OpenArray BIOIBERICA: Expression Analysis

sense imputar missings

May 29, 2019

1.1 Gestió de dades

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.).

Tractament dels NA's

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. Un cop eliminades les columnes i files segons els criteris anteriors, es considera que els NA's restants són MAR (*missing at random*) i s'imputen amb la llibreria `mice`. **No s'aplica aquet apartat, es treballa amb valors perduts sense imputar cap valor.**

En acabar el tractament dels NA, es pren logaritme decimal de les dades.

1.2 ANOVA diferències entre tractaments, per gens

	gen-func	statistic	p-value	p-value FDR
TFF3	BF	4.289870	0.035226	0.135033
HNMT	EH	14.110654	0.000441	0.010137
ANPEP	EH	4.577323	0.029541	0.135033
GCG	EH	3.030333	0.080627	0.185441
IFNGR1	IR	4.660779	0.028093	0.135033
GBP1	IR	5.711002	0.015361	0.117770
SLC15A1	NT	3.540070	0.056988	0.145635
SLC13A1	NT	3.644385	0.053191	0.145635
SLC11A2	NT	3.862233	0.046159	0.145635
SOD2	OX	11.029692	0.001330	0.015292
HSPA4	S	2.735336	0.099364	0.207761

Table 1.1: P-valors de l'ANOVA entre tractaments pel teixit **Jejunum**, es mostren només els gens amb diferències significatives. Les dades faltants (NA) només afecten al gen on hi ha valors perduts per als quals no es realitza el test.

Important: **FDR** significa *false discovery rate*. Per evitar els *falsos positius*, i decidir si un **p-valor té significació 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.

	gen-func	statistic	p-value	p-value FDR
TFF3	BF	3.999960	0.062501	0.280902
MUC2	BF	5.048447	0.038189	0.278312
MUC13	BF	4.619160	0.046385	0.278312
GCG	EH	7.192223	0.016315	0.278312
GBP1	IR	5.384009	0.033013	0.278312
SLC15A1	NT	3.770276	0.070225	0.280902

Table 1.2: P-valors de l'ANOVA entre tractaments pel teixit **Ileum**, es mostren només els gens amb diferències significatives. Les dades faltants (NA) només afecten al gen on hi ha valors perduts per als quals no es realitza el test.

	statistic	p.value	p.BH
TFF3	4.289870	0.035226	0.135033
OCLN	1.688072	0.220406	0.334975
ZO1	2.247876	0.142361	0.272859
CLDN15			
MUC2	1.425364	0.273253	0.369696
MUC13	1.619238	0.233026	0.334975
SI			
DAO1	1.228128	0.322537	0.412130
HNMT	14.110654	0.000441	0.010137
ANPEP	4.577323	0.029541	0.135033
GCG	3.030333	0.080627	0.185441
IGF1R	0.188897	0.829947	0.908989
ALPI	0.401128	0.677019	0.816796
TLR4			
TGFbeta1			
CCL20			
IFNGR1	4.660779	0.028093	0.135033
REG3G			
PPARGC1alfa			
FAXDC2	0.000397	0.999603	0.999603
GBP1	5.711002	0.015361	0.117770
IL8	2.065524	0.163658	0.289548
SLC5A1	0.350626	0.710258	0.816796
SLC16A1	0.007725	0.992309	0.999603
SLC15A1	3.540070	0.056988	0.145635
SLC13A1	3.644385	0.053191	0.145635
SLC11A2	3.862233	0.046159	0.145635
SLC30A1			
SLC39A4			
GPX2			
SOD2	11.029692	0.001330	0.015292
HSPB1	1.719791	0.214854	0.334975
HSPA4	2.735336	0.099364	0.207761
NR3C1			

Table 1.3: P-valors de l'ANOVA entre tractaments pel teixit **Jejunum**. Es mostren tots els gens, sigui significaius o no. Les files buides corresponen als gens que tenen valors perduts.

	statistic	p.value	p.BH
TFF3	3.999960	0.062501	0.280902
OCLN			
ZO1	1.551361	0.269552	0.462089
MUC2	5.048447	0.038189	0.278312
MUC13	4.619160	0.046385	0.278312
SI	0.425341	0.667502	0.767556
DAO1	0.815407	0.476111	0.714166
HNMT	0.044331	0.956871	0.984280
ANPEP			
GCG	7.192223	0.016315	0.278312
IGF1R			
ALPI	2.511450	0.142406	0.388282
TLR4			
TGFbeta1	2.177256	0.175816	0.388282
CCL20			
IFNGR1	1.982830	0.199808	0.399616
REG3G			
GBP1	5.384009	0.033013	0.278312
IL8	0.467250	0.642804	0.767556
SLC5A1	2.313421	0.161132	0.388282
SLC16A1			
SLC15A1	3.770276	0.070225	0.280902
SLC13A1	1.292670	0.326243	0.521988
SLC11A2	1.843052	0.219625	0.405461
SLC30A1	0.643417	0.550667	0.734223
SLC39A4	2.270356	0.165604	0.388282
GPX2	0.015876	0.984280	0.984280
SOD2	0.418557	0.671611	0.767556
HSPB1	0.028564	0.971939	0.984280
HSPA4	2.158540	0.177963	0.388282
NR3C1	0.645271	0.549789	0.734223

