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

11 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 i genes, i 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, presentació y distribució d'informació relativa a les dades biològiques o mèdiques.

Aquest treball s'ha centrat en l'anàlisis de bases de dades d'expressió genètica a partir de 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 significativament sota condicions experimentals establertes.

Amb aquesta premissa, el treball també s'ha enfocat en crear un aplicatiu web capaç de fer un anàlisi estadístic 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. Tot el conjunt de l'aplicació ha sigut organitzada i compartida 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 empleats.

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, 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 els nivells de producte 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 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.

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, a més del biaix estadístic que es pot generar.

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 **Nombre del grupo**.

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 al creixement dels porcins.

Les dades utilitzades en aquest treball han sigut proporcionades per **Nombre del grupo** 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 del 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, la idea ha sigut generalitzar el codi per poder utilitzar-ho amb altres dissenys experimentals.

[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.

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 i definim $n = \frac{N}{k}$. 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. 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 la mitjana 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, és a dir:

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

Si això és cert, el terme d'error per a la diferència de grups queda definit com:

$$\bar{x_i} - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n} = \bar{\sigma}^2)$$

Es pot mesurar la quantitat total de variabilitat entre observacions sumant els quadrats de les diferències entre cadascun $\bar{x}.$ i x_{ij}:

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

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. Si no existeixen diferencies, entre els grups, esperaríem que les mitjanes quadràtiques

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

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

siguin similars. El test estadístic 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. 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 minimitzarà la potència estadística en comparació amb un disseny balancejat. El poder estadístic no és més que la probabilitat que l'estudi detecti un efecte si realment hi ha un efecte. És inversament proporcional a l'error Tipus II. Si el poder estadístic és alt, 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.
A l'annex C s'expliquen mètodes per tractar un disseny desbalancejat.

2.2 Correcció per multiplicitat de contrastos

Un problema comú què ens podem trobar a qualsevol investigació és voler comparar més de 2 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 $G = 1, .., N$ 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:

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

$$P(\text{No cometre l'error}) = 1 - \alpha$$

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

$$P(\text{No cometre l'error en m tests}) = (1 - \alpha)^m$$

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

Per exemple, si tenim 1 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.

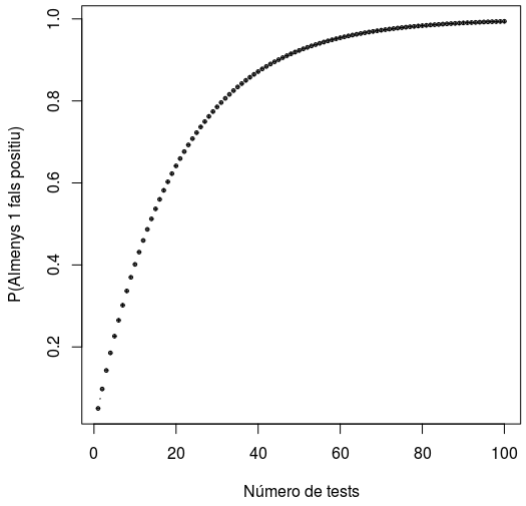

Figura 1: Gràfic contraposant el nombre de tests amb la probabilitat d'obtenir almenys 1 fals positiu. S'observa un increment en la probabilitat quan augmenta el nombre de tests.

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 False discovery Rate

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 d'aquests mètodes és massa conservador i impedeix que es detectin diferències reals. Una alternativa és controlar el false discovery rate.

El false discovery rate (FDR) es defineix com la proporció esperada de falsos positius d'entre tots els tests considerats com significatius. L'objectiu de controlar el false discovery rate es establir un límit de significació per a un conjunt de tests tal que, d'entre tots els tests considerats com significatius, la proporció d'hipòtesis nul·les (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 són realment falsos positius.
La primera aproximació per controlar el FDR va ser descrita per Benjamini % Hochberg en 1995. D'acord amb la seva publicació 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 *pvalor* (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 *pvalors* 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:

	Pvalor	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: A l'exemple de la taula hi ha una columna amb les *pvalors* sense corregir, un altre amb la correcció de Bonferroni i per últim una amb la correcció proposada per Benjamini&Hochberg, per a un total de 55 *pvalors* (Encara que només es mostren els 10 primers). Observem que Bonferroni és un mètode molt més conservador i cap *pvalor* és significatiu. (Dels 55 *pvalors*, 0 són significatius amb Bonferroni). En canvi, amb el mètode de Benjamini&Hochberg, observem que del total de *pvalors*, 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. En els estudis exploratoris és d'esperar que la proporció d'hipòtesis nul·les falses, és a dir, tests que són realment significatius, sigui alta. Per tant, mètodes que només depenen del nombre de tests no són els més potents per aquest tipus d'estudis, 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:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \mu_1 \neq \mu_2 \neq ... \neq \mu_k$$

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, també segueix una distribució del rang estudentitzat. L'estadístic compara les mitjanes dels grups j i k, i ve donat per la següent equació:

$$HSD = q_{I,N-I}(\alpha)\sqrt{\frac{\hat{S}^2}{n}},$$

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

on $\hat{X}_j$ 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

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

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

$$IC(\mu_j - \mu_k)_{(1-\alpha)} = (\bar{X}_j - \bar{X}_k) \pm q_{I,N-I,1-\alpha}\sqrt{\frac{\hat{S}^2}{n}}$$

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 unes covariables. 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. El color i la intensitat de les caselles s'utilitzen per representar canvis (no en valor absolut) de l'expressió gènica.

