Effects of Vanlafaxine and MDMA on differents brain regions of Dark agouti rats

University of Trento 2019

Data Analysis and Exploration

Sébastien CARARO

207 984

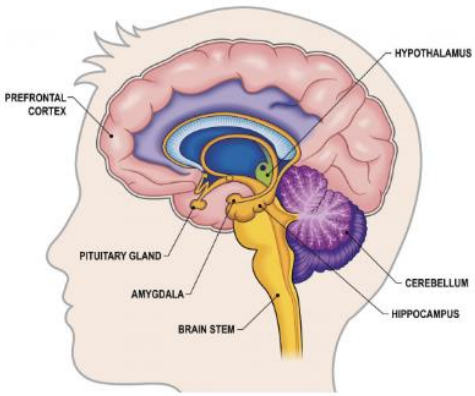
Table of contents

I – Introduction and presentation of the dataset 2
II – Exploratory analysis 3
PCA 3
K-means 4
Hierarchical clustering 6
III – Effects of MDMA : study on the three zones 7
Random Forests and heatmaps 7
LDA 9
SCUDO 10
Functional enrichment analysis 12
Network Analysis with Cytoscape 15
IV – Effects of Vanlafaxine : study on the three zones 16
Random Forests and heatmaps 16
LDA 18
SCUDO 19
Functional enrichment analysis 21
Network Analysis with Cytoscape 23
V – Other considerations and conclusion 23
Could we distinguish between zones ? 23
A brief study of the combined effects of MDMA and VLX combined (ex_TEST) 24
Conclusions 25
VI – Appendices & references 26

Following professor's instructions, this report has been tried to be summed up and to contain the very relevant information. It was aimed to be around 15 pages long, which resulted in a choice of information to display. In particular the introduction and the "other perspectives" parts were cut off in order not to get over 25 pages. Also, the codes and additionnal results will be provided with this report.

I – Introduction and presentation of the dataset

MDMA is a psychoactive drug primarly used as a recreational drug. The effects may include euphoria, happiness and self-confidence but to a long-term use it can lead to major depressive episodes, brain damage and cardiac problems. Venlafaxine is one of the most prescribed antidepressant worlwide (sold as "*Effexor*"), it is widely used to cope with depression and is widely legal and prescribed by doctors. However, multiple cases of undesirable effects and overdose incidents have been observed with this medication. A priori, one common point between these two drugs is the fact that they artificially feed the serotonin circuit inside the user, creating a temporary state of happiness and satisfaction. Anyway, it has been proved that such drugs can actually aggravate illnesses such as depression.



Three brain regions were studied during the experiments, each one of them has specific functions in the brain :

- Frontal Cortex : Intelligence, concentration, self awareness, Personality, planning and problem solving.
- Hippocampus : Memory, emotions, space representation, behavior.
- Dorsal Raphe : Serotonergic nucleus, narcolepsy.

The hippocampus shrinkage is linked with disease such as Alzheimer or depression [3].

The dataset GSE 47541 is composed of 22 523 genes observations, measured on 46 different Rattus Norvergicus subjects. The experiments aimed to study the effects of the two substances which are MDMA and Venlafaxine, plus also to capture the effects of the combined treatment. The MDMA injection consists in a single dose of 15 mg/kg whereas the Venlafaxine treatment consisted in a 3-weeks long regular injection of 40 mg/kg. The control subjects were injected Saline instead of toxic substance, in order to distinguish the effects of the injection in itself from the effects of the treatment. The measures were obtained via transciption profilling via array, and from the total 48 pooled samples 2 were excluded because of quality control reasons.

The subjects were split into 4 treatments :

First choice \ Second choice	VLX	SALINE
MDMA	MDMA/VLX	MDMA/SALINE
SALINE	SALINE/VLX	SALINE/SALINE

In this report the 4 treatments are named SALSAL (=control subjects), MDMASAL, SALVLX, MDMAVLX. The samples were renamed in the form *Region_Treatment_n°.* As the dataset is split into 4 treatments and into 3 zones, it is quite difficult to raise general models for classification as the training set for each method is limited, however we will see that the methods are accurate and the features selected correspond each time with the expected cognitive functions.

One might notice that the main purpose of the study is to determine the consequences of these two drugs, so the Functional Enrichment Analysis is the principal point of the study, and we will use the classification methods to determine the accurate genes. The goal corresponds to real life problematic, as we are more interested in the functions threatened rather than to simply determine which drug was used.

The main question I was trying to answer was :

In which way do MDMA and Venlafaxine expose their users to brain damage, and should we be concerned about the wide prescription of Venalafaxine ?

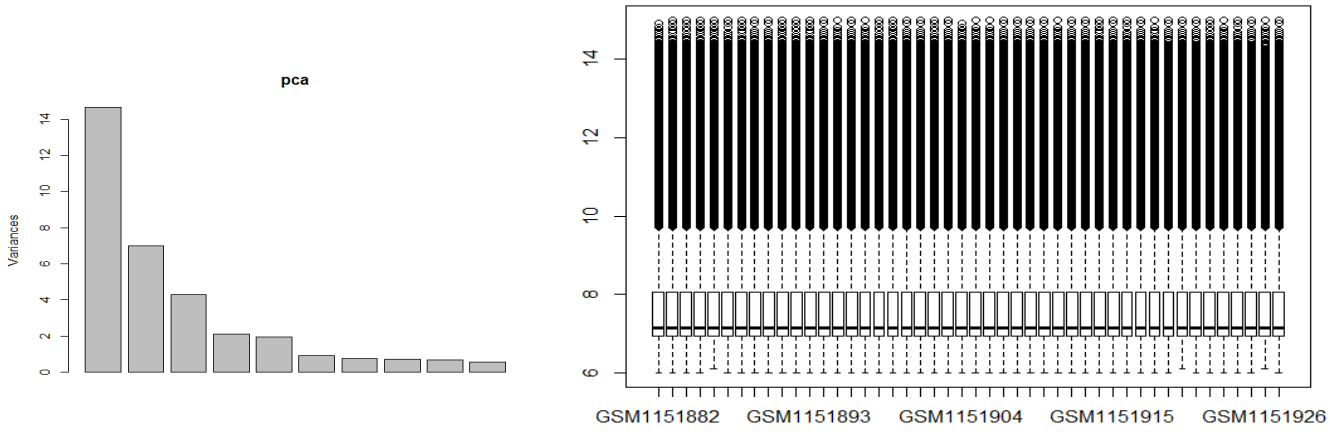
The process of study of the presented dataset consists first in an exploratory phase – involving methods such as PCA, K-means and Hierarchical clustering – in order to get a clear overview of the data. This phase is rather subjective but

was useful during the project in the sense that it guided the future paths of study. Then, considering both the division into three zones and the division into treatments, I decided first to focus on each asministrated drug separately to perceive the inherent effects, and draw conclusions. Finally I also focused on the cumulativity of the treatments and on comparing the results.

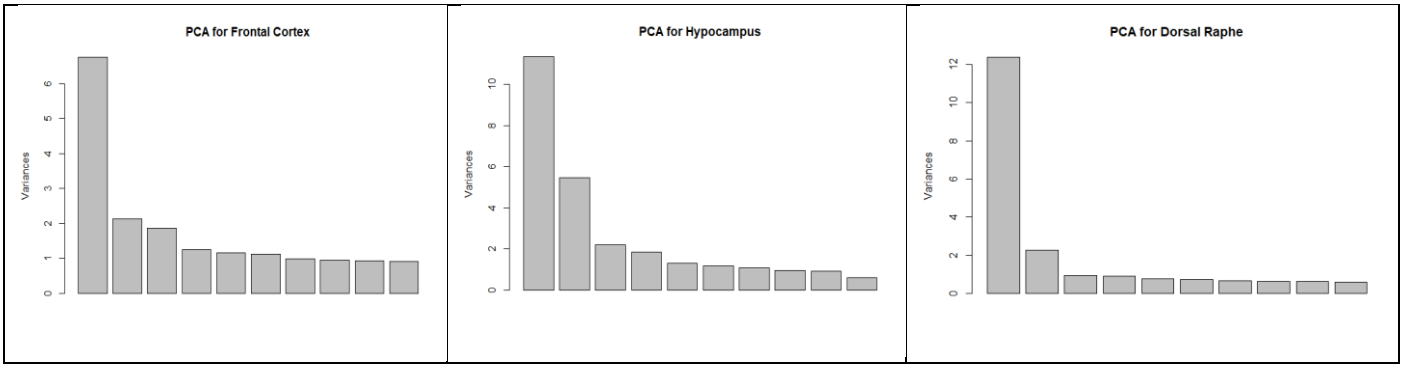
II – Exploratory analysis

PCA

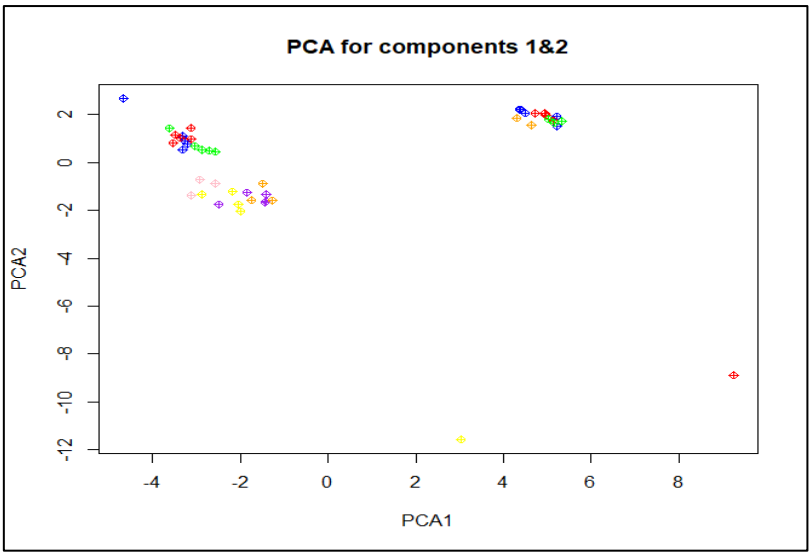
As a first step into our analysis we can perform PCA on the dataset and observe the decreasing of the variance explained by each Principal Component :



A huge part of the variance is contained in the first 3 Principal Components, which is a good sign for the pursuit of our analysis. Furthermore, when we perform PCA on each specific zone we can notice that the Dorsal Raphe region yields a very relevant result.