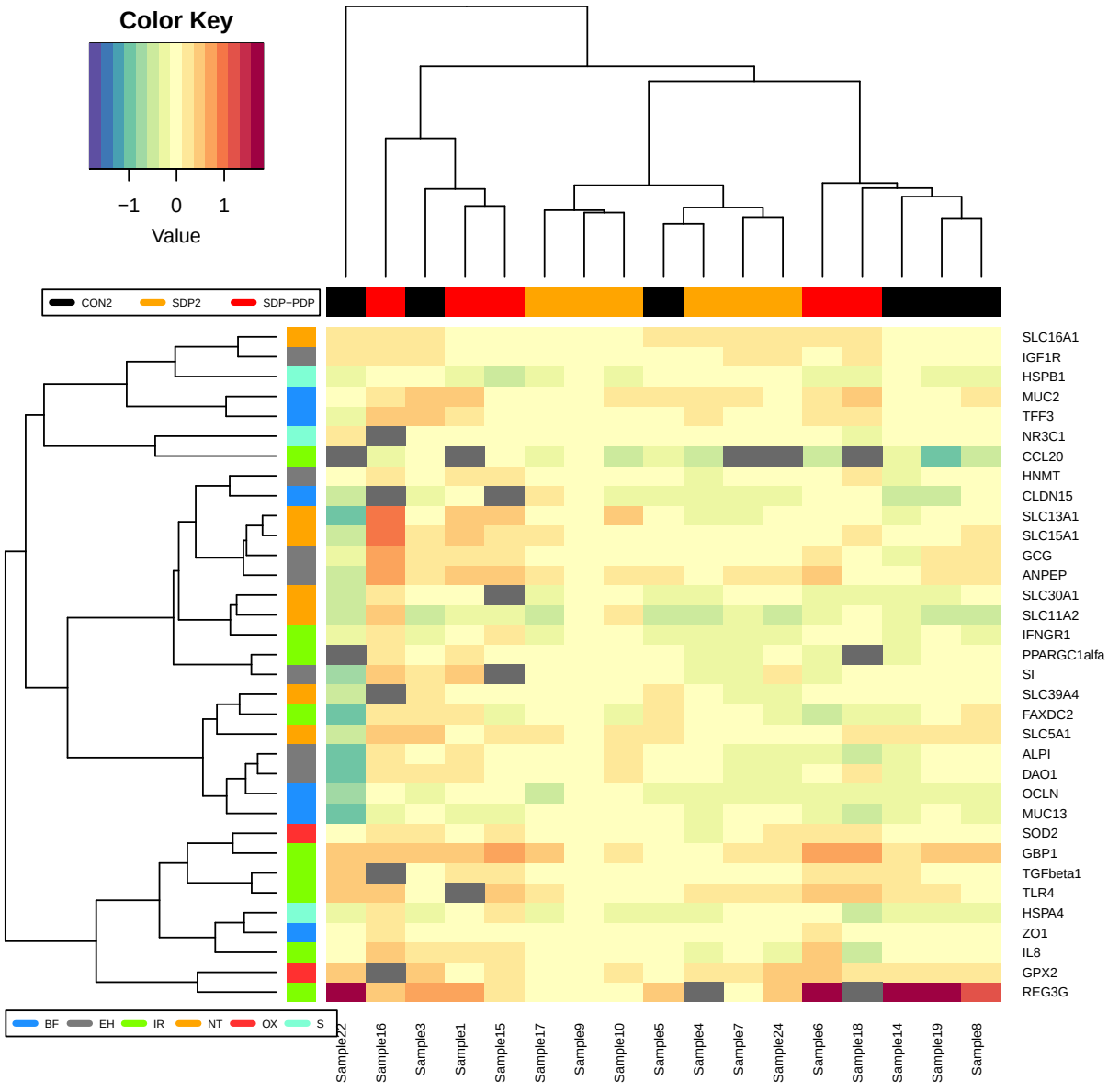
Table 1.4: P-valors de l'ANOVA entre tractaments pel teixit **Ileum**. Es mostren tots els gens, sigui significaius o no. Les files buides corresponen als gens que tenen valors perduts.

1.3 Heatmap

La distància entre genss depèn del coeficient de correlació, concretament $d = \frac{1}{2}(1-r)$. Això és fonamental per no comparar valors de nivells d'expressió entre gens sinó la correlació entre aquests nivells. 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.

1.3.1 Jejunum

La clau de colors correspon als nivells d'expressió. Els valors NA destacats de color gris fosc.



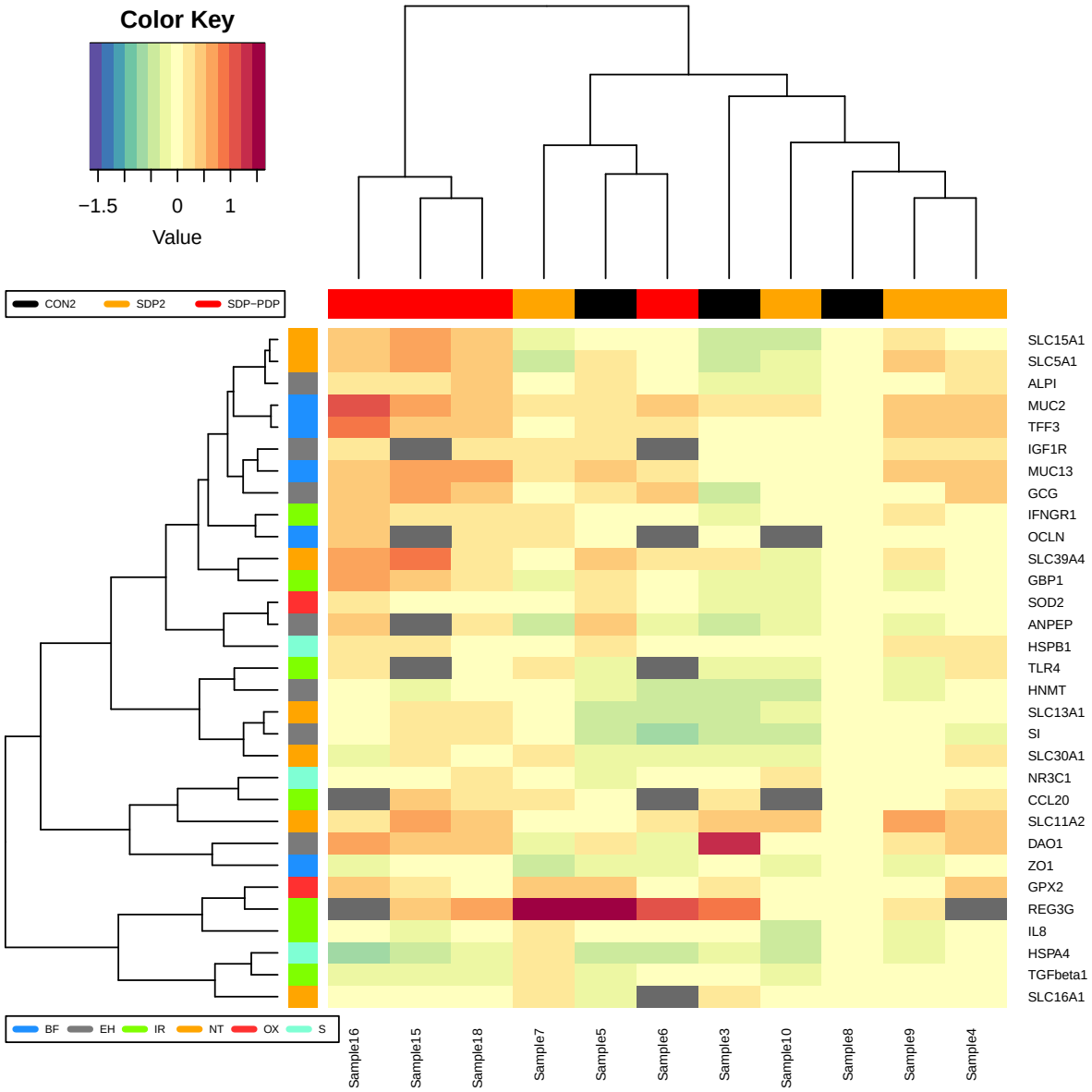
Al dendrograma de l'esquerra de la imatge hi ha els clústers dels gens, els 6 colors corresponen a les 6 funcionalitats gèniques. No es veu cap agrupació clara per funcionalitat (els colors del dendrograma estan força barrejats). Hi ha un gen outlier REG3G amb nivells d'expressió més alts en algunes 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 les dues mostres singulars (22 i 16), 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.