Tenim el següent exemple que podem obtenir a l'aplicatiu:

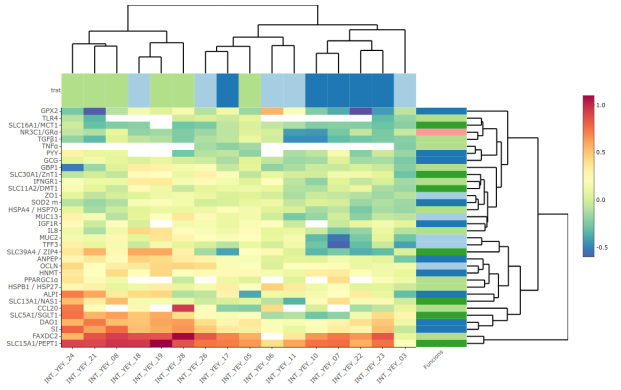

Figura 2: `Heatmap` generat a partir de 34 gens i 16 mostres. A les files trobem els gens i a les columnes les mostres. A més, es troben agrupats tant per gens com per mostres per mitjà d'un clúster jeràrquic. Per obtenir més informació, s'ha afegit variables fora de la matriu, concretament, a quin tractament 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 *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 optar 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 variables sobre l'i-èssim individu del k-èssim clúster, suposant que aquest clúster 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:

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

- 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.

Observació. El mètode de Ward és un dels més utilitzats en la pràctica; posseeix gairebé tots els avantatges del mètode de la mitjana i sol ser més discriminant en la determinació dels nivells d'agrupació. Una investigació duta a terme per Kuiper i Fisher va provar que aquest mètode era capaç d'encertar millor amb la classificació òptima que altres mètodes.

2.4.2 Anàlisi per mitjà del Biplot

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 més enfocada al biplot, un tipus de gràfic exploratori on s'analitzen les correlacions entre les variables.

L'objectiu amb les Components Principals

Donades p variables $X_1, X_2, ..., X_p$, volem construir p noves variables $Z_1, Z_2, ..., Z_p$ (que anomenarem components principals) amb les següents condicions:

1. Les noves variables han de contenir la mateixa informació que les variables originals
2. Les noves variables han de ser incorrelacionades i de variància 1
3. Les noves variables han d'estar ordenades de major a menor importància

Les variables originals han de ser una mitjana ponderada de les components principals, on les variables més importants tinguin un major pes:

$$X_j = w_{j1}Z_1 + w_{j2}Z_2 + ... + w_{jp}Z_p$$

per a $i = 1, 2, ..., p$.

L'element clau d'una anàlisi de components principals és, doncs, la matriu de pesos que permet expressar les variables originals com a combinació lineal de les noves variables.

$$\begin{bmatrix} w_{11} & w_{12} & ... & w_{1p} \\ w_{21} & w_{22} & ... & w_{2p} \\ ... & ... & ... & ... \\ w_{31} & w_{32} & ... & w_{33} \end{bmatrix}$$

Com que hem construït les variables $Z_1, Z_2, ..., Z_p$ de manera que fossin incorrelacionades, es verifica que la suma dels quadrats dels pesos de la fila i és igual a la variància de X_i:

$$w_{i12} + w_{i22} + ... + w_{ip2} = s_{i2}$$

També es demostra que la suma dels quadrats dels pesos de la columna j és la variància total que captura Z_j , que anomenarem variància explicada per la component j i la denotarem per λ_j :

$$w_{1j}^2 + w_{2j}^2 + ... + w_{pj}^2 = \lambda_j$$

Com que hem posat la condició que les noves variables estiguin ordenades per ordre d'importància i que les components més importants tinguin un major pes, en la matriu de pesos observem que la primera columna (que conté els pesos de la primera component principal en les variables originals) contindrà nombres en conjunt més grans que els de la segona; aquesta els contindrà més grans que els de la tercera; i així successivament. Per tant:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$$

Com que la suma dels quadrats de tots els elements de la matriu dels pesos ha de ser la mateixa tant si la suma es fa per files (Σs_i^2) com si es fa per columnes ($\Sigma \lambda_i$), les components principals $Z_1, Z_2, ..., Z_p$ contenen la mateixa informació que les dades originals i la suma dels seus valors propis coincideix amb la variància total de les variables $X_1, X_2, ..., X_p$:

$$\lambda_1 + \lambda_2 + ... + \lambda_p = s_1^2 + s_2^2 + ... + s_p^2 = \text{Variació Total}$$

Per tant, el percentatge de la variació total del model que està continguda en la component j és igual a:

$$\frac{\lambda_j}{\Sigma_{i=1}^{p} \lambda_i} \cdot 100$$

Biplot

El biplot aproxima la distribució d'una mostra multivariant en un espai de dimensió reduïda, normalment de dues dimensions, i superposa sobre la mateixa gràfica les representacions de les variables sobre les quals es mesura la mostra. Un biplot permet mostrar gràficament la informació de les files (individus, casos, etc, ...) i les columnes (variables) d'una matriu de dades multivariants. El prefix bi es refereix a aquest fet i no al fet que el gràfic es fa normalment en dues dimensions.

Podem veure un exemple a la següent figura:

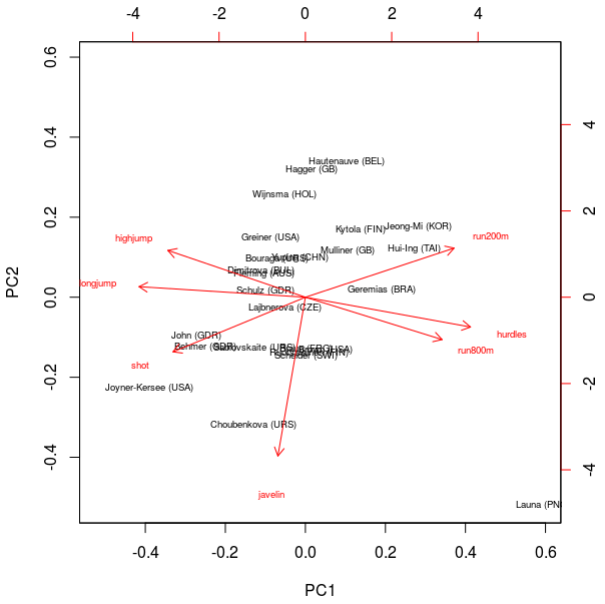

Figura 3: Biplot de les dues primeres components (dades escalades). Aquest és un exemple que es pot trobar al llibre *A Handbook of Statistical Analyses Using R. Brian S. Everitt and Torsten Hothorn.* Les dades corresponen als resultats de 25 competidors en 7 proves esportives.

