Integration of Transcriptomic data into Human Hepatocytes model

Author

Shogofa MORTAZA
M2BI Student
Paris Diderot University
75 013 PARIS
shogofa.mortaza@hotmail.fr

HEALS

Health and Environment-wide Associations based on Large population Surveys

Responsibles

Karine AUDOUZE
Inserm-UMR-S973
Paris Diderot University
karine.audouze@univ-paris-diderot.fr

Martine AGGERBECK
UMR-S1124
Paris Descartes University
martine.aggerbeck@parisdescartes.fr

Table of contents

I - Introduction

The HEALS (Health and Environment-wide Associations based on Large population Surveys), European project, focuses on the exposome effects on human health. The exposome is the total exposures to environmental factors in a human organism from conception to the end of its survival. This project is structured around several axes: cohort analysis to assess internal and external exposures, data modeling to reconstruct the exposure from partial data and, the analysis of the signaling pathways involved in pathologies of interest after exposure to some of the pollutants found in the cohorts.

As part of this project, this proposed topic focuses on the integration of transcriptomic data obtained in a human hepatocytes model (HepaRG line) treated with a mixture of pollutants (3 phthalates and 2 heavy metals) in order to study biological enrichment in the presence of these compounds and their relationship with pathologies.

II - Materiel & Method

All data and the complete analysis (figures and codes) are available on my Github : https://github.com/SMORTAZA/My_project

1 - Data

The data are CEL files derived from a transcriptomic microarray experiment. The experiment consisted in differentiating the HepaRG cells and then treating them under different conditions for three weeks. These conditions are :

- Baseline, corresponding to the differentiated cells before the treatments,
- Contrôle, corresponding to the cells without any treatment
- HNO3, the heavy metals solvent
- M1, the mixture with the lowest concentrations of 3 phthalates and heavy metals
- M2, the mixture with the highest concentrations of 3 phthalates and heavy metals (10 times more concentrated)

For each of these conditions, five biological replicates are made and correspond to different prinking out of cells. Thus, in total there are 53 617 genes / probes for 25 samples.

2 - RStudio

In order to perform statistical analysis of these data, we use RStudio (or R), a dedicated tool for this. Moreover, different packages are to be installed like oligo, pd.hugene.2.0.st, hugene20sttranscriptcluster.db, genefilter and limma.
This tool allows us to read the data, normalize them and make various statistical tests (Ttests, ANOVA) to select the most significant genes in the comparisons.
The different comparisons made are :

- Ctrl vs HNO3
- HNO3 vs M1
- HNO3 vs M2
- M1 vs M2

- Baseline vs Ctrl

For statistical tests, Ttest was used to calculate pvalues and fold changes (fc) in order to have volcanoplots. In a statistical test, the pvalue is the probability of getting the same (or even more extreme) value of the test if the null hypothesis was true. The fold change is the difference between two conditions. From these plots, thresholds were chosen following a conventional approach. Thus, for each case, 3 analyzes are carried out:

- pval < 0,05 and fc<1 ;
- pval < 0,05 and fc<0.5 ;
- pval < 0,05 only, not selection.

Then, the pvalues are recalculated with ANOVA. Depending on the thresholds, the most significant genes are selected.

3 - Biological Enrichment

Biological enrichment consists of discovering pathways and / or GO (Gene Ontology) terms with lists of significant genes. Online tools have been used for this part like DAVID, Panther and Gorilla.
AMIGO 2 is also used to define some GO terms.

