CGS4144 Bioinformatics - Assignment 2

By: Liron Naccache, Seggev Haimovich, Jayden Johnson, and William Manuel

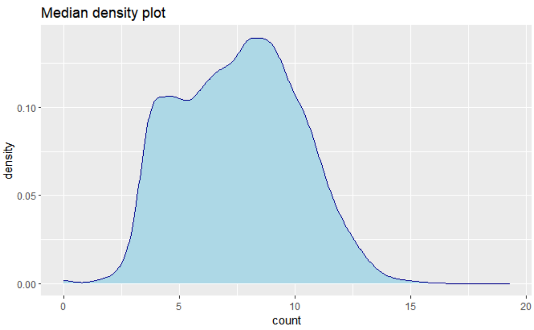
Github

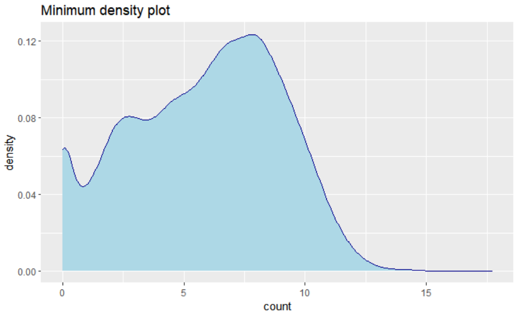
1) **Expression matrix size**: 15929 rows and 69 columns,

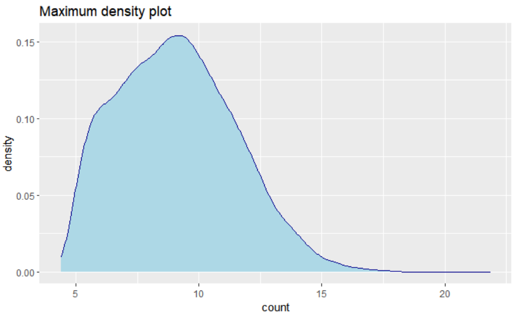
   The number of genes: Each row represents a gene, meaning there are 15929 genes.

   Variations in the data: We observed that there are some genes with more counts than others, however, there aren't many genes with significantly high counts.

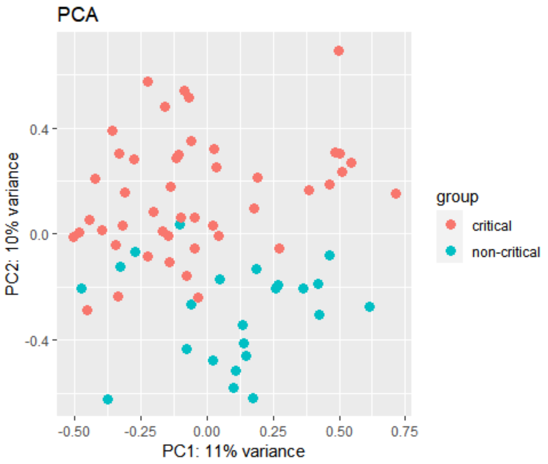
   We calculated for each gene the median, minimum, and maximum values of counts and produced the following density plots.





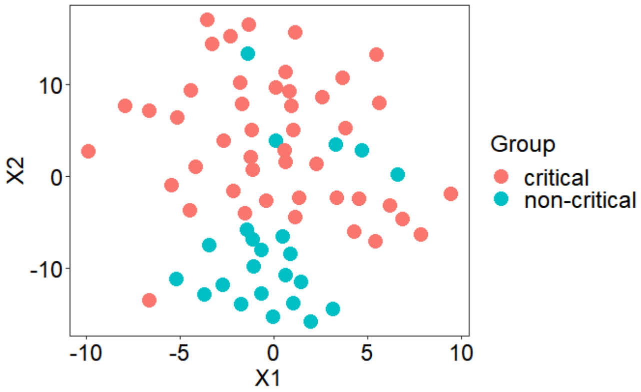


2) In this section, we used dimensional reduction algorithms to check if there's a signal in our data. As mentioned above, each sample has more than 15,000 features (aka genes) meaning it's very hard to learn something from the raw data. That's why we reduced the data dimensions to 2. Using PCA we got the following result:



From the graph, we can infer that the main components that describe our two groups are different.

