I hate coming up with titles, please don't make me do this

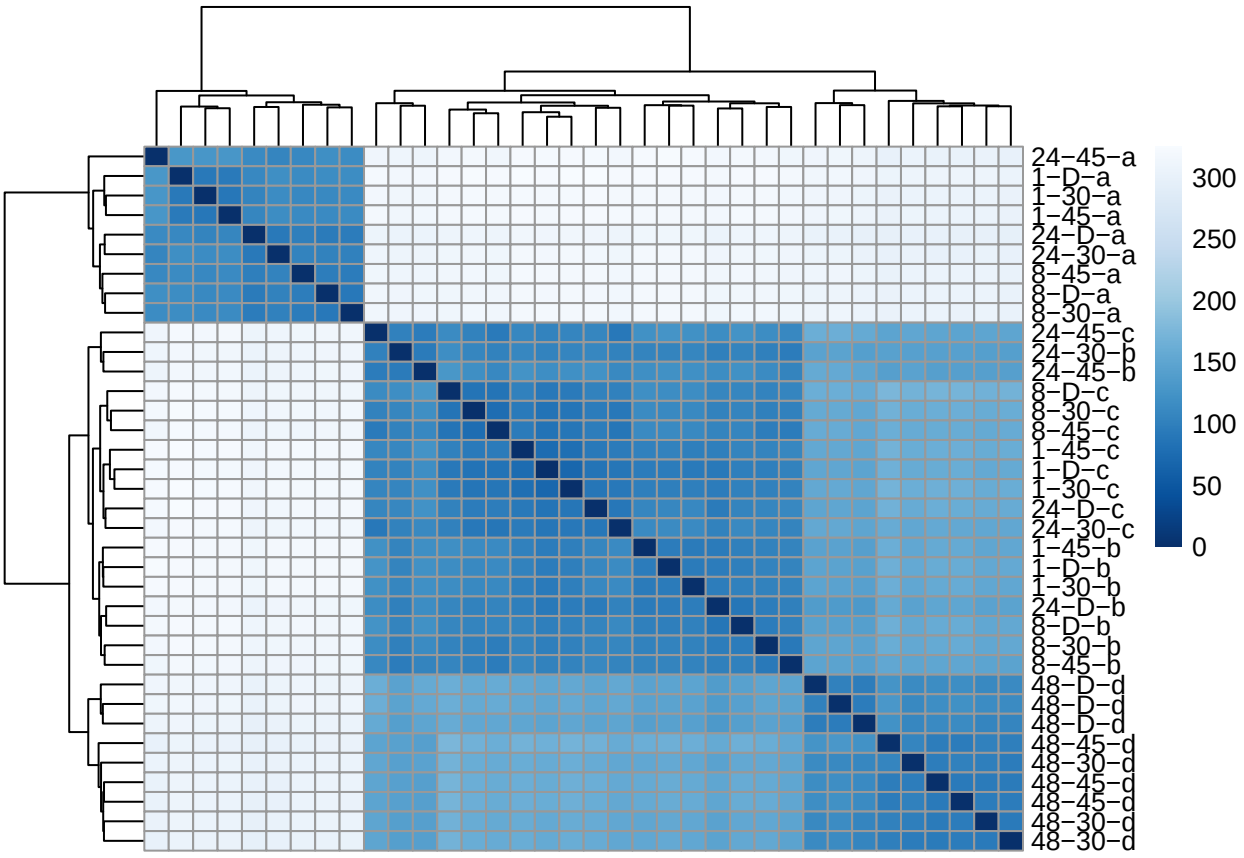
But it's the analysis of RNA-Seq of OVCAR3 cells in the presence of atovaquone up to 48 hours

AJ Fagan

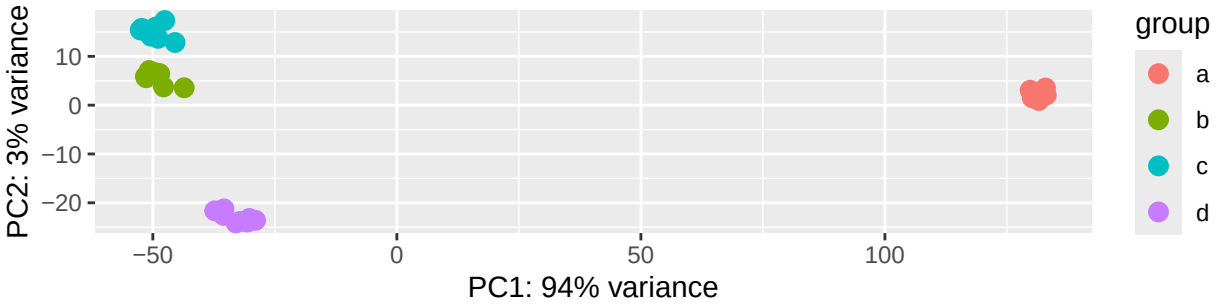
12 Sept, 2024

Data visualization

First, we plot a hierarchical clustering of the data according to the raw estimated number of counts. The plot shows that batches a, b/c, and d have huge differences between them, which cause the batch effect to be much more prevalent than either the treatment or time effect. In particular, batch a seems much further than the others, while batch d seems fairly close to b/c. Batch d contains only samples after 48 hours, so differences between it and b/c are unsurprising.



A PCA plot supports this notion. This one seems to suggest that the vast majority of variance in the data can be attributed to variation between batch a, and the others. After that, the next largest contributor to variance seems to be differentiating between groups b/c and d, and, finally, between b and c.



After normalizing, we model our data as

$$\log_2 y_{gitjk} = \mu_g + \text{Treatment}_{gi} + \text{Time}_{gt} + \text{TreatmentTime}_{git} + \text{BatchA}_{gj} + \varepsilon_{gitjk},$$

a linear model with an interaction between time and treatment, which includes a batch effect for batch A to account for varying levels of gene g in DMSO after 1 hour in the batch A group compared to the rest. The null model is set as

$$\log_2 y_{gitjk} = \mu_g + \text{Time}_{gt} + \text{BatchA}_{gj} + \varepsilon_{gitjk},$$

which still contains the batch adjustment, but now asserts that there is no treatment effect on the log-counts.

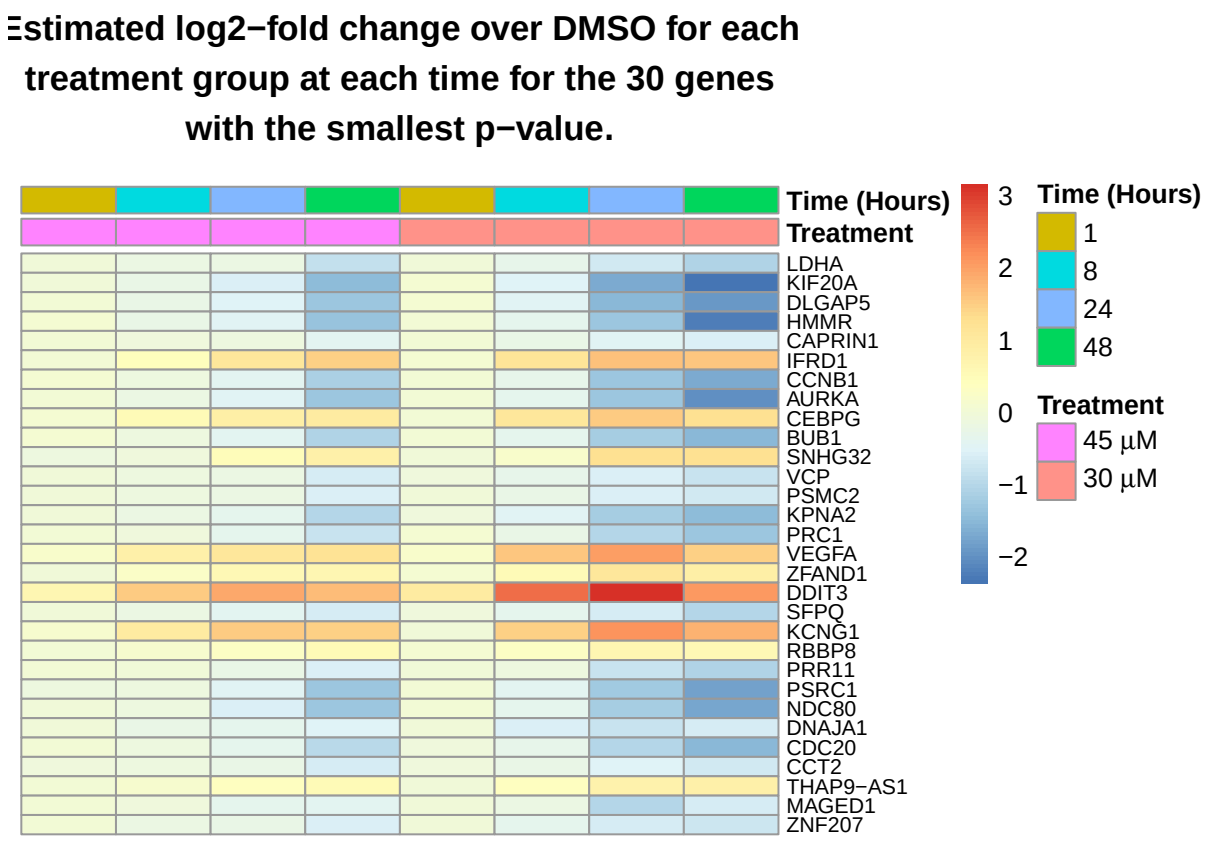
For each gene, we conduct a LRT to compare the full model to the reduced, null model, and, from that obtain a p-value. Those p-values are then adjusted using the Benjamini-Hochberg correction, to obtain an estimated False Discovery Rate. False Discovery Rate cutoff is set at 0.05, and a total of 4181 genes are found to have a treatment effect.

