Anàlisi de la incidència de l'esforç d'alta intensitat en la generació de lncRNA

TFM àrea Intel·ligència Artificial

Ariadna Rius Soler
UOC. TFM. Curs 2013-2014
Consultor: Samir Kanaan Izquierdo

Agraïments

- A en Josep M., per la seva paciència i la força que em dóna.
- A la meva mare, per creure en mi.
- A la Laura E., pels moments d'esbarjo.
- A la resta d'amics i familiars, pel seu recolzament.
- Als professors i companys d'estudis del màster, per la bona experiència.
- A l'Elena G., que fa anys va vaticinar que m'interessaria per l'àmbit de la intel·ligència artificial.
- A en Jordi A. i en Léonard J., per les facilitats que m'han donat per realitzar aquest màster i el seu recolzament, així com al Tecnocampus Mataró-Maresme en general.
- Al projecte SUMMIT, per proporcionar les dades a analitzar en el present TFM.
- A tots els desenvolupadors i investigadors que han definit paquets de R utilitzats en aquest TFM i/o que han redactat els articles i fonts consultades.
    - En especial a l'Helena B., per la seva tesi doctoral.

- I un especial agraïment a en Samir Kanaan, per tot el que m'ha ensenyat, per introduir-me en l'àmbit de la intel·ligència artificial i pel seu suport en aquest TFM.

Índex

Índex de figures

1 Elecció del tema del TFM

El tema escollit pel TFM és l'aplicació de tècniques d'intel·ligència artificial a fragments llargs de non-coding RNA (lncRNA) per descobrir criteris de classificació que permetin relacionar-ho amb funcions, propietats i/o el desenvolupament de malalties.

Els lncRNA són fragments d'ARN que no es tradueixen en proteïnes però que estan implicats en molts processos cel·lulars durant tota o part de la vida de les cèl·lules. Aquestes parts no codificades poden, per exemple, condicionar la quantitat en què es tradueixen les proteïnes. Alguns estudis experimentals suggereixen que els lncRNA estan directament relacionats amb l'envelliment i l'aparició de malalties degeneratives i oncològiques.

El present projecte té per objectiu aplicar tècniques de Machine Learning a fragments lncRNA per intentar descobrir noves relacions amb malalties, funcions o propietats biològiques. El coneixement actual d'aquestes parts es basa en estudis manuals, de manera que una aproximació computacional permetrà analitzar, classificar i relacionar més dades, aplicant tècniques i algoritmes que ja han ajudat a conèixer altres tipus de parts de l'ADN. Per tal de validar els resultats obtinguts, es compararan amb estudis existents i s'extrapolaran a altres hipòtesis per generar nou coneixement.

2 Objectius del TFM

El present TFM té per objectiu aplicar tècniques d'Intel·ligència Artificial per analitzar la incidència de l'esforç d'alta intensitat en la generació de lncRNA. Com a fites concretes s'han fixat les següents:

- Estudi de la incidència de determinats fenotips en la generació de lncRNA i de proteïnes.
- Anàlisi de la correlació entre els nivells d'expressió de proteïnes amb els dels lncRNA.
- Relació de les proteïnes del pathway de la Glicòlisis/Gluconeogènesi correlacionades amb els lncRNA per trobar evidència de la incidència dels lncRNA en aquest pathway.

Utilitzant les dades d'un projecte real (el projecte SUMMIT descrit als apartats posteriors) i tenint en compte les mesures d'individus:

- Que practiquen esport o bé entre 3 i 10 hores a la setmana o més.
- De diferents gèneres.
- Abans i després d'una cursa d'alta intensitat.
- I que han participat en activitats de llargades diverses.

3 Introducció

La genètica és la branca de la biologia que té per objecte d'estudi l'herència biològica entre generacions, permetent entendre allò que succeeix durant el cicle cel·lular i la reproducció dels organismes vius. En aquest sentit cal diferenciar entre les característiques biològiques (genotips) i les físiques (fenotip) d'aparença o, fins i tot, personalitat. El principal objecte d'estudi són els gens, formats per segments d'ADN i ARN.

Una molècula és un conjunt de dos o més àtoms enllaçats covalentment i que conformen un sistema estable i de càrrega elèctrica neutre. Quan aquestes molècules tenen una massa molecular elevada, és a dir que estan formades per un gran nombre d'àtoms, s'anomenen macromolècules. Els polímeres com l'ADN, són macromolècules formades per la unió de molècules més petites. Els àcids nucleic són grans polímers formats per la repetició de monòmers (molècules de massa reduïda) anomenats nucleòtids, i units mitjançant enllaços fosfodiésters, un tipus d'enllaç covalent. Aquest tipus d'unió dóna lloc a llargues cadenes que poden arribar a tenir milions de nucleòtids encadenats.

Els àcids nucleics emmagatzemen la informació genètica dels organismes vius i són responsables de la transmissió hereditària, n'existeixen dos tipus bàsics:

L'**ADN**: conté instruccions genètiques utilitzades en el desenvolupament i funcionament de tots els organismes vius que es coneixen i d'alguns virus. El paper principal d'aquests molècules és l'emmagatzemament d'informació a llarg termini. L'ADN conté les instruccions necessàries per construir altres components de les cèl·lules com les proteïnes (molècules formades per cadenes lineals d'aminoàcids) i les molècules d'ARN. Els fragments d'ADN que contenen les instruccions per generar proteïnes s'anomenen gens. Les altres seqüències d'ADN tenen propòsits estructurals o intervenen en la regulació de l'ús de la informació genètica. Actualment encara hi ha seqüències de funció desconeguda.

L'**ARN**: format per una cadena de ribonucleòtids, és present tant a les cèl·lules que tenen nucli cel·lular com a les que no (procariotes i eucariotes respectivament) i és l'únic material genètic de molts virus. L'ARN és lineal i de cadena simple (excepte a alguns virus que és doble). L'ARN permet a l'ADN transferir informació vital durant la síntesi de les proteïnes, alguns tipus permeten regular l'expressió genètica i d'altres tenen activitat catalítica (augmentar la velocitat d'una reacció química). Als següents subapartats

s'aprofundirà més en l'ARN i l'ARN no codificant, ja que en el present projecte es treballarà amb fragments d'aquesta mena.

3.1 ARN

L'ARN consisteix en una família de molècules que duen a terme múltiples rols vitals en la codificació, descodificació, regulació i expressió dels gens. Els organismes cel·lulars utilitzen ARN missatger (ARNm) per transmetre informació genètica que regula la síntesi de proteïnes específiques. Les molècules d'ARN tenen diverses funcions cel·lulars com catalitzar reaccions biològiques, controlar l'expressió dels gens o detectar i comunicar respostes a senyals cel·lulars. En la síntesi de proteïnes, les molècules d'ARNm missatger dirigeixen l'acoblament de les proteïnes en ribosomes. Aquest procés utilitza les molècules ARN de transferència (ARNt) per lliurar els aminoàcids als ribosomes, on l'ARN ribosòmic connecta els aminoàcids per formar les proteïnes. D'aquesta manera, els ARNm serveixen de model químic per crear la proteïna, transportant la informació codificada als ribosomes, on el polímer d'àcid nucleic és traduït a un d'aminoàcids (proteïna).

La informació genètica de l'ARN està codificada en la seqüència de 4 nucleòtids agrupats en codons de tres bases cadascun. Cada codó comença per un aminoàcid específic, excepte el STOP que indica el final de la síntesi proteica. L'ARNt reconeix el codó i aporta l'aminoàcid corresponent a la cadena proteica en síntesi mentre que l'ARNr fan la traducció en sí.

Segons el dogma de la biologia molecular la traducció d'ADN a proteïnes és unidireccional, ja que a partir de la informació continguda a l'ADN, es pot generar ARN i és possible sintetitzar proteïnes. Aquesta afirmació només té una excepció coneguda en el cas dels retrovirus, on es produeix l'anomenada transcripció inversa. Els retrovirus integren una còpia del seu genoma, que està composat per ARN, al genoma de l'hoste, modificant-ne així la informació per generar proteïnes.

L'expressió genètica és el procés pel qual la informació d'un gen s'utilitza en la síntesi d'una molècula d'ARN o d'una proteïna. El nivell d'expressió és diferent per cada cèl·lula, de manera que el seu estudi permet comparar cèl·lules malaltes i sanes, amb medicació o sense, amb diferents condicions d'estrès, estadis de creixement, etc.

La imatge següent mostra la comparació de l'expressió genètica de cèl·lules sanes i canceroses de fetge i pàncrees. Els nivells d'expressió estan representats en vermell (més baixos) i verd (més elevats). L'eix

vertical representa les cèl·lules de fetge i pàncrees mentre que l'horitzontal permet diferenciar els nivells d'expressió de les cèl·lules sanes (primera meitat) de les canceroses (segona meitat):

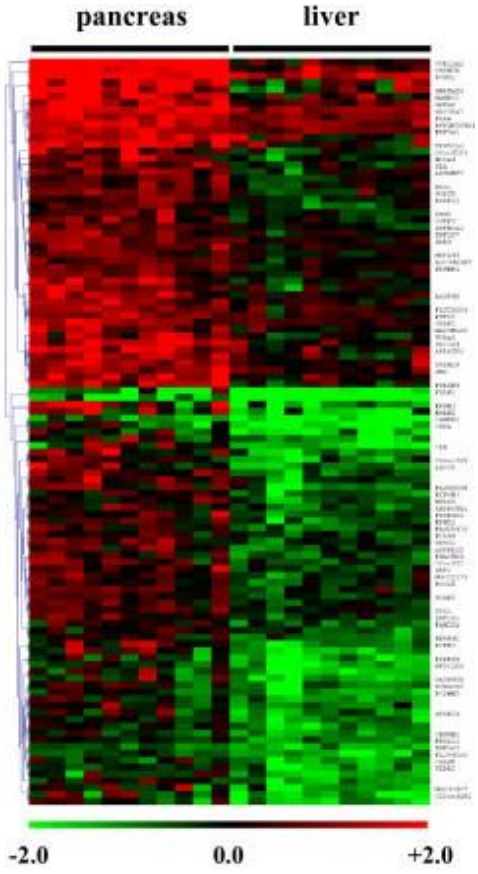


Figura 1. Nivells d'expressió de cèl·lules sanes i canceroses de pàncrees i fetge.

No totes les molècules d'ARN es codifiquen en proteïnes, com és el cas dels anomenats ARNnc (ARN no codificant), que inclouen els ARNt i els ARNr (ARN ribosòmic). Fins fa poc es considerava que els fragments no codificants no tenien cap funció però s'està descobrint la seva implicació en varis processos cel·lulars. No es coneix el nombre d'ARNnc però estudis recents suggereixen que n'hi ha milers, molts d'ells desconeguts o de funció no confirmada. Quan un ARNnc té més de 200 nucleòtids s'anomena ARNnc llargs (lncRNA) els quals són l'objecte del present TFM.

3.2 lncRNA

Des que es va reconèixer la participació activa dels fragments lncRNA en la biologia cel·lular el seu estudi s'ha centrat cada vegada més en la contribució que poden tenir en la causa de determinades malalties. Diversos estudis han evidenciat la implicació dels lncRNA amb el desenvolupament i l'aparició de malalties neurològiques i oncològiques. Comparant les expressions d'ARNnc en cèl·lules tumorals amb d'altres de

sanes, s'han evidenciat diferències que permeten conèixer millor i abans l'evolució que tindrà el pacient. Malgrat s'han identificat nombrosos efectes dels ARNnc en el càncer, la funció d'aquests fragments i el seu rol potencial respecte els tumors és força desconegut. L'ARNnc també s'ha relacionat amb els estats de diferents malalties i l'envelliment.
L'habilitat dels fragments lncRNA per regular les proteïnes codificades dels gens pot contribuir en el desenvolupament d'una malaltia o variar-ne els efectes si es desregula la síntesi d'una proteïna amb significat clínic. L'alteració de l'expressió d'ARNnc també pot intervenir en els canvis a nivell epigenètic per afectar l'expressió de gens i contribuir en la causa d'una malaltia.

El fet d'evidenciar la relació de fragments lncRNA amb l'aparició i el desenvolupament de malalties es pot dur a terme a través de l'agrupament de mostres per diferents criteris. Agrupant individus (i la seva informació) i observant si comparteixen o no fragments de lncRNA es poden detectar incidències d'aquest tipus de fragments en la característica observada. Per dur a terme aquests anàlisis, és necessari tenir també un grup de contrast, que no tingui el tret objectiu. La dificultat d'aquesta mena d'estudis és, per una banda, triar la o les característiques que compartiran els elements de la mostra i, per l'altre, identificar els lncRNA. També cal tenir en compte com incideix el context de cada individu a la presència del lncRNA i als altres trets.

Els estudis que s'han realitzat al voltant dels lncRNA han estat força manuals, de manera que no s'han pogut analitzar grans quantitats de dades. Aquest projecte té per objectiu aplicar tècniques d'intel·ligència artificial per superar aquesta limitació i incorporar aprenentatge respecte els estudis ja realitzats, aprofitant-ne l'experiència i els coneixements obtinguts. Aquest plantejament també permetrà fer proves amb agrupacions per criteris diferents, així com analitzar interaccions i dependències entre característiques observades. També s'espera que l'aprenentatge realitzat a partir d'experiències reals es pugui extrapolar i aplicar a nous paradigmes.

4 Planificació

Aquest apartat presenta la planificació realitzada del TFM a través d'una taula de fites, la relació de lliurables i d'un diagrama de Gantt:

4.1 Taula de fites

La taula següent mostra les fites temporalitzades que del TFM:

Fita		Durada	Inici	Fi
Elecció del tema del TFM		7 dies	18/09/2013	24/09/2013
Planificació del TFM		12 dies	25/09/2013	**6/10/2013**
Preparació		28 dies	7/10/2013	**3/11/2013**
	Ampliació coneixements sobre el domini del TFM	6 dies	7/10/2013	12/10/2013
	Estat de l'art d'algoritmes i eines a aplicar i elecció	6 dies	13/10/2013	18/10/2013
	Recerca de BBDD i elecció de les dades a treballar	5 dies	19/10/2013	23/10/2013
	Recerca d'estudis experimentals	5 dies	24/10/2013	28/10/2013
	Tractament de les dades i transformació a format adequat	6 dies	29/10/2013	3/11/2013
Execució		42 dies	4/11/2013	**15/12/2013**
	Aplicació dels algoritmes i ús d'eines	21 dies	4/11/2013	24/11/2013
	Comparació de resultats amb estudis experimentals realitzats	9 dies	25/11/2013	3/12/2013
	Extrapolació de resultats i prediccions	12 dies	4/12/2013	15/12/2013
Anàlisi de resultats i conclusions		14 dies	16/12/2013	29/12/2013
Redacció de la memòria		105 dies	25/9/2013	**7/01/2014**
Preparació de la presentació i de la defensa		9 dies	30/01/2014	**7/01/2014**
TOTAL:		112 dies	18/09/2013	07/01/2014

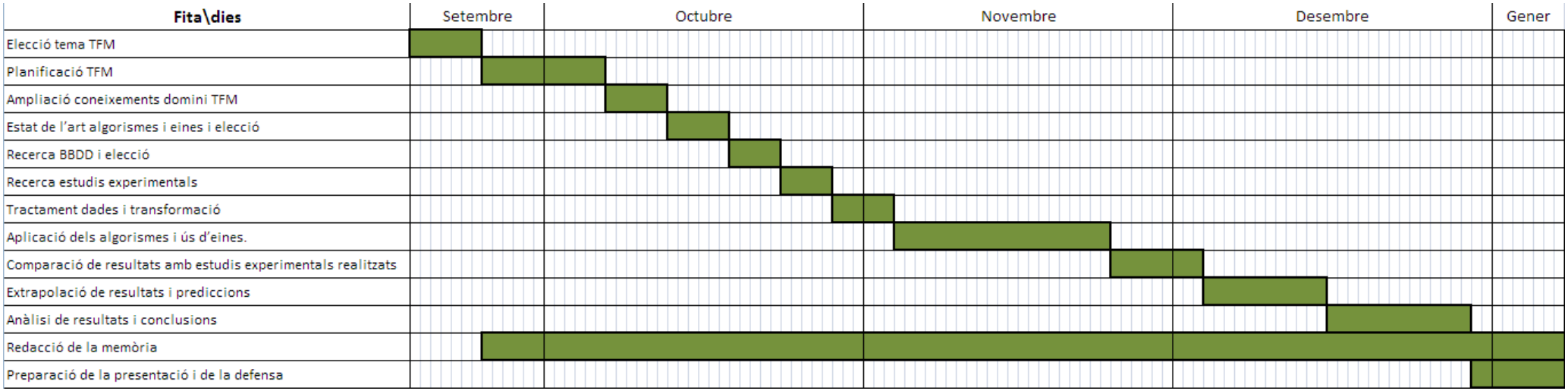
4.2 Lliurables

Els lliurables que es preveuen a les fases del treball es presenten a la taula següent:

Fase	Lliurable	Data
Planificació del TFM	Pla de treball temporalitzat	6/10/2013
Preparació	Parts de la memòria del TFM corresponents a: • Objectius concrets del TFM i descripció del problema. • Introducció sobre el domini del TFM (genètica). • Algoritmes i eines a aplicar. • BBDD a utilitzar. • Estudis experimentals trobats. • Tractament de dades aplicat. Aquests apartats inclouran un resum de les diferents recerques fetes, així com justificació de les decisions preses.	3/11/2013
Execució	Part de la memòria explicant els algoritmes i eines aplicades, comparació de resultats obtinguts amb estudis experimentals i extrapolació de resultats i prediccions. Inclourà l'explicació del codi font utilitzat.	15/12/2013
Tancament	Memòria, incloent la part de resultats, conclusions i possibles ampliacions a realitzar.	7/01/2014
Tancament	Presentació	7/01/2014

4.3 Diagrama de Gantt

A continuació es presenta el diagrama de Gantt corresponent a la planificació presentada als apartats anteriors:



4.4 Modificacions respecte la planificació inicial

En aquest apartat es recullen les modificacions realitzades respecte la planificació inicial prèviament presentada:

- Les subfases "estat de l'art d'algoritmes i eines a aplicar i elecció" i "recerca de BBDD i elecció de les dades a treballar" s'han executat simultàniament, ja que depenen molt l'una de l'altre.
- La subfase "recerca estudis experimentals" ha acabat formant part de la subfase "recerca de BBDD i elecció de les dades a treballar", ja que els estudis que s'utilitzaran són els que han permès identificar els pathways (proteïnes que treballen plegades per dur a terme una funció determinada). D'aquesta manera, la subfase consisteix en trobar BBDD de pathways (veure apartat 4.2 "Bases de Dades").
- També hi ha hagut una desviació temporal, ja que ha calgut més temps per dur a terme la fase de planificació i la de preparació del que inicialment es va preveure.
- L'entrega final de TFM s'ha realitzat el 14 de gener (una setmana més tard del que inicialment es va preveure, tot i que a la planificació inicial ja es comptava amb aquest cert marge).

5 Preparació

Aquest apartat té per objectiu recollir tots els passos previs a l'aplicació dels algoritmes a les dades genètiques, des de l'elecció d'eines fins al tractament dels fragments d'ARN per tal que tinguin un format adequat. La fase de preparació també inclou l'elecció d'algoritmes i de les BBDD a utilitzar, així com la selecció d'estudis experimentals que es tindran en compte en l'aprenentatge i extrapolació de resultats:

5.1 Eines

El llenguatge de programació seleccionat per realitzar aquest TFM ha estat R, ja que és el que ofereix més paquets especialitzats per treballar amb dades genètiques. El web bioconductor.org, per exemple, conté una gran varietat de recursos per analitzar i comprendre dades d'aquest àmbit. Com a alternativa es podrien utilitzar els llenguatges Python o LISP, ja que són adequats per desenvolupar algoritmes d'intel·ligència artificial, així com l'entorn i el llenguatge MatLab, indicat pel càlcul numèric i l'explotació de dades des del punt de vista matemàtic i estadístic.

Amb l'objectiu de facilitar el desenvolupament amb R, també s'utilitzarà el RStudio IDE (Integrated Development Environment). Aquest entorn ofereix diferents funcionalitats de suport com el ressaltat de sintaxis, l'autocompletat de codi, l'organització per projectes, la navegació entre scripts o l'exportació a pdf. RStudio és un programari no distribuït, gratuït i obert, que té versió per Windows, Linux i Mac i permet treballar amb qualsevol versió de R a partir de la 2.11.1. Una altra característica interessant de RStudio és la capacitat de generar aplicacions web que utilitzin les funcionalitats de R i permetin realitzar operacions amb les dades. Aquesta accessibilitat es pot aconseguir utilitzant el paquet Shiny (integrat a RStudio), la versió servidor de l'entorn (RStudio Server) o un programari independent anomenat rApache (que utilitza el servidor web Apache).

En aquest TFM s'utilitzarà Shiny, donat que permet treballar en local (separa el client del servidor), està integrat al RStudio i té versió per Windows. El Shiny Server sí que s'ha d'instal·lar sobre un sistema operatiu Linux però en aquest projecte no s'utilitzarà, es durà a terme un desenvolupament en mode local. Rstudio Server i rApache també requereixen utilitzar el servidor sobre un sistema operatiu Linux i el segon no està integrat amb RStudio. A partir de la documentació de suport de les diferents eines també es percep que Shiny és la més usable i fàcil d'adoptar.

5.2 Bases de Dades

Per poder relacionar el tret característic d'una cèl·lula (com una malaltia o si el pacient es pren una determinada medicació) amb un lncRNA és necessari comptar amb diferents tipus de dades:

Nivells d'expressió de les cèl·lules a estudiar: Es corresponen amb la mesura de concentració d'una proteïna o lncRNA a una cèl·lula (o directament a la sang), en un moment determinat del temps. Les variacions d'aquestes concentracions permeten estudiar com canvia la resposta de la cèl·lula. El nivell d'expressió és un valor numèric que sol estar acotat a un rang tancat (per exemple -2,0 i 2,0 en l'exemple presentat al punt 2.1 "ARN" referent a les cèl·lules de fetge i pàncrees).

Pathways: Conjunts de proteïnes que treballen plegades per dur a terme una acció biològica. S'han identificat de manera manual, gràcies a estudis i projectes d'investigació. Per exemple, la imatge següent mostra el pathway de la síntesi i degradació de cossos cetònics (tipus de compost químic):

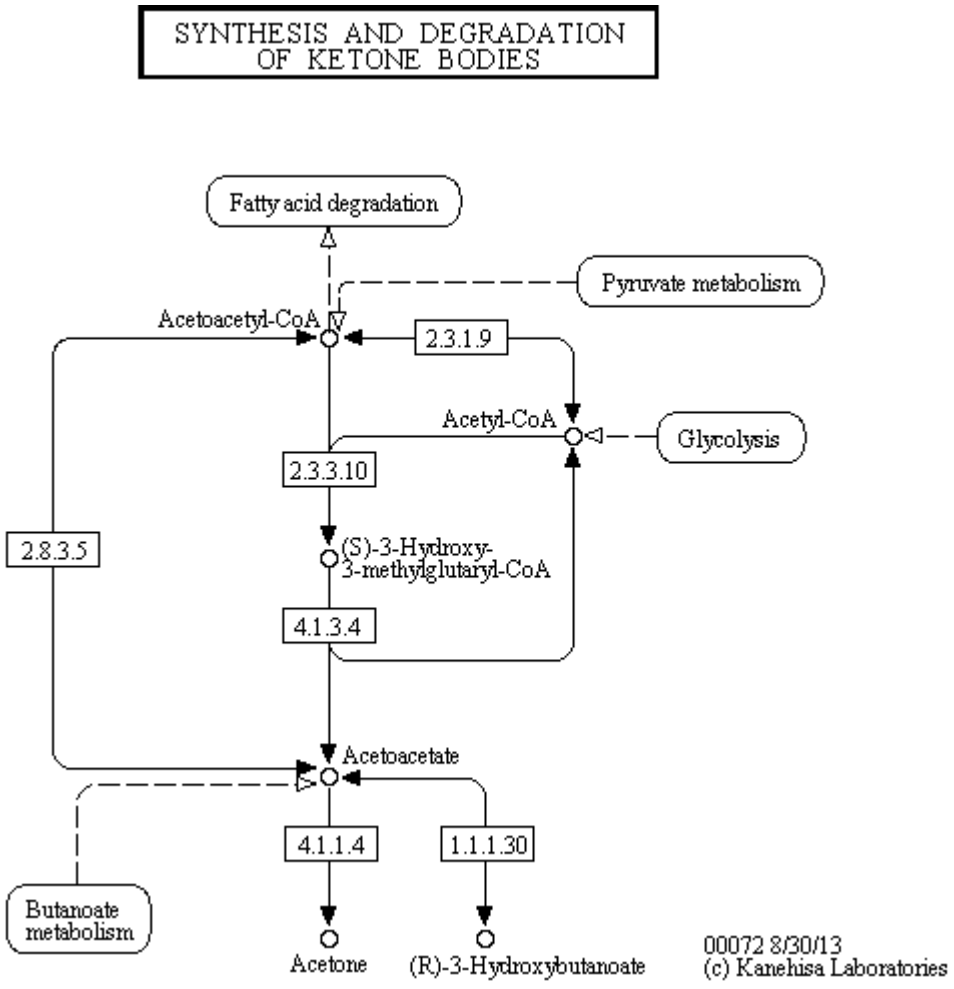


Figura 2. Pathway síntesi i degradació de cossos cetònics.