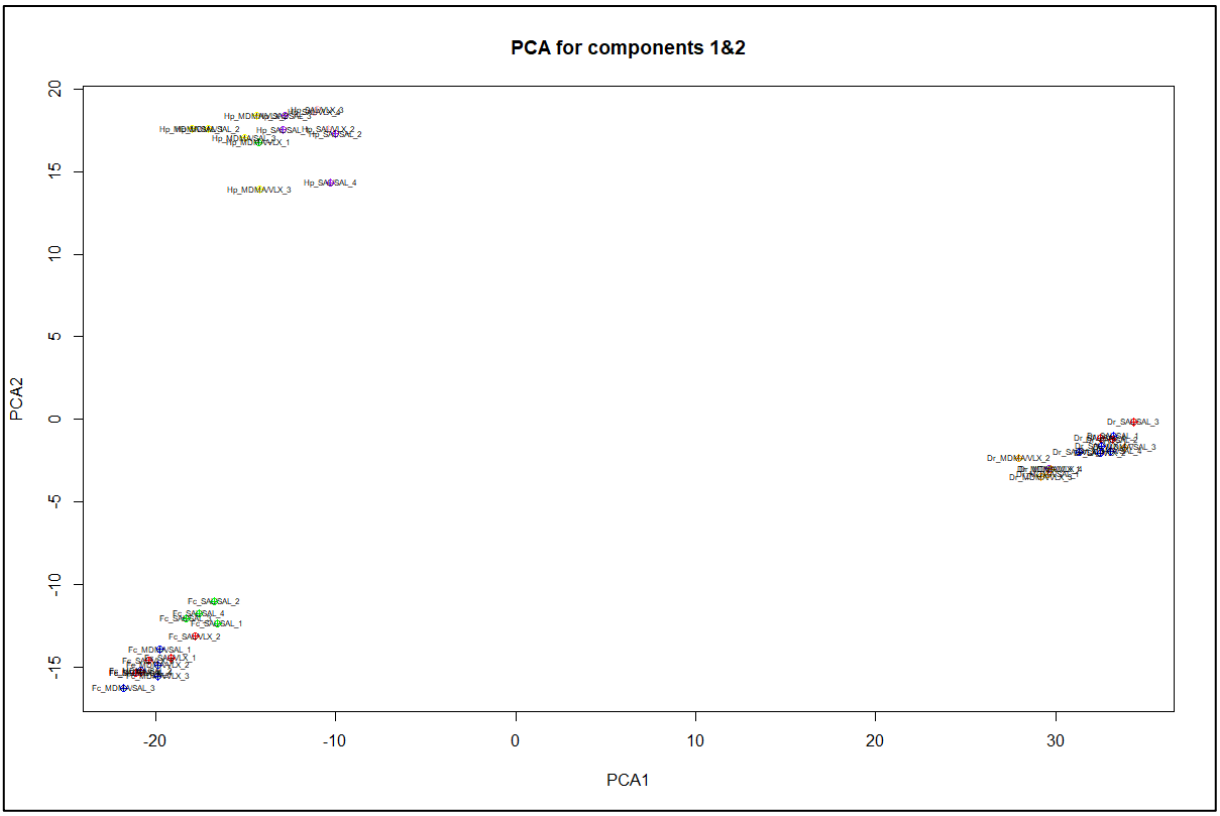


Variance explained by Principal Component, on every region



PCA on the whole dataset

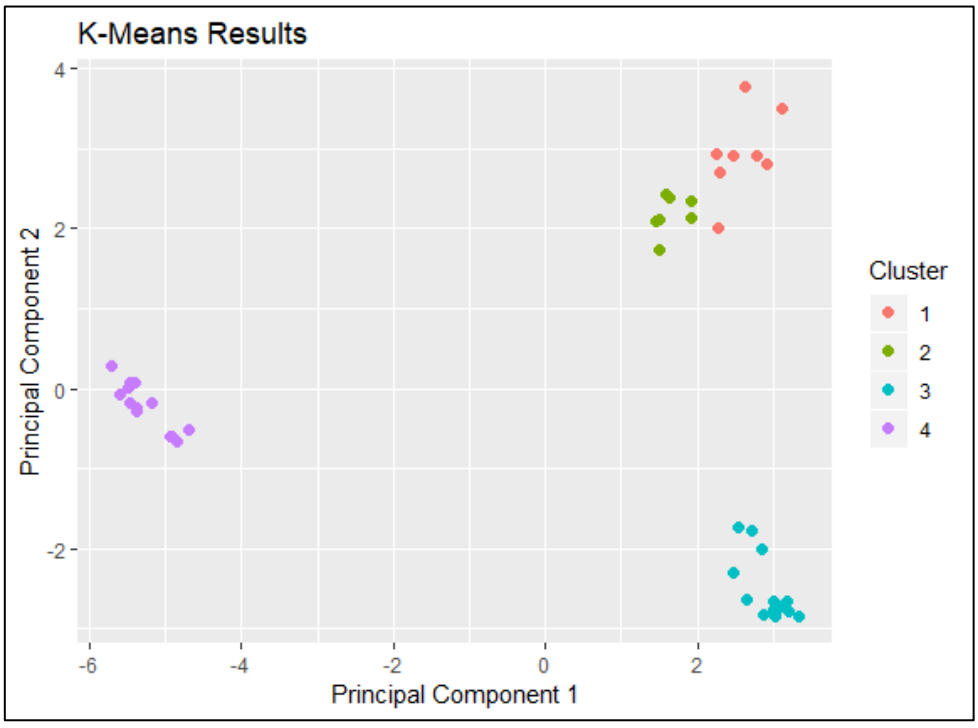
Then after removing a couple outliers we can perceive very clearly the separation between the three zones of study.



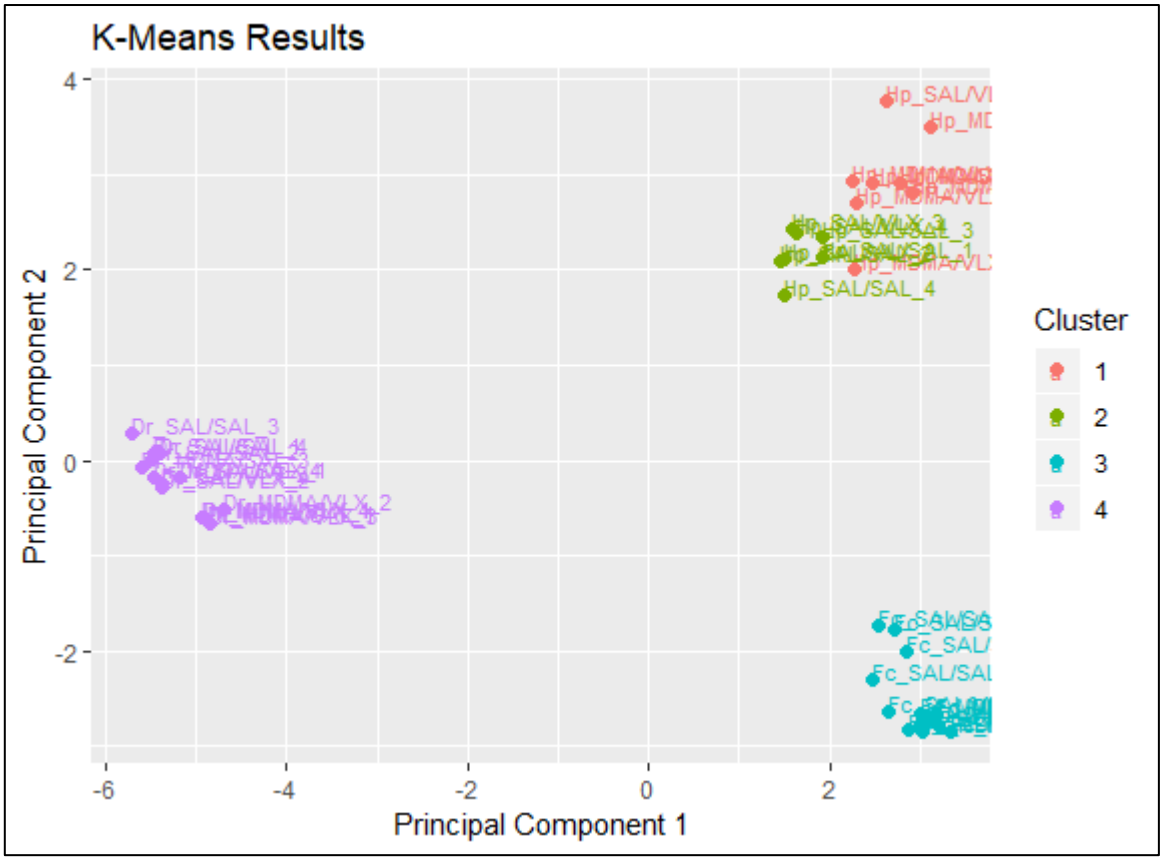
PCA on the whole dataset, after removing a couple outliers

K-means

Using the PCA, if we perform K-means on the whole dataset the samples are just clustered by zone. This unsupervised method can't accurately distinguish the treatments when we perform it on the whole dataset, it is intuitively normal as it is not the sames genes which are affected depending on the zone.



K-means on the whole dataset, without labels



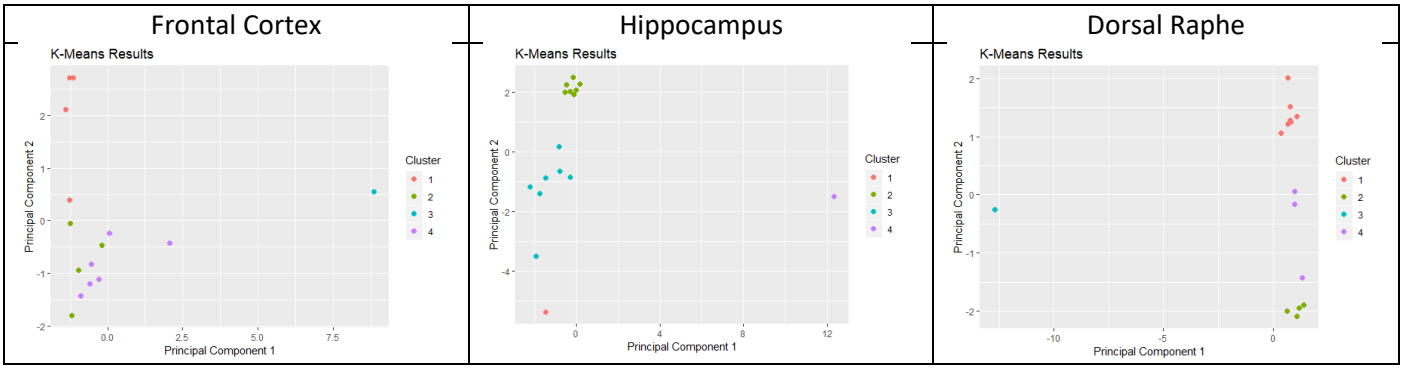
K-means on the whole dataset, with labels

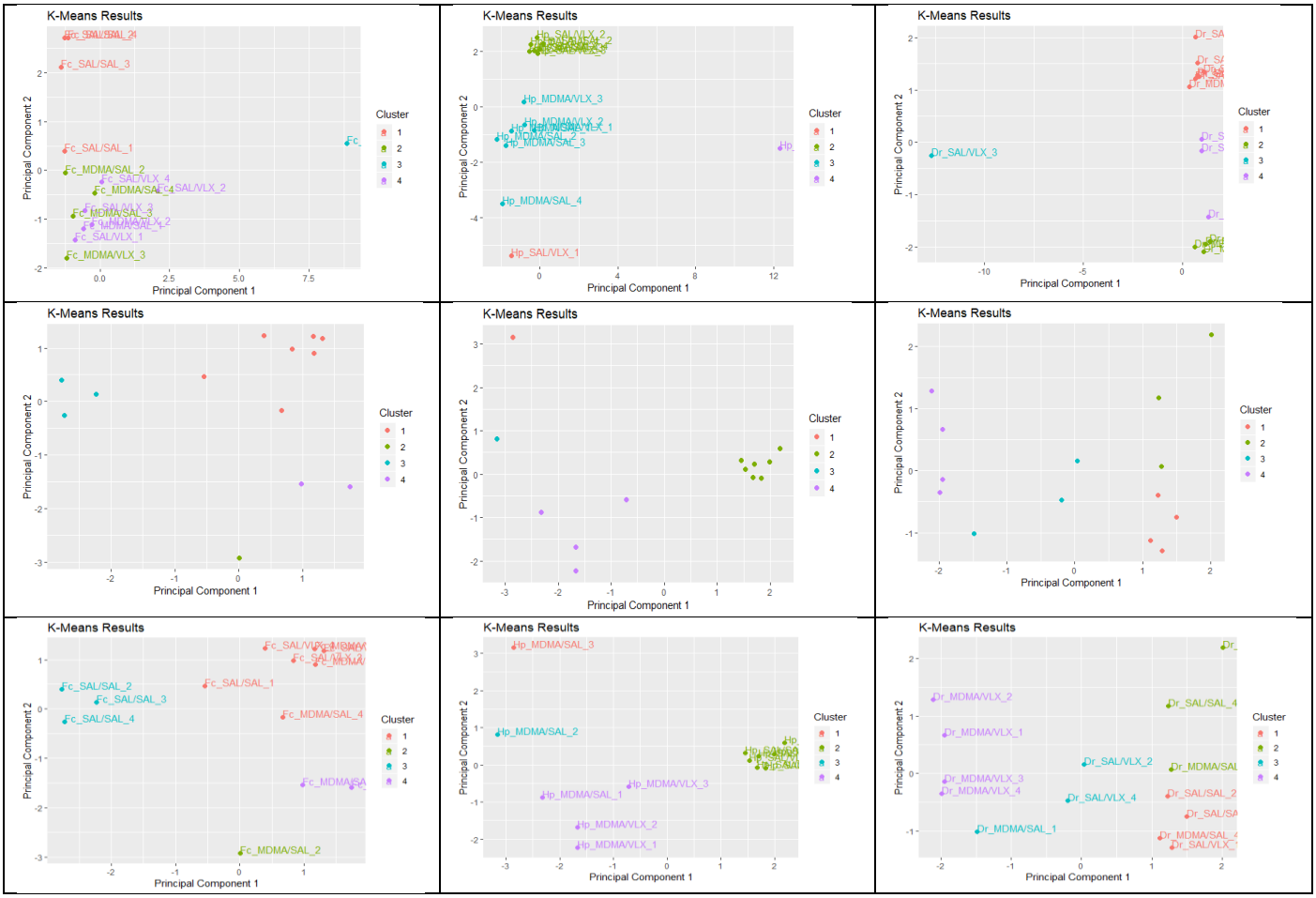
When we perform K-means on each specific region (k=4), we can notice some clusters and we see that the different treatments are distinguished with this unsupervised method. However, we also see a lot of overlappings and the presence of outliers is falsing the results. After removing the outliers (rows 3&4 in the table) we can observe a clearer separation between treatments. One constant observation is that the saline samples are well separated from the other subjects.

We can notice that for the Dorsal Raphe region, the treatments are very clearly distinguished, with no overlapping between subjects. It is interesting because both drugs are activating the serotonin circuit, and the Dorsal Raphe is the largest serotonergic nucleus and provides a substantial proportion of the serotonin innervation to the forebrain. That's why the Dorsal Raphe is very clearly clustering the subjects, because it is the main zone of action for the drugs. The two other zones also yield clear distinction between the treatments, but the samples are sparsed even if we remove the outliers.