Using T-SNE we got the graph below:

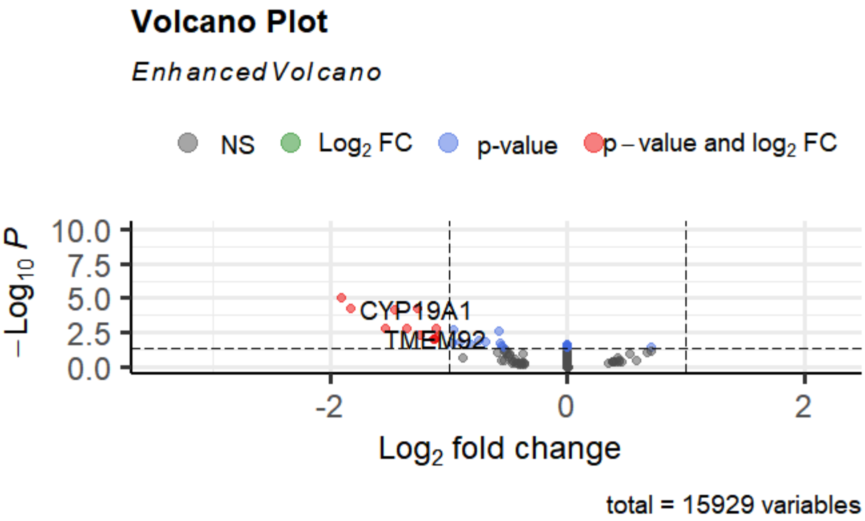


Here, the separation is less clear but still exists.

We see that in both graphs the groups are almost linear separatable and in both graphs, the critical group's values are higher than the non-critical group's values.

Using both methods, we infer that there might be a signal in our data. Nevertheless, it's known that when reducing dimension one can lose data, so it might be a false signal due to a reduction that was too radical and destroyed it.

3) In this section, we created a table of differentially expressed genes and plotted the volcano graph:



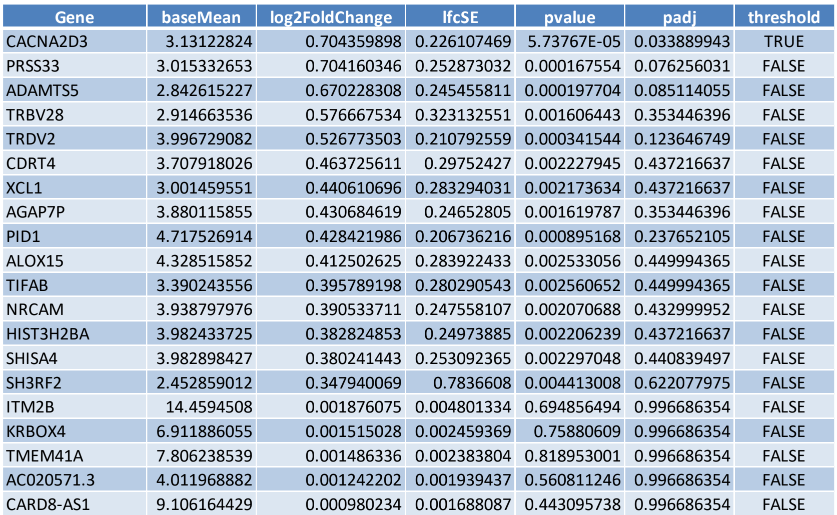
The y-axis represents the -log10 of the P-value: if a gene is statistically significant, its p-value will be close to zero, which means a higher position in the y-axis on the Volcano plot.
The x-axis represents the log2 of the fold change, meaning the ratio between the counts of a gene in the group of critical patients and the group of non-critical patients.

We can see that only 11 genes passed the thresholds.

Our threshold for log_2 fold change is the default threshold (1), we didn’t change it because it seems like a small change will add a small number of genes passing the threshold and a big change will cause too many genes to pass the threshold and make the results irrelevant.

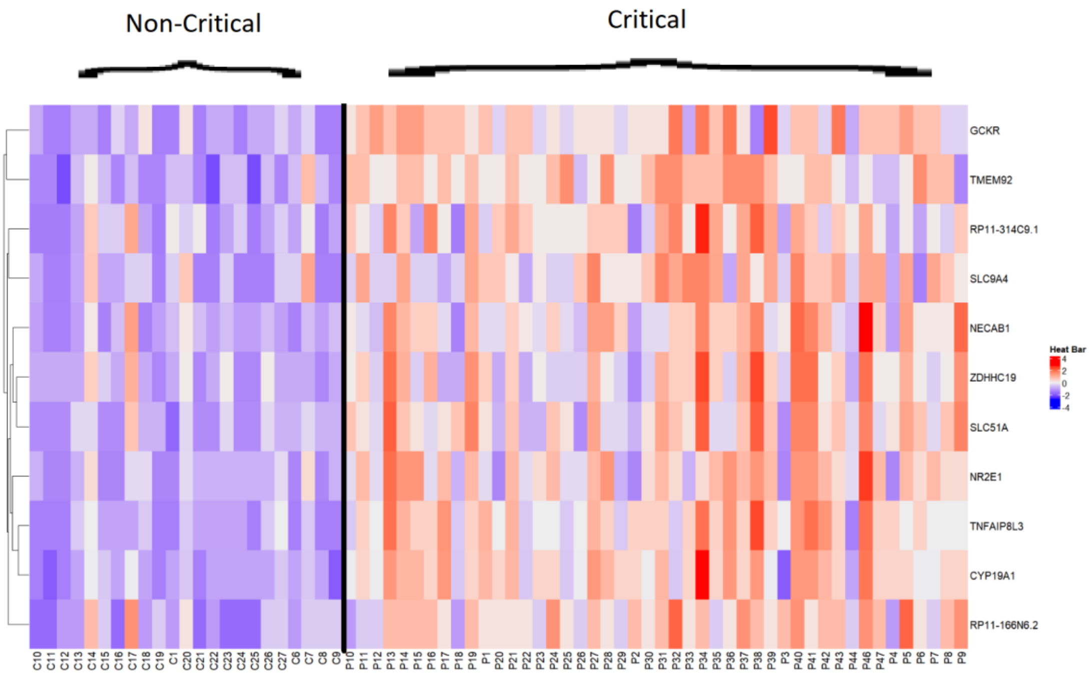
Our p-value threshold is “p<0.05”. We followed the tutorial in the exercise PDF, which indicated that due to the use of adjusted p-values we can loosen the default p-value threshold.

