An analysis of differentially expressed genes in Huntington's disease patients and control patients, followed by a gene clustering of the diseased patients

Luca Bracone

June 2021

1 Introduction

Huntington's disease is a hereditary and currently incurable neurodegenerative disease whose symptoms usually first appear between 30 to 50 years of age. The first symptoms are usually mood disorders and a general lack of coordination. They devolve into dementia and a total inability to perform coordinated movement.

In this report, we will perform two analyses using Affymetrix™ HG-U133A microchips with 22283 probe sets, of which 30 are control and 35 are diseased. These arrays have "probe" strands of RNA on which we apply DNA material that is relevant to our experiment. If the DNA and RNA match, they will bind strongly, otherwise they will not. Using fluorescence we can record this information into a computer. This is the data we analyze here. Firstly, we will analyze which genes are differentially expressed between patients of this disease and control patients. Secondly, among those affected by Huntington's disease, we will perform a clustering analysis to see if we can discover subgroups of genes. The reader may find more information about the software we used in the References section.

2 Quality Assessment

We fit a "probe-level" model to the data:

$$\log_2(PM_{i,j}) = c_{i,j} + p_j + \epsilon_{i,j}$$

Where $c_{i,j}$ is the $\log_2$ intensity for probe set i on chip j and p_j represents an effect per probe. For identifiability, the model is fit with the following constraint $\sum_j p_j = 0$. The parameters are estimated using robust regression. That is, we look for solutions to the following problem:

$$\underset{p_j, c_{i,j}}{\operatorname{argmin}} \sum_{i,j} \rho\left(\frac{PM_{i,j} - p_j - c_{i,j}}{\hat{\sigma}}\right)$$

Where ρ may be any loss function but here we decide it to be the Huber function, and $\hat{\sigma}$ is a robust estimate of scale such as the median absolute deviation.

To compute such a solution, a numerical method is required. We use iterative re-weighted least squares (IRLS): let $w(u) = \frac{1}{u} \cdot \frac{d\rho}{du}(u)$, the weight function corresponding to our loss function ρ. Then,

1. Obtain initial estimates $c_{i,j}^{(0)}$ and $p_j^{(0)}$.
2. Compute $u_i^{(0)} = \left(PM_{i,j} - c_{i,j}^{(0)} - p_j^{(0)}\right) / MAD\left(PM_{i,j} - c_{i,j}^{(0)} - p_j^{(0)}\right)$.
3. Obtain weights $w_i^{(0)} = w\left(u_i^{(0)}\right)$.
4. Perform weighted least squares with the $w_i^{(0)}$ to find new estimates $c_{i,j}^{(1)}$ and $p_j^{(1)}$.
5. Iterate until satisfied.

Then, chips that have a large number of low-weighted probes stand out from the other chips. They are considered to be of bad quality and discarded from the rest of the analysis. On Figure 1 we illustrate the results of this procedure.

Figure 1: Pseudo-images of the weights associated to each probe. The greener a pixel is, the more of an outlier this probe is.

Another measure of quality is the normalized unscaled standard error (NUSE): For each gene i on array j, using the standard error estimates for $c_{i,j}$ obtained in the probe-level model we compute the ratio of the standard error of $c_{i,j}$ with the median standard error of $c_{i,j}$ across

chips.

$$NUSE(c_{i,j}) = \frac{SE(c_{i,j})}{\text{med}_j(SE(c_{i,j}))}.$$

This gives a set of values that we group by array and we present it in the form of box plots on Figure 2.