Seqüències de lncRNA: Representen les agrupacions de nucleòtids que conformen el lncRNA. Per exemple, la seqüència del lncRNA 21A_ és:
"AAATAGTTGACCAAGTGTGGTGGCTCACGTAGTCCCAGCACTTTGGGAGGCTGAGGCAGGAGGATCACTTGAGCCC
AGGAATTTGAGACCAGCTTGGGCAACATAGTGAGACCTCATCTCTTAAAAAAAAAAATTAGCTGGGTGTGGTAGTGC
ACACCTGTGGTCCCAGCTACTTTAGAGGCTGAGGTAGAGGATTGCTTGAGCCTGGGAAGTTGGGGCTGTAGTGAGC
TTTGATTGCATCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGTCTCTAAAAAATTAAATAAATAATAAAAAA
ATTAAAAAGTAACTCCC".

Combinant l'estudi d'aquests elements és possible determinar:

- Si un lncRNA s'ha generat a partir d'una cèl·lula concreta (comparant les seqüències de lncRNA generades a partir de l'ADN de la cèl·lula amb les seqüències de lncRNA identificades). Aquesta acció permet relacionar el tret característic de la cèl·lula (malaltia o situació) amb la producció d'un lncRNA concret.
- Si una situació o acció modifica la resposta d'una cèl·lula (mesurant els canvis dels nivells d'expressió).
- Funcions biològiques de lncRNA (analitzant la correlació amb proteïnes i pathways identificats).

De manera que per realitzar el TFM cal:

- Obtenir els nivells d'expressió de cèl·lules amb característiques prèviament determinades i situacions de canvi.
- Com a mínim una BBDD de seqüències de lncRNA.
- Com a mínim una BBDD de pathways.

Als següents subapartats es presentaran possibles fonts d'aquestes dades, triant-ne per a l'execució del projecte:

5.2.1 Nivells d'expressió

Els nivells d'expressió que s'utilitzaran són els recollits al projecte SUMMIT, que té objectiu determinar si la població que practica exercici de llarga durada i gran intensitat té més risc sobre la salut que la sedentària i/o la moderadament activa. Les dades a utilitzar seran les mesures dels nivells d'expressió obtingudes de les tres mostres d'individus següents:

- Grup de persones poc actives: Composat per 20 homes i 20 dones que fan entre 0 i 3 hores setmanals d'exercici físic.

- Grup de persones físicament actives: Format per 20 dones i 20 homes que fan entre 3 i 10 hores d'exercici físic a la setmana.
- Grup de persones físicament molt actives i esportistes d'elit: 20 dones i 20 homes que fan exercici físic més de 10 hores a la setmana.

I en diferents moments de l'activitat d'alta exigència (curs amb modalitats de 10, 35 i 85 km realitzada al juny del 2013), és a dir, abans de començar l'exercici, durant i un cop finalitzat.

El fet de comptar amb aquestes dades permetrà treballar en un àmbit de recerca poc explorat: com afecta l'esforç físic extrem a la generació de lncRNA en individus dels tres grups descrits. Concretament aquest TFM té per objectius:

- Esbrinar quins lncRNA estan més expressats amb esforç i quins sense.
    - Analitzar les diferències per individu, per grau d'activitat, gènere, etc.
    - Tenir en compte també les mesures durant la cursa i analitzar-ne l'evolució.
- Calcular la correlació dels lncRNA que tinguin altes concentracions amb proteïnes, per tal d'esbrinar quina funció deuen tenir.
- Relacionar les proteïnes amb alta correlació amb els pathways per trobar possibles funcions biològiques dels lncRNA.

També es podrien utilitzar dades de BBDD de nivells d'expressió com NERD, que conté els nivells d'expressió de cèl·lules de diferents parts del cos d'humans i ratolins. Resulta difícil, però, trobar dades de lncRNA que, a més a més, siguin referents a malalties o que tinguin en compte situacions complexes.

El conjunt de dades disponible amb el qual es treballarà en aquest projecte no inclou totes les mesures indicades, donat que l'experiment no es va completar. Concretament, s'utilitzaran 28 fitxers amb:

- 16 mesures fetes abans de la cursa i 12 després.
- 8 dones i 20 homes.
- 4 mostres de 25 km, 1 de 28 km, 2 de 33 km, 2 de 42 km, 10 de 50 km 5 de 82 km i una de la qual se'n desconeix la distància.
- 18 del grup de persones actives i 10 d'esportistes d'elit.

Aquestes dades ja han passat un control de qualitat que quedava fora de l'abast del present treball.

5.2.2 BBDD de seqüències de lncRNA

S'han identificat els següents repositoris de seqüències de lncRNA i que en permeten la descàrrega:

- Lncrna db: Conté informació sobre lncRNA de cèl·lules eucariotes com les característiques, possibles funcions, l'expressió, la seqüència, referències a literatura, components associats, etc. Permet cercar per diferents criteris, entre ells, l'espècie.
- LncRNADisease database: Repositori de lncRNA que conté possibles relacions amb malalties, interaccions i prediccions. De cada element n'emmagatzema la informació bàsica (nom, descripció, espècie, la seqüència, referències literàries etc.), així com el tipus de funció que realitza i la malaltia associada. Permet navegar per lncRNA o per malaltia, i realitzar cerques.
- NONCODE: BBDD de ncRNA (excepte tRNA i rRNA) (també permet cercar i navegar només pels lncRNA). Conté un enllaç a la fitxa del nucleòtid del NCBI (National Center for Biotechnology Information) d'Estats Units.
- LNCipedia: Conté informació sobre lncRNA d'humans, els quals es poden cercar o navegar-hi. De cada element s'emmagatzema el nom, id del gen, localització, cadena, transcripcions, literatura disponible, seqüència, etc.

En aquest TFM s'utilitzaran dades dels quatre repositoris citats, eliminant possibles repeticions (veure apartat 4.4 "Tractament de les dades"):

- Lncrna db: Realitzant una cerca filtrant per espècie "Homo sapiens (Human) (127)" i a través de l'opció "Export Seq." És possible descarregar els lncRNA d'humans que conté la BBDD.
- LncRNADisease database: A l'apartat de descàrregues es descarregarà el fitxer corresponent al recurs "The RNA sequence of lncRNA (txt)".
- NONCODE: A la plana de descàrregues es descarregarà l'arxiu comprimit "lncRNA_human.zip" corresponent a "Homo sapiens (sequence and mapping)".
- LNCipedia a la secció de descàrregues es pot descarregar la BBDD de seqüències (recurs FASTA).

Finalment la comparació entre lncRNA i proteïnes s'ha realitzat a través de la correlació dels nivells d'expressió (veure apartat 4.3 "Algoritmes").

5.2.3 BBDD de pathways

Com s'ha comentat prèviament, els pathways són agrupacions de seqüències de proteïnes i les interaccions entre elles, que s'ha identificat que actuen conjuntament per dur a terme una funció biològica concreta. Després de realitzar una cerca s'ha pogut comprovar que les representacions que se'n poden fer

són diverses (gràfiques, XML, formats propietaris, ontologies, etc.) i sovint les planes web obertes no ofereixen opcions de descàrrega de les dades complertes. Per aquests motius s'han cercat directament les BBDD a les quals es pot accedir a través de paquets de R. Al web bioconductor.org s'han trobat els següents:

- NCIgraph: Ofereix mètodes per carregar els pathways de la BBDD del NCI en objectes de R grafs i per reformular-los.
- KEGGgraph: Interfície entre els pathways de la BBDD KEGG i objectes grafs, així com eines per realitzar-hi anàlisis, visualitzacions, etc. mantenint els atributs dels pathways.

Pel present TFM només s'utilitzaran dades de la BBDD KEGG, ja que la del NCI requereix de programari extern addicional per poder-la utilitzar (Cytoscape) i el paquet per usar-la no és tant complert com l'altre (se'n poden consultar exemples d'aplicació amb dades de prova al manual de referència i al següent article).

5.3 Algoritmes

En aquest TFM s'utilitzaran tècniques de Machine Learning no supervisat, ja que no existeixen dades per poder conformar un conjunt de validació. En concret s'utilitzaran les tres tècniques següents:

- Agrupació (clustering): Agrupació dels valors dels nivells d'expressió de lncRNA i de proteïnes per esbrinar si hi ha diferències. S'utilitzarà l'algoritme PAM (Partitioning Around Medoids).
- Correlació: S'usarà la correlació de Pearson per mesurar la correlació entre els nivells d'expressió dels lncRNA i els de les proteïnes.
- ICA (Independent Component Analysis): Ús d'ICA per comparar les proteïnes que composen el pathway de la glucosa amb les correlacionades amb els nivells d'expressió dels lncRNA.

Els algoritmes citats, així com l'ús que se'n farà, es detallen als següents subapartats:

5.3.1 Agrupació

L'agrupació dels nivells d'expressió de lncRNA per diferents criteris pot permetre identificar indicis sobre la incidència dels criteris en la generació dels lncRNA i de les proteïnes. D'aquesta manera, s'utilitzarà l'algoritme PAM d'agrupació per esbrinar si hi ha diferències en la generació de nivells d'expressió de lncRNA i de proteïnes per:

- Distància de la cursa d'alt nivell (de menys de 40 km, d'entre 40 i 60 km i de més de 60 km).
- Gènere.

- Moment de la cursa (abans de començar i un cop acabada).
- Nivell d'activitat física (actives i molt actives/esportistes d'elit).

Per dur a terme aquestes agrupacions es podria usar l'algoritme k-mitjanes (k-means). Aquesta tècnica es basa en triar un punt a l'atzar (anomenat centroide) per cada clúster a crear, situar les dades al grup que tingui el centroide més proper i recalcular els centroides i les situacions de les dades fins que les assignacions als grups no variïn. És important destacar que el centroide no té perquè ser un punt de les mateixes dades, només estar dins el mateix interval. El nombre de clústers a crear és un paràmetre d'entrada, l'algoritme no és capaç de determinar-ho. En funció del tipus de problema a treballar, l'algoritme pot fer agrupacions incorrectes, ja que tendeix a crear esferes similars que particionen l'espai. D'aquesta manera, alguns punts poden estar clarament més propers a un centroide però assignar-se a un altre grup per estar situats (a l'eix de coordenades) a l'esfera d'aquell altre clúster. També és un algoritme poc estable, ja que el resultat varia a cada execució en funció dels centroides generats aleatòriament a l'inici.

L'algoritme PAM funciona de manera similar al k-mitjanes, ja que també trenca les dades en un nombre de grups indicat i treballa amb iteracions minimitzant l'error. La diferència entre PAM i k-mitjanes és que el primer no treballa amb centroides, sinó amb medoides. D'aquesta manera, PAM utilitza un punt de les dades per representar els clústers. L'elecció inicial dels medoides es realitza de manera aleatòria i es va corregint amb varies iteracions fins que no varia i la disseminació de les dades agrupada és mínima.

S'ha triat l'algoritme PAM perquè és més estable que el k-mitjanes i la representació que es fa dels clústers (medoides i no centroides), així com l'agrupació final, resulta més adequada. Per aplicar la tècnica s'utilitzarà el paquet de R "cluster", que conté l'algoritme PAM.

L'algoritme d'anàlisi de components principals (PCA, Principal Component Analysis) permet simplificar les dades identificant un nou conjunt de variables en què se'n maximitza la dispersió i utilitzant-lo com a representació de dimensionalitat reduïda. Aquesta simplificació s'aplicarà a la matriu que contingui els nivells d'expressió de les proteïnes i els lncRNA abans de realitzar l'agrupament, de manera que els clústers ja es realitzin amb aquestes noves variables. Reduir la dimensionalitat de les dades permet minimitzar el cost computacional i augmentar la precisió de les agrupacions. S'utilitzarà el paquet de R "pcaMethods" per executar l'algoritme en diferents mètodes i esbrinar quants components calen per reproduir la variabilitat de les dades.

5.3.2 Correlació Pearson

Amb l'objectiu d'analitzar la possible correlació entre els nivells d'expressió, obtinguts del projecte SUMMIT, dels lncRNA i els de les proteïnes, s'utilitzarà el coeficient de correlació de Pearson com a mesura de similitud entre ambdós conjunts de dades. Una altra mètrica de similitud que es podria utilitzar és la distància Euclidiana però s'ha descartat perquè és molt sensible a l'escala i al nombre de dimensions, comportament que portaria a distorsionar els resultats. Usant la correlació de Pearson aquests problemes es resolen.

S'utilitzarà la funció cor.test(...) del paquet de R "stats" que permet mesurar la correlació entre dos estructures de dades de la mateixa mida amb varis mètodes, entre els quals, Pearson.

5.3.3 Anàlisi de Components Independents

Un dels objectius del present TFM és identificar la correlació entre nivells d'expressió de lncRNA i els de proteïnes per analitzar posteriorment la incidència de l'esforç d'alt rendiment en el pathway de la glucosa. Per assolir aquest objectiu s'utilitzarà el concepte de metafenotip introduït i descrit a la tesi doctoral "*Genetic association analysis of complex diseases through information theoretic metrics and linear pleiotropy*". El concepte de metafenotip es correspon amb una variable fenotípica sintètica obtinguda a partir d'un conjunt real de fenotips utilitzant un model matemàtic. Aquesta nova variable ha de ser capaç de capturar l'estructura de les dades originals per descriure-la com un tot.

Seguint amb el plantejament utilitzat a la tesi, s'aplicarà el mètode estadístic d'Anàlisi de Components Independents ICA per construir un metafenotip del pathway de la glucosa. S'ha descartat l'ús d'altres mètodes de factorització matricial com PCA o SVD (Singular Value Decomposition), donat que els factors que permeten obtenir no tenen perquè ser necessàriament independentment i en el cas d'ICA sí. Aquesta característica es correspon amb el tipus d'interacció entre proteïnes al pathway.

La tècnica ICA permet identificar fonts estadísticament independents a partir d'un conjunt de senyals barrejats. D'aquesta manera, permet descompondre un senyal en les fonts independents que el composen. En aquest cas la descomposició s'aplicarà per identificar quines proteïnes (correlacionades amb els lncRNA mesurats) afecten al pathway de la glucosa, tenint en compte la incidència de les seves interaccions. El fet de no haver de considerar les proteïnes a nivell individual permetrà reduir tant el soroll (falsos positius en la incidència de la proteïna al pathway) com el silenci de les dades (una proteïna per si sola pot no tenir massa afectació sobre el pathway però sí quan interactua amb alguna altra). D'aquesta

manera, el pathway es correspon amb la barreja de senyals i les proteïnes amb cada senyal (independents però en interacció entre elles).

L'algoritme s'aplicarà utilitzant el paquet de R "fastICA". Aquesta implementació és la que més s'utilitza perquè combina la facilitat d'ús amb una bona eficiència computacional.

Havent obtingut el metafenotip amb l'algoritme ICA, s'hi aplicarà un test hipergeomètric per calcular la probabilitat que surtin les proteïnes del pathway correlacionades amb els lncRNA d'entre les proteïnes del pathway i del total de les mesurades.

Per explicar el test hipergeomètric es sol utilitzar el símil amb una urna de la qual s'extrauen boles: En una urna on hi ha boles blanques i negres barrejades, es calcula la probabilitat que surtin boles determinades, sense que hi hagi una reposició de boles després de cada extracció. En el cas del pathway, l'urna conté les proteïnes del pathway (boles blanques) i totes les que s'han mesurat a l'estudi (boles negres); el test retorna la probabilitat que s'extreguin les proteïnes correlacionades amb lncRNA i que estiguin al pathway de la glucosa.

Per aplicar el test s'utilitzarà la funció phyper del paquet stats de R.

5.4 Tractament de les dades

Per cada tipus de dada (nivells d'expressió, seqüències de lncRNA i pathways) s'aplicarà el tractament específic descrit als següents subapartats:

5.4.1 Nivells d'expressió

Els nivells d'expressió es troben emmagatzemats als fitxers .CEL i les seves metadades a fitxers .csv. L'arxiu "Fenos_CDV.csv" conté les metadades en si (identificador participant, edat, grup d'activitat esportiva, alçada, pes, gènere, si és abans o després de la cursa, massa corporal, etc.) i "Descripcion_fenos.csv" recull el nom complet de cadascuna. En total hi ha 28 fitxers .CEL amb 53.617 mesures cadascun.

Amb el codi següent s'importa el contingut dels fitxers .CEL utilitzant el paquet oligo:

```
> library(oligo)
> library(pd.mogene.2.0.st)
> setwd("TFM/Nivells")
> celFiles <- list.celfiles()
```

```
> celFiles <- list.celfiles()
> affyRaw <- read.celfiles(celFiles)
Reading in : 1_PRE_(HuGene-2_0-st).CEL
Reading in : 15_POST_(HuGene-2_0-st).CEL
Reading in : 15_PRE_(HuGene-2_0-st).CEL
Reading in : 163_POST_(HuGene-2_0-st).CEL
Reading in : 163_PRE_(HuGene-2_0-st).CEL
Reading in : 164_POST_(HuGene-2_0-st).CEL
Reading in : 164_PRE_(HuGene-2_0-st).CEL
Reading in : 197_POST_(HuGene-2_0-st).CEL
Reading in : 197_PRE_(HuGene-2_0-st).CEL
Reading in : 2_POST_(HuGene-2_0-st).CEL
Reading in : 2_PRE_(HuGene-2_0-st).CEL
Reading in : 256_POST_(HuGene-2_0-st).CEL
Reading in : 256_PRE_(HuGene-2_0-st).CEL
Reading in : 3_POST_(HuGene-2_0-st).CEL
Reading in : 3_PRE_(HuGene-2_0-st)_2.CEL
Reading in : 577__PRE_(HuGene-2_0-st).CEL
Reading in : 577_POST_(HuGene-2_0-st).CEL
Reading in : 7_POST_(HuGene-2_0-st).CEL
Reading in : 7_PRE_(HuGene-2_0-st).CEL
Reading in : 735_POST_(HuGene-2_0-st).CEL
Reading in : 735_PRE_(HuGene-2_0-st).CEL
Reading in : 775_PRE_(HuGene-2_0-st).CEL
Reading in : 805_POST_(HuGene-2_0-st).CEL
Reading in : 805_PRE_(HuGene-2_0-st).CEL
Reading in : 838_PRE_(HuGene-2_0-st).CEL
Reading in : 9_PRE_(HuGene-2_0-st).CEL
Reading in : 957_POST_(HuGene-2_0-st).CEL
Reading in : 957_PRE_(HuGene-2_0-st).CEL

> affyRaw
GeneFeatureSet (storageMode: lockedEnvironment)
assayData: 2598544 features, 28 samples
  element names: exprs
protocolData
  rowNames:  1_PRE_(HuGene-2_0-st).CEL  15_POST_(HuGene-2_0-st).CEL  ...  957_PRE_(HuGene-2_0-
st).CEL (28 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames:  1_PRE_(HuGene-2_0-st).CEL  15_POST_(HuGene-2_0-st).CEL  ...  957_PRE_(HuGene-2_0-
st).CEL (28 total)
  varLabels: index
  varMetadata: labelDescription channel
featureData: none
experimentData: use 'experimentData(object)'
Annotation: pd.hugene.2.0.st
```

Per representar gràficament els nivells d'expressió es pot utilitzar:

```
> image(affyRaw, 1, col=gray((64:0)/64))
```

El segon paràmetre indica quin dels fitxers representar (en aquest cas el 1 "1_PRE_(HuGene-2_0-st).CEL").

Les imatges següents mostren els nivells d'expressió de proteïnes i RNA del primer i el segon fitxer .CEL[1]:

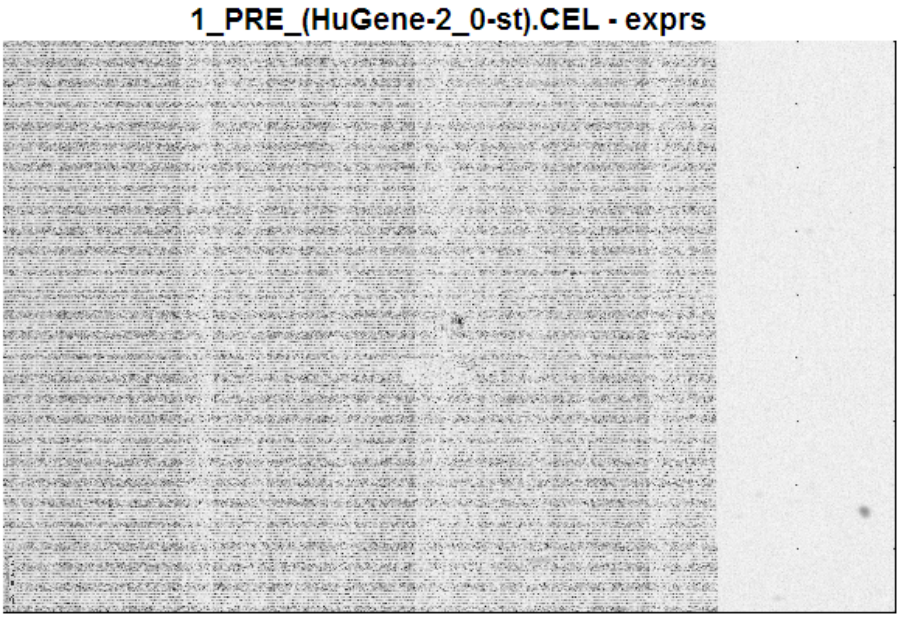


Figura 3. Nivells d'expressió del fitxer "1_PRE_(HuGene-2_0-st).CEL".

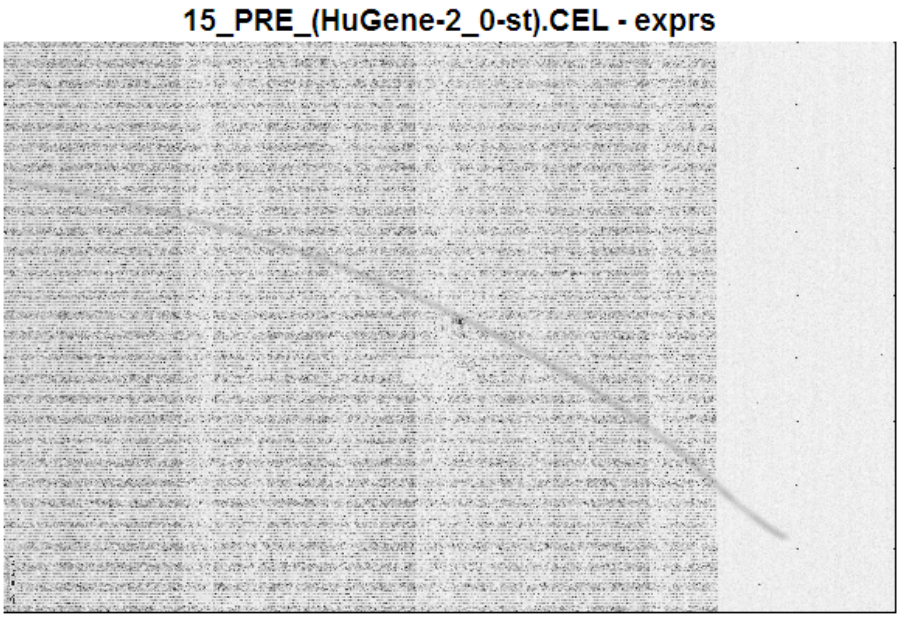


Figura 4. Nivells d'expressió del fitxer "15_PRE_(HuGene-2_0-st).CEL".

[1] Les altres imatges es poden consultar a la carpeta "NivellsExpressioIMG" del fitxer en el qual es lliura aquest projecte.

