TREBALL FINAL DE MÀSTER

EINES BIOINFORMÀTIQUES APLICADES A L'ESTUDI DE L'ESPERMATOZOIDE I EL FLUID SEMINAL

Estudiant: David Delgado Dueñas
Màster universitari en bioinformàtica i bioestadística
Àrea: Anàlisi de dades òmiques

Professor Col·laborador: Guillem Ylla Bou
Tutora Externa: Meritxell Jodar Bifet
Prof. Responsable Assignatura: Carles Ventura Royo
David Merino Arranz

2 de Gener de 2018

David Delgado Dueñas

Santa Coloma de Gramenet (BCN) – España

ddelgadodu@uoc.edu

Imatge de portada: Creada per Iconicbestiary- Freepik.com

Título del trabajo:	*EINES BIOINFORMÀTIQUES APLICADES A L'ESTUDI DE L'ESPERMATOZOIDE I EL FLUID SEMINAL*
Nombre del autor:	*David Delgado Dueñas*
Nombre del consultor/a:	*Guillem Ylla Bou (UOC)* *Meritxell Jodar Bifet (Extern)*
Nombre del PRA:	*Carles Ventura Royo - David Merino Arranz*
Fecha de entrega (mm/aaaa):	01/2018
Titulació:	*MÀSTER UNIVERSITARI EN BIOINFORMÀTICA I BIOESTADÍSTICA*
Área del Trabajo Final:	*ANÀLISI DE DADES ÒMIQUES*
Idioma del trabajo:	*CATALÀ*
Palabras clave	*PROTEÒMICA, TRANSCRIPTÒMICA, PROGRAMACIÓ.*

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

Aquest treball cerca per establir nous mètodes pel tractament de dades de proteòmica provinents de l'ús de TMT 10plex i cromatografia líquida seguida d'espectrometria de masses en tàndem (LC-MS/MS) i el seu estudi posterior, així com per l'estudi de RNAs i circRNAs d'espermatozoide provinents de la tecnologia d'alt rendiment, RNA-seq.

Pel que fa a proteòmica es va optar pel desenvolupament d'un algorisme nou que ens permet el treball amb taules de correlació, per obtenir relacions entre proteïnes .

Aplicant aquest nou algorisme, es va obtenir un llistat de les relacions entre proteïnes altament correlacionades Un ampliació de l'algorisme va permetre trobar diferencies dintre de les poblacions aplicant regles estadístiques per extreure informació.

Els resultats obtinguts ens indiquen que la població de pacients normozoospèmics es bastant homogènia pel que fa a nivells de proteïna però la resta de grups de pacients infèrtils presenta un alta heterogeneïtat.

Pel que fa a la transcriptòmica, s'estableix un procediment per treballar amb aquest tipus de mostres en el laboratori que passa per l'ús d'eines com Bowtie, TopHat, Cufflinks o DESeq i les bases per fer ús d'eines per l'estudi de circRNA com ara KNIFE.

En conclusió aquest treball s'aconsegueix crear un nou mètode per analitzar mostres de TMT seguides de LC-MS/MS amb problemes de quantificació i establir les bases per posar en marxa un punt de anàlisi bioinformàtic per a transcriptòmica.

Abstract (in English, 250 words or less):

This work seeks to establish new methods for the analysis of proteomic data derived from the use of TMT 10plex and liquid chromatography followed by tandem mass spectrometry (LC-MS / MS) and its subsequent study, and for the study of RNA and circRNAs of sperm derived from RNA-seq technology.

Regarding proteomics analysis we have developed a new algorithm that allow us to work with correlation tables, to obtain the relationships between proteins.

Applying this new algorithm, a list of relationships between highly correlated proteins was obtained. Another associated algorithm allow to find differences within populations by applying statistical rules to extract information

The results obtained indicate that the population of normozoospermics patients is quite homogeneous in terms of protein levels but the rest of the groups of infertile patients has a high heterogeneity.

Regarding the transcriptomics, on the one hand, a procedure has been established to work with the RNA-seq data in the laboratory by the use of tools such as Bowtie, TopHat, Cufflinks or DESeq. On other hand, the bases to use of tools for circRNA study such as KNIFE has been established.

In conclusion, this work achieves the creation of a new method for analyzing TMT data followed by LC-MS / MS with quantification problems and establishing the bases to launch a bioinformatic analysis for RN-seq data.

ÍNDEX

LLISTA DE FIGURES

LLISTA D'ALGORISMES

LLISTA DE TAULES

1.INTRODUCCIÓ

A. CONTEXT I JUSTIFICACIÓ DEL TREBALL

A.1. Descripció general

El treball final de màster (TFM) que he dut a terme pren com a punt de partida l'anàlisi de dades òmiques provinents de l'anàlisi de RNA i proteïnes presents a l'espermatozoide i al fluid seminal en diferents estudis realitzats per un mateix grup de recerca[1–3].

Les línies d'investigació i les fites més significatives aconseguides del grup de recerca receptor son les següents:

1. Proteòmica de l'espermatozoide: El grup liderat pel Dr. Oliva és grup de referència a nivell internacional en l'estudi de la proteòmica de l'espermatozoide mitjançant l'ús de tècniques d'alt rendiment. El grup ha estat pioner en la caracterització d'una part substancial de les proteïnes presents en l'espermatozoide, a més d'aportar informació valuosa sobre els mecanismes moleculars bàsics que regulen la producció, la maduració i el trànsit dels espermatozoides i, d'alguns dels mecanismes patogènics implicats en la infertilitat masculina [4–7]. Recentment, també han identificat un conjunt de proteïnes de l'espermatozoide clínicament rellevants, que podrien predir la taxa d'èxit de certes tècniques de reproducció assistida (TRA)[8].

2. Genòmica de l'espermatozoide: El grup és expert en estudis mutacionals de gens associats a la infertilitat masculina, i forma part del *European Molecular Quality* Network per a l'estudi de les microdeleccions del cromosoma Y [9–11].

3. Transcriptòmica de l'espermatozoide: L'aplicació de tècniques transcriptòmiques d'alt rendiment, com són els *arrays* d'expressió i la seqüenciació massiva, ha permès al grup determinar possibles mecanismes patogènics que afecten a la motilitat de l'espermatozoide i ha proporcionat un dels llistats més complets dels petits RNAs no codificants presents en l'espermatozoide humà [2,12–14].

4. Epigenètica de l'espermatozoide: Els estudis del grup en el camp de l'epigenètica han permès corroborar que la cromatina espermàtica està diferenciada en dos grans dominis segons la seva unió a dos grans famílies de proteïnes nuclears, (i) el domini nucleoprotamina (85-95% del DNA de l'espermatozoide) i (ii) el domini nucleohistona (5-15%). El resultats del grup han mostrat que aquests dos dominis estan associats tan a un patró gènic com proteic diferencial, no donat a l'atzar, que podrien modular l'expressió gènica en l'embrió primerenc [5,15].

Com es pot observar en els punts anteriors, els estudis del grup han estat centrats en l'estudi de l'espermatozoide, tot i que els espermatozoides només representen el 5% de l'ejaculat. El 95% restant correspon a les secrecions de les glàndules accessòries masculines que sembla que juguin un paper molt més important del que ser simplement

un mitjà de transport pels espermatozoides. Per aquesta raó, una de les línies actuals d'investigació del grup és la caracterització proteòmica i transcriptòmica del plasma seminal a través de tecnologies d'alt rendiment, per a completar el coneixement dels aspectes fonamentals de composició, regulació i funció del semen i, al mateix temps, identificar nous biomarcadors clínicament rellevants [3,16]. L'altra gran línia actual del grup és l'aplicació del poder de les tecnologies d'obtenció de dades òmiques i de la biologia de sistemes per identificar i desenvolupar nous biomarcadors que permetin estratificar els pacients infèrtils en subgrups amb fenotips moleculars i clínics ben definits, per així poder establir un diagnòstic precís i personalitzat que permeti oferir al pacient el tractament de fertilitat més adequat.

Personalment m'he incorporat a l'anàlisi biocomputacional de les dades òmiques derivades de proteòmica i RNA-seq. Per aquesta raó el meu projecte es pot dividir en dos subapartats:

Proteòmica del fluid seminal

El mètode estàndard per la quantificació de proteïnes de mostres d'espermatozoides és el marcatge proteic amb *tamdem mass tag* (TMT) i la identificació i quantificació mitjançant espectrometria de masses (LC-MS/MS). No obstant, en les nostres mostres de 16 pacients (4 pacients amb paràmetres seminals normals (normozoospèrmics), 4 amb alteracions en la mobilitat dels espermatozoides (astenozoospèrmics), 4 amb alteracions en la concentració de espermatozoides a fluid seminal (oligozoospèrmics) i 4 amb absència de espermatozoides a fluid seminal (azoospèrmics)) es va detectar que el mètode estàndard fallava perquè la quantificació de les proteïnes en el pool que es fa servir com a control i havia de permetre la quantificació de les proteïnes en la resta de mostres, havia estat incorrecta.

Per la qual cosa el meu esforç en aquest apartat es centra en trobar un mètode alternatiu per controlar l'expressió obtinguda mitjançant la comparació de les mostres quan la quantificació inicial del pool de control no ha estat correcta i no es pot fer servir el mètode estàndard.

TMT 10-plex

El marcatge diferencial isobàric amb TMT 10-plex consisteix en un total de 10 isòtops diferents que s'uneixen de manera covalent a l'extrem N- o C-terminal dels residus de lisina de les proteïnes. Aquests diferents isòtops presenten una massa monoisotòpica diferent, la qual pot ser detectada utilitzant espectrometria de masses d'alta resolució, de tal manera que permet la quantificació de les proteïnes de mostres independents.

ESPECTOMETRIA DE MASSES

La espectrometria de masses d'alta resolució permet determinar la distribució de les molècules en funció de la seva massa i càrrega, de tal manera que permet la quantificació de les proteïnes de mostres independents

La rellevància d'aquest primer apartat es troba en el establiment d'un nou procediment que pot donar resultats que ofereixen més informació que el procediment estàndard i que a part pot ser utilitzat quan hi hagin hagut algun problema durant el transcurs de la tècnica.

Transcriptòmica d'espermatozoide

En una segona vessant, el TFM versa sobre el disseny, establiment i consolidació dels algorismes necessaris per l'anàlisi de *longRNAs* (>200 nucleòtids) i *circRNA* d'espermatozoides.

El grup de recerca ha dut a terme un estudi pilot per tal de validar resultats preliminars que mostraven que diferències en el processament inicial de la mostra de semen feia disminuir l'abundància d'RNAs específics d'exosomes seminals, cosa que suggereix que el processament inicial de la mostra de semen afecta a la unió dels exosomes amb l'espermatozoide.

Aquest estudi es basa en l'anàlisi de 4 mostres d'espermatozoide procedents d'individus amb paràmetres seminals normals que es van processar totes quatre de dues maneres diferents. A partir d'aquestes mostres s'ha realitzat la construcció de 8 llibreries de *longRNA* que s'han analitzat mitjançant RNA-seq.

La meva tasca en aquest projecte és dur a terme l'anàlisi diferencial dels RNAs als 2 tipus de processament inicial de la mostra amb eines bioinformàtiques de l'RNA a nivell de transcriptoma. El control de qualitat de les llibreries essent especialment crític s'ha realitzat al centre "Centro Nacional de Análisis Genómico (CNAG)". A continuació es duu a terme l'alineament de les seqüències, tot fent una selecció de les millors alternatives per dur a terme aquesta tasca. Es fa una visualització de l'alineament i es passa al recompte de seqüències i normalització.

Així mateix, es va voler fer un estudi de RNA circular (*circRNA*), un tipus d'RNA no codificant més estable que el linear i que presenta patrons d'expressió específics de tipus cel·lular. A més, s'ha observat que poden regular l'expressió gènica tan a nivell transcripcional com post-transcripcional [17].

Per l'anàlisi d'aquest tipus de RNA havíem planejat utilitzar els scripts de KNIFE (Known an Novel IsoForm Explorer). No obstant aquest algorisme únicament permet detectar *circRNA* que es troben en exons. Addicionalment modificarem els scripts per poder detectar aquells *circRNA* que es troben a introns ja que les dades preliminars del grup suggereixen que aquests *circRNA* intrònics poden ser molt abundants i importants en espermatozoides.

La rellevància d'aquesta part del treball es trobava en obtenir unes eines i un procediment de treball que pugui se utilitzat per tothom que estigui interessat en l'estudi dels *circRNAs* de l'espermatozoide. Per desgracia per l'ajustat de la planificació aquesta part ha quedat fora de l'abast d'aquest TFM.

A.2. Justificació del TFM

El grup de recerca del Dr. Oliva, tot i la gran experiència en òmiques es troba en un punt on requereix el suport de un bioinformàtic per desenvolupar les seves investigacions. L'establiment de noves eines i protocols bioinformàtics són necessitats del grup de treball, donant-me l'oportunitat de realitzar aquesta tasca en forma de treball final de màster.

D'altra banda, aquest TFM pot arribar a permetrem establir nous mètodes pel tractament de dades de proteòmica provinents de l'ús de TMT 10plex i cromatografia líquida seguida d'espectrometria de masses en tàndem (LC-MS/MS) i el seu estudi posterior. Igualment, em trobo amb el repte de buscar nous mètodes per l'estudi de RNA i *circRNAs* d'espermatozoide. Ens trobem amb un camp d'estudi nou on aplicar tècniques bioinformàtiques existents de una manera diferent i on potser caldrà resoldre les anàlisis amb noves tècniques i algorismes.

L'elecció de l'àrea de treball, l'anàlisi de dades òmiques es troba en un moment on la diversificació de les seves branques comença a fer difícil tenir una visió general de totes. Aquest treball em permet establir contacte amb dades que requereixen coneixements en proteòmica i transcriptòmica. La seva rellevància actualment i la possibilitat d'establir contacte a diferents nivells ha estat part decisiva en la elecció d'aquest TFM. La elecció d'aquest TFM a nivell professional em permet entrar d'una manera rellevant en el món de l'anàlisi de dades sense deixar de banda les ciències de la salut. El fet de poder arribar a publicar articles amb el grup de recerca em permet una projecció professional a nivell mundial. És sens dubte un dels fets més rellevants per escollir dur a terme aquesta tasca.

B. OBJECTIUS

B.1. Objectius generals

En aquest apartat trobem dos objectius bàsics:

1. Desenvolupar un algorisme bioinformàtic per a poder pal·liar defectes en la quantificació de proteïnes previ a la seva detecció i quantificació mitjançant marcatge amb TMT 10plex i cromatografia líquida seguida d'espectrometria de masses en tàndem (LC-MS/MS) on no es pot fer servir mètodes estàndards. Ho durem a cap explorant la utilitat de les correlacions dels pèptids intramostrals per l'anàlisi de les dades proteòmiques obtingudes mitjançant TMT 10plex i LC-MS/MS.

2. Establir el procediment d'anàlisi bioinformàtica de dades de *LongRNA* d'espermatozoide provinents de RNA-seq.

Aquests dos objectius bàsics donaran lloc a un treball bioinformàtic associat que llistarem en el següent apartat. La documentació d'aquest treball associat, el podem concebre com un tercer objectiu:

3. Establir les pautes de treball d'anàlisi bioinformàtic aplicable a l'anàlisi de dades òmiques provinents de l'espermatozoide o el fluid seminal.

B.2 Objectius Específics

1. Per assolir el primer objectiu ha calgut assolir els següents objectius específics:

   - ✓ Determinar l'aproximació més adient per tractar les taules de resultats dins de R.
   - ✓ Obtenir les correlacions entre tots els pèptids detectats i quantificats en una mostra (mitjançant una funció d'R).
   - ✓ Obtenir les proteïnes altament correlacionades amb un tant per cent de pèptids altament correlacionats associat (mitjançant una funció d'R).
   - ✓ Aplicar les funcions a diferents subpoblacions fent diferents conjunts.
   - ✓ Obtenir resultats: Anàlisi per detectar desregularitzacions de proteïnes que apareixen als grups estudiats.

2. En el cas del segon objectiu podem dividir els objectius específics que ha calgut assolir que afecten a *longRNA*:

   - ✓ Determinar el programari a fer servir per l'anàlisi de dades RNA-seq.
   - ✓ Procedir a l'anàlisi de les dades.
   - ✓ Obtenir alineament.
   - ✓ Identificar i quantificar l'expressió de longRNA.
   - ✓ Analitzar estadísticament els RNA diferencialment expressats.

3. En el cas del tercer objectiu ens ha calgut assolir els següents objectius específics:

- ✓ Instal·lar tot el programari necessari per dur a terme una anàlisi bioniformàtica "estàndard" al grup de recerca. (Això es duu a terme en part en els objectius anteriors).
- ✓ Deixar operativa una estació de treball.

C. ENFOCAMENT I MÈTODE A SEGUIR

La divisió clara del treball que he abordat en aquest TFM fa que calgui una divisió entre objectius per poder determinar l'enfocament i el mètode a seguir.

C.1 Primer objectiu

En primer lloc per dur a terme el primer objectiu, arribar a un mètode bioinformàtic per poder pal·liar defectes de quantificació de les mostres de TMT 10plex i cromatografia líquida seguida d'espectrometria de masses en tàndem (LC-MS/MS), podem determinar 2 enfocaments:

- ✓ Intentar determinar els errors comesos i aplicar un factor per corregir la quantificació.
- ✓ Determinar la quantificació a partir de les relacions intramostrals.

L'enfocament que dins del grup de recerca s'ha escollit per considerar-lo més adient ha estat el segon.

El mètode a seguir passa per partint de zero crear un algorisme "*de novo*" per fer correlacions entre pèptids detectats i quantificats dins de la mostra mitjançant el programari R que pugui assolir el objectiu de correlacionar tots els valors dels pèptids per cada mostra i pèptid a pèptid.

Una vegada obtingudes aquestes correlacions, es va fer un estudi que comparava i buscava correlacions entre proteïnes per tal d'obtenir aquelles que tenen més del 95% dels seus pèptids altament correlacionats (coeficient de Pearson >0.9).

C.2.A Segon objectiu

El segon objectiu passa obligatòriament per seguir un mètode normalitzat d'anàlisi de dades RNA-seq per anar-se adaptant a les peculiaritats dels espermatozoides. Es planteja doncs la dicotomia:

- ✓ Plantejar un procediment especialitzat.
- ✓ Seguir el procediment habitual d'anàlisi de dades RNA-seq i anar adaptant-lo a les peculiaritats.

En aquest cas tan jo com el grup de recerca hem arribat a la conclusió que calia adaptar la metodologia però que ens desviariem de la convencional quan sigui necessari per la seva menor dificultat a l'hora d'abordar el problema.

Doncs ens ha calgut establir, documentar i protocol·litzar el procediment bàsic d'anàlisi de dades de RNA-seq (Control de qualitat, mapeig i visualització, visualització, recompte i normalització...) i detectar els punts on cal fer canvis en la metodologia.

C.3 Tercer objectiu

El tercer objectiu, l'enfocament podia ser:

- ✓ Documentar allò que es faci servir en funció del seu ús.
- ✓ Fer un escaneig i determinar quines eines serien útils.

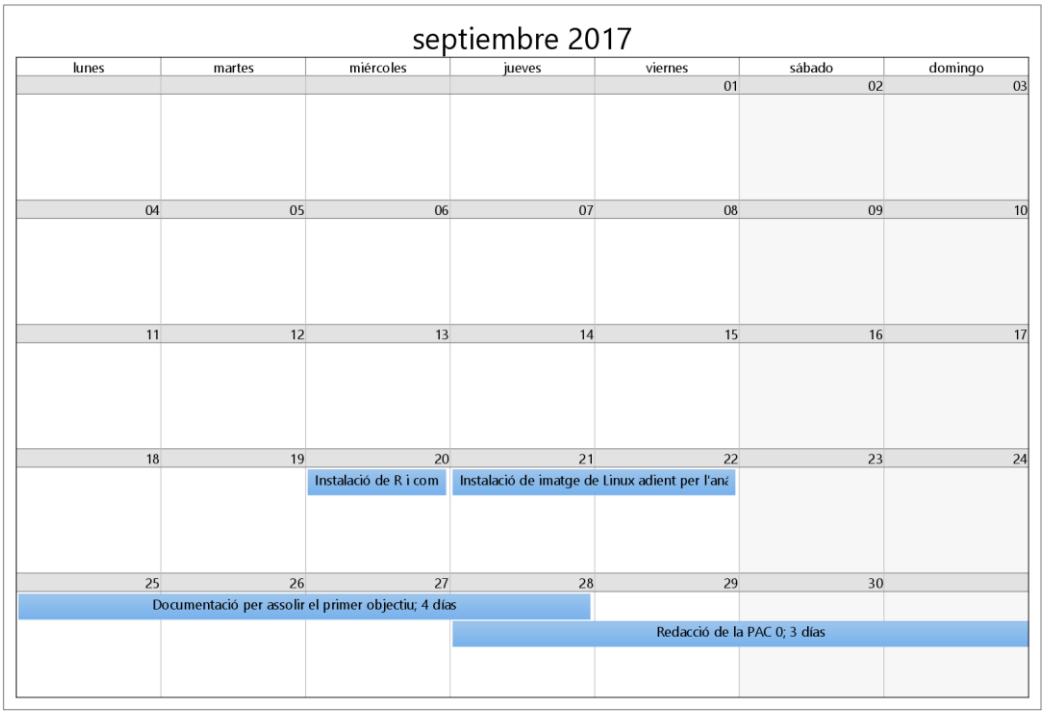
En el nostre cas donat el caràcter d'aquest treball vam fer una junció de ambdues aproximacions i fent un escaneig previ de les tècniques disponibles documentar aquelles que ens siguin útils (o fins i tot aquelles que podrien resultat d'utilitat en casos similars al nostre).

D. PLANIFICACIÓ AMB FITES I TEMPORITZACIÓ

D.1 Tasques

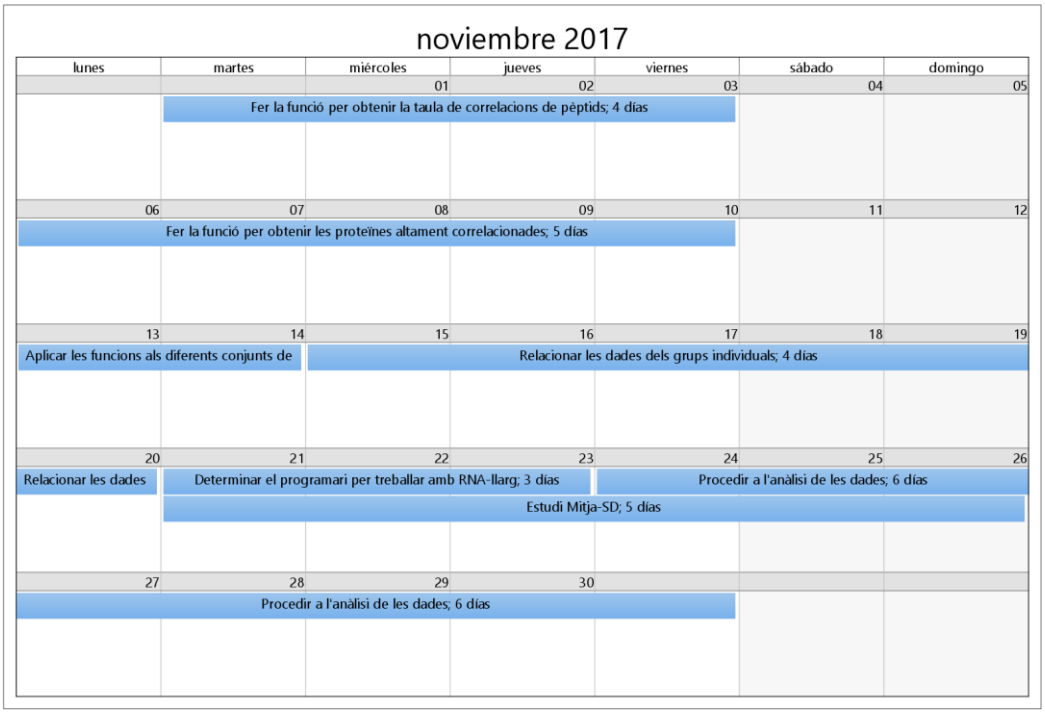
- **PAC 0: Definició dels continguts del treball**
  - Instal·lació de R i complements
  - Instal·lació de imatge de Linux adient per l'anàlisi bioinformàtica
  - Documentació per assolir el primer objectiu
  - Redacció de la PAC 0
- **PAC 1: Pla de Treball**
  - Documentació per Redacció del Pla de Treball
  - Redacció del Pla de Treball
- **PAC 2: Desenvolupament del Treball- Fase I**
  - Realitzar l'estudi de forma convencional
  - Determinar la forma de treballar amb les taules
  - Fer la funció per obtenir la taula de correlacions de pèptids
  - Fer la funció per obtenir les proteïnes altament correlacionades
  - Aplicar les funcions als diferents conjunts de poblacions
  - Relacionar les dades dels grups individuals
- **PAC 3: Desenvolupament del Treball- Fase II**
  - Obtenir i comentar els resultats
  - Estudi Mitja-SD
  - Determinar el programari per treballar amb RNA-llarg
  - Procedir a l'anàlisi de les dades
  - Obtenir l'alineament
  - Anàlisi estadístic dels RNA diferencialment expressats.
- **PAC 4: Redacció de la Memòria**
  - Acabar anàlisi estadístic dels RNA diferencialment expressats.
  - Tancament dels anàlisis
  - Documentar el programari associat a les anteriors Fases
  - Redacció Final de la Memòria
- **PAC 5A: Elaboració de la Presentació**
  - Elaboració de la Presentació
- **PAC 5B: Defensa Pública**
  - Defensa Pública

D.2 Calendari



Il·lustració 1: Setembre 2017

Il·lustració 2: Octubre 2017

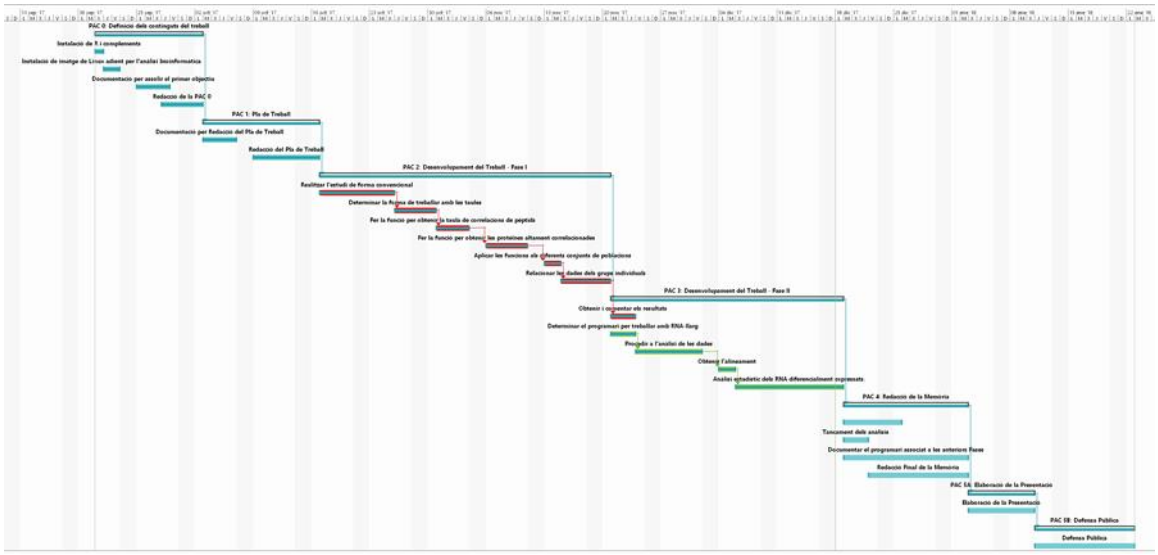


Il·lustració 3: Novembre 2017

Il·lustració 4: Desembre 2017

Il·lustració 5: Gener 2018

Diagrama de Gantt



Il·lustració 6: Diagrama de Gantt

Es pot trobar al document adjunt "Delgado_Dueñas_David_TFM.gif" i "Delgado_Dueñas_David_TFM.mpp" amb millor resolució aquest mateix diagrama.

D.3 Fites

He considerat les Fites les entregues de les PACs, donat que són un punt de control on podrem veure l'avenç del treball i posar-lo en comú amb els tutors.

PAC 0: Realitzar l'entrega de la definició de continguts.	02/10/2017
PAC 1: Realitzar l'entrega del pla de treball.	16/10/2017
PAC 2: Tenir acabat l'objectiu 1 i avançat l'objectiu 2A.	20/11/2017
PAC 3: Tenir acabat l'objectiu 2A, 2B(o bé molt avançat) i el 3.	18/12/2017
PAC 4: Tenir assolits tots els objectius i entregada la Memòria.	02/01/2018
PAC 5A: Haver fet la presentació.	10/01/2018
PAC 5B: Haver dut a terme la defensa pública.	22/01/2018

D.4 Anàlisi de riscos

En fer una ullada als riscos habituals dels projectes, els que poden afectar aquest principalment són:

1. Temporització massa ajustada: en fer la temporització he intentat seguir una distribució de temps realista per no haver de fer grans canvis. Tot i així, poden presentar-se inconvenients de diferents tipus que facin retardar alguna de les tasques portant a una cascada de retards que caldrà afrontar.

2. El abast del projecte massa ampli: al fer la proposta de TFM hem deixat fora de l'abast del mateix una sèrie de anàlisis que el grup de recerca durà a terme per no comprometre l'assoliment dels objectius. Per exemple, el grup està realitzant el processament i estudi proteòmic mitjançant TMT així com les llibreries *smallRNA* (<200nucleòtids) de les mostres estudiades en la segona part del TFM.

3. Inexperiència amb la tecnologia: es evident que la meva experiència amb les diferents tecnologies és limitada, tot i el meu esforç podem trobar problemes per aquesta inexperiència que haurem de suplir mitjançant la documentació i possiblement amb hores no contemplades a la temporització per poder assolir els objectius a temps.

4. Tasques estimades d'una manera no realista: la documentació prèvia a la realització de la temporització ha intentat ser suficientment exhaustiva per donar lloc a una temporització realista, però és evident que la inexperiència en alguns camps pot dur a que les tasques hagin estat malament estimades.

E. RESULTATS ESPERATS

En acabar el TFM haurem que obtingut:

- Pla de treball: document que un cop acabat determinarà uns objectius, un enfocament, una metodologia i una temporització determinats pel nostre TFM.

- Memòria: document on donarem context a la nostra recerca i s'explicarà detalladament la manera com s'ha arribat als objectius.

- Producte: Obtindrem alguns scripts desenvolupats en R També tindrem productes de l'anàlisi de dades com planes Excel de sortida de dades. Molt possiblement el grup de recerca pugui publicar entre 2 i 3 articles relacionats, tot i que quedaran fora del TFM per raons temporals.

- Presentació virtual: Un vídeo o presentació digital on s'explicarà tot el procés per l'obtenció dels resultats exposats a la memòria.

- Autoavaluació del treball: un document on críticament valoraré el meu treball.

F. ESTRUCTURACIÓ DEL PROJECTE

En una aproximació al que haurà de contenir el meu treball un cop completat podrem trobar:

1. Introducció: bàsicament el contingut de la PAC 1 amb alguna lleu modificació per adaptar-lo a la resta de la documentació.

PART 1: Proteòmica del fluid seminal

1. Introducció a la proteòmica del fluid seminal: Una breu explicació de allò que estudia la proteòmica concretant en els fets rellevants de la proteòmica del fluid seminal.
2. Estudi d'eines bioinformàtiques útils per l'estudi de dades provinents de l'estudi de proteòmica (TMT-MS/MS): En aquest apartat partirem dels resultats que ens avoca la proteòmica (raw-data) i buscarem quines eines bioinformàtiques son les adients per obtenir el resultat esperat.
3. Mètode bioinformàtic aplicat a l'estudi: En aquest apartat discutirem quina serà la manera d'obtenir els resultats un cop escollida la eina en l'apartat anterior.
4. Resultats obtinguts: Presentació dels resultats obtinguts en l'estudi dut a terme.
5. Discussió dels resultats: En aquest apartat analitzarem els resultats obtinguts.

PART 2: Transcriptòmica de l'espermatozoide

1. Introducció a la transcriptòmica d'espermatozoide: Una breu explicació de allò que estudia la transcriptòmica concretant en els fets rellevants de la transcriptòmica d'espermatozoide.
1. Estudi d'enes bioinformàtiques útils per l'estudi de dades provinent de l'estudi de transcriptòmica (*longRNA*): En aquest apartat analitzarem les dades que hem de tractar i buscarem les eines bioinformàtiques que podem aplicar.
2. Mètode bioinformàtic aplicat a l'estudi: En aquest apartat indicarem quin és el procediment analític que seguirem per obtenir els resultats esperats.
3. Resultats obtinguts: Presentació dels resultats que hem obtingut.
4. Discussió dels resultats: Anàlisi dels resultats obtinguts.

2. PROTEÒMICA DEL FLUID SEMINAL

A. INTRODUCCIÓ A LA PROTEÒMICA DEL FLUID ESPERMÀTIC

El proteoma és el conjunt de proteïnes expressades per un genoma, cèl·lula, teixit o organisme en un moment i circumstàncies determinades. Per dur a terme l'estudi d'aquest conjunt de proteïnes i el seu efecte s'han desenvolupat una sèrie de procediments i mètodes d'alt rendiment que permeten la identificació, quantificació i caracterització d'un gran nombre de proteïnes en un sol experiment.

Degut a la disponibilitat de bases de dades amb seqüències genòmiques i als mètodes d'ionització de pèptids, l'espectrometria de masses (MS) s'ha convertit en el mètode d'elecció per dur a terme els anàlisis proteòmics a gran escala. Aquesta tècnica es basa en la fragmentació de les proteïnes en pèptids, la seva separació per cromatografia líquida i la seva identificació a partir de les relacions exactes de massa/càrrega dels ions generats.

En l'actualitat, un total de 2064 proteïnes han estat identificades en el plasma seminal mitjançant l'ús de tècniques d'alt rendiment com és la MS. El plasma seminal es un fluid complex que està compost per les secrecions provinents del testicle (1-2%), epidídim (2-4%) i les glàndules sexuals accessòries presents en el tracte urogenital masculí, que inclou les vesícules seminals (65-75%), la pròstata (25-30%) i les glàndules bulbouretrals (<1%).

La funció més bàsica del plasma seminal és la d'actuar com a medi de transport dels espermatozoides durant el procés d'ejaculació i de trànsit pel tracte femení fins arribar a l'oòcit, no obstant recents estudis mostren que els components del plasma seminal també són crucials per la supervivència i funció dels espermatozoides [18].

L'accessibilitat i l'alta concentració de proteïnes del plasma seminal suggereix que el plasma seminal pot ser una font rica de biomarcadors per a la fertilitat masculina, de la mateixa manera que s'han identificat en l'espermatozoide [8]. A més, alteracions de les proteïnes del plasma seminal s'han relacionat amb diversos aspectes implicats en la infertilitat masculina [19–22]

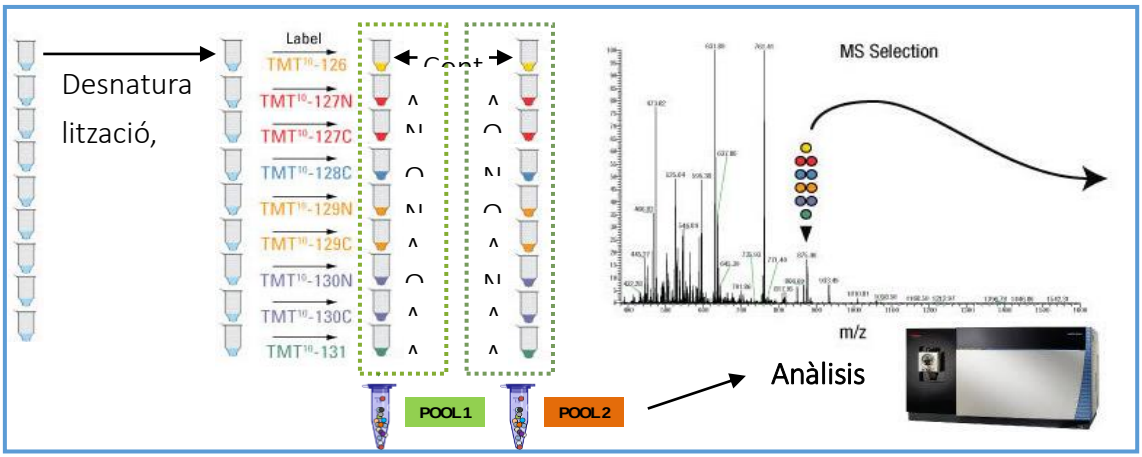
Per tal d'avaluar el potencial rol de les proteïnes del plasma seminal en la funcionalitat i integritat dels espermatozoides, el grup de recerca ha dissenyat un estudi per tal de caracteritzar i comparar el proteoma del plasma seminal mitjançant espectrometria de masses (LC-MS/MS) de diferents subtipus de pacients infèrtils d'acord amb els seus paràmetres seminals:

(i) 4 pacients normozoospèrmics (NZ; paràmetres seminals normals)
(ii) 4 pacients astenozoospèrmics (AS; espermatozoides amb motilitat alterada)
(iii) 4 pacients oligozoospèrmics (OZ; baixa concentració d'espermatozoides)
(iv) 4 pacients azoospèrmics (AZ; absència d'espermatozoides)

Per tal de poder quantificar i comparar la expressió de les proteïnes s'ha emprat la tecnologia TMT (*tandem mass tag*) combinat amb LC-MS/MS. Aquest mètode es basa en

el marcatge de proteïnes de diferents mostres amb marcadors isobàrics específics (amb la mateixa massa) que un cop es fragmenten alliberen un senyal iònic únic que es visualitza a l'espectre MS/MS. Les intensitats d'aquests senyals iònics són usades per calcular la quantitat relativa d'una proteïna entre les diferents mostres.

Breument, es va utilitzar el kit TMT10-plex que disposa de 10 *tags* diferents i ens permet l'estudi de 8 mostres diferents en un mateix experiment de LC-MS/MS . El dos canals restants, en un no s'afegeix mostra i es la que ens determinarà el soroll de fons de la tècnica y l'altre correspon al control intern que correspon a un pool proteïnes provinents de la mateixa quantitat de les 16 mostres a estudiar. El disseny experimental de l'estudi es mostra en la figura 7.



Il·lustració 7: Disseny Experimental TMT combinat amb LC-MS/MS

El anàlisis quantitatiu estàndard del mètode TMT es basa en l'estudi de la relació de la quantitat de proteïna en una mostra concreta amb la quantitat de proteïna present en el control intern. Per tant, no es una quantificació absoluta sinó una quantificació relativa. Un cop identificada una proteïna es calcula la relació entre la intensitat del *tag* d'una mostra concreta entre la intensitat del *tag* del control intern.

B. ESTUDI D'EINES BIOINFORMÀTIQUES ÚTILS PER L'ESTUDI DE DADES PROVINENTS DE L'ESTUDI DE PROTEÒMICA (TMT-MS/MS)

Les dades que ens arriben per l'estudi que es troben a la carpeta "Datos" de la carpeta anàlisi convencional, ja han patit un pretractament per intentar corregir els errors de quantificació de les mostres de TMT 10plex i cromatografia líquida. Això es fa evident al full "Data BO_sense buits i amb MPC" on es veu que al costat de cada columna de expressió hi ha una columna de color lila en la que es té el resultat d'aplicar un pretractament a les dades. Aquest pretractament ha consistit en una correcció de l'expressió obtinguda mitjançant l'aplicació de un algorisme corrector:

Genes	# Unique Peptide	Σ# PSMs	N1	N1 norm (x 1,371363)	N2	N2 norm (x 1,087514)	(...)
SEMG2	52	1862	0.900	1.235	1.163	1.265	(...)
(..)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
CLSTN1	1	2	0.756	1.037	0.835	0.908	(...)
ABLIM3	1	2	0.730	1.001	3.480	3.785	(...)
KIDINS220	1	2	1.007	1.381	1.071	1.165	(...)
LAMA3	1	2	0.375	0.514	0.534	0.580	(...)
	Total		Col N1	Multiplicador N1	Col N2	Multiplicador N2	(...)
Promedio	1.034257421		0.75418242	1.371363	0.9510292	1.087514	(...)
Desv est	0.188815708		0.17921607		0.27740807		(...)

Il·lustració 8: Dades d'origen de l'estudi

1. Procedim a fer la mitjana per columna (Col N1, Col N2,..).
2. Fem el mitjana d'aquests valors (Total).
3. Dividim la mitjana per columnes per aquest valor (Multiplicador N1, MultiplicadorN2, ...).
4. Multipliquem tots els valors de les mostres per el resultat d'aquestes divisions. (N1 norm, N2 norm,...).

Aquesta correcció ve donada per la reflexió que van fer els investigadors que els va dur a la conclusió que el mitjana de l'expressió de totes les proteïnes identificades i quantificades hauria de ser 1.

Un cop establert l'origen de les dades es va abordar com treballar amb elles. D'un bon començament es va adreçar l'anàlisi cap a R. Encara i així potser hauria estat possible dur-lo a terme amb Python i les seves eines matemàtiques, inclús podent integrar R [23–26]. Però donat el caràcter estadístic dels tractaments que es volien dur a terme i l'experiència per la meva part en l'ús de R molt major que amb Python o altres solucions estadístiques ens va fer decantar per R.

Es va escollir R-Studio com a solució integrada per la gestió de desenvolupament mitjançant R [27]. En una primera aproximació durant el plantejament del projecte es va optar pel desenvolupament d'un algorisme nou que ens hauria de permetre el treball amb taules de correlació, però com a requisit previ, en el grup es va parlar d'adoptar en part el primer enfocament que es va fer per solucionar el problema, es a dir dur a terme

un estudi aplicant factors de correcció. D'aquesta manera disposarem d'un referent, potser amb les seves mancances, però amb el que podrem extreure certa informació.

Per dur a terme aquest enfocament i donada la naturalesa dels resultats que volíem obtenir que no donaven lloc a taules excessivament grans i sí a un gran nombre d'elements gràfics, es va optar per fer servir un script en R-markdown.

Un cop fet aquest anàlisi passarem al anàlisi contemplat en la planificació que es basa en obtenir una aproximació de la quantificació de les mostres mitjançant les seves relacions intramostrals.

Aquest segon anàlisi es centrarà en taules de gran mida pel que l'adient es fer servir sortides en MS Excel i si s'escau algun gràfic li podrem donar sortida mitjançant un arxiu d'imatge.

Un cop obtingudes les proteïnes amb una relació alta, passarem a veure com s'expressen en els diferents grups poblacionals que graficarem en un diagrama pel que farem servir un script d'R que ens doni com a sortida un diagrama de Veen.

Amb tot això, passarem a fer un estudi determinant quin és el nivell d'expressió de les proteïnes destacades en la població normal per enfrontar-ho a les poblacions amb alteracions als paràmetres seminals. Per dur a terme aquesta feina novament farem servir R.

Com a feina derivada d'aquest treball es va crear una funció que obtenia dades des de la web d'Uniprot (http://www.uniprot.org/), a partir d'aquesta funció al grup va aparèixer la necessitat d'obtenir a partir d'un llistat de proteïnes en codi d'entrada Uniprot els GO separats per tipus (F,C,P) i la funció de la proteïna. Es va ampliar la funció amb R, a més, aquesta funció per fer-la més accessible als components que no són afins a la programació en R es va decidir fer servir Shiny i presentar-la com una webApp. El script d'R i la webApp generada a partir d'aquest pot ser consultat en la documentació adjunta a aquesta Memòria, la webApp es pot trobar a: https://uniproter.shinyapps.io/uniproter/.

C. MÈTODE BIOINFORMÀTIC APLICAT A L'ESTUDI

C.1 ANÀLISI CONVENCIONAL

En aquest primer script sense fer majors correccions que les aplicades pels investigadors es va estudiar la distribució de la mostra mitjançant els test de Shaphiro-Wilk [28] i el test de Anderson-Darling[29] aplicant el *algorisme 1* amb una sortida encapçalada per la *taula 1*, que ens indica que segons les proves els dos test no afirmen que la distribució sigui normal.

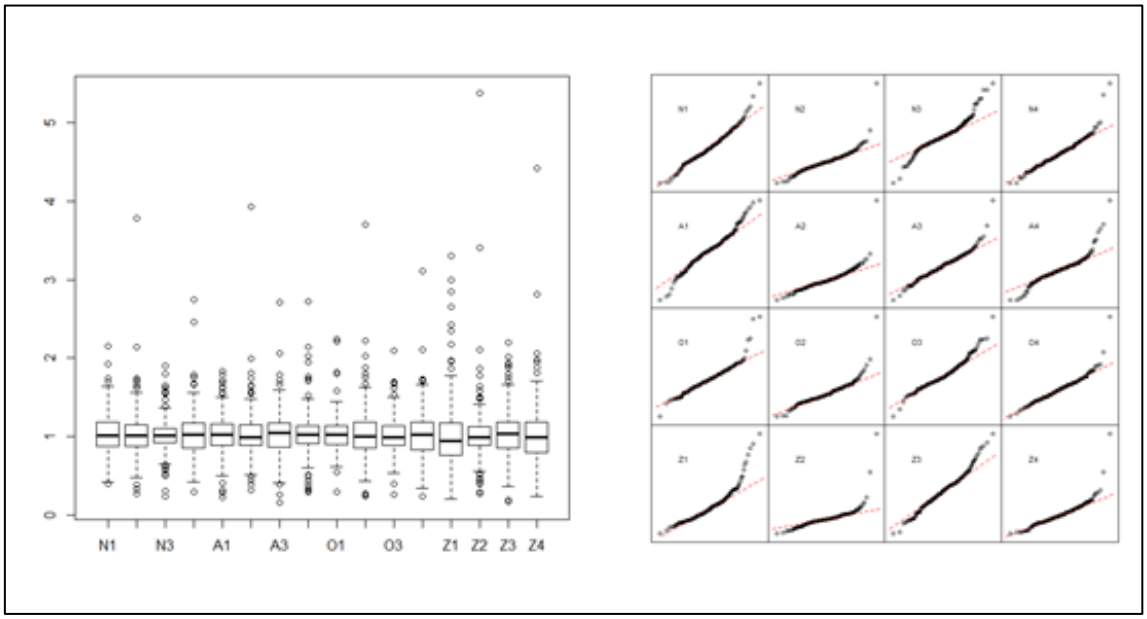
```
## Realizamos test de normalidad shapiro-Wilk y Anderson-Darling
Normal<- function(E){
  LoQueEsNormal <- data.frame(Proteina=character(), S.W=character(),
pvalor=numeric(), A.D= character(), porcentaje=numeric(),
stringsAsFactors=FALSE) #Creamos una tabla vacía

  for (i in 1:ncol(E)){
    # Realizamos el test
    Alpha <- shapiro.test(as.array(E[,i]))
    P <- Alpha$p.value
    ADT <- ad.test(E[,i])
    AD <- ADT$p.value

    if (P>=0.05){
      Wii <- data.frame(Proteina = colnames(E)[i], S.W= "Sí",
pvalor=round(P, 4), A.D= "", porcentaje = AD*100)
    }else{
      Wii <- data.frame(Proteina = colnames(E)[i], S.W = "No",
pvalor=round(P, 4), A.D= "", porcentaje = AD*100)
    }

    if(AD > 0.5){
      Wii$A.D = "Sí"
    }else{
      Wii$A.D = "No"
    }

    LoQueEsNormal <- rbind(LoQueEsNormal, Wii)
  }

LoQueEsNormal
}
```

Algorisme 1: Funció Normal

Proteina	S.W	pvalor	A.D	porcentaje
N1	No	1e-04	No	0.3188338
N2	No	0e+00	No	0.0000000
N3	No	0e+00	No	0.0000000

Taula 1: Part de la sortida de la funció "Normal"

Seguidament es va dur a terme un boxplot per veure la distribució dels elements i una gràfica Q-Q (*Il·lustració 9*) per poder veure les distribucions que segueixen els nivells d'expressió dins de cada proteïna en cadascuna de les mostres. Amb això podrem determinar quin és el tipus d'anàlisi adient pel nostre conjunt de dades (paramètrics o no paramètrics.)



Il·lustració 9: Boxplot i gràfica Q-Q en dades d'origen

Amb aquests test van poder veure que la distribució de les mostres no era normal, cosa que ens avoca a dos tractaments possibles: Normalitzar-les o bé deixar-les com estan.

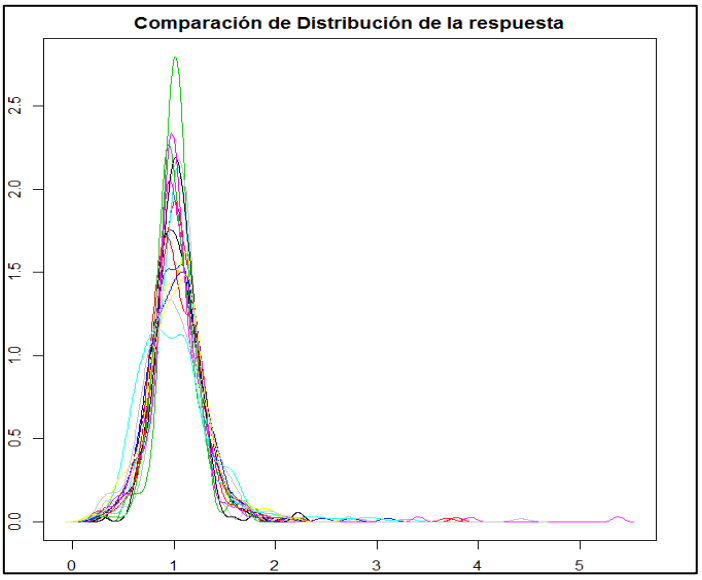
Per deixar-les com estan i fer servir test no paramètrics ens cal una certa coneixença de la distribució que ens dona el gràfic de densitats. Es va aplicar la següent funció per obtenir-la:

```
## Realizamos test de densidad de una matriz
densited <- function(E){

  par(mfrow=c(1,1))
  par(mar=c(2,2,2,2))

  dens <- apply(E, 2, density)
  plot(NA, xlim=range(sapply(dens, "[", "x")),
ylim=range(sapply(dens, "[", "y")), ylab="Densidad",
main="Comparación de Distribución de la respuesta")
  mapply(lines, dens, col=1:length(dens))
}
```

Algorisme 2: Funció densited

Al gràfic de densitats podem veure que tot i que hi ha elements divergents les

distribucions mostrals son aproximadament iguals. Cal veure que hi ha algunes mostres que tenen una distribució més irregular però centrada en la mateixa zona que la resta.



Il·lustració 10: Distribució de la resposta abans de normalitzar

Amb els valors obtinguts, i amb la certesa que la mostra no és normal però que la distribució és bastant semblant com per ser comparada vaig establir un primer estudi de les dades.

