la correlació entre les dues tecnologies no és perfecta, provocant que les capacitats predictives observades en els *microarrays* no sempre es reprodueixin en qPCR.

El tercer objectiu d'aquesta tesi ha sigut seleccionar un subconjunt de 35 gens a traslladar a qPCR, on la metodologia proposada ha tingut en compte tres rellevàncies (biològica, estadística i multivariant) dels gens per tal de maximitzar la reproductibilitat de la informació al canviar de plataforma. Degut a que la metodologia que s'ha utilitzat per a la selecció de gens és complexa d'aplicar a la pràctica, s'ha proposat lassoVoting, un mètode més senzill i que s'ha demostrat que és adequat quan l'estructura de correlació està ordenada. Amb els 35 gens seleccionats, s'ha utilitzat una mostra de 44 pacients dels 159 inicials per construir el predictor en qPCR. Aquest predictor ha inclòs 8 gens i *s*'ha provat exitosament en una cohort de validació independent de 63 pacients.

Degut a la incapacitat de l'expressió gènica per distingir les tres entitats *Miscellaneous*, el quart objectiu d'aquesta tesi ha sigut avaluar si altres trets moleculars i genètics publicats en la literatura permeten distingir-les. S'han identificat diversos trets que, un cop s'han descartat les altres sis entitats en base a l'expressió gènica, les poden discriminar.

Finalment, en una cohort de 64 B-CLPD, NOS, s'han classificat a una entitat específica 29 casos utilitzant només l'expressió gènica i 14 utilitzant l'expressió i trets moleculars/genètics. En els 21 restants, només s'han pogut descartar sis entitats.

Agraïments

En primer lloc, vull expressar la meva gratitud als dos directors d'aquesta tesi: la Sílvia Beà, qui ha tingut una enorme paciència per ensenyar-me tot allò que he necessitat del desconegut camp de la biologia, i l'Alex Sánchez, qui m'ha sabut guiar per anar donant forma a aquest treball. També vull dirigir un gran agraïment a l'Alba Navarro, amb qui he treballat colze amb colze durant aquest llarg camí. I per suposat, a l'Elías Campo, qui m'ha donat l'oportunitat de formar part d'un equip excepcional. Per últim, a la gent del laboratori, que els seus dubtes s'han convertit en el meu aprenentatge.

A nivell personal, no puc estar més agraït amb la Laura, qui m'ha acompanyat des dels meus primers passos en l'estadística, m'ha donat més suport que ningú i m'ha fet millor persona a cada pas. També estic profundament agraït al petit Biel, qui va venir en el moment que més ho necessitava per donar-me l'impuls final per acabar aquesta etapa. A la Rosa i el Tomás, que gràcies a la seva ajuda he pogut dedicar més temps al Biel i a aquesta tesi del que m'hagués pogut permetre. Finalment, no pot faltar un sincer agraïment als meus pares, Anna i Salvador, i germans, Arnau i Ferran.

Índex de Continguts

Índex de Figures

Índex de Taules

Abreviatures

A	Adenina
B-CLPD	Síndromes limfoproliferatius de cèl·lules B
B-CLPD, NOS	B-CLPD, *not otherwise specified*
C	Citosina
CBS	*Circular binary segmentation*
cDNA	DNA complementari
CLL	Leucèmia limfàtica crònica
cMCL	MCL convencional
cRNA	RNA complementari
DLDA	Anàlisi lineal discriminant diagonal
DNA	Àcid desoxiribonucleic
FC	*Fold-Change*
FDR	*False discovery rate*
FISH	*Fluorescence in situ hybridization*
FL	Limfoma fol·licular
fRMA	*Frozen* RMA
FT	*Fluorescence threshold*
FWER	*Family-wise error rate*
G	Guanina
GEO	*Gene Expression Omnibus*
HCL	Tricoleucèmia
HCLv	Tricoleucèmia variant
HTT	Tecnologies *high-throughput*
IQR	Rang interquartílic
LAR	*Least angle regression*
lasso	*Least absolute shrinkage and selection operator*

LDA	Anàlisi lineal discriminant
LPL	Limfoma limfoplasmacític
LRR	*Log R Ratio*
LS-SVM	*Least squares* SVM
LTT	Tecnologies *low-throughput*
MCL	Limfoma de les cèl·lules del mantell
mRNA	RNA missatger
NCBI	*National Center for Biotechnology Information*
NMC	*Nearest mean classification*
nnMCL	MCL no-nodal
NSC	*Nearest shrunken centroids*
PAM	*Prediction Analysis for Microarrays*
qPCR	*Quantitative polymerase chain reaction*
RF-VIMP	*Random forests variable importance*
RMA	*Robust multiarray analysis*
RMSE	Error quadràtic mitjà
RNA	Àcid ribonucleic
ROC	*Receiver operating characteristic*
SDRPL	Limfoma esplènic de la polpa vermella
SMZL	Limfoma esplènic de la zona marginal
SVM	*Support vector machine*
SVM-RFE	SVM *recursive feature elimination*
T	Timina
U	Uracil
UHR	*Universal Human Reference*
WHO	*World Health Organization*

1 Introducció

Els síndromes limfoproliferatius crònics de cèl·lula B (B-CLPD) engloben diverses entitats de leucèmies i limfomes que s'originen en les cèl·lules B madures. El pronòstic i tractament depèn de l'entitat del síndrome, el qual es pot considerar indolent (creixement lent) o agressiu. El diagnòstic d'aquestes entitats es pot establir fàcilment en la majoria dels casos mitjançant característiques morfològiques, fenotípiques i moleculars. Tot i així, dintre de cada una hi pot haver variacions en l'evolució clínica, la morfologia i les lesions genètiques. Aquesta heterogeneïtat dificulta el diagnòstic dels pacients, el qual posa de manifest la falta d'un marcador específic per tal d'identificar l'entitat de cada pacient. La tecnologia dels *microarrays* ha obert la porta a descobrir marcadors genètics específics de cada entitat, facilitant el diagnòstic en pacients que a nivell clínic no mostren característiques úniques d'aquesta.

1.1 Síndromes limfoproliferatius crònics de cèl·lula B

Les neoplàsies hematològiques són malalties que afecten la sang, la medul·la òssia i els diferents teixits limfoides (ganglis limfàtics, melsa i timus). Aquests tres sistemes estan units mitjançant el sistema immunològic. Antigament, les neoplàsies es dividien segons: leucèmies, en cas que afectessin la medul·la òssia i la sang, o limfomes, en cas que presentessin afectació en els ganglis. En l'actualitat s'ha descobert que la separació entre limfoma i leucèmia no és tan clara ja que tant un com l'altre poden afectar els tres

sistemes. Per aquest motiu, el diagnòstic no depèn de la localització sinó del tipus de cèl·lula.

La classificació de la *World Health Organization* (WHO) [1] recull i descriu moltes formes de neoplàsies descobertes des de l'any 1832, quan Thomas Hodgkin va descriure per primera vegada un càncer de sang [2]. Una d'aquestes agrupacions distingeix les neoplàsies que afecten les cèl·lules B madures, com és el cas de la leucèmia limfàtica crònica (CLL), el limfoma fol·licular (FL), la tricoleucèmia (HCL), el limfoma limfoplasmacític (LPL), el limfoma de cèl·lules del mantell (MCL), el limfoma esplènic de polpa vermella difusa (SDRPL) i el limfoma esplènic de zona marginal (SMZL).

No hi ha una única causa coneguda comuna per tots els subtipus o entitats de neoplàsies, però en tots els casos són el resultat de mutacions o alteracions en el codi genètic [3]. Fins i tot dintre d'un mateix subtipus de síndrome les causes que fan desenvolupar la malaltia poden variar [4,5]. El correcte diagnòstic del subtipus de limfoma és especialment important degut a les diferències en el tractament de cada un, així com en el pronòstic del pacient. Tot i que hi ha heterogeneïtat en l'evolució clínica pacient a pacient, alguns dels subtipus presenten un comportament clínic més agressiu [6] mentre d'altres presenten un comportament més indolent, on els pacients poden arribar a no rebre tractament durant anys [7].

Actualment, el diagnòstic dels diferents subtipus de B-CLPD es basa en diferents criteris recollits en la classificació WHO [1]:

- **Morfològics**: trets estructurals de les cèl·lules, com per exemple, la grandària de les cèl·lules o la forma del nucli d'aquestes.
- **Fenotípics**: presència de diverses proteïnes a la membrana de les cèl·lules tumorals.
- **Moleculars**: presència de mutacions en un gen concret o d'anomalies cromosòmiques, com per exemple, translocacions (desplaçament d'un segment de cromosoma a un altra localització del genoma).
- **Clínics**: localitzacions afectades per la malaltia. Per exemple, si afecta només la

sang o també altres teixits.

Habitualment aquests trets són suficients per diagnosticar el subtipus de síndrome d'un pacient, però un 15% dels pacients presenten variacions dificultant o impossibilitant un diagnòstic definitiu, ja sigui per la mancança d'un tret específic d'un dels subtipus o la presència de trets específics de diferents subtipus [8–10]. Aquest subgrup de casos se'ls categoritza com B-CLPD, *not otherwise specified* (B-CLPD, NOS*)*.

1.2 Breu introducció a la genètica: DNA, RNA i proteïnes

L'àcid desoxiribonucleic (DNA) codifica la informació necessària que necessita un organisme viu per funcionar i desenvolupar-se. L'estructura del DNA es pot interpretar com una llarga cadena de lletres (nucleòtids) unides. Cada nucleòtid es representa amb una de quatre possibles lletres (A, C, G o T), la qual depèn de la base nitrogenada d'aquest (A: Adenina, C: Citosina, G: Guanina, T: Tiamina). La disposició seqüencial d'aquestes lletres en paraules (gens) al llarg de la cadena és el que codifica la informació genètica. A la vegada, aquestes paraules (gens) s'estructuren en 24 paràgrafs (cromosomes) per formar el genoma complet d'un humà. Cada cèl·lula del cos humà conté en el seu nucli el genoma complet, el qual té més de 3000 milions de bases i aproximadament 20000 gens.

El DNA està format per dues cadenes lligades entre elles en forma de doble hèlix creant parells de bases [11]. Aquestes dues cadenes estan lligades de forma determinista, aparellant les bases A amb bases T, i les bases C amb bases G. Per tant, coneixent una cadena es pot deduir fàcilment la cadena complementària. Aquesta estructura s'aprofita en el procés de replicació del DNA. A la Figura 1.1 hi ha una representació gràfica simplificada d'una cèl·lula.

Les instruccions codificades en el DNA s'han de transportar del nucli de la cèl·lula al citoplasma, que és la part de la cèl·lula que envolta el nucli. Al citoplasma hi ha els ribosomes, que són unes molècules especialitzades en produir i sintetitzar les proteïnes

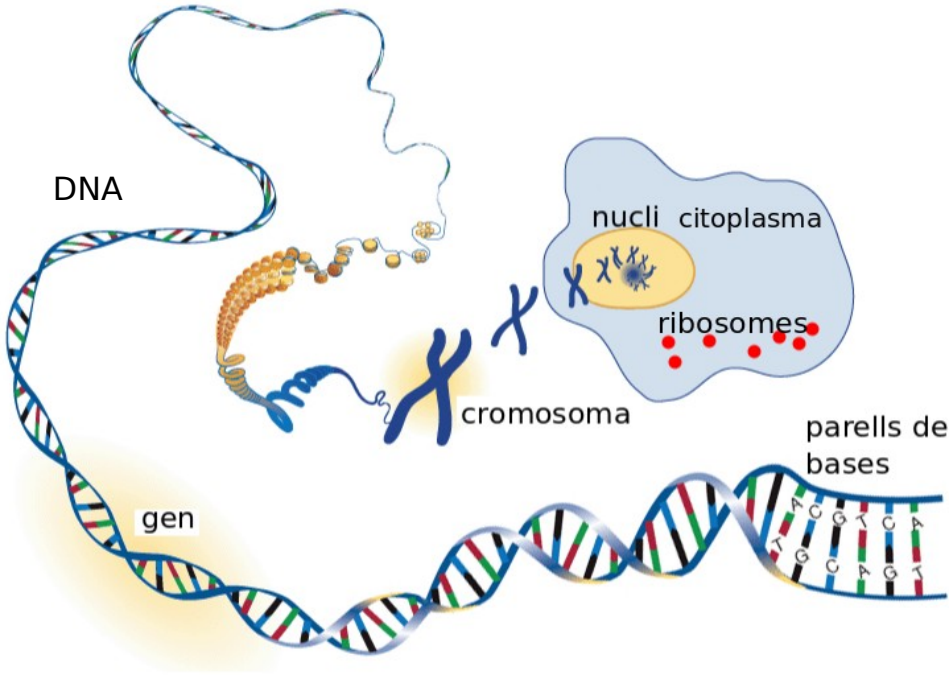

Figura 1.1: Representació gràfica d'una cèl·lula. *En la figura hi ha representat com s'estructuren les bases, els gens i els cromosomes en la cèl·lula, juntament amb la localització dels ribosomes en el citoplasma. Imatge adaptada del National Human Genome Research Institute.*

necessàries. Per tal de que la informació continguda en el DNA pugui fer aquest viatge, es transcriuen seccions de DNA en àcid ribonucleic missatger (mRNA), i és l'mRNA el que porta la informació genètica als ribosomes per tal de generar les proteïnes. No tota la informació del DNA s'inclou a l'mRNA, sinó que a través d'un procés anomenat *splicing* es suprimeixen seqüències de nucleòtids contingudes dintre dels gens (introns), a la vegada que es lliguen els extrems de les seqüències no suprimides (exons). L'mRNA també es diferencia del DNA per només tenir una cadena en la qual s'ha substituït la Tiamina (T) per l'Uracil (U).

Tot i que totes les cèl·lules del cos humà contenen còpies del mateix genoma, cada cèl·lula requerirà diferents quantitats de cada proteïna per realitzar la funció que li pertoca. Els gens que controlen la producció d'aquestes proteïnes treballaran més o menys depenent de les necessitats de la cèl·lula. Quan hi ha molta quantitat d'un gen en l'mRNA es diu que el gen està expressat, si hi ha poca quantitat es diu que no està expressat i la cèl·lula no produeix la proteïna que el gen codifica. Una explicació més detallada sobre genètica es pot trobar a Alberts et al. [12].

1.3 Tecnologies d'anàlisi del genoma

L'estudi dels diversos nivells d'informació genètica (DNA, mRNA i proteïna) és molt important per entendre, diagnosticar i tractar les malalties d'origen genètic. Per exemple, en el codi genètic hi pot haver la informació necessària per saber el pronòstic d'un malalt de càncer o si el pacient serà resistent a un cert tractament. Per tal d'identificar aquests biomarcadors fa falta mesurar, de manera objectiva i numèrica, el que està passant als diferents nivells d'informació. En les últimes dècades han aparegut diverses metodologies i tecnologies amb la capacitat de quantificar el DNA, l'RNA o les proteïnes [13].

1.3.1 *Microarrays* d'expressió

Els *microarrays* s'han utilitzat extensivament en recerca biomèdica des de l'any 1981 [14], degut a la seva capacitat de mesurar en una mostra diverses regions del genoma simultàniament en un únic experiment. Per exemple, el *microarray GeneChip® Human Genome U133 Plus 2.0* mesura més de 50000 regions al llarg del genoma. Tenint en compte que el nombre de gens en un humà està al voltant de 20000, poder mesurar-los tots (o gairebé tots) en un únic experiment facilita la identificació de biomarcadors candidats en gran mesura. El que antigament podria suposar mesos o anys de realitzar experiments actualment es pot fer en hores o dies.

La tecnologia dels *microarrays* aprofita el fet que el DNA estigui format per dues cadenes, una complementària a l'altra. Un *microarray* es construeix enganxant en un xip seqüències de DNA d'una sola cadena, anomenades *probes* o sondes, que habitualment representen un gen concret. Amb el xip construït, es processa una mostra de cèl·lules en la que es vulgui mesurar l'expressió gènica i se n'extrau l'mRNA. Mitjançant aquest mRNA es sintetitza DNA complementari (cDNA), el qual s'utilitza per transcriure RNA complementari (cRNA). Finalment, es marca el cRNA amb fluorescència, es fragmenta i s'exposa al *microarray*. En un procés anomenat hibridació, les seqüències del cRNA de la mostra de cèl·lules s'uniran a la seva part complementària del xip (Figura 1.2).

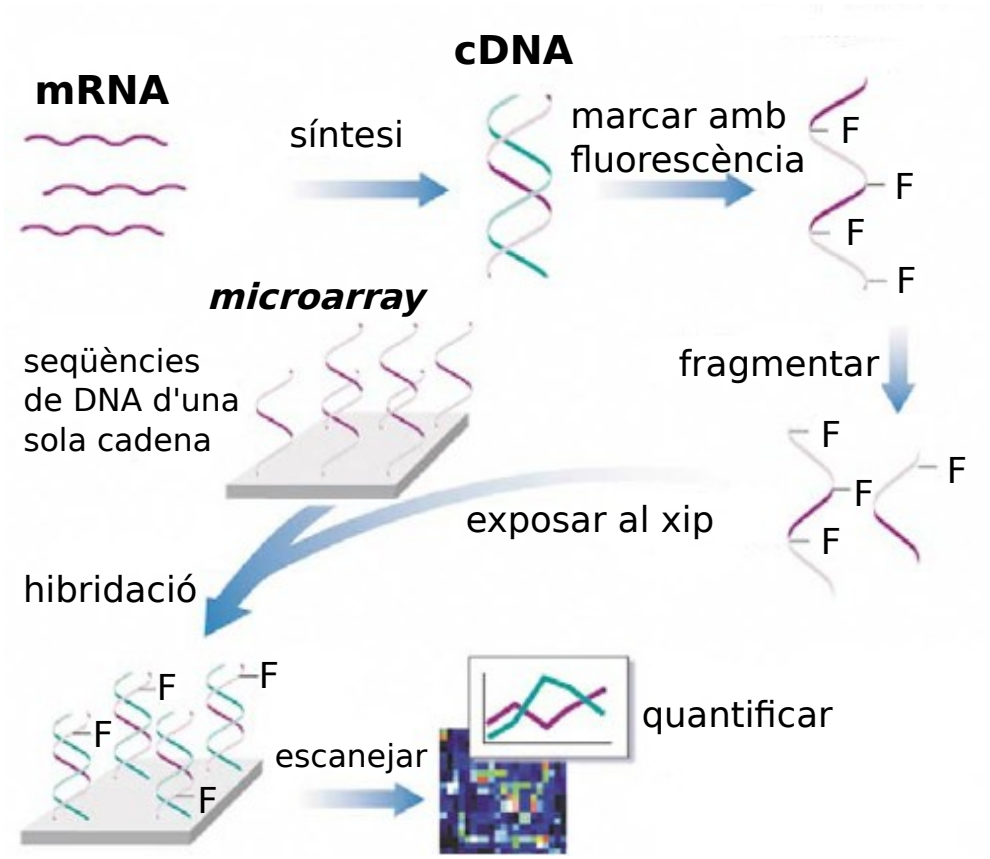

***Figura 1.2: Procés experimental d'un microarray**. En la figura es pot veure els diversos passos experimentals que s'han de realitzar per mesurar l'expressió gènica d'una mostra. Imatge adaptada d'Affymetrix.*

S'observarà una fluorescència major si una mostra de cèl·lules conté nivells elevats d'una seqüència concreta d'un gen, indicant que aquell gen està sobreexpressat. Si la mostra té una mancança d'aquella seqüència, aleshores la fluorescència serà baixa, indicant que el gent no està expressat. Una descripció més detallada es pot trobar a Bumgarner [15].

1.3.2 *Microarrays* de *copy-number*

A diferència dels *microarrays* d'expressió, que avaluen si un gen està present a l'mRNA per tal de realitzar la seva funció, els *microarrays* de *copy-number* serveixen per avaluar si en les cèl·lules analitzades hi ha guanys o pèrdues de material genètic. Aquests *microarrays* avaluen el genoma a nivell de DNA en lloc de a nivell d'RNA. La

construcció d'un *microarray* de *copy-number* és molt similar a la del *microarray* d'expressió, amb la diferència que en aquest cas s'hibrida el DNA en comptes del cRNA. Un cop s'hibrida la mostra als segments de DNA del xip, se'n mesura el senyal fluorescent. Un senyal baix significarà que el pacient té una pèrdua en aquell segment, i en cas que sigui elevat significarà que el pacient té un guany en aquell sement de DNA.

1.3.3 *Quantitative polymerase chain reaction*

La *quantitative polymerase chain reaction* (qPCR), també anomenada *real-time polymerase chain reaction*, és una tècnica que serveix per quantificar l'expressió en l'RNA [16]. El propòsit és el mateix que el dels *microarrays* d'expressió, però s'utilitza quan l'interès està en mesurar l'expressió gènica d'un únic gen o d'un petit subconjunt de gens.

Com en el cas dels *microarrays*, el primer pas és convertir l'RNA de la mostra en cDNA. Un cop convertit, el funcionament d'aquesta tècnica es basa en forçar la replicació d'una seqüència concreta de cDNA (amplificar-la). El procés d'amplificació és necessari per tal de poder comparar mostres diferents, atès que el senyal de la seqüència seria indetectable si no s'amplifiqués. A més, com que es produeix al mateix ritme en les dues mostres, és possible calcular quina té més quantitat de la seqüència d'interès en el moment inicial.

Per poder forçar que el cDNA s'amplifiqui, la qPCR utilitza l'enzim Taq polimerasa. Gràcies a aquest enzim es pot provocar que una certa seqüència d'interès, anomenada *primer*, comenci a sintetitzar una nova cadena de DNA. Aleshores, es marca la seqüència amb fluorescència i es provoquen 40 cicles de rèpliques, on a cada cicle es duplica la seqüència. Aquest procés de replicació es segueix, mesurant la quantitat de fluorescència generada a cada cicle, la qual va augmentant a mesura que les còpies de la seqüència augmenten. Si la mostra inicial està enriquida amb la seqüència en estudi, l'amplificació generarà ràpidament molta fluorescència, mentre que si la mostra no té la seqüència, la fluorescència trigarà més cicles a tenir nivells alts.

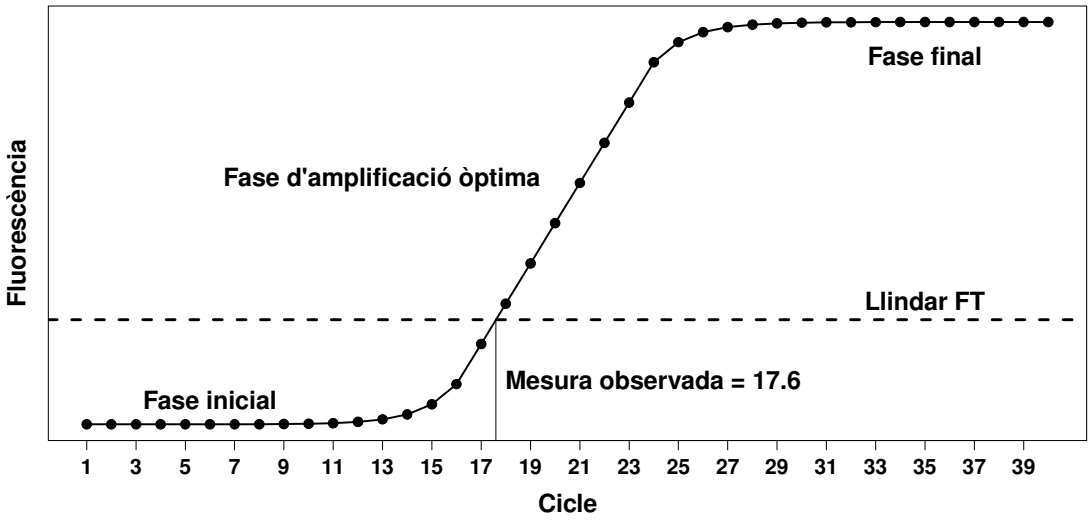

***Figura 1.3: Corba d'amplificació (qPCR)**. En la fase inicial d'una qPCR la seqüència en estudi es va replicant però la fluorescència encara és massa baixa per ser detectada. En la segona fase l'amplificació és òptima i la fluorescència es detecta clarament. La fase final comença quan la quantitat de components restants és limitada, reduint el ritme de replicació fins al punt que ja no es replica més i la fluorescència es manté estable.*

El valor numèric que s'analitza és el nombre de cicles que han fet falta per a que la quantificació de la fluorescència superi un cert llindar, anomenat *fluorescence threshold* (FT). A la Figura 1.3 es mostra la corba teòrica del procés d'amplificació en 40 cicles per una única mostra, on hi ha una fase inicial en què l'emissió de fluorescència no supera l'emissió de fons, després una fase en què l'amplificació és òptima (exponencial), per acabar amb la fase final en què els components de la reacció són limitats i la fluorescència deixa de créixer. Una descripció més detallada d'aquesta tècnica es pot trobar a Kubista et al. [17].

A canvi de només poder mesurar uns pocs gens o seqüències a la vegada, la qPCR té diferents característiques que la fan molt popular, entre les quals hi ha la simplicitat, la poca duració de l'experiment, el poc material cel·lular necessari i el reduït cost econòmic.

1.4 Consideracions estadístiques en l'entorn de la genètica

La utilització dels *microarrays* en la investigació clínica ha canviat l'estructura de les dades obtingudes en aquest entorn, on disposar de desenes de milers de mesures per a un únic pacient és habitual. Aquest canvi ha provocat que els mètodes estadístics tradicionals no puguin fer front a les noves necessitats. El principal problema que han hagut d'afrontar els analistes de dades és el conegut com la *maledicció de la dimensionalitat*. El cost d'un *microarray* és elevat i habitualment només és podrà utilitzar en un nombre reduït de pacients. Si a aquesta limitació li afegim que els *microarrays* generen milers d'observacions per cada pacient, l'analista s'ha d'enfrontar a situacions on el nombre de variables és molt més gran que el nombre de pacients. Tant si l'enfoc de l'anàlisi és univariant (un gen a la vegada), com si és multivariant (múltiples gens conjuntament) hi ha diversos aspectes a considerar a l'hora d'analitzar les dades.

1.4.1 Tests múltiples

Quan es realitza una prova d'hipòtesi es poden cometre dos errors:

- **Error de tipus I o fals positiu**: rebutjar la hipòtesi nul·la (H_0) quan H_0 és certa. La probabilitat de cometre aquest tipus d'error és α, què habitualment es fixa a 0.05.
- **Error de tipus II o fals negatiu**: no rebutjar H_0 quan H_0 és falsa. La probabilitat de cometre aquest tipus d'error és β.

El problema dels tests múltiples apareix quan es realitzen múltiples proves d'hipòtesi simultàniament. A l'apartat 1.3.1 s'ha explicat que els *microarrays* d'expressió mesuren aproximadament $M = 50000$ punts al llarg del genoma. Si en cada un dels punts es compara l'expressió entre dos grups s'estaria fent 50000 proves d'hipòtesi separades. Suposant que en cap punt del genoma hi ha diferències entre grups (H_0 certa), s'esperaria trobar $\alpha \cdot M = 0.05 \cdot 50000 = 2500$ fals positius. Aquest augment en el risc de

cometre errors és el problema dels tests múltiples.

Recollir moltes variables en un estudi és desitjable, ja que permet obtenir molta més informació, però la contrapartida és l'augment de resultats falsos. En concret, si es realitzen *M* proves d'hipòtesi independents, on H_0 és certa en *K* de les *M* proves, la probabilitat de cometre almenys un error de tipus I és de $1 - (1- \alpha)^K$. Aquesta probabilitat es coneix com *family-wise error rate* (FWER). La Figura 1.4 mostra com augmenta l'FWER a l'augmentar *K* per una α fixada a 0.05. En el gràfic es pot veure com l'FWER ràpidament supera el llindar del 50%, fins i tot amb només 70 proves d'hipòtesi ja hi ha quasi un 100% de probabilitat de que algun dels *K* tests reporti un fals positiu. Si realitzem suficients tests on H_0 és certa, fàcilment en trobarem algun amb un *P*-valor per sota d'α.

Existeixen procediments que ajusten el *P*-valor de manera que es controla l'FWER en comptes de l'error de tipus I, d'aquesta manera s'evita l'excés de falsos positius. Alguns d'aquests procediments són el de Bonferroni, el de Šidák [18] o el de Holm [19]. Aquests procediments disminueixen el valor d'α de forma proporcional a *M*, de manera que, independentment de la quantitat de proves d'hipòtesi realitzades, l'FWER sigui com a màxim igual a un valor prefixat (per exemple, 0.05). El cost de controlar l'FWER i realitzar menys errors de tipus I és que s'augmenta el risc a obtenir falsos resultats negatius (error tipus II). Per exemple, si es controla l'FWER mitjançant el mètode de Bonferroni quan només un test de 50000 correspon a H_0 falsa, aquest test hauria d'obtenir un *P*-valor $< 10^{-6}$ per declarar-lo significatiu.

En l'entorn de la genètica, on s'estudien milers de covariables, és habitual que una de les anàlisis sigui identificar tots els gens que en mitjana s'expressen diferent entre dos subgrups. L'objectiu és identificar patrons de gens amb funcions determinades i entendre les bases genètiques de la malaltia. En aquesta situació, controlar l'FWER provocaria que no s'identifiquessin molts dels gens amb expressió diferencial (error tipus II). En canvi, no fer cap tipus de control provocaria que molts gens sense expressió diferencial donessin resultats significatius. Per aquest motiu s'acostuma a controlar el *false discovery rate* (FDR), el qual es defineix com $FDR = E[V/(V+S)]$, on *V* és el

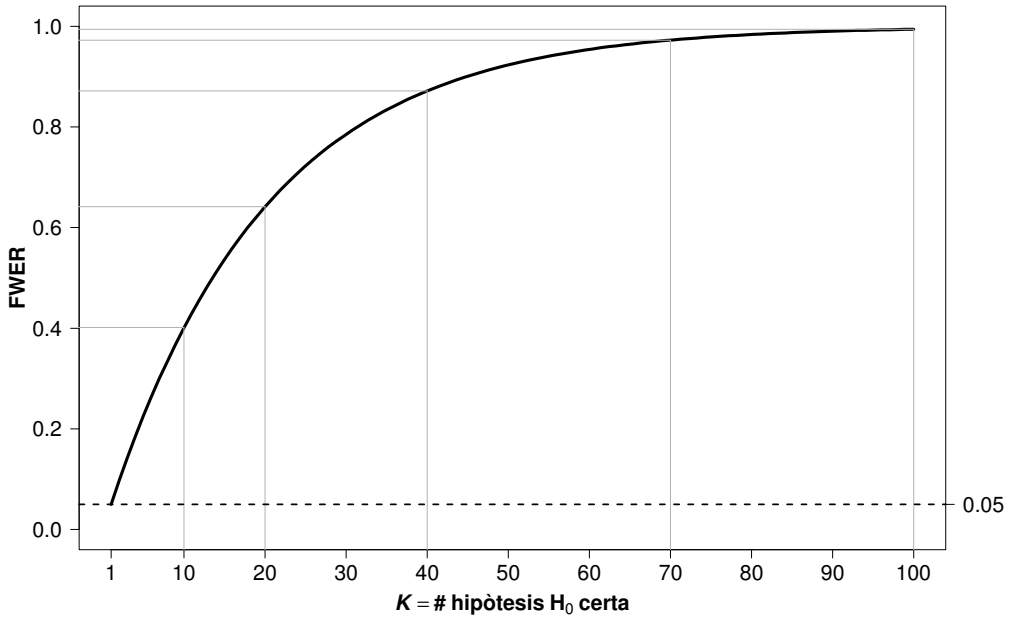

***Figura 1.4: FWER segons el número d'hipòtesis en què H_0 és certa**. Probabilitat d'obtenir almenys un P-valor < 0.05 al realitzar K proves d'hipòtesi en què H_0 és certa.*

nombre de falsos positius i *S* el nombre de veritables positius. És a dir, un FDR del 5% vol dir que s'espera que només un 5% dels tests en els que s'ha rebutjat la hipòtesi nul·la siguin incorrectes, mentre que el 95% restant serien veritables positius. Controlar l'FDR serveix per buscar un compromís entre els errors de tipus I i de tipus II. El procediment més habitual per controlar l'FDR és el de Benjamini-Hochberg [20]. Una revisió de diferents metodologies per controlar l'FWER i l'FDR es pot trobar a Dudoit et al. [21].

1.4.2 Restricció de mètodes

L'anàlisi lineal discriminant (LDA) [22] possiblement sigui el mètode de classificació més utilitzat degut a la seva senzillesa i a tenir una precisió similar a mètodes més complexos. En la formulació de l'LDA és necessari invertir la matriu de covariàncies, la qual s'estima directament de les dades quan és desconeguda. En les situacions en què el nombre de variables (p) és més gran que el nombre de mostres (n), la matriu estimada no té rang complet, provocant que no sigui invertible i, per tant, que no es pugui aplicar el mètode.

L'LDA no és la única metodologia que no es pot aplicar en situacions de $p>n$. Altres mètodes clàssics, com la família dels models lineals generalitzats, tampoc es poden utilitzar donat que no poden incloure més variables que mostres en la modelització. Per aquest motiu, en l'entorn de la genètica s'ha de treballar amb mètodes que tolerin més variables que mostres o amb modificacions dels mètodes que no ho toleren. Un exemple de mètode que tolera qualsevol quantitat de variables és el s*upport vector machine* (SVM) [23]. Un exemple de modificació simple de l'LDA és combinar-lo amb una metodologia *stepwise*.

1.4.3 *Overfitting* i selecció de variables

El sobreajust, o *overfitting* en anglès, és un problema estadístic que es troba en tots els entorns, i fa referència a que un model estadístic està modelant soroll (error aleatori) en comptes del procés real. En concret, el model construït conté massa paràmetres i s'ajusta molt bé a les dades que s'han utilitzat per construir-lo (*training set*) però no a les dades no-vistes. La principal conseqüència de l'*overfitting* és que incrementa els errors de predicció esperats del model ajustat.

En l'entorn de la genètica aquest problema s'amplifica degut a l'elevat nombre de variables disponibles per construir el model estadístic. Per exemplificar aquest fenomen s'ha utilitzat un escenari simulat, on es mostra com s'incrementa l'error quan creix el nombre de variables no-informatives recollides. En concret, s'han simulat 300 *training sets* de 50 mostres i M covariables (X_1, …, X_M), on cada covariable es distribueix segons una Normal(0,1). El model teòric de la variable resposta utilitzat ha sigut: $Y = 0.6X_1 + 0.4X_2 + 0.2X_3 + E$, on E és l'error aleatori irreductible, també distribuït segons una Normal(0,1). D'aquesta fórmula es pot veure que només s'ha donat informació sobre Y a 3 de les M covariables X_j (X_1, X_2 i X_3). Un cop simulats els 300 *training sets* s'ha ajustat, en cada un, un model lineal utilitzant les 5 covariables X_j més correlacionades amb Y. Per últim, s'ha calculat i guardat l'error quadràtic mitjà (RMSE) teòric de cada un dels 300 models ajustats. Aquest procediment s'ha dut a terme per quatre valors de M = (10, 100, 1000, 10000).

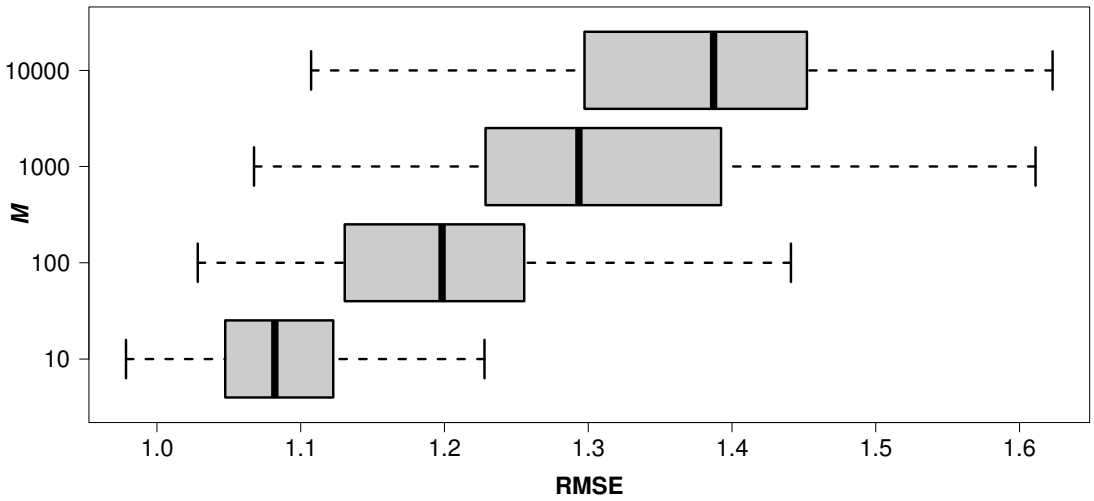

***Figura 1.5: Efecte del nombre de variables no-informatives a l'RMSE**. Cada boxplot representa l'RMSE teòric de 300 models construïts amb training sets simulats d'M covariables. Tots els models s'han construït amb 5 de les M covariables.*

A la Figura 1.5 hi ha representat els *boxplots* dels RMSE obtinguts corresponent a cada *M*, on es pot veure com l'RMSE augmenta quan també ho fa el nombre de variables no-informatives. Aquest efecte és degut a que, quan es recull una gran quantitat de variables no-informatives, s'augmenta la probabilitat de que alguna correlacioni per atzar amb *Y*. En conseqüència, el model ajustat utilitza una variable que en el *training set* funciona però no en dades no-vistes.

Una manera d'aconseguir models predictius menys propensos a l'*overfitting* és seleccionant un subconjunt de variables rellevants amb les quals construir-lo [22]. A part de reduir el risc d'*overfitting*, aquesta selecció també serveix per obtenir models més generalitzables, senzills i fàcils d'interpretar per l'investigador. Com en l'exemple de la Figura 1.5, s'ha utilitzat dades simulades per mostrar com afecta la selecció de variables a l'error de predicció esperat. Per fer-ho, s'han simulat 100 *training sets* de 40 mostres i 1000 covariables (X_1, …, X_{1000}). En aquest cas la variable resposta *Y* és binària, i 20 mostres corresponen a la classe A i 20 a la B. Les tres primeres covariables (X_1, X_2, X_3) es distribueixen segons una Normal(0,1) en la classe A i segons una Normal(1,1) en la classe B. La resta de covariables (X_4, …, X_{1000}) es distribueixen segons una Normal(0,1) en ambdues classes. Per tant, només tres variables tenen informació

rellevant sobre Y. Seguidament, en cada *training set* s'ha construït un model predictiu utilitzant les k = (2, 3, 10, 100, 1000) covariables X_j més associades amb Y. S'ha utilitzat el mètode SVM per construir els models, ja que tolera qualsevol quantitat de variables. L'associació entre les covariables X_j i Y s'ha mesurat amb l'estadístic T. Per últim, s'ha calculat i guardat l'error de classificació teòric de cada un dels models ajustats.

La Figura 1.6 mostra els *boxplots* d'aquests errors pels diferents valors de k. Al gràfic es pot veure com l'error de classificació es redueix considerablement quan es limiten el nombre de variables amb les quals es construeix el model, disminuint del 45% (k = 1000) al 30% (k = 2).

En la Figura 1.6 també es pot veure que l'error obtingut per k = 2 és més elevat que per k = 3, el qual és degut a que en l'escenari utilitzat hi havia tres variables amb informació (X_1, X_2, X_3). Aleshores, els models de tres variables han pogut incloure, en mitjana, més variables informatives. Quan s'utilitzen menys variables informatives s'obté models infraajustats, mentre que si s'utilitzen més variables no-informatives s'obtenen models sobreajustats, aquest problema és conegut en anglès com el *bias-variance tradeoff* [24]. Una metodologia que no identifica suficients variables predictores relacionades amb la variable resposta tindrà error degut al biaix (*bias*), mentre que una metodologia que tendeix a utilitzar moltes variables no-informatives tindrà error degut a la variància (*variance*). Qualsevol metodologia aplicada a un *training set* patirà aquests dos errors. L'analista de dades ha de trobar la metodologia que permeti obtenir un model en el qual la suma dels dos errors sigui el mínim possible, és a dir, un model que contingui el màxim número de variables rellevants i, a la vegada, el mínim número de variables no-informatives.

No solament la metodologia de selecció de variables afecta a l'intercanvi entre biaix i variància, sinó que la flexibilitat del model a l'hora de relacionar les variables predictives amb la resposta també hi intervé. Per exemple, un model de regressió lineal assumeix una relació lineal entre la resposta i el predictor, una estructura molt rígida que pateix de biaix quan el model que ha generat les dades no compleix aquesta relació, però amb poca variància donat que en repeticions successives del mateix experiment

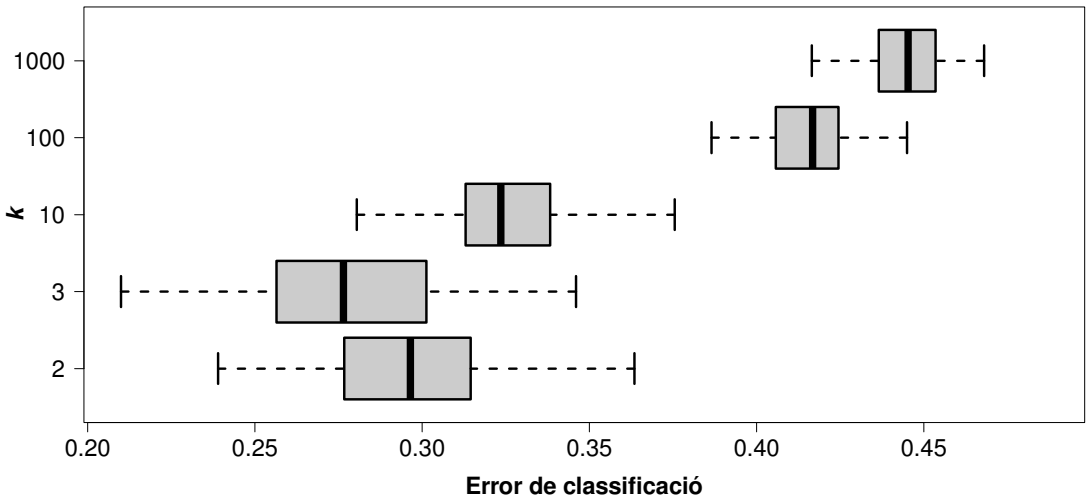

Figura 1.6: Efecte de la selecció de variables a l'error de classificació. *Cada boxplot representa l'error de classificació teòric dels 100 models construïts amb k covariables. Cada model s'ha construït amb un training set simulat de 1000 covariables.*

s'obtindrien models similars. Si s'utilitza una metodologia més flexible, com són els mètodes de regressió local, els models obtinguts ajustaran la relació entre la variable predictora i la resposta correctament en mitjana, però repeticions successives obtindrien models molt diferents. En resum, un model amb molts paràmetres efectius (flexible i amb moltes covariables) patirà majoritàriament de variància, mentre un model amb pocs paràmetres efectius (rígid i poques covariables) patirà de biaix [24].

1.4.4 Trasllat de plataforma

Els *microarrays* formen part del conjunt de tecnologies que generen dades d'alt rendiment (*high-throughput*, HTT). Les HTT permeten estudiar milers de variables o gens simultàniament, generant una gran quantitat d'informació sobre una malaltia en un únic experiment i facilitant l'etapa d'*screening* per identificar possibles biomarcadors. Un inconvenient de les HTT és el seu elevat cost econòmic, el qual dificulta aplicar-les a la rutina clínica per a cada pacient. Si es limita el nombre de gens interessants a una petita quantitat, es poden utilitzar tecnologies més barates i àgils (*low-throughput*, LTT) per realitzar mesures en els pacients.

Agafant d'exemple els *microarrays* d'expressió, el següent procés de tres etapes és el que habitualment es segueix per crear un biomarcador utilitzable a nivell clínic [25–31]:

1) **Descobriment de biomarcadors candidats**: mitjançant *microarrays* (HTT) es mesura l'expressió de milers de gens en el *training set*. S'identifica un subconjunt de gens amb bones capacitats predictives.
2) **Trasllat a qPCR:** en el *training set* es mesura, mitjançant qPCR (LTT), l'expressió del subconjunt de gens seleccionats en la primera etapa. Amb aquestes mesures es construeix el predictor final.
3) **Validació:** en una sèrie de validació independent de pacients es mesura, mitjançant qPCR, l'expressió del subconjunt de gens inclosos en el predictor de la segona etapa. S'aplica el predictor en aquesta nova sèrie per comprovar la seva validesa.

Aquest procés afegeix una nova dimensió al problema de crear un predictor en aquest entorn en el qual $p>>n$, donat que la tecnologia utilitzada per seleccionar el subconjunt de variables és diferent a la tecnologia utilitzada per la creació del predictor. Les dues principals conseqüències d'aquest canvi de tecnologia són:

- **Limitació del nombre de variables:** per tal de poder fer el canvi de *high* a *low-throughput* el nombre de variables a mesurar s'ha de limitar. La quantitat depèn de diferents factors, però habitualment no arriba a la desena i rarament sobrepassa la cinquantena de variables.
- **Correlació entre tecnologies**: tot i que les HTT i les LTT mesuren el mateix, no ho fan exactament de la mateixa manera. Aquestes diferències en les especificacions tècniques provoquen que la correlació entre diferents tecnologies no sigui perfecta i, en alguns casos puntuals, pot arribar a ser nul·la [32–35].

Aquests dos punts afecten negativament a la precisió de les prediccions de les LTT respecte les HTT. Per exemple, un predictor construït amb dades de *microarrays* que minimitza l'error utilitzant 90 gens, al traslladar-lo a una LTT no solament se n'haurà de reduir el nombre incrementant l'error, sinó que algun dels gens podria tenir baixa correlació entre les dues tecnologies perjudicant encara més la precisió. Habitualment, per adreçar aquest problema es traslladen més gens dels necessaris de les HTT a les LTT

[29,30,36–38]. D'aquesta manera si algun dels gens no té un bon rendiment en les LTT es pot substituir per un altre sense necessitat de repetir tot el procés.

En el procés de selecció de gens es poden tenir en compte factors coneguts que poden influenciar la correlació entre les dues tecnologies, de manera que es maximitzi la informació que es trasllada de l'HTT a l'LTT. Per exemple, es pot prioritzar gens que tinguin un gran canvi entre les dues condicions en estudi per sobre de gens que tinguin un *P*-valor més petit [32]. També es pot prioritzar gens en què la sonda del *microarray* i la sonda de qPCR n'interroguen la mateixa seqüència.

1.4.5 Preprocessament de dades

Les mesures obtingudes de les diverses tecnologies que quantifiquen la informació genètica no es poden analitzar directament, sinó que necessiten d'una sèrie de transformacions per tal de convertir-les en dades amb sentit biològic. Un cop transformades ja es poden utilitzar per realitzar els tests estadístics corresponents. Agafant d'exemple els *microarrays*, el que s'obté de l'experiment és una imatge escanejada com la mostrada en la Figura 1.7. A partir d'aquesta imatge s'han d'aplicar múltiples procediments per tal d'extreure la informació sobre l'expressió gènica. A aquest conjunt de procediments se'ls anomena preprocessament.

El principal objectiu del preprocessament és eliminar diverses fonts de variabilitat, no relacionades amb la variabilitat biològica, que afecten les mesures obtingudes en l'experimentació. En el cas dels *microarrays* d'expressió el preprocessament pot incloure els següents passos [39]:

1) **Preprocessament d'imatge**: en aquest pas es transforma la imatge obtinguda de l'experiment en valors numèrics per cada *probe*. També es corregeix possibles defectes en la imatge.
2) **Correcció del soroll de fons**: la intensitat del senyal de cada *probe* és proporcional a l'abundància d'una seqüència d'RNA específica que hi ha en la mostra, però és habitual que també s'hi enllacin seqüències no-complementàries,

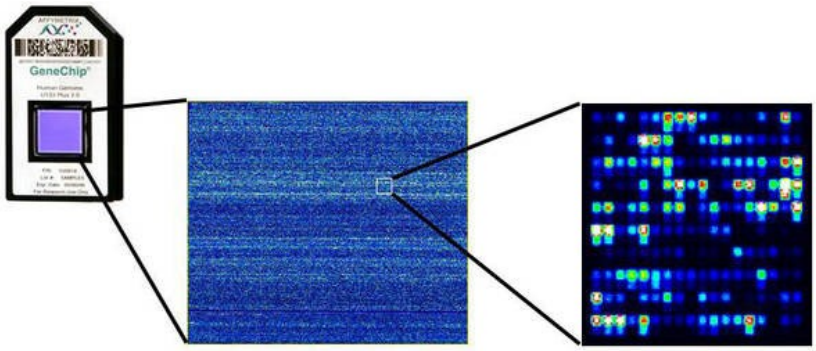

***Figura 1.7: Resultat experimental d'un microarray**. Cada punt correspon a una probe diferent. La intensitat lumínica és proporcional a l'expressió de la seqüencia de DNA que mesura la probe. Imatge obtinguda de German Cancer Research Center.*

les quals pertorben el senyal d'interès. En aquest pas s'intenta estimar el senyal degut a aquests enllaços no-específics per, posteriorment, sostreure'l. Simultàniament es corregeix la fluorescència deguda a la superfície del xip.

3) **Normalització**: en aquest pas s'intenten minimitzar les diferències no-biològiques entre diferents xips. Per exemple, els *microarrays* mesurats en un escàner concret podrien obtenir mesures de la intensitat més elevades que els *microarrays* mesurats en un escàner diferent.
4) **Resum**: en l'elaboració d'un *microarray* s'inclouen grups de *probes* que interroguen la mateixa seqüència de 25 bases d'RNA. Les *probes* d'un mateix grup es diferencien entre elles per estar desplaçades entre 0 i varies bases respecte la seqüencia d'interès. A cada grup se l'anomena *probeset*. En el resum es combinen les mesures dels grups de *probes* per obtenir una única mesura per cada *probeset*.

Existeixen múltiples metodologies per realitzar cadascun dels passos. Aquesta diversitat provoca que hi hagi més d'un centenar de combinacions diferents per realitzar el preprocessament dels *microarrays* d'expressió. Davant de la gran quantitat d'opcions, la selecció de la metodologia a utilitzar es pot basar en la literatura o en les recomanacions del fabricant. En el Capítol 2 es descriurà la metodologia que s'ha utilitzat en les dades d'aquesta tesi.

En el cas dels *microarrays* de *copy-number*, el preprocessament de les mesures obtingudes és similar al dels *microarrays* d'expressió. En aquests *microarrays* es realitza el **preprocessament d'imatge**, la **normalització** i el **resum**. La diferència més gran radica en afegir un últim pas, la **segmentació**. La principal informació que es pot obtenir dels *microarrays* de *copy-number* és identificar regions del genoma d'un pacient perdudes o amb còpies extra. Per fer-ho, s'utilitzen algoritmes de segmentació [40] que identifiquen seqüències de *probes* adjacents amb valors per damunt o per sota de l'esperat. Per exemple, si una mostra té una còpia perduda d'un segment del cromosoma 13, els valors de les *probes* que mesuren les diferents seqüències d'aquest segment quedaran distribuïdes al voltant d'un valor més baix que les de fora el segment.

La Figura 1.8 mostra els valors preprocessats, anomenats *Log R Ratio*, d'una part del cromosoma 13 d'un pacient de B-CLPD. En la figura es pot veure l'estructura de correlació espacial al llarg de les bases, *probes* consecutives tenen valors de *Log R Ratio* distribuïts al voltant del mateix valor. També es pot veure com en diversos punts del cromosoma hi ha salts en el valor mitjà del *Log R Ratio*. Els algoritmes de segmentació serveixen per identificar la regió del cromosoma en què es produeixen aquests salts.

Per últim, en el cas de la qPCR el preprocessament de dades involucra dos passos, la **selecció del llindar FT** i la **normalització** (o quantificació relativa) [41]. A l'apartat 1.3.3 s'ha explicat que el valor numèric a analitzar d'un experiment de qPCR és el nombre de cicles que la quantificació de la fluorescència necessita per superar el llindar FT. L'objectiu d'aquest llindar és distingir a partir de quin cicle la fluorescència observada en l'experiment està per damunt de la fluorescència associada al soroll de fons.

La selecció d'FT es pot realitzar de forma automàtica mitjançant algoritmes o de forma manual mitjançant gràfics [42,43]. Per fer-ho manualment s'ha de representar en un mateix gràfic les corbes d'amplificació de totes les mostres analitzades, aleshores, el llindar es fixaria en el punt on la majoria de mostres comencen la fase d'amplificació òptima (exponencial). Tot i que només hi ha representada la corba teòrica d'una única

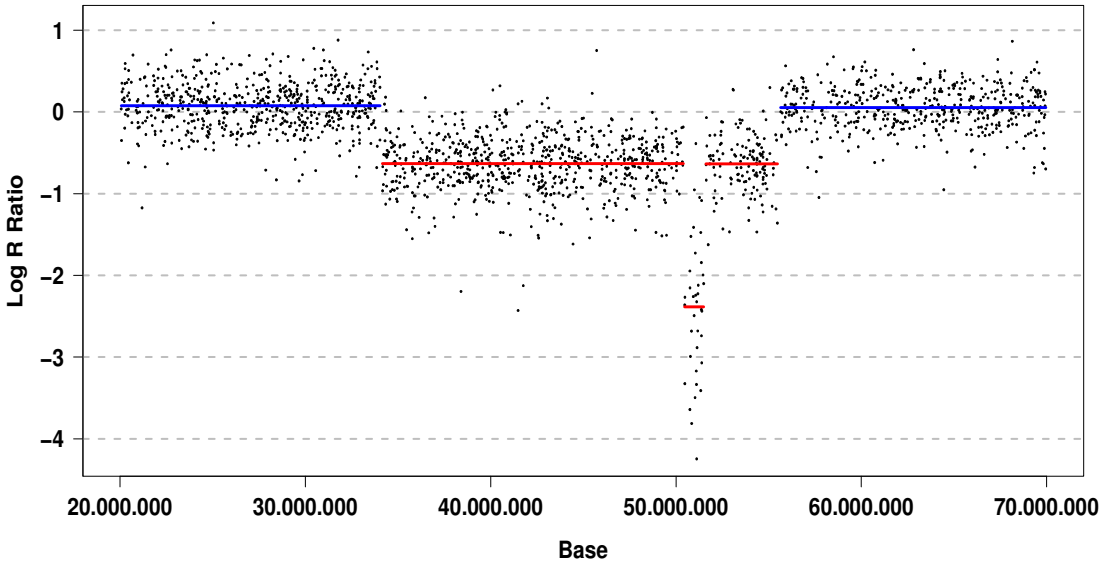

***Figura 1.8: Log R Ratio d'un microarray de copy-number**. En la figura hi ha representat els valors preprocessats (Log R Ratio) d'una part del cromosoma 13 d'un pacient de B-CLPD. Els segments marcats en vermell (Log R Ratio<0) indiquen pèrdues de material genètic en aquella regió del DNA.*

mostra, a la Figura 1.3 el llindar FT està fixat, aproximadament, al principi d'aquesta fase. En cas de mesurar diversos gens, s'ha de fixar un llindar per cada un de manera independent.

El segon pas, la normalització, té el mateix propòsit que en el cas dels *microarrays*: minimitzar fonts de variabilitat no desitjades en la mesura de l'expressió gènica. D'aquesta manera es pot comparar mostres diferents el més acuradament possible. La metodologia més utilitzada per normalitzar dades de qPCR involucra la utilització de gens *housekeeping*. Els gens *housekeeping* s'expressen a nivells constants sota diferents condicions biològiques, característica que els fa útils per mesurar i controlar factors tècnics que poden influenciar l'expressió obtinguda, com per exemple, la diferència en l'extracció de l'mRNA entre mostres o el rendiment del procés de mesura.

Tot i no formar part del preprocessament, hi ha un últim aspecte a considerar en l'anàlisi de dades de qPCR: com tractar els experiments que no han arribat a superar el llindar FT en els 40 cicles de la replicació. Aquests experiments són *missings* informatius. Tot i

que no se'ls hi ha pogut observar un valor numèric, no superar el llindar pot significar que l'expressió del gen mesurat en aquella mostra és molt baixa. No hi ha consens en com analitzar aquests experiments, l'estratègia més habitual és assignar-los el valor del nombre de cicles que s'han realitzat (per exemple 40), però existeixen algoritmes més complexes que eviten crear tant de biaix [44].

1.4.6 Control de qualitat

L'etapa d'hibridació d'un *microarray* pot no funcionar a ple rendiment, provocant que les mesures d'expressió gènica de la mostra hibridada es vegin influenciades i, per tant, no siguin comparables amb les mostres en les quals el procés d'hibridació ha funcionat correctament. Eliminar aquestes mostres de les anàlisis estadístiques millora la potència dels resultats obtinguts [45,46], aleshores, és important utilitzar eines que permetin identificar-les.

En el cas de la qPCR també és important identificar les mostres en les quals l'eficiència de la reacció és baixa o la degradació de l'RNA ha influenciat el procés d'amplificació, entre d'altres. El control de qualitat en experiments de qPCR es fa mitjançant la inspecció de les corbes d'amplificació (Figura 1.3), representades simultàniament per totes les mostres en un mateix gràfic. Aquelles mostres que tenen una corba molt diferent a la resta es marquen com a candidates a ser excloses. La inspecció de les corbes es pot fer de manera visual o mitjançant algoritmes de detecció de corbes atípiques [47,48].

2 Objectius

El treball d'aquesta tesi està emmarcat dins d'un projecte més global que analitza extensivament les diverses entitats de leucèmies i limfomes englobades en els B-CLPD. En el primer apartat d'aquest capítol es descriuen els objectius d'aquest projecte, mentre que en el segon apartat es descriuen els objectius específics de la tesi.

2.1 Objectius del projecte B-CLPD

Dintre dels B-CLPD hi ha una gran diversitat de subtipus degut, en part, a la complexitat cel·lular i funcional de les poblacions de cèl·lules en el sistema immunològic. Cada subtipus té un pronòstic i tractament diferent, però fins i tot dins de cadascun hi ha una gran heterogeneïtat en l'evolució clínica dels pacients, la biologia i el fenotipus. El propòsit del projecte B-CLPD és el d'entendre diferents aspectes d'aquests síndromes per tal de millorar el diagnòstic, el tractament i la supervivència dels pacients. Alguns d'aquests aspectes són què els origina, què provoca que un pacient progressi clínicament, quina és la causa que un pacient sigui resistent al tractament i, a nivell diagnòstic, com es poden distingir els diferents subtipus. Per contestar totes aquestes preguntes és important la identificació de gens diana rellevants en la biologia d'aquests tumors.