Per fer-ne l'histograma es passa el GeneFeatureSet a ExpressionSet (classe adient per posteriorment usar les dades als algoritmes):

```
> affyRawExp <- ExpressionSet(assayData(affyRaw))
> hist(expr(affyRawExp))
```

La imatge següent es correspon amb l'histograma generat, que mostra les freqüències dels valors dels nivells d'expressió:

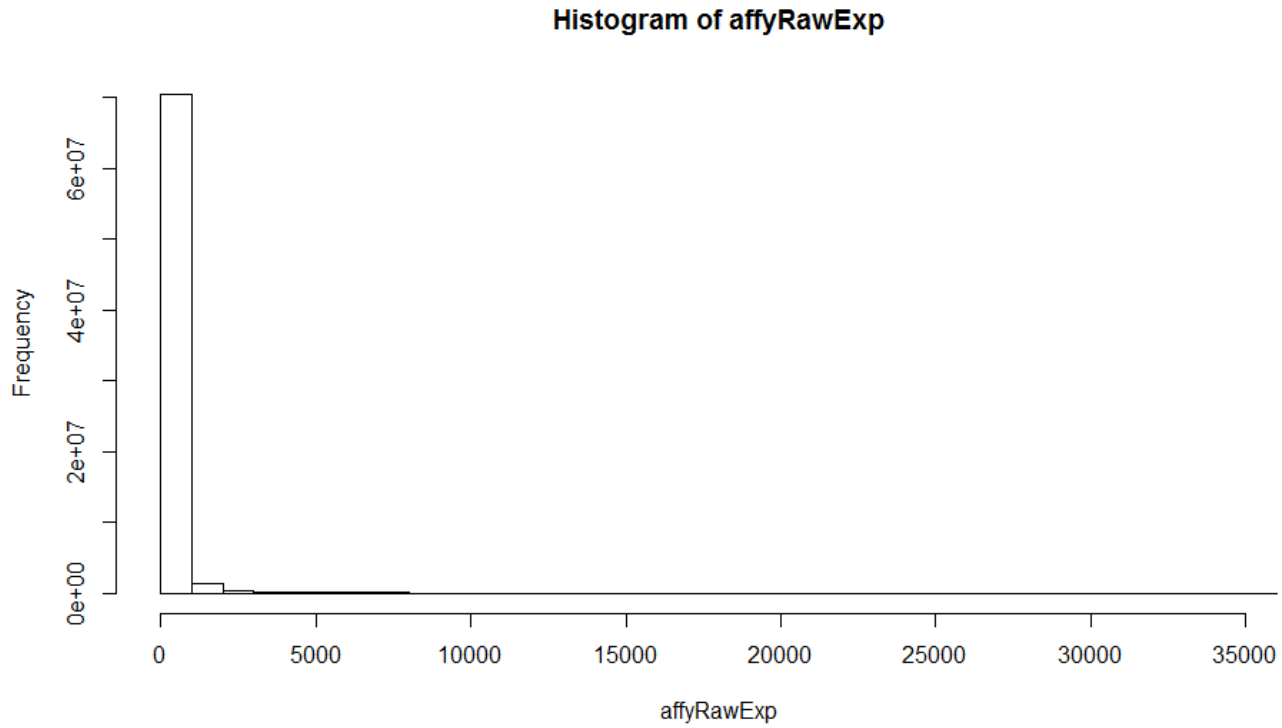


Figura 5. Histograma dels nivells d'expressió.

Reprenent els GeneFeatureSet carregat es normalitza aplicant el mètode rma (que consisteix en aplicar sostracció de fons, normalització de quartils i resum a les dades), de manera que els valors més grans no tinguin més falsa influència quan s'apliquin els algoritmes:

```
> eset  <- rma(affyRaw)
Background correcting
Normalizing
Calculating Expression
```

Es passa el GeneFeatureSet resultant a ExpressionSet i se'n fa l'histograma:

```
> affyRawExp <- ExpressionSet(assayData(eset))
> hist(exprs(affyRawExp))
```

Tal i com es pot observar a la imatge següent, els nivells s'han normalitzat:

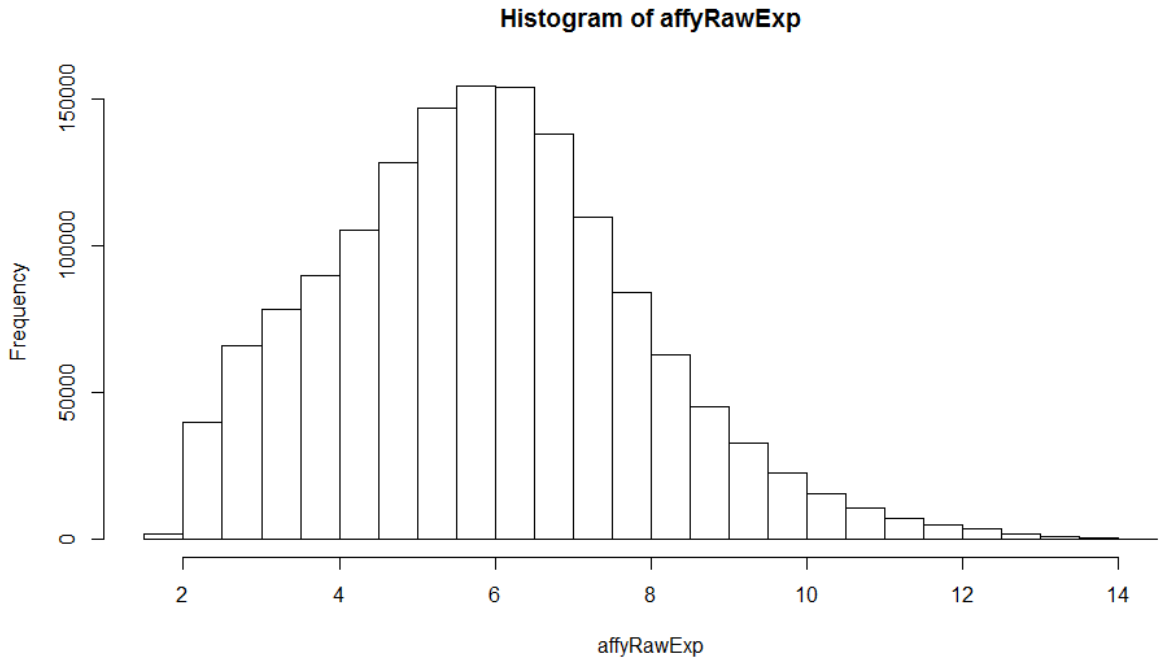


Figura 6. Histograma dels nivells d'expressió normalitzats.

Amb els dades normalitzades es pot repetir la representació dels nivells d'expressió:

```
> image(exprs(affyRawExp),ylab='samples',xlab='expr levels',axes=FALSE)
```

La imatge següent mostra els nivells d'expressió de tots els fitxers i amb els valors normalitzats:

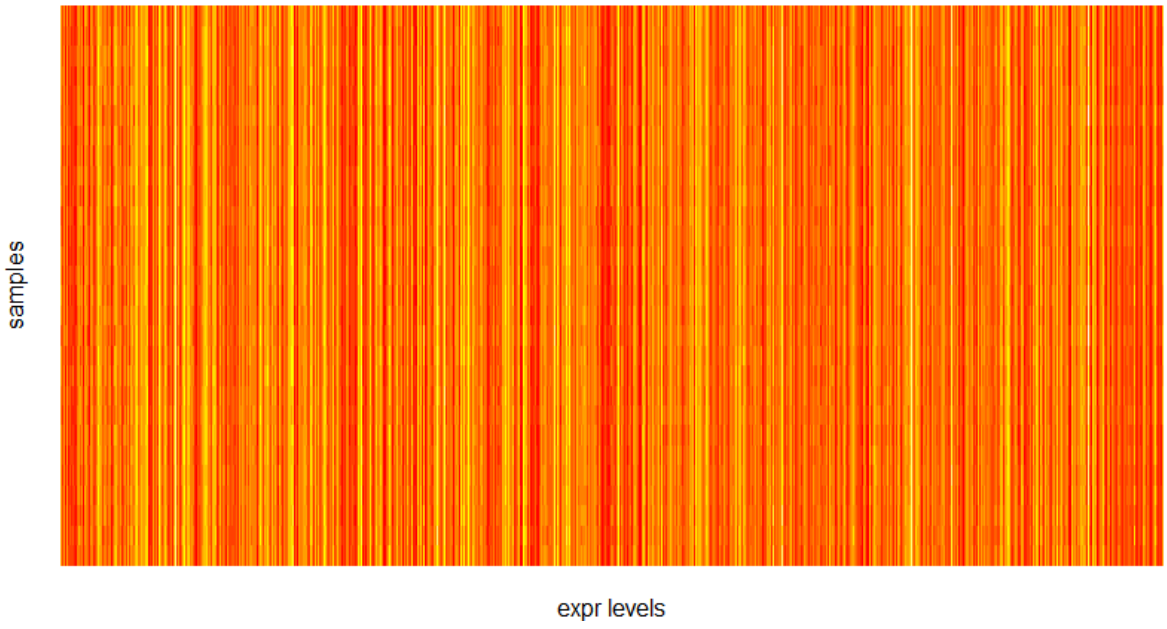


Figura 7. Nivells d'expressió normalitzats de tots els fitxers.

La funció exprs() retorna una matriu que conté, per cada proteïna o lncRNA el valor del nivell d'expressió de tots els individus, de manera que s'obté a la primera columna els ids de les proteïnes o lncRNA i a la resta els valors dels respectius fitxers.

Per tal d'identificar la proteïna o lncRNA concret de cada identificador dels fitxers carregats, és necessari descarregar el fitxer amb les definicions i atributs de la plana web d'affyMetrix. S'ha descarregat el fitxer "HuGene-2_0-st-v1.na33.2.hg19.transcript.csv" amb 53.618 registres i utilitzant el programari Microsoft Excel s'han aplicat diversos filtres textuals per obtenir només els registres corresponents a proteïnes i lncRNA. Concretament s'han descartat els registres que contenen:

- "small nucleolar RNA", "microRNA", "mRNA", "antisense RNA", "intronic transcript 1 (non-protein coding)", "intronic transcript", "testis-specific transcript", la resta amb "non-protein coding" i la resta amb "RNA" al camp "gene_assignment".
- "mRNA", "small nuclear RNA", "miRNA", "snRNA", "microRNA", "snoRNA", "rRNA", "misc_RNA", la resta amb "non-coding RNA", la resta amb "RNA", "antisense" i "transcript" al camp "mrna_assignment".

I agafat els que contenen:

- "long intergenic non-protein coding RNA" i "protein" al camp "gene_assignment".
- "lincRNA", "lncRNA" i protein al camp "mrna_assignment".

Els filtres s'han obtingut a través d'observacions aleatòries del fitxer i del web "Annotation of ncRNAs".

El fitxer resultant "idsProteineslncRNA.csv" conté 9.856 registres dels quals 8.656 són lncRNA i 1.200 són proteïnes. Només s'han desat els camps "probesetid" i un de generat que indica si el registre es correspon amb una proteïna "protein" o un lncRNA "lncRNA". S'ha desat una altra versió del fitxer que també conté els camps "gene_assignment" i "mrna_assignment".

L'arxiu resultant s'ha importat a RStudio utilitzant l'assistent d'importació de fitxers de text.

Partint de la matriu que conté les expressions dels nivells d'expressió normalitzats, s'eliminaran els que no són al fitxer carregat "idsProteineslncRNA.csv":

```
> affyRaw <-exprs(ExpressionSet(assayData(affyRaw)))
> idsProteineslncRNA <-as.matrix(idsProteineslncRNA)
> setwd("C:/Users/Admin/Documents/TFM")
```

```
> source("funcionsTFM.R")
> nivells <-generaMatriuReduida(idsProteineslncRNA,affyRaw)
```

On s'utilitza la funció[2] definida:

```
generaMatriuReduida <- function(idsProteineslncRNA,affyRaw){
  matriuR <-affyRaw
  capE=FALSE
  j <-1
  while(capE==FALSE){
    capE=TRUE
    mida <-dim(matriuR)[1]
    for(i in j:mida){
      if(length(grep(row.names(matriuR)[i], idsProteineslncRNA))==0){
        matriuR <-matriuR[-i,]
        capE=FALSE
        j <-j--
        break
      }
      j <-j+1
    }
  }
  matriuR <-as.matrix(matriuR)
  matriuR
}
```

Amb aquesta reducció la matriu amb els nivells d'expressió es redueix de 53.617 registres a 9.856.

Important dues vegades el mateix fitxer csv però només amb els registres corresponents a proteïnes i la segona vegada només amb els corresponents a lncRNA s'obtindran dues matrius que permetran separar ambdós tipus de nivells d'expressió:

```
> idslncRNA <-as.matrix(idslncRNA)
> idsProteines <- as.matrix(idsProteines)
```

Amb aquest codi s'han generat dues matrius més, una amb els identificadors de les 1.200 proteïnes i una altra amb els dels 8.656 lncRNA.

[2] Totes les funcions que s'han definit es poden trobar al fitxer "funcionsTFM.R" de la carpeta "FitxersTreball" de l'arxiu en el qual es lliura aquest treball.

Pel tractament del fitxer de metadades "fenos_CDV" s'han eliminat les columnes que no corresponien als fenotips a analitzar amb el programari Microsoft Excel, mantenint les següents:

- MomentCursa: Moment de la cura en què s'ha realitzat la mesura. Si conté el valor PRE es refereix a abans de la cursa i POST es correspon amb després.
- GrupActivitat: "A" per individus actius i "E" per esportistes d'elit.
- Distància: Nombre de quilòmetres. En aquest atribut cal tractar els valors absents (es substituiran per la moda 50km[3]) i assignar cada registre a un dels tres grups definits (1-"menys de 40 km", 2-"entre 0 i 60 km" i 3-"més de 60 km").
- Gènere: Gènere de l'individu, "Male" per home i "Female" per dona.

L'arxiu resultant s'ha carregat a RStudio utilitzant l'assistent d'importació de fitxers i s'ha definit la funció "filtreCriteri" per filtrar les mostres dels nivells d'expressió pels diferents fenotips:

```
filtreCriteri <- function (camp,valor,nivells){
 matriuF <-nivells
 capE=FALSE
 while(capE==FALSE){
  capE=TRUE
  if(length(dim(matriuF))>0){
   for(i in 1:dim(matriuF)[2]){
    nomColumna <-strsplit(colnames(matriuF)[i],"_")[[1]]
    id <-nomColumna[1]
    moment <-nomColumna[2]
    seleccio <-subset(Fenos_CDV,Id==id & MomentCursa==moment)
    if(seleccio[camp]!=valor){
     matriuF <-matriuF[,-i]
     capE=FALSE
     break
    }
   }
  }
  else{matriuF<-matrix()}
 }
 matriuF <-as.matrix(matriuF)
 matriuF
}
```

Amb aquesta funció es poden crear les diferents matrius que continguin els nivells d'expressió de les mostres corresponents a un criteri determinat (dones, homes, abans de la cursa, grup d'activitat d'elit, etc.), per exemple per la de dones seria[4]:

[3] Amb la mitjana es substituiria per 42 que seguiria estant al grup d'entre 40 i 60 km.

```
> source("funcionsTFM.R")
> Fenos_CDV <-as.matrix(Fenos_CDV)
> colnames(nivells)[16]<-"577_PRE_(HuGene-2_0-st).CEL"
> dones <-filtreCriteri("Genere", "Female", nivells)
```

Els fitxers importats es poden trobar a la carpeta "fitxersTreball" de l'arxiu en el qual es lliure aquest projecte.

5.4.2 Seqüències de lncRNA

El tractament que es realitzarà a les seqüències de lncRNA es presenten a continuació:

- Deixar només els camps dels fitxers corresponents al nom (i id en el cas de human.fa per aquelles que no tenen nom) i a la seqüència.
- Descartar els registres que no tinguin seqüència (N/A).
- Descartar els registres que corresponguin a lncRNA de ratolí (mouse o mus musculus).
- Passar les seqüències a majúscules.
- Eliminar possibles repeticions de seqüències que hi puguin haver a les BBDD.

Inicialment s'han eliminat els comentaris (línies que començaven per #) dels fitxers que en tenien (human.fa), així com les columnes que no eren el nom o l'id (també del fitxer human.fa i amb el programa Microsoft Office Excel).

Utilitzant el paquet de R Biostrings i el mètode de càrrega readRNAStringSet es carrega el contingut del fitxer en format FASTA indicat, passant automàticament les seqüències en minúscules a majúscules i reconeixent el nom i la llargada de la seqüència, a més de la seqüència en si. Per carregar les dades dels fitxers descartant els lncRNA de ratolins i els que no tenen seqüència, s'ha definit la següent funció:

```
tractamentlncRNA <- function(ruta){
 seqInicial<-readRNAStringSet(ruta, format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
 sequenciaFinal<-RNAStringSet()
 for(i in 1:length(seqInicial)){
  if(width(seqInicial[i])>2){
   if(length(grep("mouse",names(seqInicial[i]),TRUE))==0 && length(grep("mus
musculus",names(seqInicial[i]),TRUE))==0){
     sequenciaFinal<-append(sequenciaFinal,seqInicial[i],after=length(sequenciaFinal))
   }} }
 sequenciaFinal
}
```

[4] S'ha corregit el nom d'un dels fitxers eliminant un "_" sobrant.

Que s'utilitza per carregar les dades dels fitxers lncRNAdbExportSeq.fa i lncRNADiseases.txt (el human.fa i el lncipedia.fasta.txt només contenen lncRNA d'humans, no cal aplicar el filtre per eliminar els ratolins i no tenen seqüències buides). Aquests fitxers es poden consultar a la carpeta "sequenciesIncRNA" de l'arxiu en què es lliura aquest projecte.

Per obtenir un sol conjunt de seqüències sense repeticions (entre diferents fonts ni d'un mateix fitxer) s'usa el mètode union:

```
> source("funcionsTFM.R")
>primera<-readRNAStringSet("TFM/human.fa", format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
>segona<-readRNAStringSet("TFM/lncipedia.fasta.txt", format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
> tercera<-tractamentlncRNA("TFM/lncRNAdbExportSeq.fa")
> quarta<-tractamentlncRNA("TFM/lncRNADiseases.txt")

> length(primera)
[1] 33829

> length(segona)
[1] 21488

> length(tercera)
[1] 114

> length(quarta)
[1] 127

> sequencies<-union(primera,segona,tercera,quarta)
> length(sequencies)
[1] 38943
```

5.4.3 Pathways

En el cas dels pathways no s'utilitzaran les BBDD senceres, sinó que es carregaran pathways determinats. Es començarà utilitzant els pathways relacionats amb la glucosa, ja que és el glúcid que s'utilitza com a font d'energia i s'espera que els nivells de glucosa dels individus de la mostra variïn al llarg de la cursa d'alta exigència. Per carregar un pathway determinat es pot cercar a la plana web de KEGG i obtenir l'URL del fitxer XML amb la seva definició utilitzant el paquet KEGGgraph de R i l'operació getKGMLurl().

Per exemple, per obtenir els pathways ens els quals està involucrada la glucosa, es cerca a la plana principal el terme "glucose" indicant com a organisme "hsa" corresponent a "Homo Sapiens (human)

[hsa]". Si s'accedeix a la informació de detall del primer element, es pot obtenir l'id del pathway, tal i com es ressalta a la imatge següent:

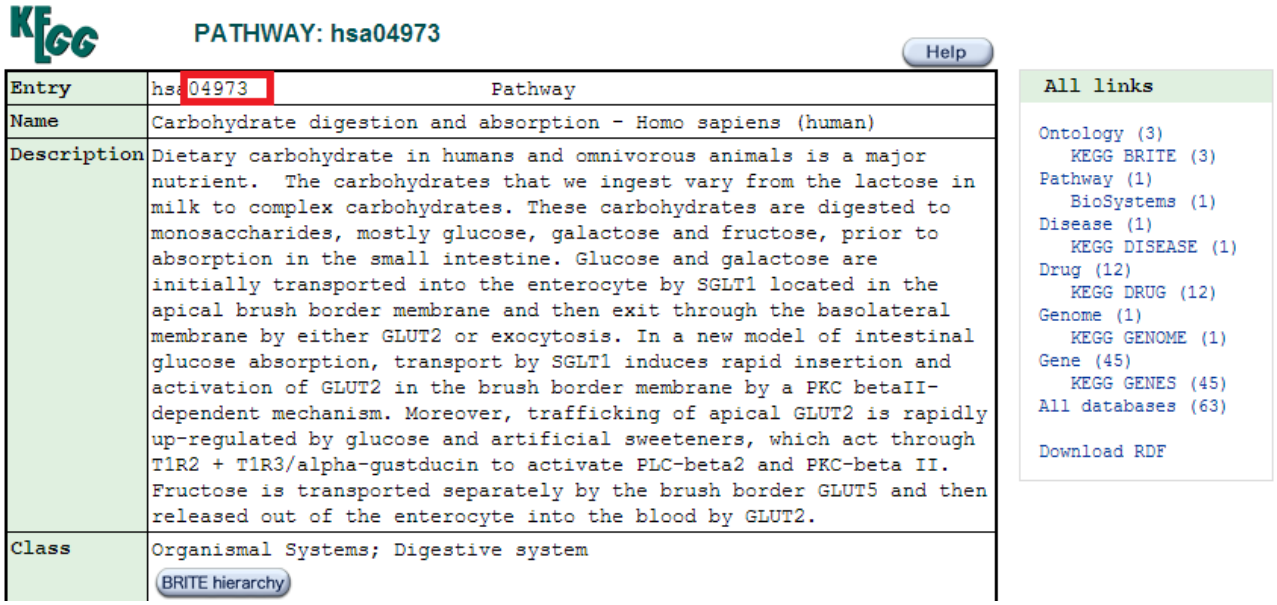


Figura 8. Detall d'informació d'un pathway al web KEGG.

Amb aquest codi es pot utilitzar el mètode prèviament indicat i descarregar el fitxer XML de definició del pathway emmagatzemat a la URL que es retorna:

```
> library(KEGGgraph)
> getKGMLurl("04973", organism = "hsa")
[1] "http://www.genome.jp/kegg-bin/download?entry=hsa04973&format=kgml"
```

El fitxer XML descarregat es desa a la carpeta "extdara" del paquet KEGGgraph: "C:\Program Files\R\R-3.0.2\library\KEGGgraph\extdata".

Per passar-lo a un graf de R es pot usar el codi presentat a continuació:

```
> mapkKGML <- system.file("extdata/hsa04973.xml", package="KEGGgraph")
> map <- parseKGML(mapkKGML)
> map
KEGG Pathway
[ Title ]: Carbohydrate digestion and absorption
[ Name ]: path:hsa04973
[ Organism ]: hsa
[ Number ] :04973
[ Image ] :http://www.kegg.jp/kegg/pathway/hsa/hsa04973.png
[ Link ] :http://www.kegg.jp/kegg-bin/show_pathway?hsa04973
```

```
------------------------------------------------------------
Statistics:
        70 node(s)
        9 edge(s)
        0 reaction(s)
------------------------------------------------------------
```

Aquest paquet també permet realitzar la càrrega de tota la BBDD del KEGG però la comparació amb tots els pathways queda fora de l'abast del present TFM i, a més a més, per tenir-hi accés cal realitzar una subscripció de pagament.

També es pot utilitzar el paquet pathview per descarregar directament el pathway de la BBDD KEGG i formatar-lo. En aquest exemple es repeteix la cerca per glucosa a la plana web i es tria el pathway amb id 04911 (hsa04911):

```
> library(pathview)
> download.kegg(pathway.id = "00010", species = "hsa", kegg.dir = "C:/Program Files/R/R-3.0.2/library/pathview/extdata")
[1] "Downloading xml files for hsa04911, 1/1 pathways.."
[1] "Downloading png files for hsa04911, 1/1 pathways.."
 hsa04911
"succeed"
> xml.file=system.file("extdata", "hsa04911.xml", package = "pathview")
> node.data=node.info(xml.file)
> names(node.data)
 [1] "kegg.names" "type"      "component"  "size"      "labels"    "shape"
 [7] "x"         "y"         "width"     "height"
> data(gse16873.d)
> plot.data.gene=node.map(mol.data=gse16873.d[,1], node.data,node.types="gene")
> head(plot.data.gene)
  kegg.names  labels type   x   y width height   mol.data
2       6513  SLC2A1 gene 174 283    46     17 -0.2675475
3       3767  KCNJ11 gene 688 130    46     17        NA
4        775 CACNA1C gene 836 130    46     17 -0.1771607
5       1131   CHRM3 gene 174 583    46     17        NA
6       2740   GLP1R gene 174 357    46     17 -0.0791128
7       5330   PLCB1 gene 354 583    46     17  0.3581203
```

Per tal d'identificar les proteïnes que pertanyen al pathway a analitzar es pot utilitzar el fitxer descarregat d'AffyMetrix, i obtenir el llistat aplicant filtres de text al camp "pathway".