III - Results

1 - Data normalization and visualization

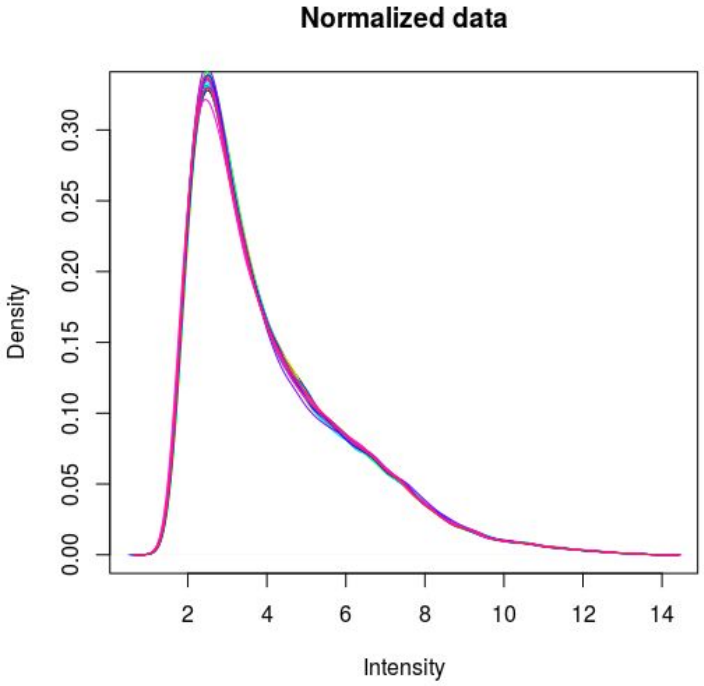


Figure 1 : Density plot showing data normalization. All the curves are superimposed.

In order to evaluate the quality of microarray data there are many things that can be done. One such thing is the density plot (Figure 1). Another good way of evaluating the data

is to do a Singular Value Decomposition (SVD), which is kind of the same as a Principal Component Analysis (PCA), just designed to take matrice that have many more rows than columns. An SVD analysis kind of breaks my data into 25 components (= number of samples). The first three component contains the most variation, and the last the least (none) (Figure 2). With the plot of SVD from the second and the third components, we can see that there are not outliers (Figure 3). Three groups stand out :

1. Baseline,
2. M1 et M2,
3. HNO3 et Ctrl.

The separation into these groups can be explained. The first group corresponds to the "young" cells, and without any treatment. The other two groups correspond three weeks old cells. Group 2 corresponds to the group treated with pollutants whereas group 3 is untreated. Figure 3 could show the influence of aging cells and the influence of the treatment by the mixture of pollutants.

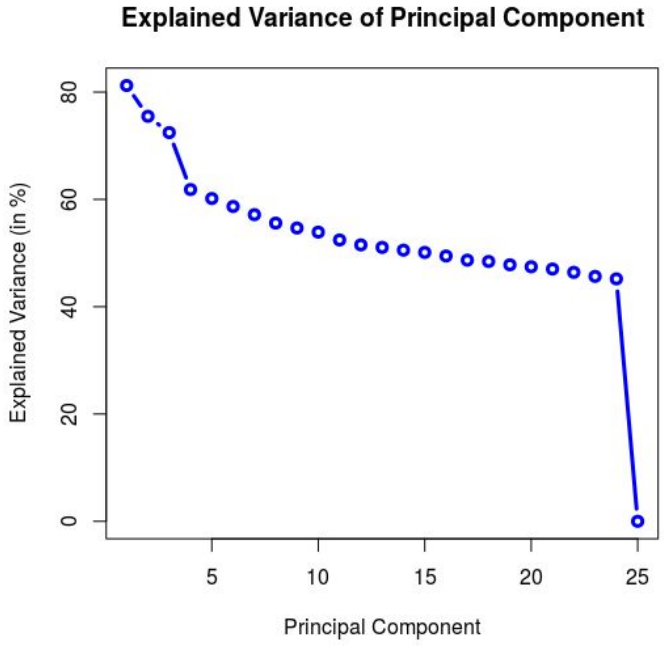
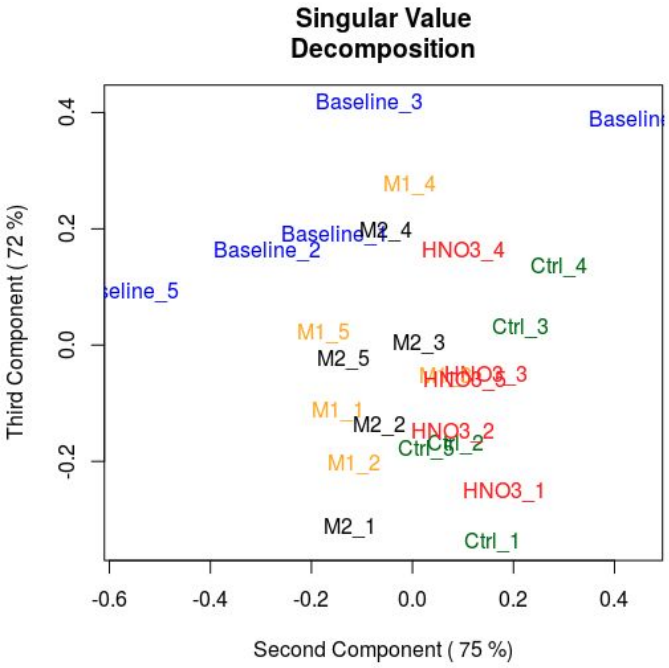