Vam dur a terme el test de Kruskal-Wallis (*Algorisme 3*) a les mostres per veure si s'hi observaven diferencies en les variàncies que poguessin ser ressenyables, donat que al no ser normals no s'hauria d'aplicar ANOVA . Podem veure l'encapçalament de la sortida a la *taula 2*.

```
# Columna por columna hacemos KW vs los Grupos
SignificantProteinKW <- data.frame(Proteina = character(), P.valor =
character(),stringsAsFactors = FALSE)  #Creamos una tabla vacía
NotSignificantKW <- data.frame(Proteina = character(), P.valor =
character(),stringsAsFactors = FALSE)  #Creamos una tabla vacía

# Empezamos un bucle for desde 2 hasta la última columna
for (i in 2:(ncol(raw.work.t) - 1)) {
    # Realizamos el test
     Alpha <- kruskal.test(raw.work.t[, i] ~ Grupos, raw.work.t)
     # Nos quedamos el p.valor
     P <- unlist(Alpha)["p.value"]
    Wii <- data.frame(Proteina = colnames(raw.work.t)[i], P.valor =
round(as.numeric(P),
          digits = 3))
    # Si el p.valor es inferior a 0.05 anotamos la proteína y
    # mostramos su p-valor
    if (P < 0.05) {
          SignificantProteinKW <- rbind(SignificantProteinKW, Wii)
    } else {
          NotSignificantKW <- rbind(NotSignificantKW, Wii)
   }
}
```

Proteina	P.valor
NPC2	0.009
CAMP	0.046
PTGDS	0.050

Taula 2: Part de la sortida del codi KW

Seguidament vaig dur a terme una correlació de la Concentració de espermatozoides amb els nivells d'expressió de les proteïnes i el mateix amb la motilitat. Per últim vaig corregir els valors aplicant FDR per evitar falsos positius (*algorisme 4*). Podem veure l'encapçalament de la sortida a la *taula 3*.

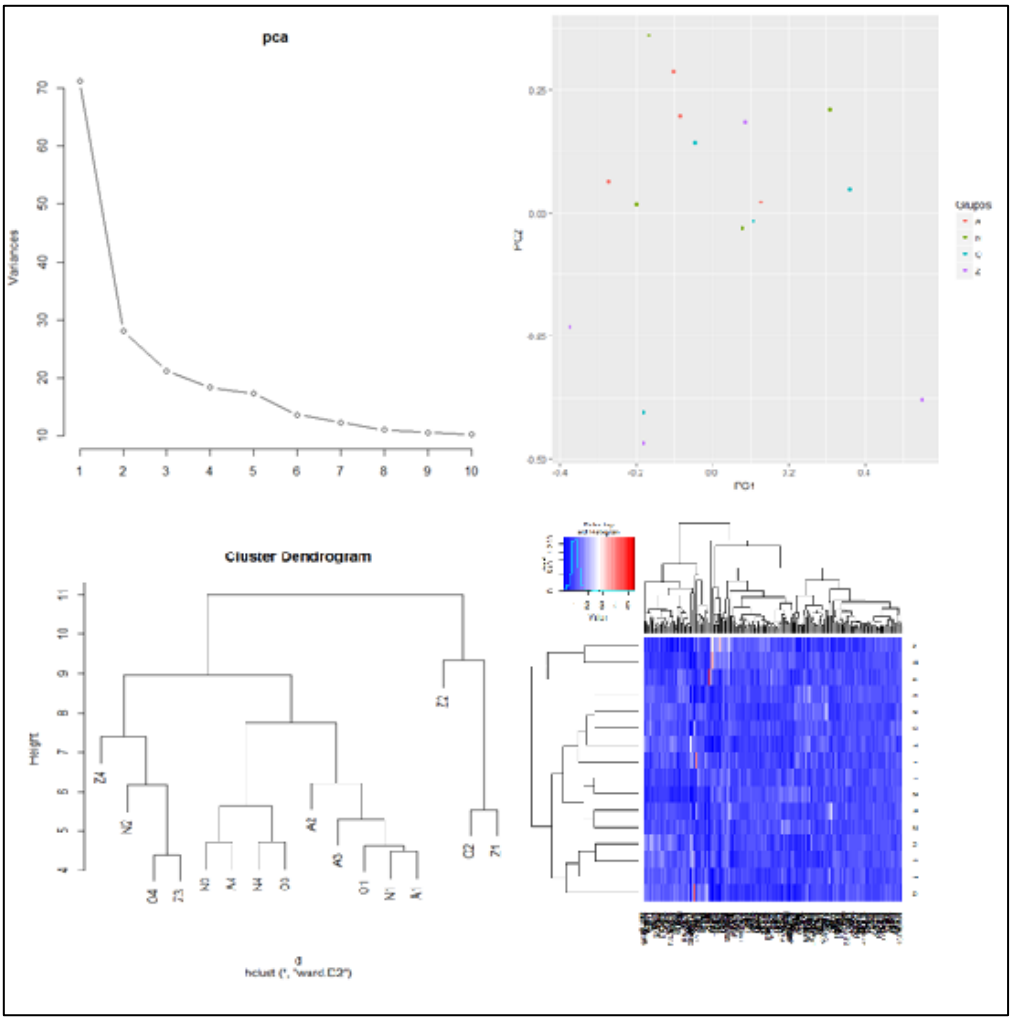
```
# CORRELATION: C VS PROTS
M<-PcaTotalF[3:18, 4]
C<-PcaTotalF[3:18, 3]
Pear.C.raw <- cbind(C,raw.work.t[c(-1)])

## Los Azo tienen el mismo valor al correlacionar no obtendremos
valores exactos porque hay "ties" (coincidencias) en la variable
dependiente
cor.c.raw <- correlate(Pear.C.raw, "spearman")
head(cor.c.raw[with(cor.c.raw, order(corrA, decreasing=T)),][1:3])
head(cor.c.raw[with(cor.c.raw, order(corr)),])

# cORRELATION: MOVILIDAD VS PROTS
Pear.M.raw <- cbind(M,raw.work.t[c(-1)])
cor.m.raw <- correlate(Pear.M.raw,"spearman")
head(cor.m.raw[with(cor.m.raw, order(corrA, decreasing=T)),][1:3])

# FDR
cor.c.raw.fdr <-pHCorr(cor.c.raw,"fdr")
head(cor.c.raw.fdr[with(cor.c.raw.fdr, order(corrA,
decreasing=T)),][c(1,2,3,5)],10)

cor.m.raw.fdr <-pHCorr(cor.m.raw,"fdr")
head(cor.m.raw.fdr[with(cor.m.raw.fdr, order(corrA,
decreasing=T)),][c(1,2,3,5)],10)
```

Algorisme 4: Tros de codi que calcula Correlació i FDR

	Prot	sign	corr	FDR
rho	C	0.000	1.000	0.00000
rho181	LDHC	0.000	0.853	0.00000
rho10	NPC2	0.000	0.790	0.00000
rho39	ECM1	0.001	0.763	0.06250

Taula 3: Exemple de sortida FDR

Seguidament es duu a terme un PCA que no avoca cap distribució ni agrupació rellevant. Per últim es va fer un dendrograma y un heatmap de la distribució de l'expressió.



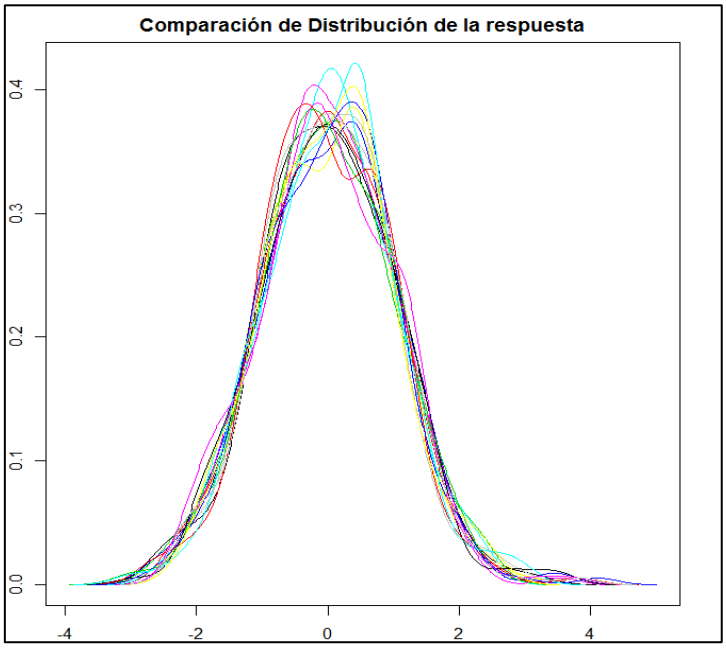
Il·lustració 11: PCA (variació), PCA (distribució), Dendograma i Heatmap

L'altra opció que trobàvem en tenir la distribució de mostres allunyada de la normal era normalitzar-la. Donada la manipulació prèvia feta de les dades en un primer moment no vaig voler dur a terme més manipulacions, però al final vaig fer una normalització de les mostres mitjançant la transformació de Johnson que es basa en la aplicació de t-test per normalitzar les mostres.[30]

```
## Normalizar una matriz mediante el método de Johnson
JohnM <- function(E){

  for (i in 1:ncol(E)){
    x_johnson <- RE.Johnson(E[,i])
    E[,i] <- as.data.frame(x_johnson$transformed)
  }
 return(E)
}
```

Algorisme 5: Normalització per Johnson

mostres (Il·lustració 12). Tot i així, la manipulació de les mostres prèvia afegida a aquesta transformació ens dur a patir per la possible pèrdua d'informació rellevant pel que continuem tenint en compte els resultats obtinguts al primer estudi.



Il·lustració 12: Distribució de la resposta després de la normalització

Un cop comprovada la distribució amb les eines utilitzades en l'anàlisi anterior podem veure que la distribució s'ha normalitzat en la majoria de les mostres.

Tornem a fer l'anàlisi però aquest cop amb proves paramètriques ANOVA i trobem resultats molt similars als anterior tot i les reticències que trobàvem al fer el primer anàlisi.

```
#Empezamos un bucle for desde 2 hasta la última columna
for (i in 2:(ncol(work.t)-1)){

    if (shapiro.test(work.t[,i])$p.value>=0.05){

      # Realizamos el test
      Alpha <- oneway.test(work.t[,i]~Grupos, work.t)
      # Nos quedamos el p.valor
      P <- unlist(Alpha)["p.value"]
      Wii <- data.frame(Test="AN", Proteina = colnames(work.t)[i],
P.valor = round(as.numeric(P), digits=3 ))
      # Si el p.valor es inferior a 0.05 anotamos la proteína y
mostramos su anova
      if (P<0.05){
        SignificantProtein <- rbind(SignificantProtein, Wii)
      }else{
        NotSignificant <- rbind(NotSignificant, Wii)
      }
}else{
        # Realizamos el test
        Alpha <- kruskal.test(work.t[,i]~Grupos, work.t)
```

```
        # Nos quedamos el p.valor
        P <- unlist(Alpha)["p.value"]
        Wii <- data.frame(Test="KW", Proteina = colnames(work.t)[i],
P.valor = round(as.numeric(P), digits=3 ))
        # Si el p.valor es inferior a 0.05 anotamos la proteína y
mostramos su anova
        if (P<0.05){
          SignificantProtein <- rbind(SignificantProtein, Wii)
        }else{
          NotSignificant <- rbind(NotSignificant, Wii)
        }

    }
}

# Imprimimos la tabla de Proteínas significativas con su p-valor
SignificantProtein
head(NotSignificant[with(NotSignificant, order(P.valor)),])
```

Algorisme 6: ANOVA

Amb el *algorisme 6* vam obtenir una sortida com el encapçalament que es mostra:

Test	Proteina	P.valor
AN	NPC2	0.019
AN	ANPEP	0.017
AN	SCGB1A1	0.013

Taula 4: Sortida algorisme per aplicar ANOVA.

Aquest cop apliquem tests Post-Hoc, com Duncan i Tukey aplicant els *algorisme 7* i *l'algorisme 8* obtenint unes taules de treball que contenen les correccions de ANOVA per ambdós mètodes.

```
## [TEST EXPLORATORIO!!!] Test de Tuckey Post-HOC
sin <- as.vector(SignificantProtein[2])
Tukey.work.t <- data.frame(Proteina=character(), Vs=character(),
P.valor=character(), stringsAsFactors=FALSE)

for(i in 1:nrow(sin)){
  a<-as.character(sin[i,1])
  work.tA <- aov(get(a)~Grupos, work.t)
  at<- TukeyHSD(work.tA, 'Grupos',  conf.level=0.95)

  for (j in 1:nrow(at$Grupos)){
    P <- at$Grupos[j,4]
    if(P<0.05){
      Wii <- data.frame( Proteina=a, Vs=row.names(at$Grupos)[j],
P.valor=round(as.numeric(P), digits=3 ))
      Tukey.work.t <-rbind(Tukey.work.t,Wii)
    }
  }
}

Tukey.work.t
```

Algorisme 7: Tukey

```
## Test de Duncan

# Post-Hoc en principio toca Duncan (consideramos significante las
diferencias observadas en anova)
Duncan.work.t <- data.frame(Proteina= character(), type=character(),
groups=character(), stringsAsFactors = FALSE)
Dun.G <- data.frame(Proteina= character(), type=character(),
groups=character(), stringsAsFactors = FALSE)
Dun.T <- data.frame(Proteina= character(), vs=character(),
stringsAsFactors = FALSE)
Duncan.work.tF <- data.frame(Proteina= character(), vs=character(),
stringsAsFactors = FALSE)
Dun <- as.array(NA)
sin <- as.vector(SignificantProtein[2])

for(i in 1:nrow(sin)){

  a<-as.character(sin[i,1])
  work.tA <- aov(get(a)~Grupos, work.t)
  at<-  duncan.test(work.tA, 'Grupos',  alpha=0.05)

  for (j in 1:nrow(at$groups)){
    Dun=as.character(at$groups[j,3])
    Dun.G <- data.frame(Proteina=a,type=at$groups[j,1], groups=Dun)
    Duncan.work.t <- rbind(Duncan.work.t, Dun.G)
  }

}

Duncan.work.t
```

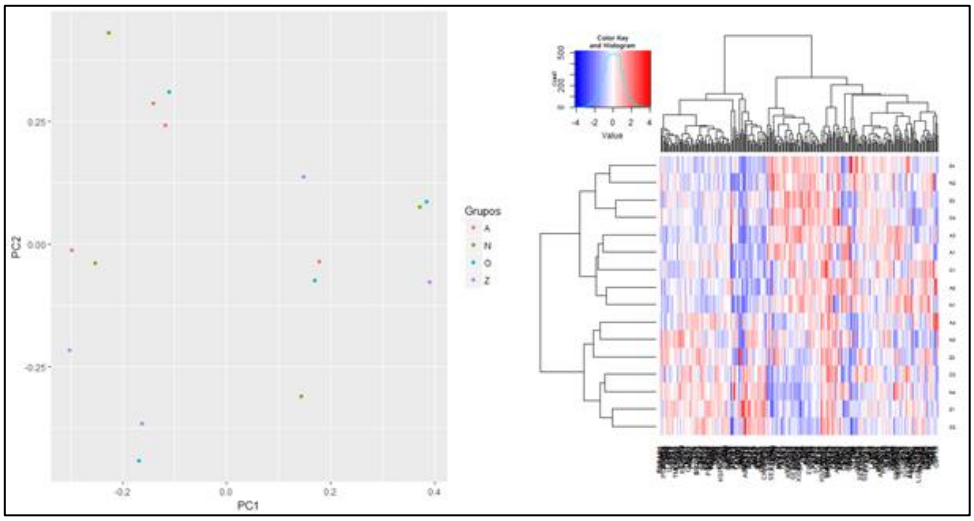
Algorisme 8: Duncan

Vam repetir la resta de l'anàlisi trobant moltes similituds també a les correlacions i la seva corresponent FDR, en aquest punt, cal destacar que les correlacions de les proteïnes LDHC i ECM1 amb la concentració d'espermatozoides ha estat confirmada recentment per el grup de recerca en un número major de mostres mitjançant la tècnica de Western Blot.

	Prot	sign	corr	FDR
rho	C	0.000	1.000	0.00000
rho181	LDHC	0.000	0.853	0.00000
rho10	NPC2	0.000	0.790	0.00000
rho39	ECM1	0.001	0.763	0.06250
rho138	SPINT3	0.004	0.677	0.14286
rho140	SHISA5	0.004	-0.677	0.14286
rho24	CAMP	0.004	0.674	0.14286
rho50	CRISP1	0.008	0.636	0.25000
rho61	PTGDS	0.009	0.628	0.25000
rho217	TSPAN1	0.014	0.601	0.35000

Taula 5: Correlació i FDR contra concentració

El PCA tampoc no retorna cap distribució rellevant com en l'anterior anàlisi però el heatmap revela una distribució més realista de l'expressió.



Il·lustració 13: PCA i heatmap amb dades normalitzades

Els anàlisis complets es poden trobar a l'Annex 1 i Annex 2

C.2. ANÀLISI NOU

Un cop duts a cap els anàlisis que faríem amb una mostra que no patís cap error de quantificació, es va passar a fer les anàlisis per intentar pal·liar aquest defecte.

En primer lloc es va fer un tractament de les taules de resultats per quedar-nos únicament amb els valors que ens són d'interès. Ens trobàvem amb una taula on es presentaven 2 *runs* de TMT amb 16 mostres del nostre interès. Vam procedir a quedar-nos amb aquestes dades des de les taules d'origen i a seleccionar aquells pèptids que no donaven lloc a ambigüitat i que eren únics en la seva quantificació:

```
# Importamos tabla Excel con datos
ProteinRaw <-readXL("Datos/Copia de 2017.06.12_Exported PSM Layer
1_Multiconsensus from 2 Reports_NET.xlsx", rownames=FALSE,
header=TRUE, na="NA", sheet="Sheet1", stringsAsFactors=TRUE)

# Procedemos a limpiar la tabla de aquellos datos que no usaremos.
ProteinRaw<- ProteinRaw[c( 1, 2, 3, 4, 7, 8, 15, 26:35)]
ProteinRaw <- ProteinRaw[ProteinRaw$PSM.Ambiguity == "Unambiguous",]
ProteinRaw <- ProteinRaw[ProteinRaw$Quan.Info == "Unique",]
```

Algorisme 9: Importació i filtratge de dades

D'aquest filtratge vam obtenir les taules RunA i RunB que podem trobar a la documentació adjunta dins de l'arxiu *ProteinPairs.xlsx* a la carpeta de anàlisi nou. Per fer la selecció dels pèptids que havien d'aparèixer en cada Run ens vam basar en la seva intensitat que s'indica a la columna A9 i B9 , escollint només els que teníem valor "High". Es va unir la seqüència dels pèptids amb les seves modificacions originades pel marcatge per tal d'obtenir una identificació única. Seguidament vam passar a nombrar les columnes tal i com es va fer a l'experiment:

```
#Creamos los Runs A y B
RunA <- subset(ProteinRaw, A9 == "High")
RunA <- RunA[-2]
RunA <- RunA[c(-1,-3,-6, -7, -10)]

RunA <- cbind(paste(RunA$Sequence,RunA$Modifications), RunA)
RunA <-RunA[c(-2,-4)]
colnames(RunA) <-
c("Identificador","Proteina","A1","N1","O1","N2","Z1","O2","A2","Z2")

RunB <- subset(ProteinRaw, B9 == "High")
RunB <- RunB[-1]
RunB <- RunB[c(-1,-3,-6, -7,-10)]

RunB <- cbind(paste(RunB$Sequence,RunB$Modifications), RunB)
RunB <-RunB[c(-2,-4)]
colnames(RunB) <-
c("Identificador","Proteina","Z3","O3","N3","O4","A3","N4","Z4","A4")
```

Algorisme 10: Creació del RunA i RunB

En aquest punt volíem unir els dos *Runs* en un *RunAB* on quedessin representats aquells pèptids amb les mateixes modificacions presents a tots dos Runs. Per fer-ho es va fer ús de les propietats de les bases de dades SQL que permeten ser aplicades sobre taules en format *data.table.* Es va assignar el valor de 0.0001 (molt baix) als NA i és va a dur a terme la mitjana dels valors amb mateix identificador i proteïna a cada *Run*. Finalment vam practicar un inner join [31] sobre les dues taules *RunA* i *RunB* i vam obtenir *RunAB* on quedaven representats únicament les files d'interès. En un últim pas vam eliminar la tripsina de porc de la nostra taula per si estava representada ja que s'utilitza de forma estàndard en la metodologia per tallar les proteïnes.

```
# Transformamos los data.frame en data.table para poder asignarle
clave.
RunA <- data.table(RunA, key = "Identificador")
RunB <- data.table(RunB, key = "Identificador")

# Gestión de los NA == 0, ==low value, === NA????
RunA[is.na(RunA)] <- 0.0001
RunB[is.na(RunB)] <- 0.0001

# Calculamos la media de las filas que tienen duplicado la id del
peptido
RunA <- RunA[ ,lapply(.SD, mean), by=c("Identificador","Proteina")]
RunB <- RunB[ ,lapply(.SD, mean), by=c("Identificador","Proteina")]

# Hacemos un Inner Join de las dos tablas (Inner Join = solo las key
que están en las dos tablas en nuestro caso "Identificador")
RunAB <- RunA[RunB, nomatch = 0]

# Eliminamos la tripsina de cerdo por si acaso está en el listado de
proteínas
RunAB <- RunAB[Proteina!="P00761"]
```

Algorisme 11: Obtenció de RunAB

Arribats a aquesta taula, el grup investigador em va fer saber que la identificació de les proteïnes els era més familiar amb el nom curt de la mateixa per comptes del codi Uniprot. Per això vaig dissenyar una funció que obtenia el nom curt a partir de la referencia i addicionalment vaig afegir una columna de funció. Per fer-ho es connecta a la web d'Uniprot de cada proteïna en la seva versió XML i obté les dades. L'entrada de la funció és un llistat de proteïnes en format *dataframe* i en columna tal i com es mostra en la següent il·lustració:

	A	B	
1	P02768		
2	P80723		
3	P62258		

Il·lustració 14: Entrada getNamesUniprot

El codi d'aquesta funció va donar lloc a una modificació i la programació d'una aplicació web que també es presenta en la documentació adjunta a aquesta memòria.

```
getNamesUniprot <- function(x){

  protNames <- data.frame( ProtID = character(), ProtName =
character(), ProtFunct = character())

  for (i in 1:dim(x)[1]){

    url <- paste ("http://www.uniprot.org/uniprot/", x[i,], ".xml",
sep="")
    try(data <- xmlParse(url))
    xml_data <- xmlToList(data)

    ProtName <- "Not available"

    try(ProtFunct<- as.character(xml_data$entry$comment$text[1]))
    try(ProtName <-
as.character(xml_data$entry$gene$name$text[xml_data$entry$gene$name$
.attrs=="primary"]))

    if (length(ProtFunct) == 0L){ProtFunct <- "not listed"}

    name <-data.frame( ProtID =
as.character(x[i,]),ProtName=ProtName, ProtFunct=ProtFunct)
    protNames <- rbind(protNames, name)

    cat("     \r",round(i/(dim(Proteinas)[1])*100, 2), "%")

  }

  return(protNames)

}
```

Algorisme 12: GetNamesUniprot

La sortida d'aquesta funció es una taula com la que es mostra en la següent il·lustració:

ProtID	ProtName	ProtFunct
P02768	ALB	Serum albumin, the main protein of plasma, has a good binding capacity for water, Ca(2+), Na(+), K(+), fatty acids, hormones, biliru pressure of blood. Major zinc transporter in plasma, typically binds about 80% of all plasma zinc.
P80723	BASP1	Associated with the membranes of growth cones that form the tips of elongating axons.
P62258	YWHAE	Adapter protein implicated in the regulation of a large spectrum of both general and specialized signaling pathways. Binds to a larg phosphothreonine motif. Binding generally results in the modulation of the activity of the binding partner (By similarity). Positive cytoplasm (PubMed:12917326).

Il·lustració 15; Sortida de GetNamesUniprot

Un cop fets aquests preparatius de les dades, vam passar a l'anàlisi de les dades totals i per poblacions. Donat que el anàlisi és el mateix per tots els grups, es va crear una funció que engloba la creació de les taules de alta correlació per poder aplicar els algorisme d'una manera més senzilla.

Però per poder fer la funció (que s'anomena *Datos),* primer es van crear totes les funcions que aquesta altra hauria de cridar de manera automatitzada.

En primer lloc es va dissenyar una funció per poder correlacionar les taules de pèptids quantificats. Aquesta funció es va simplificar gràcies a les indicacions del meu consultor per donar lloc a una funció molt més concisa. Aquesta funció aprofita la capacitat de la crida *cor* per correlacionar matrius que personalment desconeixia. El mètode utilitzat primàriament va ser Pearson per això he mantingut aquest, però donada la natura de les mostres podríem haver optat per Spearman amb total correcció. Així, després d'aplicar les millores a aquesta funció vam obtenir la següent:

```
# Función que crea una tabla de correlación a partir de una matriz de
datos.
# @ param: x - data frame
# $ return: cor.completa - data frame
#
correlationTable <- function(x){
  aa<-x[, c(-1,-2)]
  cor.completa<-cor(t(aa), method = "pearson")

  # Indentificamos las columnas y las filas añadiendo además la
proteina a la que pertenece el péptido.
  cor.completa<- as.data.frame(cor.completa)
  colnames(cor.completa)<- x$Identificador
  cor.completa <- cbind(x$Identificador, x$Proteina, cor.completa)
  colnames(cor.completa)[1:2] <- c("Identificador", "Proteina")
  cat("\n")

  #Devolvemos cor.completa como retorno de la función
  return(cor.completa)
}
```

Algorisme 13: CorrelationTable

Com a resultat d'aplicar aquesta funció sobre les dades obtenim una taula que es pot veure als adjunts a la carpeta CorrelacionsCSV, donat que las vaig haver de exportar en .csv perquè el *plugin* d'exportació a Excel no suportava una taula tan gran en fer la exportació .

Identificador	Proteina	aAcLLPk N-Tr	aAEAAAAAPA	aAFDDAIAEL	aAFTEccQAA	aAFTEccQAA	aAFTEccQAA	aALAcSk N-T	aAQEEYVk N	aASQPGELk I	aATSDLEHYD	acP5HQPDIS	aDAVTLDGGI
1 aAcLLPk N-Term(TMT10Plex); C3(Carbamidomethyl); K7(TMT10Plex)	P02768	1	0.66845546	0.57191544	0.46719866	0.22891808	0.91059402	0.71536037	0.66330775	0.66277999	-0.32578477	0.74663834	0.020129
2 aAEAAAAPAESAAPAAGEEPSkEEGEPk N-Term(TMT10Plex); K22(TMT10Plex); K28(TMT10Plex)	P80723	0.66845546	1	0.93679083	0.57353884	0.3041193	0.82821415	0.76825248	0.80306391	0.8240656	-0.27852954	0.92554985	-0.10759342
3 aAFDDAIAELDTLSEESYk N-Term(TMT10Plex); K19(TMT10Plex)	P62258	0.57191544	0.93679083	1	0.53292383	0.25765413	0.76946785	0.75490201	0.78591233	0.8854419	-0.34398166	0.85806301	-0.13171087
4 AAFTEccQAADk C6(Carbamidomethyl); C7(Carbamidomethyl); K12(TMT10Plex)	P02768	0.46719866	0.57353884	0.53292383	1	0.93355437	0.60409447	0.75871035	0.31864792	0.55894593	-0.43522166	0.61865117	-0.10078584
5 aAFTEccQAADK N-Term(TMT10Plex); C6(Carbamidomethyl); C7(Carbamidomethyl)	P02768	0.22891808	0.3041193	0.25765413	0.93355437	1	0.35258753	0.60949064	0.01850098	0.33571866	-0.40544049	0.35242536	-0.0723056
6 aAFTEccQAADk N-Term(TMT10Plex); C6(Carbamidomethyl); C7(Carbamidomethyl); K12(TM1	P02768	0.91059402	0.82821415	0.76946785	0.60409447	0.35258753	1	0.80787323	0.65195577	0.78170695	-0.4951891	0.87249669	-0.00851798
7 aALAcSk N-Term(TMT10Plex); C5(Carbamidomethyl); K7(TMT10Plex)	P15144	0.71536037	0.76825248	0.75490201	0.75871035	0.60949064	0.80787323	1	0.55437253	0.87720771	-0.55450778	0.70006583	-0.07102765
8 aAQEEYVk N-Term(TMT10Plex); K8(TMT10Plex)	P04075	0.66330775	0.80306391	0.78591233	0.31864792	0.01850098	0.65195577	0.55437253	1	0.6412446	0.05640082	0.80921446	0.05174643
9 aASQPGELk N-Term(TMT10Plex); K9(TMT10Plex)	P16870	0.66277999	0.8240656	0.8854419	0.55894593	0.33571866	0.78170695	0.87720771	0.6412446	1	-0.50193637	0.74376715	-0.27781782
10 aATSDLEHYDk N-Term(TMT10Plex); K11(TMT10Plex)	P80303	-0.32578477	-0.27852954	-0.34398166	-0.43522166	-0.40544049	-0.4951891	-0.55450778	0.05640082	-0.50193637	1	-0.24964096	0.03624281
11 acP5HQPDISSGLELPTPGVPTLDNIk N-Term(TMT10Plex); C2(Carbamidomethyl); K28(TMT10P	Q16610	0.74663834	0.92554985	0.85806301	0.61865117	0.35242536	0.87249669	0.70006583	0.80921446	0.74376715	-0.24964096	1	-0.11404781
12 aDAVTLDGGFIYEAGLAPYk N-Term(TMT10Plex); K20(TMT10Plex)	P02788	0.020129	-0.10759342	-0.13171087	-0.10078584	-0.0723056	-0.00851798	-0.07102765	0.05174643	-0.27781782	0.03624281	-0.11404781	1
13 aDAVTLDGGFIYEAGLAPYkLRPVAAEVYGTER K20(TMT10Plex)	P02788	0.24475385	0.42616525	0.41956117	0.1906438	0.12315165	0.32690475	0.39466605	0.4925658	0.26227247	-0.03754648	0.38705281	0.71880874
14 aDAVTLDGGFIYEAGLAPYkLRPVAAEVYGTER N-Term(TMT10Plex); K20(TMT10Plex)	P02788	-0.57196286	-0.39042105	-0.35282829	-0.53016293	-0.44552749	-0.63031458	-0.62417869	-0.0362676	-0.61666616	0.64070529	-0.34252404	0.39933457
15 aDDkETcFAEEGk N-Term(TMT10Plex); K4(TMT10Plex); C7(Carbamidomethyl); K13(TMT10Ple	P02768	0.79165577	0.86790892	0.85437918	0.60789162	0.33922074	0.96311031	0.75078972	0.6445115	0.79145556	-0.49037573	0.90703013	-0.0577452
16 aDDkETcFAEEGkk N-Term(TMT10Plex); K4(TMT10Plex); C7(Carbamidomethyl); K13(TMT10Pl	P02768	-0.4283277	-0.36454243	-0.39735096	-0.47026996	-0.4672628	-0.56009468	-0.72273256	-0.01855535	-0.6637946	0.85179477	-0.31951022	0.0696087
17 aDEGWYWcGVk N-Term(TMT10Plex); C8(Carbamidomethyl); K11(TMT10Plex)	P01833	0.35519434	0.30847363	0.29781618	0.38438064	0.38447789	0.3215284	0.412933	0.28198838	0.38504555	-0.00690499	0.3222875	0.16128274
18 aDIEEIk N-Term(TMT10Plex); K7(TMT10Plex)	P80303	0.77869053	0.84773849	0.78913715	0.54582442	0.34059975	0.85089895	0.88106458	0.63447657	0.92584388	-0.45695476	0.79199631	-0.24484799
19 aDSQAQLLLSTVVGVFTAPGLHLk N-Term(TMT10Plex); K24(TMT10Plex)	P01019	0.26339313	0.60173673	0.62297034	0.68342866	0.58184399	0.46596639	0.5211775	0.37952967	0.43309098	-0.24503922	0.61495912	-0.05113732
20 aEEELDAEVLEVFHPTHEWQALQPGQAVPAGSHVR N-Term(TMT10Plex)	Q9H173	0.75455667	0.82384717	0.79618573	0.45884024	0.21640992	0.79334862	0.80814928	0.79151356	0.80558554	-0.37199103	0.75253633	0.19604991
21 aEEYEFLTPVEEAPk N-Term(TMT10Plex); K15(TMT10Plex)	P52565	0.75445218	0.91837923	0.93875187	0.60656067	0.33921487	0.92209755	0.84433503	0.71209913	0.91077553	-0.48191982	0.87880264	-0.0414169
22 AEFAEVSk K8(TMT10Plex)	P02768	0.43910289	0.62640355	0.57866281	0.97822675	0.9074521	0.62385423	0.73127712	0.34617853	0.55110142	-0.38048045	0.67736242	-0.05948439
23 aEFAEVSk N-Term(TMT10Plex); K8(TMT10Plex)	P02768	0.82203856	0.8778671	0.85506416	0.68403747	0.43114841	0.97659243	0.83542997	0.65361785	0.83298262	-0.51480227	0.90433585	-0.06790551
24 aEGSDVANAVLDGADcIMLSGETAk N-Term(TMT10Plex); C16(Carbamidomethyl); K25(TMT10Pl	P14618	0.09203349	0.61498562	0.67146379	0.36497241	0.18308967	0.25886012	0.24583184	0.65881753	0.31698338	0.21101948	0.57244482	0.09113736
25 aEMIIEQNTDGVNFYNILTk N-Term(TMT10Plex); K20(TMT10Plex)	Q9H840	0.23637558	0.70966696	0.75365815	0.22371748	-0.01628081	0.39940227	0.32208945	0.68612517	0.47208087	0.20186695	0.61867907	-0.14406761
26 aFQVVSDVTPLR N-Term(TMT10Plex)	P08253	-0.22891088	0.29686834	0.2424593	0.18614501	0.10431699	-0.11220579	-0.10819445	0.36676712	-0.09185482	0.28799972	0.2461403	-0.02350807
27 aGAVNPTVk N-Term(TMT10Plex); K9(TMT10Plex)	P27487	0.75672959	0.83582063	0.89610348	0.58123008	0.31798644	0.87991385	0.85191323	0.67960482	0.94272678	-0.50732081	0.80139654	-0.21500071
28 aGEVQEPELR N-Term(TMT10Plex)	P25311	0.66512652	0.93394352	0.96144282	0.55143289	0.26901552	0.86576661	0.73908232	0.72377612	0.86698507	-0.37498379	0.90801798	-0.15276239
29 aGGVLAYELLPALDEVLASDSR N-Term(TMT10Plex)	P54802	0.08863987	0.56758251	0.5917848	0.06855032	-0.16056783	0.220286	0.08530563	0.68102054	0.18881847	0.31383167	0.51502203	0.21086418
30 aGkPVIcATQMLESMIk N-Term(TMT10Plex); K3(TMT10Plex); C7(Carbamidomethyl); K17(TMT:	P14618	0.36224295	0.58129418	0.53706565	0.17530298	-0.06596097	0.37656579	0.17067219	0.81976894	0.28513314	0.31313546	0.64207953	-0.00927767

Il·lustració 16: Sortida de correlationTable

Un cop obtinguda aquesta taula passem a buscar les proteïnes les quals els seus pèptids tenen una alta correlació mitjançant una altra funció que tindrà com entrada la taula de correlació i com a sortida un llistat de proteïnes aparellades amb el tant per cent de pèptids correlacionats altament i el número de pèptids per proteïna que s'han estudiat. Per aquesta funció, que té una execució que requereix molt de temps, vaig intentar aplicar

```
# Función que crea una tabla con los datos de las proteínas altamente
correlacionadas (con una correlación mayor a 0.9 entre
# más del 95% de los péptidos que aparecen en cor.completa. .
# @ param: x - matriz de datos
# $ return: highCorrelation - data frame
#
highCorrelation <- function(x,Proteinas){
   High.corr <- data.frame( Proteina1 = character(), Proteina2 =
character(), Percent =  numeric(), Peptidos= character())
   cat("     Progreso altamente correlacionadas:\n")

   for(i in 1:(dim(Proteinas)[1]-1)){
    aN <- as.character(Proteinas[i,2])
    a <- as.character(Proteinas[i,1])
    Prot1 <- subset(x, Proteina == a)
    v=i+1

    for(s in v:dim(Proteinas)[1]){
      ProtVS <- NULL
      bN <- as.character(Proteinas[s,2])
      b <- as.character(Proteinas[s,1])
      Prot2 <- subset(x, Proteina == b

      for (t in 1:dim(Prot2)[1]){

        if (is.null(ProtVS)){
          ProtVS <- cbind(Prot1[,1:2],
Prot1[as.character(Prot2$Identificador[t])])

        }else{
          ProtVS <- cbind(ProtVS,
Prot1[as.character(Prot2$Identificador[t])])
        }
      }

      sum = 0
   # Obtenemos la cantidad de peptidos con correlación mayor que 0.9
      sum <- sum(ProtVS[,c(-1,-2)]>0.9)
      lim <- dim(ProtVS)[1]*((dim(ProtVS)[2])-2)

      if ((lim*0.95) < sum){
        h <- data.frame(Proteina1 = aN, Proteina2 = bN, Percent =
sum/lim*100,
PeptidosTotales=paste(dim(ProtVS)[1],"X",((dim(ProtVS)[2])-2)))
        High.corr <- rbind (High.corr, h)
      }
      ProtVS <- NULL
      }
     cat("     \r",round(i/(dim(Proteinas)[1])*100, 2), "%")
  }
  cat("     \r 100% \n")
  return(High.corr)
}
```

Algorisme 14: highCorrelation

certs canvis, que no em van resultar possibles donat que afectaven greument a la informació de la sortida. La taula de sortida de *highCorrelation* és simètrica però per poder obtenir una taula temporal amb la informació que es mostra a la *il·lustració 15,* requeria aquesta simetria donat que tal i com es mostra en la part inferior la sortida final fruit d'aquestes taules temporals s'havia de mostrar el nom de les dues proteïnes, el tant per cent de pèptids amb una correlació elevada i les dimensions de la relació, és a dir, el número de pèptids implicats entre cada relació de proteïnes.

		Proteïna 2		
		Pèptid A	Pèptid B	Pèptid C
Proteïna 1	Pèptid A	0.95	0.7	0.99
	Pèptid B	0.9	0.5	0.56
	Pèptid C	0.5	0.8	0.77

PROTEÏNA 1 x PROTEÏNA 2	33%	3 X 3

Il·lustració 17: Taula temporal de relació entre proteïnes

Si bé, l'ús del for en aquesta funció de ben segur que fa que la execució sigui més lenta, em va permetre un control del flux més acurat que alguna variant de *apply* i a més cal tenir en compte que fem servir 3 *fors* aniuats que recorren més de 200 proteïnes fent comparacions una contra una.

En veure el que trigava en processar la funció, vaig decidir posar un comptador per informar al usuari del progrés de la mateixa.

Aquesta funció en un principi estava pensada per donar la alta correlació si les proteïnes tenien relacionat més del 50% dels pèptids amb una correlació superior a 0.9. Es van fer proves més estrictes que es poden trobar a la documentació adjunta i finalment s'ha optat per determinar que altament correlacionats són aquelles proteïnes que tenen correlacionats els pèptids amb un valor superior al 0.9 en un 95% des casos.

Com a sortida d'aquesta funció obtenim una taula amb el format que podem veure a continuació en un tast de la sortida real:

Proteina1	Proteina2	Percent	PeptidosTotales
BASP1	ARHGDIA	100	2 X 1
BASP1	AZGP1	52.94118	2 X 17
BASP1	PARK7	66.66667	2 X 3

Taula 6: Sortida de HighCorrelation

Un cop establertes les bases de l'anàlisi es va ampliar el mateix amb una funció que llista les correlacions en format de text :

```
RelationsNamed <- function(Prot, Correlacio){

  t <-  data.frame( ProtID = character(), Relations = numeric(),
ProteinasRelacionades = character())
  h  <- 2

  for(i in 1:nrow(Prot)){

    a<-subset(Correlacio, Proteina1 == as.character(Prot[i,h]))[,-1]
    b<-subset(Correlacio, Proteina2 == as.character(Prot[i,h]))[,-2]

    colnames(a)[1]<- "Proteína"
    colnames(b)[1]<- "Proteína"

    c<-as.data.frame(rbind(a,b))
    colnames(c)[1]<-"Proteína"
    c<-c[with(c, order(-c[2])), ]

    f<-''

    for(d in 1:nrow(c)){

      if(d==nrow(c)){
        f <- paste(paste(f,
as.character(c[d,1])),paste("(",round(c[d, 2],2),").",sep=''),sep='')
      }else{
        f<- paste(paste(f, as.character(c[d,1])),paste("(",round(c[d,
2],2),"),",sep=''),sep='')
      }

    }

    t <- rbind(t, data.frame(ProtID = Prot$ProtName[i], Relations=
nrow(c), ProtRelacionades = f))

  }

  t<-t[with(t, order(-t[2])), ]
  return(t)
}
```

Aquesta funció té una sortida similar a aquesta (nom, relacions, proteïnes relacionades):

LDHB	9	GSTP1(100), MCFD2(100), CPZ(100), PRCP(100), LIPG(100), TPP1(100), SERPINF2(100), CHID1(100), CRISPLD1(100).
CRISPLD1	9	SPINT3(100), IGFBP4(100), SERPING1(100), SMPD1(100), SIAE(100), LAMA5(100), HYOU1(100), REG3G(100), LDHB(100).

Taula 7: Sortida RelationsNamed

D'aquesta manera la lectura de les relacions es més senzilla per poder extreure conclusions.

Un cop definides les funcions a les que cridaria *Datos*, la funció es va crear com es mostra en *l'Algorisme 16:*

```
Datos<- function(Name, Run, Prot){

  cat("Realizando cálculos...\n")

  #Llamamos a las funciones para correlacionar y extraer relaciones
  corr <- correlationTable(Run)
  highCorr <- highCorrelation(corr,Prot)
  relation <- RelationsNamed(Prot,highCorr)

  # Guardar las correlaciones en csv para evitar el crasheo de la
librería de excel.
  cat("     Generando csv de correlación...\n")
  write.csv(corr , paste("Resultados09/CorrelacionCSV/Correlacion
",Name,".csv", sep=""))
  cat("     Generando xlsx de altamente correlacionados...\n")
  write.xlsx(highCorr, "Resultados09/AltaCorrelacion.xlsx",
sheetName=Name, append=TRUE, row.names=FALSE)
  cat("     Generando xlsx de Relaciones con nombre de
proteina...\n")
  write.xlsx(relation, "Resultados09/Relaciones.xlsx",
sheetName=Name, append=TRUE, row.names=FALSE)

  #Creamos un nuevo objeto para el retorno
  cat("     Generando lista de datos...\n")
  listar <- list(corr, highCorr,relation)
  return(listar)

}
```

Algorisme 16: Funció Datos

La sortida per una banda és fa en arxius Excel o CSV i per un altra i donat que R no permet retorn múltiple es va optar per encabir les dades en una llista on es guarden les 3 taules de sortida.

Arribats a aquest punt ja es va passar a fer l'estudi per subgrups fent-lo per:

- Una única població (N,A,Z,O)
- Per dues poblacions (NA, NO, NZ, AO, AZ, OZ)
- Per tres poblacions(NAO, NAZ, NOZ, AOZ)

Per prosseguir amb l'anàlisi es va voler veure en un diagrama de *Veen* quines relacions de proteïnes es mantenien entre els diferents grups de forma aïllada. Per fer el diagrama per veure les relacions, mitjançant la funció *ProteinSwap* es van posar totes les relacions en el mateix ordre que es trobaven en les altres seguint el ordre de si es trobaven a N té preferència la seva ordenació, seguida de A i per últim Z. Així si una relació té lloc sempre ens la indicarà de la mateixa manera: "Proteïna1-Proteïna2" per comptes d'aparèixer com "Proteïna2-Proteïna1".
Un cop fet el processat de les relacions vam procedir a crear les taules necessàries per fer el diagrama de Veen. En primer lloc fem una taula on hi siguin totes les relacions aparegudes en cada població única i fem que les relacions duplicades es presentin un sol

cop:

```
# Creamos una tabla de valores únicos
Ti <-rbind(Ni, Ai,Zi,Oi)
Ti <- unique(Ti[c(1, 2)])
```

Algorisme 17: Creació de taula de valors únics

Seguidament afegim una columna nova a les taules de relacions úniques amb un 1 i transformem totes les taules en *data.tables.* Realitzem un *outer join* [32] de les taules per obtenir una taula final on si un parell de proteïnes es troba relacionat per la població tindrà un 1 i si no tindrà un 0:

```
#Llenamos las tablas con una columna de unos
Ni$N <- rep(1,nrow(Ni))
Ai$A <- rep(1,nrow(Ai))
Zi$Z <- rep(1,nrow(Zi))
Oi$O <- rep(1,nrow(Oi))

# Convertimos als tablas en data.tables
Ti <- data.table(Ti, key = c("Proteina1","Proteina2"))
Ni <- data.table(Ni, key = c("Proteina1","Proteina2"))
Ai <- data.table(Ai, key = c("Proteina1","Proteina2"))
Zi <- data.table(Zi, key = c("Proteina1","Proteina2"))
Oi <- data.table(Oi, key = c("Proteina1","Proteina2"))

#Realizamos una Full Outer Join con las tablas
Ti <- merge(Ti,Ni, all=TRUE)
Ti <- merge(Ti,Ai, all=TRUE)
Ti <- merge(Ti,Zi, all=TRUE)
Ti <- merge(Ti,Oi, all=TRUE)

#Convertimos los NA en 0
Ti[is.na(Ti)]<- 0
```

Algorisme 18: Transformació a data.tables

Amb això haurem creat una taula *T*i que serà quelcom la *il·lustració 18* i ens permetrà dur a terme un diagrama de Veen de les dades que tenim:

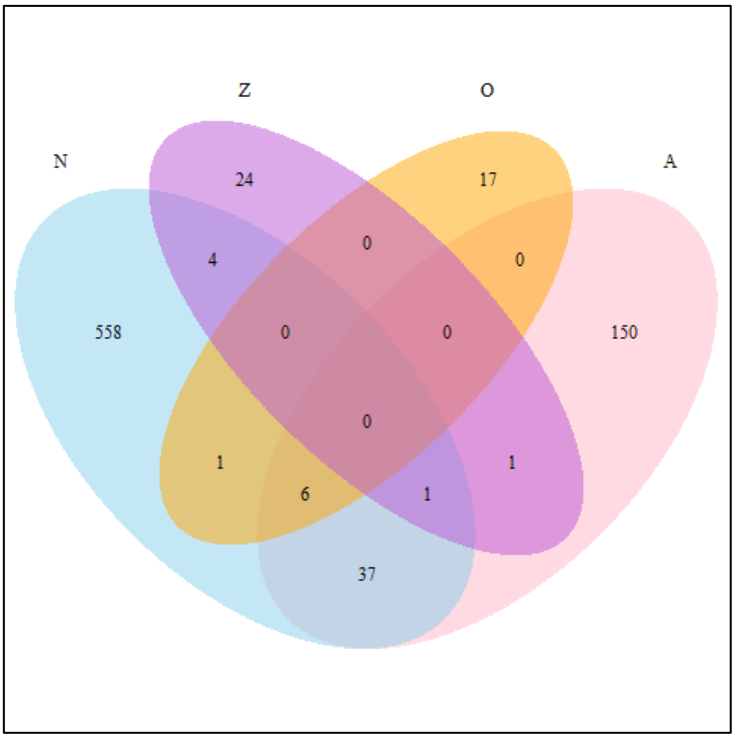
Proteïna 1	Proteïna 2	N	A	Z	O
Proteïna A	Proteïna B	1	1	1	0
Proteïna A	Proteïna C	1	1	0	1
Proteïna C	Proteïna B	0	0	1	0

Il·lustració 18: Taula Ti

Amb la taula anterior passem a crear un diagrama de Veen de 4 àrees que exportarem a un arxiu png, quedant un diagrama com el de la *il·lustració 19*:

```
#Creamos un diagrama de Veen
png(filename="Resultados09/VeenDiagramNAZO.png")
grid.newpage()
draw.quad.venn(area1 = nrow(Ti[Ti$N==1,]),
               area2 = nrow(Ti[Ti$A==1,]),
               area3 = nrow(Ti[Ti$Z==1,]),
               area4 = nrow(Ti[Ti$O==1,]),
               n12 = nrow(Ti[Ti$N==1 & Ti$A==1,]),
               n13 = nrow(Ti[Ti$N==1 & Ti$Z==1,]),
               n14 = nrow(Ti[Ti$N==1 & Ti$O==1,]),
               n23 = nrow(Ti[Ti$A==1 & Ti$Z==1,]),
               n24 = nrow(Ti[Ti$A==1 & Ti$O==1,]),
               n34 = nrow(Ti[Ti$Z==1 & Ti$O==1,]),
               n123 = nrow(Ti[Ti$N==1 & Ti$A==1 & Ti$Z==1,]),
               n124 = nrow(Ti[Ti$N==1 & Ti$A==1 & Ti$O==1,]),
               n134 = nrow(Ti[Ti$N==1 & Ti$Z==1 & Ti$O==1,]),
               n234 = nrow(Ti[Ti$A==1 & Ti$Z==1 & Ti$O==1,]),
               n1234 = nrow(Ti[Ti$N==1 & Ti$A==1 & Ti$Z==1 &
Ti$O==1,]),
               category = c("N", "A","Z", "O"),
               lty = "blank",
               fill = c("skyblue", "pink1", "mediumorchid",
"orange"))
dev.off()
```

Algorisme 19: Crida per dibujar el diagrama de Veen



Il·lustració 19: Diagrama de Veen

A partir de les dades obtingudes per fer aquest diagrama es va obtenir també una taula on es llistaven les proteïnes que apareixien en relació en cada grup amb el número de aparicions que presentaven. En el algorisme següent podem veure la creació de la taula pels pacients normozoospèrmics. La funció a la que es crida, *binder*, duu a terme una unió de les columnes de *Proteïna 1* i *Proteïna 2* en una sola columna. Seguidament s'obté una relació de les proteïnes segons les seves aparicions indicant la freqüència en que han format part d'una relació.

```
# Obtenemos las proteina que aparecen altamente relacionadas solo
para cada grupo
nN <- Ti[Ti$N==1 & Ti$A==0 & Ti$Z==0 & Ti$O==0,]

inN <- binder(nN)
onlyInN <- as.data.frame(table(unlist(inN)))
onlyInN <- onlyInN[onlyInN$Freq>0,]
onlyInN <- onlyInN[with(onlyInN, order(-onlyInN[2])), ]
```

Algorisme 20: Obtenció del llistat de proteïnes amb relacions úniques per grup

Amb tot l'anterior es van obtenir unes taules de similars a les que es mostren a continuació:

Var1	Freq
CDH1	32
LDHC	30
PGK1	30
NME2	28
AHCY	27
RAB27B	27
HSPA1A	26
PRSS8	26
BASP1	25
PRCP	25

Taula 8: Resultat de relacions per grup

C.3. AMPLIACIÓ: ESTUDI MITJANA-SD

Després d'obtenir les dades anterior per a cada grup amb les proteïnes que diferencialment es trobaven expressades i fent un estudi preliminar de resultats, la Dra. Jodar va determinar que hi havia una via per la qual es podria continuar l'estudi fent-lo més "personalitzat" per cada pacient.