We also extract from that model the MLE of the $\log_2$-fold change between DMSO and IC75 Atovaquone, each at 48 hours, to model the effect size of the treatment. The 10 genes with the lowest FDR are reported here.

	log2FoldChange	pvalue	padj
LDHA	-1.09	1.05e-134	1.67e-130
KIF20A	-2.38	7.93e-85	6.28e-81
DLGAP5	-1.93	9.28e-83	4.9e-79
HMMR	-2.23	3.79e-78	1.5e-74
CAPRIN1	-0.572	4.5e-75	1.43e-71
IFRD1	1.57	1.34e-74	3.54e-71
CCNB1	-1.68	2.35e-69	5.33e-66
AURKA	-2	5.61e-69	1.11e-65
CEBPG	1.25	2.19e-68	3.86e-65
BUB1	-1.54	9.39e-68	1.49e-64

Heatmaps of fold-change expression

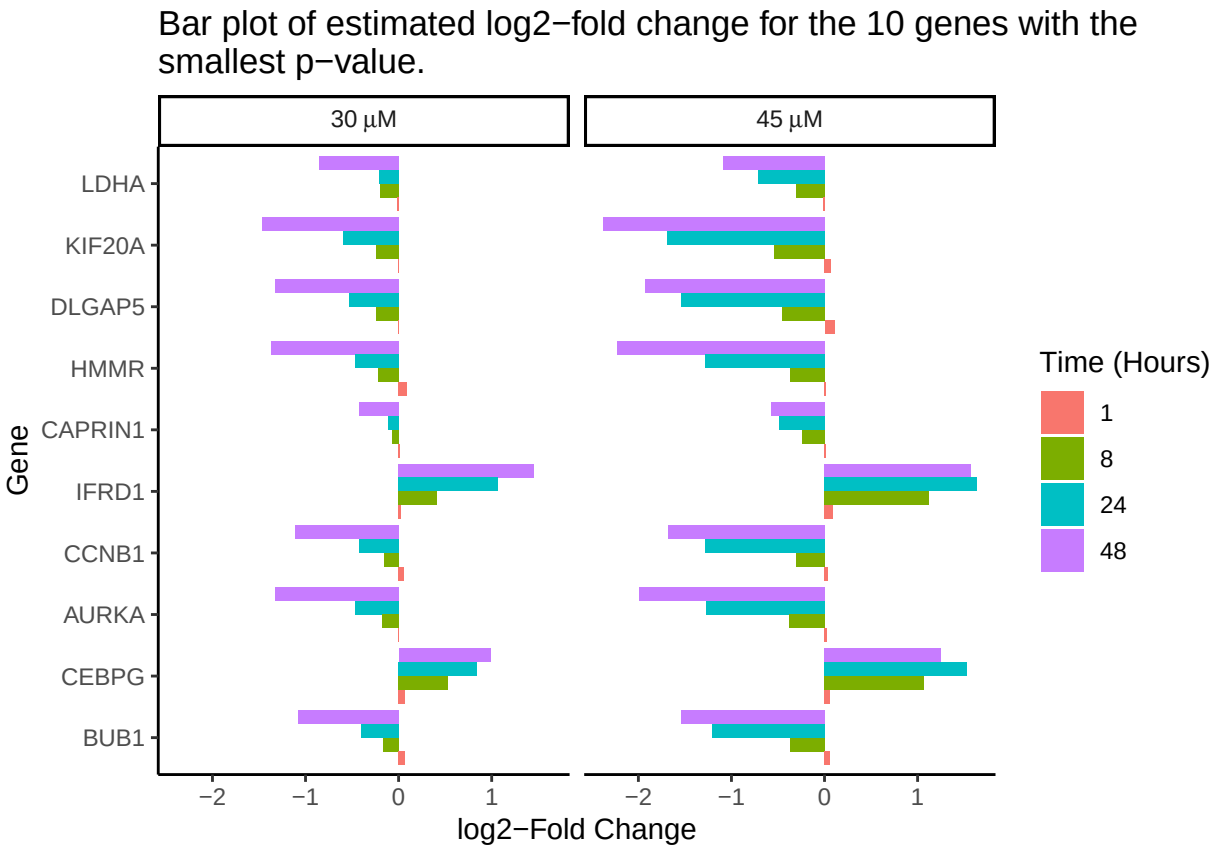
Changed the fontsizes Only 30 is displayed here. Both 30 and 50 can be found in `comparisons/lfc-heatmaps-*.png`.



Bar plots of lfc

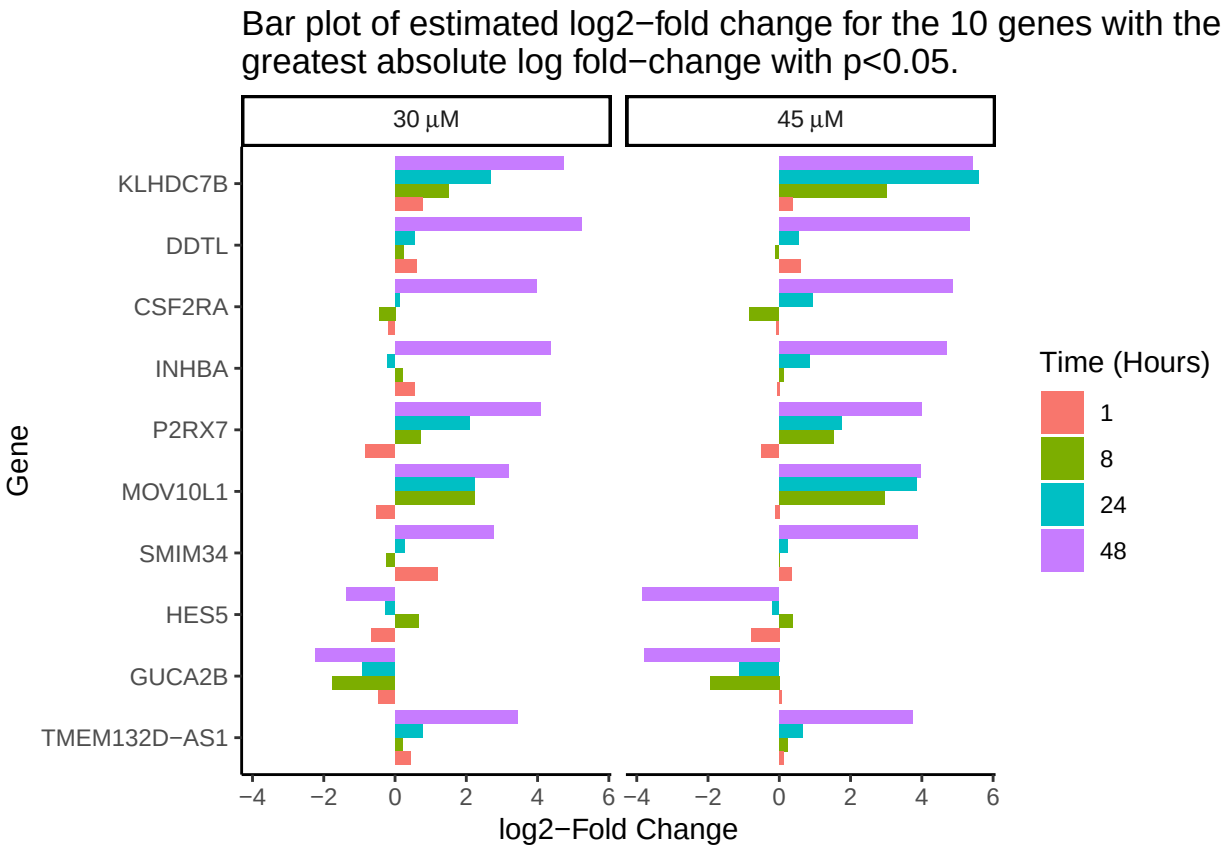
Sorted by p

Only 10 genes are shown here. Comparisons can be found in `comparisons/lfc-barplots-by_p-*.png`.



