Report_Group4

Mariam

7/13/2022

Introduction

Mouse development and organogenesis occurs as early as compaction and the formation of a blastocyst before preimplantation of the mouse embryo. After the fertilization the most important stages of development are the two cell; four cell; eight cell stadium just as the morula and the blastocyst which already contains three different cell types: trophectoderm; epiblast and the endoderm (Kojima, Tam, and Tam 2014). The blastula is roughly reached after 72 hours (Ciemerych and Sicinski 2005). The very first two cycles after fertilization have a lengthened duration compared to the fourth and eighth cell stage. This is due to the chromatin remodeling and the decondensation of maternal and parental chromatin in order to gain a functional nucleus (Ciemerych and Sicinski 2005). The dynamic cell changes are controlled by so called D-cyclins, many transcription factors, and mainly performed by DNA- and Histone methylases and demethylases (Mihajlović and Bruce 2017), (Sha et al. 2019). In the one cell stage and right after the end of the two-cell stage entering the fourth cell stage, the minor and the major Zygote Genome Activation (ZGA) onsets (Mihajlović and Bruce 2017). This implies that from now on the development will be directed by the zygote's genome transcripts, while the maternal mRNA transcripts will be degraded, and thus the expression pattern will drastically change (AOKI 2022). We will concentrate on the dynamic change of the gene expression especially in the fourth cell stage. In comparison ZGA takes place between the fourth and the eight-cell stage in humans (Xie et al. 2010). Since mammalian embryos develop under low oxygen conditions, managing these conditions and providing enough oxygen for morphogenesis and cell proliferation and tissue formation is essential. In order to prevent hypoxia, a low oxygen condition, while embryogenesis, there are hypoxia sensitive genes which will be activated (Dunwoodie 2009). One of the most important factors for this matter is the Hypoxia Inducing Factor (HIF). HIF binds to the HIF-Responsive element, which is encoded by three genes. Whenever HIF is absent or epigenetically silenced, the morphogenesis of the heart is impaired. Especially affected is the formation of the endothelium in the cardiovascular muscles and the chamber formation of the heart. To conclude, in order to develop a healthy cardiovascular system HIF is essential (Krishnan et al. 2008). A rather hidden and enigmatic role plays tissue restricted antigens (TRAs) in embryonic development. With the aim of establishing functioning T cells, which recognize intruders as pathogens via T cell receptors (TCRs), the T cells need to be trained (Alberts et al., 2015). The positive and negative selection in the thymus allows T cells to recognize self-antigens which are displayed by MHC molecules on the cell surface. The expression and regulation are controlled by AIRE autoimmune regulator and Fezf2 (Monteleone-Cassiano et al. 2022). The role of TRAs in the crucial stages of embryonic development is yet unknown, just as the immune suppressive impact of Fezf2 regulator in those cells (Takaba and Takayanagi 2017).

Materials

1. R and RStudio

This project was entirely done in R(R Core Team 2022) version 4.2.0 (2022-04-22) and RStudio (RStudio Team 2021) version 2021.09.0.

2. Affy Packages

The microarray chips used in the the research of Xie et al. are Affymatrix GeneChips. In order to process and analyse these chips we used the affy package (Gautier et al. 2004) that was installed using Bioconductor. Affy is an R package that is used to analyse gene chips of the affymatrix type. Some of its many functions are to read in data and do quality control checks. The data are read in as .CEL files.

3. Brainarray and loading the Chip Describtion Files of mouse and bovine

The chip descritption files (CDF) of our 2 data sets (mouse and bovine) were downloaded using BrainArray (Dai et al. 2005). Brainarray is an online data bank that gathers re-analyzed existing Affymatrix Genechip data "with updated probe set definitions," (Dai et. al, 2005) to offer custom cdf files with better gene annotations and calculations.

4. Bioconductor

Bioconductor (Morgan 2022) gathers different packages that are used in R, in order to widen the analysis of gene expression data sets. Most of the packages that we used in our project are installed through Bioconductor, this includes: limma, affy,vsn, GSEA and AnnotationDbi.

5. Tidyverse

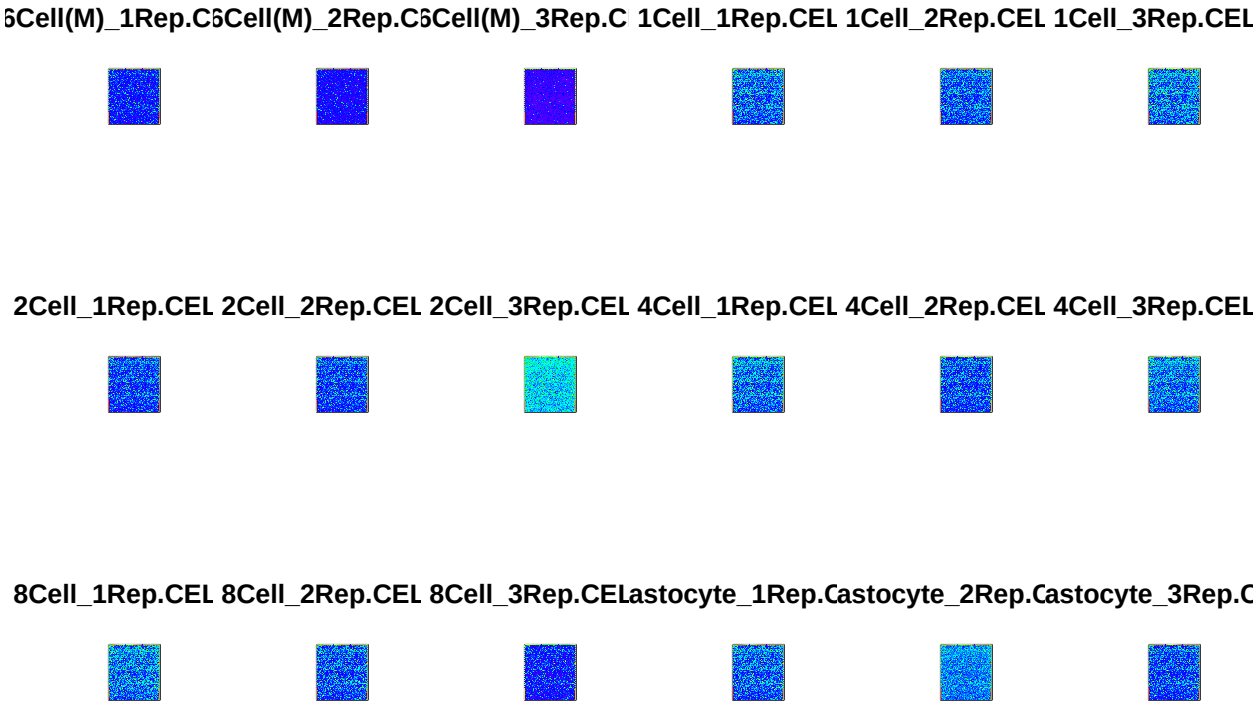
Tidyverse is a collection of packages used for "data import, tidying, manipulation, visualisation, and programming" (Wickham et al. 2019). It is analog to Bioconductor.