1.3.2 Ilenum

La clau de colors correspon als nivells d'expressió. Els valors NA destacats de color gris fosc.



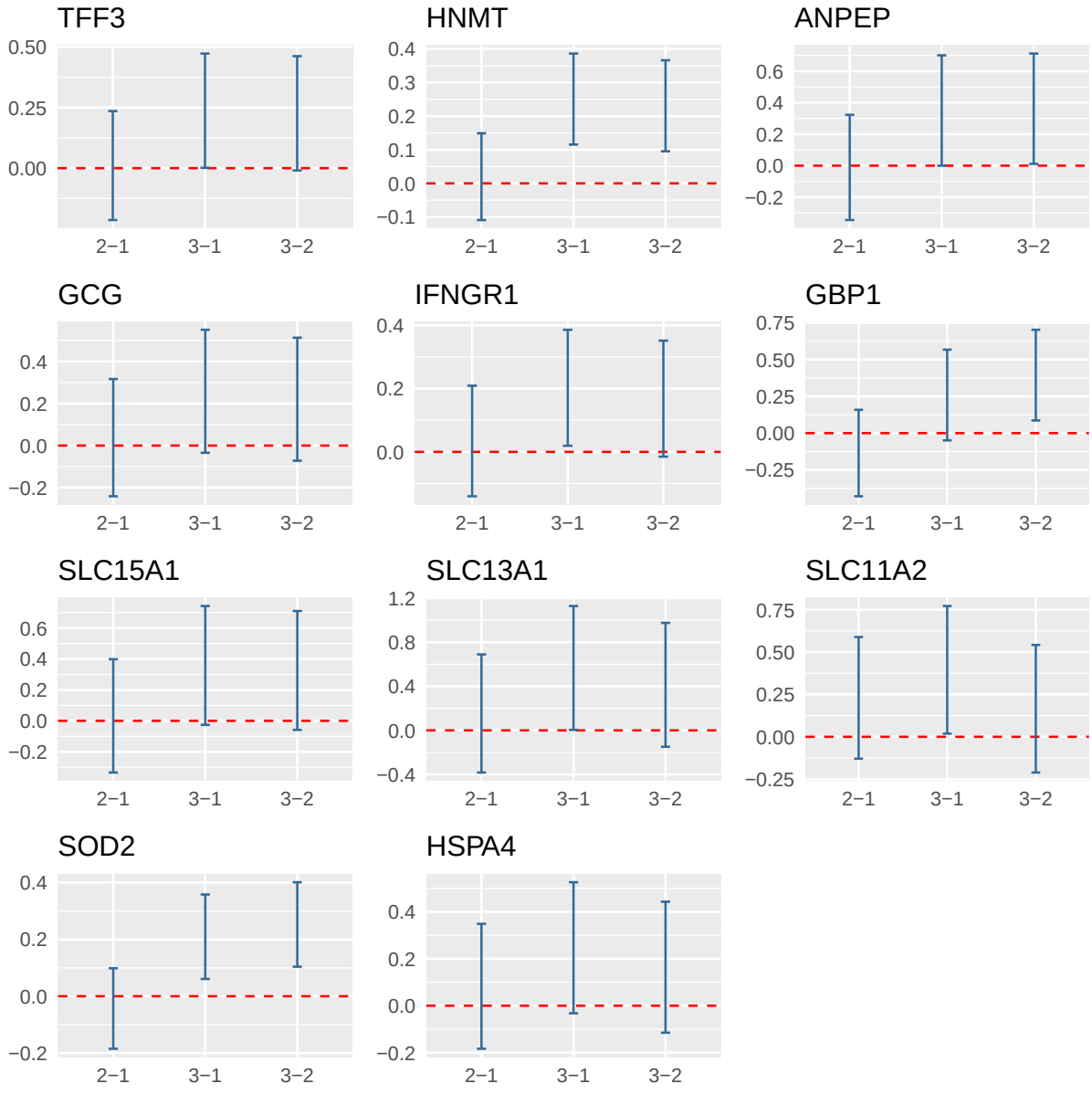
1.4 Tukey

1.4.1 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
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.IFNGR1	-0.175	0.125	-0.140	0.106	0.028	0.115	0.865	0.030	0.074
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 1.5: Tukey:comparacions múltiples Gene-Tractament



Nota: Veiem que les diferències significatives a la majoria de gens són entre SDP-PDP i els altres

