OpenArray BIOIBERICA: Expression Analysis

amb aov() que tolera valors NA

May 29, 2019

Les dades han de llegir-se en format .csv o .xlsx i han de veure's al full1 del fitxer. Cap nom pot contenir caràcters estranys (ni ni % ni) ni lletres gregues; recomanem substituir α per alf, β per bet, γ per gam, etc. La primera fila ha de contenir el nom de la variable. Al full2 del fitxer, hi ha d'haver una columna amb el nom del gen (exactament igual que com apareix a la primera fila del full1) i una altra columna amb dues lletres que representin la funcionalitat del gen (BF, IR, etc.).

Si hi ha zeros, es transformen en NA.

Tractament dels NA's i logaritme

El tractament dels NA's és el següent:

1. S'eliminen aquelles i files (mostres) sense cap observació d'expressió vàlida.
2. S'eliminen aquelles columnes (gens) que, per a algun tractament, tinguin el 50% o més de rèpliques missing (si `del.badRows=T`) o un nombre de rèpliques mínim especificat a `noNAmin` a la funció `gestioNA( data, remove0=TRUE, del.badRows=TRUE, noNaMin=NULL)`.
3. No s'imputen els NA's fins a components principals (apartat 6.2) on s'aplica l'anomenat *EM-PCA algorithm* amb la funció `imputePCA()` de la llibreria `missMDA`. Cal tenir en compte que, per defecte, PCA() ja substitueix els valors pel valor mitjà, la qual cosa empobreix els resultats i és preferible aplicar un mètode adhoc.

4. En acabar el tractament dels NA, s'aplica logaritme decimal a les dades.

1.1 ANOVA diferències entre tractaments, per gens

Important: **FDR** significa *false discovery rate*. Per evitar els *falsos positius*, i decidir si un **p-valor té significacíó experimental** en el conjunt de tests que es fan, se solen fer correccions. Una de les més utilitzades és la de *Benjamini-Hochberg* que veieu a la columna *FDR p-value*. Interpretació: a nivell experimental només s'haurien de considerar significatius del conjunt experimental aquells tests en els que el *FDR p-value* estigui per sota de determinat threshold, per exemple 0.1. En aquest cas, a l'Ili no n'hi ha cap d'experimentalment significatiu, mentre que al Jejú sí.

No obstant, ens guiarem per la singificació a cada gen (ANOVA p-value) i mostrarem el FDR com a informació addicional. En l'ANOVA, els p-valors són significatius sí són < 0.05 i quasi-significatius si són < 0.1.

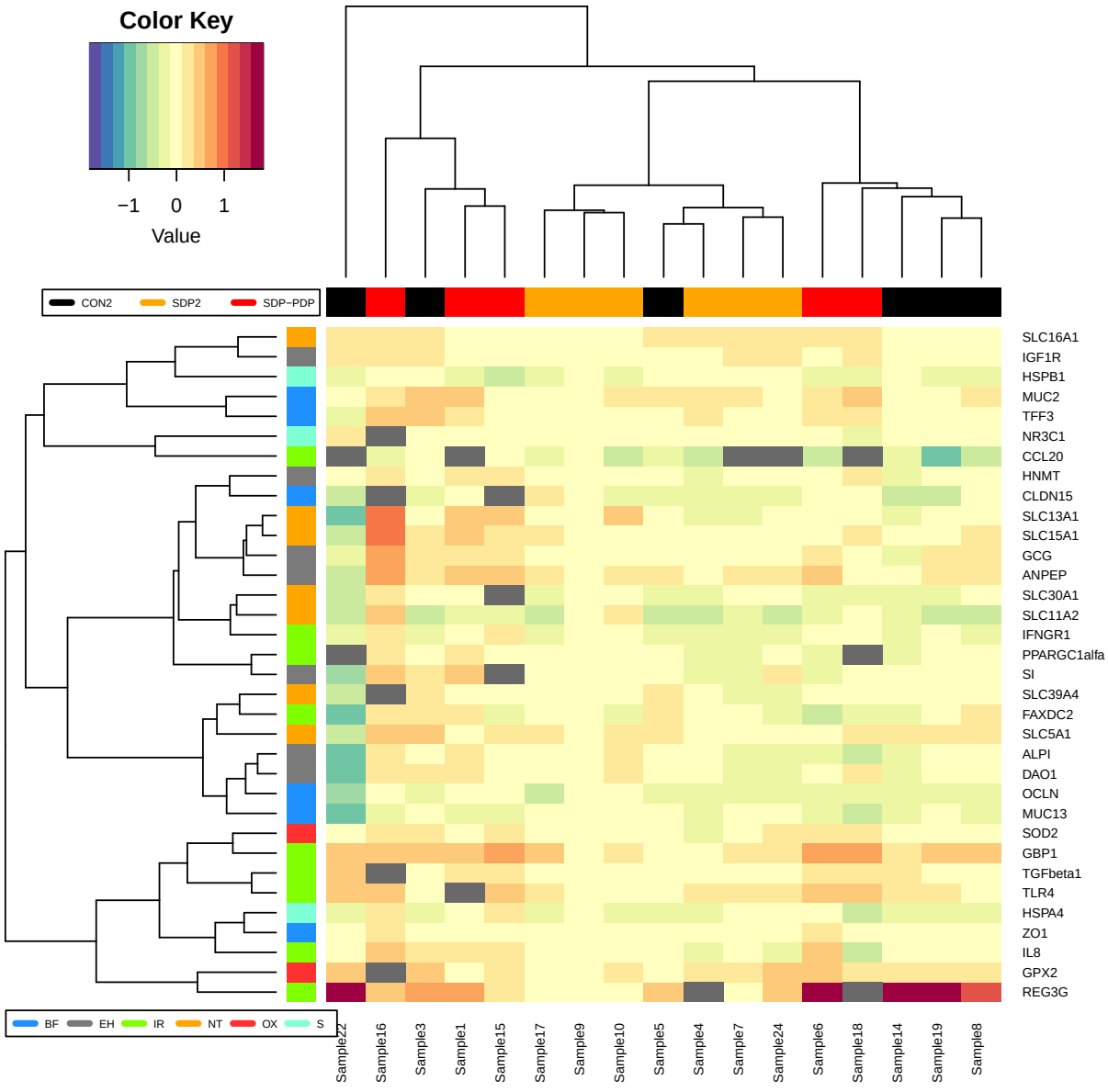
Seguidament, mostrem les taules amb tots els gens, siguin significatius o no.

1.2 Heatmap

La distància entre genss depèn del coeficient de correlació, concretament $d = \frac{1}{2}(1-r)$. La distància entre mostres és l'Euclidiana. El mètode d'enllaç jeràrquic és l'anomenat *complete* (veí més llunyà) per als gens i *ward.D2* per a les mostres.

1.2.1 Jejunum

La clau de colors correspon als nivells d'expressió. El color gris fosc correspon als NAs, que no semblen aleatoris perquè afecten al tractament SPD-PDP preferentment, i potser més als gens de funció IR.