Els objectius del projecte són:

1) Determinar el perfil d'expressió de diversos subtipus de síndromes limfoproliferatius mitjançant *microarrays* d'expressió.
2) Caracteritzar el perfil d'alteracions genètiques dels diversos subtipus de síndromes.
3) Determinar l'impacte de les alteracions cromosòmiques sobre els perfils d'expressió.
4) Crear un model predictiu pel diagnòstic dels diferents subtipus de síndromes mitjançant dades de *microarrays.*
5) Crear un model predictiu pel diagnòstic dels diferents subtipus de síndromes mitjançant dades de qPCR.
6) Determinar si els models predictius dels punts 4 i 5 també serveixen per diagnosticar els pacients categoritzats com B-CLPD, NOS, els quals no es poden diagnosticar mitjançant els criteris descrits en la classificació de la WHO [1].
7) Identificar gens diana que puguin estar implicats en el desenvolupament i progressió tumoral.
8) Avaluar l'impacte clínic de les alteracions genètiques i els perfils d'expressió per tal de poder predir la supervivència dels pacients.
9) Avaluar l'impacte clínic de les alteracions genètiques i els perfils d'expressió per tal de poder predir la resposta al tractament.

2.2 Objectius de la tesi

En aquesta tesi s'abordaran únicament els punts 4, 5 i 6 de l'apartat 2.1. L'objectiu d'aquests tres punts és el de desenvolupar models diagnòstic predictius, mitjançant informació del DNA i l'RNA, que ajudin a diagnosticar els diferents subtipus de síndromes, especialment en aquells pacients en què els marcadors morfològics, fenotípics i moleculars no siguin suficients per donar un diagnòstic clar (B-CLPD, NOS).

Actualment, la construcció de models de predicció mitjançant dades basades en *microarrays* és relativament senzilla, donat que hi ha hagut molta recerca en aquest

camp i s'han publicat programaris que permeten aplicar diferents metodologies amb relativa facilitat [49,50]. En el cas concret d'aquesta tesi hi ha diferents consideracions a tenir en compte que dificulten la utilització d'aquests programaris. La primera és que es disposa de dues fonts d'informació, una a nivell de DNA i l'altra a nivell d'RNA, les quals s'han d'integrar per construir un predictor únic. La segona és la quantitat de síndromes diferents a discriminar, el qual afecta a la metodologia de construcció del predictor. La tercera i última consideració és que es volen construir dos predictors, un en dades de *microarrays* i un en dades de qPCR. El primer predictor serviria per aquells pacients en què hi ha interès en obtenir dades genètiques completes mitjanant *microarrays*, atès que seria una despesa de material i cost econòmic innecessari obtenir també dades mitjançant qPCR. El segon es podria implementar a nivell clínic gràcies a la senzillesa de la tècnica i a que suposa un menor cost econòmic i de material.

Tenint en compte els aspectes acabats de descriure, els objectius principals d'aquesta tesi són:

1) **Construir un model de predicció que integri la informació obtinguda mitjançant *microarrays* d'expressió i *microarrays* de *copy-number*.** Cadascuna de les fonts d'informació té una estructura única i conté milers de variables que podrien ser indicatives d'un subtipus de síndrome concret. En el procés de construcció d'aquest predictor serà important valorar si les dues fonts aporten informació rellevant, donat que cada una s'obté mitjançant una metodologia diferent, duplicant el cost econòmic i de material. La modelització haurà de tenir en compte que s'estudien fins a nou subtipus diferents de síndromes. A més, el predictor resultant ha de ser interpretable des d'un punt de vista biològic.

2) **Construir dos models de predicció basats en expressió gènica, un en dades de *microarrays* i l'altre en dades de qPCR.** El predictor obtingut del primer objectiu estarà construït amb informació de dues plataformes diferents de *microarrays*. Degut al cost econòmic, és convenient comparar el rendiment d'aquest amb el rendiment d'un predictor que només requereixi obtenir

informació de la plataforma d'expressió. De la mateixa manera, un predictor que només requerís l'expressió mesurada per qPCR tindria un cost encara més reduït i seria més senzill d'implementar a nivell clínic.

3) **Plantejar una metodologia de selecció de gens i alteracions a mesurar mitjançant qPCR.** El desenvolupament de mètodes classificadors per dades d'alta dimensionalitat s'ha abordat extensivament en els últims anys. En menor grau, també hi ha hagut interès en estudiar i desenvolupar metodologies de selecció de variables, però sempre plantejant la selecció com un procés previ a la construcció del predictor. En aquesta tesi es realitzarà la selecció d'un subconjunt de variables en dades de *microarrays*, per després traslladar-les a una tecnologia diferent (qPCR). Aquest procediment requereix un plantejament metodològic diferent.

4) **Valorar si mutacions o alteracions genètiques descrites en la literatura es poden combinar amb els models de predicció d'aquesta tesi per tal de millorar el diagnòstic.** Tot i que en la literatura no s'han estudiat tantes entitats de B-CLPD conjuntament, si que s'ha fet de manera més específica en cadascuna. D'aquests estudis es pot extreure informació que, combinada amb l'obtinguda en aquesta tesi, ajudi a diagnosticar de manera més precisa el subtipus de B-CLPD en nous pacients.

5) **Valorar si els models de predicció poden ajudar a establir un diagnòstic precís en els pacients categoritzats com B-CLPD, NOS**. Les prediccions d'aquests models afegirien una nova capa d'informació en el diagnòstic d'aquests pacients que per criteris diagnòstics convencionals no es poden diagnosticar.

3 Material i Mètodes

En el primer apartat d'aquest capítol es descriuen les dades que s'han utilitzat com a base per a enfrontar els objectius d'aquesta tesi. En el segon apartat s'explica que són R i Bioconductor, dues eines molt importants en l'anàlisi de dades d'origen genètic. En el tercer apartat s'expliquen les metodologies estadístiques que s'han utilitzat per preprocessar les dades crues obtingudes de les diferents tecnologies. En el quart apartat s'expliquen les metodologies estadístiques utilitzades per la selecció de variables i construcció dels predictors. En el cinquè apartat s'explica la *cross-validation*, una tècnica que serveix per estimar l'error esperat d'un predictor en dades futures. En el sisè apartat es resumeix el procés que s'ha seguit per crear dos predictors: un basat en dades de *microarrays* i un altre basat en dades de qPCR. Finalment, en l'últim apartat s'expliquen altres metodologies estadístiques utilitzades en diferents anàlisis de la tesi.

3.1 Descripció de les dades

Els resultats d'aquesta tesi es basen en informació obtinguda de mostres de sang extretes de pacients diagnosticats d'alguna de les entitats de B-CLPD. Com s'ha explicat en el primer capítol, aquests síndromes poden afectar un o més dels següents tres sistemes: sang, medul·la òssia i/o els diferents teixits limfoides. Tots els pacients inclosos en aquest estudi són leucèmics, és a dir, són pacients en què la sang està afectada per la malaltia. Crear un model predictiu mitjançant mostres de sang té dues avantatges:

- Facilita el diagnòstic en pacients que no tenen els ganglis afectats.
- L'extracció de sang és un procediment no invasiu.

Els pacients analitzats, què majoritàriament provenen de l'Hospital Clínic de Barcelona, estan dividits en tres cohorts. La primera cohort (*training*) consta de 159 pacients diagnosticats d'una entitat definida de B-CLPD, i serà la cohort que s'utilitzarà per construir el predictor en dades de *microarray* i el predictor en dades de qPCR. La segona cohort (validació) està formada per 63 pacients, també amb un diagnòstic definit, i s'utilitzarà per validar el predictor construït mitjançant dades de qPCR. La tercera i última cohort està formada per 64 pacients en els quals no se'ls hi ha pogut diagnosticar el subtipus i, per tant, s'han categoritzat com a B-CLPD, NOS. Els pacients de les dos primeres cohorts estan diagnosticats d'alguna de les següents nou entitats:

- **CLL: Leucèmia limfàtica crònica.**
- **FL: Limfoma fol·licular.**
- **HCL: Tricoleucèmia.**
- **HCLv: Tricoleucèmia variant.**
- **LPL: Limfoma limfoplasmacític.**
- **cMCL: Limfoma de les cèl·lules del mantell convencional.**
- **nnMCL: Limfoma de les cèl·lules del mantell no-nodal.**
- **SDRPL: Limfoma esplènic de la polpa vermella.**
- **SMZL: Limfoma esplènic de la zona marginal.**

Nota: Existeixen més subtipus de B-CLPD, però s'ha limitat l'estudi als més freqüents i que presenten més solapament en el seu diagnòstic.

En totes les 159 mostres de la cohort *training* s'ha obtingut informació mitjançant *microarrays* d'expressió, en 114 s'ha obtingut informació de *copy-number* i en 44 s'ha mesurat l'expressió gènica de 35 gens mitjançant qPCR. En les 63 mostres de la cohort de validació només s'ha disposat de l'expressió gènica d'aquests 35 gens. En la tercera cohort, formada per 64 B-CLPD, NOS, s'ha obtingut informació mitjançant *microarrays* d'expressió en 30 mostres, mentre que en les 34 restants s'ha obtingut de qPCR. En 16 de les 64 mostres s'ha obtingut informació de *copy-number*. La Taula 3.1 conté la

Cohort	Plataforma	N	CLL	FL	HCL	HCLv	LPL	cMCL	nnMCL	SDRPL	SMZL
Training	**Expressió**	159	54	12	4	4	4	30	24	4	23
	Copy-Number	114	50	3	1	4	4	20	13	4	15
	qPCR	44	8	4	2	3	3	6	6	2	10
Validació	**qPCR**	63	14	10	0	0	2	13	16	0	8
B-CLPD, NOS	**Expressió**	30									
	Copy-Number	16									
	qPCR	34									

Taula 3.1: Distribució dels pacients en les diferents cohorts. *Nombre de pacients de cada entitat i cohort mesurats mitjançant les diferents tecnologies.* ***N****: nombre total de pacients de la cohort mesurats mitjançant la respectiva tecnologia.*

distribució dels pacients de cada cohort entre les nou entitats de B-CLPD i les tres fonts d'informació (expressió, *copy-number* i qPCR).

La plataforma comercial de *microarrays* d'expressió utilitzada ha sigut l'Affymetrix *GeneChip® Human Genome U133 Plus 2.0*, la qual mesura més de 50000 seqüències al llarg del genoma. La plataforma de *microarrays* de *copy-number* utilitzada ha sigut l'Affymetrix *Genome-Wide Human SNP Array 6.0*, amb gairebé dos milions de mesures. Finalment, la plataforma de qPCR utilitzada per mesurar l'expressió de 35 gens ha sigut la *qPCR Fluidigm BioMark 48.48 Dynamic Array (Fluidigm®)* amb *Taqman® Gene Expression Assays (Applied Biosystems)*.

3.1.1 Grandària mostral i distribució de casos entre cohorts

El nombre de casos de cada entitat inclosos en aquesta tesi s'ha vist afectat per diverses limitacions. La primera i més important: la freqüència de pacients de cada subtipus diagnosticats a l'Hospital Clínic de Barcelona. Les entitats HCL, HCLv i SDPRL tenen una incidència molt baixa, el qual dificulta incloure més d'uns pocs casos. La segona limitació és la disponibilitat de material de cada un dels pacients. Cada experiment requereix utilitzar una certa quantitat de sang, aleshores, degut a que en aquesta tesi s'han utilitzat diverses plataformes (*microarrays* d'expressió, *microarrays* de *copy-number* i qPCR, entre d'altres), només s'ha pogut incloure en l'estudi pacients dels quals es disposava d'un gran volum de material. La tercera limitació és econòmica, a Desembre de 2008 els preus per un sol *microarray GeneChip® Human Genome U133*

Plus 2.0 era aproximadament de 600 euros, el qual limita el nombre d'experiments que es poden realitzar.

L'assignació de casos entre les cohorts *training* i de validació s'ha fet segons: *i*) quantitat de material, en la cohort *training* s'utilitzen *microarrays* i qPCR, motiu pel qual una mostra amb poc material no pot incloure's en aquesta cohort; *ii*) mostres no-purificades només s'han inclòs en la cohort de validació; *iii*) un cop tancada la cohort *training*, pacients posteriors s'han inclòs en la cohort de validació. Tot i que aquesta estratègia no assigna els casos a les diferents cohorts a l'atzar, no es creu que els criteris que s'han seguit hagin creat un biaix entre cohorts.

3.1.2 Limitacions

La limitada quantitat de mostres en la cohort de validació perjudica l'estimació de la precisió dels predictors, especialment per les entitats HCL, HCLv i SDRPL de les quals no es disposa de cap mostra. Una sèrie de validació apropiada hauria d'estar formada per un alt nombre de mostres i, per garantir un alt poder de generalització, aquestes mostres haurien de ser externes (per exemple, altres hospitals) [51]. Donat que aquestes entitats de B-CLPD són molt poc freqüents, obtenir-ne una grandària mostral apropiada és molt difícil. Una manera de comprovar que els descobriments no són falsos és comparant-los amb altres descobriments publicats en la literatura. Addicionalment, les funcions biològiques conegudes dels gens amb els quals s'ha construït el predictor també poden ser indicatives de la veracitat dels resultats. Per exemple, un predictor construït amb el gen *XIST*, un gen associat al sexe, és biològicament menys plausible que un construït amb un oncogen (gens responsables de transformar cèl·lules normals en tumorals).

3.2 R i Bioconductor

R és un entorn i llenguatge de programació enfocat a l'anàlisi estadística de dades. En aquesta tesi s'han realitzat totes les anàlisis mitjançant aquest programari, amb alguna

excepció indicada en els apartats corresponents. R forma part d'un projecte obert i col·laboratiu on els usuaris poden publicar noves funcions per realitzar diferents anàlisis estadístiques i representacions gràfiques, entre d'altres.

Bioconductor és un projecte de codi obert enfocat a l'anàlisi de dades òmiques (genòmica, proteòmica,...). Està basat principalment en contribucions d'usuaris mitjançant paquets d'R, tot i que també es pot contribuir amb funcions en altres llenguatges de programació. A part de paquets enfocats a l'anàlisi de dades, també inclou paquets d'anotació, com són bases de dades amb la funció i localització dels gens en el cromosoma o amb informació sobre quina seqüència de DNA mesura cada *probeset* d'un *microarray*.

Al llarg de la tesi s'indiquen els paquets d'R i Bioconductor utilitzats per realitzar les anàlisis pertinents. En cas d'utilitzar un altre programari també s'ha indicat.

3.3 Preprocessament de dades

En l'apartat 1.4.5 s'ha explicat que les dades crues obtingudes de les diferents tecnologies poden estar influenciades per factors tècnics aliens als factors biològics d'interès. Aquestes fonts de variabilitat tècniques es poden reduir mitjançant el preprocessament de les dades, el qual és l'aplicació de diferents algoritmes que les intenten ajustar.

3.3.1 Preprocessament dels *microarrays* d'expressió

En aquesta tesi s'ha pogut disposar de 189 (159 *training* + 30 B-CLPD, NOS) mostres de sang mesurades mitjançant *microarrays* d'expressió. En el primer capítol s'ha explicat que el resultat experimental d'un *microarray* és una imatge, la qual s'ha de transformar i processar mitjançant diversos algoritmes per tal d'obtenir valors numèrics amb sentit biològic. El pas de transformar la imatge escanejada en dades numèriques

(**preprocessament d'imatge**) s'ha dut a terme amb el programari *GeneChip® Command Console® Software*, seguint les recomanacions del fabricant dels xips (Affymetrix). Una descripció més detallada d'aquest pas es pot trobar a Artega-Salas et al. [52]. Les dades obtingudes del preprocés d'imatge es consideren les dades crues.

Els següents tres passos (correcció soroll de fons, normalització i resum) s'han realitzat mitjançant el mètode *frozen robust multiarray analysis* (fRMA) [53]. Per tal d'explicar el funcionament d'aquest algoritme és més natural començar explicant el funcionament del *robust multiarray analysis* (RMA) [54], donat que l'fRMA és basa en una modificació d'aquest.

Robust multiarray analysis

Per dur a terme la **correcció del soroll de fons**, l'RMA assumeix que la fluorescència observada d'una *probe* del *microarray* segueix el següent model:

$$\begin{aligned} O_{ij} &= B_{ij} + S_{ij} \\ B_{ij} &\sim N(\mu_i, \sigma_i) \\ S_{ij} &\sim \exp(\alpha_i), \end{aligned}$$

on O_{ij} és la fluorescència de la *probe j* en la mostra *i*, B_{ij} és el soroll de fons provinent de les diverses fonts (*background*) i S_{ij} és el senyal d'interès (*signal*). Notem que els paràmetres μ_i, σ_i i α_i són iguals per totes les *probes* de la mostra *i*. L'objectiu d'aquest pas és desxifrar el valor d'S_{ij}, el qual representa la fluorescència de la *probe* un cop extret l'efecte del soroll de fons.

El valor esperat d'S_{ij} correspon a

$$\begin{aligned} E(S_{ij} | O_{ij} = o_{ij}) &= o_{ij} - a_i + \sigma_i \frac{\varphi\left(\frac{o_{ij}-a_i}{\sigma_i}\right) - \varphi\left(\frac{a_i}{\sigma_i}\right)}{\Phi\left(\frac{o_{ij}-a_i}{\sigma_i}\right) + \Phi\left(\frac{a_i}{\sigma_i}\right) - 1}, \\ a_i &= \mu_i + \sigma_i^2 \alpha_i, \end{aligned}$$

on φ i Φ són les funcions de densitat i distribució d'una normal estàndard,

respectivament. Gràcies a l'assumpció de que totes les *probes* d'una mateixa mostra comparteixen els paràmetres μ_i, σ_i i α_i, aquests es poden estimar utilitzant les distribucions observades de fluorescència en totes les *probes* de la mostra. Aleshores, es poden obtenir els senyals corregits del soroll de fons (Y_{ij}) introduint aquestes estimacions a la fórmula anterior $\left(Y_{ij} = E\left(S_{ij} | \hat{O}_{ij} = o_{ij}\right)\right)$.

Un cop obtinguts els Y_{ij}, el següent pas és la **normalització** dels *microarrays*. L'objectiu d'aquest pas és fer comparables les diferents mostres, ja que una mostra podria obtenir fluorescències més altes que una altra degut, per exemple, a que s'han analitzat en escàners diferents. L'RMA ho fa mitjançant la normalització per quantils [55], un algoritme que força que la funció de distribució de totes les mostres sigui igual. Aquesta normalització assumeix que la majoria de les *probes* d'un *microarray* no estan influenciades per factors biològics, és a dir, assumeix que les diferències biològiques entre les diferents mostres afecten un percentatge baix de *probes*.

L'algoritme és senzill d'aplicar, el valor normalitzat (N) de la *probe j* en la mostra *i* és

$$N_{ij} = \frac{1}{n} \sum_{t=1}^{n} Y_{t(R[i,j])},$$

on $R[i,j]$ és la posició de la *probe j* dins la mostra *i*, i $Y_{t(R[i,j])}$ és el senyal de la *probe* amb posició $R[i,j]$ dintre les *probes* de la mostra *t*. De la fórmula es pot veure que totes les *probes* que tenen la mateixa posició en les diferents mostres obtenen el mateix valor, el qual correspon a la mitjana d'aquestes. La Figura 3.1 mostra la distribució de densitat dels senyals de tres mostres simulades (línies discontínues), i la distribució resultant de la normalització (línia sòlida). En la figura es pot veure que la distribució normalitzada és una combinació de les altres tres.

Finalment, el pas del **resum** té com objectiu obtenir un únic valor per cada *probeset*, què es defineix com el conjunt de *probes* que mesuren la mateixa seqüència d'RNA. El següent model s'ajusta a les *K probes* d'una única *probeset*:

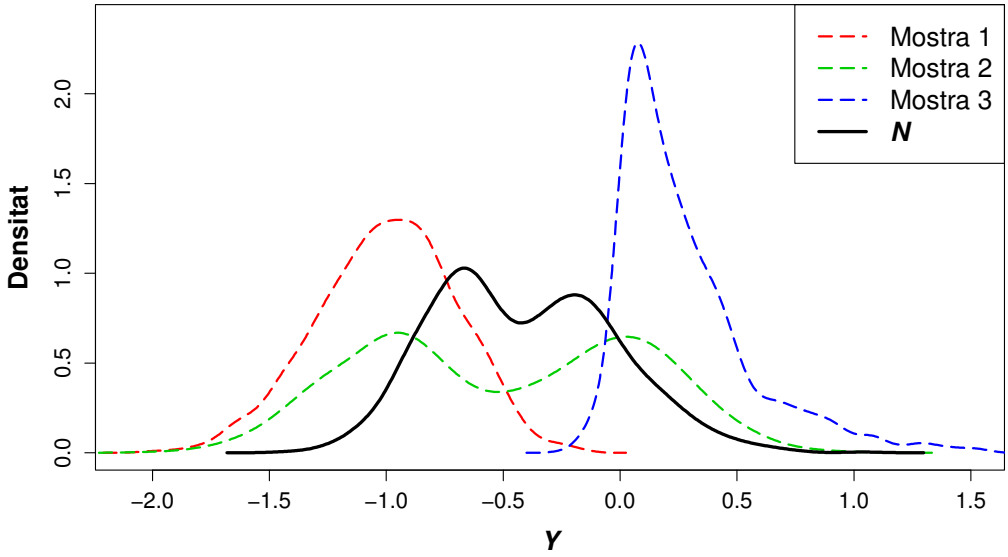

***Figura 3.1: Normalització per quantils**. Cada línia discontínua representa la distribució de densitat dels senyals en tres mostres diferents. La línia negra sòlida (**N**) representa la distribució de densitat de les tres mostres un cop transformades mitjançant la normalització per quantils.*

$$\log_2(N_{ik}) = \mu_i + \alpha_k + e_{ik},$$
$$\sum_{k=1}^{K} \alpha_k = 0,$$
$$V(e_{ik}) = \sigma^2,$$

on N_{ik} és el valor normalitzat de la *probe k* en la mostra *i*, μ_i és l'expressió de la *probeset* en la mostra *i*, α_k és l'efecte de la *probe k*, i e_{ik} és l'error aleatori. Per identificabilitat dels paràmetres es força que la suma de les α_k sigui 0. L'estimació de μ_i és el valor d'expressió d'interès i que s'utilitzarà per a les subsegüents anàlisis. Tot i que per simplicitat no s'ha inclòs l'índex corresponent a la *probeset*, notem que cada una obté una estimació diferent del conjunt de paràmetres. A través d'α_k, aquest model té en compte que algunes *probes* de la mateixa *probeset* obtenen sistemàticament valors més alts o més baixos. Tenir en compte aquest efecte millora l'estimació de l'expressió [56].

Per estimar els valors d'interès (μ_i) s'utilitza el mètode *median polish* proposat per Tukey [57]. Definim $\log_2(N)$ com la matriu que conté el logaritme (base 2) del valor normalitzat de les *n* mostres i *K probes*. L'algoritme de *median polish* correspon a:

1) Es calcula la mediana de cada fila de la matriu $\log_2(N)$, seguidament es calcula la matriu *E* extraient aquestes medianes a tots els valors de la fila respectiva.
2) S'actualitza *E* utilitzant el mateix procediment per columnes.
3) Es repeteixen els passos 1 i 2 fins que les medianes per files i columnes de *E* són igual o molt pròximes a zero.
4) Es calcula $\hat{\mu}_i = \frac{1}{K}\sum_{k=1}^{K}(\log_2(N_{ik}) - E_{ik})$.

La matriu *E* resultant és una estimació dels errors e_{ik}, per això al sostreure-la d'*N* es pot obtenir una estimació d'α i μ. Aquest algoritme s'ha d'aplicar independentment a cada *probeset*. La plataforma de *microarrays* Affymetrix *GeneChip® Human Genome U133 Plus 2.0* utilitzada en aquesta tesi està formada per més d'un milió de *probes*, les quals queden resumides en 54675 *probesets* després d'aquest pas. L'ús del mètode *median polish* per estimar els paràmetres del model es justifica degut a la presència habitual de dades atípiques en dades de *microarrays*, situació en què la mediana no es veu tan afectada com la mitjana.

En el pas de la normalització i el resum l'RMA utilitza models que preprocessen totes les mostres conjuntament, ja que s'ha demostrat que l'eliminació de fonts de variabilitat no-biològiques és més acurada si es fa conjuntament que mostra a mostra [55,56]. L'inconvenient d'aquesta estratègia és que conjunts de mostres preprocessades de forma separada no són comparables. Un dels objectius d'aquesta tesi és construir un predictor amb dades de la cohort *training* per aplicar-lo després a noves mostres. Per tant, un cop construït el predictor, el preprocessament dels *microarrays* de les noves mostres s'ha de fer de forma separada a la cohort *training*. Per aquest motiu l'RMA no s'adapta bé en l'entorn d'aquesta tesi.

Frozen robust multiarray analysis

El mètode fRMA [53] va crear-se per tal de solucionar l'inconvenient de preprocessar totes les mostres conjuntament. McCall et al. van utilitzar 850 *microarrays* provinents de diferents estudis (*batches*) per obtenir models de referència, els quals s'utilitzen de

base per fer el preprocessament de les noves mostres. D'aquesta manera s'evita que les mostres preprocessades conjuntament utilitzin la seva pròpia mitjana de base.

Per crear aquestes referències van realitzar la correcció del soroll de fons de les 850 mostres de la mateixa manera que l'RMA, què només depèn de la pròpia mostra. Posteriorment, la normalització la van fer per quantils obtenint una distribució de referència. Finalment, pel pas del resum van ajustar el següent model a cada conjunt de *K probes* contingudes en una mateixa *probeset*:

$$\log_2(N_{ikb}) = \mu_i + (\alpha_k + \gamma_{kb}) + e_{ikb},$$
$$\sum_{k=1}^{K} \alpha_k = 0,$$
$$V(\gamma_{kb}) = \tau_k^2,\ V(e_{ikb}) = \sigma_k^2,$$

on N_{ikb} és el valor normalitzat de la *probe k* en la mostra *i* del *batch b*, μ_i és l'expressió de la *probeset* en la mostra *i*, $\alpha_k + \gamma_{kb}$ és l'efecte de la *probe k* en el *batch b*, γ_{kb} és un efecte aleatori amb variància τ_k^2, i e_{ikb} és l'error aleatori amb variància σ_k^2 per la *probe k*. De la mateixa manera que en l'RMA, no s'ha inclòs l'índex corresponent a la *probeset*, però cada una obté una estimació diferent del conjunt de paràmetres. En aquest model es té en compte que l'efecte específic d'una *probe* pot canviar estudi a estudi. En el pas del resum de l'RMA la variabilitat de l'error aleatori s'assumia igual en totes les *probes*, en aquesta nova modelització s'ajusta una diferent per cada *probe* (σ_k^2).

Per estimar els diversos paràmetres del model van fer el següent:

1) Mitjançant un procediment robust (estimador-M [58]), van estimar els paràmetres μ_i i α_k ajustant el mateix model que en l'RMA:

$$\log_2(N_{ik}) = \mu_i + \alpha_k + e_{ik}.$$

2) Van calcular els residus segons:

$$r_{ikb} = \log_2(N_{ikb}) - (\hat{\mu}_i + \hat{\alpha}_k).$$

3) Utilitzant els residus, van estimar els paràmetres τ_k i σ_k segons

$$\hat{\tau}_k^2 = \frac{1}{K}\sum_{k=1}^{K}(\bar{r}_{\cdot kb}-\bar{r}_{\cdot k\cdot})^2,$$
$$\hat{\sigma}_k^2 = \frac{1}{n\cdot K}\sum_{k=1}^{K}\sum_{i=1}^{n}(r_{ikb}-\bar{r}_{\cdot kb})^2,$$

on

$$\bar{r}_{\cdot kb} = \frac{1}{n}\sum_{i=1}^{n} r_{ikb},$$
$$\bar{r}_{\cdot k\cdot} = \frac{1}{n\cdot K}\sum_{k=1}^{K}\sum_{i=1}^{n} r_{ikb}.$$

És a dir, les estimacions corresponen a les variàncies de la *probe k* entre i intra *batches*, respectivament.

D'aquests càlculs van obtenir una distribució de referència per la normalització per quantils i van fixar les estimacions d'α, τ, i σ de cada *probeset*. El preprocessament d'una nova mostra mitjançant l'fRMA es fa segons:

1) **Correcció del soroll de fons**: es fa de la mateixa manera que en l'RMA, donat que aquest pas només depèn de la pròpia mostra.
2) **Normalització**: es normalitza per quantils la nova mostra, projectant-la a la distribució de referència obtinguda dels 850 microarrays. Definim Z_j com el valor normalitzat de la *probe j* en la nova mostra.
3) **Resum**: per tal d'obtenir un únic valor per cada *probeset*, primer es calcula el valor ajustat per l'efecte probe (α) segons $X_j = \log_2(Z_j)-\hat{\alpha}_j$, on $\hat{\alpha}_j$ s'ha fixat amb els 850 *microarrays*. Després, utilitzant les *K probes* contingudes en la *probeset*, s'estima μ (valor d'interès de cada *probeset*) del model $X_k = \mu+\gamma_k+e_k$ segons

$$\hat{\mu} = \frac{\sum_{k=1}^{K} w_k \frac{1}{v_k} X_k}{\sum_{k=1}^{K} w_k \frac{1}{v_k}},$$
$$v_k = \hat{\tau}_k^2+\hat{\sigma}_k^2,$$

on v_k correspon a la suma de les variàncies intra i entre *batches* fixades amb els 850 *microarrays*, i w_k són pesos obtinguts d'un procediment d'estimació-M. Aquesta formulació és força intuïtiva: les *probes* amb més variabilitat entre estudis o entre mostres es penalitzen amb menys pes en la ponderació. També les *probes* que tenen valors atípics surten penalitzades (mitjançant els pesos *w*).

En resum, el mètode fRMA permet preprocessar els *microarrays* de forma separada sense perjudicar l'obtenció de bones estimacions de l'expressió per cada *probeset*. Per mostrar l'efecte del preprocessament s'ha representat, a la Figura 3.2, els senyals de dues mostres abans (esquerra, *probes*) i després (dreta, *probesets*) d'aplicar el mètode fRMA. Els gràfics superiors comparen el senyal de 1000 *probes*/*probesets* seleccionades a l'atzar entre dues mostres, on es veu que la normalització en linealitza la relació. En els gràfics inferiors es compara la distribució global de tots els senyals, on al de la dreta es veu com s'han igualat les distribucions de les dues mostres. L'alta correlació observada en els gràfics superiors és deguda a que l'estat basal d'alguns gens és més actiu i, per tant, totes les mostres els expressen a nivells més elevats.

A part de l'RMA i de l'fRMA existeixen altres mètodes per realitzar el preprocés de les dades crues, com són el MAS5 [56], el PLIER [59] i l'MBEI [60]. A més, dintre de cada mètode hi ha certa flexibilitat, com per exemple, en l'RMA es podria utilitzar la normalització per *loess* [61] en comptes de la normalització per quantils. En aquesta tesi s'ha utilitat l'fRMA pels avantatges ja comentats. Aquest mètode està implementat en el paquet *frma* de Bioconductor.

3.3.2 Filtratge de *probesets*

En el primer capítol s'ha explicat que mesurar milers de variables simultàniament augmenta el risc de trobar un fals resultat positiu. Les diferents metodologies d'ajust del *P*-valor serveixen per disminuir aquest risc, penalitzant el *P*-valor de forma proporcional al nombre de tests realitzats. Agafant el mètode de Bonferroni com exemple, un *P*-valor = 0.01 s'ajustaria a 0.03 si s'han realitzat tres tests d'hipòtesi, en

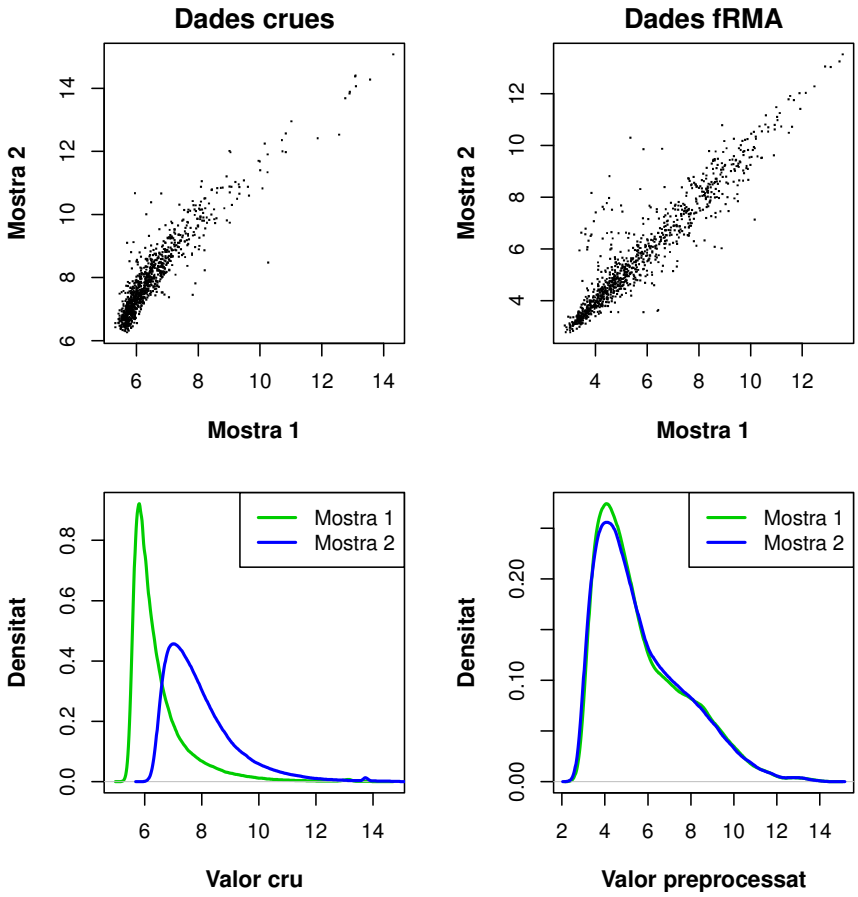

Figura 3.2: Preprocessament fRMA de dues mostres. *Els gràfics superiors comparen els valors crus (esquerra) i preprocessats (dreta) de 1000 senyals en dues mostres, mentre que els inferiors en comparen la distribució de totes els senyals.*

canvi, si s'han realitzat 20 tests s'ajustaria a 0.2, un ajust bastant més gran. Tot i que tenir moltes variables disponibles és un avantatge per trobar nova informació, augmentar el nombre de tests afegint variables sense motiu també pot dificultar la identificació de la informació rellevant.

En estudis de *microarrays*, on es disposa d'una gran quantitat de variables mesurades, és possible reduir el nombre de tests a realitzar sense perdre informació rellevant. Per exemple, en la plataforma de *microarrays GeneChip® Human Genome U133 Plus 2.0* hi ha 62 *probesets* incloses com a control de qualitat. Aquestes *probesets* recullen informació sobre l'eficiència de l'experiment en lloc d'informació biològica d'interès. Excloure-les de les anàlisis és recomanable ja que es redueixen la quantitat de tests a realitzar. Hi ha altres estratègies per filtrar *probesets* que no interessen en les anàlisis,

però en cap cas s'ha de fer aquest filtratge sense tenir en compte els objectius de l'estudi. En aquesta tesi el filtratge de *probesets* ha inclòs diferents procediments. Primer, s'ha eliminat les 62 *probesets* corresponents al control de qualitat. Segon, en la plataforma de *microarrays* utilitzada hi ha 12265 *probesets* que mesuren seqüències de l'RNA que no tenen anotades un únic gen en la base de dades *Entrez Gene* [62], un repositori d'informació específica de gens del *National Center for Biotechnology Information* (NCBI). Aquests *probesets* també s'han eliminat, ja que no tenen interpretació biològica i tampoc es poden traslladar de plataforma al no tenir un únic gen associat.

La tercera estratègia de filtratge ha sigut seleccionar una *probeset* per gen, donat que hi ha moltes *probesets* que mesuren seqüències de DNA contingudes en un mateix gen. A l'estar situades a una localització molt similar, aquestes *probesets* poden mostrar una correlació molt alta i, per tant, la informació que aporten entre elles és redundant. Per altra banda, la localització de la seqüència dintre el gen pot provocar que alguna *probeset* només mesuri soroll en comptes de l'efecte biològic.

Per exemplificar aquest fenomen s'ha representat, a la Figura 3.3, l'expressió mesurada en la cohort *training* de les 8 *probesets* associades al gen *XIST*, un gen localitzat al cromosoma X i que la seva expressió es veu afectada pel sexe del pacient. En la figura es pot veure com la majoria de *probesets* identifiquen un grup de casos amb valors alts i un grup amb valors baixos, en canvi, la *probeset* 243712_at (marcada en verd) té valors d'expressió baixos en tots els casos i no mostra la bimodalitat de les altres, degut a que només està mesurant soroll. Addicionalment, l'elevada correlació entre les *probesets* 224588_at i 227671_at mostra com les dues contenen la mateixa informació. La selecció d'una *probeset* per gen s'ha fet seleccionant la que té un rang interquartílic (IQR) més gran, donat que *probesets* que estiguin mesurant l'expressió correctament tendiran a tenir més variabilitat. Després d'aquest pas el nombre de *probesets* a analitzar són 20546, un 37.6% del total.

El quart i últim mètode de filtratge utilitzat correspon a seleccionar un cert percentatge de les 20546 probesets amb més IQR. Aquest tipus de filtratge pot incrementar la potència de les anàlisis [63], però depenent del test utilitzat hi ha risc en incrementar el

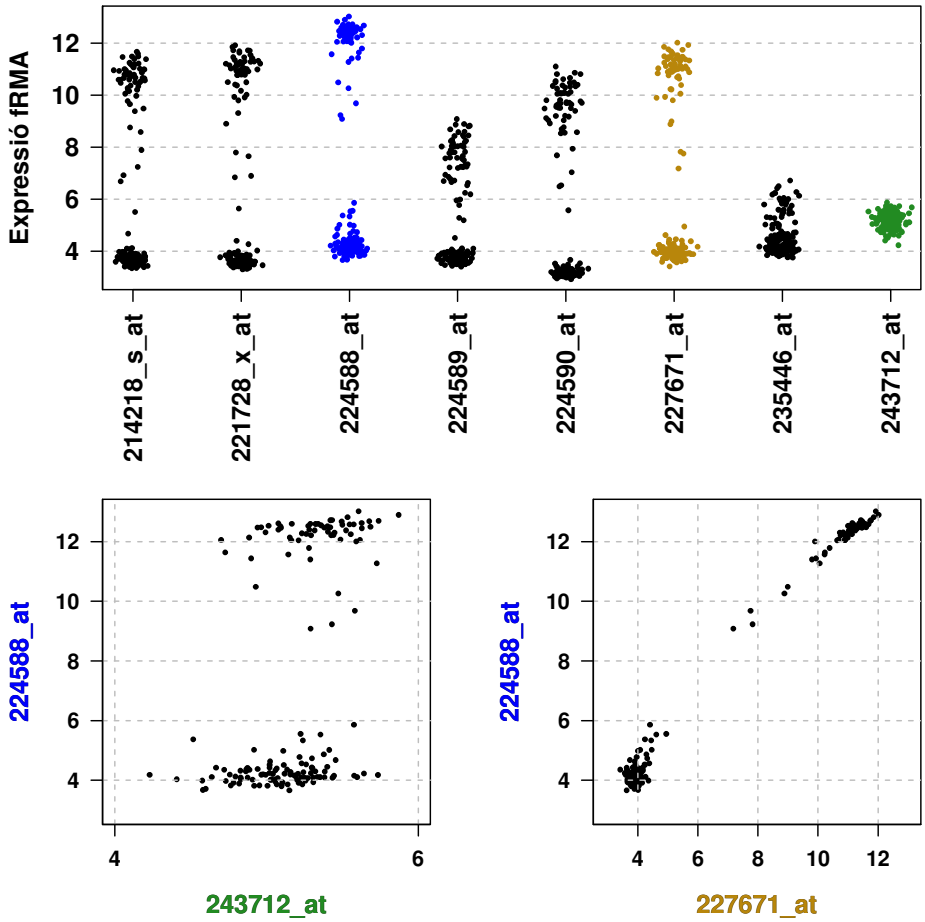

Figura 3.3: Expressió de les probesets associades al gen XIST. *El gràfic superior mostra l'expressió de les 8 probesets. Els gràfics inferiors mostren la relació de l'expressió de la probeset 224588_at amb les probesets 243712_at (esquerra) i 227671_at (dreta).*

nombre de falsos positius [64]. La càrrega computacional d'alguna de les metodologies utilitzades en aquesta tesi és molt elevada, així que també s'ha utilitzat aquesta quarta metodologia de filtratge per reduir la quantitat de variables a analitzar i, en conseqüència, el temps de computació.

Els tres primers mètodes de filtratge s'han aplicat en totes les anàlisis, mentre que aquest últim s'ha ajustat depenent de l'objectiu d'aquesta. Quan s'ha aplicat limma (apartat 3.4.3) s'ha eliminat un 0% de les probesets (p = 20546), quan s'ha aplicat el mètode de Dziuda (apartat 3.4.4) se n'ha eliminat un 50% (p = 10273), i per la resta de les anàlisis se n'ha eliminat un 25% (p = 15409).

3.3.3 Preprocessament dels *microarrays* de *copy-number*

El preprocessament de les 130 mostres en què es disposa de *microarrays* de *copy-number* s'ha fet mitjançant el programari *Nexus Copy Number* (*BioDiscovery*). Aquest programari és de codi tancat i l'algoritme que utilitza per preprocessar els *microarrays* no està descrit, tot i així, la idea general és similar que la dels *microarrays* d'expressió. Per exemple, el mètode CRMA v2 [65] ajusta els següents efectes:

- **Diafonia al·lèlica**: els *microarrays Genome-Wide Human SNP Array 6.0* estan construïts mitjançant parelles de *probes*. Aquestes parelles tenen com a objectiu la mateixa seqüència de DNA, però amb una base de diferència (per exemple, si una és TCCGT**A**GTTAT l'altra és TCCGT**T**GTTAT). Aquesta similitud provoca que els senyals de les dues *probes* es vegin pertorbats entre si. A més, la pertorbació no és la mateixa si la parella diferent és, per exemple, (A,C) de si és (A,T).
- **Efecte de la *probe***: similar als *microarrays* d'expressió (apartat 3.3.1), algunes *probes* obtenen senyals sistemàticament més alts o més baixos que d'altres.
- **Resum**: cada seqüència es mesurada per un conjunt de *probes*, les quals es poden combinar per obtenir-ne un únic valor.
- **Efecte de la llargada dels fragments**: en el procés d'hibridació del *microarray* s'ha de fragmentar el DNA de la mostra. Com més llargs són els fragments, menys del DNA objectiu s'hibrida, és a dir, el senyal obtingut és més baix quan els fragments són més llargs.

Un cop realitzades aquestes correccions s'obté una estimació del senyal per cada mostra i localització. Definim θ_{ij} com l'estimació del senyal en la localització j de la mostra i. Aleshores, es calcula el número de còpies relatiu (C_{ij}) a un senyal de referència (θ_{Rj}) segons

$$C_{ij} = 2\frac{\theta_{ij}}{\theta_{Rj}},$$

però és més habitual treballar amb el valor relatiu en escala logarítmica, calculat segons

$$M_{ij} = \log_2\left(\frac{\theta_{ij}}{\theta_{Rj}}\right).$$

La referència θ_{Rj} es pot obtenir de fer la mitjana robusta de les θ_{ij} de totes les mostres en la localització *j*. Els valors M_{ij} són els que s'utilitzen en les posteriors anàlisis i s'anomenen *Log R Ratio* (LRR). Aquests *microarrays* serveixen per buscar desviacions en el número de còpies de diferents segments al llarg del cromosoma, el qual és de dues en una cèl·lula normal (els humans tenim genomes diploides, és a dir, tenim dues còpies de cada cromosoma). La Figura 1.8 del primer capítol mostra els valors d'LRR d'una part del cromosoma 13 d'un pacient de B-CLPD.

Segmentació

Un cop obtinguts els valors d'LRR només falta la segmentació, la qual serveix per identificar localitzacions consecutives amb senyals més alts o més baixos de l'esperat (Figura 1.8). La segmentació es duu a terme mitjançant algoritmes que busquen punts de canvi en el valor mitjà dels LRR al llarg de les localitzacions. Aquests algoritmes s'apliquen de manera independent a cada mostra. En les dades d'aquesta tesi la segmentació s'ha fet mitjançant el mètode r*ank segmentation*, inclòs dintre el programari *Nexus Copy Number* (*BioDiscovery*). Aquest mètode és una versió robusta del mètode c*ircular binary segmentation* (CBS) [66].

La base del CBS és un problema de punts de canvi. Les dades estan estructurades de forma ordenada en l'espai i el que interessa és detectar a partir de quina localització hi ha un canvi en la distribució de les observacions. Definim M_j com l'LRR de la localització *j* (per simplicitat s'ha eliminat l'índex corresponent a la mostra). Aleshores, *v* és un punt de canvi si es compleix que:

$$\begin{aligned} &M_1, \ \ldots, \ M_v \sim F_0, \\ &M_{v+1}, \ \ldots, \ M_p \sim F_1, \\ &F_0 \neq F_1, \end{aligned}$$

on F_0 i F_1 són funcions de distribució i *p* és el total de localitzacions. És a dir, *v* és un

punt de canvi si la distribució de probabilitat dels valors d'LRR en les localitzacions situades fins a *v* és diferent que la distribució dels LRR en les localitzacions situades després de *v*.

L'algoritme de CBS amplia aquest problema a un número indefinit de punts de canvi. Definint les sumes parcials $S_t = M_1 + \ldots + M_t$, $1 \leq t \leq p$, es calcula l'estadístic:

$$Z_{ij} = \frac{\left(\frac{S_j - S_i}{j-i} - \frac{S_p - (S_j - S_i)}{p-(j-i)}\right)}{\sqrt{\left(\frac{1}{j-i} + \frac{1}{p-(j-i)}\right)}},$$

on $S_j - S_i$ és la suma dels LRR situats en les localitzacions contingudes en $[i+1, j]$, i $j - i$ és la quantitat de localitzacions en aquest segment. De la mateixa manera, $S_p - (S_j - S_i)$ és la suma dels LRR situats en la resta de localitzacions i $p - (j - i)$ és la quantitat de localitzacions en aquesta segona suma. Per tant, el numerador és una diferencia de mitjanes i el denominador és l'error estàndard assumint que la variància és igual a 1 en tots els segments. Notem que si s'assumeix la mateixa variància en tots els segments, el valor de l'estadístic sempre serà proporcional a Z_{ij}:

$$Z_{ij}^{*}(\sigma) = \frac{1}{\sigma} Z_{ij},$$

on σ es podria estimar de les dades en cas de no assumir-ho.

Un cop calculat Z_{ij} per cada parella $\{i, j\}$, el CBS calcula l'estadístic

$$Z_C = max_{1 \leq i < j \leq p} |Z_{ij}|,$$

i el compara amb el valor crític sota la hipòtesi nul·la (H_0: no existeixen punts de canvi). En cas de rebutjar-la, s'estimaria que la parella $\{i, j\}$ que compleix $Z_C = |Z_{ij}|$ correspon a dos punts de canvi (en *i* i en *j*). La distribució sota H_0 es pot obtenir mitjançant simulacions. Aquest procediment s'aplica de forma recursiva als nous segments identificats fins que no s'obté cap Z_C significatiu. La principal diferencia del mètode *rank segmentation* respecte el CBS és que el primer utilitza la posició d'M_j dintre la

mostra en comptes d'M_j, el qual millora la capacitat de detecció de punts de canvi [67].

Els segments obtinguts del procés de segmentació no sempre són rellevants des del punt de vista biològic. Per exemple, el mètode de segmentació podria detectar un fragment amb una mitjana poc per damunt de 0, aquesta lleugera diferència respecte el valor basal podria estar relacionada amb factors tècnics més que biològics, atès que, en cas d'haver-hi un guany, l'increment seria més elevat. També es podria identificar un segment molt curt o un guany/pèrdua en una zona del cromosoma mal representada pel *microarray*, dos motius per creure que el segment està relacionat amb factors tècnics. Analitzar automàticament aquestes desviacions mitjançant un algoritme és complex, degut a que depèn de la malaltia en estudi. Per aquest motiu, és habitual que múltiples observadors familiaritzats amb la malaltia realitzin una inspecció visual dels segments per descartar aquells que per diferents criteris (localització, llargada, grandària del canvi,...) no són rellevants.

En alguns estudis la informació dels segments es por resumir encara més. Per exemple, si en un estudi la meitat dels pacients tenen una deleció que comença entre les bases [100-200] i que acaba entre les bases [800-900], la informació biològica és pràcticament la mateixa encara que els punts de canvi no coincideixin en tots els casos. Aquesta informació es podria categoritzar en una única variable binària. En aquesta tesi s'ha seguit aquesta estratègia, un cop detectat i refinat els segments perduts o guanyats en cada pacient, s'ha resumit la informació mitjançant variables binàries. Per exemple, en el braç *q* del cromosoma 13 (els 24 cromosomes humans estan formats per dos braços, anomenats *p* i *q*) s'ha identificat segments perduts en 19 dels 114 casos de la cohort *training*. L'inici i final del segment de cada cas és diferent, però els segments de tots 19 intersequen en una certa regió (regió mínima alterada comuna).

Resumint, la informació obtinguda després de preprocessar i analitzar els *microarrays* de *copy-number* és un conjunt de variables binàries, on cada una indica si el pacient té una pèrdua/guany de material genètic en el braç corresponent del cromosoma. Concretament, en aquesta tesi s'ha reduït el gairebé dos milions de mesures d'aquests *microarrays* a 28 variables binàries.

3.3.4 Preprocessament de les dades de qPCR

El preprocessament de les dades de qPCR és més simple que el preprocessament de les dades de *microarrays*. Només involucra dos passos: la selecció del llindar FT de cada gen i la normalització. Recordem que el llindar FT serveix per distingir a partir de quin cicle el senyal obtingut no està influenciat pel soroll de fons, és a dir, és el cicle en què el valor de fluorescència està mesurant l'expressió i no soroll blanc (apartat 1.3.3).

La **selecció d'FT** s'ha fet mitjançant la representació conjunta de les corbes d'amplificació de totes les mostres analitzades. Aleshores, s'ha situat manualment al punt on la majoria de corbes comencen la fase d'amplificació òptima. Com a exemple, a la Figura 3.4 s'ha representat les corbes del gen FMOD en 11 mostres. En la figura es pot veure com cap mostra creix en els primers cicles, degut a que el senyal obtingut prové del soroll de fons i no de la fluorescència de la mostra. En la fase d'amplificació les corbes són paral·leles i, per últim, en la fase final totes les mostres es saturen a un valor similar. Aquest pas s'ha dut a terme amb el programari *Fluidigm Real-Time PCR Analysis*.

Mètode ΔΔCT

La **normalització** s'ha fet amb quantificació relativa mitjançant el mètode ΔΔCT [68]. Aquesta normalització de dades de qPCR utilitza gens *housekeeping*. Aquests gens tenen la característica que la seva expressió no canvia sota diferents condicions biològiques, i, per tant, serveixen de base per controlar els factors tècnics de l'experiment. La normalització es fa tenint en compte que la qPCR funciona replicant la seqüència d'interès. L'objectiu de la replicació és augmentar la quantitat del gen d'interès fins al punt on es pot mesurar l'expressió, ja que en el moment inicial l'escàner no pot mesurar la fluorescència de manera fiable. Al provocar la replicació la seqüència del gen creixerà molt ràpid si aquest està expressat, o més lentament si no ho està.

L'amplificació exponencial de la qPCR en el gen g d'una mostra d'interès x es pot

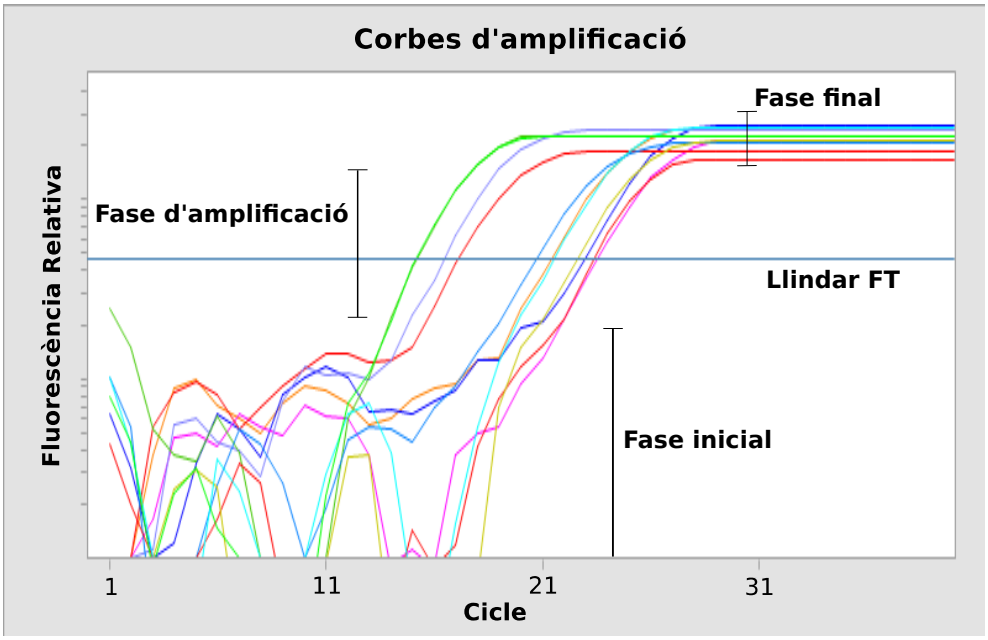

***Figura 3.4: Corbes d'amplificació del gen FMOD en 11 mostres**. En les corbes es pot identificar les tres etapes: la inicial, on el soroll de fons domina les mesures, la d'amplificació òptima, on el creixement és lineal, i la fase final en què totes les mostres es saturen. El llindar FT està situat a la primera meitat de la segona fase.*

descriure segons

$$X_{gc} = X_{g0}(1+E)^{c},$$

on X_{gc} és el nombre de seqüències del gen g al cicle c, X_{g0} és el nombre de seqüències del gen al moment inicial, i E és l'eficiència de l'amplificació. Assumint que l'eficiència és perfecta (igual a 1), significaria que a cada cicle que passa es duplica la seqüència d'interès. Per simplificar, s'assumirà que l'eficiència és 1.

Utilitzant l'equació de l'amplificació podem calcular el cicle CT_{xg} en què el nombre de seqüències del gen g de la mostra x és igual a K_g, essent K_g un valor suficientment gran com per poder-lo mesurar sense la influencia del soroll de fons. És a dir,

$$X_{g,CT_{xg}} = K_g = X_{g0}2^{CT_{xg}},$$

d'on podem aïllar X_{g0} segons

$$X_{g0} = K_g 2^{-CT_{xg}},$$

fórmula amb la qual podríem estimar, en la mostra d'interès *x*, el valor d'expressió del gen *g* en el moment inicial (X_{g0}), per així comparar-lo amb altres mostres.

El problema és que X_{g0} no és comparable entre mostres, ja que, degut a factors tècnics, tots els gens d'una mostra podrien tenir més quantitat de material al moment inicial. Per poder controlar aquests factors s'utilitza els gens *housekeeping*. Aquests gens, els quals la seva expressió només es veu influenciada per factors tècnics, serveixen per obtenir mesures normalitzades comparables entre mostres. Similar al gen *g*, la quantitat inicial del gen *housekeeping h* en la mostra *x* es pot calcular segons

$$X_{h0} = K_h 2^{-CT_{xh}}.$$

Aleshores, l'expressió inicial del gen *g* normalitzada pel gen *housekeeping h* en la mostra *x* correspon a

$$\frac{X_{g0}}{X_{h0}} = \frac{K_g 2^{-CT_{xg}}}{K_h 2^{-CT_{xh}}} = \frac{K_g}{K_h} 2^{-\Delta CT_{xg}},$$
$$\Delta CT_{xg} = CT_{xg} - CT_{xh},$$

valor que es pot comparar entre mostres.

Un segon problema és que podria existir un efecte global que provoqués que tots els experiments d'un mateix lot (dia, laboratori, màquina, ...) es veiessin influenciats de forma sistemàtica. Aquest fet forçaria a utilitzar valors de K_g diferents en cada lot i, en conseqüència, que mostres de diferents lots no fossin comparables. Per tal de controlar aquests efectes entre lots es fa un segon nivell de normalització. Aquest segon nivell consisteix en utilitzar una mostra de referència o calibradora (*r*), la qual es mesuraria juntament amb les altres mostres d'un mateix lot. S'assumeix que les diferències d'expressió d'aquesta mostra entre lots són degudes a factors tècnics. L'expressió inicial del gen *g* normalitzada pel gen *housekeeping h* en la mostra calibradora *r* correspon a

$$\frac{R_{g0}}{R_{h0}} = \frac{K_g 2^{-CT_{rg}}}{K_h 2^{-CT_{rh}}} = \frac{K_g}{K_h} 2^{-\Delta CT_{rg}},$$
$$\Delta CT_{rg} = CT_{rg} - CT_{rh}.$$

Finalment, el càlcul de l'expressió relativa de la mostra d'interès *x* respecte la mostra calibradora *r* es fa segons

$$N_g = \frac{X_{g0}/X_{h0}}{R_{g0}/R_{h0}} = \frac{(K_g/K_h)2^{-\Delta CT_{xg}}}{(K_g/K_h)2^{-\Delta CT_{rg}}} = 2^{-\Delta\Delta CT_{xg}},$$
$$\Delta\Delta CT_{xg} = \Delta CT_{xg} - \Delta CT_{rg} = (CT_{xg} - CT_{xh}) - (CT_{rg} - CT_{rh}),$$

on N_g correspon al valor normalitzat del gen *g* en la mostra d'interès *x*. Un cop realitzats els dos nivells de normalització, les dades ja són comparables i es poden utilitzar en les anàlisis estadístiques corresponents.

Una assumpció important d'aquest mètode de normalització és que l'eficiència *E* de l'amplificació és la mateixa en totes les mostres, en cas que una o més mostres presentessin una eficiència menor s'hauria de repetir l'experiment o normalitzar el conjunt de mostres mitjançant altres metodologies [69]. Comprovar que l'eficiència és similar en totes les mostres es pot fer mitjançant el gràfic conjunt de les corbes d'amplificació, on totes haurien de mostrar un patró de creixement similar, només diferenciant-se per un desplaçament en l'eix X (cicles). A la Figura 3.4 es pot veure com el patró de totes les corbes és similar.

La mostra calibradora utilitzada en aquesta tesi ha sigut la *Universal Human Reference* (UHR), una mostra que està formada per una barreja de línies cel·lulars (poblacions de cèl·lules). El gen *housekeeping* utilitzat ha sigut el *B2M*.