Methods

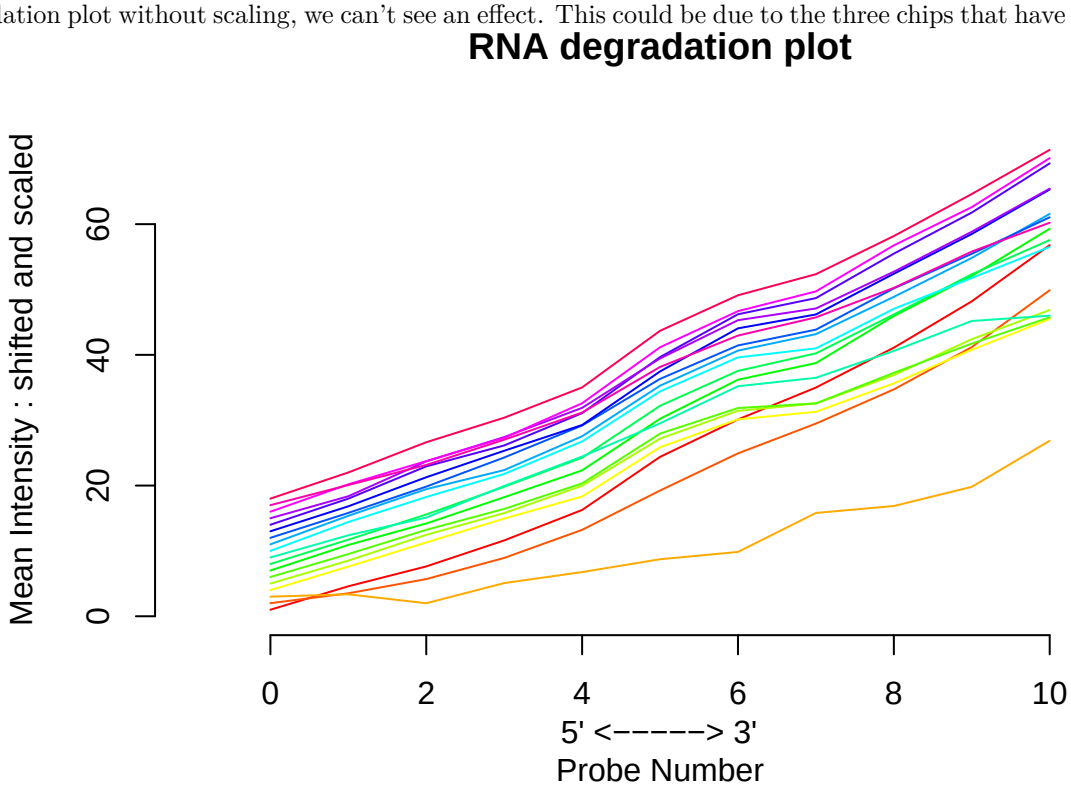
1.Quality Control

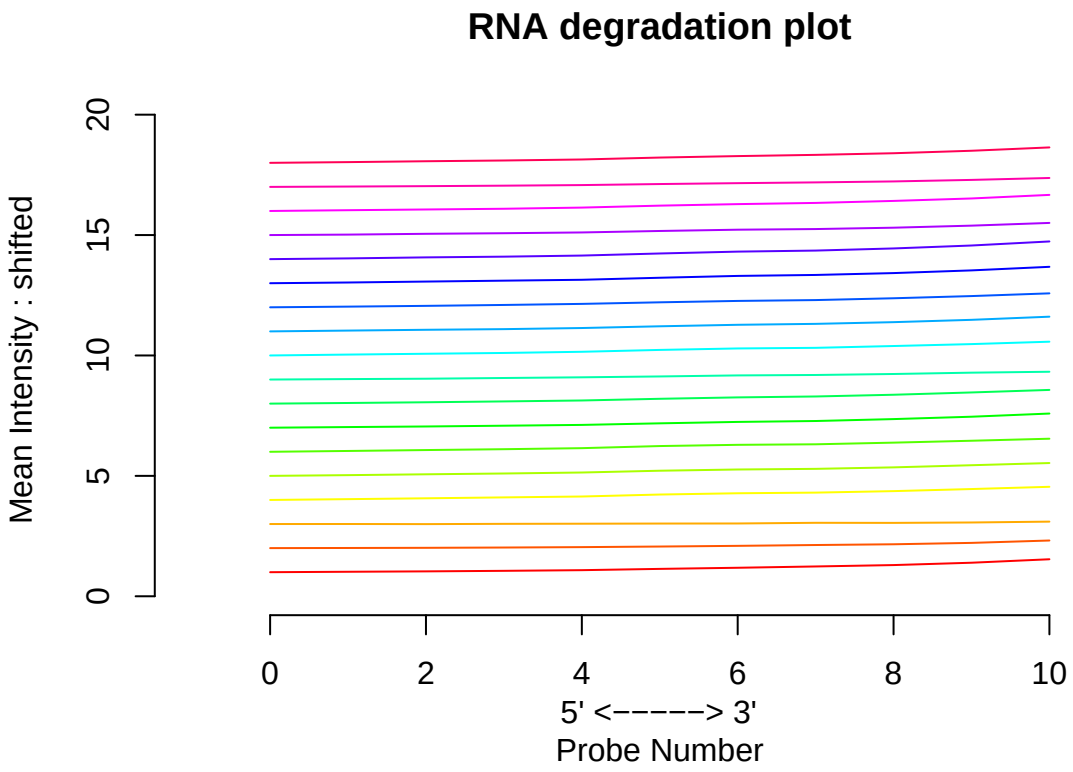
Mouse chips

After reading in the data, we examined the chips of the mouse data set to see if any of the chips have quality issues. This was done using different objectives. Firstly, we read the chips as images in order to see if they differ from the overall expression trend. We noticed three chips that seemed to differ. The first chip,2 Cell 3rd replicate, is distinctly over-expressed and the other two, morula 2nd and 3rd replicate,were under-expressed.



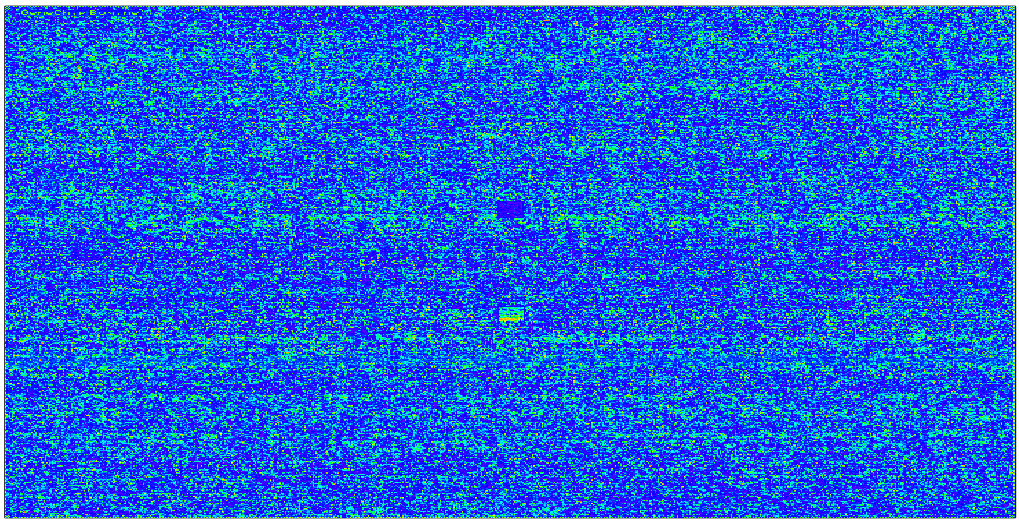
The second step in the quality control was done through an RNA degradtaion plot on the data set, that is shifted and scaled. The RNA degradation plot, follows the degradation of the RNA by targeting the probe set in different regions of the selected transcript, the central section, the 3 prime and the 5 prime. This allows assessing the degradation rate of individual transcripts by examining the 3'/5' probe-set signal ratios. A good RNA degradation plot would show a steady upward trend with minimal crossing. In our case we can see that the orange line follows a different trend than the others and that there is crossing.On the other hand, if we only shift the RNA degradation plot without scaling, we can't see an effect. This could be due to the three chips that have low quality.





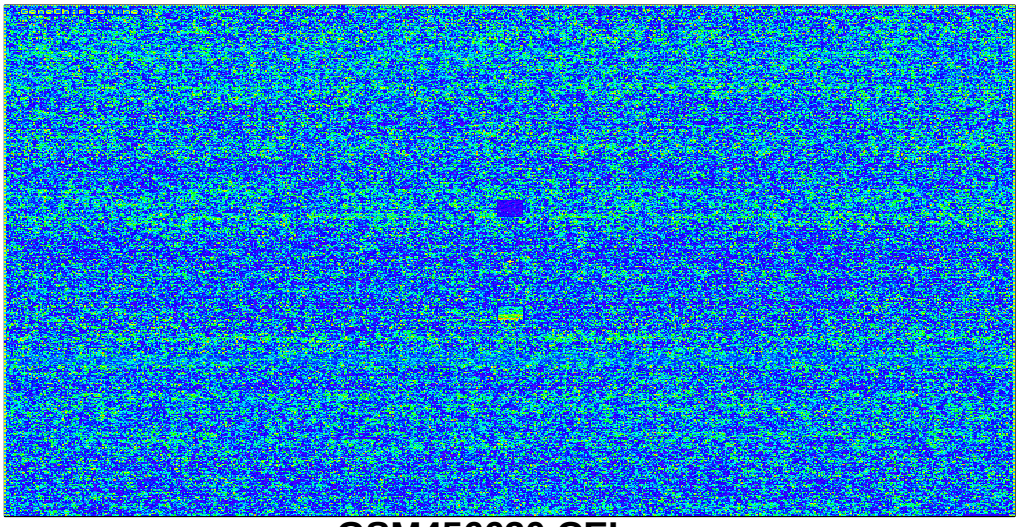
Bovine

The same procedure was done for the bovine data set. Through the quality control of the bovine chips, we saw that the last chip had quality issues, as the dye showed a difference from the rest. This can also be seen by plotting the RNA degradation plot of the 16 chips, as the 16th chip (GSM456642) (blastocyst, second replicate)

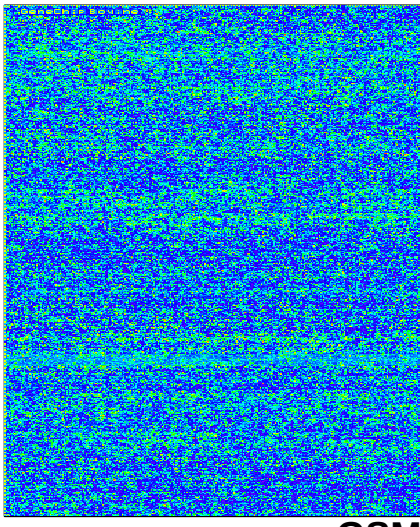


crosses the rest of the chips.