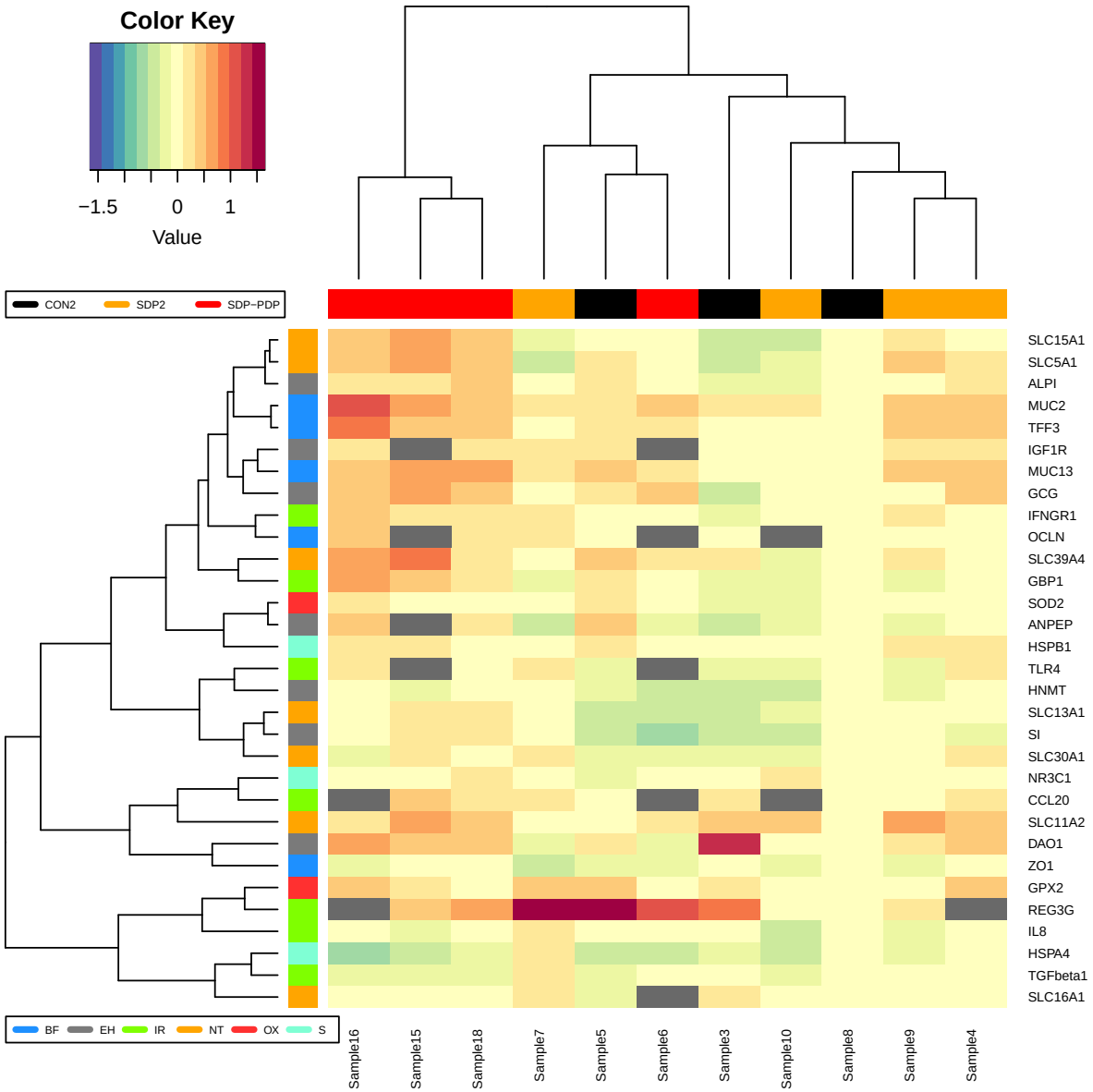
Al dendrograma de l'esquerra de la imatge hi ha els clústers dels gens, els 6 colors corresponen a les 6 funcionalitats gèniques. No es veu cap agrupació clara per funcionalitat (els colors del dendrograma estan força barrejats). Hi ha un gen outlier REG3G amb nivells d'expressió més alts en algunes de les mostres. El dendrograma a la part superior de la imatge correspon amb els colors negre, taronja i vermell segons la llegenda. La mostra Sample 22 (CON2) forma un clúster separat que destaca per nivells d'expressió més

baixos (colors blaus). Sample 16 (SDP-PDP) també destaca lleugerament per tenir nivells d'expressió més alts. Llevat de la mostra singular (sample22), es podrien considerar tres clústers amb les mostres de CON2 i SPD-PDP més barrejades. Les mostres tractades amb SDP2 queden més agrupades entre sí formant el clúster central i es caracteritzen per nivells d'expressió més intermedis. En general, els nivells d'expressió més elevats pertoquen a les mostres de SDP-PDP i els més baixos a les mostres de CON2.

Les similituds i diferències entre tractaments que veiem en un clúster (dendrograma) no són del mateix tipus que hem tractat en un test anova. Més explícitament, en l'ANOVA és comparen les mitjanes dels tractaments a cada gen per separat (tenint òbviament en compte les desviacions típiques). En l'anàlisi de clústers, les agrupacions entre tractaments es basen en el comportament en el conjunt dels gens comparant les distàncies entre mostres en tots els gens alhora.

1.2.2 Ileum

La clau de colors correspon als nivells d'expressió. El color gris fosc correspon als NAs, que no semblen aleatoris perquè afecten al tractament SPD-PDP preferentment, i potser més als gens de funció IR.