These are the first 20 genes in our table (the full table is on GitHub)



Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj	threshold
CACNA2D3	3.13122824	0.704359898	0.226107469	5.73767E-05	0.033889943	TRUE
PRSS33	3.015332653	0.704160346	0.252873032	0.000167554	0.076256031	FALSE
ADAMTS5	2.842615227	0.670228308	0.245455811	0.000197704	0.085114055	FALSE
TRBV28	2.914663536	0.576667534	0.323132551	0.001606443	0.353446396	FALSE
TRDV2	3.996729082	0.526773503	0.210792559	0.000341544	0.123646749	FALSE
CDRT4	3.707918026	0.463725611	0.29752427	0.002227945	0.437216637	FALSE
XCL1	3.001459551	0.440610696	0.283294031	0.002173634	0.437216637	FALSE
AGAP7P	3.880115855	0.430684619	0.24652805	0.001619787	0.353446396	FALSE
PID1	4.717526914	0.428421986	0.206736216	0.000895168	0.237652105	FALSE
ALOX15	4.328515852	0.412502625	0.283922433	0.002533056	0.449994365	FALSE
TIFAB	3.390243556	0.395789198	0.280290543	0.002560652	0.449994365	FALSE
NRCAM	3.938797976	0.390533711	0.247558107	0.002070688	0.432999952	FALSE
HIST3H2BA	3.982433725	0.382824853	0.24973885	0.002206239	0.437216637	FALSE
SHISA4	3.982898427	0.380241443	0.253092365	0.002297048	0.440839497	FALSE
SH3RF2	2.452859012	0.347940069	0.7836608	0.004413008	0.622077975	FALSE
ITM2B	14.4594508	0.001876075	0.004801334	0.694856494	0.996686354	FALSE
KRBOX4	6.911886055	0.001515028	0.002459369	0.75880609	0.996686354	FALSE
TMEM41A	7.806238539	0.001486336	0.002383804	0.818953001	0.996686354	FALSE
AC020571.3	4.011968882	0.001242202	0.001939437	0.560811246	0.996686354	FALSE
CARD8-AS1	9.106164429	0.000980234	0.001688087	0.443095738	0.996686354	FALSE

4) In this section, we extracted the 11 differentially expressed genes and created a heatmap:

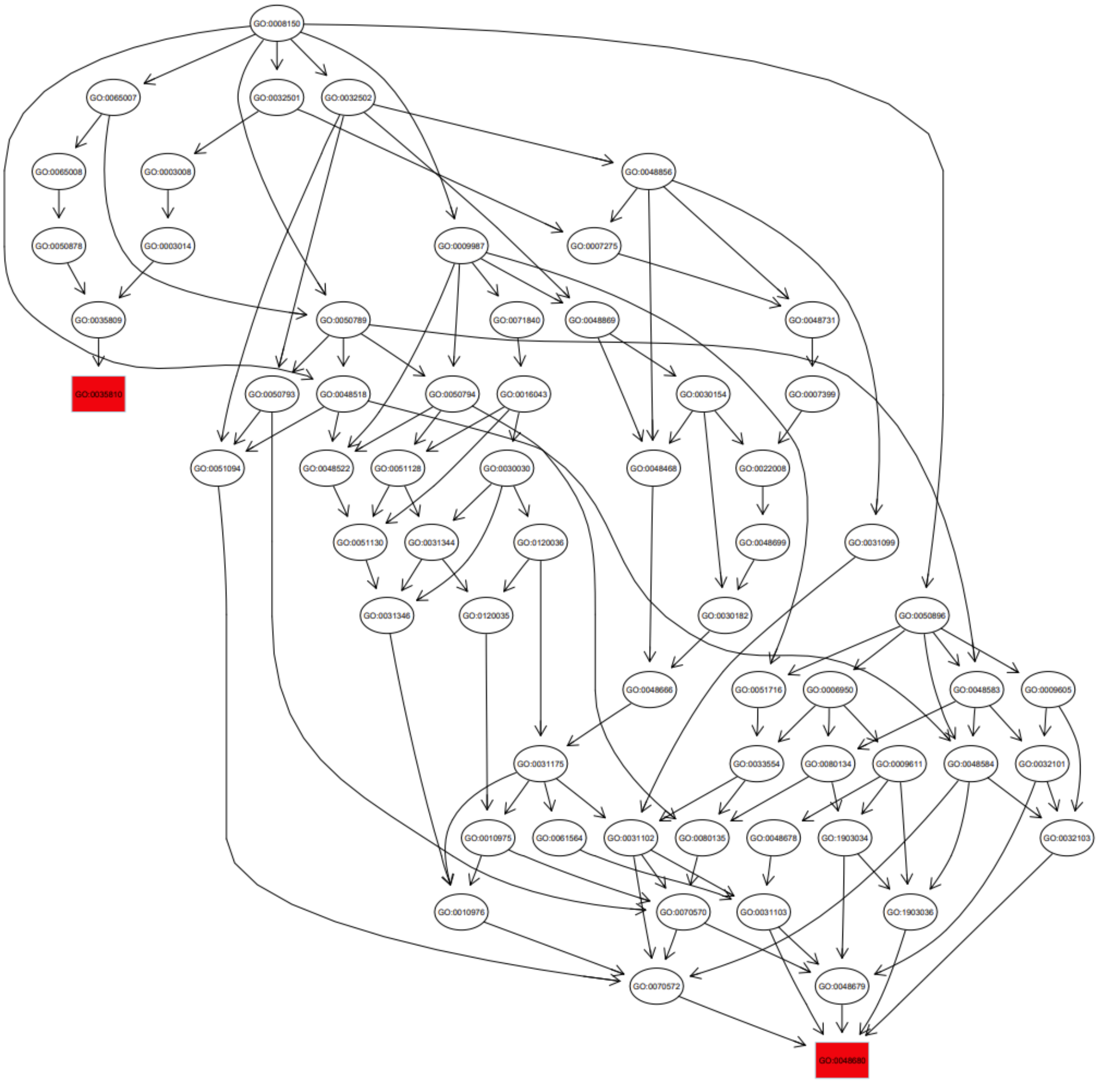


There aren't so many differentially expressed genes but in the little ones that we do have, the difference in the expression is very clear. The black line represents the separation between the two sample groups (critical and non-critical patients).

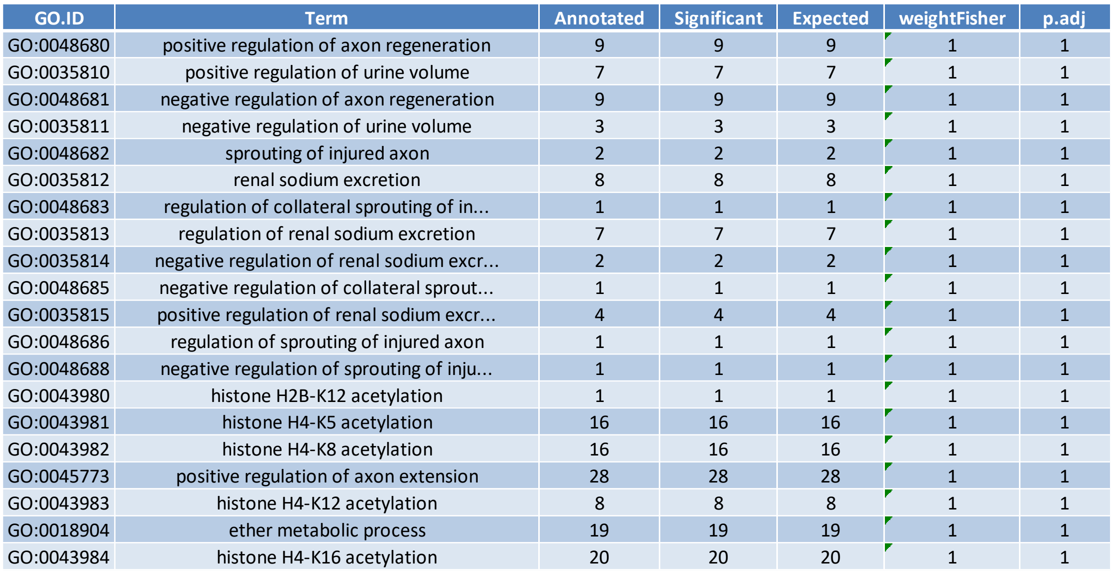
5+6) In this section, we ran an enrichment analysis with 4 different methods:

i. topGo

Using the topGO method, we generated the below GO graph with 63 nodes and 122 edges of GO terms. The two most significant terms are GO:0035810 and GO:0048680, which are for "positive regulation of urine volume" and "positive regulation of axon regeneration" respectively. This is interesting as there have been some studies linking organ damage from COVID-19 to dissemination through peripheral nerves as well as kidney damage affecting the individual's urine.