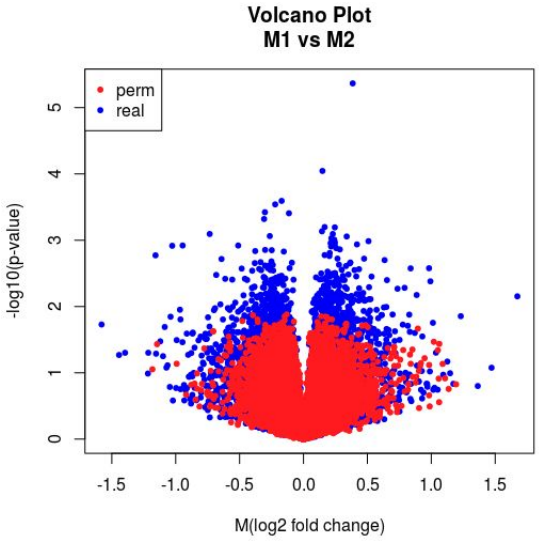
Figure 2 : Scree plot of the variance explained by each principal component. The first three are the interesting principal component.



<u>Figure 3 :</u> Plot of the Singular Value Decomposition with the second and the third components. Both of them contains important variation (respectively 75 % and 72%).

3 - Statistical analysis

In a volcanoplot, each point represents a gene. The genes at the top of the volcanoplot are the most significant . A fold change (fc) less than 0 shows the downregulated genes and a fc greater than 0 shows the upregulated genes (Figure 4).



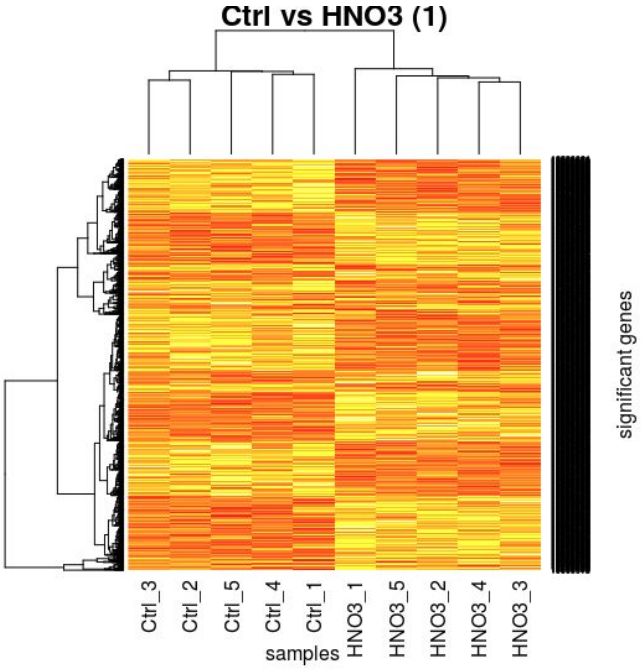
<u>Figure 4 :</u> Volcanoplot of M1 vs M2 comparison. Each point represents a gene. The blue dots are the real data and the red dots are the simulated data.