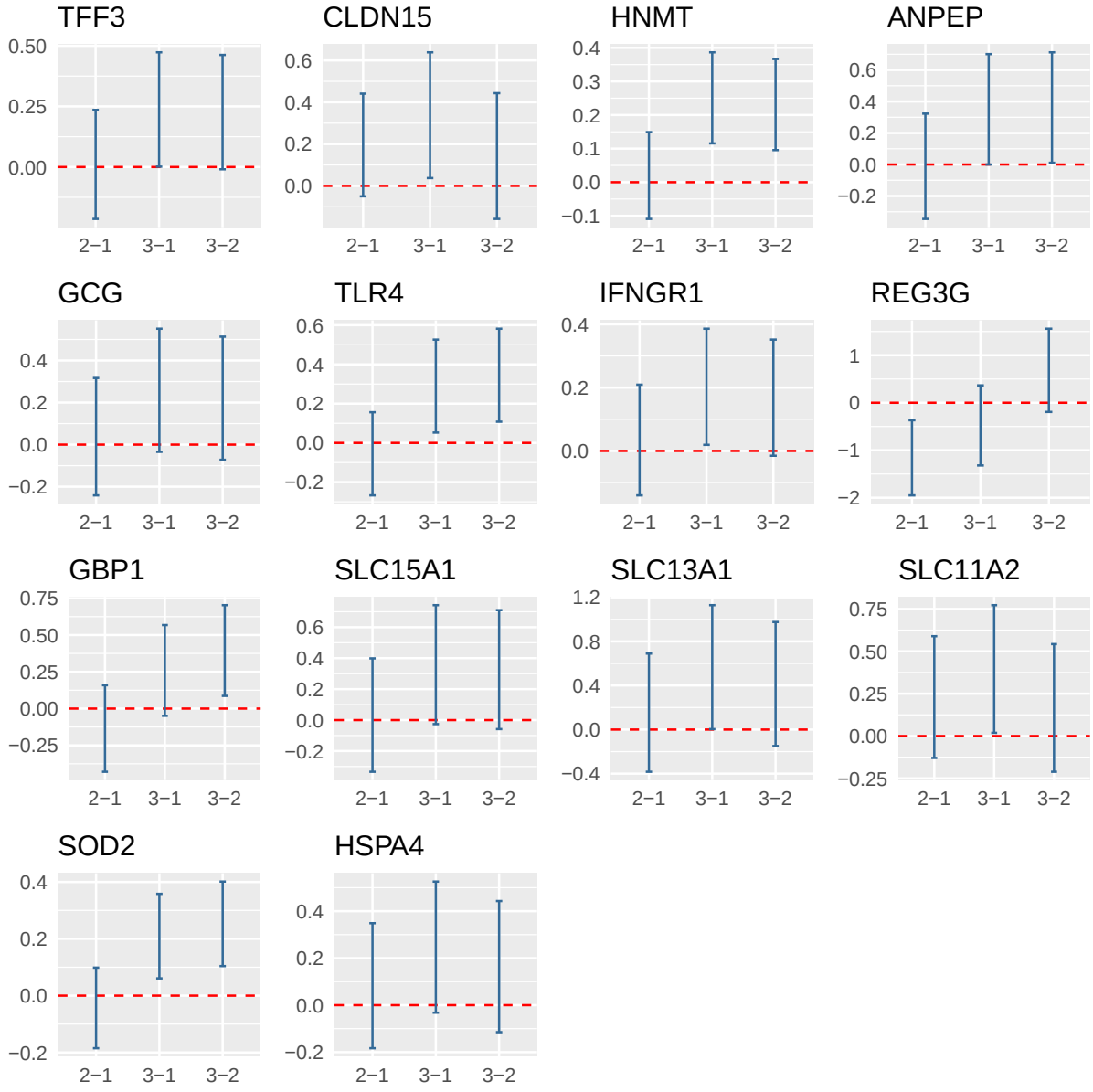
1.3 Tukey

1.3.1 Jejunum

Només s'han comparat dos a dos els gens amb diferències significatives a l'ANOVA de la Table 1.1.

	T1.mean	T1.sd	T2.mean	T2.sd	T3.mean	T3.sd	2-1	3-1	3-2
BF.TFF3	0.012	0.215	0.023	0.072	0.249	0.116	0.991	0.049	0.061
BF.CLDN15	-0.319	0.166	-0.123	0.180	0.019	0.055	0.126	0.028	0.440
EH.HNMT	-0.069	0.070	-0.049	0.086	0.182	0.101	0.914	0.001	0.001
EH.ANPEP	0.143	0.285	0.132	0.119	0.493	0.228	0.996	0.050	0.043
EH.GCG	0.005	0.174	0.043	0.057	0.264	0.278	0.934	0.087	0.156
IR.TLR4	0.158	0.171	0.103	0.123	0.448	0.099	0.772	0.017	0.005
IR.IFNGR1	-0.175	0.125	-0.140	0.106	0.028	0.115	0.865	0.030	0.074
IR.REG3G	1.282	0.530	0.122	0.251	0.805	0.638	0.005	0.321	0.136
IR.GBP1	0.346	0.199	0.211	0.224	0.606	0.141	0.470	0.105	0.012
NT.SLC15A1	-0.012	0.253	0.021	0.093	0.347	0.339	0.971	0.069	0.102
NT.SLC13A1	-0.176	0.408	-0.023	0.255	0.390	0.389	0.739	0.048	0.169
NT.SLC11A2	-0.439	0.066	-0.210	0.241	-0.044	0.346	0.249	0.039	0.500
OX.SOD2	0.013	0.073	-0.031	0.096	0.222	0.112	0.711	0.006	0.001
S.HSPA4	-0.220	0.053	-0.137	0.134	0.027	0.287	0.701	0.086	0.303

Table 1.1: Tukey:comparacions múltiples Gene-Tractament

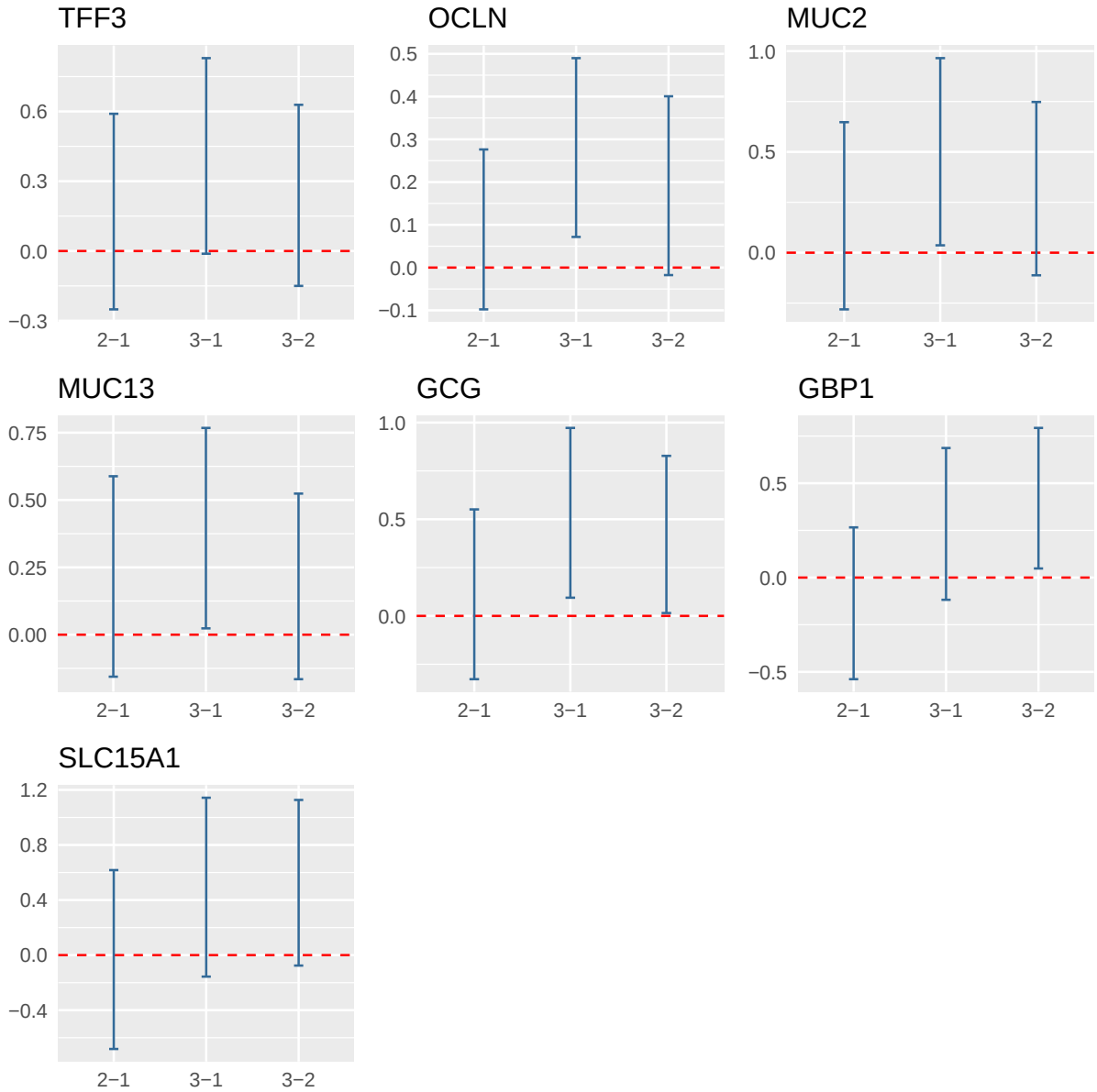


Nota: Veiem que les diferències significatives a la majoria de gens són entre SDP-PDP i els altres tractaments (un d'ells o ambdós); llevat del gen REG3G, no hi ha diferències significatives entre CON2 i SDP2.