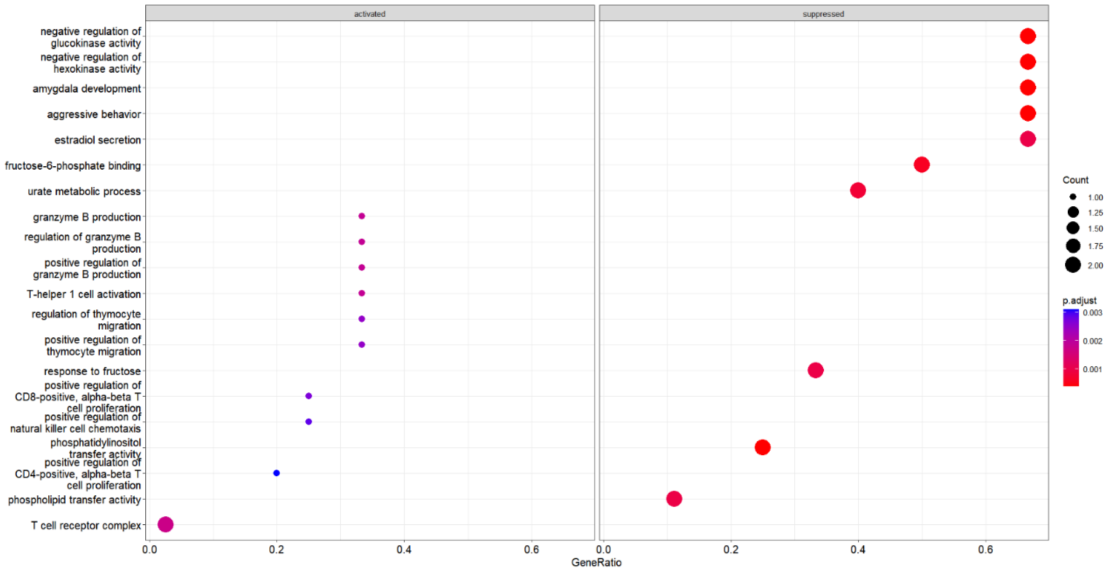
In the table below, we can see the significant GO terms as well as their level of significance, with 9 being the most significant and 1 the least significant.



GO.ID	Term	Annotated	Significant	Expected	weightFisher	p.adj
GO:0048680	positive regulation of axon regeneration	9	9	9	1	1
GO:0035810	positive regulation of urine volume	7	7	7	1	1
GO:0048681	negative regulation of axon regeneration	9	9	9	1	1
GO:0035811	negative regulation of urine volume	3	3	3	1	1
GO:0048682	sprouting of injured axon	2	2	2	1	1
GO:0035812	renal sodium excretion	8	8	8	1	1
GO:0048683	regulation of collateral sprouting of in...	1	1	1	1	1
GO:0035813	regulation of renal sodium excretion	7	7	7	1	1
GO:0035814	negative regulation of renal sodium excr...	2	2	2	1	1
GO:0048685	negative regulation of collateral sprout...	1	1	1	1	1
GO:0035815	positive regulation of renal sodium excr...	4	4	4	1	1
GO:0048686	regulation of sprouting of injured axon	1	1	1	1	1
GO:0048688	negative regulation of sprouting of inju...	1	1	1	1	1
GO:0043980	histone H2B-K12 acetylation	1	1	1	1	1
GO:0043981	histone H4-K5 acetylation	16	16	16	1	1
GO:0043982	histone H4-K8 acetylation	16	16	16	1	1
GO:0045773	positive regulation of axon extension	28	28	28	1	1
GO:0043983	histone H4-K12 acetylation	8	8	8	1	1
GO:0018904	ether metabolic process	19	19	19	1	1
GO:0043984	histone H4-K16 acetylation	20	20	20	1	1

ii. <u>clusterProfiler</u>

By looking at the graph below, we see that some of our genes relate to families that have a connection with the Immune System and are activated and that some genes that have a connection with the Metabolism Process that is suppressed. Since our database relates to Covid-19, it's not surprising to see that it has something to do with the Immune System.