3.4 Mètodes de classificació i selecció de variables

Un cop aplicats els diversos mètodes de preprocessament i eliminat les fonts de variabilitat causades per les diferents tecnologies, les dades ja es poden utilitzar per mesurar les diferencies biològiques entre les mostres. Els mètodes estadístics presentats en aquest apartat han servit per dos propòsits: *i*) construir models predictius per classificar nous casos, i *ii*) seleccionar gens en dades de *microarrays* per traslladar-los a qPCR, un pas necessari per tal d'obtenir un predictor fàcil d'utilitzar a nivell clínic.

3.4.1 Mètodes *kernel* per a la integració de dades

La integració de dades òmiques és un problema que ha pres interès en els últims anys a causa de la gran quantitat d'informació que s'obté de diferents fonts [70,71]. En aquesta tesi es disposa de dues fonts d'informació òmiques: *microarrays* d'expressió i *microarrays* de *copy-number*. Combinar aquestes dues fonts amb l'objectiu de crear un únic predictor no és trivial. Existeixen tres maneres de fer aquesta combinació [72]:

- **Integració prematura**: aquesta integració es basa en concatenar el conjunt de variables de cada font en una única matriu de covariables. Per exemple, si disposes de 20 variables d'una font i 15 d'una altra, les anàlisis estadístiques es farien com si només es tingués una única font de 35 variables.
- **Integració intermèdia**: aquest tipus d'integració es basa en resumir la informació de cada font de dades de forma separada. Després es combinen aquestes informacions en un únic predictor.
- **Integració tardana**: aquesta integració es basa en obtenir un predictor per cada font d'informació, aleshores es combina les prediccions obtingudes de cada un en una única predicció final.

La integració intermèdia té un gran avantatge respecte a les altres dues: permet ponderar la importància global de cada font d'informació, una característica molt rellevant en aquesta tesi. Els *microarrays* són experiments cars de realitzar, així que poder valorar de forma global la informació que aporta cada tipus de *microarrays* (expressió i *copy-number*) és molt rellevant per identificar si fa falta obtenir informació dels dos tipus o només d'un.

La família dels mètodes *kernel* [73] permet realitzar la integració intermèdia de manera molt natural. Per explicar el funcionament d'aquesta família de mètodes començarem suposant la situació en què només es disposa d'una única font d'informació, en concret, suposarem que volem distingir dues classes amb p variables predictores mesurades en n mostres.

Funcions *kernel*

Els mètodes *kernel* comencen per projectar la matriu de p covariables, mesurades en l'espai original O^p, a un nou espai m-dimensional S^m, on $m>p$. L'objectiu d'aquesta projecció és que en S^m les dades quedin millor estructurades per distingir les dues classes que en O^p. El punt interessant d'aquests mètodes és que no fa falta conèixer el nou espai, el qual pot arribar a ser infinit, sinó que només cal saber el producte escalar entre totes les parelles de mostres en S^m.

Recordem que el producte escalar entre el vector $\boldsymbol{x} = \{x_1, x_2, \ldots, x_p\}$ i el vector $\boldsymbol{y} = \{y_1, y_2, \ldots, y_p\}$ en l'espai original p-dimensional és

$$\boldsymbol{x}\cdot\boldsymbol{y} \;=\; \boldsymbol{x}^T\boldsymbol{y} \;=\; \sum_{i=1}^{p} x_i\, y_i,$$

d'on s'obté un únic valor per cada parella de mostres.

Els mètodes *kernel* calculen el producte escalar entre dues mostres en l'espai desconegut S^m a través de la funció *kernel* $K(\boldsymbol{x},\boldsymbol{y})$. Suposem que la funció φ projecta les mostres de l'espai O^p a l'espai S^m, aleshores, la funció *kernel* calcula

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \varphi(\boldsymbol{x})\cdot\varphi(\boldsymbol{y}) \;=\; \sum_{i=1}^{m} \varphi(\boldsymbol{x})_i\,\varphi(\boldsymbol{y})_i,$$

amb la peculiaritat que es defineix $K(\boldsymbol{x},\boldsymbol{y})$ sense definir i aplicar φ.

Per exemple, suposem $p = 2$ i definim la funció *kernel* com

$$K(\boldsymbol{x},\boldsymbol{y}) \;=\; \left(1+\sum_{i=1}^{2}(x_i\,y_i)\right)^2,$$

si expandim l'expressió obtenim

$$\begin{aligned} K(\boldsymbol{x},\boldsymbol{y}) \;&=\; \left(1+\sum_{i=1}^{2}(x_i\,y_i)\right)^2 \;=\; (1+x_1\,y_1+x_2\,y_2)^2 \\ &=\; 1+x_1^2\,y_1^2+x_2^2\,y_2^2+2\,x_1\,y_1+2\,x_2\,y_2+2\,x_1\,x_2\,y_1\,y_2, \end{aligned}$$

què correspon al producte escalar dels vectors

$$\boldsymbol{x}^* = \{1, x_1^2, x_2^2, \sqrt{2}\,x_{1,}\sqrt{2}\,x_2, \sqrt{2}\,x_1 x_2\},$$
$$\boldsymbol{y}^* = \{1, y_1^2, y_2^2, \sqrt{2}\,y_{1,}\sqrt{2}\,y_2, \sqrt{2}\,y_1 y_2\},$$

d'on es pot deduir que la funció φ transporta els punts de l'espai O^2 de 2 dimensions a l'espai S^6 de 6 dimensions segons

$$\varphi(\boldsymbol{x}) \;=\; \varphi(\{x_1, x_2\}) \;=\; \{1, x_1^2, x_2^2, \sqrt{2}\,x_{1,}\sqrt{2}\,x_2, \sqrt{2}\,x_1 x_2\}.$$

En aquest exemple s'ha pogut calcular, a través de $K(\boldsymbol{x},\boldsymbol{y})$, el producte escalar en S^6 sense projectar cada mostra segons φ, ni tan sols es coneixia aquest espai abans de desenvolupar l'expressió. Per explicar com aprofiten aquesta propietat els mètodes *kernel* s'agafarà d'exemple el mètode més conegut d'aquesta família, el s*upport vector machine* (SVM) [23].

Support vector machine

L'SVM és un mètode de classificació que busca l'hiperplà que millor separa les mostres de dues classes. En la Figura 3.5 hi ha un exemple amb 20 mostres repartides en 2 classes ($y_i \in \{-1, 1\}$) i 2 variables ($\boldsymbol{x_i} = \{x_{i1}, x_{i2}\}$). L'hiperplà M de la figura és el que maximitza la separació, o marge m, entre les dues classes i és de la forma:

$$M\colon\ \boldsymbol{w}^T\boldsymbol{x}+b \;=\; 0,$$

on $\boldsymbol{w}$ és el vector normal o perpendicular a l'hiperplà M, i b és una constant que indica la distància perpendicular d'aquest a l'origen.

En cas de que la separació lineal de les classes sigui perfecta, aleshores existeixen els dos hiperplans H definits per la forma

$$H_1\colon\ \boldsymbol{w}^T\boldsymbol{x}+b \;=\; -1$$
$$H_2\colon\ \boldsymbol{w}^T\boldsymbol{x}+b \;=\; 1,$$

els quals compleixen les següents restriccions per $i = 1, ..., n$:

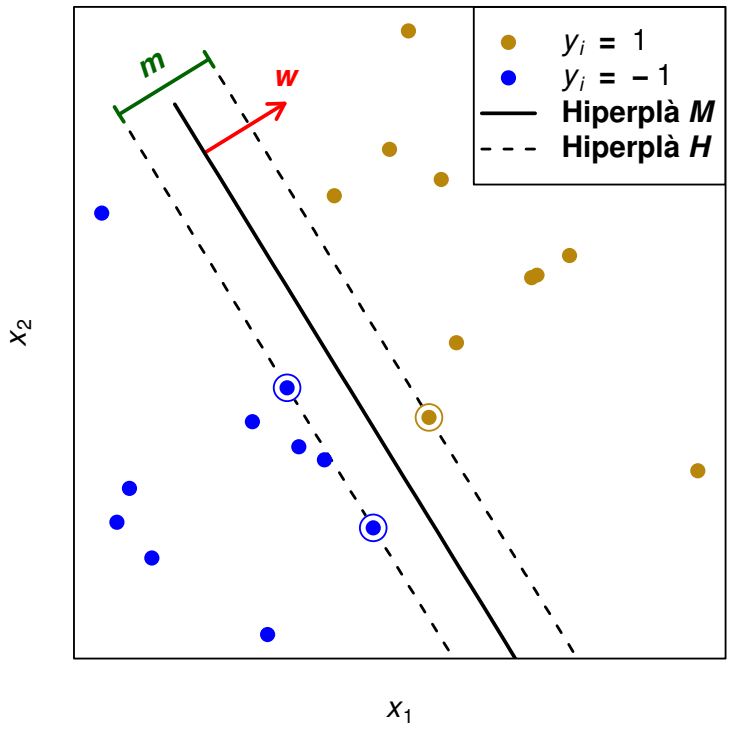

Figura 3.5: Hiperplans d'SVM en dues variables. *Hiperplans i marge (m) del mètode SVM que discrimina dues classes.*

$$H_1: \ \boldsymbol{w}^T \boldsymbol{x_i} + b \ \geq \ 1, \text{si } y_i = 1$$
$$H_2: \ \boldsymbol{w}^T \boldsymbol{x} + b \ \leq \ -1, \text{si } y_i = -1,$$

és a dir, totes les mostres del grup y = 1 queden per damunt de l'hiperplà H_2 i totes les mostres del grup y = -1 queden por sota de l'hiperplà H_1. Aquestes restriccions es poden reescriure segons

$$y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b) \ \geq \ 1.$$

Un cop definits els hiperplans H, el mètode SVM busca el vector $\boldsymbol{w}$ que maximitza el marge m entre els dos. La distància geomètrica d'aquest marge correspon a

$$m \ = \ \frac{2}{\|\boldsymbol{w}\|},$$

per tant, per maximitzar aquesta distància s'ha de solucionar el següent problema

d'optimització, el qual anomenarem ***hard-margin***:

$$\begin{aligned} &\text{Minimitzar} && \|\boldsymbol{w}\| \\ &\text{subjecte a} && y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b) \geq 1, \ \forall i, \end{aligned}$$

on els paràmetres $\boldsymbol{w}$ i b que el solucionen són els que s'utilitzen per predir noves dades. En concret, la predicció d'una nova mostra ($\boldsymbol{x_{new}}$) es fa en funció de l'hiperplà M, si es compleix que $\boldsymbol{w}^T\boldsymbol{x_{new}} + b > 0$, aleshores la predicció és $\hat{y}_{new} = 1$, en cas de que $\boldsymbol{w}^T\boldsymbol{x_{new}} + b < 0$, aleshores la predicció seria $\hat{y}_{new} = -1$.

L'optimització *hard-margin* assumeix que les classes són perfectament separables. En les situacions que no ho són, s'ha de canviar el problema d'optimització de la següent manera, què anomenarem ***soft-margin***:

$$\begin{aligned} &\text{Minimitzar} && \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{subjecte a} && y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b) \geq 1 - \xi_i, \ \forall i \\ & && \xi_i \geq 0, \ \forall i, \end{aligned}$$

on s'afegeix dos termes nous: C i $\boldsymbol{\xi} = \{\xi_1, ..., \xi_n\}$. El vector de paràmetres $\boldsymbol{\xi}$ relaxa la restricció de que les mostres de les dues classes quedin perfectament separades. En concret, com més gran el valor de ξ_i d'una mostra, més allunyada queda aquesta de l'hiperplà H en el costat "incorrecte". Si $\xi_i = 0$ significa que la mostra queda en el costat "correcte" de l'hiperplà H. El paràmetre C determina l'intercanvi entre l'amplada del marge m i l'esforç de que les mostres quedin en el costat correcte dels hiperplans H. Si el valor de C és molt gran, l'optimització *soft-margin* prioritzaria que els components de $\boldsymbol{\xi}$ siguin pròxims a 0. En cas contrari, si C és pròxima a 0 l'optimització prioritzaria un marge m més gran.

La Figura 3.6 mostra l'efecte de dos valors de $C = \{1000, 1\}$ en dos exemples, un linealment separable i un altre que no. En ambdós exemples els marges són més amplis quan $C = 1$, si més no, el nombre de mostres entre els dos hiperplans H també és més elevat. El valor de C s'ha d'especificar a priori, una manera de fer-ho és seleccionant el valor que minimitza l'error de classificació estimat mitjançant *cross-validation*.

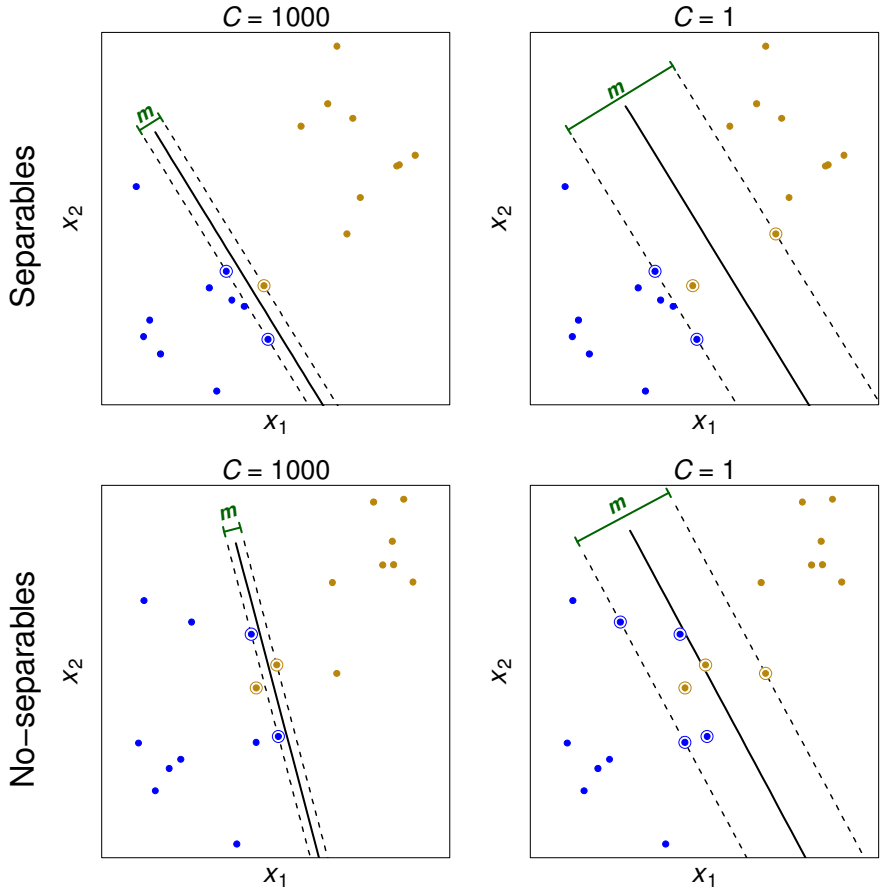

***Figura 3.6: Efecte del paràmetre C en SVM**. Els gràfics superiors corresponen a dues classes perfectament separables, en els inferiors ho són parcialment. S'explora C=1000 (esquerra) i C=1 (dreta).*

En les formulacions *hard* i *soft-margin* presentades no es pot aprofitar la funció *kernel*, donat que el problema no inclou el producte escalar entre mostres. Per aprofitar-ho s'ha de transformar el problema d'optimització *soft-margin* a la seva versió dual de Lagrange. Aquesta transformació canvia el problema sense canviar els valors objectius, així que els valors dels paràmetres $\boldsymbol{w}$ i b que solucionen un són els mateixos que els que solucionen l'altre. La versió **dual** és:

$$\begin{aligned} &\text{Maximitzar} && \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \\ &\text{subjecte a} && 0 \leq \alpha_i \leq C \\ & && \sum_{i=1}^{n} \alpha_i y_i = 0, \end{aligned}$$

d'on s'obté el valor del vector de paràmetres $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_n\}$ que maximitza la funció

objectiu. A les mostres *i* que compleixen $\alpha_i > 0$ se les anomena *support vectors*, i són les mostres situades exactament damunt dels hiperplans *H* o en el costat "incorrecte" d'aquests. En les Figures 3.5 i 3.6 els *support vectors* són els punts marcats amb una circumferència.

Per recuperar el vector de paràmetres $\boldsymbol{w}$ en aquesta nova formulació es pot fer segons

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i,$$

i per recuperar *b* s'ha d'agafar una mostra *k* que compleixi $\alpha_k > 0$ i calcular amb aquesta mostra

$$b = y_k - \boldsymbol{w}^T \boldsymbol{x}_k, \quad k : \alpha_k > 0.$$

En la versió dual es pot veure que les covariables ($\boldsymbol{x}$) només s'inclouen en la formulació a través del producte escalar entre mostres ($\boldsymbol{x_i} \cdot \boldsymbol{x_j}$). Aleshores, és molt senzill aprofitar la funció *kernel K* en aquesta versió. Suposem que s'ha aplicat la transformació φ per traslladar les dades a S^m, la funció objectiu a maximitzar en aquest espai seria:

$$\begin{aligned} \text{Maximitzar} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\varphi(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j)) \\ & = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \end{aligned}$$

on només fa falta conèixer la funció *kernel K* per tal d'identificar els hiperplans *M* i *H* que separen linealment les classes en S^m.

El vector $\boldsymbol{w}$ en S^m correspon a

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \varphi(\boldsymbol{x}_i),$$

el qual no es pot recuperar sense aplicar la funció φ a les dades. Tot i així, per realitzar prediccions no cal calcular-lo explícitament, ja que les prediccions d'una nova mostra *z* es poden obtenir segons

$$p(\boldsymbol{z}) = \boldsymbol{w}^T \varphi(\boldsymbol{z}) + b = \left(\sum_{i=1}^{n} \alpha_i y_i (\varphi(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{z})) \right) + b = \left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{z}) \right) + b,$$

si $p(\boldsymbol{z}) > 0 \Rightarrow \hat{y}_{new} = 1,$
si $p(\boldsymbol{z}) < 0 \Rightarrow \hat{y}_{new} = -1,$

on b es pot recuperar seleccionant una mostra k que compleixi $\alpha_k > 0$ i calcular:

$$b = y_k - \boldsymbol{w}^T \varphi(\boldsymbol{x}_k) = y_k - \sum_{i=1}^{n} \alpha_i y_i (\varphi(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_k)) = y_k - \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_k)$$
$$k : \alpha_k > 0.$$

Integració de múltiples fonts

Amb l'SVM i les funcions *kernel* definides, només falta explicar com es pot realitzar la integració intermèdia mitjançant aquest família de mètodes. Per fer-ho s'agafarà d'exemple les dades d'aquesta tesi, on es disposa de dues fonts d'informació heterogènies: *microarrays* d'expressió (***E***) i *microarrays* de *copy-number* (***D***). A cada font s'aplica les funcions *kernel* K_E i K_D, respectivament. Aleshores, la integració intermèdia s'obté combinant les matrius segons [72]

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K_E(\boldsymbol{E}_i, \boldsymbol{E}_j) + K_D(\boldsymbol{D}_i, \boldsymbol{D}_j),$$

on $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ correspon al valor del *kernel* combinat entre les mostres *i* i *j*.

Per integrar expressió gènica amb altres fonts, Daemen et al. [74–76] proposen el model

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \beta K_E(\boldsymbol{E}_i, \boldsymbol{E}_j) + (1-\beta) K_D(\boldsymbol{D}_i, \boldsymbol{D}_j), \quad 0 \le \beta \le 1,$$

on β és un paràmetre que controla el pes que es dóna a cada font d'informació. Si β=1, aleshores només les dades d'expressió influeixen en el predictor, en canvi, si β=0 només ho farien les dades dels *microarrays* de *copy-number*.

La funció *kernel* que proposen Daemen et al. per les diferents fonts és

$$K_E(\boldsymbol{x}_i, \boldsymbol{x}_j) = K_D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)}{\sqrt{(\boldsymbol{x}_i \cdot \boldsymbol{x}_i)(\boldsymbol{x}_j \cdot \boldsymbol{x}_j)}},$$

la qual correspon a la versió normalitzada del *kernel* lineal ($\varphi(\boldsymbol{x}) = \boldsymbol{x}$).

Les estructures de les dades de les dues fonts d'informació són molt diferents, per exemple, la font d'expressió està formada per milers de variables que prenen valors entre 2 i 18, mentre que la font de *copy-number* està formada per menys d'un centenar de variables binàries. L'ordre del producte escalar entre dos mostres en el primer cas és $\geq 10^4$, mentre que en el segon cas és de 10^1. Aquesta diferència provocaria que les dades de *copy-number* fossin irrellevants en la suma de *kernels* no-normalitzats, excepte per valors de β pròxims a 0. Aleshores, la versió normalitzada del kernel lineal projecta les dades a l'esfera unitària, d'aquesta manera l'ordre de la matriu *kernel* resultant és el mateix en les dues fonts.

Finalment, Daemen et al. proposen utilitzar una versió modificada de l'SVM, el l*east squares* SVM (LS-SVM) [77], el qual modifica el problema d'optimització *soft-margin* segons:

$$\begin{aligned} &\text{Minimitzar} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\frac{1}{2}\sum_{i=1}^{n}\xi_i^2 \\ &\text{subjecte a} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x_i} + b\right) = 1-\xi_i, \ \forall i, \end{aligned}$$

on la solució es troba mitjançant un sistema lineal en comptes de mitjançant un problema de programació quadràtica. La principal avantatge és que en dades d'alta dimensionalitat el temps computacional es redueix considerablement, sense perdre precisió en les prediccions [78].

En aquesta tesi s'ha fet servir l'SVM, atès que s'ha pogut tolerar l'increment del temps computacional entre aquest i l'LS-SVM. La funció *kernel* utilitzada ha sigut la lineal normalitzada. Utilitzar funcions *kernel* més complexes no millora la precisió del predictor quan el nombre de variables supera el de mostres [79]. A més a més, la interpretació dels resultats obtinguts mitjançant un *kernel* lineal és més simple. Els paràmetres C i β s'han estimat mitjançant *cross-validation*. L'SVM està implementat en el paquet *kernlab* d'R.

3.4.2 *Nearest shrunken centroids*

En l'apartat 3.4.1 s'ha presentat un mètode que permet construir un predictor mitjançant la combinació de diferents fonts d'informació. Tot i així, també és rellevant considerar la possibilitat de construir el predictor amb només la font d'informació d'expressió gènica, ja que es disposa de més mostres en aquesta font i en la literatura hi ha molts predictors validats basats en expressió gènica [6,25,29,80]. En el cas que el predictor basat en expressió gènica obtingués precisions similars a les del predictor basat en les dues fonts d'informació, es prioritzaria el primer, donat que seria menys complex i fàcil d'interpretar des d'un punt de vista biològic. Com ja s'ha comentat anteriorment, donar sentit biològic a un predictor alleugera l'inconvenient del reduït nombre de mostres.

Existeixen un gran nombre de mètodes de classificació disponibles per construir el predictor basat en dades d'expressió, fins i tot tenint en compte la restricció de mètodes explicada a l'apartat 1.4.2. Identificar quin obtindria un error de predicció significativament més petit en les dades d'aquesta tesi requeriria un nombre de mostres força més elevat del que es disposa. Si més no, en la literatura s'han explorat un gran nombre de metodologies en diferents entorns. Diversos autors han observat que mètodes de classificació simples, els quals no tenen en compte l'estructura de correlació entre les variables predictores, poden tenir bons resultats en situacions d'alta dimensionalitat, especialment quan el nombre de mostres és baix [81–85].

Aquest fenomen és degut a que l'estructura de correlació no es pot estimar correctament quan $n<<p$, provocant que el model ajustat pateixi d'*overfitting*. És a dir, en aquesta situació és preferible ajustar un model que pateixi de biaix, provocat per l'assumpció de correlacions nul·les entre variables, que un model sense biaix però menys estable. Si, a més, el nombre de mostres és baix en termes absoluts (per exemple, $n<10$), el risc d'*overfitting* es pot reduir encara més si s'aplica mètodes de regularització o de *shrinkage* [86,87].

Un mètode molt utilitzat en estudis d'expressió gènica que compleix aquests dos

requisits és el *nearest shrunken centroids* (NSC) [88], també referenciat com *Prediction Analysis for Microarrays* (PAM) en la literatura. A part de complir els dos requisits esmentats, aquest mètode es pot aplicar per classificar més de dues classes, una altra característica convenient en aquesta tesi en què es volen discriminar nou classes diferents.

L'NSC es pot interpretar com una modificació de l'anàlisi lineal discriminant diagonal (DLDA) [81], i la seva aplicació és molt senzilla. Primer definim x_{ij} com l'expressió de la mostra *i* en la *probeset* o gen *j*, on el conjunt de dades està format per *n* mostres i *p* gens. Les *n* mostres estan distribuïdes en *K* classes, i C_k indica els índexs de les n_k mostres de la classe *k*. Aleshores, per cada classe *k* es calcula els *p* components del corresponent centroide segons

$$\bar{x}_{jk} = \sum_{i \in C_k} \frac{x_{ij}}{n_k},$$

és a dir, cada component correspon a la mitjana de l'expressió d'un gen en les mostres de la classe *k*. De la mateixa manera es calcula els components del centroide global segons

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n}.$$

Utilitzant els centroides prèviament definits, es calcula l'estadístic

$$d_{jk} = \frac{\bar{x}_{jk} - \bar{x}_j}{m_k (s_j + s_0)},$$

on

$$\begin{aligned} s_j^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_{jk})^2, \\ m_k &= \sqrt{\frac{1}{n_k} - \frac{1}{n}}, \\ s_0 &= \text{Mediana}(\{s_1, \ldots, s_p\}), \end{aligned}$$

és a dir, d_{jk} és una diferència de mitjanes dividida per un error estàndard, per tant, es pot

interpretar com una espècie d'estadístic T que compara la classe k amb el centroide global. La constant s_0 s'afegeix per evitar que gens amb baixa expressió obtinguin per atzar valors alts de d_{jk}, donat que habitualment les expressions baixes no són fiables i es veuen més influenciades pel soroll de fons.

De la fórmula anterior podem aïllar el component j del centroide de la classe k segons

$$\bar{x}_{jk} = \bar{x}_j + m_k(s_j + s_0)d_{jk},$$

on l'NSC aplica l'encongiment (*shrinkage*) del centroide apropant tots els valors de d_{jk} a zero. Quan es disminueix d_{jk} els centroides de cada classe s'aproximen al centroide global, en conseqüència, també s'aproximen els centroides de cada classe entre si. Aquest efecte d'encongiment evita l'*overfitting*, ja que provoca que els centroides no quedin tant centrats als valors estimats en el *training set*.

El **centroide reduït** s'obté segons

$$\bar{x}'_{jk} = \bar{x}_j + m_k(s_j + s_0)d'_{jk},$$

on

$$d'_{jk} = \begin{cases} 0, & \text{si } |d_{jk}| - \Delta \leq 0 \\ sign(d_{jk})(|d_{jk}| - \Delta), & \text{altrament} \end{cases}$$

i Δ és la quantitat de reducció. Quan $\Delta>0$ els centroides de les diferents classes s'aproximen al centroide global. Si Δ fos suficientment gran podria provocar que el component d'un gen fos igual en totes les classes i, per tant, no influenciés en la predicció de noves mostres.

Aquesta propietat de l'NSC provoca que, mitjançant Δ, realitzi la selecció de variables de manera intrínseca, una propietat molt rellevant tal com s'ha vist en l'apartat 1.4.3. A la Figura 3.7 es mostra com canvia la localització dels centroides de cada classe quan s'incrementa el valor de Δ. Les dades representades en la figura corresponen a mostres de dues classes ($y_i \in \{-1, 1\}$) i a dues variables (x_1, x_2). Les fletxes indiquen en quina

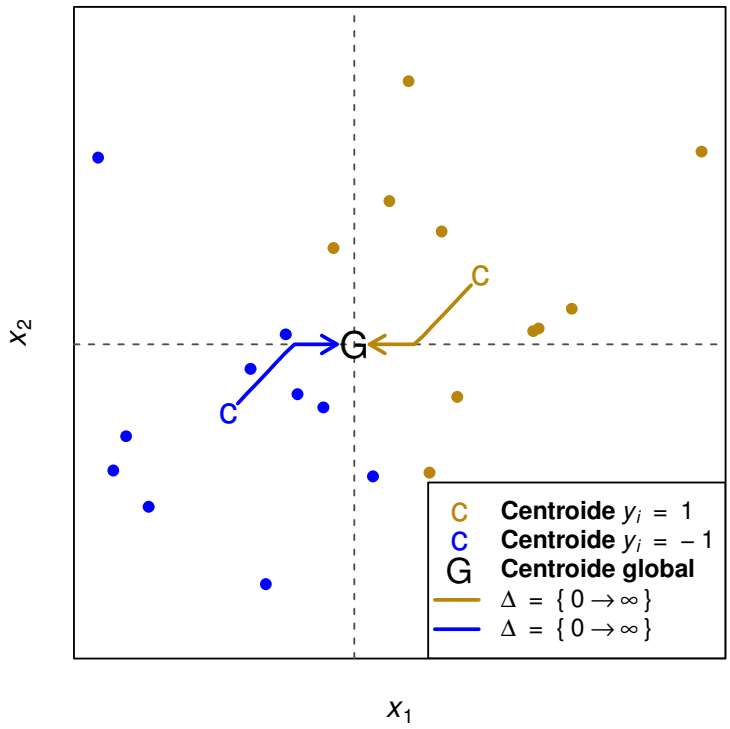

Figura 3.7: Efecte del paràmetre Δ en NSC. *Les línies de color groc i blau marquen la direcció en què es mou el centroide de cada classe a l'augmentar el valor del paràmetre Δ. A partir d'un cert valor de Δ els centroides de les dues classes coincideixen amb el centroide global.*

direcció es desplaça el centroide de cada classe a l'anar incrementant Δ. En la figura es pot veure com els centroides s'aproximen cada vegada més al centroide global. També es veu que, a partir d'un cert valor de Δ, el component corresponent a la variable x_2 és igual en les dues classes (coincidint amb el centroide global).

La **predicció de noves mostres** es fa al centroide reduït més pròxim, estandarditzant per l'error estàndard ($s_j + s_0$) i corregint pel nombre relatiu de mostres en cada classe. La distància corregida d'una nova mostra amb nivells d'expressió $\boldsymbol{z} = \{z_1, ..., z_p\}$ al centroide reduït de la classe k és

$$\delta_k(\boldsymbol{z}) = \sum_{j=1}^{p} \frac{\left(z_j - \bar{x}_{jk}^{(s)}\right)^2}{\left(s_j + s_0\right)^2} - 2\log \pi_k,$$

on π_k és la probabilitat a priori de la classe *k*. La nova mostra es classificaria a la classe *k* que minimitza $\delta_k(\boldsymbol{z})$.

Habitualment, la probabilitat a priori π_k no és coneguda. Si la cohort recollida representa correctament la població, aleshores π_k es pot estimar de les dades segons $\pi_k = n_k/n$. En cas de que la cohort no sigui representativa de la població (per exemple, la cohort està enriquida en alguna classe menys freqüent), es pot fixar $\pi_k = 1/K$ per tot *k*.

De manera similar a l'LDA, es pot obtenir una estimació de la probabilitat de que la nova mostra ($\boldsymbol{z}$) pertanyi a la classe *k* segons

$$\hat{p}_k(\boldsymbol{z}) = \frac{e^{(-1/2)\delta_k(\boldsymbol{z})}}{\sum_{l=1}^{K} e^{(-1/2)\delta_l(\boldsymbol{z})}}.$$

Observant les fórmules de l'NSC es pot veure l'analogia que té amb l'LDA, on l'NSC es diferencia per tres aspectes: *i*) assumeix que la matriu de covariàncies és diagonal, *ii*) fa una correcció de les desviacions estàndard a través de s_0, i *iii*) fa una reducció dels centroides a través de Δ. Totes tres diferències serveixen per evitar que el model s'ajusti perfectament al *training set*, és a dir, per disminuir el risc d'*overfitting*. La primera redueix el nombre de paràmetres a estimar pel model, la segona evita que cap gen pugui tenir molt més pes que la resta en els càlculs, i la tercera torna a reduir el nombre de paràmetres al només considerar els gens que la diferència d'expressió entre classes superi un cert llindar.

El mètode NSC està implementat en el paquet *pamr* de R. L'estimació del paràmetre Δ s'ha fet mitjançant *cross-validation*.

3.4.3 *Linear models for microarray data*

El segon objectiu d'aquesta tesi és plantejar una metodologia de selecció de gens a mesurar mitjançant una plataforma de qPCR. Aquesta selecció es pot fer en base a les

dades de *microarrays*, on es mesuren una gran quantitat de gens i se'n pot extreure informació sobre quins són rellevants per, entre d'altres, distingir els diferents subtipus de B-CLPD. Una manera senzilla d'avaluar quins gens són més rellevants és ordenant-los segons el *P*-valor corresponent a l'estadístic *T* (en cas de comparar 2 subtipus) o l'estadístic *F* (en cas de comparar més de 2 subtipus). Els gens al capdamunt de la llista serien els que tenen les mitjanes d'expressió entre subtipus més diferenciades, per tant, es podria traslladar els *r* primers a la plataforma de qPCR.

Quan el nombre de mostres és baix i el nombre de variables és alt apareixen dos problemes. El primer és que s'augmenta la quantitat de gens amb la variància infraestimada, el qual provoca que alguns estadístics *T* estiguin molt inflats. El segon problema és la falta de potència al disposar de poques mostres. A la pràctica, utilitzar estadístics *T* modificats adreça aquests problemes, augmentant la precisió en la identificació de gens rellevants [89,90].

Dintre la família dels estadístics *T* modificats, el mètode que ha mostrat millors resultats ha sigut el *linear models for microarray data* (limma) [91], el qual ajusta models lineals en un entorn empíric Bayesià per permetre l'intercanvi d'informació entre gens. Aquest intercanvi d'informació millora l'estimació de les variàncies, especialment quan el nombre de mostres és petit. Al basar-se en models lineals, limma es pot utilitzar en un rang més divers de situacions com, per exemple, comparar l'expressió mitjana entre més de dos subtipus de B-CLPD. Aquesta metodologia està implementada al paquet *limma* de Bioconductor.

Limma assumeix que l'esperança de l'expressió (estandarditzada) de cada gen segueix un model lineal, és a dir, pel gen *g* assumeix que

$$E(\boldsymbol{y}_g) = \boldsymbol{X}\boldsymbol{\beta}_g,$$

on $\boldsymbol{y}_g = \{y_{1g}, y_{2g}, \ldots, y_{ng}\}$ és el vector que conté les expressions del gen *g* en les mostres *i* = 1:*n*, la matriu $\boldsymbol{X}$ és la matriu del disseny (amb *n* files i *m* columnes) i $\boldsymbol{\beta}_g = \{\beta_{1g}, \beta_{2g}, \ldots, \beta_{mg}\}$ és el vector de paràmetres desconeguts del model lineal ajustat pel gen *g*. En cas de comparar dos subtipus, la matriu del disseny estaria formada per una primera columna

de 1's i una segona columna *dummy*, amb 0's en les files que corresponen a les mostres d'un subtipus i 1's en les files que corresponen a l'altre subtipus.

L'estimador per mínims quadrats del vector de paràmetres $\boldsymbol{\beta}_g$ és

$$\hat{\boldsymbol{\beta}}_g = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}_g,$$

mentre que l'estimador de la variància residual (σ_g^2) del model lineal correspon a

$$\begin{aligned} s_g^2 = \hat{\sigma}_g^2 &= \frac{(\boldsymbol{y}_g - \hat{\boldsymbol{y}}_g)^T (\boldsymbol{y}_g - \hat{\boldsymbol{y}}_g)}{d_g} = \frac{\sum_{i=1}^{n} (y_{ig} - \hat{y}_{ig})^2}{d_g}, \\ \hat{\boldsymbol{y}}_g &= \boldsymbol{X} \hat{\boldsymbol{\beta}}_g, \\ d_g &= n - m, \end{aligned}$$

on d_g representa els graus de llibertat residuals i $\hat{\boldsymbol{y}}_g = \{\hat{y}_{1g}, \hat{y}_{2g}, \ldots, \hat{y}_{ng}\}$ el valor predit del model lineal.

Assumint normalitat dels residus, definits com $e_{ig} = y_{ig} - \hat{y}_{ig}$, la distribució d'$s_g^2$ és proporcional a la distribució khi-quadrat (χ^2) amb d_g graus de llibertat segons

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2.$$

La diferència de limma respecte el model lineal clàssic és que, mitjançant l'ús d'un entorn Bayesià, limma assumeix distribucions a priori pel paràmetre σ_g^2. La distribució a priori que limma utilitza és

$$\sigma_g^2 \sim \frac{d_0 s_0^2}{\chi_{d_0}^2},$$

la qual, un cop especificats els paràmetres d_0 i s_0^2, descriu com s'espera que variïn les variàncies residuals gen a gen a priori.

La distribució a posteriori de σ_g^2 donades les dades (s_g^2) és

$$\sigma_g^2 | s_g^2 \sim \frac{d_0 s_0^2 + d_g s_g^2}{\chi^2_{d_0+d_g}},$$

on es pot veure que és una combinació de la informació a priori (d_0 i s_0^2) i la informació de les dades (d_g i s_g^2). En concret, es pot considerar que la distribució a posteriori combina dues estimacions de la variància residual, una prèvia (s_0^2) i una de les dades actuals (s_g^2), ponderades pels graus de llibertat en les estimacions d'aquestes (d_0 i d_g). Com més gran d_0 respecte d_g, menys variable serà la distribució a priori al voltant d's_0^2 i, per tant, la distribució a posteriori quedarà més pròxima a s_0^2 que a s_g^2. Per altra banda, si d_0 és més petit que d_g, la informació a priori no tindrà pes i la distribució a posteriori estarà dominada per les dades (s_g^2).

L'esperança de la distribució a posteriori de σ_g^2 es pot calcular segons

$$E\left(\frac{1}{\sigma_g^2} | s_g^2\right) = \frac{1}{\widetilde{s}_g^2},$$
$$\widetilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g},$$

on es pot veure que és una mitjana ponderada de les dues variàncies. Limma utilitza el mateix valor dels paràmetres d_0 i s_0^2 per tots els gens, per tant, d'aquesta fórmula es pot deduir que les variàncies mostrals s'encongeixen (en major o menor grau, depenent de d_0) cap a un valor comú s_0^2.

Finalment, limma fa una combinació de inferència freqüentista i inferència Bayesiana per tal d'obtenir, pel paràmetre β_{jg}, el següent **estadístic *T* moderat**:

$$\widetilde{t}_{jg} = \frac{\hat{\beta}_{jg}}{\widetilde{s}_g^2 \sqrt{V_{jj}}},$$
$$\boldsymbol{V} = (\boldsymbol{X}^T \boldsymbol{X})^{-1},$$

el qual correspon a l'estadístic *T* clàssic substituint la variància mostral (s_g^2) per l'esperança de la variància a posteriori ($\widetilde{s}_g^2$). Sota la hipòtesi nul·la (H_0: $\beta_{jg} = 0$) aquest estadístic segueix una distribució *T* amb $d_0 + d_g$ graus de llibertat. En general, qualsevol

estadístic derivat del model lineal es pot convertir en un estadístic moderat si es substitueix la variància residual de la mostra per la variància a posteriori (amb el subseqüent augment en els graus de llibertat).

Per calcular l'estadístic *T* de limma s'han de fixar els paràmetres d_0 i s_0^2. Limma estima aquests paràmetres de les dades (d'aquí que es consideri un mètode empíric Bayesià) en comptes de fixar-los manualment o de considerar-los hiperparàmetres en un model Bayesià jeràrquic. Per fer-ho, primer calcula $z_g = \log s_g^2$. z_g es distribueix segons una constant més una distribució *Z* de Fisher [92], donat que s_g^2 es distribueix segons una distribució *F* escalada. Aleshores, limma estima d_0 i s_0^2 igualant la mitjana i variància teòriques de la distribució *Z* a la mitjana i variància mostrals de z_g. Els càlculs concrets d'aquest procediment es poden trobar a Smyth [91].

A la Figura 3.8 es mostra com es diferencia la variància residual clàssica de la variància moderada de limma en unes dades d'exemple. En concret, s'han simulat unes dades que consisteixen de 16 mostres (8 per classe) i 100 variables. Cada variable té una variància residual diferent. Al gràfic s'ha representat la variància de limma respecte la variància clàssica en aquestes 100 variables, on es pot veure com les variàncies mostrals més petites sofreixen un augment considerable quan s'estimen mitjançant limma. Simultàniament, les variàncies més grans sofreixen un decrement molt lleuger.

Existeixen altres metodologies que comparteixen informació entre gens per tal de millorar l'estimació de les variàncies, com el SAM [93] o l'RVM [94]. En aquesta tesi s'ha utilitzat limma ja que obté resultats més acurats per qualsevol nombre de mostres [89,90]. Notem que el mètode NSC, explicat a l'apartat 3.4.2, també realitza una correcció de les variàncies. Aquesta estratègia és habitual en l'entorn de la genètica, donat que compensa la poca informació de les mostres (*n* petita) amb compartir informació entre la gran quantitat de variables (*p* gran).

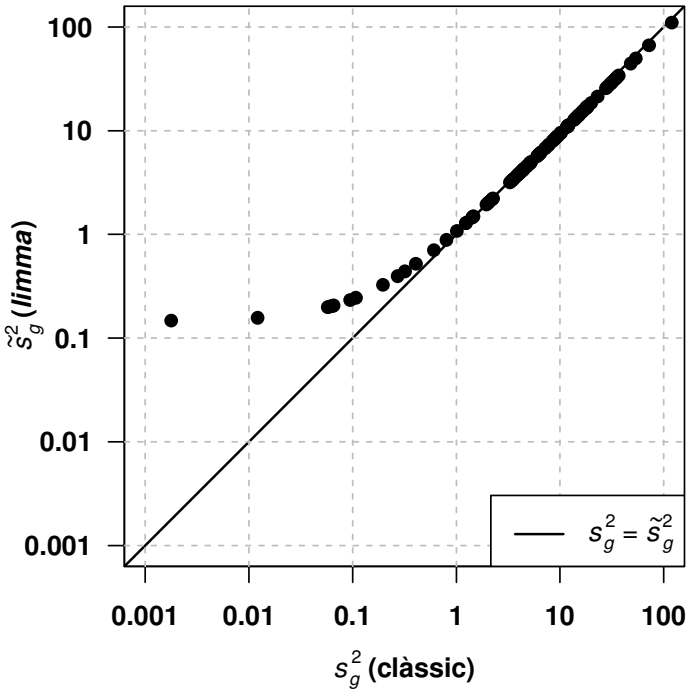

Figura 3.8: Comparació de la variància clàssica i l'estimada segons limma*. En la figura es compara la variància residual de 100 variables estimades per dos mètodes diferents: la variància mostral clàssica i la variància segons limma.*

3.4.4 Mètode de Dziuda

Tal com s'ha explicat a l'apartat 3.4.2 (NSC), les estratègies que no tenen en compte l'estructura de correlació entre *probesets*/gens obtenen, en general, models predictius més precisos en problemes d'alta dimensionalitat i baix nombre de mostres. El mètode limma (apartat 3.4.3), el qual tampoc té en compte aquesta estructura al basar-se en estadístics *T,* s'ha presentat com un mètode útil per a seleccionar variables dels *microarrays* a traslladar a qPCR.

Existeix una diferència important entre el problema de crear un predictor i el problema de seleccionar gens a traslladar: la limitació en el nombre de variables. En la construcció del predictor en dades de *microarrays* no cal limitar el nombre de gens a un número preespecificat, simplement s'utilitza la quantitat que obtingui la millor precisió. En el cas de construir el predictor en dades de qPCR no es té aquesta comoditat per dos

motius: *i*) el nombre de gens a traslladar com a possibles candidats és limitat, i *ii*) el predictor final ha de ser fàcil d'utilitzar a nivell clínic i, per tant, el nombre de gens que aquest pot contenir com a màxim són 2 o 3 per subtipus de B-CLPD. Aleshores, quan l'objectiu és traslladar gens, l'interès està en seleccionar uns pocs que continguin el màxim d'informació possible en quant a discriminació de classes. En aquest cas, els mètodes que tenen en compte l'estructura de correlació podrien ajudar a maximitzar aquesta informació.

A part de limma, en el procés de selecció de gens d'aquesta tesi també s'ha utilitzat un mètode que té en compte l'estructura de correlació. El mètode multivariant utilitzat ha sigut el de Dziuda [95], el qual mesura la rellevància multivariant dels gens mitjançant la combinació de tres metodologies: l'estadístic traça de Lawley-Hotelling (T^2) [96], una metodologia *stepwise* (*stepwise hybrid feature selection with* T^2) i una tècnica de remostratge (m*odified bagging schema*). Primer es definiran cada una d'aquestes tres metodologies, per després explicar com es combinen.

Estadístic traça de Lawley-Hotelling

Aquest estadístic es pot utilitzar per mesurar el poder discriminant de p variables per discriminar K classes, en concret, quantifica com de separats estan els centroides de les K classes. La seva interpretació és similar a l'estadístic F d'un procediment ANOVA, però en el cas que el nombre de variables numèriques dependents és major que 1. Definim $\boldsymbol{x}_{ik} = \{x_{i1k}, x_{i2k}, ..., x_{ipk}\}$ com el vector dels valors d'expressió dels p gens en la mostra i de la classe k, i definim n_k com el nombre de mostres en la classe k. L'estadístic traça de Lawley-Hotelling (T^2) es calcula segons

$$
\begin{aligned}
T^2 &= tr\left(\boldsymbol{H}\,\boldsymbol{E}^{-1}\right), \\
H &= \sum_{k=1}^{K} n_k \left(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\right)^T, \\
E &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k\right)\left(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k\right)^T,
\end{aligned}
$$

on

$$\bar{\boldsymbol{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{ik},$$
$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \boldsymbol{x}_{ik},$$

i *tr*(·) és la funció traça d'una matriu (suma dels elements de la diagonal). La matriu $p \times p$ *H* mesura com de diferents són els centroides de les diferents classes al centroide global, és a dir, descriu la variabilitat entre-classes. La matriu $p \times p$ *E* mesura com de dispers són les mostres dintre la classe, és a dir, descriu la variabilitat intraclasses.

Notem l'analogia amb l'estadístic *F* d'ANOVA, on en ambdós casos es divideix un terme que mesura la dispersió entre les classes respecte un terme que mesura la dispersió dintre les classes. En cas d'assumir normalitat, es pot veure que maximitzar l'estadístic T^2 es tradueix en minimitzar la superposició entre les *K* classes. Aquest estadístic no es pot calcular quan $p > n$ i tampoc penalitza la inclusió de variables, és a dir, sempre creix o es manté quan s'inclou una variable més.

Stepwise hybrid feature selection with T^2

Per tal d'identificar el subconjunt d'*m* gens que maximitza la discriminació entre classes es podria calcular l'estadístic T^2 per cada combinació possible. El problema d'aquesta estratègia és que requereix una enorme càrrega computacional quan el nombre de variables és elevat. Per exemple, buscar el millor subconjunt de 3 gens dintre de 15000 disponibles equival a calcular $3.37 \cdot 10^{12}$ estadístics, si cada càlcul tarda 10^{-6} segons, el temps total seria de 39 dies. Una manera d'identificar un subconjunt de gens amb alt poder discriminant, sense haver d'avaluar totes les combinacions, és mitjançant una estratègia *stepwise*, la qual redueix en gran mesura el nombre de càlculs. L'inconvenient d'aquesta estratègia respecte la primera és que el subconjunt identificat podria correspondre a un màxim local i no el global.

El següent algoritme descriu la metodologia *stepwise* que utilitza el mètode de Dziuda per tal d'identificar un subconjunt d'*m* gens, on $\boldsymbol{v}$ correspon al subconjunt provisional de variables durant els diferents passos de l'algoritme, i *s* al nombre de variables incloses

en ***v***:

1) S'inicia ***v*** amb una variable seleccionada a l'atzar de les *p* disponibles (*s* = 1).
2) S'inclou a ***v*** la variable que maximitza T^2 al combinar-la amb la seleccionada en (1) (*s* = 2).
3) Es repeteixen els següents passos fins que *s* = *m*:
   i) S'afegeix a ***v*** la variable que maximitza T^2 al combinar-la amb les *s* actuals (*s* = *s* + 1).
   ii) Per cada variable de ***v***, es calcula l'estadístic T^2 amb les *s* – 1 restants. Es guarda m*axT* com el màxim dels estadístics T^2 calculats en aquest pas.
   iii) Si *maxT* és més gran que l'anterior estadístic T^2 identificat per s – 1 variables, aleshores s'iguala ***v*** a les *s* – 1 variables amb les que s'ha calculat *maxT* (*s* = *s* – 1).
   iv) Si *s* < *m*, es torna a (i).

Modified bagging schema

El m*odified bagging schema* és un procediment de remostratge que genera nous *training sets* a partir del original. Quan *p*>>*n* el subconjunt de variables identificat mitjançant una metodologia *stepwise* pot patir d'*overfitting* i de ser inestable [97]. En aquesta situació generar nous *training sets* ajuda a identificar subconjunts més robustos evitant aquests dos problemes. El m*odified bagging schema* és un procediment que genera, a partir del *training set* original, *B* nous *training sets* mitjançant mostreig estratificat aleatori sense reemplaçament. Cadascun dels nous *B sets*, els quals anomenarem Monte Carlo *sets*, conté $(1 - \gamma_{OOB})\cdot n_k$ mostres de cada una de les *K* classes (arrodonit a la baixa).

Combinació de les tres metodologies

El primer pas del mètode de Dziuda consisteix en utilitzar l'*stepwise hybrid feature selection with* T^2 per tal d'identificar l'***Informative Set of Genes***, definit com el subconjunt de gens que contenen tota la informació rellevant per la diferenciació de les *K* classes. Tenint en compte aquesta definició es pot argumentar que limma també es

podria utilitzar per identificar aquest subconjunt, en concret, limma inclouria en l'*Informative Set of Genes* tots els gens amb diferències significatives en l'expressió mitjana de les diverses classes. Metodologies que, com limma, miren els gens un a un poden passar per alt aquells que la seva importància només es pot veure de forma multivariant. L'*Informative Set of Genes*, detectat mitjançant el mètode de Dziuda, té en compte la correlació entre gens i, per tant, presenta una nova capa d'informació útil per entendre la malaltia i els processos biològics que diferencien els subtipus de B-CLPD.

El primer pas per detectar l'*Informative Set of Genes* consisteix en **generar *M* biomarcadors** de *m* gens, on un biomarcador es defineix com un subconjunt de gens amb possible poder discriminant. Per generar-los s'utilitza el següent algoritme:

1) S'identifica en el *training set* un biomarcador de *m* gens mitjançant l'*stepwise hybrid feature selection with* T^2.
2) S'actualitza el *training set* eliminant els *m* gens inclosos en el biomarcador identificat en (1).
3) Es repeteix (1) i (2) fins que s'han generat *M* biomarcadors diferents.

A cada iteració de l'algoritme es redueix la quantitat d'informació disponible en el *training set* sobre la diferenciació de classes, de manera que a partir d'una certa iteració s'ha exhaurit tota la informació i els biomarcadors que es generen ja no tenen poder discriminant. El paràmetre *M* es fixa manualment i ha de ser suficientment gran com per arribar a generar biomarcadors sense informació.

El següent pas consisteix en **seleccionar els biomarcadors que contenen informació** rellevant i descartar la resta. Per fer aquesta selecció s'ajusta el model lineal

$$t_i^2 = \beta_0 + \beta_1 \ln(i) + e_i,$$

on t_i^2 és el valor de l'estadístic T^2 del biomarcador obtingut en la iteració *i* de l'algoritme. Aquest model captura correctament el decreixement de l'estadístic T^2 quan es van eliminant variables rellevants del *training set*. A la Figura 3.9 hi ha representat, com a exemple, el decreixement obtingut en unes dades d'aquesta tesi, on es veu com el model

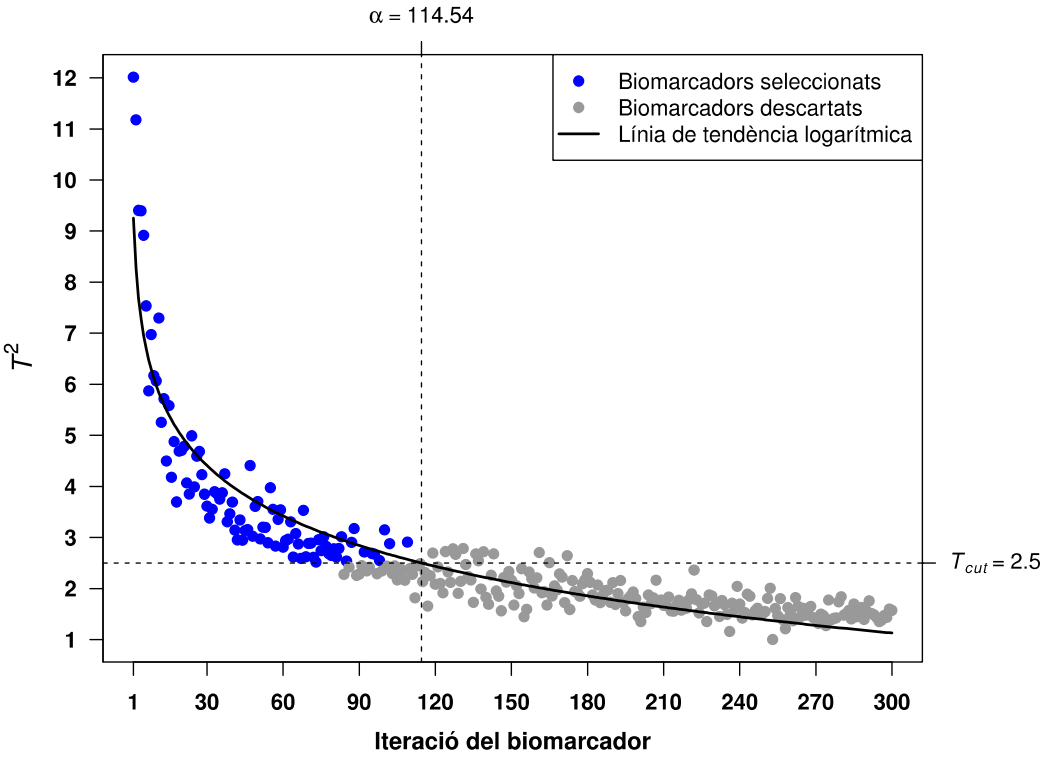

Figura 3.9: Decreixement de l'estadístic T^2. *A la figura hi ha representat el valor de l'estadístic T^2 dels biomarcadors identificats en cada iteració de l'algoritme (mètode de Dziuda). Es pot veure com l'estadístic decreix a mesura que es van eliminant variables del training set al llarg de les iteracions.*

logarítmic (línia negra) ajusta força bé el decreixement de T^2 al llarg de les iteracions.

Amb les estimacions dels paràmetres β_0 i β_1 del model lineal es calcula

$$\alpha = \exp\left(\frac{T_{cut} - \hat{\beta}_0}{\hat{\beta}_1}\right),$$

on T_{cut} s'especifica manualment i ha de representar bé a partir de quin valor de l'estadístic T^2 s'ha exhaurit la informació dels biomarcadors. De la fórmula es pot veure que α correspon a la iteració en què el model lineal ajustat prediu un valor igual a T_{cut} (Figura 3.9). El fet que α pugui prendre valors no enters no és important, ja que no afectarà als càlculs posteriors. L'especificació de T_{cut} es fa mitjançant la inspecció visual del gràfic resultant equivalent al de la Figura 3.9.

Aleshores, es considera que els biomarcadors que contenen informació rellevant són aquells que compleixen $i < \alpha$ i $t_i^2 > T_{cut}$. Els gens continguts en aquests biomarcadors són els que s'inclouen en l'*Informative Set of Genes*. En l'exemple de la Figura 3.9 els biomarcadors rellevants són els que corresponen als punts blaus, la resta de biomarcadors es descarten.

Un cop identificat l'*Informative Set of Genes*, el mètode de Dziuda utilitza el *modified bagging schema* per **crear *B* Monte Carlo *sets*** de dos conjunts de dades diferents: un corresponent al *training set* original (l'anomenarem *original set*) i l'altre corresponent al *training set* que només inclou els gens de l'*Informative Set of Genes* (l'anomenarem *INF set*). És a dir, es creen $B \cdot 2$ Monte Carlo *sets*, la meitat provinents de l'*original set* i l'altra meitat de l'*INF set*. En cada un dels Monte Carlo *sets* creats s'identifica *m* gens amb poder predictiu mitjançant l'*stepwise hybrid feature selection with* T^2.

L'últim pas del mètode de Dziuda consisteix en **calcular, per cada gen, dos *scores***. El primer *score* correspon al percentatge de vegades que el gen ha estat seleccionat en els *B* Monte Carlo *sets* provinents de l'*original set* (S^{ALL}). El segon *score* es calcula igual però amb els *sets* provinents de l'*INF set* (S^{INF}). Els gens candidats a traslladar-se a qPCR segons el mètode de Dziuda són aquells que tenen un valor elevat en S^{ALL} i S^{INF}. En aquesta tesi s'ha fixat $B = 1000$, $\gamma_{OOB} = 0.2$, $M = 300$, $T_{cut} = 2.5$ i $m = 3$. A causa del cost computacional d'aquest mètode, s'ha filtrat un 50% de les *probesets* (apartat 3.3.2). L'Annex A conté el codi d'R per calcular S^{INF} i S^{ALL}.

Depenent de l'objectiu de l'estudi és important comprovar si s'han inclòs la majoria dels gens amb informació en l'*Informative Set of Genes*. La identificació d'aquest conjunt depèn de dos paràmetres fixats per l'usuari (M i T_{cut}) i que, per tant, podrien estar mal especificats perjudicant la selecció dels gens. Una manera de fer-ho és comparant dos errors de predicció, un estimat amb el *training set* sense incloure'ls (*no-INF set*) i l'altre estimat amb l'*original set*. Si la diferència és gran significaria que la majoria dels gens rellevants s'han seleccionat. Aquest pas és especialment important en situacions en què, per exemple, es vol aplicar un g*ene set enrichment analysis* [98], un tipus d'anàlisis que permet detectar si un conjunt de gens està enriquit en algun tipus concret de funció

cel·lular. No incloure tots els gens rellevants podria perjudicar la detecció d'aquestes funcions. En aquesta tesi, on l'objectiu és seleccionar uns pocs gens a traslladar a qPCR, només prenen importància aquells amb més poder discriminant i, per tant, no és indispensable realitzar aquesta comprovació. Agafant d'exemple el cas de la Figura 3.9, encara que el valor de T_{cut} hagués deixat gens amb informació en els biomarcadors descartats, els gens amb més poder discriminant i més probables de ser traslladats estaran continguts en les primeres iteracions.