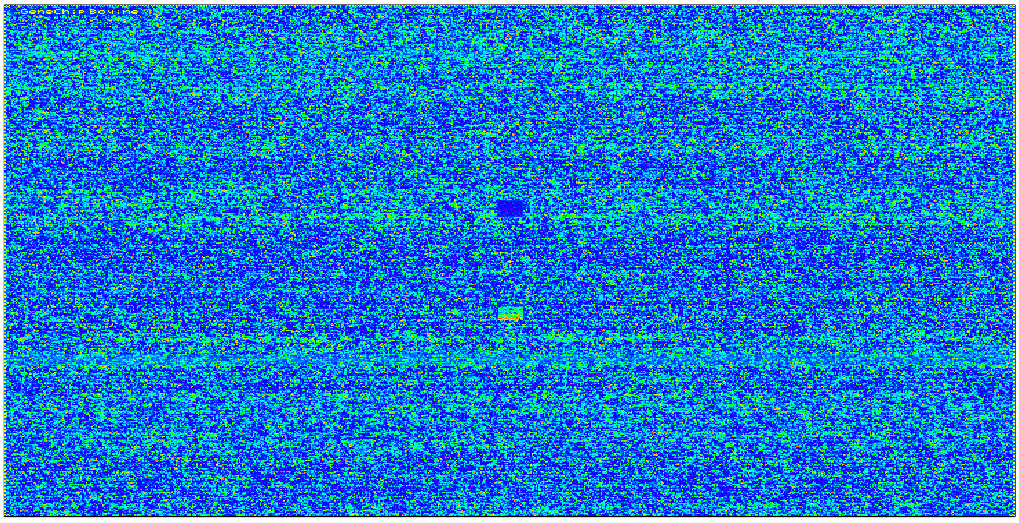
GSM456628.CEL.gz



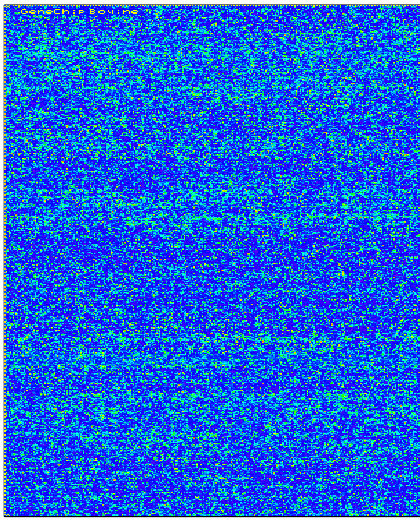
GSM



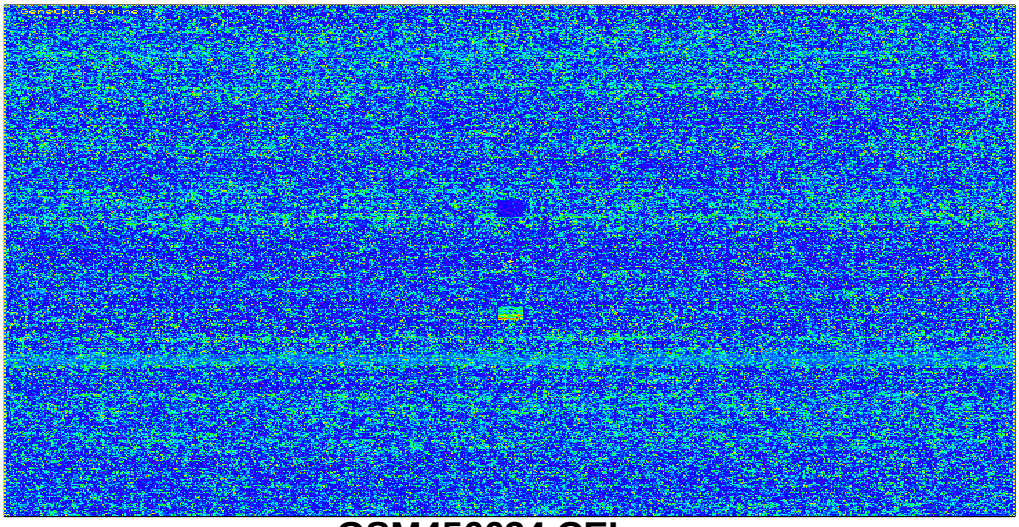
GSM456630.CEL.gz

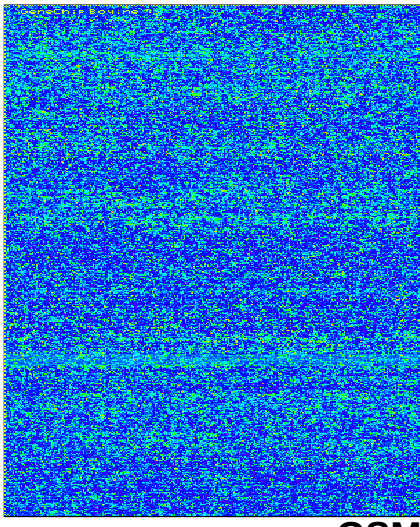


GSM

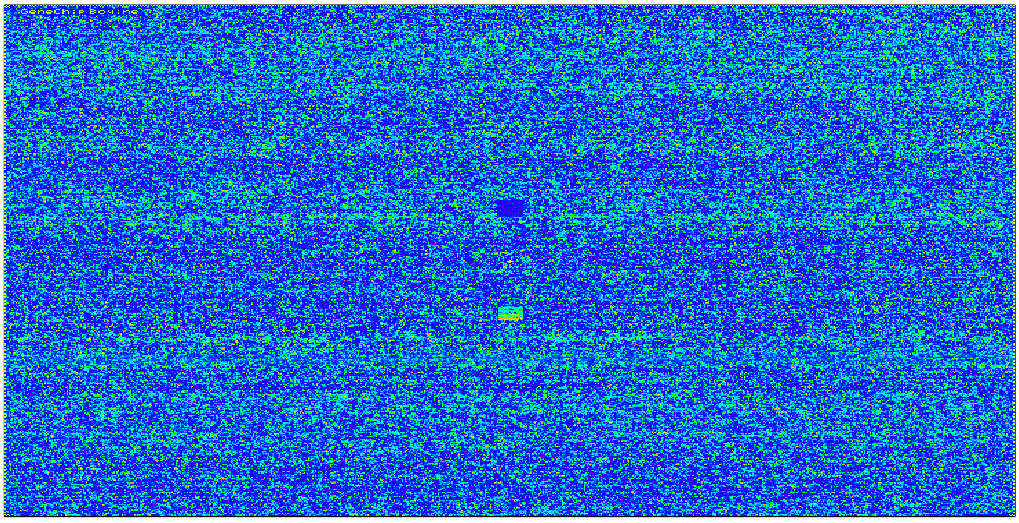


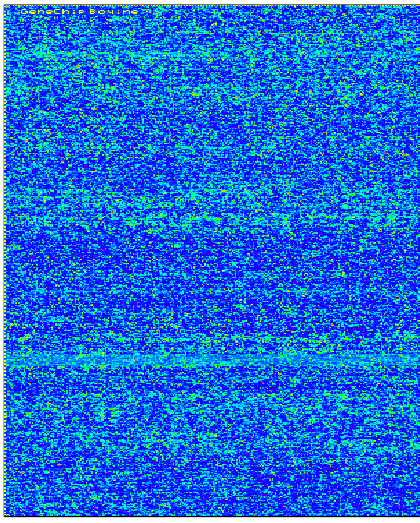
GSM456632.CEL.gz **GSM**



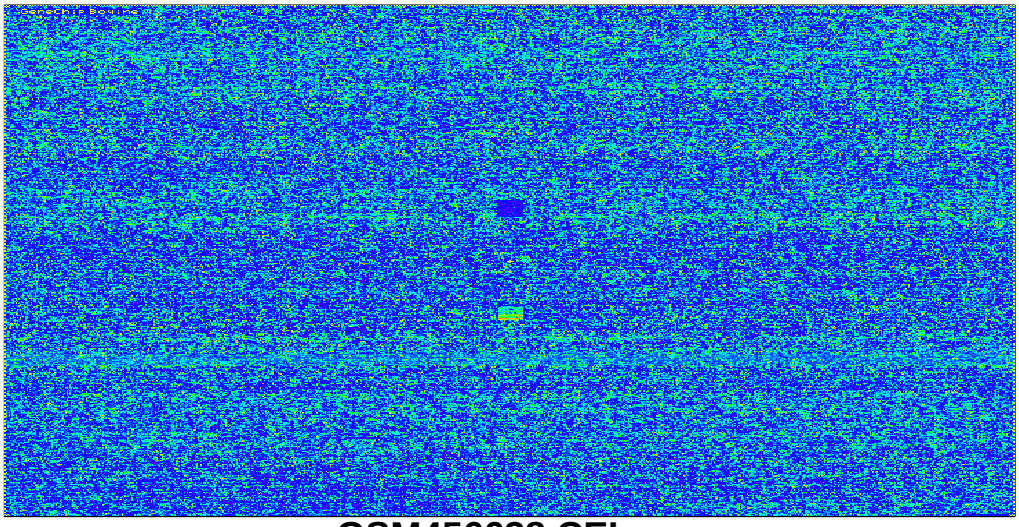


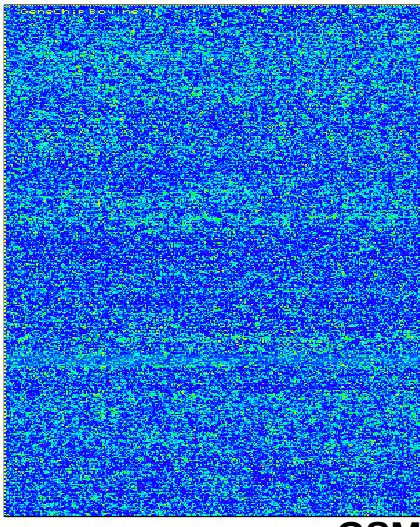
GSM456634.CEL.gz **GSM**



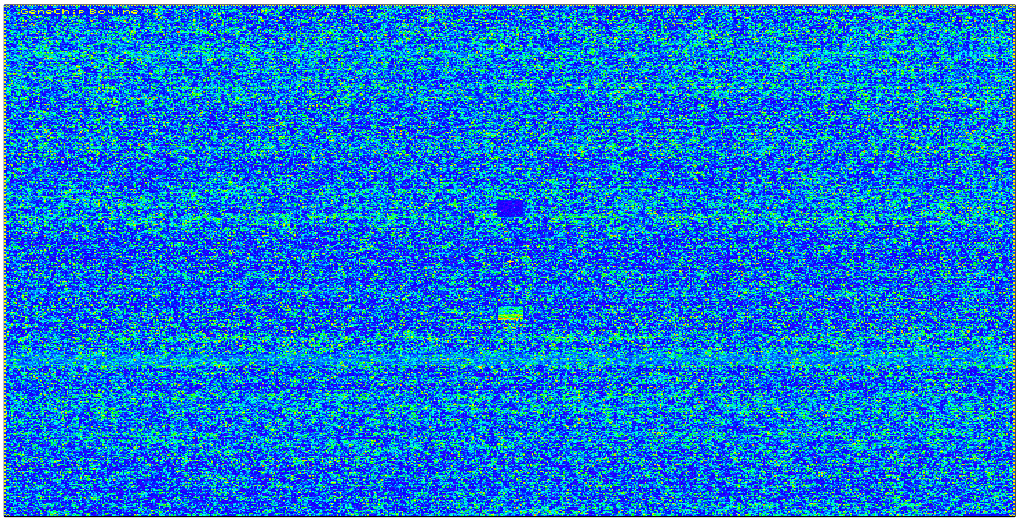


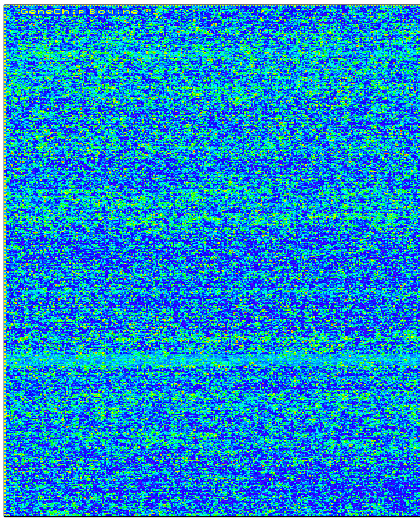
GSM456636.CEL.gz **GSM**

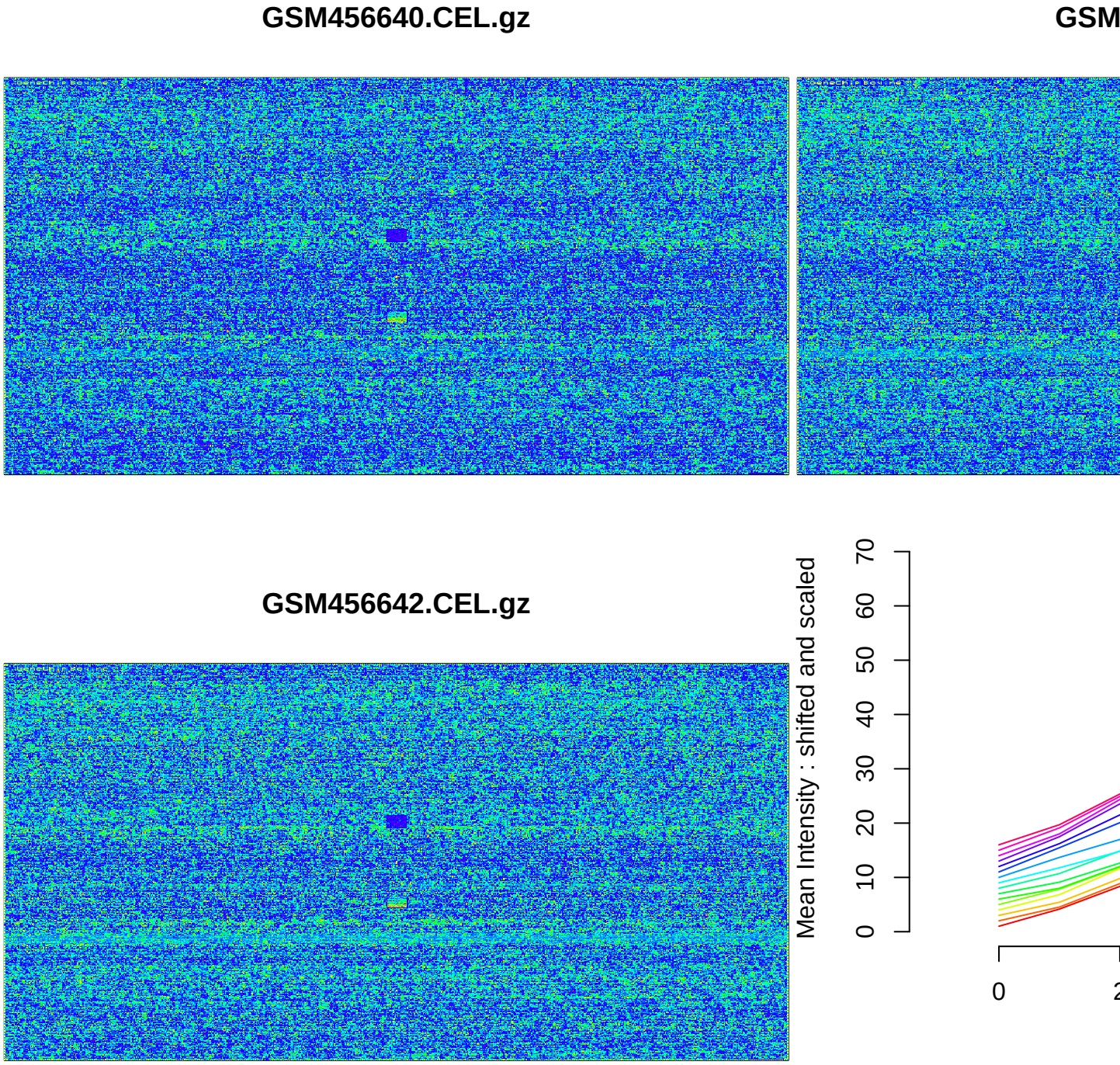


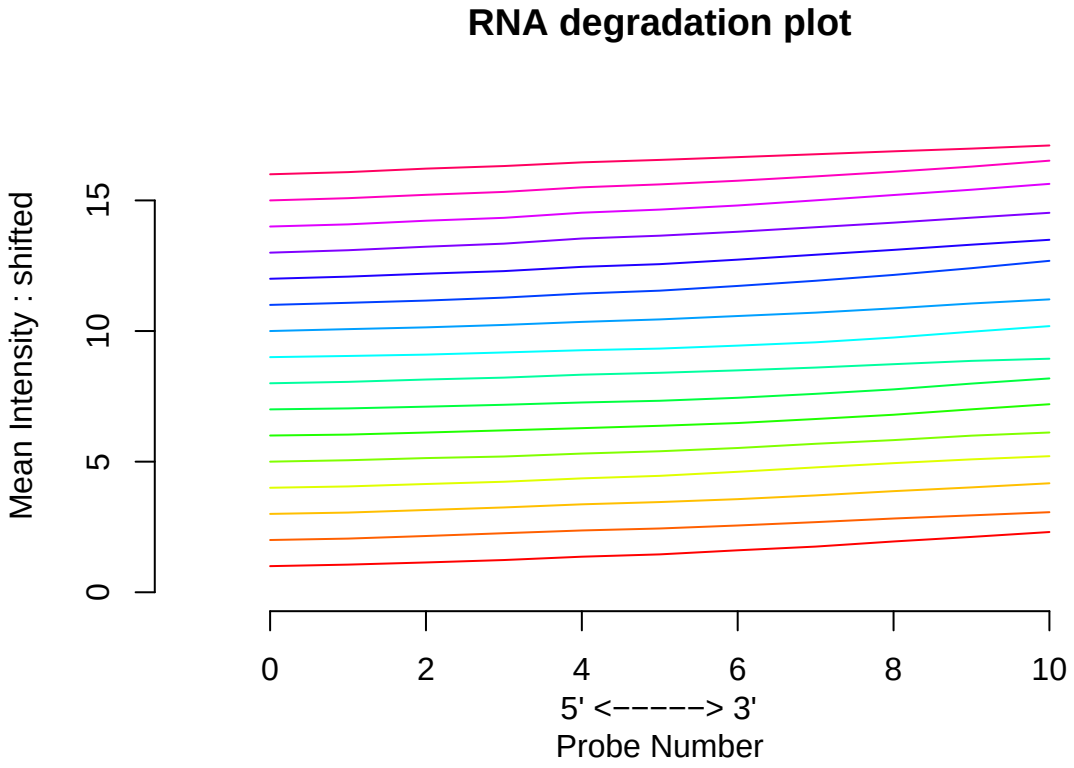


GSM456638.CEL.gz **GSM**






GSM456640.CEL.gz
GSM
GSM456642.CEL.gz
Mean Intensity : shifted and scaled
0
10
20
30
40
50
60
70
0



2.Variance Stabilization Normalization

Variance Stabilization Normalization (VSN) is a stastistical method,developed by Huber et al. (Huber et al. 2002) that is used for micro-arrays to reduce background noise, optical illusions and dye irregularities. It is done through a log transformation in order to get a better concept of perception. It includes three main steps, the normalization, which is done through data calibration, the mean- variance- dependance of the the model, and a variance stabilizing transformation. (Huber et. al) For both data sets (mouse and bovine), the vsn was visualized using different plotting techniques.

1. Mean versus Standard deviation plot

The quality of the VSN can be visualized using the mean versus standard deviation plot (meanSDplot). The standard deviation should not have a strong correlation to the mean/variance and thus the red line of the median estimator should be horizontal. (Dinkelacker 2019) This was done for both data sets.

2. Density Plot