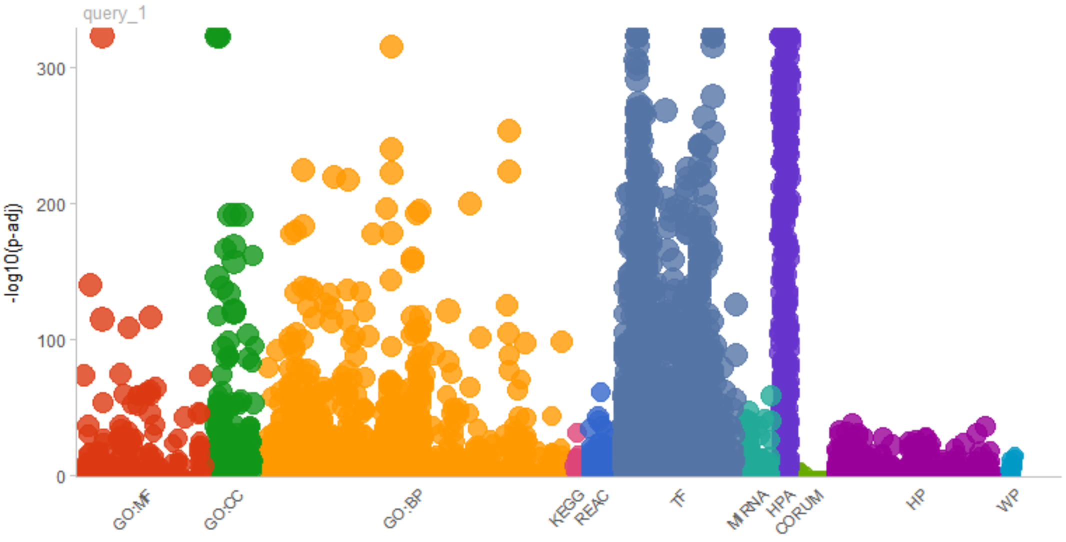


The table below represents the different families, sorted by p-value (The full table is on GitHub). We can see for each family the Ontology it belongs to, the official ID, and more features like set size and qvalue.

ONTOLOGY	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank	leading_edge	core_enrichment
MF	GO:0008526	phosphatidylinositol transfer activity	8	-0.999985587	-1.76206721	0.000383877	0.0003839	0.618007265	2	tags=25%, list=0%, signal=25%	OSBPL2/TNFAIP8L3
BP	GO:0002118	aggressive behavior	3	-0.999865829	-1.572135916	0.000387973	0.000388	0.618007265	4	tags=67%, list=0%, signal=67%	AVPR1A/NR2E1
BP	GO:0021764	amygdala development	3	-0.999870457	-1.572143193	0.000387973	0.000388	0.618007265	4	tags=67%, list=0%, signal=67%	NF1/NR2E1
BP	GO:0033132	negative regulation of glucokinase activity	3	-0.999932266	-1.572240379	0.000387973	0.000388	0.618007265	3	tags=67%, list=0%, signal=67%	MIDN/GCKR
BP	GO:1903300	negative regulation of hexokinase activity	3	-0.999932266	-1.572240379	0.000387973	0.000388	0.618007265	3	tags=67%, list=0%, signal=67%	MIDN/GCKR
MF	GO:0070095	fructose-6-phosphate binding	4	-0.999933646	-1.637718098	0.000584795	0.0005848	0.618007265	3	tags=50%, list=0%, signal=50%	PFKM/GCKR
BP	GO:0046415	urate metabolic process	5	-0.999927105	-1.684769629	0.000777152	0.0007772	0.618007265	3	tags=40%, list=0%, signal=40%	PNP/GCKR
MF	GO:0120014	phospholipid transfer activity	18	-0.999933152	-1.825275706	0.000954745	0.0009547	0.618007265	2	tags=11%, list=0%, signal=11%	PLEKHA8P1/TNFAIP8L3
BP	GO:0009750	response to fructose	6	-0.999883795	-1.720625197	0.000959325	0.0009593	0.618007265	3	tags=33%, list=0%, signal=33%	PTGS2/GCKR
BP	GO:0035938	estradiol secretion	3	-0.999680751	-1.57184491	0.000969932	0.0009699	0.618007265	7	tags=67%, list=0%, signal=67%	SPP1/CYP19A1
BP	GO:2000864	regulation of estradiol secretion	3	-0.999680751	-1.57184491	0.000969932	0.0009699	0.618007265	7	tags=67%, list=0%, signal=67%	SPP1/CYP19A1
BP	GO:0021960	anterior commissure morphogenesis	5	-0.999864869	-1.684664768	0.00097144	0.0009714	0.618007265	4	tags=40%, list=0%, signal=40%	FBXO45/NR2E1
BP	GO:0040034	regulation of development, heterochronic	4	-0.999848788	-1.637579115	0.000974659	0.0009747	0.618007265	4	tags=50%, list=0%, signal=50%	RBPJ/NR2E1
BP	GO:0048505	regulation of timing of cell differentiation	4	-0.999848788	-1.637579115	0.000974659	0.0009747	0.618007265	4	tags=50%, list=0%, signal=50%	RBPJ/NR2E1
BP	GO:0033131	regulation of glucokinase activity	7	-0.999915481	-1.742771823	0.001153624	0.0011536	0.618007265	3	tags=29%, list=0%, signal=29%	PFKFB2/GCKR
BP	GO:0002677	negative regulation of chronic inflammatory response	3	-0.999663997	-1.571818567	0.001163919	0.0011639	0.618007265	7	tags=67%, list=0%, signal=67%	IL10/CYP19A1
BP	GO:0048712	negative regulation of astrocyte differentiation	5	-0.99985556	-1.684649084	0.001165728	0.0011657	0.618007265	4	tags=40%, list=0%, signal=40%	LDLR/NR2E1
BP	GO:1903299	regulation of hexokinase activity	8	-0.999911206	-1.761936142	0.001535509	0.0015355	0.618007265	3	tags=25%, list=0%, signal=25%	PFKFB2/GCKR
BP	GO:0006710	androgen catabolic process	4	-0.999673332	-1.637291749	0.001559454	0.0015595	0.618007265	7	tags=50%, list=0%, signal=50%	HSD17B11/CYP19A1
CC	GO:0042101	T cell receptor complex	79	0.999094994	1.985460568	0.001719057	0.0017191	0.618007265	5	tags=3%, list=0%, signal=3%	TRBV28/TRDV2