<u>Outliers removed by zone :</u>

- "Fc_MDMAVLX_1", "Fc_SALVLX_2"
- "Hp_MDMAVLX_4","Hp_MDMASAL_4","Hp_SALVLX_1"
- "Dr_SALVLX_3"

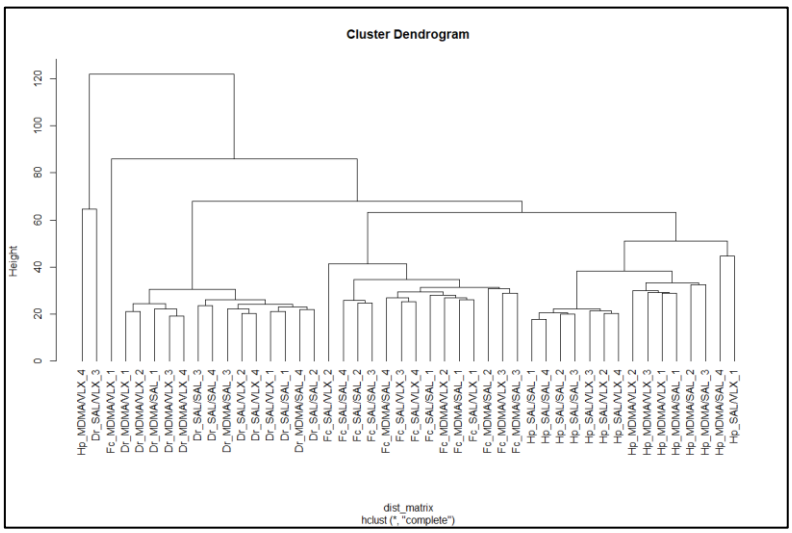




4-means algorithm performed on every region[Row 1&2 = with/without labels, rows 3&4 = after removing outliers]

Hierarchical clustering

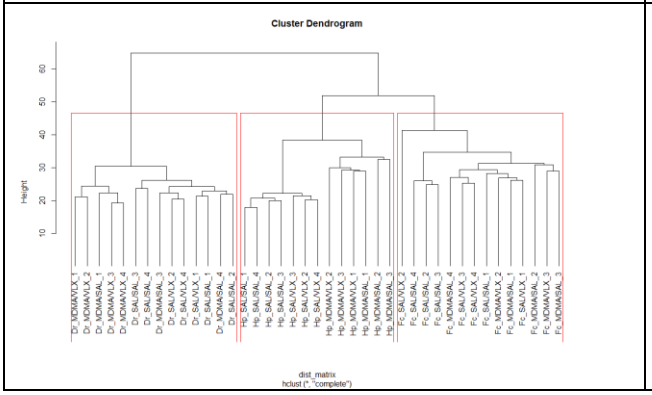
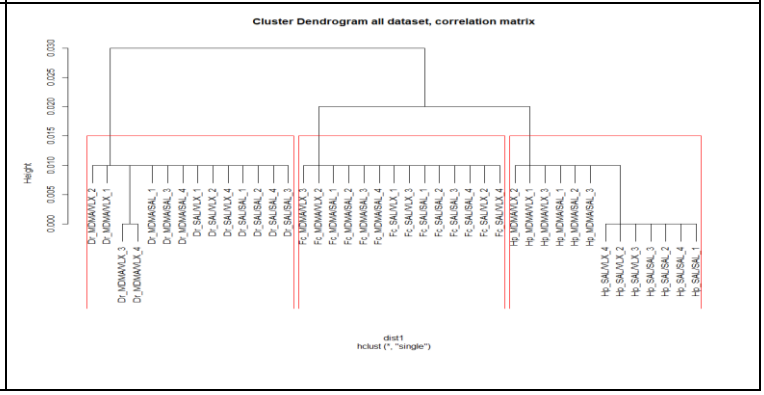
Another way to represent the data is to use a dendrogram, this method allows to visualize at once every possible clustering. Actually, we can notice the presence of the very same outliers which we identified during PCA, and we can also see that within the zones, the samples are often clustered by treatment. However these unsupervised methods are not enough precise to classify the data and that's why we will advance towards more evolved methods.



Hierarchical clustering with the whole dataset

After removing 5 outliers the classification is correct for k=3. Otherwise the outliers are clustered apart, creating overlaps between the three zones. This illustrates the fact that clustering algorithms such as hierarchical clustering are not robust to outliers. The dendrogram also allows to observe that within specific zones, treatments are often well identified are clustered together (we see the similarity between samples on the vertical axis*, i.e.* if two samples are joined at a low height it means they are closely related).

Then after removing the outliers, we can have a clearer representation, using either euclidean distance or correlation matrix :

Euclidean Distance	Correlation Matrix
	

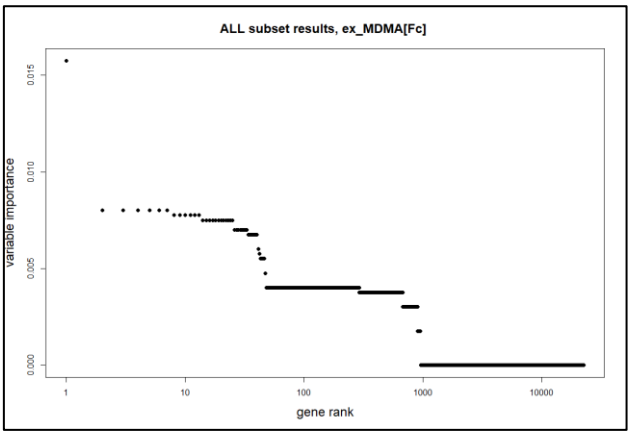
III – Effects of MDMA : study on the three zones

For the prupose of studying the inherent effects of the MDMA molecul, we first focus on the treatments MDMASAL and SALSAL (control subjects). We will first compare the samples in every specific region, then we will study the difference of effects of the MDMASAL treatment among the three differents zones. We will use the Random Forest algorithm, then the LDA and the SCUDO software. We will perform features selection at some point, in order to guide the methods and hence to yield more accurate results. However, one has to be careful about the value of the p-value for features selection, as a low value could result in overfitting the data.

Finally, we will study the selected features via each method, in order to determine which cognitive functions are affected in each brain zone. For that we will use the DAVID website and the JEPETTO plugin for Cytoscape.

Random Forests and heatmaps

In order to represent the Random Forest classification in a heatmap, we first take a look at the variable importance plots in order to choose the accurate number of genes :



Variable importance plot concerning Random Forest, for Frontal Cortex

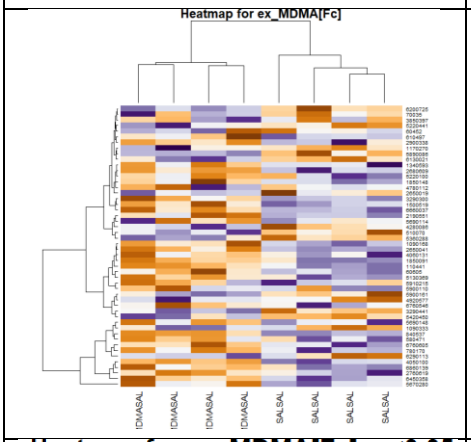
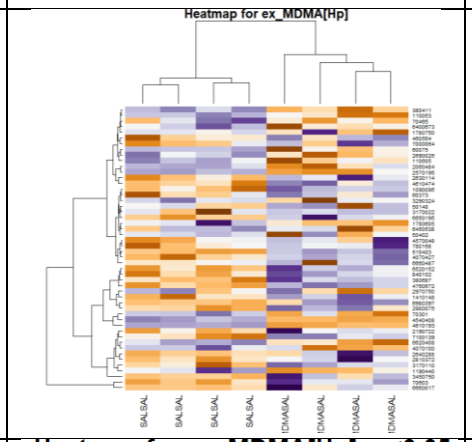
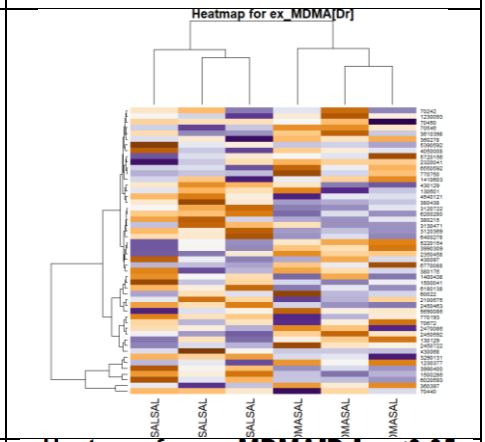
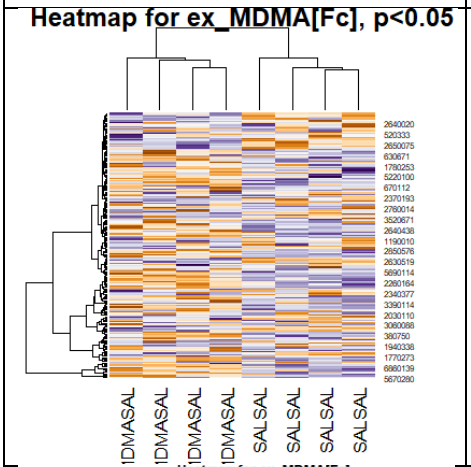
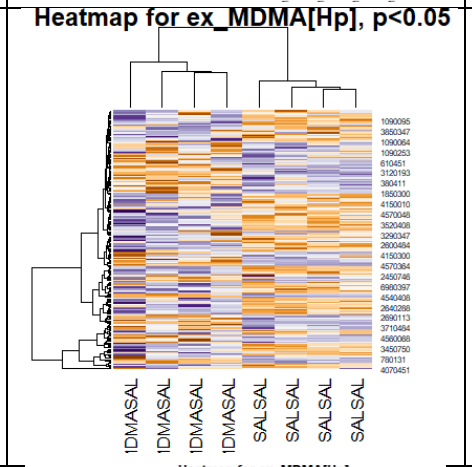
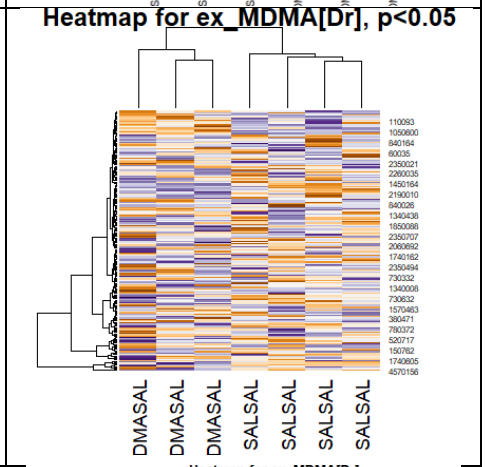
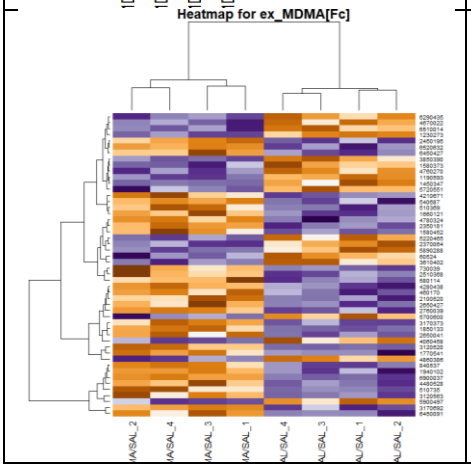
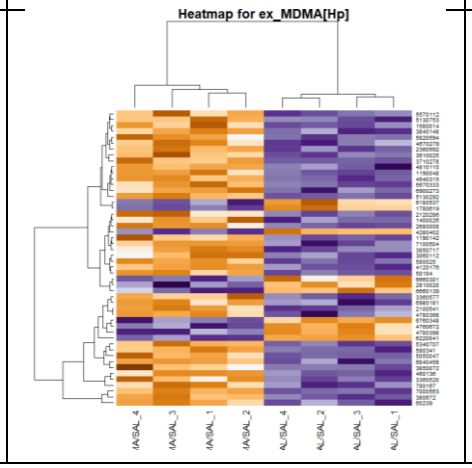
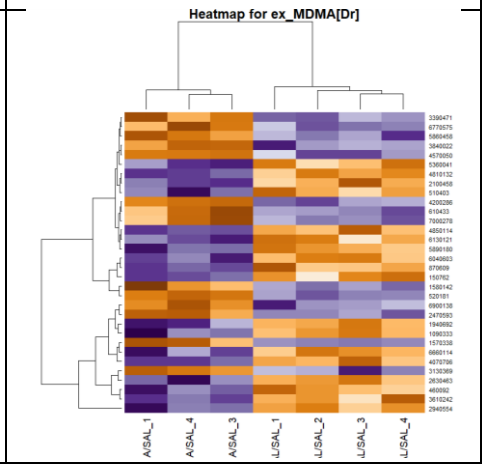
We can notice that there is an elbow at around 50 genes, so we decide to keep 50 genes for selection. The two other zones yielded similar results concerning this plot, with an elbow at around 50 genes. The Random Forest algorithm usually chooses among the square root of the numbers of features at each node. To get multiple results, we decide to perform the algorithm with different thresold values for features selection. The risk of having a too low value is being confronted to overfitting the data, which is really a risk in our dataset because we don't have so many samples to check our fitted models on. We will then check the accuracy of the features selection in the functional enrichment part. For the optimal value of the threshold ($p<0.005$), we will collect the 200 genes so that we have more genes to study in the functional enrichment analysis.

Thereafter are displayed the result of the heatmap for three parametrizations :

- No features selection : 22523 features
- Features selection : p<0.05. Around 1500 genes left by region. It is the oprimal thresold value for p.
- Features selection : p<0.001. Around 100 genes left by region .

On the following plots one needs to pay attention to the name of the columns in the heatmaps, as the order treated/control is often changed.

Interpretation of the results : Comparing the treated subjects and the saline subjects, we can see that for Frontal Cortex and Hippocampus, the MDMA figures show an up-regulation of the most imprtant genes, while in the Dorsal Raphe there are about half of the important genes which are down-regulated and the other half which is up-regulated. The heatmap without features selection (row1) is not very convincing, while the results with a lower p-value displays more contrast between the treatments. While the models without features selection yield OOB error rates of respectively 12%, 25% and 8%, the models using features selection yield an error rate of 0%. This could be exlained by the fact that our number of samples is very limited (4 treated samples & 4 control samples per zone), and so we could be overfitting the data. However, we will see in the functional enrichment analysis that the selected genes correspond to the affected function in each zone. We can also notice that it is rarely the same genes which are affected in the different regions of study.

Frontal Cortex	Hippocampus	Dorsal Raphe
		
		
		

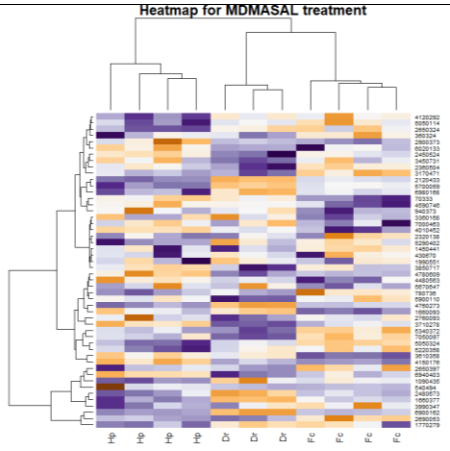
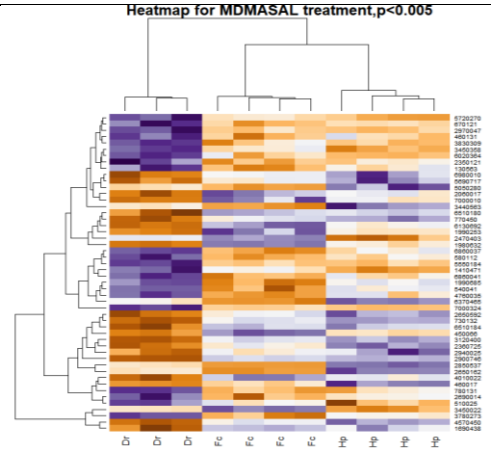
Heatmaps produced after we perform Random Forest on the three zones, with different p-value for features selection : No features selection, p<0.05, p<0.001

Then we can also split the class into the three zones of study, in order to see which genes are modified on specific regions. For this we consider only the subjects affected by the treatment MDMASAL, in order to see the differences of effects among the brain regions. Without any surprise the classification is perfect – which was obtained by unsupervised methods too.