The density plot is used, to plot the density function against the log intensity of each chip. This way we can confirm if the vsn was succesful or not. If the curves are well adjusted after the vsn, this would mean that the normalization was successful. As the QC showed, chips GSM456666,GSM456674,GSM456675 are of low quality.That's why we will be disregarding them for the rest of our analysis.

3. Principal Component Analysis

Principal component analysis (PCA) is used for dimension reduction. This is a method to summarize the information given in a data set. We first define the number of principal components that explain the total variance of the data sets. Through the plotting of the data sets the correlation of the data points can be measured by their distance to each other in the graph. Data points with high correlation will be closer to each other. This method is used to see if the different chips are similar to each other or not

4. Hierarchical Clustering

After performing the QC, we proceeded to cluster the 15 chips. We created a distance matrix using the euclidean distance. After that the hierarchical clustering was done using the average linkage method. This was plotted and a dendogram was formed. The bigger the hight difference the more different the groups are.

5. Finding TRAs in our data set

Through the available data set provided by Dinkelacker, we were able to match the TRAs with our mouse data set using R.

6. Differential Gene Expression Analysis

The differential gene expression (DGE) analysis "refers to the analysis and interpretation of differences in abundance of gene transcripts within a transcriptome." (Conesa et al. 2016) (Conesa et al., 2016) It is done in R using the limma package provided by Bioconductor. (Phipson et al. 2016) Limma uses the linear model