iii. gProfiler2

Using this graph, we wanted to see if we can connect our genes with specific families of genes, for further exploration. The y-axis is like in the volcano plot, meaning the higher the point is, the gene it represents is more statistically significant. The x-axis represents the different ontologies and pathways. We can infer that many of our statistically significant genes relate to the TF and HPA pathways.

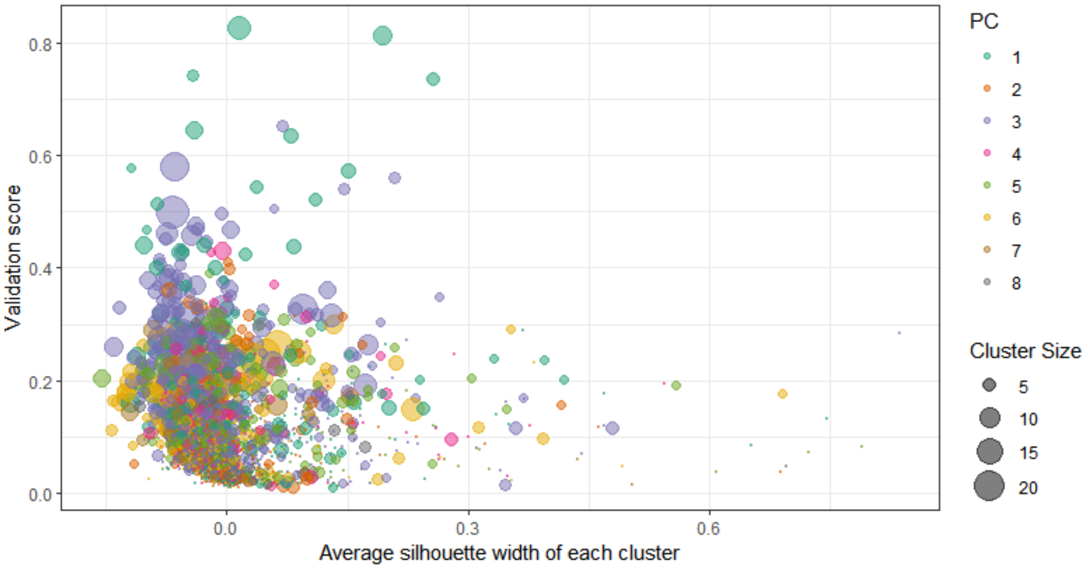


In the table below we can see all the terms and their relevant values such as p-value and source.