Distinction between zones : We consider the MDMASAL treatment over the tree zones (12 samples). As we already saw during the exploratory phase, clustering the samples based on the zones is not very difficult, as the genes affected by the treatment are not the same : the functions involved in the reaction to the drug differ according to the region.

```
> rf

Call:
 randomForest(x = t(small.eset), y = as.factor(group), ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 149

        OOB estimate of  error rate: 0%
Confusion matrix:
   Dr Fc Hp class.error
Dr  3  0  0           0
Fc  0  4  0           0
Hp  0  0  4           0
```

Summary of the Random Forest algorithm to distinguish between the zones in the MDMASAL treatment

Without features selection	Features selection p<0.005
	

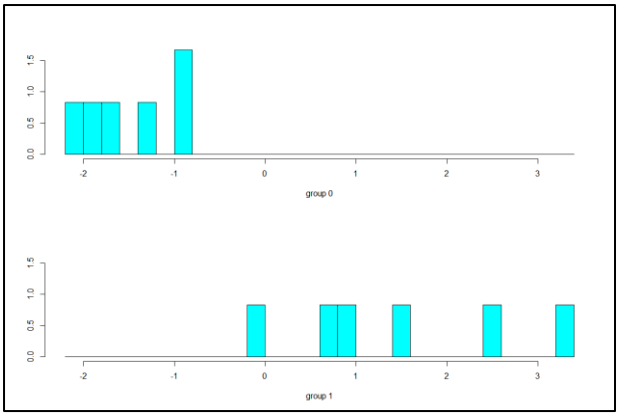
Heatmaps produced after we performed Random Forest on the MDMASAL treated samples

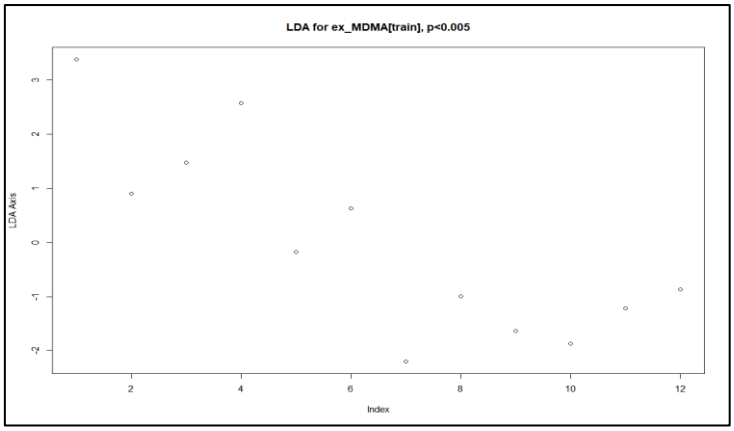
LDA

The interest of the LDA method is limited in the context of our study, as it only classifies the samples but does not yield any information about the relevancy of genes. We did performed it on the three regions concerning the MDMA treatment, but as there was not enough samples then we could not draw accurate conclusions concerning its classification power. Furthermore, the Random Forest algorithm already yielded very accurate results concerning classification. Instead, we decided to perform LDA on the MDMASAL & SALSAL subjects, considering all of the three zones at once. This made a total of 22 samples, we then took 12 for training and 10 for testing – the class were equally partitioned. After features selection the results seemed impressive, as the two classes were clearly delimited. The model was then able to predict the treatment, independently from the origin of the gene sample.

We can reasonably think the classification was accurate – and that we weren't overfitting the training data – because the classification performed perfectly on the test samples : there are a few genes changes which are really characteristic of the treatment.

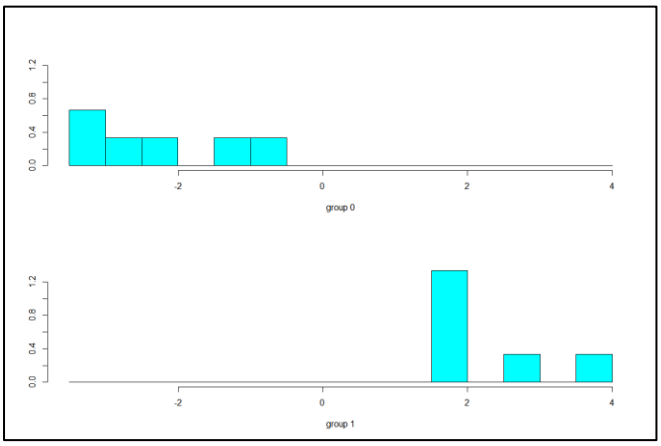
P< 0.005 : There were only 144 genes left after the features selection

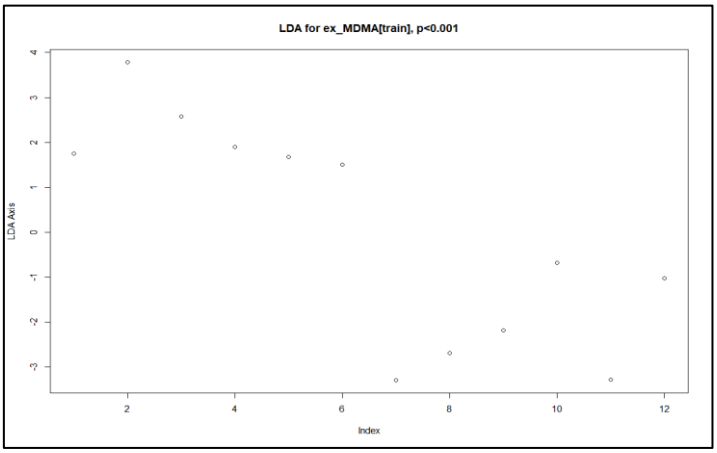





```
> preds$class      #predictions
 [1] 1 1 1 1 1 0 0 0 0 0
Levels: 0 1
> dat$Y[test]      # real values
 [1] 1 1 1 1 1 0 0 0 0 0
```

P<0.001 : there were only 26 genes after the features selection.





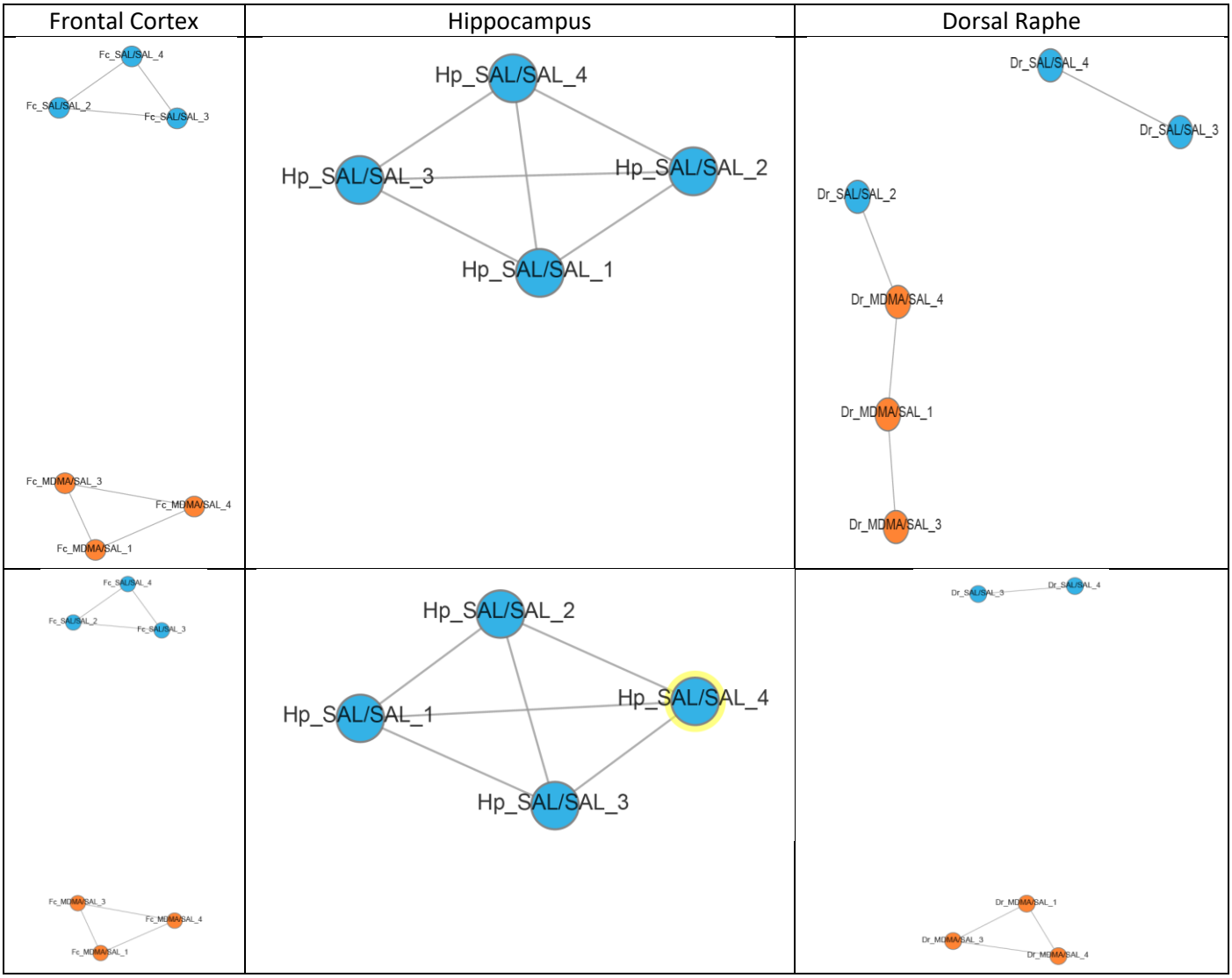

```
> preds$class      #predictions
 [1] 1 1 1 1 1 0 0 0 0 0
Levels: 0 1
> dat$Y[test]      # real values
 [1] 1 1 1 1 1 0 0 0 0 0
```

SCUDO

We will use SCUDO for 2 main purposes :

1. Obtain a graphical representation of the samples, which reveals the closeness between clusters but also among clusters
2. Obtain the most important genes for classification, which we will study in the Functional Enrichment Analysis

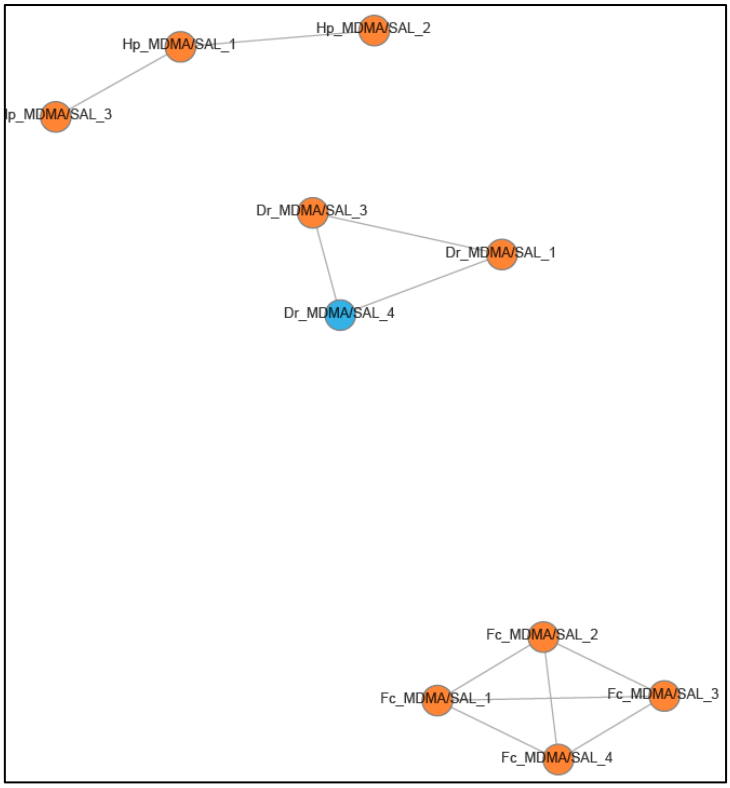
We then used SCUDO in unsupervised mode (p<0.99) and in supervised mode (p<0.05), the parameters were n1=n2=250, N% = 20. The unsupervised mode is more likely to display all the samples, whereas the supervised mode tends to exclude some samples in the network if they don't fit enough. On the other hand the supervised mode will yield more accurate signatures. We performed the algorithm on each region, the first row of the following table correponds to the unsupervised mode. We can notice that the plots are quite similar, however in the Dorsal Raphe the unsupervised mode created some overlapping between the treatments.



SCUDO networks by zone, comparing the MDMASAL & SALSAL treatments

Then we can study the MDMASAL treatment in unsupervised mode in order to seek for the interesting genes inter-zones, such that we can collect the signatures genes to analyze them in Functional Enrichment Analysis. We can oberve that the three brain zones are well separated on the graph, with no overlapping.

NB : As SCUDO only asks for control/ill subjects and that the graphs contains 3 clusters, I had to arbitrarly choose one sample as control subject (blue dot) but the display colour is meaningless.