Sorted by absolute-lfc

Only 10 genes are shown here. Comparisons can be found in `comparisons/lfc-barplots-by_lfc-*.png`.



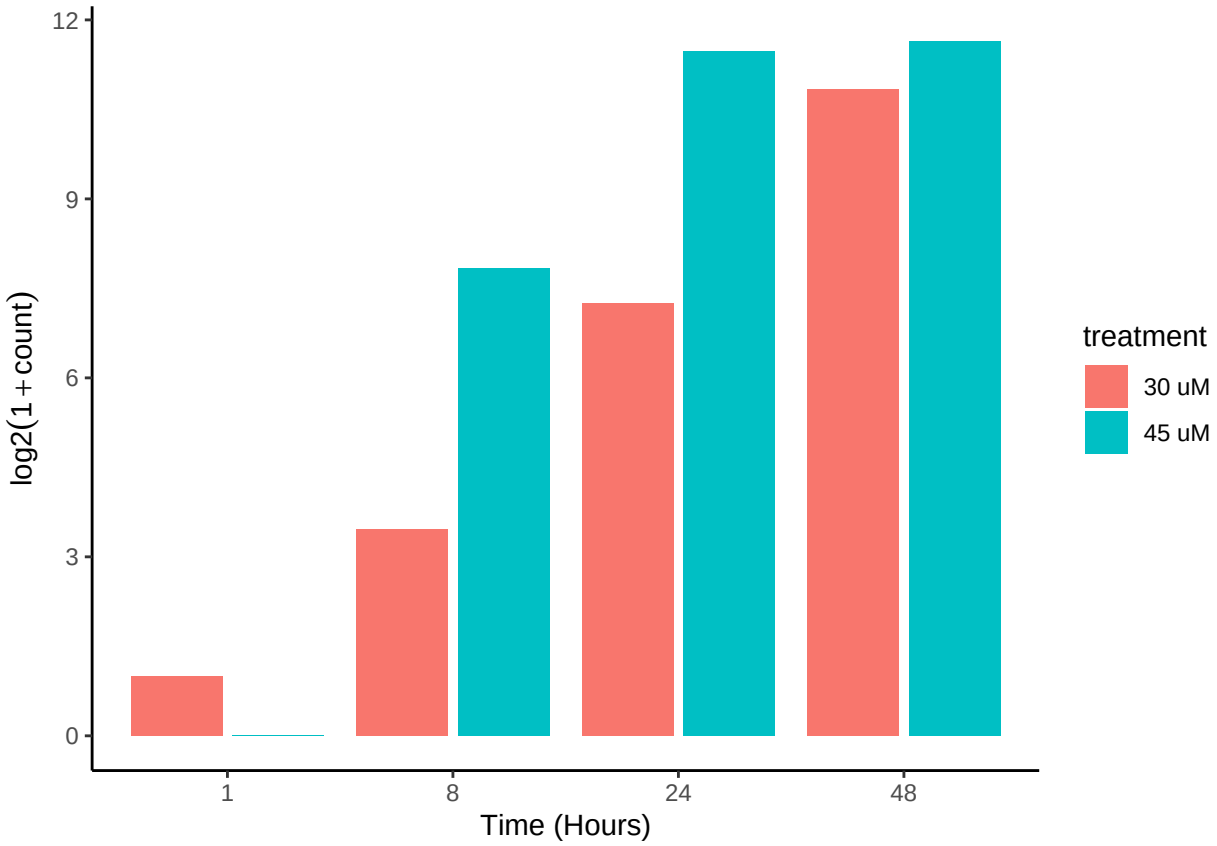
Number DE vs time

The plot below shows the number of DE genes (when compared to DMSO at the same time) over time. Note that, since we aren't much interested in the 1 hour time point, we keep a Bonferroni adjustment multiplier of 6, and we simply avoid inference on the 1 hour time point.

We see that the number of DE genes does seem to increase over time, with the higher dose of atovaquone having typically more DE genes than the lower. However, after 24 hours in the 45 uM group, the rate of new DE genes drops off, gaining only 371 genes by 48 hours, while the 30 uM group gained 1671 new DE genes in the same time.

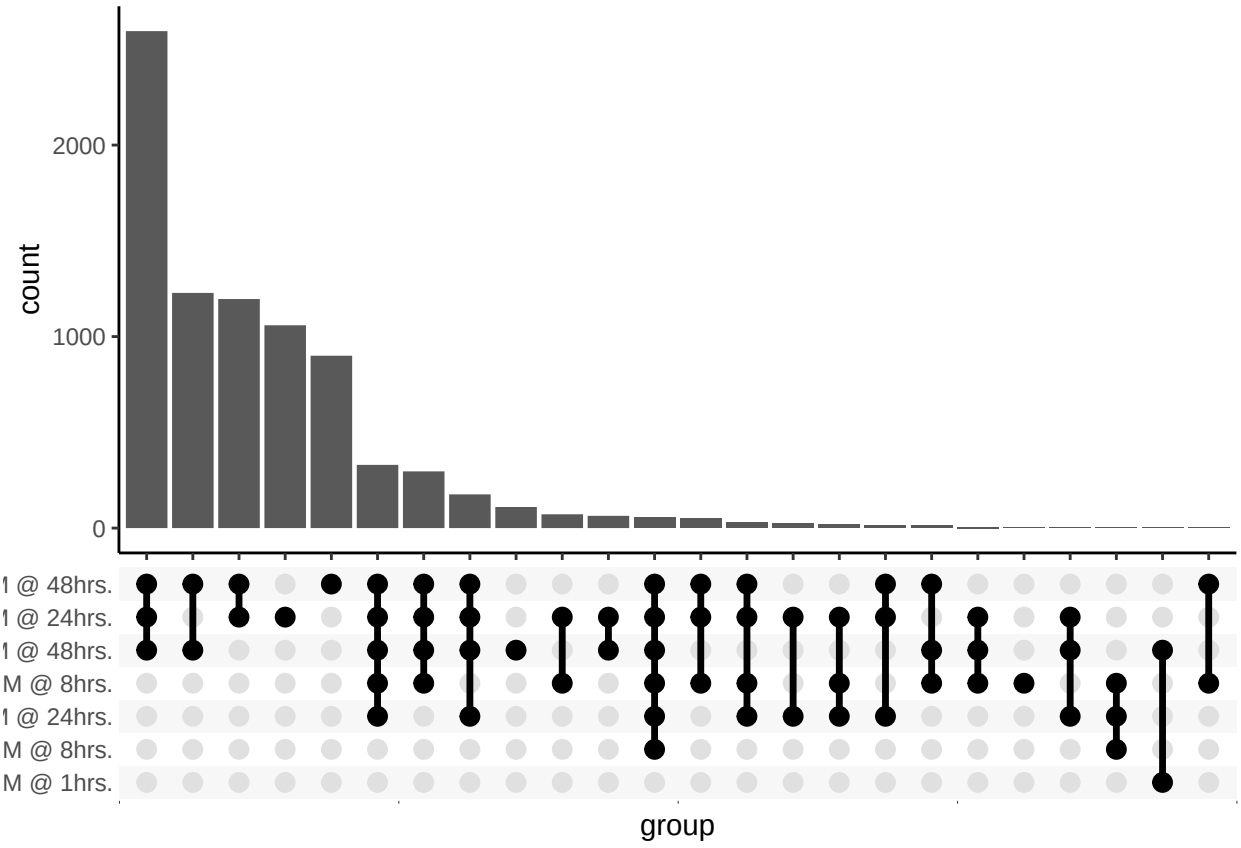
This image can be found in `figures/DE-counts.png`.

```
## pdf
##   2
```



This upset plot helps further elucidate the intersection pattern of these genes. This plot can be found in `figures/DE-upset.png`.

```
## pdf
##   2
```



Comparison in pairwise comparisons

Below is a table that summarizes the number of genes that are DE in each group compared to DMSO at the same time.