En cas de realitzar aquesta comprovació es pot utilitzar el següent algoritme, aplicat a l'*original set* i al *no-INF set* per separat, per estimar els dos errors de predicció:

1) Es crea *B* Monte Carlo *sets* mitjançant el *modified bagging schema*. Per cadascun es guarden les mostres que no s'hi han inclòs, les quals anomenarem mostres *out-of-bag* (OOB).
2) S'aplica a cada Monte Carlo *set*:
   i) Es selecciona *m* gens mitjançant l'*stepwise hybrid feature selection with* T^2.
   ii) S'ajusta un model predictiu LDA utilitzant els *m* gens obtinguts a (i).
   iii) S'aplica el model ajustat en (ii) en les respectives mostres OOB del Monte Carlo *set*. Es guarda el percentatge d'errors en aquestes mostres (discrepància entre el grup predit i el grup real).
3) S'estima l'error de predicció com la mitjana dels *B* percentatges calculats en (2).

Abans de calcular els dos *scores* (S^{ALL} i S^{INF}), el mètode de Dziuda realitza un procediment addicional per refinar la selecció de l'*Informative Set of Genes*. El primer pas d'aquest procediment consisteix en agrupar els gens de l'*Informative Set of Genes* en *C* clústers, de forma que gens molt correlacionats quedin agrupats conjuntament. Aleshores, per cada clúster, es calcula el percentatge de vegades que algun gen contingut en aquest s'ha seleccionat en els *B* Monte Carlo *sets* provinents de l'*INF set*. De la mateixa manera es calcula el percentatge segons els Monte Carlo *sets* provinents de l'*original set*. El mètode de Dziuda assumeix que els clústers amb percentatges més alts podrien contenir informació biològica més rellevant sobre la diferència entre subtipus. La llista de gens candidats es podria reduir als continguts en aquests clústers. Tenint en compte l'objectiu de seleccionar uns pocs gens, aquest pas no és necessari,

donat que els gens amb *scores* més alts sempre formaran part de clústers amb percentatges alts.

Existeixen altres metodologies que avaluen les variables de manera multivariant, com per exemple l'RFE-SVM [99] i les diverses mesures d'importància de variables del mètode r*andom forests* [100]. El motiu per prioritzar el mètode de Dziuda davant d'altres és la informació addicional que se'n obté a través de l'*Informative Set of Genes*, el qual facilita la interpretació biològica i la identificació de gens similars, una característica rellevant per adreçar els problemes exposats a l'apartat 1.4.4.

3.5 Estimació del rendiment d'un predictor: *Cross-validation*

El principal objectiu d'aquesta tesi és construir dos predictors, un en dades de *microarrays* i un en dades de qPCR, que classifiquin nous pacients de B-CLPD a una de les entitats estudiades. Una pregunta crucial quan es construeix un predictor és com de precís serà en aquests nous pacients en què s'apliqui, no solament per valorar-ne la utilitat a la pràctica, sinó també per poder comparar diferents models predictius entre si. Quan el procés que ha generat les dades és desconegut, aquesta precisió s'ha d'estimar a través de les dades disponibles.

La metodologia clàssica per fer-ho és construir el predictor amb les dades del *training set*, per després aplicar-lo a un segon conjunt independent de mostres (*test set*) on es compara la classe predita amb la real. El problema d'aquesta estratègia és que requereix un nombre molt elevat de mostres a l'haver-les de separar en dos *sets* (*training* i *test*), motiu pel qual no es pot utilitzar en entorns on el nombre de mostres és molt limitat. Un baix nombre de mostres al *training set* perjudica l'estimació dels paràmetres del model, mentre que un baix nombre de mostres al *test set* perjudica l'estimació de la precisió.

La *cross-validation* [101] és una metodologia que permet estimar la precisió del predictor sense necessitat de dividir el reduït nombre de mostres en dos *sets*. Suposem que un cert *training set T* inclou *N* casos i X covariables. La variable d'interès *Y* és

dicotòmica. Utilitzant T s'ha construït un predictor $f(\cdot)$ per predir Y a través d'X. Aleshores, l'interès està en calcular l'error de classificació esperat d'$f(\cdot)$, definit com la $P(Y_t \neq f(X_t))$ en una nova mostra $\{Y_t, X_t\}$ seleccionada a l'atzar de la població. El següent algoritme descriu com aplicar la *cross-validation* en el *training set* T per estimar-la:

1) Es separen les N mostres de T en K divisions (*folds*) a l'atzar, de manera que cada *fold* contindrà, aproximadament, N/K casos. Definim T_{-k} com el *set* resultant d'excloure les mostres de T corresponents al *fold* $k = \{1, 2, \ldots, K\}$, i definim T_k com el *set* complementari a T_{-k}.
2) S'aplaca a cada *fold* k:
   i) Es construeix el predictor $f_k(\cdot)$, el qual s'obté d'utilitzar la mateixa metodologia que per construir $f(\cdot)$ però només utilitzant el *set* T_{-k}.
   ii) Es calcula, per cada mostra j del *set* T_k, $\hat{Y}_j = f_k(X_j)$.
3) Finalment, s'estima la probabilitat segons

$$P\left(Y_t \neq \hat{f}\left(X_t\right)\right) = \frac{1}{N}\sum_{j=1}^{N} I\left(Y_j \neq \hat{Y}_j\right).$$

És a dir, mitjançant la *cross-validation* s'obté una predicció per cadascuna de les mostres de T amb un predictor construït sense utilitzar la respectiva mostra. Aleshores s'estima l'error de classificació comparant $\hat{Y}$ amb els valors Y observats.

A part de per estimar l'error de classificació, la *cross-validation* també es pot utilitzar per estimar paràmetres dels diferents mètodes de classificació. Per exemple, el mètode NSC (apartat 3.4.2) depèn del paràmetre Δ, què ha de ser definit per l'usuari. En aquesta situació la *cross-validation* es pot utilitzar per estimar l'error de classificació que s'obtindria per cadascun dels diferents valors de Δ. Aleshores, es fixaria el paràmetre al valor que minimitzés l'error. La *cross-validation* també es pot utilitzar per comparar l'error de predictors construïts mitjançant diferents mètodes de classificació.

En el procés de la *cross-validation* hi ha algunes consideracions a tenir en compte, com per exemple, el nombre de *folds* K i si s'utilitza una estratègia estratificada o no, on estratificar significa repartir les N mostres entre els K *folds* de manera que, en cada *fold*,

la proporció de casos de cada classe és similar a la proporció del *training set* sencer. En aquesta tesi s'ha decidit estratificar i fixar $K = 10$, donat que obté bons resultats a la pràctica [102]. En les anàlisis on alguna classe C ha complert $n_C < 10$, s'ha utilitzat $K = \min_C(n_C)$. L'última consideració que s'ha tingut en compte és que s'ha repetit el procediment de la *cross-validation* múltiples vegades, on en cada repetició la repartició dels casos en els diferents *folds* ha sigut diferent i a l'atzar. L'estimació final s'obté de la mitjana de les estimacions obtingudes en cada repetició. Repetir el procediment augmenta el cost computacional però redueix la variabilitat de l'estimació [103,104].

Tot i que, en general, les estimacions de l'error de predicció mitjançant *cross-validation* no pateixen de biaix, si ho fan de variància [102,103]. Aquest fet encara s'agreuja més quan el nombre de mostres és reduït [105], el qual pot provocar que en un estudi s'obtingui una estimació de l'error esperat molt per sobre o molt per sota del real. En aquests estudis s'extraurien conclusions molt inexactes i allunyades de la realitat. Per aquest motiu és important validar els predictors en sèries de casos independents [106].

Existeix una segona família de mètodes, basats en *bootstrap* [107], que també serveixen per estimar l'error de predicció i no pateixen de tanta variància. Dintre aquesta família, el mètode *bootstrap* 0.632+ [108] ha mostrat un bon rendiment a la pràctica, però, tot i que obté estimacions menys variables, en situacions d'alta dimensionalitat pateix un fort biaix [103,109]. Per tant, en aquesta tesi no és recomanable utilitzar la família de mètodes *bootstrap* davant la família *cross-validation*.

En la literatura es poden trobar múltiples estudis, especialment en l'entorn de la genètica, on s'ha aplicat incorrectament la *cross-validation* [110]. En aquests estudis no s'ha tingut en compte part del procediment utilitzat per construir el predictor, el qual provoca que les estimacions dels errors estiguin esbiaixades [111,112]. La manera correcta d'aplicar-la és replicant tot el procés utilitzat per construir $f(\cdot)$ en la construcció dels $f_k(\cdot)$ de cada *fold*. Per exemple, si en el procés de construcció d'$f(\cdot)$ es seleccionen les 5 millors variables d'entre 1000 disponibles segons un determinat criteri, aquest criteri de selecció també s'ha d'aplicar a cada *fold k*. El biaix apareix quan la selecció de les 5 variables es fa amb el *training set* sencer en comptes de a cada *fold*

independentment. Un segon exemple, suposem que s'ha construït el predictor $f_{NSC}(\cdot)$ mitjançant el mètode NSC, on s'ha optimitzat el paràmetre Δ mitjançant *cross-validation*. Per estimar l'error de predicció d'$f_{NSC}(\cdot)$ s'hauria d'aplicar una *cross-validation* interna a cada *fold k* de la *cross-validation* externa, de manera que la interna s'utilitzaria per estimar el paràmetre Δ de manera independent a cada *fold*. En aquest escenari es construirien $K_{ext} \cdot K_{int}$ predictors, és a dir, un predictor per cada un dels K_{int} *folds* interns de cada K_{ext} *fold* extern.

En aquesta tesi la *cross-validation* s'ha utilitzat per estimar les següents mesures de rendiment:

- **Sensibilitat**: probabilitat de que un pacient de l'entitat d'interès es predigui com a tal.
- **Especificitat**: probabilitat de que un pacient d'una entitat diferent a la d'interès es predigui com a tal.
- **Error de classificació**: probabilitat de que l'entitat predita d'un pacient no correspongui a la real. Aquesta mesura no es pot utilitzar com a estimació de l'error en dades futures si la freqüència de casos de cada entitat en el *training set* és diferent a la freqüència poblacional.
- **Precisió balancejada**: en un predictor que discrimina dues classes es defineix com (sensibilitat + especificitat) / 2. En un predictor que discrimina múltiples classes es defineix com la mitjana de les sensibilitats de cada classe.
- **Error de classificació balancejat**: S'obté de fer 1 – Precisió balancejada.

Hi ha una última consideració a ressaltar quan s'utilitza la *cross-validation*. Definim les mesures:

- $Err_T = P(Y_0 \neq f(X_0)|T)$.
- $Err = E_T[P(Y_0 \neq f(X_0)|T)] = E_T[Err_T]$.

és a dir, Err_T correspon a la probabilitat d'error d'un predictor construït mitjançant el *training set T* fixat, mentre que *Err* correspon a fer la mitjana dels Err_T obtinguts de cada possible *training set T* de la població. A la pràctica, l'interès està en estimar correctament Err_T, donat que és l'error específic del predictor construït mitjançant l'únic

training set recollit. Malauradament la família de mètodes de la *cross-validation* (i de *bootstrap*) estimen millor *Err* que Err_T, donat que l'estimació de la segona és complicada quan només es disposa de les dades del propi *training set T* [22]. Aquest fet no té gaires implicacions a la pràctica, especialment quan només s'utilitza la *cross-validation* per comparar predictors, però és un detall que es pot tenir en compte quan també hi ha l'opció d'estimar l'error mitjançant una sèrie de validació, on aleshores si que s'estima Err_T.

3.6 Disseny de l'estudi

La Figura 3.10 resumeix el procés utilitzat per construir els dos predictors (*microarrays* i qPCR). Els quadres blaus fan referència a les sèries de pacients, els verds als mètodes estadístics, els taronges als resultats obtinguts dels mètodes estadístics i en blanc els detalls experimentals.

El primer pas de l'estudi ha sigut mesurar l'expressió gènica dels 159 pacients inclosos en el *training set* mitjançant *microarrays* d'expressió, formant així el *training set* de *microarrays*. Aquest *set* s'ha utilitzat per construir un predictor (GEP55, branca esquerra de la figura), el qual s'ha utilitzat per predir una entitat en 30 B-CLPD, NOS. A l'apartat 4.4 es pot trobar detallat el procés de construcció d'aquest predictor, el qual s'ha basat en el mètode NSC.

El segon pas ha sigut re-analitzar el *training set* de *microarrays* per seleccionar 35 gens a traslladar a qPCR (branca dreta), el qual s'ha fet mitjançant limma i el mètode de Dziuda. Més detalls de la selecció de gens es poden trobar a l'apartat 4.5.

El tercer pas ha sigut mesurar, mitjançant qPCR, l'expressió dels 35 gens en 44 dels 159 casos inclosos en el *training set*, creant així el *training set* de qPCR. Aquest *set* s'ha utilitzat per construir un predictor (qPCR 8-*gene*) basat en corbes ROC, el qual s'ha provat en una sèrie independent de 63 mostres i, finalment, s'ha aplicat en 34 B-CLPD, NOS. Més detalls d'aquest segon predictor es poden trobar a l'apartat 4.6.

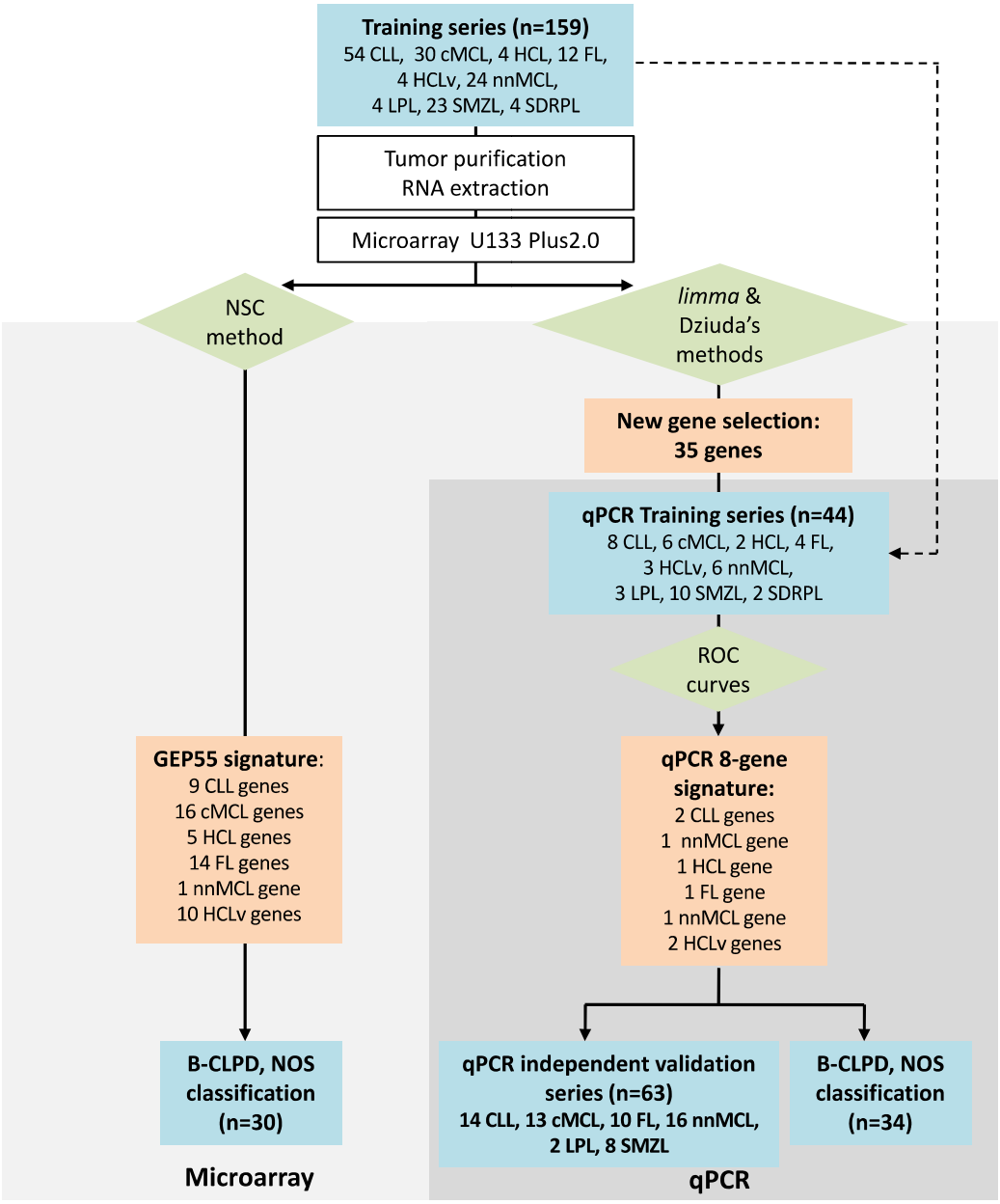

Figura 3.10: Representació esquemàtica del disseny de l'estudi. *Els quadres blaus fan referència a sèries de pacients, els verds a mètodes estadístics, els taronges a resultats obtinguts dels mètodes estadístics i en blanc detalls experimentals. Figura mantinguda en anglès com en l'article on s'ha publicat (apartat 4.8).*

A la figura 3.10 no s'han inclòs totes les anàlisis que s'han realitzat, sinó aquelles relacionades amb l'obtenció final dels dos predictors basats en expressió. Als apartats 4.1 i 4.2 es detallen anàlisis descriptives de les dades de *microarrays* d'expressió i de *copy-number*, respectivament. A l'apartat 4.3 es detalla la construcció del predictor que integra aquestes dues fonts. A l'apartat 4.7 es detalla com combinar informació

molecular i d'alteracions amb els predictors d'expressió (GEP55 i qPCR 8-*gene*) per millorar les prediccions d'algunes entitats de B-CLPD.

3.7 Mètodes estadístics addicionals

3.7.1 Ajust dels *P*-valors

A l'apartat 1.4.1 s'ha explicat que realitzar simultàniament una gran quantitat de tests d'hipòtesi augmenta en gran mesura la probabilitat d'obtenir falsos resultats positius. Per adreçar-ho es poden transformar els *P*-valors de manera que controlin l'FDR o l'FWER.

En aquesta tesi s'ha utilitzat el mètode de Benjamini-Hochberg [20], el qual controla l'FDR a almenys un nivell α. Aquest procediment també controla l'FDR en escenaris on hi ha dependència [113], similars als que es poden trobar en dades de *microarrays*. L'algoritme per calcular els *P*-valors ajustats d'*m* tests d'hipòtesi segons aquest mètode és el següent:

1) S'ordena els *m* *P*-valors de major a menor. Definim p_i com el *P*-valor en la posició *i*, per tant, p_1 és el *P*-valor més gran i p_m el més petit.
2) Es transforma els *P*-valors segons $p_i = p_i(m/(m-i+1))$. Si algun $p_i > 1$, aleshores $p_i = 1$.
3) Es calcula el *P*-valor ajustat (q_i) de la posició *i* segons $q_i = \min(p_1, \ldots, p_i)$.

3.7.2 Control de qualitat en *microarrays* d'expressió

A l'apartat 1.4.6 s'ha exposat la importància de detectar *microarrays* en què el procés d'hibridació no ha funcionat correctament, donat que excloure'ls de les anàlisis augmenta la potència estadística [45,46]. En aquesta tesi s'ha utilitzat dues mesures per detectar *microarrays* d'expressió de baixa qualitat, la NUSE i l'RLE [114], amb les que no s'ha detectat problemes de qualitat en cap dels 189 *microarrays*.

4 Resultats: construcció dels predictors

En aquest capítol es presenten els resultats obtinguts d'aplicar les metodologies descrites al Capítol 3. El primer i segon apartat consisteixen en les anàlisis descriptives de les dades provinents dels *microarrays* d'expressió i *microarrays* de *copy-number*, respectivament. En el tercer apartat es presenta el predictor obtingut d'integrar les dues fonts de dades. En el quart apartat es descriu i presenta el predictor obtingut de les dades d'expressió. En el cinquè apartat es presenta els resultats de la selecció de gens. El sisè apartat correspon a la construcció del predictor en dades de qPCR. En el setè apartat es presenta com combinar la informació d'expressió amb informació molecular i d'alteracions per millorar el diagnòstic dels pacients de B-CLPD. El vuitè apartat referencia l'article on s'han publicat part dels resultats presentats al llarg del capítol.

4.1 Anàlisi descriptiva: *microarrays* d'expressió

Una primera anàlisi descriptiva de les dades permet identificar, en cas d'existir, patrons d'expressió no-aleatoris en els casos estudiats. Un cop identificats aquests patrons es pot avaluar si estan relacionats amb els diferents subtipus de B-CLPD. En aquesta tesi les dades dels *microarrays* d'expressió s'han preprocessat segons el mètode fRMA, d'on s'ha obtingut una matriu formada per 189 casos (159 de la cohort *training* i 30 B-CLPD, NOS) i, després del filtratge, 15409 *probesets*. La representació conjunta d'aquesta informació no es pot fer directament i s'ha d'utilitzar algun tipus de metodologia que

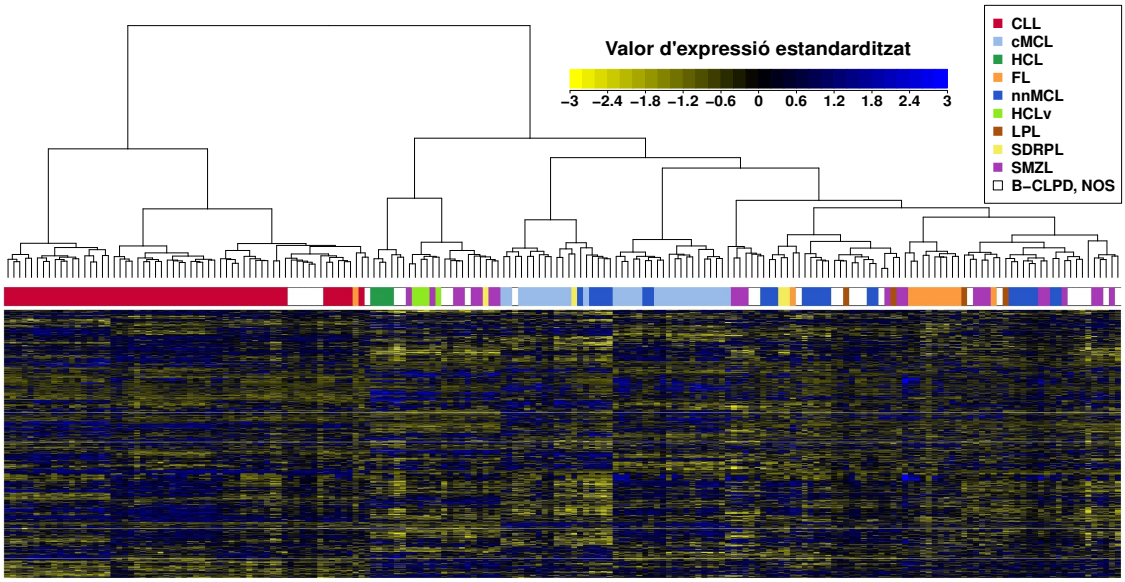

Figura 4.1: Clúster i heatmap de l'expressió per microarrays dels 189 casos. *Les files del heatmap corresponen a les probesets i les columnes a les mostres. El clúster s'ha calculat amb totes les probesets (p = 15409), mentre que al heatmap només s'hi ha representat l'expressió estandarditzada del 15% amb més IQR d'aquestes (p = 2311). Expressions altes de la probeset es representen amb blau i baixes amb groc.*

permeti resumir-la. Les tècniques de clusterització, juntament amb un *heatmap* dels valors d'expressió, són les més utilitzades a la literatura per fer-ho.

La Figura 4.1 mostra un clúster jeràrquic, construït segons el criteri de Ward [115] i la distància euclidiana, de les 189 mostres (columnes). Els càlculs del clúster s'han realitzat amb totes les *probesets* (files), mentre que al *heatmap* només s'hi ha representat l'expressió normalitzada del 15% amb més IQR d'aquestes (p = 2311). A la figura es pot veure una gran branca formada quasi exclusivament per mostres de l'entitat CLL, juntament amb 1 FL i 7 BCLPD-NOS. En l'altra gran branca es pot veure com les mostres de cMCL, HCL, HCLv i FL queden majoritàriament agrupades juntes segons els seus diagnòstics. En canvi, les mostres de nnMCL, SMZL, LPL i SDRPL es distribueixen en diferents branques del segon gran clúster.

La primera conclusió és que els patrons d'expressió més importants de les dades semblen estar relacionats amb els tipus de B-CLPD, el qual suggereix que pot existir un predictor basat en l'expressió gènica capaç de distingir aquestes entitats. La segona conclusió és que hi ha algunes entitats que tenen un perfil d'expressió global molt

diferenciat respecte les altres, en especial la CLL, on les 54 mostres diagnosticades d'aquesta entitat queden agrupades en la mateixa branca. En l'altre extrem es troba les mostres de l'entitat SMZL, les quals queden repartides en múltiples branques. L'última conclusió és que algunes mostres sense un diagnòstic clar (B-CLPD, NOS) formen part de branques formades quasi exclusivament per un únic subtipus de B-CLPD, el qual suggereix que l'expressió gènica podria ajudar a diagnosticar mostres que per altres criteris no ha sigut possible. Addicionalment, els patrons d'expressió de les 2311 *probesets* representades en el *heatmap* mostren diferències entre les diferents branques, però podrien no ser les més representatives per discriminar les classes.

4.2 Anàlisi descriptiva: *microarrays* de *copy-number*

L'estructura de les dades provinents dels *microarrays* de *copy-number* és molt diferent de la d'expressió. En aquest cas la matriu està formada per 130 casos (114 de la cohort *training* i 16 B-CLPD, NOS) i 28 variables binàries, obtingudes mitjançant el preprocessament descrit a l'apartat 3.3.3. Recordem que les variables indiquen si la mostra té o no una alteració en diferents punts del genoma. Per exemple, la variable 3q+ indica si la mostra té un guany de material genètic (+) al braç *q* del cromosoma 3.

A la Figura 4.2 s'ha representat, per cada cas (columna) i variable (fila), si està alterat (vermell) o no (gris). Per similitud a l'apartat anterior també s'ha aplicat un mètode de clusterització, en aquest cas la distància utilitzada ha sigut la de Manhattan, equivalent a comptabilitzar quantes discrepàncies hi ha entre dos casos. Les variables estan ordenades segons la freqüència de casos alterats. A la dreta de la figura s'ha representat el percentatge de casos alterats de cada variable en els subtipus CLL, cMCL, nnMCL i SMZL. S'ha decidit no representar els altres subtipus ja que només es disposa d'informació de 4 o menys casos.

El primer punt a destacar de la figura és que algunes alteracions són molt poc freqüents i només s'han detectat en 2 o 3 casos. També es pot veure que en les alteracions més freqüents hi ha diferències en el percentatge entre els subtipus de B-CLPD, per

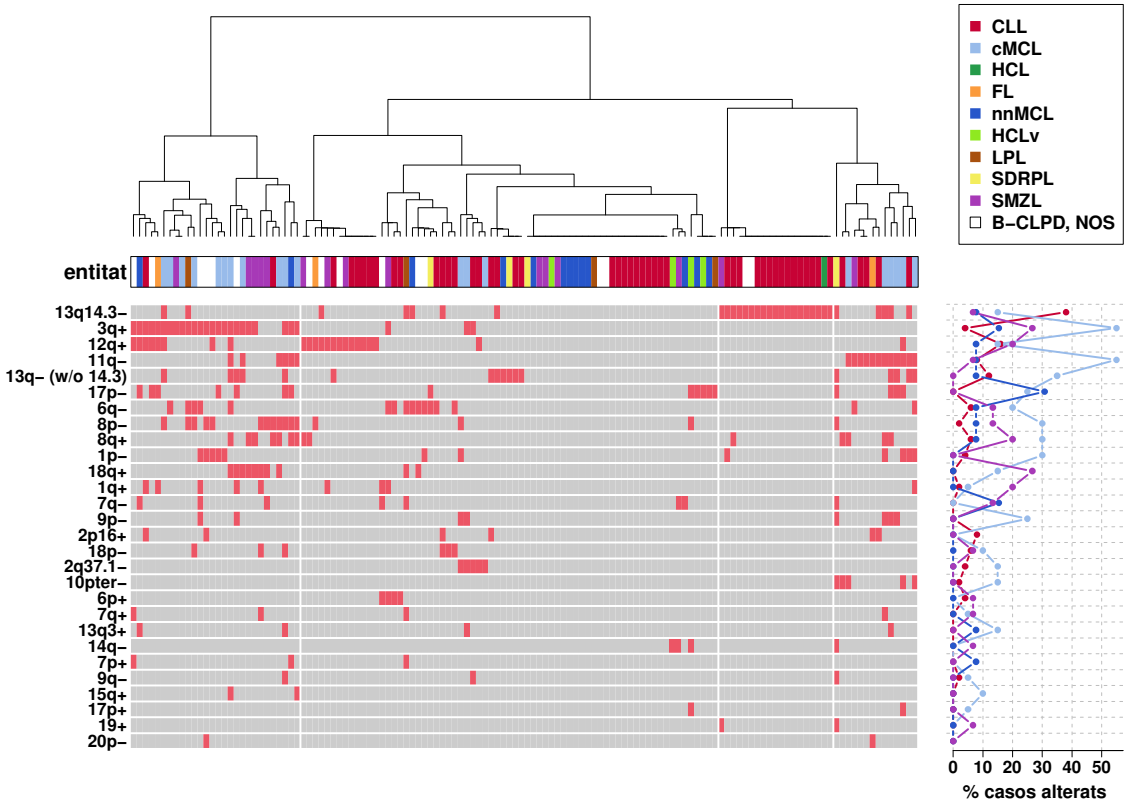

***Figura 4.2: Alteracions per microarrays de copy-number dels 130 casos**. El heatmap indica, per cada pacient, si té (vermell) o no (gris) cadascuna de les alteracions. A la dreta hi ha representat el percentatge de casos alterats en les entitats CLL, cMCL, nnMCL i SMZL.*

exemple, 13q14.3- està present en gairebé un 40% de les mostres CLL, mentre que per les altres entitats no arriba al 20%. En canvi, les alteracions en 3q+ i 11q- són més freqüents en cMCL. Si es mira, a través del *heatmap* i el clúster, el perfil global d'alteracions no s'identifica un patró clar relacionat amb les diferents entitats. A priori sembla que les dades provinents dels *microarrays* de *copy-number* tenen menys informació que les d'expressió en quant a discriminació de classes, tot i així, algunes alteracions podrien complementar la informació d'expressió.

4.3 Predictor basat en la integració d'expressió i *copy-number*

Un punt important a contestar en la construcció del predictor en dades de *microarrays* és si utilitzar les dues tecnologies, expressió i *copy-number,* augmenta considerablement la precisió del predictor respecte a utilitzar-ne una. Els mètodes *kernel*, explicats en

l'apartat 3.4.1, permeten avaluar si la informació que aporta cada tecnologia per a la distinció de classes és rellevant i complementària a l'altra. En aquesta tesi s'ha utilitzat el mètode SVM amb la funció *kernel* lineal normalitzada, ponderant les fonts d'informació a través del paràmetre β. En concret, un valor de $\beta = 1$ significaria que el predictor s'ha construït exclusivament amb les dades d'expressió, mentre que un valor de $\beta = 0$ significaria que s'ha construït només amb les dades de *copy-number*.

Hi ha diversos aspectes a considerar abans de poder aplicar aquesta metodologia:

1) El mètode no implementa de forma natural la discriminació multiclasse, sinó que només permet discriminar dues classes.
2) Dos paràmetres del mètode s'han d'especificar per l'usuari: $C \in (0, \infty)$, el qual indica com de permissius som en deixar mostres fora dels hiperplans, i $\beta \in [0, 1]$, el qual indica el pes de les dades d'expressió.
3) La quantitat de variables a utilitzar de cada font (p_E i p_D) per construir el predictor també s'ha d'especificar a priori. En les dades d'expressió hi ha 15409 variables disponibles i en les dades de *copy-number* n'hi ha 28. A més de decidir quantes, també s'ha d'especificar quines utilitzar concretament.

Per adreçar el primer punt s'ha transformat el problema multiclasse a múltiples binaris, de manera que a cada problema binari sí es pot aplicar el mètode. En concret, s'han plantejat els següents quatre problemes binaris:

- P1: CLL ($n = 50$) vs la resta d'entitats ($n = 64$).
- P2: cMCL ($n = 20$) vs la resta d'entitats ($n = 94$) .
- P3: nnMCL ($n = 13$) vs la resta d'entitats ($n = 101$).
- P4: SMZL ($n = 15$) vs la resta d'entitats ($n = 99$).

Per les altres 5 entitats (FL, HCL, HCLv, LPL, SDRPL) només es disposa d'entre 1 i 4 casos amb informació completa, una quantitat molt limitada per construir un predictor amb una estructura tant complexa.

La selecció de paràmetres dels punts 2 i 3 (C, β, p_E i p_D) s'ha adreçat mitjançant l'ús de la

cross-validation, la qual permet estimar la precisió de les prediccions per cada combinació possible de C, β, p_E i p_D. Un cop obtingudes les precisions, es pot estudiar i fixar els valors més adequats per utilitzar en el predictor final. En concret, s'ha explorat els següents valors de cada paràmetre: C = $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, β = {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1} i les parelles $[p_E, p_D]$ = {[5, 2], [50, 3], [100, 5], [1000, 15]}. El fet de valorar els paràmetres p_E i p_D en parella és per simplificar la xarxa de les possibles combinacions, a més, facilita la interpretació dels resultats, ja que els valors explorats representen un model de menor a major complexitat. El procés de la *cross-validation* s'ha realitzat amb 10 *folds* i s'ha repetit 20 vegades.

Finalment, la selecció de variables s'ha adreçat mitjançant estadístics univariants que les ordenin, per després utilitzar les p_E i p_D primeres. L'estadístic utilitzat en el cas de la dades d'expressió ha sigut limma (apartat 3.4.3), mentre que per les dades de *copy-number* ha sigut el test exacte de Fisher.

A la Figura 4.3 hi ha representada, per cadascun dels quatre problemes binaris, la precisió balancejada estimada pels diferents valors dels paràmetres. Per cada combinació de β i $\{p_E, p_D\}$ s'ha representat la precisió màxima al llarg dels valors de C, donat que aquest paràmetre no té interpretació ni interès més enllà d'optimitzar-lo.

El resultat més rellevant és que, en tots quatre predictors, hi ha una tendència clara d'augment de la precisió quan s'incrementa β, indicant que les dades d'expressió tenen més poder discriminant per distingir els diferents subtipus de B-CLPD que les de *copy-number*. Si només es té en compte les dades de *copy-number* (β = 0), no es pot distingir nnMCL o SMZL de la resta d'entitats, atès que la precisió estimada en aquestes dues entitats és aproximadament del 50%. En canvi, quan es vol distingir CLL i cMCL les precisions són aproximadament del 70% i 75%, respectivament. Si només es té en compte les dades d'expressió (β = 1), les entitats CLL i cMCL tenen precisions gairebé perfectes, el qual no deixa marge a les de *copy-number* a complementar-les. En les entitats nnMCL i SMZL les precisions amb β = 1 arriben a 87% i 82%, respectivament. Només en l'entitat SMZL s'observa que β = 0.9 té una precisió lleugerament més elevada que β = 1. Per últim, en les entitats CLL, cMCL i nnMCL el rendiment màxim

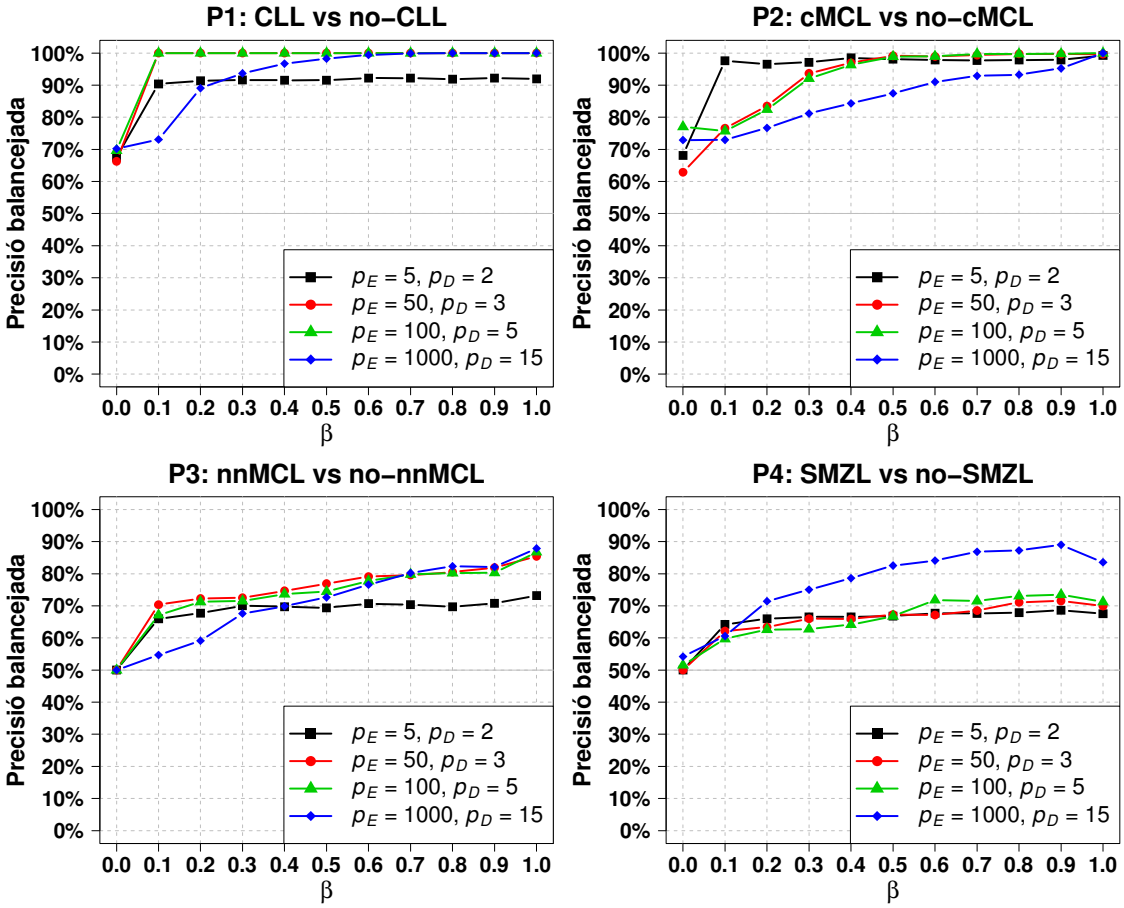

Figura 4.3: Precisions dels models integradors de dades. *En la figura hi ha representada la precisió balancejada per diferents valors dels paràmetres dels quatre problemes binaris. La precisió s'ha estimat mitjançant 10-fold cross-validation repetida 20 vegades. Per cada combinació de β i {p_E, p_D}, només s'ha representat la precisió estimada més alta al llarg dels valors de C.*

ja s'aconsegueix en valors de p_E i p_D relativament baixos ($p_E \leq 50$). En SMZL el rendiment és màxim quan l'estructura inclou $p_E = 1000$ i $p_D = 15$ variables.

Els resultats obtinguts d'aquesta metodologia suggereixen que **per construir el predictor només fa falta utilitzar les dades d'expressió**. Aquesta font té molt poder per distingir les entitats CLL i cMCL, per tant, encara que es disposi de les dades de *copy-number*, ja no fa falta considerar-les en la modelització al no poder millorar la precisió de les prediccions. En les entitats SMZL i nnMCL la situació és diferent, on cap de les dues fonts té tant poder discriminant. Les dades d'expressió per si soles arriben a obtenir una precisió del 90%, mentre que en les de *copy-number* és del 50%, equivalent a predir l'entitat a l'atzar. Encara que en aquestes entitats hi hagi marge de maniobra per

millorar les prediccions, les dades de *copy-number* no tenen informació suficient i, per tant, no són capaces de complementar l'expressió. Les freqüències de les alteracions en SMZL i nnMCL són baixes, dificultant l'obtenció d'informació útil.

L'única situació en la que combinar les fonts incrementa la precisió del predictor és quan es discrimina l'entitat SMZL utilitzant un gran nombre de variables (p_E = 1000 i p_D = 15), on amb β = 0.9 es maximitza la precisió balancejada al 87% aproximadament (corresponent a una sensibilitat del 78% i una especificitat del 97%). El fet que es necessitin tantes variables i que l'especificitat sigui força més elevada que la sensibilitat podria indicar que en realitat no s'està discriminant SMZL, sinó que a cada variable que s'inclou s'està utilitzant informació que discrimina les altres entitats de mica en mica, notem el poc canvi que hi ha entre $\{p_E = 5, p_D = 2\}$ i $\{p_E = 50, p_D = 3\}$.

En conclusió, sembla adequat construir el model predictor només amb les dades d'expressió, donat que la millora que s'aconsegueix al combinar les dues fonts és nul·la o molt petita. Un avantatge addicional de fer el predictor només amb dades d'expressió és que es disposa de més mostres en la majoria d'entitats, amb la subseqüent millora de poder analitzar totes les entitats i obtenir estimacions més precises dels paràmetres.

4.3.1 Limitacions

Els resultats presentats mostren que les dades de *copy-number* no complementen les d'expressió, però s'ha de tenir en compte que tenen una estructura que requereix un nombre elevat de mostres per poder-les valorar i utilitzar de forma fiable. Les dades analitzades es tracten de variables binàries que indiquen si el pacient té una alteració o no a una determinada localització del genoma. Les freqüències de casos alterats en la majoria d'aquestes variables són relativament baixes, així que si s'afegeix que el nombre de mostres també ho és, s'acaba disposant de pocs casos alterats per estudiar si existeix una relació amb les diferents entitats de B-CLPD. Aquesta situació s'agreuja especialment en el cas d'estudiar relacions multivariants.

En aquesta tesi, per tal d'augmentar aquestes freqüències, s'ha resumit les dades agrupant localitzacions properes, per exemple, els casos amb alteració en 3q+ tenen el guany en diferents localitzacions del braç *q* del cromosoma 3, però s'han agrupat en una única variable per no tenir la informació tant dispersa. Tot i així, les freqüències obtingudes d'aquest procediment no són suficientment altes com per valorar la majoria d'alteracions de manera fiable. S'ha de tenir en compte que aquesta agrupació es pot fer si les alteracions dels casos agrupats estan situades en una regió comuna aproximada.

Addicionalment, l'estructura multivariant de les alteracions podria ser rellevant. Per exemple, en la Figura 4.2 es pot veure un subgrup de pacients que no tenen cap alteració excepte en 13q14.3-. Aquest subgrup està format per 13 CLL, 1 HCL i 2 B-CLPD, NOS, pel que es pot concloure que tenir aquesta alteració com a única és molt indicatiu de CLL, un resultat que es pot verificar a la literatura ja que s'ha descrit en molta major freqüència en aquesta entitat [5]. En aquesta tesi s'ha pogut identificar aquest perfil multivariant gràcies a que 13 de 50 (26%) mostres diagnosticades com a CLL el comparteixen. Aleshores, perfils similars podrien existir en la resta d'entitats, però al disposar de menys de 20 mostres en cadascuna és molt difícil de valorar si no la presenten la majoria dels casos.

4.3.2 Validesa de la integració intermèdia

La metodologia d'integració intermèdia utilitzada assumeix que no hi ha correlació entre variables provinents de diferents fonts (entrefonts), donat que només modelitza l'estructura dins una mateixa font (intrafont). Quan el nombre de mostres és limitat les estimacions d'estructures complexes no són precises, així que simplificar-les ajuda a evitar l'*overfitting*. Tot i així, si les dues fonts estiguessin fortament relacionades es podria plantejar utilitzar una estratègia d'integració prematura, és a dir, es podria calcular la matriu *kernel* resultant de concatenar les dues bases de dades en una única. Aquesta modelització tindria en compte la redundància d'informació en les dues fonts, una característica que no té la integració intermèdia. Precisament, la integració intermèdia obté precisions més elevades quan l'estructura de correlació intrafont és més

forta que l'entrefonts [72].

Les dades d'aquesta tesi podrien no complir aquest requisit, donat que es disposa d'una font que mesura el DNA i una altra que mesura l'RNA, dos nivells que estan biològicament lligats. Aleshores, és convenient considerar la possibilitat que si un pacient té una pèrdua en, per exemple 11q, l'expressió dels gens en aquesta localització es vegi afectada. En aquest cas, la informació del DNA i de l'RNA seria redundant i el model no ho aprofitaria.

Estudiar la relació entre les dues fonts en aquesta tesi és difícil a causa, principalment, del limitat nombre de mostres de cada entitat i la baixa freqüència de les alteracions. A més, en la literatura s'han publicat estudis on s'identifica correlació a aquests dos nivells [116] i estudis on no [117,118]. Aquesta discrepància provoca que sigui difícil basar-se en altres estudis per prendre una decisió, donat que podria haver-hi un efecte que depengués del tipus de teixit, del càncer analitzat o de la metodologia experimental. Tot i així, es disposa de 19 CLL amb l'alteració 13q14.3- i 31 CLL sense, una quantitat suficientment gran per, almenys, estudiar-ho en aquesta localització i entitat.

La Figura 4.4 compara l'expressió de 20 gens localitzats en la regió q14.3 del cromosoma 13 entre els casos que tenen perdut aquest segment i els que no. No s'han representat tots els gens de la regió, sinó només aquells continguts en els segments alterats de tots els 19 casos (mínima regió comuna). El gràfic superior mostra, per cada gen, un *boxplot* dels casos amb l'alteració (vermell) i un dels casos sense (gris), on, en general, es veu força superposició. El gràfic inferior compara la mitjana del grup alterat (x_1) respecte la mitjana del grup no alterat (x_0) en cada gen, en concret, s'ha representat el valor mitjà de les mitjanes (eix x) respecte la diferència d'aquestes (eix y). En aquest segon gràfic es veu una lleugera tendència global a valors negatius, suggerint que els casos amb la deleció tenen menys expressió, especialment en el gen *DLEU2*.

Aquests resultats mostren que hi ha correlació entre les dues fonts d'informació, però també mostren que és relativament baixa a l'haver-hi força solapament en els *boxplots* i un canvi mitjà d'expressió força petit en tots els gens. Per tant, intentar modelitzar-la és

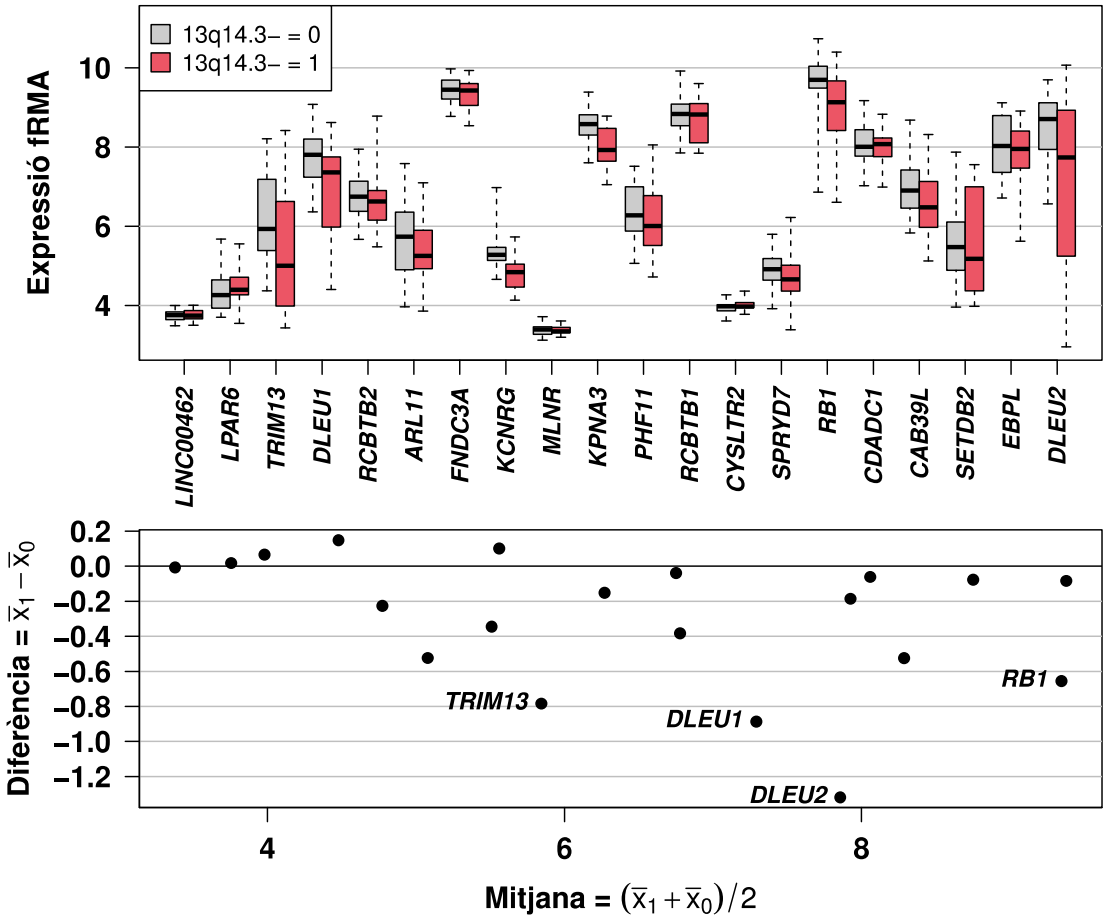

Figura 4.4: Efecte de l'alteració 13q14.3- en l'expressió. *El gràfic superior mostra boxplots dels nivells d'expressió de 20 gens en casos CLL amb l'alteració 13q14.3- (vermell) i sense (gris). Els 20 gens estan localitzats a la banda q14.3 del cromosoma 13. El gràfic inferior mostra el valor mitjà de les mitjanes de cada grup respecte la diferència d'aquestes mitjanes.*

difícil amb el baix nombre de mostres i s'augmentaria el risc d'*overfitting*. Si s'extrapola aquest resultat a la resta d'alteracions i entitats significaria que **es pot aplicar la metodologia d'integració intermèdia davant la d'integració prematura**.

4.4 Predictor basat en dades de *microarrays* d'expressió

Els resultats del model integrador de l'apartat 4.3 mostren que les dades d'expressió tenen molt poder discriminant per distingir els diferents subtipus de B-CLPD, mentre que les de *copy-number* no aporten informació suficient com per complementar-les. En vista d'aquests resultats és adequat plantejar-se construir un model que utilitzi només les

dades d'expressió, el qual tindria dos avantatges importants: es disposa de més mostres per construir-lo i és més fàcil d'interpretar. El mètode per construir-lo ha sigut el *nearest shrunken centroids* (NSC, apartat 3.4.2), el qual té dues propietats convenients: implementa de forma natural la selecció de variables i la discriminació multiclasse.

4.4.1 Enfocament multiclasse

El *training set* disponible consisteix de 54 CLL, 30 cMCL, 4 HCL, 12 FL, 24 nnMCL, 4 HCLv, 4 LPL, 23 SMZL i 4 SDRPL. Les mesures de rendiment del predictor per diferents valors del paràmetre Δ s'han estimat mitjançant una estratègia de *cross-validation* estratificada amb 4 *folds* i repetida 300 vegades. A causa de l'estratificació, el màxim nombre de *folds* que es pot utilitzar correspon al nombre de mostres en l'entitat que menys en té. S'han explorat 60 valors de Δ equidistants continguts en la regió [0, 22.13], on 22.13 és la quantitat de reducció que anul·la tots els gens.

L'NSC també assumeix que les probabilitats a priori de cada classe (π_k) són conegudes. Aquestes s'han fixat a les proporcions mostrals ($\pi_k = n_k/n$), donat que la distribució de casos en la cohort *training* aproxima la distribució poblacional. Per exemple, la incidència de l'entitat CLL és força més elevada que les incidències de la resta d'entitats, així com la incidència de HCL i HCLv és molt més baixa.

A la Figura 4.5 s'ha representat com varia la precisió balancejada global (línia negra) i la sensibilitat de les 9 classes (línies de colors) a mesura que augmenta Δ, on es pot veure que la precisió global és d'aproximadament del 68% quan s'utilitzen tots els gens (Δ = 0). Aquesta precisió augmenta lleugerament fins el 72% amb Δ = 2.25, per després anar disminuint gradualment fins que la reducció no deixa cap gen al model.

A la figura també es pot veure que les sensibilitats de les entitats LPL i SDRPL estan al voltant del 0% per qualsevol valor de Δ, indicant que, independentment dels gens utilitzats, sempre es confonen amb altres entitats. En l'altre extrem es troba la CLL, la qual sempre manté la sensibilitat pròxima al 100%, fins i tot quan no s'utilitza cap gen

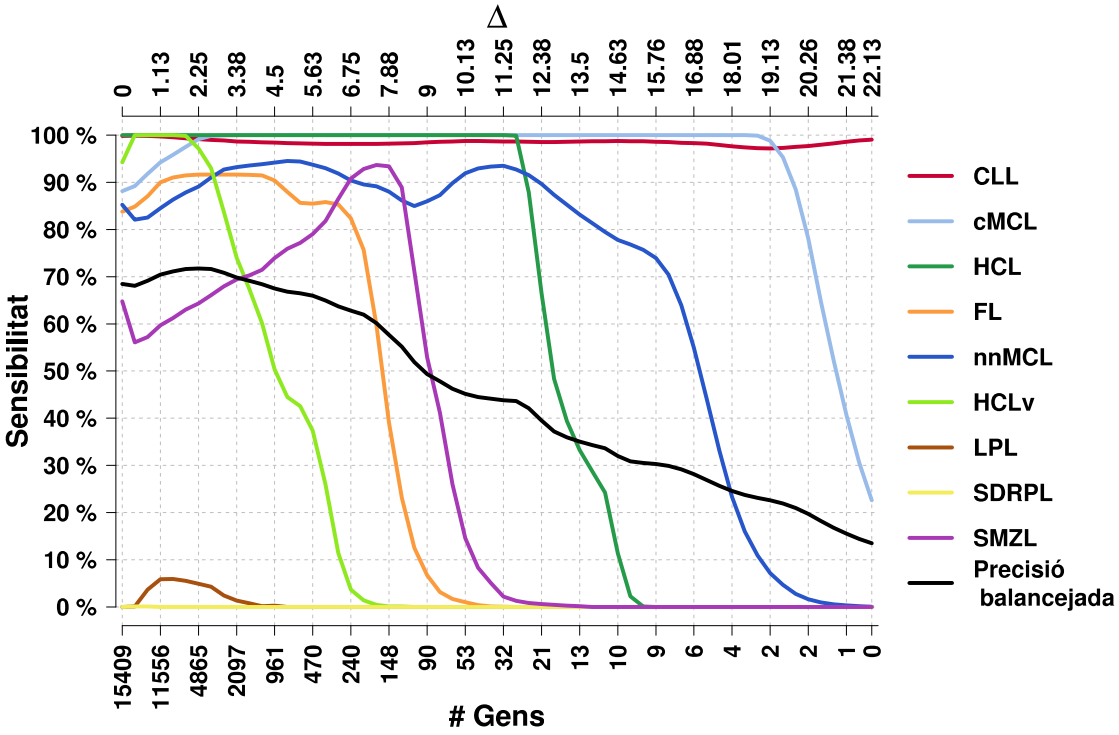

Figura 4.5: Rendiment del model NSC multiclasse. *Sensibilitat i precisió balancejada de cada classe segons el paràmetre Δ. A l'eix x s'indica el nombre de gens que utilitza el model amb el valor corresponent de Δ.*

(Δ = 22.13). Al ser l'entitat amb π_k més gran, quan no es disposa d'informació les prediccions s'assignen a CLL. Les entitats HCLv, FL, HCL, nnMCL i cMCL comencen amb valors relativament alts de sensibilitat i, a mesura que va augmentant Δ, arriba un punt que comencen a disminuir fins al 0%. En concret, a partir de Δ = 2.25 la sensibilitat d'HCLv comença a decréixer, la segueix la sensibilitat d'FL a partir de Δ = 6.75, després la d'HCL i la de nnMCL ho fan a partir de Δ = 11.25, per últim la de cMCL a Δ = 19.13.

El comportament de les sensibilitats observat en la Figura 4.5 té una explicació senzilla. El paràmetre Δ es pot interpretar, aproximadament, com una reducció als estadístics *T* que comparen l'expressió d'una entitat respecte la resta d'entitats en cada gen. Per exemple, la sensibilitat de l'entitat FL és del 0% quan Δ = 10, indicant que cap gen té un estadístic *T* superior a 10 quan es compara les mostres d'FL respecte la resta de mostres. D'aquí es pot concloure que les entitats CLL i cMCL tenen diferències molt grans en algun gen i que, per això, mantenen sempre sensibilitats properes al 100%. En canvi, les mostres de l'entitat HCLv no tenen valors d'expressió tant diferents en cap gen i una

reducció de 7 ja els iguala tots a 0.

La selecció del valor de Δ per construir el predictor final es pot fer en base a la Figura 4.5. Concretament, sembla adequat Δ = 2.25, ja que minimitza la precisió balancejada i s'obtenen sensibilitats elevades per moltes classes. El problema és que el model amb Δ = 2.25 inclou al voltant de 5000 gens, un nombre suficientment elevat com per sospitar que **el predictor està utilitzant molts gens innecessaris**. A més, augmentar Δ per tal de disminuir el nombre de gens perjudicaria la precisió balancejada.

4.4.2 Enfocament multiclasse escalat segons θ

El comportament de les sensibilitats de la Figura 4.5 suggereix que seria adequat utilitzar un valor de Δ diferent en cada entitat. Per exemple, un valor de Δ = 19 deixa suficient informació com per discriminar les entitats CLL i cMCL, aleshores, si s'utilitza un Δ < 19 s'estarien incloent al predictor gens que discriminen aquestes dues entitats però que no són necessaris, ja que amb els primers ja es fa correctament. Un cop discriminada una entitat l'interès està en discriminar la següent i, a la vegada, en deixar d'incloure gens al predictor que donen informació sobre entitats ja diferenciades.

El propi mètode NSC es pot modificar per tal d'adaptar-lo a aquesta situació, tal com està descrit a [119]. Recordem que, segons la formulació original, l'estadístic que comparava els components *j* del centroide de la classe *k* i del centroide global era

$$d_{jk} = \frac{\bar{x}_{jk} - \bar{x}_j}{m_k (s_j + s_0)},$$

el qual es pot adaptar segons

$$d^*_{jk} = \frac{\bar{x}_{jk} - \bar{x}_j}{\theta_k m_k (s_j + s_0)} = \frac{d_{jk}}{\theta_k},$$

on el vector de paràmetres $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$ escala tots els estadístics de cada classe *k*. Per exemple, assignar $\theta_{HCL} = 2$ provocaria que els estadístics de l'entitat HCL es

dividissin per dos, amb la conseqüència de que es necessitaria la meitat de reducció (Δ) per anul·lar els gens rellevants d'aquesta entitat. A través de $\boldsymbol{\theta}$ es pot igualar l'efecte de la reducció en totes les classes per tal d'evitar que s'incloguin molts gens innecessaris.