1.3.2 Ilenum

	T1.mean	T1.sd	T2.mean	T2.sd	T3.mean	T3.sd	2-1	3-1	3-2
BF.TFF3	0.063	0.120	0.233	0.198	0.472	0.224	0.512	0.056	0.244
BF.OCLN	0.015	0.081	0.104	0.064	0.296	0.060	0.346	0.016	0.067
BF.MUC2	0.152	0.139	0.336	0.185	0.653	0.271	0.525	0.036	0.149
BF.MUC13	0.097	0.218	0.313	0.163	0.492	0.138	0.277	0.038	0.346
EH.GCG	-0.021	0.310	0.091	0.203	0.512	0.050	0.754	0.021	0.043
IR.GBP1	-0.007	0.225	-0.143	0.074	0.277	0.226	0.616	0.169	0.029
NT.SLC15A1	-0.083	0.236	-0.115	0.278	0.410	0.349	0.989	0.137	0.085

Table 1.2: Tukey: comparacions múltiples Gene-Tractament

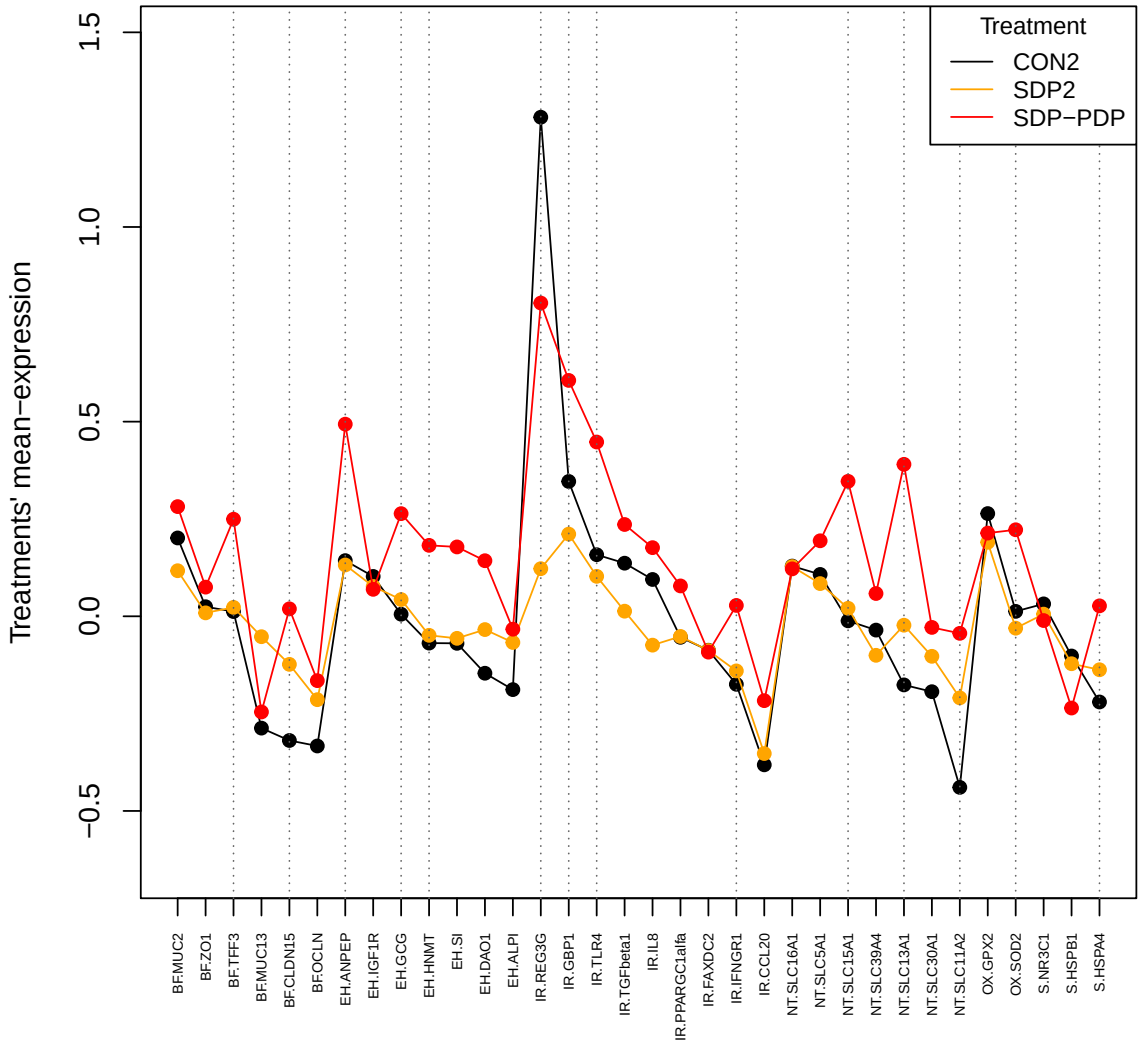


Nota: Com al Jejú, veiem diferències significatives a la majoria de gens entre SDP-PDP i els altres tractaments (un d'ells o ambdós).