SCUDO graph for the MDMASAL treatment

Functional enrichment analysis

Random Forest Genes : We need to precise that as the original dataset depicted the genes via unregistered names, I had to do a translation in order to get the genes in "Symbol" nomenclature. Unfortunaly some features couldn't be exploited as they were not repertoried into the Symbol database (they were called LOCxxx and often they could not be identified by the DAVID website).Then it was possible to perform Functional Enrichment analysis on DAVID. The genes used were the output of the RandomForest algorithm*, i.e.* the genes that were the most chosen in nodes during the construction of the trees. They are then considered as the most relevant to distinguish between treatments (here MDMASAL vs SALSAL).

SCUDO : We easily obtain the signatures in both down-regulation and up-regulation by accessing the signatures file produced by the software.

I used the different signatures in order to do a cross-validation of the result, a very good sign is that the results match, so we can positively draw safe conclusions. There down is a recap of the main functions returned when we enter the signatures in DAVID and in JEPETTO (to get a more precise view one can enter the signatures attached to the project directly onto the website, but the main functions are listed down there). Many of the functions were repeated, the table tries to show the spectific functions of each signatures set too. The results are commented just under the table.

MDMASAL vs SALSAL	Frontal Cortex	Hippocampus	Dorsal Raphe
Random Forest P<0.005	- Regulation of cell death, cell projection - Glycosylation, Glucose metabolic processes, carbohydrate processes - Regulation of synaptic plasticity, nerve impulse - Sensory organ development (inner-ear,ear, olfactory activity, sexual reproduction) - Inflammatory response, defense process	- Heart development - Digestive system, gut development - Respiratory system, lung development - Immune system response, defense - Blood vessel development - Regulation of cell death, apostosis - Protein kinase, kinase, ATP binding, synapse - Response to stress, neurotransmittor binding - Memory	- Inflammatory response, stimulus - Cell death, apostosis - Calcium, ion transport - Dopamine metabolic process - Synasptic plasticity, synapse, response to stimulus, nerve impulse, kinase, serine - Olfactory receptor, learning - Sexual reproduction
SCUDO P<0.01, Up-regulated	- Cell projection - Protein kinase, ATP binding, prostate cancer, serine protein kinase - Regulation of neurogenesis and of cell death - Long-term depression - Vesicle and synapse - Olfactory receptor activity	- Synapse, vesicle - Pancreatic , colorectal cancers. Diabete - Cell mobility - Protein kinase, synaptic transmission - Cell death, apostosis - Reproductive structure, olfactory receptor activity	- Synapse, vesicle, cell motion & migration, kinase - GTPase activity, GTP binding, ATP binding - Melanoma, prostate, pancreatic, colorectal cancers. - Reproductive structure, sensory perception - Cell death, apostosis - Blood vessel
SCUDO P<0.01, Down-regulated	- Regulation of lipid metabolic process. Polysaccharide/carbohydrate binding - Cellular/ion homeostasis - Cell death - Response to wounding, defense response, inflammatory response	- Synaptic plasticity, transmission, synapse, kinase - Metal, ion, ATP binding - Reponse to insulin, hormone - Blood vessel development - Sexual reproduction, spermatogenesis, sensory reception, cognition	- Cell adhesion, development, neurogenesis - Synapse, dendrite, protein kinase, serine, vesicle - Immunoglobin - Insulin stimulus - Cell migration - Spermatogenesis

JEPETTO Using the Random Forests genes

Pathway or Process	XD-score
Starch and sucrose metabolism	0,35612
Glycosylphosphatidylinositol...	0,26088
Porphyrin and chlorophyll m...	0,21699
Selenoamino acid metabolism	0,21046
Pentose phosphate pathway	0,17516
Amino sugar and nucleotide ...	0,13516
Base excision repair	0,10016
DNA replication	0,09638
Vasopressin-regulated wate...	0,08327
Lysosome	0,08183
Aldosterone-regulated sodiu...	0,08043
Small cell lung cancer	0,07273
Nucleotide excision repair	0,07040
Vibrio cholerae infection	0,07040
Glycolysis / Gluconeogenesis	0,07040
Basal cell carcinoma	0,06607
Hedgehog signaling pathway	0,06607
Amyotrophic lateral sclerosis...	0,06027
Amoebiasis	0,05680
Cardiac muscle contraction	0,05209
Long-term depression	0,04534

Pathway or Process	XD-score
Other glycan degradation	0,36393
Sphingolipid metabolism	0,32756
Ether lipid metabolism	0,21393
RIG-I-like receptor signaling ...	0,18211
Shigellosis	0,17821
Glycerolipid metabolism	0,16393
Dorso-ventral axis formation	0,16393
Allograft rejection	0,12393
Fatty acid metabolism	0,11777
Adipocytokine signaling path...	0,10428
Type I diabetes mellitus	0,10186
Toll-like receptor signaling p...	0,09726
Pancreatic cancer	0,07821
Glycerophospholipid metabol...	0,07821
Hypertrophic cardiomyopath...	0,07504
Arginine and proline metabol...	0,07203
Bladder cancer	0,06919
Dilated cardiomyopathy	0,06782
TGF-beta signaling pathway	0,06660
Small cell lung cancer	0,06149

Pathway or Process	XD-score
Glycerolipid metabolism	0,35640
Nitrogen metabolism	0,24212
Proximal tubule bicarbonate ...	0,17863
Pentose phosphate pathway	0,15640
Starch and sucrose metabolism	0,14688
Phototransduction	0,14327
Olfactory transduction	0,12085
Amino sugar and nucleotide ...	0,11640
ErbB signaling pathway	0,09926
Long-term depression	0,09676
Phosphatidylinositol signalin...	0,08544
Long-term potentiation	0,08339
Gastric acid secretion	0,08339
T cell receptor signaling pat...	0,07514
Chronic myeloid leukemia	0,07396
Alzheimer's disease	0,07315
B cell receptor signaling pat...	0,07235
Glycerophospholipid metabol...	0,07069
Salivary secretion	0,06752
Vasopressin-regulated wate...	0,06451
Neurotrophin signaling path...	0,05650

Interpretation :

We can observe that among the three zones there is a development of apostosis and regulation of death cells. We also notice the synaptic behavior modification (kinase, synaptic plasticity, serine) and the involvement of the immune system, the nervous system reacts to some perturbation (inflammatory response, response to stimulus, defense process, response to wounding). Then there are some zone-specific functions which are involved :

- Frontal Cortex = Glucose, carbohydrate and lipid metabolic processes. Modification of the sensory circuits (inner-ear, ear, olfactory, sexual reproduction). Apparition of long-term depression. We also see Cardiac muscle contraction.
- Hippocampus = Some organs are involved such as Heart development (acceleration of heart rate), digestive system, respiratory system, blood vessel development. We also notice some diseases such as cancer cells (pancreatic, colorectal, bladder) and also diabete. Some cognitive functions are also altered (memory, sensory reception)
- Dorsal Raphe = We notice the activation of the dopamine metabolic process, but also the augmentation of cell transport (cell motion, cell migration) which we can link to the fact that Dorsal Raphe is a serotonin production region (highly activated during the treatment).

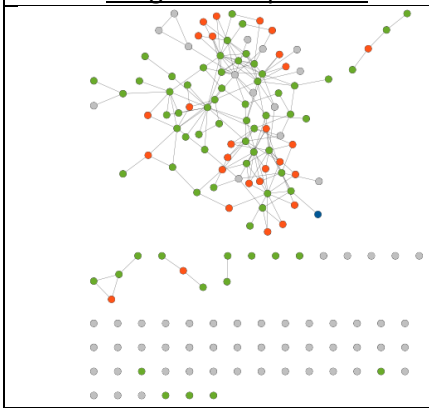
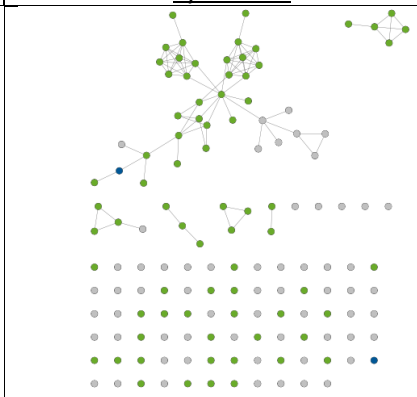
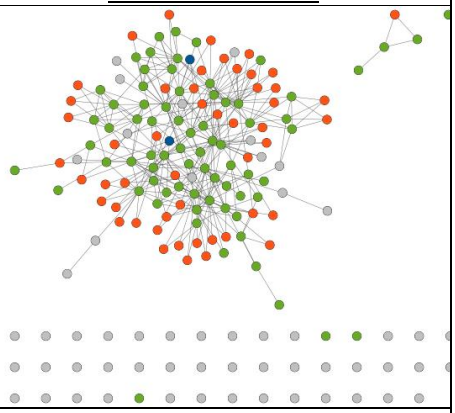
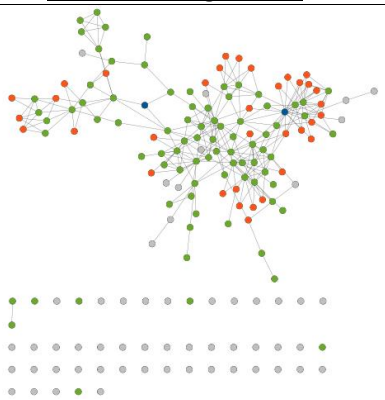
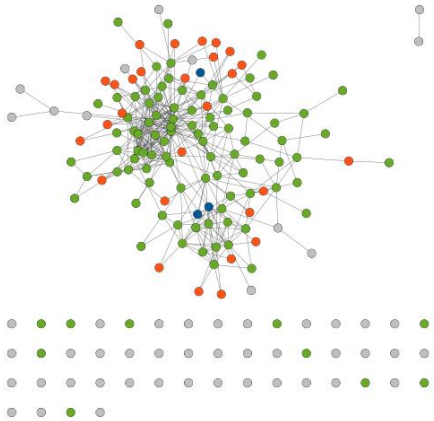
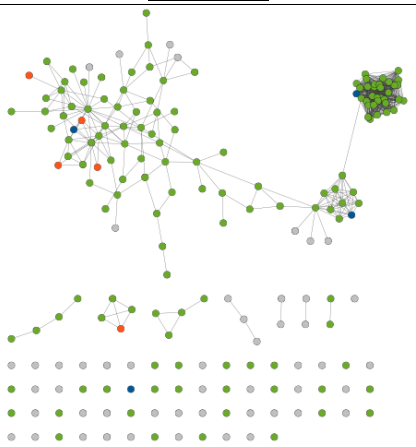
Distinction between zones :

	MDMASAL treatment only
Random Forests	- Cell surface, plasma membrane - Cell death, regulation of apostosis - Neuroactive ligand-receptor interaction - Signal, receptors, - Response to organic substance, glycoprotein

The signatures don't yield very interesting result, we only notice general functions which are involved.

Network Analysis with Cytoscape

When we enter the SCUDO networks into Cytoscape, it only display the already visualized networks, so we can use Cytoscape to visualize the characteristic functions involved in the treatment :

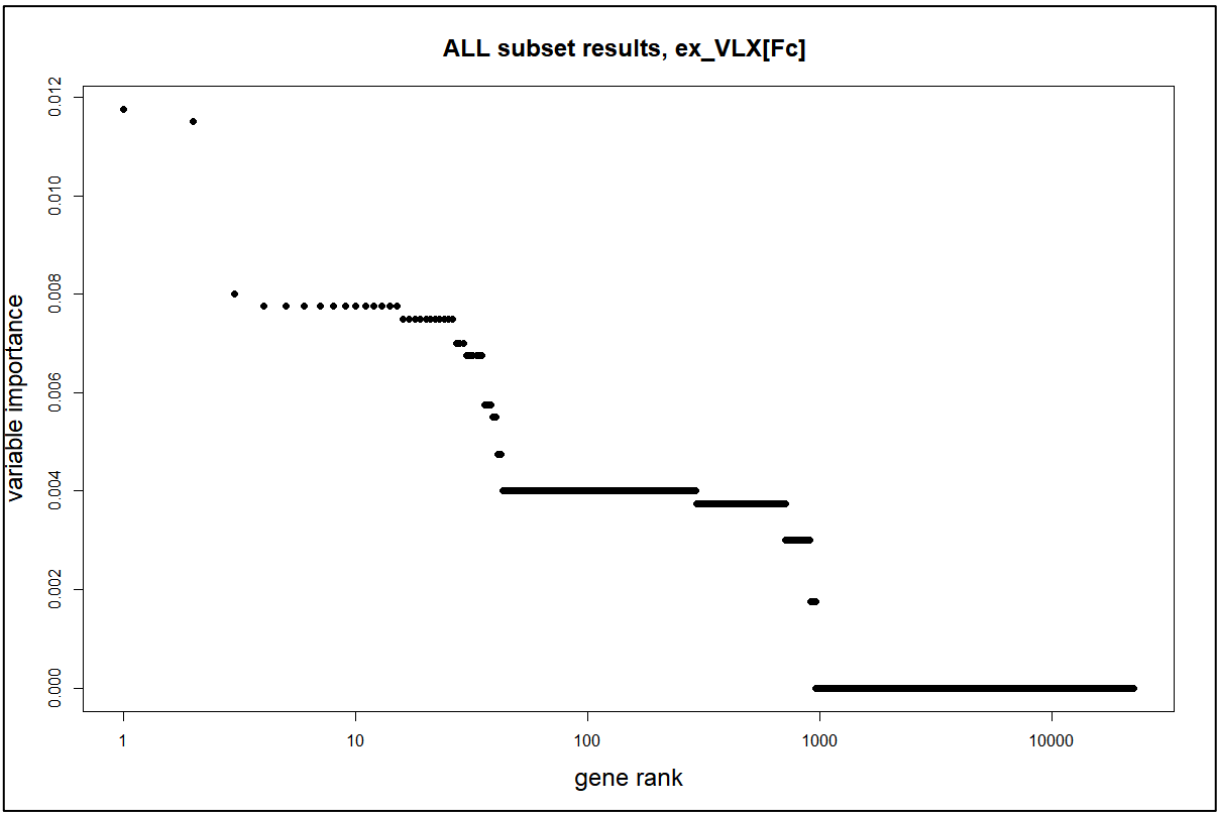
Long-term depression	Lysosome	Pancreatic cancer
		
Small cell lung cancer	**T cell receptor signaling pathway**	**Alzheimer**
		

IV – Effects of Vanlafaxine : study on the three zones

Here we repeat the same process of analysis that we used during the MDMA treatment study, by comparing the effects on each specific brain region. The Venlafaxine treatment was a bit different from the MDMA treatment, as the MDMA subjects took one single dose at the beginning of the experiment, while the VLX subjects took regular doses during the three weeks of experimentation. However, this was done to fit with the usual condition of utilization, as Venlafaxine is a daily antidepressant while MDMA users do drugs once in a while. The dataset we study here is composed of the treated subjects (SALVLX treatment) and of the control subjects (SALSAL treatment). We will perform Random Forest, LDA, SCUDO and finally Functional Enrichment Analysis based on the signatures provided by Random Forests and SCUDO. For that we will use the DAVID website to perform clustering gene analysis, and also the JEPETTO package in Cytoscape.

