Single cell RNA sequencing of the medial Amygdala to determine sex and parenting cell type specific gene expression differences

Shachar Katz, Dorin Shteyman
Supervisor: Dr. Amit Zeisel

August 2022

Abstract

With the use of advanced data manipulation, the GABAergic neurons of the Amygdala area were shown to contain differentially expressed genes between females, males, parents, and naïve mice.

1 Dataset

The data on which the analysis and conclusions were derived from, was collected from 16 distinct mice, distributed equally over the four features: sex and sexual maturity (Figure 1). The samples were separately stored for each mouse, withholding for each of the 16 mice, around 5000 Amygdala cell samples (Figure 2). The large amount of singe-cell RNA Amygdala samples for each mouse was used to avoid biased conclusions and improve generalization. Each cell sample contained all 27,998 mouse genes and the amount their expression within the specific cell.

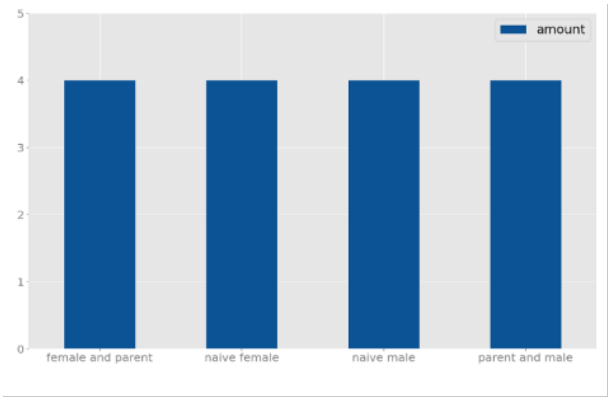

Figure 1: Dataset distribution over: males, females, parent and naive mice groups

2 Data Preparation

2.1 Cell Samples Filtering

In order to extract valuable conclusions from the dataset, we first had to clean and prepare it for further analysis. As a first step, we filtered all cell samples who were too sparse. We set the threshold at total gene expression to be 3,000 and filtered all samples with less overall gene expression than the given threshold. Another filtering criteria was samples who were less expressive, meaning that the total genes who were expressed was less than a threshold of 2,500. In addition, all 'Gaba' genes were also filtered in this initial stage. After the sample filtering stage, we filtered about half of the total samples (Figure 3).

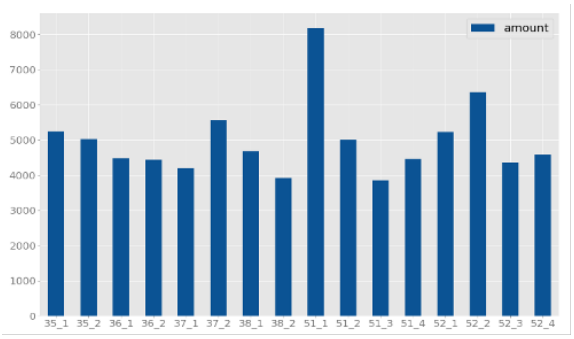

Figure 2: Dataset sample size per mice (number of cells for each sample)

2.2 Data Gene Expression Analysis

From modeling certain aspects of the group samples, we can infer the variability of the data given from each mouse (meaning, sample group). In order to infer conclusions over all the samples combined, we must first validate that they are distributed similarly in certain aspects such as: molecules number (Figure 4) and the mitochon-

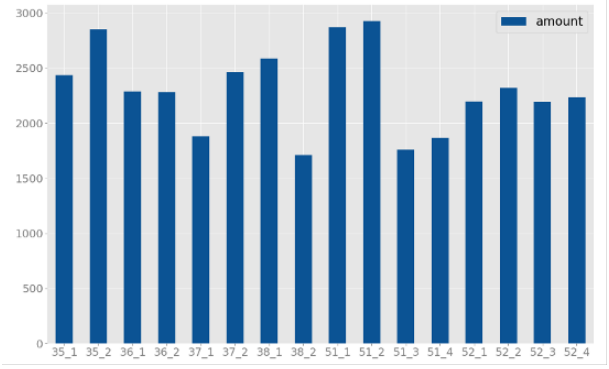

Figure 3: Dataset sample size per mice after filtering

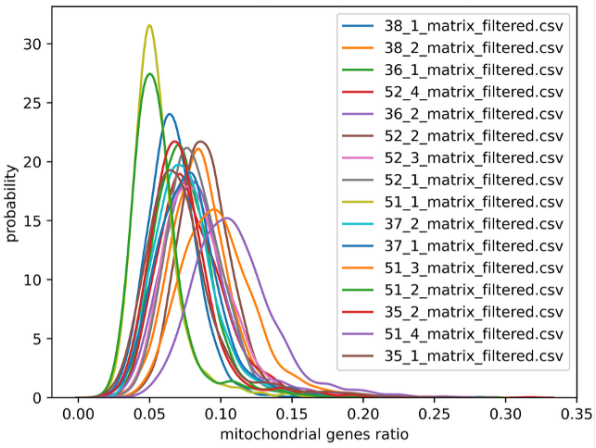

Figure 5: PDF of mitochondrial genes expression ration per sample

drial genes ratio from the data (Figure 5). The demonstrated plots indeed confirm that overall, the sample groups show similar trends of these measured aspects.

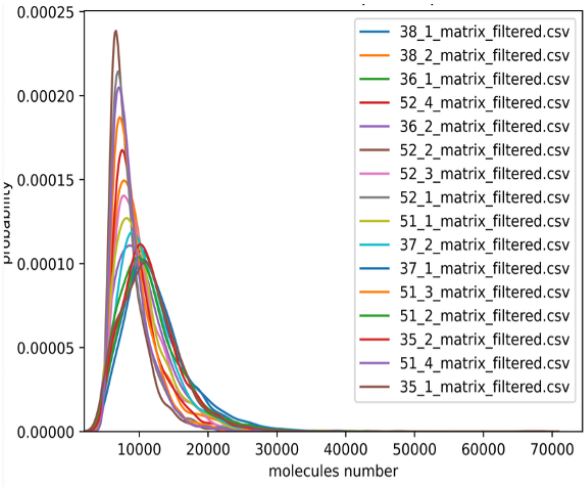

Figure 4: PDF of molecules per sample

2.3 Genes Filtering

The next conducted step was normalization on the gene expression of each sample, in order to bring all genes to the same scale.

Another element of the data we addressed was the female versus male aspect of the samples. After splitting the data based on this criterion, we analyzed which genes were prevalent in males significantly more than females and vice versa.

We marked the 20 genes with the greatest difference between their average expression in females versus in males and the 20 genes with the greatest difference between their percentage of expression in females versus in males and marked them (Figure 6, Figure 7).

These plots (Figure 6, Figure 7) coherently expose the genes expressed only for females or male mice, such as: Ddx3y, Eif2s3y, Uty, Kdm5d for males, and Tsix, Xist for females. For the next stages of the data processing, we remove these genes to investigate neural genes and not sex genes. In contrast to previous filtering of non-informative samples, now we filter non-informative genes. On top of the 6 sex genes mentioned above, we also filter genes who appear in more than half of the cell samples (of all mice) and genes who appear in less than 5 cell samples. In this gene filtering stage, we removed 10,499 genes, leaving us with 17,499 out of the total 27,998 genes of a mouse.

Our next interest was to find genes who are the most expressive. We measured their expressiveness by their deviation from their mean expression across all cell samples we obtain by this point. Meaning, the farthest the difference between the coefficient of variation $C_v = \frac{\sigma}{\mu}$ and the mean μ, the more expressive we consider the gene to be (Figure 8). By sorting the genes in ascending order of calculated $C_v - \mu$ for each one, we noticed that the values decay and converge to zero in a rate similar to y(x)=1/x. For this reason, finding the knee point (meaning, the point on the graph of the genes' $C_v - \mu$ value in ascending order closest to the origin (0,0)) can set a good threshold to indicate genes with significant difference of $C_v - \mu$.

With the knee point as a threshold, we filtered all genes with $C_v - \mu$ value lower than the knee point (Figure 9). With a knee point value of 0.1274, we filtered 17, 045 genes and were left with 454 genes.

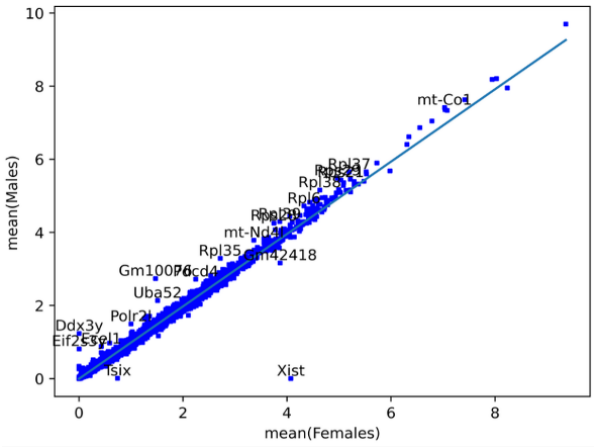

Figure 6: Females vs Males genes mean: expressiveness by deviation from mean expression across all cell samples