Aquest gràfic és important ja que ens dóna molta informació:

"*La longitud de cada fletxa ens informa del percentatge de variació de cada variable explicada per les dues components principals.*"

De fet, la longitud és exactament igual a l'arrel quadrada de la comunalitat de l'extracció (que és el percentatge de la variació de cada variable que queda explicada). En efecte, el teorema de Pitàgores ens diu que la suma dels quadrats dels catets és igual al quadrat de la hipotenusa i, en aquest cas, els catets són els pesos i la hipotenusa la longitud del vector. Per tant, tenim que:

$$(longituddelvector)^2 = w_{11}^2 + w_{12}^2$$

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 desllletament. 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 46 casos.
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 46 casos, queden 23 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.

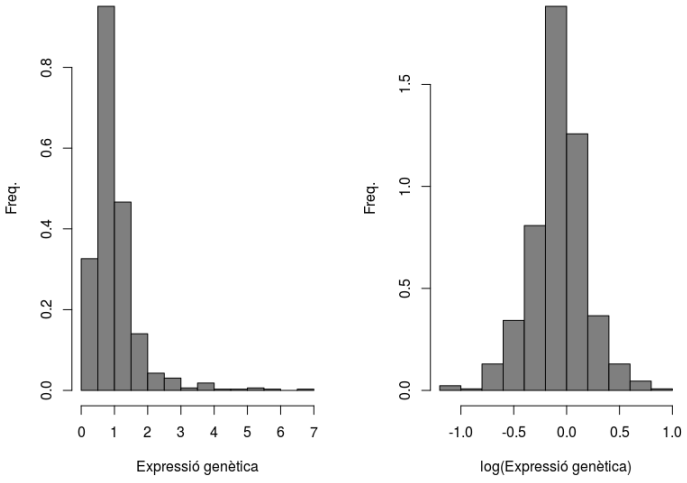

Figura 4: (1) El gràfic de l'esquerra mostra les dades sense cap tipus de transformació.(2) El gràfic de la dreta mostra les dades amb una transformació logarítmica.

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 pvalors 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 pvalor, 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 pvalors 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 pvalors 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.

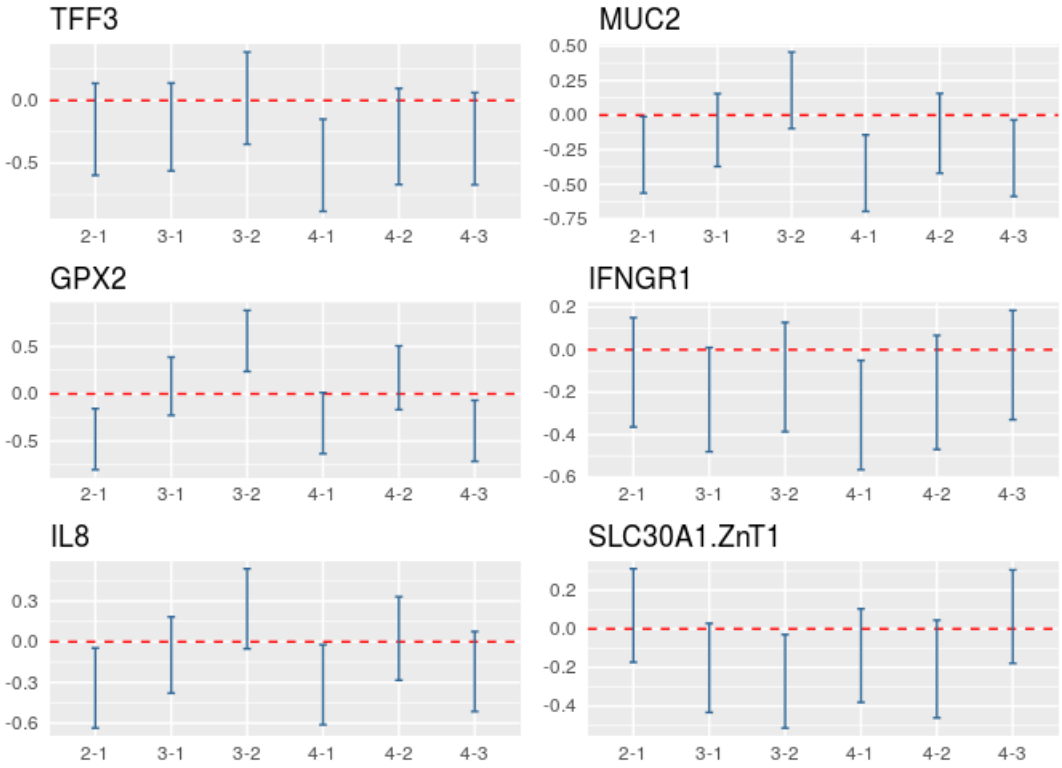

Figura 5: Intervals de confiança del 95% per a la diferència de mitjanes per a cada comparació i per a cada gen del teixit **Ilenum**. Els intervals que no contenen el 0, es mostren com comparacions estadísticament significatives.

En aquest teixit, 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 pvalors quasi significatius o significatius.

