Taula de continguts

Abstract

Aquest informe correspon a l'estudi on s'ha dut a terme una anàlisi de fosfoproteòmica utilitzant dades públiques extretes de (https://github.com/nutrimetabolomics/metaboData/tree/main) on apareixen mostres riques amb fosfolípids que tenen diferents modificacions en les seqüències. Aquestes mostres corresponen a dos grups tumorals MSS i PD. Primer, s'ha creat un objecte de classe SummarizedExperiment de bioconductor per estructurar les dades d'abundància dels fosfopèptids, conjuntament amb la informació de les mostres i les seves metadades. Després, s'ha dut a terme, mitjançant Rstudio una anàlisi que inclou l'anàlisi de components principals o PCA, representada gràficament que indica certa separació entre els dos grups experimentals. Posteriorment, s'ha utilitzat el model estadístic de limma per identificar les mostres diferencialment expressades i s'ha representat el resultat en un gràfic del tipus volcanoplot.

Objectius

1. Diferènciació de dos grups tumorals del dataset: phosphoproteinomics (https://github.com/nutrimetabolomics/metaboData/tree/main)
    1.1. Descripció breu de les dades
    1.2. Summairzed experiment
    1.3. Principal component analisis (PCA)
    1.4. Limma analisis i representació en volcanoplot
2. Controlar versions amb Git i creació repositori de GitHub: https://github.com/laorfl/Ortega_Flores_Laia_PEC1

Mètodes

1. Descàrrega dataset

Es descarrega el conjunt de dades de fosfoproteinomics del GitHub proporcionat per l'enunciat de la PAC.

L'arxiu inclou un conjunt de dades que s'han obtingut mitjançant un experiment de fosfoprotòmica que s'ha realitzat per analitzar (3 + 3) models PDX de dos subtipus diferents utilitzant mostres enriquides amb Fosfopeptids. En cada mostra s'ha realitzat l'anàlisi LC-MS de 2 duplicats tècnics. El conjunt de resultats va consistir en abundàncies normalitzades de senyals d'EM per a prop de 1400 fosfopèptids.

L'objectiu d'estudi és la cerca de dos grups tumorals diferenciats.

L'arxiu conté dues pàgines d'Excel, la primera amb les dades o valors d'abundància i el segon amb les dades dels targets.

1.1- Tractament de duplicats.

La presència de dades duplicades a causa de la presència de dos duplicats tècnics, s'ha gestionat amb la mitjana dels dos duplicats per treballar amb només una còpia de les dades.

2. Creació nou projecte R-studio sota control del git

Es crea un nou repositori sota el control de versions de Git per poder veure les diferents versions del codi del projecte, el codi es troba disponible al rmd del següent repositori de GitHub: https://github.com/laorfl/Ortega_Flores_Laia_PEC1

3. PCA

Es realitza un Principal Component Analisi (PCA) per veure patrons dominants i agrupacions de les dades, s'utilitza el paquet base de R, i la funció prcomp per realitzar el PCA, després es representa mitjançant un ggplot propi del paquet de R ggplot2.

4. Anàlisi amb limma

Es realitza una anàlisi de limma per poder buscar aquells fosfopeptids diferencialment expressats respecta la resta, s'ha escollit aquesta metodologia degut a experiència prèvia utilitzant aquesta anàlisi disponible al rmd del repositori de GitHub: https://github.com/LaiaOrFl/ATAC_embryos_mayflies_TFM.

S'han establert les dues condicions per a l'estudi MSS i PD, per veure les mostres que estan diferencialment expressades.

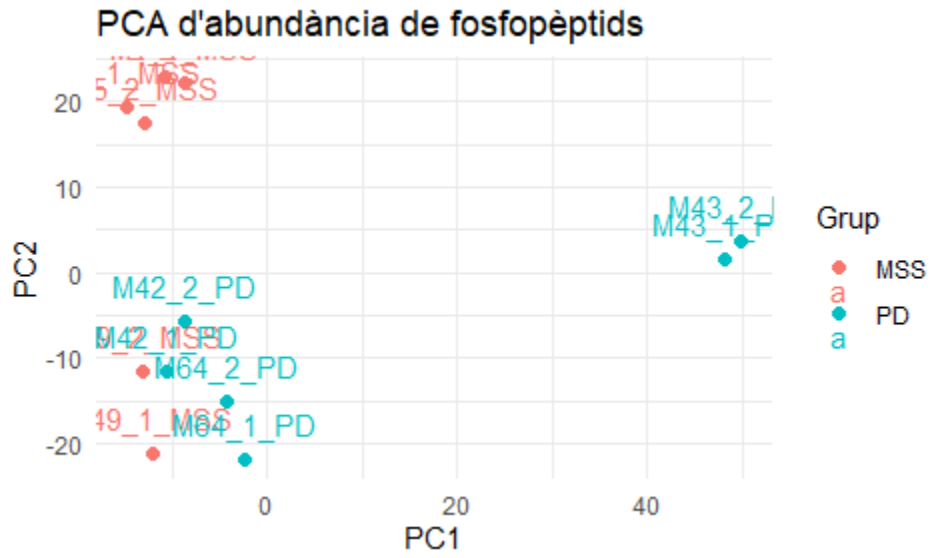
4.1. Procediment

Prèviament s'han establert les dues condicions d'estudi a MSS vs PD, després s'agafen les dades ja tractades i fem que s'ajustin a un model linial mitjançant la funció lmFit, seguidament es crea la matriu de contrasts i s'usa per veure els contrasts per les dues condicions d'estudi. S'aplica l'estadísitca de Bayes i el False Discovery Rate (FDR) per detectar falsos positius i augmentar la precisió de selecció de les dades diferenciades. Finalment, obtenim la topTable on hi ha la llista de fosfopèptids diferencialment expressats, dins del dataframe resultats_model. D'aquí es filtra els que tenen el p valor inferior a 0.05 i es representen en un gràfic del tipus volcanoplot.

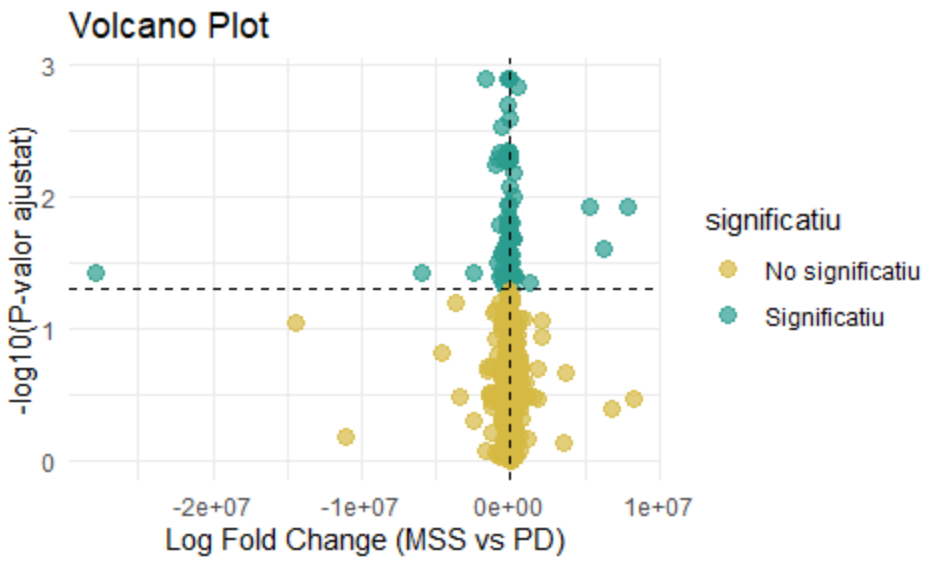
Resultats

L'anàlisi de les dades biològiques, s'ha inciat per un SummarizedExperiment que conté 1437 línies de fosfolípids amb modificacions, aquestes modificacions corresponen a metilacions, fosforilacions… entre d'altres. L'experiment conté Assays corresponent a la matriu de counts on hi ha les dades d'abundància de fosfopeptids i ColData que correspondria a la taula de metadades que conté informació de les mostres i les classifica en 2 grups MSS i PD.

Un cop creat l'experiment, es fa una Anàlisi de components principals (PCA) on mirem la distribució de les dades amb l'objectiu de determinar si hi ha una separació entre els 2 grups.



L'anàlisi de les dades biològiques, s'ha iniciat per un SummarizedExperiment que conté 1437 línies. Al PCA podem veure una diferenciació marcada al PC1 entre les mostres MSS i PD cosa que podria indicar patrons diferents en les modificacions de les seqüències, però hi ha un cluster amb dades de PD i MSS barrejades, aquest fet es pot explicar argumentant que hi ha similituds entre els patrons de modificacions de seqüències, però també pot deure's a una mala gestió de les dades, en la seva normalització.



Respecte a l'anàlisi diferencial de l'abundància amb limma s'ha representat gràficament amb un volcanoplot. L'eix X representa el Log fold change(LFC) indica si cada mostra està representada positivament o negativament (up o downregulated), essent el LDC positiu el més abundant en PD i el LFC negatiu el més abundant en MSS. L'eix Y representa la significació estadística amb el P-valor, els punts verds són els significatius (p-valor <0.05) i en daurat els no significatius, es a dir, que no superen el llindar establert.

La majoria de les mostres significatives, es troben sobreexpressats a PD respecte al grup MSS, això pot indicar una associació entre les modificacions seqüencials dels fosfolípids del grup i la seva progressió tumoral, però s'haurien de fer experiments addicionals per poder determinar si aquest fet correspon a un fet biològic rellevant o no.

Discussió

L'anàlisi de les dades mitjançant un PCA i una anàlisi diferència amb limma ha revelat que podrien haver diferències importants entre els dos grups tumorals MSS i PD, el PCA ha mostrat una diferenciació al PC1 dels dos grups i l'anàlisi ha identificat una sèrie de mostres que estan diferencialment expressades amb un p-valor < 0.05 dins del grup PD. Tot això suggereix que podria haver modificacions en la fosforilació de les seqüències que podrien explicar la progressió del tumor especialment en el grup PD, com per exemple l'activació d'una via de senyalització metabòlica que afavoreixi la progressió del tumor. També s'ha detectat una sèrie de mostres que indicaven patrons similars de modificacions seqüencials, sobretot si ens fixem en el PCA, això podria significar també que hi ha hagut error en el tractament de les dades, es podria fer un altre procés de normalització de les dades per fer-ne la comprovació.

Per interpretar biològicament aquests resultats, és necessària més informació sobre la investigació d'on s'han obtingut les dades i segurament més experiments addicionals, per exemple, comparacions d'aquests resultats amb altres bases de dades o fer un experiment on es vegi el patró de creixement cel·lular on hi hagi detectats aquestes modificacions que s'expressen diferencialment. També es podria fer una anàlisi funcional, per exemple rutes de senyalització, si es disposa de més dades, provinents d'altres òmiques, com per exemple gens es podria utilitzar gProfiler: https://biit.cs.ut.ee/gprofiler/gost, entre altres webs d'anàlisi interessants.

Conclusions

Si bé l'estudi ha permès veure una sèrie de mostres expressades diferencialment que podrien explicar certs patrons de progressió tumoral i sí que s'han observat diferenciacions entre els dos grups PD i MSS, no es poden fer grans conclusions sense disposar de més dades i informació de l'estudi que s'ha seleccionat i es necessitaria informació i investigació addicional on s'analitzin específicament aquestes mostres.

Referències

- Bonnin, S. (2020, marzo 9). *19.11 Volcano plots*. Github.io. https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html
- Jaadi, Z. (2021, abril 1). *Principal Component Analysis (PCA): A step-by-step explanation*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
- *Limma*. (s/f). Bioconductor. https://www.bioconductor.org/packages/release/bioc/html/limma.html
- *metaboData: A repository with a few public metabolomics datasets borrowed from different public open sources*. (s/f).
- Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (s/f). *SummarizedExperiment for coordinating experimental assays, samples, and regions of interest*. Bioconductor.org.https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html
- Pedro, L. (s/f). *Pedro Luque – Cómo usar Git/GitHub con R*. http://destio.us.es/calvo/asignaturas/ge_esco/tutorialusargitgithubrstudio/UsarGitGithubconRStudio.html
- *Resumen y configuración*. (s/f). Github.io. https://swcarpentry.github.io/git-novice-es/
- MATERIAL PROPI DE: Universitat Oberta de Catalunya. (s/f). *UOC*. Uoc.edu. https://uoc.edu/