6 Execució

A continuació es presenta l'aplicació dels algoritmes indicats a l'apartat anterior a les dades seleccionades:

6.1 Agrupació

Amb l'algoritme PCA es simplifiquen les dades a agrupar i analitzar, obtenint un conjunt de dimensionalitat reduïda. Per executar la tècnica s'ha utilitzat la funció pca del paquet de R pcaMethods, la definició de la qual es presenta tot seguit:

```
pca(object, method, nPcs = 2, scale = c("none", "pareto", "vector", "uv"), center = TRUE, completeObs
= TRUE, subset = NULL, cv = c("none", "q2"), ...)
```

I que contempla els paràmetres següents:

- object: Dades a analitzar o simplificar. En aquest cas, s'indica la matriu amb els nivells d'expressió de les proteïnes i els lncRNA.
- method: Mètode del PCA a escollir entre: "svd", "nipals", "rnipals", "bpca", "ppca", "svdImpute", "robustPca" o "nlpca". Es sol utilitzar el svd per matrius sense valors absents i, si n'hi ha, el mètode es tria en funció de l'acció a realitzar per substituir-los. Els mètodes disponibles es poden obtenir amb al funció listPcaMethods() del mateix paquet de R.
- nPcs: Nombre de components principals a calcular. Inicialment es deixa en 28 (nombre de variables de cada registre) per esbrinar quants components es poden eliminar del conjunt de dades inicial.
- scale: Permet indicar si cal escalar les dades abans d'aplicar PCA o no. En aquet cas no és necessari, ja que les dades han estat prèviament normalitzades.
- center: Indica si la matriu ha de ser centrada en la mitjana o no. També es pot indicar un vector amb les mitjanes a utilitzar. Es fixa a TRUE.
- completeObs: TRUE si al resultat final s'inclou un conjunt amb totes les observacions inicials havent substituït els valors absents. S'indica a TRUE.
- subset: Subconjunt de variables per calcular el model. Poden ser noms de columnes o índexs. Es deixa a NULL per no indicar-ne.
- cv: Tipus de validació creuada que es durà a terme. No es realitza deixant el valor a "none", ja que en alguns algoritmes produeix errors i ens els que funciona no varia el resultat.

Al codi següent s'executa la funció pca usant els mètodes "svd":

```
> library (pcaMethods)

> pcaComp <-pca(nivells, "svd", 28, "none", TRUE, TRUE, NULL, "none")

> pcaComp
svd calculated PCA
Importance of component(s):
      PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9    PC10    PC11    PC12
R2  0.968 0.00611 0.00316 0.0024 0.00173 0.00145 0.00137 0.0013 0.00114 0.00111 0.00107 0.00102
Cumulative  R2  0.968  0.97414  0.97730  0.9797  0.98143  0.98288  0.98425  0.9856  0.98669  0.98780
0.98887 0.98989
           PC13    PC14    PC15    PC16  PC17    PC18    PC19    PC20  PC21    PC22    PC23    PC24
R2           0.0009 0.00089 0.00076 0.00075 0.0007 0.00069 0.00068 0.00064 0.0006 0.00057 0.00055
0.00053
Cumulative  R2  0.9908  0.99168  0.99244  0.99319  0.9939  0.99458  0.99526  0.99590  0.9965  0.99707
0.99762 0.99815
           PC25    PC26    PC27    PC28
R2          0.0005 0.00048 0.00044 0.00043
Cumulative R2 0.9987 0.99913 0.99957 1.00000
28         Variables
9856       Samples
0          NAs ( 0 %)
28         Calculated component(s)
Data was mean centered before running PCA
Data was NOT scaled before running PCA
Scores structure: [1] 9856   28
Loadings structure: [1] 28 28
```

S'ha repetit l'execució de la funció utilitzant els altres mètodes i s'ha triat finalment el svd. Com que no hi ha valors absents a la matriu sobre la qual aplicar el PCA, els resultats d'utilitzar un mètode o un altre no han de variar massa. La taula següent resumeix els resultats obtinguts:

Mètode	Acumulat amb PCA1	Component amb què s'arriba al 99% (o màxim) i valor	
svd	96,8%	13	99,08%
nipals	96,8%	13	99,07%
rnipals	96,8%	13	99,07%
ppca	96,8%	13	99,07%
svdImpute	96,8%	13	99,07%
robustPca	96,8%	1	96,8%
nlpca[5]	15,49%	11	96,06%
bpca	96,8%	16	99,02%

[5] S'ha limitat a 2000 iteracions (afegint a la funció pam el paràmetre maxSteps =2000) per reduir el cost computacional (amb 2000 triga unes 8 hores). Com a molt arriba al 96,06% al component 11 i decreix.

L'execució es repeteix amb 13 components (amb les qual s'arriba al 99%):

```
> pcaComp <-pca(nivells, "svd", 13, "none", TRUE, TRUE, NULL, "none")

> pcaComp
svd calculated PCA
Importance of component(s):
      PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9    PC10    PC11    PC12
R2  0.968 0.00611 0.00316 0.0024 0.00173 0.00145 0.00137 0.0013 0.00114 0.00111 0.00107 0.00102
Cumulative  R2  0.968  0.97414  0.97730  0.9797  0.98143  0.98288  0.98425  0.9856  0.98669  0.98780
0.98887 0.98989
           PC13
R2          0.0009
Cumulative R2 0.9908

28        Variables
9856      Samples
0         NAs ( 0 %)
13        Calculated component(s)
Data was mean centered before running PCA
Data was NOT scaled before running PCA
Scores structure:
[1] 9856   13
Loadings structure:
[1] 28 13
```

S'utilitza l'algoritme PAM per agrupar els nivells d'expressió de lncRNA i de proteïnes amb l'objectiu de trobar indicis d'incidències dels fenotips següents:

- Moment de la cursa: agrupació en dos clústers i representació de PRE (abans de la cursa) i POST (després).
- Gènere: agrupació en dos clústers i representació d'homes i dones.
- Distància: agrupació en tres clústers i representació de menys de 40 km, entre 40 i 60 km i de més de 60 km.
- Grup de nivell d'activitat: agrupació en dos clústers i representació de si són actius/molt actius o bé si són esportistes d'elit.

Donada la definició de la funció pam del paquet de R cluster:

```
pam(x, k, diss = inherits(x, "dist"), metric = "euclidean",  medoids = NULL, stand = FALSE, cluster.only = FALSE, do.swap = TRUE, keep.diss = !diss && !cluster.only && n < 100, keep.data = !diss && !cluster.only, pamonce = FALSE, trace.lev = 0)
```

L'algoritme presenta els paràmetres següents:

- x: matriu amb les dades a agrupar. En aquest cas s'utilitza la matriu dels nivells d'expressió de proteïnes i lncRNA normalitzats.
- k: nombre de clústers a crear. S'aplica l'algoritme per generar una agrupació amb 2 clústers i una altra amb 3 clústers (corresponents al nombre de valors dels fenotips a analitzar).
- diss: permet indicar si es tracta d'una matriu de dissimitud (TRUE) o d'una matriu d'observacions i variables (FALSE). En aquest cas es correspon amb una matriu d'observacions i es deixa a FALSE.
- metric: mètrica a utilitzar per calcular les dissimilituds entre els punts en cas que es tracti d'una matriu d'observacions de variables. Inicialment s'utilitzen ambdós mètodes per poder-los comparar.
- menoids: medoides inicials a considerar (o null si es calcularan durant l'execució). En aquest cas es deixa a null.
- stand: permet indicar si cal o no estandarditzar les dades abans d'aplicar l'algoritme. Es fixa a FALSE, donat que les dades ja han estat prèviament normalitzades.
- cluster.only: TRUE si només es computa i retorna el clustering en sí. S'indica a FALSE per incloure'ls.
- do.swap: Indica si es duu a terme la fase d'intercanvi o no. No es duu a terme per reduir el cost computacional.
- keep.diss, keep.data: Permet indicar si la matriu amb les dades a agrupar es manté o no al resultat. Es fixa a TRUE per mantenir-la.
- pamonce: Permet indicar dreceres als algoritmes, en aquest cas es deixa el valor per defecte FALSE.
- trace.lev: Estableix el nivell de traçabilitat en funció de la quantitat d'informació a mostrar per pantalla. Es deixa el valor per defecte 0 (i s'augmentaria si l'algoritme no finalitzés correctament i no es trobés l'error).

El codi utilitzat per agrupar les mostres en dos i tres clústers es presenta a continuació (parteix de les puntuacions dels components principals calculats amb PCA):

```
> library (cluster)
> clas2 <-pam(scores(pcaComp), 2, FALSE, "manhattan", NULL, FALSE, FALSE, FALSE, TRUE)
> clas3 <-pam(scores(pcaComp), 3, FALSE, "manhattan", NULL, FALSE, FALSE, FALSE, TRUE)
> clas2E <-pam(scores(pcaComp), 2, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
> clas3E <-pam(scores(pcaComp), 3, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
> library(fpc)
```

```
> plotcluster(scores(pcaComp)[,1:2],clas2$cluster)
> plotcluster(scores(pcaComp)[,1:2],clas3$cluster)
> plotcluster(scores(pcaComp)[,1:2],clas2E$cluster)
> plotcluster(scores(pcaComp)[,1:2],clas3E$cluster)
```

Les imatges següents mostren representacions dels dos components principals de les agrupacions en:

- Dos clústers:

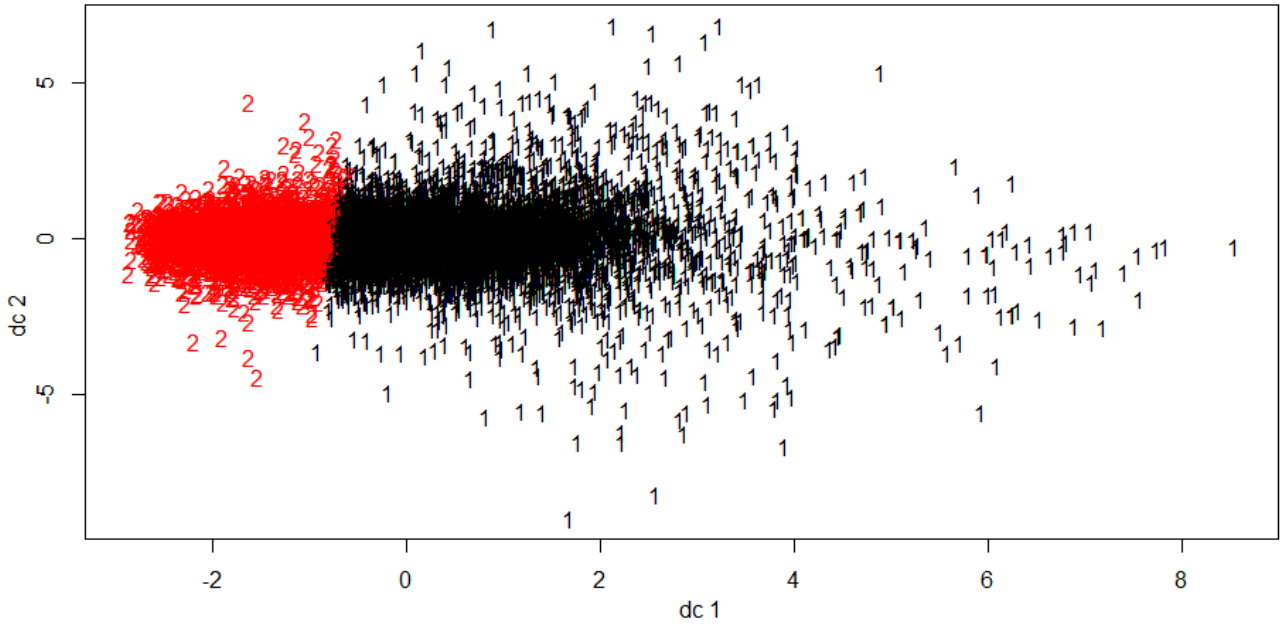


Figura 9. Dos clústers amb el mètode Manhattan.

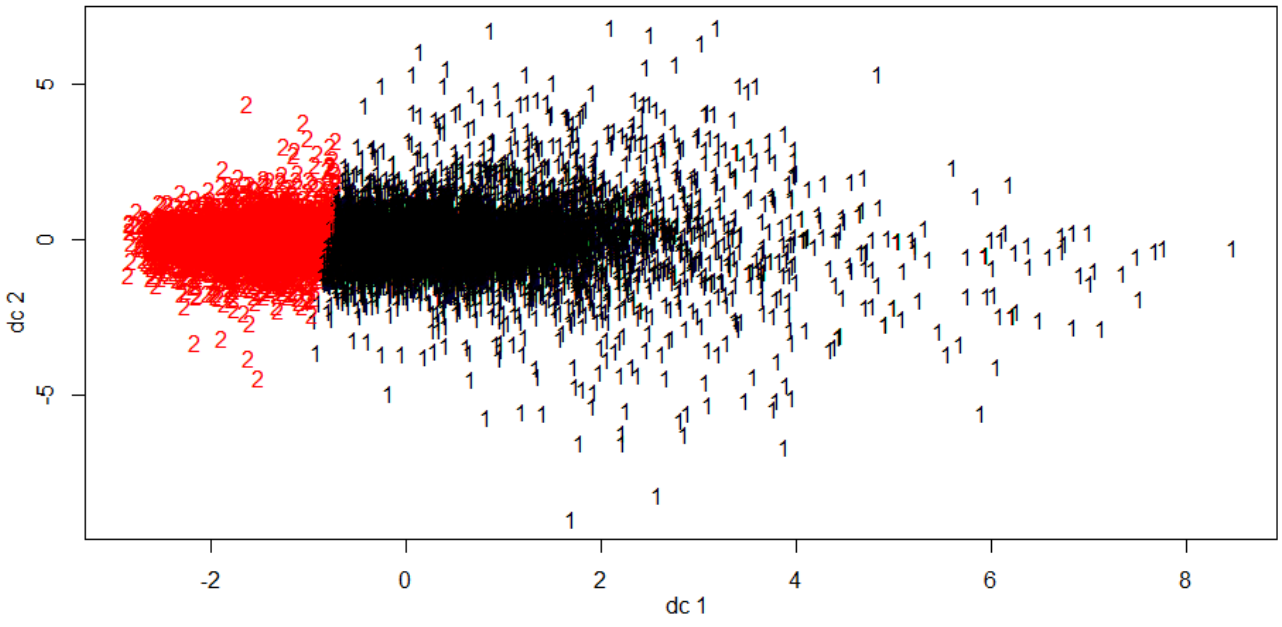


Figura 10. Dos clústers amb el mètode Euclidean.

- Tres clústers:

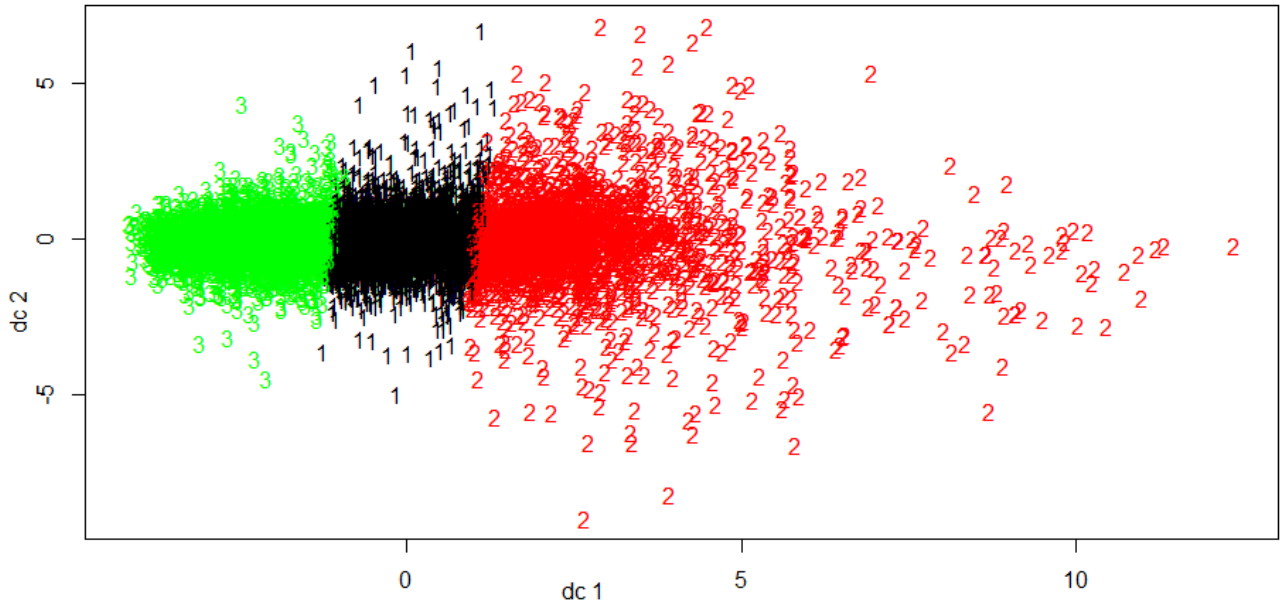


Figura 11. Tres clústers amb el mètode Manhattan.

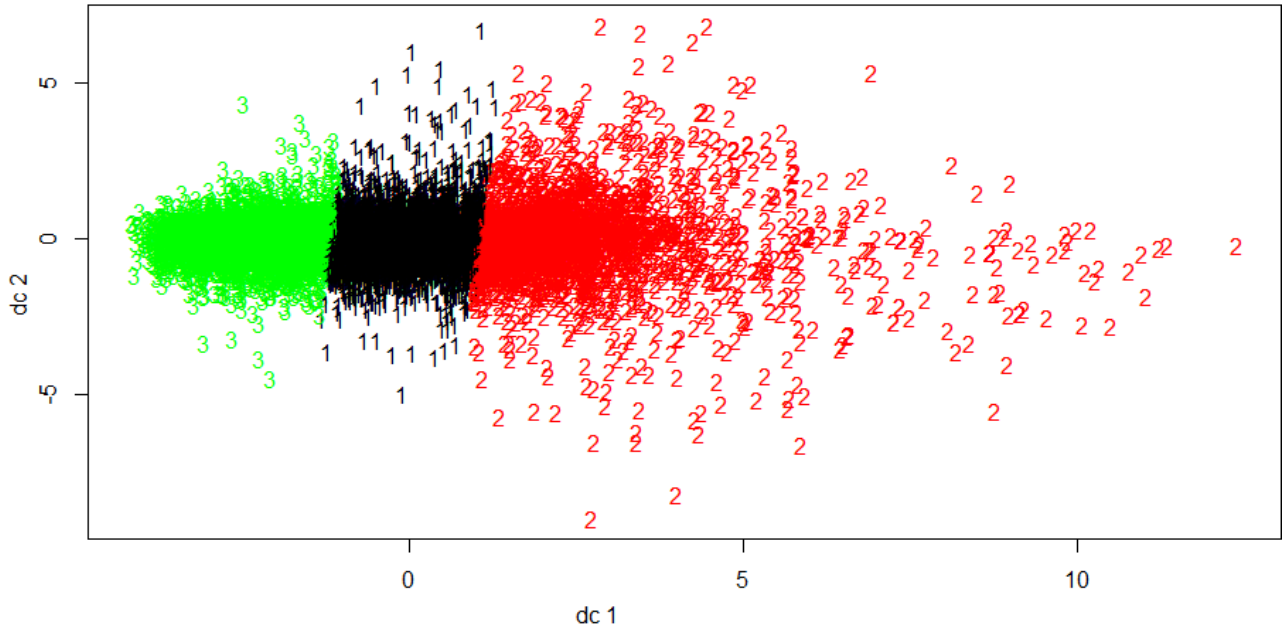


Figura 12. Tres clústers amb el mètode Euclidean.

```
> clusplot(scores(pcaComp)[,1:2],clas2$clustering,color=TRUE)
> clusplot(scores(pcaComp)[,1:2],clas3$clustering,color=TRUE)
> clusplot(scores(pcaComp)[,1:2],clas2E$clustering,color=TRUE)
> clusplot(scores(pcaComp)[,1:2],clas3E$clustering,color=TRUE)
```

- Dos clústers:

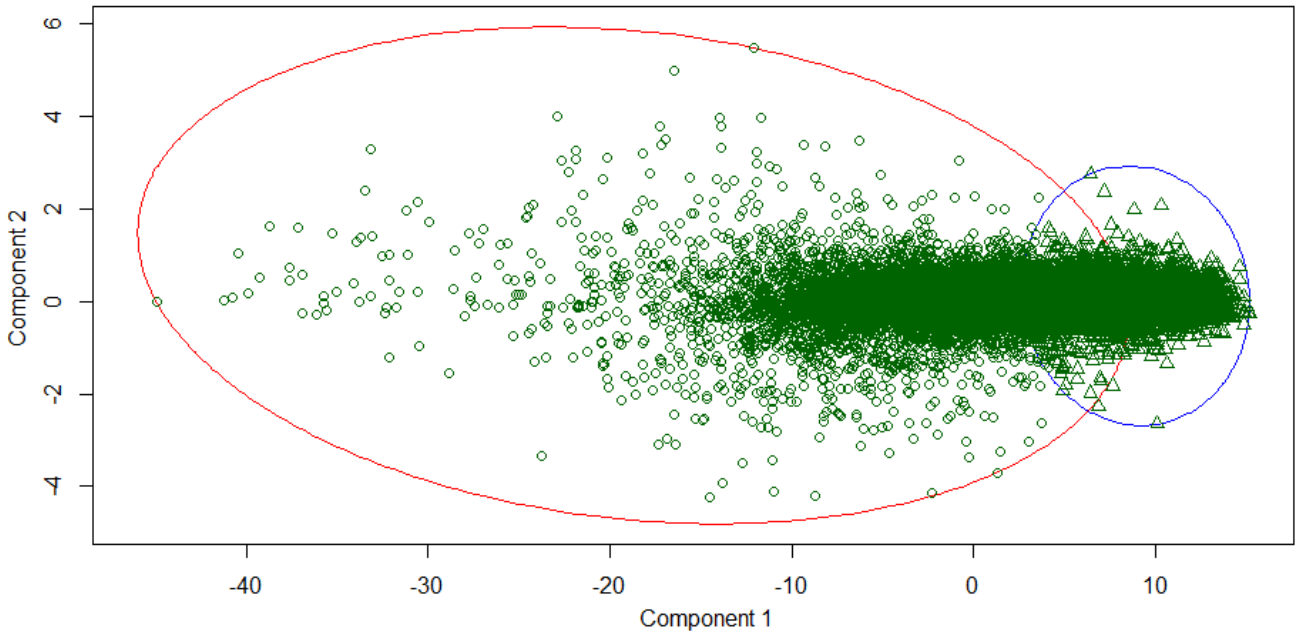


Figura 13. Dos clústers amb el mètode Manhattan.

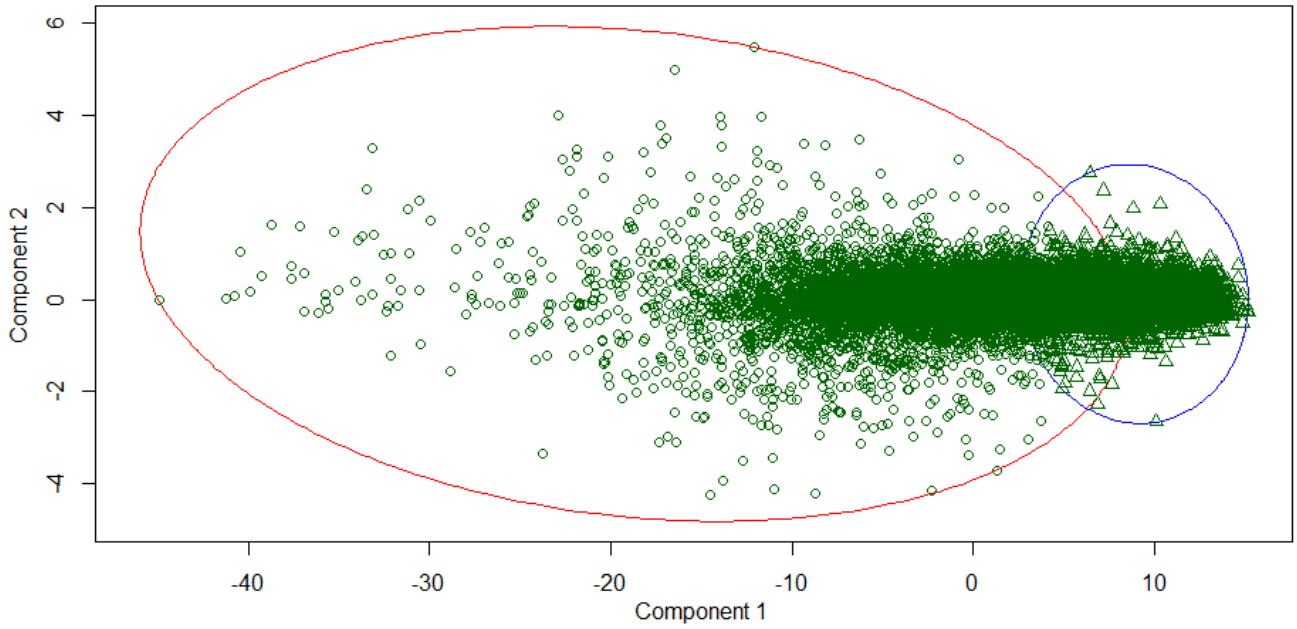


Figura 14. Dos clústers amb el mètode Euclidean.