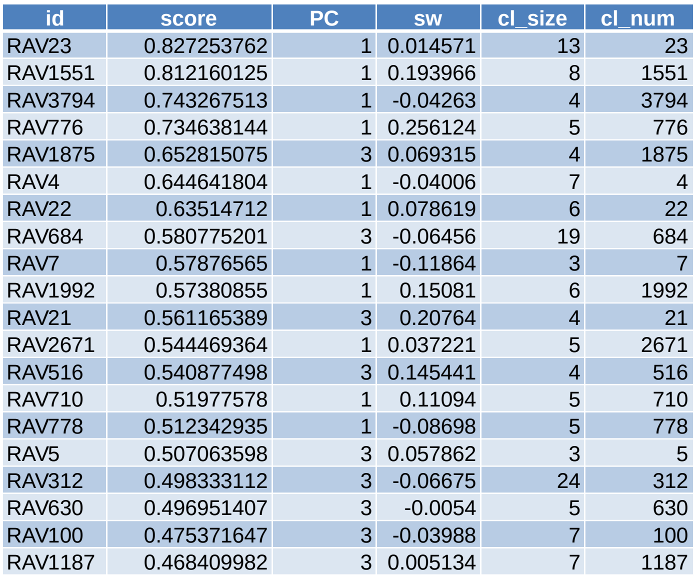
term_id	significant	p_value	term_size	query_size	intersection_size	precision	recall	source	term_name	effective domain size	source_order
CORUM:306	TRUE	2.08159E-05	80	3109	80	0.025731747	1	CORUM	Ribosome, cytoplasmic	3627	183
CORUM:320	TRUE	0.000408862	78	3109	77	0.024766806	0.987179	CORUM	55S ribosome, mitochondrial	3627	191
CORUM:351	TRUE	0.00044113	141	3109	135	0.043422322	0.957447	CORUM	Spliceosome	3627	205
CORUM:324	TRUE	0.003172443	48	3109	48	0.015439048	1	CORUM	39S ribosomal subunit, mitochondrial	3627	193
CORUM:308	TRUE	0.003709126	47	3109	47	0.015117401	1	CORUM	60S ribosomal subunit, cytoplasmic	3627	184
CORUM:305	TRUE	0.032916437	33	3109	33	0.010614345	1	CORUM	40S ribosomal subunit, cytoplasmic	3627	182
CORUM:230	TRUE	0.038457781	32	3109	32	0.010292699	1	CORUM	Mediator complex	3627	124
CORUM:1181	TRUE	0.04291746	79	3109	75	0.024123512	0.949367	CORUM	C complex spliceosome	3627	659
CORUM:338	TRUE	0.044929883	31	3109	31	0.009971052	1	CORUM	40S ribosomal subunit, cytoplasmic	3627	199
GO:0044260	TRUE	0	5766	11376	4317	0.379483122	0.748699	GO:BP	cellular macromolecule metabolic process	21100	12699
GO:1901564	TRUE	2.322E-254	6446	11376	4603	0.404623769	0.714086	GO:BP	organonitrogen compound metabolic process	21100	24853
GO:0044267	TRUE	1.2258E-240	4882	11376	3630	0.319092827	0.743548	GO:BP	cellular protein metabolic process	21100	12703
GO:0009058	TRUE	1.4099E-225	5861	11376	4192	0.368495077	0.715236	GO:BP	biosynthetic process	21100	3661
GO:1901576	TRUE	3.8792E-225	5774	11376	4139	0.363836146	0.716834	GO:BP	organic substance biosynthetic process	21100	24864
GO:0044249	TRUE	1.7662E-223	5700	11376	4091	0.359616737	0.717719	GO:BP	cellular biosynthetic process	21100	12692
GO:0019538	TRUE	2.6333E-220	5487	11376	3956	0.347749648	0.720977	GO:BP	protein metabolic process	21100	6769
GO:0031323	TRUE	6.0046E-218	6004	11376	4260	0.374472574	0.709527	GO:BP	regulation of cellular metabolic process	21100	8255
GO:0080090	TRUE	9.3476E-201	5812	11376	4106	0.360935302	0.706469	GO:BP	regulation of primary metabolic process	21100	20855
GO:0043412	TRUE	7.6906E-197	3936	11376	2954	0.25966948	0.750508	GO:BP	macromolecule modification	21100	12301
GO:0051171	TRUE	2.8434E-195	5654	11376	3999	0.351529536	0.707287	GO:BP	regulation of nitrogen compound metabolic process	21100	15640

iv. GenomicSuperSignature

In this method, we used 8 principal components from our data and tried to cluster them. The x-axis represents the silhouette width of each cluster, which is a way of understanding how close the points of each cluster are – as being close to 0 is the best. The y-axis represents the Validation score, which is a way of understanding if our prediction is good – being close to 1 is the best. With that being said, we only have a few clusters with small silhouette widths and high validation scores. We think we can not infer much from this graph.



Below, we can see the first 20 rows of the data table used to create the graph above (The full chart is on GitHub). Each line represents a cluster and has information like cluster size and the PC it relates to.



id	score	PC	sw	cl_size	cl_num
RAV23	0.827253762	1	0.014571	13	23
RAV1551	0.812160125	1	0.193966	8	1551
RAV3794	0.743267513	1	-0.04263	4	3794
RAV776	0.734638144	1	0.256124	5	776
RAV1875	0.652815075	3	0.069315	4	1875
RAV4	0.644641804	1	-0.04006	7	4
RAV22	0.63514712	1	0.078619	6	22
RAV684	0.580775201	3	-0.06456	19	684
RAV7	0.57876565	1	-0.11864	3	7
RAV1992	0.57380855	1	0.15081	6	1992
RAV21	0.561165389	3	0.20764	4	21
RAV2671	0.544469364	1	0.037221	5	2671
RAV516	0.540877498	3	0.145441	4	516
RAV710	0.51977578	1	0.11094	5	710
RAV778	0.512342935	1	-0.08698	5	778
RAV5	0.507063598	3	0.057862	3	5
RAV312	0.498333112	3	-0.06675	24	312
RAV630	0.496951407	3	-0.0054	5	630
RAV100	0.475371647	3	-0.03988	7	100
RAV1187	0.468409982	3	0.005134	7	1187