Random Forests and heatmaps

To use the Random Forests algorithm, we first run it once to display the gene importance plot :

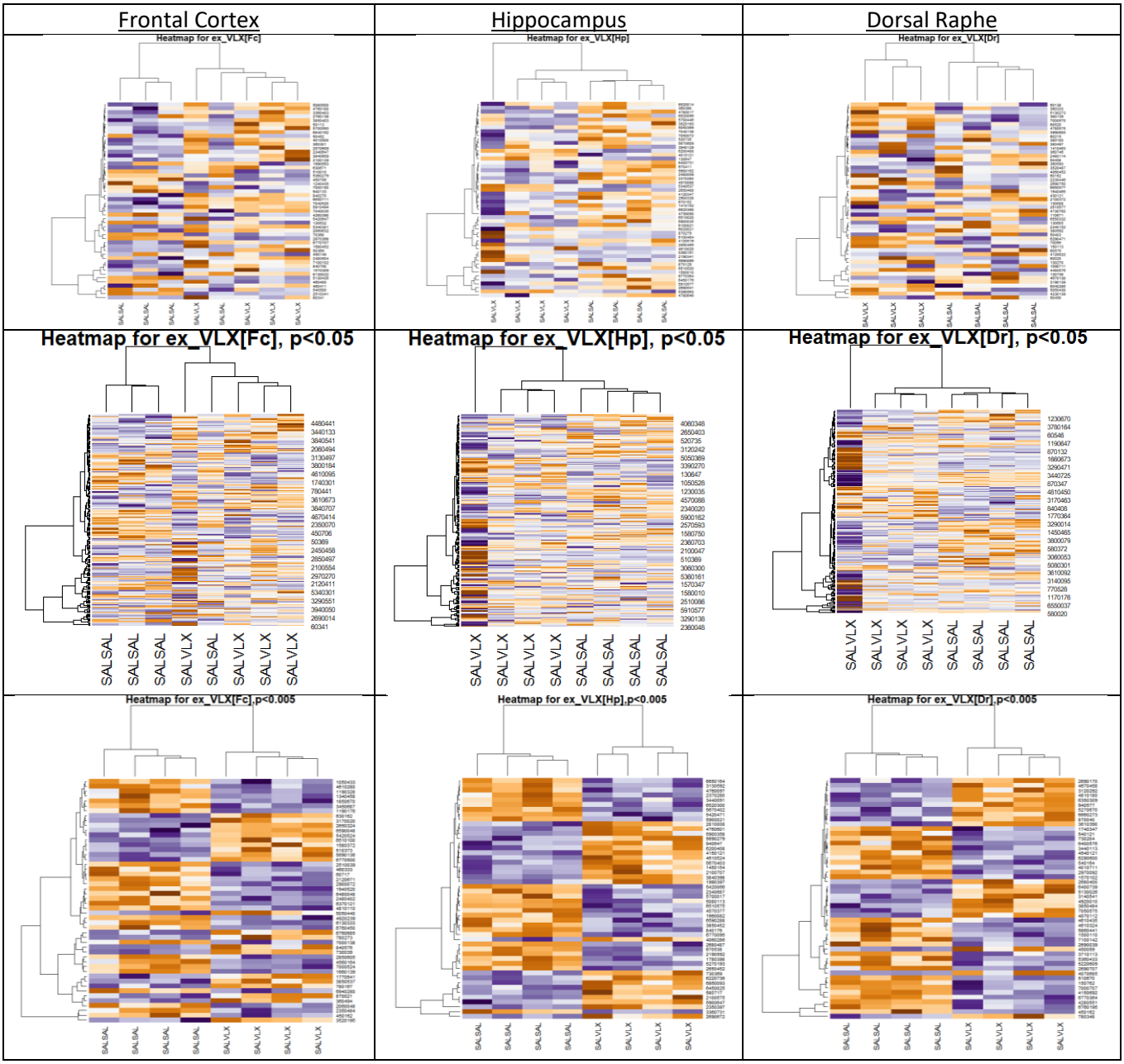


Variable importance plot after Random Forest on Frontal Cortex, for the SALVLX & SALSAL samples classification

We can notice that there is an elbow at around 50 genes, so we decide to keep 50 genes for selection.We took 200 genes for the parametrization p<0.05, as it corresponds to the best features selection, so that we could have more genes to analyze in the Functional Enrichment Analysis (due to the structure of the data, some signatures couldn't be interpreted as they were named LOCxxx and not recognized by DAVID). The two other zones yielded similar results concerning this plot. Thereafter are displayed the result of the heatmap for three parametrizations :

- No features selection : OOB error rates 37.5%[Fc], 75%[Hp], 42.86%[Dr].
- Features selection : p<0.05. OOB Error rates 37.5% [Fc], 75% [Hp], 50% [Dr]
- Features selection : p<0.01. OOB error rates 0%

With features selection : (p<0.01). The results are very interesting and we have about 100 genes each times – which is close to perfect . We can notice that the number of genes after features selection (for the same value of threshold) is way lower than for the effects of MDMA – we could interpret this as the idea that MDMA affects more specific genes than VLX.



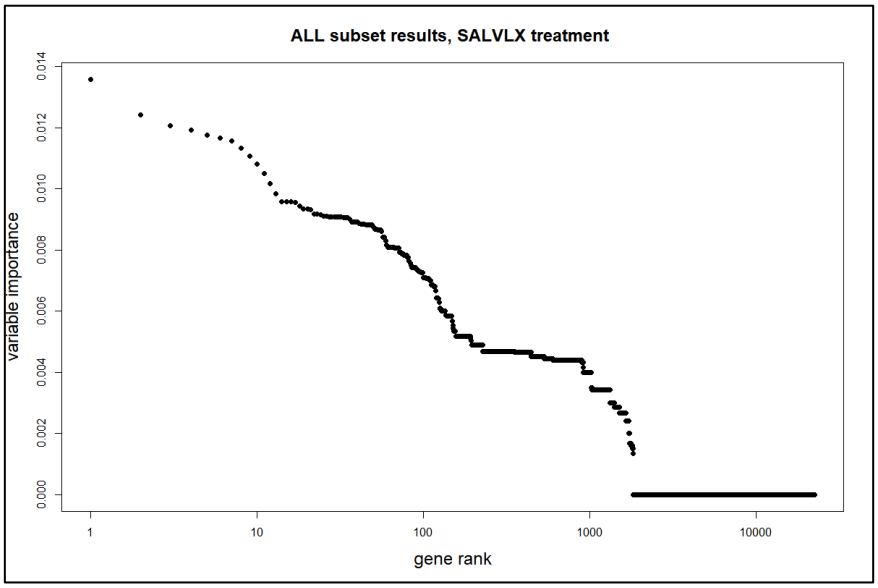
Heatmaps produced after we perform Random Forest to distinguish between SALVLX and SALSAL treatments, by zones (rows correspond to no features selection, p<0.05, p<0.01)

We then performed random forests on the SALVLX samples, in order to cluster the samples according to the zone of study. The OOB error rate is very low, which is not a surprise as zone identification is not difficult :

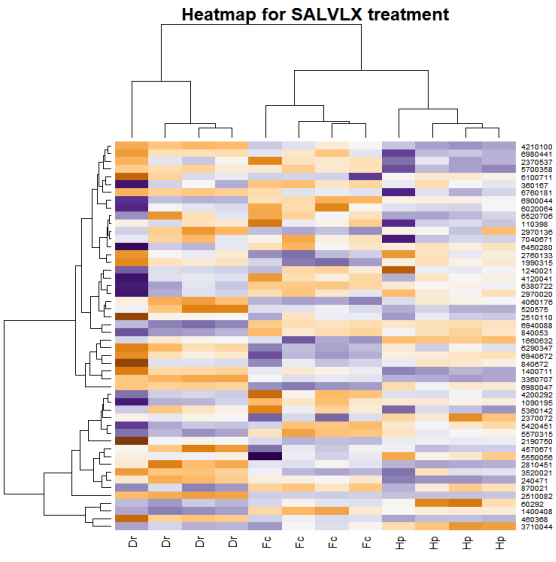
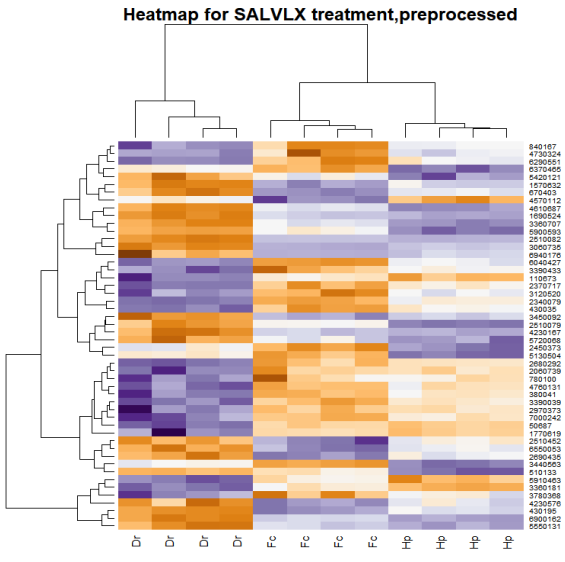
```
> rf

Call:
 randomForest(x = t(small.eset), y = as.factor(group), ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 149

        OOB estimate of  error rate: 8.33%
Confusion matrix:
   Dr Fc Hp class.error
Dr  4  0  0        0.00
Fc  0  4  0        0.00
Hp  0  1  3        0.25
```

Summary of the Random Forest to distinguish zones for the SALVLX treatment

Considering the following features importance plot, I chose to represent the first 90 features. With the features selection the curve is quite similar so I chose 90 genes for the heatmap aswell.



We can notice that the Dorsal Raphe regions shows the most extreme variations among the selected genes :

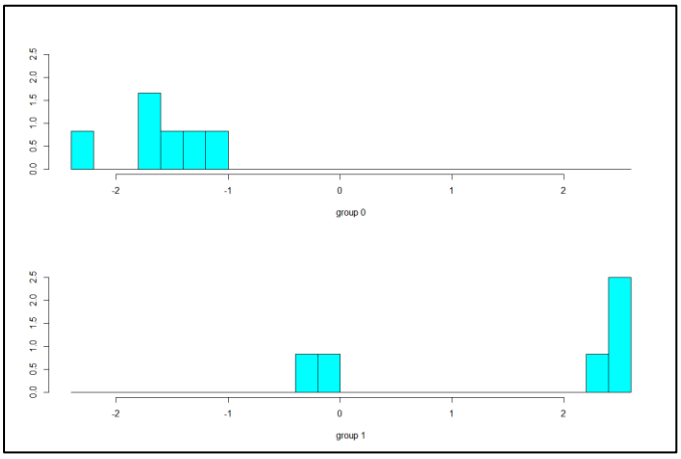
Without features selection	Features selection : p<0.005
	

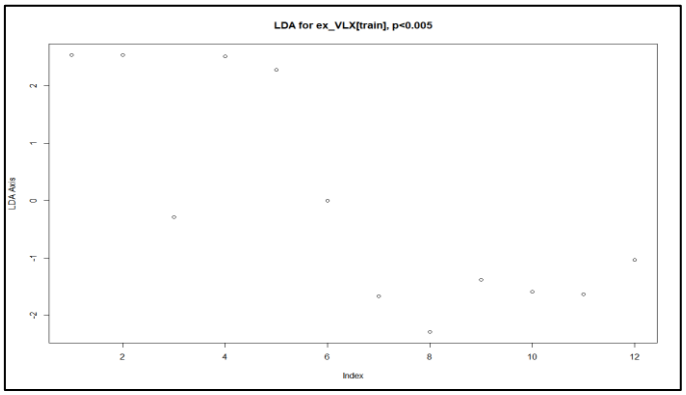
Heatmaps produced after we performed Random Forest on the SALVLX treated samples

LDA

As we did during the MDMA study, we performed LDA in order to get an inter-zones classifier. This time, the results were a little bit less accurate , as with a feature selection considering p<0.005 and only 53 features, there were still some overlaps between the classes. However, with an even tighter threshold of p<0.001 and only 14 features we were able to classify the samples, independently from there original brain zone. We took 12 samples (6+6) for training and 10 samples for testing.