1.4 Línies de mitjanes dels gens, per tractament

1.4.1 Jejunum

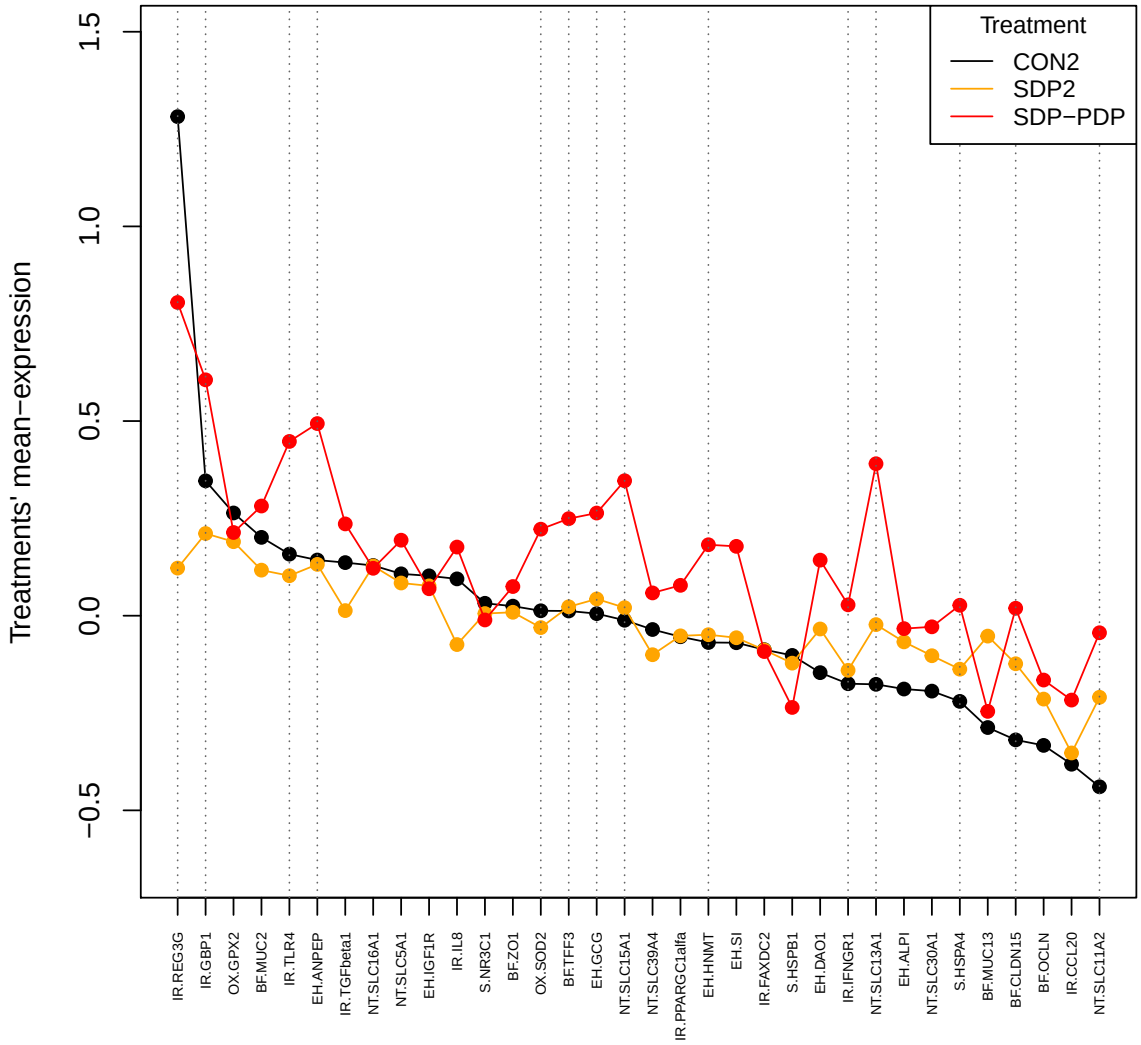
Ordenat per família gènica primer i dins de la família per T1 decreixent



Les línies puntejades indiquen els gens en els quals hi les ha diferències significatives més rellevants entre les mitjanes dels tractaments. No es visualitzen les desviacions típiques i, per tant, la significació pot no coincidir amb les distàncies verticals més elevades.

El tractament SDP-PDP tendeix a tenir nivells més elevats d'expressió en mitjana, de fet és superior en tots els gens amb diferències significatives (exceptuant REG3G).

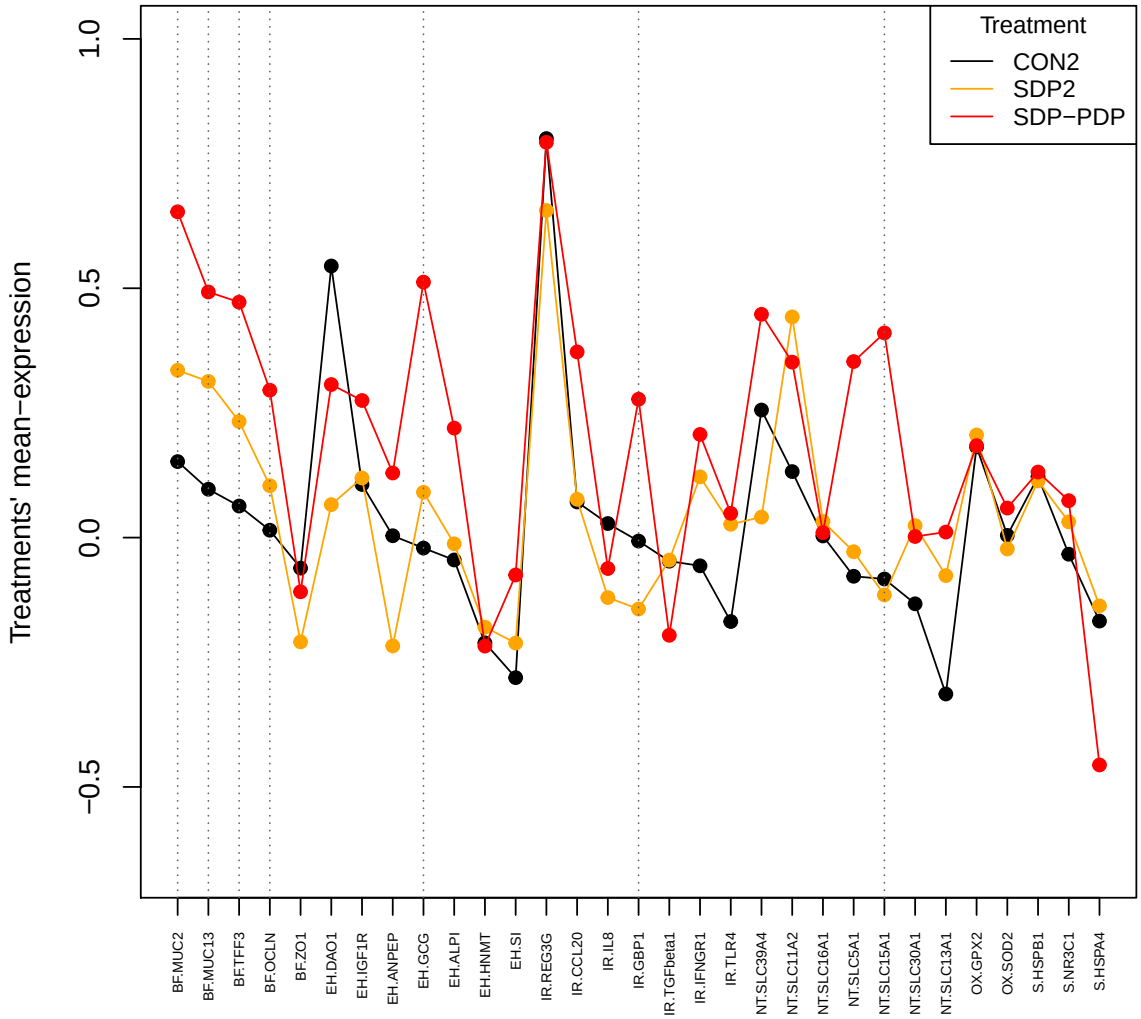
Ordenat per ordre decreixent d'expressió en el tractament T1



Com a la imatge anterior, però encara més clarament, el tractament SDP-PDP té valors més elevats en mitjana en tots els gens que presenten diferències significatives, menys en un d'ells.