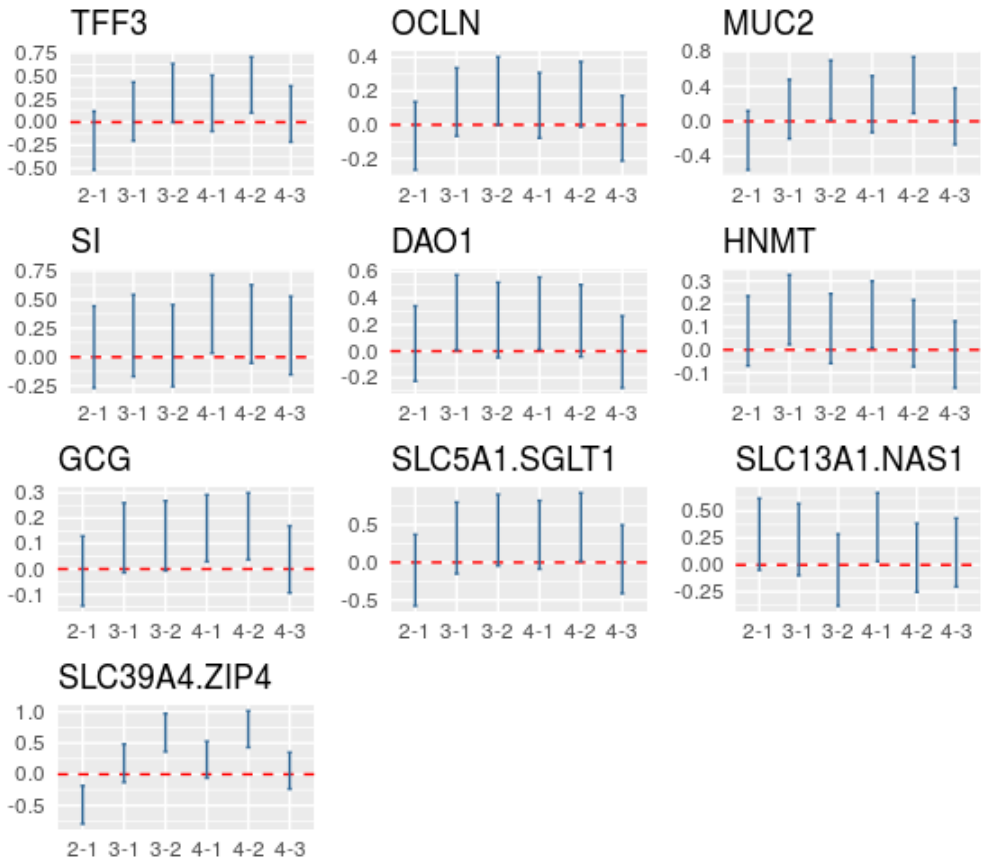

Figura 6: Intervals de confiança del 95% per a la diferència de mitjanes per a cada comparació i per a cada gen del teixit **Jejunum**. Els intervals que no contenen el 0, es mostren com comparacions estadísticament significatives.

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 *biplot* per a cada teixit:

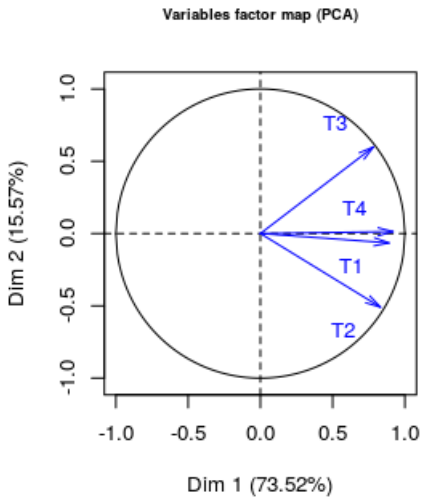

Figura 7: Biplot de les mitjanes per cada gene-tractament del teixit **Ilenum**.

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.

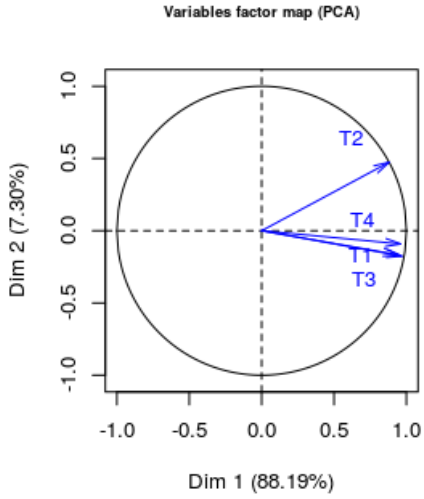

Figura 8: Biplot de les mitjanes per cada gene-tractament del teixit **Jejunum**.

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.

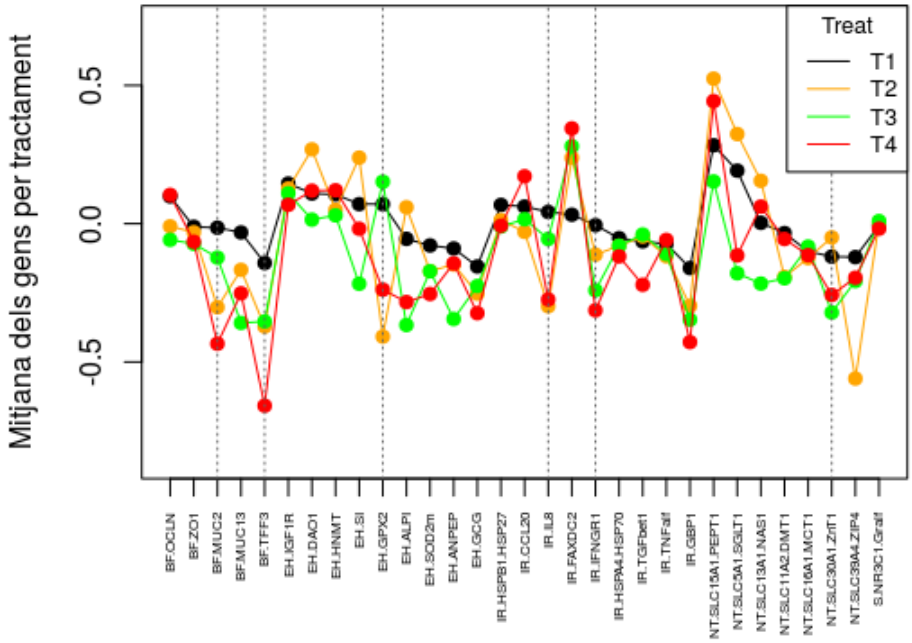

Figura 9: Mitjana dels gens per tractament, teixit **Ilenum**. El gràfic esta ordenat per família gènica primer i dins de la familia per el tractament 1 de forma decreixent. El nom de cada gen esta acompanyat de la seva funció dins de l'organisme, com per exemple, *BF.OCLN* = funció del gen: *Barrer Function*, nom del gen: *OCLN*