The main diagonal indicates how many are in that group. For example, the top-left entry indicates that after 48 hours in 45 uM atovaquone, 3204 genes were DE, while the bottom right says that after 8 hours in 30 uM, only 10 genes were DE.

The off-diagonal entries indicate the size of overlap between the respective groups indicated by the row and column. For example, the bottom left corner (or, equivalently, the top-right corner) indicates that there are 9 genes that are DE both in the 48 hours in 45 uM atovaquone vs 48 hours in DMSO, and in the 8 hours in 30 uM atovaquone vs 8 hours in DMSO. Since 8 hours in 30 uM has 9 DE genes, this indicates that there is only 1 gene DE in 30 uM after 8 hours that is *NOT* DE in 45 uM after 48 hours.

A full summary of these intersections can be found in `results/gene-pairwise-significance.csv`.

	48 Hours\\		24 Hours\\		8 Hours\\	
Treatment	45 uM	30 uM	45 uM	30 uM	45 uM	30 uM
48 Hours						
45 uM	3204	1676	1685	131	178	9
30 uM	1676	1822	1093	120	155	9
24 Hours						
45 uM	1685	1093	2833	150	217	9
30 uM	131	120	150	151	89	10
8 Hours						
45 uM	178	155	217	89	228	10
30 uM	9	9	9	10	10	10

Gene Set Analysis

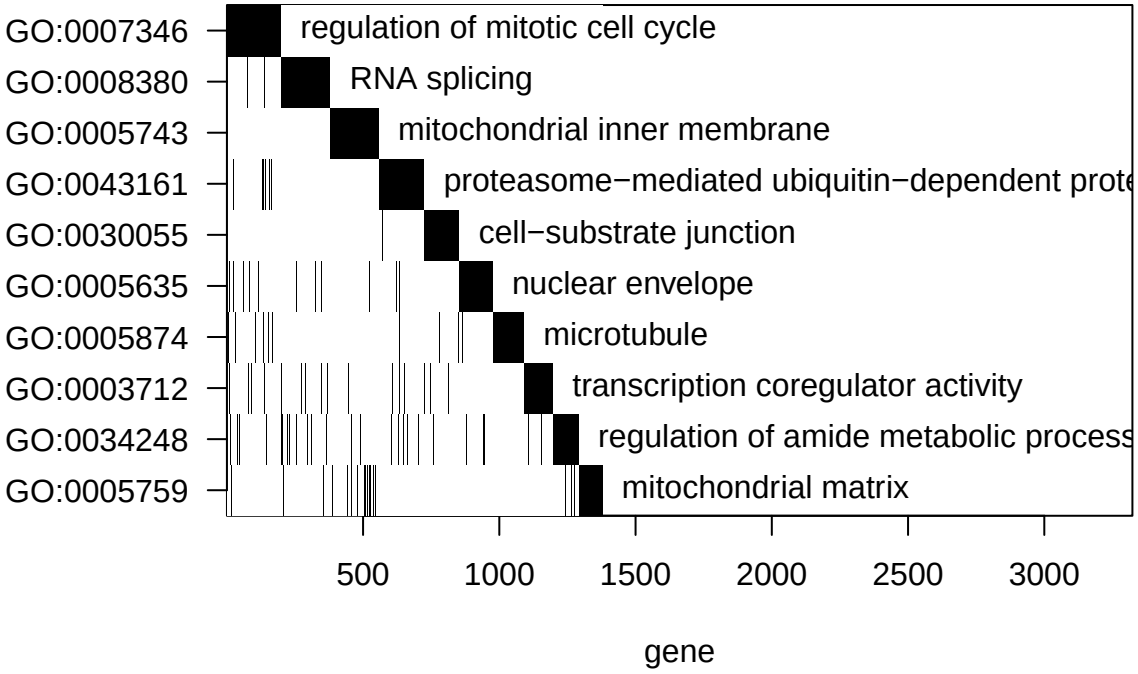
Allez - Binary

To try and make sense of the gene-level results from our model, we apply a gene set analysis. First, we use the *Allez* model, which, for each gene set, checks if that set has an unusually high number of stat. sig. DE genes for a set of its size.

After a Bonferonni correction, 286 gene sets with a number of genes between 5 and 500 are found to be "overly active." The 10 such sets with the highest z-score are displayed here, and a complete list can be found in `results/allez-sigsets.csv`.

Allez makes no attempt to consider the overlap between these gene sets. It does, however, provide a waterfall plot that helps visualize the results. The plot works by first taking the stat. sig. gene set that explains the greatest number of genes, placing it at the top, and plotting those genes along the x-axis. Then, iteratively, it takes the next stat. sig. gene set that explains the greatest number of *yet unexplained* genes, and plots them similarly, while also marking overlap between it and previously included genes.

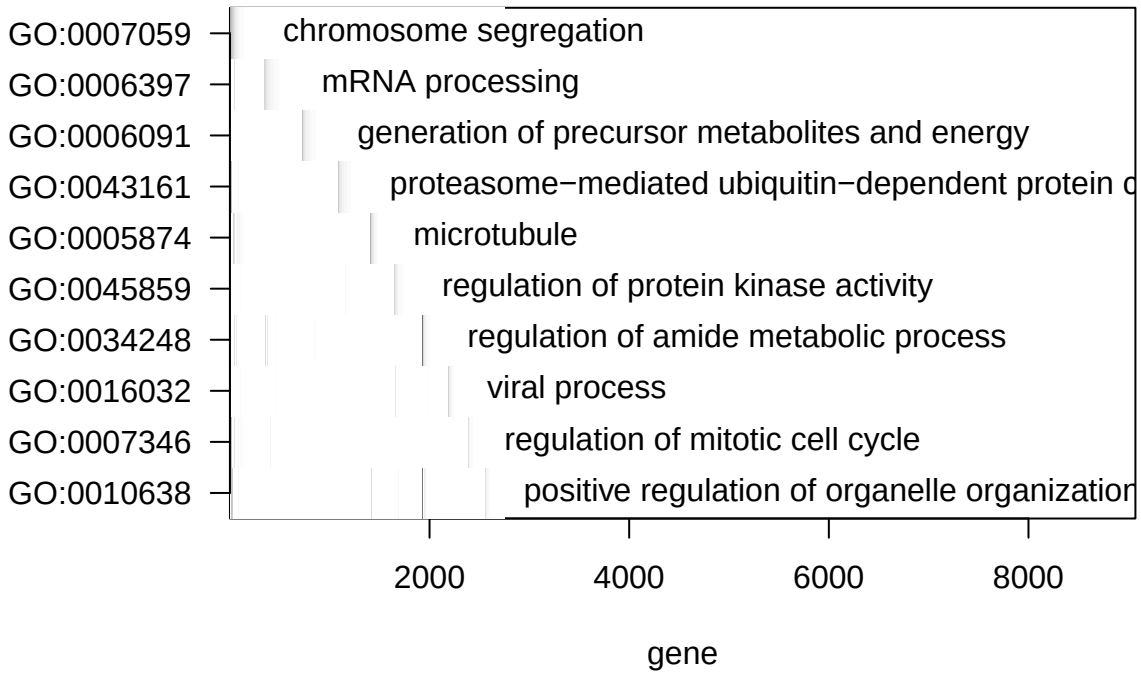
> Note to Mayra: I have a "beta" updated version of this code that may make this more visually appealing. Let me know if that's of interest.



Allez - $-\log(p)$

Allez is also very flexible, and can use any score for each gene as input. If we want to prioritize more significant genes more strongly, for example, we could use $-\log(p)$ as the score. Using this scoring method, the top 10 gene sets are similarly reported, and all 462 stat. sig. gene sets are saved in `results/allez-logp-sigsets.csv`.

And a corresponding Allez plot is also provided.