Tibshirani et al. [119] proposen un algoritme per identificar un conjunt de valors adequats pel vector $\boldsymbol{\theta}$, el qual a l'aplicar-lo a les dades d'aquesta tesi s'ha obtingut $\boldsymbol{\theta} = \{\theta_{CLL} = 1.11, \theta_{cMCL} = 1.11, \theta_{HCL} = 1.11, \theta_{FL} = 1.11, \theta_{nnMCL} = 1.11, \theta_{HCLv} = 1.11, \theta_{LPL} = 1.11, \theta_{SDRPL} = 1.11, \theta_{SMZL} = 1\}$. El vector proposat per l'algoritme és un canvi mínim respecte la formulació original i, per tant, no suposa una millora en la reducció dels gens inclosos en el predictor. Aquest comportament pot ser degut a que l'algoritme intenta, de forma iterativa, reduir l'error de la classe amb menor sensibilitat al llarg de Δ, aleshores, si a cada iteració l'entitat de menor sensibilitat sempre és la mateixa, l'algoritme no arriba a canviar el valor de θ_k de la resta.

Encara que aquest algoritme no obté una solució que redueixi la quantitat de gens, s'ha buscat de forma heurística un vector $\boldsymbol{\theta}$ que obtingués un bon compromís entre error i quantitat de gens. La Figura 4.6 mostra la precisió balancejada i les sensibilitats obtingudes a l'utilitzar el vector $\boldsymbol{\theta} = \{\theta_{CLL} = 2.8, \theta_{cMCL} = 2.8, \theta_{HCL} = 1.8, \theta_{FL} = 1, \theta_{nnMCL} = 1, \theta_{HCLv} = 0.8, \theta_{LPL} = 1, \theta_{SDRPL} = 1, \theta_{SMZL} = 0.9\}$. A la figura es pot veure com les entitats SDRPL, LPL i CLL tenen el mateix perfil que en la Figura 4.5. En canvi, la resta d'entitats tenen les corbes superposades, en concret, les sensibilitats sofreixen una forta disminució entre $6.24 < \Delta < 7.68$. En aquest cas, la precisió balancejada augmenta ràpidament a l'augmentar el nombre de gens, obtenint una precisió de gairebé el 70% amb 195 gens, quan el model anterior requeria 2097 per obtenir una precisió similar. Per tant, **utilitzar una quantitat de reducció diferent per cada entitat gairebé no perjudica la precisió a canvi de reduir en gran mesura el nombre de gens**.

La identificació del vector $\boldsymbol{\theta}$ òptim en aquesta tesi és complexa, no solament degut a que s'han d'optimitzar 9 paràmetres simultàniament, sinó també perquè l'òptim depèn del compromís entre la quantitat de gens i la precisió del model. Per exemple, a la Figura 4.6 l'òptim en quan a precisió es troba amb 4865 gens (70.56%), però amb 195 ja s'aconsegueix un valor pràcticament igual (69.33%), una situació preferible per tal

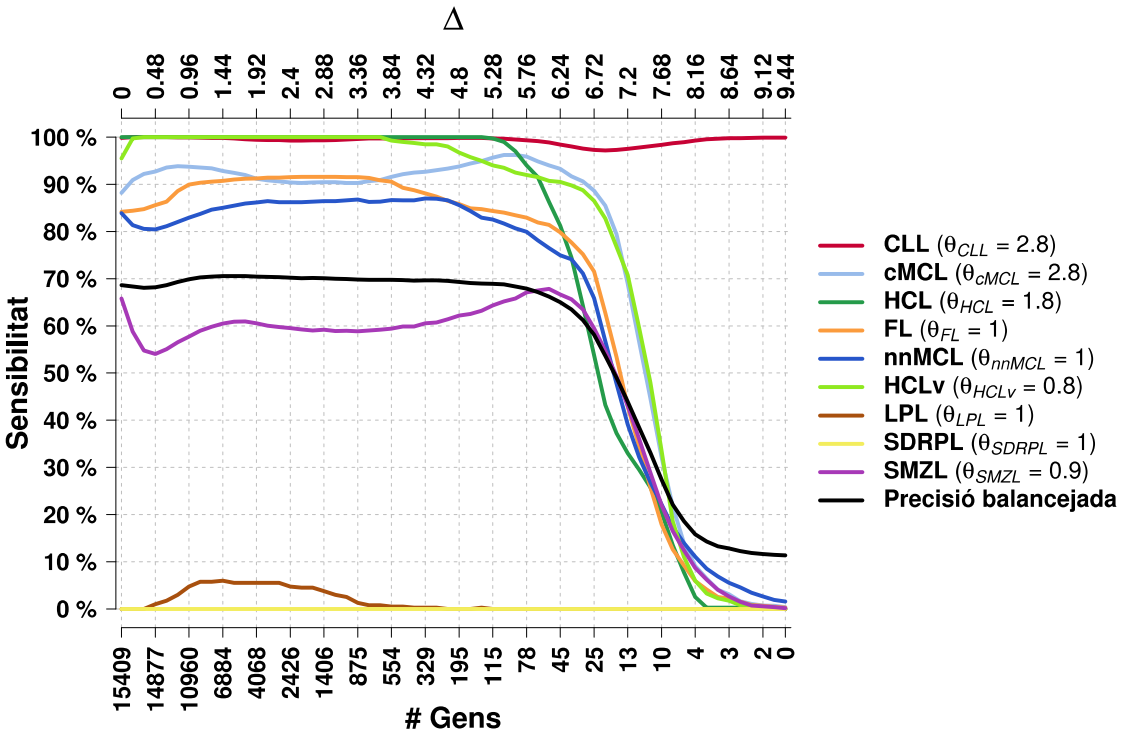

***Figura 4.6: Rendiment del model NSC multiclasse escalat segons* θ**. *Sensibilitat i precisió balancejada de cada classe segons el paràmetre* Δ. *Els centroides de cada classe s'han escalat segons el vector* **θ** indicat a la llegenda. *A l'eix x s'indica el nombre de gens que utilitza el model amb el valor corresponent de* Δ.

d'obtenir un predictor interpretable. Un segon inconvenient és que el procediment d'escalar els components del centroide no facilita el control del nombre de gens que es seleccionen per entitat. Per exemple, quan Δ = 5.28 s'inclouen 115 gens al predictor, on 25 tenen el component diferent de zero en l'entitat nnMCL i 34 en l'entitat HCLv, unes quantitats suficientment elevades com per plantejar-se reduir-les més si la sensibilitat no en surt molt perjudicada.

4.4.3 Enfocament *multi-step*

Amb l'objectiu de simplificar encara més el predictor, s'ha utilitzat un altre enfocament que també permet l'assignació d'un valor de Δ diferent a cada classe (Δ_C). Aquest enfocament consisteix en construir el predictor per passos (*multi-step*), tal com descriu el següent algoritme:

1) Es selecciona *C*, corresponent a l'entitat que es vol discriminar.

2) En el *training set*, es construeix un predictor que discrimini les mostres de l'entitat *C* respecte les mostres de la resta d'entitats agrupades conjuntament. Mitjançant *cross-validation* s'obté el valor de Δ_C òptim.
3) S'actualitza el *training set* excloent les mostres de l'entitat *C*.
4) Es repeteix els passos (1)-(3) fins que el *training set* només contingui mostres d'una entitat.

El fet d'excloure l'entitat en comptes d'utilitzar una metodologia d'un-vs-tots, com en l'apartat 4.3, és degut a la possibilitat que existeixin gens que tenen nivells d'expressió elevats en múltiples classes. A l'agrupar entitats amb nivells d'expressió heterogenis en un mateix grup es reduiria la diferència de mitjanes, provocant que aquests gens no es detectessin. Al fer servir l'estratègia *multi-step*, l'heterogeneïtat es va reduint a mesura que es van eliminat entitats del *training set*, fins que en alguna iteració ja no s'agrupen entitats amb mitjanes diferents. Per altra banda, els inconvenients d'aquesta estratègia són que es perd potència a cada iteració al reduir el nombre de casos i que, respecte l'estratègia multiclasse, no s'aprofita al màxim la informació aportada per aquests gens amb nivells d'expressió heterogenis entre classes.

L'entitat discriminada a cada pas s'ha seleccionat en base a les dades, en concret, a cada pas s'ha discriminat l'entitat que domina la separació. En les Figures 4.1 i 4.5 l'entitat CLL mostra una influència més gran en l'anàlisi, concretament, en la primera figura les mostres de CLL formen la primera gran branca i en la segona és l'entitat que manté sempre una sensibilitat més elevada. Per tant, és raonable que en el primer pas es discrimini l'entitat CLL respecte la resta. L'estratègia utilitzada en la resta de passos s'ha basat en aplicar l'NSC multiclasse a les entitats que encara no s'han discriminat. Concretament, s'ha comparat les prediccions *cross-validated* respecte l'entitat real quan s'utilitza el valor Δ que minimitza la precisió balancejada. Aleshores, l'entitat *C* seleccionada ha sigut la més diferenciada de la resta, és a dir, la que té més sensibilitat i, a la vegada, un menor nombre de mostres de les altres entitats es prediuen com *C*.

Seguint la metodologia descrita, l'ordre de les primeres sis entitats a discriminar ha sigut: CLL, cMCL, HCL, FL, nnMCL i HCLv. La Figura 4.7 mostra la precisió

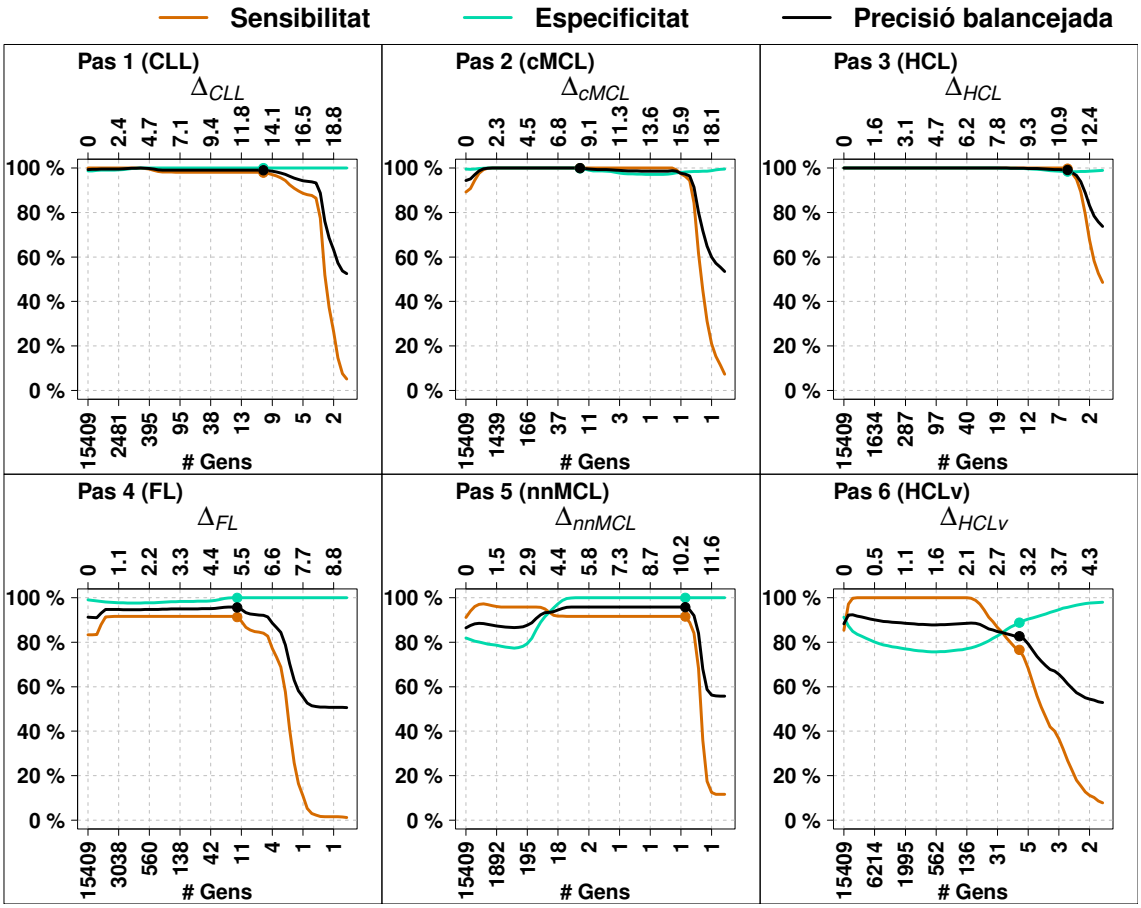

***Figura 4.7: Rendiment dels models NSC en l'enfocament multi-step**. Sensibilitat, especificitat i precisió balancejada en funció de Δ_C en els sis primers passos de l'algoritme. El punt de cada gràfic indica el valor de Δ_C seleccionat per construir el predictor final.*

balancejada (negre), la sensibilitat (verd) i l'especificitat (marró) pels diferents valors de Δ_C del model NSC construït en cadascun d'aquests passos de l'algoritme. En les tres primeres entitats (CLL, cMCL i HCL) es pot identificar valors de Δ_C amb precisions pròximes al 100%, a més, el nombre de gens necessaris per arribar a aquesta precisió és força baix. El quart i cinquè passos (FL i nnMCL) mostren un patró similar, però maximitzant la precisió al 95%, similar a la que s'observava a la Figura 4.5. Notem que per l'entitat nnMCL només és necessari un gen per maximitzar la precisió. En el sisè pas, on es distingeix HCLv de {LPL, SDRPL i SMZL}, la discriminació no és tant gran i les precisions no arriben al 90%.

Un cop diferenciades aquestes sis entitats, s'ha intentat discriminar LPL, SDRPL i SMZL en el setè pas. La Figura 4.8 mostra la precisió balancejada i la sensibilitat del

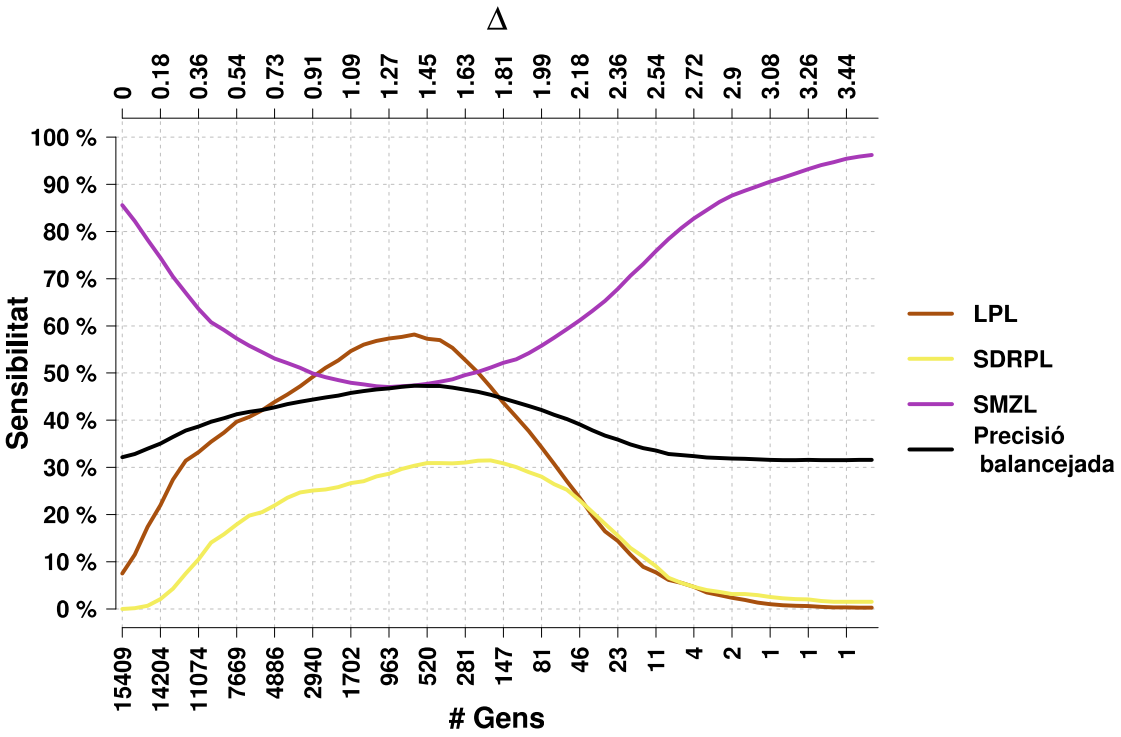

Figura 4.8: Rendiment del model NSC multiclasse en el setè pas. *Sensibilitat i precisió balancejada del model NSC multiclasse segons* Δ. El model inclou *les tres classes {LPL, SDRPL i SMZL} a discriminar al setè pas de l'algoritme multi-step.*

model NSC multiclasse quan es consideren només aquestes tres entitats. En la figura es pot veure com l'augment de la sensibilitat en qualsevol entitat és a expenses de la disminució en una altra, obtenint, com a màxim, una precisió balancejada del 45%. Degut a que **no s'han pogut distingir les entitats LPL, SDRPL i SMZL entre si** de manera fiable, s'han agrupat en un grup anomenat *Miscellaneous*.

L'enfocament *multi-step* permet valorar fàcilment en cada entitat com afecta el nombre de gens a la precisió, on, en general, amb 5 o menys gens ja s'obtenen precisions molt elevades (Figura 4.7). Els valors de Δ_C que s'han utilitzat per construir el model final han sigut valors amb precisions elevades i que no requerissin una gran quantitat de gens, representats amb un punt en la figura. L'única entitat en què s'ha perjudicat la precisió a canvi d'utilitzar menys gens ha sigut HCLv, on la precisió s'hagués pogut millorar un 4% a canvi d'incloure 99 gens més. Tenint en compte que per aquesta entitat només es disposa de 4 mostres, s'ha decidit prioritzar l'opció menys complexa.

La Taula 4.1 resumeix el model NSC de cada pas, on s'indica l'entitat discriminada, el valor del paràmetre Δ_C i diverses mesures corresponents a aquest paràmetre (nombre de gens que sobreviuen la reducció, sensibilitat i especificitat). També s'indica *K*, corresponent al nombre de *folds* del procés de *cross-validation* estratificada, que correspon al mínim entre el nombre de mostres de l'entitat a discriminar i 10.

La Figura 4.9 mostra, per cada pas, els components dels centroides reduïts estandarditzats que sobreviuen a la reducció, on en color s'indica els components que pertanyen a la classe discriminada i en gris els components de la resta de classes agrupades. Aquest gràfic mostra que alguns gens tenen més influència que d'altres a l'hora de predir noves dades. Per exemple, el gen *SOX11* en cMCL és el més rellevant, mentre que el gen *CRIM1* gairebé no afecta. La quantitat de gens en total són 55, una reducció considerable respecte els milers inicials aconseguida mitjançant un algoritme molt simple. Una avantatge addicional d'aquesta metodologia és que la interpretació dels gens també és més simple, donat que a cada pas els gens identificats són específics d'una entitat concreta.

Degut al baix nombre de mostres i la gran quantitat de variables, és recomanable, mentre sigui possible, contrastar la informació obtinguda amb informació publicada en la literatura. Un dels avantatges dels predictors simples és que aquest contrast és més directe. Per exemple, els gens *FMOD* i *LEF1*, utilitzats en el primer pas, s'han relacionat amb l'entitat CLL en altres estudis [120,121]. De la mateixa manera, s'ha descrit el gen CCND1 en nnMCL i cMCL [122], mentre que el gen *SOX11* només s'ha descrit en el segon [123]. També s'ha relacionat *IL1R2* amb HCL [124] i *MME*, també anomenat *CD10*, amb FL [125]. L'estudi de l'entitat HCLv és menys extensa en la literatura, el qual dificulta contrastar la informació, però també s'ha descrit que expressa *CXCR4* a nivells diferents que altres B-CLPD [126]. Aquests contrasts posen de manifest que el predictor no està modelant soroll o efectes de confusió, sinó que realment està fent servir informació de l'expressió gènica per distingir els subtipus de B-CLPD.

Pas	*C* (Entitat B-CLPD)	Δ_C	# Gens	*K*	Sensibilitat (%)	Especificitat (%)
1	CLL	13.45	9	10	97.92	100
2	cMCL	8.41	16	10	100	99.79
3	HCL	11.30	5	4	99.75	98.39
4	FL	5.34	14	10	91.28	99.98
5	nnMCL	10.38	1	10	91.51	100
6	HCLv	3.05	10	4	76.58	88.76
7	LPL-SDRPL-SMZL	3.57	0	4	-	-

Taula 4.1: Resum del predictor NSC construït a cada pas. *Per cada pas de l'algoritme multi-step s'indica l'entitat discriminada, la reducció Δ_C, el nombre de gens que s'utilitzen, el nombre de folds (K) del procés de cross-validation, la sensibilitat i l'especificitat.*

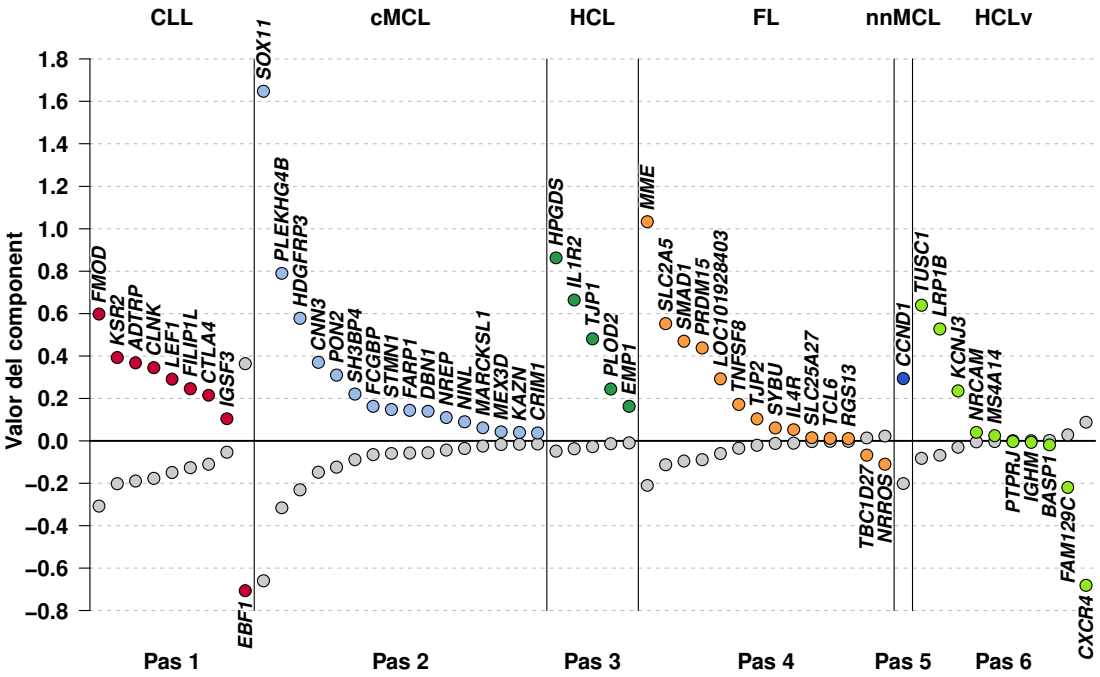

Figura 4.9: Centroides reduïts estandarditzats de l'NSC multi-step. *Components dels centroides estandarditzats reduïts que sobreviuen la reducció Δ_C a cada pas del model multi-step. En color s'ha representat els components corresponents a l'entitat discriminada, mentre que en gris s'ha representat els components corresponents a la resta d'entitats agrupades del pas.*

4.4.4 Predicció dels B-CLPD, NOS

La predicció d'una nova mostra *Z* segons el model *multi-step* es pot fer utilitzant les probabilitats predites dels sis models NSC. Definim $p_s(Z=C_s)$ com la probabilitat que la mostra *Z* pertanyi a la classe discriminada al pas *s* (C_s). Aleshores, les probabilitats de que *Z* pertanyi a cada una de les 7 classes {CLL, cMCL, HCL, FL, nnMCL, HCLv o

Miscellaneous} es poden calcular segons

$$\begin{aligned}
&\hat{\boldsymbol{p}}(\boldsymbol{Z}=\mathbf{CLL}) \;=\; \hat{p}_1(Z=\mathrm{CLL}),\\
&\hat{\boldsymbol{p}}(\boldsymbol{Z}=\mathbf{cMCL}) \;=\; (1-\hat{p}_1(Z=\mathrm{CLL}))\cdot\hat{p}_2(Z=\mathrm{cMCL}),\\
&\hat{\boldsymbol{p}}(\boldsymbol{Z}=\mathbf{HCL}) \;=\; (1-\hat{p}_1(Z=\mathrm{CLL}))\cdot(1-\hat{p}_2(Z=\mathrm{cMCL}))\cdot\hat{p}_3(Z=\mathrm{HCL}),\\
&\ldots\\
&\hat{\boldsymbol{p}}(\boldsymbol{Z}=\boldsymbol{Miscellaneous}) \;=\; (1-\hat{p}_1(Z=\mathrm{CLL}))\cdot(1-\hat{p}_2(Z=\mathrm{cMCL}))\cdot(1-\hat{p}_3(Z=\mathrm{HCL}))\\
&\qquad\qquad\cdot(1-\hat{p}_4(Z=\mathrm{FL}))\cdot(1-\hat{p}_5(Z=\mathrm{nnMCL}))\cdot(1-\hat{p}_6(Z=\mathrm{HCLv})),
\end{aligned}$$

on la mostra *Z* s'assignaria a la classe amb una probabilitat més alta.

El model *multi-step* proposat assumeix que el perfil d'expressió de *Z* només s'assembla al d'una de les entitats, però, a la pràctica, aquest requisit podria no complir-se en un petit percentatge dels pacients. Per exemple, si *Z* tingués un perfil d'expressió similar a CLL i també a HCL, el predictor assignaria una probabilitat alta a CLL i baixa a HCL, donat que aquesta segona depèn de les probabilitats dels dos primers passos. Aleshores, convé comprovar que el perfil d'expressió dels 55 gens en nous casos només s'assembla a una de les entitats. Atès que s'espera que en la majoria dels casos les probabilitats només siguin altes en una de les entitats, es pot simplificar la predicció de l'entitat amb el següent algoritme:

1) S'inicia $s = 1$.
2) Es calcula $p_s(Z = C_s)$.
3) Si $p_s(Z = C_s) \geq 0.5$, la mostra *Z* s'assigna a la classe C_s.
4) Si $p_s(Z = C_s) < 0.5$, $s = s + 1$.
5) Si $s = 7$, la mostra *Z* s'assigna a la classe *Miscellaneous* (LPL, SDRPL o SMZL).
6) Si la mostra no s'ha assignat a cap entitat, es torna a (2).

Un dels objectius de la construcció d'aquest predictor és classificar els 30 B-CLPD, NOS amb informació de *microarrays* d'expressió. Recordem que els B-CLPD, NOS són pacients en els que no se'ls hi ha pogut diagnosticar una entitat mitjançant altres criteris, aleshores, la informació de l'expressió gènica pot servir per afegir una nova capa d'informació a tenir en compte per tal d'obtenir un diagnòstic final. Un cop aplicat el predictor, 7 casos s'han predit com a CLL, 1 com cMCL, 2 com HCL i 20 com a *Miscellaneous*. En tots els casos amb probabilitats superiors al 90% excepte en dos, en

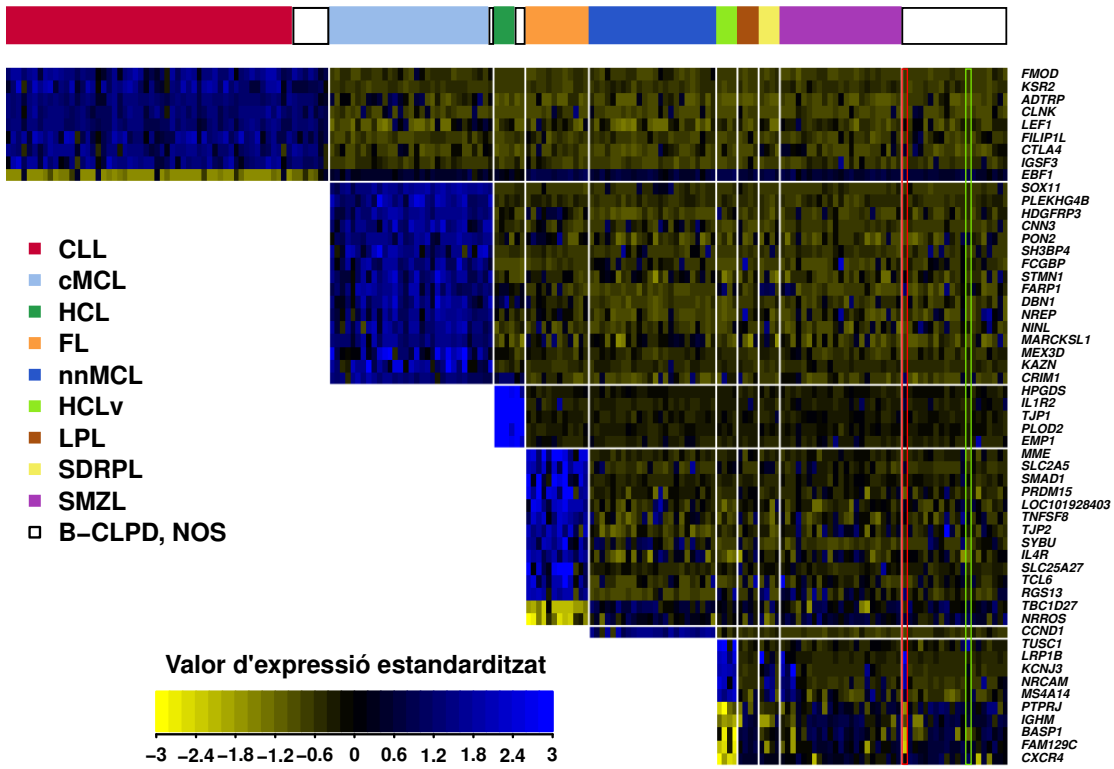

Figura 4.10: Heatmap dels 55 gens inclosos en el predictor multi-step. *Les mostres B-CLPD, NOS, indicades en blanc, s'han representat juntament amb l'entitat en la que s'han predit. El gràfic s'ha construït segons el model multi-step. En vermell i verd s'han senyalat dues mostres en les quals la predicció està confosa entre HCLv i Miscellaneous.*

els quals s'ha observat certa confusió entre HCLv i *Miscellaneous*.

La Figura 4.10 mostra el *heatmap* de l'expressió estandarditzada dels 55 gens en les 189 mostres (159 de el *training set* i 30 B-CLPD, NOS), el qual s'ha construït segons el model *multi-step*. Per exemple, en les mostres CLL no s'han representat els gens de cMCL, ja que la primera es discrimina abans que la segona i, per tant, les seves mostres no són rellevants per a la selecció dels gens del segon pas. Les 30 mostres B-CLPD, NOS s'han representat juntament amb les mostres de l'entitat en la que es prediuen, i les dues mostres amb certa confusió s'han marcat amb vermell i verd.

En la figura es pot veure uns perfils d'expressió molt clars en cada entitat, així com en els B-CLPD, NOS. En el sisè pas, la mostra marcada en vermell ha obtingut p_6(HCLv) = 42% i p_6(*Miscellaneous*) = 58%, degut a que expressa els gens *CXCR4* i *TUSC1* a nivells diferents que les mostres d'HCLv, però expressant la resta de gens específics

d'HCLv a nivells similars. La mostra marcada en verd ha obtingut p_6(HCLv) = 20% i p_6(*Miscellaneous*) = 80%, en aquest cas degut a diferències en l'expressió del gen *TUSC1*. En general, la figura mostra que, **a través del perfil d'expressió, es pot predir una entitat específica en els B-CLPD, NOS** o, almenys, delimitar l'entitat a la qual pertany, una informació útil per considerar simultàniament amb la resta de trets morfològics, fenotípics i moleculars en el diagnòstic.

4.4.5 Rendiment del predictor

El predictor construït no solament es pot utilitzar per predir les mostres B-CLPD, NOS, sinó que també pot facilitar el diagnòstic dels pacients evitant l'estudi de les altres capes d'informació. Per aquest motiu, és important avaluar com de precís és, donat que un model que s'espera que s'equivoqui en, per exemple, un 20% de les prediccions, no seria útil a nivell clínic tot i contenir informació rellevant.

En aquesta tesi no es disposa d'una sèrie de validació en *microarrays* per tal d'estimar l'error de predicció del predictor. La *cross-validation* és una alternativa per estimar-lo, però, degut a la quantitat de mostres d'algunes entitats, no es pot aplicar de manera fiable. S'ha de tenir en compte que s'hauria d'utilitzar un procés de *cross-valdation* intern i extern (apartat 3.5), on l'intern serviria per optimitzar Δ_C a cada iteració de l'extern. Al només disposar de quatre mostres en algunes entitats, els càlculs en cada *fold* de la *cross-valdation* interna s'haurien de fer amb només dues mostres.

Tot i aquests inconvenients, sí que es disposa d'una sèrie de validació pel predictor de qPCR. El predictor de *microarrays* i el de qPCR estan molt correlacionats per tres motius: *i*) l'origen en ambdós casos és l'expressió gènica, *ii*) els gens candidats a incloure's al predictor de qPCR són gens que s'han detectat com a rellevants en les dades de *microarrays*, i *iii*) els 44 casos inclosos en el *training set* de qPCR també ho estan en el *training set* de *microarrays*. Aleshores, es pot obtenir una aproximació de l'error esperat en *microarrays* a través de l'error estimat en la sèrie de validació de qPCR.

4.5 Selecció de gens a mesurar mitjançant qPCR

Els resultats presentats a l'apartat 4.4 posen de manifest que l'expressió gènica conté molta informació rellevant per poder distingir la majoria d'entitats de B-CLPD, amb l'excepció de les tres categoritzades com a *Miscellaneous* (LPL, SDRPL i SMZL). Addicionalment, les precisions representades en la Figura 4.7 mostren que limitar el nombre de gens a 5 o menys per entitat no tindria un gran impacte en aquestes. Per exemple, amb 4 gens la precisió balancejada del pas 4, on es discrimina l'entitat FL, és del 90%, un 5% menys que la precisió obtinguda amb els 14 gens utilitzats en el predictor de *microarrays*. Un patró similar es pot observar en la resta d'entitats. Aquest fet suggereix que és possible construir un predictor més simple basat en dades de qPCR, una tècnica que, a canvi de limitar el nombre de gens a mesurar de manera simultània, facilitaria la implementació d'una eina diagnòstica a nivell clínic.

A l'apartat 1.4.4 s'ha explicat el procés habitual per construir un predictor simple utilitzable a nivell clínic. El primer pas per construir-lo és seleccionar un subconjunt de gens a mesurar mitjançant qPCR. A diferència dels *microarrays*, què en un únic experiment mesuren un espectre molt ampli de gens, la qPCR generalment ho pot fer en un nombre limitat. Aleshores, la selecció d'aquest subconjunt és molt important, donat que si no és apropiada i es traslladen gens subòptims el predictor final en sortirà perjudicat.

Les dades dels *microarrays* d'expressió serveixen per estudiar i valorar quins gens són els millors candidats a traslladar-se. La metodologia habitual per seleccionar aquests gens és ordenant-los segons un cert criteri o *score*, per després traslladar els r primers. Per exemple, en Reis et al. [31], després de fer una preselecció de gens, els ordenen segons el coeficient obtingut d'aplicar el mètode lasso, i seleccionen els 4 amb el coeficient més gran. En Watanabe et al. [30] fan l'ordre segons el P-valor obtingut de comparar pacients que responen al tractament respecte els que no responen, on seleccionen els 18 gens més significatius.

En aquesta tesi, el primer *score* candidat a utilitzar-se per ordenar els gens és el centroide reduït estandarditzat presentat a la Figura 4.9. Donat que 55 gens és una quantitat molt elevada per traslladar a qPCR, es podria seleccionar els $r < 55$ amb valors més allunyats de zero. Utilitzar aquest *score* té diversos inconvenients:

- Els gens s'avaluen de manera univariant. Quan es disminueix el paràmetre Δ del mètode NSC s'inclouen més gens al predictor, i l'ordre en què s'inclouen depèn d'un estadístic *T* que compara l'expressió del gen entre dos grups. Aleshores, dos gens molt correlacionats entre si s'afegirien al model pràcticament al mateix moment, augmentant el risc de traslladar molta informació redundant si es seleccionen els *r* primers. Aquest aspecte gairebé no perjudica el predictor en *microarrays*, donat que al disposar de tota la informació només cal disminuir el paràmetre Δ per acabar introduint els gens que aporten informació complementària. En canvi, quan es trasllada un nombre determinat de gens a qPCR, el màxim d'informació que es tindrà disponible en aquesta plataforma serà limitat. Aleshores, cada gen redundant que s'inclou redueix aquest màxim. Per exemple, el màxim d'informació disponible al traslladar tres gens molt correlacionats és el mateix que al traslladar-ne només un dels tres.
- La selecció és fa en base a un estadístic *T*, equivalent a ordenar els gens en base al *P*-valor de la comparació. Aquest ordre no té en compte la grandària de la diferència, la qual podria ser rellevant al caviar de plataforma [32]. En general, davant de *P*-valors similars és preferible escollir gens on les diferencies en l'expressió siguin majors.
- Els 55 gens inclosos en el predictor de *microarrays* estan poc balancejats entre entitats. Per exemple, el predictor inclou 16 gens per discriminar l'entitat cMCL, mentre que només n'inclou un per discriminar nnMCL. Tenint en compte que el predictor final en qPCR s'hauria de construir, com a màxim, amb 2 o 3 gens per entitat, és convenient que se'n traslladin entre 4 i 6 per cadascuna. El fet de traslladar-ne més dels necessaris és degut a que el rendiment d'algun gen en qPCR podria ser molt menor [34], per tant, convé tenir alternatives a l'hora de construir el predictor final.

En el procés de selecció de gens s'ha tingut en compte aquests inconvenients. Per adreçar-ho s'ha calculat, per cada gen, els següents tres paràmetres:

- ***Fold-Change* (FC, rellevància biològica)**: és la mesura més utilitzada en l'entorn de la genètica per mesurar la grandària de la diferència entre dues condicions. Hi ha diferents maneres de calcular-ho, s'ha decidit utilitzar $\log_2(FC) = \bar{x}_A - \bar{x}_B$, on A i B indiquen dos grups diferents. És a dir, l'FC seria la potència en base 2 de la diferència de mitjanes entre dues condicions.
- ***P*-valor (rellevància estadística)**: el FC per si sol no és una bona mesura del poder discriminant d'un gen, el qual es quantifica millor mitjançant un *P*-valor [127]. Els *P*-valors s'han calculat amb limma (apartat 3.4.3) i, degut al problema dels tests múltiples (apartat 1.4.1), s'han ajustat segons el mètode de Benjamini-Hochberg (apartat 3.7.1).
- ***Scores* de Dziuda (rellevància multivariant)**: el mètode de Dziuda (apartat 3.4.4) serveix per identificar gens que aporten informació complementària a d'altres. Els dos *scores* del mètode (S^{ALL} i S^{INF}) tenen en compte el poder univariant i multivariant del gen.

Tot i que en la selecció dels gens s'han tingut en compte tots tres paràmetres, s'ha donat més pes a l'FC i el *P*-valor. La lògica per donar més pes a aquests dos paràmetres univariants és simple: quan la diferència entre grups no és gran, hi ha més risc de que l'estructura recollida per la qPCR sigui diferent que l'estructura recollida en els microarrays [32,34]. És a dir, un gen que no té un gran poder discriminant en *microarrays* és més probable que al traslladar-lo a qPCR deixi d'aportar informació. A més a més, que un gen mostri poder multivariant però no univariant podria ser degut a l'*overfitting*, donat que s'espera que els gens rellevants per una certa entitat sempre mostrin un cert canvi en l'expressió. Recordem que les estimacions d'estructures complexes són poc robustes quan el nombre de mostres és molt limitat.

El càlculs d'aquests paràmetres s'han fet seguint el mateix procediment *multi-step* que en l'apartat 4.4.3. Tant limma com el mètode de Dziuda es podrien utilitzar en un entorn multiclasse, però sofririen el mateix problema que l'NSC i dificultaria la selecció

balancejada de gens. Per exemple, es podria utilitzar limma per obtenir un *P*-valor global de les diferències en l'expressió entre les nou classes, però aquests *P*-valors estarien molt condicionats per l'entitat CLL, la qual té més mostres i més diferències globals (Figura 4.1). El mètode de Dziuda es veuria afectat pels mateixos motius, atès que la metodologia *stepwise hybrid feature selection with* T^2 prioritzaria els gens de CLL abans de començar a seleccionar els gens rellevats de les altres entitats. Així doncs, mitjançant l'ús del procediment *multi-step* es pot controlar de manera més eficient la quantitat de gens específics de cada entitat que es traslladen a la qPCR.

El percentatge de *probesets* filtrades segons l'IQR ha sigut del 0% per limma i del 50% pel mètode de Dziuda. Recordem que no filtrar *probesets* en base a l'IQR quan s'aplica limma és perquè obté resultats menys acurats quan es realitza [64]. Per altra banda, el mètode de Dziuda té una càrrega computacional molt elevada i fer un filtratge més estricte alleugereix el temps de computació. En els següents dos apartats es presenten els resultats de limma i del mètode de Dziuda, respectivament.

4.5.1 Expressió diferencial segons limma

Per tal de quantificar si globalment hi ha moltes diferències en l'expressió gènica, a la Figura 4.11 s'ha representat, per als sis primers passos, els $\log_2$(FC) respecte els *P*-valors dels 20546 gens inclosos en l'anàlisi (*volcano plot*). Els punts marcats en gris senyalen els gens que no tenen diferències significatives entre els dos grups, on el criteri de significació estadística utilitzat ha sigut obtenir un *P*-valor ajustat menor a 0.05.

A cada gràfic de la figura s'observa que hi ha almenys un gen situat a l'extrem superior dret o esquerre, localització que indica molta rellevància estadística (*P*-valor) i biològica (FC). També es pot veure que al primer pas s'obtenen una gran quantitat de gens significatius al comparar CLL contra la resta d'entitats, concretament, hi ha 8972 significatius. Pas a pas es va reduint la quantitat fins que, al sisè pas, només s'obtenen 66 gens amb expressió diferencial entre HCLv i els *Miscellaneous*. Aquest comportament és esperable per dos motius: *i*) s'ha construït l'aproximació *multi-step* de

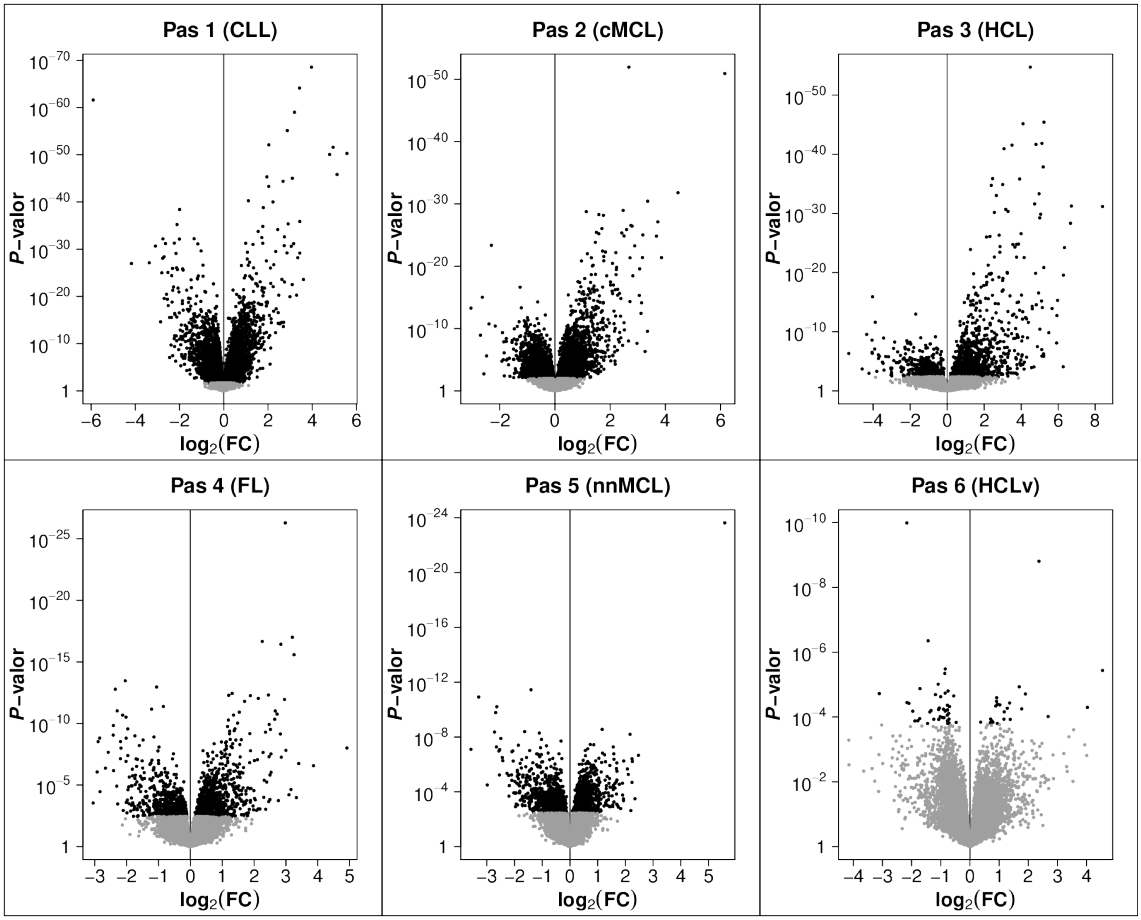

Figura 4.11: Volcano plot dels sis primers passos del limma multi-step. *En cada gràfic hi representat el log_2(FC) respecte el P-valor del corresponent pas. Els punts marcats en negre corresponen als gens que compleixen P-valor ajustat < 0.05.*

manera que es comença discriminant les entitats més fàcils de distingir, i *ii*) a cada pas es va reduint el nombre de mostres en la comparació. Una tercera observació és que en els tres primers passos hi ha una tendència a obtenir log_2(FC) positius, indicant que els gens específics de les entitats CLL, cMCL i HCL tendeixen a estar sobreexpressats en aquestes entitats. Finalment, el gen *CCND1* en nnMCL té molta més expressió diferencial que la resta de gens, resultat que concorda amb l'observat al pas 5 del predictor en *microarrays* d'expressió (Figura 4.7), on feia falta una reducció molt gran (Δ_{nnMCL}) per a que entrés més d'un gen al predictor.

En l'apartat 4.4.3 s'ha conclòs que les entitats LPL, SDRPL i SMZL no es poden distingir de manera fiable. Tot i així, s'ha decidit fer dues comparacions addicionals entre aquestes entitats mitjançant limma. En concret, en la primera s'ha comparat les

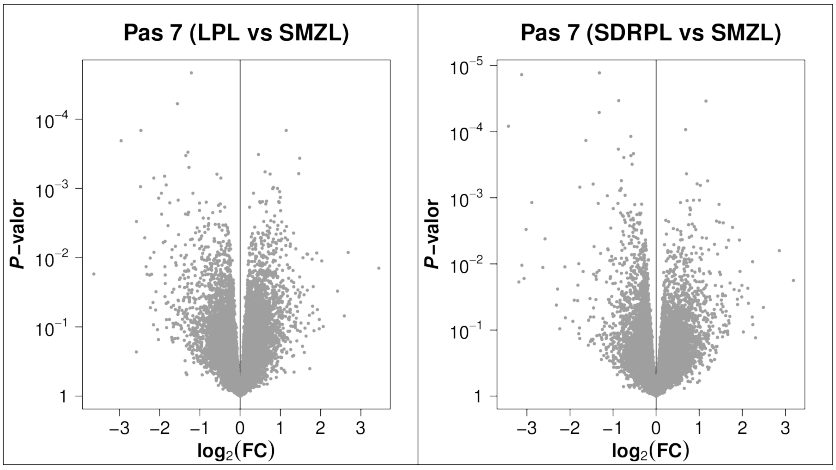

***Figura 4.12: Volcano plot al setè pas del limma multi-step**. En cada gràfic hi ha representat el log_2(FC) respecte el P-valor en les comparacions LPL vs SMZL i SDRPL vs SMZL.*

mostres de l'entitat LPL contra les mostres de l'entitat SMZL i en la segona les de SDRPL contra les de SMZL. La comparació de les entitats LPL i SDRPL no s'ha realitzat ja que només es disposa de quatre mostres en cadascuna d'aquestes entitats. La Figura 4.12 mostra el *volcano plot* d'aquestes dues comparacions, on cap gen ha obtingut un *P*-valor ajustat menor a 0.05.

D'aquesta anàlisi s'han extret dos dels paràmetres a considerar en la selecció dels gens: l'estimació de l'FC i el *P*-valor ajustat. En concret, els gens més interessants són aquells que mostren un FC elevat amb una clara significació estadística.

4.5.2 *Scores* segons el mètode de Dziuda

El tercer criteri a tenir en compte per avaluar quins gens traslladar a qPCR correspon als dos *scores* obtinguts segons el mètode de Dziuda (S^{ALL} i S^{INF}). En aquest anàlisi també s'ha utilitzar l'enfocament *multi-step*, incloent les comparacions LPL vs SMZL i SDRPL vs SMZL. A diferència de limma, què és un mètode senzill i directe d'aplicar, el mètode de Dziuda requereix que es fixin diversos paràmetres a priori. Per consistència s'ha utilitzat el mateix valor d'aquests a tots els passos de l'enfocament.

La següent llista resumeix que representa cada paràmetre i quin valor se l'hi ha assignat:

- *M*: el nombre de biomarcadors que es generen de manera iterativa. Se n'han generat 300. Tenint en compte que l'objectiu és seleccionar gens, és assumible que els que tenen més poder discriminant estan inclosos en algun dels 300.
- *m*: el nombre de gens inclosos en cadascun dels *M* biomarcadors. S'ha fixat a 3. La quantitat de mostres d'algunes entitats no permet estudiar estructures gaire complexes, i fer-ho provocaria *overfitting*. Aleshores, s'ha considerat que 3 és suficientment petit com per no amplificar el problema de l'*overfitting* i, a la vegada, permet estudiar estructures multivariants simples.
- T_{cut}: els biomarcadors amb un valor de l'estadístic $T^2 < T_{cut}$ es considera que no contenen informació. S'ha assignat a 2.5. Aquest paràmetre és el que més impacte té en definir el subconjunt dels *Informative Set of Genes*, tot i així, la seva elecció no és tant crítica en aquesta tesi. Mentre no sigui massa elevat no afectarà gaire a l'ordre dels *scores* finals (S^{ALL} i S^{INF}). S'ha utilitzat 2.5 ja que ha mostrat un comportament adequat en tots els passos.
- *B*: el nombre de Monte Carlo *sets* que es creen mitjançant el *modified bagging schema* per cada conjunt de dades (l'*original set* i l'*INF set*). Se n'han creat 1000. Aquests *sets* serveixen per calcular el percentatge de vegades que un gen concret s'ha seleccionat, per tant, 1000 és un nombre suficientment gran com per poder calcular percentatges de manera fiable.
- γ_{OOB}: el percentatge de mostres de cada classe que no s'inclouen als Monte Carlo *sets*, arrodonit a l'alça. S'ha fixat $\gamma_{OOB} = 20\%$. S'ha utilitzat el mateix valor que en Dziuda [95]. Mentre no sigui un valor gaire elevat no tindrà un efecte molt gran als *scores* finals.

Identificació de l'*Informative Set of Genes*

El paràmetre T_{cut} té un gran impacte en l'*Informative Set of Genes*, per tant, és important comprovar que el valor utilitzat sigui adequat en les vuit comparacions. Un valor massa elevat deixaria gens informatius fora de l'*Informative Set of Genes*, per altra banda, un valor massa petit inclouria al conjunt molts gens sense informació. Per avaluar la selecció de T_{cut}, a la Figura 4.13 s'ha representat el valor de l'estadístic T^2 al llarg dels

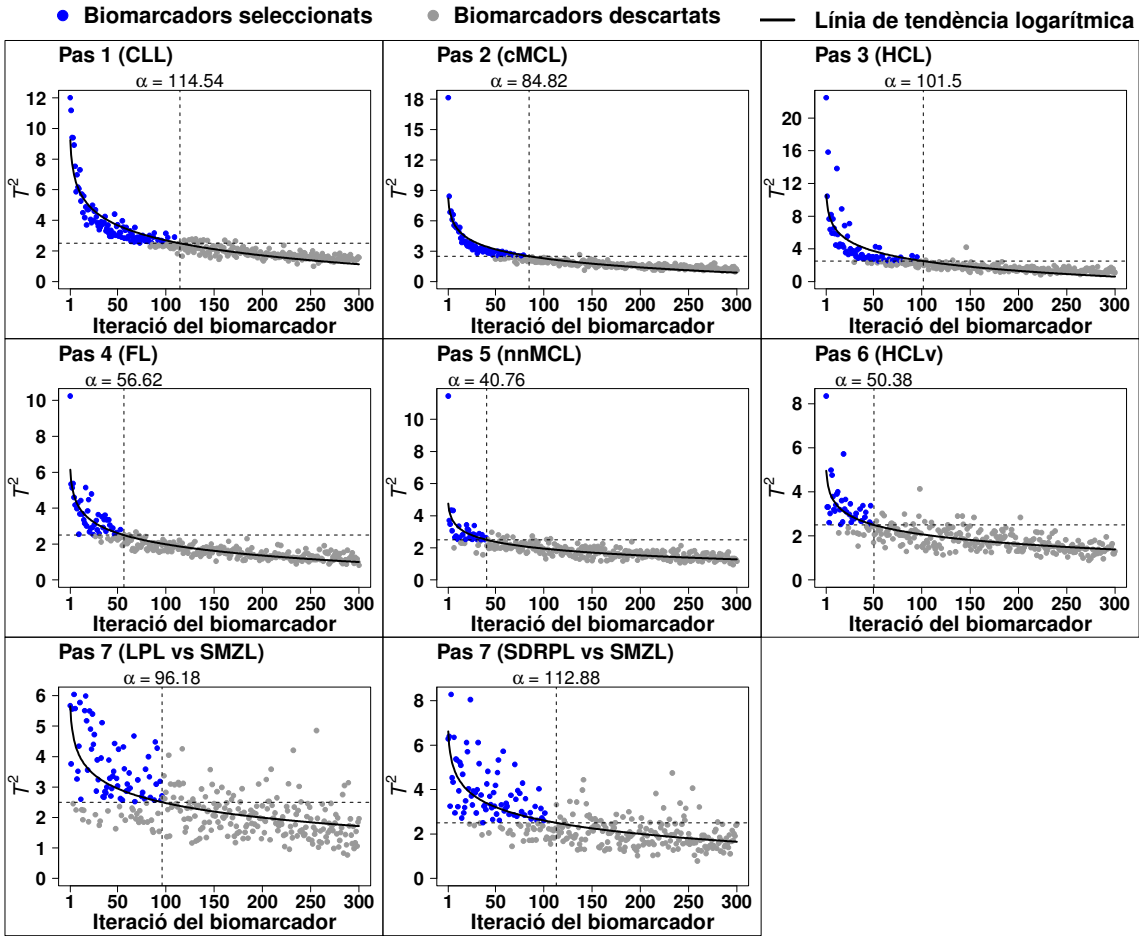

***Figura 4.13: Estadístic T^2 dels M biomarcadors de cada pas**. En cada gràfic hi ha representat, pels diferents passos del procés multi-step, l'estadístic T^2 dels M biomarcadors respecte la iteració en la que s'han identificat. Els gens inclosos en els biomarcadors marcats en blau són els que formen l'Informative Set of Genes del pas corresponent. La línia negra estima el decreixement logarítmic de l'estadístic al llarg de les iteracions.*

300 biomarcadors construïts a cada pas. En cada gràfic també s'ha afegit la línia de tendència logarítmica que recull el decreixement de l'estadístic al llarg de les iteracions, construïda de la mateixa manera que la de la Figura 3.9 de l'apartat 3.4.4. A part de per avaluar la selecció de T_{cut}, aquest gràfic també serveix per estudiar les capacitats discriminants globals.

En la figura es pot veure que a cada pas que s'avança els punts són més variables al voltant de la línia de tendència logarítmica. Aquest fenomen és indicatiu de la quantitat d'informació disponible per discriminar les classes. Un patró poc variable significa que a cada iteració queda menys informació rellevant en el conjunt de dades, en canvi, un

patró molt variable significa que els gens eliminats a cada iteració tenen poca informació rellevant. Els dos gràfics corresponents al setè pas, on s'analitzen les entitats *Miscellaneous*, s'observa un patró especialment variable, indicant que hi ha poca informació rellevant. Aquest resultat coincideix amb les anàlisis presentades en els altres apartats d'aquest capítol.

Les línies de tendència logarítmica dels tres primers gràfics de la figura no s'ajusten perfectament al decreixement, mentre que en la resta de gràfics no s'observa biaix en cap punt de la corba. Aquesta desviació entre els valors observats i la línia ajustada podria provocar que gens amb informació no s'incloguessin a l'*Informative Set of Genes*, tot i així, el valor de T_{cut} utilitzat captura suficients biomarcadors com per a que els gens amb més poder predictiu estiguin inclosos en el subconjunt. Per últim, en els passos 2, 4 i 5 s'observa un biomarcador amb molta més capacitat discriminant que els restants, un resultat esperat en el pas 5 (*CCND1* en nnMCL), però no en els passos 2 i 4. Globalment sembla que el valor de T_{cut} és adequat en tots els passos, donat que es seleccionen els biomarcadors amb més informació i, a la vegada, no s'inclouen molts d'innecessaris.

Un cop comprovat que el paràmetre T_{cut} s'ha definit correctament, s'ha inclòs en l'*Informative Set of Genes* de cada pas els gens corresponents als biomarcadors marcats en blau de la Figura 4.13. Aquests biomarcadors són els que compleixen les següents dues condicions: *i*) el seu estadístic $T^2 < T_{cut}$ i *ii*) la iteració en què s'ha detectat és anterior a la iteració en què la línia logarítmica prediu un valor igual a T_{cut}.