as an approach for the DGE, by simply forming a design matrix "which indicates in effect which RNA samples have been applied to each array" (Phipson et al., 2016) and a contrast matrix, where we define which objectives will be compared to each other. In our case the contrast matrix compares cell stages to each other and the design matrix, designs a matrix that groups the chips by the cell stage they belong to. After that a linear model will be fitted to our design matrix, and in the end the contrast matrix will be fit with the linear model. Limma uses the Bayes method in order to use probability to represent all uncertainty within the model. Here it moderates the standard errors of the estimated log-fold changes. It is calculated using the Bayesian Theorem, which is then used for hypothesis testing, in our case a t-test. The differential gene expression was performed for the mouse and bovine data sets respectively.

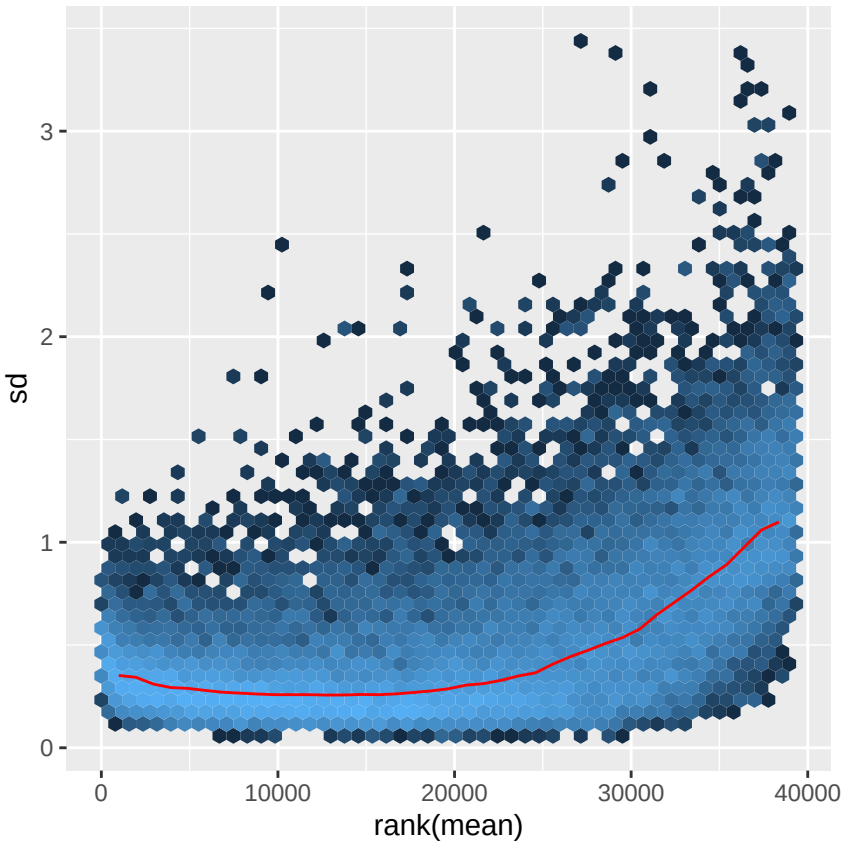
7. Gene Set Enrichment Analysis

A gene set enrichment analysis (GSEA) is used to identify if a set of genes, is enriched in expression. The analysis uses previous knowledge in order to see if a set of genes is related by shared criteria. This criteria can be a certain pathway or a functional classification. (**gsea?**) The GSEA is based on the results from the DGE that includes the results of the t-test and the p-values. Additionally we use The Molecular Signatures Database (MSigDB), which is a resource that contains annotated gene sets for our species and pathway analysis. (Dolgalev 2022) Using the annotation packages for Mus musculus and Bos taurus gives us information from different identifieres (Carlson 2022) For us the GSEA will help us with enriching pathways that might play a role in tissue formation.

