Report_Group4

Mariam

7/13/2022

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 degraadtion 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 (blastocyst, second replicate) crosses the rest of the chips.

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 6,14,15 are of low quality.That's why we will be disregarding them for the rest of our analysis.

3. 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.

4. 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.

5. 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, desgins a matrix that groups the chips by the cell stage they belong to. After that a linear model will 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 done for the mouse and bovine data sets respectively.

6. 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 are related by a 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 identifires (Carlson 2022) For us the GSEA will help us enriching pathways that might play a role in tissue formation.

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

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

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.

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.

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/.

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.