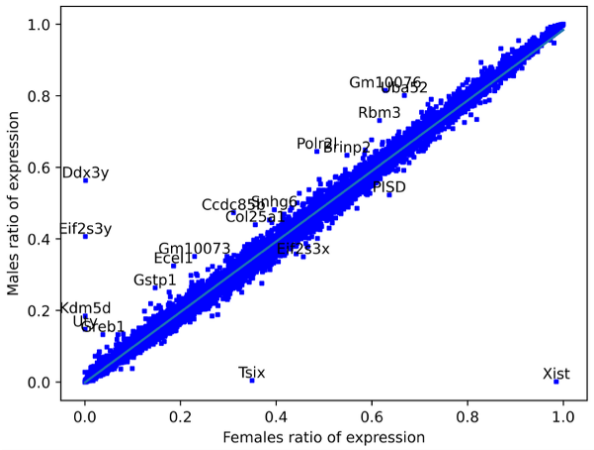

Figure 7: Females vs Males genes expression ratio

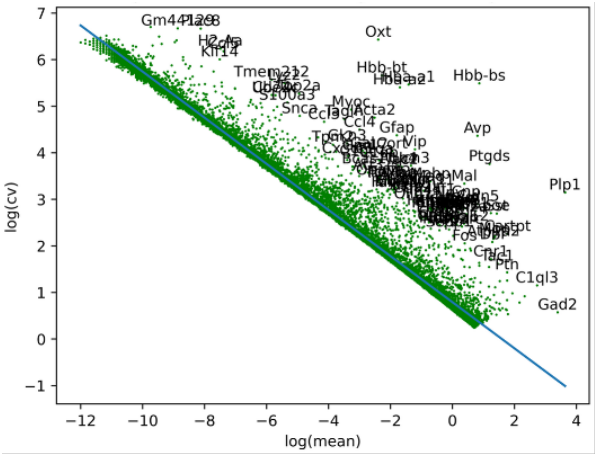

Figure 8: log(mean) as function of log(cv) for each gene

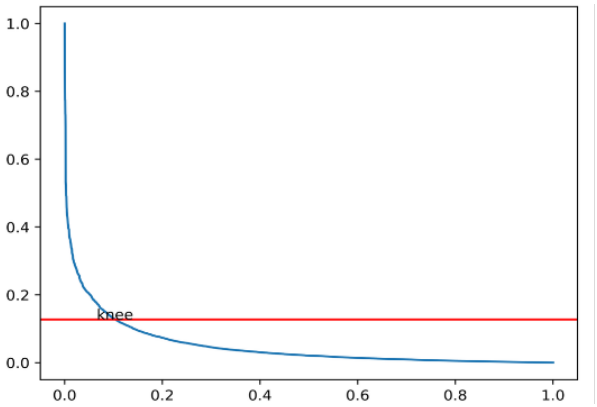

Figure 9: CV distance (absolute) density. recommend threshold=0.1274

2.4 Dimension Reduction

In order to efficiently apply clustering methods, we must first reduce the dimension of our data. We achieve this by applying first PCA and TSNE methods.

PCA reduces dimension in the most variance preserving way. This method first calculates the eigenvalues and eigenvectors of the covariance matrix. Afterwards, we must pass the value of desired dimension. As seen in the C_v calculation, the eigenvalues of the covariance matrix deteriorate and converge to zero in a rate of y(x)=1/x . From the same considerations as before, we'll determine the dimension of the PCA output by the knee value of the eigenvectors (Figure 10). Meaning, we'll choose the eigenvectors corresponding to the eigenvalues above the knee value to span our new sample space. With a knee point value of 0.0334, we determined our sample space to be of dimension 13.

We applied the TSNE algorithm on the new cell sample space (Figure 12). TSNE algorithm reduces the dimension from 13 to 2. We can see (Figure 12) that the that cell samples with different feature types (male/female, parent/naïve) are dispersed almost equally in the TSNE scatter.

2.5 Cluster Analysis

Next, we wish to cluster the TSNE scatter results into informative clusters using the DBSCAN clustering algorithm.

The best resolution for a clustering problem has no close solution. With respect to the problem, we most find the right measure of epsilon and the minimum amount of samples in each cluster to avoid over or under expressiveness of the DBSCAN result.

We calculated epsilon by calculating for each sample the distance to her 20 closest neighbors. Epsilon was set to be the 70 percentile out of all distances measured, which resulted to be 1.47246 (Figure 11). The minimum amount of samples to

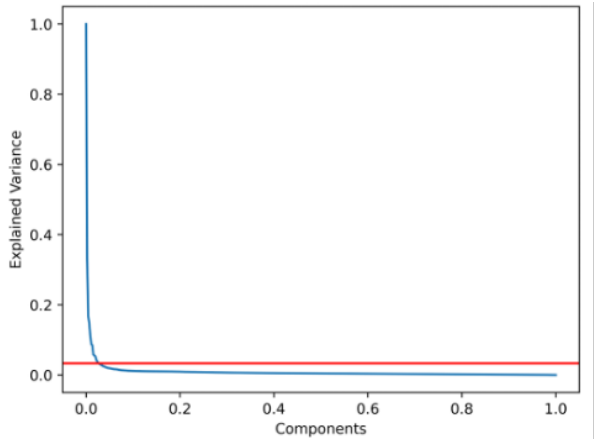

Figure 10: PCA explained variance: knee=0.0334. Only 13 values are bigger than the knee value

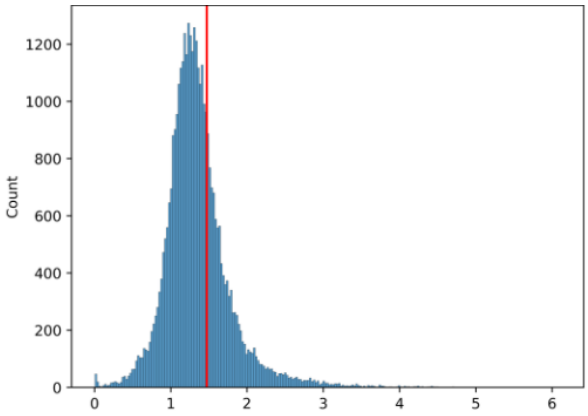

Figure 11: The distance to the 20th closest neighbor. The 70 quartile is 1.47146

form a cluster was set to 20. As a result, we receive 86 clusters (Figure 13).

Next, we'll want our clusters to have a more explanatory indexing. That is to say, indexing the clusters by their similarity (proximity) to each other. For this, we use our processed dataset as computed before applying PCA. We compute the average gene expression for all 454 (number of genes we were left with before the PCA stage) genes to model each cluster's average cell sample. Afterwards, we reduce the span from 454 to 10 by using once again the PCA algorithm. Now obtaining a condensed version of the clusters' average cell sample, we calculate the linkage to determine the clusters' indexing based on the proximity of their average cell sample.

The connections and hierarchy produced by the Linkage algorithm ((Figure 14)) is used to replace the indexing of the clusters given by the DBSCAN stage.

Having the linkage indexing of clusters allows us to produce an informative Heatmap to present the marker genes of each cluster. The marker genes are those who vary the most with respect to the average cell across all cell samples of all clusters. A marker gene is unique thus can't serve as marker gene for more than one cluster.

We sort all cell samples grouped by the cluster they belong to and sort the clusters by their linkage index on the x-axis. The marker genes for each cluster in the same order on the y-axis ((Figure 16)). The heatmap clearly shows how cells within the same cluster share similar expression of the marker genes.

Our last stage within the scope of the processing of the entire data is the clustering based on neural genes. Marking each cluster with one of the following neural genes: 'Gad2' (GABAergic gene), 'Slc17a7' (Glut 1), 'Slc17a6' (Glut 2).

Using the average gene expression of all cell samples within each cluster, we determined which of the three neural genes listed above is predominant. If the difference in expressiveness wasn't significant, we classified the cluster as 'Doublet'. If other non-neural listed genes were dominant in the cluster more than any neural gene, we classified the cluster to be 'non-neural'. The resulted outcome can be seen in (Figure **??**). From this point on, we continue our data processing only on cell samples from clusters where 'Gad2' was the dominant neural gene.