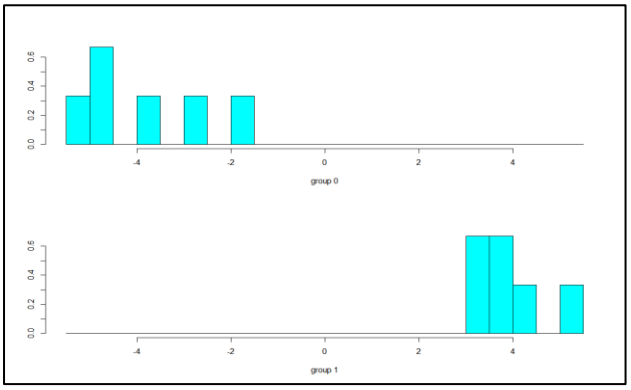
P<0.005 : There were 53 features left. Some overlapping between classes.

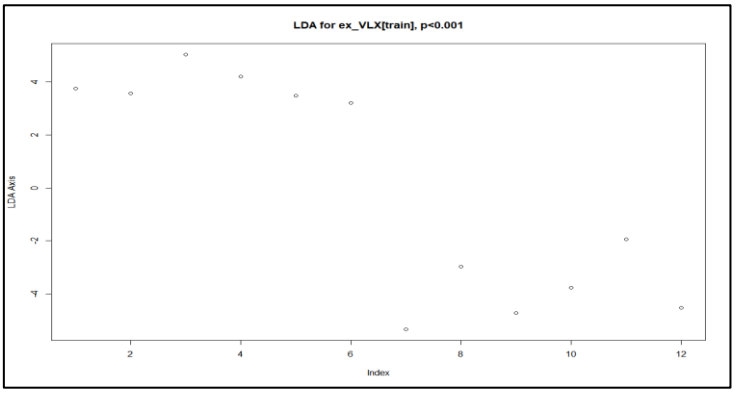





```
> preds$class      #predictions
 [1] 1 1 1 1 1 0 0 0 0 0
Levels: 0 1
> dat$Y[test]      # real values
 [1] 1 1 1 1 1 0 0 0 0 0
```

P<0.001. There were 14 features left.



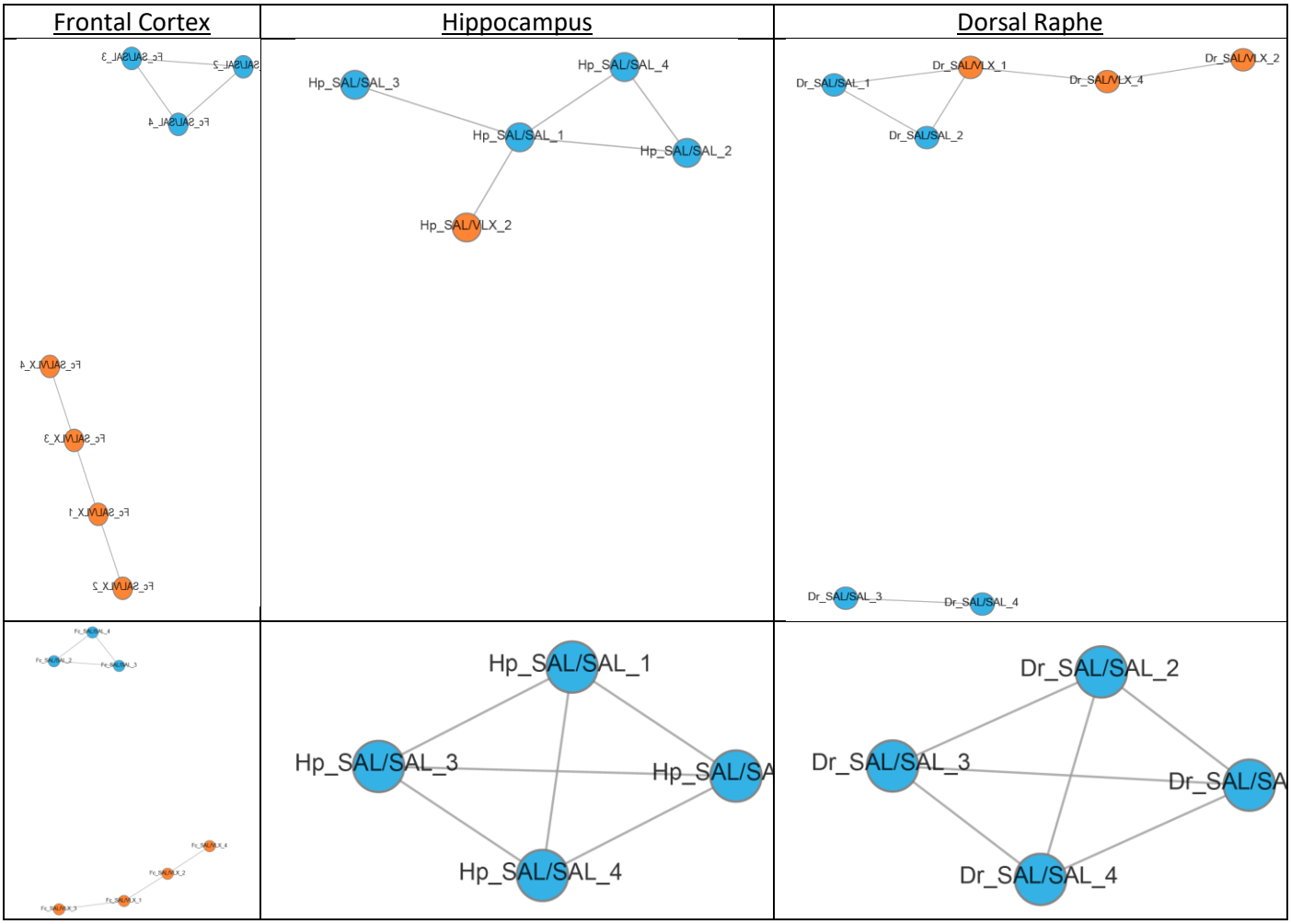



```
> preds$class      #predictions
 [1] 1 1 1 1 1 0 0 0 0 0
Levels: 0 1
> dat$Y[test]      # real values
 [1] 1 1 1 1 1 0 0 0 0 0
```

SCUDO

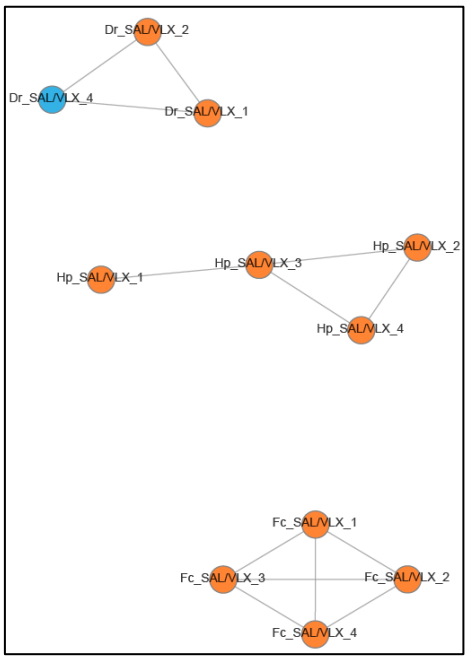
We will then use the SCUDO software for the same reasons than during the study of the MDMA effects.

We then first used SCUDO in unsupervised mode (p<0.99) and in supervised mode (p<0.1), the parameters were n1=n2=250, N% = 20. We performed the algorithm on each region, the first row correponds to the unsupervised mode. We can notice that the results are very similar for Frontal Cortex, whereas for Hippocampus and Dorsal Raphe the unsupervised mode yields low accuracy graphs as there is overlapping between the treatments. In the supervised mode for Hippocampus and Dorsal Raphe, only the control samples were displayed by the software.



Then we can study the treatment in unsupervised mode in order to seek for the interesting genes inter-zones [done in Functional Enrichment analysis]. We can see that the three brain zones are well separated, with no overlapping.

NB : As SCUDO only ask for control/ill subjects and that the graphs contains 3 clusters, I had to arbitrarly choose one sample as control subject (blue dot) but the colour display is meaningless once again.



Functional enrichment analysis

We reapeat the method used for the MDMA treatment :

Effects of Venlafaxine	Frontal Cortex	Hippocampus	Dorsal Raphe
Random Forest	- Sugar, carbohydrate binding - Ras protein, ATP, adenyl - Cytosolic part - Synaptic transmission, nerve impulse, signal, kinase, vesicle - Immune response, immunoglobin - Apostosis, cell death - Cognition, sensory perception, olfactory process	- Blood vessel, tube, neural tube development - Blood coagulation, response to wounding - Sensory perception, cognition - Response to stress - Death cell, apostosis - Cytoplasmic vesicle, locomotion, nerve impulse, kinase	- Long-term potentiation, melanogenesis - Glucose metabolysm - Protein kinase, serine, ATP, vesicle, synapse, signal - Surface receptor, cognition, olfactory - Cell death, apostosis
SCUDO P<0.01, Up-regulated	- Synapse, cytosckeleton, NMDA receptor, neurotransmittor, synaptic plasticity, signal, kinase, serine - Long-term potentiation, ATP - Neuromuscular process, Alzeimer's disease, cell development - Cell death, apostosis - Insulin stimulus, Ras protein - Sensory perception	- Chemical stimulus, sensory perception, olfactory, cognition - Cytoskeleton, immunoglobin - Placenta development - Negative neurogenesis, cell death, apostosis - Kinase, synapse - Ion transport, ATP, cell motion - Blood vessel morphogenesis, inflammatory response, defense	- Cell death, apostosis - Blood vessel, vasculature development - Responses to : vitamin D, wounding, drug, heat, temperature stimulus, inflammatory - Alzheimer's disease - Protein kinase, synapse, memory - Sensory perception, olfactory
SCUDO, P<0.01, Down-regulated	- Negative regulation of cell growth, dimerization activity - Protein complex assembly - Reponses to vitamin, nutrient, stimulus - Immune system - Cell death, apostosis - Dendrite, synapse, hormone stimulus - Sensory perception, cognition, olfactory receptor	- Response to stress, nucleoplasm, response to oxygen levels - Hormone activity, regulation of insulin - Response to organic, hormone stimulus - Regulation of blood pressure & circulation, heart development, muscle genesis - Cognition, sensory perception, olfactory - Pathways in cancer	- Sensory perception of smell, cognition - Hormone activity, regulation of system process - Immune response - Hypoxia, signal, hormone

JEPETTO Using the Random Forests genes

Pathway or Process	XD-score
Other glycan degradation	0,77877
One carbon pool by folate	0,37877
Sphingolipid metabolism	0,16058
Taste transduction	0,16058
Aminoacyl-tRNA biosynthesis	0,15268
Amino sugar and nucleotide ...	0,13877
Olfactory transduction	0,13877
Ribosome	0,13461
Peroxisome	0,12162
Base excision repair	0,10377
Lysosome	0,08543
TGF-beta signaling pathway	0,08144
Gap junction	0,07877
Pathogenic Escherichia coli i...	0,06387
Parkinson's disease	0,05957
Cardiac muscle contraction	0,05569
Shigellosis	0,05019
Long-term depression	0,04894
Melanoma	0,04507
Leishmaniasis	0,04328
Regulation of actin cytoskel...	0,04294

Pathway or Process	XD-score
Ether lipid metabolism	0,22532
Renin-angiotensin system	0,22532
Protein export	0,21061
Tyrosine metabolism	0,21061
Maturity onset diabetes of t...	0,18584
Aminoacyl-tRNA biosynthesis	0,14923
N-Glycan biosynthesis	0,13532
Glutathione metabolism	0,12916
Autoimmune thyroid disease	0,10032
Vasopressin-regulated wate...	0,08342
Lysosome	0,08198
ECM-receptor interaction	0,07921
Melanogenesis	0,07288
Vibrio cholerae infection	0,07055
Notch signaling pathway	0,07055
Basal cell carcinoma	0,06622
Hedgehog signaling pathway	0,06622
Inositol phosphate metabolism	0,06042
Chagas disease	0,05612

Pathway or Process	XD-score
Olfactory transduction	0,27775
Glycosylphosphatidylinositol...	0,24347
Porphyrin and chlorophyll m...	0,19305
Propanoate metabolism	0,16828
Maturity onset diabetes of t...	0,16828
Glycerolipid metabolism	0,15775
Melanogenesis	0,15287
Glycolysis / Gluconeogenesis	0,14823
Long-term potentiation	0,14823
Gastric acid secretion	0,14823
Oocyte meiosis	0,12797
Thyroid cancer	0,11775
RNA polymerase	0,11160
Citrate cycle (TCA cycle)	0,11160
GnRH signaling pathway	0,10233
ErbB signaling pathway	0,10061
Adipocytokine signaling path...	0,09810
Fructose and mannose meta...	0,09568
Glioma	0,09109

Interpretation :

We can observe that - like during the MDMA study - among the three zones there is a development of apostosis and regulation of death cells. We also notice the synaptic behavior modification (kinase, synaptic plasticity, serine) and the involvement of the immune system (immunoglobin, response to stress), the nervous system reacts to some perturbation (inflammatory response, response to stimulus, defense process, response to wounding). We can notice that many of the functions involved in the VLX treatment were also present in the MDMA treatment. Then there are some zone-specific functions which are involved :

- Frontal Cortex = Sugar, carbohydrate binding, glycan degradation. Some cancerous cells such as Alzeihmer's and Parkinson.
- Hippocampus = Modification of some organs functionment (tube, neural tube, placenta, heart). Sensory modification (sensory and olfactory). Regulation of insulin and hormone activity regulation.
- Dorsal Raphe = Long-term potentiation. Responses to : vitamin D, wounding, drug, heat, temperature stimulus, inflammatory.

As a general observation we can notice that the sensor system is pertubated, along with the organ functionment and the nervous circuit.