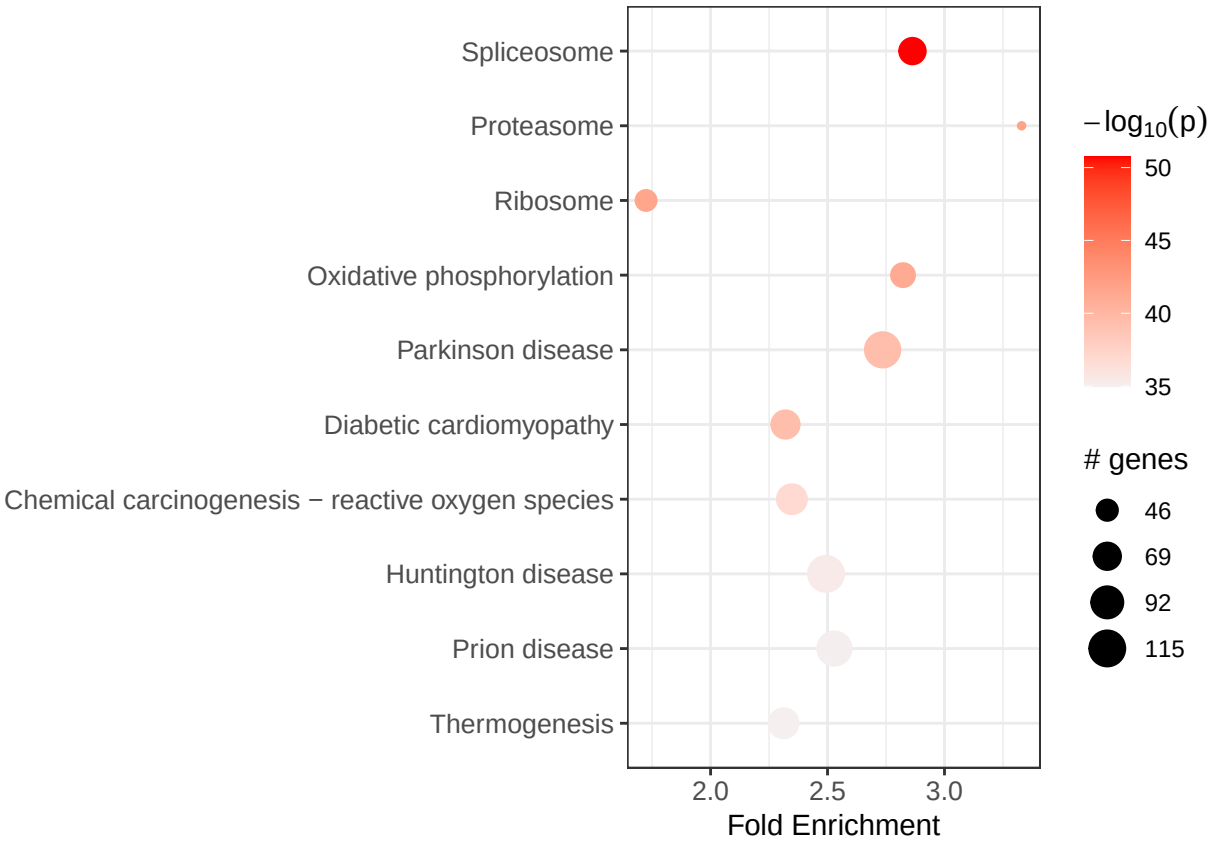


pathfindR

With 15

We also include pathway analysis using the pathfindR model. This model looks for active subnetworks of a Protein Interaction Network (PIN), and then conducts Pathway Enrichment Analysis on KEGG pathways to find those that are enriched.

This method found 247 enriched KEGG terms, of which, the 10 with the lowest p-value are displayed below. All such terms are contained in `results/pathfindR-active-sets.csv`.

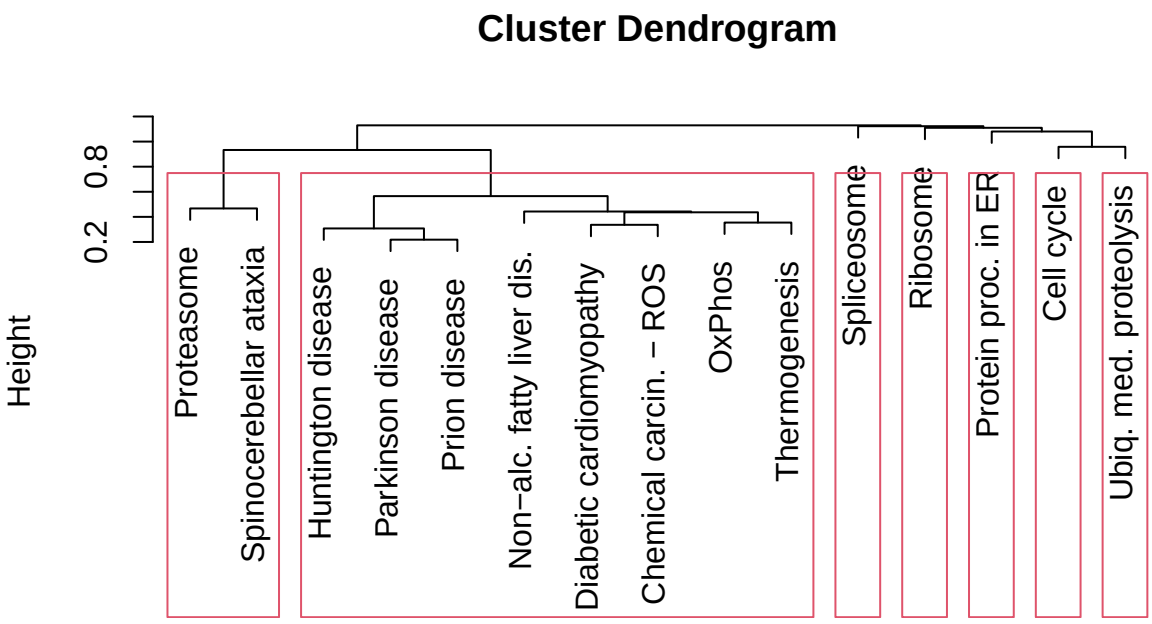


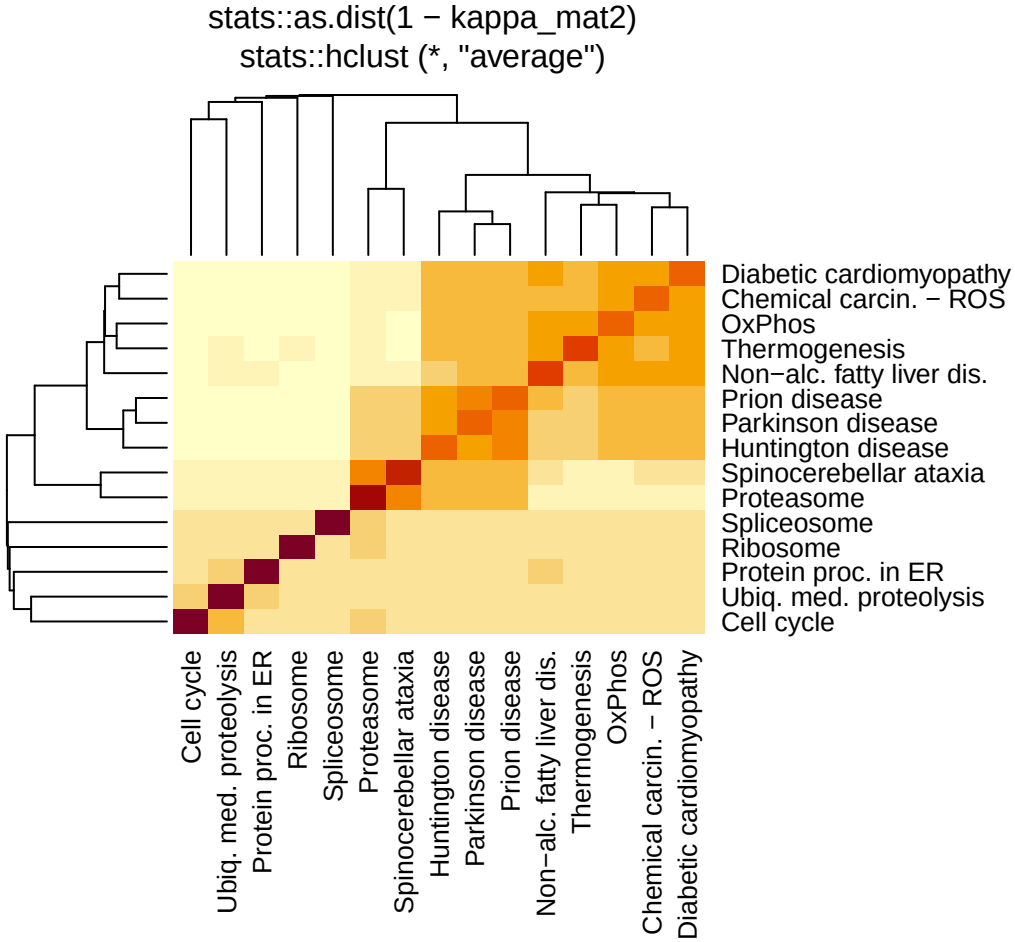
A hierarchical clustering of these terms shows 7 main clusters

1. Spliceosome
2. Proteasome + Spinocerebellar ataxia
3. Ribosome
4. OxPhos + Parkinson + Diabetic Cardiomyopathy + Chemical carcin. ROS + Huntington + Prion + Thermogenesis + Non-alc. fatty liver dis.
5. Cell cycle
6. Protein Processing in ER
7. Ubiq. med. proteolysis

This indicates that the other 8 gene sets share a great deal of genes in common, which may help explaining why, e.g., Prion Disease shows up in a dataset where we would not particularly expect it to — the KEGG pathway for Prion Disease may simply be quite similar to other pathways, such as that of OxPhos, which we would expect to show up.