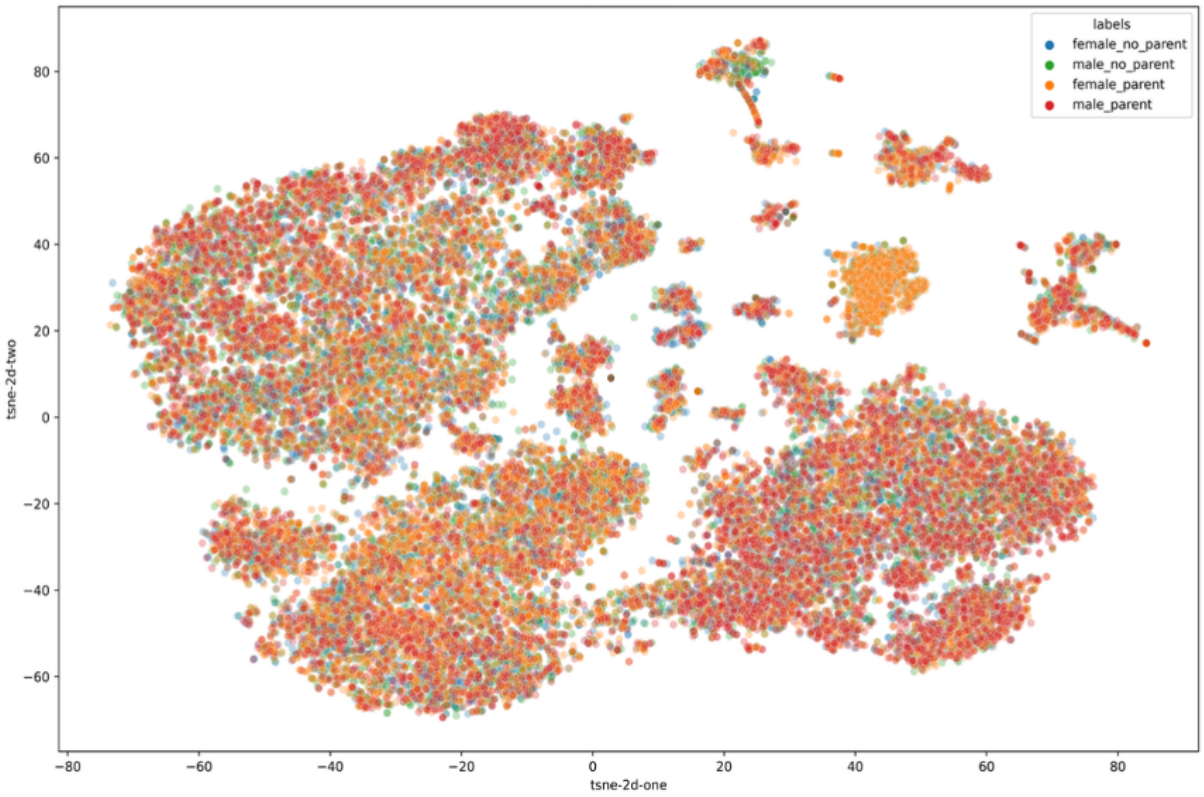

Figure 12: t-SNE in 2d. Colored by sex and sexual maturity

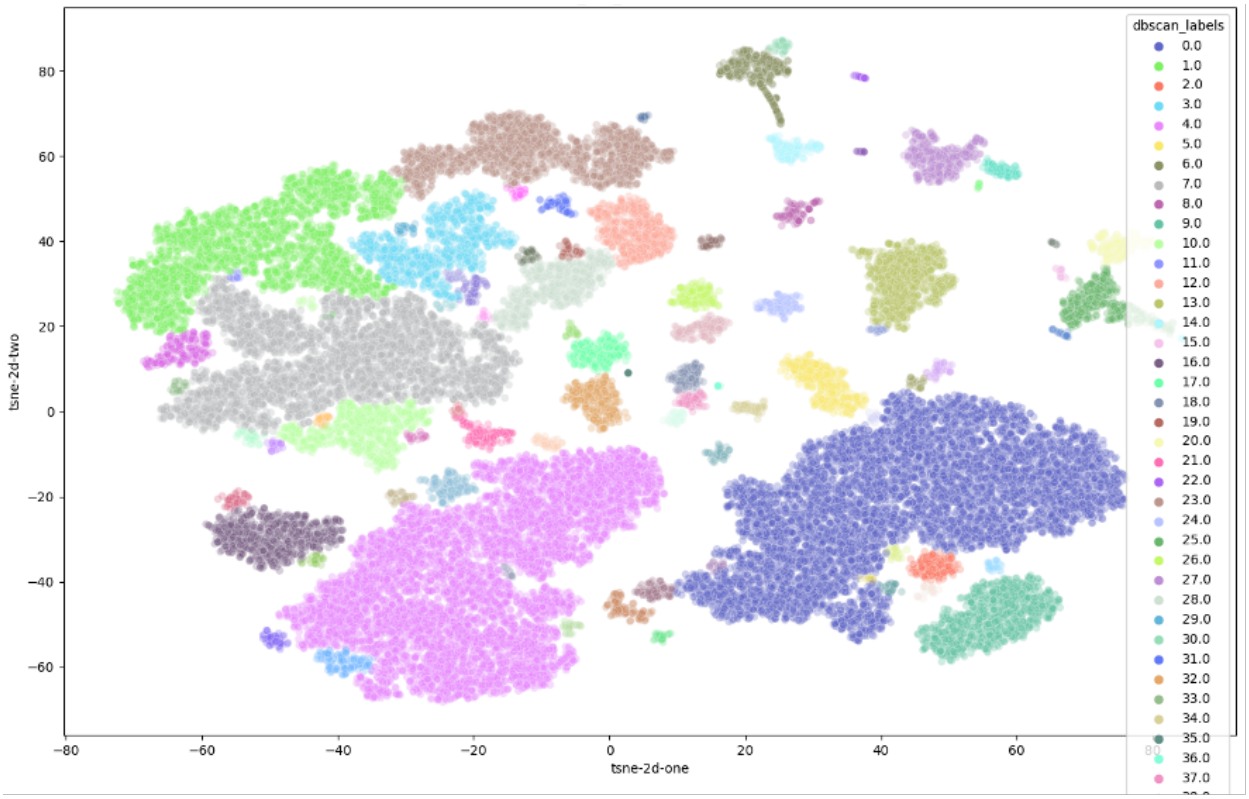

Figure 13: DBScan according to eps=1.472. Clustered 86 clusters

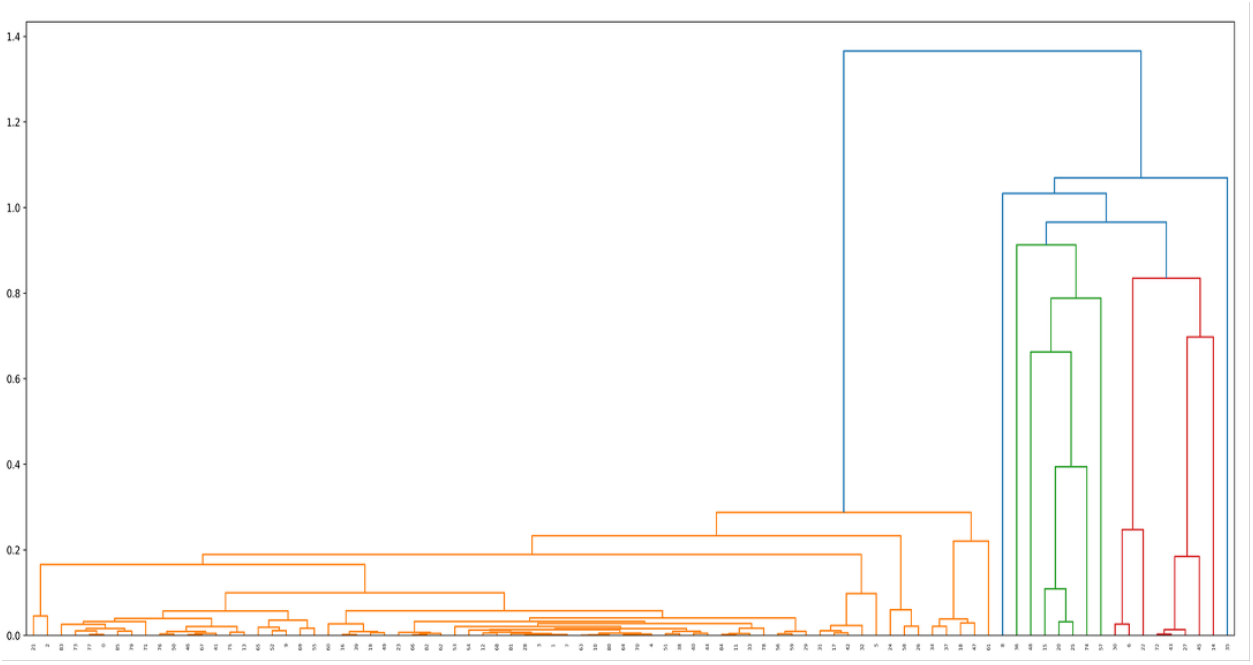

Figure 14: Linkage on clusters: finding similarities and connections between different clusters

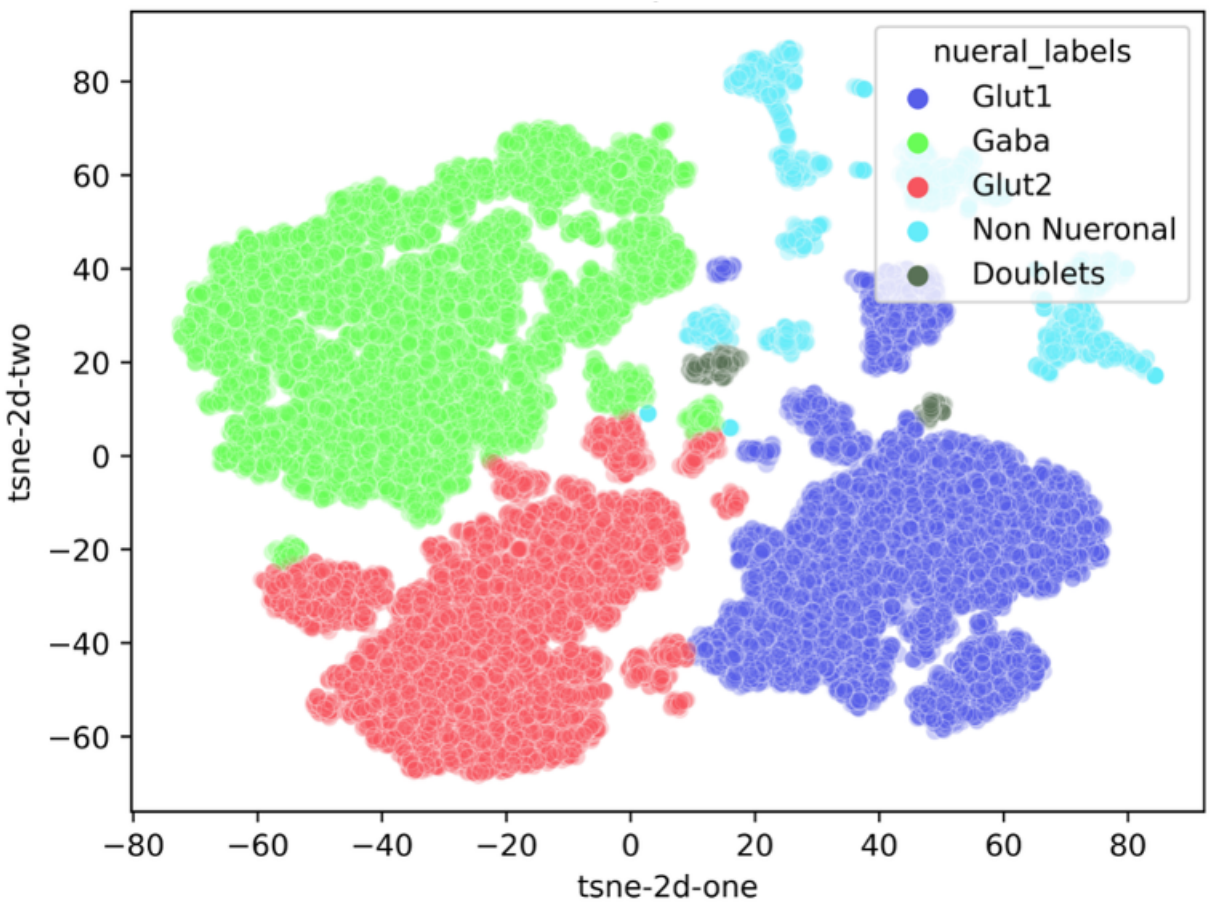

Figure 15: Classify clusters according to predominant neural genes, based on average gene expression of all cell samples withing each cluster