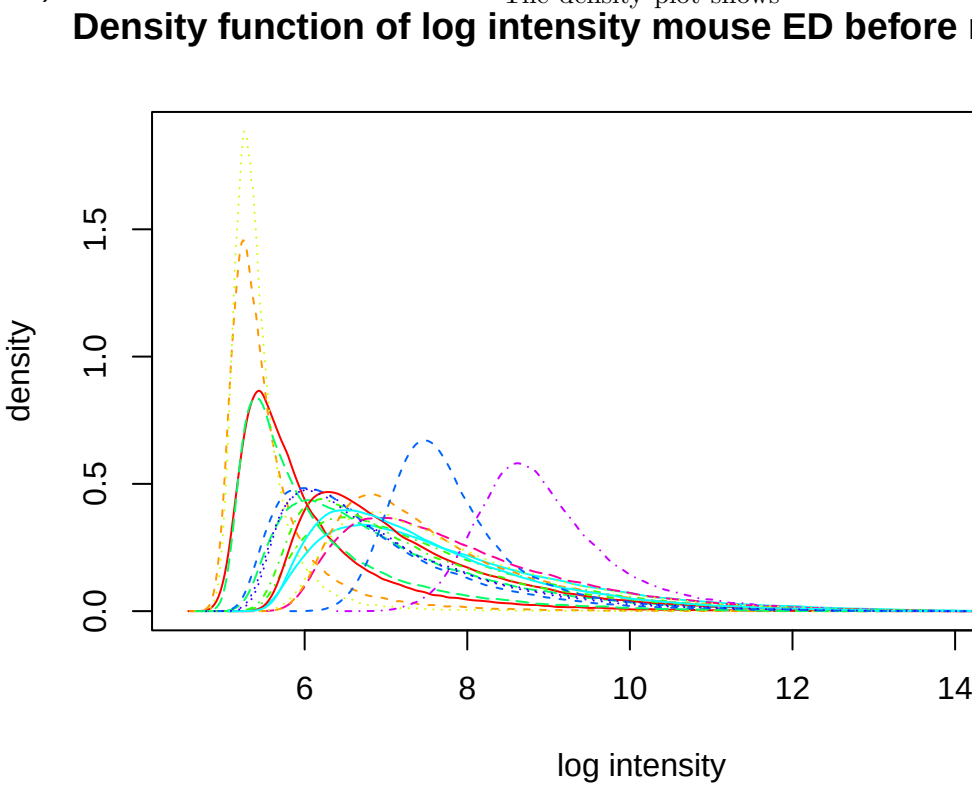
Results

1. VSN

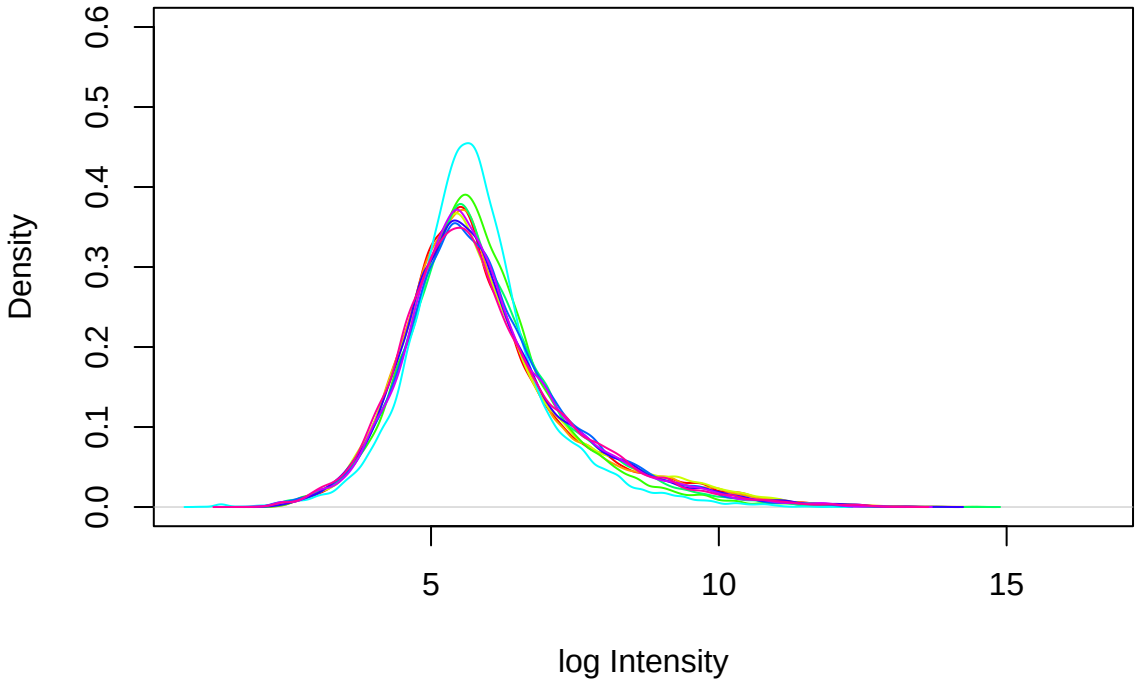
After the QC, we proceeded to normalize our data using the variance stabilization method. To see if the VSN was successful we used two plotting techniques, the mean versus standard deviation plot and the density probability against the log intensity. The mean versus standard deviation, shows a slight upward trend of the red line which represents the median estimator. If all chips were of good quality, the median estimator would show a horizontal line.



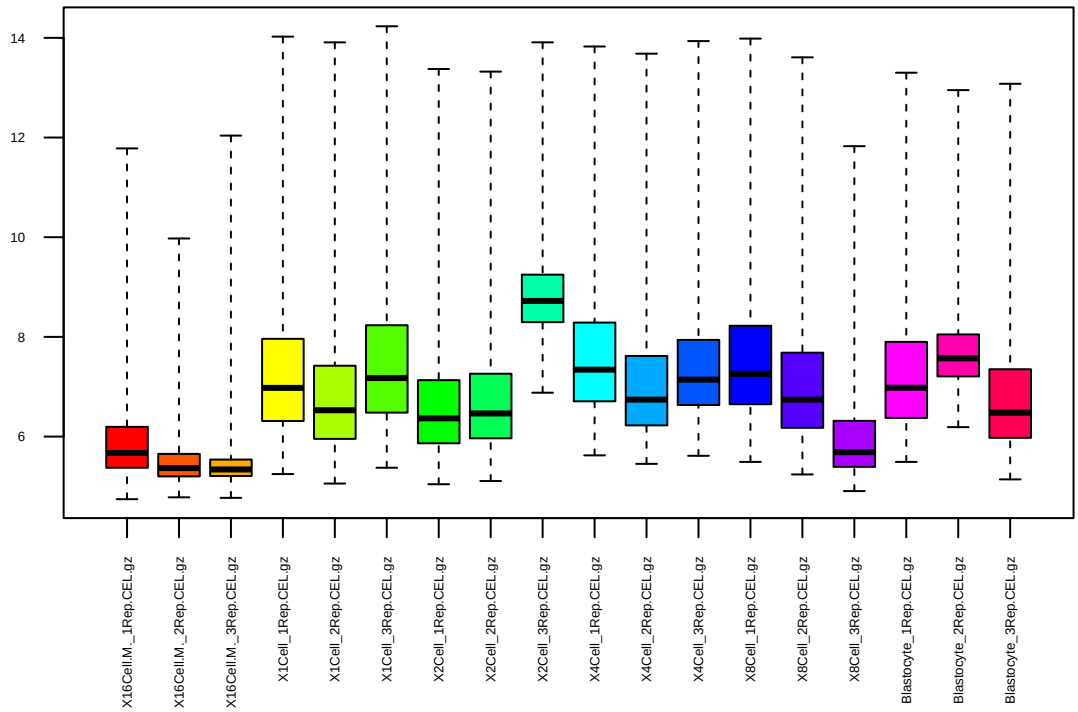
The density plot shows

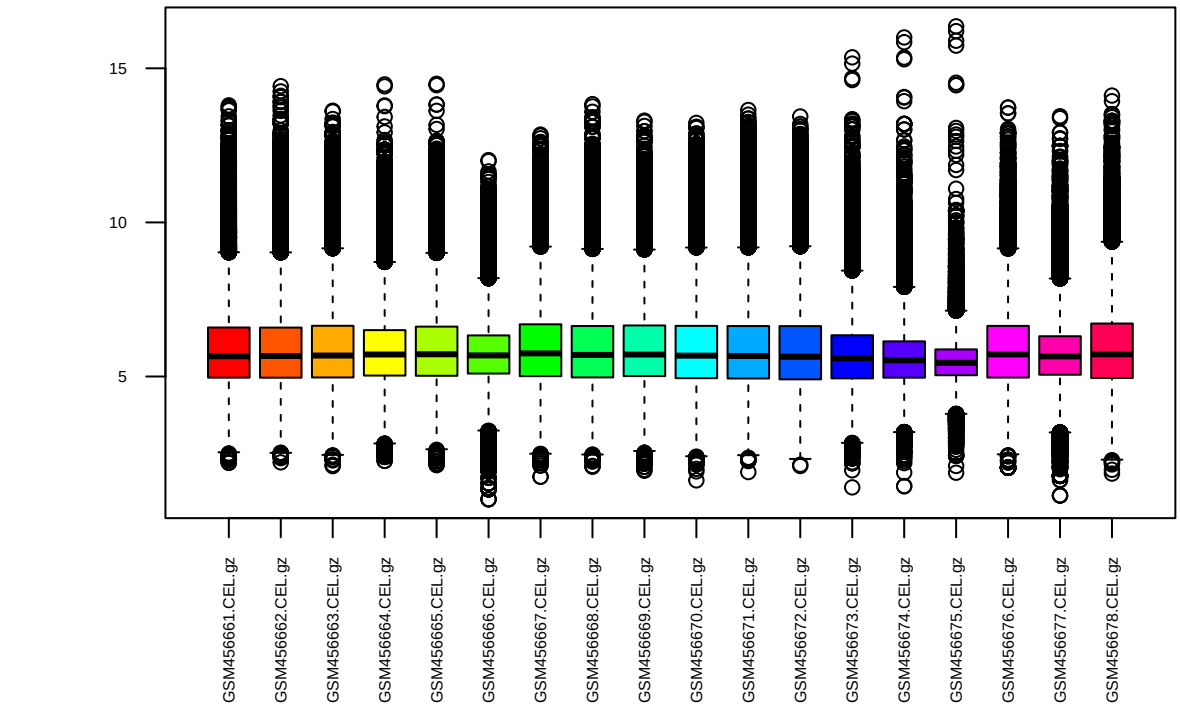


that all chips align with slight differences.