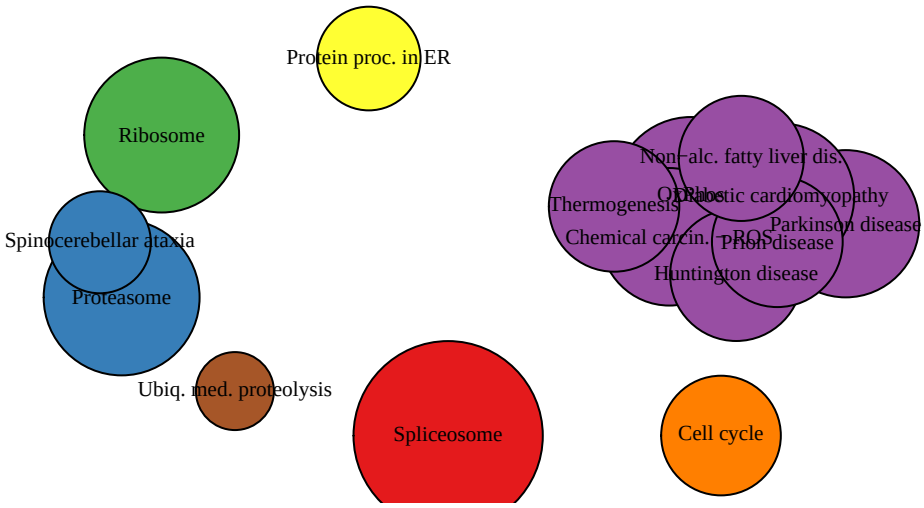
First, the clustering as a dendrogram.



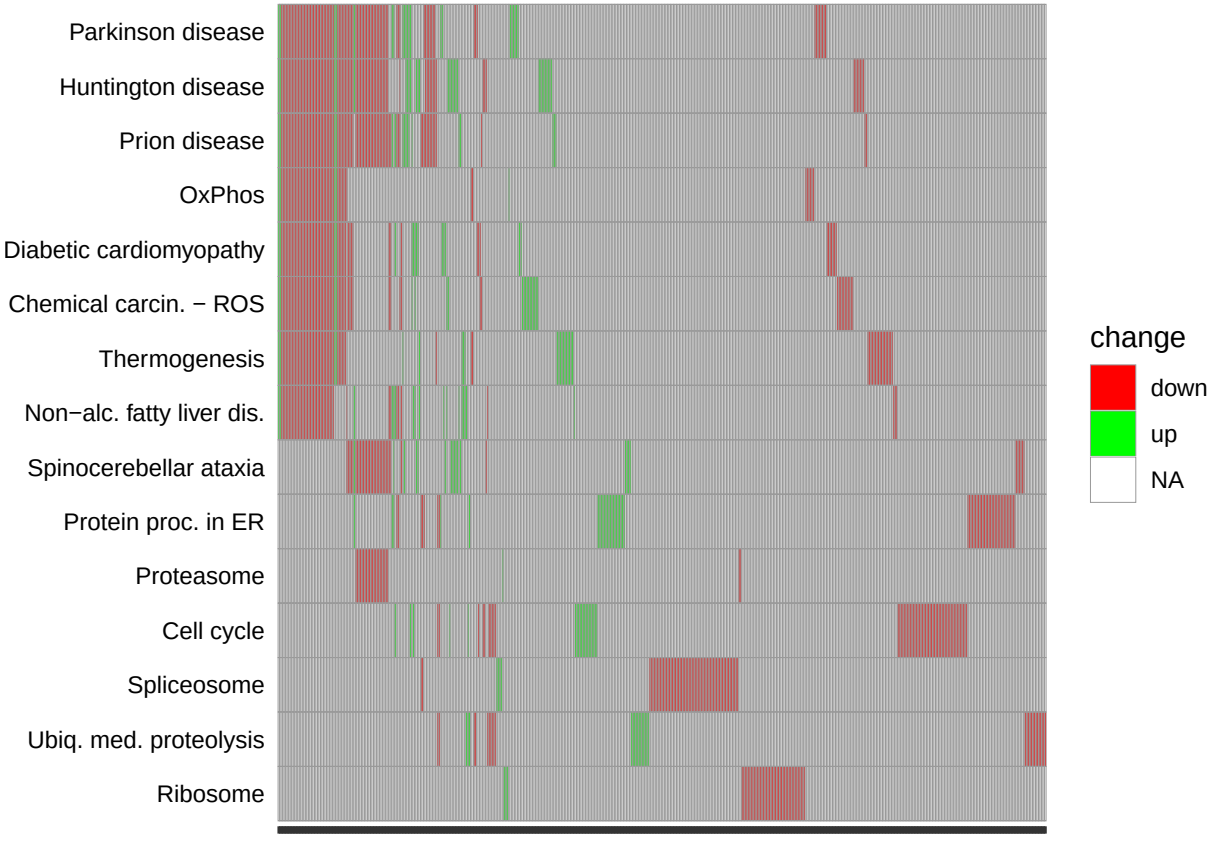


Now as a heatmap.

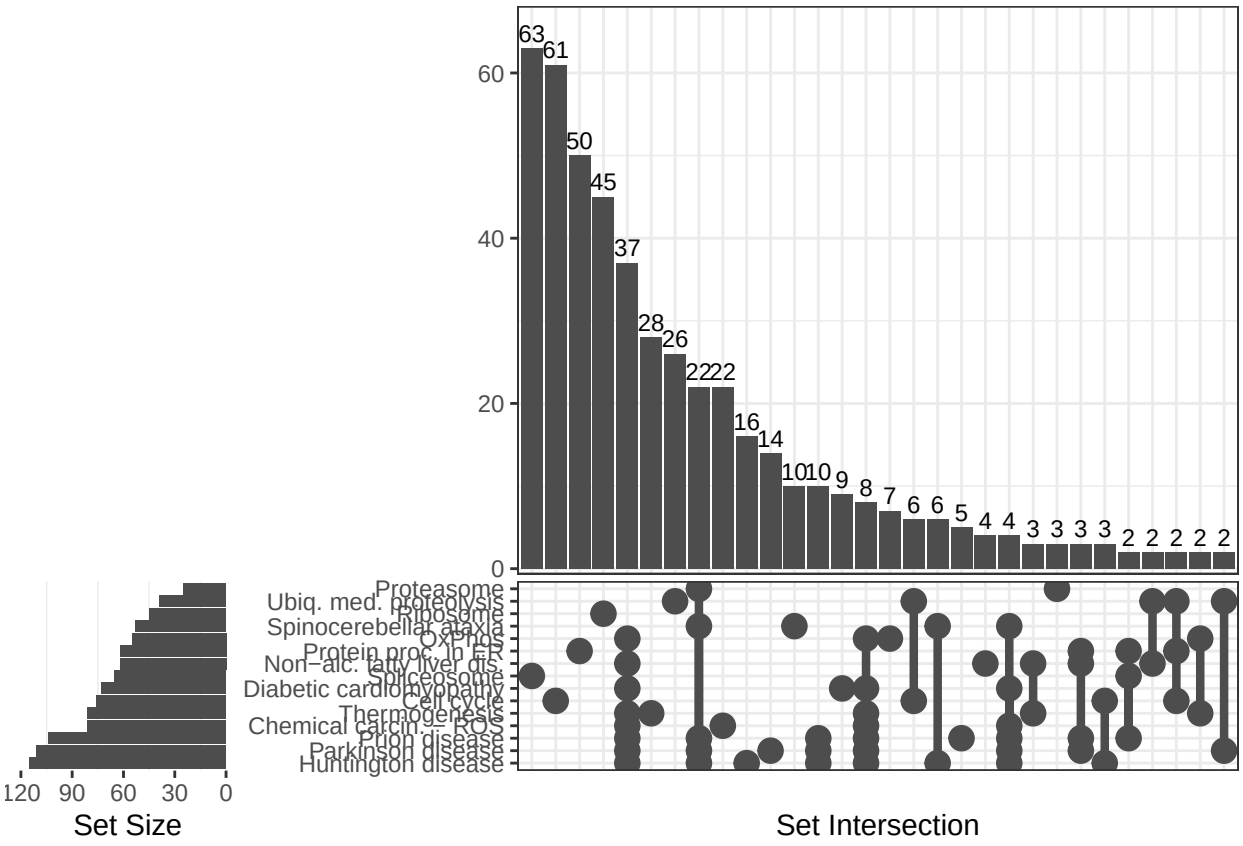
And now as a this thing.