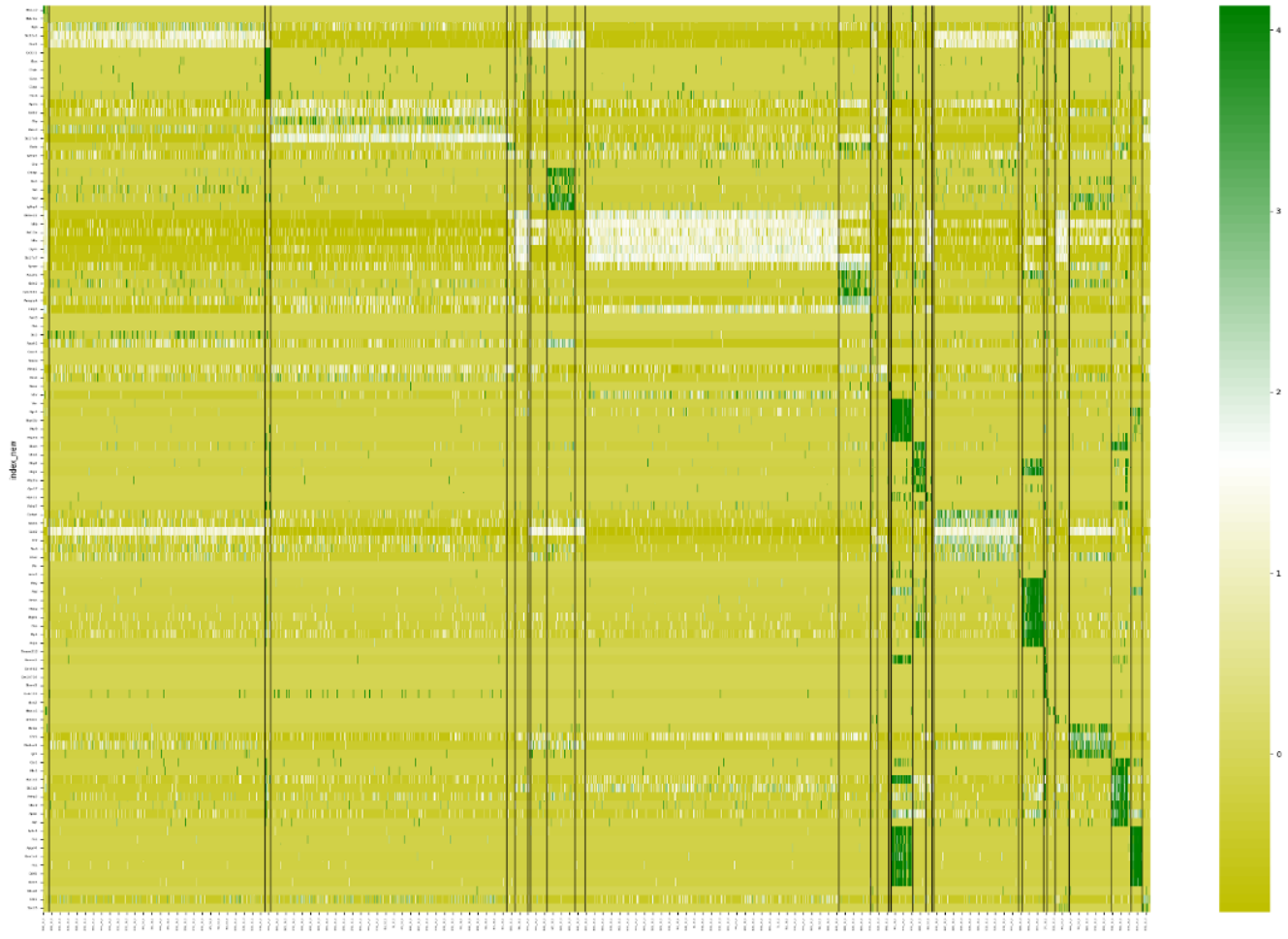

Figure 16: Heatmap: Marker genes expression of all clusters ordered by Linkage

3 Gaba Samples Analysis

In a similar manner described in stages 2.1 – 2.5, we perform the same analysis on cells identified as GABAergic. Their amount is roughly a third of the original size dataset, dispersed among all mouse samples (Figure 17).

Detailed explanation to the motives and significance of each step can be found in the equivalent steps in part 2.

3.1 Cell Samples Filtering

The cell samples filtering didn't yield any change of the dataset size.

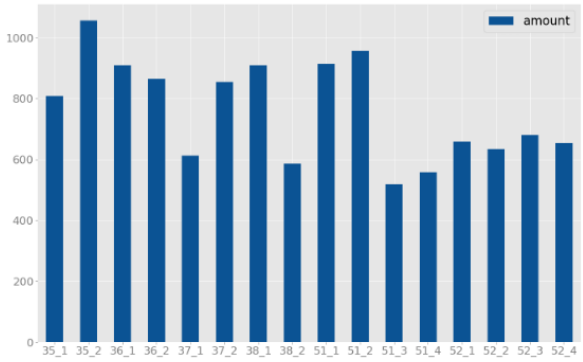

Figure 17: Dataset of only GABA (GABAergic) cells: sample size per mice (number of cells for each sample)

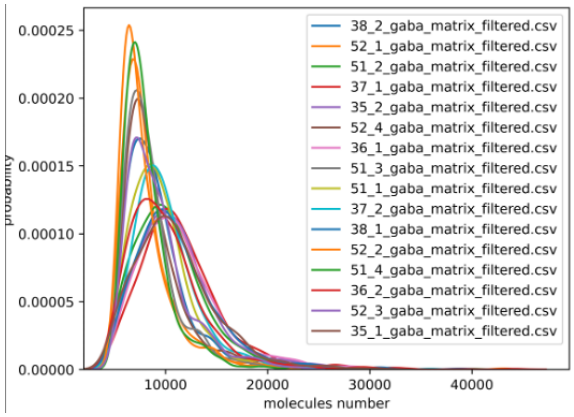

Figure 18: PDF of molecules per sample (only GABAergic cells)

3.2 Data Gene Expression Analysis

In order to infer conclusions over all the samples combined, we must first validate that they are distributed similarly in certain aspects such as: molecules number (Figure 18), mitochondrial genes ratio from the data (Figure 19) and the expression of each gene across the different sample groups.

The demonstrated plots indeed confirm that overall, the sample groups show similar trends of these measured aspects, just like in this analysis previously conducted on the entire dataset of samples.

3.3 Genes Filtering

After normalizing the samples, the comparison of gene expression between females and males showed almost identical results to the previous analysis on the entire dataset (Figure 19, Figure 20).

The same sex genes stood out and for the same motives described in 2.3, we filter these genes alongside others who fall below the knee value cal-

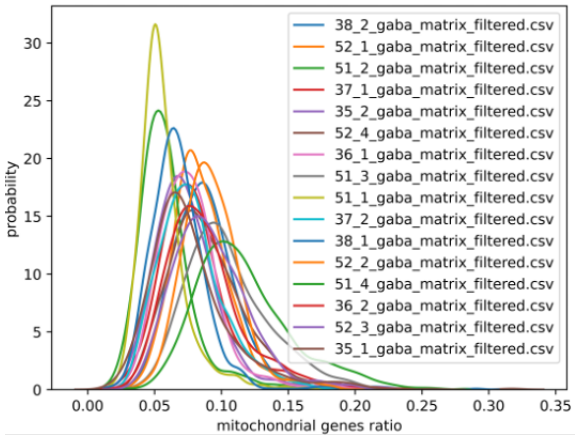

Figure 19: PDF of mitochondrial genes expression ration per sample (only GABAergic cells)

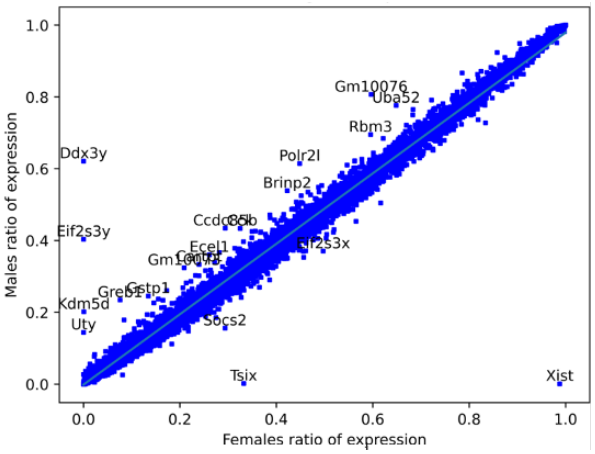

Figure 20: Females vs Males genes mean: expressiveness by deviation from mean expression across all cell samples

culated for the $C_v = \frac{\sigma}{\mu}$ value of each cell sample (elaborated explanation in 2.3) Applying the filtering by C_v value and the most common and rare genes (elaborated explanation in 2.3).

narrows our genes amount to 393 out of the original 27,998 genes of a mouse. As expected, the knee value (Figure 23) and the labeled genes (Figure 22) are almost identical to the values calculated for the entire dataset.