Comparisons	pval < 0.05 and fc < 1	pval < 0.05 and fc < 0.5	pval < 0.05
Ctrl vs HNO3	1 909	1 821	1 920
HNO3 vs M1	2 726	2 565	2 745
HNO3 vs M2	2 871	2 713	2 898
M1 vs M2	2 022	1 932	2 028
Baseline vs Ctrl	4 953	4 634	4 995

Table 1 : Number of significant genes for different analysis and for different comparisons.

For each comparison, different numbers of significant genes are obtained (Table 1). The more precise the analysis (pval and fc weak), the fewer differential genes there will be. This difference of a few tens to a few hundred genes is visible on the heatmaps, in particular between "pval <0.05 and fc <0.5" and "pval <0.05" (Figure 5).
In addition, among these comparisons:

- Ctrl vs HNO3 have the least different genes. These two conditions would therefore be considered equivalent.
- There are many more genes in comparisons between HNO3 and M1 / M2 conditions. The pollutants would therefore have an influence on the treated cells.
- There are even more different genes between Baseline and Ctrl conditions. This would show the influence of the culture time.



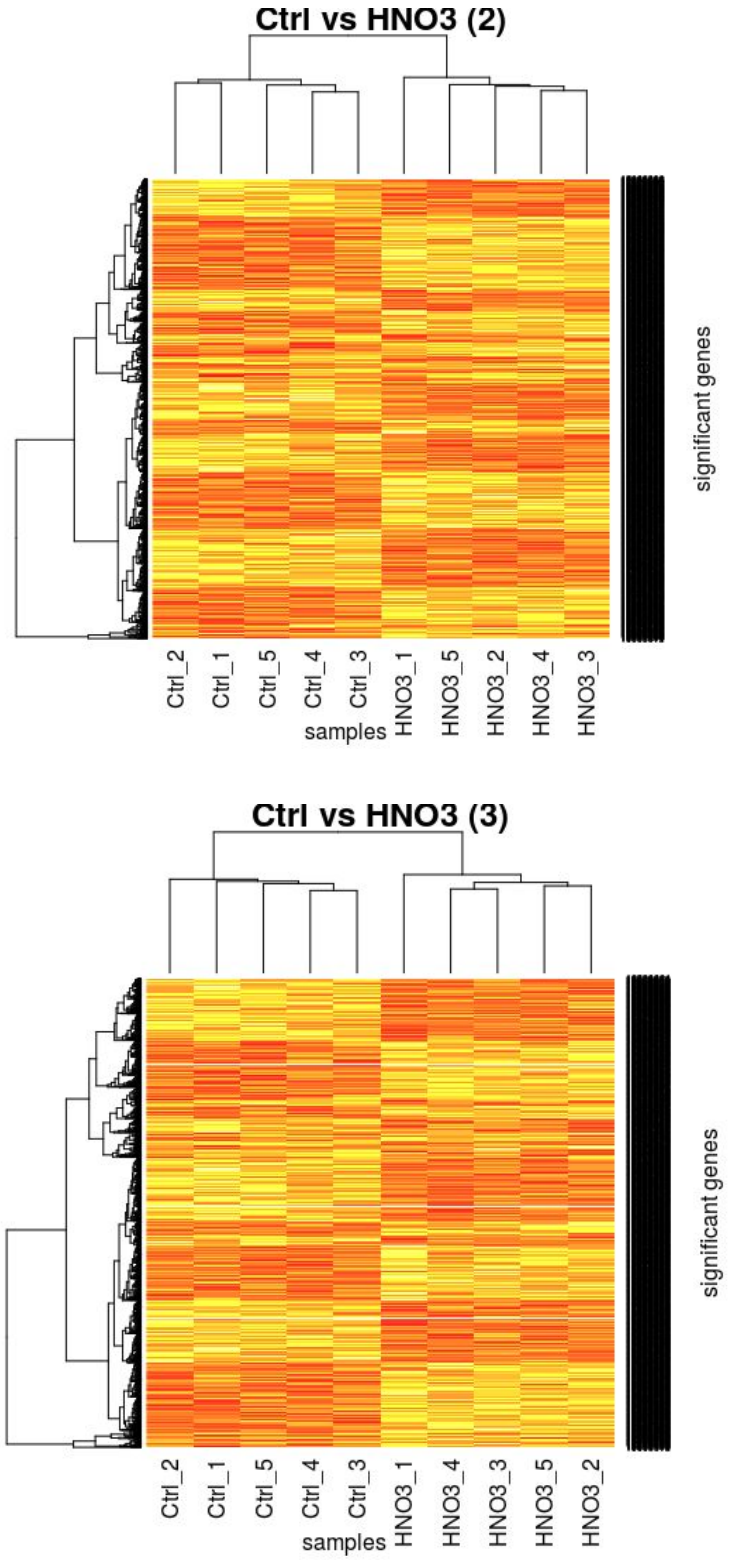


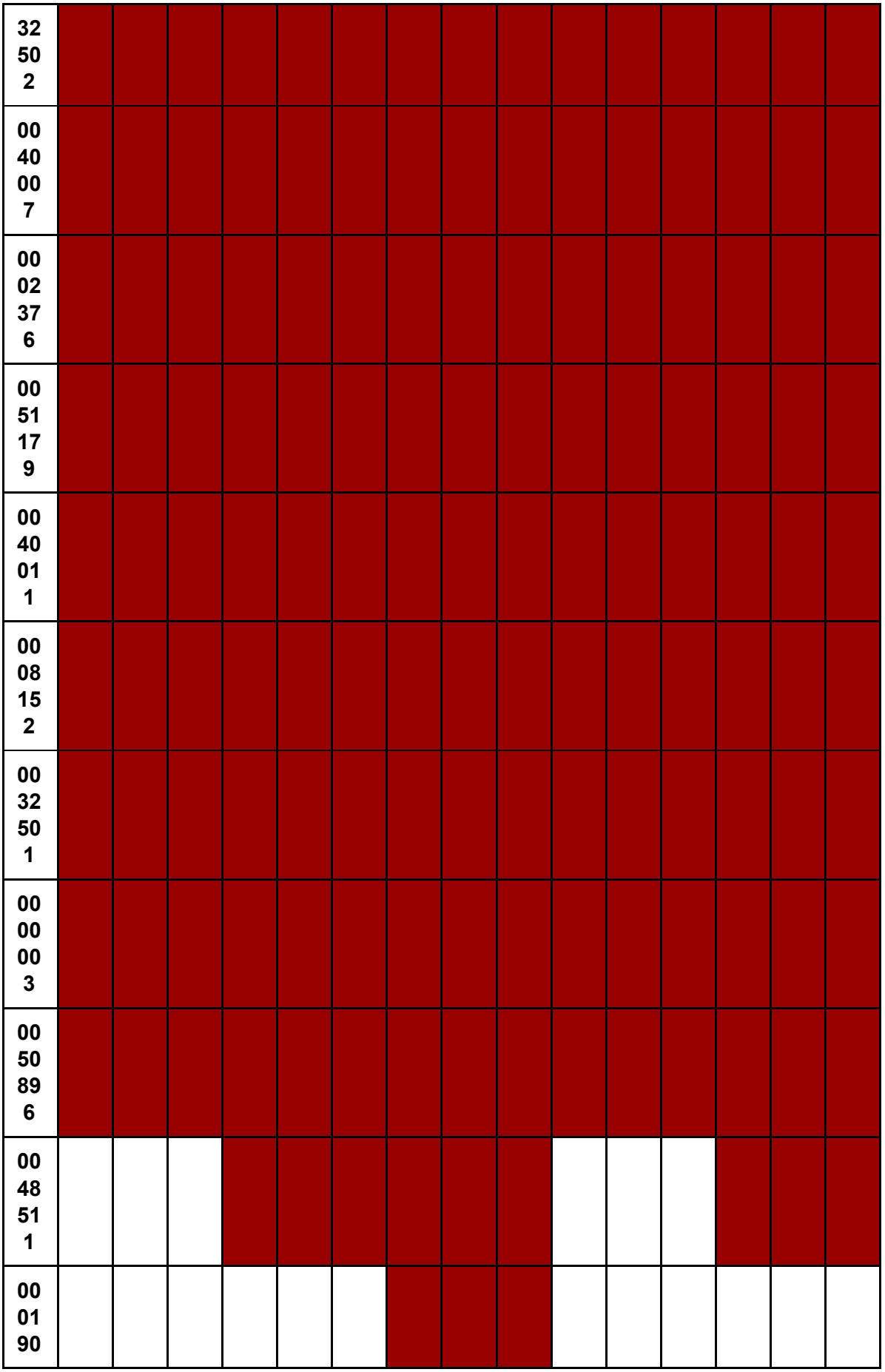
Figure 5: Heatmaps of the Ctrl vs HNO3 comparison according to the three analyzes carried out. (1) pval <0.05 and fc <1. (2) pval <0.05 and fc <0.5. (3) pval <0.05. The probes of significant genes are not visible because there are many. See the TOP50 on the Github.