Amb aquesta idea en ment i assumint la escassa quantitat de mostres que teníem per poder estandarditzar, 4 pacients en cada grup poblacional, es va dissenyar un estudi. L'objectiu es deixar en evidencia quines de les proteïnes que van establir relacions només presents en pacients normozoospèrmics, presenten disminucions o augments de la senyal en la resta de pacients de manera individualitzada. A més, aquest estudi deixaria en evidencia quines de les mostres en concret presenten alteracions respecte al pacients normozoospèrmics.

Per dur a terme els rangs de normalitat, es va optar per fer la mitjana de la divisió de les senyals dels pèptids de les proteïnes enfrontades de les mostres normozoospèrmiques. Després de fer un estudi parcial de les mostres, es va determinar que el rang de normalitat es trobaria entre mitjana menys 3 desviacions estàndards i mitjana més 3 desviacions estàndards. Tots aquells ratis que quedessin fora d'aquest rang es marcarien com alterats. I finalment determinaríem el número de senyals dintre del rang per mostra.

Un cop determinades les condicions de l'estudi, es va procedir a fer l'algorisme que tracta les dades de senyal com al primer experiment [Algorisme 9, Algorisme 10, Algorisme 11 i Algorisme 12].

A partir d'aquest punt el codi s'ha generat de nou i sempre amb la intenció d'adequar-me al estil de programació d'R he intentat substituir els bucles *for* per la funció *apply* (i derivats) el màxim possible.

En aquest punt es fa la importació de la relació de proteïnes enfrontades. Es va decidir fer una entrada des d'una plana Excel per poder facilitar el canvi de les dades de estudi.
Seguidament es va fer servir de nou la funció *binder* per unir les dues columnes de proteïnes i es va obtenir les funcions de les proteïnes implicades a *ProteinasVSUnique*

```
ProteinasVS <- read.xlsx("Datos/Proteinas enfrentadas.xlsx",
sheetName="Hoja1", header=0)

ProteinasVSUnique <- unique(binder(ProteinasVS))
ProteinasVSUnique<-ProteinasN[( ProteinasN$ProtName %in%
ProteinasVSUnique$Proteína) ,]
```

Algorisme 21: Obtenció de les funcions de les proteïnes implicades

Tot seguit es va procedir a canviar l'entrada de noms curts a número Uniprot per identificar les proteïnes. Seguidament es va assignar a *onlyResults* un TRUE, d'aquesta manera la funció dóna lloc a un sortida resumida de les dades obtingudes que facilita la

seva manipulació posterior per interpretar els resultats.

Seguidament es va crear un llibre Excel per encabir els resultats així com la taula que els ha d'encabir durant l'execució.

```
ProteinasVS <- apply(ProteinasVS, 1:2, function(x)
ProteinasVSUnique[as.character(ProteinasVSUnique$ProtName)==as.charac
ter(x), 1])

onlyResults = TRUE

wb = createWorkbook()
Totales <-t(as.data.frame(rep("",times=14)))
colnames(Totales)<- c(" ","
","A1","A2","A3","A4","O1","O2","O3","O4","Z1","Z2","Z3","Z4")
```

Algorisme 22: Preparació per l'estudi

Al codi que es mostra a continuació (Algorisme 23), s'entra al primer *for* on es recorre *ProteinaVS* fila a fila per obtenir tots els pèptids de cadascuna de las dos proteïnes enfrontades. Tot seguit s'obtenen per una banda els valors dels normozoospèrmics (R1 i R2) i per un altra el valors de la resta de pacients (O1 i O2). Es crea *tRange*, el rang de normalitat, fent la divisió (R1/R2) de cada senyal per cada pèptid de pacient normozoospèrmic i seguidament creant els intervals de mitja més 3 SD i mitja menys 3 SD (Amb això obtenim per cada parella de proteïnes una columna amb dos valors per cada parella de pèptids. P.ex: si tenim 2 pèptids per una banda i 2 per l'altra tindrem una matriu de 2x4).

	V1	V2	V3	V4
1	5.422109	5.404282	-1.278167	-1.778168
2	14.173232	12.414040	27.785918	25.974689

Il·lustració 20: tRange (Valors x+3sd, x-3sd)

Seguidament obtenim *oRange* els valors de la divisió de senyal per la resta de grups.

V1	V2	V3	V4
4.533634	0.3410792	3.505279	0.2637129
3.042676	0.2882256	2.066234	0.1957295
2.915543	0.5449096	2.083738	0.3894467
3.514097	0.3525397	2.618720	0.2627141
4.123920	0.2672126	2.226805	0.1442875
3.435593	0.3790633	2.143225	0.2364709
2.969889	0.1956596	2.830776	0.1864946
3.359235	0.4153972	3.030918	0.3747980
1.859388	0.1625305	1.882955	0.1645905
5.297659	0.4377558	4.242165	0.3505384
4.483559	0.3328483	3.380696	0.2509745
2.916652	0.3744580	1.761891	0.2262026

Il·lustració 21: oRange (Valors de la divisió pèptid a pèptid per pacient)

```
 for (i in 1:nrow(ProteinasVS)){

  P1<-
RunABVS[as.character(RunABVS$Proteina)==as.character(ProteinasVS[i,1]
),]
  P2<-
RunABVS[as.character(RunABVS$Proteina)==as.character(ProteinasVS[i,2]
),]

  R1 <- P1[,c(4,6,13,16)]
  R2 <- P2[,c(4,6,13,16)]
  O1 <- P1[,-c(1,2,4,6,13,16)]
  O2 <- P2[,-c(1,2,4,6,13,16)]

  tRange<-apply(matrix(apply(R1, 1, function(x) apply(R2,1,
function(y) x/y)),nrow=dim(R2)[2]), 2, function(x) c(mean(x)-3*sd(x),
mean(x)+3*sd(x)))
  oRange<-matrix((apply(O1, 1, function(x) apply(O2,1, function(y)
x/y))),nrow=dim(O2)[2])
```

Algorisme 23: Obtenció dels valors de referència i els valors a estudi

Seguidament entrem en un nou for per poder obtenir una matriu binaria on un 0 indiqui que el pèptid està fora de rang i 1 que és troba al nostre rang de normalitat.

```
  for (j in 1:ncol(oRange)){

    inf=tRange[1,j]
    sup=tRange[2,j]

    oRange[,j][inf>oRange[,j] | sup<oRange[,j]]<-0
    oRange[,j][oRange[,j]!=0]<-1
  }
```

Algorisme 24: Transformació en matriu binaria de oRange

En aquest punt ens queda informar els resultats per aquesta parella de proteïnes, i aquí es fa evident la meva predisposició a usar la funció *apply*, fent un *apply* una mica estrany però efectiu. Per cada filera el valor de la suma està per sobre del 75% del número de columnes tindrà assignat una alta correspondència amb el valor si no una baixa amb el valor. Seguidament posem les etiquetes a les columnes, i files de la matriu sortint i endrecem els resultats per grups de població.

```
Results<-apply(oRange,1,function(x)
if(sum(x)>=(ncol(oRange)*0.75)){c("ALTA",
round(100*(sum(x)/ncol(oRange)),digits =
2))}else{c("BAJA",round(100*(sum(x)/ncol(oRange)),digits = 2))})
  colnames(Results)<-
c("A1","O1","Z1","O2","A2","Z2","Z3","O3","O4","A3","Z4","A4")
  rownames(Results)<-c("Relación", "Valor")

  Results <- t(Results[,c(1,5,10,12,2,4,8,9,3,6,7,11)])
```

Algorisme 25: Aplicacio de les regles i sortida de resultats per parella de proteïnes

En la següent il·lustració podem veure la sortida de resultat per parella de proteïnes:

	Relación	Valor
A1	BAJA	50
A2	BAJA	50
A3	ALTA	100
A4	ALTA	100
O1	ALTA	100
O2	ALTA	100
O3	BAJA	50
O4	ALTA	100
Z1	ALTA	100
Z2	ALTA	75
Z3	ALTA	75
Z4	ALTA	75

Il·lustració 22: Sortida per parella de proteïnes

Un cop tenim el resultat l'afegim a la resta per preparar la sortida final, si hem marcat onlyResults com TRUE farà una sortida resumida on posarà el nom de les dues proteïnes enfrontades i els valors de la taula Results. Per últim esborrarà el noms de columna dels primers dos valors de Lets per evitar conflictes a l'hora de fer la unió de tots els resultats.

```
  if(onlyResults==TRUE){

    Lets<-
cbind.data.frame(rbind.data.frame(c(as.character(ProteinasVSUnique$Pr
otName[ProteinasVSUnique$ProtID==ProteinasVS[i,1]]),
as.character(ProteinasVSUnique$ProtName[ProteinasVSUnique$ProtID==Pro
teinasVS[i,2]]))), t(Results[,2]))
    colnames(Lets)[c(1,2)]<- c(" "," ")
    Totales<-rbind(Totales , Lets )
```

Algorisme 26: Preparació de la sortida de totes les proteïnes en mode Resumit

Per acabar sortirà del bucle *if* i entrarà en un altre que crearà la sortida en un document Excel amb tant sols els resultats.

```
if(onlyResults==TRUE){
  sheet = createSheet(wb, "Resultados")
  addDataFrame(Totales, sheet=sheet, startRow=1, row.names=FALSE)
  sheet = createSheet(wb, "Proteinas")
  addDataFrame(ProteinasN, sheet=sheet, startRow=1, row.names=FALSE)
}

saveWorkbook(wb, "Resultados/Resultados.xlsx")
```

Algorisme 27: Sortida Final

Per fer-ho més entenedor vam aplicar un format condicional mitjançant Excel que dona color a les cel·les en funció del valor creant un *pseudo-heatmap* que permet donar una ullada ràpida a les relacions que desapareixen completament, per exemple:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			A1	A2	A3	A4	O1	O2	O3	O4	Z1	Z2	Z3	Z4
2	BASP1	YWHAE	100	100	75	75	100	100	100	100	75	100	100	100
3	BASP1	ALDOA	83.33	100	83.33	100	100	100	100	83.33	83.33	83.33	100	100
4	BASP1	ECM1	95.45	95.45	95.45	72.73	95.45	63.64	86.36	86.36	27.27	90.91	59.09	54.55
5	BASP1	PKM	100	100	66.67	100	100	100	100	100	100	100	100	100
6	BASP1	SCPEP1	100	100	83.33	100	100	100	100	100	100	100	100	100
7	BASP1	SOD3	100	100	100	50	100	100	100	100	100	100	100	75
8	BASP1	AOC1	100	100	100	100	100	75	100	100	100	100	75	100
9	BASP1	GAA	100	100	100	100	100	75	100	100	100	100	75	100
10	BASP1	TF	100	100	100	75	96.43	85.71	100	100	78.57	60.71	57.14	100
11	BASP1	CD9	75	100	75	75	100	100	100	100	75	100	100	100
12	BASP1	LGALS3BP	85	100	80	85	95	75	90	80	65	95	75	85
13	BASP1	PARK7	66.67	83.33	16.67	100	83.33	66.67	83.33	66.67	66.67	83.33	100	16.67
14	BASP1	PTGDS	100	100	100	100	100	100	0	100	50	25	0	0
15	BASP1	CST3	80	100	100	100	80	50	100	80	60	90	80	100
16	BASP1	APOH	50	100	100	100	100	100	100	100	100	100	100	100
17	BASP1	LSAMP	100	100	50	87.5	100	87.5	87.5	87.5	62.5	100	87.5	50
18	BASP1	S100A11	100	100	50	100	100	100	100	100	100	100	100	100
19	BASP1	SPOCK1	66.67	100	50	83.33	83.33	83.33	83.33	100	83.33	100	50	66.67

Il·lustració 23: Resultat final de Estudi mitja-sd

La opció de onlyResults en FALSE dóna accés a les taules que s'han utilitzat en el càlcul per poder comprovar que tot es correcte. Aquesta opció només la vam usar per fer un estudi preliminar amb 11 parelles de proteïnes donat que allarga molt la execució sobre tot en la creació de una plana Excel per proteïna. En aquesta sortida es pot veure en primer lloc oRange [Il·lustració 21], seguidament tRange [Il·lustració 20] i per últim una combinació de oRange (transposada i en format binari) i Results [Il·lustració 22].

	X1.1	X1.2	X1.3	X2.1	X2.2	X2.3	X3.1	X3.2	X3.3	X4.1	X4.2	X4.3
A1	0.755833	1.522733	1.196673	1.253275	2.221669	4.47587	3.517459	3.683834	0.287254	0.578715	0.454796	0.476307
A2	0.712562	1.404273	1.150041	0.943982	2.294978	4.522802	3.703985	3.040323	0.331163	0.652635	0.534481	0.438715
A3	0.695714	1.352564	1.155633	1.002468	2.156472	4.19248	3.58206	3.107301	0.167886	0.326394	0.278871	0.24191
A4	1.595049	2.794717	3.158529	2.268782	3.306525	5.79343	6.54761	4.70317	0.335611	0.588031	0.664579	0.47737
O1	0.810535	1.627369	1.269696	1.346365	2.393844	4.80629	3.749934	3.976368	0.286451	0.575128	0.448723	0.475818
O2	1.100736	2.706466	2.383667	2.154669	3.200126	7.868404	6.929943	6.264186	0.291501	0.716737	0.631252	0.570608
O3	0.770454	1.576824	1.47824	0.944429	2.582371	5.285124	4.954698	3.165493	0.366731	0.750558	0.703633	0.449542
O4	0.626879	1.117981	1.465544	1.049462	1.516651	2.704806	3.545689	2.539035	0.168124	0.299833	0.393046	0.281457
Z1	1.246422	3.577301	1.914912	2.47302	2.742375	7.870768	4.213185	5.441132	0.350984	1.007343	0.539226	0.696385
Z2	1.00713	1.674649	2.368377	1.945047	2.614308	4.347053	6.147833	5.048953	0.280721	0.466781	0.660146	0.54215
Z3	0.756362	1.280482	1.335684	1.111843	1.203452	2.037382	2.125215	1.76906	0.171044	0.289568	0.302052	0.251432
Z4	0.661014	1.009828	1.177934	0.949354	2.111471	3.225683	3.762665	3.032511	0.175129	0.267543	0.312081	0.251521
	X1.1	X1.2	X1.3	X2.1	X2.2	X2.3	X3.1	X3.2	X3.3	X4.1	X4.2	X4.3
inf	-0.70664	-0.64085	-0.76613	-0.25262	-0.03709	2.100016	1.346587	1.872228	0.155307	0.452113	0.411894	0.407738
sup	2.946111	5.179098	4.842688	3.434487	5.166798	8.544007	8.152886	5.643871	0.503608	0.94614	0.827975	0.581063
	A1	A2	A3	A4	O1	O2	O3	O4	Z1	Z2	Z3	Z4
1.1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1	1	1	1	1	1
1.3	1	1	1	1	1	1	1	1	1	1	1	1
2.1	1	1	1	1	1	1	1	1	1	1	1	1
2.2	1	1	1	1	1	1	1	1	1	1	1	1
2.3	1	1	1	1	1	1	1	1	1	1	0	1
3.1	1	1	1	1	1	1	1	1	1	1	1	1
3.2	1	1	1	1	1	0	1	1	1	1	0	1
3.3	1	1	1	1	1	1	1	1	1	1	1	1
4.1	1	1	0	1	1	1	1	0	0	1	0	0
4.2	1	1	0	1	1	1	1	0	1	1	0	0
4.3	1	1	0	1	1	1	1	0	0	1	0	0
Relación	ALTA	ALTA	ALTA	ALTA	ALTA	ALTA	BAJA	ALTA	ALTA	ALTA	ALTA	ALTA
Valor	100	100	83.33	91.67	100	100	58.33	100	75	75	75	100

Il·lustració 24: Sortida Final amb onlyResults=FALSE

El codi que genera aquesta segona opció es troba a l'altra banda de *l'if* que presenta l'Algorisme 26 i es pot veure en el següent tros de codi, simplement consisteix en una sèrie de manipulacions de les taules per adequar la sortida de tal manera que resulti fàcilment interpretable pel lector. De fet el seu objectiu va ser adequar i establir les normes per l'estudi final veient una sèrie de relacions de manera ampliada.

```
Results<-t(apply(oRange,1,function(x)
if(sum(x)>=(ncol(oRange)*0.75)){c("ALTA",
round(100*(sum(x)/ncol(oRange)),digits =
2))}else{c("BAJA",round(100*(sum(x)/ncol(oRange)),digits = 2))}))
    rownames(Results)<-
c("A1","O1","Z1","O2","A2","Z2","Z3","O3","O4","A3","Z4","A4")
    colnames(Results)<-c("Relación", "Valor")

    Results<-t(Results)

    rownames(oRange)<-
c("A1","O1","Z1","O2","A2","Z2","Z3","O3","O4","A3","Z4","A4")
    colnames(oRange)<-apply(expand.grid(1:nrow(R1), 1:nrow(R2)), 1,
function(x) paste(x[2], x[1], sep="."))
    oRange <- oRange[c(1,5,10,12,2,4,8,9,3,6,7,11),]
    rownames(oRangeS)<-
c("A1","O1","Z1","O2","A2","Z2","Z3","O3","O4","A3","Z4","A4")
    colnames(oRangeS)<-apply(expand.grid(1:nrow(R1), 1:nrow(R2)), 1,
function(x) paste(x[2], x[1], sep="."))
    oRangeS<- oRangeS[c(1,5,10,12,2,4,8,9,3,6,7,11),]

    rownames(tRange)<- c("inf","sup")
    colnames(tRange)<-apply(expand.grid(1:nrow(R1), 1:nrow(R2)), 1,
function(x) paste(x[2], x[1], sep="."))

    oRange <- rbind(t(oRange), Results)

    sheet = createSheet(wb,
paste(as.character(ProteinasVSUnique$ProtName[ProteinasVSUnique$ProtI
D==ProteinasVS[i,1]]),
as.character(ProteinasVSUnique$ProtName[ProteinasVSUnique$ProtID==Pro
teinasVS[i,2]]),sep="-" ))
    addDataFrame(oRangeS, sheet=sheet, startRow=1, row.names=TRUE)
    addDataFrame(tRange, sheet=sheet, startRow=15, row.names=TRUE)
    addDataFrame(oRange, sheet=sheet, startRow=19, row.names=TRUE)
```

Algorisme 28: Creació de les taules per la Sortida Ampliada

A la carpeta de resultats de "Estudio media-sd" podem veure una sortida per a cada tipus de execució. "Resultados11.xlsx" ens deixa veure una sortida amb totes les dades, per contra "Resultados939.xlsx" ens ofereix una versió resumida i amb format condicional dels resultats a més la segona plana d'aquest Excel ens permet veure la sortida de *getnamesUniprot* [Il·lustració 15] per les proteïnes implicades.

D. RESULTATS OBTINGUTS

En els dos primers annexos (A i B) podem trobar els resultat de l'estudi convencional amb la mostra sense normalitzar i normalitzada mitjançant la aplicació de el mètode de Johnson. Van ser uns informes fets en Rmarkdown que mostren els resultats en aplicar la metodologia explicada en el apartat "Anàlisi convencional" sobre les dades amb les modificacions practicades pels investigadors sense cap canvi i amb una normalització esmentada. Aquests informes volien confirmar, per una banda, allò que els investigadors del grup ja havien calculat o obtingut amb programes estadístics molt menys flexibles que R que no els havien permès una manipulació de les dades tan eficient i per una altra, veure si les dades normalitzades podien donar més informació.

El següent annex (C) és el llistat de les relacions entre proteïnes que es van trobar altament correlacionades pel grup poblacional de pacients normozoospèrmics. Si bé hi ha molts més resultats en aquest estudi que es poden veure en el material suplementari (com per exemple aquest mateix estudi per totes les poblacions), aquest resultat ha estat el més transcendent donat que ha portat a establir les bases per l'ampliació de l'estudi.

Per últim al annex D, es pot trobar el resultat de l'estudi mitja-sd. A partir d'aquest estudi el grup investigador ha pogut extreure la majoria de les conclusions que es presenten en següent apartat.

E. DISCUSSIÓ DELS RESULTATS

L'estudi de les correlacions de proteïnes amb els paràmetres establerts (95% dels pèptids tinguessin una correlació de *Pearson* superior al 0.9) va derivar en la identificació de 933 correlacions de proteïnes en pacients normozoospèrmics, 94 en pacients astenozoospèrmics, 0 en pacients oligozoospèrmics i 5 en pacients azoospèrmics.

Aquests resultats ens indiquen que mentre la població de pacients normozoospèmics és bastant homogènia pel que fa a nivells de proteïna, la resta de grups de pacients infèrtils presenta un alta heterogeneïtat que dificulta la identificació de parelles de proteïnes altament correlacionades en tots els pacients estudiats. No obstant això, es va procedir a analitzar a nivell biològic les diferencies entre les correlacions trobades en pacients normozoospermics i astenozoospermics.

Un total de 74 proteïnes estan implicades en les 933 correlacions determinades per pacients normozoospèrmics. Aquestes proteïnes estan relacionades en vies metabòliques que ja a priori es coneix que són altament importants en el fluid seminal com per exemple la glicòlisis, gluconeogènesis, proteòlisis de la matriu extracel·lular, protecció davants els radicals d'oxigen i el sistema immune. No obstant, únicament 37 proteïnes estan involucrades en les 94 correlacions determinades en els pacients astenozoospermics on les úniques vies metabòliques sobrerrepresentades son la glicòlisis i la gluconeogènesis. [Il·lustració 25]

Com podem observar a la Il·lustració 25, hi ha proteïnes que tenen un gran número de correlacions en els pacients normozoospèrmics, les quals gairebé desapareixen en els pacients astenozoospèrmics, com per exemple la proteïna BASP1. La funció de BASP1 no es coneix ben bé però en altres estudis l'han identificat en les vesícules extracel·lulars que estan presents en el fluid seminal. BASP1 sembla estar altament correlacionada amb 54 proteïnes diferents en pacients normozoospèrmics, no obstant únicament una correlació es manté en pacients astenozoospèrmics. Aquests resultats suggereixen que la proteïna BASP1 pot ser de gran importància en el funcionament normal del fluid seminal.

També com a resultat interessant dins del grup investigador destacar dues proteïnes la MSLN i la CPM que únicament mostren correlacions significatives en pacients astenozoospèrmics però no en pacients normozoospèrmics. Quan s'han analitzat quines vies metabòliques estaven enriquides per les proteïnes correlacionades amb les proteïnes MSLN i CPM el grup ha observat que rutes implicades en l'apoptosi estan altament enriquides. Aquests resultats ens fan pensar que l'apoptosi podria estar sobrerrepresentada en tots els nostres pacients astenozoospèrmics

Proteïnes involucrades en les 933 correlacions determinades en pacients normozoospèrmics	
Proteïnes	*Número de correlacions en les que estan implicades*
BASP1	54
HPX	54
SPOCK1	50
YWHAZ	50
ALDOA	49
PARK7	48
PRDX6	47
GAA	46
CD9	46
ANXA1	45
ANXA3	43
LSAMP	42
HSPA8	41
SOD1	41
IGHG1	41
YWHAE	39
PKM	39
PTGDS	39
EZR	39
GAPDH	39
GPI	39
LCP1	39
ANXA5	38
HSP90AA1	38
ECM1	36
APOH	36
CNDP2	36
DBI	36
C3	36
PSMA2	35
GDI2	34
APLP2	33
ENO1	33
EDDM3B	33
AOC1	32
CTSL	31
IGKC	29
IGHG2	28
SOD3	26
RNASE4	26
TMPRSS2	24
FSTL1	23
SCPEP1	22
S100A11	21
ELSPBP1	19
LDHA	17
CST3	16
TF	14
SORD	14
TIMP1	14
CRTAC1	14
FUCA1	13
ORM2	10
LGALS3BP	9
TWSG1	9
LAMB2	6
PLA1A	6
HEXA	6
QSOX1	5
PEBP4	5
TGM4	4
MME	4
LRG1	4
OS9	4
PATE1	3
SDF4	3
MMP2	2
VTN	2
TPP1	2
CPE	1
CPM	1
ASAH1	1
NUCB1	1
MATN2	1

Proteïnes involucrades en les 94 correlacions determinades en pacients astenozoospèrmics	
Proteïnes	*Número de correlacions en les que estan implicades*
YWHAZ	20
YWHAE	17
MSLN	16
GDI2	13
CPM	12
SPOCK1	12
CD9	11
GPI	11
OS9	8
PARK7	7
SOD1	7
ALDOA	6
ANXA5	5
LDHA	4
HPX	4
VTN	3
ANXA3	3
LAMB2	2
APLP2	2
HSPA8	2
FSTL1	2
FUCA1	2
SERPINF2	2
PAEP	2
GAPDH	2
ANXA1	2
BASP1	1
LSAMP	1
S100A11	1
CD59	1
HEXA	1
APOH	1
IGHG1	1
PEBP4	1
DBI	1
ADAMTS1	1
TPP1	1

Il·lustració 25: Proteïnes implicades en les correlacions (Normozoospèrmics vs.

Astenozoospèrmics)

Com ja hem mencionat anteriorment, el estudi de correlacions entre proteïnes ens porta a pensar que els pacients amb paràmetres seminals alterats tenen una alta heterogeneïtat a nivell proteic. Això ens va portar a plantejar l'estudi del que succeeix per les 933 correlacions determinades en pacients normozoospèrmics en els diferents pacients a nivell individual.

Com s'ha explicat a l'apartat d'estudi mitja-sd, primer es van establir uns valors de referència per cada correlació mitjançant el càlcul de la mitjana +/- 3 desviacions estàndard per totes les correlacions de pèptids implicades en aquesta correlació de proteïnes. Després mostra a mostra es van mirar quantes correlacions de pèptids es trobaven dins els valors de referència. Per tal de fer una primera valoració es van buscar aquelles correlacions en les quals menys del 25% de les correlacions de pèptids estaven dins dels valors de referència determinats.

Un cop fet ,això s'ha determinat quines de les 74 proteïnes implicades en les 933 correlacions han perdut més correlacions amb altres proteïnes.

Proteines	Normos Correlacions	A1 Correlacions	A2 Correlacions	A3 Correlacions	A4 Correlacions	O1 Correlacions	O2 Correlacions	O3 Correlacions	O4 Correlacions	Z1 Correlacions	Z2 Correlacions	Z3 Correlacions	Z4 Correlacions
BASP1	54	52	54	47	54	54	54	53	54	54	52	53	50
HPX	54	52	54	50	54	53	52	53	53	52	54	51	46
IGHG1	41	34	41	41	41	39	41	41	41	41	40	41	38
PTGDS	39	38	38	35	39	39	39	34	38	39	31	28	32
LCP1	39	37	39	36	39	39	39	39	33	39	38	34	37
HSP90AA1	38	38	38	37	38	38	38	38	38	38	37	35	33
APOH	36	35	36	36	36	36	36	36	36	36	36	36	31
C3	36	31	36	36	36	36	36	36	36	36	32	35	33
AOC1	32	25	32	32	32	32	32	31	32	32	28	31	31
ELSPBP1	19	9	19	19	19	19	19	19	19	17	14	19	16
ORM2	10	10	10	10	10	10	10	10	10	10	5	10	10

Il·lustració 26: Comparació de correlacions perdudes

El grup d'investigació està acabant d'analitzar aquestes dades, però per exemple el pacient astenozoospèrmic A1 ha perdut 5 de les 36 correlacions determinades per la proteïna C3 en pacients normozoospèrmics. Específicament aquestes correlacions perdudes estan involucrades en la senyalització de la proteïna quinasa A, la qual es coneix estar implicada en el moviment flagel·lar.

3. TRANSCRIPTÒMICA DE L'ESPERMATOZOIDE

A. INTRODUCCIÓ A LA TRANSCRIPTÒMICA D'ESPERMATOZOIDE

La transcriptòmica és l'estudi del transcriptoma, el conjunt complet de transcripcions de RNA produïdes pel genoma, en circumstàncies específiques o en un mètode específic de cèl·lula específica, com l'anàlisi de *microarrays*. La comparació de transcriptomes permet la identificació de gens expressats diferencialment entre diferents mostres biològiques.

En l'espermatozoide madur, la transcripció i la traducció es troben aturades a causa de la compactació de la cromatina, duta a terme en l'última fase de l'espermatogènesi per l'empaquetament de les protamines i per l'eliminació de la major part del citoplasma que conté la maquinària traduccional [2]. En un principi, es pensava que els espermatozoides no presentaven RNAs ja que aquest no es podien detectar amb les metodologies d'estudi de RNA existents en el anys 80. No obstant, les millores en les tecnologies van permetre augmentar la sensitivitat de les tècniques de detecció de RNAs, permetent així la detecció de RNAs en els espermatozoides. Recentment, l'aplicació de tecnologies d'alt rendiment com la RNA-seq ha revelat l'existència d'una complexa població de RNAs, tan codificants (mRNAs) com no codificants (long i small RNAs). Tanmateix s'ha descrit una població estable de RNAs en els espermatozoides que es manté constant en les diferents mostres estudiades, el que suggereix que no són únicament romanents de l'espermatogènesi sense cap funció, sinó que hi ha una retenció selectiva de RNAs durant l'espermatogènesi amb un potencial rol en el desenvolupament embrionari primerenc [2].

Per tal de posar a punt la tècnica i l'anàlisi del RNA-seq d'espermatozoide en el grup de recerca, es va dissenyar un estudi pilot per tal de validar resultats preliminars que mostraven que el processament inicial de la mostra afectava la població de RNAs presents en els espermatozoides.

Els espermatozoides representen únicament el 5% de l'ejaculat mentre que el 95% restant correspon a les secrecions de les glàndules sexuals accessòries situades al llarg del tracte reproductor masculí (epidídim, pròstata i vesícules seminals). Aquest plasma seminal és altament ric en RNAs, proteïnes, ions i sucres que poden trobar-se lliures o bé encapsulats en el gran número (10^{11}-10^{12}) d'exosomes presents en el semen.

La hipòtesis del grup de recerca suggereix que els diferents pre-processaments de la mostra afecta a la unió dels exosomes present en el fluid seminal a l'espermatozoide. Per tant, l'estudi pilot es basa en l'anàlisi de 4 mostres procedents d'individus normozoospèrmics on per cada mostra es van realitzar dos tipus diferents de pre-processament. Per tant, es va extreure RNA de qualitat de 10 milions d'espermatozoides de cada mostra processada de dues formes diferents, tenint un total de 8 mostres de RNA d'espermatozoide.

Degut a la poca quantitat i a les peculiaritats del RNA d'espermatozoide no es poden utilitzar les metodologies estàndards per la construcció de llibreries de RNA per la seva

posterior seqüenciació. No obstant es va seguir la metodologia descrita per Jodar et al., 2015[13] on per a la retrotranscripció de la fracció de long-RNAs i l'amplificació del cDNA resultant s'utilitza el *SeqPlex RNA Amplification kit* (Sigma-Aldrich). A continuació, utilitzant el *NEBNext Ultra DNA Library Prep Kit* (Illumina, New England Biolabs®) es van construir les llibreries de cDNA per a la posterior seqüenciació massiva.

Per l'anàlisi de les dades derivades del RNA-seq d'espermatozoide hem de tenir en compte dos factors:

(1) Hi ha dos passos d'amplificació, un anterior a la construcció de llibreries i un altre durant la construcció d'aquestes, cosa que fa que a la posterior seqüenciació el número de *reads* duplicats augmenti (*reads* que s'alineen a la mateixa regió del genoma)

(2) Les llibreries es van construir a partir de RNA total, sense cap tipus de selecció de RNAs poliA ni tampoc s'han eliminat els RNAs ribosomals.

B. ESTUDI D'ENES BIOINFORMÀTIQUES ÚTILS PER L'ESTUDI DE DADES PROVINENT DE L'ESTUDI DE TRANSCRIPTÒMICA

Per dur a terme l'estudi podríem haver optat per dues opcions bàsicament, tot i que les eines disponibles són les mateixes.

La primera opció és la instal·lació en una estació de treball local. Aquesta opció requereix una estació amb molta potencia de càlcul però per contra permet controlar les nostres dades i no haver de compartir cap dada en servidors que no són de la nostra propietat o d'accés exclusiu. En segon lloc un altra problema podria ser la configuració dels paquets d'eines que haurem de fer servir, tot i que aquest problema queda esmorteït en part per distribucions com bioLinux que ens permet fer ús de les eines més comuns en l'anàlisi de dades provinents de l'estudi de transcriptòmica un cop instal·lat el sistema operatiu.

Donada la manca de potencia de càlcul, i la necessitat de modificar una estació fent-la d'ús exclusiu per anàlisi, es va optar per fer servir solucions *cloud* per dur a terme l'anàlisi. En començar la cerca de solucions *cloud*, la Dra. Jodar havia tingut contacte amb la plataforma Seven Bridges. Per tant vam optar per posar-nos en contacte amb aquesta plataforma que ens va indicar que podríem encabir les nostres anàlisis dins de la seva plataforma educativa anomenada *Cancer Genomics Cloud*.

Un cop solucionat el proveïdor de serveis, vam optar per fer una instal·lació paral·lela de bioLinux en una màquina virtual que ens servis per fer conversions o visualitzacions de dades amb una necessitat de còmput baix. Així tindrem una manera de tractar les dades en local complementaria al núvol on es duran a terme els càlculs que requereixin major poder de còmput.

Per fer la selecció d'eines que farem servir en aquest anàlisi ens vam basar en l'experiència prèvia de la Dra. Jodar, que en anteriors anàlisis va fer servir:

1. FASTQC: Aquest programa permet fer una sèrie de controls de qualitat sobre les dades en brut provinents de seqüenciació d'alt rendiment. Mitjançant aquest programa podem tenir una visió global i ràpida de la qualitat de les nostres dades.

2. Bowtie2: Aquesta eina permet fer alineaments de fragments seqüenciats amb un genoma de referència. Es particularment eficient fent alineaments curts sobre genomes llargs.

3. TopHat2: Aquesta eina permet fer mapejar *reads* ràpidament provinents de RNA-seq. Aquest programa permet l'alineació de *reads* a genomes tan extensos com els de mamífer i a continuació analitza els resultats dels mapatge per identificar llocs d'unió entre exons.

4. SAMtools/BAMtools: SAM és un format genèric per emmagatzemar *reads* alineats a un genoma. BAM és un format binari anàleg a SAM però pensat per ser llegit per computadores i més encarat a obtenir una compressió raonable de les dades,

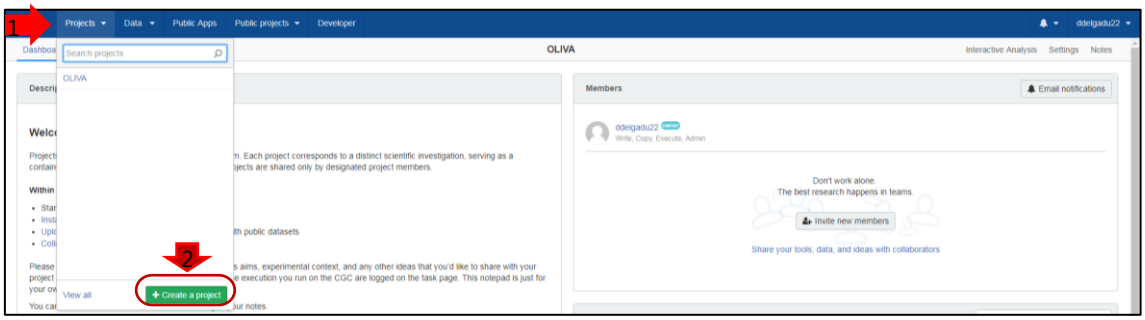
sacrificant el format que resulta il·legible pels humans. Les eines incloses en SAMtools i BAMtools són una suite per la manipulació d'aquests dos formats per emmagatzemar *reads* alineats a un genoma que permet ordenar, visualitzar, indexar i convertir d'un format al altre les dades.

5. Cufflinks: Tot i que és una eina que pot acoblar transcripcions, en el nostre cas el fem servir per les altres dues possibilitats que ofereix que són estimar les abundàncies i proves d'expressió diferencial i regulació en mostres d'RNA-Seq. Accepta RNA-Seq alineats llegeix i munta les alineacions en un conjunt parsimoniós de transcripcions. Cufflinks estima les abundàncies relatives d'aquestes transcripcions, tenint en compte els biaixos en els protocols de preparació de la biblioteca. Pel que fa a la normalització que vam fer servir va estar la predeterminada per Cufflinks *classic-fpkm* donat que segons la documentació és la que ens permetia, si ens resultés escaient, continuar l'anàlisi amb Cuffdiff.

C. MÈTODE BIONFORMÀTIC APLICAT A L'ESTUDI

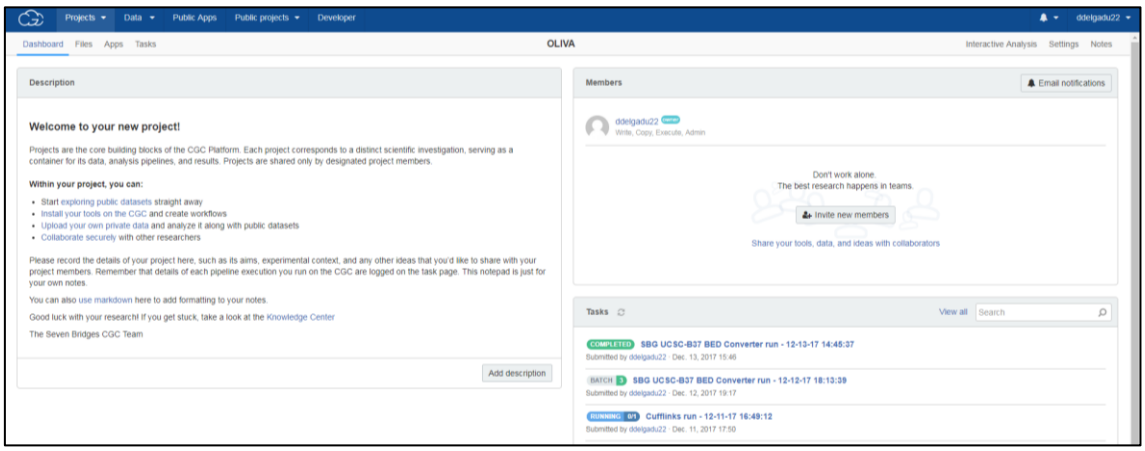
Un cop decidides les eines es va procedir a dur a terme l'estudi. Es va crear un nou compte a la plataforma CancerGenomicsCould i vam procedir a crear el projecte. Les opcions que trobem son:

1. *Projects* (on haurem de clicar) des d'on es gestionen els projectes.
2. *Data* d'on entre altres coses podem visualitzar dades públiques, importar dades de referència, etc.
3. *Public* Apps des d'on tindrem accés a un gran nombre d'aplicacions per l'anàlisi que s'han generat per diversos anàlisis i que s'han fet públiques per l'ús de tota la comunitat.
4. *Public projects* des d'on trobarem projectes de gran envergadura que es duen a terme amb la col·laboració de la comunitat.
5. *Developer* des d'on trobarem les opcions per accedir a la API i aplicacions de gestió de les dades dintre de CGC.



Il·lustració 27: Pàgina de selecció de projecte de CGC

Vam assignar un nom al projecte i vam entrar a al tauler del projecte des d'on podem accedir als arxius, les aplicacions i les tasques que associarem al nostre estudi.



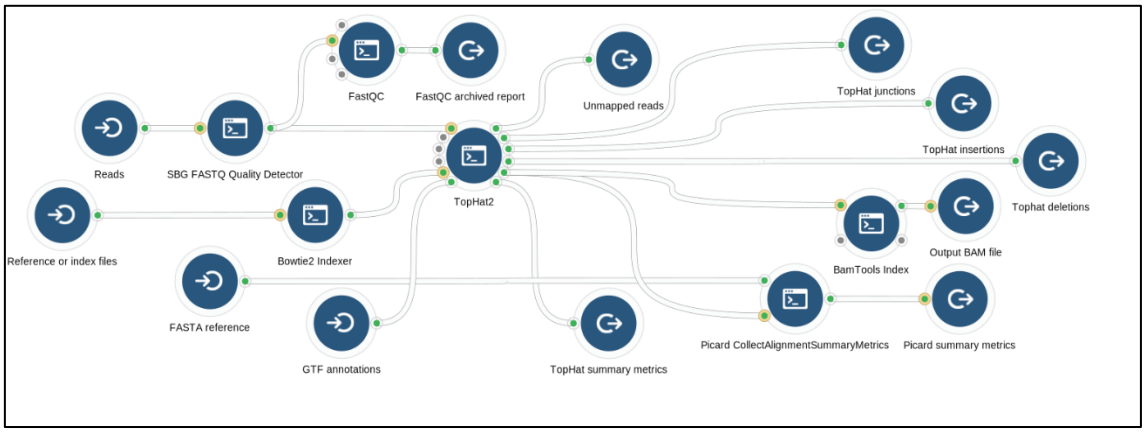
Il·lustració 28: Pàgina principal de projecte

En primer lloc es van pujar les dades dels *reads* que es van fer al "Centro Nacional de Anàlisis Genòmico" (CNAG-CRG). Aquestes dades presenten peculiaritats com que donat

el seu origen les mostres tenen moltes repeticions i son molt fragmentades.

Les dades es van pujar al núvol mitjançant l'aplicació d'escriptori per pujades massives. Es van pujar 16 lectures aparellades de dos en dos.

Un cop pujades les mostres es va procedir a terme l'alineament. Per dur-lo a terme es va importar l'aplicació pública RNA-seq Alignment–TopHat que és un procediment estandarditzat per dur a terme l'alineament i els controls de qualitat sobre les mostres. L'aplicació segueix el procediment de treball que veiem més a baix:



Il·lustració 29: Procediment RNA-seq / TopHat

Com a inputs trobem els reads (o les lectures de la seqüència a estudi), els arxius de referència o índex per *Bowtie* que és una indexació del genoma de referència per l'espècie per tal d'agilitzar i reduir els còmputs, l'arxiu FASTA del genoma de referència i les anotacions GTF que són arxius que contenen els noms de seqüències del genoma a de referència amb una sèrie de dades com ara localització cromosòmica, de cadena, etc.

Com a sortida trobem un arxiu de report FASTQC, el qual es composa d'un arxiu web i el seus complements per poder mostrar les gràfiques a l'informe. Així obtenim un arxiu con el que es veu en la imatge que mostra una sèrie de taules i gràfiques que valoren la qualitat de la mostra:

Il·lustració 30: FASTQC Report

Com a segona sortida trobem les mètriques de TopHat en un arxiu *txt* amb una text semblant al següent on s'explica com han alineats els reads de dreta i esquerra així com les parelles alineades i el seu percentatge d'alineament respecte al total de la mostra:

```
Left reads:
          Input     : 547760845
                           Mapped   : 272436149 (49.7% of input)
            of these:   6950818 ( 2.6%) have multiple alignments (291696 have >20)
Right reads:
          Input     : 547760845
           Mapped   : 264018891 (48.2% of input)
            of these:   6767319 ( 2.6%) have multiple alignments (291394 have >20)
49.0% overall read mapping rate.

Aligned pairs: 229944567
     of these:   5513552 ( 2.4%) have multiple alignments
                10437958 ( 4.5%) are discordant alignments
40.1% concordant pair alignment rate.
```

Il·lustració 31: Mètriques TopHat

Com a tercera sortida trobem un arxiu BAM amb els fragments no alineats. La quarta sortida són els fitxers BAM indexats amb els fragments alineats. Aquests arxius BAM, com hem indicat anteriorment, contenen la informació en format binari dels *reads* alineats a genoma, es a dir a partir d'aquests arxius podríem usar eines com UCSC Genome Browser per poder fer una visualització de les zones amplificades en els nostres *reads*. Però han de ser indexats per obtenir una visualització òptima.

La cinquena sortida son els arxius BED de les juncions, delecions i insercions. En aquests arxius que també es poden usar per la visualització, trobem modificacions del genoma que no son presents al de referència.

CB_accepted_hits.bam	14/12/2017 20:42	Archivo BAM	16,795,484 ...
CB_accepted_hits.bam.bai	14/12/2017 19:03	Archivo BAI	9,228 KB
CB_accepted_hits.summary_metrics.txt	14/12/2017 19:01	Archivo TXT	3 KB
CB_align_summary.txt	14/12/2017 19:01	Archivo TXT	1 KB
CB_deletions.bed	14/12/2017 19:01	Archivo BED	16,947 KB
CB_insertions.bed	14/12/2017 19:01	Archivo BED	4,385 KB
CB_junctions.bed	14/12/2017 19:01	Archivo BED	18,360 KB
CB_unmapped.bam	14/12/2017 21:10	Archivo BAM	20,207,223 ...

Il·lustració 32: Llista d'arxius generats

Per últim trobem una sèrie de mètriques provinents de l'aplicació Picard que permeten valorar la qualitat de l'alineament:

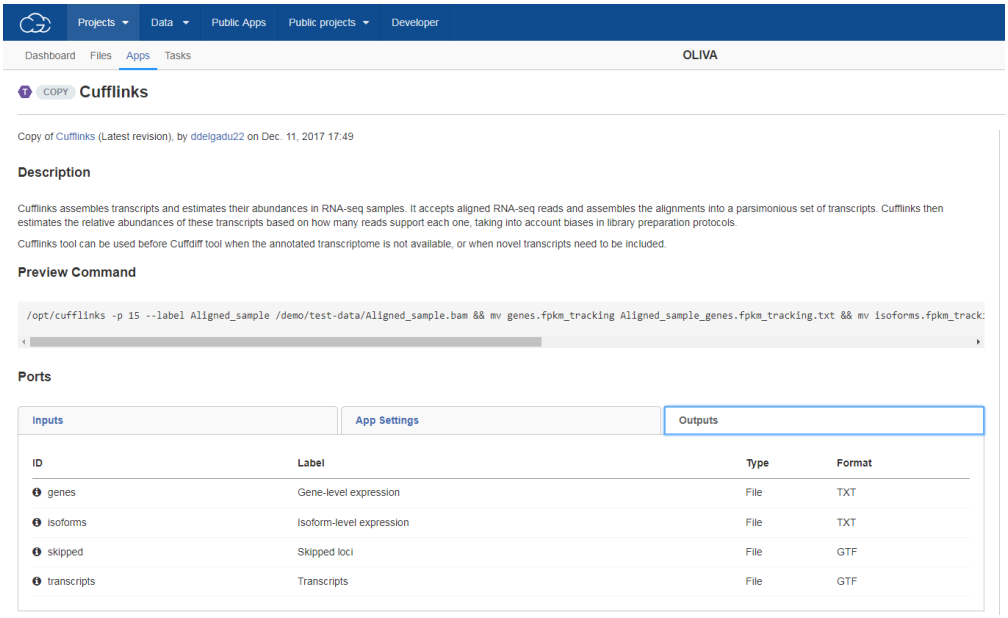
```
## htsjdk.samtools.metrics.StringHeader
# picard.analysis.CollectAlignmentSummaryMetrics REFERENCE_SEQUENCE=/sbgenomics/Projects/288ca407-0d19-43e7-
9a15-9c18c57b5650/human_g1k_v37_decoy.phiX174.fasta INPUT=/sbgenomics/Projects/288ca407-0d19-43e7-9a15-
9c18c57b5650/workspace/dd5736ca-65de-48b5-a66a-fbcec99a9222/rna-seq-alignment-
tophat_TopHat2/tophat_out/CB_accepted_hits.bam OUTPUT=CB_accepted_hits.summary_metrics.txt
VALIDATION_STRINGENCY=SILENT    MAX_INSERT_SIZE=100000
ADAPTER_SEQUENCE=[AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT,
AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT,
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG,
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT,
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG] METRIC_ACCUMULATION_LEVEL=[ALL_READS]
IS_BISULFITE_SEQUENCED=false ASSUME_SORTED=true STOP_AFTER=0 VERBOSITY=INFO QUIET=false COMPRESSION_LEVEL=5
MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json
## htsjdk.samtools.metrics.StringHeader
# Started on: Thu Dec 14 06:07:20 UTC 2017
```

```
## METRICS CLASS    picard.analysis.AlignmentSummaryMetrics
CATEGORY  TOTAL_READS         PF_READS  PCT_PF_READS        PF_NOISE_READS      PF_READS_ALIGNED
          PCT_PF_READS_ALIGNED          PF_ALIGNED_BASES    PF_HQ_ALIGNED_READS PF_HQ_ALIGNED_BASES
          PF_HQ_ALIGNED_Q20_BASES       PF_HQ_MEDIAN_MISMATCHES       PF_MISMATCH_RATE    PF_HQ_ERROR_RATE
          PF_INDEL_RATE       MEAN_READ_LENGTH    READS_ALIGNED_IN_PAIRS        PCT_READS_ALIGNED_IN_PAIRS
          BAD_CYCLES          STRAND_BALANCE      PCT_CHIMERAS        PCT_ADAPTER         SAMPLE    LIBRARY
          READ_GROUP
FIRST_OF_PAIR       272436149 272436149 1         3268928   272436149 1         8493541134          265485331
          8131368830          7874364908          0         0.694243  0.694296  0.000675  76        229944567
          0.844031  0         0.501595  0.045456  0
SECOND_OF_PAIR      264018891 264018891 1         3222303   264018891 1         8277398555          257251572
          7924872211          7612011789          0         0.694002  0.694069  0.000678  76        229944567
          0.87094   0         0.491466  0.045456  0
PAIR      536455040 536455040 1         6491231   536455040 1         16770939689         522736903
          16056241041         15486376697         0         0.694124  0.694184  0.000676  76        459889134
          0.857274  0         0.49661   0.045456  0
```

Il·lustració 33: Valoració de qualitat per Picard

Un cop seguit aquest procediment per dur a terme el alineament, es duu a terme la valoració de la transcripció mitjançant *Cufflinks* que ens permetrà, valorar les abundàncies i fer test d'expressió diferencial sobre la mostra.

En la imatge podem veure que obtindrem una quantificació dels gens i les isoformes així com una anotació sobre els loci saltats degut a la seva abundància i una altra dels fragments de la transcripció



Il·lustració 34: Pantalla principal de App Cufflinks

D. RESULTATS OBTINGUTS

El treball en aquest apartat no s'ha arribat a dur a terme dintre del període del TFM. El temps per dur a terme la part de transcriptòmica era molt ajustat. A més la meva inexperiència amb l'ús de les eines i la manca de temps per documentar-me han fet que ens topéssim amb dos problemes afegits:

1. Al treballar en un núvol bioinformàtic, la execució de les comandes no es transparent com en una estació de treball amb terminal, pel que he requerit un cert temps per "traduir" les comandes de terminal a el front-end de *CancerGenomicsCloud*.