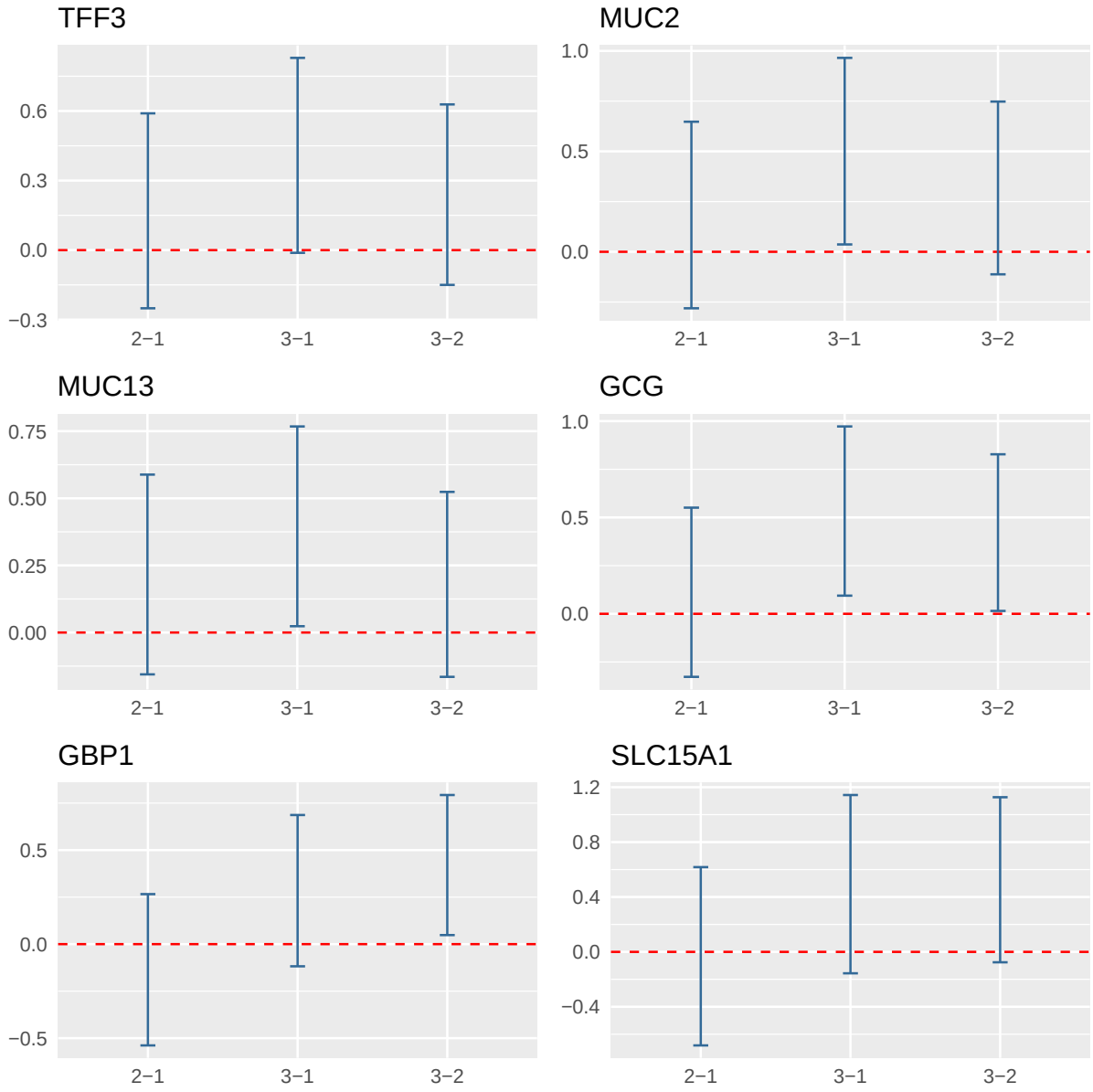
tractaments (un d'ells o ambdós), mai no hi ha diferències significatives entre CON2 i SDP2.

1.4.2 Ilenum

Només s'han comparat dos a dos els gens amb diferències significatives a l'ANOVA de la Table 1.2.

	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.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 1.6: Tukey: comparacions múltiples Gene-Tractament

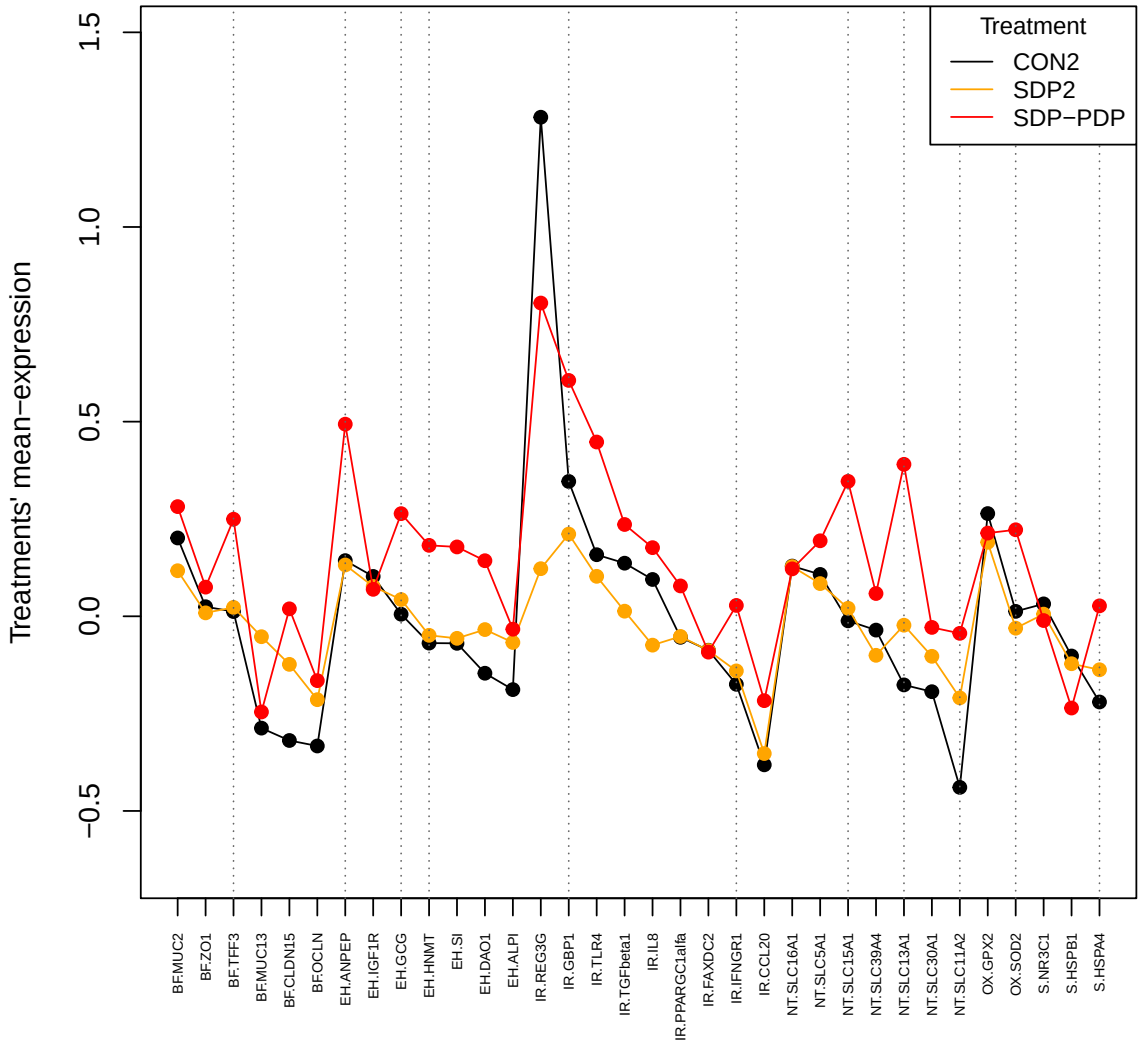


Nota: Com al Jejú, veiem diferències significatives a la majoria de gens entre SDP-PDP i els altres tractaments (un d'ells o ambdós).