Càlcul dels *scores*

Per últim, s'ha creat B = 1000 Monte Carlo *sets* provinents de l'*original set* (tots els gens) i B = 1000 provinents de l'INF *set* (gens de l'*Informative Set of Genes*). En cadascun d'aquests *sets* s'ha aplicant l'*stepwise hybrid feature selection with* T^2 amb m = 3 variables. D'aquí s'ha obtingut l'*score* S^{INF} del gen g com el percentatge de vegades que aquest gen s'ha seleccionat en els Monte Carlo *sets* provinents de l'INF *set*. De la mateixa manera s'ha calculat l'*score* S^{ALL}, però utilitzant els provinents de l'*original set*.

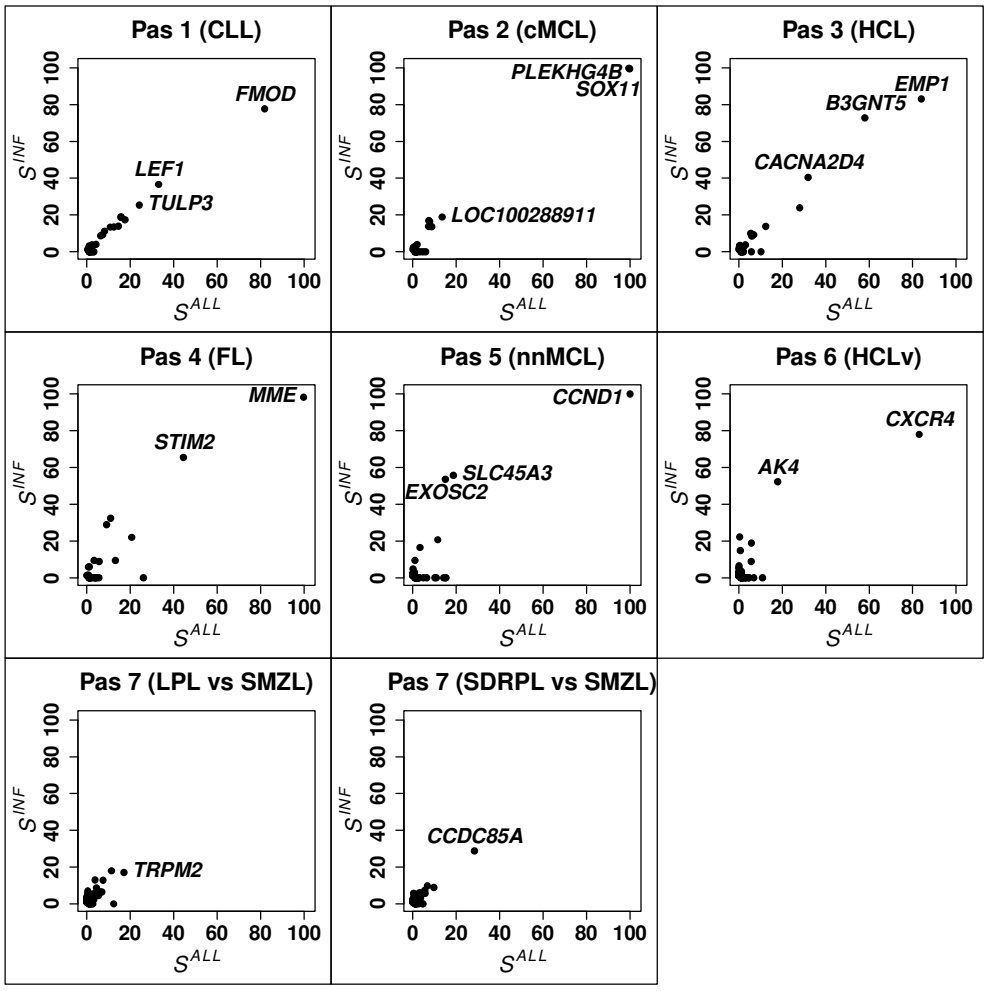

***Figura 4.14: Scores segons el mètode Dziuda**. Per cada pas del procediment multi-step s'ha representat l'score S^{ALL} respecte l'S^{INF} de cada gen. S'han exclòs els que han obtingut valors dels dos scores per sota d'1.*

La Figura 4.14 mostra, per cada pas, l'*score* S^{ALL} respecte l'*score* S^{INF} dels gens en què almenys un dels dos *scores* és superior a 1. Els gens amb més poder predictiu multivariant són els que queden situats a l'extrem superior dret del gràfic. Per exemple, al primer pas els gens seleccionats més vegades en els Monte Carlo *sets* han sigut *FMOD*, *LEF1* i *TULP3*, on el gen *FMOD* obté aproximadament un valor del 80% en els dos *scores*. En les dues comparacions del setè pas no es veu cap gen que domini la discriminació de classes, donat que cap arriba al 30% en els *scores*. S'ha considerat que els gens que compleixen $S^{INF} > 1\%$ i $S^{ALL} > 1\%$ tenen poder predictiu multivariant segons el mètode de Dziuda i, per tant, són candidats a traslladar-se. La Taula 4.2 resumeix, per cada pas, el nombre de gens inclosos en l'*Informative Set of Genes* i el nombre de gens candidats.

Subconjunt	Pas 1	Pas 2	Pas 3	Pas 4	Pas 5	Pas 6	Pas 7	Pas 7
	CLL	cMCL	HCL	FL	nnMCL	HCLv	LPL vs SMZL	SDRPL vs SMZL
Informative Set of Genes	350	268	318	252	316	803	1319	1334
Candidats (S^{INF}>1%, S^{ALL}>1%)	31	32	28	30	48	66	94	107

Taula 4.2: Quantitat de gens candidats segons el mètode de Dziuda. *Nombre de gens inclosos en el subconjunt de l'Informative Set of Genes i nombre de gens que es consideren amb poder predictiu segons el mètode de Dziuda.*

Notem que aquest mètode fa servir dos *scores* quan tindria l'opció de només fer servir l'S^{INF}, el qual es calcula utilitzant el subconjunt de gens que es defineix com el que té la informació per a discriminar les classes i, per tant, sembla més adequat per identificar els gens més rellevants. S'ha de tenir en compte que el procediment per calcular els *scores* és susceptible a l'*overfitting* per dos motius: *i*) estima estructures multivariants i *ii*) es força que a cada biomarcador s'incloguin $m = 3$ gens, quan un o dos podrien ser suficients. Aleshores, quan el procediment s'aplica a un subconjunt reduït com l'*Informative Set of Genes*, hi ha un risc alt que algun gen s'acabi seleccionant diverses vegades sense aportar informació. Aquest fet és més difícil que passi quan s'aplica el procediment a l'*original set*, on, al disposar de molts més gens, és més improbable que el mateix es seleccioni múltiples vegades per culpa de l'*overfitting*. Per altra banda, quan s'utilitzen totes els gens és més probable que a cada iteració se n'acabi seleccionant algun sense poder predictiu degut a l'*overfitting*. És a dir, l'*score* S^{ALL} té més risc a repartir els *scores* a gens sense informació.

4.5.3 Combinació d'informacions per a la selecció de gens

Amb els tres paràmetres calculats (FC, *P*-valor univariant i *scores* de Dziuda) ja es pot estudiar quins gens seleccionar per traslladar a qPCR. A part d'aquests paràmetres obtinguts de les dades, també és convenient tenir en compte altres criteris biològics i tècnics per a la selecció de gens, com són:

- La funció biològica del gen. Un gen associat al cicle cel·lular té més rellevància biològica que un gen associat, per exemple, al gust. També hi ha gens que no estan anotats i no es coneix la seva funció. Aquests gens s'anoten amb un nom que comença per *LOC* seguit per un número, com el que es veu al pas 2 de la

Figura 4.14 (*LOC100288911*). Quan es disposa de múltiples gens candidats a traslladar, és preferible seleccionar els que tenen funcions conegudes i rellevants.

- Si la seqüencia del gen està ben representada en la plataforma de *microarray*. Per exemple, les *probesets* acabades en *_x_at* contenen *probes* que mesuren una seqüència que és igual o molt similar a una seqüència no relacionada amb la d'interès. Aquesta similitud entre les dues seqüències no-relacionades pot influenciar al senyal de la *probeset* d'interès, provocant que l'expressió del gen no sigui fiable.
- L'adequació de les sondes prefabricades de qPCR per mesurar el gen d'interès. Les sondes comercials de les diferents plataformes de qPCR podrien no mesurar la mateixa regió del gen que el *microarray*, o fins i tot podria no existir una sonda pel gen. Una alternativa en aquests casos és dissenyar una sonda pròpia, però si en les dades de *microarrays* s'han identificat múltiples gens amb poder discriminant similar, seria més senzill limitar la selecció a aquells que estan ben representats en sondes comercials.

A la Taula 4.3 s'indica els 35 gens finals seleccionats per traslladar a la qPCR, a més de: la *probeset* del *microarray* associada al gen, l'entitat de B-CLPD que discrimina el gen, el pas del procediment *multi-step* en què s'ha detectat, el $\log_2$(FC), l'estadístic *T* moderat de limma, el *P*-valor ajustat segons el mètode de Benjamini-Hochberg i els dos *scores* obtinguts del mètode de Dziuda.

A la taula es pot veure que els $\log_2$(FC) són elevats en general, on 28 gens tenen valors per damunt d'1.5 (o per sota de -1.5), indicant una forta rellevància biològica. Els estadístics *T* també mostren molta rellevància estadística, amb *P*-valors ajustats per sota de 0.05 en 31 dels gens. Per altra banda, els *scores* de Dziuda indiquen poder discriminant multivariant (S^{INF} > 1% i S^{ALL} > 1%) en 21 gens, un nombre lleugerament menor degut a que s'han prioritzat els paràmetres univariants davant dels *scores* multivariants. Per exemple, a la Figura 4.14 es pot veure com el gen *AK4* té un elevat poder discriminant multivariant al sisè pas, però aquest gen té $\log_2$(FC) = 0.11, un valor molt baix quan hi ha altres gens alternatius amb valors superiors a 2. Per altra banda, el

Probeset	Gen	Entitat	Pas	log_2(FC)	Estadístic T (limma)	*P*-valor ajustat	S^{INF}	S^{ALL}
202709_at	*FMOD*	CLL	1	3.97	31.02	0.00	77.7	81.9
230551_at	*KSR2*	CLL	1	3.43	28.76	0.00	13.9	14.6
227646_at	*EBF1*	CLL	1	-5.91	-27.55	0.00	19.0	15.8
210191_s_at	*PHTF1*	CLL	1	2.04	23.14	0.00	4.0	4.3
221558_s_at	*LEF1*	CLL	1	4.79	22.32	0.00	36.6	33.1
209583_s_at	*CD200*	CLL	1	3.30	10.91	0.00	0.0	0.0
230441_at	*PLEKHG4B*	cMCL	2	2.68	29.44	0.00	99.5	99.8
204914_s_at	*SOX11*	cMCL	2	6.15	28.86	0.00	99.8	99.5
209524_at	*HDGFRP3*	cMCL	2	4.46	17.35	0.00	2.0	0.7
223627_at	*MEX3B*	cMCL	2	1.59	15.53	0.00	0.0	0.0
218412_s_at	*GTF2IRD1*	cMCL	2	1.76	15.48	0.00	0.6	0.2
200953_s_at	*CCND2*	cMCL	2	-0.88	-3.23	0.01	0.0	0.0
201324_at	*EMP1*	HCL	3	6.72	20.15	0.00	83.1	84.0
205403_at	*IL1R2*	HCL	3	8.40	20.12	0.00	0.0	0.0
201798_s_at	*MYOF*	HCL	3	6.34	15.44	0.00	10.0	5.5
205508_at	*SCN1B*	HCL	3	4.02	10.80	0.00	0.1	0.0
224499_s_at	*AICDA*	HCL	3	5.14	9.68	0.00	0.1	0.2
201012_at	*ANXA1*	HCL	3	4.88	6.18	0.00	0.1	0.0
203435_s_at	*MME*	FL	4	2.98	17.18	0.00	98.2	99.8
204430_s_at	*SLC2A5*	FL	4	3.20	11.45	0.00	9.4	13.2
230777_s_at	*PRDM15*	FL	4	2.84	11.12	0.00	22.0	20.7
227798_at	*SMAD1*	FL	4	3.26	10.65	0.00	0.7	0.6
206105_at	*AFF2*	FL	4	1.31	8.88	0.00	8.9	5.7
208712_at	*CCND1*	nnMCL	5	5.59	17.02	0.00	100.0	100.0
208072_s_at	*DGKD*	nnMCL	5	-1.03	-6.56	0.00	0.2	0.0
228696_at	*SLC45A3*	nnMCL	5	0.55	4.72	0.00	55.7	18.7
211919_s_at	*CXCR4*	HCLv	6	-2.16	-9.01	0.00	78.0	83.0
202190_at	*CSTF1*	HCLv	6	-1.44	-6.14	0.00	0.6	0.0
219643_at	*LRP1B*	HCLv	6	4.55	5.53	0.01	8.9	5.7
212765_at	*CAMSAP2*	HCLv	6	1.69	5.11	0.02	0.6	0.0
229510_at	*MS4A14*	HCLv	6	2.68	4.43	0.03	0.0	0.0
205708_s_at	*TRPM2*	LPL	7 (LPL/SMZL)	0.81	3.64	0.73	17.1	17.1
235228_at	*CCDC85A*	SDRPL	7 (SDRPL/SMZL)	3.12	5.34	0.10	31.3	28.4
207853_s_at	*SNCB*	SDRPL	7 (SDRPL/SMZL)	0.87	4.84	0.16	7.5	5.5
221933_at	*NLGN4X*	SDRPL	7 (SDRPL/SMZL)	3.43	4.67	0.18	7.2	5.7

Taula 4.3. Llista dels 35 gens seleccionats per traslladar a qPCR. *A la taula s'inclou la probeset del microarray juntament amb el gen que representa, l'entitat de B-CLPD que identifica, el pas del algoritme multi-step en què s'ha detectat i els paràmetres utilitzats (FC, P-valor, S^{INF} i S^{ALL}) en el procés de selecció dels gens.*

gen *LOC100288911* identificat en el segon pas s'ha descartat degut a no estar ben caracteritzat. Els gens *ANXA1*, *AICDA*, *CCND2* i *CD200* no s'han inclòs en el conjunt a traslladar degut a les seves capacitats predictives, sinó perquè en altres estudis s'han descrit com rellevants per distingir algunes entitats [128–131].

També s'han seleccionat quatre gens relacionats amb les entitats *Miscellaneous* (LPL, SDRPL i SMZL), els quals mostren, en els seus respectius paràmetres, menys indicis de les seves capacitats discriminants que la resta de gens. El motiu és que aquestes tres

entitats estan força confoses per criteris morfològics, clínics, fenotípics i genètics [132], aleshores, una font d'informació addicional seria molt important per ajudar a distingir-les. S'han seleccionat els gens que mostraven en algun dels paràmetres una lleugera capacitat predictiva.

4.6 Predictor basat en dades de qPCR

Els 35 gens seleccionats a l'apartat 4.5 són els que s'han traslladat a la plataforma *qPCR Fluidigm BioMark 48.48 Dynamic Array*, la qual s'ha utilitzat per mesurar l'expressió gènica en 44 dels 159 pacients de la cohort *training*. Aquests 44 pacients (8 CLL, 4 FL, 2 HCL, 3 HCLv, 6 cMCL, 6 nnMCL, 3 LPL, 2 SDRPL i 10 SMZL) formen el *training set* de qPCR i, per tant, s'han utilitzat per construir el predictor basat en aquest tipus de dades. Idealment els dos *training sets* (*microarrays* i qPCR) haurien d'estar formats per dues cohorts de pacients independents, però la baixa freqüència d'algunes entitats de B-CLPD i les restriccions de material no permeten obtenir suficients pacients com per separar-los. La reducció de 159 a 44 pacients també és deguda a les restriccions de material. Els 63 nous pacients en els que se'ls hi ha pogut mesurar l'expressió s'han inclòs a la cohort de validació, de manera que s'ha maximitzat la quantitat de casos en aquesta per a que l'estimació de l'error de predicció fos la més precisa possible.

Les mesures de l'expressió s'han preprocessat segons la metodologia descrita a l'apartat 3.3.4. Un cop preprocessades, s'ha de decidir com tractar les mesures en què la corba d'amplificació no supera el llindar FT en els 40 cicles de l'experiment (apartat 1.4.5). El problema amb aquestes mesures és que no se'ls hi pot assignar un valor numèric. Els dos principals motius per a què una corba no superi el llindar són: *i*) un problema tècnic que ha provocat que la mostra no amplifiqués correctament, o *ii*) l'expressió del gen en la mostra analitzada és extremadament baixa. En el primer cas la mesura es pot tractar com a *missing*, a més, un problema tècnic es pot detectar comparant múltiples corbes d'amplificació. En el segon cas el *missing* és informatiu, donat que precisament indica un valor d'expressió baix. En aquesta tesi, un cop s'han descartat problemes tècnics, s'ha assignat a aquestes mesures un valor de zero ($N = 2^{-\Delta\Delta CT} = 0$).

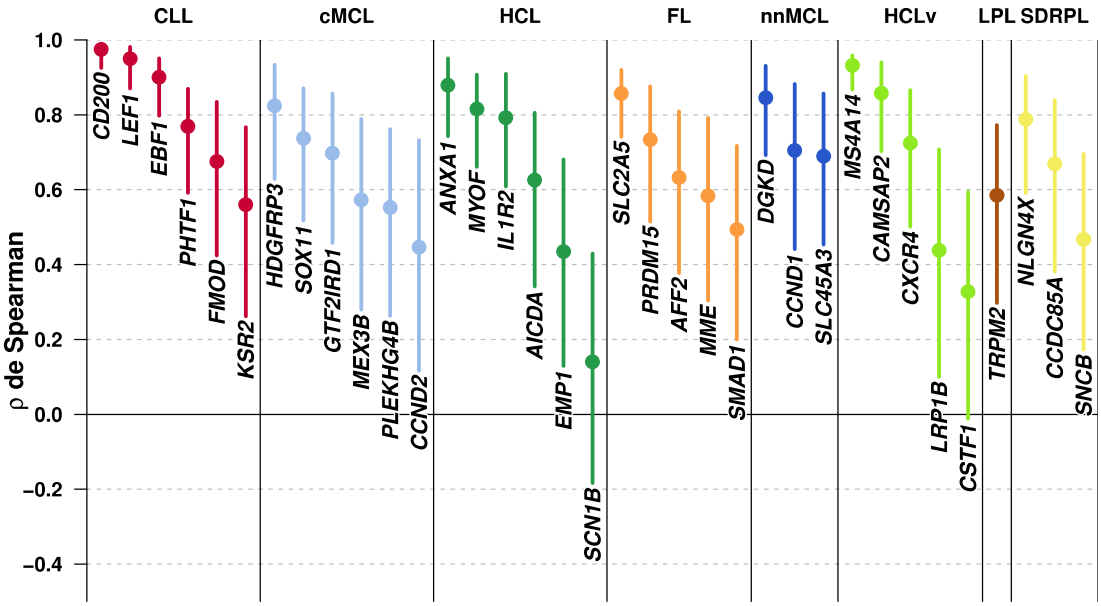

Figura 4.15: Correlació entre qPCR i microarrays dels 35 gens. *Correlació entre l'expressió mesurada amb qPCR i mesurada amb microarrays en els 35 gens seleccionats. El mètode utilitzat per estimar-la ha sigut la ρ d'Spearman. Els punts representen l'estimació puntual mentre que els segments representen l'interval de confiança del 95% aproximat per bootstrap.*

4.6.1 Correlació microarray-qPCR

La primera anàlisi realitzada amb aquestes dades ha sigut estudiar la correlació entre les dades de *microarrays* i les dades de qPCR en els 44 pacients amb informació de les dues plataformes. El principal objectiu d'aquesta anàlisi és verificar que les dues plataformes estan mesurant el mateix. Una forta correlació descarta que els resultats identificats en els *microarrays* siguin causats per la variabilitat tècnica. A la Figura 4.15 hi ha representada la correlació d'Spearman dels 35 gens entre les dues plataformes, juntament amb el seu interval de confiança obtingut a través d'una aproximació *bootstrap*.

A la figura es pot veure una **elevada correlació entre les dues plataformes**. En general, les correlacions gen a gen són positives i superiors a 0.4, tot i que hi ha certa variabilitat entre gens, on la més baixa és de 0.14 i la més alta de 0.97. Aquest resultat indica que el predictor de *microarrays* i el predictor que es construirà de qPCR estaran forçosament correlacionats, un resultat que alleugera l'inconvenient de no disposar de

sèrie de validació en els *microarrays*. Si el predictor en qPCR obtingués precisions altes en aquesta sèrie independent es podria concloure que el predictor de *microarrays* també ho faria.

4.6.2 Construcció del predictor

Per construir el predictor basat en dades de qPCR s'ha combinat l'enfocament *multi-step* amb una metodologia més simple que la utilitzada en dades de *microarrays*. En concret, s'ha basat el predictor en punts de tall (*cutoffs*) identificats mitjançant l'ús de corbes ROC (de l'anglès r*eceiver operating characteristic*). Hi ha tres motius per aquesta simplificació:

1) El predictor no pot construir-se amb els 35 gens mesurats, donat que una eina diagnòstica utilitzable a nivell clínic ha d'incloure el mínim de gens possibles.
2) El nombre de mostres del *training set* no permet utilitzar metodologies gaire complexes, donat que les estimacions dels paràmetres serien poc robustes.
3) A la pràctica, predictors senzills d'aplicar són més fàcils d'implementar a nivell clínic [133].

Recordem que l'enfocament *multi-step* discrimina les entitats en el següent ordre: 1-CLL, 2-cMCL, 3-HCL, 4-FL, 5-nnMCL, 6-HCLv, 7.1-{LPL vs SMZL}, 7.2-{SDRPL vs SMZL}. Definim C_s com l'entitat discriminada al pas $s = \{1, \ldots, 7.2\}$. Aleshores, el següent algoritme és el que s'ha utilitzat a cada pas s:

1) S'ha eliminat les mostres del *training set* que corresponen a entitats discriminades en passos previs a s.
2) Per cada gen rellevant per discriminar C_s, s'ha construït la corba ROC resultant de discriminar C_s de la resta d'entitats segons diferents llindars de l'expressió del gen. Els llindars que s'han utilitzat han sigut els punts mitjos entre els valors d'expressió del gen ordenats. Per exemple, si els valors d'expressió ordenats fossin {1, 5, 6, 7}, aleshores els llindars candidats serien {3, 5.5, 6.5}.
3) Per cada gen, s'ha fixat el *cutoff* al llindar de (2) que maximitza la suma de la sensibilitat i l'especificitat.

4) Un cop identificats els *cutoffs*, s'han utilitzat diversos criteris per decidir quins gens utilitzar en el predictor final. En general, si només un gen ha mostrat una separació perfecta o molt elevada en el *cutoff* seleccionat, aleshores s'ha utilitzat aquest gen en el pas corresponent del predictor. Si més d'un gen ha mostrat separacions elevades o perfectes, aleshores s'ha utilitzat altres criteris, com per exemple: poder discriminant en les dades de *microarrays*, nivells d'expressió i variabilitat.

A la Figura 4.16 s'ha representat l'expressió dels sis gens associats a l'entitat CLL, juntament amb el *cutoff* de cada un obtingut de discriminar aquesta entitat de la resta. La línia negra continua mostra el *cutoff* (*C*), on també s'indica si el gen està sobreexpressat (+) o infraexpressat (-) en les mostres de CLL. A la figura es pot veure que la majoria dels gens tenen una capacitat discriminant molt forta, on les 8 mostres de CLL estan perfectament separades en 4 (*FMOD*, *KSR2*, *LEF1* i *PHTF1*) i gairebé en 2 (*CD200* i *EBF1*). Atès que en aquest pas del predictor s'ha identificat diversos gens amb fortes capacitats predictives, s'ha decidit utilitzar-ne dos, *FMOD* i *KSR2*. Les noves mostres que n'expressin un o els dos per damunt del respectiu *cutoff* es prediran com a CLL.

Els motius per prioritzar *FMOD* i *KSR2* han sigut: *i*) separen perfectament les mostres de CLL, *ii*) les mostres de CLL tenen nivells d'expressió absoluts més elevats que en els altres gens, i *iii*) mostren molt poder discriminant en els paràmetres estimats en les dades de *microarrays* de la Taula 4.3. Addicionalment, la correlació observada en les mostres de CLL entre *FMOD* i *KSR2* no és molt elevada en qPCR (ρ = 0.40) ni en *microarrays* (ρ = 0.43), indicant poca redundància entre els dos gens.

En les entitats cMCL, HCL, FL i nnMCL (passos 2 al 5) no s'han identificat múltiples gens amb molt poder predictiu, així que, per cada una, només s'ha inclòs al predictor el gen que més n'ha mostrat. En concret, els gens inclosos han sigut *SOX11*, *MYOF*, *MME* i *CCND1*, respectivament. La Figura 4.17 mostra l'expressió d'aquests gens en les 44 mostres del *training set*, on s'ha indicat el *cutoff* del gen (línia negra) i s'ha representat en gris les mostres corresponents a entitats ja discriminades en l'enfocament *multi-step*. A la figura es pot observar que el gen *CCND1* l'expressen les entitats cMCL i nnMCL,

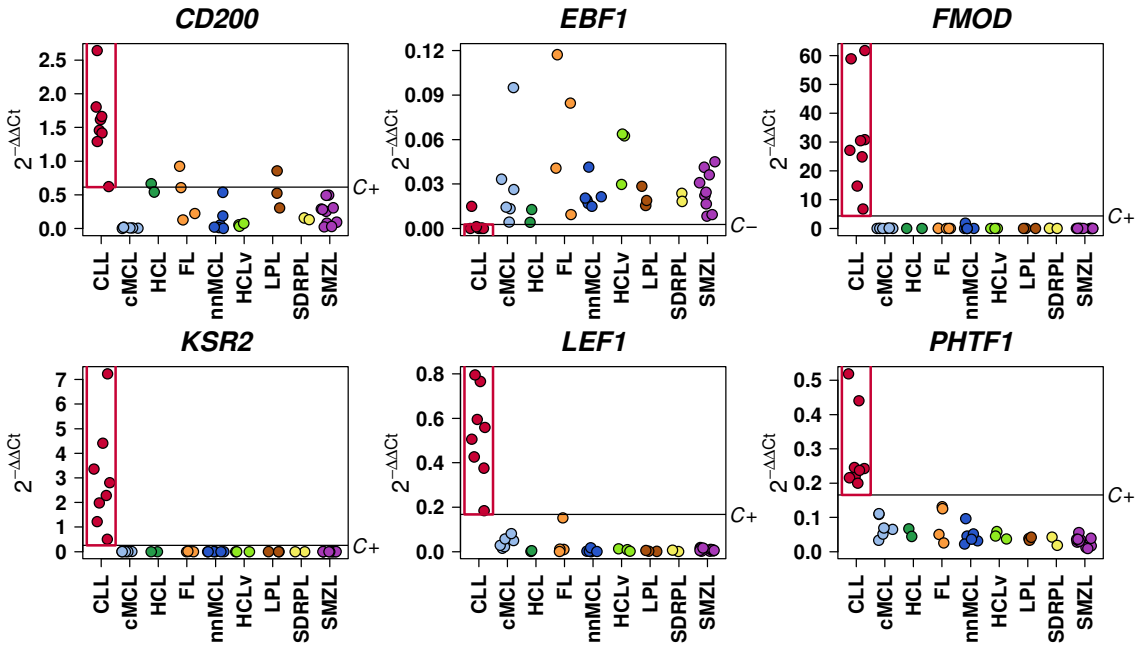

***Figura 4.16: Expressió dels gens de CLL en el training set de qPCR**. La línia negra indica el cutoff i si el gen està sobreexpressat (C+) o infraexpressat (C-) en CLL.*

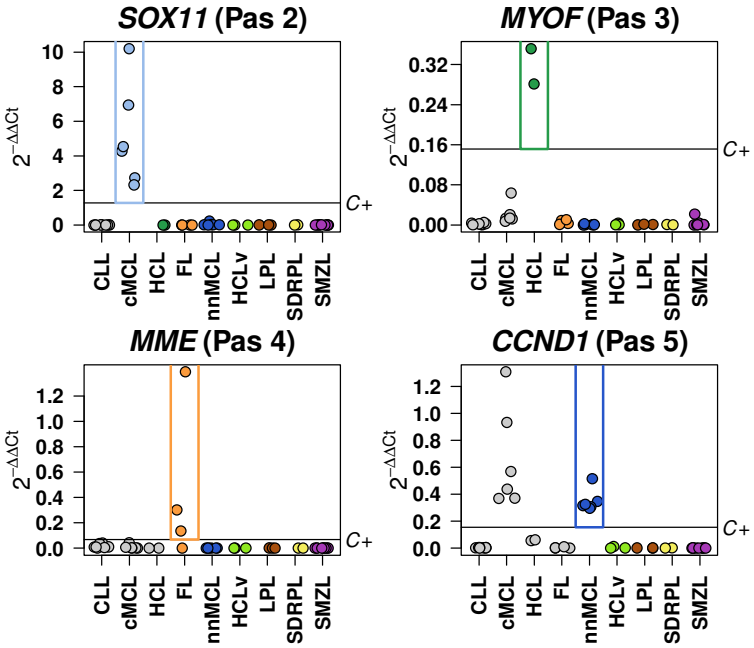

***Figura 4.17: Expressió en qPCR dels gens corresponents als passos 2 al 5**. Les mostres representades en gris corresponen a mostres d'entitats ja discriminades en passos previs. La línia negra indica el cutoff i si el gen està sobreexpressat (C+) o infraexpressat (C-) en l'entitat que discrimina.*

però no hi ha confusió entre les dues degut a que la primera s'identifica a través del gen *SOX11*.

Notem que una mostra de l'entitat FL té un valor de 0 en el gen *MME*. Aquesta mostra no ha superat el llindar FT en quatre dels cinc gens associats a aquesta entitat, entre ells *MME*, així que se'ls li ha assignat un valor de 0. Aquest comportament podria ser degut a un problema tècnic que no s'ha detectat mitjançant l'estudi de les corbes d'amplificació, però s'ha decidit no excloure-la de les anàlisis per dos motius: *i*) en les dades de *microarrays* també mostrava uns nivells d'expressió relativament baixos, i *ii*) en estudis en què s'ha analitzat mostres de ganglis de pacients d'FL, s'ha descrit que la majoria expressa *MME*/*CD10* però no tots [134].

En el sisè pas, on es discrimina l'entitat HCLv del grup *Miscellaneous*, dos gens han separat perfectament les mostres dels dos grups, *CAMSAP2* i *CXCR4*. L'expressió d'aquests gens en el *training set* està representada en els dos primers gràfics de la Figura 4.18. S'ha decidit incloure els dos al predictor, però, a diferència del primer pas, on el criteri per predir una nova mostra com a CLL era que expressés un o l'altre per damunt del *cutoff*, aquí s'ha considerat que la nova mostra ha de sobrepassar el *cutoff* en els dos gens per a que s'assigni a l'entitat HCLv. Hi ha dos motius per aquest canvi, el primer és que, tal com es pot veure als gràfics, mostres d'entitats discriminades a passos anteriors mostren nivells d'expressió similars a les mostres d'HCLv. El segon és que aquesta entitat és poc freqüent en la població, aleshores, exigir que les expressions dels dos gens han de traspassar el *cutoff* penalitza la probabilitat que una nova mostra es predigui en aquesta entitat. Addicionalment, la correlació observada entre els dos gens és baixa en qPCR (ρ = -0.19) i en *microarrays* (ρ = -0.10).

En el setè i últim pas s'ha identificat un gen prometedor, *TRMP2*, el qual està sobreexpressat en les mostres de l'entitat LPL i en gairebé cap mostra de la resta d'entitats, tal com es pot veure al gràfic superior dret de la Figura 4.18. En l'entitat SDRPL, per la qual s'han inclòs 3 gens dels 35 mesurats mitjançant qPCR, no s'ha identificat cap amb fortes capacitats predictives. En els tres gràfics inferiors de la Figura 4.18 es mostra l'expressió d'aquests gens, on només *NLGN4X* té una expressió elevada

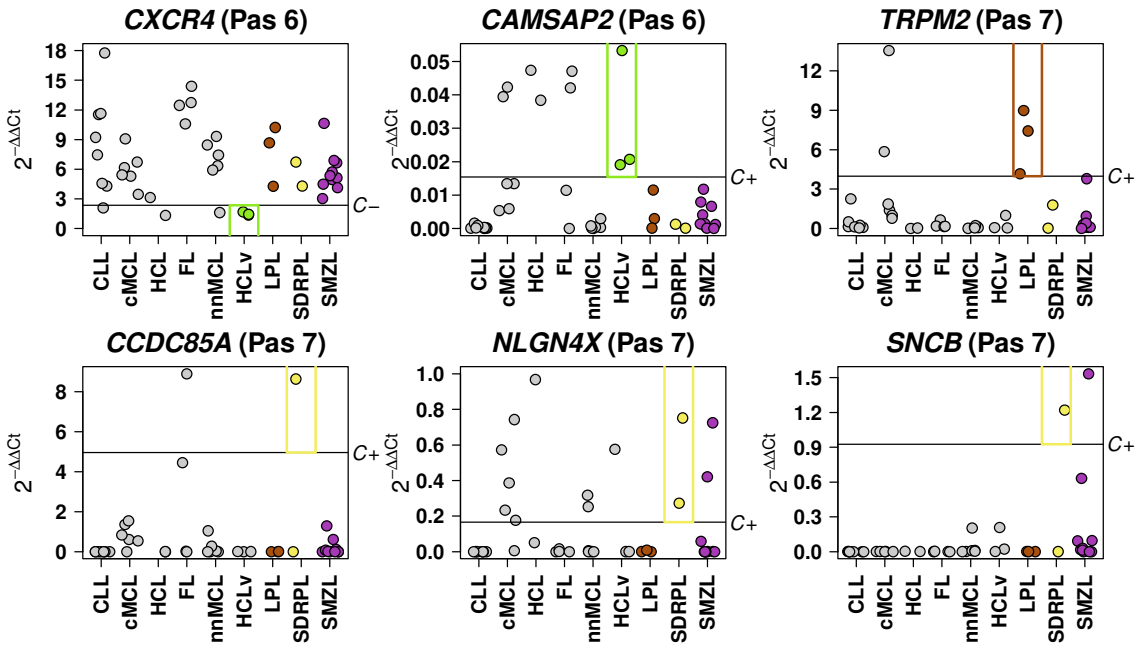

***Figura 4.18: Expressió en qPCR dels gens associats a HCLv, LPL i SDRPL**. La línia negra indica el cutoff i si el gen està sobreexpressat (C+) o infraexpressat (C-) en l'entitat que discrimina. Mostres discriminades en passos anteriors de l'enfocament multi-step s'han representat en gris.*

en les dues mostres de l'entitat SDRPL. Tot i així, s'ha decidit no incloure'l al predictor final per diversos motius: *i*) només es disposa de dues mostres de SDRPL, *ii*) dues de les 10 mostres de SMZL expressen el gen per sobre del *cutoff*, i *iii*) múltiples mostres d'entitats discriminades a passos previs també l'expressen per damunt del *cutoff*.

Recapitulant, **el predictor simple basat en qPCR s'ha construït amb un o dos gens per entitat**. En concret, inclou els següents 9 gens: *FMOD* (CLL), *KSR2* (CLL), *SOX11* (cMCL), *MYOF* (HCL), *MME* (FL), *CCND1* (nnMCL), *CXCR4* (HCLv), *CAMSAP2* (HCLv) i *TRMP2* (LPL). La predicció de noves mostres es fa mitjançant l'enfocament *multi-step*, on només es té en compte el *cutoff* dels gens respectius de cada pas. En concret, a cada pas es comprova si la nova mostra expressa el gen corresponent al pas per damunt del *cutoff*, si ho fa, s'assigna a la classe discriminada en aquell pas, si no, es mira el gen del següent pas. L'única excepció és que en el primer pas només cal que expressi *FMOD* o *KSR2* per a assignar-la a CLL, i en el sisè ha de infraexpressar *CXCR4* i sobreexpressar *CAMSAP2* per assignar-la a HCLv.

Pas	Entitat discriminada	*FMOD* (>4.33)		*KSR2* (>.26)	*SOX11* (>1.27)	*MYOF* (>.15)	*MME* (>.07)	*CCND1* (>.15)	*CXCR4* (<2.35)		*CAMSAP2* (>.015)
1	CLL	(+	o	+)							
2	cMCL	-		-	+						
3	HCL	-		-	-	+					
4	FL	-		-	-	-	+				
5	nnMCL	-		-	-	-	-	+			
6	HCLv	-		-	-	-	-	-	(-	i	+)
7	*Miscellaneous* {LPL,SDRPL,SMZL}	-		-	-	-	-	-	+		-

Taula 4.4: Resum del predictor en qPCR. *Per cada pas s'indica l'entitat B-CLPD que es discrimina i el perfil d'expressió que ha de complir per a que una mostra es predigui com a tal. + indica que el valor d'expressió del gen ha d'estar per damunt del cutoff, - per sota. En el primer pas sobrepassar el cutoff en FMOD o KSR2 ja el classifica com a CLL, en el sisè ha de sobrepassar en CXCR4 i CAMSAP2.*

4.6.3 Rendiment del predictor i predicció dels B-CLPD, NOS

Un cop construït el predictor de set passos s'ha aplicat a la sèrie de validació. Aquesta sèrie inclou 14 CLL, 13 cMCL, 10 FL, 16 nnMCL, 2 LPL i 8 SMZL. Malauradament no s'ha pogut disposar de mostres de les entitats HCL, HCLv i SDRPL, degut a que en aquestes entitats la incidència de casos amb afectació a la sang perifèrica és molt baixa. En aquesta sèrie el predictor ha assignat totes les mostres a l'entitat correcta, excepte en l'últim pas en què es discrimina LPL de {SDRPL, SMZL}, on ha assignat una de les vuit mostres d'SMZL a LPL i la resta (2 LPL i 7 SMZL) al grup {SDRPL, SMZL}. Per tant, **el predictor de qPCR és capaç de discriminar correctament les entitats CLL, cMCL, FL i nnMCL, i no és capaç de discriminar les entitats SDRPL, LPL i SMZL entre sí**. Aquest resultat concorda amb el de l'apartat 4.4, on el predictor en *microarrays* mostrava unes capacitats predictives similars. La manca de mostres d'HCL i d'HCLv no permet valorar si el predictor és capaç d'identificar-les correctament, però és rellevant ressaltar que cap mostra de la sèrie de validació s'ha classificat en aquestes dues entitats, indicant que els respectius *cutoffs* tenen una alta especificitat.

La Taula 4.4 resumeix el predictor final, on s'ha eliminat el setè pas ja que no ha mostrat capacitat predictives en la sèrie de validació.

La validació positiva del predictor en varies entitats significa que es pot aplicar a les 34

mostres B-CLPD, NOS amb informació sobre expressió gènica mesurada mitjançant qPCR. Aquestes prediccions serveixen per afegir una nova capa d'informació a valorar en el diagnòstic definitiu d'aquests pacients. De les 34 mostres, 14 s'han predit com a CLL, 1 com a FL, 3 com a HCLv, 1 com a nnMCL i 15 com a *Miscellaneous*.

4.7 Combinació de diferents capes d'informació per al diagnòstic

En aquesta tesi s'han estudiat 64 mostres de pacients en què no es va poder obtenir un diagnòstic de cap entitat utilitzant els criteris habituals, motiu pel qual se'ls va categoritzar com a B-CLPD, NOS. En 30 d'aquestes mostres se'ls ha mesurat l'expressió gènica mitjançant *microarrays* d'expressió, mentre que en les 34 restants s'ha fet mitjançant qPCR. El principal objectiu dels predictors construïts en els apartats 4.4 i 4.6 és identificar a quina de les nou entitats s'aproxima més el perfil d'expressió d'aquestes mostres. La Figura 4.19 resumeix l'enfocament *multi-step* utilitzat per construir aquests predictors, el nombre de mostres de cada entitat disponibles en els *training sets*, el nombre de mostres en la sèrie de validació de qPCR, i els gens utilitzats a cada pas del predictor.

De les 64 mostres B-CLPD, NOS, 21 han mostrat un perfil d'expressió pròxim a l'entitat CLL, 1 a cMCL, 2 a HCL, 1 a FL, 1 a nnMCL, 3 a HCLv i 35 a cap de les anteriors i que, per tant, concorden amb les entitats agrupades com a *Miscellaneous*. Les 29 mostres assignades a entitats amb perfils d'expressió ben definits (no-*Miscellaneous*) també han mostrat altres trets associats, però no únics, a aquestes entitats, reforçant el diagnòstic obtingut mitjançant el perfil d'expressió. Com per exemple, en l'únic B-CLPD, NOS predit com a cMCL també s'hi va detectar un reordenament *IGL*/*CCND2*, un tipus de reordenament descrit en MCL [130], mentre que en un dels B-CLPD, NOS predits com a HCL s'hi va detectar una mutació en el gen BRAF, associada a aquesta entitat [135].

En les diferents anàlisis del Capítol 4 s'ha vist que l'expressió gènica no ajuda a distingir les entitats LPL, SDRPL i SMZL, motiu pel qual s'han agrupat en un únic grup

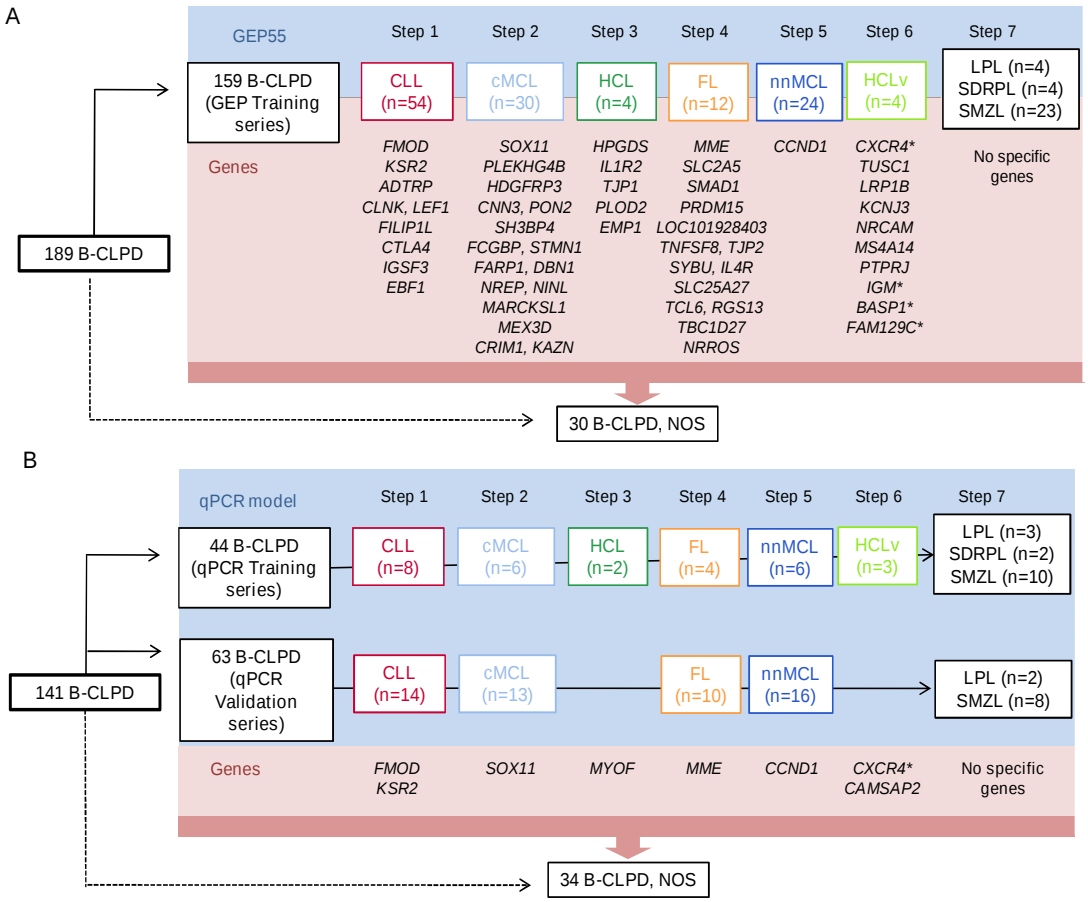

Figura 4.19: Resum de l'enfocament multi-step utilitzat en els predictors. *Per cada entitat s'indica: a quin pas de l'enfocament multi-step es discrimina, el nombre de mostres en el training set i en la sèrie de validació, i els gens que la identifiquen. El panell superior (A) correspon a les dades utilitzades pel predictor d'expressió (GEP55) i el inferior (B) pel predictor en qPCR (qPCR model). Figura mantinguda en anglès com en l'article on s'ha publicat (apartat 4.8).*

anomenat *Miscellaneous*. Per tal de precisar més el diagnòstic de les 35 mostres B-CLPD, NOS assignades a aquest grup segons l'expressió gènica, s'ha estudiat altres característiques moleculars i genètiques publicades en la literatura. En concret, mitjançant seqüenciació Sanger s'ha realitzat l'anàlisi mutacional de 7 gens (*BRAF*, *MAP2K1*, *MYD88*, *NOTCH1*, *NOTCH2*, *SF3B1* i *TP53*) associats a diferents entitats de B-CLPD, i s'ha investigat alteracions cromosòmiques mitjançant tres tècniques diferents: *microarrays* de *copy-number*, citogenètica convencional o f*luorescence in situ hybridization* (FISH) amb sondes específiques.

Aquesta informació, que només s'ha pogut obtenir en un subgrup de les 35 mostres, ha

ajudat a precisar el diagnòstic en 14 d'aquestes. Concretament:

- En 6 mostres s'ha identificat mutacions en el gen *NOTCH2* amb o sense delecions al braç *q* del cromosoma 7, ambdues associades a l'entitat SMZL [136,137].
- En 2 s'ha detectat trisomia 12 (guany de material genètic en tot el cromosoma 12) i guanys parcials al cromosoma 3, alteracions associades a SMZL [137,138].
- En 6 s'ha identificat mutacions en el gen *MYD88*, associat a l'entitat LPL [139], juntament amb presència de banda monoclonal IgM (IgM *paraprotein*), una condició que es detecta en la sang mitjançant tests clínics estàndards i que també està associada a LPL [1].

És important tenir en compte que aquests 14 casos no es podrien diagnosticar com a SMZL o LPL si no s'haguessin descartat altres entitats mitjançant l'expressió gènica, donat que les característiques descrites no són exclusives d'SMZL i LPL en l'espectre global, però si ho són en el petit subconjunt dels *Miscellaneous*. Per exemple, mutacions dels gens *MYD88* i *NOTCH2* també s'han detectat en altres entitats, concretament en pacients de CLL el primer [5] i d'FL el segon [140].

Resumint, **gràcies a la combinació d'expressió gènica, alteracions cromosòmiques i mutacions s'ha pogut diagnosticar una entitat concreta en 43 dels 64 B-CLPD, NOS** (67%), mentre que en els 21 restants (33%) s'han descartat 6 de les entitats (CLL, cMCL, HCL, FL, nnMCL i HCLv). La Figura 4.20 conté un resum del predictor que combina expressió gènica, mutacions i alteracions cromosòmiques (panell A), així com la predicció dels B-CLPD, NOS en la sèrie de *microarrays* (panell B) i en la sèrie de qPCR (panell C).

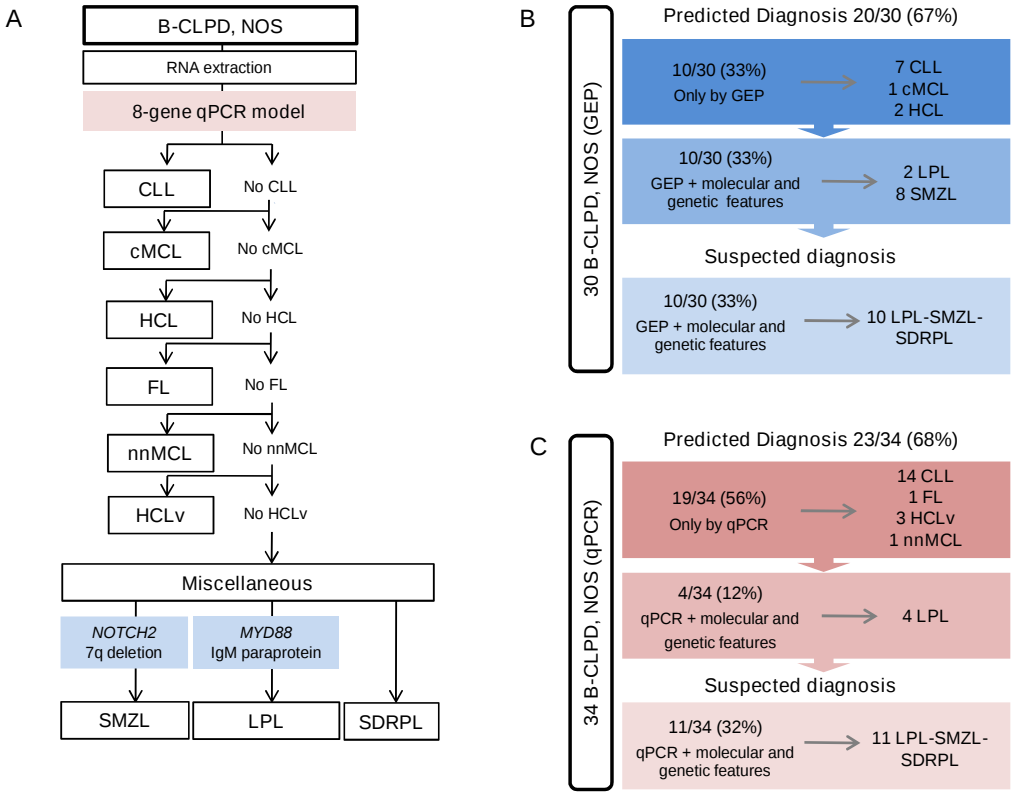

Figura 4.20: Resum de les prediccions de B-CLPD, NOS. *(A) Diagrama de flux del predictor que combina expressió gènica, mutacions i alteracions cromosòmiques. (B) Classificació dels 30 B-CLPD, NOS inclosos en la sèrie de microarrays d'expressió (GEP). (C) Classificació dels 34 B-CLPD, NOS inclosos en la sèrie de qPCR. Figura mantinguda en anglès com en l'article on s'ha publicat (apartat 4.8).*

4.8 Publicació

Part dels resultats presentats en el Capítol 4 d'aquesta tesi han estat prepublicats a la revista *Haematologica* (publicació definitiva al Setembre de 2017):

- Navarro A*, **Clot G***, Martínez-Trillos A, Pinyol M, Jares P, González-Farré B, Martínez D, Trim N, Fernández V, Villamor N, Colomer F, Costa D, Salaverria I, Martín-Garcia D, Erber W, López C, Jayne S, Siebert R, Dyer MJ, Wiestner A, Wilson WH, Aymerich M, López-Guillermo A, Sánchez A, Campo E, Matutes E, Beà S. Improved classification of leukemic B-cell lymphoproliferative disorders using a transcriptional and genetic classifier. *Haematologica*. 2017 May 18. pii: haematol.2016.160374.
  * Primera autoria compartida

5 Resultats: lassoVoting

La metodologia de selecció de gens presentada en el Capítol 3 i utilitzada en l'apartat 4.5 és complexa d'aplicar, donat que combina dos mètodes estadístics diferents, un dels quals té uns requisits computacionals elevats. En aquest capítol es presenta lassoVoting, una metodologia més senzilla d'aplicar i què és adequada en entorns similars als d'aquesta tesi, és a dir, quan l'objectiu és seleccionar un nombre limitat de gens per traslladar-los a una altra plataforma. En el primer apartat d'aquest capítol s'explica lasso, el mètode estadístic en què es basa lassoVoting. En el segon apartat es presenta el propi lassoVoting. En el tercer i quart apartat es compara lassoVoting amb altres mètodes de selecció de variables en entorns simulats i en conjunts de dades reals, respectivament.

5.1 Least absolute shrinkage and selection operator

El mètode *least absolute shrinkage and selection operator* (lasso) [141], implementat en el paquet *glmnet* d'R, es pot interpretar com una modificació dels models lineals generalitzats. Suposem primer el cas en què es vol predir una variable resposta y numèrica en funció de múltiples variables explicatives x, on el model clàssic de regressió lineal es defineix com

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \\ e_i &\sim N(0, \sigma). \end{aligned}$$

Utilitzant el mètode dels mínims quadrats ordinaris, l'estimació del vector de paràmetres $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$ en un *training set* d'N casos s'obté de minimitzar la funció objectiu

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

La modificació de lasso consisteix en afegir una restricció a aquesta minimització de la següent manera

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$
$$\text{subjecte a } \sum_{j=1}^{p} |\beta_j| \leq c$$
$$c \geq 0,$$

és a dir, la restricció consisteix en ficar un límit (c) a la suma dels coeficients en valor absolut. Un dels efectes d'aquest límit és alleugerir l'*overfitting*, ja que no deixa que els coeficients s'adaptin totalment al *training set*.

El problema de minimització es pot reescriure com

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$
$$\lambda \geq 0,$$

on λ es pot interpretar com un paràmetre d'encongiment (*shrinkage*) en les estimacions dels coeficients $\boldsymbol{\beta}$. Notem que la funció objectiu està formada per la suma de dos elements, on el primer és la suma dels residus al quadrat i el segon és la suma dels coeficients en valor absolut. Aleshores, el paràmetre λ controla el pes relatiu que es dóna a cadascun d'aquests elements. Un valor elevat de λ donarà més pes a la suma dels coeficients, en conseqüència, per minimitzar la funció objectiu els coeficients β_j hauran de prendre valors petits. En cas contrari, si λ s'aproxima a zero, els coeficients estimats seran els que minimitzen la suma dels residus al quadrat. Quan $\lambda = 0$, la solució de la funció objectiu és la mateixa que per mínims quadrats ordinaris. λ es pot estimar mitjançant la *cross-validation*.

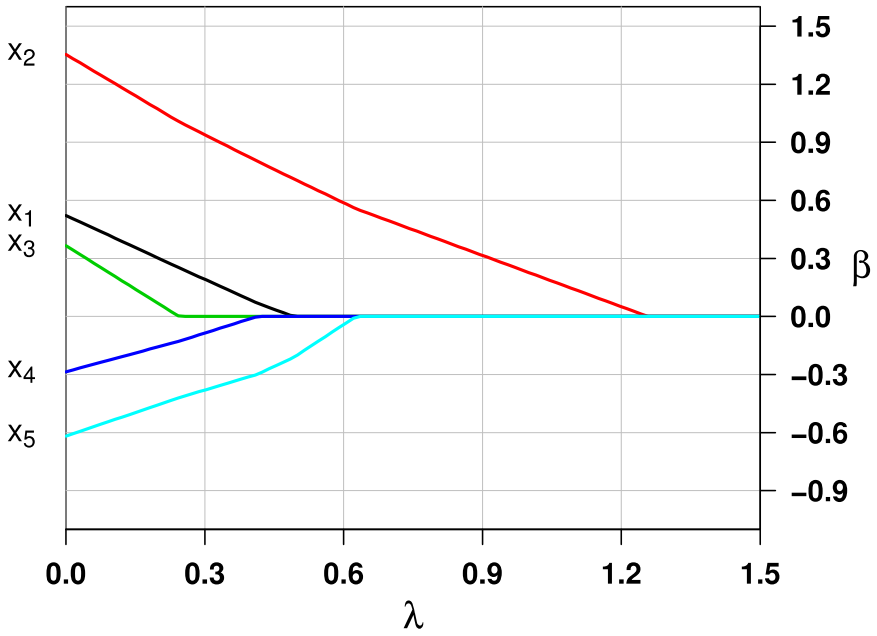

***Figura 5.1: Efecte del paràmetre λ de lasso als coeficients β**. Quan λ = 0, els coeficients β estimats coincideixen amb els estimats per mínims quadrats ordinaris. A mesura que s'augmenta λ els coeficients es van reduint.*

La Figura 5.1 mostra un exemple de com afecta λ als coeficients β_j d'un model de regressió que inclou 5 variables. Al gràfic es pot veure com l'augment en λ es tradueix en aproximar els coeficients a 0, fins que a un cert valor s'anul·len tots (aproximadament a λ = 1.3).

La mateixa restricció es pot aplicar a qualsevol model de la família dels models lineals generalitzats. En aquesta tesi, on s'han realitzat múltiples comparacions binàries, la família binomial (regressió logística) és l'adequada per construir el predictor. El model clàssic quan la variable resposta *y* és binària correspon a

$$y_i \sim Bernoulli(\pi_i),$$
$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

on $\pi_i = P(y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta})$.

Del model anterior es pot aïllar π_i segons

$$\pi_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1+\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}.$$

La funció de versemblança d'una *training set* d'*N* mostres segons aquest model correspon a

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

on, un cop aplicat el logaritme i substituït π_i per la seva expressió, es pot utilitzar el mètode de la màxima versemblança per estimar el vector de paràmetres $\boldsymbol{\beta}$. L'estimació s'obtindria de maximitzar la funció objectiu

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmax} \left\{ \sum_{i=1}^{N} \left(y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) - \ln(1+\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})) \right) \right\}.$$

De la mateixa manera que en el model lineal, lasso afegeix la penalització als coeficients segons

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmax} \left\{ \sum_{i=1}^{N} \left(y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) - \ln(1+\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})) \right) - \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

$\lambda \geq 0$.

Lasso té tres propietats molt importants en entorns d'alta dimensionalitat. La primera és que, a diferència dels models lineals clàssics, es pot aplicar quan $p>n$. La segona, a través del paràmetre λ incorpora la selecció de variables de manera natural, evitant l'ús d'estratègies *stepwise*. A la Figura 5.1 es veu que, depenent del valor de λ, alguns coeficients prenen el valor de 0, el qual significa que el model no faria servir aquestes variables per fer prediccions. La tercera és que, gràcies a l'algoritme *least angle regression* (LAR) [142], es pot obtenir la solució del problema d'optimització per diferents valors de λ simultàniament, reduint en gran mesura el temps computacional.

5.2 lassoVoting

En la literatura s'ha observat que, en entorns d'alta dimensionalitat, es pot evitar l'*overfitting* mitjançat l'ús de mètodes *ensemble* [143]. Aquests mètodes es basen en combinar les prediccions de múltiples predictors, on cada predictor es construeix a partir de remostratges del *training set* original. Existeixen múltiples mètodes de predicció *ensemble*, com el *random forests* [144], l'AdaBoost [145] i el randomGLM [146]. Tot i que en la literatura ha pres menys atenció, els mètodes *ensemble* també es poden utilitzar per augmentar l'estabilitat i la robustesa de la selecció de variables [147,148]. El mètode proposat en aquest capítol, lassoVoting, i el mètode de Dziuda (apartat 3.4.4), formen part d'aquesta última categoria.