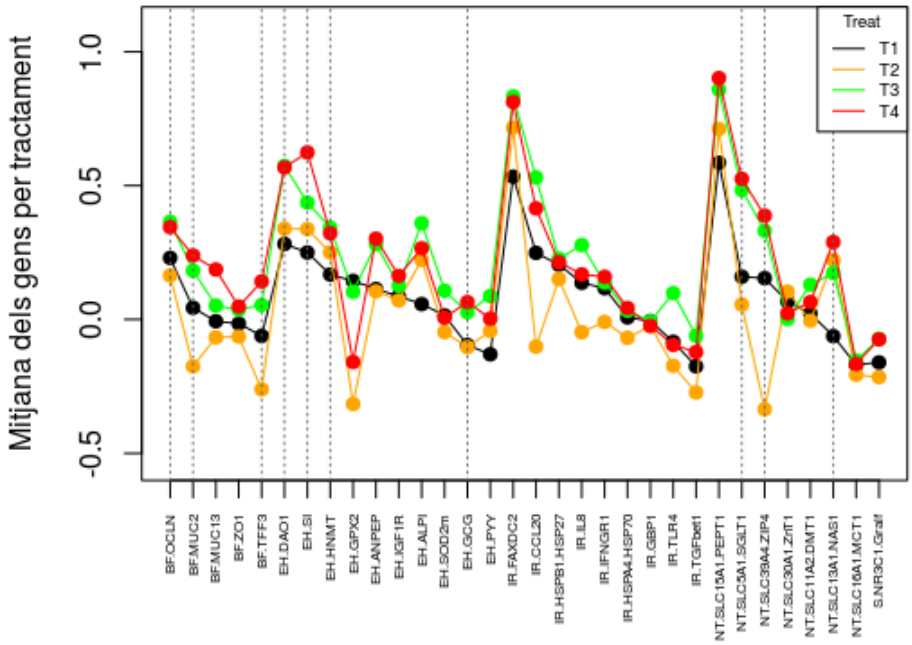

Figura 10: Mitjana dels gens per tractament, teixit **Jejunum**. El gràfic esta ordenat per família gènica primer i dins de la familia per el tractament 1 de forma decreixent.

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 (BT) dins del organisme.

Per últim, es presenten els *heatmap* per a cada teixit. Les conclusions d'aquest apartat queden incloses en les conclusions finals del cas d'estudi.

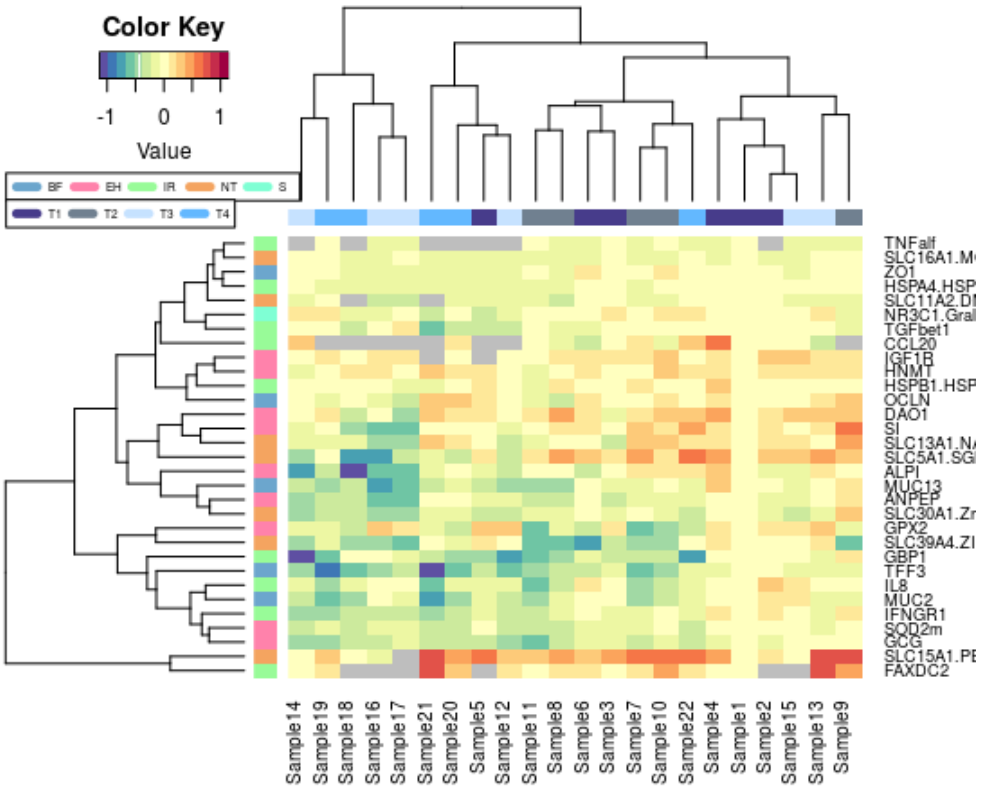

Figura 11: Heatmap del teixit**Ilenum**.