- Tres clústers:

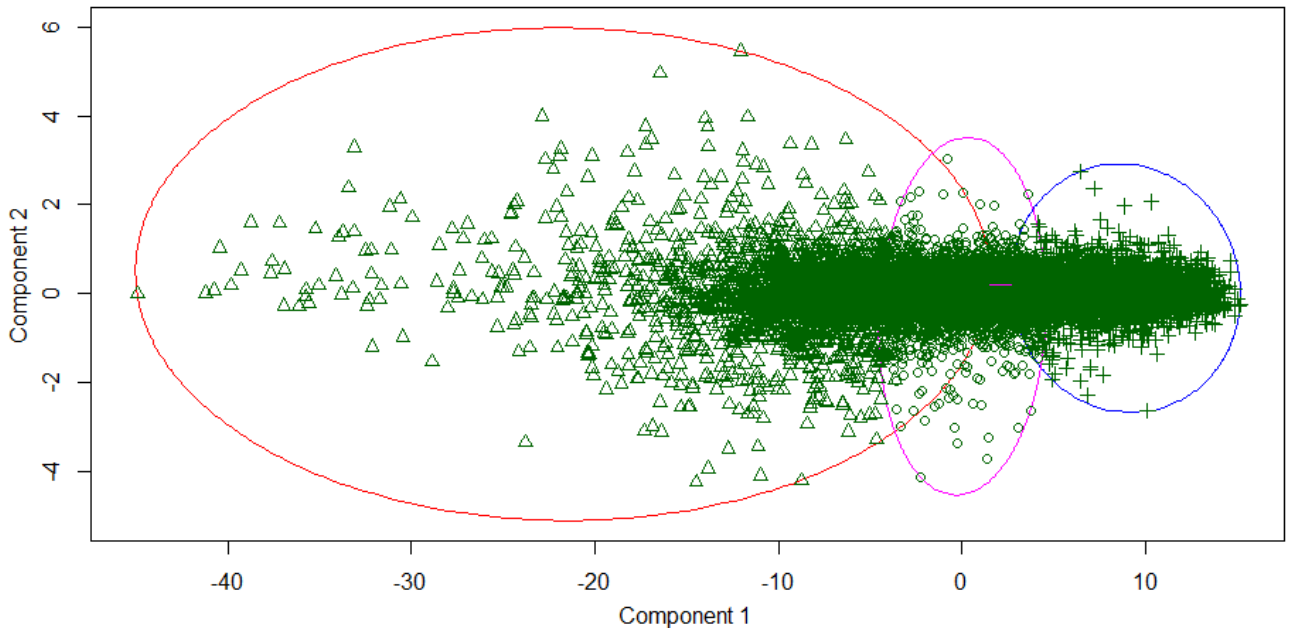


Figura 15. Tres clústers amb el mètode Manhattan.

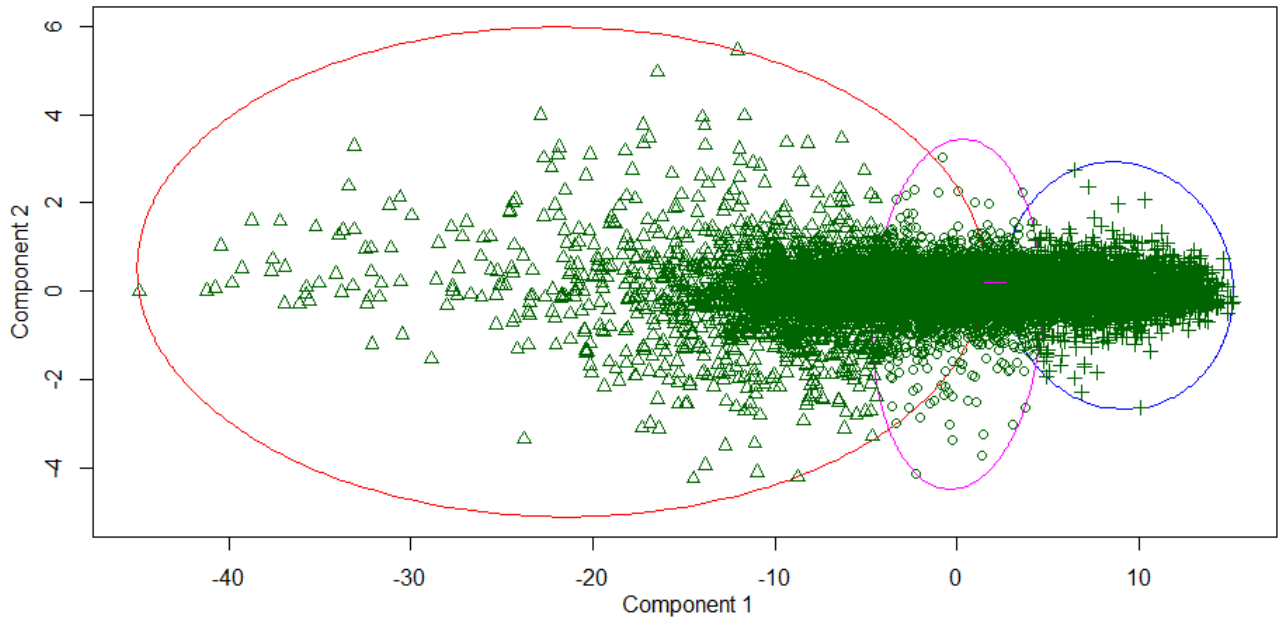


Figura 16. Tres clústers amb el mètode Euclidean.

Com que els resultats d'utilitzar Euclidean i Manhattan són molt similars (si es calcula la correlació d'ambdós agrupaments en el cas dels dos clústers és de 97,77% i dels de tres 96,87%) [6]es triarà Euclidean degut a que resulta més eficient (es percep més ràpid en execució).

[6] Veure apartat 5.2

Per analitzar la incidència dels fenotips es filtraran els nivells d'expressió per cadascun i es repetirà l'anàlisi PCA, així com l'agrupament amb l'algoritme PAM per comparar-ne els resultats i mitjançant la funció definida "generaAgrupaments[7]":

```
generaAgrupaments <-function(nivells){
 fenotips <-
matrix(c("Dones","Homes","Elit","Actius","Pre","Post","Menys40","Entre40i60","Mes60",
"Genere","Genere",
"GrupActivitat","GrupActivitat","MomentCursa","MomentCursa","Distancia","Distancia","Distancia
","Female", "Male","E","A","PRE","POST","1","2","3",4,10,5,9,7,7,4,8,3), nrow=9, ncol=4)
 colnames(fenotips) <-c("filtre","camp","valor","componentArribaA99%")
 correlacions <-matrix(nrow=0, ncol=4)
 colnames(correlacions) <-c("clusters", "fentorip1", "fenotip2", "correlacioAgrupaments")

 pcaComp <-pca(nivells, "svd", 13, "none", TRUE, TRUE, NULL, "none")
 clas2E <-pam(scores(pcaComp), 2, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
 clas3E <-pam(scores(pcaComp), 3, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
 png("Barreja2C.png",width = 962, height = 553)
 plotcluster(scores(pcaComp)[,1:2],clas2E$cluster, main="Barreja", xlim=c(-15, 15), ylim=c(-10, 10))
 dev.off()
 png("Barreja3C.png",width = 962, height = 553)
 plotcluster(scores(pcaComp)[,1:2],clas3E$cluster, main="Barreja", xlim=c(-15, 15), ylim=c(-10, 10))
 dev.off()

 for(i in c(1,3,5,7)){
  j <-i+1
  filtre1 <-filtreCriteri(fenotips[i,2], fenotips[i,3], nivells)
  filtre2 <-filtreCriteri(fenotips[j,2], fenotips[j,3], nivells)
  pca1 <-pca(filtre1, "svd", as.integer(fenotips[i,4]), "none", TRUE, TRUE, NULL, "none")
  print(pca1)
  pca2 <-pca(filtre2, "svd", as.integer(fenotips[j,4]), "none", TRUE, TRUE, NULL, "none")
  print(pca2)
  clas12C <-pam(scores(pca1), 2, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
  clas13C <-pam(scores(pca1), 3, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
  clas22C <-pam(scores(pca2), 2, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
  clas23C <-pam(scores(pca2), 3, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
  png(paste(fenotips[i,1],"2C.png", sep = ""),width = 962, height = 553)
  plotcluster(scores(pca1)[,1:2],clas12C$cluster, main=fenotips[i,1], xlim=c(-15, 15), ylim=c(-10, 10))
  dev.off()
  png(paste(fenotips[i,1],"3C.png", sep = ""),width = 962, height = 553)
  plotcluster(scores(pca1)[,1:2],clas13C$cluster, main=fenotips[i,1], xlim=c(-15, 15), ylim=c(-10, 10))
  dev.off()
  png(paste(fenotips[j,1],"2C.png", sep = ""),width = 962, height = 553)
  plotcluster(scores(pca2)[,1:2],clas22C$cluster, main=fenotips[j,1], xlim=c(-15, 15), ylim=c(-10, 10))
  dev.off()
```

[7] La darrera columna de la matriu "fenotips" conté el nombre de components de l'anàlisi PCA amb els quals s'arriba al 99% (comprovat prèviament aplicant la funció pca a cada filtre).

```
   png(paste(fenotips[j,1],"3C.png", sep = ""),width = 962, height = 553)
   plotcluster(scores(pca2)[,1:2],clas23C$cluster, main=fenotips[j,1], xlim=c(-15, 15), ylim=c(-10, 10))
   dev.off()

   correlacions <-rbind(correlacions,c("2",fenotips[i,1],fenotips[j,1],cor.test(clas12C$clustering,
clas22C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   correlacions <-rbind(correlacions,c("2",fenotips[i,1],"Barreja",cor.test(clas12C$clustering,
clas2E$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   correlacions <-rbind(correlacions,c("2",fenotips[j,1],"Barreja",cor.test(clas22C$clustering,
clas2E$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   correlacions <-rbind(correlacions,c("3",fenotips[i,1],fenotips[j,1],cor.test(clas13C$clustering,
clas23C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   correlacions <-rbind(correlacions,c("3",fenotips[i,1],"Barreja",cor.test(clas13C$clustering,
clas3E$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   correlacions <-rbind(correlacions, c("3",fenotips[j,1],"Barreja", cor.test(clas23C$clustering,
clas3E$clustering, alternative="greater",method="pearson")$estimate[[1]]))

   if(i==7){
    k <-j+1
    filtre3 <-filtreCriteri(fenotips[k,2], fenotips[k,3], nivells)
    pca3 <-pca(filtre3, "svd", as.integer(fenotips[k,4]), "none", TRUE, TRUE, NULL, "none")
    print(pca3)
    clas32C <-pam(scores(pca3), 2, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
    clas33C <-pam(scores(pca3), 3, FALSE, "euclidean", NULL, FALSE, FALSE, FALSE, TRUE)
    png(paste(fenotips[k,1],"2C.png", sep = ""),width = 962, height = 553)
    plotcluster(scores(pca3)[,1:2],clas32C$cluster,main=fenotips[k,1], xlim=c(-15, 15), ylim=c(-10, 10))
    dev.off()
    png(paste(fenotips[k,1],"3C.png", sep = ""),width = 962, height = 553)
    plotcluster(scores(pca3)[,1:2],clas33C$cluster,main=fenotips[k,1], xlim=c(-15, 15), ylim=c(-10, 10))
    dev.off()

    correlacions <-rbind(correlacions,c("2",fenotips[k,1],fenotips[i,1],cor.test(clas32C$clustering,
clas12C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
    correlacions <-rbind(correlacions,c("2",fenotips[k,1],fenotips[j,1],cor.test(clas32C$clustering,
clas22C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
    correlacions <-rbind(correlacions,c("2",fenotips[k,1],"Barreja",cor.test(clas32C$clustering,
clas2E$clustering, alternative="greater",method="pearson")$estimate[[1]]))
    correlacions <-rbind(correlacions,c("3",fenotips[k,1],fenotips[i,1],cor.test(clas33C$clustering,
clas13C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
    correlacions <-rbind(correlacions,c("3",fenotips[k,1],fenotips[j,1],cor.test(clas33C$clustering,
clas23C$clustering, alternative="greater",method="pearson")$estimate[[1]]))
    correlacions <-rbind(correlacions,c("3",fenotips[k,1],"Barreja",cor.test(clas33C$clustering,
clas3E$clustering, alternative="greater",method="pearson")$estimate[[1]]))
   }
  }
  correlacions
}
```

Que s'utilitza amb el codi:

```
> source("funcionsTFM.R")
> correlacionsClus <-generaAgrupaments(nivells)
```

La matriu resultant s'ha exportat al fitxer "correlacionsClus.txt" que es pot trobar a la carpeta "fitxersTreball" de l'arxiu en el qual es lliure aquest projecte.

6.2 Correlació Pearson

Per tal de separar els nivells d'expressió de les proteïnes i els dels lncRNA s'utilitza la funció "generaMatriuReduida" presentada prèviament i les dues matrius amb els ids de les proteïnes i dels lncRNA carregades a partir dels fitxers csv. La matriu reduïda es generarà separant les puntuacions dels components PCA de les proteïnes i les dels lncRNA:

```
> proteines <-generaMatriuReduida(idsProteines,scores(pcaComp))
> lncRNA<-generaMatriuReduida(idslncRNA, scores(pcaComp))
```

Havent separat ambdós conjunts de dades, s'aplicarà la funció definida "detectarCorrelacio" per analitzar la correlació de cada proteïna amb cada lncRNA i emmagatzemar les que en tinguin una de superior a 0,9:

```
detectarCorrelacio <- function (proteines,lncRNA){
  mida <-dim(proteines)[1]
  mida2 <-dim(lncRNA)[1]
  resultat <-matrix(nrow=0, ncol=3)
  resultatp <-matrix(nrow=0, ncol=3)

for(i in 1:mida){
    for(j in 1:mida2){
      correlacio <-cor.test(proteines[i,], lncRNA[j,], alternative="greater",method="pearson")
      if(correlacio$estimate[[1]]>0.9){
        teCor <-c(rownames(proteines)[i],rownames(lncRNA)[j],correlacio$estimate[[1]])
        resultatp<-rbind(resultatp, teCor)
      }
    }
    if(i%%100==0){
      resultat<-rbind(resultat, resultatp)
      resultatp <-matrix(nrow=0, ncol=3)
    }
    print(i)
  }
  resultat <-as.matrix(resultat)
  resultat
}
```

Que es crida amb el codi següent:

```
> source("funcionsTFM.R")
> matriuCorrelacions <-detectarCorrelacio(proteines,lncRNA)
```

La funció detectarCorrelacio utilitza la cor.test del paquet de R stats:

```
cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)
```

Que té com a paràmetres:

- x, y: Estructures de dades a comparar que han de tenir la mateixa mida. En aquest cas es fixa una fila de la matriu de proteïnes i una altra de la matriu de lncRNA i es van comparant totes les proteïnes amb tots els lncRNA.
- alternative: Hipòtesi alternativa a escollir entre “greater”, “less” o “two.side”. Es fixa a més gran per indicar associació positiva.
- method: Mètode amb el qual es calcularà la distància en la correlació. Es fixa a Pearson com ja s’havia indicat als apartats anteriors.
- exact: Booleá que indica si es calcula el valor exacta de p-value o no. Es deixa a null per no augmentar el cost computacional (s’utilitzarà el valor cor del resultat, no el p-value).
- conf.level: Nivell de confiança de l’interval de confiança retornat. Es deixa el valor per defecte 0,95 que ja es percep prou acurat.
- continuity: Booleá que permet indicar si s’utilitza una correcció de continuïtat pels mètodes “Kendall” i “Spearman”. Es deixa a False perquè s’utilitza Pearson.
- formula: Fórmula que proporcioni els valors de les mostres, les mostres han de ser de la mateixa mida. No s’utilitza, donat que es passen les dades directament.
- data: Matriu opcional que contingui les variables de la fórmula anterior. En aquest cas no s’utilitza.
- subset: Vector de dades que especifica un subconjunt de les dades a utilitzar. No s’utilitza.
- na.action: Funció que indica l’acció a realitzar si hi ha valors absents. No s’utilitza degut a que no n’hi han.

La correlació de Pearson també s’ha utilitzat per comparar els agrupaments fets a l’apartat anterior en la funció “generaAgrupaments” i en la comparació dels agrupaments amb el mètode Manhattan i Euclidean:

```
> cor.test(clas2E$clustering,clas2$clustering, alternative="greater",method="pearson")
> cor.test(clas3E$clustering,clas3$clustering, alternative="greater",method="pearson")
```

6.3 Anàlisi de Components Independents

Com s'ha indicat a l'apartat anterior, s'utilitzarà el mètode ICA per construir el metafenotip del pathway de la glucosa i analitzar la incidència que hi té l'esforç d'alta intensitat. Per assolir aquest objectiu es seguirà la mateixa metodologia que a la a la tesi doctoral "Genetic association analysis of complex diseases through information theoretic metrics and linear pleiotropy":

1. Es seleccionarà el conjunt de fenotips del pathway per dur a terme l'anàlisi.
2. Es construirà el metafenotip.
3. Es farà un anàlisi general amb les noves variables.

Els nivells d'expressió amb els quals es construirà el metafenotip són els corresponents a les proteïnes que formen part del pathway de la glucòlisi / gluconeogènesi i que tenen una alta correlació amb els lncRNA.

Per obtenir el llistat de les proteïnes que formen part del pathway es filtra el camp "pathway" seleccionant els registres que continguin el literal "Glycolysis_and_Gluconeogenesis" del fitxer descarregat d'AffyMetrix amb les anotacions de tots els probesets. A la carpeta "FitxersTreball" del fitxer en el qual es lliure aquest projecte es pot trobar una versió de la llista obtinguda amb els principals camps descriptius emmagatzemada a l'arxiu "proteines_Glycolysis_and_Gluconeogenesis.xsl". La relació carregada a Rstudio amb l'assistent d'importació de fitxers de text es proporciona a la mateixa carpeta amb el nom "IDsGlycolysis_and_Gluconeogenesis.csv" i només conté els identificadors (probesetids):

```
> IDsGlycolysis_and_Gluconeogenesis <-as.matrix(IDsGlycolysis_and_Gluconeogenesis)
```

Es calculen les correlacions de les proteïnes del pathway amb els lncRNA i s'emmagatzemen les que tinguin un valor superior a 0,9[8]:

```
> source("TFM/funcionsTFM.R")
> pathwaylncRNA <-as.matrix(pathwaylncRNA)
> nivellspathwaylnc <-generaMatriuReduida(pathwaylncRNA,affyRaw)
> pcaPathwaylncRNA <-pca(nivellspathwaylnc, "svd", 14, "none", TRUE, TRUE, NULL, "none")
> sL<-generaMatriuReduida(idslncRNA, scores(pcaPathwaylncRNA))
> sP<-generaMatriuReduida(IDsGlycolysis_and_Gluconeogenesis, scores(pcaPathwaylncRNA))
> matriuCorrelacionsPathways <-detectarCorrelacio(sP,sL)
```

S'obtenen 135.907 registres.

[8] Als anàlisis PCA es seleccionen 14 components, donat que són els necessaris per arribar al 99% (comprovat prèviament).
S'ha modificat la funció detectarCorrelacio traient la fragmentació de la matriu de resultats final de 100 en 100.

La imatge següent mostra el pathway al que pertanyen les proteïnes correlacionades "Glycolysis / Gluconeogenesis - Homo sapiens (human)":

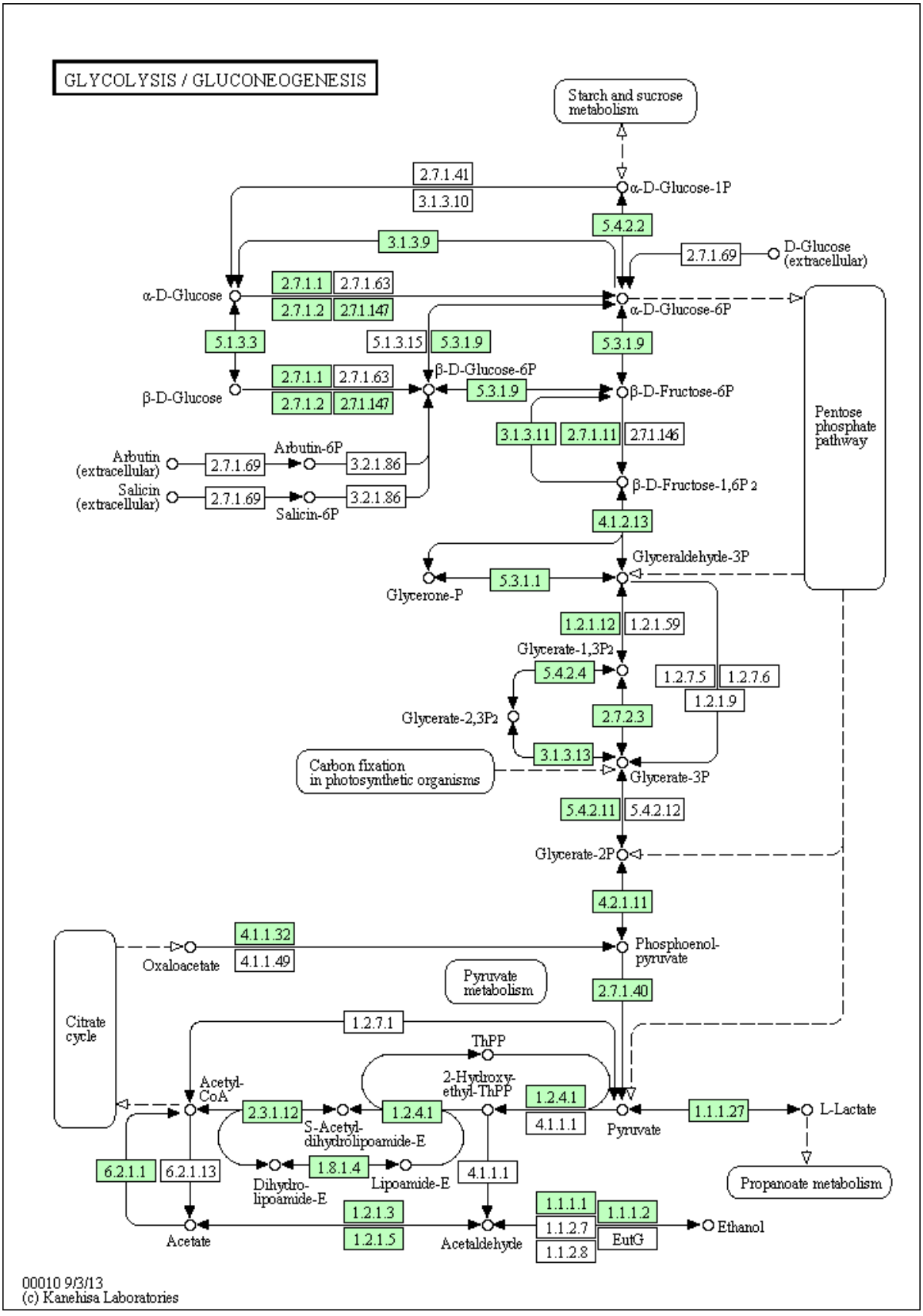


Figura 17. Glycolysis/Gluconeogenesis

Aplicant el tractament de dades desenvolupat a l'apartat anterior i mostrant el contingut de la variable map se'n obté la informació:

```
KEGG Pathway
[ Title ]: Glycolysis / Gluconeogenesis
[ Name ]: path:hsa00010
[ Organism ]: hsa
[ Number ] :00010
[ Image ] :http://www.kegg.jp/kegg/pathway/hsa/hsa00010.png
[ Link ] :http://www.kegg.jp/kegg-bin/show_pathway?hsa00010
------------------------------------------------------------
Statistics:
        96 node(s)
        84 edge(s)
        34 reaction(s)
------------------------------------------------------------
```

Es seleccionen els nivells d'expressió de les proteïnes del pathway amb una correlació superior a 0,999 amb els lncRNA (162 resultats corresponents a 19 proteïnes):

```
> nivellsCP0999 <-subset(nivellspathwaylnc,rownames(nivellspathwaylnc) %in% subset(matriuCorrelacionsPathways,matriuCorrelacionsPathways[,3]>0.999)[,1])
```

Per construir el metafenotip s'aplica l'algoritme ICA als nivells obtinguts corresponents a proteïnes correlacionades amb els lncRNA i pertanyents al pathway.

La funció fastICA està definida de la següent manera:

```
fastICA(X, n.comp, alg.typ = c("parallel","deflation"), fun = c("logcosh","exp"), alpha = 1.0, method = c("R","C"), row.norm = FALSE, maxit = 200, tol = 1e-04, verbose = FALSE, w.init = NULL)
```

Que inclou els paràmetres:

- X: La matriu amb les observacions i les variables, en aquest cas, els nivells d'expressió corresponents a les proteïnes correlacionades i del pathway.
- n.comp: Nombre de components a extraure. Es fixa a 4, que és el nombre de components necessaris per arribar al 99% si es fa un anàlisis PCA dels nivells d'expressió de les proteïnes correlacionades.
- alg.typ: Tipus d'algoritme en paral·lel o obtenció d'un component a la vegada. Es deixa el valor per defecte "parallel".