LassoVoting combina limma (poder univariant) i lasso (poder multivariant) amb tècniques de remostratge per tal de calcular un *score* que permeti ordenar les variables a traslladar de més a menys rellevants. Lasso té dues propietats que el fan molt adequat com a eina de mesura del poder multivariant. La primera, té un cost computacional reduït gràcies a l'algoritme LAR, una característica rellevant quan s'utilitzen tècniques de remostratge. La segona, lasso tendeix a seleccionar una única variable de blocs correlacionats, el qual serveix per minimitzar la redundància en el conjunt de variables.

El següent algoritme descriu com calcular els *scores* del mètode lassoVoting quan es vol comparar dos classes en una *training set* amb N casos i M variables:

1) Es repeteix B vegades:
   1.1) Es construeix un Monte Carlo *set* obtingut de seleccionar $M \cdot \gamma_m$ variables i $N \cdot \gamma_n$ casos del *training set* original a l'atzar, ambdós arrodonits a la baixa. Els casos es seleccionen per mostreig estratificat aleatori sense reemplaçament, assegurant que la proporció de cada classe en el Monte Carlo *set* sigui similar a la proporció en el *training set* original. Per l'actual repetició b, es calcula

$$F_{jb} = \begin{cases} 1 & \text{si la variable } j \text{ s'inclou en el Monte Carlo } set \text{ de la repetició } b \\ 0 & \text{altrament} \end{cases}.$$

1.2) S'ordenen les variables del Monte Carlo *set* segons el *P*-valor obtingut d'aplicar limma i es seleccionen les *r* primeres que compleixen $|\log_2(FC)| > T_{FC}$. En cas de que hi hagin menys d'*r* variables que ho compleixin, es selecciona les *r* variables amb més $|\log_2(FC)|$.

1.3) Es construeix el Monte Carlo *set* reduït incloent, únicament, les *r* variables. Per tant, aquest *set* contindrà $N \cdot \gamma_n$ casos i *r* variables.

1.4) S'aplica lasso al Monte Carlo *set* reduït amb $\lambda = \lambda_{1se}$. Definim λ_{min} com el valor de λ que minimitza l'error (e_{min}) estimat mitjançant 10-*fold cross-validation* estratificada. Definim λ_{1se} com el valor de λ més gran tal que el respectiu error (e_{1se}) compleix: $e_{1se} < (e_{min} + sd(e_{min}))$, on $sd(e_{min})$ és l'error estàndard d'e_{min} . Del model lasso s'obté l'estimació de $\boldsymbol{\beta}$ per l'actual repetició.

1.5) Per l'actual repetició *b*, es calcula

$$S_{jb} = \begin{cases} 1 & \text{si} |\beta_j| > 0 \\ 0 & \text{altrament} \end{cases}.$$

2) Es calcula l'*score* de cada variable segons

$$Score_j = \frac{\sum_{b=1}^{B} S_{jb}}{\sum_{b=1}^{B} F_{jb}}.$$

En resum, l'*score* d'una variable *j* correspon al percentatge de vegades que ha estat seleccionada per limma i lasso en els *B* Monte Carlo *sets* en què estava present. En el pas 1.2 es té en compte la rellevància estadística (*P*-valor) i la biològica (FC), així com la rellevància multivariant en el pas 1.3, on lasso procura seleccionar una sola variable de blocs correlacionats. En el pas 1.1 es selecciona un subconjunt de variables i casos a l'atzar, un procediment similar al que fa el mètode *random forests*. Aquesta selecció a l'atzar afegeix aleatorietat al *set*, el qual ajuda a que s'explorin diferents combinacions de variables per identificar les més robustes.

L'Annex A conté el codi en R per calcular l'*score* de lassoVoting en el cas concret de comparar dues classes. Els següents paràmetres del mètode els ha de definir l'usuari:

- *B*: el nombre de Monte Carlo *sets* que es creen. Com més gran millor. En *microarrays* d'expressió es recomana un mínim de 500, mentre que en *training sets* amb més variables es recomana augmentar-ho.
- γ_m: percentatge de variables seleccionades a l'atzar que s'inclouen a cada Monte Carlo *set*. Es recomana entre 0.2 i 0.5.
- γ_n: percentatge de casos de cada classe seleccionats a l'atzar que s'inclouen a cada Monte Carlo *set*. Es recomana entre 0.1 i 0.2.
- *r*: nombre de variables que passen el filtre de limma i del $\log_2$(FC). Aquest paràmetre restringeix el nombre de variables que es consideren amb poder predictiu univariant. Es recomana entre 50 i 200, tot i que dependrà de la quantitat de variables que s'espera que continguin informació.
- T_{FC}: valor mínim del $\log_2$(FC) d'una variable per a que passi el filtre de $\log_2$(FC). Dependrà de la correlació entre les dues plataformes. En *microarrays* d'expressió es recomana entre 1 i 1.5.

Notem que aquesta metodologia es pot adaptar molt fàcilment a diferents tipus de variables resposta gràcies a la versatilitat de lasso. A part de la família dels models lineals generalitzats i l'anàlisi de la supervivència [149], existeixen modificacions que serveixen per situacions més concretes. Per exemple, *group* lasso [150] força que grups de variables predefinits entrin o surtin conjuntament del model, una característica útil quan s'agrupen gens amb funcions biològiques similars. En cas que les variables tinguin un ordre temporal o espacial es pot utilitzar *fused* lasso [151], el qual suavitza els coeficients al llarg de l'estructura.

Una característica important d'aquesta metodologia és que, a través dels paràmetres *r* i T_{FC}, es pot canviar el pes que es dóna a les tres rellevàncies (estadística, biològica, multivariant). Un valor petit d'*r* no dóna opció a lasso a estudiar la importància multivariant, per tant, hi hauria un gran pes en la rellevància estadística univariant. Per altra banda, un valor de $T_{FC} = 0$ serveix per anul·lar el pes de la rellevància biològica. Notem que els càlculs associats a aquests dos paràmetres, limma i $\log_2$(FC), són especialment adequats en *microarrays* d'expressió, per tant, en altres plataformes

s'haurien de canviar per estadístics adequats a la plataforma corresponent.

5.3 Comparació de lassoVoting amb altres mètodes

Un cop definit el mètode, convé comparar-lo amb altres per comprovar si el seu comportament és adequat per seleccionar gens en *microarrays* a traslladar a qPCR. L'estratègia ideal per fer la comparació seria, primer, identificant diferents subconjunts de gens mitjançant diversos mètodes de selecció, després, mesurant cada subconjunt en qPCR i, finalment, avaluant quin subconjunt ha reproduït millor la informació dels *microarrays*. El mètode amb millor rendiment seria el que ha seleccionat el subconjunt més ben reproduït. El problema d'aquesta estratègia és que requereix un cost experimental elevat per dos motius: *i*) s'ha de mesurar una gran quantitat de gens al combinar diversos subconjunts, i *ii*) s'ha de disposar d'una gran quantitat de casos per poder avaluar-ho amb precisió. A més a més, hi ha l'inconvenient afegit de que la correlació *microarray*-qPCR podria dependre del tipus de càncer o teixit analitzat, dificultant la generalització dels resultats. Una segona opció seria utilitzar conjunts aparellats de dades (*microarrays*-qPCR) publicats en la literatura. En el cas dels *microarrays* existeixen repositoris públics de dades, com el Gene Expression Omnibus (GEO) [152], però les dades de qPCR no s'acostumen a fer públiques. La tercera opció, i la que s'ha utilitzat en aquesta tesi, és fer-ho només amb dades de *microarrays*.

L'estratègia per fer la comparació en *microarrays* consisteix en, primer, identificar diferents subconjunts de k variables obtinguts amb diferents mètodes de selecció i, segon, calcular l'error de predicció que s'obtindria amb cada subconjunt. Si un mètode obté errors de predicció menors significa que els subconjunts de variables que identifica contenen més informació sobre la discriminació de classes. Aquests mètodes traslladarien més informació a qPCR, augmentant així la probabilitat de reproduir els resultats en aquesta segona plataforma.

Aquesta estratègia assumeix que el canvi de plataforma afecta de la mateixa manera a tots els mètodes de selecció, el qual, en dades de *microarrays* d'expressió, no és cert,

atès que no tenir en compte el $\log_2$(FC) augmenta el risc a no reproduir correctament la informació [32,35] (apartat 1.4.4). Tot i així, les incongruències conegudes entre les HTT i les LTT són fàcils d'adaptar a qualsevol mètode de selecció. No s'ha cregut que les conclusions extretes amb aquesta estratègia de comparació estiguin allunyades de les que s'obtindrien mitjançant alguna de les altres dues.

En els següents subapartats d'aquest apartat es comparen diversos mètodes de selecció en dades simulades i en dades reals de *microarrays*, respectivament. Els mètodes que s'han comparat han sigut:

- LassoVoting amb $B = 500$, $\gamma_m = 0.2$, $\gamma_n = 0.1$, $r = 80$, $T_{FC} = 0$.
- LassoVoting amb $B = 500$, $\gamma_m = 0.5$, $\gamma_n = 0.1$, $r = 80$, $T_{FC} = 0$.
- Limma.
- *Support vector machine recursive feature elimination* (SVM-RFE) [99]. S'ha utilitzat la implementació d'SVM del paquet *e1071*, amb el *kernel* lineal i $C = 10$.
- *Random forests variable importance* (RF-VIMP) [144] amb B = 500 i $\gamma_m = \text{sqrt}(M)/M$. Implementat en el paquet *randomForest* d'R.
- RandomGLM amb $B = 500$, $\gamma_m = 0.2$, $\gamma_n = 0.1$, $r = 80$. Implementat en el paquet *randomGLM* d'R.
- RandomGLM amb $B = 500$, $\gamma_m = 0.5$, $\gamma_n = 0.1$, $r = 80$.

S'ha cregut que aquesta selecció de mètodes és suficient per avaluar lassoVoting, ja que les bases teòriques de cadascun són molt diferents. A més a més, comparar-ne més augmentaria la càrrega computacional en gran mesura. Limma és el representant dels mètodes univariants, mentre que l'SVM-RFE és el dels mètodes *stepwise*. Per últim, hi ha dos representats dels mètodes *ensemble*: RF-VIMP, donat que és el més utilitzat en la literatura, i randomGLM, el qual té unes característiques similars a lassoVoting.

Els diversos paràmetres de cada mètode s'han fixat a un valor concret, donat que estudiar quins valors obtenen millors resultats suposaria una gran càrrega computacional. Els valors utilitzats serveixen per donar una idea del comportament de

cada mètode. Degut a que no hi ha canvi de plataforma, s'ha fixat T_{FC} = 0 en lassoVoting. Malauradament, en el llistat de mètodes no s'ha pogut incloure la metodologia utilitzada en l'apartat 4.5 (limma + Dziuda), ja que és difícil d'automatitzar i requereix una forta càrrega computacional.

A part dels mètodes de selecció de variables, també s'han considerat diferents mètodes de classificació per construir els predictors, ja que hi podria haver sinergia entre els dos. Per exemple, la selecció segons SVM-RFE podria obtenir millors resultats amb el mètode SVM, el qual té en compte l'estructura de correlació, que amb el DLDA, que no la té. De la mateix manera, limma podria combinar millor amb DLDA. Per aquest motiu s'han utilitzat els següents quatre mètodes, els quals estan ordenats segons la complexitat que permeten en l'estructura de correlació:

- *Nearest mean classification* (NMC). Predictor que només té en compte la distància euclidiana als centroides de cada classe. Implementada en la funció *nm* del paquet *klaR* d'R.
- DLDA. Versió de l'LDA que assumeix que les correlacions entre variables són nul·les.
- LDA. Assumeix normalitat i matrius de covariàncies iguals en les dues classes. Implementat en la funció *lda* del paquet *MASS* d'R.
- SVM. El mètode de classificació que té en compte estructures més complexes. Implementat en la funció *svm* del paquet *e1071* d'R. La funció *kernel* utilitzada s'ha fixat a la lineal, mentre que per *cross-validation* s'ha estimat el valor òptim de $C = \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

L'última consideració a l'hora de comparar els diferents mètodes ha sigut el nombre de variables (k) incloses en el predictor. Tenint en compte l'entorn d'aquesta tesi, on l'interès està en seleccionar unes poques variables a traslladar a qPCR, s'ha comparat el rendiment en $k = \{1, 2, 3, 5, 7, 10, 15\}$. Resumint, en cada *training set* s'ajustaran (7 mètodes de selecció) · (4 mètodes de classificació) · (7 quantitats de variables) = 196 predictors.

5.3.1 Escenaris simulats

Per tal de fer la comparació en dades simulades s'ha considerat el problema de discriminar dues classes (A i B) en cinc escenaris diferents. Per cada escenari s'han simulat 100 *training sets* de 40 casos (20 de la classe A i 20 de la B) i 3000 variables. En cadascun d'aquests *training sets* s'han ajustat els 196 predictors per, finalment, avaluar-los en *test sets* simulats de 10000 casos (5000 de cada classe). La mesura utilitzada per comparar el rendiment dels predictors ha sigut la mitjana de l'error de predicció obtinguda en les 100 simulacions. Notem que aquesta mesura correspon al valor real d'*Err* (apartat 3.5), atès que amb les 10000 mostres del *test set* calculem el valor d'Err_T corresponent a cada *training set*.

En tots cinc escenaris, les 3000 variables s'han simulat segons la distribució normal multivariant, on les estructures dels centroides de les classes A i B es poden resumir segons

$$\begin{aligned}\boldsymbol{\mu_A} &= \{\mu_1,\ \mu_2,\ \ldots,\ \mu_{40},\ \mu_{41}=0,\ \ldots,\ \mu_{3000}=0\},\\ \boldsymbol{\mu_B} &= \{\mu_1=0,\ \ldots,\ \mu_{3000}=0\},\end{aligned}$$

és a dir, només 40 de les 3000 variables tenen informació sobre discriminació de classes. Les estructures de les matrius de covariàncies es poden resumir segons

$$\begin{aligned}\boldsymbol{\Sigma_A} &= \begin{pmatrix}\Sigma_{a,40\times 40} & 0\\ 0 & I_{2960}\end{pmatrix},\\ \boldsymbol{\Sigma_B} &= \begin{pmatrix}\Sigma_{b,40\times 40} & 0\\ 0 & I_{2960}\end{pmatrix},\end{aligned}$$

és a dir, només hi haurà correlacions no-nul·les en les 40 variables que tenen informació. En tots els escenaris, les variàncies (diagonals de les matrius $\boldsymbol{\Sigma_A}$ i $\boldsymbol{\Sigma_B}$) s'han fixat a 1.

Escenari 1 (sense correlació)

El primer escenari es caracteritza per no tenir correlació entre variables i una diferència d'1 unitat entre les classes en les 40 variables amb informació:

$$\boldsymbol{\mu}_{\boldsymbol{A}} = \{\mu_1 = 1,\ \ldots,\ \mu_{40} = 1,\ \mu_{41} = 0,\ \ldots,\ \mu_{3000} = 0\},$$
$$\boldsymbol{\Sigma}_{\boldsymbol{A}} = \boldsymbol{\Sigma}_{\boldsymbol{B}} = I_{3000}.$$

La Figura 5.2 mostra la mitjana dels errors obtinguts en les 100 simulacions dels predictors construïts mitjançant el mètode DLDA, el qual s'ha representat per si sol ja que ha obtingut errors menors que els mètodes NMC, LDA i SVM. Al gràfic de l'esquerra s'ha representat, per cadascun dels set mètodes de selecció, l'error respecte el nombre de variables (*k*) en el predictor, on es pot veure com l'error sempre disminueix a l'augmentar *k*. Al gràfic de la dreta hi han representats els mateixos errors un cop se'ls hi ha restat l'error corresponent a limma, on es pot veure que limma i els dos lassoVoting obtenen errors similars i menors que la resta de mètodes. Notem que SVM-RFE i RF-VIMP arriben a obtenir gairebé un 5% més d'error que limma.

Escenari 2 (un únic bloc de variables correlacionades)

En aquest escenari s'han utilitzat els mateixos centroides que en el primer escenari. En canvi, la correlació entre les 40 variables amb informació és de 0.75 per la classe A, i de 0.6 per la classe B. És a dir, les 40 variables amb informació formen un únic bloc fortament correlacionat, mentre la resta no tenen informació ni estan correlacionades.

La Figura 5.3 mostra els resultats en aquest cas, on s'ha representat l'error pels quatre mètodes de predicció. Als gràfics es pot veure que limma és més precís per qualsevol nombre de variables. Aquest resultat és degut a que els mètodes que tenen en compte l'estructura de correlació no estimen correctament aquesta estructura (*overfitting*). En aquest escenari el biaix de limma (originat d'ignorar la correlació) és menor que la variància dels altres mètodes.

Notem també que en els gràfics de l'LDA i de l'SVM l'error augmenta a partir de 7 variables. Aquests mètodes intenten estimar l'estructura de correlació, però, al només disposar de 40 mostres, les estimacions són poc robustes i, en conseqüència, la precisió en surt perjudicada.

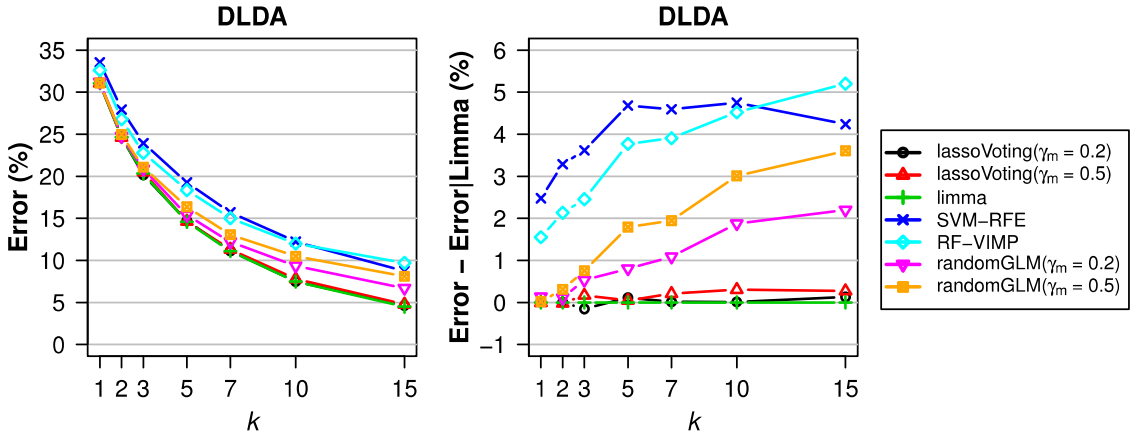

Figura 5.2: Error de classificació en el primer escenari simulat. *En el gràfic de l'esquerra s'ha representat l'error segons el nombre de variables (k) per diferents mètodes de selecció. En el de la dreta s'ha representat l'increment de l'error respecte limma. Els errors representats corresponen a predictors construïts mitjançant el mètode de classificació DLDA.*

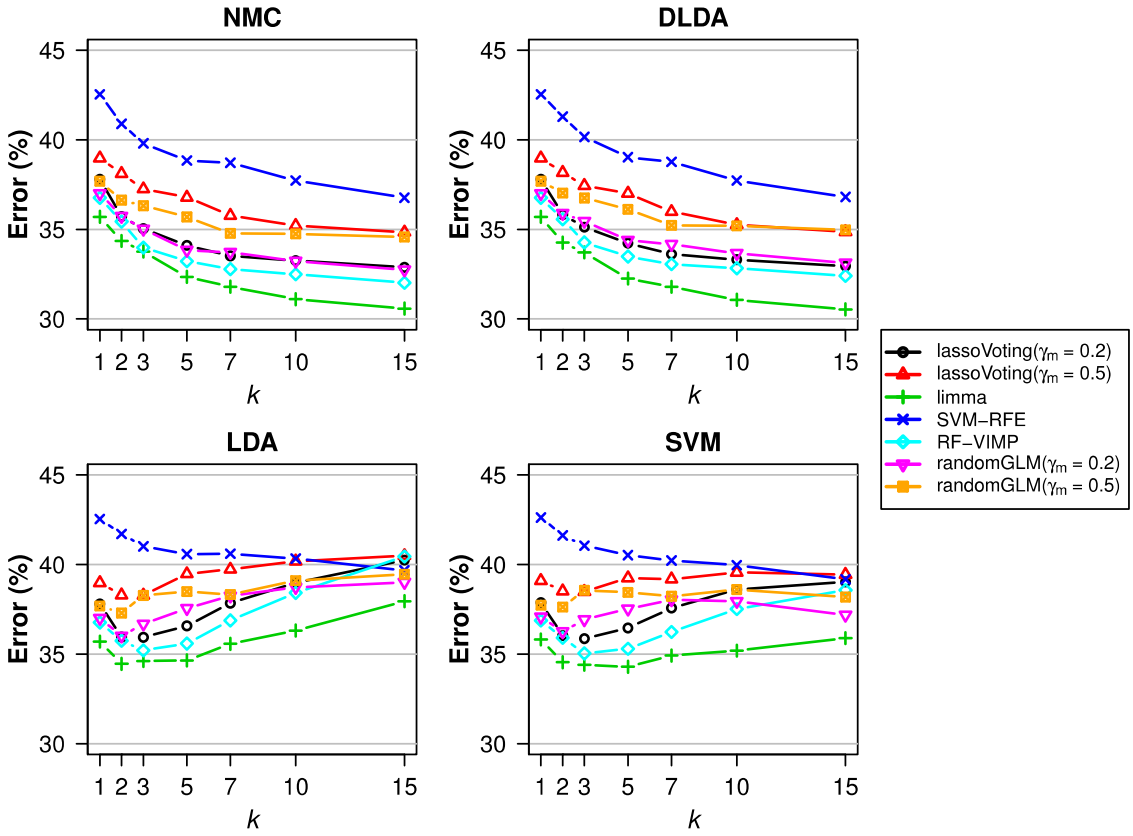

Figura 5.3: Error de classificació en el segon escenari simulat. *Error segons el nombre de variables (k) per diferents mètodes de selecció. Cada gràfic correspon a un mètode de classificació diferent.*

Escenari 3 (blocs de variables amb correlació moderada)

Aquest escenari és més complex que els dos anteriors, on les 40 variables amb informació estan estructurades en 10 blocs de 4. Les variables incloses en un mateix bloc estan correlacionades entre si (intrabloc), mentre que la resta de correlacions són nul·les. En la classe A la correlació intrabloc és de 0.4, mentre que en la classe B és de 0.28. Els 40 components informatius del centroide de la classe A varien entre 0.61 i 1.88.

La Figura 5.4 mostra l'error del mètode de classificació DLDA en aquest escenari, on, de manera similar al primer escenari, ha obtingut errors més petits que els altres tres mètodes de classificació. Al gràfic de l'esquerra s'ha representat l'error respecte el nombre de variables (*k*) en el predictor per cadascun dels set mètodes de selecció, on es pot veure com l'error sempre disminueix a l'augmentar *k*. Al de la dreta hi han representats els mateixos errors un cop se'ls hi resta l'error segons limma, on es pot veure un perfil diferent que en la Figura 5.2. En aquest escenari limma i lassoVoting obtenen errors més petits en general, on lassoVoting ho fa lleugerament millor per k = {1, 5, 7} i limma en la resta. Tot i així, les diferències entre els dos mètodes són menors de l'1%. Novament, SVM-RFE obté errors força més elevats.

Escenari 4 (blocs de gens amb correlació forta)

Aquest escenari és igual que el tercer, on només canvia la correlació intrabloc. Concretament, és de 0.85 per la classe A i de 0.6 per la classe B.

A la Figura 5.5 hi ha representat l'error segons el mètode DLDA, atès que ha tornat a obtenir millors resultats. Al gràfic de l'esquerra s'ha representat l'error respecte el nombre de variables en el predictor (*k*) per cadascun dels set mètodes de selecció, on es pot veure que, novament, l'error disminueix a l'augmentar *k*. Al de la dreta hi han representats els mateixos errors un cop se'ls hi resta l'error segons limma, on es pot veure que lassoVoting obté fins a un 2% menys d'error entre 2 i 10 variables. Per 1 i 15 variables limma és el mètode amb menys error.

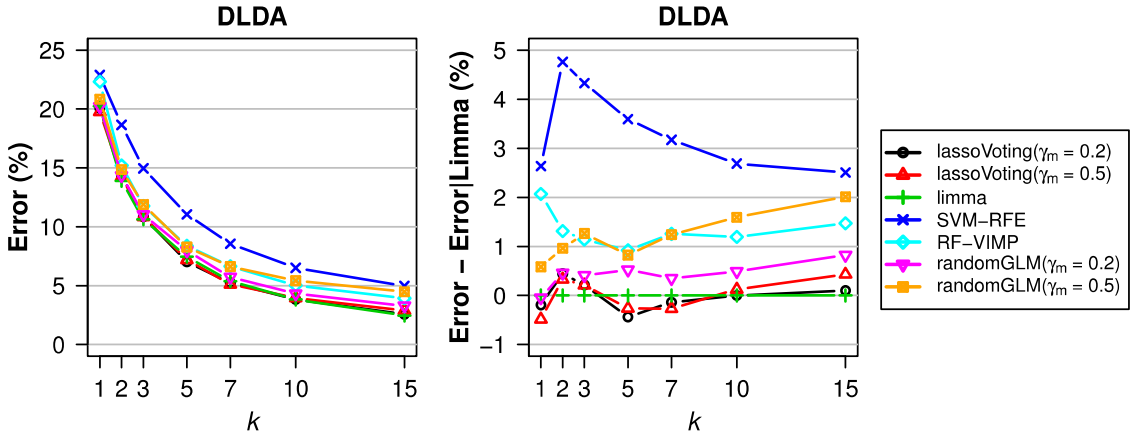

***Figura 5.4: Error de classificació en el tercer escenari simulat**. En el gràfic de l'esquerra s'ha representat l'error segons el nombre de variables (k) per diferents mètodes de selecció. En el de la dreta s'ha representat l'increment de l'error respecte limma. Els errors representats corresponen a predictors construïts mitjançant el mètode de classificació DLDA.*

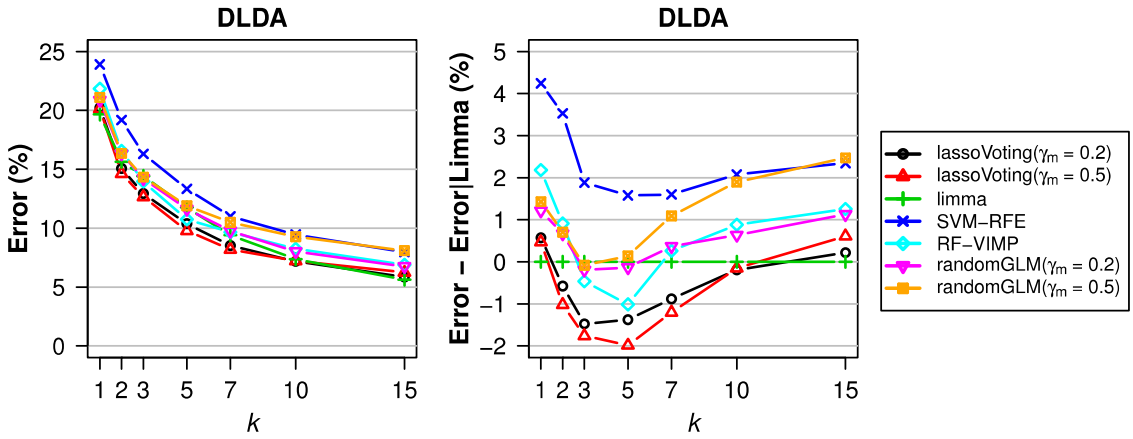

***Figura 5.5: Error de classificació en el quart escenari simulat**. En el gràfic de l'esquerra s'ha representat l'error segons el nombre de variables (k) per diferents mètodes de selecció. En el de la dreta s'ha representat l'increment de l'error respecte limma. Els errors representats corresponen a predictors construïts mitjançant el mètode de classificació DLDA.*

Escenari 5 (efecte gran i blocs de gens amb correlació forta)

L'últim escenari té la mateixa estructura de correlació que l'escenari 4, però els 40 components amb informació del centroide de la classe A varien entre 2.11 i 3.38.

A la Figura 5.6 hi ha representat l'error de manera equivalent a les Figures 5.2, 5.4 i 5.5. Al gràfic de l'esquerra es pot veure que, per qualsevol mètode de selecció, l'error arriba gairebé al 0% quan s'inclouen 10 o més variables. En aquest escenari limma obté menys

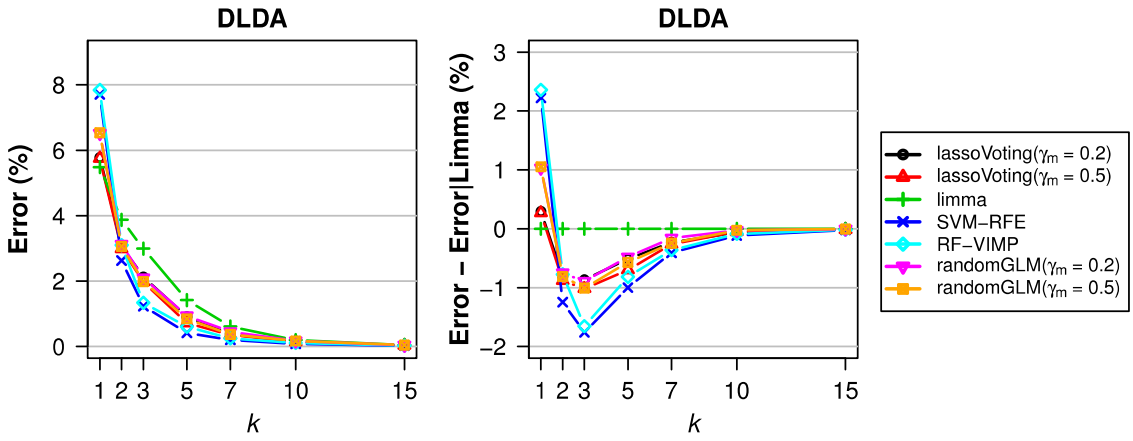

***Figura 5.6: Error de classificació en el cinquè escenari simulat**. En el gràfic de l'esquerra s'ha representat l'error segons el nombre de variables (k) per diferents mètodes de selecció. En el de la dreta s'ha representat l'increment de l'error respecte limma. Els errors representats corresponen a predictors construïts mitjançant el mètode de classificació DLDA.*

error quan només s'utilitza una variable, però a partir de 2 n'obté més que tots els mètodes multivariants. A diferència de la resta d'escenaris, en aquest SVM-RFE obté millor rendiment, especialment per $k = 3$.

Resum

Dels diversos escenaris simulats es poden extreure les següents conclusions:

- DLDA i NMC han obtingut molt bon rendiment respecte LDA i SVM. Tot i que no s'han presentat tots els gràfics i no és l'objectiu d'aquest estudi, és rellevant tenir-ho en compte. Disposar de *training sets* amb 20 mostres per classe dificulta obtenir estimacions robustes de l'estructura de correlació, perjudicant mètodes més complexos. NMC podria haver obtingut resultats similars a DLDA gràcies a que les variàncies de totes les variables s'han fixat al mateix valor.
- Si es volen seleccionar moltes variables (>15), limma és la millor metodologia. En tots els escenaris l'error mínim per $k = 15$ s'ha aconseguit amb limma. És raonable extrapolar aquest resultat a quantitats de variables encara més elevades.
- Si es vol seleccionar un nombre moderat de variables (<15), lassoVoting rendeix molt bé si l'estructura de correlació entre variables està organitzada per blocs. Del tercer al cinquè escenari, lassoVoting té molt bon rendiment en general. Només en el cinquè, amb $k = 3$, és superat significativament per SVM-RFE i

RF-VIMP, els dos mètodes que, en general, han obtingut pitjor rendiment en la resta d'escenaris.

- Quan l'estructura de correlació no està organitzada per blocs, el millor mètode és limma. Tot i que lassoVoting ha obtingut un rendiment similar quan la correlació és nul·la (primer escenari), limma ha superat a la resta de mètodes de manera molt clara en el segon escenari, on hi havia un únic bloc de variables molt correlacionades. És raonable concloure que a més correlació no-estructurada, pitjor ho fan els mètodes que intenten estimar la correlació. Els mètodes multivariants funcionen millor quan poden seleccionar variables de blocs diferents, atès que variables de diferents blocs contenen informació complementària.
- El paràmetre γ_m de lassoVoting i randomGLM té un efecte en l'error obtingut. Amb $\gamma_m = 0.2$, randomGLM obté millors resultats en tots els escenaris, excepte en el cinquè, on obté lleugerament millor rendiment $\gamma_m = 0.5$. Per lassoVoting sembla millor $\gamma_m = 0.5$, donat que ho fa igual o millor en tots els escenaris excepte en el segon, on és menys important degut a que el rendiment dels dos valors ha sigut insuficient.

Els resultats d'aquestes simulacions han permès obtenir una idea aproximada del comportament d'aquests mètodes de selecció en diferents escenaris, però s'han de tenir en compte dues limitacions. La primera, els *training sets* simulats han consistit de 20 mostres per classe, aleshores, les conclusions extretes només es poden extrapolar a grandàries mostrals similars. S'ha decidit utilitzar aquesta quantitat de mostres ja que és una quantitat habitual a la pràctica. La segona, els diferents mètodes de selecció, excepte limma, tenen paràmetres que es poden optimitzar per adaptar-los millor a cada situació. Degut al cost computacional de fer-ho, s'han fixat a un valor concret.

5.3.2 Conjunts de dades reals

Un cop estudiats els diferents mètodes de selecció en escenaris simulats, s'ha avaluat lassoVoting en diversos conjunts de dades reals d'expressió gènica. La mesura utilitzada

per la comparació ha sigut l'error de predicció balancejat, el qual s'ha estimat mitjançant 10-*fold cross-validation* estratificada i repetida 20 vegades. Notem que en els escenaris simulats (apartat 5.3.1) s'ha obtingut el valor real d'*Err*, mentre que en aquest apartat s'obté una estimació d'*Err*.

S'han utilitzat vuit conjunts de dades, quatre provinents de les dades d'aquesta tesi i quatre provinents de conjunts de dades d'altres estudis publicats en la literatura. L'estructura de cada conjunt és la següent:

- B-CLPD CLL: Consisteix de 20546 gens i 159 mostres (54 CLL i 105 no-CLL).
- B-CLPD cMCL: Consisteix de 20546 gens i 105 mostres (30 cMCL i 75 no-cMCL).
- B-CLPD FL: Consisteix de 20546 gens i 71 mostres (12 FL i 59 altres B-CLPD).
- B-CLPD nnMCL: Consisteix de 20546 gens i 59 mostres (24 nnMCL i 35 no-nnMCL).
- *Colon* [153]: Consisteix de 2000 gens i 62 mostres (22 tumorals i 40 normals).
- *Leukemia* [154]: Consisteix de 7129 gens i 72 mostres (47 ALL i 25 AML).
- *Prostate* [155]: Consisteix de 12161 gens i 102 mostres (52 tumorals i 50 normals).
- *Sclerosis* [156]: Consisteix de 20546 gens i 27 mostres (12 MS i 15 controls).

A la Figura 5.7 hi ha representat l'error balancejat estimat en els quatre conjunts de dades d'aquesta tesi, mentre que en la Figura 5.8 hi ha representat l'error balancejat en els quatre conjunts corresponents a altres estudis. A diferència dels escenaris simulats, cap dels quatre mètodes de classificació (NMC, DLDA, LDA, SVM) ha obtingut sistemàticament millor rendiment. Per aquest motiu, s'ha representat per cada mètode de selecció l'error corresponent al mètode de classificació que el minimitza. Per exemple, en el conjunt B-CLPD CLL, el mètode de selecció limma ha obtingut menys error quan s'ha aparellat amb el mètode de classificació LDA, mentre que RF-VIMP ho ha fet quan s'ha aparellat amb NMC. En les dues figures la columna de l'esquerra correspon a representar l'error estimat respecte el nombre de variables (k), mentre que la columna dreta correspon a l'error de cada mètode menys l'error de limma. Per altra banda, cada

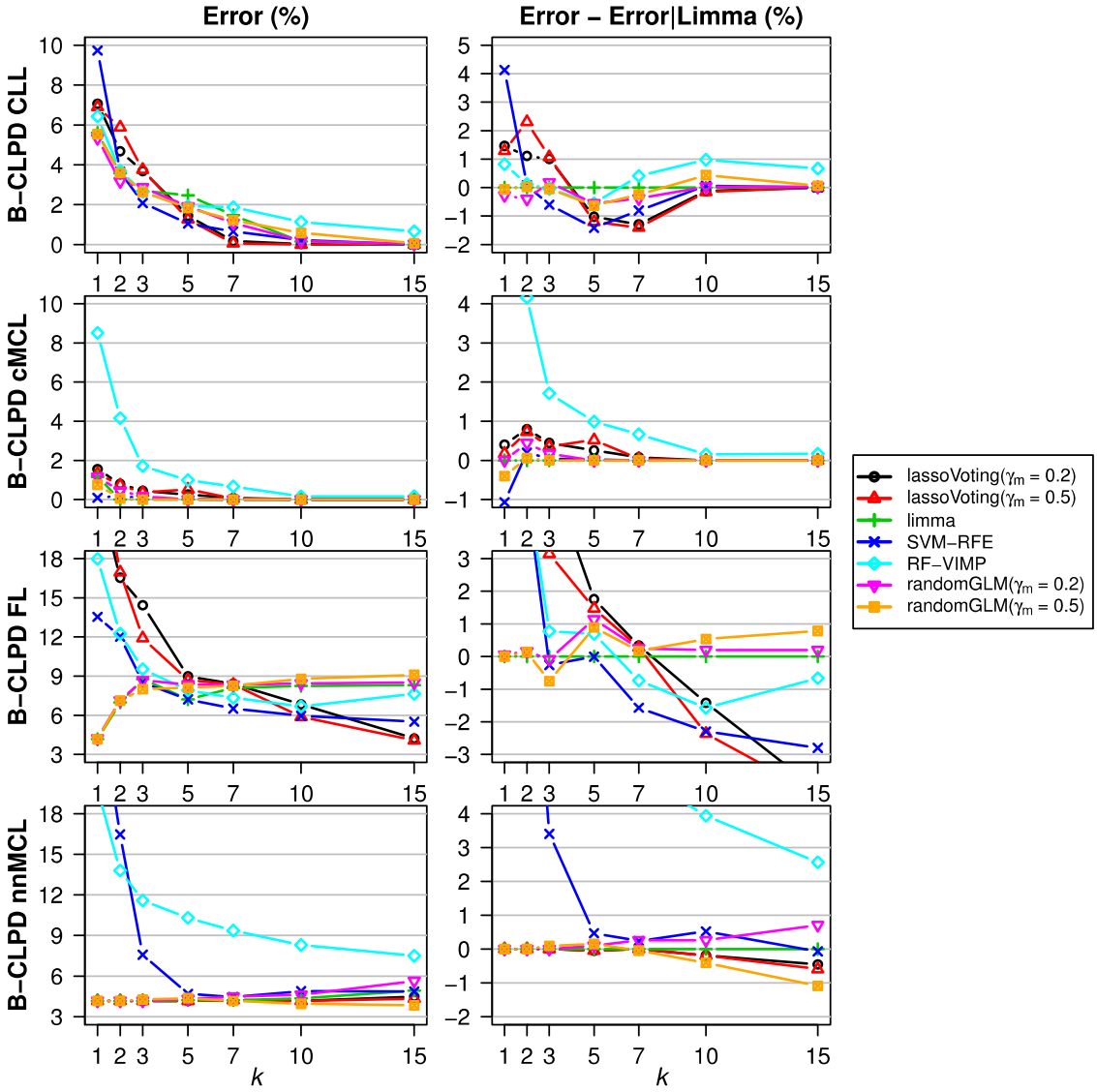

***Figura 5.7: Error estimat en els quatre conjunts de dades d'aquesta tesi**. En els gràfics de la columna esquerra, l'eix x correspon a l'error balancejat, mentre que l'eix y correspon al nombre de variables (k) incloses en el predictor. En els de la dreta l'eix x és el mateix, mentre que l'eix y correspon al mateix error un cop se li resta l'error corresponent al mètode de selecció limma. Per cada mètode de selecció només s'ha representat l'error del mètode de classificació que el minimitza. L'error balancejat s'ha estimat mitjançant 10-fold cross-validation estratificada i repetida 20 vegades.*

fila de les figures correspon a un conjunt de dades diferent.

Les conclusions que es poden extreure per cada conjunt de dades són:

- B-CLPD CLL: lassoVoting i SVM-RFE obtenen els errors més grans en un inici, però a l'anar augmentant *k* acaben obtenint millor rendiment que la resta de mètodes. En concret, lassoVoting (γ_m = 0.5) obté bons rendiments quan k ≥ 5 i

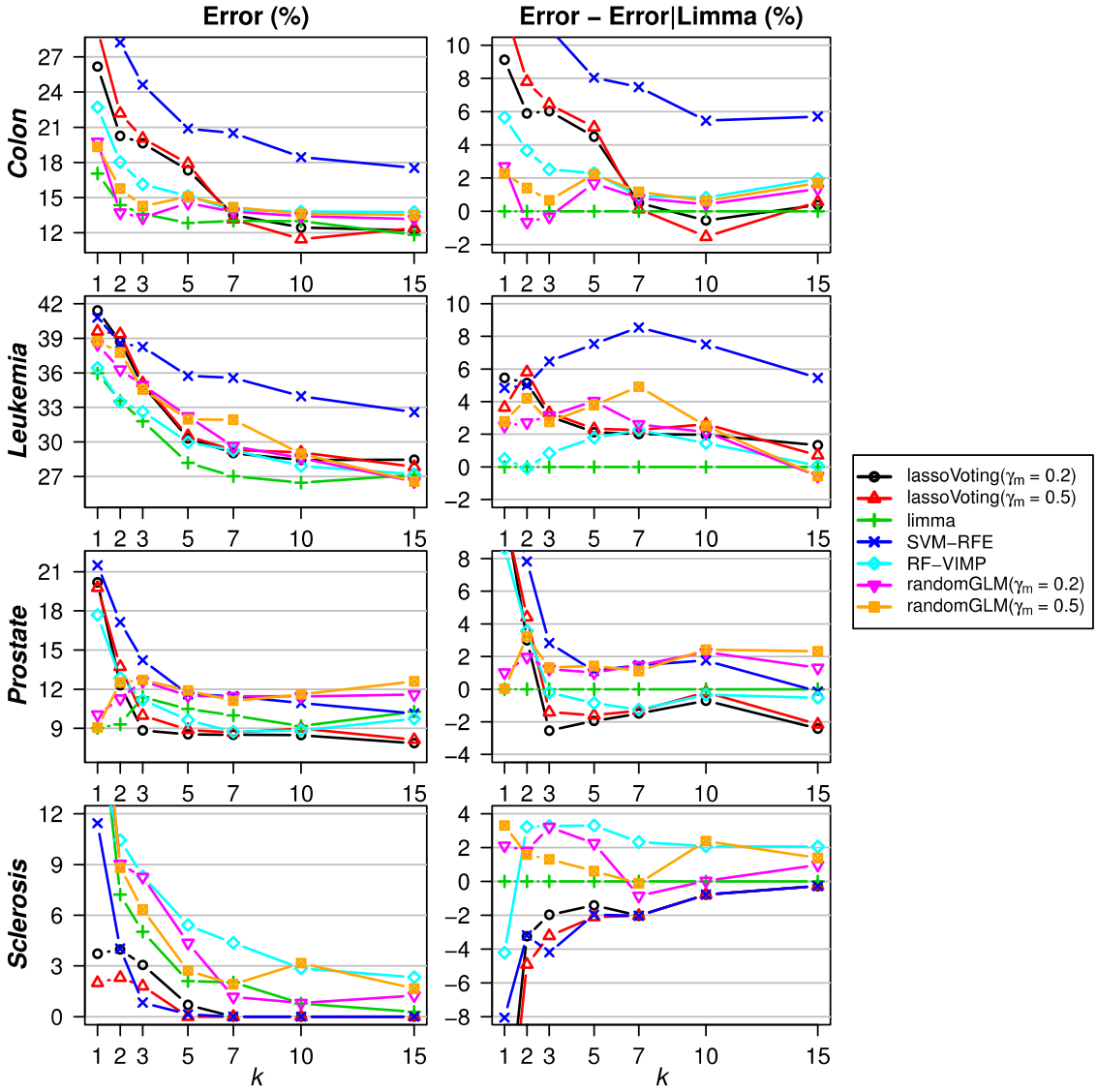

Figura 5.8: Error estimat en els quatre conjunts de dades d'altres estudis. *En els gràfics de la columna esquerra, l'eix x correspon a l'error balancejat, mentre que l'eix y correspon al nombre de variables (k) incloses en el predictor. En els de la dreta l'eix x és el mateix, mentre que l'eix y correspon al mateix error un cop se li resta l'error corresponent al mètode de selecció limma. Per cada mètode de selecció només s'ha representat l'error del mètode de classificació que el minimitza. L'error balancejat s'ha estimat mitjançant 10-fold cross-validation estratificada i repetida 20 vegades.*

dolents quan k ≤ 3, essent el primer mètode en aconseguir errors pròxims a zero ($k = 7$).

- B-CLPD cMCL, B-CLPD FL i B-CLPD nnMCL: en aquests tres conjunts limma té, en general, menys error que la resta de mètodes, per tant, no s'obté una millora de considerar l'estructura multivariant. En el conjunt B-CLPD cMCL l'error és pràcticament del 0% amb molt poques variables. En el B-CLPD FL

utilitzar més d'una provoca *overfitting*. En B-CLPD nnMCL l'error és pròxim a 4% amb un gen i no disminueix a partir d'aquí.

- *Colon*: limma obté bons rendiments per qualsevol *k*, però és lassoVoting el que minimitza l'error quan $k = 10$, obtenint quasi un 2% menys que limma. Tot i així, lassoVoting obté un rendiment força dolent quan $k < 7$.
- *Leukemia*: en aquest conjunt no hi ha gaire justificació d'utilitzar mètodes complexes, atès que limma obté bons rendiments per qualsevol *k*.
- *Prostate*: en aquest conjunt s'obté un error molt baix amb limma quan $k = 1$. LassoVoting ($\gamma_m = 0.2$) obté una lleugera millora respecte aquest error quan $k \geq 3$.
- *Sclerosis*: lassoVoting i SVM-RFE obtenen errors més baixos en general. LassoVoting minimitza l'error per qualsevol nombre de variables, excepte $k = 3$.

En general limma ha obtingut molt bon rendiment en tots els conjunts de dades, excepte en el conjunt *Sclerosis*, on els errors dels diversos mètodes suggereixen que l'estructura multivariant és molt important. També en els conjunts B-CLPD CLL, *Colon* i *Prostate* hi ha un lleuger benefici de considerar aquesta estructura. En aquests quatre conjunts, el mètode lassoVoting ha sigut el que ha minimitzat l'error. De la mateixa manera que en els escenaris simulats, utilitzar $\gamma_m = 0.5$ obté millor rendiment que $\gamma_m = 0.2$ en la majoria dels conjunts, però no hi ha gairebé diferència entre els dos. Tot i que en general lassoVoting obté bons rendiments per $5 < k < 10$, és important tenir en compte que obté errors molt elevats en algunes situacions, així que és recomanable comparar sempre el seu rendiment amb limma.

6 Discussió

El treball d'aquesta tesi doctoral està emmarcat dins d'un projecte que estudia els síndromes limfoproliferatius crònics de cèl·lula B (B-CLPD), un grup de neoplàsies hematològiques que inclou diverses entitats que s'originen a les cèl·lules B. El diagnòstic de l'entitat es fa, habitualment, mitjançant l'estudi de la informació derivada de la morfologia, la citometria de flux i la citogenètica dels limfòcits de la sang perifèrica [1,8–10]. Tot i així, en un 15% dels pacients aquestes informacions no són suficients per establir un diagnòstic precís (B-CLPD, NOS), principalment per tres motius: *i*) la manca d'un tret específic robust de cada entitat [157], *ii*) l'heterogeneïtat biològica amb superposició de característiques genètiques i immunofenotípiques [9,158], i *iii*) la manca de mostres histològiques per analitzar.

En aquesta tesi s'ha investigat si l'expressió gènica, juntament amb altres característiques moleculars i genètiques, pot facilitar el diagnòstic dels pacients, una capacitat que s'ha demostrat en altres estudis de discriminació d'entitats de neoplàsies limfoides i tumors sòlids [121,159–161]. Les dades provenen de pacients que tenen afectació per algun d'aquests síndromes a la sang perifèrica. En l'apartat 1.1 s'ha explicat que els B-CLPD no sempre afecten tots els sistemes del sistema immunològic (sang, medul·la òssia i altres teixits limfoides), sinó que poden arribar a afectar-ne només un. Aleshores, els resultats obtinguts són especialment interessants per ajudar en el diagnòstic d'aquells pacients que tenen exclusivament afectació en sang perifèrica. Amb l'objectiu d'afegir una nova capa d'informació, s'ha estudiat el DNA i l'RNA en mostres

de sang provinents de pacients diagnosticats de nou entitats diferents de B-CLPD (CLL, cMCL, HCL, FL, nnMCL, HCLv, LPL, SDRPL, SMZL), el qual no s'ha fet de manera simultània en cap altre estudi.

El primer enfocament utilitzat per tal de construir un model diagnòstic en sang ha sigut el de combinar la informació obtinguda mitjançant dues plataformes de *microarrays* diferents, una que mesura l'expressió gènica i una altra que identifica les alteracions cromosòmiques (apartat 4.3). En el procés de construcció d'aquest model s'ha observat com les dades d'expressió tenen un gran poder discriminant, on, amb uns pocs gens, s'han pogut diferenciar clarament algunes de les entitats. Per altra banda, les dades de *copy-number* no han mostrat perfils d'alteracions tan específics i discriminants de les entitats. A més a més, aquesta segona font d'informació (*copy-number*) no ha sigut capaç de complementar la primera (expressió) en les entitats que no estaven tant ben discriminades per aquesta. Malauradament, no s'ha pogut disposar de suficients mostres per avaluar totes les entitats, però és raonable extrapolar les conclusions extretes a tots els subtipus de B-CLPD. Concretament, s'ha observat com les entitats amb més informació en una plataforma (per exemple, CLL) també són les que tenen més informació en l'altra.

És rellevant remarcar que el model integrador construït té dos inconvenients. El primer és que l'estructura de les dades de *copy-number* és complexa, es tracta de variables binàries on el percentatge de casos alterats per la majoria d'alteracions és relativament baix. A més a més, podrien existir estructures multivariants rellevants, com per exemple, l'associació de l'alteració 13q14.3- amb CLL quan aquesta es troba com a alteració única [5]. Aleshores, el *kernel* lineal utilitzat podria no ser el més adequat per recollir aquesta estructura. El segon inconvenient és que el model d'integració intermèdia utilitzat assumeix que no hi ha correlació entre les dues fonts d'informació. A la Figura 4.4 s'ha pogut veure que la correlació entre les dues fonts no és molt forta, però tampoc és inexistent. Degut al baix nombre de mostres no es poden estimar de manera fiable estructures complexes, per tant, intentar adreçar aquests dos inconvenients augmentaria en gran mesura el risc d'*overfitting*. En una sèrie més gran es podrien explorar diferents funcions *kernel*, així com permetre correlacions parcials, és a dir, es podrien utilitzar

estructures que estimessin la correlació entre les alteracions i les expressions dels gens continguts en les mateixes localitzacions.

Tot i aquests dos inconvenients, la metodologia utilitzada ha permès avaluar el poder predictiu de la combinació de les dues fonts. Es podria argumentar que una estratègia d'integració prematura també ho permetria, però al tenir unes estructures tant diferents hi hauria un gran risc de beneficiar una font per davant de l'altra. En aquest cas, es beneficiaria l'expressió al disposar de moltes més variables numèriques. Un dels objectius més importants d'aquesta tesi era quantificar la millora en la precisió del diagnòstic quan s'utilitzen dades de les dues tecnologies (*microarrays* d'expressió i *microarrays* de *copy-number*), atès que utilitzar les dues duplica el cost econòmic. La metodologia utilitzada ha permès quantificar la millora en un entorn en què les dues fonts estan en igualtat de condicions. La principal conclusió d'aquesta anàlisi és que, per construir un predictor capaç de diferenciar les diverses entitats de B-CLPD, és suficient amb les dades d'expressió, donat que les dades de *copy-number* no tenen un gran poder predictiu independent de l'expressió.

Els resultats del model integrador han motivat la construcció d'un model basat només en *microarrays* d'expressió, on, a més, s'ha pogut utilitzar el conjunt total dels casos de la cohort *training*. En la construcció d'aquest predictor s'ha proposat l'ús d'una metodologia per passos (*multi-step*) per davant de metodologies multiclasse o un-vs-tots. En l'apartat 4.4.3 s'ha vist com l'estratègia *multi-step* ha facilitat la identificació de gens específics de cada entitat, així com ha permès el poder minimitzar el nombre de gens utilitzats pel predictor de manera senzilla i sense perjudicar-ne la precisió. En l'entorn de la genètica, on generalment es disposa de poques mostres i moltes variables, la identificació de gens específics és important per tal d'aconseguir un model amb interpretació biològica i, a més, ajuda a contrastar la informació amb la literatura, un requisit important per tal d'evitar l'*overfitting*. El predictor final, el qual s'ha construït amb l'expressió de només 55 gens dels 20546 inicials, ha sigut capaç de discriminar sis de les entitats. Les tres entitats confoses (LPL, SDRPL i SMZL) s'han categoritzat com a *Miscellaneous*. En conclusió, una metodologia *multi-step* és adequada quan es volen identificar signatures específiques de les diferents classes, contrastar la informació a la

literatura i minimitzar el nombre de gens inclosos en el predictor.

En les dades de *microarrays* no ha sigut possible disposar d'una sèrie de validació externa independent, el qual, combinat amb el limitat nombre de mostres d'algunes de les entitats en el *training set*, no ha permès estimar l'error del predictor mitjançant una estratègia de validació externa ni tampoc mitjançant una d'interna [51], com és la *cross-validation* explicada a l'apartat 3.5. A més, les sensibilitats i especificitats presentades a la Taula 4.1 no es poden utilitzar com a estimacions de l'error, ja que, en cas de fer-ho, patirien de biaix degut a que no s'han calculat tenint en compte tot el procés de creació del predictor (apartat 3.5).

Fins i tot si es volguessin estimar els errors de les entitats amb major nombre de casos, s'ha de diferenciar que utilitzar la *cross-validation* per optimitzar un paràmetre és un problema diferent a utilitzar-la per estimar l'error esperat del predictor. En aquest segon cas s'han d'incloure en el procés de la *cross-validation* tots els passos per construir el predictor, és a dir, en cada *fold* del procés de *cross-validation* també s'haurien de tenir en compte aspectes com:

- El filtratge de *probesets*. Aquest pas depèn de les dades ja que es seleccionen *probesets* en base a l'IQR.
- L'enfocament per construir el predictor. L'enfocament *multi-step* no ha sigut l'única metodologia considerada, sinó que s'ha comparat amb altres, els resultats de les quals també depenen de les dades.
- L'ordre de discriminació de les entitats també s'ha fet en base a les dades.
- La selecció de Δ.

Aquest fet provoca que, independentment de la quantitat de mostres disponibles, sigui molt complex utilitzar la *cross-validation* per estimar l'error del predictor quan la construcció d'aquest no s'ha preespecificat a priori. De tota manera, utilitzar una sèrie de validació independent és la metodologia més recomanable per diversos motius [51,106]:

- Les estimacions de l'error mitjançant *cross-validation* no pateixen de biaix, però sí que pateixen de variància. L'error estimat podria ser força més baix o força

més elevat que el real. Aquest fet és especialment greu quan el nombre de mostres és baix [105].

- Qualsevol metodologia de validació interna (*split*, *cross-validation*, *bootstrap*, ...) té fonts de biaix potencials que poden afectar l'estimació, com, per exemple, la manipulació de les mostres de sang o els reactius utilitzats en els experiments.
- Els pacients inclosos en l'estudi podrien no representar correctament la diversitat real de la població en la que s'utilitzaria el predictor.

Tot i que no s'ha pogut disposar de sèrie de validació externa en *microarrays*, sí que se n'ha pogut disposar pel predictor construït amb les dades de qPCR. El predictor de *microarrays* i el de qPCR estan molt correlacionats, donat que els gens que utilitza el segon són gens que s'han identificat com rellevants en les dades del primer. A la Figura 4.15 s'ha pogut veure l'elevada correlació entre les dues tecnologies. Per aquest motiu, validar el predictor de qPCR també valida el de *microarrays* i, per tant, no és necessari mesurar l'expressió en una segona cohort mitjançant *microarrays*, el qual incrementaria el cost econòmic en gran mesura i innecessàriament.

El primer pas per tal de construir un predictor de qPCR que es pugui implementar a la rutina clínica és el de seleccionar, en base a la informació dels *microarrays* d'expressió, un subconjunt de gens a traslladar a qPCR. La metodologia de selecció de gens utilitzada en aquesta tesi ha tingut en compte diferents aspectes (apartat 4.5): la rellevància biològica (FC), la rellevància estadística (*P*-valor) i la rellevància multivariant (*scores* de Dziuda). Les rellevàncies biològica i estadística serveixen per maximitzar la probabilitat de que es reculli la mateixa informació en les dues plataformes [32–35], és a dir, que les diferències observades entre les entitats en les dades de *microarrays* es mantinguin en les dades de qPCR. La minimització de la redundància (rellevància multivariant) és especialment important quan el nombre de gens a seleccionar és limitat, donat que incloure un gen redundant en el petit subconjunt de qPCR té un gran impacte en la quantitat d'informació independent que es pot aconseguir en aquesta plataforma. Tot i així, s'han prioritzat els criteris de rellevància biològica i rellevància estadística per davant del criteri de redundància multivariant,

donat que estimar l'estructura de correlació en entorns d'alta dimensionalitat és poc robust i té risc d'*overfitting*. A més, s'ha utilitzat la mateixa estratègia *multi-step* que en el predictor de *microarrays*, fet que ha facilitat la selecció balancejada entre entitats i el càlcul dels tres paràmetres (FC, *P*-valor, *scores* de Dziuda) per cada gen.

Un cop seleccionats els 35 gens rellevants en les dades de *microarrays*, se n'ha mesurat l'expressió mitjançant qPCR. La majoria dels gens amb poder discriminant en els *microarrays* l'han mantingut en la qPCR, fet que suggereix que tenir en compte la rellevància estadística i la biològica pot ajudar a mantenir les diferències en aquesta plataforma. Tot i que, degut al baix nombre de mostres, és difícil avaluar si els *scores* multivariants milloren la redundància en qPCR, les dues parelles de gens utilitzades en el predictor de qPCR, {*FMOD*, *KSR2*} per CLL i {*CXCR4*, *CAMSAP2*} per HCLv, no han mostrat correlacions elevades. Es pot concloure que una metodologia de selecció de gens que té en compte els diferents criteris pot contribuir a maximitzar la reproductibilitat de la informació al canviar de plataforma.