1.5 Línies de mitjanes dels gens, per tractament

1.5.1 Jejunum

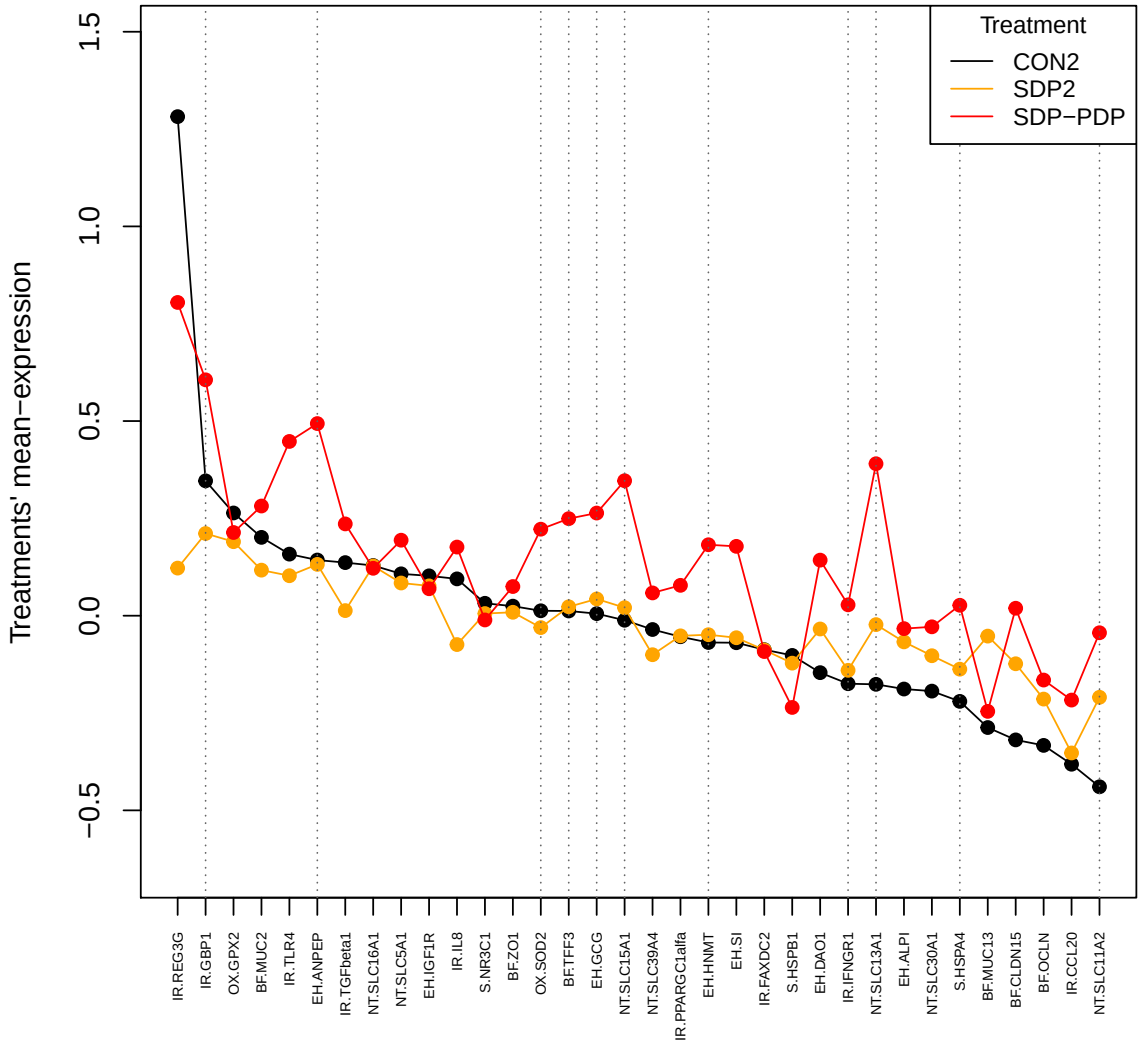
Ordenat per família gènica primer i dins de la família per T1 decreixent



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.