- fun: Forma de la funció d'aproximació a la neguentropia (aproximació a l'ordre previsible partint d'un origen desendreçat i aleatori). Es fixa a "logcosh".
- alpha: Constant utilitzada en l'aproximació a la neguentropia si el paràmetre anterior es fixa a "logcosh". Es deixa el valor per defecte 1.0.
- method: Si la computació es realitza només en R ("R") o en C (C) habilitant una execució més ràpida i utilitzant codi optimitzat. S'utilitza el valor per defecte R, donat que ja va molt ràpid en executar l'algoritme.
- row.norm: Permet indicar si les dades s'han d'estandarditzar abans d'aplicar l'algoritme. Es deixa a FALSE donat que les dades ja han estat prèviament tractades.
- maxit: Nombre màxim d'iteracions a realitzar, es deixa el valor per defecte 200.
- tol: Escalar positiu que indica la tolerància en la qual es considera que la matriu "no barrejada" ha convergit. Es manté el valor per defecte "1e-04".
- verbose: Booleà que determina el nivell d'informació que ofereix l'algoritme per pantalla en temps d'execució. S'utilitza el valor per defecte FALSE i si es produïssin errors es canviaria a TRUE per obtenir més informació.
- w.init: Matriu no barrejada inicial en aquest cas es deixa a null per agafar l'estructura de dades per defecte.

I s'utilitza amb el codi:

```
> library (fastICA)
> resultatICA <-fastICA(nivellsCP0999, 4, fun="logcosh")
```

La funció phyper es defineix de la següent manera:

```
phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
```

On:

- q: Són les proteïnes correlacionades amb els lncRNA que pertanyen al pathway. Concretament, se'n passa la variable S del resultat d'aplicar ICA (la matriu no barrejada).
- m: És el nombre de boles blanques, el nombre de proteïnes del pathway, és a dir, 47.
- n: És el nombre de boles negres, el nombre de boles total considerades en el present estudi menys les boles blanques (47), és a dir, 1200 proteïnes identificades.

- k: nombre de boles extretes de l'urna, en aquest cas les 275 correlacionades amb els lncRNA (19 del pathway més 256 de la resta[9]).
- lower.tail: Booleà que indica si les probabilitats s'expressen com P[X ≤ x] si es fixa a TRUE o, si és FALSE, com P[X > x]. Es deixa el valor per defecte TRUE.
- log.p: Booleà que permet indicar si es retorna una probabilitat p (FALSE) o log(p) (TRUE). Es deixa a FALSE per tal que es retorni la probabilitat i no el logaritme.

El test s'aplica amb el codi següent:

```
> testHipergeometric <-phyper(resultatICA$S, 47,1200,275)
> testHipergeometric
            [,1]        [,2]        [,3]        [,4]
16681370 0.000000e+00 0.000000e+00 6.381950e-06 6.381950e-06
16701841 0.000000e+00 6.381950e-06 6.381950e-06 6.381950e-06
16723593 0.000000e+00 7.010406e-04 6.381950e-06 0.000000e+00
16731169 6.381950e-06 9.546046e-05 6.381950e-06 6.381950e-06
16747338 0.000000e+00 0.000000e+00 6.381950e-06 0.000000e+00
16747529 6.381950e-06 0.000000e+00 9.546046e-05 6.381950e-06
16811372 0.000000e+00 0.000000e+00 6.381950e-06 6.381950e-06
16817790 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
16826985 6.381950e-06 6.381950e-06 0.000000e+00 9.546046e-05
16832438 6.381950e-06 0.000000e+00 0.000000e+00 6.381950e-06
16860709 6.381950e-06 0.000000e+00 0.000000e+00 6.381950e-06
16861033 9.546046e-05 6.381950e-06 9.546046e-05 0.000000e+00
16881838 6.381950e-06 0.000000e+00 0.000000e+00 0.000000e+00
16922993 6.381950e-06 0.000000e+00 0.000000e+00 9.546046e-05
16955511 0.000000e+00 9.546046e-05 6.381950e-06 6.381950e-06
17006378 6.381950e-06 0.000000e+00 0.000000e+00 0.000000e+00
17047461 6.381950e-06 9.546046e-05 0.000000e+00 6.381950e-06
17057218 9.546046e-05 0.000000e+00 9.546046e-05 0.000000e+00
17105047 0.000000e+00 6.381950e-06 6.381950e-06 0.000000e+00
```

[9] De les 1200 proteïnes analitzades inicialment 66 tenen una correlació superior a 0,9995 amb els lncRNA.

7 Resultats obtinguts

7.1 Agrupació

La taula següent mostra els resums dels anàlisis PCA dels nivells d'expressió filtrats pels fenotips:

Gènere				
		Dones	Homes	Barrejat
Nombre de mostres		8	20	28
PCA	% 1er component	97,63%	96,79%	96,8%
	component per arribar al 99%	4	10	13
	valor al component que supera el 99%	99,139%	99,06%	99,08%
Grup d'activitat				
		Elit	Actiu	Barrejat
Nombre de mostres		10	18	28
PCA	% 1er component	97,09%	96,9%	96,8%
	component per arribar al 99%	5	9	13
	valor al component que supera el 99%	99,10%	99,05%	99,08%
Moment de la cursa				
		POST	PRE	Barrejat
Nombre de mostres		12	16	28
PCA	% 1er component	96,83%	97,52%	96,8%
	component per arribar al 99%	7	7	13
	valor al component que supera el 99%	99,06%	99,01%	99,08%

Distància					
		> 60 km	< 40 km	> 40 i <60 km	Barrejat
Nombre de mostres		5	7	16	28
PCA	% 1er component	97,46%	96,93%	97,08%	96,8%
	component per arribar al 99%	3	4	8	13
	valor al component que supera el 99%	99,29%	99,03%	99,11%	99,08%

El fenotip que presenta una major variabilitat és el de distància superior a 60 km (requereix de 3 components per explicar-ne 5) mentre que el d'abans de la cursa pre esdevé el més homogeni (7 components per representar-ne 16).

La taula presentada a continuació mostra les correlacions entre els agrupaments realitzats:

Gènere			
Clústers	Fenotip2	Fenotip2	Correlació agrupaments
2	Dones	Homes	0,955963707643745
2	Dones	Barreja	0,970097120728657
2	Homes	Barreja	0,985049667868279
3	Dones	Homes	0,929711150180189
3	Dones	Barreja	0,949726946510871
3	Homes	Barreja	0,977931353687618
Grup d'activitat			
Clústers	Fenotip2	Fenotip2	Correlació agrupaments
2	Elit	Actius	0,957542516471966
2	Elit	Barreja	0,972426210003841
2	Actius	Barreja	0,984582300127206
3	Elit	Actius	0,933794201134697
3	Elit	Barreja	0,955814058315562
3	Actius	Barreja	0,97709392948154
Moment de la cursa			
Clústers	Fenotip2	Fenotip2	Correlació agrupaments
2	Pre	Post	0,941138317459857
2	Pre	Barreja	0,965191196105094
2	Post	Barreja	0,965560100588764
3	Pre	Post	0,902663853352039
3	Pre	Barreja	0,940198902251357
3	Post	Barreja	0,944393906782428
Distància			
Clústers	Fenotip2	Fenotip2	Correlació agrupaments
2	Menys40	Entre40i60	0,951930279782311
2	Menys40	Barreja	0,960364614660355
2	Entre40i60	Barreja	0,985352805955457
2	Mes60	Menys40	0,936213055925521
2	Mes60	Entre40i60	0,952189569063092
2	Mes60	Barreja	0,962006549692591
3	Menys40	Entre40i60	0,924942095353444
3	Menys40	Barreja	0,938520294625972
3	Entre40i60	Barreja	0,977006257964972
3	Mes60	Menys40	0,896518288674553
3	Mes60	Entre40i60	0,9229082942147
3	Mes60	Barreja	0,936365597176374

La gràfica següent mostra el percentatge de correlació entre agrupaments realitzats per fenotips considerats oposats (més de 60 km i menys de 40, dones i homes, pre i post, menys de 40 km i entre 40 i 60, més de 60 km i entre 40 i 60 km i elit i actiu) per dos i tres grups:

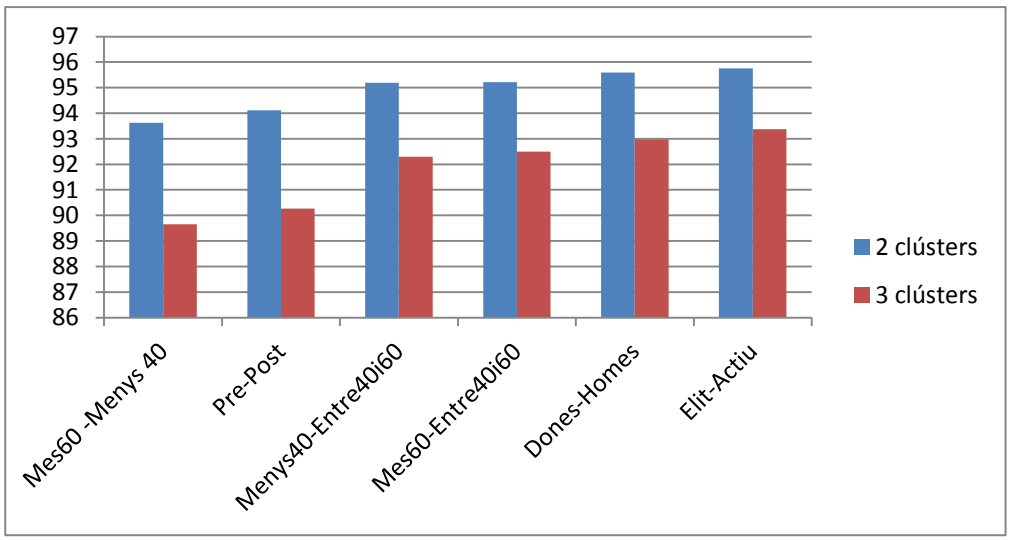


Figura 18. Correlacions entre fenotips oposats.

Independentment del nombre de clústers, els fenotips que impliquen agrupaments més similars són els del grup d'activitat elit-actius (95,75% de correlació en 2 clústers i 93,37% en 3), mentre que les distancies de menys de 40 km i més de 60 proporcionen els agrupaments més diferents (93,62% en 2 grups i 89,65 en 3).

Realitzant la mateixa comparació entre cadascun dels fenotips i els agrupaments realitzats amb tots els nivells d'expressió (barreja):

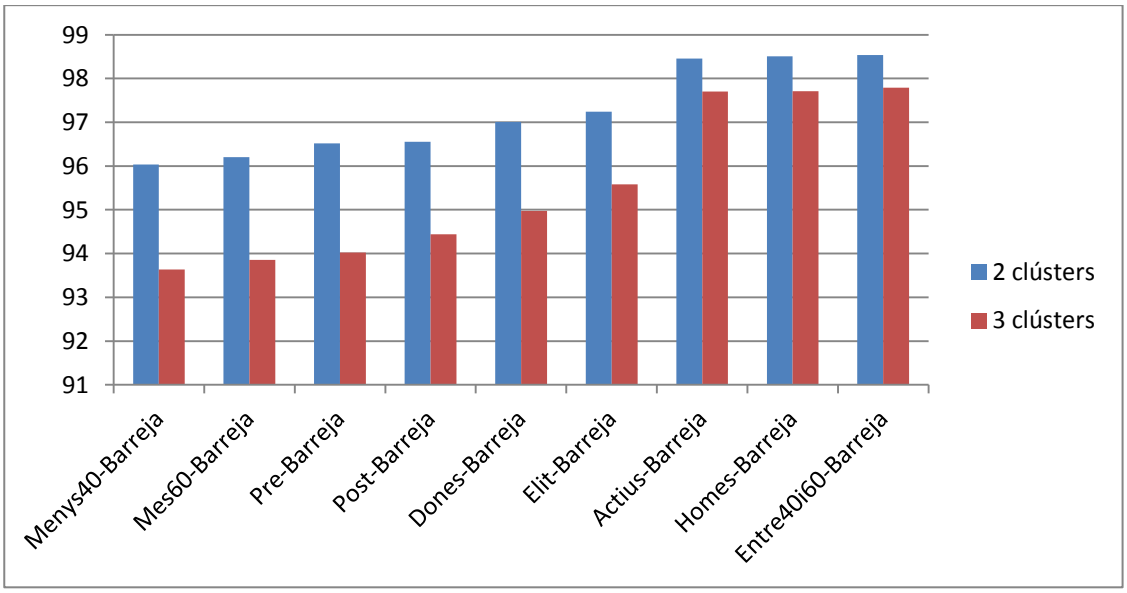


Figura 19. Correlacions entre fenotips i barreja.

Es pot observar que els agrupaments dels nivells d'expressió corresponents a una distància d'entre 40 i 60 km són els més similars als de la barreja en el cas de 2 clústers (98,53%) i els dels homes (97,79%) en el de 3. Els agrupaments que menys s'assemblen als de la barreja són els oferts per distància de menys de 40 km (96,03%) en 2 grups i els de més de 60 km (93,63%) en 3.

La similitud dels agrupaments es veu condicionada pel nombre de mostres de cada fenotip, de manera que s'ha buscat una mesura unitària per comparar-los: dividint la diferència de correlació del fenotip i la barreja entre nombre de mostres (100-correlació en% dividit entre el nombre de mostres):

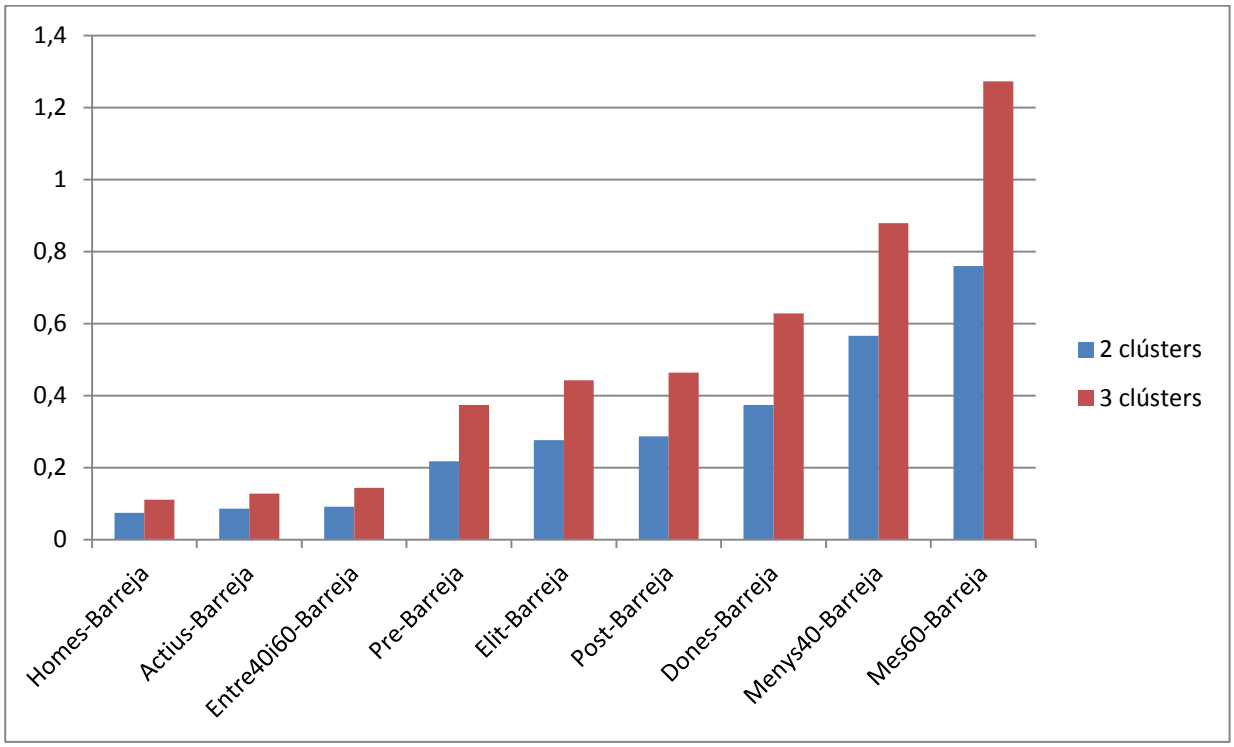


Figura 20. Diferències unitàries entre fenotips i barreja.

Amb aquesta mesura unitària, els agrupaments més similars als de la barreja són els dels homes, tant en 2 clústers com en 3 i els que menys s'assemblen els de distància de més de 60 km.

Les representacions de les diferents agrupacions per fenotip es presenten tot seguit (utilitzant els mateixos rangs dels eixos (x:(-15, 15) i y(-10,10)) per oferir una comparació més clara):

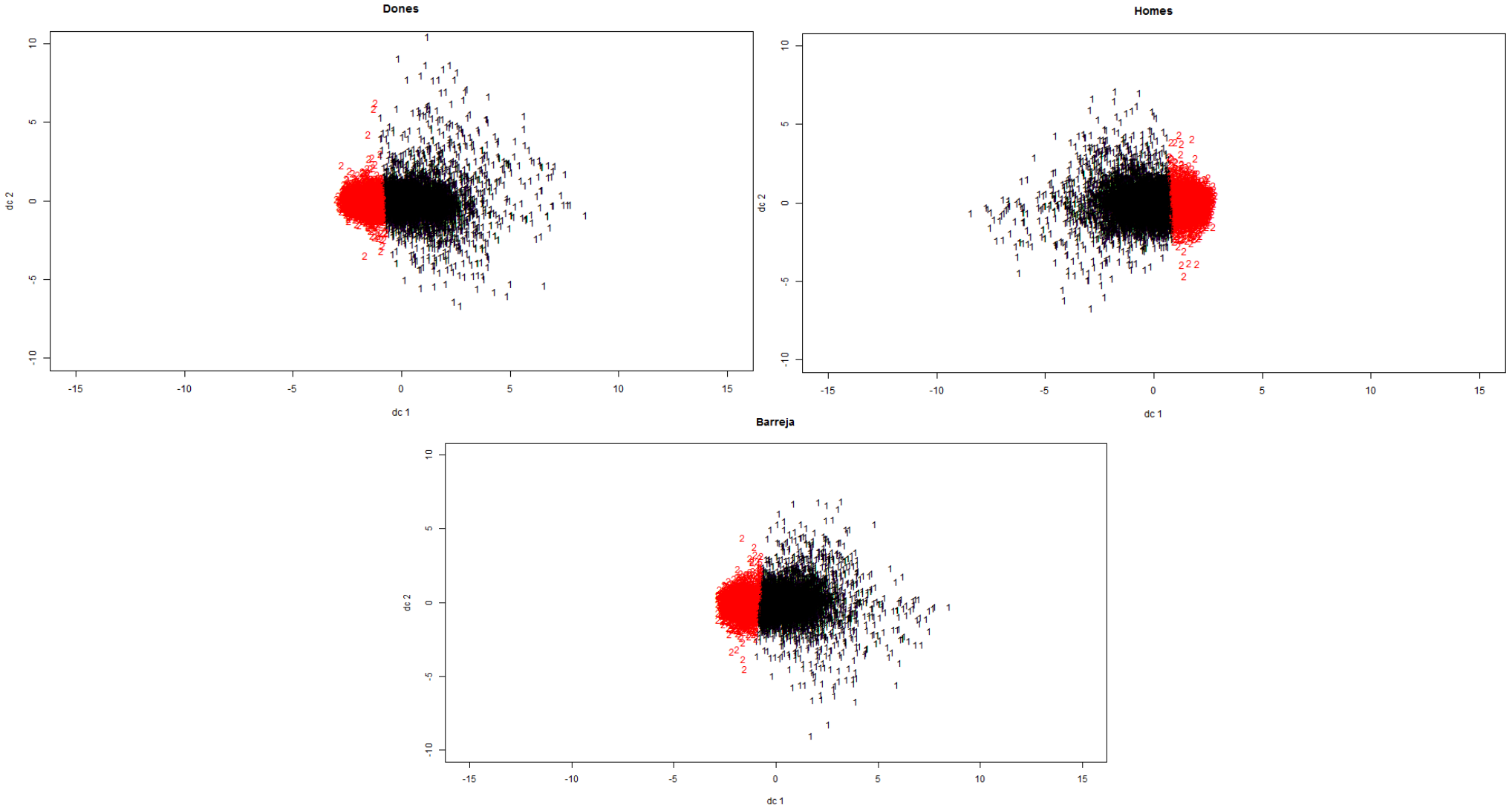


Figura 21. Comparativa dels agrupaments en dos clústers de dones, homes i ambdós.

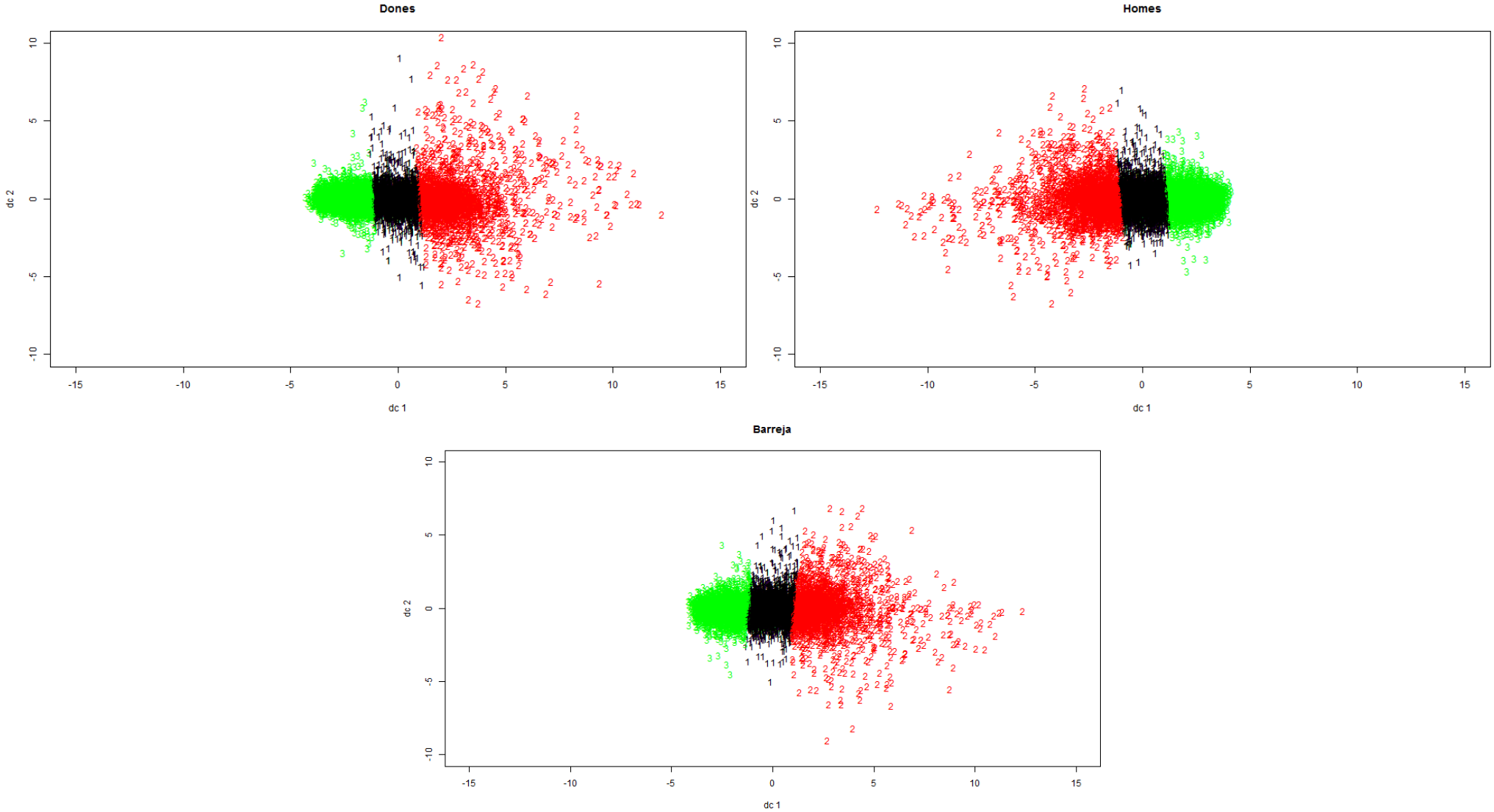


Figura 22. Comparativa dels agrupaments en tres clústers de dones, homes i ambdós.