3.4 Dimension Reduction

When performing dimensional reduction on the GABAergic genes, we noticed that the knee value calculated for PCA was approximately 3 times bigger than the one calculated in 2.4, resulting about 3 times more elements to span the newly, reduced cell sample space (Figure 25).

For the TSNE dimension reduction on the reduced space yielded by the PCA algorithm, we of course receive different scatter from the one re-

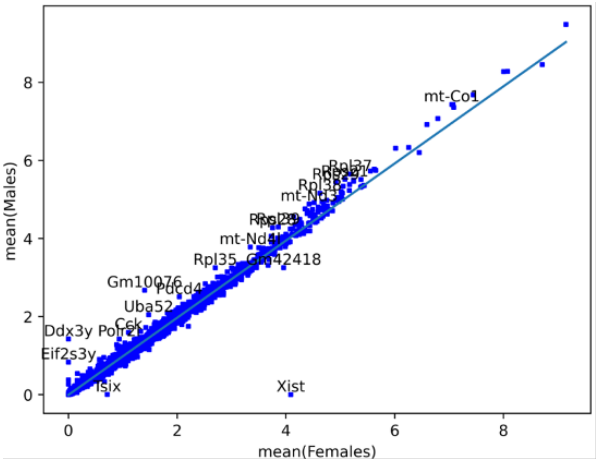

Figure 21: GABA cells: Females vs Males genes expression ratio

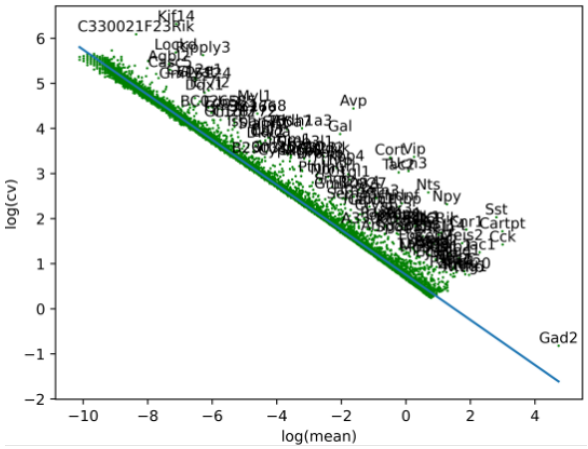

Figure 22: GABA cells: log(mean) as function of log(cv) for each gene

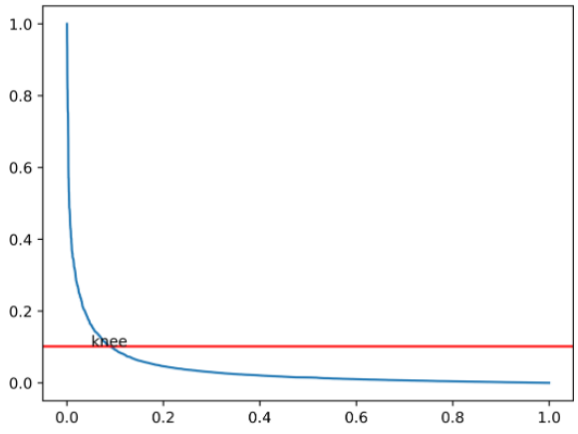

Figure 23: GABA cells: CV distance (absolute) density. recommend threshold=0.1274

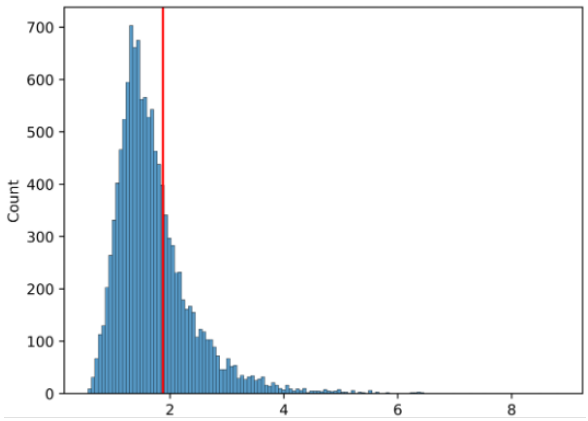

Figure 24: The distance to the 20th closest neighbor. The 70 quartile is 1.87431

ceived in 2.4 (Figure 12). The PCA transform on the data produced a completely different sub-space and therefore the TSNE algorithm applied on this sub-space (Figure 25) is different as well.

On a general note, from now on we won't expect any of our data processing outputs on GABAergic genes to resemble previous results on the entire dataset since the sample space has changed.