2. La meva experiència en l'ús de les eines en mode terminal no era gran cosa, per tan m'ha requerit ja de per si documentació.

Amb això l'anàlisi ha quedat inconclús degut a que qualsevol configuració errònia donava lloc a una pèrdua de temps molt gran que no hem pogut esmorteir pel que finalment no hem arribat a tancar les anàlisis a temps i solament podem presentar els resultats intermedis del estudi.

Aquests resultats intermedis, tot i l'esforç per part del grup d'investigació no han donat lloc a cap resultat final. Tot i tenir una sèrie de regions cromosòmiques localitzades per obtenir la informació que corrobori les teories plantejades a l'estudi, no s'ha pogut extreure cap conclusió amb els elements intermedis que tenim.

Els resultats que he obtingut han arribat fins a la sortida de Top-Hat però un error a la hora de programar la execució de la tasca en *batch* de l'alineament de les diferents mostres pre y post rentat, ens va dur a tenir un alineament de totes les mostres juntes per comptes de separades. Així, tot i tenir un alineament fet, els resultats no són els correctes. Aquest alineament no és útil per dur a terme les observacions que el grup investigador volia portar a terme i un cop detectada la incidència és va descartar l'alineament i detenir tot el anàlisi posterior amb Cufflinks. La manca de temps com s'ha comentat anteriorment ha fet que aquest error es tornés insalvable per l'execució del projecte en els terminis establerts.

Per altra banda, aquest no tancament dels anàlisis a temps pot fer replantejar el *pathway* de l'anàlisi gràcies a les indicacions del meu consultor que ens recomanava no fer servir Cufflinks ja que es un mètode que es troba en desús. Segons les seves indicacions el anàlisi el continuarem amb DESeq2 com era intenció de la Dra. Jodar en un primer moment i que més tard va canviar per *Cufflinks* veient que el temps se'ns esgotava i ella havia dut a terme experiments prèviament amb aquest programari.

Per aquesta raó, el apartat de discussió de resultats ha estat canviat per un apartat de anàlisis ulteriors, en que s'intenta fer una explicació de com continuarà el nostre anàlisi des del punt en que ens trobem avui fins arribar a la consecució dels resultats que ens permetin confirmar o descartar les hipòtesis establertes en l'experiment.

Si bé, és una mala noticia no poder acabar l'anàlisi en el període del TFM, és un bon punt de partida per continuar vinculat al món de la bioinformàtica i continuar aprenent amb una temporització molt més relaxada.

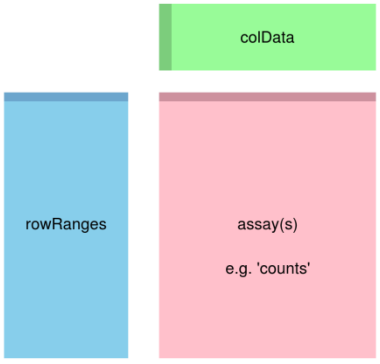
E. ANÀLISI ULTERIOR

El anàlisi de la part de transcriptòmica presentada en aquest treball haurà de continuar seguint el següent patró:

1. Procediment RNA-seq Alignment – TopHat aplicat en *batch* per cadascuna de les mostres individualment per tal d'obtenir 8 sortides SAM que permetin fer una visualització de les zones on es codifiquen les protamines a estudi. Per cadascun dels parells rentat/no rentat podrem comprovar individualment la presència\absència de la seqüència que dona lloc a les proteïnes a estudi.

2. Cufflinks/DESeq2/EdgeR: amb aquests programes/llibreries podrem quantificar el número de *reads* i fer l'expressió diferencial.
   Pel que fa aquesta part un cop descartat Cufflinks, sembla quasi bé segur que es farà ús de *DESeq2*, donat que personalment he fet ús abans de la llibreria en R tot i que ha estat en pràctiques dirigides.

   a) Per dur a terme l'anàlisi partirem d'arxius BAM que haurem d'obtenir mitjançant el punt anterior i el ús de *SAMtools.*
   b) Tot seguit passarem a crear un taula d'anotacions *TxDb* mitjançant la importació d'un arxiu d'anotacions que podrem ordenar per gens. Aquest pas també es pot dur a terme mitjançant la descarrega des d' *AnnotationHub*.
   c) A continuació caldrà crear una taula de *counts* d'RNA-seq. Per fer-ho es procedirà a llegir el disseny experimental. En el nostre cas la matriu es senzilla, tractament contra no tractament.
   d) Importarem els arxius BAM per poder crear mitjançant *summarizeOverlaps* un objecte de tipus *SummarizedExperiment* que contindrà les dades del experiment tal i com es mostra en la il·lustració. Al bloc blau trobarem informació sobre els blocs genòmics, al bloc verd trobarem informació sobre les mostres i l'experiment i per últim al bloc rosa trobarem els valors sumaritzats



Il·lustració 35: Objecte Summarized Experiment [32]

e) A partir d'aquest punt podrem fer diferents anàlisis dintre del que ofereix DESeq2, amb el disseny experimental que volem aplicar o amb variacions que el grup investigador consideri adients. Podrem començar per una lectura exploratòria (recompte, distàncies, PCA,...) , continuar per l'expressió diferencial amb les sortides gràfiques escaients, i acabar per decidir quin és el millor format per presentar les sortides de l'anàlisi.

A partir d'aquest punt ampliarem l'anàlisi a com el havíem previst inicialment en les fases primerenques del TFM amb l'estudi de circ-RNA.

3. Ús de KNIFE[33]/ Machete[34] per detectar isoformes de circRNA tan anotades como *de novo*.

4. Aprofundir en l'anàlisi amb la inclusió de small-RNA en el nostre estudi.

Aquests estudis en un principi es duran a terme en la mateixa plataforma *cloud* que s'ha fet servir durant el TFM.

4. CONCLUSIONS

Aquest treball de fi de Màster, tot i que no he pogut assolir tots els objectius ens deixa una sèrie de resultats que poden donar lloc a la redacció d'articles per publicar en revistes d'investigació.

Per una banda com s'ha comentat, ens trobem amb els resultats de la part de proteòmica de fluid seminal on s'han extret les conclusions que es presenten en l'apartat de interpretació dels resultats. Tot i així, queda clar que encara es pot treure més informació dels estudis realitzats pel que el tancament fet en vista a la presentació d'aquest TFM pot ser temporal reobrint l'anàlisi per poder obtenir més informació més endavant. A més es fa evident que caldrà valorar i confirmar les conclusions que hem extret d'aquesta part del treball fent els experiments necessaris.

En aquest sentit un dels projectes que s'ha comentat és l'obtenció de forma massiva del GO i les vies metabòliques que presenten les proteïnes involucrades en les nostres conclusions per aquest TFM o l'automatització de certes tasques que duen a terme els investigadors que prenen gran quantitat de temps i que un programa informàtic podria fer de forma massiva en un temps molt curt.

Pel que fa el meu objectiu en aquest punt [punt 1 pàg.12], he arribat a obtenir una sèrie d'algorismes que poden pal·liar els defectes de quantificació, però com he comentat caldrà valorar i confirmar els resultats obtinguts per poder assegurar al 100% que els algorismes creats donen resultats vàlids.

La part de transcriptòmica, tot i que no ha arribat a bon port m'ha donat la possibilitat d'aprendre de les fallades i posar-me mans a l'obra per obtindré els resultats i aprofundir en la utilització de la plataforma Cancer Genomics Cloud, però a més tindre la possibilitat de fer ús d'una estació dedicada dins del laboratori que arribarà en dates properes, pel que un dels objectius que inicialment es van plantejar en aquest TFM, i que per tenir una temporització molt ajustada es va descartar, el enllestir un estació de treball dedicada a l'estudi bioinformàtic es farà realitat.

Si tot va bé i es confirmen les hipòtesis que el grup investigador planteja, la part de transcriptòmica, que aquest TFM ha deixat encarada per obtenir resultats valorables, podrà donar lloc a la publicació d'un article derivat.

Pel que fa el meu objectiu en aquesta part ha estat assolit [punt 2, pàg.12]. Estrictament parlant, he determinat un procediment per l'anàlisi de mostres procedents de RNA-seq. Encara que per la no finalització del procediment no l'he pogut mostrar afinat i més detallat, cal dir que en aquest TFM es presenta un procediment de treball per obtenir expressió diferencial a partir de mostres de RNA-seq.

El tercer objectiu, [punt 3, pàg. 12] no ha estat assolit en tota la magnitud amb que es va plantejar donat que aquest objectiu contava amb un apartat propi de documentació del programari que s'ha vist reduït a la part d'Estudi d'eines aplicables de cada capítol

d'aquest TFM. Si bé, tot i no haver-ho documentat profusament, si que s'ha arribat en part al objectiu presentat. En aquest apartat, la feina continuarà donat que em caldrà la documentació d'aquest programari i la redacció de pautes de treball si vull exercir com a bioinformàtic.

En conclusió, com he indicat en aquestes línies, aquest TFM ha estat un èxit. No per haver assolit les fites que es van plantejar en un primer moment, ben al contrari he hagut de modificar el projecte i adequar-lo a la temporització i als problemes sorgits. L'he retallat per sobre del que m'hagués agradat, però tot i així, com he dit considero que ha estat exitós.

L'èxit d'aquest TFM ha estat per mi el poder vincular-me al món de la investigació i la bioinformàtica de tal manera que un cop acabat el període lectiu, podré continuar estant en contacte amb el grup d'investigació, continuant el meu aprenentatge i col·laborant en les investigacions que duguin a terme.

5. GLOSARI

TMT 10-PLEX: El marcatge diferencial isobàric amb TMT 10-plex consisteix en un total de 10 isòtops diferents que s'uneixen de manera covalent a l'extrem N- o C- terminal dels residus de lisina de les proteïnes. Aquests diferents isòtops presenten una massa monoisotòpica diferent, la qual pot ser detectada utilitzant espectrometria de masses d'alta resolució, de tal manera que permet la quantificació de les proteïnes de mostres independents.

ESPECTOMETRIA DE MASSES: La espectrometria de masses d'alta resolució permet determinar la distribució de les molècules en funció de la seva massa i càrrega, de tal manera que permet la quantificació de les proteïnes de mostres independents.

RNA-SEQ: aproximació de seqüenciació de nova generació (NGS) que quantifica la presència de mRNA en una mostra biològica. S'empra per fer el mapeig els canvis en el transcriptoma cel·lular.

NORMOZOOSPÈRMIA: Terme utilitzat com a diagnòstic d'un seminograma que significa que les característiques seminals compleixen els criteris de normalitat que estableix l'Organització Mundial de la Salut (OMS).

AZOOSPÈRMIA: Absència d'espermatozoides en el semen. Afecta aproximadament al 2% dels homes. Pot ser degut a que no es produeixen els espermatozoides al testicle o a una obstrucció en els conductes de la via seminal que impedeix que arribin a ejacular-se.

OLIGOZOOSPÈRMIA: Terme utilitzat com a diagnòstic d'un seminograma que significa que la concentració d'espermatozoides és menor de 15 milions per mil·lilitre o bé que a la totalitat de l'ejaculat hi ha menys de 40 milions d'espermatozoides.

ASTENOZOOSPÈRMIA: Es defineix així la disminució del nombre d'espermatozoides mòbils. És l'alteració seminal més freqüent.

6. BIBLIOGRAFIA

[1] **Amaral, A.; Castillo, J.; Ramalho-Santos, J.; Oliva, R.** The combined human sperm proteome: cellular pathways and implications for basic and clinical science. *Hum Reprod Update*., **2014**, 20, 40–62.

[2] **Jodar, M.; Selvaraju, S.; Sendler, E.; Diamond, MP.; Krawetz, SA.** The presence, role and clinical use of spermatozoal RNAs. *Hum Reprod Update*., **2013**, 19, 604–24.

[3] **Jodar, M.; Soler-Ventura, A.; Oliva, R.** Semen proteomics and male infertility. Vol. 162, Journal of Proteomics. **2017**, p. 125–34.

[4] **Amaral, A.; Castillo, J.; Estanyol, JM.; Ballesca, JL.; Ramalho-Santos, J.; Oliva, R.** Human Sperm Tail Proteome Suggests New Endogenous Metabolic Pathways. *Mol Cell Proteomics*., **2013**, 12, 330–42.

[5] **Castillo, J.; Amaral, A.; Oliva, R.** Sperm nuclear proteome and its epigenetic potential. *Andrology*., **2014**, 2, 326–38.

[6] **Amaral, A.; Paiva, C.; Attardo Parrinello, C.; Estanyol, JM.; Ballescà, JL.; Ramalho-Santos, J.; et al.** Identification of proteins involved in human sperm motility using high-throughput differential proteomics. *J Proteome Res*., **2014**, 13, 5670–84.

[7] **Bogle, OA.; Kumar, K.; Attardo-Parrinello, C.; Lewis, SEM.; Estanyol, JM.; Ballescà, JL.; et al.** Identification of protein changes in human spermatozoa throughout the cryopreservation process. *Andrology*., **2017**, 5, 10–22.

[8] **Azpiazu, R.; Amaral, A.; Castillo, J.; Estanyol, JM.; Guimerà, M.; Ballescà, JL.; et al.** High-throughput sperm differential proteomics suggests that epigenetic alterations contribute to failed assisted reproduction. *Hum Reprod*., **2014**, 29, 1225–37.

[9] **Jodar, M.; Oriola, J.; Mestre, G.; Castillo, J.; Giwercman, A.; Vidal-Taboada, JM.; et al.** Polymorphisms, haplotypes and mutations in the protamine 1 and 2 genes. *Int J Androl*., **2011**, 34, 470–85.

[10] **Jodar, M.; Oliva, R.** Genetic Damage in Human Spermatozoa. **2014**, 83-102 p.

[11] **Krausz, C.; Giachini, C.; Xue, Y.; O'Bryan, MK.; Gromoll, J.; Meyts, ER-d.; et al.** Phenotypic variation within European carriers of the Y-chromosomal gr/gr deletion is independent of Y-chromosomal background. *J Med Genet*., **2008**, 46, 21–31.

[12] **Pantano, L.; Jodar, M.; Bak, M.; Ballescà, JL.; Tommerup, N.; Oliva, R.; et al.** The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. *RNA*., **2015**, 21, 1085–95.

[13] **Jodar, M.; Sendler, E.; Moskovtsev, SI.; Librach, CL.; Goodrich, R.; Swanson, S.; et al.** Absence of sperm RNA elements correlates with idiopathic male infertility. *Sci Transl Med*., **2015**, 7, 295re6.

[14] **Jodar, M.; Kalko, S.; Castillo, J.; Ballescà, JL.; Oliva, R.** Differential RNAs in the sperm cells of asthenozoospermic patients. *Hum Reprod*., **2012**, 27, 1431–8.

[15] **Castillo, J.; Amaral, A.; Azpiazu, R.; Vavouri, T.; Estanyol, JM.; Ballescà, JL.; et al.** Genomic and proteomic dissection and characterization of the human sperm chromatin. *Mol Hum Reprod*., **2014**, 20, 1041–53.

[16] **Jodar, M.; Sendler, E.; Krawetz, SA.** The protein and transcript profiles of human semen. *Cell Tissue Res*., **2016**, 363, 85–96.

[17] **Barrett, SP.; Salzman, J.** Circular RNAs: analysis, expression and potential functions. *Development*., **2016**, 143, 1838–47.

[18] Camargo, M.; Intasqui, P.; Bertolla, R. Proteomic profile of seminal plasma in adolescents and adults with treated and untreated varicocele. *Asian J Androl* [Internet]., 2016, 18, 194. Available from: http://www.ajandrology.com/text.asp?2016/18/2/194/168788

[19] Giacomini, E.; Ura, B.; Giolo, E.; Luppi, S.; Martinelli, M.; Garcia, RC.; et al. Comparative analysis of the seminal plasma proteomes of oligoasthenozoospermic and normozoospermic men. *Reprod Biomed Online*., 2015, 30, 522–31.

[20] Gilany, K.; Minai-Tehrani, A.; Savadi-Shiraz, E.; Rezadoost, H.; Lakpour, N. Exploring the human seminal plasma proteome: An unexplored gold mine of biomarker for male Infertility and male reproduction disorder. Vol. 16, Journal of Reproduction and Infertility. 2015, p. 61–71.

[21] Drabovich, AP.; Saraon, P.; Jarvi, K.; Diamandis, EP. Seminal plasma as a diagnostic fluid for male reproductive system disorders. *Nat Rev Urol*., 2014, 11, 278–88.

[22] Vilagran, I.; Yeste, M.; Sancho, S.; Castillo, J.; Oliva, R.; Bonet, S. Comparative analysis of boar seminal plasma proteome from different freezability ejaculates and identification of Fibronectin 1 as sperm freezability marker. *Andrology*., 2015, 3, 345–56.

[23] Gundersen, VB. R: MATLAB commands in numerical Python (NumPy). *Manual* [Internet]., 2006, 0, 17. Available from: papers2://publication/uuid/14EC328F-448B-49D5-B1DF-C02B9D156CA4

[24] Jain, K. SAS Vs. R Vs. Python- Which Analytics Tool Should I Learn? *Anal Vidya* [Internet]., 2014, 1–47. Available from: https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

[25] Wayner, P. Python vs. R: The battle for data scientist mind share [Internet]. InfoWorld. 2017,. Available from: http://www.infoworld.com/article/3187550/data-science/python-vs-r-the-battle-for-data-scientist-mind-share.html

[26] Mcclelland, M.; Wang, Y. PypeR, A Python Package for Using R in Python. *J Stat Softw*., 2010, 35, 1--8.

[27] RStudio Team,-. RStudio: Integrated Development for R. *[Online] RStudio, Inc, Boston, MA URL http//www rstudio com*., 2016, RStudio, Inc., Boston, MA.

[28] Shapiro, SS.; Wilk, MB. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* [Internet]., 1965, 52, 591. Available from: http://www.jstor.org/stable/2333709?origin=crossref

[29] Pettitt, AN. Testing the Normality of Several Independent Samples Using the Anderson-Darling Statistic. *J R Stat Soc Ser C (Applied Stat* [Internet]., 1977, 26, 156–61. Available from: http://www.jstor.org/stable/2347023%5Cnhttp://www.jstor.org.ezproxy.library.tufts.edu/stable/pdfplus/2347023.pdf?acceptTC=true

[30] Razali, NM.; Wah, YB. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal*., 2011, 2, 21–33.

[31] Data, R. SQL Joins [Internet]. w3schools.com. 2014 [cited 2017 Dec 30], p. 1. Available from: https://www.w3schools.com/sql/sql_join.asp

[32] Love, M.; Anders, S.; Huber, W. Differential expression, manipulation, and visualization of RNA-seq reads [Internet]. 2015 [cited 2018 Jan 7],. Available from: https://www.bioconductor.org/help/course-

materials/2015/BioC2015/bioc2015rnaseq.html

[33] Szabo, L.; Morey, R.; Palpant, NJ.; Wang, PL.; Afari, N.; Jiang, C.; et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol* [Internet]., 2015 [cited 2018 Jan 7], 16, 126. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26076956

[34] Hsieh, G.; Bierman, R.; Szabo, L.; Lee, AG.; Freeman, DE.; Watson, N.; et al. Statistical algorithms improve accuracy of gene fusion detection. *Nucleic Acids Res* [Internet]., 2017 [cited 2018 Jan 7], 45, e126. Available from: http://academic.oup.com/nar/article/45/13/e126/3852042

ANNEX A- TESTS CONVENCIONALS: MOSTRA NO NORMALITZADA

HIGH-TROUGHPUT PROTEOMICS DETECTS DIFERENTIALLY EXPRESSED SEMINAL PLASMA PROTEINS CORRELATED TO SPERM PARAMETERS

TEST EN MUESTRA NO NORMALIZADA (DATOS ESTANDARIZADOS)

David Delgado Dueñas

Noviembre 2017

Contenido

Gestión de los datos

Importamos las tablas y las adecuamos a nuestras necesidades de análisis:

```
> # Importamos tabal Excel con datos sobre quantifiación de
> # proteínas
> PcaTotalG <- readXL("Datos/Libro1.xlsx", rownames = TRUE, header = TRUE,
+     na = "NA", sheet = "Hoja2", stringsAsFactors = FALSE)
> colnames(PcaTotalG) <- c("Unique Peptides", "PSMs", "N1", "N2",
+     "N3", "N4", "A1", "A2", "A3", "A4", "O1", "O2", "O3", "O4",
+     "Z1", "Z2", "Z3", "Z4")
> names <- row.names(PcaTotalG)
> E <- PcaTotalG[, c(-1, -2)]
> E <- E[complete.cases(E), ]
>
>
> PcaTotal <- readXL("Datos/Libro1.xlsx", rownames = FALSE, header = TRUE,
+     na = "NA", sheet = "Hoja1", stringsAsFactors = TRUE)
> PcaTotalF <- PcaTotal[2:10]
> PcaTotalF <- rbind(`Unique Peptides` = NA, PcaTotalF)
> PcaTotalF <- rbind(PSMs = NA, PcaTotalF)
> rownames(PcaTotalF) <- c("Unique Peptides", "PSMs", "N1", "N2",
+     "N3", "N4", "A1", "A2", "A3", "A4", "O1", "O2", "O3", "O4",
+     "Z1", "Z2", "Z3", "Z4")
```

Test sobre la igualdad de las distribuciones de las muestras

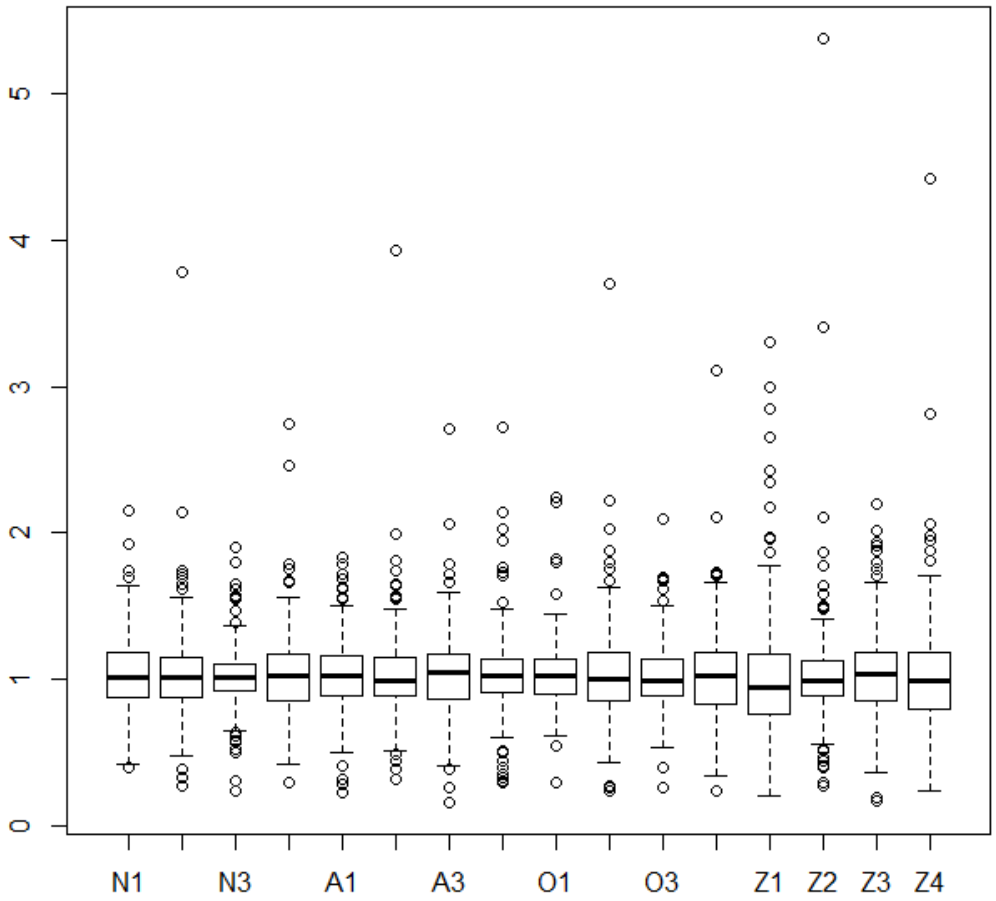
Comenzaremos estudiando la normalidad de la expresión proteica en las 16 muestras con la ayuda del test de Shapiro y el test de Anderson-Darling. Seguidamente, mediante un diagrama de cajas, una gráfica de densidades comparadas y la relación de cuantil-cuantil estudiamos la dstribución de la muestras.

Teniendo en cuenta la aplicación de una estandarización previa a los datos, se desaconseja la aplicación de más transformaciones que pueden atenuar las señales.

```
> Normal(E)
```

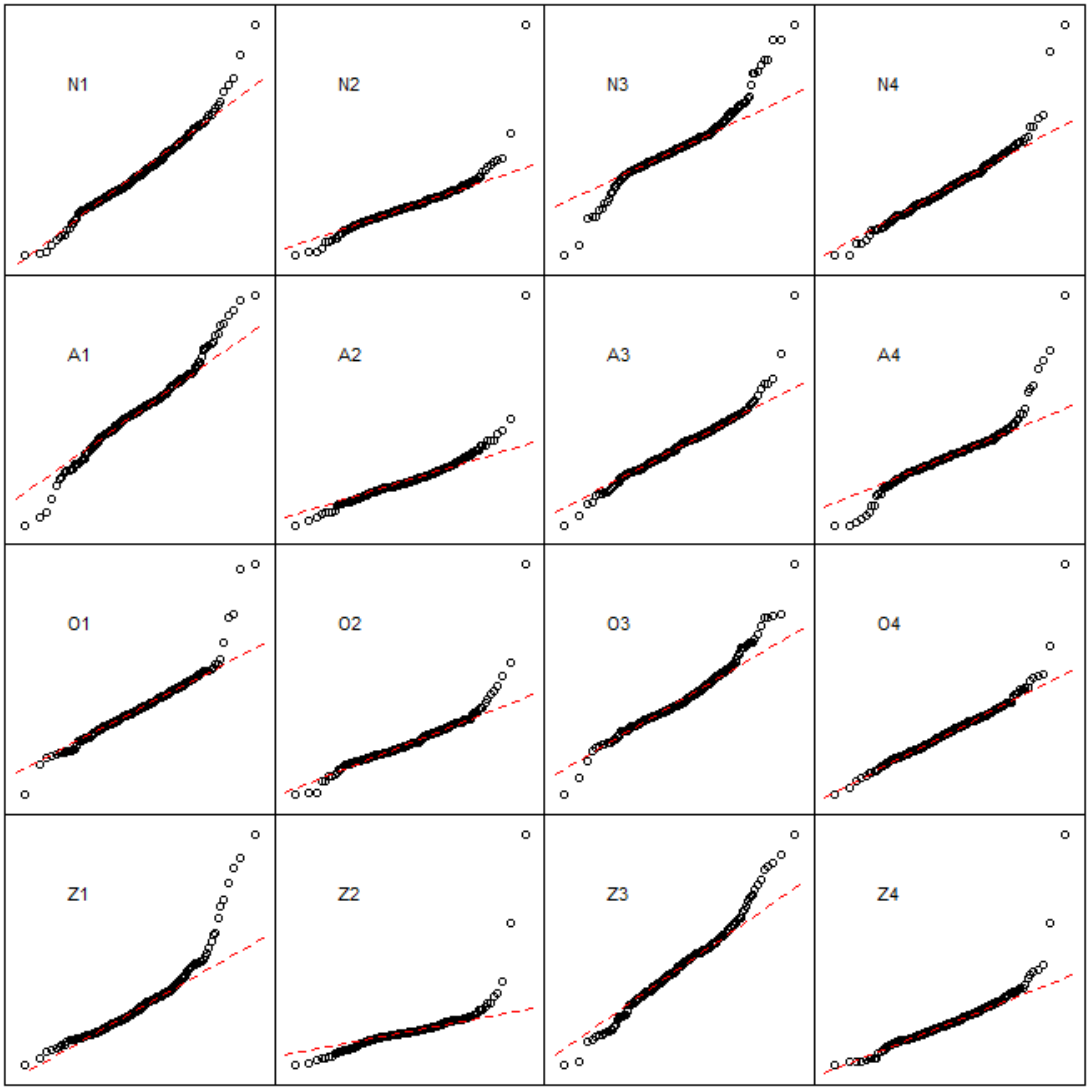
Proteina	S.W	pvalor	A.D	porcentaje
N1	No	1e-04	No	0.3188338
N2	No	0e+00	No	0.0000000
N3	No	0e+00	No	0.0000000
N4	No	0e+00	No	0.0088358
A1	No	3e-04	No	0.0118067
A2	No	0e+00	No	0.0000000
A3	No	0e+00	No	0.0549600
A4	No	0e+00	No	0.0000000
O1	No	0e+00	No	0.0001446
O2	No	0e+00	No	0.0000000
O3	No	0e+00	No	0.0000138
O4	No	0e+00	No	0.0144133
Z1	No	0e+00	No	0.0000000
Z2	No	0e+00	No	0.0000000
Z3	No	7e-04	No	0.2058470
Z4	No	0e+00	No	0.0000000

```
> boxplot(E)
```



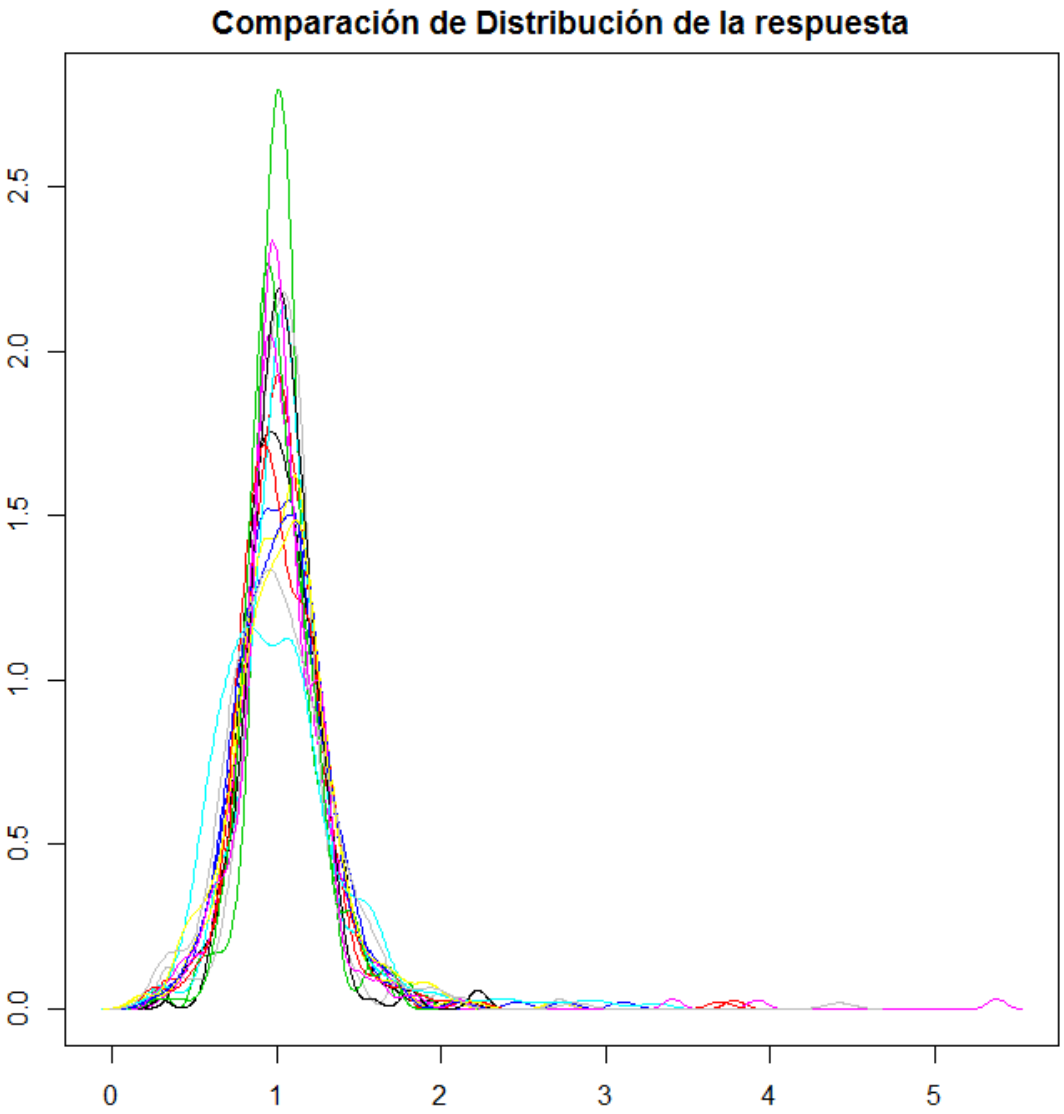
El boxplot nos muestra que las distribuciones siguen un patrón similar pero algunas tienen gran número de elementos divergentes (outliers).

```
> qqploted(E)
```



El gráfico de cuantiles nos muestra que las luminancias obtenidas no muestran una distribución normal, en principio la expresión proteica debería seguir una distribución normal.

```
> densited(E)
```



Como se puede observar por el gráfico de densidades, existen una serie de outliers que determinan gravemente la distribución. Por otro lado las distribuciones no son todo lo similar que se esperaría para poder hacer un estudio significativo.

Test de Kruskal-Wallis

Aplicamos Kruslkal-Wallis sobre las luminancias de las proteínas para las diferentes muestras:

```
> names <- row.names(E)
> raw.work.t <- as.data.frame(t(E))
> colnames(raw.work.t) <- names
> raw.work.t$myfactor <- factor(row.names(raw.work.t))
>
> Grupos <- c("N", "N", "N", "N", "A", "A", "A", "A", "O", "O",
+       "O", "O", "Z", "Z", "Z", "Z")
> raw.work.t <- cbind(Grupos, raw.work.t)
>
> # Columna por columna hacemos KW vs los Grupos
>
> SignificantProteinKW <- data.frame(Proteina = character(), P.valor = character(
),
+       stringsAsFactors = FALSE)  #Creamos una tabla vacía
> NotSignificantKW <- data.frame(Proteina = character(), P.valor = character(),
+       stringsAsFactors = FALSE)  #Creamos una tabla vacía
>
> # Empezamos un bucle for desde 2 hasta la última columna
> for (i in 2:(ncol(raw.work.t) - 1)) {
+       # Realizamos el test
+       Alpha <- kruskal.test(raw.work.t[, i] ~ Grupos, raw.work.t)
+       # Nos quedamos el p.valor
+       P <- unlist(Alpha)["p.value"]
+       Wii <- data.frame(Proteina = colnames(raw.work.t)[i], P.valor = round(as.n
umeric(P),
+             digits = 3))
+       # Si el p.valor es inferior a 0.05 anotamos la proteína y
+       # mostramos su p-valor
+
+       if (P < 0.05) {
+             SignificantProteinKW <- rbind(SignificantProteinKW, Wii)
+       } else {
+             NotSignificantKW <- rbind(NotSignificantKW, Wii)
+       }
+ }
>
> # Imprimimos la tabla de Proteínas significativas con su
> # p-valor
> SignificantProteinKW
```

Obtenemos un listado de las proteínas con mayor significancia por su p-valor:

Proteina	P.valor
NPC2	0.009
CAMP	0.046
PTGDS	0.050
PLA2G2A	0.029
SPINT3	0.034
SDCBP	0.048
LDHC	0.016
LAMC1	0.021

Y por otro lado imprimimos también aquellas que han obtenido un p-valor superior a 0.05 pero aún así se tratan de valores muy bajos.

```
> head(NotSignificantKW[with(NotSignificantKW,  order(P.valor)),
+        ])
```

	Proteina	P.valor
48	CRISP1	0.052
43	SCGB1A1	0.053
44	EDDM3B	0.053
192	SIAE	0.056
37	ECM1	0.057
135	SHISA5	0.060

Correlación concentración contra luminancia de las proteínas

```
> # cORRELATION: C VS PROTS
>
> M <- PcaTotalF[3:18, 4]
> C <- PcaTotalF[3:18, 3]
>
> Pear.C.raw <- cbind(C, raw.work.t[c(-1)])
> ## Los Azo tienen el mismo valor al correlacionar no
> ## obtendremos valores exactos porque hay 'ties'
> ## (coincidencias) en la variable dependiente
> cor.c.raw <- correlate(Pear.C.raw, "spearman")
> head(cor.c.raw[with(cor.c.raw, order(corrA, decreasing = T)),
+       ][1:3])
```

	Prot	sign	corr
rho	C	0.000	1.000
rho181	LDHC	0.000	0.853
rho10	NPC2	0.000	0.790
rho39	ECM1	0.001	0.763
rho138	SPINT3	0.004	0.677
rho140	SHISA5	0.004	-0.677

Correlación motilidad contra luminancia de las proteínas

```
> # cORRELATION:  MOTILIDAD  VS  PROTS
>
> Pear.M.raw <- cbind(M,  raw.work.t[c(-1)])
>
> cor.m.raw <- correlate(Pear.M.raw,  "spearman")
> head(cor.m.raw[with(cor.m.raw,  order(corrA,  decreasing  =  T)),
+        ][1:3])
```

	Prot	sign	corr
rho	M	0.000	1.000
rho38	MSMB	0.016	0.692
rho106	ALDOA	0.020	-0.671
rho209	LAMC1	0.022	-0.664
rho35	ANPEP	0.043	0.601
rho204	TMPRSS2	0.043	0.601

Corrección de la concentración contra luminancia de las proteínas

```
> # FDR
>
> cor.c.raw.fdr <- pHCorr(cor.c.raw, "fdr")
> head(cor.c.raw.fdr[with(cor.c.raw.fdr, order(corrA, decreasing = T)),
+       ][c(1, 2, 3, 5)], 10)
```

	Prot	sign	corr	FDR
rho	C	0.000	1.000	0.00000
rho181	LDHC	0.000	0.853	0.00000
rho10	NPC2	0.000	0.790	0.00000
rho39	ECM1	0.001	0.763	0.06250
rho138	SPINT3	0.004	0.677	0.14286
rho140	SHISA5	0.004	-0.677	0.14286
rho24	CAMP	0.004	0.674	0.14286
rho50	CRISP1	0.008	0.636	0.25000
rho61	PTGDS	0.009	0.628	0.25000
rho217	TSPAN1	0.014	0.601	0.35000

Corrección de la motilidad contra luminancia de las proteínas

```
> cor.m.raw.fdr <- pHCorr(cor.m.raw, "fdr")
> head(cor.m.raw.fdr[with(cor.m.raw.fdr, order(corrA, decreasing = T)),
+       ][c(1, 2, 3, 5)], 10)
```

	Prot	sign	corr	FDR
rho	M	0.000	1.000	0.00000
rho38	MSMB	0.016	0.692	0.96822
rho106	ALDOA	0.020	-0.671	0.96822
rho209	LAMC1	0.022	-0.664	0.96822
rho35	ANPEP	0.043	0.601	0.96822
rho204	TMPRSS2	0.043	0.601	0.96822
rho133	HSPA8	0.046	-0.594	0.96822
rho86	CD177	0.049	0.587	0.96822
rho168	PRCP	0.049	-0.587	0.96822
rho33	PPIB	0.052	-0.580	0.96822

PCA

Realizamos el estudio de los principales componentes (PCA):

```
> pca <- prcomp(raw.work.t[, c(-1, -251)], center = TRUE, scale. = TRUE)
```

Sumarizamos los resultados del estudio:

```
> summary(pca)

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     8.439 5.2998 4.61009 4.28464 4.16869 3.69345
Proportion of Variance 0.286 0.1128 0.08535 0.07373 0.06979 0.05479
Cumulative Proportion  0.286 0.3988 0.48419 0.55791 0.62770 0.68249
                          PC7    PC8    PC9   PC10   PC11    PC12
Standard deviation     3.51108 3.3212 3.25151 3.20443 2.95132 2.76971
Proportion of Variance 0.04951 0.0443 0.04246 0.04124 0.03498 0.03081
Cumulative Proportion  0.73200 0.7763 0.81876 0.85999 0.89497 0.92578
                          PC13   PC14    PC15      PC16
Standard deviation     2.66802 2.4849 2.27742 6.004e-15
Proportion of Variance 0.02859 0.0248 0.02083 0.000e+00
Cumulative Proportion  0.95437 0.9792 1.00000 1.000e+00
```

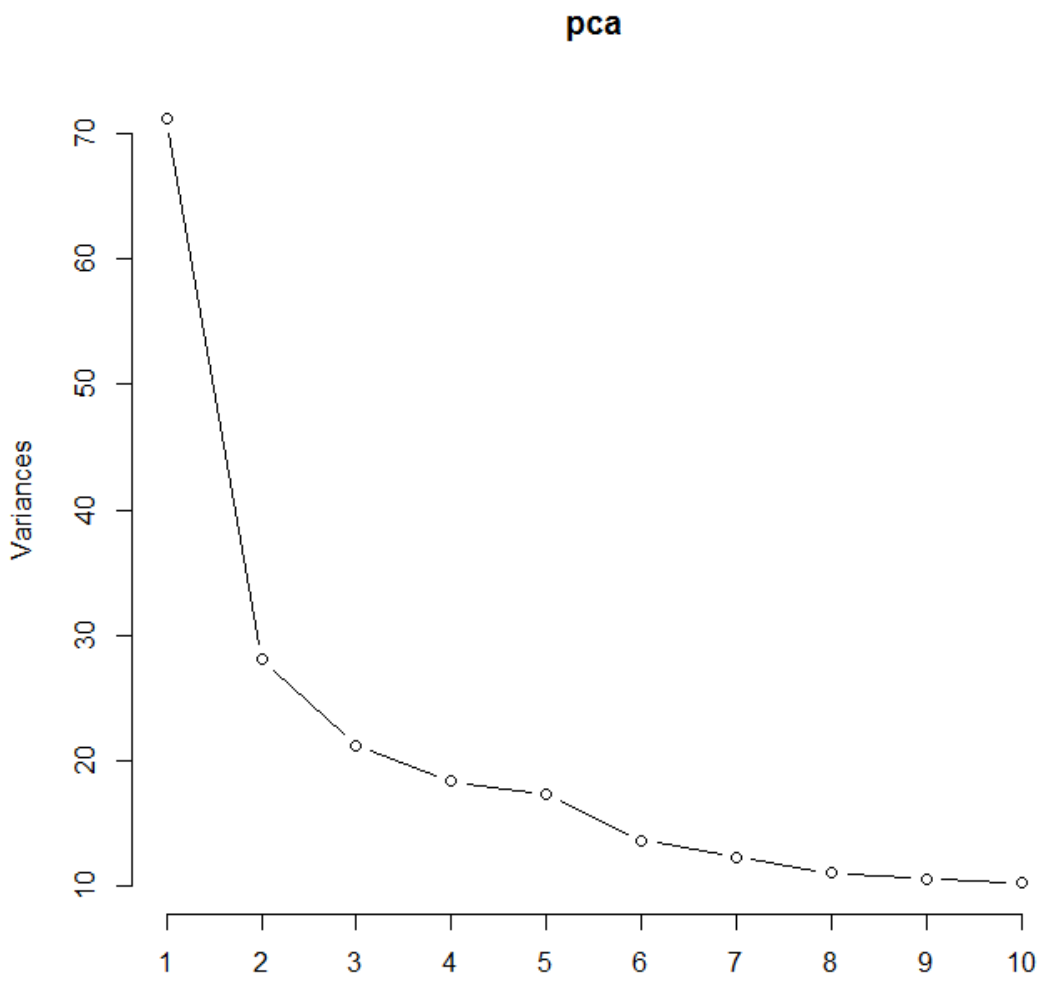
Mostramos las revoluciones:

```
> pca$x
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
N1	-5.658767	7.6477344	-4.0424870	3.3715980	-2.7404540	2.4608165	-5.0526378	0.3259358	-1.024835	0.5774525	-1.4759519	-5.0761161	2.9835606	0.5938734	3.3295662	0
N2	-6.748800	0.3461712	-0.1404597	4.5421332	5.9745266	0.9296952	1.8755189	1.7349068	-4.229017	7.0436632	-1.1851795	1.5430745	-4.0253302	-0.4909115	1.2642190	0
N3	2.642406	-0.6517129	2.9980408	-4.4023933	2.7248020	-2.1604200	0.3313102	2.0012794	4.196604	6.0782157	-0.7647025	-3.1662102	3.8784859	-2.5251762	-2.8315116	0
N4	10.450157	4.4454761	0.0778021	-9.2716336	0.8183842	-0.1020215	-0.7370261	-3.5917746	-2.545438	1.5256521	-2.1684213	0.7253093	-1.0170118	5.1497287	0.0972582	0
A1	-3.444509	6.0498953	3.6352300	0.8499462	-3.8368292	-6.1375403	-0.7576363	-1.6394375	-1.253381	1.3345642	4.8987615	5.0892359	2.7991089	-0.9643914	1.3047138	0
A2	-2.893731	4.2102645	-5.6788805	2.4516732	-3.6433871	5.2272866	6.7078608	-4.8037973	3.791500	0.6243081	-1.3368345	2.0716642	0.6516595	-0.5429632	-1.8922207	0
A3	-9.271068	1.3715381	-2.2372181	-4.6834989	-4.3323683	-4.2237250	-3.0554027	0.9207399	3.077430	-1.7510680	-2.8853310	0.1286802	-5.7384600	-2.8862605	-0.2148399	0
A4	4.264913	0.4659748	1.5365926	-3.0642435	-2.2895360	3.2405308	4.8240481	7.3700086	2.970810	-1.0061221	3.9203175	-0.5290945	-1.3643522	1.5188664	3.1588950	0
O1	-1.592217	3.0310446	0.9480345	3.8660181	-1.5205659	-4.4203056	3.8949629	1.9077054	-4.392987	-3.2219564	0.9941505	-4.0347610	-0.8987808	1.9950789	-4.4903182	0
O2	12.103944	1.0074035	4.1994241	2.3679655	1.3625116	-0.5166306	1.7552898	2.0899202	-1.716500	-4.2120165	-6.6190827	2.3083183	2.0318721	-2.8378667	1.5220668	0
O3	3.608018	-0.3709823	-5.1118104	-4.2572348	2.5494339	5.5272876	-3.2365775	0.9004301	-4.718986	-2.3324870	3.9477108	1.0377692	0.3675753	-3.9920585	-2.2263777	0
O4	-6.139236	-8.5818551	4.9044465	-2.5345861	0.4242906	0.0677800	3.3840490	-6.4854056	-1.806059	-1.6203745	1.3170428	-3.6899553	-0.1112636	-1.7902044	2.8966049	0
Z1	18.542249	-8.0136765	-7.2945212	5.3275409	-1.8450866	-3.9213027	-1.7934349	-1.3882189	1.583953	1.5610587	1.5917192	-0.5875970	-1.1978003	0.2895628	0.7711142	0
Z2	2.926852	3.8919281	7.5272798	4.8510117	5.8233979	2.8354162	-4.4356196	-2.5463392	5.012660	-2.4651730	2.0734080	-0.1145077	-2.4468119	1.0361228	-1.1429580	0
Z3	-6.114538	-9.9233479	4.8685252	1.4724230	-7.1189341	4.1462834	-3.7374105	1.8044292	-1.108307	1.3378074	-1.7413274	2.1872685	1.1264483	2.4921753	-1.9685408	0
Z4	-12.675672	-4.9258559	-6.1899987	-0.8867198	-7.6498145	-2.9531507	0.0327057	1.3996175	2.162554	-3.4735245	-0.5662795	2.1069218	2.9611003	2.9544242	0.4223289	0

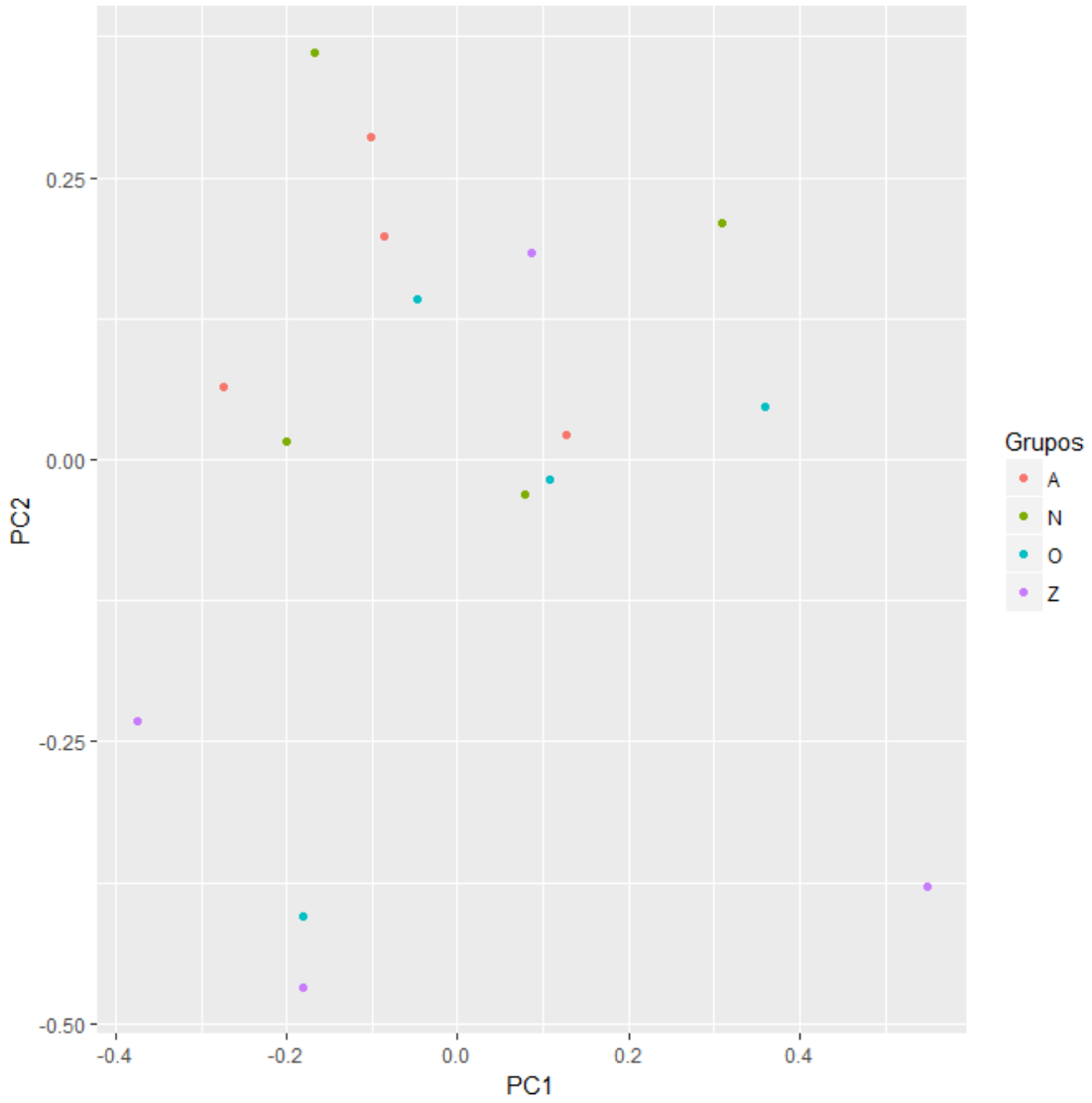
Gráficamos los pesos de los componenetes sobre la varianzas y vemos que el primer componente supone más del 70 por ciento de la varianza entre muestras:

```
> plot(pca, type = "l")
```



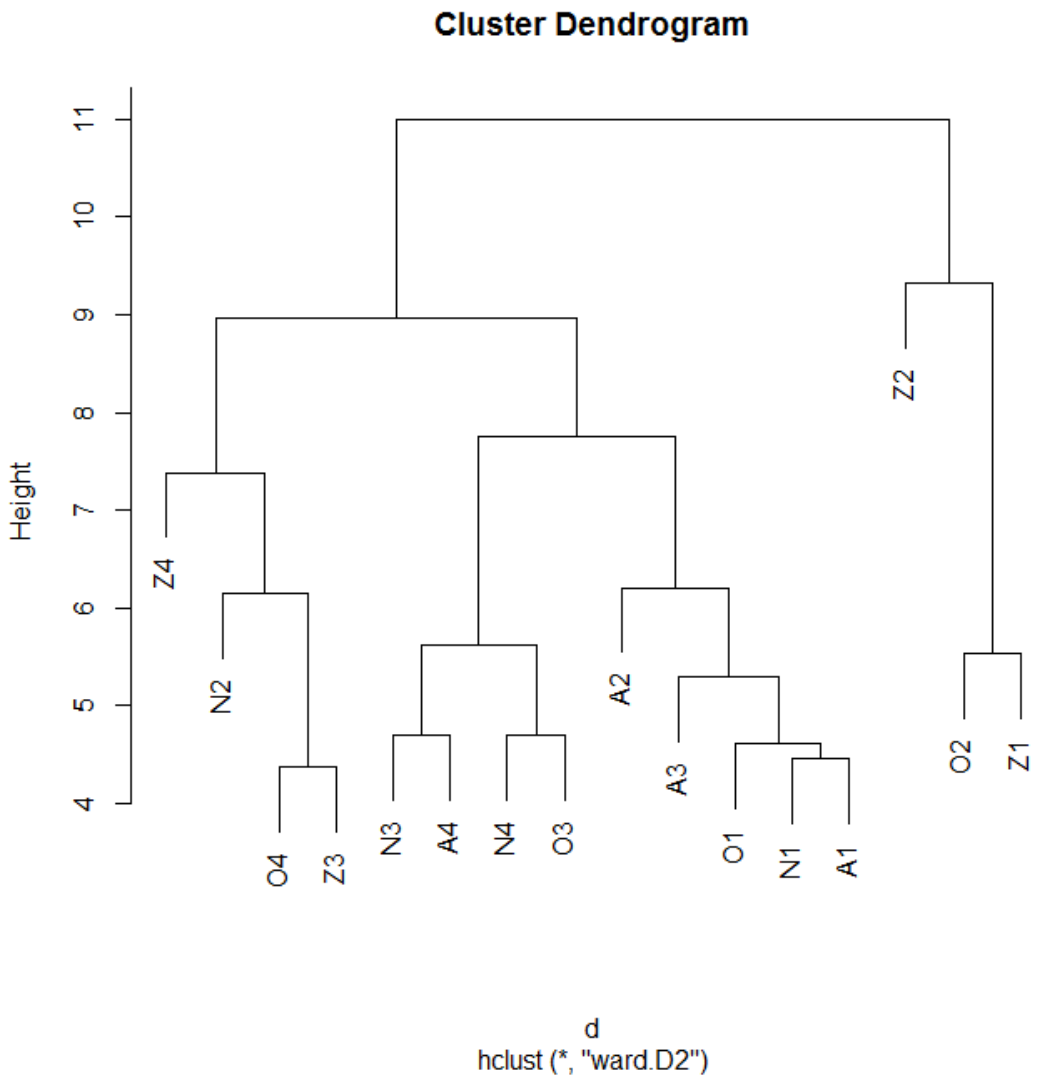
Mostramos una gráfica de distribución de las muestras en función del primer y segundo componente:

```
> autoplot(pca, data = raw.work.t, colour = "Grupos")
```



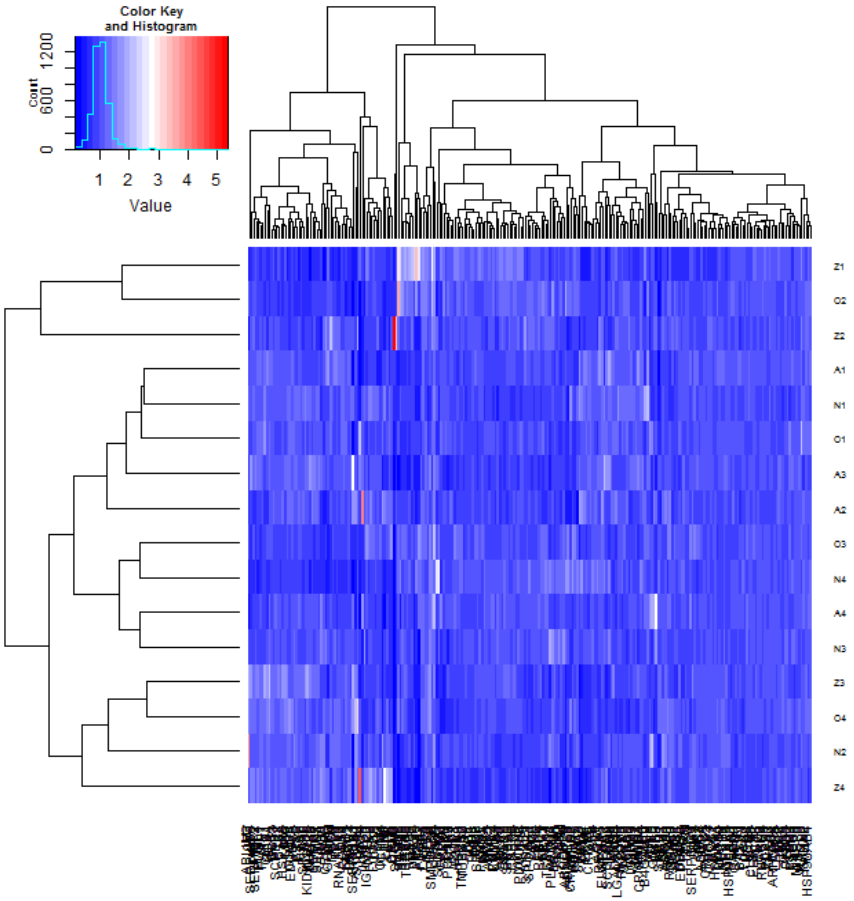
Dendograma

```
> ##  plot(pca$x[c(1:16),3:5],col=raw.work.t[,1])

> d <- dist(raw.work.t[, c(-1, -251)], method = "euclidean", diag = FALSE,
+       upper = TRUE)
> c1 <- hclust(d, method = "ward.D2", members = NULL)
> plot(c1)
```



Heatmap

```
> my_palette <- colorRampPalette(c("blue","white","red"))(n = 25)
>
> d2 <- dist(t(raw.work.t[,c(-1,-251)]),method = "euclidean", diag = FALSE, upper
= TRUE)
> c2 <- hclust(d2, method = "ward.D2", members = NULL)
>
> # Plot heatmap with heatmap.2
> par(cex.main=0.75)

> heatmap.2(as.matrix(raw.work.t[,c(-1,-251)]),
+           Rowv=as.dendrogram(c1),
+           Colv = as.dendrogram(c2),
+           density.info="histogram",
+           trace="none",
+           col = my_palette
+           cexRow=0.5,cexCol=0.75
```



ANNEX B - TESTS CONVENCIONALS: MOSTRA NORMALITZADA

HIGH-THROUGHPUT PROTEOMICS DETECTS DIFFERENTIALLY EXPRESSED SEMINAL PLASMA PROTEINS CORRELATED TO SPERM PARAMETERS

TEST EN MUESTRA NORMALIZADA (DATOS ESTANDARIZADOS Y NORMALIZADOS)

David Delgado Dueñas

Noviembre 2017

Contenido

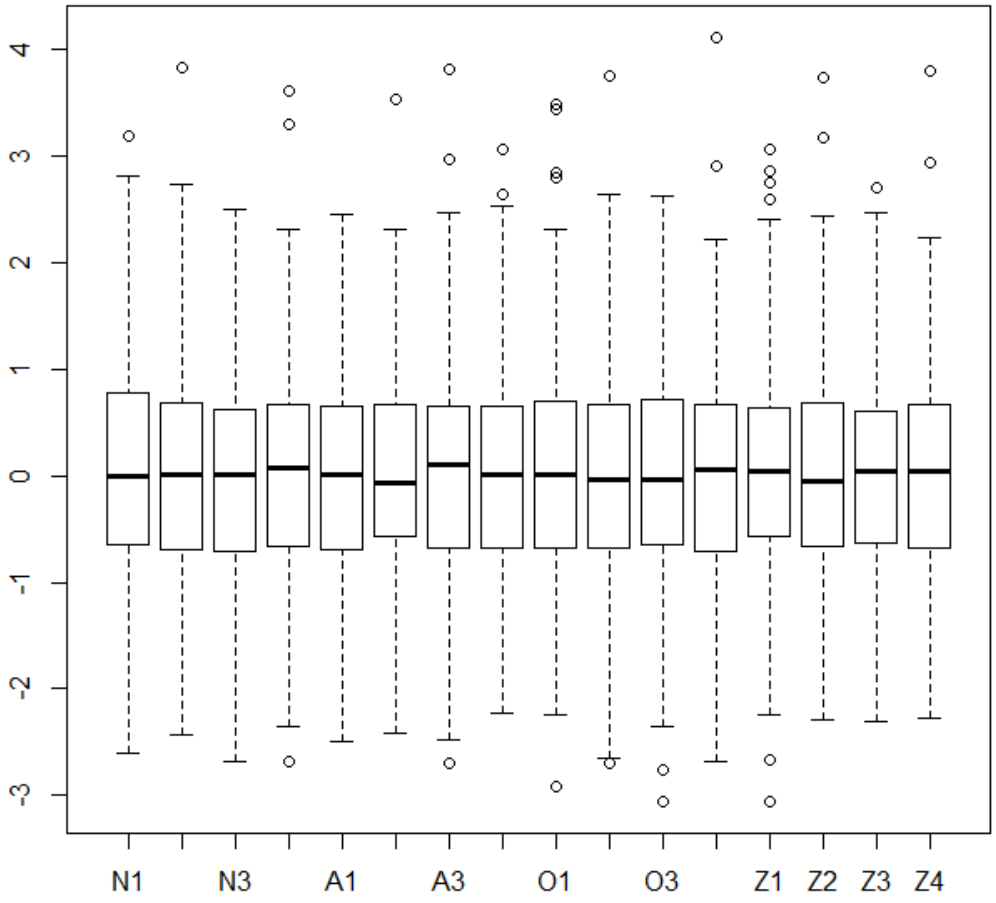
Gestión de los datos

Importamos las tablas y las adecuamos a nuestras necesidades de análisis:

```
# Importamos tabal Excel con datos sobre cuantificación de
> # proteínas
> PcaTotalG <- readXL("Datos/Libro1.xlsx", rownames = TRUE, header = TRUE,
+         na = "NA", sheet = "Hoja2", stringsAsFactors = FALSE)
> colnames(PcaTotalG) <- c("Unique Peptides", "PSMs", "N1", "N2",
+         "N3", "N4", "A1", "A2", "A3", "A4", "O1", "O2", "O3", "O4",
+         "Z1", "Z2", "Z3", "Z4")
> names <- row.names(PcaTotalG)
> E <- PcaTotalG[, c(-1, -2)]
> E <- E[complete.cases(E), ]
>
>
> PcaTotal <- readXL("Datos/Libro1.xlsx", rownames = FALSE, header = TRUE,
+         na = "NA", sheet = "Hoja1", stringsAsFactors = TRUE)
> PcaTotalF <- PcaTotal[2:10]
> PcaTotalF <- rbind('Unique Peptides' = NA, PcaTotalF)
> PcaTotalF <- rbind(PSMs = NA, PcaTotalF)
> rownames(PcaTotalF) <- c("Unique Peptides", "PSMs", "N1", "N2",
+         "N3", "N4", "A1", "A2", "A3", "A4", "O1", "O2", "O3", "O4",
+         "Z1", "Z2", "Z3", "Z4")
```

Como forma de exploración se aplica un método de normalización basado en la Transformación de Johnson. Este método ya de por sí puede conllevar la atenuación de la señal diferencial, deberemos por ello ser especialmente cuidadosos en el contraste de los resultados obtenidos.

```
## Normalizamos mediante la transformación de Johnson
> john.E <- JohnM(E)
> boxplot(john.E)
```



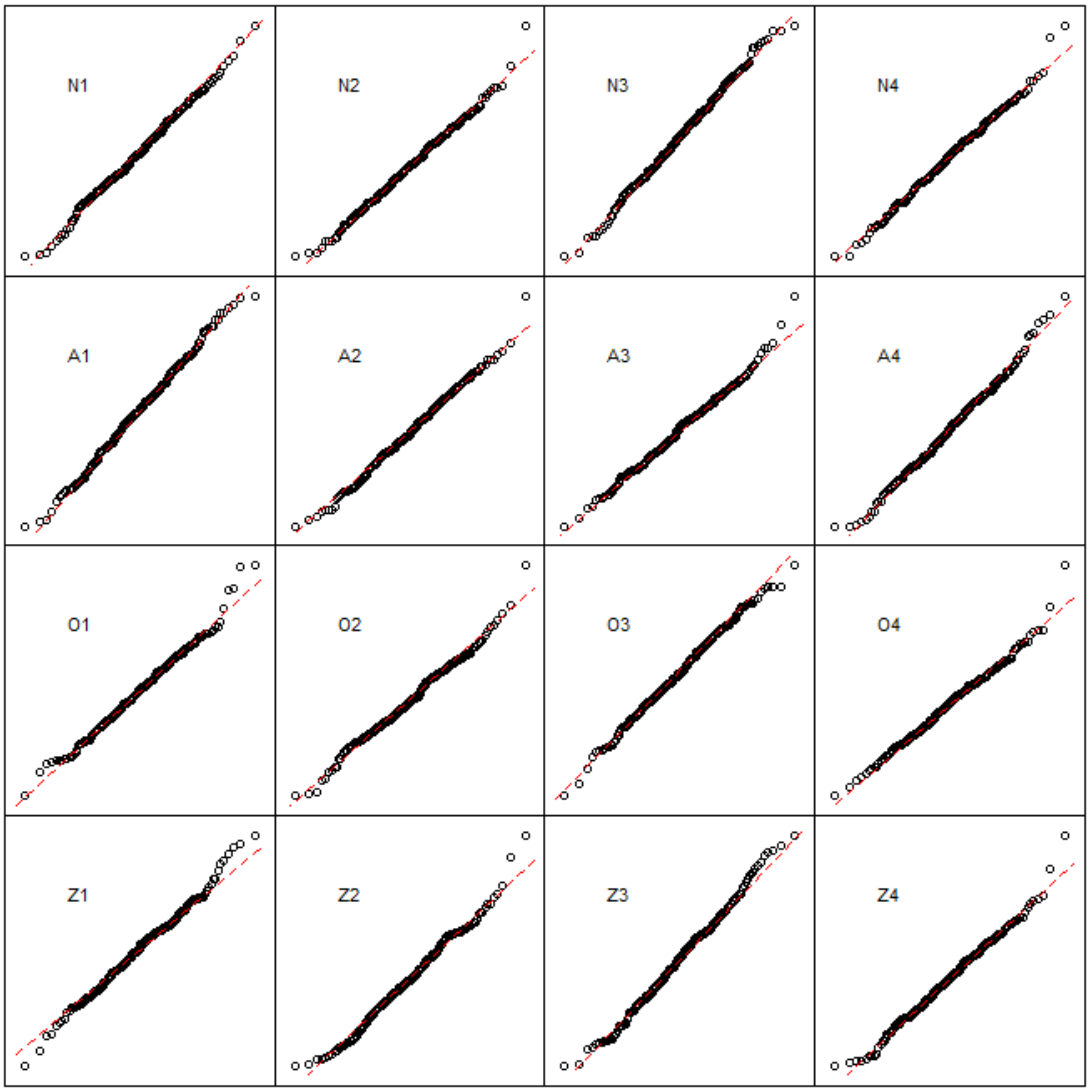
Comprobamos la Normalidad de la expresión proteica en las diferentes muestras. Se indica la proteína, el resultado del test de Shapiro-Wilk con su p-valor y Anderson-Darling con su porcentaje de certidumbre

```
Normal(john.E)
```

Proteina	S.W	pvalor	A.D	porcentaje
N1	Sí	0.8405	Sí	83.43337
N2	Sí	0.5094	Sí	95.68592
N3	Sí	0.8625	Sí	98.20236
N4	Sí	0.5243	Sí	79.00285
A1	Sí	0.8054	Sí	92.15532
A2	Sí	0.7114	Sí	95.14029
A3	Sí	0.3545	No	46.18683
A4	Sí	0.3379	Sí	59.94350
O1	Sí	0.1989	Sí	85.92962
O2	Sí	0.1704	No	23.29667
O3	Sí	0.8936	Sí	92.33742
O4	Sí	0.1692	Sí	53.73782
Z1	Sí	0.4520	No	31.95319
Z2	Sí	0.0923	No	37.92992
Z3	Sí	0.5213	Sí	66.46257
Z4	Sí	0.2197	Sí	86.12870

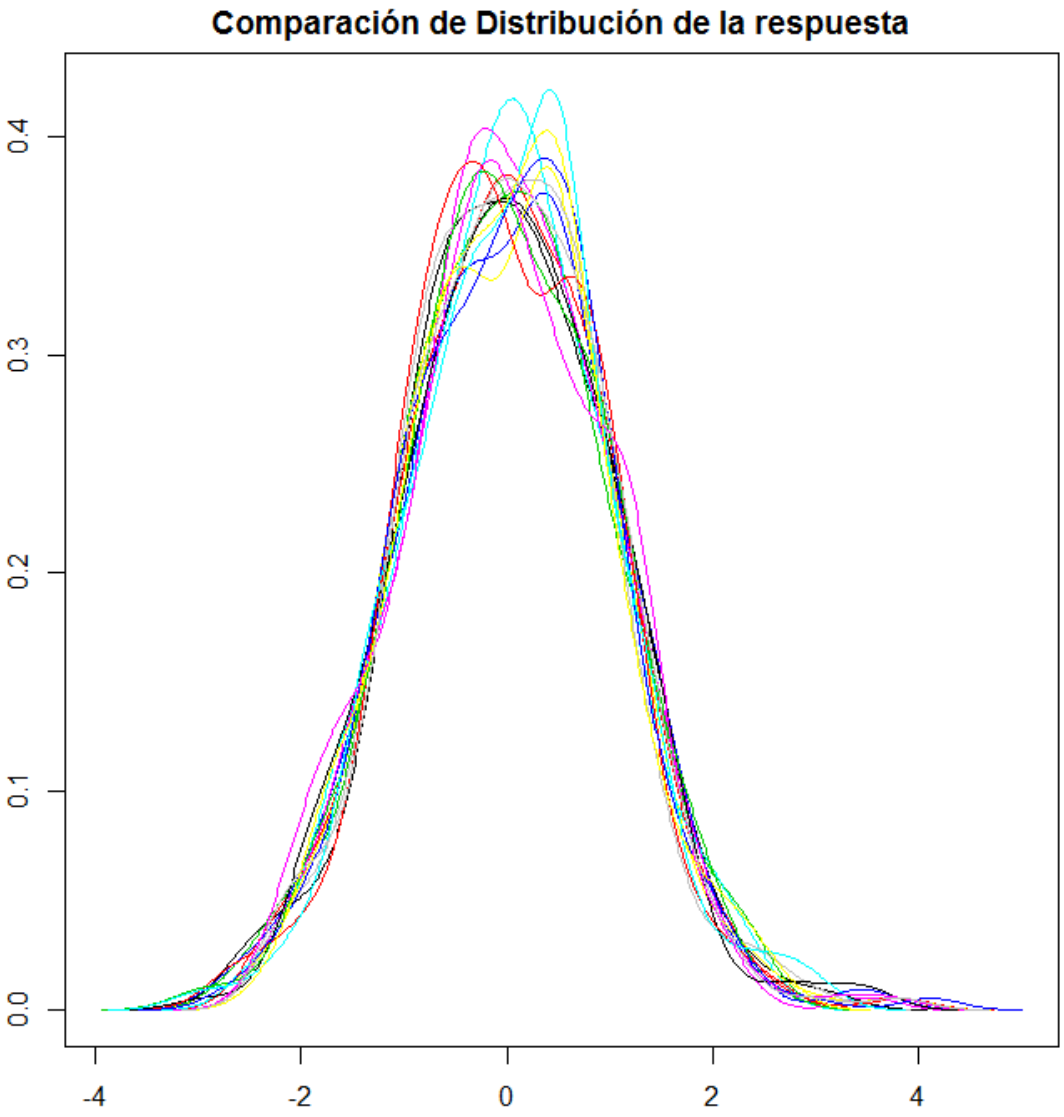
Se muestra la distribución de las luminancias de las proteínas presentes en las muestras.

```
qqploted(john.E)
```



Se compara las respuestas de las diferentes muestras.

```
densited(john.E)
```



Test de ANOVA o Kruskal-Wallis para determinar significancia en la varianza entre grupos

Según si pasa el test Shapiro de normalidad para la distribución de muestra de cada proteína, usaremos el test paramétrico o no paramétrico para realizar el descubrimiento de proteínas con cambios significativos entre grupos.

```
## ANOVA
>
> names <- row.names(john.E)
> work.t <- as.data.frame(t(john.E))
> colnames(work.t) <- names
> work.t$myfactor <- factor(row.names(work.t))
>
> Grupos <- c("N", "N", "N", "N", "A", "A", "A", "A", "O", "O",
+         "O", "O", "Z", "Z", "Z", "Z")
> work.t <- cbind(Grupos, work.t)
>
> SignificantProtein <- data.frame(Test = character(), Proteina = character(),
+         P.valor = character(), stringsAsFactors = FALSE)    #Creamos una tabla vacía
> NotSignificant <- data.frame(Test = character(), Proteina = character(),
+         P.valor = character(), stringsAsFactors = FALSE)    #Creamos una tabla vacía
>
>
> # Empezamos un bucle for desde 2 hasta la última columna
> for (i in 2:(ncol(work.t) - 1)) {
+
+         if (shapiro.test(work.t[, i])$p.value >= 0.05) {
+
+                 # Realizamos el test
+                 Alpha <- oneway.test(work.t[, i] ~ Grupos, work.t)
+                 # Nos quedamos el p.valor
+                 P <- unlist(Alpha)["p.value"]
+                 Wii <- data.frame(Test = "AN", Proteina = colnames(work.t)[i],
+                         P.valor = round(as.numeric(P), digits = 3))
+                 # Si el p.valor es inferior a 0.05 anotamos la proteína y
+                 # mostramos su anova
+                 if (P < 0.05) {
+                         SignificantProtein <- rbind(SignificantProtein, Wii)
+                 } else {
+                         NotSignificant <- rbind(NotSignificant, Wii)
+                 }
+         } else {
+                 # Realizamos el test
+                 Alpha <- kruskal.test(work.t[, i] ~ Grupos, work.t)
+                 # Nos quedamos el p.valor
+                 P <- unlist(Alpha)["p.value"]
+                 Wii <- data.frame(Test = "KW", Proteina = colnames(work.t)[i],
+                         P.valor = round(as.numeric(P), digits = 3))
+                 # Si el p.valor es inferior a 0.05 anotamos la proteína y
+                 # mostramos su anova
+                 if (P < 0.05) {
+                         SignificantProtein <- rbind(SignificantProtein, Wii)
+                 } else {
+                         NotSignificant <- rbind(NotSignificant, Wii)
```

```
+         }
+
+     }
+ }
>
> # Imprimimos la tabla de Proteínas significativas con su
> # p-valor
> SignificantProtein
```

Test	Proteina	P.valor
AN	NPC2	0.019
AN	ANPEP	0.017
AN	SCGB1A1	0.013
AN	CRISP1	0.003
AN	EEF1A1	0.016
AN	CNDP2	0.026
AN	IGHG2	0.027
AN	PLA2G2A	0.000
AN	SPINT3	0.008
AN	LDHC	0.000
AN	LAMC1	0.018

```
head(NotSignificant[with(NotSignificant, order(P.valor)), ])
```

	Test	Proteina	P.valor
194	AN	TMPRSS2	0.053
148	AN	ACE	0.060
218	AN	CDH1	0.061
43	AN	EDDM3B	0.062
57	AN	PTGDS	0.066
224	AN	LAMA5	0.072

Test de Duncan

Se realiza el test de multicomparación de Duncan. El resultado nos indica el tipo perteneciente a cada grupo muestral.

```
# Post-Hoc en principio toca Duncan (consideramos
> # significante las diferencias observadas en anova)
> Duncan.work.t <- data.frame(Proteina = character(), type = character(),
+       groups = character(), stringsAsFactors = FALSE)
> Dun.G <- data.frame(Proteina = character(), type = character(),
+       groups = character(), stringsAsFactors = FALSE)
> Dun.T <- data.frame(Proteina = character(), vs = character(),
+       stringsAsFactors = FALSE)
> Duncan.work.tF <- data.frame(Proteina = character(), vs = character(),
+       stringsAsFactors = FALSE)
> Dun <- as.array(NA)
> sin <- as.vector(SignificantProtein[2])
>
> for (i in 1:nrow(sin)) {
+
+       a <- as.character(sin[i, 1])
+       work.tA <- aov(get(a) ~ Grupos, work.t)
+       at <- duncan.test(work.tA, "Grupos", alpha = 0.05)
+
+       for (j in 1:nrow(at$groups)) {
+             Dun = as.character(at$groups[j, 3])
+             Dun.G <- data.frame(Proteina = a, type = at$groups[j,
+                   1], groups = Dun)
+             Duncan.work.t <- rbind(Duncan.work.t, Dun.G)
+       }
+
+ }
>

Duncan.work.t
```

Proteina	type	groups
NPC2	A	a
NPC2	N	ab
NPC2	O	bc
NPC2	Z	c
ANPEP	O	a
ANPEP	N	a
ANPEP	Z	ab
ANPEP	A	b
SCGB1A1	A	a
SCGB1A1	Z	b

SCGB1A1	N	b
SCGB1A1	O	b
CRISP1	A	a
CRISP1	N	a
CRISP1	O	ab
CRISP1	Z	b
EEF1A1	O	a
EEF1A1	Z	a
EEF1A1	A	a
EEF1A1	N	a
CNDP2	Z	a
CNDP2	O	a
CNDP2	A	b
CNDP2	N	b
IGHG2	N	a
IGHG2	A	ab
IGHG2	Z	ab
IGHG2	O	b
PLA2G2A	Z	a
PLA2G2A	N	a
PLA2G2A	O	ab
PLA2G2A	A	b
SPINT3	A	a
SPINT3	N	a
SPINT3	O	ab
SPINT3	Z	b
LDHC	N	a
LDHC	A	ab
LDHC	O	bc
LDHC	Z	c
LAMC1	A	a
LAMC1	Z	a
LAMC1	O	ab
LAMC1	N	b

Tabla resumen de las diferencias observadas:

NPC2	A vs O, A vs Z, N vs Z
ANPEP	O vs A, N vs A
SCGB1A1	A vs Z, A vs N, A vs O
CRISP1	A vs Z, A vs N, A vs O
CNDP2	A vs Z, A vs O, N vs Z, N vs O
IGHG2	N vs O
PLA2G2A	A vs Z, A vs N
SPINT3	A vs Z, A vs N
LDHC	A vs Z, N vs Z, N vs O
LAMC1	A vs N, Z vs N

[TEST EXPLORATORIO] Test de Tukey Post-HOC

```
sin <- as.vector(SignificantProtein[2])
> Tukey.work.t <- data.frame(Proteina = character(), Vs = character(),
+         P.valor = character(), stringsAsFactors = FALSE)
>
>
> for (i in 1:nrow(sin)) {
+
+         a <- as.character(sin[i, 1])
+         work.tA <- aov(get(a) ~ Grupos, work.t)
+         at <- TukeyHSD(work.tA, "Grupos", conf.level = 0.95)
+
+         for (j in 1:nrow(at$Grupos)) {
+               P <- at$Grupos[j, 4]
+               if (P < 0.05) {
+                     Wii <- data.frame(Proteina = a, Vs = row.names(at$Grupos)[j],
+                           P.valor = round(as.numeric(P), digits = 3))
+                     Tukey.work.t <- rbind(Tukey.work.t, Wii)
+               }
+         }
+ }
>
> Tukey.work.t
```

Proteina	Vs	P.valor
NPC2	Z-A	0.008
NPC2	Z-N	0.043
SCGB1A1	O-A	0.049
CRISP1	Z-A	0.023
PLA2G2A	N-A	0.030
PLA2G2A	Z-A	0.018
SPINT3	Z-A	0.038
LDHC	Z-A	0.014
LDHC	Z-N	0.007
LAMC1	N-A	0.010
LAMC1	Z-N	0.014

Correlación concentración contra luminancia de las proteínas

```
CORRELACIÓN: C VS PROTS
>
> ## Normal
>
> C <- PcaTotalF[3:18, 3]
> Pear.C <- cbind(C, work.t[c(-1)])
>
> cor.c <- correlate(Pear.C, "pearson")
> head(cor.c[with(cor.c, order(corrA, decreasing = T)), ][1:3])
```

	Prot	sign	corr
cor	C	0.000	1.000
cor181	LDHC	0.001	0.753
cor138	SPINT3	0.002	0.714
cor10	NPC2	0.002	0.709
cor39	ECM1	0.002	0.709
cor235	LAMA5	0.003	-0.695

Correlación motilidad contra luminancia de las proteínas

```
> # CORRELACION: M VS PROTS
>
> ## normal
>
> M <- PcaTotalF[3:18, 2]
> Pear.M <- cbind(M, work.t[c(-1)])
>
> cor.m <- correlate(Pear.M, "pearson")
> head(cor.m[with(cor.m, order(corrA, decreasing = T)), ][1:3])
```

	Prot	sign	corr
cor	M	0.000	1.000
cor90	NUCB1	0.006	-0.654
cor154	LCP1	0.008	0.639
cor214	SPINT1	0.009	0.631
cor57	NUCB2	0.009	-0.630
cor1	SEMG2	0.018	-0.584

Corrección de la correlaciones

```
FDR
>
> cor.c.fdr <- pHCorr(cor.c, "fdr")
> head(cor.c.fdr[with(cor.c.fdr, order(corrA, decreasing = T)),
+         ][c(1, 2, 3, 5)], 10)
```

	Prot	sign	corr	FDR
cor	C	0.000	1.000	0.00000
cor181	LDHC	0.001	0.753	0.09375
cor138	SPINT3	0.002	0.714	0.09375
cor10	NPC2	0.002	0.709	0.09375
cor39	ECM1	0.002	0.709	0.09375
cor235	LAMA5	0.003	-0.695	0.09375
cor61	PTGDS	0.003	0.694	0.09375
cor96	PRDX2	0.003	-0.686	0.09375
cor93	CNDP2	0.006	-0.655	0.16667
cor140	SHISA5	0.009	-0.633	0.22500

```
> cor.m.fdr <- pHCorr(cor.m, "fdr")
> head(cor.m.fdr[with(cor.m.fdr, order(corrA, decreasing = T)),
+         ][c(1, 2, 3, 5)], 10)
```

	Prot	sign	corr	FDR
cor	M	0.000	1.000	0.00000
cor90	NUCB1	0.006	-0.654	0.45000
cor154	LCP1	0.008	0.639	0.45000
cor214	SPINT1	0.009	0.631	0.45000
cor57	NUCB2	0.009	-0.630	0.45000
cor1	SEMG2	0.018	-0.584	0.59091
cor225	FUCA2	0.018	0.582	0.59091
cor11	CST4	0.022	-0.565	0.59091
cor237	IGF2R	0.023	0.564	0.59091
cor31	ORM1	0.024	0.559	0.59091

PCA

Realizamos el estudio de los principales componentes (PCA):

```
pca <- prcomp(work.t[, c(-1, -251)], center = TRUE, scale. = TRUE)
```

Sumarizamos los resultados del estudio:

```
summary(pca)

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     7.953 5.1791 4.80620 4.32556 4.27874 3.87025
Proportion of Variance 0.254 0.1077 0.09277 0.07514 0.07352 0.06016
Cumulative Proportion  0.254 0.3617 0.45450 0.52965 0.60317 0.66333
                           PC7    PC8    PC9    PC10    PC11    PC12
Standard deviation     3.77982 3.5948 3.3436 3.15390 3.07056 2.99018
Proportion of Variance 0.05738 0.0519 0.0449 0.03995 0.03786 0.03591
Cumulative Proportion  0.72070 0.7726 0.8175 0.85745 0.89531 0.93122
                          PC13    PC14    PC15     PC16
Standard deviation     2.56908 2.39637 2.18706 2.19e-15
Proportion of Variance 0.02651 0.02306 0.01921 0.00e+00
Cumulative Proportion  0.95773 0.98079 1.00000 1.00e+00
```

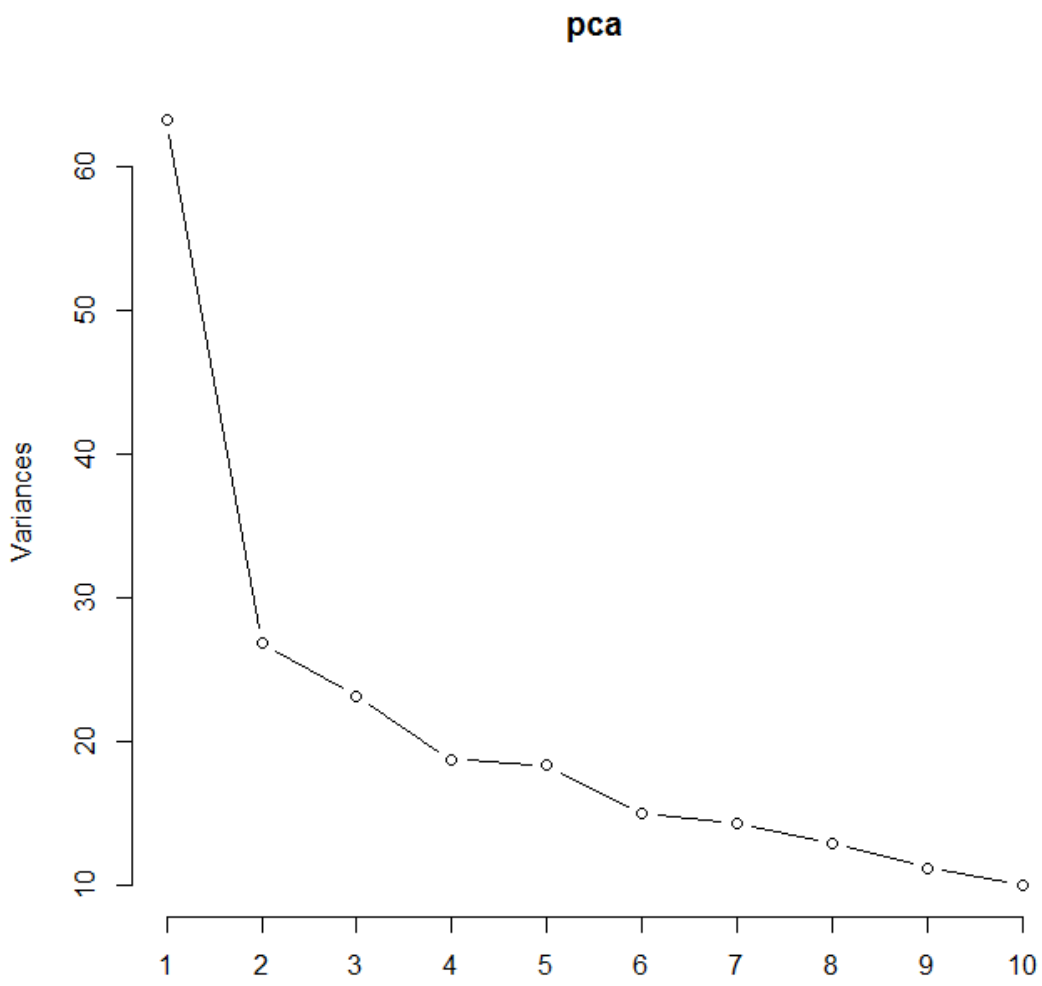
Mostramos las revoluciones:

```
pca$x
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
N1	-7.212648	8.9084603	-3.4918826	0.4608162	-4.113192	1.3930215	-0.1339487	1.4590585	-5.4504566	-0.8140522	-3.7852396	-3.2614002	1.6769691	-3.6996245	-1.1566599	0
N2	-8.071587	-0.8017344	2.9485648	-4.8781179	-4.928719	-3.6930461	6.4417296	0.7221559	0.9519538	-2.9386090	5.8186387	-2.3852656	1.1692731	1.2441072	-1.5845753	0
N3	4.555193	-6.3990774	3.6884146	9.8513126	-3.049666	-2.2551866	5.9836045	2.0602682	-1.9446326	-1.2825009	-4.0733136	0.8169105	-1.2047715	0.9735595	0.3590914	0
N4	11.796590	1.5543550	-3.8584831	4.0426851	-1.560951	-3.2587598	-2.9634781	0.7275802	5.0313194	0.2064511	1.6200564	-4.5811007	3.7598468	-0.3632726	2.8495766	0
A1	-4.485334	5.9441517	3.8436967	4.8749508	2.067755	-2.7538834	-5.9192269	1.5710502	0.7885145	-5.5591564	3.0544608	4.5444260	-1.7118769	-1.1210073	-0.2138761	0
A2	-3.762558	5.0230079	-4.7445127	-1.1781137	0.685015	6.9560592	3.8863825	5.3246864	6.1541380	-0.3071944	-1.7210974	1.5856596	-1.9125627	1.8667103	0.9498307	0
A3	-9.491231	-0.2395254	-3.5677679	5.5623057	1.434513	-0.7699001	-3.2785491	0.2238701	-1.3288213	6.5570187	2.3849975	-1.3680308	-0.7275991	3.7724797	-2.8279069	0
A4	5.646263	-0.7195277	-1.2377028	3.8789928	4.868627	7.4627571	4.0675241	-6.0992200	-1.3065658	-0.6791225	4.1194826	0.9211298	1.8646241	-2.2264343	-1.0045599	0
O1	-3.511728	6.4052478	3.5774140	-2.6657547	8.643463	-5.0209299	3.0362658	-4.7175771	0.0424490	0.6165755	-3.8929871	-0.3199716	1.1975714	2.1738635	1.6588632	0
O2	12.226745	1.8032199	3.6048426	-3.1009132	1.991917	-0.5914366	0.0281410	0.9654081	-0.6261950	2.1595494	1.2617025	-4.0151864	6.7342879	-2.0786864	-0.6544440	0
O3	5.379729	-1.5423905	-9.5107353	-3.1525253	-5.190217	-1.7324498	-1.9265237	-6.6532677	-0.5950935	-3.1656469	-1.6975530	2.0363578	-2.5496307	2.2085984	-0.5752983	0
O4	-5.351095	-9.1298893	2.0829425	-1.9909306	2.736589	0.4013658	-3.2453831	-0.5978209	5.3584334	-1.3318467	-4.0691978	-1.5532102	1.1147958	-2.4541167	-3.7296065	0
Z1	12.416773	-1.5980138	-2.7600078	-5.1194291	3.136072	-2.4252138	0.1472341	6.2689721	-3.4113669	1.4246589	0.2593706	4.5510538	3.1796996	0.1519860	-1.8584461	0
Z2	4.690029	2.8331313	10.6121495	-2.2998554	-7.467257	4.6006920	-3.3128369	-2.3239158	0.5552885	3.1264149	-1.0536203	2.0774989	1.7885932	1.7089471	0.2030119	0
Z3	-5.181989	-7.5604509	0.8284775	-2.7803668	2.841062	4.3038850	-3.8914264	2.1556173	-4.9125105	-3.0773845	0.6276288	-2.8401411	-0.0149151	2.1034261	3.8766311	0
Z4	-9.643152	-4.4809645	-2.0154101	-1.5050564	-2.095012	-2.6169745	1.0804912	-1.0868655	0.6935455	5.0648450	1.1466708	3.7912701	-0.8957293	-4.3605361	3.7083682	0

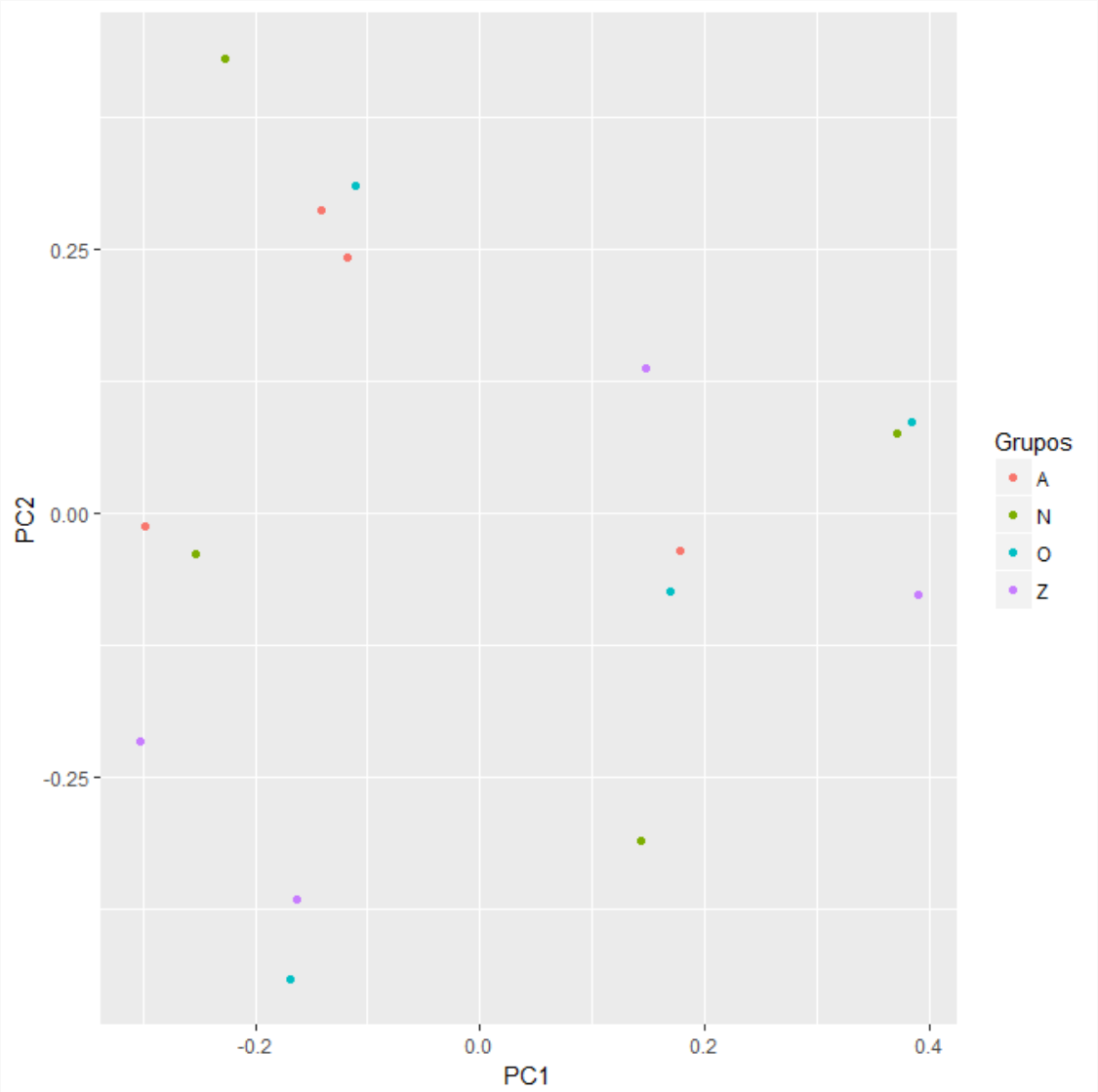
Graficamos los pesos de los componentes sobre la varianzas y vemos que el primer componente supone más del 60 por ciento de la varianza entre muestras:

```
plot(pca, type = "l")
```



Mostramos una gráfica de distribución de las muestras en función del primer y segundo componente:

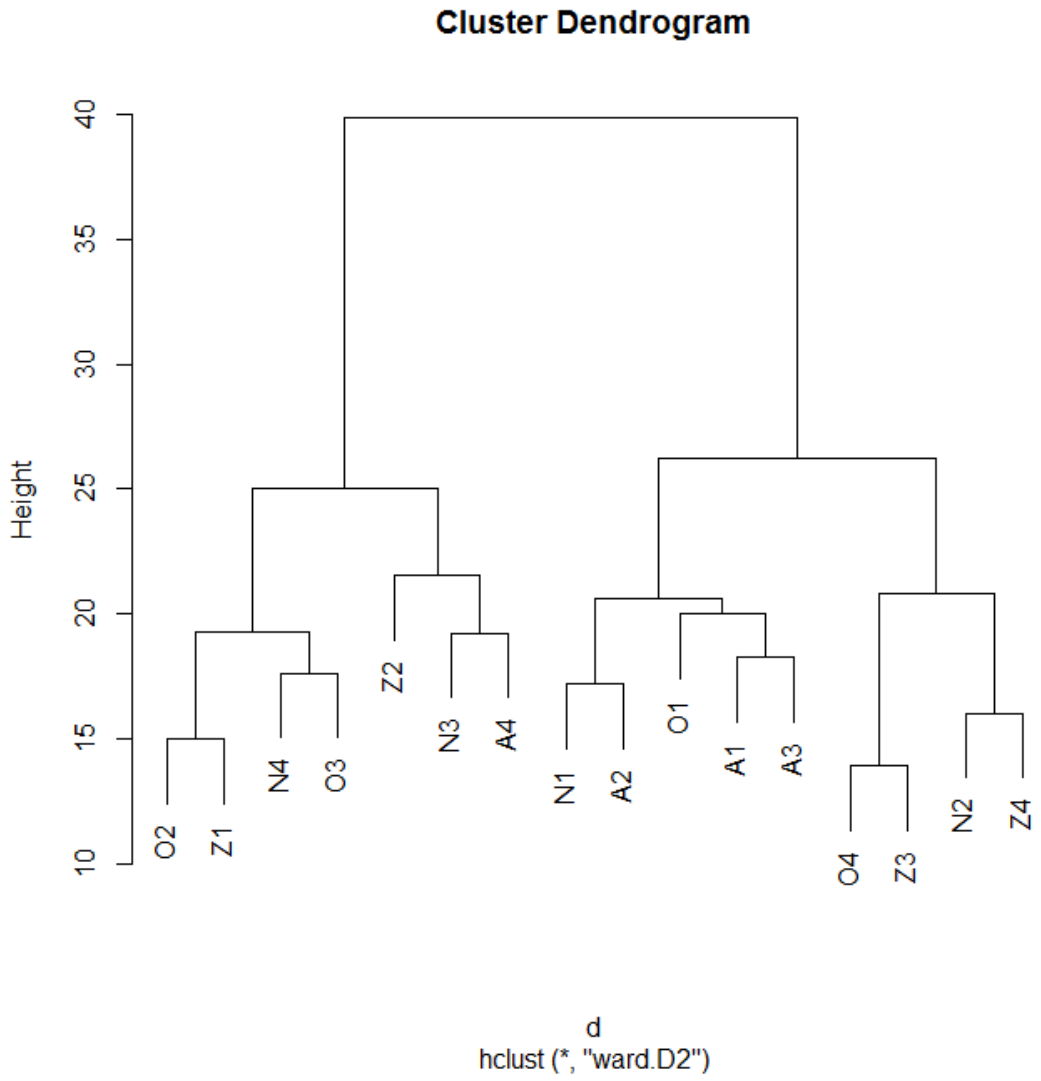
```
autoplot(pca, data = work.t, colour = "Grupos")
```



No aparece ninguna distribución de muestras que las estratifique.

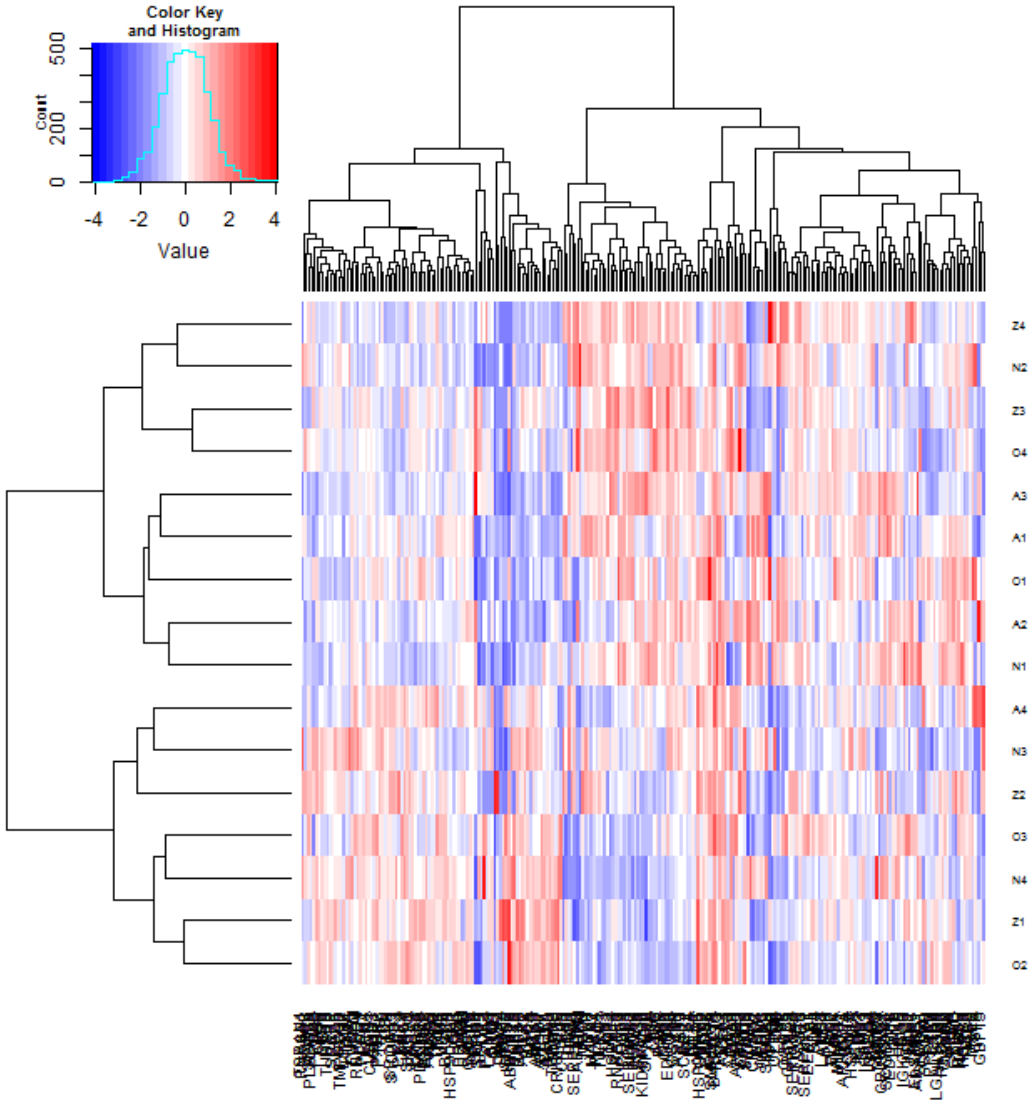
Dendograma

```
d <- dist(work.t[,c(-1,-251)], method = “euclidean”, diag = FALSE, upper = TRUE)
> c1 <- hclust(d, method = “ward.D2”, members = NULL)
> plot(c1)
```



Heatmap

```
d2 <- dist(t(work.t[,c(-1,-251)]),method = "euclidean", diag = FALSE, upper = TRUE)
> c2 <- hclust(d2, method = "ward.D2", members = NULL)> my_palette <-
colorRampPalette(c("blue","white","red"))(n = 25)
> par(cex.main=0.75)
> heatmap.2(as.matrix(work.t[,c(-1,-251)]), Rowv=as.dendrogram(c1),   Colv =
as.dendrogram(c2),
+                density.info="histogram",   trace="none,   col = my_palette,
+                cexRow=0.5,cexCol=0.75)
```



ANNEX C- RESULTAT D'ANÀLISI NOU

Proteïnes amb alta correlació amb més d'un pèptid implicat en pacients normozoospèrmics.