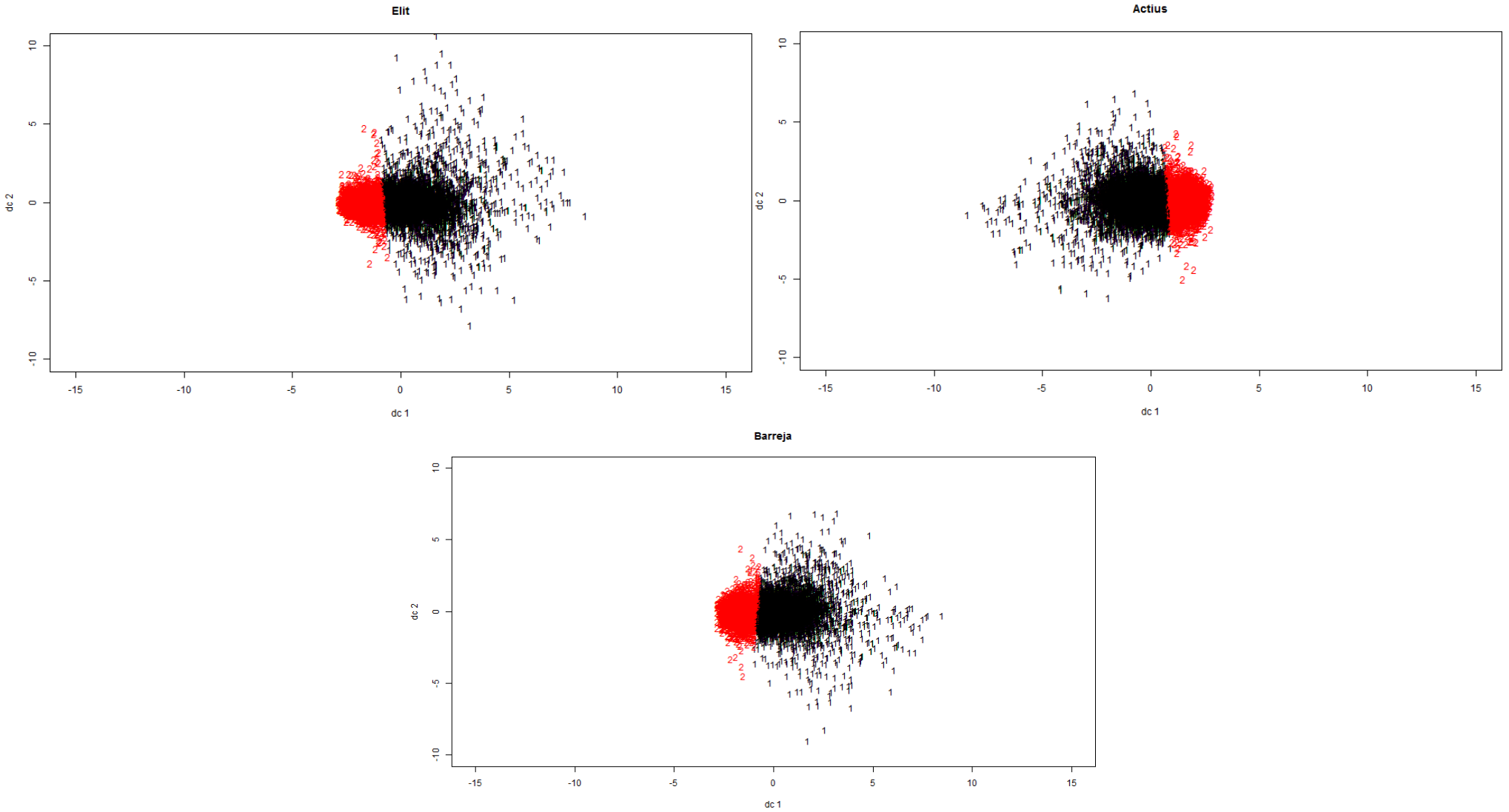


Figura 23. Comparativa dels agrupaments en dos clústers del grup d'elit, l'actiu i ambdós.

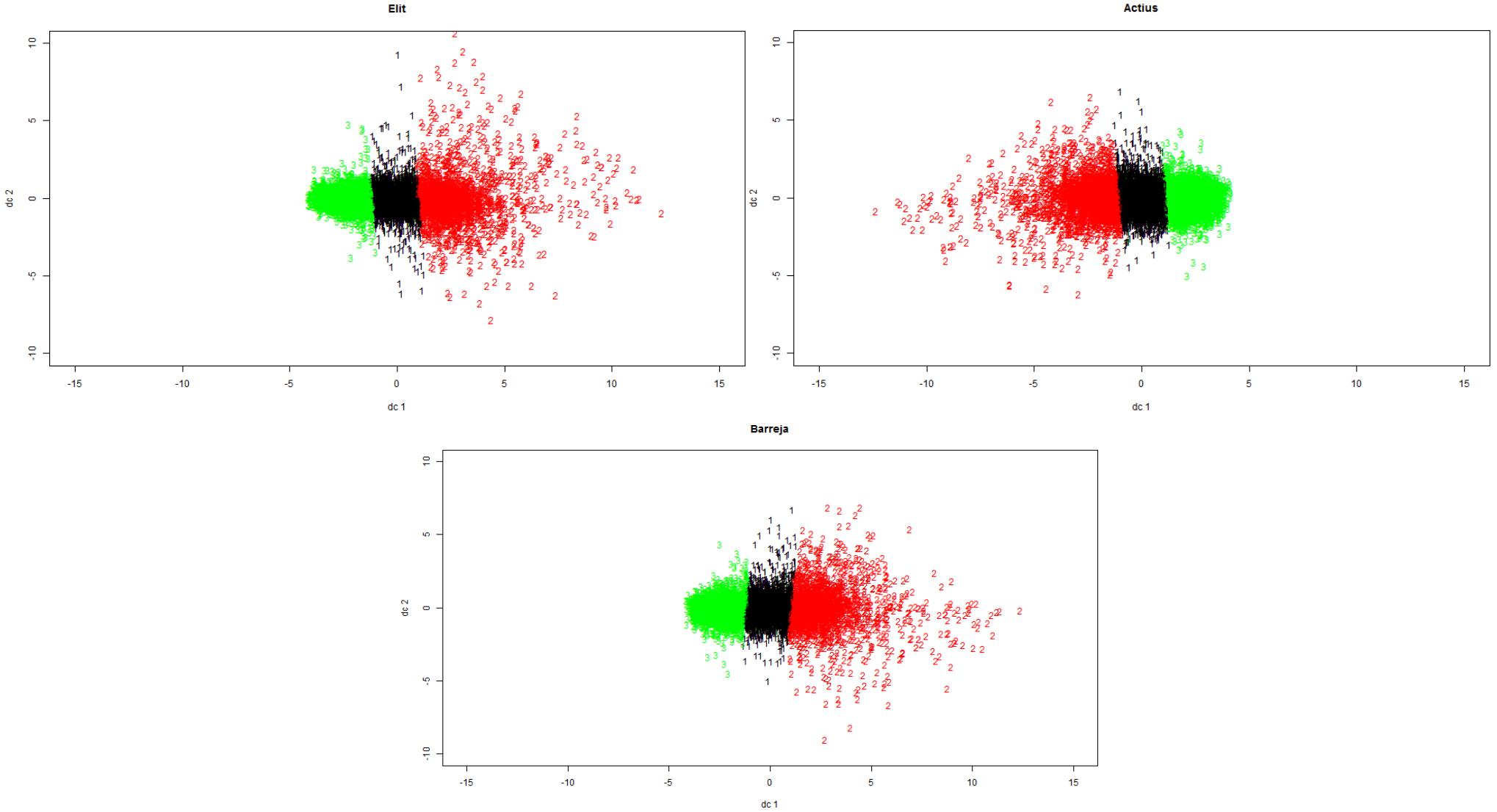


Figura 24. Comparativa dels agrupaments en tres clústers del grup d'elit, l'actiu i ambdós.

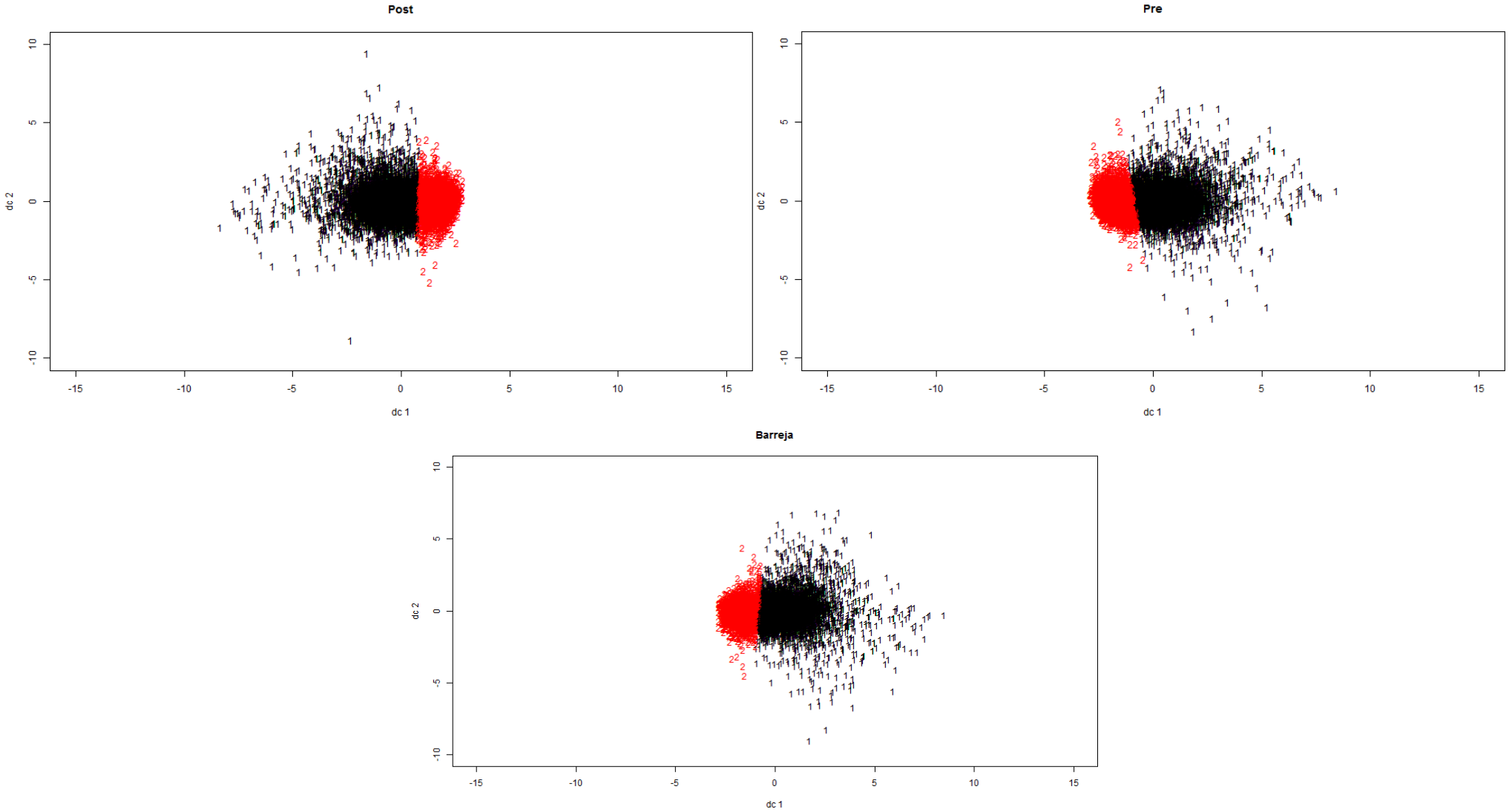


Figura 25. Comparativa dels agrupaments en dos clústers de mesures abans de la cursa PRE, després POST i ambdós.

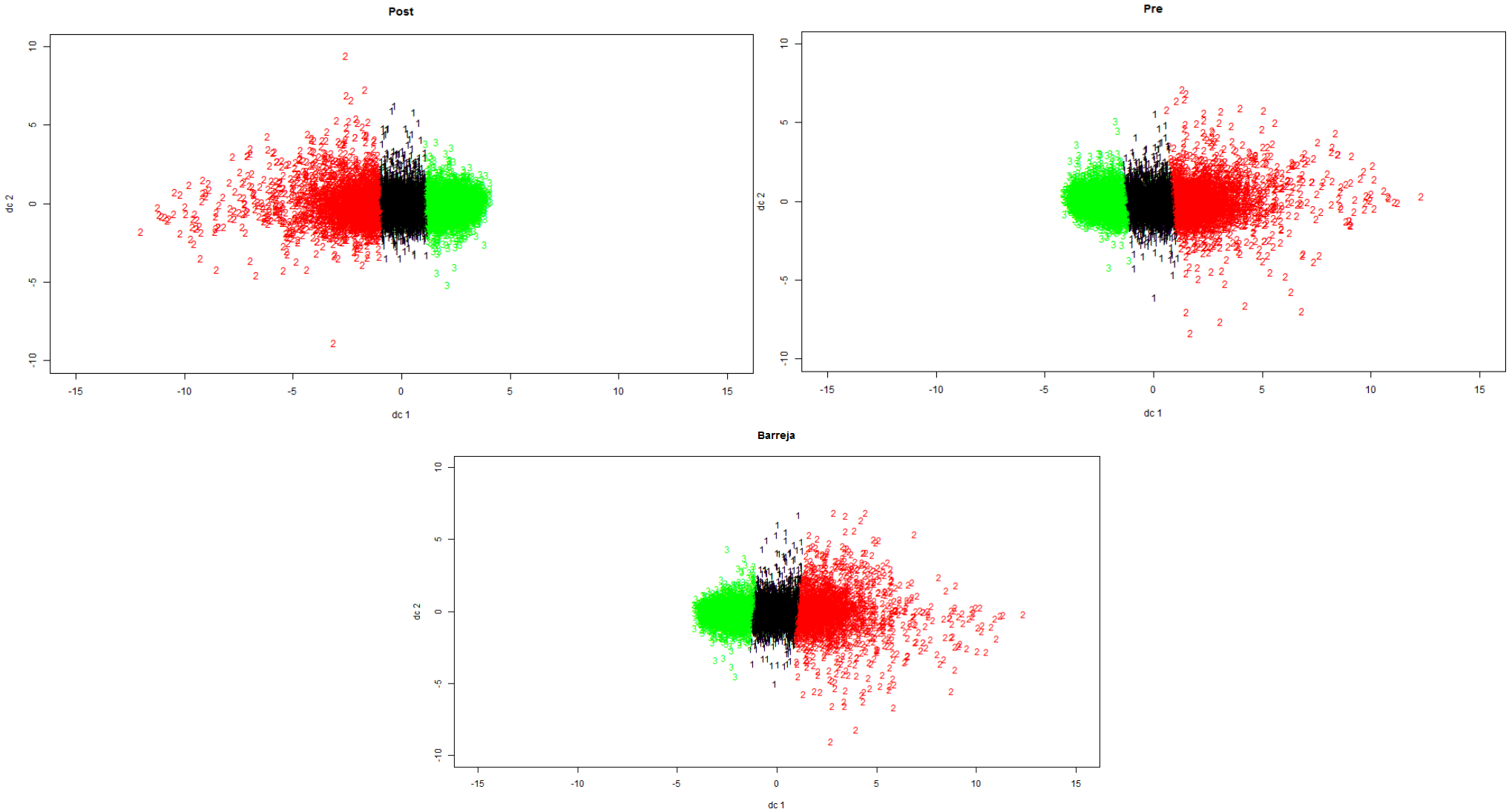


Figura 26. Comparativa dels agrupaments en tres clústers de mesures abans de la cursa PRE, després POST i ambdós.

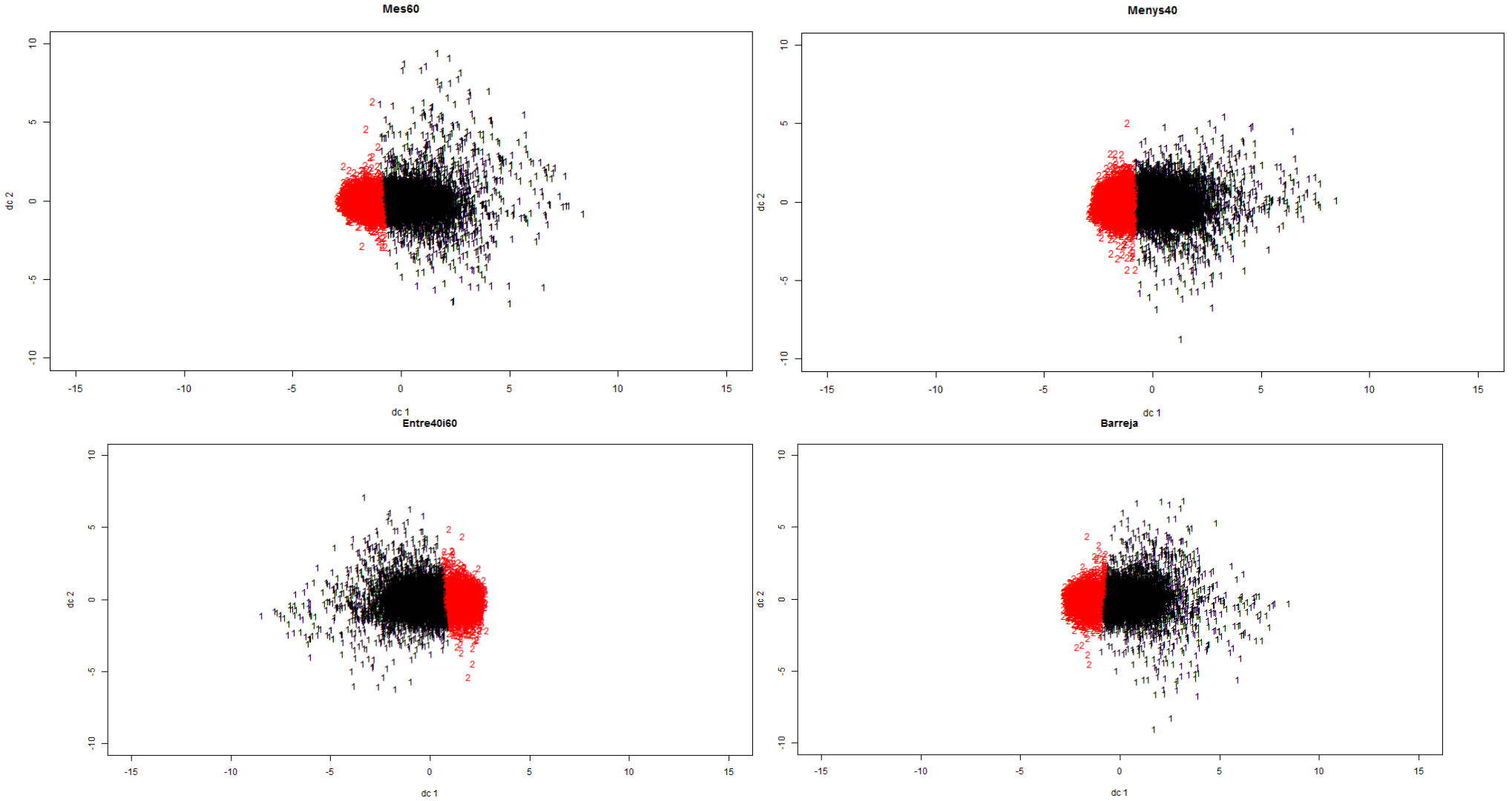


Figura 27. Comparativa dels agrupaments en dos clústers dels grups de distàncies i barrejat.

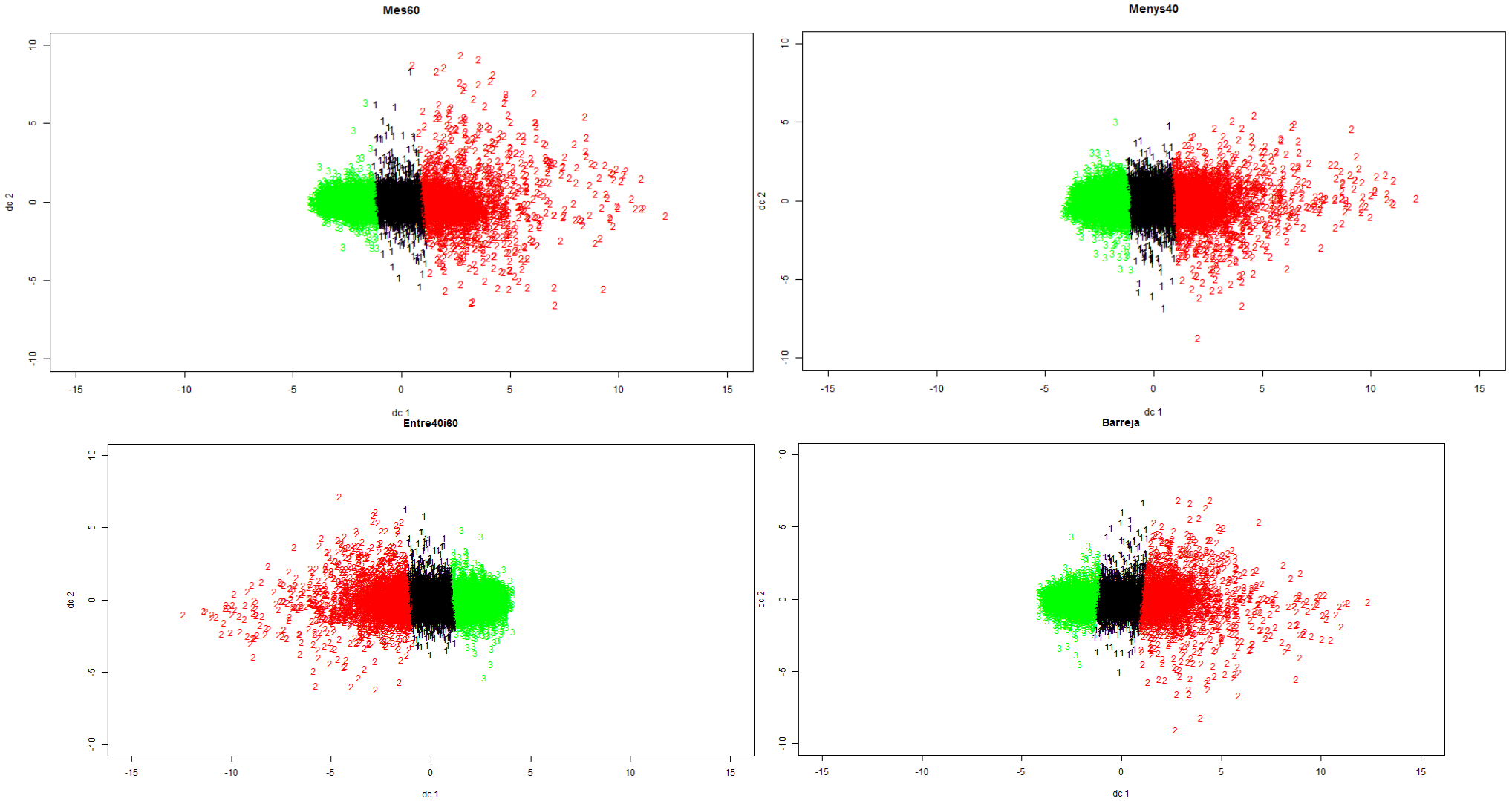


Figura 28. Comparativa dels agrupaments en tres clústers dels grups de distàncies i barrejat.

L'objectiu inicial al qual dóna resposta aquest subapartat és el d'identificar indicis d'incidència de diferents fenotips en la generació de proteïnes i lncRNA. En aquest sentit, s'indica que s'han trobat indicis de la incidència dels següents fenotips:

- Distància inferior a 40 km i superior a 60km, donat que ofereixen els agrupaments més diferents entre si i les mesures dels nivells d'expressió que s'hi corresponen presenten una alta variabilitat. Els agrupaments d'aquests fenotips també són els que es diferencien més de la barreja.

- Moment de la cursa pre i post, ja que són els segons que ofereixen agrupaments més diferents entre si (amb un valor de correlació força per sota de la resta), el grup de pre és el que presenta una major homogeneïtat i el de post és dels que té més variabilitat. Respecte la comparació amb els agrupaments de la barreja, no són dels que més en divergeixen però la diferència és prou elevada.