1.4.2 Ilenum

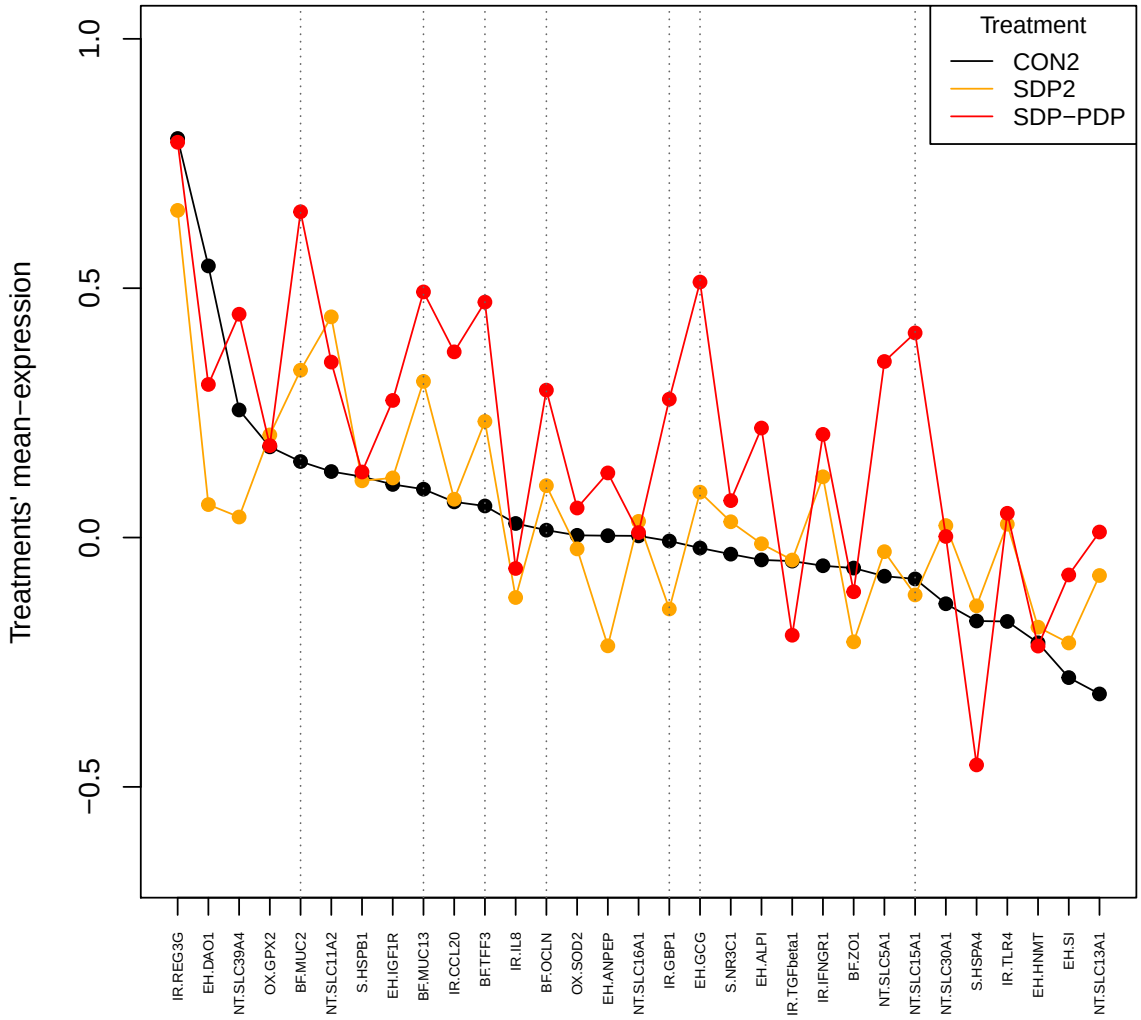
Ordenat per família gènica primer i dins de la família per CON2 decreixent



Compte que l'escala vertical no és la mateixa que pel Jejú, per això les oscil.lacions aparenten ser més significatives quan realment no ho són !

Tal i com passa en el Jejú, també a l'Ili el tractament SDP-PDP té valors més elevats en mitjana en tots els gens que presenten diferències significatives.

Ordenat per ordre decreixent d'expressió en el tractament T1



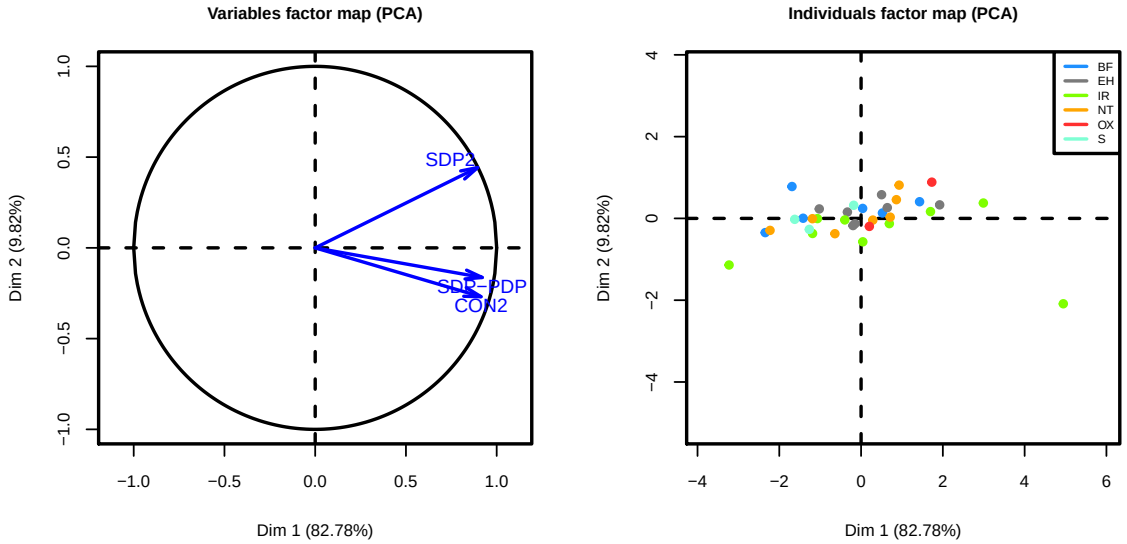
Mateixos comentaris que al gràfic anterior.

1.5 Components principals

1.5.1 Variables=mitjanes de tractaments, casos=gens

Una primera versió de components principals considera les variables=mitjanes dels tractaments (3 variables en aquest cas) i casos=gens. No és la versió estàndard. S'òbvia el detall de les mostres i només ens quedem amb els valors mitjans.

