Marta Domingo Domènech

Anàlisis de dades òmiques: PAC 1.

Taula de continguts

Abstract

Aquest informe presenta una anàlisi exploratòria de dades de fosfoproteòmiques obtingudes de models derivats de pacients (PDX/PD) i models analitzats amb espectrometria de masses (MSS). Utilitzant tècniques de visualització com histogrames, boxplots, PCA i heatmaps, es va examinar la distribució i les diferències en les expressions de fosforilació entre els dos grups. Els resultats mostren una clara separació entre les mostres MSS i PD, amb diferències significatives en les expressions de fosforilació. Aquestes troballes suggereixen que les variacions observades no són degudes a l'efecte batch, sinó a diferències biològiques reals entre els grups. Les dades utilitzades per l'anàlisi exploratori han sigut desengranades, separant les dades de les metadades, creant un objecte de la classe SummarizedExperiment.

Objectius

L'objectiu principal d'aquest treball és explorar i analitzar les dades de fosfoproteòmica obtingudes de models PDX i MSS per identificar diferències significatives en les expressions de fosforilació entre els dos grups i reflexionar sobre les limitacions de l'estudi en el context del problema biològic del treball. Això inclou:

- Visualitzar la distribució de les dades mitjançant histogrames i boxplots.
- Aplicar una transformació logarítmica per tractar les dades.
- Realitzar una anàlisi multivariant mitjançant PCA per identificar patrons de variació.
- Utilitzar heatmaps per visualitzar les distàncies entre les mostres i confirmar la separació entre els grups.
- Objectius secundaris: saber fer ús de la classe SummarizedExperiment i saber-ne les diferències en comparació amb la classe ExpressionSet.

Mètodes

Les dades utilitzades en aquest informe provenen d'un experiment de fosfoproteòmica que analitza models PDX i MSS. Les dades inclouen mesures de fosforilació en diferents mostres. La metodologia emprada inclou:

Origen i naturalesa de les dades: Les dades provenen d'un experiment de fosfoproteòmica dissenyat per analitzar 6 models PDX (3 de cada subtipus) utilitzant mostres enriquides amb fosfopèptids. Aquest tipus d'experiment ajuda a entendre millor les modificacions post-traduccionals de les proteïnes en diferents subtipus de càncer, utilitzant models derivats de pacients per obtenir resultats més rellevants per a la recerca biomèdica.

Metodologia emprada:

- Preparació de mostres: Les mostres biològiques (cèl·lules, teixits) es preparen per a l'anàlisi, incloent la lisi cel·lular per alliberar les proteïnes.
- Digestió proteica: Les proteïnes es digereixen en pèptids mitjançant enzims com la tripsina.
- Enriquiment de fosfopèptids: Les mostres es tracten per enriquir-les amb fosfopèptids.
- Separació i identificació de proteïnes fosforilades: Utilitzant espectrometria de masses (MS) per obtenir dades detallades sobre la massa molecular, la seqüència dels pèptids, les modificacions post-traduccionals com la fosforilació, i la quantificació de l'abundància dels pèptids i proteïnes.
- Eines estadístiques i bioinformàtiques: R per a la manipulació i anàlisi de dades, ggplot2 per a la visualització, i funcions específiques per a la realització de PCA i heatmaps.

Procediment general d'anàlisi:

- Visualització inicial de les dades amb histogrames i boxplots.
- Aplicació de la transformació logarítmica.
- Anàlisi multivariant mitjançant PCA.
- Visualització de distàncies amb heatmaps.

Creació d'un objecte de la classe SummarizedExperiment:

- Creació del dataframe row_data: Es crea un dataframe anomenat "row_data" amb la informació de cada feature (fila). Aquest dataframe recull les metadades de cada feature: "SequenceModifications, Accession, Description, Score, CLASS, PHOSPHO".
- Construcció de la matriu assay_matrix: Es construeix una matriu amb les columnes del dataset que contenen les dades numèriques.
- Creació del dataframe colData: Es classifica la informació de les mostres entre els dos mètodes d'anàlisi emprats: MSS i PD.
- Creació de l'objecte SummarizedExperiment: Es crea un objecte SummarizedExperiment que conté les dades i les metadades.
- Revisió de l'objecte creat: Es revisa l'objecte SummarizedExperiment per assegurar-se que conté les dades correctament estructurades.