Additionally, we plotted the VSN reuslts using boxplots. The boxplots show us that the median of all chips align on one line, but comparing this boxplot to the boxplot before normalization one can see that the boxplot after VSN has more outliers.





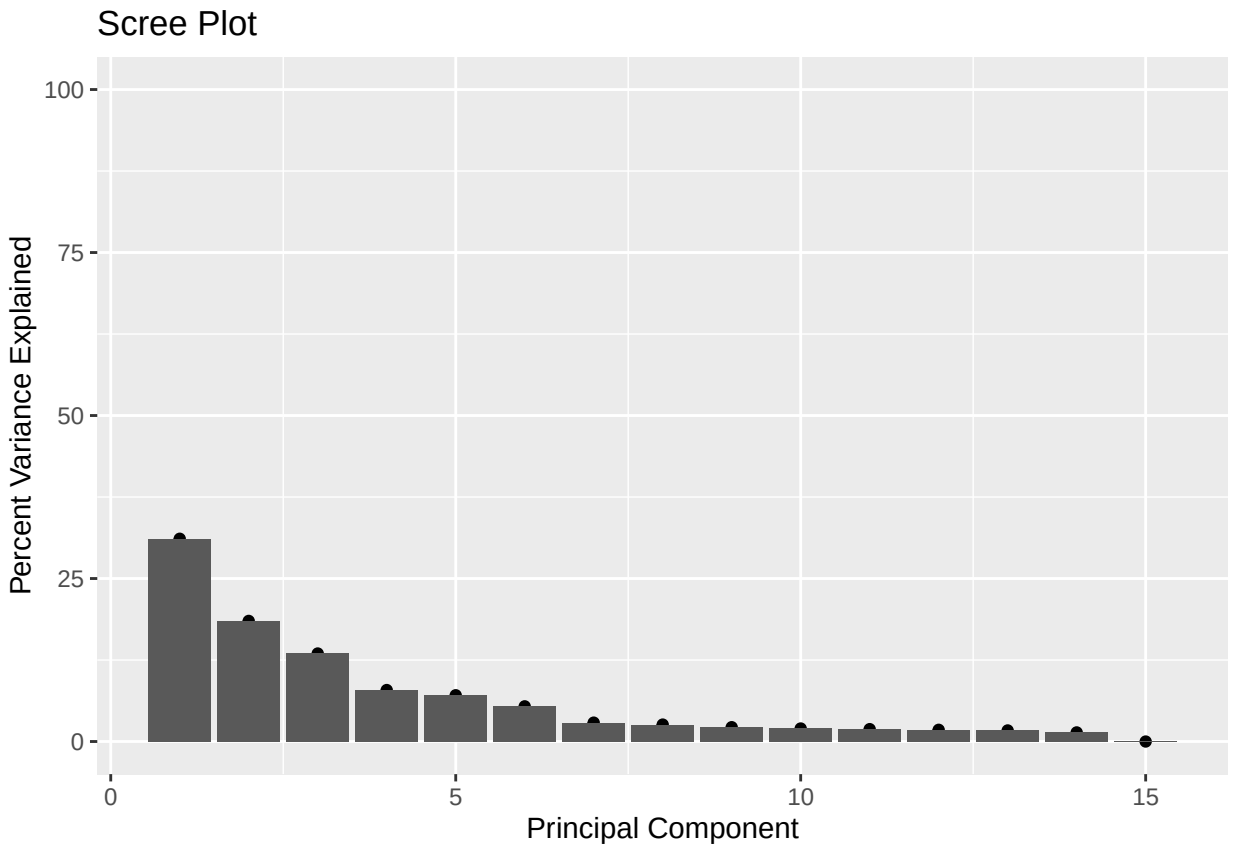
After all those steps were done, we decided to disregard 3 chips in total (see Quality Control)

2. Principal Component Analysis

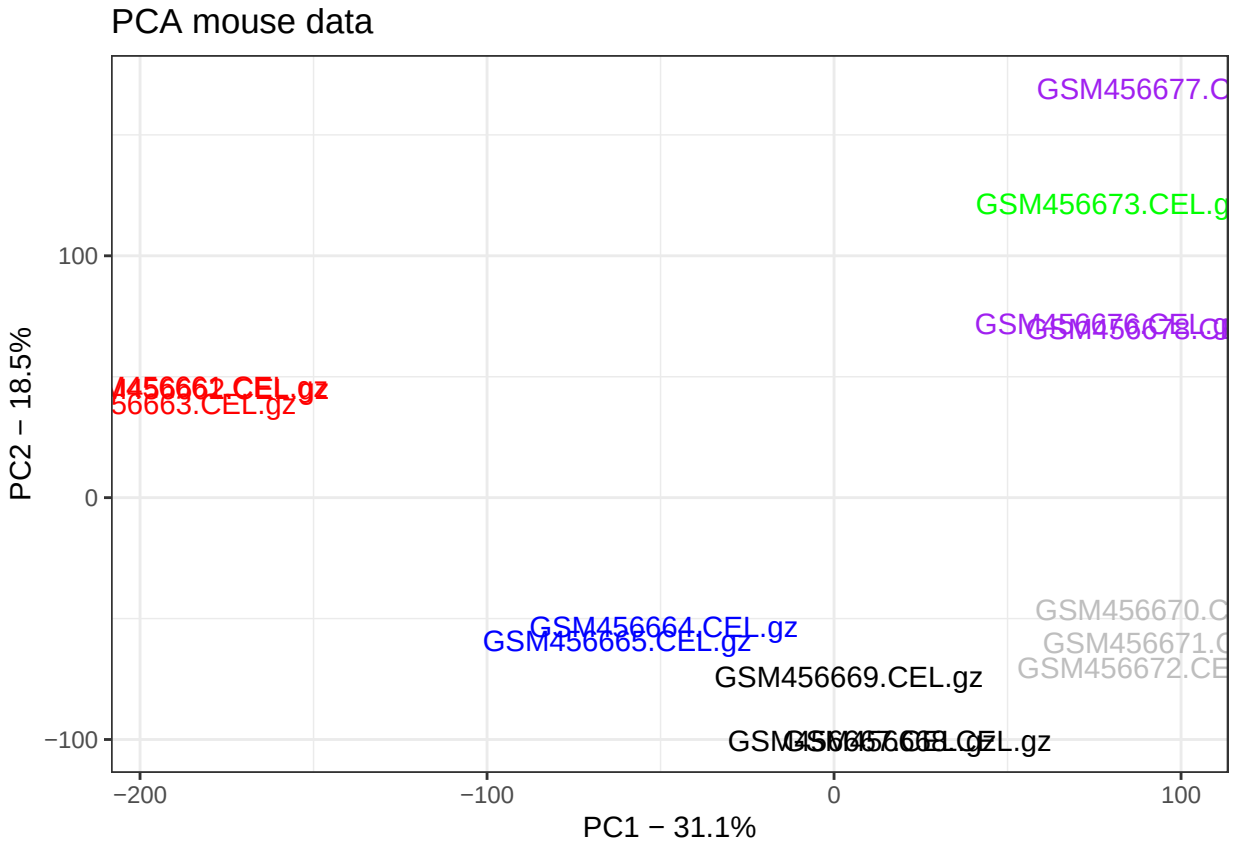
We wanted to see if the different replicates of the chips would show high correlation if they belonged to the same stage and to see if we can reduce the dimension of our data set. This is why we performed a principal component analysis. Here the variables are the different genes expressed (39281) and we have 15 samples. We transposed the matrix in order to have the columns as the genes and the samples as the rows. After that a scree plot was done in order to see how many principal components are needed, which are 2 as those explain around 50% of all the data variance. A PCA was done using the PCA function in R. Through the ggplot2 we can see that the first two samples (one cell stage, first and second replicate) have excellent correlation which means one sample and explain the entirety of the second sample. Sample 3 (GSM456663) is also close to them. This is also the case in sample 7 and 8 (GSM456667 and GSM456668) but here the third replicate is farther than the case of sample 3, which hints at a difference between the gene expressions although it's in the same cell stage. The blastocyst first and third replicate (GSM456676 and GSM456678) are closer to each other which hints at good correlation. What one can notice is that the second replicate of the blastocyst (GSM456677) is a lot more farther away than the other two replicates, which could mean that there are errors in the gene expression itself as the variation is way too high for it to be in the same cell stage.

```
setwd(paste(projectPath,"Sessions","RDS_Files",sep = "/"))
data_ohneaffx = readRDS("data_ohneaffx.RDS")
pca = prcomp(t(data_ohneaffx), scale = TRUE)
pca_var = pca$sdev^2    #calculating variation
pca_var_per = round(pca_var/sum(pca_var)*100,1)    # variation into percentage
##### Scree plot
qplot(c(1:length(pca_var_per)), pca_var_per) +
  geom_col()+
  xlab("Principal Component") +
  ylab("Percent Variance Explained") +
  ggtitle("Scree Plot") +
```

```
  ylim(0, 100)
```