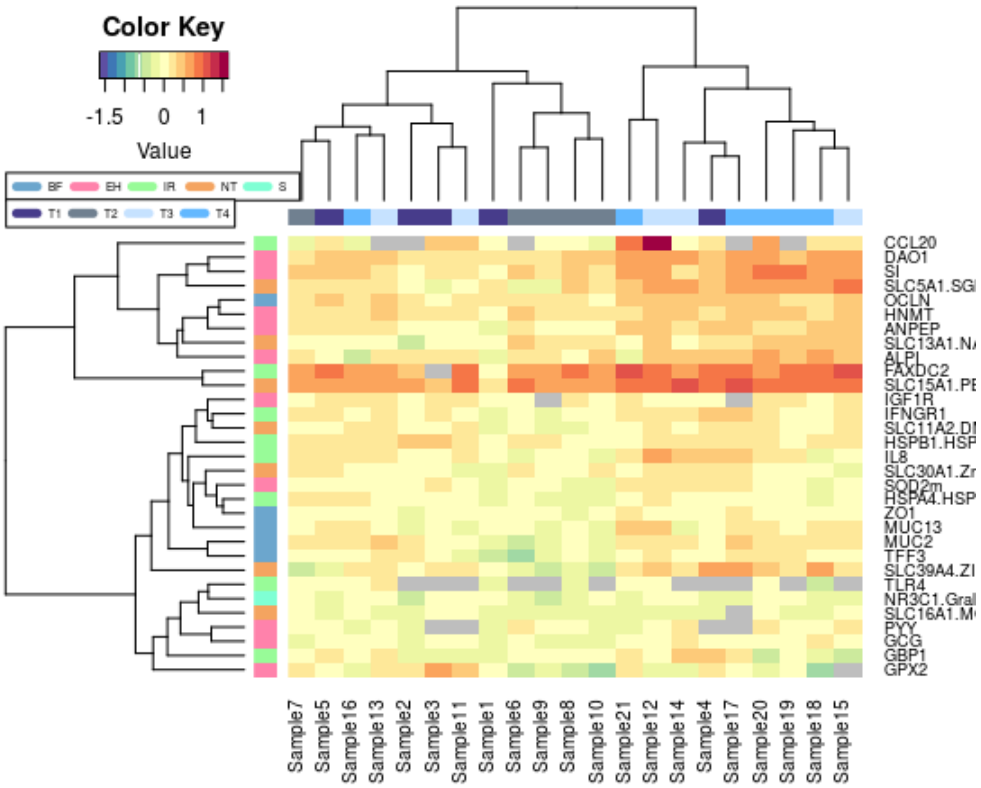

Figura 12: Heatmap del teixit **Jejunum**.

3.3 Conclusions

MIRAR QUE CONCLUYE EN EL TFG DE GENES.

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` (*A user interface object*)
- `server.R` (*A server function*)
- `App.R` (*A 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/tree/UserInterface`. A l'enllaç 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:

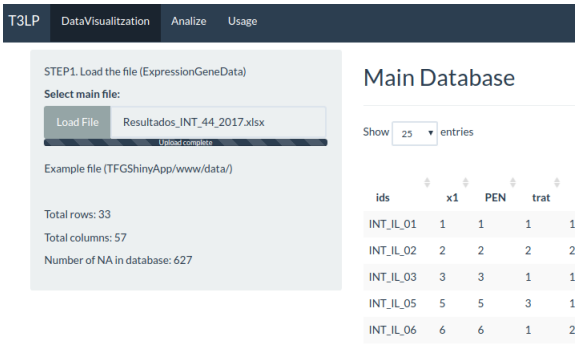

Figura 13: Imatge de l'aparença de l'aplicació i la seva estructura de pestanyes.

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 el posterior protocol d'anàlisi.
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 D 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:

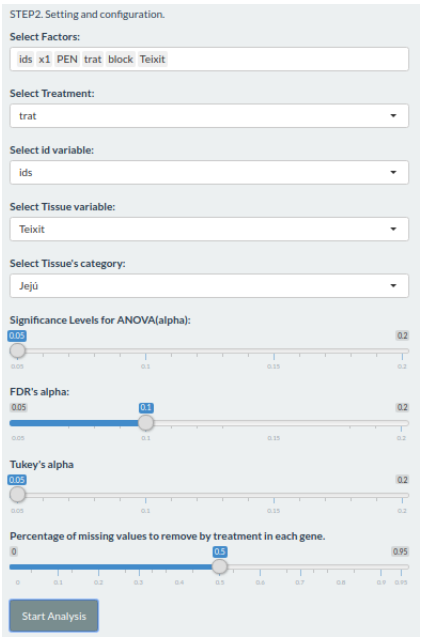

Figura 14: Panell de configuració dels paràmetres

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
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 al 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.

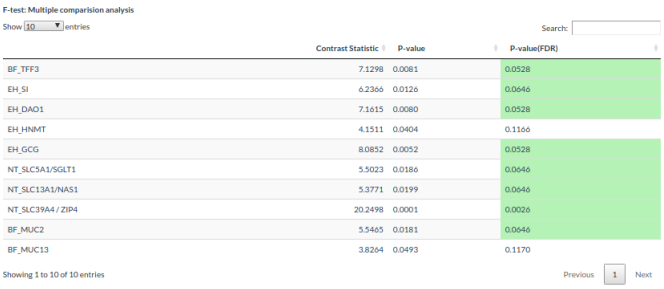

Figura 15: Taula de resultats de l'ANOVA

Figura 16: 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

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.