Protein a1	Protein a2	Percent	PeptidosTotales
BASP1	YWHAE	100	2 X 2
BASP1	ALDOA	100	2 X 3
BASP1	ECM1	100	2 X 11
BASP1	PKM	100	2 X 3
BASP1	SCPEP1	100	2 X 3
BASP1	SOD3	100	2 X 2
BASP1	AOC1	100	2 X 2
BASP1	GAA	100	2 X 2
BASP1	TF	100	2 X 14
BASP1	CD9	100	2 X 2
BASP1	LGALS3BP	100	2 X 10
BASP1	PARK7	100	2 X 3
BASP1	PTGDS	100	2 X 2
BASP1	CST3	100	2 X 5
BASP1	APOH	100	2 X 2
BASP1	LSAMP	100	2 X 4
BASP1	S100A11	100	2 X 2
BASP1	SPOCK1	100	2 X 3
BASP1	APLP2	100	2 X 2
BASP1	HSPA8	100	2 X 2
BASP1	FSTL1	100	2 X 3
BASP1	QSOX1	100	2 X 5
BASP1	PATE1	100	2 X 4
BASP1	SOD1	100	2 X 2
BASP1	YWHAZ	100	2 X 2
BASP1	LDHA	100	2 X 2
BASP1	FUCA1	100	2 X 3
BASP1	ANXA5	100	2 X 7
BASP1	IGKC	100	2 X 5
BASP1	CNDP2	100	2 X 5
BASP1	HPX	100	2 X 4
BASP1	ANXA3	100	2 X 3
BASP1	ENO1	100	2 X 4
BASP1	ORM2	100	2 X 4
BASP1	SORD	100	2 X 4
BASP1	PRDX6	100	2 X 3
BASP1	ELSPBP1	100	2 X 4
BASP1	RNASE4	100	2 X 4
BASP1	GDI2	100	2 X 2
BASP1	EZR	100	2 X 2
BASP1	GAPDH	100	2 X 4
BASP1	ANXA1	100	2 X 3
BASP1	CRTAC1	100	2 X 4
BASP1	EDDM3B	100	2 X 3
BASP1	PSMA2	100	2 X 2
BASP1	GPI	100	2 X 2
BASP1	LCP1	100	2 X 2
BASP1	HSP90A1	100	2 X 2
BASP1	CTSL	100	2 X 2
BASP1	DBI	100	2 X 3
BASP1	IGHG1	100	2 X 2
BASP1	IGHG2	100	2 X 2
BASP1	TMPRSS2	100	2 X 2
BASP1	C3	100	2 X 2
YWHAE	ALDOA	100	2 X 3
YWHAE	ECM1	100	2 X 11
YWHAE	PKM	100	2 X 3
YWHAE	SCPEP1	100	2 X 3
YWHAE	SOD3	100	2 X 2
YWHAE	GAA	100	2 X 2
YWHAE	CD9	100	2 X 2
YWHAE	PARK7	100	2 X 3
YWHAE	LAMB2	100	2 X 5
YWHAE	PTGDS	100	2 X 2
YWHAE	CST3	100	2 X 5
YWHAE	LSAMP	100	2 X 4
YWHAE	SPOCK1	100	2 X 3
YWHAE	APLP2	100	2 X 2
YWHAE	HSPA8	100	2 X 2
YWHAE	SOD1	100	2 X 2
YWHAE	YWHAZ	100	2 X 2
YWHAE	FUCA1	100	2 X 3
YWHAE	ANXA5	100	2 X 7
YWHAE	IGKC	100	2 X 5
YWHAE	HPX	100	2 X 4

YWHAE	ANXA3	100	2 X 3
YWHAE	ENO1	100	2 X 4
YWHAE	PRDX6	100	2 X 3
YWHAE	ELSPBP1	100	2 X 4
YWHAE	RNASE4	100	2 X 4
YWHAE	GDI2	100	2 X 2
YWHAE	EZR	100	2 X 2
YWHAE	GAPDH	100	2 X 4
YWHAE	ANXA1	100	2 X 3
YWHAE	HEXA	100	2 X 3
YWHAE	CRTAC1	100	2 X 4
YWHAE	GPI	100	2 X 2
YWHAE	LCP1	100	2 X 2
YWHAE	CTSL	100	2 X 2
YWHAE	DBI	100	2 X 3
YWHAE	IGHG1	100	2 X 2
YWHAE	TMPRSS2	100	2 X 2
		96.969	
ALDOA	ECM1	7	3 X 11
ALDOA	PKM	100	3 X 3
ALDOA	SCPEP1	100	3 X 3
ALDOA	SOD3	100	3 X 2
ALDOA	AOC1	100	3 X 2
ALDOA	GAA	100	3 X 2
		97.619	
ALDOA	TF	05	3 X 14
ALDOA	CD9	100	3 X 2
	LGALS3	96.666	
ALDOA	BP	67	3 X 10
ALDOA	PARK7	100	3 X 3
ALDOA	PTGDS	100	3 X 2
ALDOA	APOH	100	3 X 2
ALDOA	LSAMP	100	3 X 4
ALDOA	S100A11	100	3 X 2
ALDOA	SPOCK1	100	3 X 3
ALDOA	APLP2	100	3 X 2
ALDOA	HSPA8	100	3 X 2
ALDOA	FSTL1	100	3 X 3
ALDOA	SOD1	100	3 X 2
ALDOA	YWHAZ	100	3 X 2
ALDOA	LDHA	100	3 X 2
ALDOA	ANXA5	100	3 X 7
ALDOA	IGKC	100	3 X 5
ALDOA	CNDP2	100	3 X 5

ALDOA	HPX	100	3 X 4
ALDOA	ANXA3	100	3 X 3
ALDOA	ENO1	100	3 X 4
ALDOA	ORM2	100	3 X 4
ALDOA	SORD	100	3 X 4
ALDOA	PRDX6	100	3 X 3
ALDOA	RNASE4	100	3 X 4
ALDOA	GDI2	100	3 X 2
		95.833	
ALDOA	TIMP1	33	3 X 8
ALDOA	EZR	100	3 X 2
ALDOA	GAPDH	100	3 X 4
ALDOA	ANXA1	100	3 X 3
ALDOA	EDDM3B	100	3 X 3
ALDOA	PSMA2	100	3 X 2
ALDOA	GPI	100	3 X 2
ALDOA	LCP1	100	3 X 2
ALDOA	HSP90A1	100	3 X 2
ALDOA	CTSL	100	3 X 2
ALDOA	DBI	100	3 X 3
ALDOA	IGHG1	100	3 X 2
ALDOA	IGHG2	100	3 X 2
ALDOA	TMPRSS2	100	3 X 2
ALDOA	C3	100	3 X 2
		96.428	
CPE	ANXA5	57	20 X 7
ECM1	PKM	100	11 X 3
ECM1	SOD3	100	11 X 2
ECM1	AOC1	100	11 X 2
ECM1	GAA	100	11 X 2
ECM1	CD9	100	11 X 2
ECM1	PARK7	100	11 X 3
ECM1	PTGDS	100	11 X 2
ECM1	LSAMP	100	11 X 4
ECM1	SPOCK1	100	11 X 3
ECM1	APLP2	100	11 X 2
ECM1	HSPA8	100	11 X 2
ECM1	SOD1	100	11 X 2
ECM1	YWHAZ	100	11 X 2
ECM1	ANXA5	100	11 X 7
ECM1	IGKC	100	11 X 5
		96.363	
ECM1	CNDP2	64	11 X 5
ECM1	HPX	100	11 X 4

ECM1	ANXA3	100	11 X 3
		97.727	
ECM1	ENO1	27	11 X 4
ECM1	PRDX6	100	11 X 3
	ELSPBP	97.727	
ECM1	1	27	11 X 4
		95.454	
ECM1	RNASE4	55	11 X 4
ECM1	GDI2	100	11 X 2
ECM1	EZR	100	11 X 2
ECM1	GAPDH	100	11 X 4
ECM1	ANXA1	100	11 X 3
	EDDM3		
ECM1	B	100	11 X 3
ECM1	GPI	100	11 X 2
ECM1	LCP1	100	11 X 2
ECM1	DBI	100	11 X 3
ECM1	IGHG1	100	11 X 2
ECM1	IGHG2	100	11 X 2
		95.454	
ECM1	C3	55	11 X 2
PKM	SOD3	100	3 X 2
PKM	AOC1	100	3 X 2
PKM	GAA	100	3 X 2
PKM	CD9	100	3 X 2
	LGALS3	96.666	
PKM	BP	67	3 X 10
PKM	PARK7	100	3 X 3
PKM	PTGDS	100	3 X 2
PKM	APOH	100	3 X 2
PKM	LSAMP	100	3 X 4
PKM	SPOCK1	100	3 X 3
PKM	APLP2	100	3 X 2
PKM	HSPA8	100	3 X 2
PKM	FSTL1	100	3 X 3
PKM	SOD1	100	3 X 2
PKM	YWHAZ	100	3 X 2
PKM	ANXA5	100	3 X 7
PKM	IGKC	100	3 X 5
PKM	HPX	100	3 X 4
PKM	ANXA3	100	3 X 3
PKM	ENO1	100	3 X 4
PKM	PRDX6	100	3 X 3
	ELSPBP		
PKM	1	100	3 X 4
PKM	GDI2	100	3 X 2
PKM	EZR	100	3 X 2

PKM	GAPDH	100	3 X 4
PKM	ANXA1	100	3 X 3
	EDDM3		
PKM	B	100	3 X 3
PKM	PSMA2	100	3 X 2
PKM	GPI	100	3 X 2
PKM	LCP1	100	3 X 2
	HSP90A		
PKM	A1	100	3 X 2
PKM	DBI	100	3 X 3
PKM	IGHG1	100	3 X 2
PKM	IGHG2	100	3 X 2
PKM	C3	100	3 X 2
SCPEP1	GAA	100	3 X 2
SCPEP1	CD9	100	3 X 2
SCPEP1	PARK7	100	3 X 3
SCPEP1	LSAMP	100	3 X 4
SCPEP1	SPOCK1	100	3 X 3
SCPEP1	APLP2	100	3 X 2
SCPEP1	SOD1	100	3 X 2
SCPEP1	YWHAZ	100	3 X 2
SCPEP1	ANXA5	100	3 X 7
SCPEP1	IGKC	100	3 X 5
SCPEP1	HPX	100	3 X 4
SCPEP1	ANXA3	100	3 X 3
SCPEP1	PRDX6	100	3 X 3
		95.833	
SCPEP1	TIMP1	33	3 X 8
SCPEP1	EZR	100	3 X 2
SCPEP1	GAPDH	100	3 X 4
SCPEP1	ANXA1	100	3 X 3
SCPEP1	GPI	100	3 X 2
SCPEP1	CTSL	100	3 X 2
MMP2	PTGDS	100	6 X 2
MMP2	IGHG2	100	6 X 2
SOD3	GAA	100	2 X 2
SOD3	CD9	100	2 X 2
SOD3	PARK7	100	2 X 3
SOD3	LSAMP	100	2 X 4
SOD3	SPOCK1	100	2 X 3
SOD3	APLP2	100	2 X 2
SOD3	SOD1	100	2 X 2
SOD3	YWHAZ	100	2 X 2
SOD3	ANXA5	100	2 X 7
SOD3	IGKC	100	2 X 5
SOD3	HPX	100	2 X 4

SOD3	ANXA3	100	2 X 3
SOD3	PRDX6	100	2 X 3
SOD3	RNASE4	100	2 X 4
SOD3	GDI2	100	2 X 2
SOD3	EZR	100	2 X 2
SOD3	GAPDH	100	2 X 4
SOD3	ANXA1	100	2 X 3
SOD3	GPI	100	2 X 2
SOD3	LCP1	100	2 X 2
SOD3	CTSL	100	2 X 2
AOC1	GAA	100	2 X 2
AOC1	TF	100	2 X 14
AOC1	TGM4	100	2 X 20
AOC1	PARK7	100	2 X 3
AOC1	PTGDS	100	2 X 2
AOC1	APOH	100	2 X 2
AOC1	S100A1	100	2 X 2
AOC1	SPOCK1	100	2 X 3
AOC1	HSPA8	100	2 X 2
AOC1	FSTL1	100	2 X 3
AOC1	YWHAZ	100	2 X 2
AOC1	LDHA	100	2 X 2
AOC1	CNDP2	100	2 X 5
AOC1	HPX	100	2 X 4
AOC1	ENO1	100	2 X 4
AOC1	PRDX6	100	2 X 3
AOC1	PEBP4	100	2 X 4
AOC1	GAPDH	100	2 X 4
AOC1	ANXA1	100	2 X 3
AOC1	EDDM3B	100	2 X 3
AOC1	PSMA2	100	2 X 2
AOC1	GPI	100	2 X 2
AOC1	LCP1	100	2 X 2
AOC1	HSP90A1	100	2 X 2
AOC1	DBI	100	2 X 3
AOC1	IGHG1	100	2 X 2
AOC1	IGHG2	100	2 X 2
AOC1	C3	100	2 X 2
		96.428	
GAA	TF	57	2 X 14
GAA	CD9	100	2 X 2
GAA	PARK7	100	2 X 3
GAA	PTGDS	100	2 X 2
GAA	CST3	100	2 X 5

GAA	APOH	100	2 X 2
GAA	LSAMP	100	2 X 4
GAA	S100A1	100	2 X 2
GAA	SPOCK1	100	2 X 3
GAA	APLP2	100	2 X 2
GAA	HSPA8	100	2 X 2
GAA	FSTL1	100	2 X 3
GAA	SOD1	100	2 X 2
GAA	YWHAZ	100	2 X 2
GAA	FUCA1	100	2 X 3
GAA	ANXA5	100	2 X 7
GAA	IGKC	100	2 X 5
GAA	CNDP2	100	2 X 5
GAA	HPX	100	2 X 4
GAA	ANXA3	100	2 X 3
GAA	ENO1	100	2 X 4
GAA	PRDX6	100	2 X 3
GAA	ELSPBP1	100	2 X 4
GAA	RNASE4	100	2 X 4
GAA	GDI2	100	2 X 2
GAA	EZR	100	2 X 2
GAA	GAPDH	100	2 X 4
GAA	ANXA1	100	2 X 3
GAA	CRTAC1	100	2 X 4
GAA	EDDM3B	100	2 X 3
GAA	PSMA2	100	2 X 2
GAA	GPI	100	2 X 2
GAA	LCP1	100	2 X 2
GAA	HSP90A1	100	2 X 2
GAA	DBI	100	2 X 3
GAA	IGHG1	100	2 X 2
GAA	IGHG2	100	2 X 2
GAA	C3	100	2 X 2
TF	PTGDS	100	14 X 2
	S100A1	96.428	
TF	1	57	14 X 2
TF	HSPA8	100	14 X 2
		98.571	
TF	CNDP2	43	14 X 5
TF	EDDM3B	100	14 X 3
TF	PSMA2	100	14 X 2
TF	HSP90A1	100	14 X 2

TF	IGHG1	100	14 X 2
TF	IGHG2	100	14 X 2
TF	C3	100	14 X 2
CD9	PARK7	100	2 X 3
CD9	LAMB2	100	2 X 5
CD9	PTGDS	100	2 X 2
CD9	CST3	100	2 X 5
CD9	APOH	100	2 X 2
CD9	LSAMP	100	2 X 4
CD9	SPOCK1	100	2 X 3
CD9	APLP2	100	2 X 2
CD9	HSPA8	100	2 X 2
CD9	SOD1	100	2 X 2
CD9	YWHAZ	100	2 X 2
CD9	FUCA1	100	2 X 3
CD9	ANXA5	100	2 X 7
CD9	IGKC	100	2 X 5
CD9	CNDP2	100	2 X 5
CD9	HPX	100	2 X 4
CD9	ANXA3	100	2 X 3
CD9	ENO1	100	2 X 4
CD9	PRDX6	100	2 X 3
CD9	ELSPBP1	100	2 X 4
CD9	PLA1A	100	2 X 4
CD9	RNASE4	100	2 X 4
CD9	GDI2	100	2 X 2
CD9	EZR	100	2 X 2
CD9	GAPDH	100	2 X 4
CD9	ANXA1	100	2 X 3
CD9	HEXA	100	2 X 3
CD9	CRTAC1	100	2 X 4
CD9	TWSG1	100	2 X 2
CD9	PSMA2	100	2 X 2
CD9	GPI	100	2 X 2
CD9	LCP1	100	2 X 2
CD9	HSP90AA1	100	2 X 2
CD9	CTSL	100	2 X 2
CD9	DBI	100	2 X 3
CD9	IGHG1	100	2 X 2
CD9	TMPRSS2	100	2 X 2
CD9	C3	100	2 X 2
TGM4	PTGDS	100	20 X 2
TGM4	IGHG2	97.5	20 X 2
TGM4	C3	100	20 X 2
LGALS3BP	PARK7	100	10 X 3
LGALS3		96.666	
LGALS3BP	SPOCK1	67	10 X 3
LGALS3BP	YWHAZ	100	10 X 2
LGALS3		97.142	
LGALS3BP	ANXA5	86	10 X 7
LGALS3BP	IGKC	98	10 X 5
LGALS3BP	HPX	100	10 X 4
PARK7	PTGDS	100	3 X 2
PARK7	APOH	100	3 X 2
PARK7	LSAMP	100	3 X 4
PARK7	S100A11	100	3 X 2
PARK7	SPOCK1	100	3 X 3
PARK7	APLP2	100	3 X 2
PARK7	HSPA8	100	3 X 2
PARK7	FSTL1	100	3 X 3
PARK7	QSOX1	100	3 X 5
PARK7	SOD1	100	3 X 2
PARK7	YWHAZ	100	3 X 2
PARK7	LDHA	100	3 X 2
PARK7	ANXA5	100	3 X 7
PARK7	IGKC	100	3 X 5
PARK7	CNDP2	100	3 X 5
PARK7	HPX	100	3 X 4
PARK7	ANXA3	100	3 X 3
PARK7	ENO1	100	3 X 4
PARK7	SORD	100	3 X 4
PARK7	PRDX6	100	3 X 3
PARK7	ELSPBP1	100	3 X 4
PARK7	RNASE4	100	3 X 4
PARK7	GDI2	100	3 X 2
PARK7	EZR	100	3 X 2
PARK7	GAPDH	100	3 X 4
PARK7	ANXA1	100	3 X 3
PARK7	EDDM3B	100	3 X 3
PARK7	PSMA2	100	3 X 2
PARK7	GPI	100	3 X 2
PARK7	LCP1	100	3 X 2
PARK7	HSP90AA1	100	3 X 2

PARK7	CTSL	100	3 X 2
PARK7	DBI	100	3 X 3
PARK7	IGHG1	100	3 X 2
PARK7	IGHG2	100	3 X 2
PARK7	TMPRSS2	100	3 X 2
PARK7	C3	100	3 X 2
MME	YWHAZ	97.058	17 X 2
MME	HPX	82	17 X 4
MME	EDDM3B	97.058	17 X 3
MME	HSP90AA1	96.078	17 X 2
LAMB2	SPOCK1	43	5 X 3
LAMB2	ANXA3	100	5 X 3
LAMB2	PRDX6	100	5 X 3
LAMB2	GDI2	100	5 X 2
PTGDS	APOH	100	2 X 2
PTGDS	LSAMP	100	2 X 4
PTGDS	S100A11	100	2 X 2
PTGDS	SPOCK1	100	2 X 3
PTGDS	HSPA8	100	2 X 2
PTGDS	FSTL1	100	2 X 3
PTGDS	SOD1	100	2 X 2
PTGDS	YWHAZ	100	2 X 2
PTGDS	LDHA	100	2 X 2
PTGDS	CNDP2	100	2 X 5
PTGDS	HPX	100	2 X 4
PTGDS	ENO1	100	2 X 4
PTGDS	ORM2	100	2 X 4
PTGDS	SORD	100	2 X 4
PTGDS	PRDX6	100	2 X 3
PTGDS	EZR	100	2 X 2
PTGDS	GAPDH	100	2 X 4
PTGDS	ANXA1	100	2 X 3
PTGDS	EDDM3B	100	2 X 3
PTGDS	PSMA2	100	2 X 2
PTGDS	GPI	100	2 X 2
PTGDS	LCP1	100	2 X 2
PTGDS	HSP90AA1	100	2 X 2
PTGDS	DBI	100	2 X 3
PTGDS	IGHG1	100	2 X 2
PTGDS	IGHG2	100	2 X 2

PTGDS	C3	100	2 X 2
CPM	PLA1A	100	2 X 4
CST3	LSAMP	100	5 X 4
CST3	SPOCK1	100	5 X 3
CST3	SOD1	100	5 X 2
CST3	IGKC	100	5 X 5
CST3	HPX	100	5 X 4
CST3	ANXA3	100	5 X 3
CST3	PRDX6	100	5 X 3
CST3	RNASE4	100	5 X 4
CST3	GDI2	100	5 X 2
CST3	ANXA1	100	5 X 3
CST3	CTSL	100	5 X 2
CST3	TMPRSS2	100	5 X 2
APOH	LSAMP	100	2 X 4
APOH	S100A11	100	2 X 2
APOH	SPOCK1	100	2 X 3
APOH	APLP2	100	2 X 2
APOH	HSPA8	100	2 X 2
APOH	FSTL1	100	2 X 3
APOH	SOD1	100	2 X 2
APOH	YWHAZ	100	2 X 2
APOH	LDHA	100	2 X 2
APOH	ANXA5	100	2 X 7
APOH	CNDP2	100	2 X 5
APOH	HPX	100	2 X 4
APOH	ANXA3	100	2 X 3
APOH	ENO1	100	2 X 4
APOH	SORD	100	2 X 4
APOH	PRDX6	100	2 X 3
APOH	TIMP1	100	2 X 8
APOH	EZR	100	2 X 2
APOH	GAPDH	100	2 X 4
APOH	ANXA1	100	2 X 3
APOH	PSMA2	100	2 X 2
APOH	GPI	100	2 X 2
APOH	LCP1	100	2 X 2
APOH	HSP90AA1	100	2 X 2
APOH	CTSL	100	2 X 2
APOH	DBI	100	2 X 3
APOH	IGHG1	100	2 X 2
APOH	C3	100	2 X 2
LSAMP	SPOCK1	100	4 X 3
LSAMP	APLP2	100	4 X 2

LSAMP	HSPA8	100	4 X 2
LSAMP	SOD1	100	4 X 2
LSAMP	YWHAZ	100	4 X 2
LSAMP	ANXA5	100	4 X 7
LSAMP	IGKC	100	4 X 5
LSAMP	CNDP2	100	4 X 5
LSAMP	HPX	100	4 X 4
LSAMP	ANXA3	100	4 X 3
LSAMP	ENO1	100	4 X 4
LSAMP	PRDX6	100	4 X 3
LSAMP	ELSPBP1	100	4 X 4
LSAMP	RNASE4	100	4 X 4
LSAMP	GDI2	100	4 X 2
LSAMP	TIMP1	96.875	4 X 8
LSAMP	EZR	100	4 X 2
LSAMP	GAPDH	100	4 X 4
LSAMP	ANXA1	100	4 X 3
LSAMP	CRTAC1	100	4 X 4
LSAMP	EDDM3B	100	4 X 3
LSAMP	PSMA2	100	4 X 2
LSAMP	GPI	100	4 X 2
LSAMP	LCP1	100	4 X 2
LSAMP	HSP90AA1	100	4 X 2
LSAMP	CTSL	100	4 X 2
LSAMP	DBI	100	4 X 3
LSAMP	IGHG1	100	4 X 2
LSAMP	TMPRSS2	100	4 X 2
S100A11	SPOCK1	100	2 X 3
S100A11	HSPA8	100	2 X 2
S100A11	FSTL1	100	2 X 3
S100A11	YWHAZ	100	2 X 2
S100A11	LDHA	100	2 X 2
S100A11	CNDP2	100	2 X 5
S100A11	HPX	100	2 X 4
S100A11	EDDM3B	100	2 X 3
S100A11	PSMA2	100	2 X 2
S100A11	HSP90AA1	100	2 X 2
S100A11	IGHG1	100	2 X 2
S100A11	IGHG2	100	2 X 2
S100A11	C3	100	2 X 2
SPOCK1	APLP2	100	3 X 2
SPOCK1	HSPA8	100	3 X 2
SPOCK1	FSTL1	100	3 X 3
SPOCK1	PATE1	100	3 X 4
SPOCK1	SOD1	100	3 X 2
SPOCK1	YWHAZ	100	3 X 2
SPOCK1	FUCA1	100	3 X 3
SPOCK1	ANXA5	100	3 X 7
SPOCK1	IGKC	100	3 X 5
SPOCK1	CNDP2	100	3 X 5
SPOCK1	HPX	100	3 X 4
SPOCK1	ANXA3	100	3 X 3
SPOCK1	ENO1	100	3 X 4
SPOCK1	PRDX6	100	3 X 3
SPOCK1	ELSPBP1	100	3 X 4
SPOCK1	RNASE4	100	3 X 4
SPOCK1	GDI2	100	3 X 2
SPOCK1	EZR	100	3 X 2
SPOCK1	GAPDH	100	3 X 4
SPOCK1	ANXA1	100	3 X 3
SPOCK1	CRTAC1	100	3 X 4
SPOCK1	EDDM3B	100	3 X 3
SPOCK1	PSMA2	100	3 X 2
SPOCK1	GPI	100	3 X 2
SPOCK1	LCP1	100	3 X 2
SPOCK1	HSP90AA1	100	3 X 2
SPOCK1	CTSL	100	3 X 2
SPOCK1	DBI	100	3 X 3
SPOCK1	IGHG1	100	3 X 2
SPOCK1	IGHG2	100	3 X 2
SPOCK1	TMPRSS2	100	3 X 2
SPOCK1	C3	100	3 X 2
APLP2	HSPA8	100	2 X 2

APLP2	SOD1	100	2 X 2
APLP2	YWHAZ	100	2 X 2
APLP2	ANXA5	100	2 X 7
APLP2	IGKC	100	2 X 5
APLP2	HPX	100	2 X 4
APLP2	ANXA3	100	2 X 3
APLP2	PRDX6	100	2 X 3
APLP2	ELSPBP1	100	2 X 4
APLP2	RNASE4	100	2 X 4
APLP2	GDI2	100	2 X 2
APLP2	EZR	100	2 X 2
APLP2	GAPDH	100	2 X 4
APLP2	ANXA1	100	2 X 3
APLP2	GPI	100	2 X 2
APLP2	LCP1	100	2 X 2
APLP2	HSP90AA1	100	2 X 2
APLP2	CTSL	100	2 X 2
APLP2	DBI	100	2 X 3
APLP2	IGHG1	100	2 X 2
HSPA8	FSTL1	100	2 X 3
HSPA8	SOD1	100	2 X 2
HSPA8	YWHAZ	100	2 X 2
HSPA8	LDHA	100	2 X 2
HSPA8	ANXA5	100	2 X 7
HSPA8	CNDP2	100	2 X 5
HSPA8	HPX	100	2 X 4
HSPA8	ANXA3	100	2 X 3
HSPA8	ENO1	100	2 X 4
HSPA8	ORM2	100	2 X 4
HSPA8	SORD	100	2 X 4
HSPA8	PRDX6	100	2 X 3
HSPA8	PEBP4	100	2 X 4
HSPA8	EZR	100	2 X 2
HSPA8	GAPDH	100	2 X 4
HSPA8	ANXA1	100	2 X 3
HSPA8	EDDM3B	100	2 X 3
HSPA8	PSMA2	100	2 X 2
HSPA8	GPI	100	2 X 2
HSPA8	LCP1	100	2 X 2
HSPA8	HSP90AA1	100	2 X 2
HSPA8	DBI	100	2 X 3
HSPA8	IGHG1	100	2 X 2
HSPA8	IGHG2	100	2 X 2
HSPA8	C3	100	2 X 2
FSTL1	YWHAZ	100	3 X 2
FSTL1	LDHA	100	3 X 2
FSTL1	CNDP2	100	3 X 5
FSTL1	HPX	100	3 X 4
		95.833	
FSTL1	TIMP1	33	3 X 8
FSTL1	EDDM3B	100	3 X 3
FSTL1	PSMA2	100	3 X 2
FSTL1	LCP1	100	3 X 2
FSTL1	HSP90AA1	100	3 X 2
FSTL1	IGHG1	100	3 X 2
FSTL1	IGHG2	100	3 X 2
FSTL1	C3	100	3 X 2
QSOX1	YWHAZ	100	5 X 2
QSOX1	HPX	100	5 X 4
QSOX1	CTSL	100	5 X 2
PATE1	HPX	100	4 X 4
SOD1	YWHAZ	100	2 X 2
SOD1	FUCA1	100	2 X 3
SOD1	ANXA5	100	2 X 7
SOD1	IGKC	100	2 X 5
SOD1	CNDP2	100	2 X 5
SOD1	HPX	100	2 X 4
SOD1	ANXA3	100	2 X 3
SOD1	ENO1	100	2 X 4
SOD1	PRDX6	100	2 X 3
SOD1	ELSPBP1	100	2 X 4
SOD1	RNASE4	100	2 X 4
SOD1	GDI2	100	2 X 2
SOD1	EZR	100	2 X 2
SOD1	GAPDH	100	2 X 4
SOD1	ANXA1	100	2 X 3
SOD1	CRTAC1	100	2 X 4
SOD1	PSMA2	100	2 X 2
SOD1	GPI	100	2 X 2
SOD1	LCP1	100	2 X 2
SOD1	HSP90AA1	100	2 X 2
SOD1	CTSL	100	2 X 2
SOD1	DBI	100	2 X 3
SOD1	IGHG1	100	2 X 2
SOD1	TMPRSS2	100	2 X 2

YWHAZ	LDHA	100	2 X 2
YWHAZ	ANXA5	100	2 X 7
YWHAZ	IGKC	100	2 X 5
YWHAZ	CNDP2	100	2 X 5
YWHAZ	HPX	100	2 X 4
YWHAZ	ANXA3	100	2 X 3
YWHAZ	ENO1	100	2 X 4
YWHAZ	SORD	100	2 X 4
YWHAZ	PRDX6	100	2 X 3
YWHAZ	ELSPBP1	100	2 X 4
YWHAZ	RNASE4	100	2 X 4
YWHAZ	GDI2	100	2 X 2
YWHAZ	TIMP1	100	2 X 8
YWHAZ	EZR	100	2 X 2
YWHAZ	GAPDH	100	2 X 4
YWHAZ	ANXA1	100	2 X 3
YWHAZ	EDDM3B	100	2 X 3
YWHAZ	PSMA2	100	2 X 2
YWHAZ	GPI	100	2 X 2
YWHAZ	LCP1	100	2 X 2
YWHAZ	HSP90A1	100	2 X 2
YWHAZ	CTSL	100	2 X 2
YWHAZ	DBI	100	2 X 3
YWHAZ	IGHG1	100	2 X 2
YWHAZ	IGHG2	100	2 X 2
YWHAZ	TMPRSS2	100	2 X 2
YWHAZ	C3	100	2 X 2
LDHA	CNDP2	100	2 X 5
LDHA	HPX	100	2 X 4
LDHA	PSMA2	100	2 X 2
LDHA	LCP1	100	2 X 2
LDHA	HSP90A1	100	2 X 2
LDHA	IGHG1	100	2 X 2
LDHA	C3	100	2 X 2
LRG1	OS9	100	3 X 2
LRG1	SDF4	100	3 X 2
LRG1	CTSL	100	3 X 2
LRG1	TMPRSS2	100	3 X 2
FUCA1	HPX	100	3 X 4
FUCA1	ANXA3	100	3 X 3
FUCA1	PRDX6	100	3 X 3
FUCA1	RNASE4	100	3 X 4
FUCA1	GDI2	100	3 X 2
FUCA1	ANXA1	100	3 X 3
FUCA1	HEXA	100	3 X 3
ANXA5	IGKC	100	7 X 5
		97.142	
ANXA5	CNDP2	86	7 X 5
ANXA5	HPX	100	7 X 4
ANXA5	ANXA3	100	7 X 3
ANXA5	ENO1	100	7 X 4
ANXA5	PRDX6	100	7 X 3
ANXA5	RNASE4	100	7 X 4
ANXA5	GDI2	100	7 X 2
ANXA5	EZR	100	7 X 2
ANXA5	GAPDH	100	7 X 4
ANXA5	ANXA1	100	7 X 3
		96.428	
ANXA5	CRTAC1	57	7 X 4
	EDDM3	95.238	
ANXA5	B	1	7 X 3
ANXA5	GPI	100	7 X 2
ANXA5	LCP1	100	7 X 2
ANXA5	CTSL	100	7 X 2
ANXA5	DBI	100	7 X 3
ANXA5	IGHG1	100	7 X 2
ANXA5	C3	100	7 X 2
IGKC	HPX	100	5 X 4
IGKC	ANXA3	100	5 X 3
IGKC	PRDX6	100	5 X 3
IGKC	RNASE4	100	5 X 4
IGKC	GDI2	100	5 X 2
IGKC	EZR	100	5 X 2
IGKC	GAPDH	100	5 X 4
IGKC	ANXA1	100	5 X 3
IGKC	GPI	100	5 X 2
IGKC	LCP1	100	5 X 2
IGKC	DBI	100	5 X 3
CNDP2	HPX	100	5 X 4
CNDP2	ANXA3	100	5 X 3
CNDP2	ENO1	100	5 X 4
CNDP2	ORM2	100	5 X 4
CNDP2	SORD	100	5 X 4
CNDP2	PRDX6	100	5 X 3
CNDP2	TIMP1	97.5	5 X 8
CNDP2	EZR	100	5 X 2
CNDP2	GAPDH	100	5 X 4

CNDP2	ANXA1	100	5 X 3
	EDDM3		
CNDP2	B	100	5 X 3
CNDP2	PSMA2	100	5 X 2
CNDP2	LCP1	100	5 X 2
	HSP90A		
CNDP2	A1	100	5 X 2
CNDP2	DBI	100	5 X 3
CNDP2	IGHG1	100	5 X 2
CNDP2	C3	100	5 X 2
VTN	PLA1A	100	2 X 4
VTN	TWSG1	100	2 X 2
HPX	ANXA3	100	4 X 3
HPX	ENO1	100	4 X 4
HPX	SORD	100	4 X 4
HPX	PRDX6	100	4 X 3
	ELSPBP		
HPX	1	100	4 X 4
HPX	RNASE4	100	4 X 4
HPX	GDI2	100	4 X 2
HPX	TIMP1	100	4 X 8
HPX	EZR	100	4 X 2
HPX	GAPDH	100	4 X 4
HPX	ANXA1	100	4 X 3
HPX	CRTAC1	100	4 X 4
	EDDM3		
HPX	B	100	4 X 3
HPX	PSMA2	100	4 X 2
HPX	GPI	100	4 X 2
HPX	LCP1	100	4 X 2
	HSP90A		
HPX	A1	100	4 X 2
HPX	CTSL	100	4 X 2
HPX	DBI	100	4 X 3
HPX	IGHG1	100	4 X 2
HPX	IGHG2	100	4 X 2
	TMPRSS		
HPX	2	100	4 X 2
HPX	C3	100	4 X 2
ANXA3	ENO1	100	3 X 4
ANXA3	PRDX6	100	3 X 3
	ELSPBP		
ANXA3	1	100	3 X 4
		95.238	
ANXA3	NUCB1	1	3 X 7
ANXA3	RNASE4	100	3 X 4
ANXA3	GDI2	100	3 X 2
		95.833	
ANXA3	TIMP1	33	3 X 8
ANXA3	EZR	100	3 X 2
ANXA3	GAPDH	100	3 X 4
ANXA3	ANXA1	100	3 X 3
ANXA3	CRTAC1	100	3 X 4
ANXA3	TWSG1	100	3 X 2
ANXA3	GPI	100	3 X 2
ANXA3	LCP1	100	3 X 2
	HSP90A		
ANXA3	A1	100	3 X 2
ANXA3	CTSL	100	3 X 2
ANXA3	DBI	100	3 X 3
ANXA3	IGHG1	100	3 X 2
	TMPRSS		
ANXA3	2	100	3 X 2
ENO1	PRDX6	100	4 X 3
ENO1	EZR	100	4 X 2
ENO1	GAPDH	100	4 X 4
ENO1	ANXA1	100	4 X 3
	EDDM3		
ENO1	B	100	4 X 3
ENO1	PSMA2	100	4 X 2
ENO1	GPI	100	4 X 2
ENO1	LCP1	100	4 X 2
	HSP90A		
ENO1	A1	100	4 X 2
ENO1	DBI	100	4 X 3
ENO1	IGHG1	100	4 X 2
ENO1	IGHG2	100	4 X 2
ENO1	C3	100	4 X 2
	EDDM3		
ORM2	B	100	4 X 3
ORM2	PSMA2	100	4 X 2
	HSP90A		
ORM2	A1	100	4 X 2
ORM2	IGHG1	100	4 X 2
ORM2	C3	100	4 X 2
OS9	TWSG1	100	2 X 2
OS9	CTSL	100	2 X 2
	TMPRSS		
OS9	2	100	2 X 2
	EDDM3		
SORD	B	100	4 X 3
SORD	PSMA2	100	4 X 2
	HSP90A		
SORD	A1	100	4 X 2
SORD	IGHG1	100	4 X 2

SORD	C3	100	4 X 2
PRDX6	ELSPBP 1	100	3 X 4
PRDX6	RNASE4	100	3 X 4
PRDX6	GDI2	100	3 X 2
		95.833	
PRDX6	TIMP1	33	3 X 8
PRDX6	EZR	100	3 X 2
PRDX6	GAPDH	100	3 X 4
PRDX6	ANXA1	100	3 X 3
PRDX6	CRTAC1	100	3 X 4
PRDX6	EDDM3 B	100	3 X 3
PRDX6	PSMA2	100	3 X 2
PRDX6	GPI	100	3 X 2
PRDX6	LCP1	100	3 X 2
PRDX6	HSP90A A1	100	3 X 2
PRDX6	CTSL	100	3 X 2
PRDX6	DBI	100	3 X 3
PRDX6	IGHG1	100	3 X 2
PRDX6	IGHG2	100	3 X 2
PRDX6	TMPRSS 2	100	3 X 2
PRDX6	C3	100	3 X 2
ELSPBP 1	GDI2	100	4 X 2
ELSPBP 1	EZR	100	4 X 2
ELSPBP 1	ANXA1	100	4 X 3
ELSPBP 1	GPI	100	4 X 2
ASAH1	TMPRSS 2	100	4 X 2
PLA1A	TWSG1	100	4 X 2
PLA1A	TPP1	100	4 X 3
PLA1A	TMPRSS 2	100	4 X 2
PEBP4	PSMA2	100	4 X 2
PEBP4	IGHG1	100	4 X 2
PEBP4	C3	100	4 X 2
RNASE4	GDI2	100	4 X 2
RNASE4	EZR	100	4 X 2
RNASE4	ANXA1	100	4 X 3
RNASE4	HEXA	100	4 X 3
RNASE4	CRTAC1	100	4 X 4
RNASE4	GPI	100	4 X 2

GDI2	EZR	100	2 X 2
GDI2	GAPDH	100	2 X 4
GDI2	ANXA1	100	2 X 3
GDI2	HEXA	100	2 X 3
GDI2	CRTAC1	100	2 X 4
GDI2	TWSG1	100	2 X 2
GDI2	GPI	100	2 X 2
GDI2	LCP1	100	2 X 2
GDI2	CTSL	100	2 X 2
GDI2	TMPRSS 2	100	2 X 2
TIMP1	EZR	100	8 X 2
		95.833	
TIMP1	ANXA1	33	8 X 3
TIMP1	CTSL	100	8 X 2
		95.833	
TIMP1	DBI	33	8 X 3
EZR	GAPDH	100	2 X 4
EZR	ANXA1	100	2 X 3
EZR	PSMA2	100	2 X 2
EZR	GPI	100	2 X 2
EZR	LCP1	100	2 X 2
EZR	HSP90A A1	100	2 X 2
EZR	CTSL	100	2 X 2
EZR	DBI	100	2 X 3
EZR	IGHG1	100	2 X 2
EZR	C3	100	2 X 2
GAPDH	ANXA1	100	4 X 3
GAPDH	EDDM3 B	100	4 X 3
GAPDH	PSMA2	100	4 X 2
GAPDH	GPI	100	4 X 2
GAPDH	LCP1	100	4 X 2
GAPDH	HSP90A A1	100	4 X 2
GAPDH	CTSL	100	4 X 2
GAPDH	DBI	100	4 X 3
GAPDH	IGHG1	100	4 X 2
GAPDH	IGHG2	100	4 X 2
GAPDH	C3	100	4 X 2
SDF4	CTSL	100	2 X 2
SDF4	TMPRSS 2	100	2 X 2
ANXA1	CRTAC1	100	3 X 4
ANXA1	EDDM3 B	100	3 X 3

ANXA1	PSMA2	100	3 X 2
ANXA1	GPI	100	3 X 2
ANXA1	LCP1	100	3 X 2
ANXA1	HSP90A A1	100	3 X 2
ANXA1	CTSL	100	3 X 2
ANXA1	DBI	100	3 X 3
ANXA1	IGHG1	100	3 X 2
ANXA1	TMPRSS 2	100	3 X 2
ANXA1	C3	100	3 X 2
HEXA	TMPRSS 2	100	3 X 2
EDDM3 B	PSMA2	100	3 X 2
EDDM3 B	GPI	100	3 X 2
EDDM3 B	LCP1	100	3 X 2
EDDM3 B	HSP90A A1	100	3 X 2
EDDM3 B	DBI	100	3 X 3
EDDM3 B	IGHG1	100	3 X 2
EDDM3 B	IGHG2	100	3 X 2
EDDM3 B	C3	100	3 X 2
TWSG1	TPP1	100	2 X 3
TWSG1	CTSL	100	2 X 2
TWSG1	TMPRSS 2	100	2 X 2
PSMA2	LCP1	100	2 X 2
PSMA2	HSP90A A1	100	2 X 2
PSMA2	DBI	100	2 X 3
PSMA2	IGHG1	100	2 X 2
PSMA2	IGHG2	100	2 X 2
PSMA2	C3	100	2 X 2
GPI	LCP1	100	2 X 2
GPI	HSP90A A1	100	2 X 2
GPI	CTSL	100	2 X 2
GPI	DBI	100	2 X 3
GPI	IGHG1	100	2 X 2
GPI	IGHG2	100	2 X 2
GPI	TMPRSS 2	100	2 X 2
LCP1	HSP90A A1	100	2 X 2
LCP1	DBI	100	2 X 3
LCP1	IGHG1	100	2 X 2
LCP1	IGHG2	100	2 X 2
LCP1	C3	100	2 X 2
HSP90A A1	DBI	100	2 X 3
HSP90A A1	IGHG1	100	2 X 2
HSP90A A1	IGHG2	100	2 X 2
HSP90A A1	C3	100	2 X 2
CTSL	MATN2	100	2 X 2
CTSL	TMPRSS 2	100	2 X 2
DBI	IGHG1	100	3 X 2
DBI	IGHG2	100	3 X 2
DBI	C3	100	3 X 2
IGHG1	IGHG2	100	2 X 2
IGHG1	C3	100	2 X 2
IGHG2	C3	100	2 X 2

ANNEX D – RESULTATS ESTUDI MITJA-SD

		A1	A2	A3	A4	O1	O2	O3	O4	Z1	Z2	Z3	Z4
BASP1	YWHAE	100	100	75	75	100	100	100	100	75	100	100	100
BASP1	ALDOA	83.33	100	83.33	100	100	100	100	83.33	83.33	83.33	100	100
BASP1	ECM1	95.45	95.45	95.45	72.73	95.45	63.64	86.36	86.36	27.27	90.91	59.09	54.55
BASP1	PKM	100	100	66.67	100	100	100	100	100	100	100	100	100
BASP1	SCPEP1	100	100	83.33	100	100	100	100	100	100	100	100	100
BASP1	SOD3	100	100	100	50	100	100	100	100	100	100	100	75
BASP1	AOC1	100	100	100	100	100	75	100	100	100	100	75	100
BASP1	GAA	100	100	100	100	100	75	100	100	100	100	75	100
BASP1	TF	100	100	100	75	96.43	85.71	100	100	78.57	60.71	57.14	100
BASP1	CD9	75	100	75	75	100	100	100	100	75	100	100	100
BASP1	LGALS3BP	85	100	80	85	95	75	90	80	65	95	75	85
BASP1	PARK7	66.67	83.33	16.67	100	83.33	66.67	83.33	66.67	66.67	83.33	100	16.67
BASP1	PTGDS	100	100	100	100	100	100	0	100	50	25	0	0
BASP1	CST3	80	100	100	100	80	50	100	80	60	90	80	100
BASP1	APOH	50	100	100	100	100	100	100	100	100	100	100	100
BASP1	LSAMP	100	100	50	87.5	100	87.5	87.5	87.5	62.5	100	87.5	50
BASP1	S100A11	100	100	50	100	100	100	100	100	100	100	100	100
BASP1	SPOCK1	66.67	100	50	83.33	83.33	83.33	83.33	100	83.33	100	50	66.67
BASP1	APLP2	50	100	0	75	75	100	100	75	75	75	100	25
BASP1	HSPA8	100	100	50	100	100	100	100	100	100	100	100	100
BASP1	FSTL1	100	100	50	100	100	100	100	100	83.33	100	100	100
BASP1	QSOX1	100	100	100	100	100	100	100	100	80	100	100	100
BASP1	PATE1	100	100	75	100	100	100	100	100	100	100	100	100
BASP1	SOD1	100	75	50	75	100	100	100	75	100	100	75	100
BASP1	YWHAZ	100	100	25	100	75	75	50	50	75	100	100	50
BASP1	LDHA	100	100	75	100	100	100	100	100	100	100	100	75
BASP1	FUCA1	100	100	100	100	100	100	100	100	100	100	100	100
BASP1	ANXA5	85.71	85.71	57.14	100	100	100	92.86	92.86	78.57	100	92.86	100
BASP1	IGKC	90	100	90	100	80	70	100	100	80	100	100	100
BASP1	CNDP2	100	100	50	100	100	90	100	80	90	100	100	100
BASP1	HPX	87.5	100	100	100	87.5	50	87.5	50	50	62.5	62.5	75
BASP1	ANXA3	100	100	100	100	100	100	100	83.33	100	100	100	100
BASP1	ENO1	75	100	75	87.5	100	100	100	87.5	87.5	100	100	100
BASP1	ORM2	100	100	50	87.5	100	100	100	100	100	75	75	100
BASP1	SORD	87.5	75	25	87.5	100	100	100	87.5	100	87.5	100	50

BASP1	PRDX6	100	100	16.67	83.33	100	100	100	66.67	66.67	100	100	83.33
BASP1	ELSPBP1	0	100	100	62.5	100	75	50	62.5	37.5	25	62.5	25
BASP1	RNASE4	87.5	100	87.5	100	100	87.5	100	100	87.5	100	100	100
BASP1	GDI2	100	100	75	100	100	75	100	75	75	100	75	100
BASP1	EZR	100	100	100	75	100	100	100	100	100	100	100	100
BASP1	GAPDH	100	100	50	100	100	100	100	75	87.5	100	87.5	87.5
BASP1	ANXA1	100	100	16.67	83.33	100	83.33	100	83.33	66.67	83.33	100	50
BASP1	CRTAC1	100	100	87.5	100	100	75	100	100	62.5	87.5	100	100
BASP1	EDDM3B	100	100	100	100	100	100	100	100	83.33	100	66.67	83.33
BASP1	PSMA2	100	100	50	100	100	100	100	100	100	100	100	100
BASP1	GPI	100	100	100	100	100	100	100	100	100	100	100	100
BASP1	LCP1	100	100	0	100	100	100	75	75	75	75	50	75
BASP1	HSP90AA1	100	100	75	100	100	100	75	100	100	100	100	75
BASP1	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
BASP1	DBI	100	100	66.67	100	100	100	83.33	83.33	66.67	100	100	66.67
BASP1	IGHG1	25	100	100	100	75	100	100	75	100	100	100	75
BASP1	IGHG2	100	100	100	100	100	100	75	100	100	100	100	100
BASP1	TMPRSS2	100	100	50	100	100	100	100	100	100	100	100	100
BASP1	C3	75	100	100	100	100	100	100	100	100	100	50	100
YWHAE	ALDOA	100	100	100	100	100	100	100	100	83.33	100	83.33	100
YWHAE	ECM1	100	95.45	100	63.64	95.45	81.82	100	40.91	54.55	100	54.55	72.73
YWHAE	PKM	100	100	83.33	83.33	100	100	100	83.33	100	100	50	100
YWHAE	SCPEP1	100	100	100	100	100	100	100	83.33	100	100	100	100
YWHAE	SOD3	100	75	100	25	75	100	100	75	100	100	50	75
YWHAE	GAA	100	100	100	75	100	75	100	75	100	100	75	100
YWHAE	CD9	100	100	100	100	100	100	100	100	75	100	75	100
YWHAE	PARK7	100	83.33	83.33	100	100	100	100	83.33	83.33	100	66.67	66.67
YWHAE	LAMB2	100	100	100	100	100	100	100	100	90	90	90	100
YWHAE	PTGDS	100	100	100	100	100	100	0	25	75	25	0	25
YWHAE	CST3	100	90	100	100	60	60	100	100	70	90	100	100
YWHAE	LSAMP	100	100	87.5	62.5	75	87.5	100	100	75	87.5	87.5	75
YWHAE	SPOCK1	100	100	83.33	100	66.67	100	100	66.67	83.33	100	50	83.33
YWHAE	APLP2	100	100	50	100	100	100	100	100	100	100	75	50
YWHAE	HSPA8	100	100	100	100	100	100	100	100	100	100	75	100
YWHAE	SOD1	100	100	100	100	100	75	100	75	75	100	50	100
YWHAE	YWHAZ	100	100	100	100	100	100	100	100	50	100	75	75
YWHAE	FUCA1	100	83.33	83.33	83.33	83.33	100	83.33	66.67	83.33	100	66.67	66.67
YWHAE	ANXA5	100	85.71	100	100	100	100	100	92.86	85.71	100	85.71	100
YWHAE	IGKC	80	90	90	100	80	70	90	60	60	80	60	80
YWHAE	HPX	100	87.5	100	100	87.5	100	100	100	75	100	100	100
YWHAE	ANXA3	100	100	83.33	100	100	100	100	66.67	50	100	83.33	100