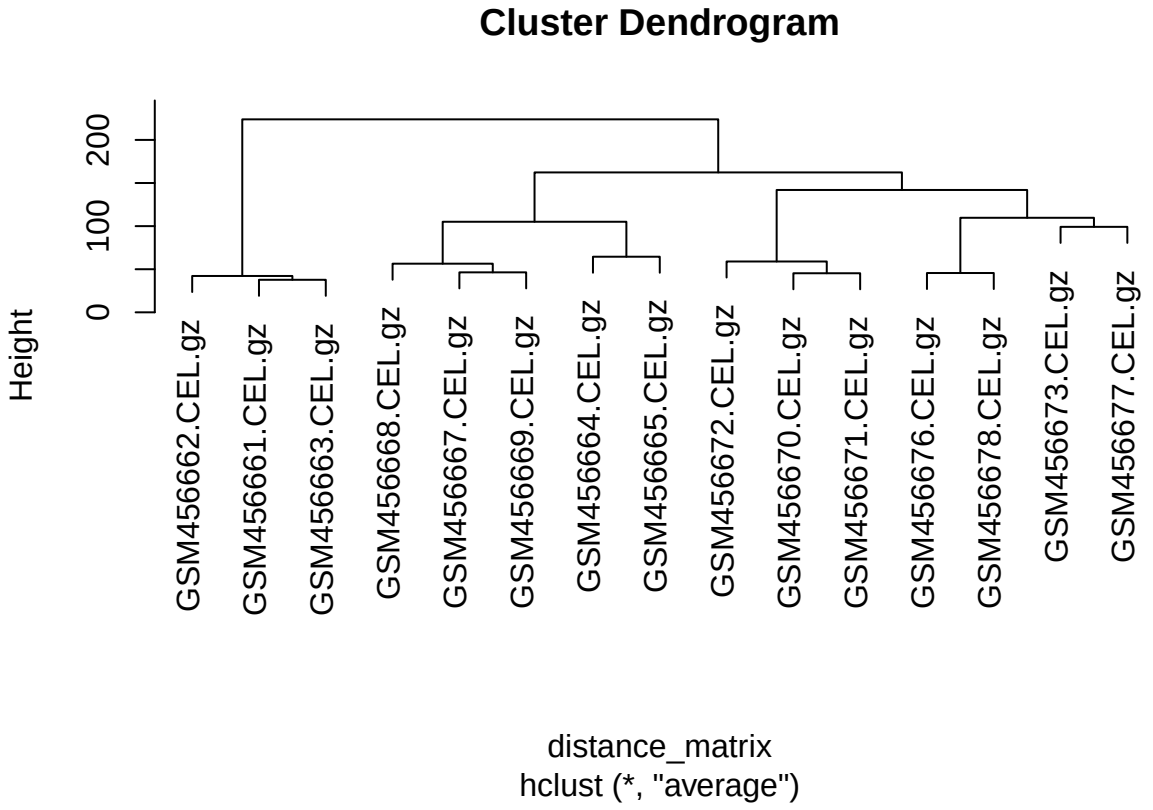



```
pca_data = data.frame(Sample = rownames(pca$x), X= pca$x[,1], Y = pca$x[,2])
ggplot(data=pca_data, aes(x=X, y=Y, label=Sample)) +
  xlab(paste("PC1 - ", pca_var_per[1], "%", sep="")) +
  ylab(paste("PC2 - ", pca_var_per[2], "%", sep="")) +
  theme_bw() +
  ggtitle("PCA mouse data") +
  geom_text(col= c("red","red","red","blue","blue","black","black","black","gray","gray","gray","green"
```



3. Hierarchical Clustering Hierarchical clustering analysis is based on an algorithm that calculates distances between the objects and forms clusters. Before we clustered we created a distance matrix using the euclidean distance. Based on the disance matrix we plotted a dendrogram in order to see which cluster differ the biggest from each other. This is based on the height of the branches. Based on the plot we can see that GSM456661 to GSM456663 differ signficantlly from the rest of the chips. The clusters with the biggest height difference is between GSM456661- GSM456663 and GSM456676-GSM456678

```
distance_matrix= dist(t(data_ohneaffx), method = "euclidean")
cluster= hclust(distance_matrix, method="average")
plot(cluster)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

AOKI, Fugaku. 2022. "Zygotic Gene Activation in Mice: Profile and Regulation." *Journal of Reproduction and Development* 68 (2): 79–84. https://doi.org/10.1262/jrd.2021-129.

Carlson, Marc. 2022. *Org.mm.eg.db: Genome Wide Annotation for Mouse.*

Ciemerych, Maria A, and Peter Sicinski. 2005. "Cell Cycle in Mouse Development." *Oncogene* 24 (17): 2877–98. https://doi.org/10.1038/sj.onc.1208608.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-seq Data Analysis." *Genome Biol.* 17 (1).

Dai, Manhong, Wang Pinglang, Andrew D. Boyd, Georgi Kostov, Brian Athey, Edward G. Jones, William E. Bunney, et al. 2005. "Evolving Gene/Transcript Definitions Significantly Alter the Interpretation of GeneChip Data." *Nucleic Acids Research* 33(20): e175.

Dolgalev, Igor. 2022. *Msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format.* https://CRAN.R-project.org/package=msigdbr.

Dunwoodie, Sally L. 2009. "The Role of Hypoxia in Development of the Mammalian Embryo." *Developmental Cell* 17 (6): 755–73. https://doi.org/10.1016/j.devcel.2009.11.008.

Gautier, Laurent, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. 2004. "affy—analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics* 20 (3): 307–15. https://doi.org/10.1093/bioinformatics/btg405.

Huber, Wolfgang, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron. 2002. "Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression." *Bioinformatics* 18 Suppl. 1: S96–104.

Kojima, Yoji, Oliver H. Tam, and Patrick P. L. Tam. 2014. "Timing of Developmental Events in the Early Mouse Embryo." *Seminars in Cell Developmental Biology* 34 (October): 65–75. https://doi.org/10.1016/j.semcdb.2014.06.010.

Krishnan, Jaya, Preeti Ahuja, Sereina Bodenmann, Don Knapik, Evelyne Perriard, Wilhelm Krek, and Jean-Claude Perriard. 2008. "Essential Role of Developmentally Activated Hypoxia-Inducible Factor a for

Cardiac Morphogenesis and Function." *Circulation Research* 103 (10): 1139–46. https://doi.org/10.1161/01.res.0000338613.89841.c1.

Mihajlović, Aleksandar I., and Alexander W. Bruce. 2017. "The First Cell-Fate Decision of Mouse Preimplantation Embryo Development: Integrating Cell Position and Polarity." *Open Biology* 7 (11): 170210. https://doi.org/10.1098/rsob.170210.

Monteleone-Cassiano, Ana Carolina, Janaina A. Dernowsek, Romario S. Mascarenhas, Amanda Freire Assis, Dimitrius Pitol, Natalia Chermont Santos Moreira, Elza Tiemi Sakamoto-Hojo, João Paulo Mardegan Issa, Eduardo A. Donadi, and Geraldo Aleixo Passos. 2022. "The Absence of the Autoimmune Regulator Gene (AIRE) Impairs the Three-Dimensional Structure of Medullary Thymic Epithelial Cell Spheroids." *BMC Molecular and Cell Biology* 23 (1). https://doi.org/10.1186/s12860-022-00414-9.

Morgan, Martin. 2022. *BiocManager: Access the Bioconductor Project Package Repository.* https://CRAN.R-project.org/package=BiocManager.

Phipson, Belinda, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. 2016. "Robust Hyperparameter Estimation Protects Against Hypervariable Genes and Improves Power to Detect Differential Expression." *Ann. Appl. Stat.* 10 (2): 946–63.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2021. *RStudio: Integrated Development Environment for r.* Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Sha, Qian-Qian, Ye-Zhang Zhu, Sen Li, Yu Jiang, Lu Chen, Xiao-Hong Sun, Li Shen, Xiang-Hong Ou, and Heng-Yu Fan. 2019. "Characterization of Zygotic Genome Activation-Dependent Maternal mRNA Clearance in Mouse." *Nucleic Acids Research* 48 (2): 879–94. https://doi.org/10.1093/nar/gkz1111.

Takaba, Hiroyuki, and Hiroshi Takayanagi. 2017. "The Mechanisms of t Cell Selection in the Thymus." *Trends in Immunology* 38 (11): 805–16. https://doi.org/10.1016/j.it.2017.07.010.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Dan, Chieh-Chun Chen, Leon M. Ptaszek, Shu Xiao, Xiaoyi Cao, Fang Fang, Huck H. Ng, Harris A. Lewin, Chad Cowan, and Sheng Zhong. 2010. "Rewirable Gene Regulatory Networks in the Preimplantation Embryonic Development of Three Mammalian Species." *Genome Research* 20 (6): 804–15. https://doi.org/10.1101/gr.100594.109.