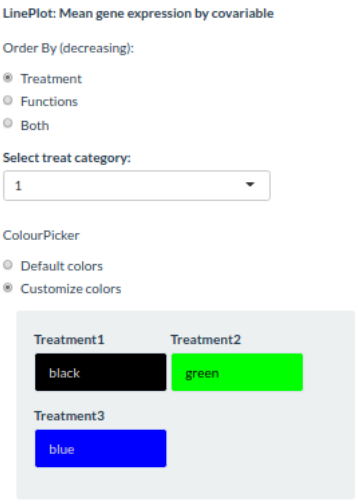

Figura 17: Panell de configuració del *LinePlot*

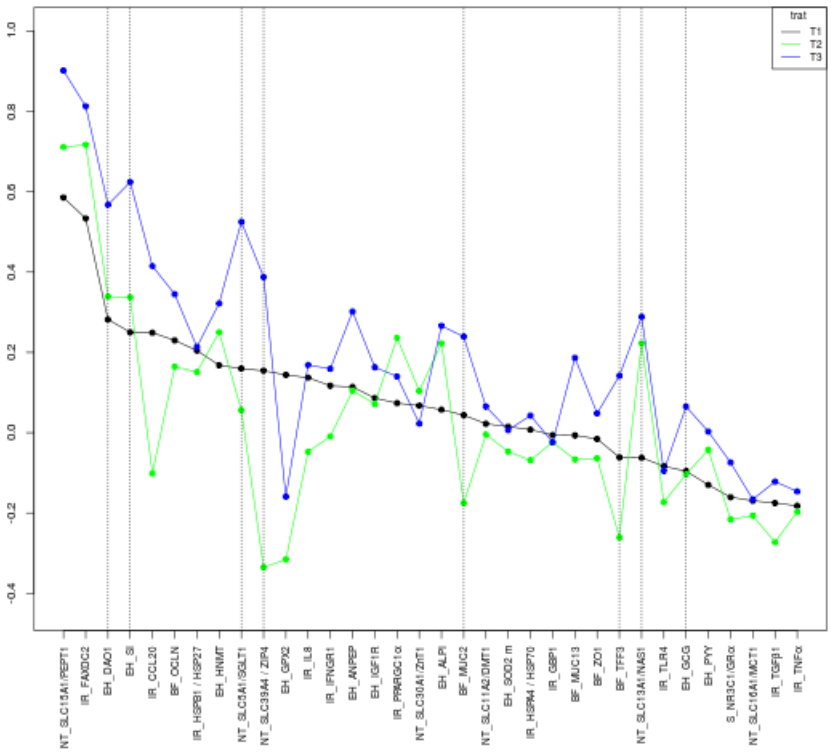

Figura 18: *LinePlot*. Exemple on ordenem els gens (eix x) de forma decreixent pel valor del tractament 1 i escollim els colors per a cada tractament.

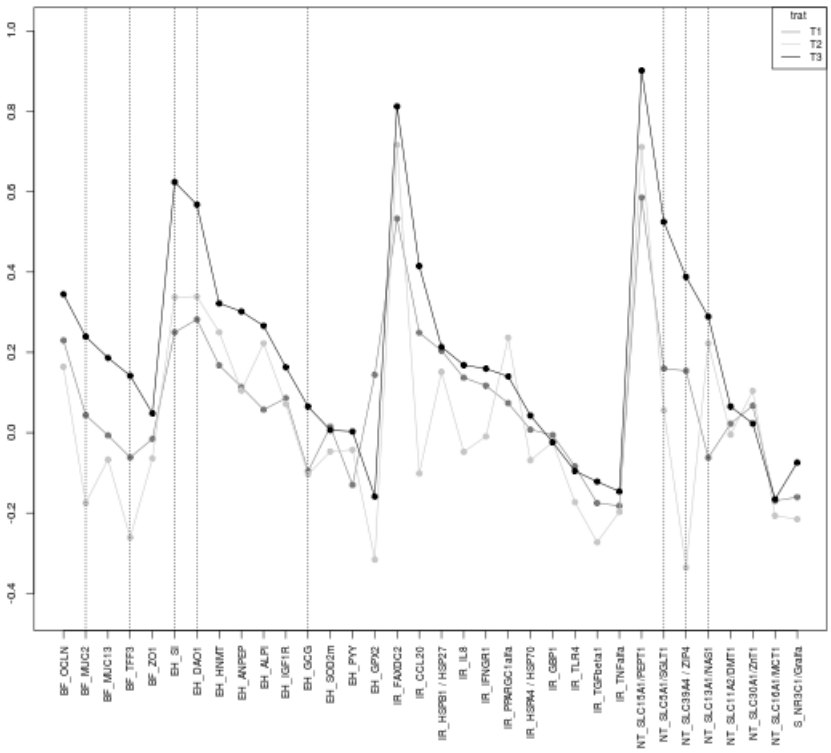

Figura 19: *LinePlot*. Exemple on ordenem els gens (eix x) de forma decreixent pel valor del tractament 3 i agrupat per la funcionalitat del gen. A més escollim una escala de grisos.

El següent gràfic que trobem és el *Biplot*, on podem comparar les correlacions entre les mitjanes dels tractaments:

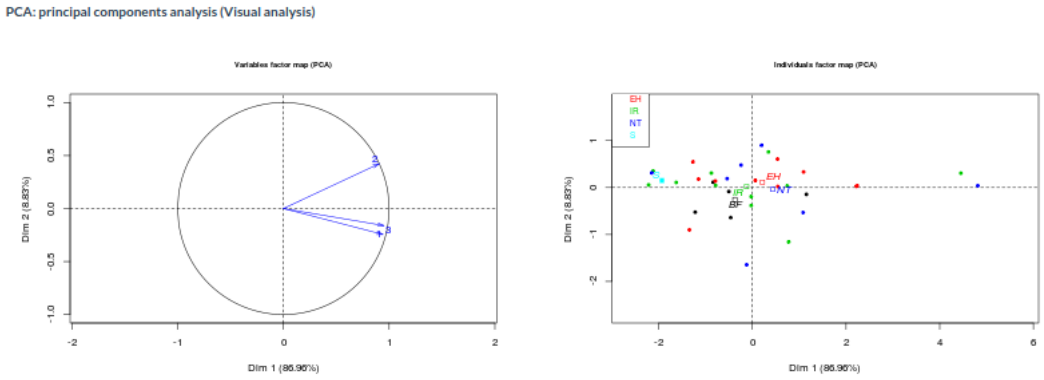

Figura 20: *Biplot*

El *Heatmap* el trobem de forma interactiva i amb moltes components interessants per analitzar:

- La matriu d'expressió, podem trobar patrons en ella.
- Els dendrogrames, tant les agrupacions per casos com les agrupacions per gens, ens poden ajudar a fer clústers.
- Les variables, *tractament* i *funció del gen*, donen molt suport descriptiu a l'anàlisi.