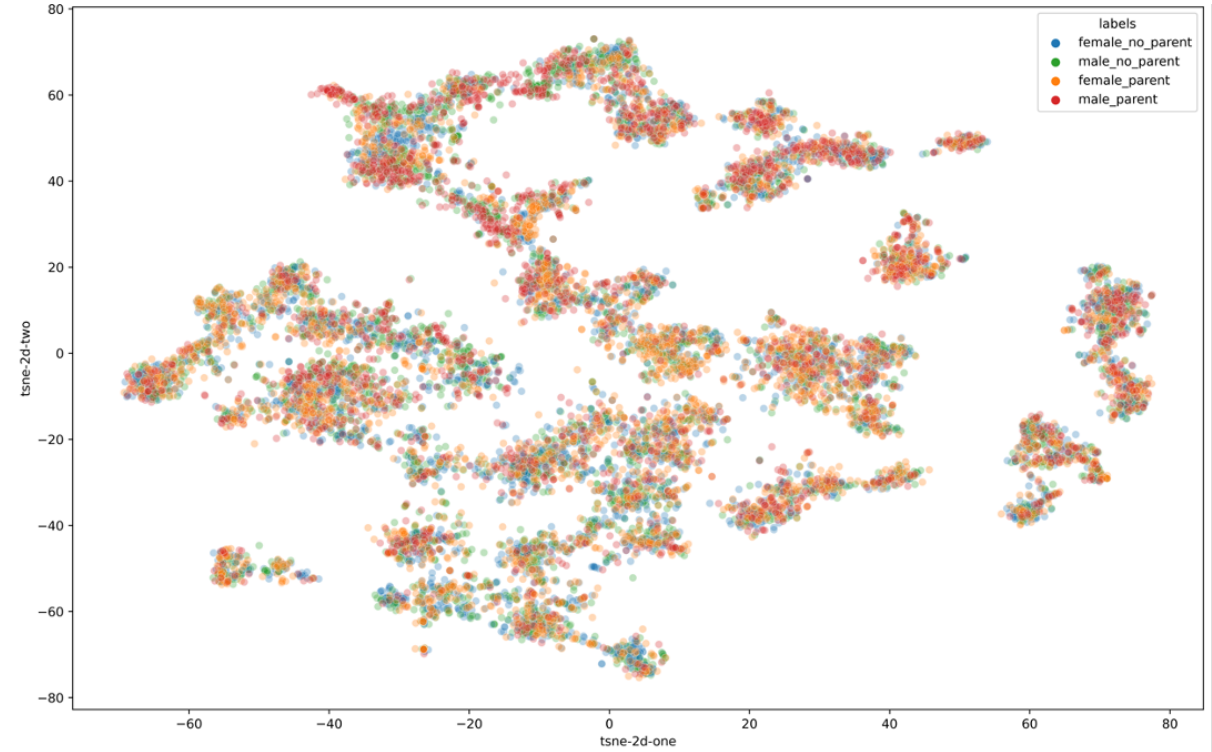

Figure 25: t-SNE in 2d on GABA cells. Colored by groups

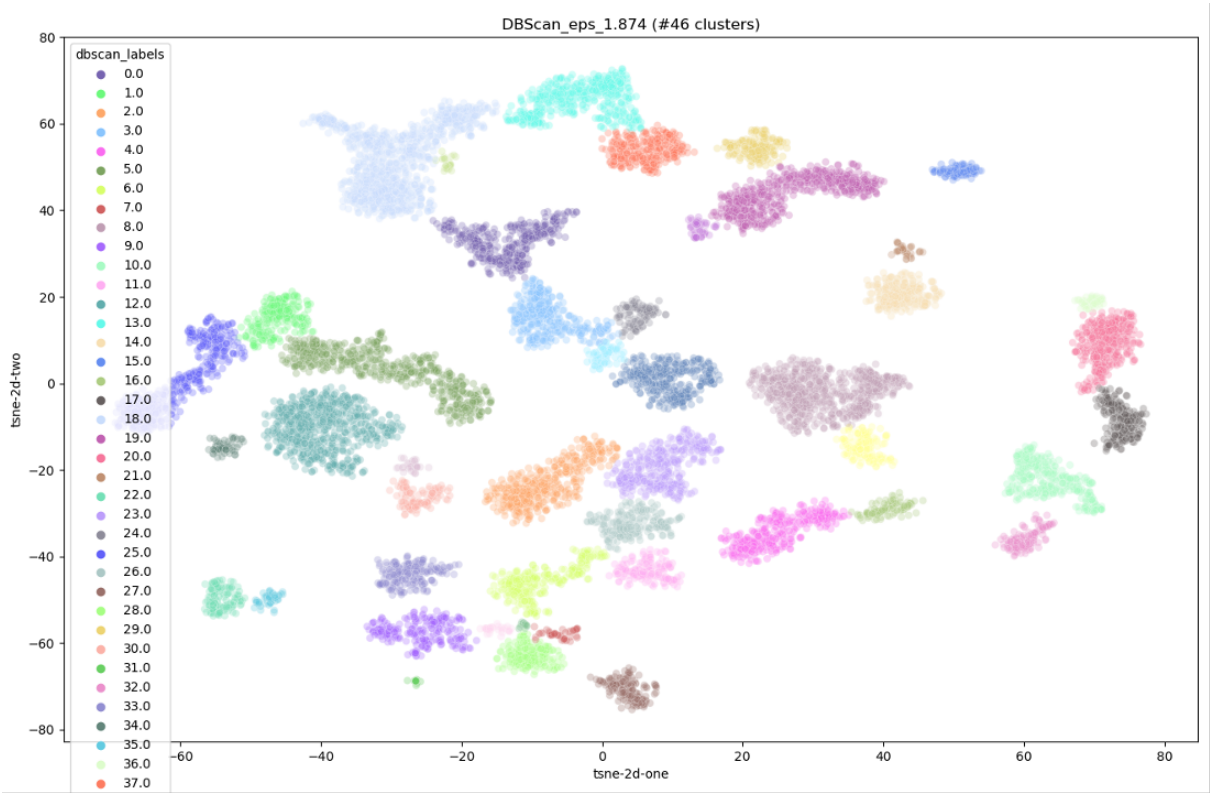

Figure 26: DBScan on GABA cells according to eps=1.874. Clustered 46 clusters

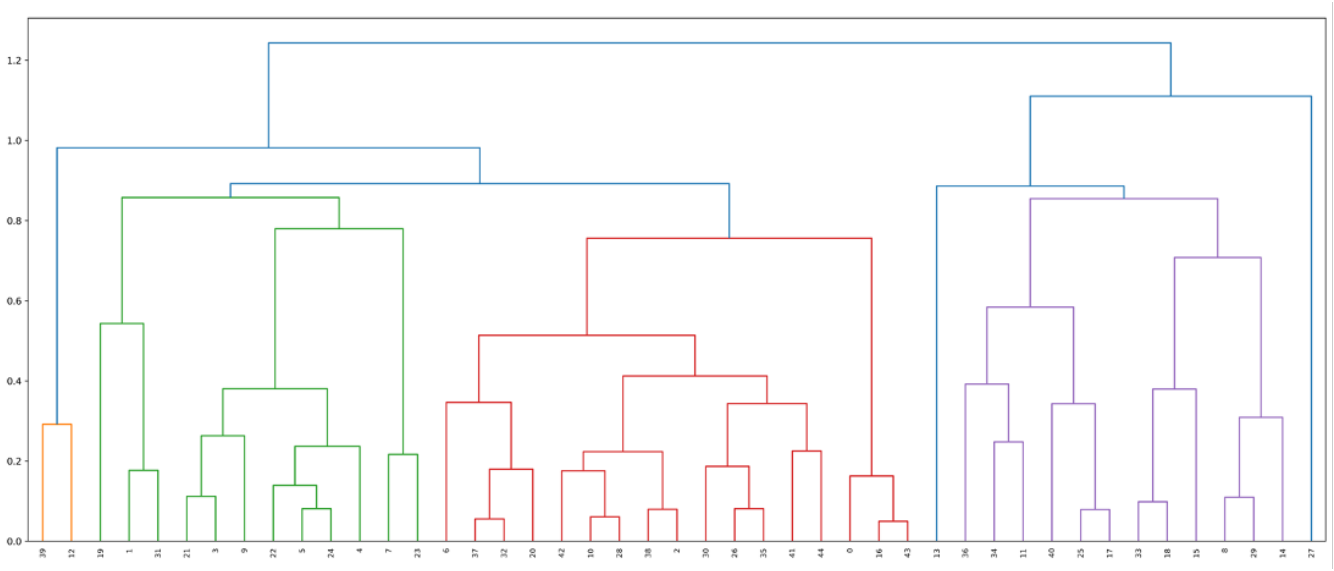

Figure 27: Linkage on clusters of GABAergic cells

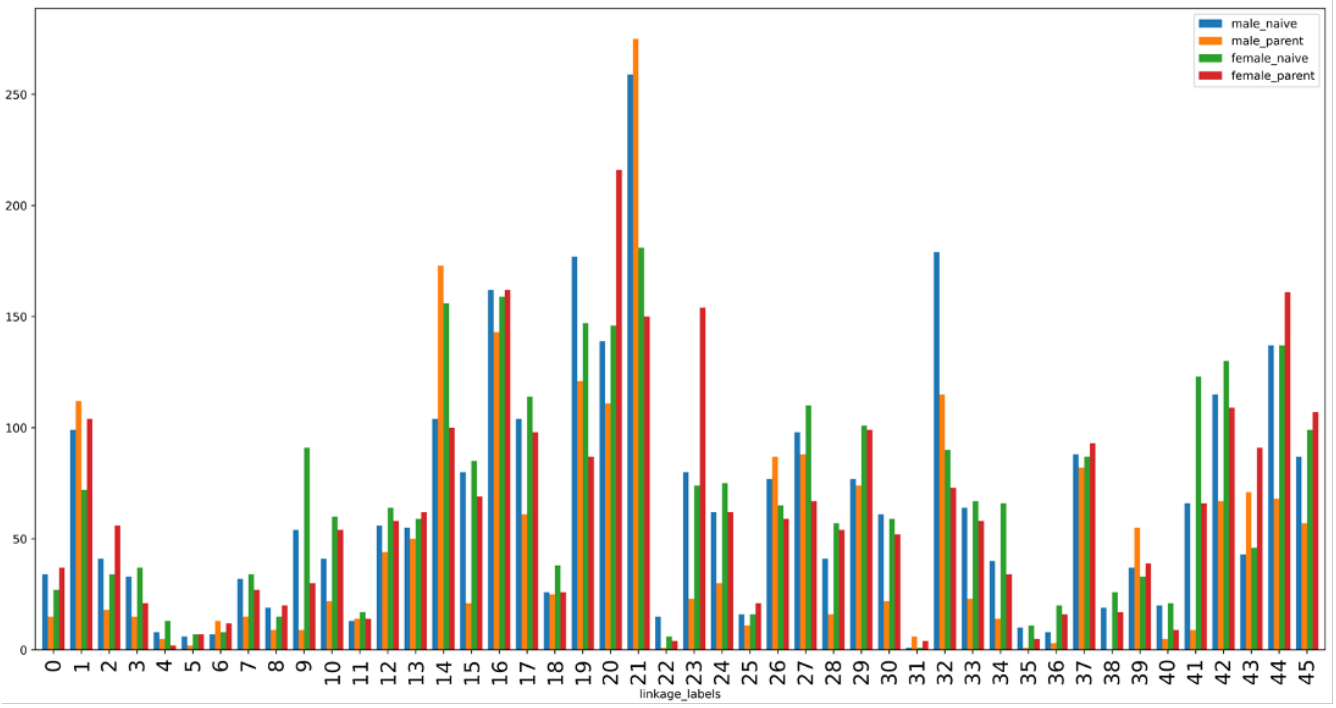

Figure 28: The cluster sizes of GABAergic cells according to the Linakge clustring indices

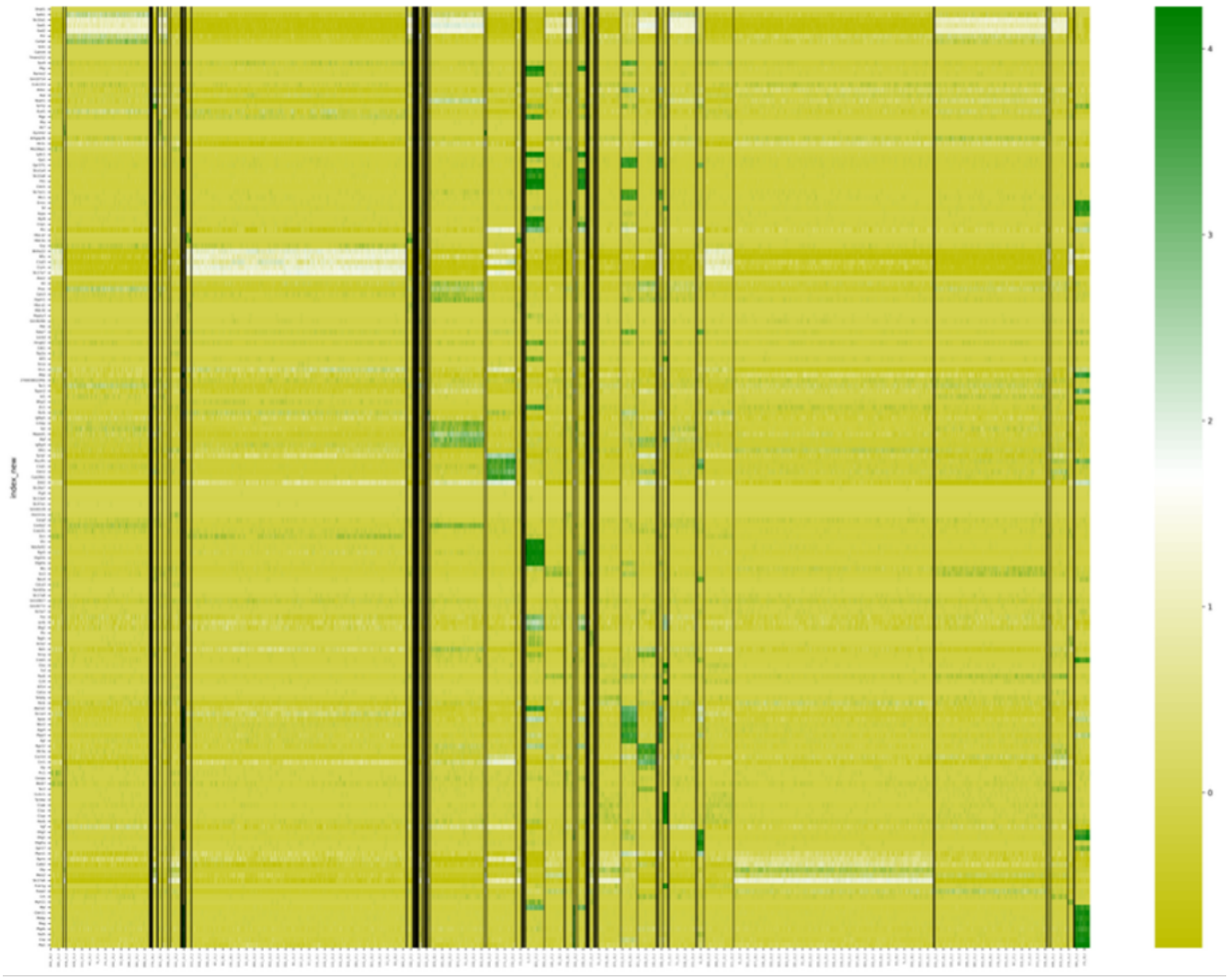

Figure 29: Heatmap for GABA cells: Marker genes expression of all clusters ordered by Linkage

3.5 Cluster Analysis

In order to apply the DBSCAN algorithm, we first calculated the epsilon value and got 1.87431 (more in detail explanation on how in 2.5).