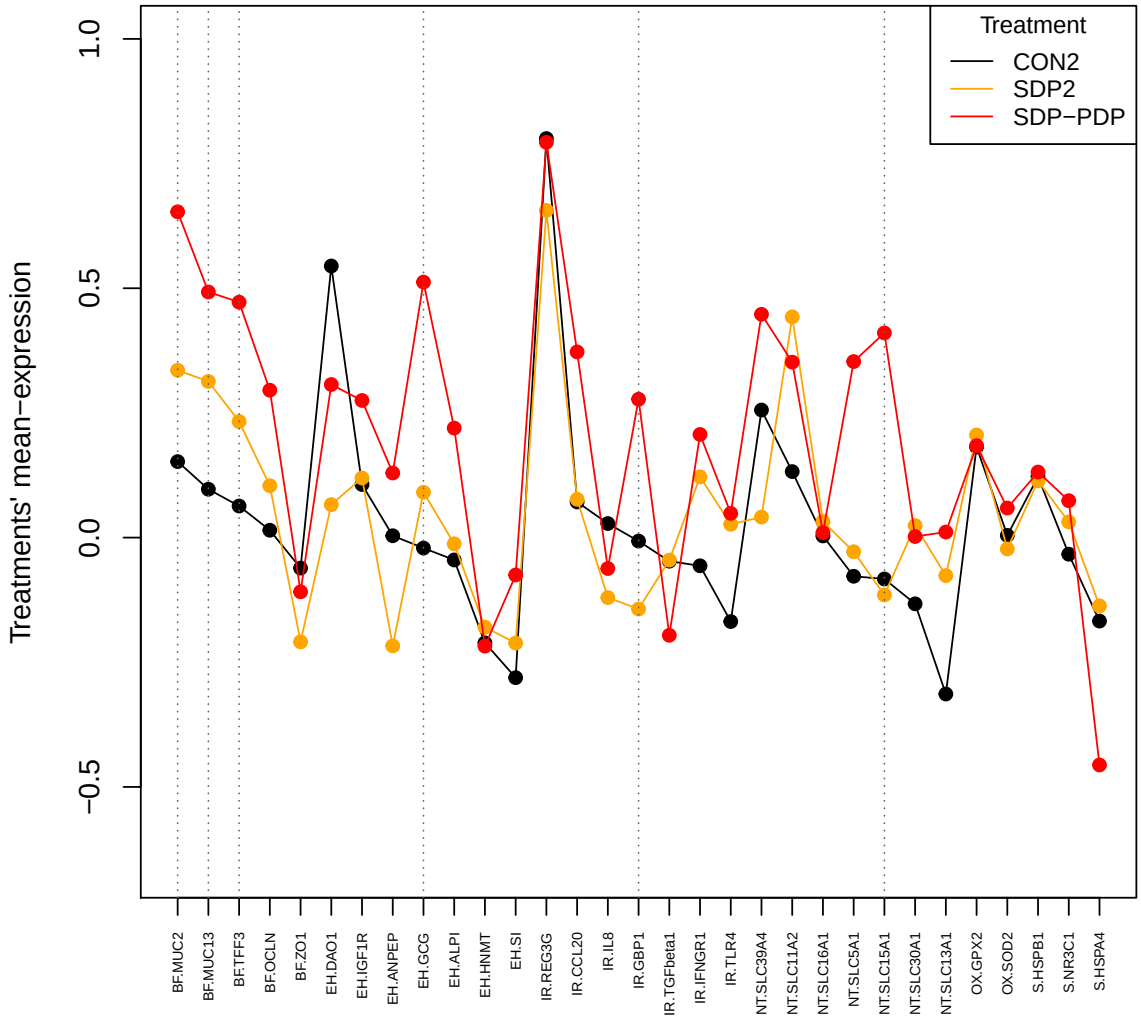
Ordenat per ordre decreixent d'expressió en el tractament T1



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.

1.5.2 Ilenum

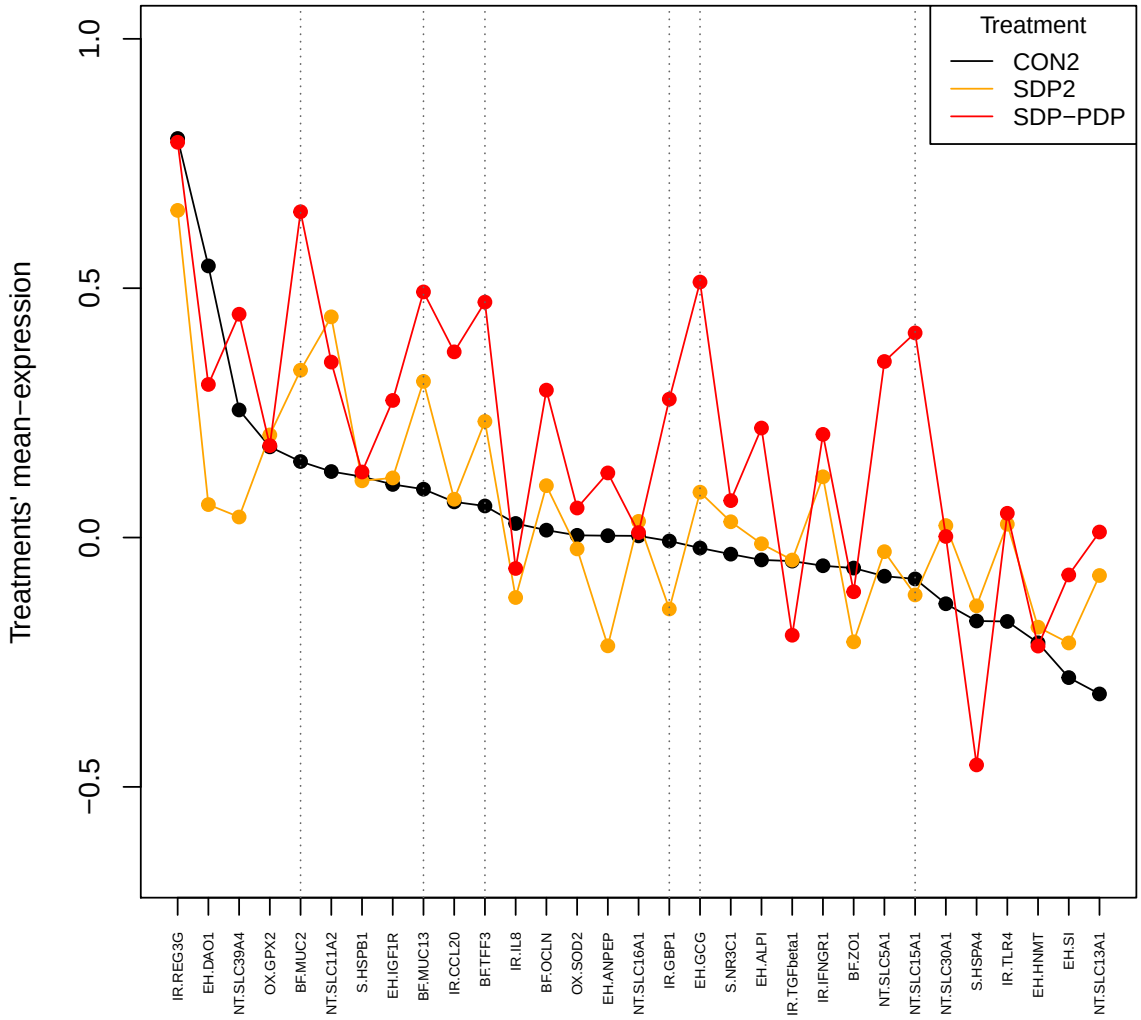
Ordenat per família gènica primer i dins de la família per CON2 decreixent



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 !

A l'Ili el tractament SDP-PDP té valors més elevats en mitjana en els gens que presenten diferències significatives.