7.2 Correlació Pearson

En total s'han trobat 3.091.321 correlacions superiors a 0,9 l'identificador de la proteïna, del lncRNA i el valor exacte de la correlació de les quals es pot consultar al fitxer "correlacions090.txt" de la carpeta "fitxersTreball" de l'arxiu en el qual es lliure aquest projecte. La taula següent mostra el nombre de resultats amb una correlació cada vegada més propera a 1:

Correlació superior a	Nombre resultats	Nombre proteïnes
0,9	3.091.321	1.138
0,95	2.367.665	1.107
0,99	603.579	911
0,995	187.076	791
0,999	2.201	256
0,9991	1.468	236
0,9992	950	193
0,9993	543	151
0,9994	243	114
0,9995	120	66
0,9996	40	30
0,9997	17	16
0,9998	6	6
0,99985	1	1
0,99988	1	1
0,99989	0	0

La taula següent mostra les 17 correlacions amb un valor superior a 0.9997[10]:

Id proteïna	Id lncRNA	Correlació
16666109	16730697	0,999812169866738
16760755	16914913	0,999703229953376
16786780	16929056	0,999741334323389
16929199	17001843	0,99981510060652
16935029	16741524	0,999807178348168
17028655	16879484	0,999744286406899
17029003	17057995	0,999745744734548
17031781	17057995	0,999755401383085
17034960	17074523	0,99978673764426
17035022	17057995	0,999795570627739
17039281	17057995	0,999772827870107
17039839	16741524	0,999719960405202
17041697	17057995	0,999887101353949
17041782	17057995	0,999824289523568
17100653	17057995	0,999742663002719
17118292	16657450	0,999774973080241

Les correlacions trobades donen resposta a l'objectiu inicial d'analitzar la relació entre la generació de proteïnes i de lncRNA, de manera que s'han trobat indicis de relació amb diferents nivells de correlació que es podrien anar estudiant progressivament (de més correlació a menys) per confirmar-los. És important destacar que als primers resultats s'han trobat molts casos en què una mateixa proteïna té correlació elevada amb varis lncRNA, de manera que els resultats obtinguts no són u a u fins que el valor de correlació no és molt més proper a 1.

7.3 Anàlisi de Components Independents

La taula següent mostra el nombre de resultats de correlacions de proteïnes del pathway amb lncRNA:

Correlació superior a	**Nombre resultats**	**Proteïnes**
0,9	135.907	45
0,99	30.139	38
0,999	162	19
0,9995	23	8
0,9996	10	5
0,9997	5	3
0,9998	3	1
0,99985	1	1
0,99988	0	0

[10] Per filtrar les correlacions s'ha usat la funció subset: subset(matriuCorrelacions,matriuCorrelacions[,3]>"0.9997"). I per obtenir el nombre de proteïnes: dim(subset(nivells,rownames(nivells) %in% subset(matriuCorrelacions,matriuCorrelacions[,3]>0.9997)[,1])).

Els valors p obtinguts del test geomètric es presenten a continuació:

```
> 1-testHipergeometric
          [,1]     [,2]     [,3]     [,4]
16681370 1.0000000 1.0000000 0.9999936 0.9999936
16701841 1.0000000 0.9999936 0.9999936 0.9999936
16723593 1.0000000 0.9992990 0.9999936 1.0000000
16731169 0.9999936 0.9999045 0.9999936 0.9999936
16747338 1.0000000 1.0000000 0.9999936 1.0000000
16747529 0.9999936 1.0000000 0.9999045 0.9999936
16811372 1.0000000 1.0000000 0.9999936 0.9999936
16817790 1.0000000 1.0000000 1.0000000 1.0000000
16826985 0.9999936 0.9999936 1.0000000 0.9999045
16832438 0.9999936 1.0000000 1.0000000 0.9999936
16860709 0.9999936 1.0000000 1.0000000 0.9999936
16861033 0.9999045 0.9999936 0.9999045 1.0000000
16881838 0.9999936 1.0000000 1.0000000 1.0000000
16922993 0.9999936 1.0000000 1.0000000 0.9999045
16955511 1.0000000 0.9999045 0.9999936 0.9999936
17006378 0.9999936 1.0000000 1.0000000 1.0000000
17047461 0.9999936 0.9999045 1.0000000 0.9999936
17057218 0.9999045 1.0000000 0.9999045 1.0000000
17105047 1.0000000 0.9999936 0.9999936 1.0000000
```

Com que cap del valors és inferior o igual al valor de significació per defecte (0,05) es conclou que no hi ha diferències estadísticament significatives de la incidència de les diferents proteïnes als metafenotips construïts.

7.4 Presentació de resultats

Per tal de presentar els resultats obtinguts en aquest TFM s'ha desenvolupat una aplicació web utilitzant el paquet Shiny de l'eina Rstudio i que presenta les següents funcionalitats:

- Visualització de la informació relativa als fitxers dels nivells d'expressió (metadades per fenotips):
    - Incloent la possibilitat de filtrar per fenotip.
- Visualització dels resultats de l'anàlisi PCA per fenotips i barreja.
- Visualització dels agrupaments realitzats al TFM per fenotips i comparació amb la barreja a diferents nivells:
    - Representació gràfica dels agrupaments en 2 i 3 clústers.
    - Gràfiques resumint les correlacions entre fenotips contraris i entre els fenotips i la barreja (mesures globals i unitàries).

Incloent el filtre per tipus de fenotip (“Distància”, “Gènere”, “Moment de la cursa” i “Grup de l’activitat”).

- Consulta del nombre de resultats de proteïnes correlacionades amb lncRNA trobades al TFM:
    - Amb valor dinàmic de correlació entre 0,9 i 1 a fixar per l’usuari.
- Visualització del nombre de resultats de proteïnes del pathway correlacionades amb lncRNA trobades al TFM.
- Visualització dels p-value de cada proteïna amb cadascun dels metafenotips.
- Presentació de conclusions del TFM.

Per mostrar aquests resultats l’aplicació utilitza les imatges prèviament generades (figures de la 18 a la 28), així com taules resum de les dades que s’hi presenten (per exemple, la taula resum “resultatsPCA.csv” en comptes de calcular el PCA de les dades en temps d’execució).

La imatge següent mostra l’aplicació desenvolupada:

Figura 29. Web de resultats del TFM.

L’aplicació s’executa en mode local amb el codi següent:

```
> library(shiny)
> runApp("~/TFM/webResultats")
```

El codi complert de l’aplicació es pot consultar a l’annex 1 d’aquest document.

8 Conclusions del TFM

En relació amb els objectius fixats del present projecte, es destaquen les següents conclusions finals:

1. En l'anàlisi de la incidència dels fenotips gènere, distància, moment de la cursa i grup d'activitat en la generació de lncRNA i proteïnes s'han trobat indicis de la incidència dels fenotips de distància (inferior a 40 km i superior a 60km) i del moment de la cursa (pre i post).

2. En l'estudi de la correlació entre els nivells d'expressió de proteïnes i els de lncRNA s'han obtingut molts resultats de correlacions entre 0,9 (1.138 proteïnes correlacionades) i 0,99988 (1 proteïna correlacionada). La majoria de les correlacions trobades han estat d'una mateixa proteïna a varis lncRNA.
    a. A partir d'aquest anàlisi es podrà investigar el paper dels lncRNA en les funcions de les proteïnes correlacionades, prioritzant els que han obtingut millor resultat (correlació més propera a 1).

3. En l'exploració de la incidència dels lncRNA en el pathway de la glucosa:
    a. S'han trobat molts resultats amb alta correlació entre les proteïnes del pathway de la Glicòlisis/Gluconeogènesi amb els lncRNA: amb 0,9 45 proteïnes del pathway correlacionades i amb 0,99985 una.
    b. No s'han trobat diferències estadísticament significatives de la incidència de les diferents proteïnes als metafenotips construïts:
        i. De manera que cap proteïna, per si sola, té més incidència als metafenotips que la resta.
        ii. I, de retruc, tampoc cap lncRNA correlacionat amb aquestes proteïnes del pathway.

9 Problemes trobats

La taula següent conté un resum dels principals problemes trobats, i la solució aplicada per superar-los:

Problema	Descripció	Solució
Domini nou	Desconeixement del domini de la genètica	Previsió d'una subfase d'aprenentatge i familiarització
Nou llenguatge de programació	Sense nocions del llenguatge R ni de les eines associades	Aprenentatge i millora progressiva del codi: primer comandes repetitives que posteriorment s'han anat substituint per funcions, codi més optimitzat i altres comandes més eficients/adequades
Dades projecte SUMMIT incompletes	Al projecte SUMMIT no es van realitzar totes les mesures inicialment indicades (veure apartat 4.2.1)	Usar les dades disponibles (28 fitxers)
Error paquet Affy per carregar els fitxers .CEL	El paquet affy que es sol utilitzar per carregar i llegir els fitxers .CEL amb els nivells d'expressió no funciona correctament amb fitxers recents com els proporcionats	Usar el paquet oligo
Limitació de memòria a utilitzar a RStudio	Degut a les limitacions de l'ordinador (32 bits amb 8 GB de RAM) utilitzat no era possible carregar tots els fitxers .CEL a la vegada (necessari per analitzar les dades conjuntament) ni aplicar l'algoritme PAM (per falta de memòria amb tots els probesets per allotjar els objectes que utilitza, inicialment quasi 11GB necessaris)	Usar un ordinador de 64 bits i més memòria RAM Augmentar el límit de memòria a usar per RStudio a 11 GB amb memory.limit(11811160064)
Identificació dels lncRNA i proteïnes	Cada BBDD de lncRNA utilitza identificadors de lncRNA i proteïnes propis, de manera que no es poden relacionar els fitxers .CEL amb els noms o tipus de lncRNA (i de proteïnes) usant l'id. A més a més, cada BBDD té camps descriptius propis tot i usar el mateix format de fitxer FASTA	Filtratge textual per camps (tant per seleccionar com per descartar registres) en el pretractament del fitxer d'affyMetrix descarregat amb els identificadors de les proteïnes i els lncRNA a relacionar amb el contingut tractat dels fitxers .CEL
Temps elevats d'execució i inicialment desconeguts	Incertesa en quant al temps d'execució d'algoritmes (per exemple l'anàlisi de correlacions entre lncRNA i les proteïnes ha trigat unes 20 hores).	Controls i còmputs per hora Optimització de codi Inclusió d'elements informatius a les funcions (impressions per pantalla per conèixer evolució)

10 Possibles ampliacions

A continuació es presenten una sèrie de possibles ampliacions del present TFM:

- Publicar a Internet l'aplicació web desenvolupada amb Shiny i RStudio, habilitant i configurant un servidor Linux amb el programari Shiny Server.
- Afegir més funcionalitats a l'aplicació web fent-la més interactiva.
- Augmentar el nombre de mostres dels nivells d'expressió per obtenir resultats més globals (i amb nombre de mostres per fenotip similar).
- Realitzar agrupació dels nivells d'expressió utilitzant més criteris (del fitxer de metadades) i combinacions.
- Investigar la incidència de l'esforç d'alt rendiment a altres pathways en base a una recerca prèvia d'evidència científica al respecte.

11 Annex 1: Aplicació web

A continuació es presenta el codi corresponent a la part d'interfície gràfica d'usuari de l'aplicació web desenvolupada (fitxer "ui.R"):

```
library(shiny)

shinyUI(
 pageWithSidebar(
  headerPanel("Resultats del TFM: Incidència de l'esforç d'alta intensitat en la generació de lncRNA."),
  sidebarPanel(
   h4("Nivells d'expressió del projecte SUMMIT (28 fitxers .CEL):"),
   selectInput("dataset", "Triar filtre per fenotip:",
   choices = c("Tots (Barreja)","Dones", "Homes", "Menys de 40 km","Entre 40 i 60 km","Més de 60 km", "Pre cursa", "Post cursa","Elit","Actius")),

   h4("Agrupaments per fenotips:"),
   selectInput("agrupament", "Triar agrupament per tipus de fenotip:",
   choices = c("Gènere", "Distància", "Moment de la cursa","Grup d'activitat")),
   selectInput("visua", "Triar visualització",
          choices = c("2 clústers", "3 clústers")),

   h4("Correlació lncRNA i proteïnes:"),
   sliderInput("correlacio",
          "Valor de la correlació:",
          min = 0.9,
          max = 1,
          value = 0.9)
  ),

  mainPanel(
   tabsetPanel(
    tabPanel("Nivells d'expressió",
     h3(textOutput("taula")),
     tableOutput("view"),
     h3("Resultat de l'anàlisi PCA:"),
     tableOutput("resultatsPCA"),
     br(),br(),br(),
     p("Ariadna Rius Soler",style="text-align: right;"),
     p("UOC. TFM. Curs 2013-2014",style="text-align: right;"),
     p("Consultor: Samir Kanaan Izquierdo",style="text-align: right;")
    ),

    tabPanel("Agrupaments",
     h3("Agrupaments per fenotips:"),
     imageOutput("agrupament1"),
     div(id="separador", style="height:300px;"),
```

```
      h3("Correlacions entre els agrupaments per fenotips oposats:"),
      imageOutput("correlacions1"),
      h3("Correlacions entre els agrupaments per fenotips i la barreja:"),
      imageOutput("correlacions2"),
      h3("Diferàncies de correlacions entre els agrupaments per fenotips i la barreja:"),
      p("Mesures unitàries: (100-correlació) / nombre mostres"),
      imageOutput("correlacions3"),
      br(),br(),br(),
      p("Ariadna Rius Soler",style="text-align: right;"),
      p("UOC. TFM. Curs 2013-2014",style="text-align: right;"),
      p("Consultor: Samir Kanaan Izquierdo",style="text-align: right;")
     ),
     tabPanel("Correlació",
      h3("Correlacions entre les proteïnes i els lncRNA:"),
      p("Hi ha un total de: ", textOutput("corProteineslncRNA"),"resultats amb correlació superior
a",textOutput("corFix")),
      br(),br(),br(),
      p("Ariadna Rius Soler",style="text-align: right;"),
      p("UOC. TFM. Curs 2013-2014",style="text-align: right;"),
      p("Consultor: Samir Kanaan Izquierdo",style="text-align: right;")
     ),
     tabPanel("Pathway",
          h3("Correlacions entre les proteïnes del pathway i els lncRNA:"),
          tableOutput("corPath"),
          h3("p-values resultants del test hipergeomètric:"),
          tableOutput("pvalues"),
          br(),br(),br(),
          p("Ariadna Rius Soler",style="text-align: right;"),
          p("UOC. TFM. Curs 2013-2014",style="text-align: right;"),
          p("Consultor: Samir Kanaan Izquierdo",style="text-align: right;")

     ),
     tabPanel("Conclusions",
          h3("Conclusions del TFM"),
          p("El fenotip amb major variablitat és el de distància major a 60 km."),
          p("El fenotip més homogeni és el de moment de la cursa PRE."),
          p("S'han trobat indicis d'incidència en la generació de proteïnes i lncRNA dels fenotips:
distància inferior a 40 km i superior a 60km i moment de la cursa PRE i POST."),
          p("No s'han trobat diferències estadísticament significatives de la incidència de les
diferents proteïnes als metafenotips construïts."),
          br(),br(),br(),
          p("Ariadna Rius Soler",style="text-align: right;"),
          p("UOC. TFM. Curs 2013-2014",style="text-align: right;"),
          p("Consultor: Samir Kanaan Izquierdo",style="text-align: right;")
     )
    )
   )
  )
)
```

I el corresponent a la part servidor (fitxer “server.R”):

```
shinyServer(function(input, output) {

 dataset <- read.csv("~/TFM/Fenos_CDV.csv", quote="")
 rPCA<-read.csv("~/TFM/webResultats/resultatsPCA.csv", sep=";", dec=",", quote="")
 correlacions <-read.csv("~/TFM/correlacions090.txt", sep=";", quote="")

 datasetInput <- reactive({
  switch(input$dataset,
       "Tots (Barreja)"=dataset,
       "Dones" = subset(dataset,Genere=="Female"),
       "Homes" = subset(dataset,Genere=="Male"),
       "Menys de 40 km" = subset(dataset,Distancia==1),
       "Entre 40 i 60 km"=subset(dataset,Distancia==2),
       "Més de 60 km"=subset(dataset,Distancia==3),
       "Pre cursa"=subset(dataset,MomentCursa=="PRE"),
       "Post cursa"=subset(dataset,MomentCursa=="POST"),
       "Elit"=subset(dataset,GrupActivitat=="E"),
       "Actius"=subset(dataset,GrupActivitat=="A")
       )
 })
 fenotipTriat <- reactive({
  switch(input$dataset,
       "Tots (Barreja)"="Barreja",
       "Dones" = "Dones",
       "Homes" = "Homes",
       "Menys de 40 km" = "<40 km",
       "Entre 40 i 60 km"="<40<60 km",
       "Més de 60 km"=">60km",
       "Pre cursa"="PRE",
       "Post cursa"="POST",
       "Elit"="Elit",
       "Actius"="Actius"
  )
 })
 output$taula <- renderText({
  paste(dim(datasetInput())[1]," fitxers:", sep = " ")
 })
 output$view <- renderTable({
  dades <-datasetInput()
  head(dades, n = dim(dades)[1])
 })
 output$resultatsPCA <- renderTable({
  dadesPCA <-subset(rPCA,Fenotip==fenotipTriat())
  head(dadesPCA, n = dim(dadesPCA)[1])
 })
```

```
  nomAgrupament <- reactive({
   switch(input$agrupament,
       "Gènere"="Genere",
       "Distància"="Distancia",
       "Moment de la cursa"="MomentCursa",
       "Grup d'activitat"="GrupActivitat"
   )
  })
  tipusVisualitzacio <-reactive({
   switch(input$visua,
        "2 clústers" ="2C.png",
        "3 clústers" ="3C.png"
   )
  })
  output$agrupament1 <- renderImage({
   list(src = paste("~/TFM/webResultats/",nomAgrupament(),tipusVisualitzacio(),sep=""))
  },deleteFile = FALSE)
  output$correlacions1 <- renderImage({
   list(src = "~/TFM/webResultats/fenotipsOposats.png")
  },deleteFile = FALSE)
  output$correlacions2 <- renderImage({
   list(src = "~/TFM/webResultats/fenotipsBarreja.png")
  },deleteFile = FALSE)
  output$correlacions3 <- renderImage({
   list(src = "~/TFM/webResultats/fenotipsBarrejaUnitari.png")
  },deleteFile = FALSE)
  output$corProteineslncRNA <- renderText({
   valorCor <-input$correlacio
   resultat <-subset(correlacions,correlacions[,3]>valorCor)
   dim(resultat)[1]
  })
 output$corFix <- renderText({
   input$correlacio
  })
  output$corPath <- renderTable({
   pathn <-read.csv("~/TFM/WebResultats/nombrePath.csv", sep=";", quote="")
   head(pathn, n = dim(pathn)[1])
  })
   output$pvalues <- renderTable({
   path <-read.csv("~/TFM/WebResultats/pvalues.csv", sep=";", quote="")
   head(path, n = dim(path)[1])
  })
 })
```

Tant el codi com els fitxers utilitzats en l'aplicació web es poden consultar a la carpeta "WebResultats" del fitxer en el qual es lliure el present TFM.

Per tal que l'aplicació funcioni correctament cal desar ambdós fitxers amb l'opció "save with Encoding" i seleccionar UTF-8, així com corregir els accents que s'hagin descodificat (o bé treure'ls tots).

12 Glossari

ADN: Àcid DesoxiriboNucleic.

ARN: Àcid RiboNucleic.

ARNm: ARN missatger.

ARNnc: ARN no codificant.

ARNt: ARN de transferència.

BBDD: Bases de Dades.

DNA: DeoxyriboNucleic Acid.

ICA: Independent Component Analysis.

IDE: Integrated development environment.

KEGG: Kyoto Encyclopedia of Genes and Genomes.

lincRNA: long intergenic non-protein coding RNA.

lncRNA: long non-coding RNA.

miRNA: micro RNA.

misc_RNA: miscellaneous other RNA.

mRNA: Messenger RNA.

NCBI: National Center for Biotechnology Information.

PAM: Partitioning Around Medoids.

PCA: Principal Component Analysis.

PICB: Partner Institute for Computational Biology.

RNA: RiboNucleic Acid.

snRNA: small nuclear RNA.

TFM: Treball Final de Màster.

13 Bibliografia

http://es.wikipedia.org (gener 2014)

http://en.wikipedia.org/ (gener 2014)

http://webinar.sciencemag.org/ (gener 2014)

http://www.r-project.org/ (gener 2014)

http://www.mathworks.es/products/matlab/ (octubre 2013)

http://www.python.org/ (octubre 2013)

http://www.cliki.net/ (octubre 2013)

http://www.bioconductor.org/ (octubre 2013)

http://www.rstudio.com/shiny/ (gener 2014)

http://www.rstudio.com/ide/download/server (octubre 2013)

http://rapache.net/ (octubre 2013)

http://nred.matticklab.com/cgi-bin/ncrnadb.pl (octubre 2013)

http://www.lncrnadb.org/ (octubre 2013)

http://202.38.126.151/hmdd/html/tools/lncrnadisease.html (octubre 2013)

http://159.226.118.44/NONCODERv3/index.htm (octubre 2013)

http://www.lncipedia.org/ (octubre 2013)

http://www.proyectosummit.com/ (novembre 2013)

http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Partitioning_Around_Medoids_(PAM) (novembre 2013)

http://en.wikipedia.org/wiki/K-medoids (novembre 2013)

http://www.cs.umb.edu/cs738/pam1.pdf (novembre 2013)

http://cran.fhcrc.org/web/packages/cluster/cluster.pdf (novembre 2013)

http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html (novembre 2013)

http://cran.r-project.org/web/packages/fastICA/fastICA.pdf (gener 2014)

http://ca.wikipedia.org/wiki/Neguentropia (desembre 2013)

http://www.bioconductor.org/packages/release/bioc/html/oligo.html (octubre 2013)

http://www.ensembl.org/info/genome/genebuild/ncrna.html (octubre 2013)

http://www.bioconductor.org/packages/release/bioc/html/affy.html (octubre 2013)

http://www.affymetrix.com/support/technical/annotationfilesmain.affx (gener 2014)

http://www.bioconductor.org/packages/release/bioc/html/pcaMethods.html (novembre 2013)

http://www.sbforum.org/cmsimages/presentations/BioC.pdf (novembre 2013)

http://www.bioconductor.org/packages/release/bioc/html/NCIgraph.html (octubre 2013)

http://www.bioconductor.org/packages/release/bioc/html/KEGGgraph.html (novembre 2013)

http://www.cytoscape.org/ (octubre 2013)

http://bioinformatics.psb.ugent.be/webtools/MotifSuite/fastaformat.php (octubre 2013)

http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml (octubre 2013)

http://en.wikipedia.org/wiki/FASTA_format (octubre 2013)

http://ca.wikipedia.org/wiki/Glucosa (desembre 2013)

http://www.bioinformatics.jp/en/keggftp.html (desembre 2013)

http://www.bioconductor.org/packages/release/bioc/html/pathview.html (desembre 2013)

http://stat.ethz.ch/R-manual/R-patched/library/stats/html/Hypergeometric.html (gener 2014).

Mòduls de l'assignatura "*Intel·ligència Artificial*" del Màster Universitari en Enginyeria Informàtica.

Mòduls de l'assignatura "*Intel·ligència Artificial Avançada*" del Màster Universitari en Enginyeria Informàtica.

"Analysis of Differential Gene Expression Studies", D. Scholtens i A. von Heydebreck: http://www.cbcb.umd.edu/~hcorrada/CMSC858B/readings/Solutions_ch14.pdf (novembre 2013).

"Analysis Overview", V. J. Carey i R. Gentleman: http://www.cbcb.umd.edu/~hcorrada/CMSC858B/readings/Solutions_ch11.pdf (novembre 2013).

"Cluster Analysis of Genomic Data", K. S. Pollard i M. J. van der Laan: http://www.cbcb.umd.edu/~hcorrada/CMSC858B/readings/Solutions_ch13.pdf. (novembre 2013).

"NRED: a database of long noncoding RNA expression", Marcel E. Dinger, Ken C. Pang, Tim R. Mercer, Mark L. Crowe, Sean M. Grimmond i John S. Mattick: http://nar.oxfordjournals.org/content/37/suppl_1/D122.full (novembre 2013).

"Reducing and Clustering high Dimensional Data through Principal Component Analysis" R.Indhumathi(College of Arts and Sciences for Women), Dr.S.Sathiyabama (College of Technology, Tiruchengode): http://www.ijcaonline.org/volume11/number8/pxc3872158.pdf (desembre 2013).

"The pcaMethods Package" Wolfram Stacklies and Henning Redestig CAS-MPG (PICB) Shanghai, P.R. China and Max Planck Institute for Molecular Plant Physiology Potsdam, Germany: http://www.bioconductor.org/packages/2.14/bioc/vignettes/pcaMethods/inst/doc/pcaMethods.pdf (desembre 2013).

Doctoral Thesis "Genetic association analysis of complex diseases through information theoretic metrics and linear pleiotropy". PhD Student: Helena Brunel Montaner, PhD Advisor: Dr. Alexandre Perera i Lluna. Universitat Politècnica de Catalunya, Programa de Doctorat en Enginyeria Biomèdica. Barcelona, 2013.

Webinar "*Targeting Noncoding RNAs in Disease: Challenges and Opportunities*".

Vídeo "Ayudantía BioCel: Expresión genética 1/3 (Introducción y tipos de RNA)": http://www.youtube.com/watch?v=eyp9xXTdiB0 (octubre 2013).