Jejunum: Variables=mitjanes de tractaments, casos=gens



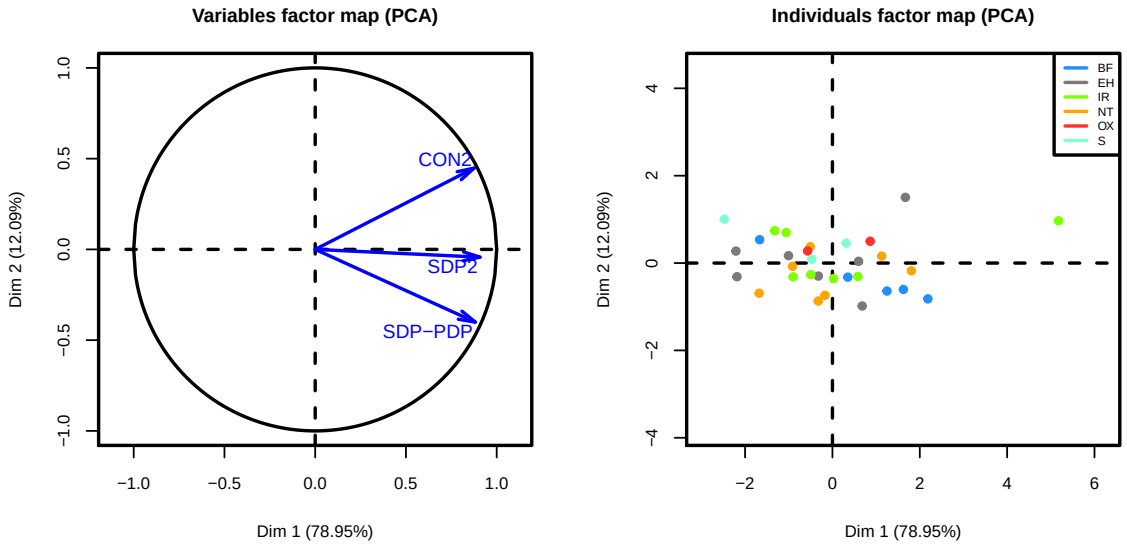
Amb dues components s'explica el 92.6% (82.78+9.82) de la variabilitat mitjana dels tractaments, en el cas del Jejú.

Important: les fletxes corresponen a les mitjanes dels tractaments en aquest cas. Les mitjanes de tractaments que tenen les fletxes amb un angle menor estan més correlacionades positivament, en el sentit que els seus nivells d'expressió mitjans en les gràfiques de línies pugen i baixen simultàniament d'una manera més evident (Fig. 1.4.1). Correlació negativa seria un angle proper a 180 graus i indicaria que quan un s'expressa més en un tractament, en l'altre menys.

En aquest cas, les mitjanes dels tres tractaments estan positivament correlacionats, però CON2 i SDP2 tenen una correlació més elevada.

A la gràfica de dispersió els punts són els gens (no les mostres) i els colors són les funcions gèniques. No s'aprecien agrupacions clares de colors.

Ilenum: Variables=mitjanes de tractaments, casos=gens

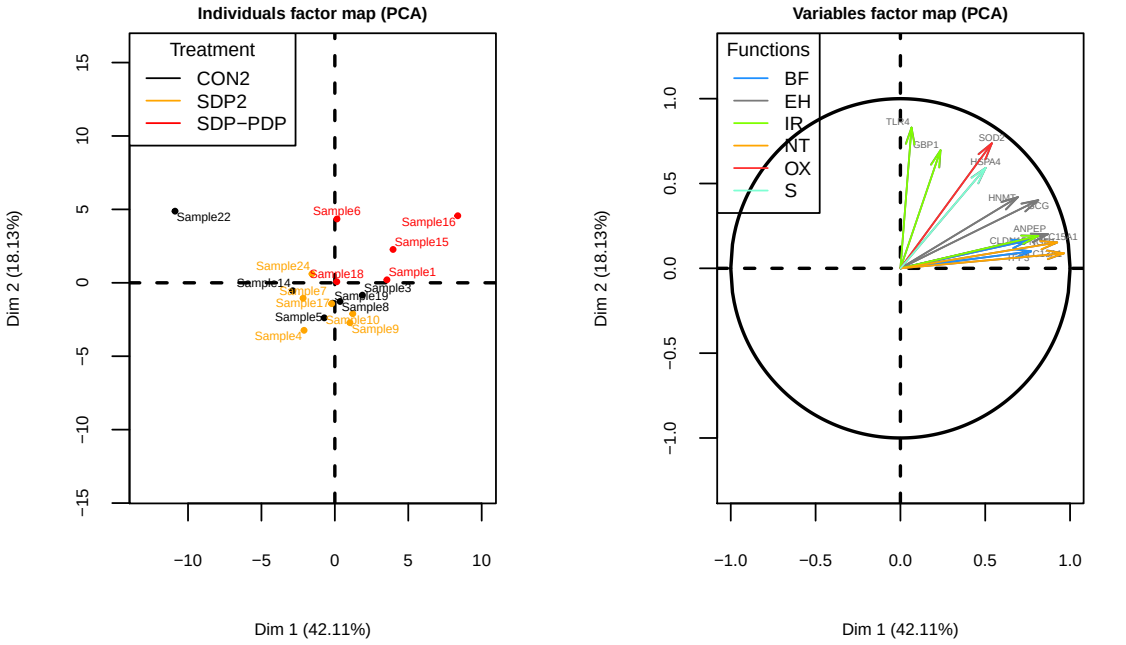


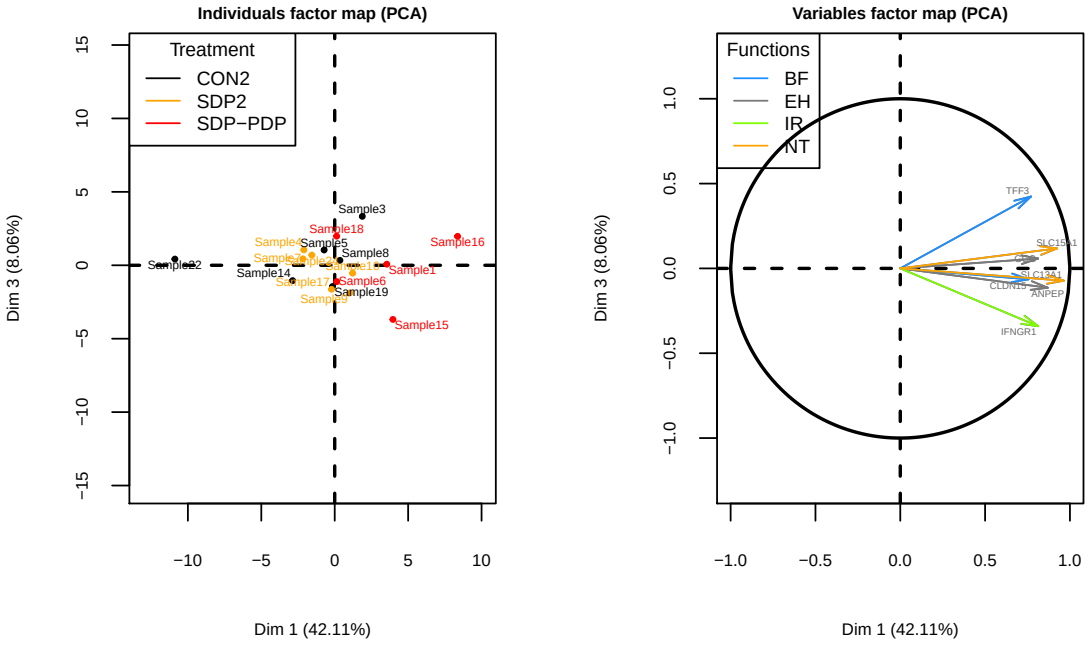
Amb dues components s'explica el 91.04% (78.95+12.09) de la variabilitat mitjana dels tractaments, en el cas de l'Ili.

1.5.2 Components principals: Variables=gens, casos=mostres

La segona versió de components principals és l'estàndard. Considera les variables=gens i casos=mostres. Només es mostren les fletxes d'aquells gens que són significatius i alhora estan ben representatts (qualitat de representació superior al 50% en el pla). Aquest mètode té globalment una qualitat de representació una mica més baixa perquè té el compte els nivells d'expressió de totes les mostres, no només els nivells mitjans en els tractaments.