These plots show a plot of the genes of each cluster of terms, along with the direction of change. As you can see, the Spliceosome and Ribosome sets share few actual genes in common, indicating that the relation between the two is a bit more complex. However, the other 8 terms share many genes in common. In fact, Parkinson, Huntington, and Prion disease pathways all look like ROS + Proteasome + Others.



An upset can be used to see the size of various intersection patterns between these 10 gene sets.

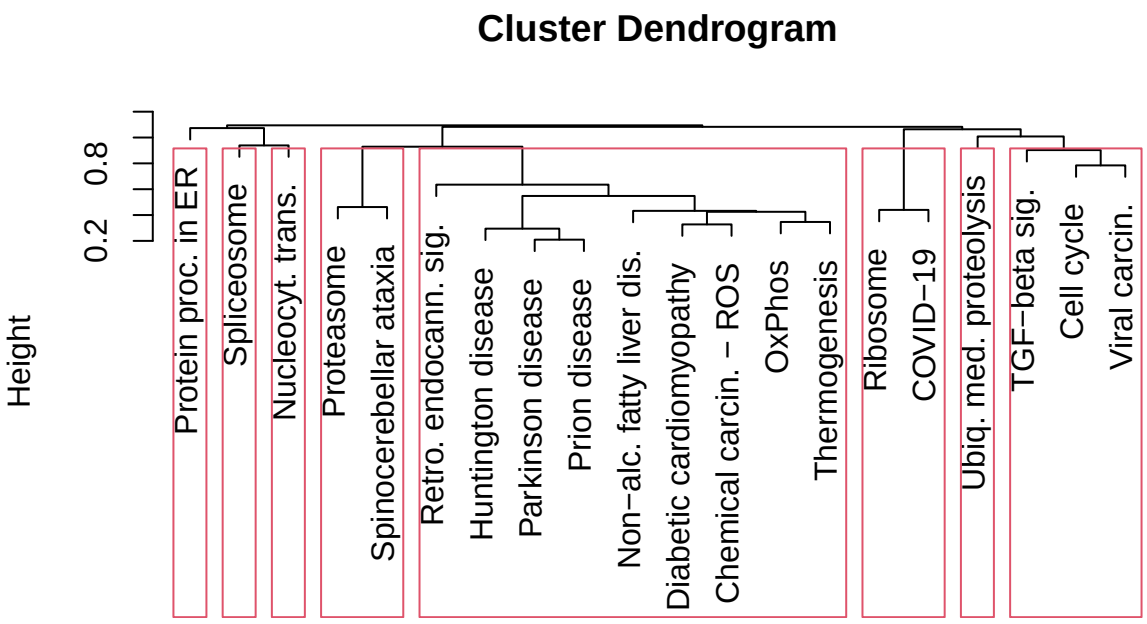


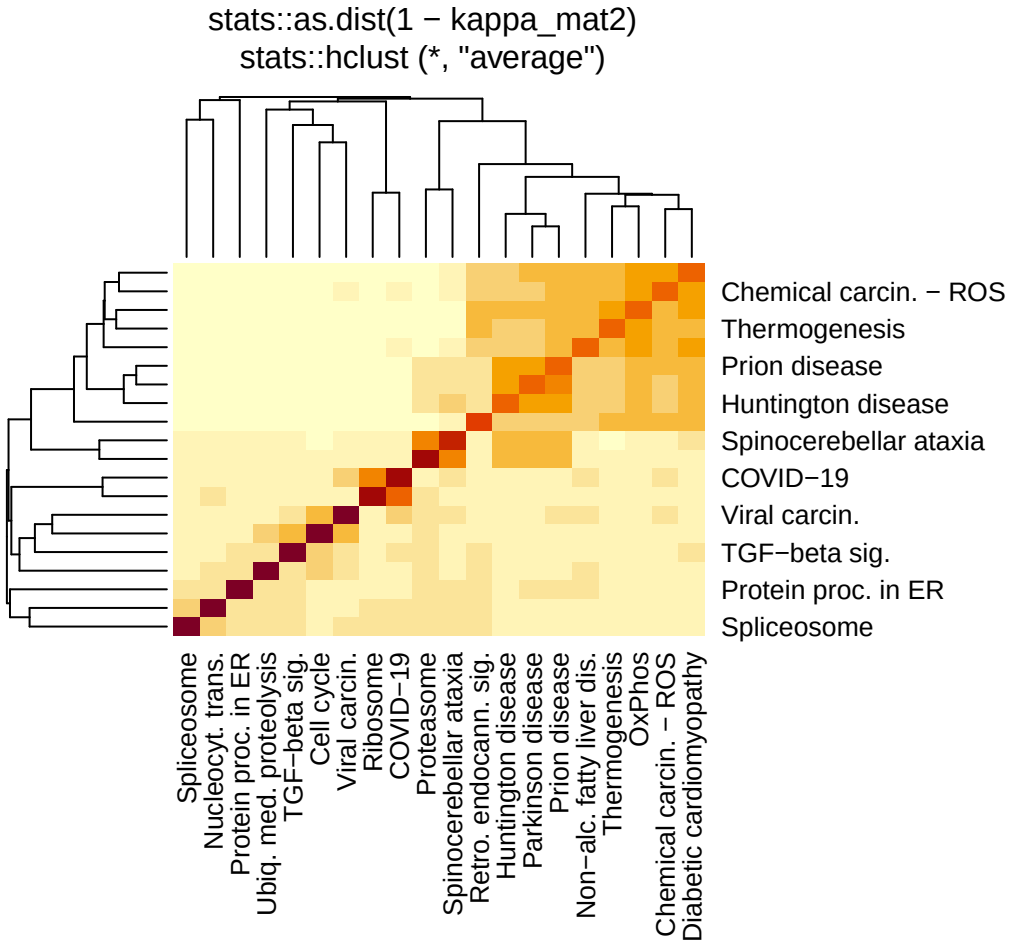
With 20

A hierarchical clustering of these terms shows 8 main clusters

1. Spliceosome
2. Proteasome + Spinocerebellar ataxia
3. Ribosome + COVID-19
4. OxPhos + Parkinson + Diabetic Cardiomyopathy + Chemical carcin. ROS + Huntington + Prion + Thermogenesis + Non-alc. fatty liver dis.
5. Cell cycle + TGF-beta sig. + Viral carcin.
6. Protein Processing in ER
7. Ubiq. med. proteolysis
8. Nucleocyoplasmic transport

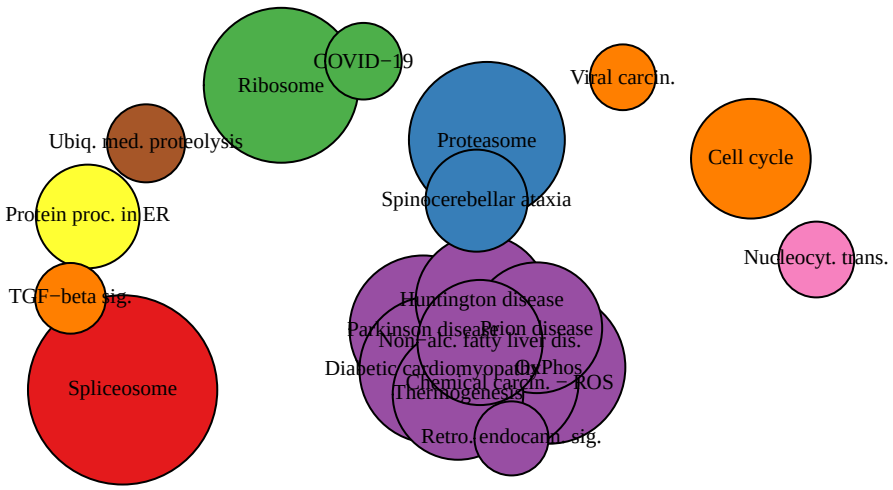
This indicates that the other 8 gene sets share a great deal of genes in common, which may help explaining why, e.g., Prion Disease shows up in a dataset where we would not particularly expect it to — the KEGG pathway for Prion Disease may simply be quite similar to other pathways, such as that of OxPhos, which we would expect to show up.