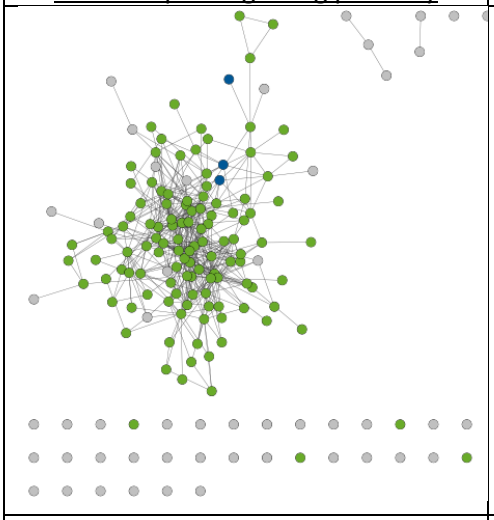
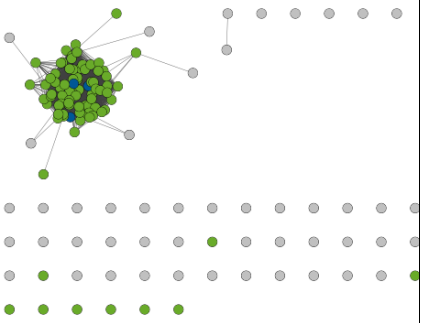
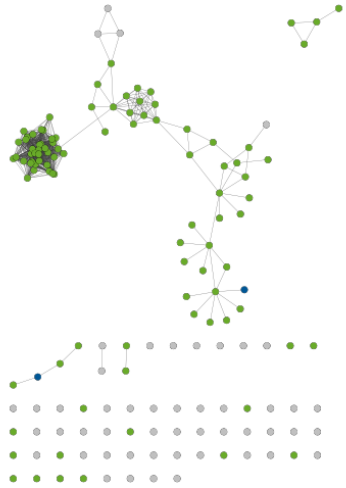
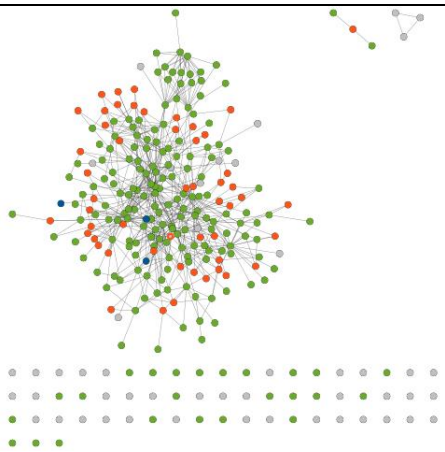
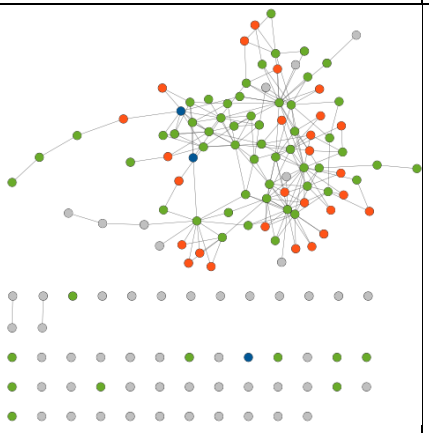
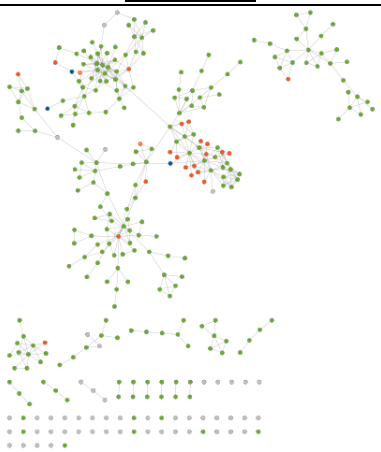
Distinction between zones :

	SALVLX treatment only
Random Forests	- Cell membrane, protein receptor - Glycoprotein, ATP binding - Kinase, signal, synapse

Interpretation : The function are not very interesting, as during the MDMASAL treatment.

Network Analysis with Cytoscape

Once again when we enter the SCUDO networks into Cytoscape, it only display the already visualized networks, so we can use Cytoscape to visualize the characteristic functions involved in the treatment :

Neurotrophin signaling pathway	Ribosome	Parkinson's disease
		
Regulation of actin cytoskeleton	**Long-term potentiation**	**Cytokine-cytokine receptor interaction**
		

V – Other considerations and conclusion

Could we distinguish between zones ?

Here we try to classify the treatment considering every zone. In this way we could be able to determine whether or not a subject has done one of the two drugs, independently of the zone of extraction.

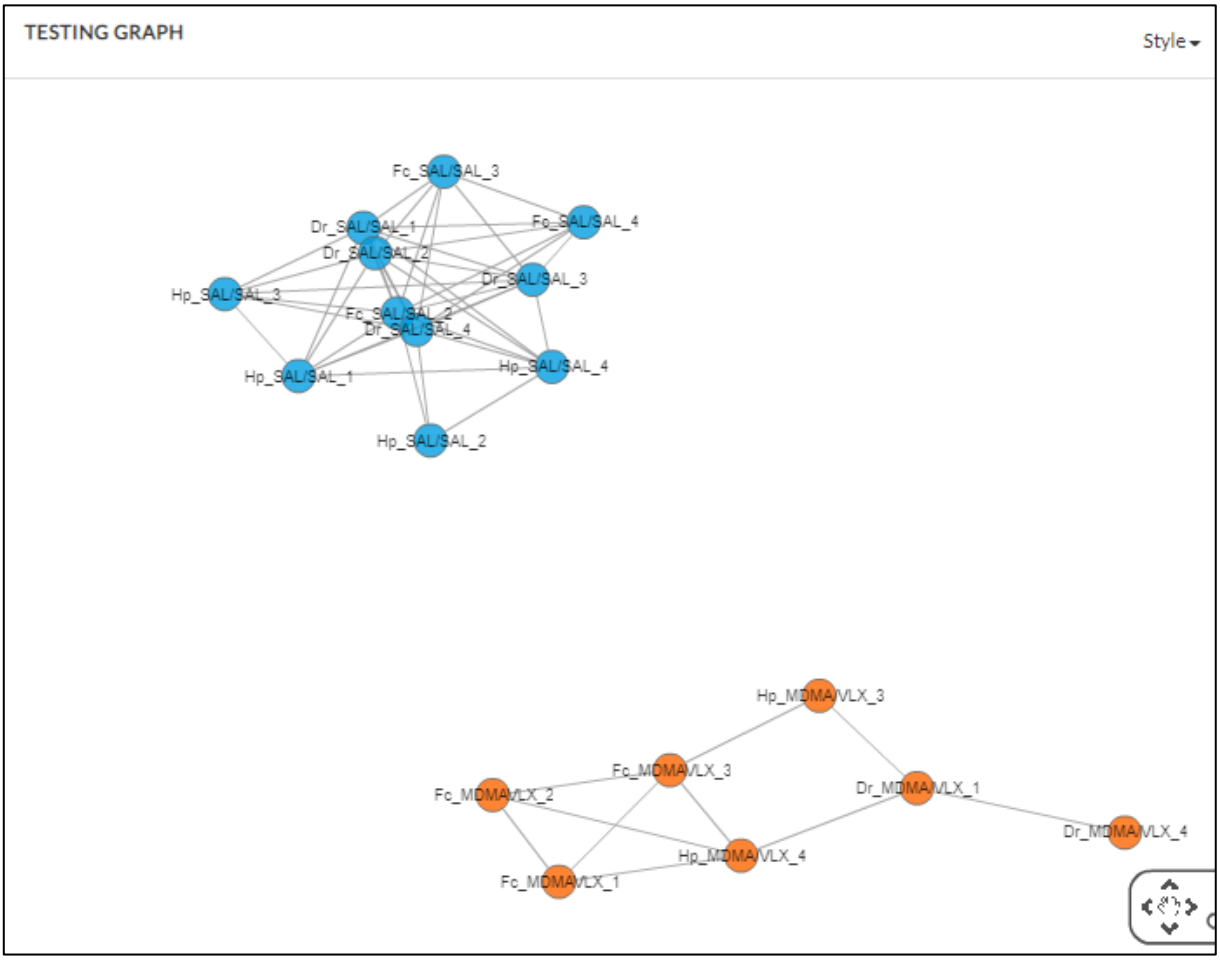
Random forests with features selection (we did 2 features selections) : we still observe a huge overlap between the MDMA and VLX, whereas the control subjects are easily identified.

```
> rf

Call:
 randomForest(x = t(small.eset), y = as.factor(group), ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 11

        OOB estimate of  error rate: 45.71%
Confusion matrix:
          MDMDASAL SALSAL SALVLX class.error
MDMDASAL         3      0      8   0.7272727
SALSAL           0     12      0   0.0000000
SALVLX           8      0      4   0.6666667
```

A brief study of the combined effects of MDMA and VLX combined (ex_TEST)

Does it yield cumulative effects of both MDMA and VLX ?

We performed functionl enrichment analysis of the ex_TEST dataset by region, which we can compare to the previous functions threatened by respectively MDMA and VLX.



SCUDO graph for the MDMAVLX treatment, compared to the control subjects. With features selection (p<0.01), the software achieve a good classification inter-zones

The signatures returned by SCUDO don't yield any improvement in the comprehension of the treatment, as the functions involved were the superposition of the functions involved in the two treatments. No function has appeared due to the combination of the drugs.

We can also try to perform random forest on the data, after pre-processing the features with a selection (p<0.05). The result is disapointing concerning the identification of the treated subjects :

```
> rf

Call:
 randomForest(x = t(small.eset), y = as.factor(group), ntree = 10000)
               Type of random forest: classification
                     Number of trees: 10000
No. of variables tried at each split: 149

        OOB estimate of  error rate: 34.78%
Confusion matrix:
        MDMAVLX SALSAL class.error
MDMAVLX       4      7  0.63636364
SALSAL        1     11  0.08333333
```

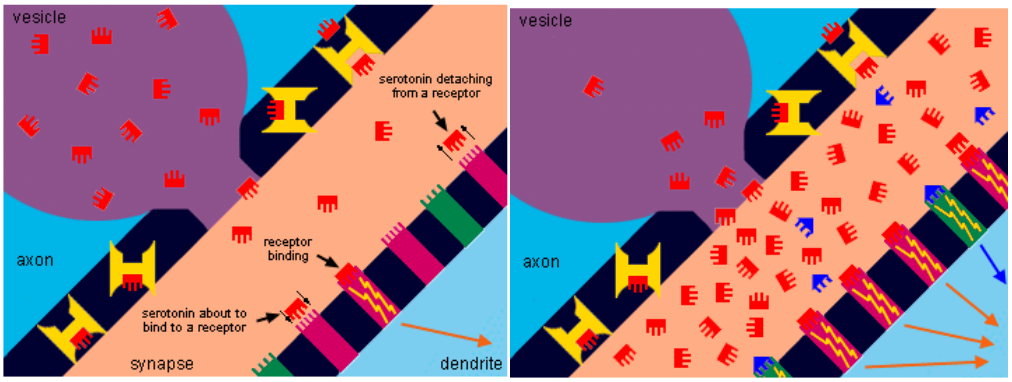
Random Forests result for the MDMAVLX treatment, when mixing zones with no features selection

Conclusions

Due to the structure of our dataset, we were limited to draw conclusions about the accuracy of our models, however thanks to the functional analysis tools we were able to identify the affected functions in the brains, according to the type of treatment and to the brain region. The fact that the dataset did not contain a sufficiently large amount of samples by treatment and by region wasn't a limit in our study, as the objectives of the project were mainly to determine the affected functions by the treatments.

Thanks to the multiple tools for functional enrichment analysis (DAVID, JEPETTO) , we were able to cross-validate the results and fortunaly the results were matching across the softwares. We can notice that multiple functions are impacted by the take of these drugs, which the main ones are concerning synaptic connections, molecul binding (characteristic of drugs)and some organs functionment modification. Some cognitive functions were also altered, such as memory skills, modification of the synaptic plasticity – which leads to depresssion , sensory reception and perception (olfactory, inner-ear, ear).

The apparition of synaptic deregulation is explained by the fact that both drugs artificially creates well-being states by having effects on nervous system :



Modelization of the perturbation inside vesicle and synaptic receptor : before/ after intake of MDMA

These nervous modifications yield dangerous consequences on the long-term, because as the different regions of the brains modify their functionment after drug intake, the serotonin levels are lowered after such brutal increases. The functional enrichment analysis particularly underlines the apparition of long-term depression.

The treatments also made appear cancerous cells, yielding frightening results such as the apparition of Alzheimer, Parkinson and multiple types of cancer. We can also notice that a huge amount of the altered functions were common to the two treatments.

This dataset has been the subject of multiple data analysis projects, in which we can find some common conclusions with this report. [1], [4], [5].

As the consequences of the Venlafaxine were far from being negligible, we might be questioned about the relevancy of making such treatment legal and widely prescribed. We could think that the consequences should at least be more displayed, so that treated patients would face the consequences with full awereness.

VI – Appendices & references

[1] Effects of Venlafaxine (study of the dataset GSE 47541)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4244101/

[2] The dorsal raphe nucleus and serotonin: implications for neuroplasticity linked to major depression and Alzheimer's disease.

https://www.ncbi.nlm.nih.gov/pubmed/18772036

[3] Description of Hippocampus

https://www.medicalnewstoday.com/articles/313295.php

[4] Paper on the effects of the Venlafaxine treatment on the dataset

https://www.researchgate.net/figure/Significantly-changed-genes-after-three-week-long-venlafaxine-administration-in-the_fig1_268874473

[5] Paper on the effects of the MDMA treatment on the dataset

https://www.researchgate.net/publication/326785039_Gene_expression_analysis_indicates_reduced_memory_and_cognitive_functions_in_the_hippocampus_and_increase_in_synaptic_reorganization_in_the_frontal_cortex_3_weeks_after_MDMA_administration_in_Dark_Ago