The DBSCAN algorithm resulted with 46 cluster groups on the GABAergic TSNE space (Figure 26) and again we used Linkage to produced the connection between the clusters (Figure 27).

4 Results

By this point, the processed GABAergic cell samples are expressive and concise. We'll use some further statistical analysis to get our desired result – configuring the dimorphic genes for each inspected feature of the studied mice group: female / male, parent / naive.

4.1 Marker Genes

In the same manner as calculated for the previously produced heatmaps, we extracted the 2 marker genes of each cluster, indexed by the Linkage stage. We received the following table output (Table 2).

4.2 Hypergeometric Cumulative Probability Distribution (CDF)

Our premise regarding our dataset is that the overall amount of females and males is evenly distributed between the clusters. For instance, if there are 1000 samples of GABAergic cells and 400 are samples of male mice and 600 are samples of female mice, We'd expect that approximately 40% of the samples of each cluster will be of male mice and 60% of female mice. Similarly for the Parent/naïve feature. However, the premise was contradicted by the calculated CDF for each cluster (Figure 30).

The distribution of males and females inside each cluster varies. From Figure 30, we can infer that there are many clusters with a majority of females or male. The same for the Parenting feature.

These findings are also supported by the bar plot (Figure 28) and by its stacked, normalized version (Figure 32). We can clearly see how the percentage of samples from all 4 labels varies between each cluster and is not uniform.

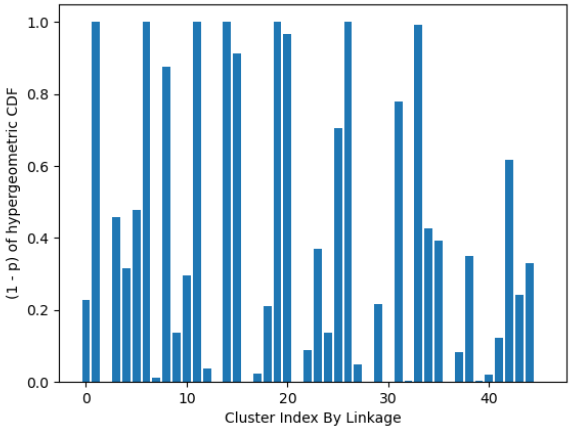

Figure 30: GABAergic clusters' group enrichment female/male

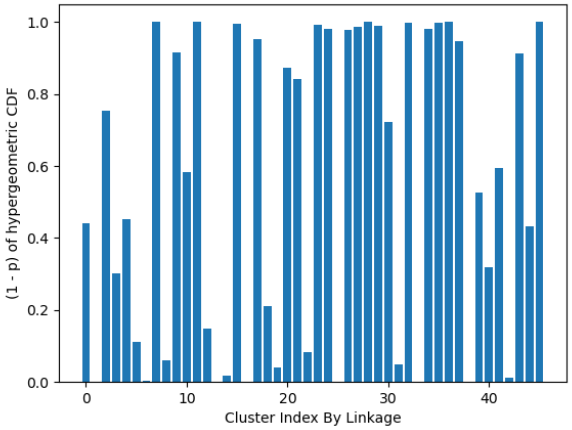

Figure 31: GABAergic clusters' group enrichment parent/naive

4.3 Gender and Parenting Gene Expression Comparison

Similar to figures 21, 22, 6, 7, we scatter the cells' genes to determine how the mean cell of a given cluster varies across different groups (sex/parent). We divided each cluster to four groups: [Female + Parent, Female + Naive, Male+ Parent, Male + Naive]. Looking each time on one of the four groups, we find the mean genes expressions. For each gene, we model its distortion from the mean genes' expressions calculated using all GABAergic samples. This shows us which genes vary the most for a certain cluster. In addition, we can determine the stability of a cluster – the less variation and genes far from y=0 the more stable the cluster will be considered. In addition, from the sub-group perspective, the plots we got helps to identify gens

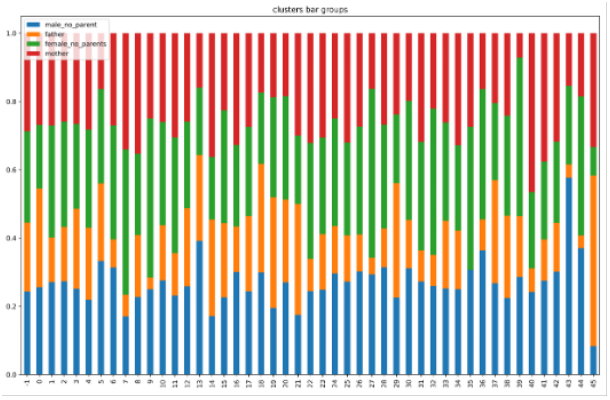

Figure 32: Cluster percentage per sub group

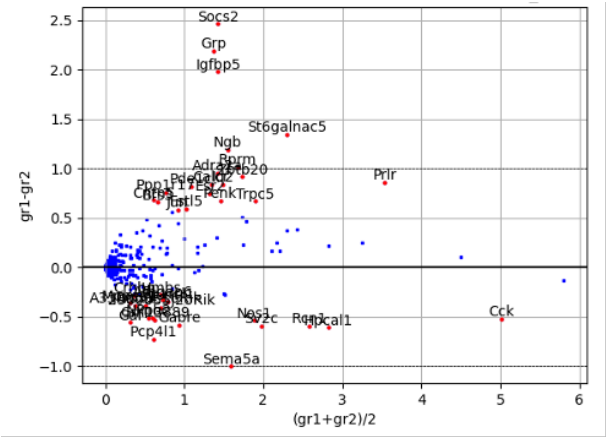

Figure 33: Gean experssion mean vs diff with famles vs males with parents of cluster 1

that are much more significant in one gender than another or for parent versus naïve mice. For example, in Figure 33, we can see that gen Socs2 is a male oriented gene while Sema5a is female oriented.

4.4 Ranksum Evaluation – Identification of Dimorphic genes

For each cluster and for each of the sub-groups described in 4.3, we also computed Wilcoxon Ranksum statistic (Rank-sum). This algorithm describes how far are two distributions from each other. In our case, our premise was that for each gene, its expression levels do not depend on the mice belonging to one of the sub-groups. Meaning, we'd expect that when ordering each gene expression within a certain cluster, samples from each of the four groups will be equally dispersed between the least expressive cell sample for this certain gene and the most expressive cell sample for this certain gene.

Genes who are far from the origin are the ones violating our premise. Their gene expression is correlated to the sub-group (one of the four Without loss of generality) they belong to. This indicates that they are Dimorphic.

Running the Ranksum algorithm for each cluster helps us to gain statistical stability, making our prediction on the Dimorphic genes for each subgroup more reliable.

For instance, in Figure 34, we can see an example of a Ranksum plot output for cluster number 32, where gene Id2 is female oriented and Hpcal1 is male oriented.

We consider a gene to be distorted, or in other words, significant, by a manual threshold of $\log_{10}(Ranksum\ p\ value) > 3$ and $|\log_{10}(males) - \log_{10}(females)| > 0.2$. We circle in red those gens. In every plot of this algorithm, we identify and extract those highlighted gens in order to find the most dimorphic genes across all clusters.

With the data we extracted for each cluster and each feature (gender/ parent) in the previous step, we wish to determine the dimorphic genes.

We gathered all genes who were circled in red (violated the threshold in both y and x axis) at least once and examined which appeared the most for each of the sub-groups.

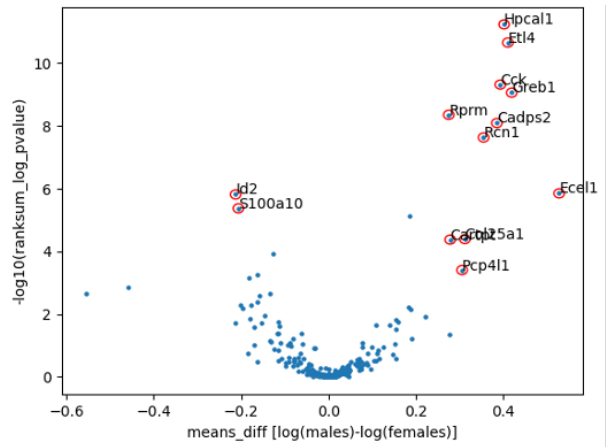

Figure 34: Ranksum for cluster 32 with naive females vs naive males

We presented the output via heatmap, where the color of (group, gen) describe in how many subgroups (sex and parent/virgin) it violated the thresholds. One heatmap (Figure 36) is binary representation for each subgroup (white color indicate the existence of gen) and the second (Figure 35) is an aggregated version of the same plot, where we sum those values across each group.