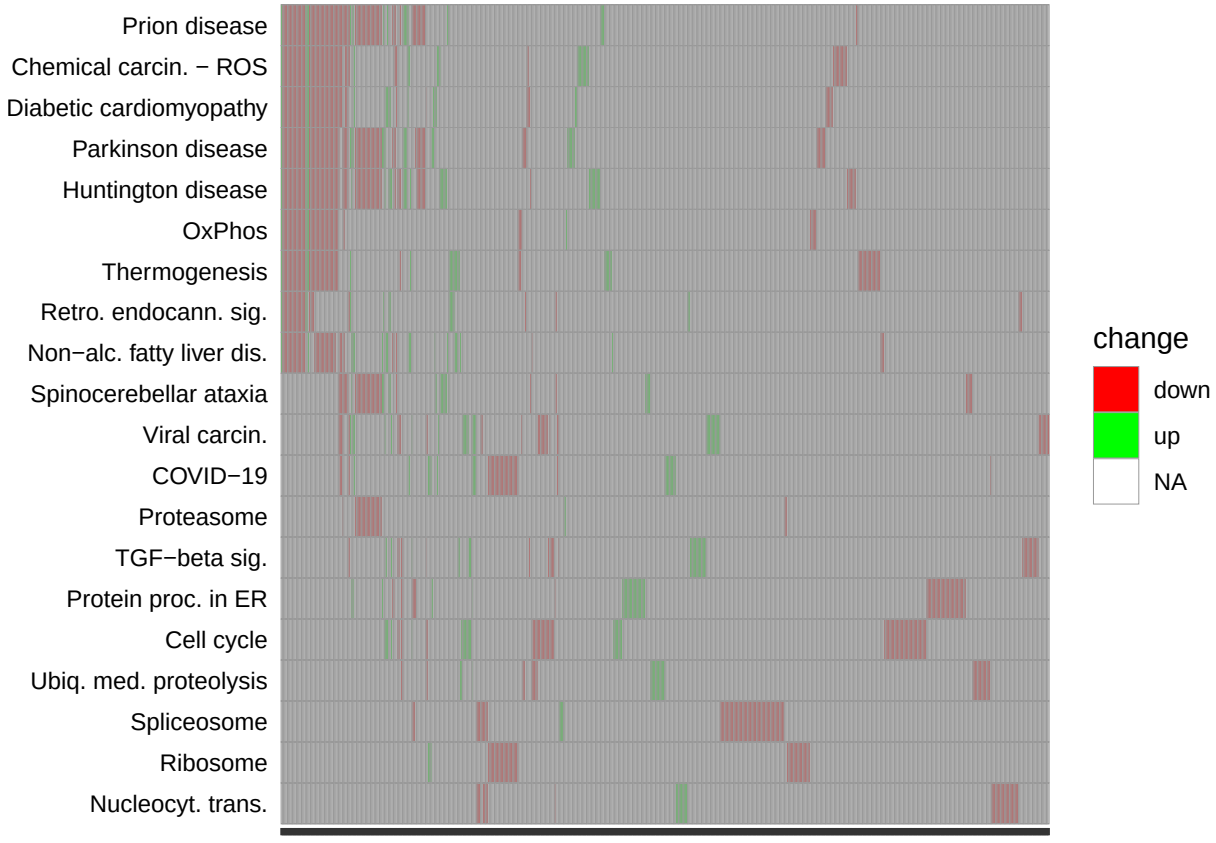


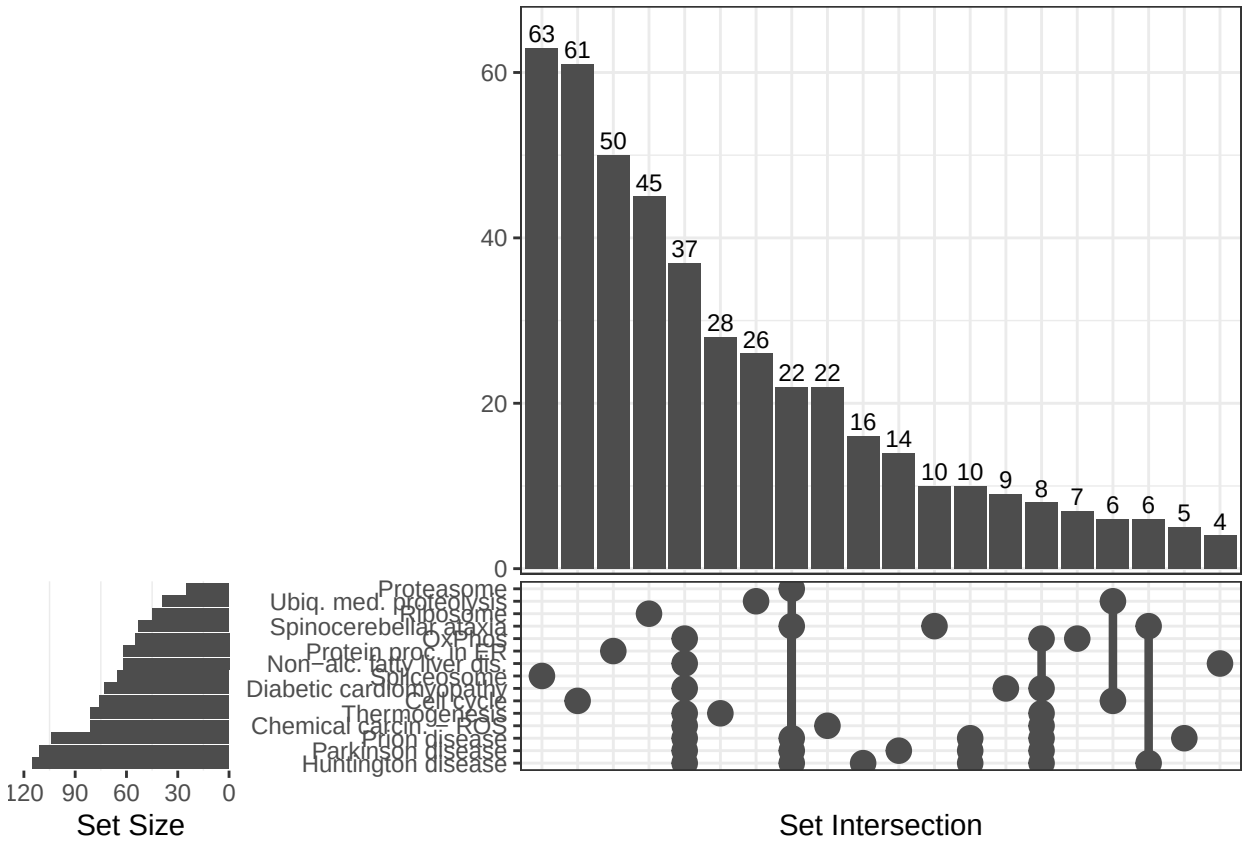
Now as a heatmap.

And now as a this thing.



These plots show a plot of the genes of each cluster of terms, along with the direction of change. As you can see, the Spliceosome and Ribosome sets share few actual genes in common, indicating that the relation between the two is a bit more complex. However, the other 8 terms share many genes in common. In fact, Parkinson, Huntington, and Prion disease pathways all look like ROS + Proteasome + Others.




63
61
50
45
37
28
26
22
22
16
14
10
10
9
8
7
6
6
5
4
60
40
20
0
Proteasome
Ubiq. med. proteolysis
Ribosome
Spinocerebellar ataxia
Oxphos
Protein proc. in ER
Non−alc. fatty liver dis.
Spliceosome
Diabetic cardiomyopathy
Cell cycle
Thermogenesis
Chemical carcin.− ROS
Prion disease
Parkinson disease
Huntington disease
120
90
60
30
0
Set Size
Set Intersection