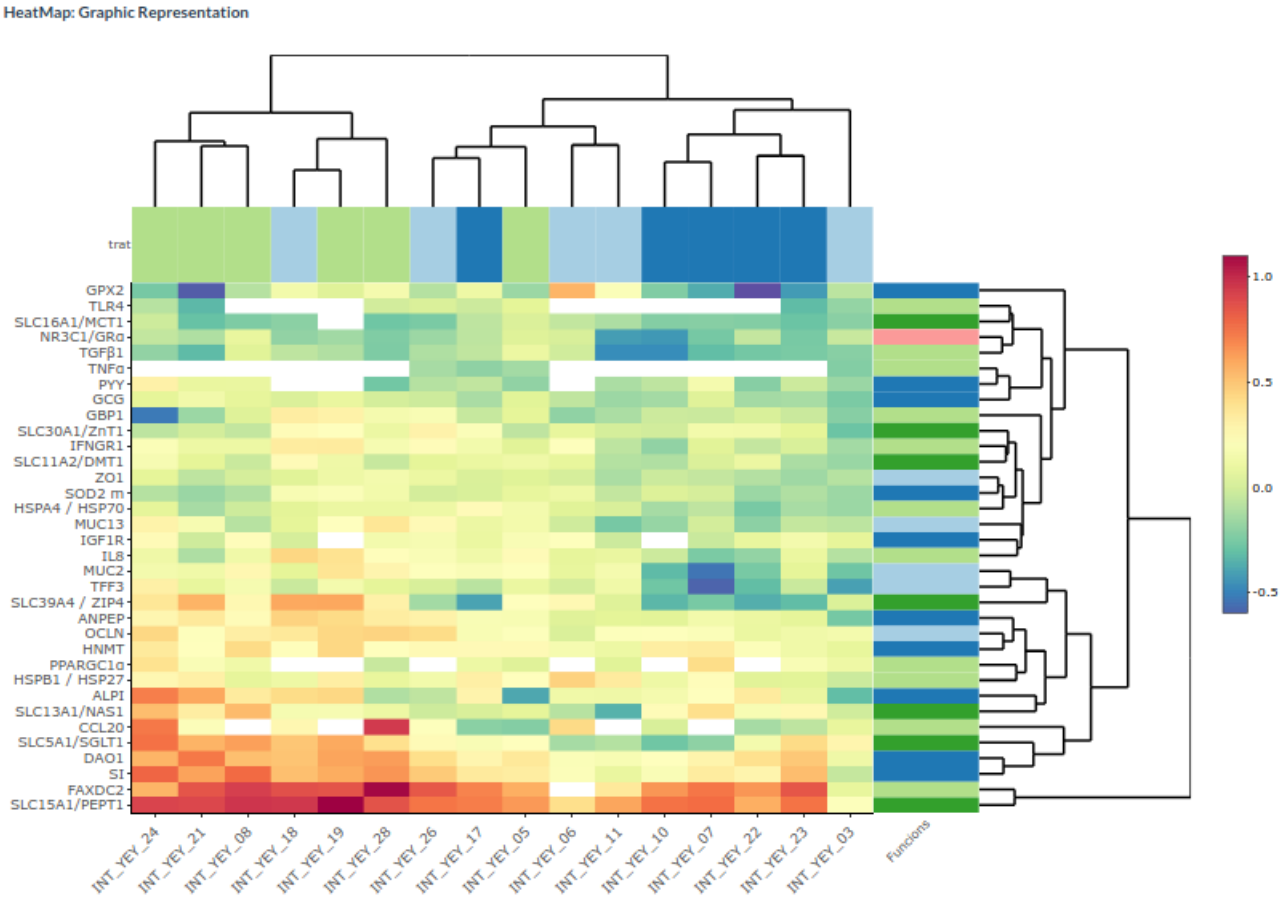

Figura 21: *Heatmap* interactiu

4.3 Desenvolupament i futures versions

A dia d'avui l'aplicació està a la versió 1.0 i en funcionament. Encara que es poden realitzar nous canvis per millorar l'experiència, com per exemple, mantenir l'aplicació en un servidor, d'aquesta manera no faria falta descarregar el codi, i directament l'aplicació estaria connectada a la xarxa. Un altre idea per a futures versions podria ser la creació automàtica d'un informe amb els resultats de l'anàlisi.

Si mirem quins són els punts febles, podríem destacar el llarg temps de càrrega a causa de les gràfiques interactives que fan la pàgina més lenta. També seria important millorar l'estructura del codi per modularitzar les tasques i no repetir codi. Per últim, comentar tot el codi utilitzant un manual d'estil, per facilitar l'entrada de nou codi en mans de futurs usuaris de l'aplicació.

Referències

[1] Mark D. Robinson, Davis J. McCarthy, Gordon K. Smyth; edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, Volume 26, Issue 1, 1 January 2010, Pages 139–140, https://doi.org/10.1093/bioinformatics/btp616

[2] K. R. GABRIEL; The biplot graphic display of matrices with application to principal component analysis, Biometrika, Volume 58, Issue 3, 1 December 1971, Pages 453–467, https://doi.org/10.1093/biomet/58.3.453

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

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

[5] Universidad de Granada: Comparaciones múltiples. http://wpd.ugr.es/∼bioestad/guia-spss/practica7

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

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

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

[9] *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
$(Gen_2, Gen_3), Gen_1$	$C_{Gen_1,Gen_2} = (10.08, 11.41, 7.45)$	$E_{Gen_2,Gen_3} = 20.79$ $E_{Gen_1} = 0$	20.79	20.79

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:

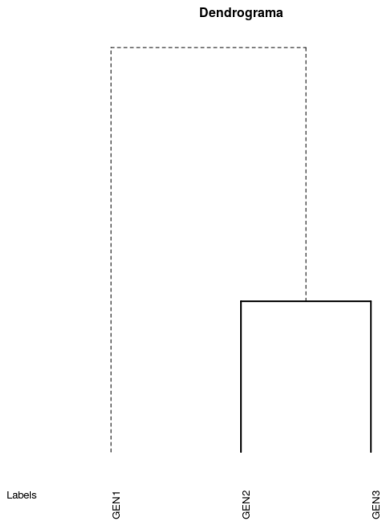

Figura 22: Dendrograma obtingut amb la funció `hclust`. Si apliquem directament la funció `hclust`, que utilitza el mètode de Ward, obtenim 2 clústers. El primer clúster amb el Gen_1 i un segon clúster amb el Gen_2 i el Gen_3. Trobem els mateixos clústers amb el procediment manual que hem calculat anteriorment.

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
	lL6	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 23: Gens *diana* i la seva funció dins de l'organisme.

C Anova unifactorial per a dissenys desbalancejats

D 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  }
```