YWHAE	ENO1	100	87.5	87.5	100	75	75	100	100	75	87.5	75	87.5
YWHAE	PRDX6	100	100	50	100	100	66.67	100	100	83.33	66.67	83.33	83.33
YWHAE	ELSPBP1	25	87.5	100	37.5	100	100	100	50	75	50	50	37.5
YWHAE	RNASE4	87.5	100	75	87.5	87.5	75	100	75	75	75	75	87.5
YWHAE	GDI2	100	75	100	100	100	75	100	100	100	75	75	75
YWHAE	EZR	100	100	100	75	100	100	100	75	50	100	100	100
YWHAE	GAPDH	100	100	87.5	100	100	100	100	100	100	87.5	75	100
YWHAE	ANXA1	100	83.33	66.67	100	83.33	83.33	100	83.33	66.67	66.67	50	66.67
YWHAE	HEXA	100	83.33	83.33	83.33	66.67	66.67	83.33	100	83.33	83.33	83.33	83.33
YWHAE	CRTAC1	87.5	100	100	87.5	87.5	75	87.5	100	62.5	100	87.5	100
YWHAE	GPI	100	100	100	100	100	100	100	75	75	75	50	100
YWHAE	LCP1	100	100	75	100	100	100	100	50	75	75	50	75
YWHAE	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
YWHAE	DBI	100	100	100	100	83.33	100	100	66.67	66.67	100	66.67	66.67
YWHAE	IGHG1	50	100	100	100	50	100	100	100	100	100	100	100
YWHAE	TMPRSS2	100	100	100	100	100	100	100	100	100	100	75	100
ALDOA	ECM1	96.97	87.88	96.97	81.82	81.82	87.88	100	66.67	51.52	90.91	75.76	78.79
ALDOA	PKM	100	100	55.56	77.78	100	100	88.89	88.89	77.78	100	88.89	77.78
ALDOA	SCPEP1	88.89	88.89	77.78	100	66.67	88.89	100	88.89	55.56	88.89	88.89	88.89
ALDOA	SOD3	100	100	100	83.33	100	100	100	83.33	100	100	83.33	66.67
ALDOA	AOC1	100	100	100	100	100	66.67	100	83.33	100	100	100	100
ALDOA	GAA	100	83.33	83.33	83.33	66.67	83.33	100	66.67	83.33	83.33	83.33	100
ALDOA	TF	100	100	100	88.1	90.48	97.62	100	100	97.62	71.43	69.05	100
ALDOA	CD9	100	100	100	100	83.33	100	100	100	83.33	100	100	100
ALDOA	LGALS3BP	100	93.33	86.67	80	83.33	90	86.67	50	80	86.67	80	86.67
ALDOA	PARK7	100	77.78	77.78	100	100	100	100	100	100	88.89	100	77.78
ALDOA	PTGDS	100	100	100	100	100	100	66.67	50	83.33	16.67	0	0
ALDOA	APOH	83.33	100	100	100	100	100	100	100	100	100	100	100
ALDOA	LSAMP	100	100	83.33	83.33	83.33	100	100	91.67	83.33	100	91.67	83.33
ALDOA	S100A11	100	100	100	100	100	100	100	83.33	100	83.33	100	100
ALDOA	SPOCK1	100	100	88.89	100	100	100	100	88.89	100	100	77.78	88.89
ALDOA	APLP2	100	100	50	83.33	66.67	100	100	83.33	66.67	83.33	100	50
ALDOA	HSPA8	100	100	100	100	100	100	100	100	100	100	100	83.33
ALDOA	FSTL1	100	100	88.89	100	88.89	100	100	88.89	88.89	100	100	88.89
ALDOA	SOD1	100	100	83.33	100	66.67	66.67	100	100	83.33	83.33	83.33	100
ALDOA	YWHAZ	83.33	83.33	83.33	100	83.33	83.33	83.33	100	83.33	83.33	100	83.33
ALDOA	LDHA	100	100	100	100	100	100	100	100	83.33	100	100	83.33
ALDOA	ANXA5	90.48	85.71	95.24	100	95.24	95.24	95.24	100	76.19	95.24	95.24	95.24
ALDOA	IGKC	66.67	73.33	73.33	86.67	60	66.67	86.67	66.67	80	73.33	80	80
ALDOA	CNDP2	100	100	86.67	93.33	100	86.67	100	93.33	73.33	86.67	100	100
ALDOA	HPX	100	91.67	100	100	75	91.67	91.67	91.67	91.67	91.67	83.33	83.33

ALDOA	ANXA3	100	100	100	100	88.89	100	100	100	100	100	100	100
ALDOA	ENO1	100	91.67	100	100	91.67	100	100	100	91.67	91.67	100	100
ALDOA	ORM2	100	100	91.67	100	75	100	100	91.67	100	66.67	66.67	91.67
ALDOA	SORD	100	50	50	100	83.33	100	100	100	91.67	100	100	58.33
ALDOA	PRDX6	100	100	88.89	100	100	100	100	100	55.56	88.89	100	100
ALDOA	RNASE4	100	91.67	100	100	100	100	100	83.33	91.67	83.33	100	100
ALDOA	GDI2	100	100	100	100	100	100	100	100	83.33	100	100	100
ALDOA	TIMP1	91.67	100	87.5	87.5	83.33	100	100	95.83	100	91.67	100	91.67
ALDOA	EZR	100	100	100	83.33	83.33	100	100	100	83.33	100	100	100
ALDOA	GAPDH	100	100	83.33	100	83.33	91.67	100	100	91.67	91.67	100	100
ALDOA	ANXA1	100	100	88.89	100	100	88.89	100	100	77.78	88.89	100	100
ALDOA	EDDM3B	100	100	100	100	88.89	100	100	100	77.78	100	44.44	77.78
ALDOA	PSMA2	100	100	66.67	100	100	100	100	83.33	66.67	100	83.33	83.33
ALDOA	GPI	100	83.33	100	100	50	83.33	100	100	83.33	100	100	100
ALDOA	LCP1	100	100	50	100	100	83.33	83.33	33.33	83.33	50	50	66.67
ALDOA	HSP90AA1	83.33	100	66.67	83.33	66.67	83.33	83.33	83.33	50	100	66.67	66.67
ALDOA	CTSL	83.33	100	100	100	83.33	83.33	100	100	83.33	100	83.33	100
ALDOA	DBI	100	100	88.89	100	100	100	100	66.67	88.89	88.89	100	66.67
ALDOA	IGHG1	50	100	100	100	50	100	100	100	83.33	100	100	83.33
ALDOA	IGHG2	100	66.67	100	100	83.33	100	100	66.67	83.33	100	100	100
ALDOA	TMPRSS2	100	100	83.33	100	100	100	100	100	83.33	100	100	83.33
ALDOA	C3	50	100	100	100	66.67	100	83.33	100	100	83.33	50	83.33
CPE	ANXA5	96.43	90.71	95.71	99.29	96.43	97.14	95	95	78.57	97.14	95.71	86.43
ECM1	PKM	96.97	96.97	51.52	96.97	93.94	90.91	96.97	96.97	84.85	96.97	93.94	96.97
ECM1	SOD3	100	90.91	100	90.91	100	95.45	100	100	95.45	100	90.91	90.91
ECM1	AOC1	90.91	100	100	100	100	95.45	100	100	100	95.45	90.91	95.45
ECM1	GAA	100	100	77.27	100	95.45	100	100	95.45	95.45	95.45	90.91	95.45
ECM1	CD9	100	95.45	95.45	100	100	95.45	100	100	77.27	100	100	100
ECM1	PARK7	100	93.94	72.73	100	93.94	93.94	96.97	87.88	84.85	93.94	84.85	90.91
ECM1	PTGDS	95.45	81.82	100	100	95.45	100	31.82	95.45	95.45	27.27	9.09	50
ECM1	LSAMP	100	95.45	86.36	97.73	100	97.73	100	97.73	97.73	100	95.45	97.73
ECM1	SPOCK1	93.94	96.97	69.7	100	100	96.97	93.94	100	87.88	96.97	96.97	93.94
ECM1	APLP2	90.91	95.45	50	100	100	81.82	100	100	63.64	86.36	86.36	86.36
ECM1	HSPA8	100	100	68.18	100	95.45	100	100	95.45	86.36	100	100	90.91
ECM1	SOD1	100	95.45	68.18	95.45	90.91	95.45	100	90.91	77.27	86.36	86.36	81.82
ECM1	YWHAZ	100	90.91	100	100	90.91	86.36	100	100	59.09	90.91	90.91	100
ECM1	ANXA5	87.01	87.01	81.82	93.51	94.81	92.21	98.7	88.31	76.62	94.81	88.31	92.21
ECM1	IGKC	81.82	96.36	85.45	98.18	90.91	83.64	90.91	92.73	81.82	87.27	87.27	81.82
ECM1	CNDP2	100	98.18	72.73	100	100	89.09	100	98.18	72.73	94.55	89.09	98.18
ECM1	HPX	93.18	90.91	97.73	95.45	95.45	97.73	93.18	77.27	86.36	97.73	72.73	72.73
ECM1	ANXA3	100	96.97	93.94	96.97	93.94	93.94	100	93.94	84.85	93.94	87.88	93.94

ECM1	ENO1	100	97.73	97.73	97.73	100	97.73	100	97.73	97.73	100	93.18	95.45
ECM1	PRDX6	100	90.91	57.58	96.97	87.88	69.7	100	78.79	57.58	78.79	72.73	84.85
ECM1	ELSPBP1	18.18	93.18	93.18	72.73	95.45	95.45	90.91	70.45	88.64	56.82	81.82	56.82
ECM1	RNASE4	84.09	93.18	88.64	100	93.18	81.82	100	93.18	75	90.91	81.82	90.91
ECM1	GDI2	100	95.45	95.45	95.45	90.91	95.45	100	81.82	90.91	90.91	77.27	77.27
ECM1	EZR	100	95.45	100	95.45	95.45	95.45	100	95.45	59.09	90.91	81.82	90.91
ECM1	GAPDH	97.73	95.45	93.18	95.45	95.45	93.18	100	93.18	79.55	95.45	88.64	93.18
ECM1	ANXA1	96.97	93.94	66.67	93.94	90.91	81.82	100	96.97	63.64	87.88	90.91	93.94
ECM1	EDDM3B	93.94	90.91	100	93.94	90.91	93.94	93.94	90.91	87.88	90.91	78.79	93.94
ECM1	GPI	95.45	95.45	100	90.91	95.45	100	100	95.45	81.82	95.45	86.36	90.91
ECM1	LCP1	95.45	95.45	45.45	100	90.91	81.82	90.91	90.91	86.36	81.82	86.36	81.82
ECM1	DBI	100	96.97	78.79	100	96.97	93.94	100	96.97	87.88	93.94	90.91	90.91
ECM1	IGHG1	54.55	95.45	100	90.91	81.82	95.45	81.82	81.82	77.27	95.45	86.36	77.27
ECM1	IGHG2	100	72.73	100	100	100	100	95.45	100	95.45	95.45	95.45	95.45
ECM1	C3	50	100	100	100	100	100	100	90.91	90.91	95.45	90.91	90.91
PKM	SOD3	100	100	100	100	100	100	100	100	100	100	100	100
PKM	AOC1	100	100	100	100	100	50	100	100	100	100	100	100
PKM	GAA	100	66.67	83.33	83.33	66.67	66.67	83.33	66.67	83.33	66.67	100	83.33
PKM	CD9	100	100	100	100	100	100	100	100	100	100	100	100
PKM	LGALS3BP	83.33	90	83.33	86.67	70	80	86.67	66.67	60	83.33	86.67	83.33
PKM	PARK7	100	100	100	100	100	100	100	100	100	100	100	100
PKM	PTGDS	100	100	100	100	83.33	100	66.67	100	66.67	0	16.67	83.33
PKM	APOH	83.33	100	100	100	100	100	100	100	100	100	100	100
PKM	LSAMP	100	100	100	100	100	100	100	100	91.67	100	100	100
PKM	SPOCK1	100	100	100	100	100	100	100	100	100	100	100	100
PKM	APLP2	100	100	83.33	100	100	100	100	100	100	100	100	100
PKM	HSPA8	100	100	100	100	100	100	100	100	100	100	100	100
PKM	FSTL1	100	100	100	100	100	100	100	100	88.89	100	100	100
PKM	SOD1	83.33	100	100	100	100	50	100	100	83.33	50	100	83.33
PKM	YWHAZ	100	100	100	100	100	100	100	100	100	100	100	100
PKM	ANXA5	95.24	85.71	95.24	95.24	100	100	100	100	85.71	100	100	100
PKM	IGKC	60	73.33	80	93.33	53.33	80	80	86.67	80	66.67	80	73.33
PKM	HPX	91.67	100	100	100	100	100	100	100	83.33	100	100	100
PKM	ANXA3	88.89	100	100	88.89	100	100	100	100	100	88.89	100	100
PKM	ENO1	100	100	91.67	100	100	100	100	100	83.33	91.67	100	100
PKM	PRDX6	100	100	100	100	100	100	100	100	77.78	100	100	100
PKM	ELSPBP1	16.67	100	100	75	100	100	91.67	75	66.67	50	100	75
PKM	GDI2	83.33	100	100	100	100	100	100	100	66.67	100	100	100
PKM	EZR	100	100	100	100	100	100	100	100	100	100	100	100
PKM	GAPDH	100	100	100	91.67	100	100	100	100	91.67	100	100	100
PKM	ANXA1	100	100	100	100	100	100	100	100	100	100	100	100

PKM	EDDM3B	100	100	55.56	100	77.78	77.78	88.89	100	66.67	88.89	100	100
PKM	PSMA2	100	100	83.33	100	100	100	100	100	100	100	100	100
PKM	GPI	100	66.67	83.33	83.33	66.67	50	100	100	83.33	83.33	100	100
PKM	LCP1	100	100	83.33	100	100	100	83.33	50	66.67	66.67	50	66.67
PKM	HSP90AA1	100	100	83.33	100	83.33	100	66.67	100	83.33	100	83.33	66.67
PKM	DBI	100	100	100	100	100	100	100	100	88.89	100	100	100
PKM	IGHG1	50	100	100	100	83.33	100	100	100	100	100	100	100
PKM	IGHG2	100	33.33	66.67	100	83.33	66.67	50	83.33	66.67	83.33	100	100
PKM	C3	33.33	100	100	100	100	100	100	100	100	50	100	100
SCPEP1	GAA	100	66.67	100	83.33	100	83.33	100	66.67	50	83.33	66.67	100
SCPEP1	CD9	100	100	100	100	100	100	100	100	100	100	100	100
SCPEP1	PARK7	100	88.89	100	100	100	100	100	100	88.89	100	100	100
SCPEP1	LSAMP	100	100	100	100	100	100	100	100	75	100	100	100
SCPEP1	SPOCK1	100	100	100	100	100	100	100	100	100	100	100	100
SCPEP1	APLP2	100	83.33	66.67	100	100	100	100	100	100	100	100	83.33
SCPEP1	SOD1	100	100	100	100	100	83.33	100	100	50	50	66.67	100
SCPEP1	YWHAZ	100	100	100	100	100	100	100	100	100	100	100	100
SCPEP1	ANXA5	95.24	80.95	100	100	95.24	100	100	100	80.95	100	95.24	95.24
SCPEP1	IGKC	73.33	73.33	86.67	100	86.67	73.33	80	73.33	66.67	53.33	40	73.33
SCPEP1	HPX	100	91.67	100	100	100	75	100	100	66.67	100	100	100
SCPEP1	ANXA3	100	88.89	100	100	100	88.89	100	100	100	66.67	100	100
SCPEP1	PRDX6	100	100	100	100	100	100	100	100	66.67	100	100	100
SCPEP1	TIMP1	95.83	91.67	91.67	87.5	100	95.83	100	100	83.33	100	87.5	95.83
SCPEP1	EZR	100	83.33	100	100	100	100	100	100	83.33	83.33	100	100
SCPEP1	GAPDH	100	100	100	100	100	100	100	100	75	100	91.67	100
SCPEP1	ANXA1	100	88.89	100	100	100	100	100	100	100	100	100	100
SCPEP1	GPI	100	50	100	100	100	66.67	100	100	50	66.67	100	100
SCPEP1	CTSL	83.33	100	100	100	100	66.67	83.33	100	16.67	66.67	100	100
MMP2	PTGDS	100	91.67	91.67	100	100	100	75	91.67	83.33	25	0	83.33
MMP2	IGHG2	75	58.33	41.67	100	91.67	91.67	83.33	83.33	91.67	83.33	41.67	100
SOD3	GAA	100	100	100	100	100	100	100	75	100	100	100	100
SOD3	CD9	50	100	100	100	75	100	100	100	50	100	100	75
SOD3	PARK7	100	83.33	100	100	83.33	83.33	100	100	100	100	100	83.33
SOD3	LSAMP	87.5	87.5	87.5	100	62.5	87.5	100	87.5	75	100	100	75
SOD3	SPOCK1	83.33	100	83.33	100	83.33	100	83.33	83.33	100	100	100	83.33
SOD3	APLP2	25	75	50	100	50	100	100	100	100	100	100	50
SOD3	SOD1	100	100	100	100	75	75	100	100	100	100	100	100
SOD3	YWHAZ	100	75	100	75	100	100	100	75	0	100	75	75
SOD3	ANXA5	71.43	85.71	100	85.71	92.86	92.86	100	85.71	78.57	92.86	71.43	85.71
SOD3	IGKC	70	80	80	90	60	80	90	70	80	90	90	80
SOD3	HPX	62.5	87.5	100	50	50	50	62.5	62.5	100	100	50	62.5

SOD3	ANXA3	83.33	66.67	83.33	66.67	66.67	83.33	83.33	83.33	83.33	100	50	50
SOD3	PRDX6	100	66.67	33.33	100	100	83.33	100	66.67	50	83.33	50	83.33
SOD3	RNASE4	100	100	100	100	100	100	100	75	87.5	100	100	100
SOD3	GDI2	75	100	100	100	100	50	100	75	100	100	50	100
SOD3	EZR	75	100	100	100	75	100	100	75	75	100	75	75
SOD3	GAPDH	87.5	100	100	87.5	87.5	87.5	100	87.5	75	100	87.5	87.5
SOD3	ANXA1	100	83.33	50	100	83.33	83.33	83.33	100	33.33	83.33	100	66.67
SOD3	GPI	75	75	100	75	75	75	75	75	75	75	75	75
SOD3	LCP1	100	100	50	75	75	100	100	50	100	75	75	100
SOD3	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
AOC1	GAA	50	75	100	100	100	100	100	100	100	50	100	100
AOC1	TF	60.71	96.43	100	96.43	75	89.29	100	100	96.43	17.86	89.29	96.43
AOC1	TGM4	55	90	97.5	100	85	95	100	100	100	97.5	97.5	100
AOC1	PARK7	33.33	100	100	100	100	100	100	100	100	83.33	100	100
AOC1	PTGDS	75	100	100	100	100	100	25	100	75	0	0	25
AOC1	APOH	25	100	100	100	100	100	100	100	100	100	100	100
AOC1	S100A11	50	100	75	100	100	75	100	100	75	100	100	100
AOC1	SPOCK1	0	100	83.33	100	100	100	100	100	100	83.33	100	100
AOC1	HSPA8	25	100	75	100	100	75	100	100	100	25	100	100
AOC1	FSTL1	33.33	83.33	66.67	100	100	50	100	100	66.67	50	100	83.33
AOC1	YWHAZ	50	100	100	100	100	75	100	100	100	50	100	100
AOC1	LDHA	75	100	75	100	100	75	100	100	100	75	100	75
AOC1	CNDP2	70	100	80	100	100	80	100	100	80	100	100	100
AOC1	HPX	25	100	100	100	87.5	87.5	100	87.5	100	62.5	87.5	100
AOC1	ENO1	50	100	100	100	100	100	100	100	87.5	87.5	100	100
AOC1	PRDX6	50	100	83.33	100	100	66.67	100	100	83.33	100	100	100
AOC1	PEBP4	37.5	100	87.5	100	100	87.5	100	87.5	87.5	75	100	75
AOC1	GAPDH	62.5	100	87.5	100	87.5	87.5	100	87.5	87.5	62.5	100	100
AOC1	ANXA1	33.33	100	83.33	100	100	66.67	100	100	83.33	100	100	100
AOC1	EDDM3B	50	100	100	100	100	100	100	100	100	100	83.33	100
AOC1	PSMA2	25	100	75	100	100	75	100	100	100	100	100	100
AOC1	GPI	50	100	100	100	100	75	100	100	75	75	100	100
AOC1	LCP1	75	100	75	100	100	100	100	100	100	100	100	100
AOC1	HSP90AA1	75	100	75	100	75	75	75	100	100	100	100	75
AOC1	DBI	33.33	100	83.33	100	100	66.67	100	100	83.33	66.67	100	83.33
AOC1	IGHG1	0	100	100	100	50	100	100	100	75	50	100	75
AOC1	IGHG2	75	75	100	100	100	100	75	100	100	75	100	100
AOC1	C3	0	75	50	100	75	50	75	50	75	25	50	50
GAA	TF	100	100	100	100	100	100	100	100	96.43	92.86	92.86	100
GAA	CD9	100	100	100	100	100	100	100	100	100	100	100	100
GAA	PARK7	100	100	100	100	100	100	100	100	100	100	100	100

GAA	PTGDS	100	100	100	100	100	100	50	100	75	25	0	25
GAA	CST3	100	100	100	100	100	100	100	100	90	100	100	100
GAA	APOH	100	100	100	100	100	100	100	100	100	100	100	100
GAA	LSAMP	100	100	100	100	100	100	100	100	87.5	100	100	100
GAA	S100A11	100	100	100	100	100	100	100	100	100	100	100	100
GAA	SPOCK1	100	100	100	100	100	100	100	100	100	100	100	100
GAA	APLP2	100	100	75	100	100	100	100	100	100	100	100	100
GAA	HSPA8	100	100	100	100	100	100	100	100	100	100	100	100
GAA	FSTL1	100	100	100	100	100	100	100	100	100	100	100	100
GAA	SOD1	100	100	100	100	100	100	100	75	100	100	100	100
GAA	YWHAZ	100	100	100	100	100	100	100	100	100	100	100	100
GAA	FUCA1	100	100	100	83.33	100	100	100	100	100	100	100	100
GAA	ANXA5	92.86	92.86	100	92.86	92.86	92.86	100	92.86	85.71	100	92.86	100
GAA	IGKC	70	90	80	90	90	80	80	80	80	90	80	80
GAA	CNDP2	100	100	100	100	100	90	100	90	90	100	100	100
GAA	HPX	100	100	100	100	100	100	100	100	100	100	100	100
GAA	ANXA3	100	100	100	100	100	100	100	100	100	100	100	100
GAA	ENO1	100	100	100	100	100	100	100	100	100	100	100	100
GAA	PRDX6	100	100	100	100	100	100	100	100	83.33	100	100	100
GAA	ELSPBP1	25	100	87.5	75	100	100	100	62.5	87.5	62.5	100	62.5
GAA	RNASE4	100	100	100	100	100	87.5	100	100	87.5	100	87.5	100
GAA	GDI2	100	100	100	100	100	100	100	100	100	100	100	100
GAA	EZR	100	100	100	100	100	100	100	100	100	100	100	100
GAA	GAPDH	100	100	100	87.5	100	100	100	87.5	100	100	100	100
GAA	ANXA1	100	100	100	100	100	100	100	100	83.33	100	100	100
GAA	CRTAC1	75	100	87.5	100	100	100	87.5	87.5	87.5	100	87.5	100
GAA	EDDM3B	100	83.33	100	100	100	100	100	83.33	100	100	83.33	83.33
GAA	PSMA2	100	100	100	100	100	100	100	100	100	100	100	100
GAA	GPI	100	100	75	75	100	100	100	75	100	100	75	100
GAA	LCP1	100	100	75	100	100	75	100	75	75	75	75	75
GAA	HSP90AA1	100	100	100	100	100	100	100	100	100	100	100	75
GAA	DBI	100	100	100	100	100	100	100	100	83.33	100	100	100
GAA	IGHG1	50	100	100	100	100	100	100	100	100	100	100	100
GAA	IGHG2	100	75	100	100	100	100	100	75	100	100	100	100
GAA	C3	50	100	100	100	100	100	100	100	100	100	100	100
TF	PTGDS	100	89.29	100	100	100	100	17.86	82.14	50	53.57	17.86	10.71
TF	S100A11	100	100	89.29	96.43	92.86	78.57	100	100	89.29	75	100	85.71
TF	HSPA8	78.57	96.43	82.14	92.86	96.43	100	100	100	92.86	92.86	89.29	96.43
TF	CNDP2	100	88.57	72.86	98.57	88.57	85.71	97.14	91.43	80	78.57	81.43	95.71
TF	EDDM3B	97.62	100	97.62	100	97.62	97.62	100	100	90.48	95.24	90.48	83.33
TF	PSMA2	75	82.14	71.43	82.14	75	78.57	96.43	89.29	71.43	67.86	78.57	64.29

TF	HSP90AA1	96.43	100	100	100	100	96.43	100	100	96.43	89.29	100	60.71
TF	IGHG1	21.43	92.86	96.43	78.57	67.86	96.43	64.29	92.86	67.86	71.43	64.29	67.86
TF	IGHG2	96.43	64.29	100	100	100	100	96.43	100	92.86	100	100	100
TF	C3	42.86	92.86	100	92.86	92.86	92.86	89.29	92.86	89.29	92.86	82.14	85.71
CD9	PARK7	100	100	100	100	100	100	100	83.33	50	100	100	50
CD9	LAMB2	100	100	100	90	100	90	100	100	50	100	100	90
CD9	PTGDS	100	100	100	100	100	100	50	100	50	50	0	50
CD9	CST3	100	100	90	90	100	70	100	100	40	100	80	90
CD9	APOH	100	100	100	100	100	100	100	100	50	100	100	100
CD9	LSAMP	100	87.5	100	100	100	87.5	100	87.5	50	100	100	75
CD9	SPOCK1	100	100	100	100	100	100	83.33	83.33	50	100	100	100
CD9	APLP2	100	75	75	100	100	100	100	100	50	100	100	50
CD9	HSPA8	100	100	100	100	100	100	100	100	50	100	100	75
CD9	SOD1	100	100	100	100	100	75	100	100	50	100	100	100
CD9	YWHAZ	75	100	100	75	75	100	100	100	50	100	100	75
CD9	FUCA1	100	100	100	100	100	100	100	100	66.67	100	100	100
CD9	ANXA5	100	92.86	92.86	92.86	100	100	92.86	100	50	100	85.71	100
CD9	IGKC	100	90	90	100	90	90	100	80	50	90	100	80
CD9	CNDP2	100	100	100	100	100	100	100	100	60	100	100	100
CD9	HPX	100	100	75	87.5	100	62.5	100	87.5	50	100	75	100
CD9	ANXA3	100	100	66.67	66.67	100	83.33	100	100	50	83.33	66.67	100
CD9	ENO1	100	100	100	100	100	100	100	100	62.5	100	100	100
CD9	PRDX6	66.67	100	100	100	100	83.33	100	100	50	83.33	83.33	83.33
CD9	ELSPBP1	75	100	100	75	100	87.5	100	75	50	75	100	62.5
CD9	PLA1A	100	100	87.5	100	100	75	100	100	50	100	87.5	100
CD9	RNASE4	100	100	100	100	100	100	100	87.5	50	100	100	100
CD9	GDI2	100	100	100	100	100	100	100	100	50	100	100	100
CD9	EZR	100	100	100	100	100	100	100	100	50	100	100	100
CD9	GAPDH	100	100	100	100	100	100	100	100	50	100	100	100
CD9	ANXA1	100	83.33	100	100	83.33	100	100	100	50	83.33	100	66.67
CD9	HEXA	100	100	100	100	100	100	100	100	50	100	100	100
CD9	CRTAC1	100	100	100	100	100	87.5	100	100	62.5	100	100	100
CD9	TWSG1	100	100	100	100	100	75	100	100	25	100	100	100
CD9	PSMA2	100	100	100	100	100	100	100	100	50	100	100	100
CD9	GPI	100	100	100	100	100	100	100	100	50	100	100	100
CD9	LCP1	100	100	75	100	100	100	100	75	50	75	75	75
CD9	HSP90AA1	100	100	100	100	100	100	75	100	50	100	100	75
CD9	CTSL	100	100	100	100	100	100	100	100	50	100	100	100
CD9	DBI	100	100	100	100	100	100	100	100	50	100	100	66.67
CD9	IGHG1	75	100	100	100	100	100	100	100	50	100	100	100
CD9	TMPRSS2	100	100	100	100	100	100	100	100	50	100	100	75

CD9	C3	100	100	100	100	100	100	100	100	50	100	100	100
TGM4	PTGDS	100	95	95	100	100	100	65	87.5	72.5	30	67.5	65
TGM4	IGHG2	100	97.5	97.5	100	100	100	95	100	97.5	95	100	100
TGM4	C3	90	97.5	95	100	100	100	97.5	100	100	87.5	97.5	100
LGALS3BP	PARK7	100	83.33	76.67	93.33	96.67	93.33	96.67	83.33	93.33	86.67	86.67	76.67
LGALS3BP	SPOCK1	93.33	93.33	80	96.67	100	90	100	83.33	93.33	90	83.33	90
LGALS3BP	YWHAZ	100	95	90	90	85	90	95	80	70	90	85	95
LGALS3BP	ANXA5	94.29	85.71	87.14	95.71	95.71	95.71	94.29	87.14	85.71	90	92.86	97.14
LGALS3BP	IGKC	86	86	82	92	86	82	92	86	74	80	78	86
LGALS3BP	HPX	97.5	95	92.5	90	95	92.5	95	65	87.5	85	85	87.5
PARK7	PTGDS	100	100	100	100	100	100	0	83.33	66.67	16.67	0	50
PARK7	APOH	83.33	100	100	100	100	100	100	100	100	100	100	100
PARK7	LSAMP	100	100	75	66.67	100	91.67	100	91.67	66.67	91.67	83.33	83.33
PARK7	S100A11	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	SPOCK1	77.78	88.89	77.78	100	88.89	100	77.78	77.78	88.89	88.89	77.78	88.89
PARK7	APLP2	100	83.33	66.67	83.33	83.33	66.67	100	83.33	33.33	83.33	100	66.67
PARK7	HSPA8	100	100	100	100	100	100	83.33	100	100	83.33	100	100
PARK7	FSTL1	100	100	100	100	100	100	88.89	100	88.89	100	100	100
PARK7	QSOX1	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	SOD1	100	66.67	83.33	83.33	83.33	83.33	100	100	66.67	83.33	83.33	83.33
PARK7	YWHAZ	100	66.67	100	100	66.67	66.67	83.33	100	50	100	100	100
PARK7	LDHA	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	ANXA5	90.48	95.24	100	100	100	95.24	95.24	100	90.48	100	90.48	100
PARK7	IGKC	93.33	100	100	100	93.33	73.33	100	93.33	93.33	93.33	100	93.33
PARK7	CNDP2	100	86.67	86.67	93.33	100	86.67	100	93.33	73.33	100	100	93.33
PARK7	HPX	91.67	83.33	58.33	100	91.67	83.33	83.33	66.67	75	66.67	50	50
PARK7	ANXA3	100	100	88.89	100	100	100	100	100	88.89	100	100	77.78
PARK7	ENO1	100	100	100	100	100	100	100	100	91.67	100	100	100
PARK7	SORD	100	83.33	75	91.67	100	83.33	91.67	91.67	75	83.33	100	83.33
PARK7	PRDX6	77.78	66.67	88.89	88.89	88.89	55.56	100	100	55.56	88.89	100	88.89
PARK7	ELSPBP1	16.67	100	66.67	50	100	100	66.67	66.67	75	25	83.33	50
PARK7	RNASE4	100	100	91.67	100	100	91.67	100	100	83.33	100	100	100
PARK7	GDI2	100	100	100	100	100	100	100	66.67	100	100	66.67	66.67
PARK7	EZR	100	100	100	83.33	100	100	100	100	100	100	100	100
PARK7	GAPDH	100	91.67	100	100	100	83.33	100	100	83.33	100	100	83.33
PARK7	ANXA1	100	100	100	100	88.89	55.56	100	100	55.56	88.89	100	100
PARK7	EDDM3B	100	100	100	100	100	100	100	100	100	100	77.78	100
PARK7	PSMA2	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	GPI	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	LCP1	66.67	83.33	66.67	100	100	66.67	66.67	66.67	83.33	66.67	66.67	83.33
PARK7	HSP90AA1	100	100	100	100	100	100	83.33	100	100	100	100	100

PARK7	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	DBI	100	77.78	100	100	100	77.78	88.89	77.78	77.78	100	100	88.89
PARK7	IGHG1	50	100	100	100	83.33	100	100	100	83.33	100	100	50
PARK7	IGHG2	100	100	100	100	100	100	83.33	100	100	100	100	100
PARK7	TMPRSS2	100	100	100	100	100	100	100	100	100	100	100	100
PARK7	C3	100	100	100	100	100	100	100	100	100	100	66.67	100
MME	YWHAZ	100	100	97.06	100	73.53	70.59	70.59	94.12	70.59	100	76.47	97.06
MME	HPX	83.82	95.59	79.41	98.53	98.53	97.06	100	82.35	48.53	92.65	98.53	73.53
MME	EDDM3B	98.04	96.08	88.24	100	100	98.04	96.08	98.04	52.94	98.04	60.78	96.08
MME	HSP90AA1	94.12	97.06	97.06	97.06	97.06	79.41	58.82	97.06	85.29	91.18	67.65	91.18
LAMB2	SPOCK1	100	100	93.33	100	100	100	100	100	93.33	100	100	100
LAMB2	ANXA3	100	100	86.67	86.67	93.33	93.33	93.33	86.67	80	86.67	100	100
LAMB2	PRDX6	80	86.67	86.67	93.33	93.33	80	100	100	60	86.67	100	86.67
LAMB2	GDI2	100	100	90	100	90	80	100	100	70	80	100	80
PTGDS	APOH	75	100	100	100	100	100	100	100	100	100	100	100
PTGDS	LSAMP	100	62.5	87.5	100	100	100	100	100	100	100	100	100
PTGDS	S100A11	100	100	50	100	100	75	75	100	75	50	75	100
PTGDS	SPOCK1	83.33	33.33	66.67	100	100	100	100	100	100	100	83.33	100
PTGDS	HSPA8	100	100	50	100	100	100	100	100	100	100	100	100
PTGDS	FSTL1	66.67	50	33.33	100	100	100	83.33	100	83.33	83.33	66.67	83.33
PTGDS	SOD1	100	100	50	100	100	100	75	100	100	100	75	100
PTGDS	YWHAZ	100	100	100	100	100	100	100	100	100	100	100	100
PTGDS	LDHA	100	75	50	100	100	100	75	100	50	50	50	100
PTGDS	CNDP2	100	80	60	100	100	100	100	100	90	80	80	100
PTGDS	HPX	87.5	37.5	100	100	100	100	87.5	87.5	100	87.5	62.5	87.5
PTGDS	ENO1	100	75	100	100	100	100	100	100	100	100	87.5	100
PTGDS	ORM2	75	100	37.5	100	87.5	100	75	100	62.5	50	62.5	75
PTGDS	SORD	87.5	25	50	100	100	87.5	87.5	100	75	75	62.5	100
PTGDS	PRDX6	100	100	16.67	100	100	100	83.33	100	83.33	66.67	33.33	100
PTGDS	EZR	100	100	100	100	100	100	100	100	75	75	75	100
PTGDS	GAPDH	100	75	75	100	100	100	87.5	100	100	87.5	87.5	100
PTGDS	ANXA1	100	50	66.67	100	100	100	100	100	83.33	83.33	83.33	100
PTGDS	EDDM3B	100	83.33	83.33	100	100	100	66.67	83.33	100	66.67	83.33	100
PTGDS	PSMA2	50	75	25	100	100	100	75	75	50	50	50	100
PTGDS	GPI	100	50	100	100	100	100	100	100	100	100	75	100
PTGDS	LCP1	100	75	25	100	100	100	100	100	100	100	100	100
PTGDS	HSP90AA1	75	100	0	100	75	75	75	75	50	50	50	75
PTGDS	DBI	100	83.33	50	100	100	100	100	100	83.33	83.33	66.67	100
PTGDS	IGHG1	50	100	100	100	75	100	50	100	50	50	50	50
PTGDS	IGHG2	100	50	100	100	100	100	100	100	100	75	75	100
PTGDS	C3	0	75	50	100	100	100	50	50	100	50	50	25

CPM	PLA1A	100	100	100	100	100	75	100	87.5	87.5	100	100	75
CST3	LSAMP	100	100	70	70	100	90	95	65	95	100	65	50
CST3	SPOCK1	100	100	60	93.33	100	80	100	53.33	73.33	93.33	46.67	53.33
CST3	SOD1	100	90	80	80	90	80	100	80	80	80	40	80
CST3	IGKC	92	88	76	88	88	88	96	40	76	92	48	84
CST3	HPX	100	95	100	100	95	85	100	95	90	95	80	85
CST3	ANXA3	86.67	86.67	100	100	93.33	93.33	100	46.67	86.67	86.67	80	73.33
CST3	PRDX6	93.33	80	53.33	86.67	80	66.67	86.67	93.33	60	80	73.33	53.33
CST3	RNASE4	100	95	85	85	100	95	100	55	95	85	65	70
CST3	GDI2	100	100	60	90	90	70	90	100	80	90	80	60
CST3	ANXA1	93.33	100	53.33	86.67	80	73.33	100	60	73.33	86.67	60	46.67
CST3	CTSL	100	100	90	80	100	90	100	100	90	100	90	90
CST3	TMPRSS2	100	100	60	100	100	100	90	80	90	100	50	50
APOH	LSAMP	100	75	75	100	100	100	75	100	87.5	100	100	50
APOH	S100A11	100	100	75	100	100	100	50	75	100	100	100	50
APOH	SPOCK1	100	100	83.33	100	100	100	100	100	100	100	100	83.33
APOH	APLP2	100	75	50	100	100	100	100	100	100	100	100	25
APOH	HSPA8	100	100	50	100	100	100	50	100	100	100	100	50
APOH	FSTL1	100	83.33	83.33	100	100	100	66.67	50	83.33	100	100	50
APOH	SOD1	100	100	100	100	100	75	100	100	75	75	100	100
APOH	YWHAZ	100	100	100	100	100	100	50	100	100	100	100	25
APOH	LDHA	100	100	100	100	100	100	100	100	100	100	100	50
APOH	ANXA5	100	78.57	92.86	100	92.86	92.86	50	92.86	85.71	100	92.86	64.29
APOH	CNDP2	100	70	80	100	100	100	80	100	100	100	100	70
APOH	HPX	100	75	100	100	100	100	100	100	87.5	100	100	100
APOH	ANXA3	100	83.33	100	100	83.33	83.33	33.33	100	100	83.33	100	100
APOH	ENO1	100	87.5	100	100	100	100	100	100	87.5	100	100	87.5
APOH	SORD	100	50	50	100	100	100	75	87.5	100	100	100	37.5
APOH	PRDX6	100	100	83.33	100	100	100	83.33	100	66.67	100	100	66.67
APOH	TIMP1	100	93.75	93.75	87.5	93.75	100	93.75	87.5	100	100	100	75
APOH	EZR	100	100	100	50	100	100	50	100	100	100	100	75
APOH	GAPDH	100	100	100	100	100	100	75	100	100	100	100	75
APOH	ANXA1	100	83.33	83.33	100	100	100	83.33	100	100	100	100	50
APOH	PSMA2	100	100	50	100	100	100	50	100	100	100	100	25
APOH	GPI	100	75	100	100	100	100	100	100	100	75	100	100
APOH	LCP1	100	100	50	100	100	100	75	50	75	100	75	25
APOH	HSP90AA1	100	100	75	100	100	100	50	100	100	100	75	25
APOH	CTSL	100	100	100	100	100	100	50	100	75	100	100	100
APOH	DBI	100	100	83.33	100	100	100	83.33	66.67	83.33	100	100	50
APOH	IGHG1	100	100	100	100	100	100	100	100	100	100	100	100
APOH	C3	100	100	100	100	100	100	100	100	100	100	100	100

LSAMP	SPOCK1	100	100	91.67	83.33	100	91.67	100	83.33	75	91.67	75	83.33
LSAMP	APLP2	87.5	87.5	37.5	100	87.5	87.5	100	75	62.5	87.5	75	75
LSAMP	HSPA8	100	100	87.5	100	100	100	100	100	100	87.5	87.5	100
LSAMP	SOD1	100	100	87.5	100	100	100	100	100	87.5	100	87.5	87.5
LSAMP	YWHAZ	100	87.5	87.5	87.5	62.5	75	100	100	50	100	75	87.5
LSAMP	ANXA5	100	96.43	92.86	96.43	92.86	92.86	100	100	85.71	100	92.86	92.86
LSAMP	IGKC	90	100	90	100	95	80	95	90	90	95	85	85
LSAMP	CNDP2	100	100	95	100	100	100	100	100	80	100	100	95
LSAMP	HPX	100	100	81.25	81.25	87.5	87.5	75	81.25	68.75	81.25	87.5	68.75
LSAMP	ANXA3	100	100	91.67	91.67	100	100	100	100	91.67	100	91.67	91.67
LSAMP	ENO1	93.75	93.75	87.5	75	93.75	93.75	93.75	93.75	81.25	87.5	87.5	87.5
LSAMP	PRDX6	100	91.67	83.33	100	100	91.67	100	100	66.67	91.67	100	100
LSAMP	ELSPBP1	37.5	100	81.25	68.75	100	100	87.5	68.75	87.5	37.5	68.75	43.75
LSAMP	RNASE4	100	100	87.5	100	100	93.75	100	87.5	93.75	93.75	81.25	93.75
LSAMP	GDI2	100	100	100	100	100	100	100	87.5	100	100	100	87.5
LSAMP	TIMP1	96.88	100	90.62	96.88	100	90.62	96.88	96.88	87.5	90.62	96.88	90.62
LSAMP	EZR	100	100	100	100	100	100	100	100	100	100	100	100
LSAMP	GAPDH	93.75	93.75	87.5	81.25	87.5	81.25	100	87.5	81.25	93.75	81.25	75
LSAMP	ANXA1	91.67	100	75	91.67	83.33	83.33	91.67	91.67	58.33	75	91.67	83.33
LSAMP	CRTAC1	100	100	87.5	93.75	100	87.5	93.75	93.75	81.25	87.5	81.25	87.5
LSAMP	EDDM3B	100	100	100	100	100	100	100	100	100	100	75	100
LSAMP	PSMA2	100	100	100	100	100	100	100	100	100	100	100	100
LSAMP	GPI	100	100	100	100	100	100	100	100	100	100	100	100
LSAMP	LCP1	87.5	100	75	87.5	100	87.5	100	75	87.5	75	50	75
LSAMP	HSP90AA1	100	100	100	100	100	100	100	100	100	100	87.5	87.5
LSAMP	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
LSAMP	DBI	100	100	91.67	83.33	100	83.33	91.67	91.67	83.33	91.67	83.33	91.67
LSAMP	IGHG1	50	100	87.5	100	100	100	87.5	100	87.5	100	100	75
LSAMP	TMPRSS2	100	100	87.5	100	100	100	100	100	87.5	100	87.5	87.5
S100A11	SPOCK1	100	100	100	100	100	83.33	100	100	100	83.33	100	100
S100A11	HSPA8	100	100	100	100	75	75	100	100	100	75	100	100
S100A11	FSTL1	83.33	83.33	100	100	100	83.33	100	100	83.33	50	100	100
S100A11	YWHAZ	100	100	100	100	100	100	100	75	75	100	100	100
S100A11	LDHA	100	100	100	100	100	100	100	100	100	100	100	100
S100A11	CNDP2	100	80	100	100	100	90	90	90	80	100	90	100
S100A11	HPX	100	100	87.5	100	100	25	100	87.5	75	62.5	75	87.5
S100A11	EDDM3B	100	100	100	100	100	83.33	100	100	83.33	100	83.33	100
S100A11	PSMA2	75	100	100	100	100	75	100	100	100	100	100	100
S100A11	HSP90AA1	100	100	100	100	75	100	100	100	100	100	100	100
S100A11	IGHG1	0	75	75	75	25	75	75	75	75	75	75	75
S100A11	IGHG2	100	75	100	100	100	75	100	100	100	100	100	100

S100A11	C3	50	100	100	100	75	50	75	75	100	50	75	75
SPOCK1	APLP2	100	100	50	83.33	100	100	83.33	100	83.33	100	83.33	66.67
SPOCK1	HSPA8	100	100	100	100	100	100	83.33	100	100	100	100	100
SPOCK1	FSTL1	100	100	100	100	100	100	88.89	100	100	100	100	100
SPOCK1	PATE1	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	SOD1	100	83.33	100	100	100	100	100	100	100	100	100	83.33
SPOCK1	YWHAZ	100	100	100	100	83.33	83.33	66.67	83.33	83.33	100	100	100
SPOCK1	FUCA1	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	ANXA5	100	90.48	100	100	95.24	100	90.48	95.24	100	100	90.48	100
SPOCK1	IGKC	93.33	100	93.33	93.33	93.33	73.33	93.33	93.33	86.67	86.67	93.33	86.67
SPOCK1	CNDP2	100	100	100	100	100	93.33	93.33	100	93.33	100	100	100
SPOCK1	HPX	100	100	75	100	100	91.67	100	75	83.33	100	58.33	75
SPOCK1	ANXA3	100	100	88.89	100	100	100	100	100	100	100	88.89	88.89
SPOCK1	ENO1	91.67	100	91.67	100	91.67	83.33	100	91.67	83.33	91.67	83.33	91.67
SPOCK1	PRDX6	88.89	88.89	88.89	88.89	88.89	88.89	66.67	88.89	66.67	88.89	77.78	88.89
SPOCK1	ELSPBP1	41.67	100	83.33	41.67	100	91.67	75	66.67	66.67	33.33	100	50
SPOCK1	RNASE4	100	100	100	100	100	83.33	100	91.67	83.33	83.33	91.67	100
SPOCK1	GDI2	100	100	83.33	100	100	100	83.33	83.33	100	100	83.33	83.33
SPOCK1	EZR	100	100	100	66.67	100	100	83.33	100	100	100	100	100
SPOCK1	GAPDH	100	100	91.67	100	100	91.67	100	83.33	91.67	100	75	83.33
SPOCK1	ANXA1	100	100	100	100	88.89	88.89	100	100	77.78	88.89	88.89	100
SPOCK1	CRTAC1	100	100	91.67	100	100	83.33	91.67	100	66.67	100	83.33	91.67
SPOCK1	EDDM3B	100	100	100	100	100	100	100	100	100	100	88.89	100
SPOCK1	PSMA2	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	GPI	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	LCP1	83.33	100	83.33	100	100	83.33	83.33	100	83.33	83.33	100	100
SPOCK1	HSP90AA1	100	100	100	100	100	100	83.33	100	100	100	100	100
SPOCK1	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	DBI	100	100	88.89	100	100	88.89	77.78	100	77.78	88.89	88.89	88.89
SPOCK1	IGHG1	66.67	100	100	100	100	100	100	100	100	100	100	66.67
SPOCK1	IGHG2	100	100	100	100	100	100	83.33	100	100	100	100	100
SPOCK1	TMPRSS2	100	100	100	100	100	100	100	100	100	100	100	100
SPOCK1	C3	100	100	100	100	100	100	100	100	100	100	100	100
APLP2	HSPA8	100	100	75	100	100	100	100	100	100	75	100	75
APLP2	SOD1	75	100	75	100	100	100	100	100	50	75	75	75
APLP2	YWHAZ	75	100	50	75	50	75	75	75	75	75	100	75
APLP2	ANXA5	92.86	92.86	78.57	100	92.86	100	100	100	71.43	100	92.86	85.71
APLP2	IGKC	100	100	100	100	100	80	100	100	80	90	90	90
APLP2	HPX	100	100	25	75	87.5	37.5	87.5	62.5	25	75	50	12.5
APLP2	ANXA3	100	100	83.33	100	100	100	100	100	100	100	100	66.67
APLP2	PRDX6	100	100	66.67	100	100	100	100	100	66.67	100	100	83.33