La metodologia proposada per construir el predictor en dades de qPCR ha estat condicionada per dos factors: el primer, la limitada quantitat de mostres en el *training set* d'aquestes dades, i el segon, que el predictor resultant fos el més simple possible per facilitar-ne la implementació a nivell clínic. La metodologia, basada en *cutoffs*, ha evitat l'estimació de molts paràmetres i ha simplificat l'aplicació del predictor.

El predictor final de qPCR, tot i només incloure 8 gens, ha obtingut bons resultats en distingir les entitats CLL, cMCL, FL, nnMCL i l'agrupació *Miscellaneous* (LPL, SDRPL i SMZL) en la sèrie de validació. Malauradament, no s'ha pogut disposar de mostres de les entitats HCL, HCLv i SDRPL en aquesta sèrie. A part de ser poc freqüents en la població, els pacients diagnosticats amb aquestes entitats no acostumen a tenir afectació en sang perifèrica, dificultant, encara més, l'obtenció de casos suficients per valorar de manera fiable la precisió del predictor. Tot i que abans d'implementar aquesta eina a nivell clínic farien falta unes cohort *training* i de validació més grans, es pot concloure que és factible construir un predictor, senzill d'aplicar i basat en expressió gènica, que ajudi en el diagnòstic de les diferents entitats de B-CLPD.

El procediment que s'ha utilitzat per seleccionar els 35 gens que es traslladen a qPCR té l'inconvenient de ser força complex, tant en els càlculs, on s'han de calcular diversos paràmetres amb diferents metodologies, com en la valoració dels resultats, on s'han de valorar conjuntament aquests paràmetres per decidir quins gens es seleccionen finalment. Addicionalment, el càlcul dels *scores* de Dziuda inclou passos difícils d'automatitzar i requereix un temps computacional elevat. Per aquests motius, en el capítol 5 s'ha proposat lassoVoting, una metodologia més senzilla d'aplicar i que combina la rellevància biològica, l'estadística i la multivariant.

LassoVoting pot, a través dels paràmetres r i T_{FC}, donar més o menys pes a cadascuna d'aquestes rellevàncies. Fixant el valor dels paràmetres, s'ha comparat el rendiment de la metodologia proposada amb altres mètodes de selecció de variables en escenaris simulats i en dades reals. En els escenaris simulats s'ha pogut veure que lassoVoting és adequat per seleccionar entre 3 i 10 variables quan l'estructura de correlacions està ordenada per blocs, mentre que limma és superior quan la correlació no està estructurada o es volen seleccionar més de 15 variables. En els diferents conjunts de dades reals, limma ha obtingut errors estimats generalment per sota de la resta de mètodes, el qual suggereix que en dades reals l'estructura de correlació no és tant ordenada. Tot i així, en quatre conjunts (SLPC CLL, *Colon*, *Prostate* i *Sclerosis*) lassoVoting ha minimitzat l'error per alguna $k \geq 3$. En general, quan limma no ha sigut el millor mètode, ho ha sigut lassoVoting.

Es poden extreure tres conclusions de la comparació de mètodes. La primera, en entorns d'alta dimensionalitat és important tenir en compte la rellevància univariant, atès que la multivariant és susceptible a l'*overfitting*. La segona, en situacions en què les variables estan estructurades, les metodologies multivariants poden augmentar la quantitat d'informació si el nombre de variables a seleccionar és limitat. La última, lassoVoting és una metodologia adequada en aquestes situacions. Tot i així, faria falta una anàlisi més extensiva del rendiment de lassoVoting en altres conjunts de dades reals i simulats, així com estudiar l'efecte que tenen els paràmetres del mètode i la quantitat de mostres en el *training set* al seu rendiment.

Els dos predictors construïts (*microarrays* i qPCR), no solament serveixen per facilitar el diagnòstic en casos típics d'entitats ben definides, sinó que també serveixen per classificar aquells casos que pels criteris habituals no es poden diagnosticar (B-CLPD, NOS). En la Figura 4.10 s'ha pogut veure que alguns pacients categoritzats com B-CLPD, NOS tenen uns perfils d'expressió clarament concordants amb l'entitat CLL. Diversos d'aquests pacients tenen una translocació entre la banda q32 del cromosoma 13 i la banda q21 del cromosoma 18 (resumida com t(14;18)(q32;q21)) o la translocació t(14;19)(q32;q13), les quals poden provocar que la morfologia, l'immunofenotip i les característiques genètiques mostrin trets atípics en pacients de CLL [162–164]. En canvi, l'expressió gènica els ha pogut identificar clarament, el qual facilitaria el diagnòstic d'aquestes CLL atípiques. En la mateixa figura es pot veure un B-CLPD, NOS amb un perfil concordant amb l'entitat cMCL. Aquest pacient no presenta l'alteració t(11;14)(q13;q32), què és un dels criteris convencionals indispensables en el diagnòstic de les entitats cMCL i nnMCL. Els cMCL han de rebre tractament agressiu des del moment del diagnòstic, per tant, diagnosticar bé aquesta entitat és especialment important. De la mateixa manera, el petit percentatge de pacients d'FL que no tenen la t(14;18)(q32;q21) o afectació als ganglis són de difícil diagnòstic, però en la sèrie de B-CLPD, NOS de qPCR un pacient ha presentat un perfil concordant amb aquesta entitat. En conclusió, l'expressió gènica es manté constant en casos que presenten característiques atípiques i, per tant, els predictors de qPCR o de *microarrays* es poden utilitzar en els B-CLPD, NOS.

Tant en *microarrays* com en qPCR no s'han pogut distingir les entitats LPL, SDRPL i SMZL entre si, tal com mostren la Figura 4.8, la Figura 4.12 i els resultats de la sèrie de validació. Aquest resultat és paral·lel a l'observat en citologia i immunofenotip [132]. Per aquest motiu, s'ha conclòs que l'expressió gènica no conté informació per diferenciar-les i s'han agrupat en la categoria *Miscellaneous*.

Tot i que l'expressió gènica no pot distingir les entitats *Miscellaneous* entre si, sí que es pot fer servir per delimitar el diagnòstic a només aquestes tres entitats. Un cop delimitades, s'han utilitzat trets moleculars i genètics descrits en la literatura per ajudar en el diagnòstic. Per exemple, mutacions en el gen *MYD88* suggereixen el diagnòstic

d'LPL, mentre que mutacions de *NOTCH2* o delecions de 7q suggereixen el diagnòstic d'SMZL. És important remarcar que les mutacions de *MYD88* i *NOTCH2* no són exclusives d'LPL i SMZL, sinó que també es poden identificar en un percentatge petit de pacients amb CLL, cMCL, FL o nnMCL [5,140,165]. De la mateixa manera, la deleció de 7q també es pot identificar en altres entitats. Aleshores, al descartar les entitats CLL, cMCL, HCL, FL, nnMCL i HCLv en base a l'expressió gènica, aquests trets serveixen per diagnosticar les entitats LPL i SMZL. La Figura 4.20A resumeix com es combina aquesta informació amb la d'expressió gènica. En conclusió, l'expressió gènica juntament amb trets moleculars i genètics poden refinar el diagnòstic de les entitats *Miscellaneous* (LPL, SDRPL i SMZL).

En tots 64 B-CLPD, NOS inclosos en aquesta tesi se'ls hi ha pogut assignar, o almenys delimitar, una entitat específica. En concret, 29 (45%) s'han pogut assignar a una entitat exclusivament en base a l'expressió, 14 (22%) en base a combinar l'expressió amb trets moleculars i/o genètics, i en els 21 (33%) restants només s'han pogut descartar diverses entitats. En conclusió, l'expressió gènica juntament amb trets moleculars i genètics poden facilitar el diagnòstic dels pacients que mitjançant criteris convencionals no és possible.

En els apartats 4.4, 4.5 i 4.6 s'ha vist que l'entitat HCLv té un perfil d'expressió diferenciat de la resta, un resultat interessant ja que està considerada com una entitat provisional en la classificació de la WHO [1]. A més a més, distingir aquesta entitat de la resta és crític, atès que el tractament dels pacients és molt diferent [166]. Els resultats d'aquesta tesi suggereixen que l'expressió gènica podria diagnosticar correctament els pacients d'HCLv, una entitat que, per altres criteris, és difícil distingir-la d'SDRPL o HCL, però faria falta una sèrie més gran de casos per confirmar el resultat.

En resum, en aquesta tesi s'ha proposat una metodologia de construcció de predictors que té en compte les dades i el coneixement biològic previ. Els resultats d'aquests predictors ressalten que la combinació de l'expressió amb trets moleculars i genètics addicionals podria millorar el diagnòstic de les entitats de B-CLPD, especialment en els casos atípics.

7 Conclusions i futures línies de treball

En els dos primers apartats d'aquest capítol es presenten les conclusions extretes al llarg d'aquesta tesi. En el primer es llisten les conclusions sobre el refinament del diagnòstic dels B-CLPD, i en el segon es llisten les conclusions extretes sobre les diverses metodologies estadístiques emprades. El tercer apartat llista les futures línies de treball.

7.1 Conclusions sobre el refinament del diagnòstic dels B-CLPD

- Els *microarrays* de *copy-number* no tenen una aportació molt significativa en el diagnòstic de les diferents entitats de B-CLPD. Recollir informació de les dues plataformes de *microarrays* (expressió i *copy-number*) pot resultar redundant.

- S'ha vist, a través de la construcció dels predictors en dades de *microarrays* i de qPCR, que l'expressió gènica en sang conté molta informació sobre discriminació d'entitats de B-CLPD. En concret:
  - S'ha provat que és capaç de distingir les entitats CLL, cMCL, nnMCL i FL de la resta.
  - Suggereix que podria distingir HCL i HCLv de la resta, però fa falta una sèrie de validació adequada per confirmar-ho.
  - No és capaç de distingir les entitats *Miscellaneous* (LPL, SDRPL i SMZL) entre elles.

- El predictor simple de 8 gens mesurats mitjançant qPCR és capaç de distingir diverses entitats, i la seva senzillesa permetria que s'implementés a la rutina clínica.

- Un cop descartades les entitats CLL, cMCL, HCL, FL, nnMCL i HCLv per expressió, es poden utilitzar altres trets moleculars i genètics (mutacions de NOTCH2, mutacions de MYD88, deleció de 7q i IgM *paraportein*) per refinar la discriminació de les entitats LPL, SDRPL i SMZL entre si.

- Els pacients considerats B-CLPD, NOS tenen perfils d'expressió en sang coincidents amb alguna entitat. Per tant, els predictors basats en expressió es poden fer servir per afegir una nova capa d'informació per al diagnòstic d'aquests pacients.

7.2 Conclusions sobre metodologies estadístiques

- Quan es disposa de múltiples fonts d'informació poc correlacionades entre si, l'estratègia d'integració intermèdia utilitzada permet avaluar la capacitat de discriminació que té cadascuna per si mateixa, així com la capacitat de discriminació complementària entre les fonts.

- En situacions on es volen discriminar més de dues classes, utilitzar una estratègia per passos (*multi-step*) té diversos avantatges:
  - Facilita la identificació de gens específics de cada classe.
  - Facilita l'assoliment d'un bon compromís entre el nombre de gens inclosos en el predictor i l'error de predicció.
  - Facilita la selecció balancejada de gens específics de cada classe.
  - Concretament en l'entorn dels B-CLPD, aquesta estratègia ha facilitat el contrast de la informació en la literatura.

- En cas de construir dos predictors, un utilitzant dades de *microarrays* i un utilitzant dades de qPCR, no sempre és indispensable validar el predictor en *microarrays*. Quan les dues tècniques estan fortament correlacionades, com és el cas d'aquesta tesi, l'error de predicció del predictor en qPCR aproxima l'error de predicció del predictor en *microarrays*.

- Tenir en compte la rellevància estadística, la biològica i la multivariant en la selecció de gens a traslladar a qPCR pot ajudar a maximitzar la informació en aquesta plataforma.

- És sabut que quan es construeix un predictor en entorns d'alta dimensionalitat (poques mostres, moltes variables), no hi ha una estratègia de selecció de gens amb millor rendiment en general, sinó que depèn de la situació. Si el nombre de mostres és baix ($n \approx 20$), es pot concloure que:
  - Una estratègia univariant com limma té un bon rendiment quan la quantitat de gens a seleccionar no està limitada, independentment de l'estructura de correlació de les variables predictores.
  - Una estratègia univariant com limma té un bon rendiment quan sí que hi ha un límit i l'estructura de correlació no està organitzada per blocs.
  - Una estratègia que té en compte l'estructura de correlació, com és lassoVoting, té un bon rendiment quan sí hi ha un límit i l'estructura de correlació està organitzada per blocs.

- S'ha provat que el mètode lassoVoting funciona en situacions potencialment comunes.

- LassoVoting és capaç de ponderar la importància relativa que es vol donar a les tres rellevàncies (biològica, estadística i multivariant).

7.3 Futures línies de treball

- Provar el predictor basat en dades de qPCR en una sèrie de validació amb suficients mostres de les entitats HCL i HCLv.

- Estudiar extensivament l'efecte dels paràmetres de lassoVoting en el seu rendiment, així com l'efecte de la grandària mostral (n) i la quantitat de variables (p).

- Comparar el rendiment de lassoVoting amb el d'altres mètodes en més conjunts de dades reals.

- Construir una *Shiny app* que faciliti la utilització de lassoVoting.

Bibliografia

[1] Swerdlow SH et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. Lyon, France: IARC Press; 2008.

[2] Hodgkin T. On some morbid appearances of the absorbent glands and spleen. *Med Chir Trans*. 1832;17:68–114.

[3] Schmitt MW, Prindle MJ, Loeb LA. Implications of genetic heterogeneity in cancer. *Ann N Y Acad Sci*. 2012;1267(1):110–6.

[4] Martinez D et al. NOTCH1 , TP53 , and MAP2K1 mutations in splenic diffuse red pulp small B-cell lymphoma are associated with progressive disease. *Am J Surg Pathol*. 2015;0(0):1–10.

[5] Puente XS et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519–24.

[6] Rosenwald A et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*. 2003;3(2):185–97.

[7] Döhner H et al. Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med*. 2000;343(26):1910–6.

[8] Healey R, Naugler C, de Koning L, Patel JL. A classification tree approach for improving the utilization of flow cytometry testing of blood specimens for B-cell non-Hodgkin lymphoproliferative disorders. *Leuk Lymphoma*. 2015;8194(October 2014):1–6.

[9] Zhang Q-Y et al. A retrospective study to assess the relative value of peripheral blood, bone marrow aspirate and biopsy morphology, immunohistochemical stains, and flow cytometric analysis in the diagnosis of chronic B cell lymphoproliferative neoplasms. *Int J Lab Hematol*. 2015;37(3):390–402.

[10] Dronca RS et al. CD5-positive chronic B-cell lymphoproliferative disorders: diagnosis and prognosis of a heterogeneous disease entity. *Cytom Part B - Clin Cytom*. 2010;78(SUPPL. 1).

[11] Watson J, Crick F. Molecular structure of nucleic acids. *Nature*. 1953;171:737–8.

[12] Alberts B et al. *Molecular Biology of the Cell*. New York, USA: Garland Science; 2014.

[13] Trevino V, Falciani F, Barrera-Saldana HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med*. 2007;13(9–10):527–41.

[14] Taub F, DeLeo J, Thompson E. Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. *Dna*. 1983;2(4):309–27.

[15] Bumgarner R. Overview of DNA microarrays: Types, applications, and their future. *Curr Protoc Mol Biol*. 2013;

[16] Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. *Genome Res*. 1996;6(10):986–94.

[17] Kubista M et al. The real-time polymerase chain reaction. Vol. 27, Molecular Aspects of Medicine. 2006.

[18] Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*. 1967;62(318):626–33.

[19] Holm S. A simple sequential rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65–70.

[20] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289–300.

[21] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*. 2003;18(1):71–103.

[22] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, USA: Springer; 2009.

[23] Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995;20(3):273–97.

[24] James G, Witten D, Hastie T, Tibishirani R. *An Introduction to Statistical Learning*. New York, USA: Springer; 2013.

[25] Bernard PS et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160–7.

[26] Queirós AC et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*. 2015;29(3):598–605.

[27] Montraveta A et al. CD69 expression potentially predicts response to bendamustine and its modulation by ibrutinib or idelalisib enhances cytotoxic effect in chronic lymphocytic leukemia. *Oncotarget*. 2016;7(5):5507–20.

[28] van der Heijden AG et al. A five-gene expression signature to predict progression in T1G3 bladder cancer. *Eur J Cancer*. 2016;64:127–36.

[29] Scott DW et al. New molecular assay for the proliferation signature in mantle cell lymphoma applicable to formalin-fixed paraffin-embedded biopsies. *J Clin Oncol*. 2017;JCO.2016.70.7901.

[30] Watanabe T et al. Prediction of response to preoperative chemoradiotherapy in rectal cancer by using reverse transcriptase polymerase chain reaction analysis of four genes. *Dis Colon Rectum*. 2014;57(1):23–31.

[31] Reis PP et al. A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer*. 2011;11(1):437.

[32] Morey JS, Ryan JC, Van Dolah FM. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online*. 2006;8(1):175–93.

[33] Chen Y, Gelfond J a L, McManus LM, Shireman PK. Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis. *BMC Genomics*. 2009;10:407.

[34] Dallas P et al. Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR - how well do they correlate?. *BMC Genomics*. 2005;6(1):59.

[35] Rajeevan MS, Vernon SD, Taysavang N, Unger ER. Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR. *J Mol Diagnostics*. 2001;3(1):26–31.

[36] McEvoy LM et al. Identifying novel hypoxia-associated markers of chemoresistance in ovarian cancer. *BMC Cancer*. 2015;15(1):547.

[37] Urquidi V et al. A candidate molecular biomarker panel for the detection of bladder cancer. *Cancer Epidemiol Biomarkers Prev*. 2012;21(12):2149–58.

[38] Carén H et al. Identification of epigenetically regulated genes that predict patient outcome in neuroblastoma. *BMC Cancer*. 2011;11(1):66.

[39] Wu Z. A review of statistical methods for preprocessing oligonucleotide microarrays. *Stat Methods Med Res*. 2009;18:533–41.

[40] Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array-CGH data. *Bioinformatics*. 2005;21(19):3763–70.

[41] Pabinger S et al. A survey of tools for the analysis of quantitative PCR (qPCR)

data. *Biomol Detect Quantif*. 2014;1(1):23–33.

[42] Rebrikov D V, Trofimov DY. Real-time PCR: A review of approaches to data analysis. *Appl Biochem Microbiol*. 2006;42(5):455–63.

[43] Caraguel CGB et al. Selection of a cutoff value for real-time polymerase chain reaction results to fit a diagnostic purpose: analytical and epidemiologic approaches. *J Vet Diagn Invest*. 2011;23:2–15.

[44] McCall MN, McMurray HR, Land H, Almudevar A. On non-detects in qPCR data. *Bioinformatics*. 2014;30(16):2310–6.

[45] McCall MN et al. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*. 2011;12(1):137.

[46] Kauffmann A, Huber W. Microarray data quality control improves the detection of differentially expressed genes. *Genomics*. 2010;95(3):138–42.

[47] Sisti D et al. Shape based kinetic outlier detection in real-time PCR. *BMC Bioinformatics*. 2010;11:186.

[48] Bar T, Muszta A. Kinetics quality assessment for relative quantification by real-time PCR. *Biotechniques*. 2005;39(3):333–40.

[49] Slawski M, Daumer M, Boulesteix A-L. CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*. 2008;9:439.

[50] Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1–26.

[51] Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*. 2005;23(29):7332–41.

[52] Arteaga-Salas JM et al. An overview of image-processing methods for Affymetrix GeneChips. *Brief Bioinform*. 2008;9(1):25–33.

[53] McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010;11(2):242–53.

[54] Irizarry RA et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.

[55] Bolstad BM, Irizarry R., Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185–93.

[56] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression

index computation and outlier detection. *Proc Natl Acad Sci U S A*. 2001;98(1):31–6.

[57] Tukey JW. *Exploratory Data Analysis*. Reading, Massachusetts, USA: Addison-Wesley; 1977.

[58] Huber PJ. *Robust Statistics*. New York, USA: Wiley; 1981.

[59] Affymetrix. Guide to probe logarithmic intensity error (PLIER) estimation. *Tech Note*. 2005;

[60] Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*. 2001;2(8):RESEARCH0032.

[61] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin*. 2002;12(1):111–39.

[62] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res*. 2011;39(Database issue):D52–D57.

[63] Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*. 2009;10(1):11.

[64] Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A*. 2010;107(21):9546–51.

[65] Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*. 2009;25(17):2149–56.

[66] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557–72.

[67] Dellinger AE et al. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res*. 2010;38(9).

[68] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2^-ΔΔCT Method. *Methods*. 2001;25:402–8.

[69] Rao X, Huang X, Zhou Z, Lin X. An improvement of the 2ˆ(-delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. *Biostat Bioinforma Biomath*. 2013;3(3):71–85.

[70] Hamid JS et al. Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics*. 2009;2009(1):869093-.

[71] Gomez-Cabrero D et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*. 2014;8 Suppl 2(2):I1.

[72] Weston J, Pavlidis P, Cai J, Grundy WN. Gene functional classification from heterogeneous data. *Proc Fifth Annu Int Conf Comput Mol Biol*. 2001;(212):1–11.

[73] Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *Ann Stat*. 2008;36(3):1171–220.

[74] Daemen A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. *Conf Proc IEEE Eng Med Biol Soc*. 2007;2007:5411–5.

[75] Daemen A et al. Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pac Symp Biocomput*. 2008;166–77.

[76] Daemen A et al. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*. 2009;1(4):39.

[77] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9(3):293–300.

[78] Ye J, Xiong T. SVM versus least squares SVM. *J Mach Learn Res - Proc Track*. 2007;2:644–51.

[79] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classication. *Natl Taiwan Univ*. 2016;

[80] Wright G et al. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A*. 2003;100(17):9991–6.

[81] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc*. 2002;97:77–87.

[82] Shieh GS, Jiang YC, Shih Y. Comparison of support vector machines to other classifiers using gene expression data. *Commun Stat - Simul Comput*. 2006;35(1):241–56.

[83] Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data Anal*. 2005;48(4):869–85.

[84] Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*. 2011;6(12).

[85] Lai C, Reinders MJT, van't Veer LJ, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*. 2006;7:235.

[86] Pang H, Tong T, Zhao H. Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*. 2009;65(4):1021–9.

[87] Zararsiz G et al. Diagonal discriminant analysis for gene-expression based tumor classification. *J Adv Inf Technol*. 2015;6(2):59–62.

[88] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*. 2002;99(10):6567–72.

[89] Jeanmougin M et al. Should we abandon the t-Test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS One*. 2010;5(9):1–9.

[90] Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*. 2006;7:359.

[91] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.

[92] Aroian LA. A study of R. A. Fisher's z distribution and the related F distribution. *Ann Math Stat*. 1941;12(4):429–48.

[93] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116–21.

[94] Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*. 2003;19(18):2448–55.

[95] Dziuda DM. *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*. New York, USA: Wiley; 2010.

[96] Hotelling H. A generalized T test and measure of multivariate dispersion. *Proc Second Berkeley Symp Math Stat Probab*. 1951;23–41.

[97] Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *Br J*

Math Stat Psychol. 1992;45(2):265–82.

[98] Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.

[99] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1–3):389–422.

[100] Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat*. 2007;1:519–37.

[101] Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc*. 1974;36(2):111–47.

[102] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Jt Conf Artif Intell*. 1995;14(12):1137–43.

[103] Molinaro A, Simon R, Pfeiffer R. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21(15):3301–7.

[104] Kim JH. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal*. 2009;53(11):3735–45.

[105] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification?. *Bioinformatics*. 2004;20(3):374–80.

[106] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *JNCI J Natl Cancer Inst*. 2003;95(1):14–8.

[107] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc*. 1983;78(382):316.

[108] Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *J Am Stat Assoc*. 1997;92(438):548.

[109] Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Stat Med*. 2007;26(29):5320–34.

[110] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*. 2002;99(10):6562–6.

[111] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079−2107.

[112] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91.

[113] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165–88.

[114] Bolstad BM et al. Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol*. 2004;60:25–58.

[115] Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236–44.

[116] Pollack JR et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*. 2002;99(20):12963–8.

[117] Huang J et al. Correlation between genomic DNA copy number alterations and transcriptional expression in hepatitis B virus-associated hepatocellular carcinoma. *FEBS Lett*. 2006;580(15):3571–81.

[118] Platzer P et al. Silence of chromosomal amplifications in colon cancer. *Cancer Res*. 2002;62(4):1134–8.

[119] Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat Sci*. 2003;18(1):104–17.

[120] McCarthy BA et al. A seven-gene expression panel distinguishing clonal expansions of pre-leukemic and chronic lymphocytic leukemia B cells from normal B lymphocytes. *Immunol Res*. 2015;63(1–3):90–100.

[121] Cornet E et al. Developing molecular signatures for chronic lymphocytic leukemia. *PLoS One*. 2015;10(6):e0128990.

[122] Jares P, Colomer D, Campo E. Molecular pathogenesis of mantle cell lymphoma. *J Clin Invest*. 2012;122(10):3416–23.

[123] Vegliante MC et al. SOX11 regulates PAX5 expression and blocks terminal B-cell differentiation in aggressive mantle cell lymphoma. *Blood*. 2013;121(12):2175–85.

[124] Pettirossi V et al. BRAF inhibitors reverse the unique molecular signature and phenotype of hairy cell leukemia and exert potent antileukemic activity. *Blood*. 2015;125(8):1207–16.

[125] Leich E et al. Similar clinical features in follicular lymphomas with and without breaks in the BCL2 locus. *Leukemia*. 2016;30(4):854–60.

[126] Wong S, Fulcher D. Chemokine receptor expression in B-cell lymphoproliferative disorders. *Leuk Lymphoma*. 2004;45(12):2491–6.

[127] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*. 2006;7(1):55–65.

[128] Falini B et al. Simple diagnostic assay for hairy cell leukaemia by immunocytochemical detection of annexin A1 (ANXA1). *Lancet*. 2004;363(9424):1869–71.

[129] Forconi F et al. Hairy cell leukemia: At the crossroad of somatic mutation and isotype switch. *Blood*. 2004;104(10):3312–7.

[130] Salaverria I et al. CCND2 rearrangements are the most frequent genetic events in cyclin D1 - mantle cell lymphoma. *Blood*. 2013;121(8):1394–402.

[131] Fan L et al. Expression patterns of CD200 and CD148 in leukemic B-cell chronic lymphoproliferative disorders and their potential value in differential diagnosis. *Leuk Lymphoma*. 2015;56(12):3329–35.

[132] Dufresne SD et al. Defining the borders of splenic marginal zone lymphoma: a multiparameter study. *Hum Pathol*. 2010;41(4):540–51.

[133] Matutes E et al. The immunological profile of B-cell disorders and proposal of a scoring system for the diagnosis of CLL. *Leukemia*. 1994;8(10):1640–5.

[134] Guo Y et al. Low-grade follicular lymphoma with t(14;18) presents a homogeneous disease entity otherwise the rest comprises minor groups of heterogeneous disease entities with Bcl2 amplification, Bcl6 translocation or other gene aberrances. *Leukemia*. 2005;19(6):1058–63.

[135] Tiacci E et al. BRAF mutations in hairy-cell leukemia. *N Engl J Med*. 2011;364(24):2305–15.

[136] Kiel MJ et al. Whole-genome sequencing identifies recurrent somatic NOTCH2 mutations in splenic marginal zone lymphoma. *J Exp Med*. 2012;209(9):1553–65.

[137] Salido M et al. Cytogenetic aberrations and their prognostic value in a series of 330 splenic marginal zone B-cell lymphomas: a multicenter study of the Splenic B-Cell Lymphoma Group. *Blood*. 2010;116(9):1479–88.

[138] Dierlamm J et al. Trisomy 3 in marginal zone B-cell lymphoma: a study based on cytogenetic analysis and fluorescence in situ hybridization. *Br J Haematol*. 1996;93(1):242–9.

[139] Gachard N et al. IGHV gene features and MYD88 L265P mutation separate the three marginal zone lymphoma entities and Waldenström

macroglobulinemia/lymphoplasmacytic lymphomas. *Leukemia*. 2013;27(1):183–9.

[140] Karube K et al. Recurrent mutations of NOTCH genes in follicular lymphoma identify a distinctive subset of tumours. *J Pathol*. 2014;234(3):423–30.

[141] Tibshirani R. Regression selection and shrinkage via the lasso. *J R Stat Soc B*. 1996;58(1):267–88.

[142] Efron B et al. Least angle regression. *Ann Stat*. 2004;32(2):407–99.

[143] Dietterich TG. Ensemble methods in machine learning. *Mult Classif Syst*. 2000;1857:1–15.

[144] Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.

[145] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1):119–39.

[146] Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*. 2013;14:5.

[147] Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. *Proc 25th Eur Conf Mach Learn Knowl Discov Databases, Part II*. 2008;313–25.

[148] Abeel T et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. 2009;26(3):392–8.

[149] Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med*. 1997;16(4):385–95.

[150] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B Stat Methodol*. 2006;68(1):49–67.

[151] Tibshirani R et al. Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B Stat Methodol*. 2005;67(1):91–108.

[152] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207–10.

[153] Alon U et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 1999;96(12):6745–50.

[154] Golub TR et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (80-)*. 1999;286(5439):531–7.

[155] Singh D et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002;1(2):203–9.

[156] Kemppinen AK, Kaprio J, Palotie A, Saarela J. Systematic review of genome-wide expression studies in multiple sclerosis. *BMJ Open*. 2011;1(1):e000053.

[157] Rossmann ED et al. Variability in B-cell antigen expression: implications for the treatment of B-cell lymphomas and leukemias with monoclonal antibodies. *Hematol J*. 2001;2(5):300–6.

[158] Gualco G, Natkunam Y, Bacchi CE. The spectrum of B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Hodgkin lymphoma: a description of 10 cases. *Mod Pathol*. 2012;25:661–74.

[159] Haferlach T et al. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group. *J Clin Oncol*. 2010;28(15):2529–37.

[160] Iqbal J et al. Gene expression signatures delineate biological and prognostic subgroups in peripheral T-cell lymphoma. *Blood*. 2014;123(19):2915–23.

[161] Piccaluga PP et al. Molecular profiling improves classification and prognostication of nodal peripheral T-cell lymphomas: results of a phase III diagnostic accuracy study. *J Clin Oncol*. 2013;31(24):3019–25.

[162] Huh YO et al. The t(14;19)(q32;q13)-positive small B-cell leukaemia: A clinicopathologic and cytogenetic study of seven cases. *Br J Haematol*. 2007;136(2):220–8.

[163] Martín-Subero JI et al. A comprehensive genetic and histopathologic analysis identifies two subgroups of B-cell malignancies carrying a t(14;19)(q32;q13) or variant BCL3-translocation. *Leukemia*. 2007;21(7):1532–44.

[164] Kelly RJ et al. t(14;19)(q32;q13) incidence and significance in B-cell lymphoproliferative disorders. *Br J Haematol*. 2008;141(4):561–3.

[165] Beà S et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc Natl Acad Sci U S A*. 2013;110(45):18250–5.

[166] Behdad A, Bailey NG. Diagnosis of splenic B-cell lymphomas in the bone marrow: A review of histopathologic, immunophenotypic, and genetic findings. *Arch Pathol Lab Med*. 2014;138(10):1295–301.

Annex A: Codi i funcions en R

Codi i funcions per calcular els scores del mètode de Dziuda

```
# especificació dels paràmetres del mètode (exemple)
M=50
m=3
Tcut=0.5
B=250
gOOB=0.2

# dades (exemple)
X=matrix(rnorm(20000),ncol=500)
Y=factor(rep(0:1,each=20))

# generar M biomarcadors de m variables
alt.biomark=T2.alternative.biomarkers(X,Y,M=M,m=m)

# identificar Informative Set of Genes
INF=plot.INF(alt.biomark,Tcut=Tcut)

# calcular scores Sall i Sinf de generar B Monte Carlo sets
scores=scores(X,Y,INF=INF,B=B,gOOB=gOOB)

# funcions utilitzades per aplicar el mètode de Dziuda

# funció que genera M biomarcadors

T2.alternative.biomarkers=function(X,Y,M=300,m=3){

        # X: matriu amb files=mostres, columnes=individus
        # Y: vector que indica la classe de cada fila d'X
        # M: quantitat de biomarcadors que es generen
        # m: quantitat de variables que s'inclouen a cada biomarcador (mínim 3)

        if (!class(X) %in% c("matrix","data.frame")){
                stop("X no és matrix o data.frame")
        }
        if (nrow(X) != length(Y)) stop("Dimensions de X i Y no corresponen")
        if (is.null(colnames(X))){
                colnames(X)=paste0("V",1:ncol(X))
        }
        if (is.null(rownames(X))){
                rownames(X)=paste0("S",1:nrow(X))
        }
        if (max(table(colnames(X)))>1) stop("X conté columnes amb nom repetit")
        if (max(table(rownames(X)))>1) stop("X conté files amb nom repetit")
        if (class(Y) != "factor"){
```

```
		Y=factor(Y)
	}
	if (m<3) stop("min(m)=3")

	aux.X=t(X)
	var.biomark=matrix(nrow=M,ncol=m)
	T2v=numeric()

	# generar biomarcadors de forma iterativa (excloent variables)
	for (k in 1:M){
		Stepwise=stepwise.T2(aux.X,Y,m=m)
		var.biomark[k,]=Stepwise$v
		T2v[k]=Stepwise$TH
		aux.X=aux.X[!(rownames(aux.X) %in% Stepwise$v),]
	}
	return(data.frame(var.biomark,T2=T2v))
}

# funció que representa l'estadístic T^2 dels M biomarcadors i retorna
# l'Informative Set of Genes segons el paràmetre Tcut

plot.INF=function(alt.biomark,Tcut=2.5){

	# alt.biomark: objecte resultant de la funció T2.alternative.biomarkers
	# Tcut: valor mínim de T^2 per a que el biomarcador pugui ser seleccionat

	T2=alt.biomark$T2

	# ajust del model logarítmic
	x=1:length(T2)
	fit=lm(T2~log(x))
	tall=(exp((Tcut-coef(fit)[1])/coef(fit)[2]))

	# gràfic
	par(mar=c(4,4,3,5))
	plot(x,T2,ylab="",xlab="",pch=19,font.axis=2,
	    col=c("grey60","blue")[((x<tall) & (T2>=Tcut))+1])
	title(xlab="Iteració del biomarcador",font.lab=2,line=2.5,cex.lab=1.1)
	title(ylab=expression(italic(T^2)),line=2.5,cex.lab=1.25)
	abline(h=Tcut,lty=2)
	lines(x,predict(fit),lwd=2)
	abline(v=tall, lty=2)
	axis(3,at=tall,round(tall,2))
	axis(4,at=Tcut,Tcut,las=2)

	# seleccio de de l'Informative Set of Genes
	var.biomark=alt.biomark[,!colnames(alt.biomark) %in% "T2"]
	cond1=T2>Tcut
	cond2=(1:length(T2))<tall
	INF=var.biomark[cond1&cond2,]
	return(INF)

}

# funció que calcula els scores del mètode de Dziuda (Sall, Sinf)
```

```
scores=function(X,Y,INF,B=1000,gOOB=0.2){

	# X: matriu amb files=mostres, columnes=individus
	# Y: vector que indica la classe de cada fila d'X
	# INF: objecte resultant de la funció plot.INF (Informative Set of Genes)
	# B: quantitat de Monte Carlo (MC) sets que es generen de original/INF sets
	# gOOB: (1-gOOB)=percentatge de mostres de cada classe a cada MC set

	if (!class(X) %in% c("matrix","data.frame")){
		stop("X no és matrix o data.frame")
	}
	if (nrow(X) != length(Y)) stop("Dimensions de X i Y no corresponen")
	if (is.null(colnames(X))){
		colnames(X)=paste0("V",1:ncol(X))
	}
	if (is.null(rownames(X))){
		rownames(X)=paste0("S",1:nrow(X))
	}
	if (max(table(colnames(X)))>1) stop("X conté columnes amb nom repetit")
	if (max(table(rownames(X)))>1) stop("X conté files amb nom repetit")
	if (class(Y) != "factor"){
		Y=factor(Y)
	}

	m=ncol(INF)
	INF=c(as.matrix(INF))
	Xinf=X[,INF]

	# generar cada MC set i aplicar-hi l'stepwise selection a cada
	MC.inf=table(mod.bagging(em=t(Xinf),g=Y,B=B,m=m,gOOB=gOOB))
	MC.all=table(mod.bagging(em=t(X),g=Y,B=B,m=m,gOOB=gOOB))

	# càlcul dels scores

	score.all=numeric(ncol(X))
	names(score.all)=colnames(X)
	score.inf=score.all

	score.all[names(MC.all)]=MC.all/B
	score.inf[names(MC.inf)]=MC.inf/B

	return(data.frame(Sall=score.all,Sinf=score.inf))
}

# altres funcions utilitzades per T2.alternative.biomarkers, plot.INF i scores

# funció que fa l'Stepwise hybrid feature selection with T2

stepwise.T2=function(X,Y,m){

	# variable inicial agafada al atzar
	v=sample(rownames(X),1)
	TH=numeric(length=m)
	TH[1]=0

	# afegim segona variable que maximitza T2
```

```
	sort=afegir.variable(X,v,Y)
	v=sort[[1]]
	TH[length(v)]=sort[[2]]

	# mentre v no tingui m variables
	while (length(v)<m){

		# afegir variable
		sort=afegir.variable(X,v,Y)
		v=sort[[1]]
		TH[length(v)]=sort[[2]]

		# treure variable si millora el T2 del pas anterior
		v=treure.variable(X,v,TH,Y,m)
		TH[length(v)]=calcT2(t(X[v,]),Y)
	}

	object=list()
	object$v=v
	object$TH=TH[m]
	return(object)
}

# funció que afegeix a v la variable que maximitza T2

afegir.variable=function(X,v,Y){
	THi=numeric(length=nrow(X))
	names(THi)=rownames(X)
	for (i in (names(THi)[!names(THi) %in% v])){
		THi[i]=calcT2(t(X[c(i,v),]),Y)
	}
	v=c(v,names(which.max(THi)))
	return(list(v,max(THi,na.rm=TRUE)))
}

# funció que treu una variable a v si millora T2 anterior (tolerància de 4 decimals)

treure.variable=function(X,v,TH,Y,m){
	THt=numeric()
	for (i in 1:length(v)){
		THt[i]=calcT2(t(X[v[-i],]),Y)
	}
	if (round(max(THt-TH[length(v)-1]),4)>0){
		treure=which.max(THt-TH[length(v)])
		v=v[-treure]
	}
	return(v)
}

# funció que calcula l'estadístic T2

calcT2=function(x,y){
	lg=table(y)
	nc=ncol(x)
	E=matrix(0,ncol=nc,nrow=nc)
	for (i in 1:length(lg)){
```

```
                mat=x[y==levels(y)[i],]
                E=E+var(mat)*(nrow(mat)-1)
        }
        V=var(x)*(nrow(x)-1)
        H=V-E
        return(sum(diag(H%*%solve(E))))
}

# funció que genera B MC sets, i fa l'stepwise selection

mod.bagging=function(em,g,B,m,gOOB){
        selected=matrix(NA,nrow=B,ncol=m)
        for (i in 1:B){

                # selecció a l'atzar de (1-gOBB) mostres de cada classe
                aux=tapply(1:ncol(em),g,sample.gOOB,gOOB=gOOB)
                OOB=do.call(c,aux)
                g.sub=g[-OOB]
                em.sub=em[,-OOB]

                # stepwise selection
                MBS=stepwise.T2(em.sub,g.sub,m=m)
                selected[i,]=MBS$v
        }
        return(selected)
}

sample.gOOB=function(x,gOOB){
        q=floor(length(x)*gOOB+0.99)
        return(sort(sample(x,q)))
}
```

Codi i funcions per calcular l'*score* de lassoVoting

```
# paquets d'R necessaris
library(limma)
library(glmnet)

# especificació dels paràmetres del mètode (exemple)
B=250
pm=0.5
pn=0.8
r=80
Tfc=0

# dades (exemple)
X=matrix(rnorm(20000),ncol=500)
Y=factor(rep(0:1,each=20))

# càlcul score
score=lassoVoting(X,Y,B=B,pm=pm,pn=pn,r=r,Tfc=Tfc)

# funcions utilitzades per aplicar lassoVoting

# funció que calcula l'score de lassoVoting quan es comparen 2 classes

lassoVoting=function(X,Y,B=1000,pm=0.5,pn=0.8,r=80,Tfc=0){

	# X:	matriu amb columnes=variables i files=mostres
	# Y:	vector que indica la classe de cada fila d'X
	# pm:	percentatge de variables seleccionades a l'atzar d'X
	# pn:	percentatge de mostres seleccionades a l'atzar d'X
	# r:	nombre de variables filtrades per limma i del log2(FC)
	# Tfc:	valor mínim del log2(FC) d'una variable per a que passi el filtre

	if (!class(X) %in% c("matrix","data.frame")){
		stop("X no és matrix o data.frame")
	}
	if (nrow(X) != length(Y)) stop("Dimensions de X i Y no corresponen")
	if (is.null(colnames(X))){
		colnames(X)=paste0("V",1:ncol(X))
	}
	if (is.null(rownames(X))){
		rownames(X)=paste0("S",1:nrow(X))
	}
	if (max(table(colnames(X)))>1) stop("X conté columnes amb nom repetit")
	if (max(table(rownames(X)))>1) stop("X conté files amb nom repetit")
	if (class(Y) != "factor"){
		Y=factor(Y)
	}
	if (nlevels(Y)>2) stop("Y té més de 2 classes")

	S=matrix(0,nrow=ncol(X),ncol=B)
	F=matrix(0,nrow=ncol(X),ncol=B)
	rownames(S)=colnames(X)
	rownames(F)=colnames(X)
```

```
	for (b in 1:B){

		# selecció a l'atzar de mostres i variables
		inds=do.call(c,tapply(1:length(Y),Y,sample.pn,pn=pn))
		vars=sample(ncol(X),floor(ncol(X)*pm))
		bootX=X[inds,vars,drop=F]
		bootY=Y[inds]

		F[colnames(bootX),b]=1

		# limma
		limma=limma.function(t(bootX),bootY)
		limma=topTable(limma,number=ncol(bootX))

		# selecció segons P-value i Tfc
		nfc=sum(abs(limma$logFC)>=Tfc)
		if (nfc>=min(ncol(bootX),r)){
			filter.pass=rownames(limma[abs(limma$logFC)>=Tfc,])[1:r]
		} else {
			limma=limma[order(abs(limma$logFC),decreasing=T),]
			filter.pass=rownames(limma)[1:min(ncol(bootX),r)]
		}
		bootX=bootX[,filter.pass,drop=F]

		# repartició estratificada dels casos en els folds de lasso
		nfolds=min(min(table(bootY)),10)
		indicador.folds=do.call(rbind,tapply(1:length(bootY),bootY,
			repartir.folds,nfolds=nfolds))
		foldid=indicador.folds[order(indicador.folds[,1]),2]

		# lasso
		cvfit = cv.glmnet(as.matrix(bootX), bootY, family = "binomial",
			type.measure = "class",foldid=foldid,standardize=FALSE,
			alpha=1)

		# S[jb]
		auxS=coef(cvfit, s = "lambda.1se")[,1]
		auxS=auxS[!names(auxS) %in% "(Intercept)"]
		S[names(auxS),b]=as.numeric(abs(auxS)>0)
	}

	scores=sort(rowSums(S)/rowSums(F),decreasing=T)
	return(scores)

}

# altres funcions requerides per la funció lassoVoting

sample.pn=function(x,pn){
	q=floor(length(x)*pn)
	return(sort(sample(x,q)))
}

repartir.folds=function(x,nfolds){
	reps=floor(length(x)/nfolds)
```

```
        folds=rep(1:nfolds,each=reps)
        folds=sample(c(folds,sample(1:nfolds,length(x)-length(folds))))
        return(data.frame(x,folds))
}

limma.function=function(x,y){
        levels(y)=make.names(levels(y))
        design <-model.matrix(~0+y)
        colnames(design)=levels(y)
        contrat.matrix=makeContrasts(contrasts=paste(rev(levels(y)),collapse="-"),
                levels=levels(y))
        fit=lmFit(x,design)
        fit=contrasts.fit(fit,contrat.matrix)
        fit<-eBayes(fit)
        return(fit)
}
```

Annex B: Articles publicats en l'entorn dels B-CLPD

La següent llista conté altres publicacions, dins l'entorn dels B-CLPD, en què he estat involucrat durant el curs d'aquesta tesi:

1. Scott DW, Abrisqueta P, Wright GW, Slack GW, Mottok A, Villa D, Jares P, Rauert-Wunderlich H, Royo C, **Clot G**, Pinyol M, Boyle M, Chan FC, Braziel RM, Chan WC, Weisenburger DD, Cook JR, Greiner TC, Fu K, Ott G, Delabie J, Smeland EB, Holte H, Jaffe ES, Steidl C, Connors JM, Gascoyne RD, Rosenwald A, Staudt LM, Campo E, Rimsza LM; Lymphoma/Leukemia Molecular Profiling Project. New Molecular Assay for the Proliferation Signature in Mantle Cell Lymphoma Applicable to Formalin-Fixed Paraffin-Embedded Biopsies. *J Clin Oncol*. 2017 May 20;35(15):1668-1677.

2. Law PJ, Berndt SI, Speedy HE, Camp NJ, Sava GP, Skibola CF, Holroyd A, Joseph V, Sunter NJ, Nieters A, Bea S, Monnereau A, Martin-Garcia D, Goldin LR, **Clot G**, Teras LR, Quintela I, Birmann BM, Jayne S, Cozen W, Majid A, Smedby KE, Lan Q, Dearden C, Brooks-Wilson AR, Hall AG, Purdue MP, Mainou-Fowler T, Vajdic CM, Jackson GH, Cocco P, Marr H, Zhang Y, Zheng T, Giles GG, Lawrence C, Call TG, Liebow M, Melbye M, Glimelius B, Mansouri L, Glenn M, Curtin K, Diver WR, Link BK, Conde L, Bracci PM, Holly EA, Jackson RD, Tinker LF, Benavente Y, Boffetta P, Brennan P, Maynadie M, McKay J, Albanes D, Weinstein S, Wang Z, Caporaso NE, Morton LM, Severson RK, Riboli E, Vineis P, Vermeulen RC, Southey MC, Milne RL, Clavel J, Topka S, Spinelli JJ, Kraft P, Ennas MG, Summerfield G, Ferri GM, Harris RJ, Miligi L, Pettitt AR, North KE, Allsup DJ, Fraumeni JF, Bailey JR, Offit K, Pratt G, Hjalgrim H, Pepper C, Chanock SJ, Fegan C, Rosenquist R, de Sanjose

S, Carracedo A, Dyer MJ, Catovsky D, Campo E, Cerhan JR, Allan JM, Rothman N, Houlston R, Slager S. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat Commun*. 2017 Feb 6;8:14175.

3. Queirós AC, Beekman R, Vilarrasa-Blasi R, Duran-Ferrer M, **Clot G**, Merkel A, Raineri E, Russiñol N, Castellano G, Beà S, Navarro A, Kulis M, Verdaguer-Dot N, Jares P, Enjuanes A, Calasanz MJ, Bergmann A, Vater I, Salaverría I, van de Werken HJ, Wilson WH, Datta A, Flicek P, Royo R, Martens J, Giné E, Lopez-Guillermo A, Stunnenberg HG, Klapper W, Pott C, Heath S, Gut IG, Siebert R, Campo E, Martín-Subero JI. Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage. *Cancer Cell*. 2016 Nov 14;30(5):806-821.

4. Schmidt J, Gong S, Marafioti T, Mankel B, Gonzalez-Farre B, Balagué O, Mozos A, Cabeçadas J, van der Walt J, Hoehn D, Rosenwald A, Ott G, Dojcinov S, Egan C, Nadeu F, Ramis-Zaldívar JE, **Clot G**, Bárcena C, Pérez-Alonso V, Endris V, Penzel R, Lome-Maldonado C, Bonzheim I, Fend F, Campo E, Jaffe ES, Salaverria I, Quintanilla-Martinez L. Genome-wide analysis of pediatric-type follicular lymphoma reveals low genetic complexity and recurrent alterations of TNFRSF14 gene. *Blood*. 2016 Aug 25;128(8):1101-11.

5. Montraveta A, Lee-Vergés E, Roldán J, Jiménez L, Cabezas S, **Clot G**, Pinyol M, Xargay-Torrent S, Rosich L, Arimany-Nardí C, Aymerich M, Villamor N, López-Guillermo A, Pérez-Galán P, Roué G, Pastor-Anglada M, Campo E, López-Guerra M, Colomer D. CD69 expression potentially predicts response to bendamustine and its modulation by ibrutinib or idelalisib enhances cytotoxic effect in chronic lymphocytic leukemia. *Oncotarget*. 2016 Feb 2;7(5):5507-20.

6. Martinez D, Navarro A, Martinez-Trillos A, Molina-Urra R, Gonzalez-Farre B, Salaverria I, Nadeu F, Enjuanes A, **Clot G**, Costa D, Carrio A, Villamor N, Colomer D, Martinez A, Bens S, Siebert R, Wotherspoon A, Beà S, Matutes E,

Campo E. NOTCH1, TP53, and MAP2K1 Mutations in Splenic Diffuse Red Pulp Small B-cell Lymphoma Are Associated With Progressive Disease. *Am J Surg Pathol*. 2016 Feb;40(2):192-201.

7. Sebastián E, Alcoceba M, Martín-García D, Blanco Ó, Sanchez-Barba M, Balanzategui A, Marín L, Montes-Moreno S, González-Barca E, Pardal E, Jiménez C, García-Álvarez M, **Clot G**, Carracedo Á, Gutiérrez NC, Sarasquete ME, Chillón C, Corral R, Prieto-Conde MI, Caballero MD, Salaverria I, García-Sanz R, González M. High-resolution copy number analysis of paired normal-tumor samples from diffuse large B cell lymphoma. *Ann Hematol*. 2016 Jan;95(2):253-62.

8. Salaverria I, Martín-Garcia D, López C, **Clot G**, García-Aragonés M, Navarro A, Delgado J, Baumann T, Pinyol M, Martin-Guerrero I, Carrió A, Costa D, Queirós AC, Jayne S, Aymerich M, Villamor N, Colomer D, González M, López-Guillermo A, Campo E, Dyer MJ, Siebert R, Armengol L, Beà S. Detection of chromothripsis-like patterns with a custom array platform for chronic lymphocytic leukemia. *Genes Chromosomes Cancer*. 2015 Nov;54(11):668-80.

9. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, Munar M, Rubio-Pérez C, Jares P, Aymerich M, Baumann T, Beekman R, Belver L, Carrio A, Castellano G, **Clot G**, Colado E, Colomer D, Costa D, Delgado J, Enjuanes A, Estivill X, Ferrando AA, Gelpí JL, González B, González S, González M, Gut M, Hernández-Rivas JM, López-Guerra M, Martín-García D, Navarro A, Nicolás P, Orozco M, Payer ÁR, Pinyol M, Pisano DG, Puente DA, Queirós AC, Quesada V, Romeo-Casabona CM, Royo C, Royo R, Rozman M, Russiñol N, Salaverría I, Stamatopoulos K, Stunnenberg HG, Tamborero D, Terol MJ, Valencia A, López-Bigas N, Torrents D, Gut I, López-Guillermo A, López-Otín C, Campo E. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015 Oct 22;526(7574):519-24.

10. Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, Beekman R, Raineri E, Esteve A, **Clot G**, Verdaguer-Dot N, Duran-Ferrer M, Russiñol N, Vilarrasa-Blasi R, Ecker S, Pancaldi V, Rico D, Agueda L, Blanc J, Richardson D, Clarke L, Datta A, Pascual M, Agirre X, Prosper F, Alignani D, Paiva B, Caron G, Fest T, Muench MO, Fomin ME, Lee ST, Wiemels JL, Valencia A, Gut M, Flicek P, Stunnenberg HG, Siebert R, Küppers R, Gut IG, Campo E, Martín-Subero JI. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet*. 2015 Jul;47(7):746-56.

11. Queirós AC, Villamor N, **Clot G**, Martinez-Trillos A, Kulis M, Navarro A, Penas EM, Jayne S, Majid A, Richter J, Bergmann AK, Kolarova J, Royo C, Russiñol N, Castellano G, Pinyol M, Bea S, Salaverria I, López-Guerra M, Colomer D, Aymerich M, Rozman M, Delgado J, Giné E, González-Díaz M, Puente XS, Siebert R, Dyer MJ, López-Otín C, Rozman C, Campo E, López-Guillermo A, Martín-Subero JI. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*. 2015 Mar;29(3):598-605.

12. Delgado J, Salaverria I, Baumann T, Martínez-Trillos A, Lee E, Jiménez L, Navarro A, Royo C, Santacruz R, López C, Payer AR, Colado E, González M, Armengol L, Colomer D, Pinyol M, Villamor N, Aymerich M, Carrió A, Costa D, **Clot G**, Giné E, López-Guillermo A, Campo E, Beà S. Genomic complexity and IGHV mutational status are key predictors of outcome of chronic lymphocytic leukemia patients with TP53 disruption. *Haematologica*. 2014 Nov;99(11):e231-4.

13. Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, Djebali S, Papasaikas P, López-Guerra M, Colomer D, Royo C, Cazorla M, Pinyol M, **Clot G**, Aymerich M, Rozman M, Kulis M, Tamborero D, Gouin A, Blanc J, Gut M, Gut I, Puente XS, Pisano DG, Martin-Subero JI, López-Bigas N, López-Guillermo A, Valencia A, López-Otín C, Campo E, Guigó R.

Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res*. 2014 Feb;24(2):212-26.

14. Enjuanes A, Albero R, **Clot G**, Navarro A, Beà S, Pinyol M, Martín-Subero JI, Klapper W, Staudt LM, Jaffe ES, Rimsza L, Braziel RM, Delabie J, Cook JR, Tubbs RR, Gascoyne R, Connors JM, Weisenburger DD, Greiner TC, Chan WC, López-Guillermo A, Rosenwald A, Ott G, Campo E, Jares P. Genome-wide methylation analyses identify a subset of mantle cell lymphoma with a high number of methylated CpGs and aggressive clinicopathological features. *Int J Cancer*. 2013 Dec 15;133(12):2852-63.

15. Beà S, Valdés-Mas R, Navarro A, Salaverria I, Martín-Garcia D, Jares P, Giné E, Pinyol M, Royo C, Nadeu F, Conde L, Juan M, **Clot G**, Vizán P, Di Croce L, Puente DA, López-Guerra M, Moros A, Roue G, Aymerich M, Villamor N, Colomo L, Martínez A, Valera A, Martín-Subero JI, Amador V, Hernández L, Rozman M, Enjuanes A, Forcada P, Muntañola A, Hartmann EM, Calasanz MJ, Rosenwald A, Ott G, Hernández-Rivas JM, Klapper W, Siebert R, Wiestner A, Wilson WH, Colomer D, López-Guillermo A, López-Otín C, Puente XS, Campo E. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc Natl Acad Sci U S A*. 2013 Nov 5;110(45):18250-5.

16. Navarro A, **Clot G**, Prieto M, Royo C, Vegliante MC, Amador V, Hartmann E, Salaverria I, Beà S, Martín-Subero JI, Rosenwald A, Ott G, Wiestner A, Wilson WH, Campo E, Hernández L. microRNA expression profiles identify subtypes of mantle cell lymphoma with different clinicobiological characteristics. *Clin Cancer Res*. 2013 Jun 15;19(12):3121-9.

17. Vegliante MC, Palomero J, Pérez-Galán P, Roué G, Castellano G, Navarro A, **Clot G**, Moros A, Suárez-Cisneros H, Beà S, Hernández L, Enjuanes A, Jares P, Villamor N, Colomer D, Martín-Subero JI, Campo E, Amador V. SOX11 regulates PAX5 expression and blocks terminal B-cell differentiation in

aggressive mantle cell lymphoma. *Blood*. 2013 Mar 21;121(12):2175-85.

18. Salaverria I, Royo C, Carvajal-Cuenca A, **Clot G**, Navarro A, Valera A, Song JY, Woroniecka R, Rymkiewicz G, Klapper W, Hartmann EM, Sujobert P, Wlodarska I, Ferry JA, Gaulard P, Ott G, Rosenwald A, Lopez-Guillermo A, Quintanilla-Martinez L, Harris NL, Jaffe ES, Siebert R, Campo E, Beà S. CCND2 rearrangements are the most frequent genetic events in cyclin D1(-) mantle cell lymphoma. *Blood*. 2013 Feb 21;121(8):1394-402.

19. Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, **Clot G**, Martínez-Trillos A, Castellano G, Brun-Heath I, Pinyol M, Barberán-Soler S, Papasaikas P, Jares P, Beà S, Rico D, Ecker S, Rubio M, Royo R, Ho V, Klotzle B, Hernández L, Conde L, López-Guerra M, Colomer D, Villamor N, Aymerich M, Rozman M, Bayes M, Gut M, Gelpí JL, Orozco M, Fan JB, Quesada V, Puente XS, Pisano DG, Valencia A, López-Guillermo A, Gut I, López-Otín C, Campo E, Martín-Subero JI. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet*. 2012 Nov;44(11):1236-42.

20. Navarro A, **Clot G**, Royo C, Jares P, Hadzidimitriou A, Agathangelidis A, Bikos V, Darzentas N, Papadaki T, Salaverria I, Pinyol M, Puig X, Palomero J, Vegliante MC, Amador V, Martinez-Trillos A, Stefancikova L, Wiestner A, Wilson W, Pott C, Calasanz MJ, Trim N, Erber W, Sander B, Ott G, Rosenwald A, Colomer D, Giné E, Siebert R, Lopez-Guillermo A, Stamatopoulos K, Beà S, Campo E. Molecular subsets of mantle cell lymphoma defined by the IGHV mutational status and SOX11 expression have distinct biologic and clinical features. *Cancer Res*. 2012 Oct 15;72(20):5307-16.

21. Royo C, Navarro A, **Clot G**, Salaverria I, Giné E, Jares P, Colomer D, Wiestner A, Wilson WH, Vegliante MC, Fernandez V, Hartmann EM, Trim N, Erber WN, Swerdlow SH, Klapper W, Dyer MJ, Vargas-Pabón M, Ott G, Rosenwald A, Siebert R, López-Guillermo A, Campo E, Beà S. Non-nodal type of mantle cell

lymphoma is a specific biological and clinical subgroup of the disease. *Leukemia*. 2012 Aug;26(8):1895-8.

Annex C: Finançament

Aquesta tesi s'ha finançat amb fons de les següents institucions: