Treball de final de grau

Protocol d'anàlisi per a dades d'expressió gènica amb Shiny

Grau d'Estadística Aplicada

Autor: Antonio Rodríguez Gómez

Supervisor: Mercè Farré

29 d'agost de 2018

Abstract (versió en català)

L'anàlisi de dades d'expressió genètica per a detectar expressions diferencials entre gens sota una condició donada és un dels grans reptes estadístics. La idea principal d'aquest treball, és la creació d'un protocol d'anàlisi interactiu de matrius d'expressió genètica. Es presenten mètodes estadístics com l'anàlisi de la variància (ANOVA) i eines de correcció per multiplicitat de contrastos (Tukey). També s'utilitzen mètodes gràfics com els mapes de calor, per visualitzar simultàniament clústers de casos i gens, i mètodes de components principals (PCA) per explorar correlacions entre tractaments en el conjunt dels gens. Tot aquest protocol d'anàlisi s'implementa en una aplicació web desenvolupada amb Shiny.

Abstract (versión en castellano)

El análisis de datos de expresión genética para detectar expresiones diferenciales entre genes mediante una condición establecida es uno de los grandes retos estadísticos. La idea principal de este trabajo, es la creación de un protocolo de análisis interactivo de matrices de expresión genética. Se presentan métodos estadísticos como el análisis de la varianza (ANOVA) y herramientas de corrección por multiplicidad de contrastes (Tukey). También se utilizan métodos gráficos como los mapas de calor, para visualizar simultániamente clústers de casos y genes, y mètodos de components principales (PCA) para explorar correlaciones entre tatamientos en el conjunto de genes. Todo este protocolo de análisis se implementa en una aplicación web desarrollada con Shiny.

Abstract (English version)

Data gene expression analysis is one of the major statistical challenges to detect differential expressions between genes under a given condition. The main idea is the creation of an interactive analysis protocol of gene expression matrix. Methods are presented for detecting differential expression using statistical hypothesis testing methods including analysis of variance (ANOVA). Methods for multiple testing correction and their application are described (Tukey). Graphical methods such as heatmaps are used in the analysis to detect clústers between genes and also between cases. Principal component analysis is used as graphical method to explore correlations between treatments in the set of genes. This analysis protocol is implemented in a web application developed in Shiny.

Índex

1 Introducció

Un dels reptes més grans de la biologia actualment és analitzar els volums massius de dades creats, per exemple, en la seqüenciació de DNA. La gran evolució de les tècniques de recollida de dades biològiques ha fet que sigui necessari el desenvolupament de metodologies eficients a l'hora de tractar i analitzar les dades. La disciplina que recull aquestes metodologies s'anomena bioinformàtica.

La bioinformàtica és un àrea emergent interdisciplinària que s'ocupa de l'aplicació de l'informàtica a la recopilació, emmagatzematge, organització, anàlisis, manipulació d'informació relativa a les dades biològiques o mèdiques.

Aquest treball s'ha centrat en l'anàlisis de bases de dades d'expressió genètica per a comparar diferents condicions experimentals. Al llarg del treball s'utilitzen técniques per analitzar aquests tipus de dissenys on l'objectiu recau en veure quins gens s'expressen de manera significativament diferenciada, sota condicions experimentals establertes.

Amb aquesta premissa, el treball també s'ha enfocat en crear un aplicatiu web capaç de fer una anàlisi estadística de les matrius d'expressió gènica. L'aplicatiu ha sigut programat amb R, per mitjà del paquet Shiny. Aquest paquet és capaç d'implementar el codi R de manera interactiva. No només implementa el codi R, sinó també és capaç d'interactuar amb diferents llenguatges com html, css o java. L'aplicació, en el seu conjunt, s'ha organitzat i compartit per mitjà d'un repositori creat en la plataforma GitHub. D'aquesta manera el codi queda a disposició de qualsevol usuari que vulgui utilitzar-lo o consultar els métodes emprats.

1.1 Introducció als conceptes bàsics de la bioinformàtica

Cada organisme es defineix pel seu material genètic, el genoma. La informació genètica la trobem emmagatzemada en una macromolècula anomenada DNA, que es troba al nucli de cada cèl·lula.

Un gen consisteix en un segment de DNA que conté el codi per a la producció d'una proteïna. Una única cadena d'ADN conté milers de gens i cadascun sintetitza una proteïna concreta. Per fer-nos a la idea, els humans tenim al voltant de 20.000 gens. La longitud i seqüència d'un gen determina la grandària i la forma de la proteïna que sintetitza, i quina funció tindrà aquesta proteïna dins de l'organisme.

La dotació de gens que presenta una espècie, s'anomena genotip, i l'aparença externa d'un caràcter genètic, l'anomenem fenotip. L'expressió del genotip ve determinat, a més de per la càrrega genètica, per l'ambient i el comportament dels éssers vius. Si un gen no s'expressa en un individu, aquest tindrà el mateix fenotip que un individu que no presenti el gen. Però com podem arribar a obtenir una mesura de l'expressió dels gens? Existeixen tècniques per quantificar-ho? La resposta és sí.

La mesura de l'expressió gènica generalment és dur a terme quantificant el material genètic del gen. Una tècnica molt utilitzada de mesura de l'expressió gènica que utilitza ARN missatger és la denominada transcripció inversa, seguida de la reacció en cadena quantitativa de la polimerasa (qPCR [1]). Una de les seves principals característiques és la seva sensibilitat, ja que només necessita una única molècula per iniciar el procés de replicació. A més, és molt robusta gràcies al fet que permet utilitzar diferents productes biològics, com cabells, teixits, mucoses, sang, etc. Aquesta tècnica és fonamental per l'anàlisi de dades d'expressió gènica, perquè per a l'obtenció d'aquestes dades es requereix una quantitat suficientment gran de producte biològic que no sempre és de fàcil obtenció, per tant, és important disposar d'una tècnica que faciliti la seva replicació de forma controlada, robusta i eficient.

Els centres de genòmica s'encarreguen de fer aquests processos i de retornar els resultats en matrius de dades on es recullen els nivells d'expressió per a cada gen. Per tant, és important fer un bon disseny experimental ja que aquests procesos són costosos i requereixen de temps.

1.2 Cas d'estudi: Protocol d'análisi d'un OpenArray

Des de la facultat de veterinaria de l'Universitat Autònoma s'han dut a terme estudis experimentals sobre l'expressió dels gens animals en certes condicions experimentals. L'aplicatiu web ha sigut creat per donar suport estadístic al grup d'investigació de la UAB i al *Servei de Nutrició i Benestar Animal* (SNiBA).

El cas d'estudi que el treball ha contemplat consisteix en un experiment amb animals, concretament, amb porcins. Durant l'experiment s'administraven diferents tractaments/dietes als porcins. D'aquesta manera es volia veure l'afectació d'alguns tractaments en la regulació intestinal i com afectava el creixement dels porcins.

Les dades utilitzades en aquest treball han sigut proporcionades per *Servei de Nutrició i Benestar Animal* (SNiBA) per mitjà de la tecnologia OpenArray. El material biològic que s'ha utilitzat per l'obtenció de les dades, han sigut diferents tipus de teixits de l'intestí. Els gens van ser escollits amb criteris científics pels investigadors i tenen un significat concret dins del funcionament de la regulació intestinal.

Encara que l'aplicatiu web s'ha creat a partir d'aquest estudi, el codi s'ha dissenyat amb funcions genè-

[1] Aquesta tècnica serveix per amplificar un fragment d'ADN i la seva utilitat rau en el fet que després de l'amplificació resulta molt més fàcil identificar material genètic amb una gran precisió.

riques que es poden aplicar a altres situacions experimentals, sempre que es vulguin comparar tractaments (dietes, medicacions, situacions experimentals, etc.) a partir de dades d'openarray de diversos gens.

2 Protocol d'anàlisi

En aquest apartat queden definits els mètodes estadístics utilitzats en el protocol d'anàlisi. Cada mètode ha sigut implementat en l'aplicatiu i més endavant es mostren els resultats del cas d'estudi.

2.1 Anàlisi de la variància (ANOVA)

L'anàlisi de la variància (ANOVA) és el mètode clàssic per comparar mitjanes entre grups, dos grups o més.

Observació. Hi ha una sèrie de supòsits que s'han de fer abans que s'apliqui l'ANOVA, la desviació en aquests supòsits portaran a resultats que poden ser enganyosos o inexactes. Aquests supòsits inclouen la independència, normalitat i variància constant dels errors. En algunes situacions, hi ha transformacions que poden ser utilitzades per evitar les violacions d'aquests supòsits, com ara la transformació logarítmica de les dades.

Suposem que tenim N observacions repartides en k grups. Llavors x_{ij} seria l'individu j corresponent al grup i. En aquest cas assumim que l'estudi és balancejat, és a dir, el nombre d'individus per grup és el mateix, $n = \frac{N}{k}$. Denotem $\bar{x}.$ com la mitjana de la mostra global, i $\bar{x_i}$ com la mitjana del grup i. Les observacions es poden tornar a escriure com:

$$x_{ij} = \bar{x}. + (\bar{x_i} - \bar{x}.) + (x_{ij} - \bar{x_i})$$

Això ens porta al següent model:

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

on μ i α_i són la mitjana global i l'efecte diferencial del grup i, respectivament. . S'assumeix que el terme d'error ϵ_{ij} és iid i segueix una distribució normal

$$\epsilon_{ij} \sim \mathcal{N}(\mu, \sigma^2)$$

La hipòtesi nul·la en un model ANOVA és que les mitjanes dels grups són iguals, que equival a:

$$\alpha_1 = \alpha_2 = ... = \alpha_k$$

Es pot mesurar la quantitat total de variabilitat entre observacions sumant els quadrats de les diferències entre :

$$SST(\text{Suma de quadrats totals}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}.)^2$$

La variabilitat total es pot desglossar en 2 termes:

1. La variabilitat entre grups:

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}.)^2$$

   amb $k-1$ graus de llibertat.

2. La variabilitat intra grups:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2$$

   amb $N-k$ graus de llibertat.

Per tant, podem escriure la suma de quadrats totals com:

$$SST = SSG + SSE$$

Si la variabilitat entre grups és gran en relació amb la variabilitat intra grups, llavors les dades suggereixen que les mitjanes de les poblacions són significativament diferents. Es consideren les mitjanes quadràtiques:

$$MSG = \frac{SSG}{(k-1)}$$

$$MSE = \frac{SSE}{(N-k)}$$

dividint pels graus de llibertat. L'estadístic de contrast de l'ANOVA es defineix com la ràtio entre les dues mitjanes quadràtiques:

$$F = \frac{MSG}{MSE}$$

L'estadístic F segueix una distribució F de Snedecor amb $k-1$ i $N-k$ graus de llibertat, sota les condicions imposades al model, i quan la hipòtesi nul.la és certa. Si la hipòtesi nul·la és certa, F seria proper a 1. D'altra banda, si la mitjana quadràtica entre grups MSG és gran, suposaria un valor gran de l'estadístic F. Bàsicament, l'ANOVA examina les dues fons de la variància total i mira quina part contribueix més. Per aquest motiu, s'anomena anàlisi de la variància encara que la intenció sigui comparar les mitjanes dels grups.

2.1.1 Anova per a dissenys desbalancejats

Si la mida d'una mostra és diferent per a cada grup o tractament, es reduirà la potència estadística en comparació amb un disseny balancejat. La potència estadística no és més que la probabilitat que l'estudi detecti un efecte si realment hi ha un efecte. És invers a l'error Tipus II, això significa que la probabilitat de cometre un error de tipus II serà petita. En altres paraules, ens constarà més trobar diferencies signficatives i els nostres resultats estaran esbiaixats.

2.2 Correcció per multiplicitat de contrastos

Un problema comú què ens podem trobar a qualsevol investigació és voler comparar més de dos grups de dades per detectar possibles diferències entre ells. La utilització de models d'ANOVA ens pot permetre detectar diferències, a escala global, entre les mitjanes involucrades, però en moltes ocasions volem detectar les diferències entre grups concrets. Aquest cas només és possible mitjançant l'ús dels Procediments de Comparacions múltiples (PCM).

En aquest treball el nostre interès no és avaluar si un o dos gens concrets s'expressen d'una forma diferencial entre les condicions considerades. Volem veure això a un nivell global i respondre a una pregunta com: Quins gens s'expressen d'una manera diferent (diferencial si utilitzem la literatura biològica) en els grups/tractaments que considerem? L'objectiu és poder contestar aquesta pregunta de manera que puguem controlar les vegades que afirmem expressions diferencials quan realment no la tenen (error de tipus I).

Si numerem els gens $i = 1, ..., N$ llavors per a l'i-èssim gen estem considerant el següent contrast:

- H_0 : El gen i no té una expressió diferencial entre les condicions considerades.
- H_1: El gen i té una expressió diferencial entre les condicions considerades.

Si plantegem aquest contrast per a cada gen, podem denotar com G el conjunt d'hipòtesis nul·les que estem avaluant. El número d'hipòtesis que avaluem és conegut a priori, ja que correspon al número de gens que volem avaluar. És important destacar que en els estudis d'investigació en intentar acceptar o rebutjar la hipòtesi nul·la (H_0) es poden cometre dos tipus d'errors:

- Error de tipus I: Rebutjar H_0 quan realment és certa.
- Error de tipus II: No rebutjar H_0 quan realment és falsa.

Imaginem que fem un test contrastant diferencies entre mitjanes, i fixem un nivell de significació $\alpha = 0.05$, i sabem que la hipòtesis nul·la és certa; llavors l'error de tipus I serà exactament el nivell de significació α. Per tant, podem definir la probabilitat de tenir un fals positiu en un test, és a dir, rebutjar H_0 quan realment és certa:

$$P(\text{Fals positiu}) = \alpha$$

$$P(\text{Veritable positiu}) = 1 - \alpha$$

Per tant, si definim m tests d'hipòtesis podem definir la probabilitat d'almenys tenir 1 fals positiu com:

$$P(\text{Veritables positius en } m \text{ tests}) = (1-\alpha)^m$$

$$P(\text{Almenys 1 fals positiu en } m \text{ tests}) = 1 - (1-\alpha)^m$$

Per exemple, si tenim un test, i fixem $\alpha = 0.05$, la probabilitat d'obtenir almenys 1 fals positiu és de:

$$P(\text{Almenys 1 fals positiu}) = 1 - (1 - 0.05) = 0.05$$

Si ara tenim 50 tests i calculem la mateixa probabilitat:

$$P(\text{Almenys 1 fals positiu}) = 1 - (1 - 0.05)^{50} = 0.92$$

Aquí recau el gran problema de la multiplicitat de contrastos, podem observar que si fem un test moltes vegades, hi ha una inflació en l'error de tipus I.

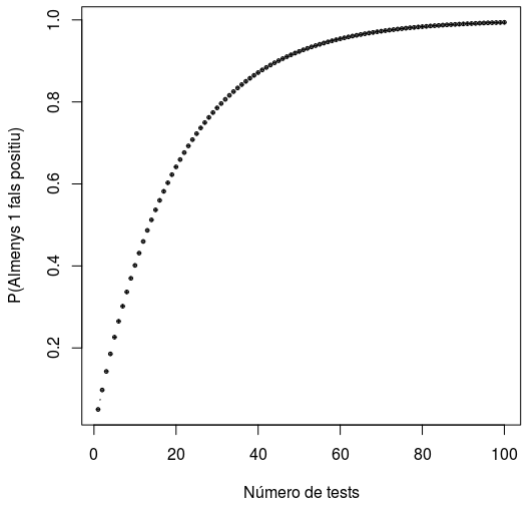


Figura 1: Gràfic contraposant el nombre de tests amb la probabilitat d'obtenir almenys 1 fals positiu. S'observa un increment en la probabilitat quan augmenta el nombre de tests.

En el nostre estudi es important tenir clar aquest problema, ja que si tenim molts gens, i no apliquem una correcció, podem caure en l'error d'afirmar que un gen s'expressa diferencialment quan realment no ho fa.

2.2.1 Ratio falsos positius (*False discovery Rate (FDR)*)

Existeixen molts mètodes per corregir el problema de la multiplicitat de contrastos. El més simple és el mètode de Bonferroni, on cada p-valor es multiplica pel nombre de tests realitzats (acotant la probabilitat màxima a 1). És un mètode molt conservador i no és el més indicat per al nostre cas d'estudi. Per a escenaris de *large-scale multiple testing* com els estudis de genòmica, els quals es realitzen milers de test de forma simultània, el resultat dels mètodes clàssics de correcció (Bonferroni, Tukey, etc.) és massa conservador i impedeix que es detectin diferències reals. Una alternativa és controlar el *FDR*.

El *FDR* es defineix com la proporció esperada de falsos positius d'entre tots els tests considerats com significatius. L'objectiu de controlar el *FDR* es establir un límit de significació per a un conjunt de tests tal que, d'entre tots els tests considerats com significatius, la proporció de falsos positius no superin un determinat

valor. Un altre avantatge afegit és la seva fàcil interpretació, per exemple, si un estudi publica resultats estadísticament significatius per a un FDR del 10%, el lector té la seguretat que, com a màxim, un 10% dels resultats considerats com a significatius poden ser realment falsos positius.
La primera aproximació per controlar el FDR va ser descrita per Benjamini & Hochberg en 1995. D'acord amb la seva publicació [5], si es desitja controlar que en un estudi amb n comparacions el FDR no superi un percentatge d hem de:

- Ordenar els n tests de menor a major p-valor (p_1,p_2,..., p_n).
- Es defineix k com l'última posició per la qual es compleix que $p_i \leq d\frac{i}{n}$.
- Es consideren significatius tots els p-valors fins a la posició k (p_1,p_2,..., p_k).

El mètode proposat per Benjamini & Hochberg assumeix a l'hora d'estimar el nombre d'hipòtesis nul·les erròniament considerades falses, que totes les hipòtesis nul·les són certes. Com a conseqüència, l'estimació del FDR està inflada i és un mètode conservador. Per poder veure l'afectació d'utilitzar un mètode com Bonferroni o utilitzar el métode Benjamini & Hochberg, tenim la següent taula d'exemple:

	P-valor	Bonferroni	Benjamini&Hochberg
1	0.0010	0.0550	0.0037
2	0.0020	0.11	0.0069
3	1	1	1
4	0.0010	0.0550	0.0037
5	0.0010	0.0550	0.0037
6	0.0010	0.0550	0.0037
7	0.25	1	0.3929
8	0.48	1	0.6286
9	0.09	1	0.1650
10	0.51	1	0.6523

Taula 1: La primera columna de la taula conté els p-valors, sense corregir ni ordenar. La segona conté els p-valors amb la correcció de Bonferroni i la tercera el FDR de Benjamini&Hochberg, per a un total de 55 p-valors (Encara que només es mostren els 10 primers). Bonferroni és un mètode molt més conservador i cap p-valor és significatiu (dels 55 p-valors, 0 són significatius). En canvi, amb el mètode de Benjamini&Hochberg, del total de p-valors, hi ha 26 significatius (en aquest cas hem fixat un $\alpha = 0.05$ i un FDR= 0.05)

A l'hora de decidir quin tipus de correcció aplicar, és important utilitzar un mètode adequat per tal d'obtenir resultats més acurats. Els mètodes que només depenen del nombre de tests no són els més potents com hem vist anteriorment.

2.3 Comparacions múltiples

Un cop realitzat l'anàlisi de la variància i si aquest confirma l'existencia de diferències significatives entre els grups o tractaments, és convenient investigar quines mitjanes són diferents. El conjunt de tècniques que tracten aquest problema es denominen *contrastos per comparacions múltiples*.

2.3.1 Métode de Tukey (*honestly-significant-difference*)

Recordem que quan el nombre de possibles comparacions és elevat, per a un nivell de significació α donat, pot conduir a una inflació de l'error de tipus I, com també hem vist quan parlàvem de multiplicitat de contrastos. Per identificar quins tractaments són significativament diferents entre ells i corregir el problema de la inflació de l'error de tipus I, hem utilitzat el mètode de Tukey i les seves hipòtesis són:

$$H_0 : \mu_i = \mu_j \text{ per a cada parella de mitjanes } i \neq j$$

$$H_1 : \mu_i \neq \mu_j \text{ almenys una parella de mitjanes } i \neq j$$

Aquest contrast es basa en la distribució del rang estudentitzat, que es defineix a partir del nombre de grups a comparar i dels graus de llibertat de l'estimador de la variància. Aquests tipus de procediments, permeten superar les dificultats que hi ha quan augmentem el nombre de grups a comparar i no podem controlar els falsos positius. En general, són mètodes conservadors, es a dir, la probabilitat real de rebutjar la hipòtesi nul·la quan és certa és menor que el nivell de significació α fixat.

Per definir el rang estudentitzat, suposem que disposem de k observacions independents $y_1, y_2, ..., y_k$ d'una distribució Normal amb mitjana μ i variància σ^2. Suposem també que disposem d'un estimador S^2 de σ^2 que té v graus de lliberat i és independent de les y_i. Definim R com el rang d'aquest conjunt d'observacions,

$$R = max(y_i) - min(y_i)$$

Sota aquestes condicions, es defineix el rang estudentitzat com el quocient,

$$\frac{max(y_i) - min(y_i)}{S} = \frac{R}{S}$$

que es denota com $q_{k,v}$. Aquest estadístic segueix una distribució que depèn dels paràmetres k i v, coneguda com la distribució del rang estudentitzat.

L'estadístic de contrast que utilitza el test de Tukey queda definit com:

$$Q_{I,n-I} = q_{I,n-I}(\alpha)\sqrt{\frac{MSE}{n}},$$

on $q_{I,n-I}$ és la quantila teòrica,

$$q_{I,n-I}(\alpha) = \frac{\bar{X}_j - \bar{X}_k}{\sqrt{\frac{\hat{S}^2}{n}}}$$

$\hat{X}_j$ i $\hat{X}_k$ són la mitjana del grup j i k, respectivament, i $\hat{X}_j > \hat{X}_k$. $\hat{S}^2$ és l'estimació de la variància del error o residual; i n és la grandària mostral per a tots els grups; on I i $n-I$ són els graus de llibertat de la distribució del rang estudentitzat (I correspon al nombre de nivells que té el factor, n correspon a la grandària mostral).En el cas de tenir grandàries mostrals diferents entre els nivells del factor, hem d'utilitzar un altre n (mitjana harmònica):

$$n_h = \frac{t}{\sum_{i=1}^{t} \frac{1}{n_i}}$$

La diferència entre mitjanes serà significativa amb un nivell de significació α si

$$|\bar{X}_j - \bar{X}_k| > HSD$$

on $HSD = Q_{I,n-I}$.

L'interval de confiança per a la diferència de mitjanes el definim com:

$$IC(\mu_j - \mu_k)_{(1-\alpha)} = (\bar{X}_j - \bar{X}_k) \pm q_{I,n-I,1-\alpha}\sqrt{\frac{MSE}{n}}$$

Si l'interval de la diferència inclou el 0, no rebutjem la hipòtesi nul·la del test i per tant, no hi ha diferències entre μ_j i μ_k.

2.4 Mètodes descriptius visuals

A l'hora de crear el protocol d'anàlisi teníem la necessitat de descriure el comportament de l'expressió dels gens segons el tractament i tenint en compte la funcionalitat de la família del gen. Una manera de veure aquests patrons era amb l'ús de tècniques visuals, que tenen una base matemàtica que explicarem en els següents apartats.

2.4.1 Heatmap

Un *heatmap* (o mapa de calor) és una representació gràfica de dades on els valors individuals continguts en una matriu es representen com a colors.

En mapes de calor, les dades es mostren en una quadrícula on cada fila representa un gen i cada columna representa una mostra o cas. El color i la intensitat de les caselles s'utilitzen per representar canvis en l'expressió gènica.

L'exemple de la Figura 2 l'hem obtingut amb l'aplicatiu que hem creat:

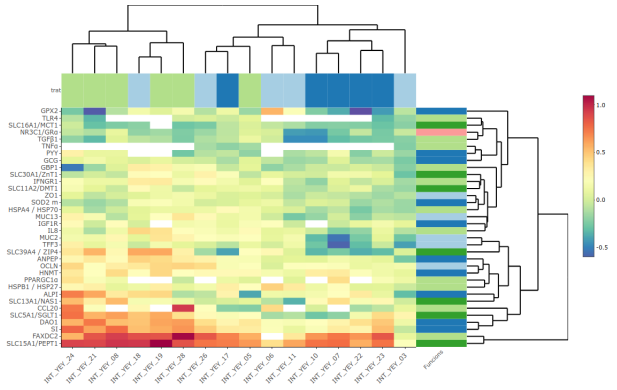


Figura 2: `Heatmap` generat a partir de 34 gens i 16 mostres. A les files trobem els gens i a les columnes les mostres.Veiem agrupacions (clústers) tant de gens com de mostres obtingudes mitjançant un algoritme jeràrquic. Per obtenir més informació, s'ha afegit variables alienes a la matriu d'expressions , concretament, a quin tractament correspon la mostra, i quina funció desenvolupa el gen.

El mapa de calor també es pot combinar amb mètodes de *clustering* que agrupen els gens i les mostres junts en funció de la similitud del seu patró d'expressió gènica. Això pot ser útil per identificar els gens que normalment s'expressen molt semblant i detectar patrons sota condicions o covariables establertes. El mètode implementat per la funció `heatmap`, utilitza l'anàlisi de clústers jeràrquics.

En el cas dels mètodes jeràrquics les dades s'ordenen en nivells de manera que els nivells superiors contenen als inferiors. La jerarquia construïda permet obtenir també una partició de les dades en grups. S'utilitza la matriu de distàncies o similituds entre els elements de la matriu original de les dades.

Els algoritmes *jeràrquics* poden ser de dos tipus: De divisió i d'aglomeració.

L'*algoritme de divisió* assumeix que en un primer pas totes les dades conformen un sol conglomerat. Aquest clúster es va dividint successivament en conglomerats més petits d'acord a algun criteri seleccionat prèviament. El resultat d'aquest procediment es representa pel dendrograma.

L'*algoritme d'aglomeració* assumeix que cada observació inicialment és un conglomerat i en cada pas s'associen els conglomerats més similars fins a arribar a un sol clúster.

La implementació `hclust` (*hierarchical cluster analysis*) de `R` utilitza el *Mètode de Ward* que calcula i actualitza a cada pas la dissimilaritat entre clústers, aquest mètode és d'aglomeració.

Mètode de Ward

Ward va proposar que la pèrdua d'informació que es produeix en integrar els diferents individus en clústers pot mesurar-se a través de la suma total dels quadrats de les desviacions entre cada punt (individu) i la mitjana del clúster en el qual s'integra. Perquè el procés de clusterització resulti òptim, en el sentit que els grups formats no distorsionin les dades originals, proposava la següent estratègia:

A cada pas de l'anàlisi, es considera la possibilitat de la unió de cada parell de grups i s'obta per la fusió d'aquells dos grups que menys incrementin la suma dels quadrats de les desviacions en unir-se.

Definim:

- x_{ij}^k com el valor de la j-èssima variable sobre l'i-èssim individu del k-èssim clúster, que té n_k individus.
- m^k com el centroide del clúster k, amb components m_j^k.
- E_k com la suma de quadrats dels errors del clúster k, és a dir, la distancia euclídea al quadrat entre cada individu del cluster k al seu centroide:

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij}^k - m_j^k)^2$$

- E com la suma de quadrats dels errors per a tots els clústers, és a dir, si suposem h clústers:

$$E = \sum_{k=1}^{h} E_k$$

El procés comença amb m clústers, cada clúster només té un sol individu, per tant, cada individu coincideix amb el centre del clúster i en aquest primer pas $E_k = 0$, per a cada clúster i això fa que $E = 0$. L'objectiu del mètode de Ward és trobar en cada etapa aquells dos clústers els quals la seva unió proporcioni el menor increment en la suma total d'errors E. Suposem que els clústers C_p i C_q s'uneixen resultant un nou clúster C_t, llavors definim l'increment de E com,

$$\triangle E_{pq} = E_t - E_p - E_q$$

el procés es repeteix fins a l'obtenció del dendrograma i l'agrupació dels individus en els diferents clústers. Un exemple aclaridor d'aquest procés iteratiu, amb dades reals, es pot trobar a l'annex A.

2.4.2 Anàlisi de components principals (ACP): gràfiques

L'anàlisi de components principals és una de les diferents maneres per analitzar l'estructura d'una matriu de correlacions donada. En aquest apartat definirem la teoria de l'anàlisi de components principals i la seva representació gràfica.

L'objectiu de l'ACP

Donades p variables $X_1, X_2, \ldots, X_p$ que suposarem centrades (sovint també normalitzades amb variància igual a 1), volem construir p noves variables $Z_1, Z_2, \ldots, Z_p$, que anomenarem compoents (les components principals seran cert subconjunt d'aquestes) amb les següents condicions:

1. Les noves variables han de ser combinació lineal de les variables originals.
2. Les noves variables han de ser incorrelacionades.
3. Les noves variables han de contenir la mateixa informaicó (variància total) que les originals, i estar ordenades de major a menor variància.

El teorema que demostra l'existència i unicitat de les components, i en dóna el procediment de càlcul és el següent:

Teorema 1. *Les components venen donades per la transformació lineal*

$$Y = V^t X$$

on la matriu V és la matriu ortogonal de vectors propis columna donada per la descomposició espectral (Jordan) de Σ - la matriu de covariàncies de X -, amb els vectors propis ordenats segons el valor propi λ_j, de més gran a més petit. A més, com que la descomposició espectral compleix

$$\Sigma = V \Lambda V^t, \tag{1}$$

amb Λ la matriu diagonal de valors propis. La descomposició (1) és equivalent a

$$\Lambda = V^t \Sigma V = \Sigma_Z. \tag{2}$$

De la igualtat (2) del teorema teorema es dedueix que les components són incorrelacionades perquè tenen matriu de covariàncies diagonal i estan ordenades de major a menor variancia. En particular tenim que

$$Var(Z_j) = \lambda_j, \quad \text{amb} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p.$$

De l'equació (2), igualant les traces, tenim que:

$$VT(Z) := \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 =: VT(X) \tag{3}$$

on VT indica la variància total. Aquesta igualtat (3) s'anomena *principi de conservació de la variància.*

De tot l'anterior, és lògic que s'escullin les k primeres components, si representen un percentatge "prou important" (70%, 80%, etc.,) de la variància total:

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} \geq 0.7 \text{ (o bé 0.8, etc.)}$$

Les k primeres components, escollides segons aquest o un altre criteri, s'anomenen *components principals*.

Pel cas més habitual en que les variables inicials estan estandarditzades (centrades i escalades), és a dir $\sigma_i^2 = 1,\ \forall i$, es té que

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = 1.$$

2.4.3 Gràfica de les variables

Suposem variables centrades i escalades. La matriu $W = V\Lambda^{1/2}$ s'anomena matriu de pesos (*loadings*) i dóna les coordenades per expressar les variables com a combinació lineal (o mitjana ponderada) de les components (veure la gràfica de les variables, a l'apartat següent). Si suposem que $k = 2$, la variable original X_i es representa amb les coordenades (w_{i1,w_i2}), en forma de "fletxa". La igualtat (1) és equivalent a

$$\Sigma = W^t W$$

Pels termes diagonals, resulta:

$$\sigma_i^2 = 1 = w_{i1}^2 + w_{i2}^2 + \cdots + w_{ip}^2$$

Per tant, els coeficients w_{ij}^2, o cosinus quadrats, sumen 1 i, si la variable està ben representada en dimensió $k = 2$, es té que

$$\text{long.vector}^2 = w_{i1}^2 + w_{i2}^2 \approx 1$$

i la fletxa d'aquesta variable s'apropa a la circumferència de radi 1.

Si una variable quedés explicada en un 100%, la longitud del vector seria 1. Si, en canvi, quedés molt mal explicada, la longitud del vector corresponent seria propera a zero. Per tant, com més properes al cercle unitat estiguin les variables, més llargues seran les fletxes i millor quedaran representades. És a dir, menor serà la pèrdua d'informació que es produeix en substituir les variables originals per les components principals. En aquest exemple, el gràfic ens mostra que les quatre variables queden molt ben representades.

"*Els angles que formen dues fletxes entre sí, ens mostren la correlació existent entre les corresponents variables*"

El cosinus de l'angle format pels dos vectors és proporcional a la seva correlació. Com més petit sigui l'angle, més gran la correlació. Així, dues variables que quedin representades en punts molt propers tindran vectors que formaran un angle proper a zero, fet que indica que tindran una correlació molt forta. En canvi, dues variables representades per punts molt separats i els vectors de les quals formin un angle proper a l'angle recte, seran pràcticament incorrelacionades.

"*Els sentits dels vectors indiquen el signe de la correlació existent entre les corresponents variables*"

Dos vectors que tinguin una mateixa direcció però sentits oposats tindran una correlació molt forta en sentit negatiu. És a dir, valors elevats en una variable es corresponen amb valors baixos en l'altra.

"*La posició de les variables en el gràfic mostra l'estructura de les dades i suggereix possibles interpretacions de les components principals*"

El fet que hi hagi un grup de variables ben representades i en posicions properes en el gràfic indica que totes elles estan molt relacionades entre sí i que ens subministren bàsicament la mateixa informació. Quan apareixen varis grups de variables vol dir que hi ha diferents aspectes descrits per la matriu de dades. També el fet que alguna variable quedi molt ben representada i que en el gràfic li correspongui un punt proper a un dels eixos de coordenades, ens està indicant que una de les dues primeres components principals està fortament correlacionada amb aquesta variable i, per tant, se li pot donar una interpretació similar.

3 Cas d'estudi

El cas d'estudi consisteix en un disseny experimental fet amb porcins que han sigut alimentats amb diferents fonts de proteïnes (diverses combinacions de productes de soja, plasma animal i mucosa) a principis del destetament. L'objectiu principal del estudi és avaluar la salut intestinal del porc poc després d'administrar les diferents dietes. A més, és vol veure si els tractaments tenen un efecte en el creixement del porc, i per tant, en la productivitat de l'empresa.

3.1 Descripció de l'estudi

L'estudi es defineix breument en els següents punts

Objectiu de l'estudi

- Analitzar i comparar els diferents tractaments i com afecten a l'expressió dels gens.
- Trobar relacions entre la funcionalitat del gen i la seva expressió sota diferents tractaments.

Disseny de l'estudi
Estudi experimental amb una mostra total de 48 casos. És un un dels primers experiments realitzats pel *SNiBA*, per tant, l'experiment es totalment exploratori i es podría definir com una prova per a futurs dissenys.
Tractaments

- **T1**: Inclusió d'aliments de soja a la dieta.
- **T2**: Inclusió de plasma animal a la dieta.
- **T3**: Inclusió de 33% de plasma animal i 66% de mucosa a la dieta.
- **T4**: Inclusió de 50% de plasma animal i 50% de mucosa a la dieta.

Variables explicatives
La principal variable explicativa és:

- Tractament

Les variables secundàries que també trobem a l'estudi:

1. ID: nom del cas.
2. Teixit: l'estudi es fa paral·lelament amb 2 tipus de teixits diferents, per tant, es farà un estudi amb cada teixit. (Dels 48 casos, queden 24 casos per teixit)
    - Jejunum
    - Ileum

Variables genètiques
Cada gen és una variable on es mesura l'abundancia de material genètic. Els gens han sigut escollits pels investigadors basant-se en les seves recerques i reben el nom de gens *diana*. A l'annex B trobem una taula de tots els gens de l'estudi amb la seva funcionalitat dins de l'organisme.

Tractament de dades faltants
El tractament dels NA's és el següent:

1. S'eliminen aquelles files i columnes sense cap observació vàlida.
2. S'eliminen aquelles columnes (gens) que tinguin més del 50% de valors perduts en algun tractament.

Mètodes estadístics
L'anàlisi estadístic ha sigut realitzat utilitzant `R` amb la versió 3.4.3.

Per a tots els tests estadístics s'ha aplicat un nivell de significació del 5% ($P < 0.05$). S'han realitzat correccions per multiplicitat de contrastos; *Benjamini&Hochberg* per l'ANOVA i *Tukey* per les comparacions 2 a 2 entre tractaments.

Tota la documentació, codi i *outputs* han sigut emmagatzemats en un repositori públic a la plataforma *GitHub*.
Tractament de les dades
Les dades han sigut validades abans de l'anàlisi, qualsevol inconsistencia en el format de les dades ha sigut eliminat o canviat. Més endavant s'expliquen quins criteris de l'estructura de la base de dades són els adequats perquè sigui funcional a l'aplicatiu.

A més, hem aplicat una tranformació logarítmica als valors d'expressió gènica per a cada gen, amb l'objectiu de treballar amb una distribució normal i poder aplicar els mètodes estadístics descrits anteriorment.

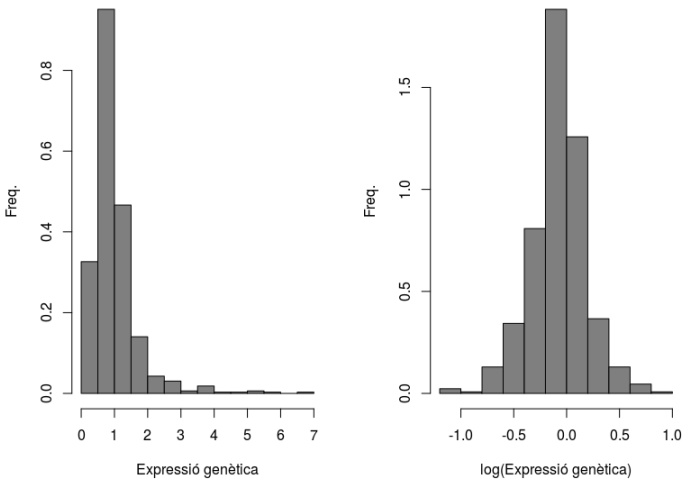


Figura 3: (1) El gràfic de l'esquerra mostra les dades sense cap tipus de transformació.(2) El gràfic de la dreta mostra les dades amb una transformació logarítmica.

3.2 Resultats

Degut al disseny de l'estudi, s'analitzen paral·lelament 2 tipus de teixits, *Ilenum* i *Jejunum*. L'anàlisi que es duu a terme consta d'una primera part on s'examinen quins gens s'expressen diferencialment en cada teixit sota els 4 tractaments, i es determinen quins tractaments són diferents entre ells per a cada gen. Després hi ha una segona part més desciptiva on s'analitzen els patrons trobats en els mètodes visuals.

3.2.1 Anàlisi de la variància (ANOVA)

Amb l'objectiu de trobar quins gens s'expressen diferencialment sota els diferents tractaments s'ha aplicat l'anàlisi de la variancia per a cada teixit:

Nom del gen	Funció del gen	Estadístic F	P-valor	P-valor (Benjamini&Hochberg)
TFF3	Barrier Function	5.2927	0.0086	0.0603*
OCLN	Barrier Function	1.2861	0.3094	0.4572
ZO1	Barrier Function	0.9128	0.4544	0.5979
MUC2	Barrier Function	7.2461	0.0022	0.0272***
MUC13	Barrier Function	2.1769	0.1261	0.2425
SI	Enzymed/Hormone	3.1867	0.0488	0.1525
DAO1	Enzymed/Hormone	1.2813	0.3109	0.4572
HNMT	Enzymed/Hormone	0.5736	0.6396	0.6953
ANPEP	Enzymed/Hormone	1.9658	0.1553	0.2773
GCG	Enzymed/Hormone	1.1972	0.3390	0.4709
IGF1R	Enzymed/Hormone			
GPX2	Enzymed/Hormone	10.5231	0.0003	0.0079***
SOD2m	Enzymed/Hormone	3.3427	0.0424	0.1516
ALPI	Enzymed/Hormone	2.1929	0.1241	0.2425
TNFα	Inmune Response			
TGFβ_1	Inmune Response	1.7353	0.1956	0.3260
CCL20	Inmune Response			
IFNGR1	Inmune Response	4.5897	0.0148	0.0741*
HSPB1.HSP27	Inmune Response	0.8298	0.4947	0.6184
HSPA4.HSP70	Inmune Response	0.6819	0.5745	0.6529
FAXDC2	Inmune Response			
GBP1	Inmune Response	0.6991	0.5647	0.6529
IL8	Inmune Response	5.1389	0.0096	0.0603
SLC5A1.SGLT1	Nutrient Transport	2.5431	0.0886	0.2013
SLC16A1.MCT1	Nutrient Transport	0.1985	0.8961	0.9334
SLC15A1.PEPT1	Nutrient Transport			
SLC13A1.NAS1	Nutrient Transport	2.8547	0.0661	0.1652
SLC11A2.DMT1	Nutrient Transport			
SLC30A1.ZnT1	Nutrient Transport	4.2445	0.0196	0.0818*
SLC39A4.ZIP4	Nutrient Transport	3.0217	0.0567	0.1575
NR3C1.Grα	Stress	0.0308	0.9925	0.9925

Taula 2: P-valors de l'ANOVA entre tractaments pel teixit **Ileum**. Els p-valors per sota de 0.05 queden marcats amb *** i per sota de 0.1 amb *.

El nivell de significació és α=0.05, però degut al conjunt de tests realitzats s'aplica una correcció (*Benjamini & Hochberg*). Per tant, només s'haurien de considerar significatius del conjunt experimental aquells tests en els que el p-valor de *Benjamini & Hochberg* estigui per sota de determinat llindar, en aquest cas hem fixat un nivell de significació del 5%, encara que en aquests tipus d'experiments poden arribar al 10%. Observem que hi ha gens sense cap p-valor, això es degut a que després de fer el tractament de dades faltants (descrit a l'apartat anterior) encara hi ha NA's, per tant, l'anàlisi de la variància no contempla aquests gens.

Nom del gen	Funció del gen	Estadístic F	P-valor	P-valor (Benjamini&Hochberg)
TFF3	Barrier Function	5.2048	0.0099	0.0635*
OCLN	Barrier Function	3.7353	0.0314	0.0919*
ZO1	Barrier Function	2.4161	0.1020	0.1962
MUC2	Barrier Function	5.0794	0.0108	0.0635*
MUC13	Barrier Function	2.1372	0.1332	0.2220
SI	Enzymed/Hormone	3.6748	0.0331	0.0919*
DAO1	Enzymed/Hormone	4.8678	0.0127	0.0635*
HNMT	Enzymed/Hormone	4.5122	0.0167	0.0697*
ANPEP	Enzymed/Hormone	2.3119	0.1126	0.2011
GCG	Enzymed/Hormone	6.6996	0.0035	0.0433***
IGF1R	Enzymed/Hormone			
PYY	Enzymed/Hormone			
GPX2	Enzymed/Hormone			
SOD2m	Enzymed/Hormone	1.4329	0.2680	0.3526
ALPI	Enzymed/Hormone	1.0417	0.3994	0.4755
TLR4	Inmune Response			
TGFβ1	Inmune Response	1.7069	0.2034	0.3178
CCL20	Inmune Response			
IFNGR1	Inmune Response	1.5036	0.2495	0.3526
HSPB1.HSP27	Inmune Response	0.3366	0.7991	0.8324
HSPA4.HSP70	Inmune Response	0.8326	0.4943	0.5617
FAXDC2	Inmune Response			
GBP1	Inmune Response	0.0084	0.9989	0.9989
IL8	Inmune Response	2.5729	0.0881	0.1835
SLC5A1.SGLT1	Nutrient Transport	4.1519	0.0223	0.0796*
SLC16A1.MCT1	Nutrient Transport			
SLC15A1.PEPT1	Nutrient Transport	2.9677	0.0613	0.1394
SLC13A1.NAS1	Nutrient Transport	3.5486	0.0368	0.0921*
SLC11A2.DMT1	Nutrient Transport	1.3570	0.2895	0.3618
SLC30A1.ZnT1	Nutrient Transport	0.4630	0.7118	0.7737
SLC39A4.ZIP4	Nutrient Transport	19.3772	0.0000	0.0003***
NR3C1.Grα	Stress	1.4606	0.2606	0.3526

Taula 3: P-valors de l'ANOVA entre tractaments pel teixit **Jejunum**. Els p-valors per sota de 0.05 queden marcats amb *** i per sota de 0.1 amb *.

Ileum

Un cop fet l'ANOVA i la corresponent correcció per multiplicitat de contrastos (FDR), s'han trobat 2 gens amb una expressió diferencial (p-valor < 0.05), *MUC2* i *GPX2*. Anomenarem quasi significatius aquells gens que es troben entre $0.5 <$ p-valor ≤ 0.1. Els gens quasi significatius són 3; *TFF3*, *IFNGR1* i *SLC30A1.ZnT*. Podem trobar aquests resultats a la *Taula 2*.

Jejunum

Un cop fet l'ANOVA i la corresponent correcció per multiplicitat de contrastos (FDR), s'han trobat 2 gens amb una expressió diferencial (p-valor < 0.05), *GCG* i *SLC39A4.ZIP4* Els gens quasi significatius són 8; *TFF3*, *OCLN*, *MUC2*, *SI*, *DAO1*, *HNMT*, *SLC5A1.SGLT1* i *SLC13A1.NAS1*. Podem trobar aquests resultats a la *Taula 3*.

3.2.2 Comparacions múltiples 2 a 2

Un cop identificats els gens que han manifestat una expressió diferencial hem aplicat comparacions 2 a 2 amb l'objectiu de veure quins tractaments són diferents entre si. L'anàlisi esta fet per a cada teixit i el mètode utilitzat per corregir la multiplicitat de contrastos ha sigut el mètode de Tukey.

	2-1	3-1	4-1	3-2	4-2	4-3
TFF3			0.0044			
MUC2	0.0399		0.0023			0.0243
GPX2	0.0028		0.0632	0.0006		0.0147
IFNGR1		0.0625	0.0160			
IL8	0.0203		0.0329			
SLC30A1.ZnT1		0.0984		0.0249		

Taula 4: Dels gens significatius per al teixit **Ilenum**, s'apliquen comparacions 2 a 2 Gene-Tractament. A la taula només es mostren aquells p-valors quasi significatius o significatius.

Podem destacar que gairebé tots els gens que surten signficatius presenten diferències estadísticament significatives entre el tractament 4 i el tractament 1. De la mateixa manera, observem que la comparació entre el tractament 4-2 no indica cap diferència significativa.

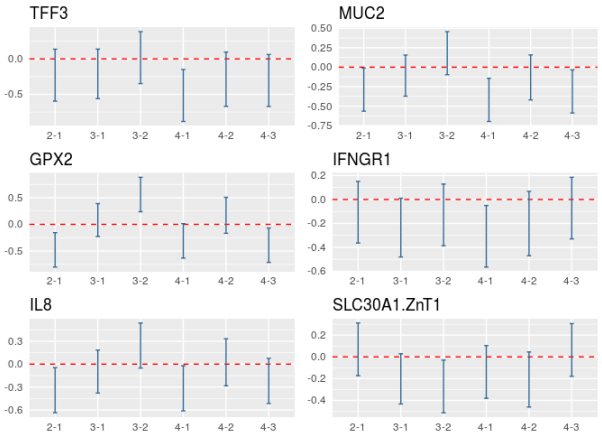


Figura 4: Intervals de confiança del 95% per a la diferència de mitjanes per a cada comparació i per a cada gen del teixit **Ilenum**. Els intervals que no contenen el 0, es mostren com comparacions estadísticament significatives.

En el teixit *Jejunum*, podem veure que cap gen presenta diferencies significatives entre el tractament 4 i 3. Les comparacions que presenten més diferencies són 3-2 i 4-2, encara que no és presenta un patró tan clar com en el teixit Ilenum.

	2-1	3-1	4-1	3-2	4-2	4-3
TFF3				0.0512	0.0074	
OCLN				0.0510	0.0713	
MUC2				0.0357	0.0098	
SI			0.0283			
DAO1		0.0422	0.0368			
HNMT		0.0211	0.0357			
GCG		0.0896	0.0142	0.0662	0.0100	
SLC5A1.SGLT1				0.0842	0.0410	
SLC13A1.NAS1			0.0286			
SLC39A4.ZIP4	0.0015			0.0001	0.0000	

Taula 5: Dels gens significatius per al teixit **Jejunum**, s'apliquen comparacions 2 a 2 Gene-Tractament. A la taula només es mostren aquells p-valors quasi significatius o significatius.

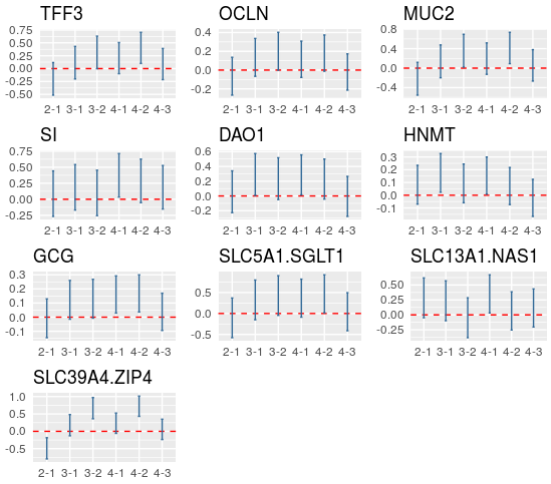


Figura 5: Intervals de confiança del 95% per a la diferència de mitjanes per a cada comparació i per a cada gen del teixit **Jejunum**. Els intervals que no contenen el 0, es mostren com comparacions estadísticament significatives.

3.2.3 Mètodes visuals

Els següents gràfics del protocol ens poden ajudar a trobar relacions i agrupacions entre els tractaments i els gens. El primer mètode analitzat ha sigut el ACP per a cada teixit:

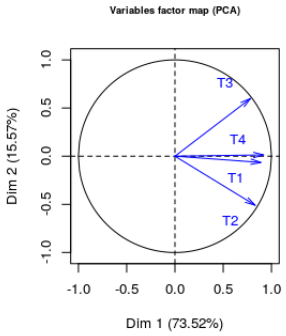


Figura 6: Representació gràfica de l'ACP de les mitjanes per cada gene-tractament del teixit **Ilenum**.

Amb aquest teixit podem observar una correlació alta entre la mitjana del tractament 4 i la mitjana del tractament 1. En canvi, trobem correlacions més febles entre el tractament 4 i els tractaments restants. L'angle entre el tractament 2 i el tractament 3, ens indica una correlació casi nul·la.

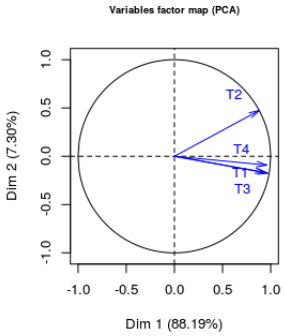


Figura 7: Representació gràfica de l'ACP de les mitjanes per cada gene-tractament del teixit **Jejunum**.

En aquest cas trobem correlacions altes entre els tractaments 1, 3 i 4. En canvi el tractament 2, sembla tenir una correlació més feble amb els altres tractaments. Recordem que per calcular el coeficient de correlació en aquest gràfic, només necesitem l'angle que formen els dos vectors, suposem que l'angle entre T4 i T2 és de 45 graus:

$$r(T4, T2) = cos(angle_{(T4,T2)}) = cos(\frac{\pi}{4}) = 0.7071$$

El següent gràfic mostra les mitjanes de cada gen per tractaments, d'aquesta manera es poden observar les diferències que observavem a l'ANOVA i a les comparacions 2 a 2.

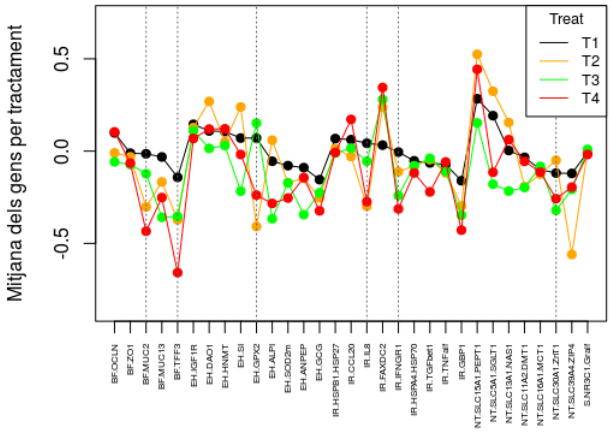


Figura 8: Mitjana dels gens per tractament, teixit **Ilenum**. El gràfic esta ordenat per família gènica primer i dins de la familia per el tractament 1 de forma decreixent. El nom de cada gen esta acompanyat de la seva funció dins de l'organisme, com per exemple, *BF.OCLN* = funció del gen: *Barrer Function*, nom del gen: *OCLN*

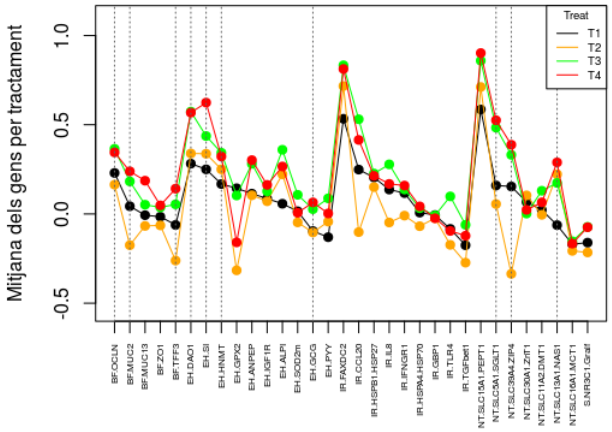


Figura 9: Mitjana dels gens per tractament, teixit **Jejunum**. El gràfic esta ordenat per família gènica primer i dins de la familia per el tractament 1 de forma decreixent.

Un fet imporant a destacar en el gràfic del texit **Jejunum**, és el gran número de gens significatius que tenen una funció barrera o protectora (BF) dins del organisme.

Per últim, es presenten els *heatmap* per a cada teixit:

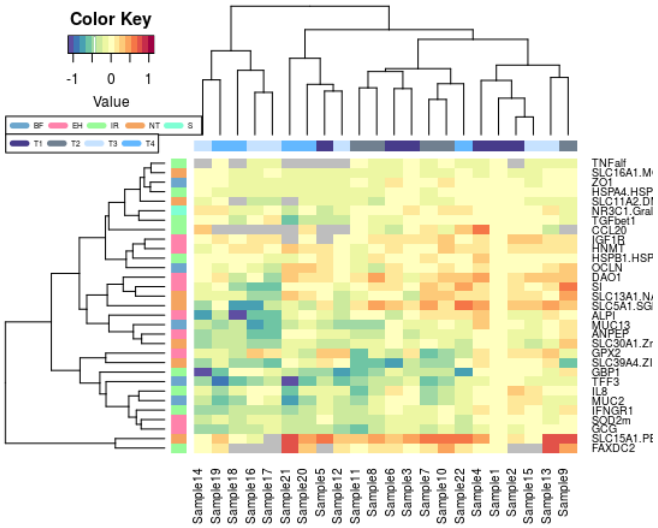


Figura 10: Heatmap del teixit **Ilenum**.

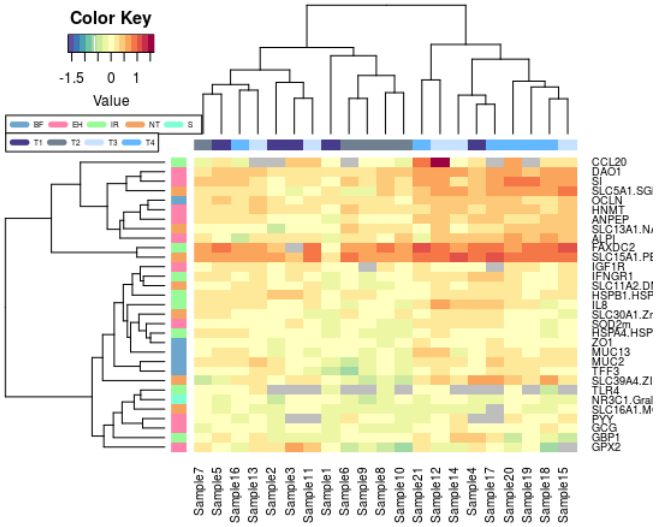


Figura 11: Heatmap del teixit **Jejunum**.

Les conclusions d'aquest apartat queden incloses en les conclusions finals del cas d'estudi.

3.3 Conclusions (cas d'estudi)

Un cop realitzat l'anàlisi per mitjà del protocol que hem desenvolupat, es l'hora de les conclusions. Recordem que és un estudi eminentment exploratori (en el sentit de que és un prototip o protocol d'actuació, sense gaires precedents en l'àmbit de la producció animal). Per tant, l'equip estadístic ha tingut un paper clau en definir les hipotesis i criteris estadístics que creiem més importants. Aquests criteris pretenen ajudar a que els investigadors interpretin resultats presents o futurs.

Els primers apartats del protocol consisteixen en l'anàlisi separada per a cada gen fixat, però amb correcció FDR, atès que les anàlisis són simultànies. La primera hipòtesi de recerca que ens plantegem és:

En algun gen (i quin o quins) hi ha diferències significatives en els nivells d'expressió entre tractaments?

Per al teixit *Ilenum*, d'un total de 31 gens, s'han trobat 2 gens amb una expressió diferencial significativa (p-valor(Benjamini & Hochberg) ≤ 0.05), *MUC2* i *GPX2*. Anomenarem quasi significatius aquells gens que es troben entre 0.5 p-valor ≤ 0.1. Els gens quasi significatius són 3: *TFF3*, *IFNGR1* i *SLC30A1.ZnT*. Podem trobar aquests resultats a la *Taula 2*.

Per al teixit *Jejunum*, d'un total de 32 gens, s'han trobat 2 gens amb una expressió diferencial (p-valor(Benjamini & Hochberg) ≤ 0.05), *SLC39A4.ZIP4* i *GCG*. Els gens quasi significatius són 8: *TFF3*, *OCLN*, *MUC2*, *SI*, *DAO1*, *SLC13A1.NAS1*, *SLC5A1.SGLT1* i *HNMT*. Podem trobar aquests resultats a la *Taula 3*.

Fixat cada gen, i mirant els gens significatius, la segona pregunta que ens plantegem és:

Entre quins tractaments (parelles) hi ha diferències significatives?

Com hem vist en l'apartat anterior, per al teixit *Ilenum*, es podría destacar que gairebé en tots els gens s'han trobat diferències significatives entre el tractament 4 i el tractament 1. A la *Taula 4* es poden trobar totes les comparacions significatives, i es podría fer un anàlisi més detallat per a cada gen (segons l'interès de l'investigador). En el cas del teixit *Jejunum*, podem veure que cap gen presenta diferencies significatives entre el tractament 4 i 3. Els resultats els podem trobar a la *Taula 5*, i segons l'interès de l'investigador es podría veure de manera més detallada quines són les comparacions significatives entre tractaments per a cada gen.

La següent part de l'anàlisi consistia en fer una exploració gràfica del conjunt total de gens. L'exploració gràfica esta dividida en tres tipus de gràfics:

I) *Amb el heatmap, explorem el doble agrupament (clustering) de gens i mostres.*

Al teixit *Ilenum* (*Figura 10*), el dendrograma vertical no mostra cap agrupació clara per funcionalitat del gen, encara que sí s'observen dos gens, *FAXDC2* i *SLC15A1.PEPT1*, que tenen una expressió alta i estan agrupats en un sol clúster (probablement siguin *outliers*). El clúster horitzontal, mostra un conjunt de casos amb una expressió baixa sota els tractaments 3 i 4, els quals en aquesta zona s'agrupen en un clúster. A l'altre part del dendograma hi ha una separació poc nítida entre tractaments. Per al teixit *Jejunum* (*Figura 11*), al dendrograma vertical no es veu cap agrupació clara per funcionalitat (els colors del dendrograma estan força barrejats). S'hi veu (al mig) un clúster amb dos gens, *FAXDC2* i *SLC15A1.PEPT1*, amb uns nivells d'expressió alts en la majoria de mostres. Aquests gens tenen un comportament outlier comparat amb els altres i convindria analitzar-lo amb més detall: revisar dades, fer noves proves per veure si es mantenen els nivells d'expressió alts, etc. El dendrograma horitzontal, corresponent a l'agrupació de mostres, observem que les mostres del tractament 2 i el tractament 4 queden bastant agrupades. Els tractaments 3 i 4 tenen nivells d'expressió més elevats sobretot en els gens que conformen el clúster superior.

II) *Amb la representació gràfica de l'ACP, explorem les correlacions entre mitjanes de tractaments al llarg dels gens.*

En el cas del teixit *Ilenum* destaquem una correlació alta i positiva entre la mitjana del tractament 4 i la mitjana del tractament 1. Per al cas del teixit *Jejunum*, trobem correlacions altes i positives entre els tractaments 1, 3 i 4. Trobem aquests gràfics a les *Figures 6* i *7*, respectivament.

III) *Amb les gràfiques de línies per a les mitjanes dels tractaments, podem comparar visualment l'evolució de les mitjanes al llarg dels gens.*

Aquests gràfics, que trobem a les *Figures 8 i 9*, realment són una representació gràfica suplementària al test ANOVA, perquè es marquen els gens que presenten diferències significatives i es pot apreciar entre quins tractaments hi ha més diferència i el signe d'aquesta. El fet més remarcable el trobem al teixit *Jejunum*, on els gens que han sortit significatius principalment són aquells que desenvolupen una funció barrera o protectora, *Barrer Function (BT)*. Les gràfiques de línia també són útils per explorar la correlació entre tractaments, en coherència amb l'ACP.

Amb les conclusions anteriors els investigadors poden veure els efectes dels tractaments en l'expressió dels gens i complementar aquesta anàlisi amb el seus criteris com especialistes en l'àmbit biocientífic. L'anàlisi del cas d'estudi consisteix doncs en la utilització de l'eina estadística que proporciona els resultats (taules i gràfics) esmentats, i ofereix un esquema bàsic de les possibles interpretacions (indicacions).

4 TL3P: Aplicatiu Web amb el paquet `Shiny` de R

Implementar una eina capaç de realitzar el protocol d'anàlisi d'una manera interactiva i generalitzada ha sigut l'objectiu principal del treball. En els següents apartats s'introdueix el software i la metodologia utilitzada per crear l'eina.

4.1 Introducció a Shiny

Amb la idea de facilitar futures investigacions als grups de recerca de la UAB, va sorgir la idea de crear una aplicació mitjançant `Shiny`.

`Shiny` és un paquet d'`R` que facilita la creació d'aplicacions web interactives (apps) directament des de `R`.

Principalment el codi està dividit en 3 scripts, anomenats:

- `ui.R` (*User interface object*)
- `server.R` (*Server function*)
- `App.R` (*Call to the* ***`shinyApp`*** *function*)

L'objecte *user interface (ui)* controla l'aparença de l'aplicació i interactua amb llenguatges com *HTML, CSS, Javascript*. La funció *server* conté els càlculs interns de l'aplicació. Finalment, la funció `shinyApp` fa una crida a l'*ui* i el *server* per obrir l'aplicació.

Això seria una petita introducció al paquet `Shiny`, la mateixa pàgina de `Rstudio` té tutorials que aprofundeixen més amb totes les funcionalitats del paquet. (`https://shiny.rstudio.com/tutorial/`)

4.1.1 Gestió i control de versions

Abans de parlar sobre l'aplicació que hem desenvolupat, caldria destacar quina ha sigut la metodologia a l'hora de guardar el codi. Habitualment treballem guardant els scripts sense cap tipus de versió i sempre sobreescrivint el treball realitzat. Aquesta manera de treballar és molt poc eficient i comporta molts problemes si es treballa en grup. Per tant, per desenvolupar l'aplicació d'una manera més eficaç i segura, hem treballat amb *GitHub*.

GitHub és una plataforma de desenvolupament col·laboratiu de programari per allotjar projectes utilitzant el sistema de control de versions *Git*.

Git és un programari de control de versions. El control de versions, resumint molt, és la gestió dels diversos canvis que es realitzen sobre un repositori (un repositori és el nom que rep el lloc on s'allotja el codi d'un projecte de desenvolupament en algun llenguatge de programació).

El repositori on es troba tot el codi de l'aplicació el tenim al següent enllaç: `https://github.com/djangosee/TFGShinyApp`. A l'enllaç hi trobem les instruccions necessàries per descarregar el repositori i obrir l'aplicació. El codi s'ha emmagatzemat de forma pública i qualsevol usuari de la plataforma pot obtenir-lo i treballar desenvolupant noves funcionalitats.

D'aquesta forma tot el codi queda centralitzat, es pot accedir de manera ràpida i fer modificacions sense afectar el funcionament de l'aplicatiu. En els següents apartats es parlarà de l'aparença de l'aplicació, les seves funcionalitats i quins canvis es podrien aplicar en futures versions.

4.1.2 Estructura de l'aplicatiu

Un cop accedim a l'aplicació, trobem la següent interfície:

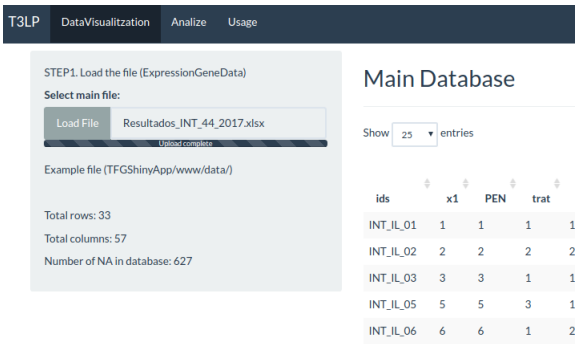


Figura 12: Imatge de l'aparença de l'aplicació i la seva estructura de pestanyes.

La idea principal a l'hora de desenvolupar l'aplicatiu ha sigut la de crear un entorn atractiu i fàcil per a l'usuari. L'aplicació està formada de 3 pestanyes:

1. *DataVisualitzation*: Pestanya on carreguem les dades i obtenim unes taules per poder fer consultes i visualitzar les dades.
2. *Analize*: Aquesta pestanya conté la configuració dels paràmetres i, un cop executat, els resultats de l'anàlisi i algunes ajudes a la interpretació.
3. *Usage*: Per últim, trobem les especificacions necessàries per obrir la base de dades i un petit tutorial de l'aplicatiu.

Observació Si existeixen problemes a l'hora d'obrir l'aplicació, a l'annex C podem trobar l'output de la funció `sessionInfo` on ens mostra la versió d'R utilitzada amb els corresponents paquets.

4.2 Funcionalitats de l'aplicació

En aquest apartat es presenten les funcionalitats que té l'aplicació en forma d'imatges i comentaris.

Observació. Les dades utilitzades en les següents imatges no corresponen a les dades utilitzades en el cas d'estudi.

4.2.1 Panell de configuració dels paràmetres

També anomenat *SidebarPanel*, en aquest panell trobem els paràmetres i variables relacionats amb l'estudi:

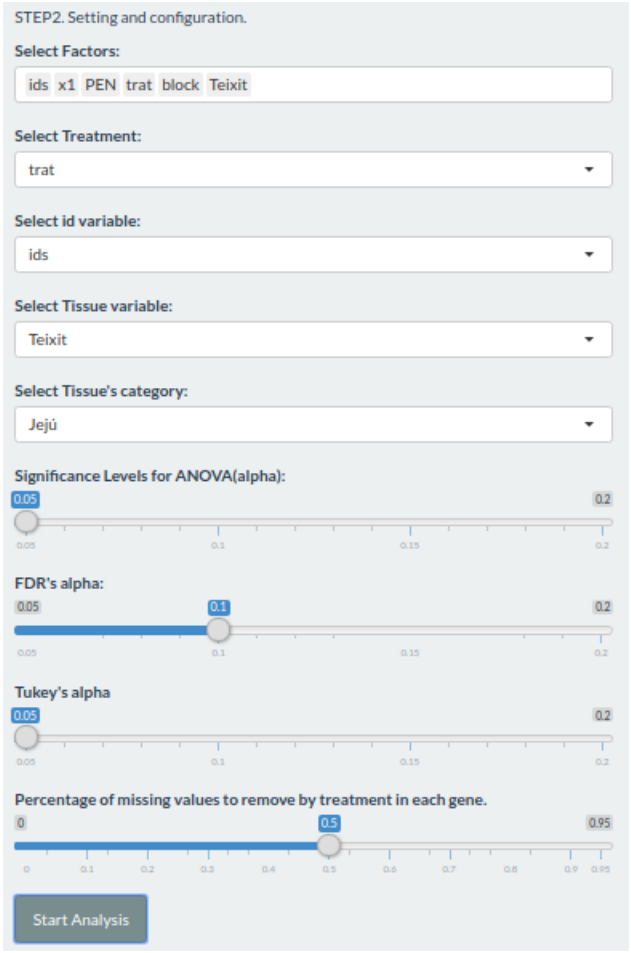


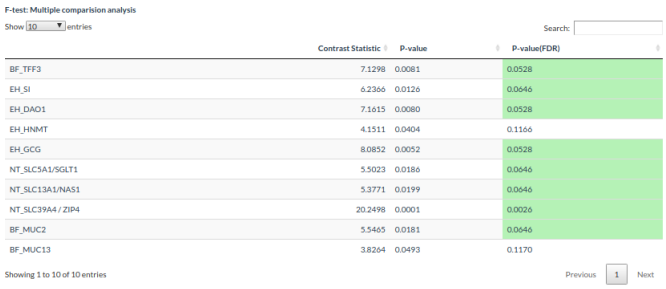
Figura 13: Panell de configuració dels paràmetres

La configuració abans de començar l'anàlisi està formada per:

1. Selecció dels factors de la base de dades.
2. Selecció de la variable tractament.
3. Selecció de la variable id.
4. Selecció de la variable teixit.
5. Selecció de la categoria de la variable teixit.
6. α per a l'ANOVA.
7. α BenjaminiHochberg(FDR).
8. α per a les comparacions 2 a 2 (Tukey).
9. Percentatge de dades faltants a eliminar per tractament per a cada gen (per defecte, el 50%).

4.2.2 Taules

A l'aplicació podem trobar 2 taules de resultats, una corresponent a l'anàlisi de la variància i l'altre corresponent a les comparacions 2 a 2 dels tractaments.


F-test: Multiple comparision analysis

	Contrast Statistic	P-value	P-value(FDR)
BF_TFF3	7.1298	0.0081	0.0528
EH_SI	6.2366	0.0126	0.0646
EH_DAO1	7.1615	0.0080	0.0528
EH_HNMT	4.1511	0.0404	0.1166
EH_GCG	8.0852	0.0052	0.0528
NT_SLC5A1/SGLT1	5.5023	0.0186	0.0646
NT_SLC13A1/NAS1	5.3771	0.0199	0.0646
NT_SLC39A4 / ZIP4	20.2498	0.0001	0.0026
BF_MUC2	5.5465	0.0181	0.0646
BF_MUC13	3.8264	0.0493	0.1170

Showing 1 to 10 of 10 entries

Figura 14: Taula de resultats de l'ANOVA


Tukey: Post-hoc

	2-1	3-1	3-2
BF_TFF3	0.2120	0.1771	0.0062
EH_SI	0.7416	0.0139	0.0583
EH_DAO1	0.7821	0.0097	0.0358
EH_HNMT	0.3351	0.0322	0.3977
EH_GCG	0.9868	0.0136	0.0100
NT_SLC5A1/SGLT1	0.7892	0.0733	0.0208
NT_SLC15A1/PEPT1	0.5675	0.0433	0.2660
NT_SLC13A1/NAS1	0.0716	0.0198	0.8254
NT_SLC39A4 / ZIP4	0.0035	0.1433	0.0001
EH_GPX2	0.0479	0.1972	0.6208

Showing 1 to 10 of 11 entries

Figura 15: Taula de les comparacions 2 a 2 amb Tukey

Les cel·les en verd mostren els p-valors que han sigut significatius. El nivell de significació s'estableix amb anterioritat per l'usuari al panell de configuració.

4.2.3 Gràfics

LinePlot

El primer gràfic que trobem correspon a les mitjanes per tractament i per gen. Aquest gràfic disposa de diferents opcions de personalització, com canviar l'ordre de les dades o fins i tot el color de les línies.

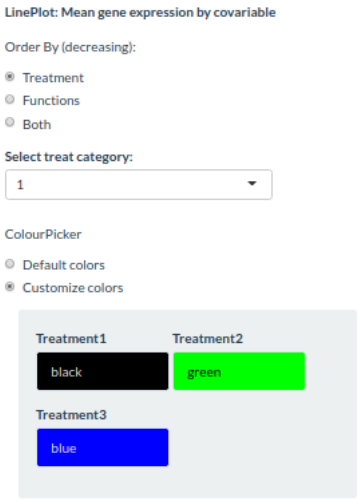


Figura 16: Panell de configuració del *LinePlot*

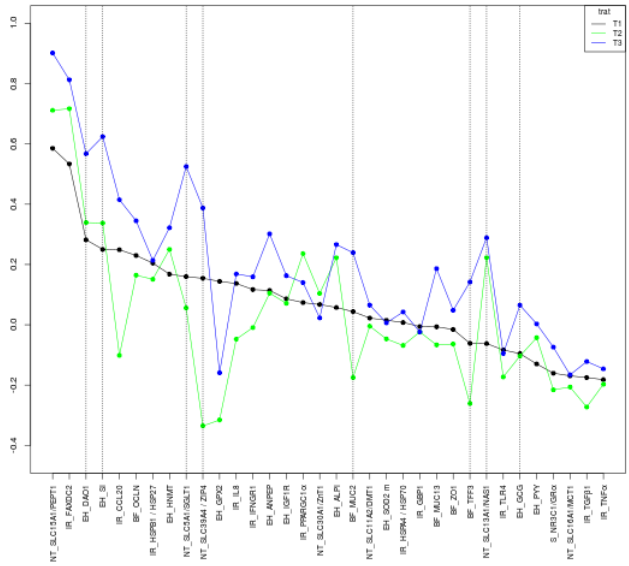


Figura 17: *LinePlot.* Exemple on ordenem els gens (eix x) de forma decreixent pel valor del tractament 1 i escollim els colors per a cada tractament.

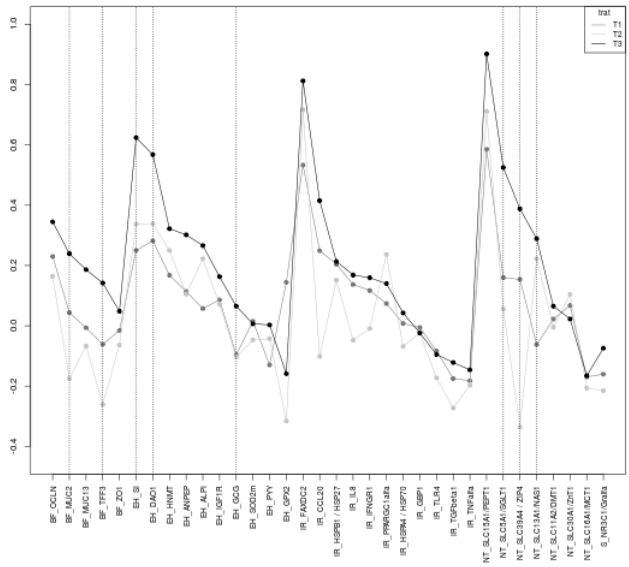


Figura 18: *LinePlot.* Exemple on ordenem els gens (eix x) de forma decreixent pel valor del tractament 3 i agrupat per la funcionalitat del gen. A més escollim una escala de grisos.

Representació gràfica de l'ACP

Els gràfics de variabes i de casos s'interpreten conjuntament. La posició de les variables en el gràfic mostra l'estructura de les dades i suggereix possibles interpretacions de les components principals. Són gràfiques exploratòries, no porten significació estadística associada.

Gràfic de variables

Per a dades escalades o tipificades (cas habitual), les variables queden representades per fletxes de longitud entre 0 i 1.

Longitud i qualitat de representació. La longitud de la fletxa indica la mida de la seva projecció en el subespai de les components (preferentment, de les dues primeres) i, per tant, la seva qualitat de representació: Una variable està ben representada en el subespai de les dues components, si la fletxa té longitud propera a 1, és a dir, s'aproxima a la circumferència de radi 1. (Només es poden interpretar les variables ben representades)

Angles i correlació. El cosinus de l'angle format pels dos vectors és proporcional a la seva correlació, per tant: *Els angles que formen dues fletxes (o variables) entre sí, ens mostren la correlació existent entre les corresponents variables.* Els sentits dels vectors indiquen el signe de la correlació. L'angle que forma una fletxa (variable) i un eix (component) indica la correlació entre la variable i la component.

Per tant,

- Si dues variables estan ben representades, com més petit (proper a zero) sigui l'angle entre les seves fletxes, més forta i positiva és la seva correlació.
- Si dues variables estan ben representades, com més proper a 180^o (π) sigui l'angle entre les seves fletxes, més forta i negativa és la seva correlació.
- Si dues variables estan ben representades i l'angle entre les dues fletxes és proper al recte, la correlació és pràcticament zero.
- Si hi ha un grup de variables ben representades i en posicions properes en el gràfic, això indica que totes elles estan molt relacionades entre sí i que ens subministren bàsicament la mateixa informació, diem que conformen *un factor latent.*
- Si apareixen varis grups de variables correlacionades, vol dir que hi ha *diversos factors latents* a la matriu de dades.
- Si una o mès variables estan ben representades i tenen un angle molt petit amb un dels eixos de coordenades, això indica que una de les components

està fortament correlacionada amb aquesta(es) variable(s) i, per tant, es pot interpretar (etiquetar) la component en base al significat d'aquesta(es) variable(s).

Gràfic de casos individuals

Els gens (en general, els casos) es representen mitjançant un diagrama de dispersió dels scores o puntuacions en l'espai de les components. Els casos amb puntuacions més elevades en una component es poden interpretar en base amb les variables més correlacionades amb aquesta component. Els colors de la gràfica de dispersió se solen utilitzar per visualitzar una variable qualitativa.

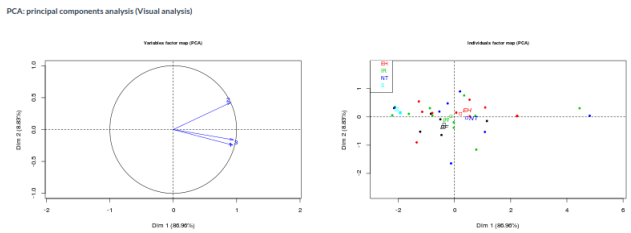


Figura 19: Representació gràfica de l'ACP

Heatmap

El *heatmap* és la respresentació conjunta de dos *dendrogrames*: un dendrograma vertical que mostra agrupacions de files (gens o variables) i un dendrograma horitzontal que mostra agrupacions de columnes (mostres o casos). Un dendrograma és un arbre que representa una jerarquia de particions (des de la màxima on cada element és un clúster fins a la mínima on tots els elements estan agrupats en un clúster únic). Per decidir quina partició s'escull, un criteri és tallar l'arbre per on les *branques són més llargues* (perquè correspon a maximitzar les distàncies entre els clústers resultants).

- Al dendrograma vertical hi ha els clústers dels gens. Després d'escollir la partició, els gens agrupats en un mateix clúster tendeixen a tenir pautes d'expressió similars en el conjunt de les mostres.
- La variable categòrica *funció del gen* es visualitza segons els colors que veiem sota del dendrograma vertical, i és útil per explorar si els gens agrupats en un mateix clúster (expressions similars) tendeixen a tenir la mateixa funcionlitat o, per contra, l'agrupació per clústers no es correspon a l'agrupacio per funcionalitats.
- Al dendrograma horitzontal hi ha els clústers de les mostres. Després d'escollir la partició, les mostres agrupades en un mateix clúster tendeixen a tenir pautes d'expressió similars en el conjunt dels gens.
- La variable *tractament* es visualitza segons els colors que veiem sota del dendrograma horitzontal, i és útil per explorar si les mostres d'un mateix clúster (expressions similars) tendeixen a correspondres a un mateix tractment o, per contra, l'agrupació per clústers no es correspon a l'agrupacio per tractaments.
- En blanc trobem els casos que són NA's.

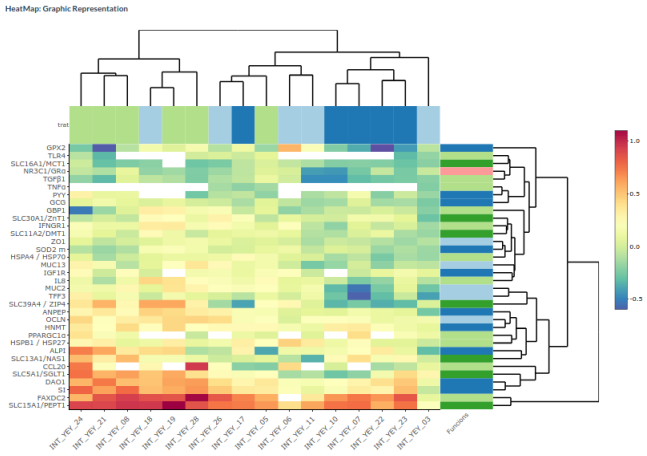


Figura 20: *Heatmap* interactiu

4.3 Futures versions de l'aplicació

A dia d'avui l'aplicació està a la versió 1.0.0. En aquest apartat es plantegen noves funcionalitats per a futures versions:

1. Utilitzar un servidor per tenir l'aplicatiu disponible en forma de pàgina web. D'aquesta manera es desvincula la necessitat que cada usuari compili l'aplicatiu des de el seu dispositiu. És una manera de facilitar als investigadors la feina, ja que no seria necessari tenir cap coneixement d'`R` per obrir l'aplicació.
2. La possibilitat de descarregar un informe, fet en `LateX`, dels resultats obtinguts.
3. Implementar métodes més acurats per a dissenys desbalancejats.
4. Implementar les gràfiques de barres d'error, per visualitzar les comparacions 2 a 2 (Tukey).
5. Fer encara més interactiva l'aplicació, amb nous gràfics de línies, si bé això podria anar en detriment del temps d'execució de l'anàlisi.

Aquestes idees i altres que es puguin tenir, es poden implementar en un futur. El codi de l'aplicació és públic, i queda obert per a usuaris que vulguin col·laborar amb el projecte. L'únic necessari és registrar-se en la plataforma de *GitHub* i sol·licitar una col·laboració a l'administrador del projecte.

5 Conclusions

En aquest treball s'ha descrit un protocol d'anàlisi capaç d'analitzar i explorar diversos aspectes d'una base de dades d'expressió gènica d'un seguit de gens, on les mostres estan classificades segons un factor, habitualment un tractament, una dieta o una altra situació experimental...A més, analitzar les dades d'expressió gènica provinents de la tecnologia *OpenArray* d'una manera senzilla i còmode per a l'usuari.

La possibilitat d'utilitzar una eina específica per aquest tipus d'estudi, però que s'ha fet prou general i automàtica, obre un ventall de possibilitats als investigadors. La replicació de l'anàlisi amb futures dades és una realitat amb l'aplicatiu. A més la seva estructura interna facilita l'eventual integració de nous mètodes estadístics dins del codi. D'aquesta manera l'eina podria anar evolucionant al llarg del temps amb la finalitat de fer una anàlisi completa i apta per a altres dissenys.

El treball de fi de grau m'ha permès aprofundir en temes bàsics de la inferència estadística (ANOVA, Tukey) i de l'anàlisi multivariant (PCA, anàlisi de clústers), entre altres. Al mateix temps, he estudiat conceptes o recursos que no havia tractat durant el grau (FDR, heatmap), dels quals he fet recerca bibliogràfica. També he pogut practicar i ampliar elements de programació (`R`,`Shiny`, gestió de versions, etc.,), alguns dels quals he après de manera autodidacta. Finalment, voldria destacar el caràcter multidisciplinar d'aquest treball, entre estadística, biomedicina i bioinformàtica, que suposa un nexe d'unió entre algunes assignatures que he cursat durant el grau.

6 Agraïments

Per portar endavant aquest projecte he necessitat l'ajuda d'algunes persones a les quals volia agrair la seva col·laboració i esforç.

Agrair la col·laboració de l'estudiant de veterinària Francesc Gonzàlez i el Dr. David Solà.

També voldria donar les gràcies a la tutora del treball, la Dr. Mercè Farré, per guiar-me durant el projecte i posar la seva confiança en mi.

Bibliografia

[1] Guillermo Ayala.(2018).Bioinformática Estadística: Análisis estadístico de datos ómicos. 4
https://www.uv.es/ayala/docencia/tami/tami13.pdf

[2] Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (Institute of Mathematical Statistics Monographs). Cambridge: Cambridge University Press. pp. 271-294, doi:10.1017/CBO9781316576533. 6

[3] Everitt, B., Hothorn, T. : An introduction to Applied Multivariate Analysis with R. Springer. 2011. 8

[4] Härdle, W., Simar, L.: Applied Multivariate Statistical Analysis. Springer. 2007. 8

[5] Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1. (1995), pp. 289-300, doi:10.2307/2346101 by Yoav Benjamini, Yosef Hochberg. 6

[6] Ramón Tamarit Agusti. Análisis de cluster no supervisados. Aplicaciones en la búsqueda y visualización de perfiles de expresión en datos de microarrays. 8
http://mural.uv.es/rata3/PECSspace.html

[7] Dong Hyun Jeong, Caroline Ziemkiewicz, William Ribarsky and Remco Chang: Understanding Principal Component Analysis Using a Visual Analytics Tool. 10
http://www.knowledgeviz.com/pdf/UKC2009.pdf

[8] Universidad de Granada: Métodos de análisis multivariante. Análisis clúster. 8
http://wpd.ugr.es/~bioestad/guia-spss/

[9] Universidad de Granada: Métodos Jerárquicos de Análisis Cluster. 8
http://www.ugr.es/~gallardo/pdf/cluster-3.pdf

[10] *NCBI, National Center of Biotechnology Information.*
https://www.ncbi.nlm.nih.gov/. USA

[11] *PubMed. US National Library of Medicine.*
https://www.ncbi.nlm.nih.gov/pubmed/.
USA

A Mètode de Ward: Exemple del mètode amb gens

Observem com funciona aquest mètode en el cas de tenir 3 gens on es mesura l'expressió gènica per a 3 mostres. Les dades són les següents:

	X_1	X_2	X_3
Gen_1	1.02	0.21	6.29
Gen_2	10.06	8.19	7.29
Gen_3	10.11	14.63	7.62

Taula 6: Expressió gènica de cada gen Gen_i per a cada mostra X_j.

Recordem que per utilitzar aquest mètode necessitem la matriu de distàncies euclidianes. La distància euclidiana entre dos punts $P = (p_1, p_2, \dots p_n)$ i $Q = (q_1, q_2, \dots q_n)$ es defineix com:

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Per tant, obtenim aquesta matriu de distàncies:

	Gen_1	Gen_2	Gen_3
Gen_1	0	12.10	17.10
Gen_2	12.10	0	6.45
Gen_3	17.10	6.45	0

Taula 7: Matriu de distàncies. La distància utilitzada és l'euclidiana.

En aquest cas, com tenim només 3 gens, tenim un total de 3 combinacions possibles de 2 elements. Per tant, calcularem $\triangle E$ per a cada combinació i escollirem el més petit com el millor clúster.

Particions	*Centroides*	E_k	E	$\triangle E$
$(Gen_1, Gen_2), Gen_3$	$C_{Gen_1,Gen_2} = (5.54, 4.19, 6.78)$	$E_{Gen_1,Gen_2} = 72.92$ $E_{Gen_3} = 0$	72.92	72.92
$(Gen_1, Gen_3), Gen_2$	$C_{Gen_1,Gen_2} = (7.06, 7.67, 7.06)$	$E_{Gen_1,Gen_3} = 112.42$ $E_{Gen_2} = 0$	112.42	112.42
$(Gen_2, Gen_3), Gen_1$	$C_{Gen_1,Gen_2} = (10.08, 11.41, 7.45)$	$E_{Gen_2,Gen_3} = 41.5850$ $E_{Gen_1} = 0$	41.5850	41.58

Taula 8: Taula resum dels càlculs proposats per obtenir l'increment de la suma de quadrats residuals. Les particions són possibles combinacions de 2 gens en un total de 3.

Podem observar que segons el criteri de Ward, escolliríem unificar el Gen_2 i el Gen_3 en un mateix clúster. Si fem el mateix però automàticament amb la funció `hclust`, obtenim el següent dendrograma:

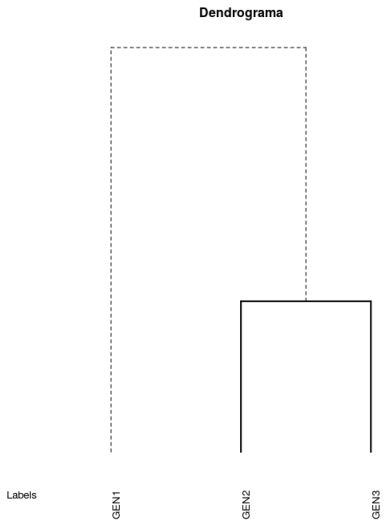


Figura 21: Dendrograma obtingut amb la funció `hclust`. Si apliquem directament la funció `hclust`, que utilitza el mètode de Ward, obtenim 2 clústers. El primer clúster amb el Gen_1 i un segon clúster amb el Gen_2 i el Gen_3. Trobem els mateixos clústers amb el procediment manual que hem calculat anteriorment. L'altura a la qual s'uneixen els clústers és $\triangle E$, en aquest cas, el Gen_2 i el Gen_3 s'uneixen a una altura de 41.58.

B Gens diana utilitzats a l'estudi

Barrier Function	OCLN	Occludin
	ZO1	Zonula occludens 1
	CLDN1	Claudin-1
	CLDN4	Claudin-4
	CLDN15	Claudin-15
	MUC2	Mucin 2
	MUC13	Mucin 13
	TFF3	Trefoil factor 3
Enzyme/Hormone	GPX2	Glutathione peroxidase 2
	SOD2 m	Superoxide dismutase
	ALPI	Intestinal alkaline phosphatase
	SI	Sucrase-isomaltase
	DAO1	Diamine oxidase
	HNMT	Histamine N-methyltransferase
	ANPEP	Aminopeptidase-N
	IDO1	Indoleamine 2,3-dioxygenase
	GCG	Glucagon
	CCK	Cholecystokinin
	IGF1R	Insulin-like growth factor 1 receptor
	PYY	Peptide tyrosine tyrosine
Nutrient transport	SLC5A1/SGLT1	Solute carrier family 5 (sodium/glucose cotransporter) member 1
	SLC16A1/MCT1	Monocarboxylate transporter 1
	SLC7A8	Solute carrier family 7 (amino acid transporter light chain, L System) member 8
	SLC15A1/PEPT1	Solute carrier family 15 (oligopeptide transporter) member 1
	SLC13A1/NAS1	Solute carrier family 13 (sodium/sulfate symporters) member 1
	SLC11A2/DMT1	Solute carrier family 11 (proton-coupled divalent metal ion transporter) member 2
	MT1A	Metallothionein 1A
	SLC30A1/ZnT1	Solute carrier family 30 (zinc transporter) member 1
	SLC39A4 / ZIP4	Solute carrier family 39 (zinc transporter) member 4

Immune response	TLR2	Toll-like receptor 2
	TLR4	Toll-like receptor 4
	IL1β	Interleukin 1β
	IL6	Interleukin 6
	IL8	Interleukin 8
	IL10	Interleukin 10
	IL17A	Interleukin 17
	IL22	Interleukin 22
	IFNγ	Interferon γ
	TNFα	Tumor necrosis factor α
	TGFβ1	Transforming growth factor β1
	CCL20	Chemokine (C-C motif) ligand 20
	CXCL2	Chemokine (C-X-C motif) ligand 2
	IFNGR1	Interferonγreceptor 1
	HSPB1 / HSP27	Heat shock protein 27
	HSPA4 / HSP70	Heat shock protein 70
	REG3G	Regenerating-islet derived protein 3γ
	PPARGC1α	Peroxisome proliferative activated receptor γ, coactivator 1α
	FAXDC2	Fatty acid hydrolase domain containing 2
	GBP1	Guanylate binding protein 1
Stress	CRHR1	Corticotropin releasing hormone receptor 1
	NR3C1/GRα	Glucocorticoid receptor
	HSD11B1	Hydroxysteroid (11-β) dehydrogenase 1
Housekeeping	GAPDH	Glyceraldehyde-phosphate-dehydrogenase
	ACTB	Actin, β
	TBP	TATA-box binding protein
	B2M	β-2-microglobulin

Figura 22: Gens *diana* i la seva funció dins de l'organisme.

C Output sessionInfo()

Listing 1: Output sessionInfo()

```
R version 3.4.3 (2017-11-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.3 LTS

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.6.0
LAPACK: /usr/lib/lapack/liblapack.so.3.6.0

locale:
 [1] LC_CTYPE=es_ES.UTF-8       LC_NUMERIC=C               LC_TIME=es_ES.UTF-8
LC_COLLATE=es_ES.UTF-8     LC_MONETARY=es_ES.UTF-8
 [6] LC_MESSAGES=es_ES.UTF-8    LC_PAPER=es_ES.UTF-8       LC_NAME=C
LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] tools     parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] stringi_1.2.2            shinymaterial_0.5.2.9000 shiny_1.0.5              readxl_1.1.0
DT_0.4
 [6] shinythemes_1.1.1        genefilter_1.60.0        RCurl_1.95-4.10          bitops_1.0-6
Biobase_2.38.0
[11] BiocGenerics_0.24.0      gplots_3.0.1             heatmaply_0.14.1         viridis_0.5.1
viridisLite_0.3.0
[16] plotly_4.7.1             ggplot2_2.2.1            shinyjs_1.0              shinycssloaders_0.2.0
RColorBrewer_1.1-2
[21] colourpicker_1.0         zoo_1.8-1                knitr_1.20               FactoMineR_1.41

loaded via a namespace (and not attached):
 [1] bit64_0.9-7          webshot_0.5.0        httr_1.3.1           prabclus_2.2-6
R6_2.2.2             KernSmooth_2.23-15
 [7] DBI_1.0.0            lazyeval_0.2.1       colorspace_1.3-2     trimcluster_0.1-2
nnet_7.3-12          gridExtra_2.3
[13] bit_1.1-12           compiler_3.4.3       TSP_1.1-6            flashClust_1.01-2
diptest_0.75-7       caTools_1.17.1
[19] scales_0.5.0         DEoptimR_1.0-8       mvtnorm_1.0-7        robustbase_0.93-0
digest_0.6.15        pkgconfig_2.0.1
[25] htmltools_0.3.6      htmlwidgets_1.2      rlang_0.2.0          RSQLite_2.1.1
bindr_0.1.1          jsonlite_1.5
[31] crosstalk_1.0.0      mclust_5.4           gtools_3.5.0         dendextend_1.8.0
dplyr_0.7.4          magrittr_1.5
[37] modeltools_0.2-21    leaps_3.0            Matrix_1.2-14        Rcpp_0.12.16
munsell_0.4.3        S4Vectors_0.16.0
[43] scatterplot3d_0.3-41 whisker_0.3-2        yaml_2.1.19          MASS_7.3-50
flexmix_2.3-14       plyr_1.8.4
[49] grid_3.4.3           blob_1.1.1           gdata_2.18.0         promises_1.0.1
miniUI_0.1.1.1       lattice_0.20-35
[55] splines_3.4.3        annotate_1.56.2      pillar_1.2.2         markdown_0.8
fpc_2.1-11           codetools_0.2-15
[61] stats4_3.4.3         XML_3.98-1.11        glue_1.2.0           gclus_1.3.1
data.table_1.11.2    httpuv_1.4.3
[67] foreach_1.4.4        cellranger_1.1.0     gtable_0.2.0         purrr_0.2.4
tidyr_0.8.0          kernlab_0.9-26
[73] assertthat_0.2.0     mime_0.5             xtable_1.8-2         later_0.7.2
survival_2.42-3      class_7.3-14
[79] seriation_1.2-3      tibble_1.4.2         iterators_1.0.9      IRanges_2.12.0
memoise_1.1.0        AnnotationDbi_1.40.0
[85] registry_0.5         bindrcpp_0.2.2       cluster_2.0.7-1  }
```