APLP2	ELSPBP1	37.5	100	50	50	100	62.5	75	75	25	12.5	75	37.5
APLP2	RNASE4	100	100	100	100	100	100	100	100	75	100	100	100
APLP2	GDI2	100	100	75	100	100	75	100	75	50	100	100	50
APLP2	EZR	100	100	75	100	100	100	100	100	100	100	100	100
APLP2	GAPDH	87.5	75	50	87.5	75	100	100	75	75	62.5	100	50
APLP2	ANXA1	83.33	100	83.33	100	83.33	83.33	100	100	83.33	83.33	100	100
APLP2	GPI	100	100	100	100	100	100	100	100	100	100	100	100
APLP2	LCP1	50	100	75	100	100	100	75	100	50	75	50	75
APLP2	HSP90AA1	100	100	100	100	100	100	100	100	100	100	100	100
APLP2	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
APLP2	DBI	83.33	100	83.33	83.33	100	100	100	100	33.33	100	100	83.33
APLP2	IGHG1	50	100	50	100	75	100	50	75	100	100	100	25
HSPA8	FSTL1	100	83.33	100	100	100	100	100	83.33	83.33	100	100	100
HSPA8	SOD1	100	100	100	100	100	75	100	100	75	75	100	100
HSPA8	YWHAZ	100	100	100	100	100	100	100	100	100	100	100	100
HSPA8	LDHA	100	100	100	100	100	100	100	100	100	100	100	100
HSPA8	ANXA5	100	85.71	92.86	100	100	100	92.86	100	78.57	100	100	92.86
HSPA8	CNDP2	100	100	90	100	100	100	90	100	90	90	100	100
HSPA8	HPX	100	87.5	100	100	87.5	50	100	100	75	100	100	100
HSPA8	ANXA3	100	83.33	83.33	100	83.33	83.33	100	100	100	83.33	100	100
HSPA8	ENO1	100	100	100	100	100	87.5	100	100	87.5	100	100	100
HSPA8	ORM2	100	100	87.5	87.5	100	87.5	87.5	100	100	62.5	87.5	100
HSPA8	SORD	100	50	87.5	100	87.5	100	100	100	100	100	100	87.5
HSPA8	PRDX6	100	100	100	100	100	100	100	100	66.67	100	100	100
HSPA8	PEBP4	100	87.5	87.5	100	100	100	87.5	62.5	87.5	100	100	75
HSPA8	EZR	100	100	75	75	100	100	100	100	75	100	100	100
HSPA8	GAPDH	100	100	100	100	100	100	100	100	100	100	100	100
HSPA8	ANXA1	100	100	100	100	100	100	100	100	83.33	100	100	100
HSPA8	EDDM3B	100	100	100	100	100	100	100	100	100	100	66.67	100
HSPA8	PSMA2	100	100	100	100	100	100	100	100	100	75	100	75
HSPA8	GPI	100	75	75	100	75	50	75	100	75	75	100	75
HSPA8	LCP1	100	100	100	100	100	100	100	75	100	100	75	100
HSPA8	HSP90AA1	100	100	100	100	100	100	100	100	100	100	100	75
HSPA8	DBI	100	100	100	100	100	100	100	83.33	83.33	100	100	100
HSPA8	IGHG1	50	100	100	100	75	100	75	100	75	100	100	50
HSPA8	IGHG2	100	75	100	100	100	100	100	100	100	100	100	100
HSPA8	C3	75	100	100	100	100	100	75	100	100	50	75	75
FSTL1	YWHAZ	100	100	100	100	100	100	100	83.33	66.67	100	100	100
FSTL1	LDHA	83.33	100	100	100	100	100	100	100	83.33	100	100	66.67
FSTL1	CNDP2	93.33	86.67	100	100	100	93.33	93.33	86.67	80	93.33	93.33	100
FSTL1	HPX	91.67	91.67	66.67	100	91.67	75	91.67	58.33	41.67	91.67	58.33	66.67

FSTL1	TIMP1	100	100	100	95.83	95.83	100	100	100	79.17	95.83	95.83	100
FSTL1	EDDM3B	100	100	100	100	100	100	100	100	55.56	100	77.78	100
FSTL1	PSMA2	100	100	83.33	100	100	100	100	100	100	100	100	100
FSTL1	LCP1	100	100	83.33	100	100	100	83.33	100	66.67	66.67	83.33	100
FSTL1	HSP90AA1	100	83.33	100	100	100	100	66.67	100	83.33	100	83.33	83.33
FSTL1	IGHG1	50	100	100	100	83.33	100	100	100	100	100	100	83.33
FSTL1	IGHG2	100	66.67	100	100	100	100	66.67	100	83.33	100	100	100
FSTL1	C3	50	83.33	83.33	100	100	100	83.33	83.33	50	83.33	66.67	16.67
QSOX1	YWHAZ	100	100	100	100	100	100	100	100	80	80	100	80
QSOX1	HPX	100	100	100	100	100	100	100	100	100	80	100	100
QSOX1	CTSL	100	70	100	100	90	60	90	70	40	20	60	70
PATE1	HPX	100	100	100	100	87.5	100	100	100	100	100	100	100
SOD1	YWHAZ	100	100	100	100	100	100	100	100	50	100	100	75
SOD1	FUCA1	100	83.33	83.33	100	83.33	83.33	83.33	83.33	66.67	83.33	100	66.67
SOD1	ANXA5	92.86	85.71	92.86	100	92.86	85.71	92.86	100	71.43	78.57	92.86	85.71
SOD1	IGKC	70	80	90	100	80	60	100	50	70	90	90	90
SOD1	CNDP2	100	90	90	100	100	90	100	100	80	90	90	100
SOD1	HPX	87.5	37.5	50	87.5	87.5	62.5	87.5	50	62.5	87.5	50	75
SOD1	ANXA3	83.33	83.33	66.67	83.33	83.33	83.33	100	83.33	66.67	83.33	66.67	83.33
SOD1	ENO1	100	75	100	100	100	100	100	100	87.5	100	100	100
SOD1	PRDX6	100	100	83.33	100	100	50	100	100	50	16.67	50	66.67
SOD1	ELSPBP1	0	87.5	75	62.5	100	75	75	75	62.5	50	100	12.5
SOD1	RNASE4	100	100	100	100	100	87.5	100	75	75	87.5	100	75
SOD1	GDI2	100	75	75	100	100	50	75	75	50	75	50	75
SOD1	EZR	100	100	100	100	100	75	50	75	50	50	75	75
SOD1	GAPDH	87.5	87.5	87.5	87.5	87.5	87.5	100	87.5	75	87.5	75	87.5
SOD1	ANXA1	100	83.33	83.33	100	100	50	100	100	33.33	50	100	66.67
SOD1	CRTAC1	87.5	87.5	100	100	100	87.5	87.5	100	62.5	100	100	87.5
SOD1	PSMA2	100	100	75	100	100	100	100	100	100	100	100	100
SOD1	GPI	100	75	75	75	75	100	100	75	50	75	75	100
SOD1	LCP1	100	75	75	100	100	100	75	50	75	50	75	75
SOD1	HSP90AA1	100	100	75	100	100	100	50	100	100	50	100	25
SOD1	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
SOD1	DBI	83.33	83.33	66.67	83.33	83.33	66.67	83.33	66.67	66.67	83.33	83.33	33.33
SOD1	IGHG1	50	100	100	100	75	100	100	100	100	100	100	100
SOD1	TMPRSS2	100	100	100	100	100	100	100	100	100	100	100	75
YWHAZ	LDHA	100	100	100	100	100	100	100	100	100	100	100	100
YWHAZ	ANXA5	92.86	78.57	92.86	100	92.86	100	100	92.86	50	100	85.71	92.86
YWHAZ	IGKC	90	80	100	100	70	70	90	70	70	90	90	90
YWHAZ	CNDP2	100	80	90	80	90	100	90	100	90	100	80	100
YWHAZ	HPX	75	62.5	25	100	0	12.5	25	75	0	62.5	25	25

YWHAZ	ANXA3	83.33	83.33	83.33	100	66.67	66.67	83.33	100	66.67	83.33	66.67	66.67
YWHAZ	ENO1	87.5	87.5	100	100	75	75	100	100	50	87.5	100	87.5
YWHAZ	SORD	75	37.5	50	87.5	50	75	100	75	87.5	75	87.5	62.5
YWHAZ	PRDX6	100	100	83.33	100	83.33	83.33	100	100	33.33	83.33	83.33	83.33
YWHAZ	ELSPBP1	0	87.5	75	50	75	50	100	62.5	12.5	0	87.5	50
YWHAZ	RNASE4	100	100	100	100	87.5	87.5	100	87.5	75	100	100	100
YWHAZ	GDI2	100	75	100	100	100	50	100	100	0	100	100	100
YWHAZ	TIMP1	93.75	81.25	100	81.25	50	68.75	93.75	87.5	50	93.75	93.75	81.25
YWHAZ	EZR	100	100	100	50	100	100	100	100	75	100	100	100
YWHAZ	GAPDH	100	100	100	100	75	87.5	87.5	100	50	100	100	100
YWHAZ	ANXA1	100	83.33	100	100	100	100	100	100	66.67	83.33	100	100
YWHAZ	EDDM3B	100	100	100	100	100	100	100	100	33.33	100	66.67	100
YWHAZ	PSMA2	100	100	100	100	100	100	100	100	75	100	100	100
YWHAZ	GPI	100	100	100	100	100	75	100	100	50	100	100	100
YWHAZ	LCP1	100	75	75	100	50	100	100	25	50	50	75	100
YWHAZ	HSP90AA1	100	100	100	100	50	100	100	100	100	100	100	100
YWHAZ	CTSL	100	100	100	100	100	100	100	100	75	100	100	100
YWHAZ	DBI	100	100	100	100	66.67	100	100	33.33	50	100	100	100
YWHAZ	IGHG1	0	100	75	100	0	100	75	100	100	100	75	25
YWHAZ	IGHG2	100	50	100	100	100	100	100	100	75	100	100	100
YWHAZ	TMPRSS2	100	100	100	100	100	100	100	100	50	100	100	100
YWHAZ	C3	50	100	100	100	100	100	100	100	50	100	50	100
LDHA	CNDP2	100	100	100	90	100	100	90	90	80	100	100	90
LDHA	HPX	87.5	87.5	87.5	100	100	75	87.5	87.5	62.5	100	100	87.5
LDHA	PSMA2	100	100	75	100	100	100	100	100	100	100	100	100
LDHA	LCP1	100	100	50	100	100	100	100	50	75	100	50	100
LDHA	HSP90AA1	75	75	50	75	75	75	75	75	100	75	25	50
LDHA	IGHG1	50	100	100	100	75	100	100	100	100	100	100	75
LDHA	C3	25	100	100	100	100	100	75	100	75	50	75	50
LRG1	OS9	83.33	100	100	83.33	83.33	83.33	100	83.33	83.33	83.33	83.33	66.67
LRG1	SDF4	100	100	100	100	100	100	100	100	50	83.33	100	83.33
LRG1	CTSL	100	83.33	100	100	83.33	100	100	100	66.67	100	83.33	66.67
LRG1	TMPRSS2	100	100	83.33	100	100	100	100	100	83.33	100	83.33	33.33
FUCA1	HPX	100	91.67	100	100	100	100	100	100	100	100	100	100
FUCA1	ANXA3	100	88.89	88.89	100	88.89	100	88.89	100	100	100	100	88.89
FUCA1	PRDX6	100	77.78	100	100	77.78	88.89	77.78	100	66.67	77.78	100	100
FUCA1	RNASE4	91.67	83.33	91.67	91.67	91.67	91.67	91.67	75	83.33	83.33	75	100
FUCA1	GDI2	100	100	100	100	83.33	100	100	83.33	100	100	100	83.33
FUCA1	ANXA1	100	100	100	100	88.89	88.89	100	100	77.78	88.89	100	100
FUCA1	HEXA	100	88.89	100	100	100	66.67	100	100	77.78	100	100	100
ANXA5	IGKC	91.43	88.57	91.43	94.29	68.57	77.14	94.29	71.43	77.14	77.14	74.29	88.57

ANXA5	CNDP2	100	100	91.43	97.14	97.14	100	94.29	100	94.29	100	94.29	100
ANXA5	HPX	92.86	96.43	85.71	100	78.57	60.71	82.14	89.29	57.14	78.57	100	82.14
ANXA5	ANXA3	100	95.24	100	95.24	95.24	85.71	100	90.48	71.43	80.95	100	95.24
ANXA5	ENO1	92.86	96.43	100	100	89.29	92.86	100	100	82.14	92.86	96.43	100
ANXA5	PRDX6	95.24	95.24	95.24	95.24	100	90.48	100	100	57.14	100	100	95.24
ANXA5	RNASE4	96.43	96.43	96.43	96.43	96.43	96.43	100	78.57	78.57	92.86	82.14	96.43
ANXA5	GDI2	100	100	100	100	100	85.71	100	100	64.29	100	100	92.86
ANXA5	EZR	100	92.86	100	57.14	85.71	92.86	100	100	71.43	92.86	100	100
ANXA5	GAPDH	96.43	100	92.86	100	89.29	96.43	100	96.43	85.71	89.29	82.14	100
ANXA5	ANXA1	100	95.24	95.24	100	100	100	100	100	76.19	100	90.48	90.48
ANXA5	CRTAC1	96.43	100	96.43	96.43	89.29	85.71	92.86	100	50	82.14	96.43	96.43
ANXA5	EDDM3B	95.24	90.48	100	100	95.24	100	100	100	71.43	100	52.38	90.48
ANXA5	GPI	100	85.71	85.71	100	78.57	78.57	92.86	100	78.57	85.71	78.57	85.71
ANXA5	LCP1	92.86	100	64.29	100	100	100	100	50	78.57	85.71	42.86	85.71
ANXA5	CTSL	100	100	100	92.86	100	85.71	100	100	85.71	100	100	100
ANXA5	DBI	100	100	95.24	100	95.24	100	100	90.48	61.9	100	71.43	80.95
ANXA5	IGHG1	35.71	100	92.86	100	42.86	100	85.71	92.86	100	100	100	85.71
ANXA5	C3	64.29	100	100	100	85.71	100	92.86	100	85.71	64.29	64.29	78.57
IGKC	HPX	100	95	100	100	95	85	100	80	75	80	85	90
IGKC	ANXA3	93.33	80	93.33	93.33	80	80	86.67	93.33	80	73.33	73.33	60
IGKC	PRDX6	86.67	86.67	93.33	100	86.67	80	86.67	86.67	66.67	86.67	86.67	46.67
IGKC	RNASE4	90	90	95	100	95	90	95	100	95	90	100	50
IGKC	GDI2	100	90	100	100	80	80	100	70	50	70	70	50
IGKC	EZR	90	80	90	100	90	80	80	90	80	90	80	50
IGKC	GAPDH	85	95	100	95	95	85	85	90	75	80	100	45
IGKC	ANXA1	80	93.33	93.33	100	80	80	93.33	100	80	86.67	93.33	53.33
IGKC	GPI	90	80	80	80	100	60	80	100	60	70	80	70
IGKC	LCP1	70	90	60	100	80	80	70	80	90	70	80	40
IGKC	DBI	93.33	93.33	93.33	100	93.33	80	86.67	100	86.67	93.33	100	46.67
CNDP2	HPX	90	90	85	100	90	45	95	100	40	80	100	85
CNDP2	ANXA3	100	100	100	100	100	73.33	100	100	80	80	100	100
CNDP2	ENO1	90	95	100	95	95	70	100	100	60	95	100	100
CNDP2	ORM2	100	100	95	100	80	70	95	95	75	75	75	95
CNDP2	SORD	100	65	90	100	95	85	100	95	70	90	90	85
CNDP2	PRDX6	100	100	100	100	100	100	100	100	66.67	100	100	100
CNDP2	TIMP1	97.5	92.5	90	100	90	62.5	92.5	92.5	45	82.5	95	90
CNDP2	EZR	100	100	90	100	100	90	90	100	80	90	100	100
CNDP2	GAPDH	100	100	100	100	95	85	100	100	75	85	95	100
CNDP2	ANXA1	100	100	100	100	100	100	100	100	100	100	100	100
CNDP2	EDDM3B	100	100	93.33	100	86.67	80	100	100	46.67	100	73.33	86.67
CNDP2	PSMA2	100	100	90	100	100	90	100	100	70	100	90	100

CNDP2	LCP1	100	100	80	100	100	90	90	60	60	90	40	80
CNDP2	HSP90AA1	100	100	90	100	80	100	70	100	80	100	60	80
CNDP2	DBI	100	100	100	100	93.33	93.33	100	73.33	53.33	100	80	86.67
CNDP2	IGHG1	30	90	70	90	60	70	70	90	100	90	100	70
CNDP2	C3	60	100	100	100	90	70	90	100	60	50	70	80
VTN	PLA1A	100	100	100	75	100	100	100	100	75	87.5	100	87.5
VTN	TWSG1	100	100	100	75	100	100	100	100	50	100	100	100
HPX	ANXA3	100	100	100	100	100	100	91.67	58.33	91.67	91.67	100	83.33
HPX	ENO1	100	100	87.5	100	100	100	93.75	93.75	81.25	100	81.25	87.5
HPX	SORD	87.5	43.75	31.25	81.25	100	68.75	68.75	50	56.25	81.25	68.75	25
HPX	PRDX6	75	83.33	8.33	91.67	91.67	50	41.67	75	41.67	66.67	83.33	16.67
HPX	ELSPBP1	12.5	100	100	43.75	100	100	50	31.25	81.25	31.25	31.25	18.75
HPX	RNASE4	93.75	100	93.75	100	93.75	93.75	100	62.5	87.5	100	62.5	81.25
HPX	GDI2	100	100	75	100	100	87.5	87.5	87.5	100	100	100	62.5
HPX	TIMP1	93.75	100	62.5	81.25	96.88	93.75	87.5	75	87.5	87.5	87.5	65.62
HPX	EZR	100	87.5	100	50	87.5	87.5	50	62.5	75	87.5	100	75
HPX	GAPDH	87.5	93.75	50	93.75	100	75	62.5	68.75	75	87.5	50	62.5
HPX	ANXA1	83.33	91.67	41.67	91.67	66.67	33.33	83.33	75	33.33	66.67	58.33	25
HPX	CRTAC1	100	100	87.5	93.75	100	93.75	87.5	93.75	68.75	87.5	75	93.75
HPX	EDDM3B	100	100	100	100	100	100	100	83.33	100	100	41.67	41.67
HPX	PSMA2	100	100	50	100	100	100	75	75	100	100	37.5	37.5
HPX	GPI	100	100	100	100	100	100	87.5	50	87.5	100	62.5	100
HPX	LCP1	75	100	0	87.5	100	62.5	50	0	87.5	50	0	0
HPX	HSP90AA1	100	87.5	62.5	100	100	87.5	50	50	87.5	87.5	12.5	0
HPX	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
HPX	DBI	100	100	50	91.67	91.67	66.67	66.67	33.33	83.33	75	33.33	33.33
HPX	IGHG1	37.5	100	100	100	87.5	100	100	100	62.5	100	100	100
HPX	IGHG2	100	87.5	100	100	100	100	50	50	100	100	75	87.5
HPX	TMPRSS2	100	100	62.5	100	100	100	75	87.5	100	100	62.5	50
HPX	C3	75	87.5	100	87.5	100	100	100	75	100	75	50	100
ANXA3	ENO1	100	83.33	91.67	100	100	100	100	91.67	75	100	91.67	100
ANXA3	PRDX6	77.78	77.78	44.44	66.67	88.89	55.56	77.78	88.89	33.33	44.44	100	44.44
ANXA3	ELSPBP1	33.33	91.67	100	50	100	100	83.33	66.67	41.67	50	66.67	58.33
ANXA3	NUCB1	100	95.24	90.48	80.95	100	80.95	85.71	100	66.67	100	85.71	95.24
ANXA3	RNASE4	100	91.67	91.67	83.33	100	83.33	100	66.67	83.33	100	66.67	91.67
ANXA3	GDI2	100	50	66.67	83.33	100	66.67	83.33	66.67	33.33	66.67	100	66.67
ANXA3	TIMP1	100	100	83.33	95.83	100	100	100	100	87.5	95.83	100	87.5
ANXA3	EZR	83.33	66.67	83.33	33.33	83.33	100	66.67	83.33	83.33	100	83.33	83.33
ANXA3	GAPDH	83.33	100	66.67	91.67	100	91.67	100	91.67	91.67	83.33	75	91.67
ANXA3	ANXA1	100	77.78	55.56	88.89	100	44.44	100	100	55.56	66.67	77.78	55.56
ANXA3	CRTAC1	100	100	91.67	91.67	91.67	100	91.67	100	50	91.67	91.67	100

ANXA3	TWSG1	100	66.67	100	83.33	100	50	100	100	50	100	100	100
ANXA3	GPI	100	83.33	83.33	83.33	83.33	83.33	100	100	50	83.33	66.67	100
ANXA3	LCP1	100	83.33	33.33	100	100	100	83.33	33.33	50	66.67	33.33	50
ANXA3	HSP90AA1	100	100	50	83.33	100	100	66.67	100	100	100	33.33	33.33
ANXA3	CTSL	100	100	100	66.67	100	100	100	100	66.67	100	100	100
ANXA3	DBI	100	100	66.67	100	100	100	100	77.78	66.67	100	55.56	55.56
ANXA3	IGHG1	50	100	100	100	83.33	100	100	100	100	100	100	100
ANXA3	TMPRSS2	100	100	66.67	83.33	100	100	100	100	66.67	100	66.67	83.33
ENO1	PRDX6	83.33	83.33	75	83.33	83.33	83.33	83.33	100	66.67	83.33	91.67	83.33
ENO1	EZR	100	87.5	100	75	100	100	87.5	100	75	100	100	100
ENO1	GAPDH	87.5	100	100	87.5	100	93.75	100	100	81.25	93.75	100	100
ENO1	ANXA1	91.67	91.67	83.33	91.67	91.67	83.33	100	83.33	83.33	83.33	100	75
ENO1	EDDM3B	91.67	100	100	100	91.67	91.67	100	100	83.33	100	75	83.33
ENO1	PSMA2	100	100	87.5	100	100	100	100	100	100	100	100	100
ENO1	GPI	100	100	100	100	100	87.5	100	100	75	100	100	100
ENO1	LCP1	100	100	62.5	100	100	100	87.5	25	75	87.5	37.5	87.5
ENO1	HSP90AA1	100	100	100	100	100	87.5	75	100	87.5	87.5	87.5	62.5
ENO1	DBI	100	100	100	100	100	100	100	75	58.33	100	100	83.33
ENO1	IGHG1	50	100	87.5	100	100	100	100	100	100	100	100	87.5
ENO1	IGHG2	87.5	62.5	100	100	87.5	87.5	62.5	87.5	87.5	87.5	100	100
ENO1	C3	75	100	100	100	100	100	100	100	87.5	100	62.5	100
ORM2	EDDM3B	91.67	100	91.67	100	100	83.33	100	100	58.33	16.67	83.33	83.33
ORM2	PSMA2	100	75	75	87.5	87.5	87.5	100	100	75	25	75	100
ORM2	HSP90AA1	100	100	75	100	100	100	75	100	100	25	87.5	75
ORM2	IGHG1	50	100	75	87.5	75	87.5	87.5	87.5	87.5	25	75	75
ORM2	C3	50	100	100	100	100	100	100	100	75	0	87.5	87.5
OS9	TWSG1	100	100	75	75	100	100	75	75	75	75	75	100
OS9	CTSL	100	100	100	100	100	100	100	75	75	100	50	100
OS9	TMPRSS2	100	100	75	75	100	75	75	75	50	75	75	50
SORD	EDDM3B	100	83.33	91.67	91.67	100	58.33	91.67	91.67	33.33	91.67	66.67	91.67
SORD	PSMA2	100	100	87.5	100	100	87.5	100	100	100	100	87.5	100
SORD	HSP90AA1	100	87.5	87.5	100	100	100	87.5	100	100	100	87.5	75
SORD	IGHG1	50	87.5	87.5	100	87.5	87.5	87.5	100	100	87.5	100	75
SORD	C3	62.5	75	75	100	100	62.5	100	100	75	50	75	50
PRDX6	ELSPBP1	0	66.67	66.67	50	100	33.33	83.33	50	41.67	0	58.33	33.33
PRDX6	RNASE4	91.67	83.33	100	100	100	83.33	100	58.33	58.33	91.67	66.67	100
PRDX6	GDI2	66.67	33.33	83.33	100	100	33.33	100	83.33	50	50	83.33	83.33
PRDX6	TIMP1	91.67	87.5	100	100	87.5	66.67	100	95.83	62.5	95.83	100	100
PRDX6	EZR	100	83.33	66.67	66.67	100	66.67	100	83.33	33.33	83.33	100	83.33
PRDX6	GAPDH	100	91.67	100	100	91.67	75	100	100	50	66.67	75	100
PRDX6	ANXA1	100	55.56	88.89	100	100	88.89	100	77.78	44.44	100	77.78	88.89

PRDX6	CRTAC1	75	58.33	100	100	91.67	41.67	83.33	100	50	66.67	91.67	91.67
PRDX6	EDDM3B	100	100	100	100	100	66.67	100	100	55.56	100	44.44	100
PRDX6	PSMA2	100	100	100	100	100	100	100	100	66.67	100	66.67	100
PRDX6	GPI	100	50	66.67	100	66.67	50	100	100	50	66.67	66.67	66.67
PRDX6	LCP1	100	50	66.67	100	83.33	83.33	83.33	16.67	50	83.33	0	100
PRDX6	HSP90AA1	100	100	100	100	83.33	100	66.67	100	50	100	50	83.33
PRDX6	CTSL	100	100	100	100	100	66.67	100	100	66.67	83.33	100	100
PRDX6	DBI	88.89	88.89	100	100	88.89	88.89	100	55.56	44.44	88.89	66.67	77.78
PRDX6	IGHG1	16.67	100	100	100	50	83.33	100	100	100	100	100	100
PRDX6	IGHG2	100	50	100	100	100	83.33	83.33	100	66.67	100	83.33	100
PRDX6	TMPRSS2	100	100	100	100	100	100	100	100	83.33	100	83.33	100
PRDX6	C3	33.33	100	100	100	83.33	66.67	100	100	66.67	33.33	50	66.67
ELSPBP1	GDI2	62.5	100	62.5	87.5	100	100	87.5	62.5	100	62.5	75	37.5
ELSPBP1	EZR	75	87.5	100	100	100	100	75	75	87.5	62.5	100	50
ELSPBP1	ANXA1	66.67	100	33.33	83.33	100	75	83.33	91.67	75	66.67	100	66.67
ELSPBP1	GPI	87.5	100	87.5	87.5	100	100	87.5	87.5	87.5	87.5	100	62.5
ASAH1	TMPRSS2	100	100	87.5	100	100	100	87.5	87.5	75	87.5	62.5	87.5
PLA1A	TWSG1	87.5	100	100	100	87.5	100	75	100	50	87.5	87.5	87.5
PLA1A	TPP1	100	100	91.67	91.67	83.33	91.67	91.67	91.67	58.33	100	75	100
PLA1A	TMPRSS2	100	100	75	100	87.5	100	87.5	87.5	62.5	100	75	75
PEBP4	PSMA2	87.5	100	87.5	100	100	87.5	100	100	87.5	100	87.5	100
PEBP4	IGHG1	37.5	87.5	75	87.5	62.5	75	87.5	75	87.5	87.5	87.5	62.5
PEBP4	C3	62.5	100	100	100	87.5	75	100	87.5	75	87.5	62.5	75
RNASE4	GDI2	87.5	100	100	100	100	87.5	100	75	75	100	100	100
RNASE4	EZR	100	100	100	100	100	87.5	100	87.5	87.5	87.5	75	100
RNASE4	ANXA1	100	100	75	100	100	100	100	91.67	91.67	100	100	91.67
RNASE4	HEXA	100	83.33	91.67	100	83.33	75	100	91.67	66.67	83.33	91.67	100
RNASE4	CRTAC1	87.5	100	100	100	87.5	81.25	87.5	93.75	75	87.5	100	93.75
RNASE4	GPI	87.5	87.5	87.5	87.5	75	87.5	100	87.5	87.5	87.5	75	87.5
GDI2	EZR	75	100	75	50	75	75	75	50	50	75	75	75
GDI2	GAPDH	87.5	100	100	100	100	100	87.5	62.5	87.5	87.5	62.5	100
GDI2	ANXA1	83.33	100	83.33	100	100	66.67	100	66.67	33.33	83.33	66.67	66.67
GDI2	HEXA	100	100	100	100	100	66.67	100	100	83.33	100	100	100
GDI2	CRTAC1	100	100	100	87.5	87.5	87.5	87.5	75	75	87.5	87.5	100
GDI2	TWSG1	100	100	100	100	100	50	100	100	50	100	100	100
GDI2	GPI	100	100	75	100	75	75	100	50	75	75	50	75
GDI2	LCP1	100	100	50	100	100	75	100	25	100	75	25	50
GDI2	CTSL	100	100	100	100	100	50	100	100	100	100	100	100
GDI2	TMPRSS2	100	100	100	100	100	100	100	50	100	100	75	100
TIMP1	EZR	93.75	100	100	100	100	100	87.5	100	93.75	100	100	100
TIMP1	ANXA1	100	100	95.83	100	95.83	91.67	100	95.83	83.33	100	100	95.83

TIMP1	CTSL	81.25	93.75	87.5	100	100	87.5	100	93.75	81.25	93.75	75	93.75
TIMP1	DBI	100	95.83	95.83	91.67	100	91.67	100	87.5	91.67	95.83	79.17	87.5
EZR	GAPDH	87.5	75	62.5	75	87.5	87.5	75	87.5	50	75	62.5	100
EZR	ANXA1	83.33	66.67	66.67	83.33	100	66.67	100	83.33	66.67	83.33	66.67	66.67
EZR	PSMA2	100	100	50	100	100	100	100	100	100	100	50	100
EZR	GPI	100	50	100	75	100	75	75	100	25	75	50	100
EZR	LCP1	100	75	50	100	100	100	100	50	50	100	0	50
EZR	HSP90AA1	100	100	50	75	75	100	75	75	100	100	25	50
EZR	CTSL	100	100	100	100	100	100	100	100	50	100	100	100
EZR	DBI	100	66.67	66.67	83.33	100	83.33	100	83.33	66.67	100	66.67	66.67
EZR	IGHG1	50	100	100	100	75	100	100	100	100	100	100	100
EZR	C3	25	75	75	75	75	100	50	75	50	50	25	50
GAPDH	ANXA1	100	100	91.67	91.67	83.33	75	100	100	75	75	100	83.33
GAPDH	EDDM3B	100	100	100	100	100	100	100	100	83.33	100	66.67	91.67
GAPDH	PSMA2	100	100	87.5	100	100	100	100	100	100	100	100	100
GAPDH	GPI	87.5	87.5	87.5	100	87.5	87.5	100	100	50	75	87.5	87.5
GAPDH	LCP1	100	100	75	100	100	100	87.5	25	75	75	25	87.5
GAPDH	HSP90AA1	100	100	100	87.5	100	100	75	100	100	100	87.5	75
GAPDH	CTSL	87.5	100	100	87.5	100	87.5	100	100	87.5	87.5	100	100
GAPDH	DBI	100	100	91.67	100	91.67	91.67	100	66.67	58.33	91.67	100	75
GAPDH	IGHG1	37.5	100	100	100	100	100	100	100	100	100	100	87.5
GAPDH	IGHG2	100	75	100	87.5	100	100	87.5	100	100	100	100	100
GAPDH	C3	50	100	100	100	100	100	87.5	100	87.5	62.5	50	75
SDF4	CTSL	50	75	75	100	50	50	25	75	0	25	75	100
SDF4	TMPRSS2	50	100	50	100	75	75	50	75	75	75	100	75
ANXA1	CRTAC1	83.33	100	75	91.67	83.33	50	91.67	91.67	25	75	100	91.67
ANXA1	EDDM3B	100	100	88.89	100	88.89	55.56	100	100	33.33	100	66.67	100
ANXA1	PSMA2	100	100	100	100	100	83.33	100	100	83.33	100	100	100
ANXA1	GPI	100	100	83.33	100	83.33	50	100	100	66.67	66.67	100	83.33
ANXA1	LCP1	100	100	100	83.33	66.67	66.67	83.33	50	66.67	83.33	33.33	100
ANXA1	HSP90AA1	83.33	100	100	100	83.33	100	66.67	100	83.33	100	83.33	100
ANXA1	CTSL	100	100	100	100	100	50	100	100	33.33	100	100	100
ANXA1	DBI	100	100	100	100	77.78	77.78	100	77.78	44.44	77.78	88.89	88.89
ANXA1	IGHG1	33.33	100	100	100	50	83.33	100	100	100	100	100	83.33
ANXA1	TMPRSS2	100	100	100	100	100	100	100	100	66.67	100	83.33	100
ANXA1	C3	50	100	83.33	100	83.33	66.67	83.33	100	50	50	50	66.67
HEXA	TMPRSS2	100	100	83.33	100	100	83.33	83.33	83.33	100	100	83.33	83.33
EDDM3B	PSMA2	100	100	50	100	100	100	100	100	83.33	100	100	100
EDDM3B	GPI	100	66.67	100	100	100	100	100	100	100	83.33	83.33	83.33
EDDM3B	LCP1	83.33	83.33	50	100	100	66.67	100	50	66.67	66.67	66.67	83.33
EDDM3B	HSP90AA1	100	100	66.67	100	100	66.67	50	83.33	66.67	83.33	83.33	83.33

EDDM3B	DBI	100	100	77.78	100	100	100	100	88.89	77.78	100	88.89	100
EDDM3B	IGHG1	50	100	100	100	100	100	100	100	83.33	100	83.33	83.33
EDDM3B	IGHG2	100	50	100	100	100	100	66.67	66.67	100	83.33	83.33	100
EDDM3B	C3	33.33	100	100	100	100	100	100	100	100	50	100	50
TWSG1	TPP1	100	100	100	100	100	100	100	100	100	100	100	100
TWSG1	CTSL	100	100	100	100	100	100	100	100	100	100	100	100
TWSG1	TMPRSS2	100	100	75	100	100	100	100	50	100	100	75	50
PSMA2	LCP1	100	100	75	100	100	100	100	75	50	100	75	100
PSMA2	HSP90AA1	100	100	100	75	100	75	75	100	100	100	75	75
PSMA2	DBI	100	100	100	100	100	100	100	100	66.67	100	100	100
PSMA2	IGHG1	50	100	100	100	75	100	75	100	75	75	100	75
PSMA2	IGHG2	100	75	100	100	100	100	100	100	75	100	100	100
PSMA2	C3	50	100	100	100	100	100	100	100	100	25	100	50
GPI	LCP1	100	100	50	100	100	100	100	50	75	100	50	50
GPI	HSP90AA1	100	100	50	75	100	75	50	100	100	100	50	25
GPI	CTSL	100	100	100	75	100	100	100	100	50	100	100	100
GPI	DBI	100	100	66.67	100	100	83.33	83.33	83.33	50	83.33	83.33	66.67
GPI	IGHG1	50	100	100	100	100	100	100	100	100	100	100	100
GPI	IGHG2	100	50	100	75	100	100	50	100	50	100	100	100
GPI	TMPRSS2	100	100	75	75	100	100	100	100	75	100	100	75
LCP1	HSP90AA1	100	100	100	100	100	100	100	100	100	75	100	100
LCP1	DBI	83.33	100	100	100	100	100	83.33	100	83.33	66.67	83.33	83.33
LCP1	IGHG1	25	100	100	100	100	100	100	100	100	50	100	75
LCP1	IGHG2	75	75	50	100	100	50	100	75	50	50	75	100
LCP1	C3	25	100	100	100	100	100	100	100	100	0	100	75
HSP90AA1	DBI	100	100	100	100	100	100	100	83.33	66.67	100	100	100
HSP90AA1	IGHG1	50	100	100	100	100	100	100	100	100	75	100	75
HSP90AA1	IGHG2	100	50	100	75	100	75	75	100	50	75	100	100
HSP90AA1	C3	50	100	100	100	100	50	100	100	75	50	75	25
CTSL	MATN2	100	75	100	100	100	75	100	100	100	75	75	100
CTSL	TMPRSS2	100	75	100	100	100	100	100	50	75	100	25	50
DBI	IGHG1	66.67	100	100	100	83.33	100	100	100	66.67	100	100	66.67
DBI	IGHG2	100	66.67	100	100	100	83.33	83.33	100	83.33	100	100	100
DBI	C3	50	100	66.67	100	83.33	66.67	83.33	66.67	66.67	66.67	50	66.67
IGHG1	IGHG2	100	50	100	100	100	100	50	100	100	100	100	25
IGHG1	C3	100	100	100	100	100	100	50	100	75	100	50	75
IGHG2	C3	50	50	100	100	100	100	50	100	100	75	75	75

ANNEX E – APLICACIÓ WEB PROGRAMADA AMB SHINY

1. Modificació de la funció

Com a feina derivada d'aquest TFM es va donar la necessitat de programar un algorisme que pogués obtenir des de la web *d'Uniprot* els GO de un llistat de proteïnes. Al grup d'investigació van veure que ampliant la funció getNamesUniprot [Algorisme 12], podíem dur a terme aquesta tasca. Així doncs en un primer moment es va procedir a modificar la funció d'R:

1. Preparem el data.*frame* on quedaran les dades, i traiem les caselles en blanc.

```
getNamesUniprot <- function(x){

  # Preparamos el data.frame para alojar los datos de salida
  protNames <- data.frame( ProtID = character(), ProtName =
character(), GO_C = character(), GO_F = character(), GO_P =
character(), ProtFunct=character())

  x<-as.data.frame(x[complete.cases(x), ])
```

Algorisme 29: GetNamesUniprot Web 1

2. Preparem un for per recórrer la columna on hi ha els codis *Uniprot*, per cada proteïna provem creem la *url* i provem la connexió amb un *if*. Si tenim connexió, procedim a obtenir l'objecte XML i dupliquem l'objecte en format llista (donat que facilita el recorregut pels nodes, passant a ser un llistat separat per "$").
Emplenem el nom i la funció de proteïna amb un "no disponible" per si de cas no podem obtenir-los i els passem a intentar aconseguir

```
  # Recorremos el data.frame de entrada para obtener los datos
  for (i in 1:dim(x)[1]){

    # Probamos la conexi?n
        url <- paste ("http://www.uniprot.org/uniprot/", x[i,],
".xml", sep="")

    if(url.exists(url)){

      # Obtenemos los datos
      try(data <- xmlInternalTreeParse(url, useInternal=TRUE),
silent=TRUE)
      # Los convertimos en lista para acceder a los nodos
      xml_data <- xmlToList(data)
      # Rellenamos el nombre por si no tiene correspondencia
            ProtName <- "Not available"
            ProtFunct<- "Not available"

            #Probamos a asignar los datos
            ProtName <-
as.character(xml_data$entry$gene$name$text[xml_data$entry$gene$name$
.attrs=="primary"])
            ProtFunct<- as.character(xml_data$entry$comment$text[1])
```

Algorisme 30: GetNamesUniprot Web 2

3. Pels GO (*Gene Onthology*) la obtenció és una mica diferent donat que hi ha molts objectes *dbReference*, hem de anar separant poc a poc, obtenir l'entrada, després el *dbReference*, seguidament els GO i per últim quedar-nos el *value* d'aquests. Un cop fet això comprovem que els GO no siguin inexistents i passem a fer un llistat separat per coma dels GO en funció del seu tipus(C,F o P) que podem trobar a la primera lletra de la cadena que els indica. Els classifiquem mitjançant GREP i mitjançant GSUB eliminem les "C:", "F:" i "P:".
Si hem comprovat que no hi ha res els posem com no disponibles. (Això es per assegurar que no hi hagi un error, en cas que no sigui un tipus quedarà en blanc però si no hi ha cap posarà el missatge de no disponible)

```
      #Obtenmos los GO
      Pr <- as.list(xml_data$entry)
      Pr <- Pr[names(Pr)=="dbReference"]
      Pr<- Pr[sapply(Pr, function(x) any(unlist(x) == "GO"))]
      Pr <- as.data.frame(sapply(Pr,function(x)
unlist(x$property["value"])))

      #Clasificamos los GO en funci?n de su vinculaci?n
      if(length(Pr) != 0L){

        c<-paste(gsub("C:","",Pr[grep("^C", Pr[,1]),]), collapse = ",
")
        f<-paste(gsub("F:","",Pr[grep("^F", Pr[,1]),]), collapse = ",
")
        p<-paste(gsub("P:","",Pr[grep("^P", Pr[,1]),]), collapse = ",
")

      }else{

        c<-"GO NOT AVAILABLE"
        f<-"GO NOT AVAILABLE"
        p<-"GO NOT AVAILABLE"

      }
}
```

Algorisme 31: GetNamesUniprot Web 3

4 Si no hem pogut fer la connexió, passem a l'altra banda de *l'if* i posem tot en blanc excepte *ProtName* on hi haurà un missatge indicant que el codi no es una proteïna:

```
    }else{
      ProtName <- "This Field is not a Protein"
      ProtFunct <- ""
      c<-""
      f<-""
      p<-""
    }
```

Algorisme 32:GetNamesUniprot Web 4

5. Amb això passem a construir una fila de la sortida, la afegim al *data.frame* de resultats i esborrem data. Mostrem el marcador de progrés i preparem el retorn:

```
      # Creamos una fila
      name <-data.frame( ProtID =
as.character(x[i,]),ProtName=ProtName,
GO_C=c,GO_F=f,GO_P=p,ProtFunct=ProtFunct)



    # A?adimos la fila a el data.frame que hemos creado
    protNames <- rbind(protNames, name)

    data<-""

    # Mostramos un mensaje de progreso
    cat("     \r",round(i/(dim(x)[1])*100, 2), "%   ")
  }

  # Devolvemos el data.frame finalizado
  return(protNames)
```

Algorisme 33: GetNamesUniprot Web 5

2. Creació del arxiu R que executi la funció

En aquest cas només ha calgut importar les llibreries, l'arxiu de funcions, l'arxiu *xlsx* i passar-lo a la funció. Addicionalment es va crear una funció de cerca que permetés als component del grup filtrar el resultat. Per últim es fa una exportació del total a Excel.

```
library(r2excel)
library(xlsx)
library(XML)
library(RCurl)

source("ownFunctions.R")

Proteinas <- read.xlsx("Proteinas.xlsx", sheetName="Hoja1", header=0)


ProteinGO<-getNamesUniprot(Proteinas)


#Filtrando el resultado
a<- filtering(ProteinGO,"axon | cytoplasm", "c")
View(a)

write.xlsx(ProteinGO, "ProteinGO.xlsx", sheetName="ProtGO",
append=TRUE, row.names=FALSE)
```

Algorisme 34: Executor de getNamesUniprot Web en local

3. Funció accessòria *filtering*

Com s'ha explicat en el punt anterior s'hagut de crear una funció de filtratge que solament es troba en local donat que la sortida de *Shiny* permet aquest filtratge. Aquesta funció requereix el objecte de sortida de GetNamesUniprot, un text de filtratge i a quin tipus de *GO* es vol buscar i retorna un *data.frame* on només hi ha els element que tenen aquest terme (es poden usar booleans per tenir més d'un terme).

```
filtering <- function(x, filter, where){

  if (where=="c"){
    # Filtrar en GO C
    f<-ProteinGO[grep(filter, ProteinGO[,3]),]

  }else if (where=="f"){
    # Filtrar en GO F
    f<-ProteinGO[grep(filter, ProteinGO[,4]),]

  }else if (where == "p"){
    # Filtrar en GO P
    f<-ProteinGO[grep(filter, ProteinGO[,5]),]

  }else{
    f <- "Use a character for 'filter' and lowercase for 'where'. Ex:
filtering(data.frame,'axon','c') "

  }

  return(f)
}
```

Algorisme 35: Funció filtering

4. Transformació en web *app Shiny*

Donat que al grup investigador hi ha components que no tenen un alt nivell en programació amb R, es va optar per fer una web *app* per facilitar-los el filtratge de la sortida. Per fer-ho es va crear un nou algorisme de R a partir de les opcions que proporciona R-*studio*. Es va cridar a les llibreries necessàries i a les funcions, que es el mateix document que en la versió local però sense *filtering*.

```
library(shiny)
library(r2excel)
library(xlsx)
library(XML)
library(RCurl)

#Llamada a el script de funciones
source("functionsWeb.R")
```

Algorisme 36: WebApp Shiny 1

Seguidament es va definir la interfície (ui). Es va fer servir un senzill *css* anomenat

bootstrap.css, es va posar el títol, un carregador d'arxius, un descarregador de resultats i la taula de resultats. A més podem veure que hi ha un missatge per quan la funció obté els GO de les proteïnes de "Calculando…":

```
# Definición de la página
ui <- fluidPage(
   #Estilos
   theme = "bootstrap.css",

   # Título
   headerPanel("UNIPROTER"),

   # Cargador de archivos
   fileInput("file", h3("Subir Archivo")),

   # Descargador de resultados
   downloadButton('downloadData', 'Download'),

   #Panel de mensaje de calculando...
   conditionalPanel(condition="$('html').hasClass('shiny-busy')",
                    tags$div("Calculando...",id="loadmessage")),

   # Mostrar la tabla de resultados
   mainPanel(dataTableOutput('dable')
  )
)
```

Algorisme 37: Interfície de la WebApp

Tot seguit trobem el servidor que es defineix amb una funció d'entrada i sortida. Primer trobem una variable global per guardar els valors reactius, seguidament es passa a una funció reactiva anomenada *GetData*, que quan pugem l'Excel calcula la taula:

```
# Definición del servidor
server <- function(input, output) {

  # Creación de una variable de guardado de variables globales
  global <- reactiveValues()

  # Obtención de datos mediante una funciÃ³n reactive que recalcule
al subir el archivo
  getData <- reactive({
    #Archivo subido
    inFile <- input$file
    #Si es nulo como al princio no retorna nada
    if (is.null(input$file))
      return(NULL)
     #Si no lee el excel
    Proteinas <- read.xlsx(inFile$datapath,sheetIndex = 1, header=0)
    #Y ejecuta la función getNamesuniprot modificada para este
proyecto
    global$ProteinasGO<-getNamesUniprot(Proteinas)
    #Devuelve la variable global que hemos creado
    return(global$ProteinasGO)
  })
```

Seguidament es crea la sortida amb una assignació reactiva (es recalcula cada cop que hi ha un canvi en l'arxiu). Seguidament trobem la crida a *getData*. A continuació veiem la creació de la variable que executarà la descàrrega.

```
    # Creación de la salida
  output$dable<- renderDataTable(
    #Llamada la función de obtención de datos
    getData()

  )

  # Preparacion de la descarga
  output$downloadData <- downloadHandler(
    #Asignación del nombre de la función
    filename = function() {
      "ProteinGo.xlsx"
    },
    # Creación del contenido de la descarga
    content = function(file) {
      write.xlsx(global$ProteinasGO, file, sheetName="Results",
append=TRUE, row.names=FALSE)
    })

}
```

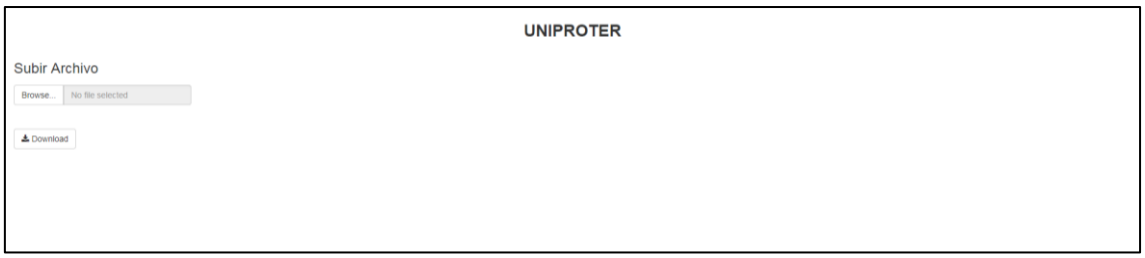
Algorisme 38: Servidor WebApp

Finalment es fa la crida a execució de la *WebApp*:

```
# Llamada a la ejecución de la función
shinyApp(ui = ui, server = server)
```

Algorisme 39: Crida a la WebApp

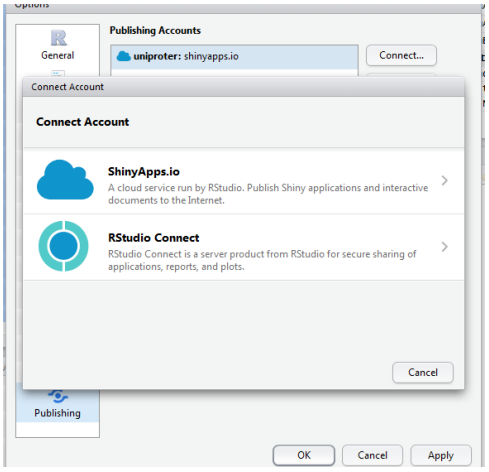
Amb això ja podem executar en local la nostra *WebApp* i comprovar que funciona:



Il·lustració 36: WebApp UNIPROTER

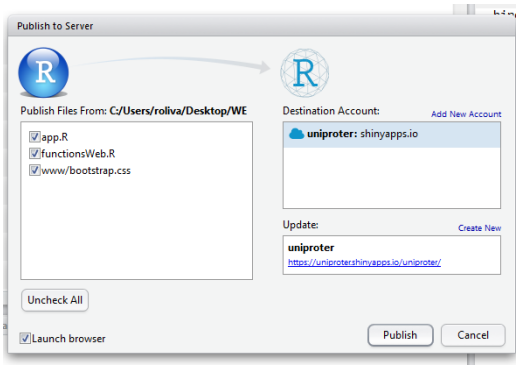
5. Pujada a la xarxa de la WebApp

La programació ha acabat però una web que no sigui a la xarxa es una mica estranya per això amb les funcions que inclou R-*studio* es va passar a pujar-la a la web amb la opció *Publish* ens va demanar crear un perfil, en el nostre cas en www.shinyapp.io seguint les passes que s'indiquen a R-*studio* vam crear el compte i el projecte:



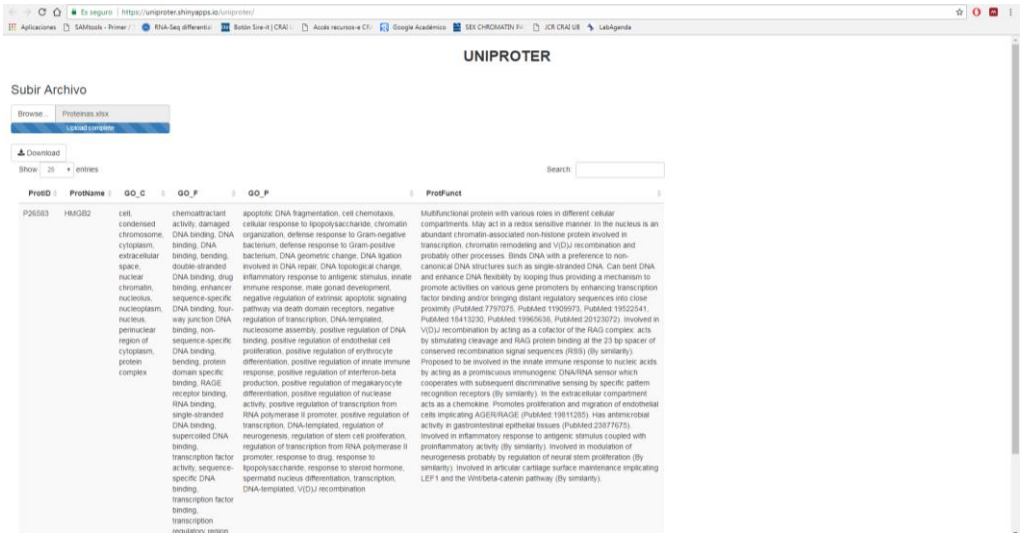
Il·lustració 37: Pujant la WebApp a la xarxa

Seguidament vam pujar els arxius amb la opció *Publish* d'R-*studio*:



Il·lustració 38: Publicació de la WebApp

I finalment tenim accessible al aplicació web a la direcció web que em designat a la creació del projecte: https://uniproter.shinyapps.io/uniproter/ on com podem veure a la aplicació està funcionant perfectament.



Il·lustració 39: Funcionament de la WebApp