4 - Biological enrichment

First, we are interested in the biological processes involved in the 15 cases (3 analyzes by comparison) with the online tool Panther (Table 2). These different processes are defined by GO terms, standard terminologies. When there is a difference when comparing two conditions, we have the GO term of the biological process affected. Overall, the more genes affected, the more biological processes involved (Table 1 and 2). In addition, the other results are as follows:

- 13 biological processes are affected in all 15 cases.
- Comparisons 2, 3 and 5 have in addition the GO term "GO: 0048511". It corresponds to the rhythmic process: Any process relevant to the generation and maintenance of rhythms in the physiology of an organism. The comparison between Baseline and Ctrl would show an influence of the culture time, and thus the possible aging of the cells. The cells of conditions M1 and M2 could have an acceleration of aging compared to normal.
- The comparison between HNO3 and M2 has in addition the GO term "GO: 0001906". It corresponds to cell killing: Any process in an organism that results in the killing of its own cells or those of another organism, including in some cases the death of the other organism. Killing here refers to the induction of death in one cell by another cell. A high concentration of pollutants would therefore induce an important acceleration of cellular aging, up to conduction at cell death.

GO	(1)			(2)			(3)			(4)			(5)		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
0022610															
0065007															
0071840															
0009987															
00															


32502
0040007
0002376
0051179
0040011
0008152
0032501
0000003
0050896
0048511
0001990

6															

Table 2: GO terms for Biological Process with significant genes for the five comparisons from Panther. (1) = Ctrl vs HNO3, (2) = HNO3 vs M1, (3) = HNO3 vs M2, (4) = M1 vs M2 and (5) = Baseline vs Ctrl. 1 = pval < 0.05 and fc < 1, 2 = pval < 0.05 and fc < 0.5, 3 = pval < 0.05.

From the same gene lists and using Gorilla, another online tool for biological enrichment, we get more GO terms. So, we can see differences between the three analyzes of a comparison. This difference is especially visible between the analyzes "pval <0.05 and fc <0.5" and "pval <0.05". But this type of results is not presented in detail in this report.

In a second time, we are interested in pathways. We understand that the larger the gene differential, the greater the number of pathways involved (Table 3).

(1)			**(2)**			**(3)**			**(4)**			**(5)**		
1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
5	5	5	31	27	31	33	21	33	5	4	5	59	51	59

Table 3: Number of KEGG Pathways with significant genes for the five comparisons from DAVID. (1) = Ctrl vs HNO3, (2) = HNO3 vs M1, (3) = HNO3 vs M2, (4) = M1 vs M2 and (5) = Baseline vs Ctrl. 1 = pval < 0.05 and fc < 1, 2 = pval < 0.05 and fc < 0.5, 3 = pval < 0.05.

IV - Conclusion & Discussion

At this stage, some conclusions can be given:

- Nitric acid (HNO3) does not appear to have a significant influence on cell cultures.
- The pollutants would have an influence on the cells.
- Pollutants (phthalates and heavy metals) could accelerate cellular aging.
- Pollutants could lead to cell death.

However, further analysis is needed to clarify the biological enrichment, see the actual difference between conditions M1 and M2. It also remains to see relationships with diseases.

V - Bibliography

1. HEALS article on going redaction
2. Karine AUDOUZE courses