Ordenat per ordre decreixent d'expressió en el tractament T1



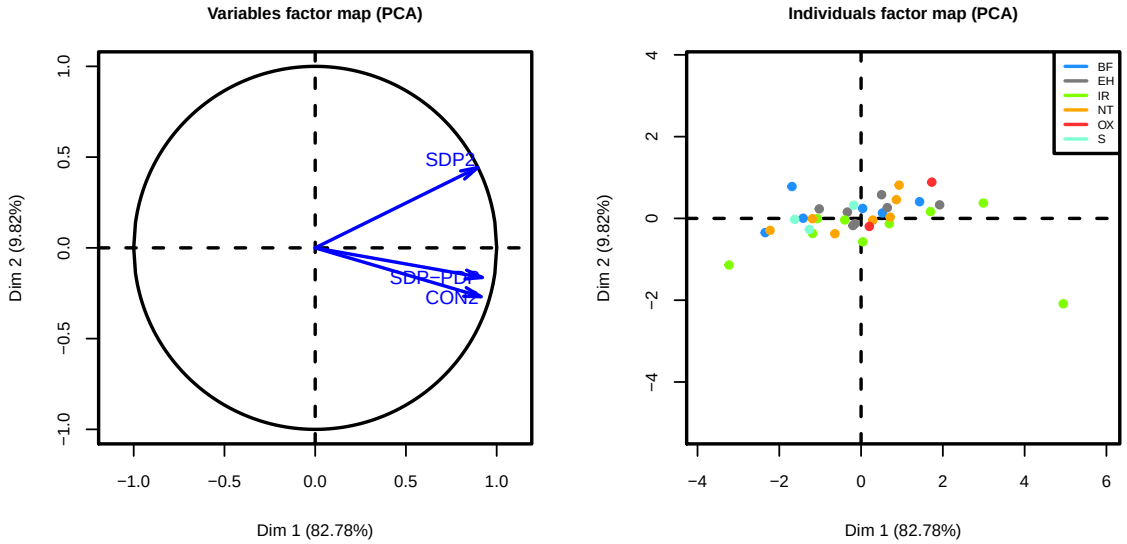
Mateixos comentaris que al gràfic anterior.

1.6 Components principals

1.6.1 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



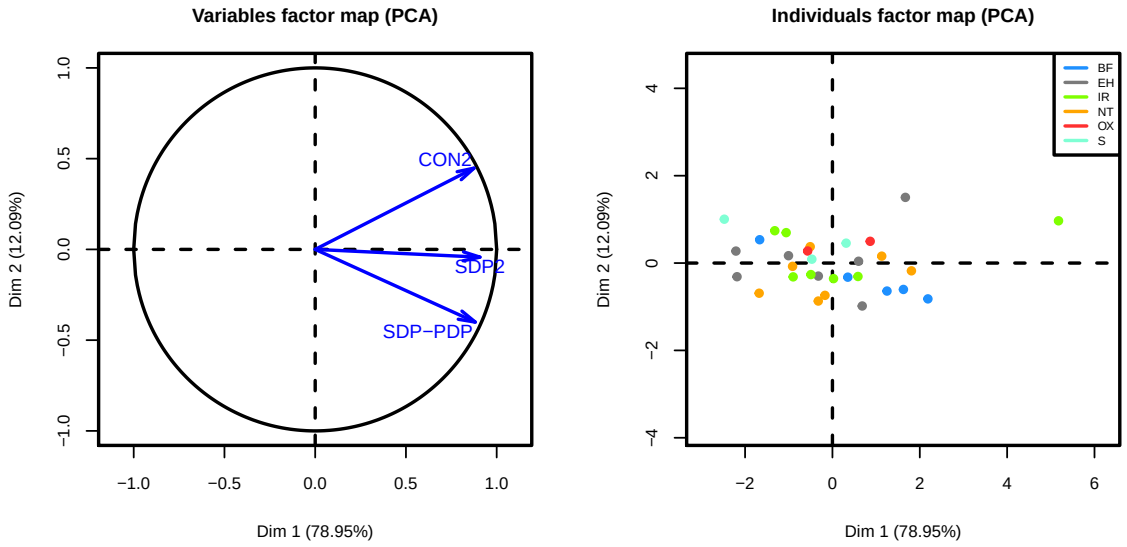
Amb dues components s'explica el 92.6% (82.78+9.82) de la variabilitat mitjana dels tractaments, en el cas del Jejú.

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. de l'apartat 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 correlacionades, però CON2 i SDP-PDP tenen una correlació més elevada. Això no implica similitud sinó correlació, poden ser (i són) significativament diferents i estar positivament correlacionades.

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 i, per tant, no smbla que els valors mitjans alts o baixos tinguin cap associació en la funció del gen.