SummarizedExperiment vs ExpressionSet

La classe SummarizedExperiment es pot considerar una evolució o substitució més recent de la classe ExpressionSet dins l'ecosistema de Bioconductor. Tot i que comparteixen la idea bàsica de contenir dades d'expressió i metadades associades, SummarizedExperiment presenta diverses diferències rellevants.

Diferències estructurals:

- Estructura d'emmagatzematge de dades:

ExpressionSet disposa d'un sol array d'expressió (exprs).

SummarizedExperiment admet múltiples "assays" en el compartiment assays, útil per gestionar conjuntament dades crues, normalitzades o d'altres tipus en un mateix objecte.

- Informació genòmica integrada (rowRanges):

ExpressionSet no incorpora informació sobre les posicions genòmiques.

SummarizedExperiment inclou el compartiment rowRanges, que emmagatzema coordenades i metadades genòmiques de cada fila, útil per a dades de seqüenciació.

- Metadades en forma de DataFrame:

ExpressionSet utilitza pData (per a mostres) i fData (per a característiques) com slots de metadades, que són data.frames base de R.

SummarizedExperiment empra colData (per a mostres) i rowData (per a característiques) que són objectes de tipus DataFrame, més flexibles i consistents amb altres paquets bioconductor.

- Versatilitat i escalabilitat:

ExpressionSet va ser concebut principalment per a dades de microarrays, amb una sola matriu d'expressió.

SummarizedExperiment és més flexible i pensat per gestionar dades d'RNA-seq, ATAC-seq, proteòmica, metabolòmica i altres dades multimodals, gràcies a la possibilitat de tenir múltiples assays i integració amb la infraestructura genòmica.

- Integració amb Bioconductor:

SummarizedExperiment s'ha convertit en l'estàndard per a moltes eines i fluxos de treball Bioconductor, especialment aquells que fan servir GenomicRanges.

ExpressionSet encara es fa servir en aplicacions i paquets antics, però la majoria de paquets nous de Bioconductor opten per SummarizedExperiment.

En resum, SummarizedExperiment és una classe més flexible i adaptada a l'anàlisi multimodal i amb integració genòmica, mentre que ExpressionSet s'enfoca en un únic array d'expressió i amb metadades més senzilles. Això fa que SummarizedExperiment sigui l'opció recomanada en la major part de pipelines de dades òmiques actuals.

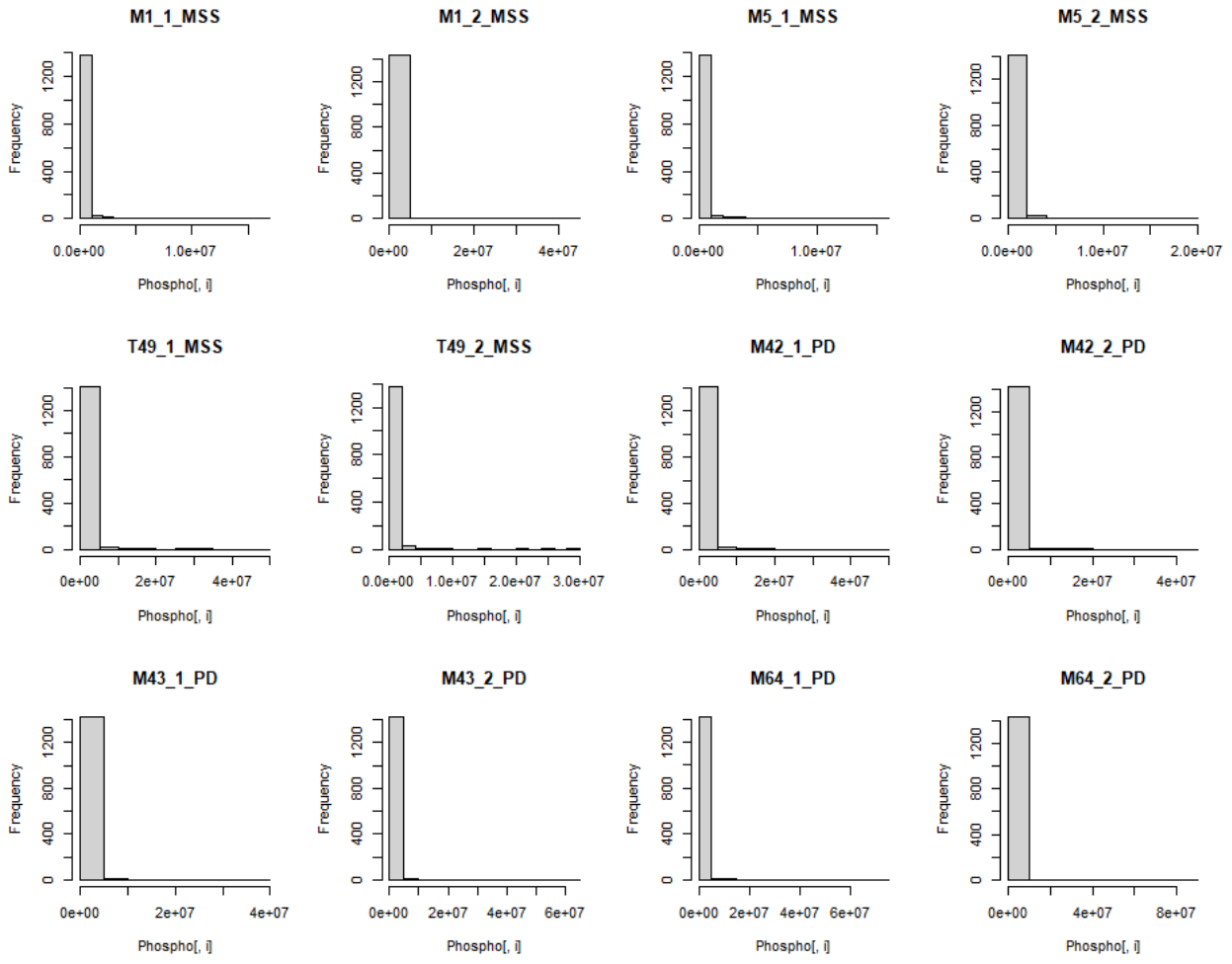
Resultats

Els resultats de l'anàlisi exploratòria es presenten a continuació:

Visualització inicial de les dades mitjançant la taula estadística (Taula 1) i els histogrames(Taula 2): Els histogrames mostren una gran variabilitat en les dades, amb valors que van des de zero fins a magnituds de 10^7. Això indica la necessitat de tractar les dades abans de continuar amb l'anàlisi.

	M1_1_MSS	M1_2_MSS	M5_1_MSS	M5_2_MSS	T49_1_MSS	T49_2_MSS	M42_1_PD	M42_2_PD	M43_1_PD	M43_2_PD	M64_1_PD	M64_2_PD
Min.	0	0	0	0	0	0	0	0	0	0	0	0
1st Qu.	5653	5497	2573	3273	9306	8611	5341	4216	19641	17299	11038	8660
Median	30682	26980	20801	26241	55641	46110	36854	30533	67945	59607	52249	47330
Mean	229841	253151	232967	261067	542449	462616	388424	333587	349020	358822	470655	484712
3rd Qu.	117373	113004	113958	130132	223103	189141	180252	152088	205471	201924	209896	206036
Max.	16719906	43928481	15135169	19631820	49218872	29240206	48177680	42558111	35049402	63082982	71750330	88912734

Taula 1. Taula estadística extreta de la funció summary()



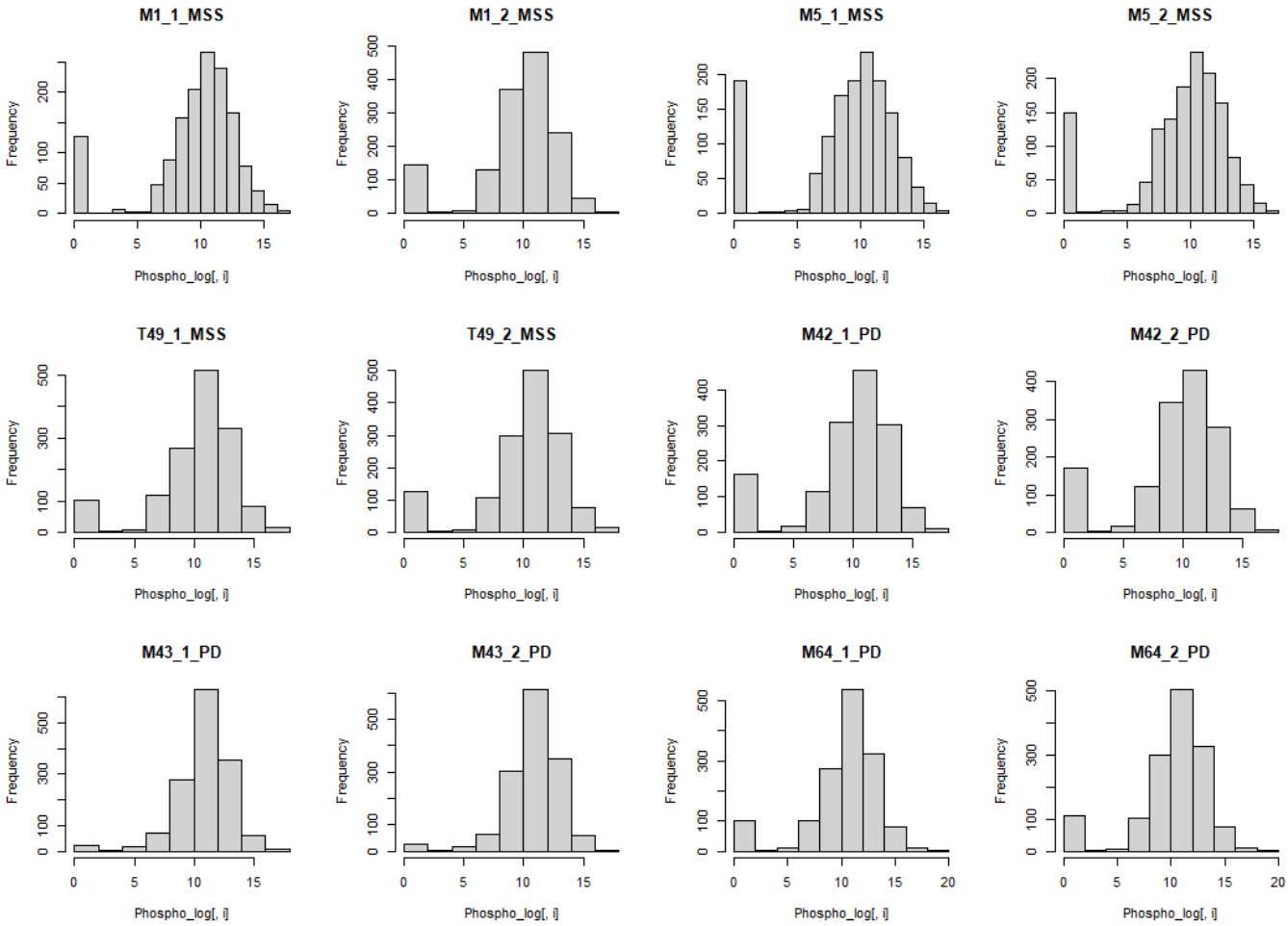
Taula 2. Histogrames del raw data.

L'histograma evidència que cal aplicar una transformació logarítmica per rebaixar els valors més alts i expandir els més petits, disminuïnt així la diferència entre ambdos valors.

En un dataset on la variable “x” pot prendre el valor zero (o valors molt petits), log(x) és matemàticament problemàtic quan x = 0, perquè log(0) no està definit (és -∞). Per això, sovint s’utilitza log(x + 1):
• Si x = 0, llavors log(1) = 0, de manera que no es produeix cap error ni valor infinit.
• Això evita haver d’eliminar les observacions que tinguin x = 0 (cosa que podria esbiaixar l’anàlisi).
• És una transformació comuna en casos on x representa coses com comptes (counts), on zero pot tenir un significat estadísticament rellevant.

És per aquesta raó que es decideix aplicar la funció logarítmica log(df+1), sent “df” el conjunt de dades “Phospho”.

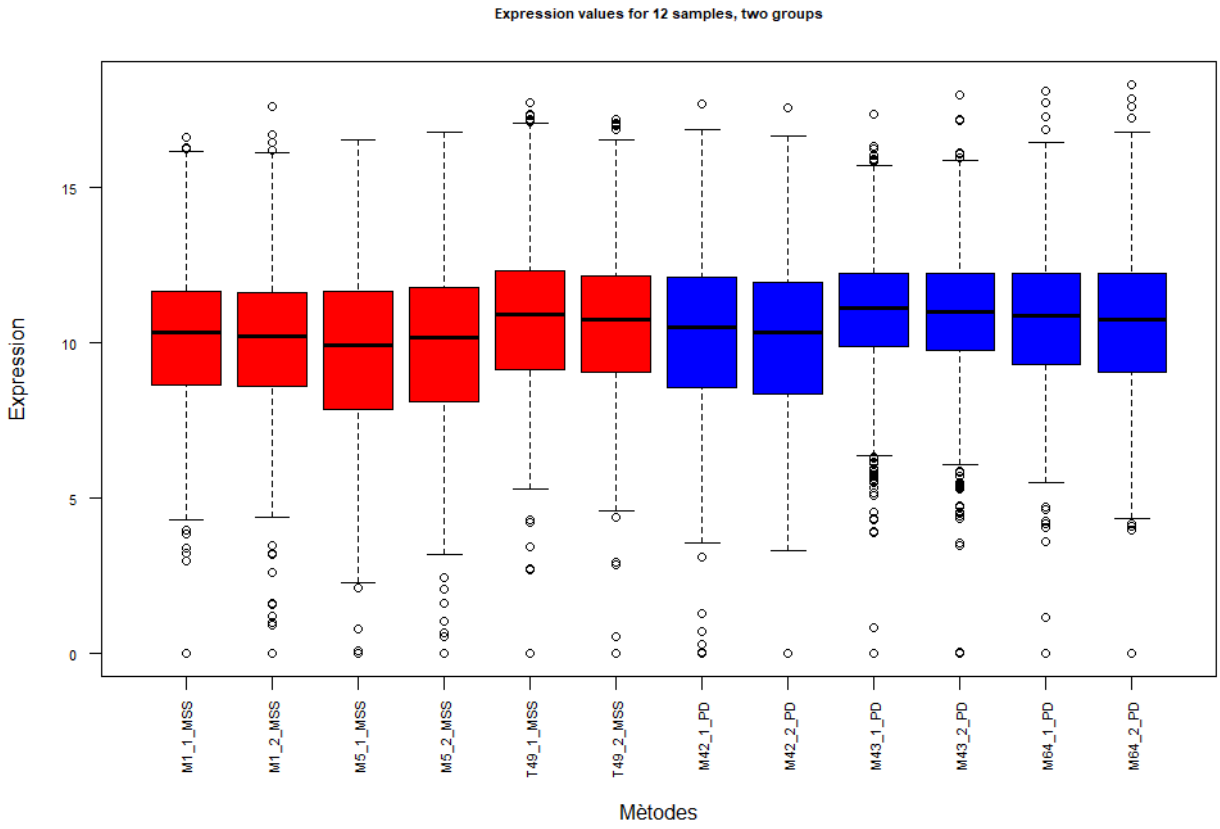
Transformació logarítmica: Aplicar una transformació logarítmica va permetre reduir els valors més alts i expandir els més petits, disminuint així la diferència entre ambdós valors. Els histogrames de les dades transformades mostren una distribució més uniforme.



Taula 3. Histogrames amb les dades transformades logarítmicament.

Com s'observa en la taula 3, un cop s'ha aplicat el logaritme les dades ja es poden apreciar de manera visual.

Una opció millor és fer servir boxplots, que permeten visualitzar totes les mostres alhora:

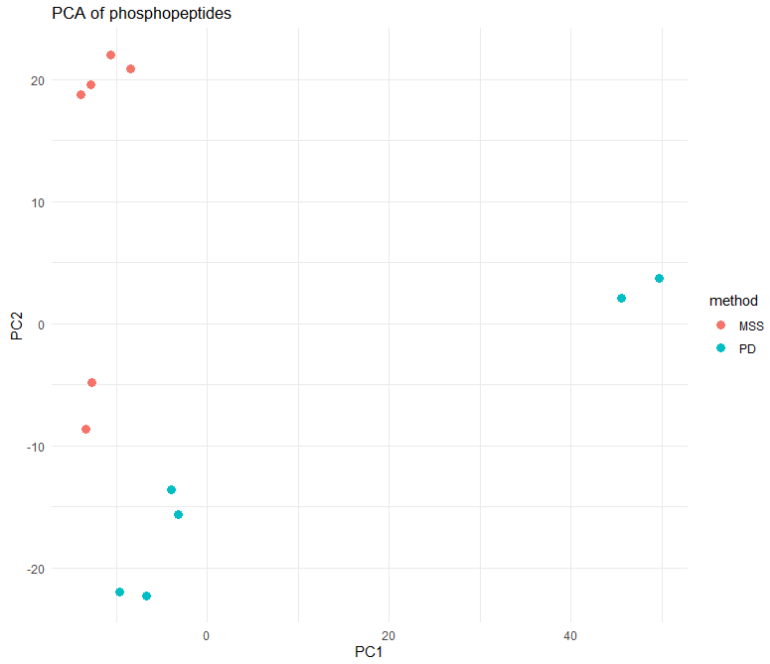


Taula 4. Gràfica bloxplot amb les dades tractades logaritmicament.

Les gràfiques boxplot són molt il·lustratives ja que ens mostren si les dades contenen outliers, o bé si les dades són assimètriques. En aquets cas podem observar com les dades tenen bastants outliers i les medianes són molt irregulars.

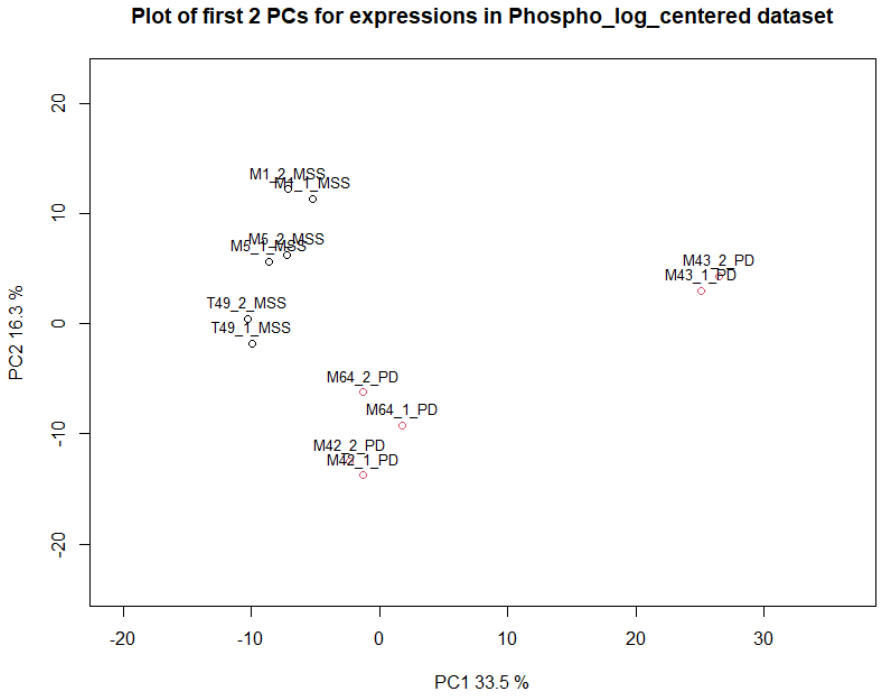
En el cas de que el boxplot es mostre amb aquesta semblança amb les dade sense tractar, l'assimètria que presenta evocaria a plantejar-se un tractament de dades per arreglar-les. En aquest cas el tractament ja s'ha aplicat, indicant que les dades per si mateixes són molt heterogènies.

Per realitzar un anàlisi multivariant és necesari primer centrar i normalitzar les dades. Un cop aquest ajust ja està fet, es pot procedir a fer l'estudi dels principals components.



Taula 5. Gràfica amb la representació dels principals components.

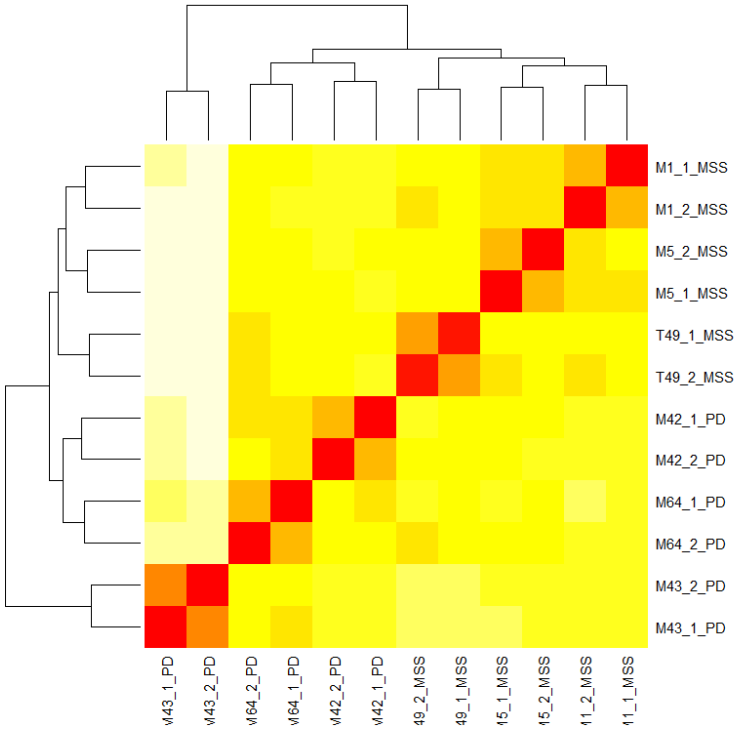
Per evitar confusions degut a l'efecte batch, és a dir, evitar variacions sistemàtiques les quals no estan relacionades amb les variables d'interès, sinó amb amb el processament de les mostres, es duen a terme una sèrie d'anàlisis. En aquest cas només es realitzaran anàlisis de detecció.



Taula 6. Representació de l'anàlisi de l'efecte Batch.

La PCA mostra una clara separació entre les mostres MSS i PD (Taula 6). Les mostres MSS es troben agrupades en una regió, mentre que les mostres PD es troben en una altra, indicant diferències significatives en les expressions de fosforilació.

Per últim executarem un estudi basat en distàncies amb la funció heatmap(). Aquest estudi consisteix en calcular una matriu de distàncies i visualitzant-la mitjançant un mapa de colors.



Taula 7. Gràfica de distàncies heatmaps de les mostres.

Visualització de distàncies amb heatmaps (Taula 7): El heatmap mostra una clara separació entre les mostres MSS i PD, amb cada grup formant clústers separats. Les mostres dins de cada grup són més similars entre elles, mentre que les mostres de diferents grups tenen distàncies més grans i menor similitud.

Discussió

L'anàlisi exploratòria de les dades de fosfoproteòmica revela diferències significatives en les expressions de fosforilació entre els models MSS i PD. Les visualitzacions amb histogrames, boxplots, PCA i heatmaps confirmen que les variacions observades no són degudes a l'efecte batch, sinó a diferències biològiques reals entre els grups. Tot i així, és important considerar les limitacions de l'estudi, com la possible presència d'outliers i la necessitat de preprocessament addicional per obtenir resultats més precisos.

Conclusions

Aquest informe ha demostrat que les dades de fosfoproteòmica obtingudes de models MSS i PD presenten diferències significatives en les expressions de fosforilació. Les tècniques de visualització i anàlisi multivariant han permès identificar patrons de variació i confirmar la separació entre els grups. Aquestes troballes són importants per comprendre millor els mecanismes de regulació de les proteïnes en el càncer i poden contribuir al desenvolupament de tractaments més efectius.

Referències

https://github.com/mdomingod/UOC.git