The second heatmap (Figure 35) shows, for ex-

Male and parent	Female and parent	Male (naïve)	Female (naïve)
Cck	Socs2	Greb1	Cntn5
Greb1	Igfbp5	Pcp4l1	Cox6a2
Ecel1	Cntn5	Ecel1	Arhgap36
Cadps2	Zbtb20	Rcn1	Igfbp5
Pcp4l1	St6galnac5	Cck	Zbtb20
Arhgap36	Thsd7a	Col25a1	Plagl1
Etl4	Dlx6os1	Cbln1	Id2
Sema5a	Ngb	Dkk3	Gpr101
Pnoc	Id2	Etl4	Sst
Nr4a1	Esr2	Nrip3	Junb

Table 1: Table of most dimorphic genes per sub-group in descending order

ample, that gen Socs2 is dimorphic for females who are parents, while Cck is is dimorphic for males who are parents.

Moreover, to better identify the dimorphic genes we gathered the 10 most dimorphic genes for each group (greatest by count) and represented it in a table (Table 1). This is a tabular representation of the most noticeable genes from figures 35 and 36.

5 Conclusion

The cleaned and extracted data of the GABAergic Single-cell RNA samples of the amygdala region of mice can be used to indicate the gender of the mouse from which the cell was sampled and whether or not he has offspring.

Based on Figure 35, the clearest correlation between a certain gene expression and the examined features is the Socs2 gene and female with offspring. Some other worth-mentioning correlations are:

- Igfbp5, Cntn5 also for female with offspring.
- Cck, Greb1 male with offspring.

As for Naive mice, the males and especially females demonstrated low correlation to any gene.

Cluster Linkage Index	First Marker Gene	Second Marker Gene
0	Gpc6	Ntng1
1	Nfib	Calb2
2	Isl1	Meis2
3	Six3	Meis2
4	Tmem132c	Nxph1
5	Gm12301	Dlk1
6	Prdm13	Cck
7	Cadps2	BC048546
8	Ngb	Sv2c
9	Lamp5	Egfr
10	Satb1	Lypd1
11	Ptn	Lhx8
12	Tshz2	Ntng1
13	Cadps2	Rgs12
14	Dlk1	Greb1
15	Nrn1	Prdm8
16	Ecel1	Sytl4
17	Egr1	Junb
18	Lrpprc	Meis2
19	Ecel1	Lhx8
20	Sst	Col25a1
21	Pvalb	C1ql1
22	Nrn1	Nrgn
23	Cryab	Slc6a1
24	Plagl1	Cbln1
25	Synpr	Plpp4
26	Satb1	Plagl1
27	Arpp21	Igfbp6
28	Bex6	Sec14l5
29	Rprm	Sec14l4
30	Lhx8	Nr4a2
31	Arpp21	Ntng1
32	Zeb2	Ndnf
33	Sst	Spon1
34	Fam159b	Gm17660
35	Nr4a2	Vgf
36	Satb1	Prkcq
37	A230065H16Rik	Six3
38	Olfr1392	Six3
39	Pde11a	Synpr
40	Ctxn3	Six3
41	Lypd1	Six3
42	Reln	Ifi27l2a
43	Crhbp	Sst
44	Gpc6	Serpina9
45	Satb1	Lypd1

Table 2: Marker genes: top two gens for each cluster

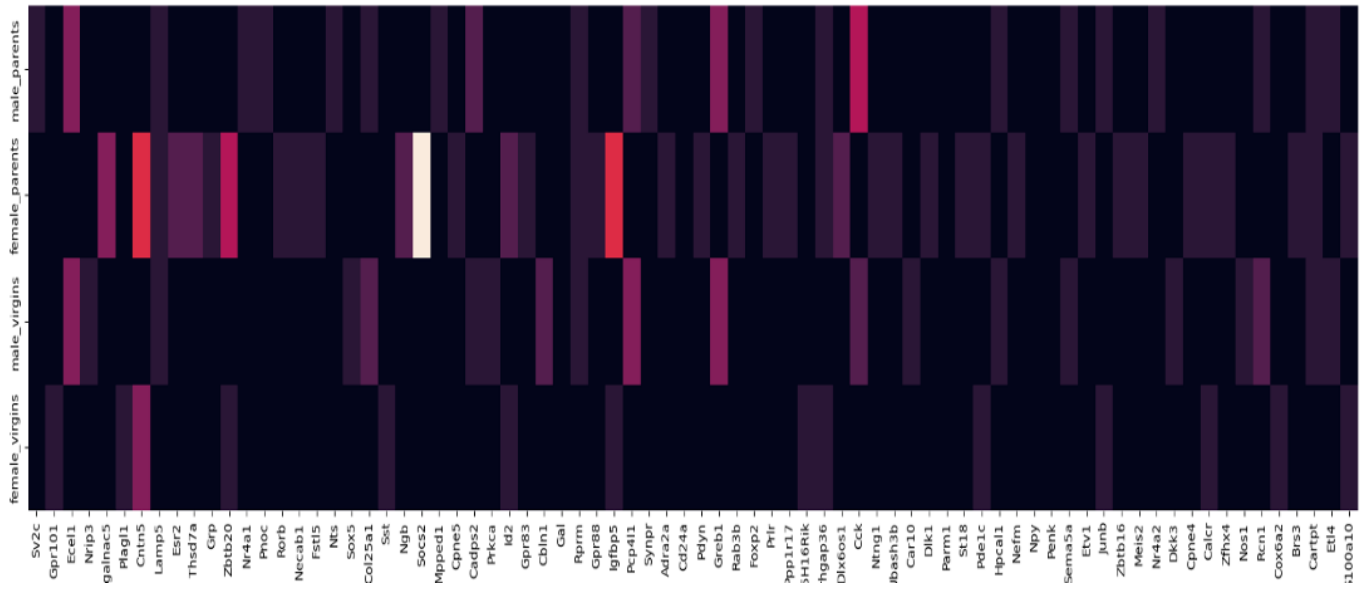

Figure 35: Summarizing for each sub group the times a gen was dimorphic. Higher number represent more dimorphic (noticeable) gen

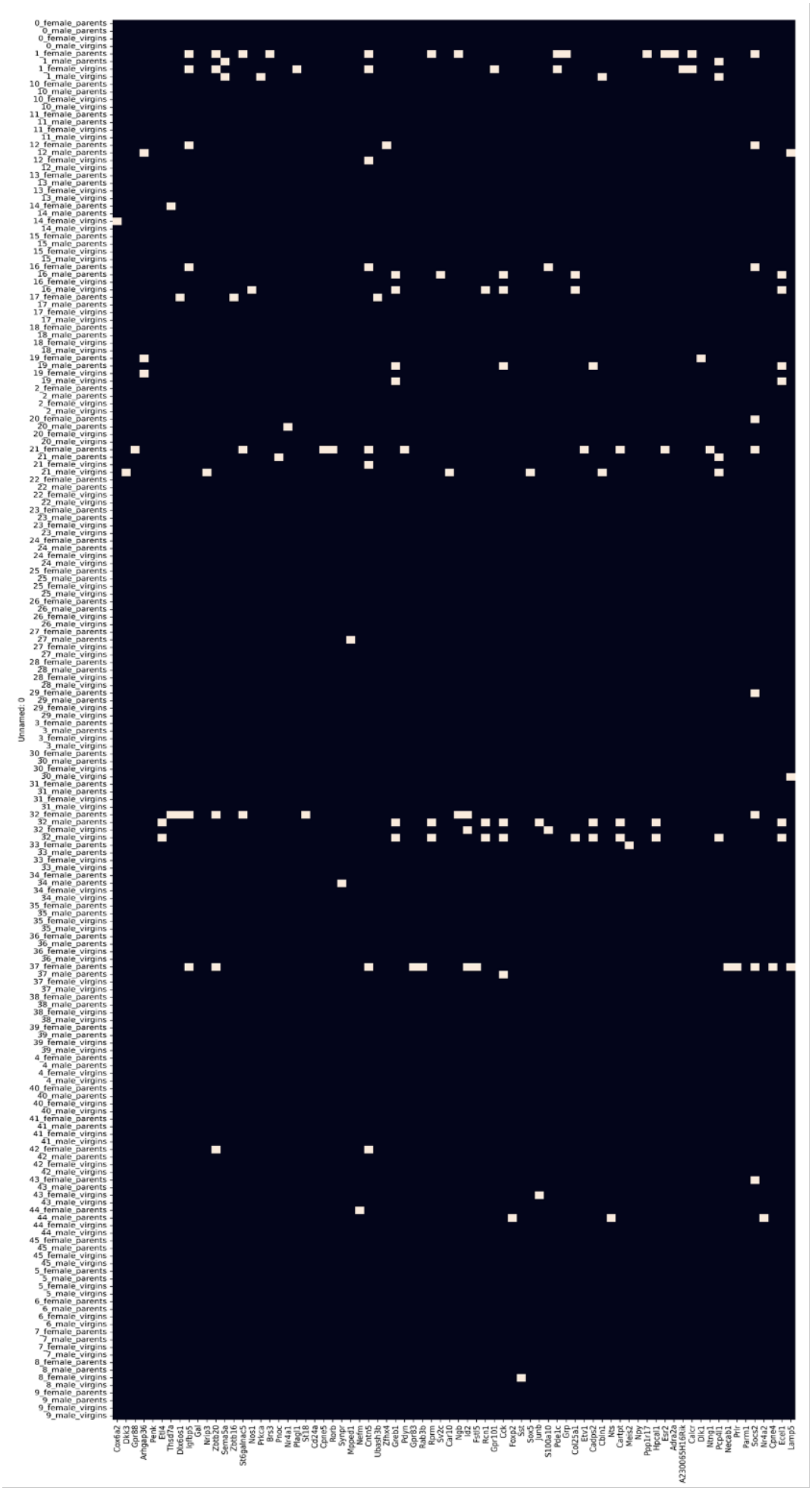

Figure 36: Heatmap-indicator for dimorphic genes of each cluster and sub group