Ilenum: Variables=mitjanes de tractaments, casos=gens

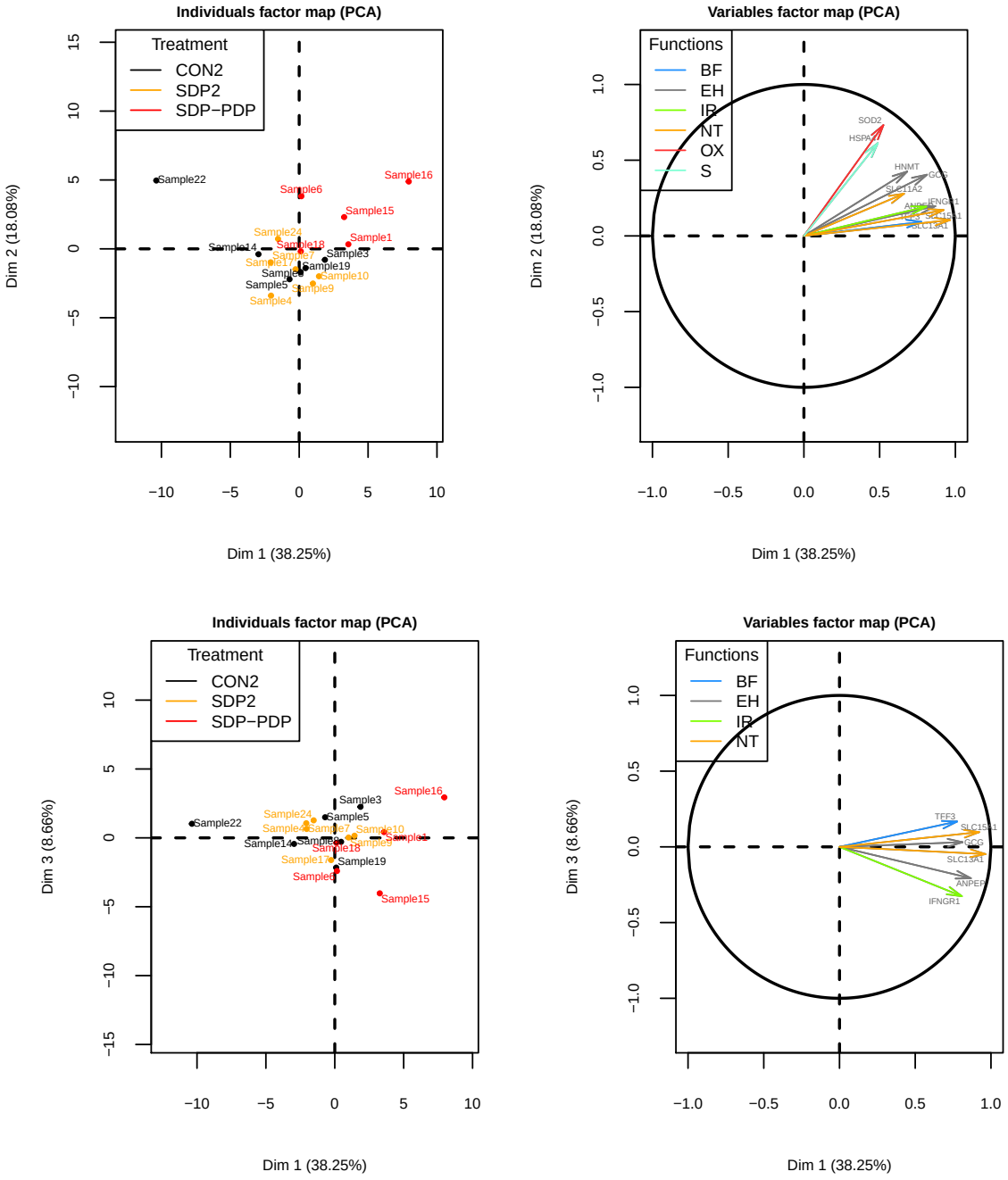


Amb dues components s'explica el 91.85% (78.95+12.90) de la variabilitat mitjana dels tractaments, en el cas de l'Ili. La resta d'interpretacions és anàloga.

1.6.2 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.

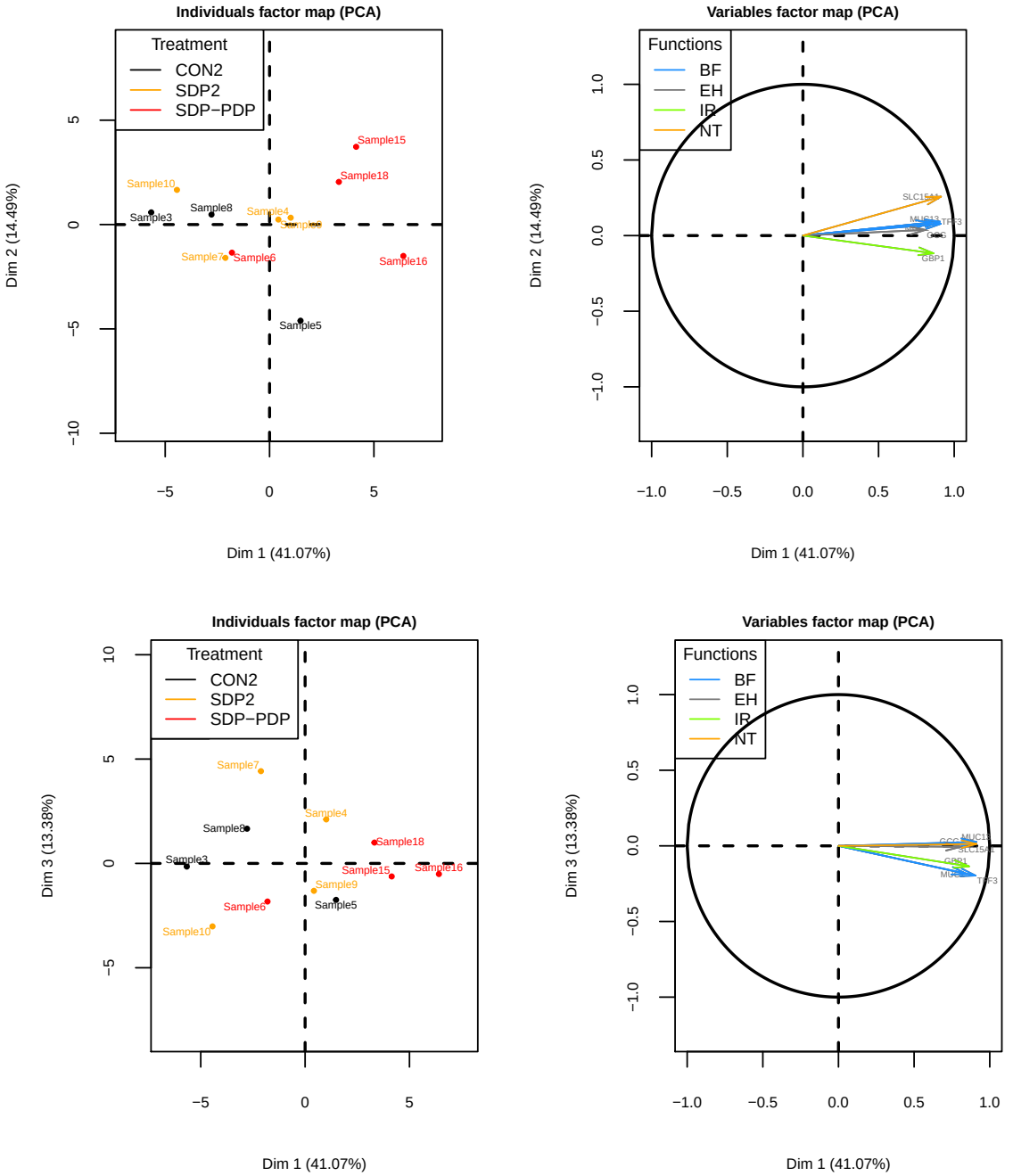
Jejunum: Variables=gens, casos=mostres



Les dues imatges superiors corresponene a les dues primeres components, les dues ingeriors a les components 1 i 3. La variabilitat explicada per les dues primeres components és del 56.33% (38.25+18.08), la

tercera component explica un 8.66%. En les dues primeres components, totes les mostres de SDP-PDP estan al primer quadrant (o quasi), indicant valors d'expressió més elevada en tots els gens significatius i amb bona qualitat de repreentació perquè 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.

Ilenum: 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).