Jejunum: Variables=gens, casos=mostres

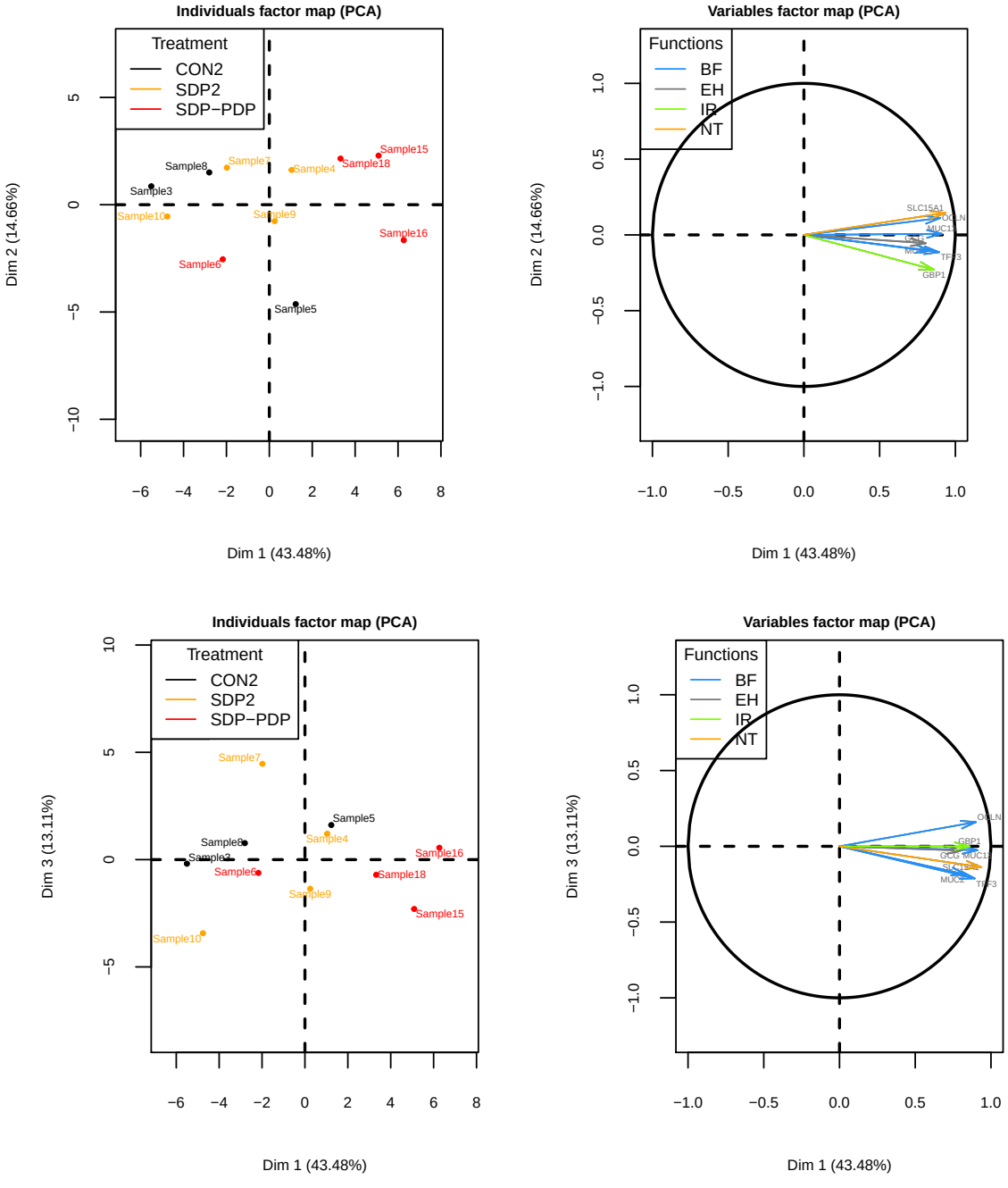




Les dues imatges superiors corresponen a les dues primeres components, les dues ingeriors a les components 1 i 3. La variabilitat explicada per les dues primeres components és del 60.24% (42.11+18.13),

la tercera component explica un 8.06%. En les dues primeres components, totes les mostres de SDP-PDP estan al primer quadrant, indicant valors d'expressió més elevada en tots els gens significatius i amb bona qualitat de repreentació (degut a que tots els gens també estan al primer quadrant). Mirant conjuntament les quatre gràfiques, també apreciem que les mostres de color taronja (SDP2) presenten nivells d'expressió mitjans en les components 1 i 3 i mitjans-baixos en la segona component.

Ileum: Variables=gens, casos=mostres



Els resultats són similars al Jejú, amb una tendència a tenir valors elevats d'expressió en les mostres de SDP-PDP (en 3 de les quatre mostres).

1.6 Conclusions

Totes les metodologies emprades (ANOVA, comparacions de Tukey, heatmaps, diagrames de línies i (dos tipus de) components principals) ens permeten concloure que:

- El tractament SPD-PDP presenta nivells d'expressió més elevats en determinats gens, més concretament, en tots els gens on hi ha diferències significatives tant a l'Ili com al Jejú. Això es veu a les taules i a les gràfiques de Tukey i de línies, i també al PCA pel cas de variables=gens i casos=mostres de l'apartat 1.6.2. El tractament SPD2 té un comportament intermedi i CON2 tendeix a presentar nivells d'expressió més baixos.
- A l'Ili hi ha menys gens on les diferències siguin significatives. En general, part de la manca de significació és atribuïble a l'escassedat de rèpliques, empitjorada per un grapat de valors d'expressió no-vàlids (missing).
- Els resultats de PCA de l'apartat 1.6.2. amb totes les mostres semblen més rellevants que els de l'apartat 1.6.1. només amb mitjanes, perquè permeten fer un mapa de tota la informació.
- Amb les dades disponibles, la funció gènica no sembla rellevant per determinar el comportament dels gens.
- Totes le eines de visualització i tests aplicats tenen present que, degut a la normalització efectuada a l'openarray, no és rellevant comparar els valors concrets d'expressió entre gens diferents sinó la seva correlació.
  - En particular, al heatmap s'ha usat una distància entre mostres basada en el coeficient de correlació de Pearson associat al mètode d'enllaç complet, mentre que entre mostres s'ha considerat la distància Euclidiana i el mètode d'enllaç de Ward, que s'associa bé a aquesta distància.
  - Els resultats del PCA de l'apartat 1.6.2 també es basen en la correlació entre gens, si bé es comparen els nivells d'expressió entre mostres.
- Tot i que com acabem de dir no sembla molt apropiat comparar nivells d'expressió en gens diferents, es pot destacar que el gen REG3G té un comportament outlier amb expressions més elevades, la qual cosa podria ser rellevant si es confirma en altres experiments.
- Els gens amb un nombre insuficient de valors vàlids (la meitat o més de rèpliques vàlides per tractament) s'han eliminat perquè suposen un disseny excessivament desequilibrat. Pels gens restants es tenen resultats de significació (p-valors de l'anova i Tukey) tot i que encara els resti algun valor NA. Així s'han recuperat alguns gens amb diferències significatives o quasi.