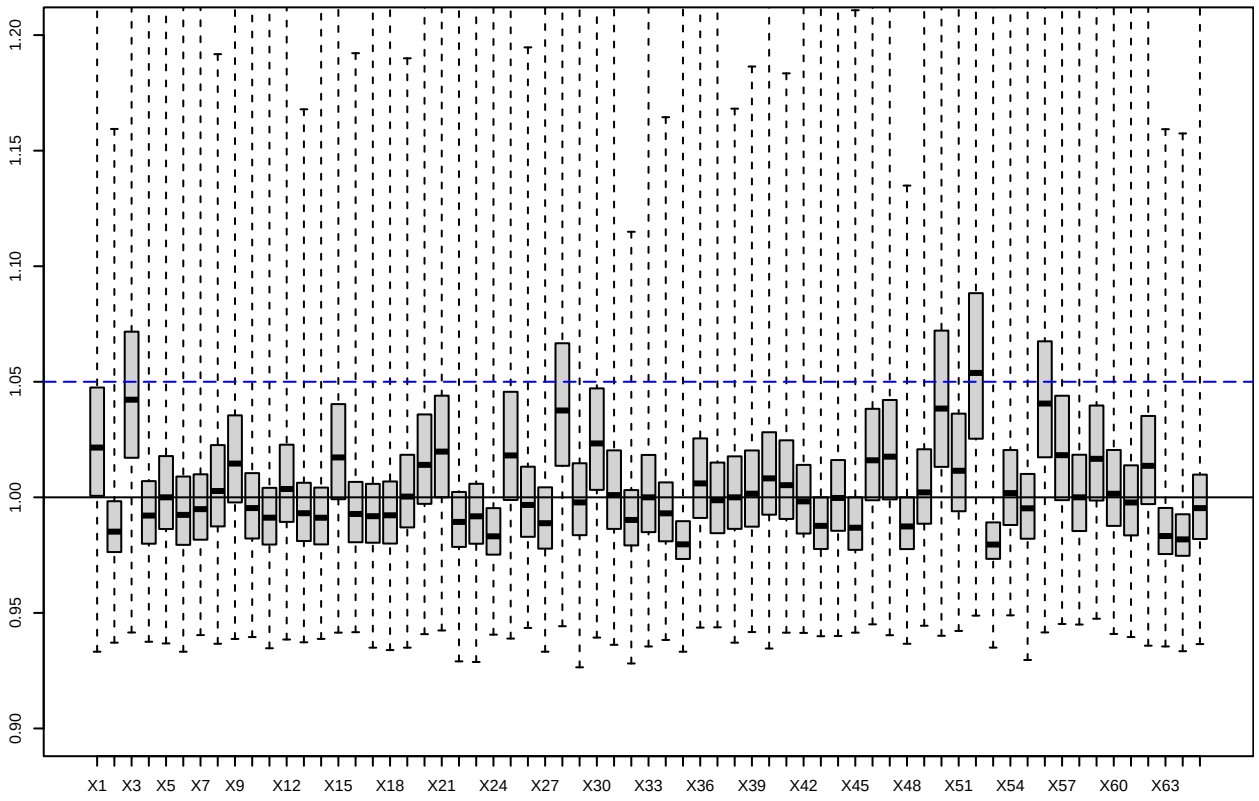

Figure 2: NUSE values of each chip

As we can see, only chip 52 exceeds the threshold of 1.05 median nuse value, so we exclude it from the rest of the study.

3 Normalization

Normalization is a three step process. First, we only consider perfect match (PM) probes. Second, we model each observed PM intensity on a "background" model: $S = X + Y$, where $X \sim Exp(\alpha)$ and $Y \sim N(\mu, \sigma^2)$. X represents the "true" intensity and Y some background noise that is the same per chip. We adjust each observed intensity to its prediction X according to this model. Finally, we move each quantile to its mean across chips. This helps to alleviate artifactual differences between chips. On Figure 3 we see the result of this process.

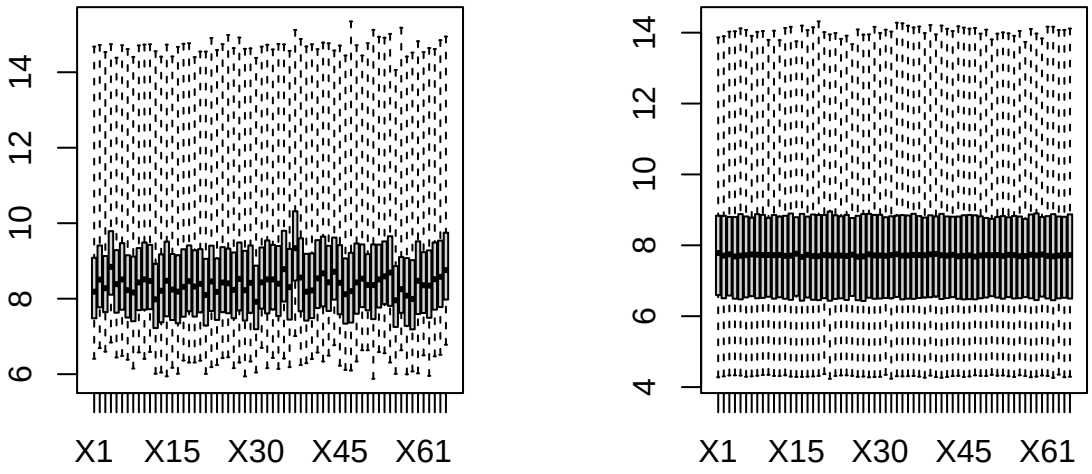

Figure 3: Boxplots of the log intensities of each chip before (left) and after (right) normalization

4 Analysis of differentially expressed genes

An MDS plot shows there is no batch effect between chips. Since we are only interested in differential expression of genes between control and Huntington disease patients, we fit a linear model that only takes this difference into account

$$y_g = \beta_0 + \beta_1 I(\text{is diseased}) + \epsilon_g.$$

Where y_g is the average log expression level of gene g across chips. Then, the design matrix is

$$A_{i,j} = \begin{cases} 1 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \text{ and } j > 30 \\ 0 & \text{otherwise.} \end{cases}$$

After fitting the model, for each gene g, we calculate the posterior variance estimate

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

where s_0^2 is the estimated average variability across all genes, s_g^2 is the estimated variability for that specific gene, d_0, and d_g are the degrees of freedom for s_0 and s_g respectively. Using this we calculate the moderated t-statistic

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}}$$

where $\hat{\beta}_g$ is the average difference of $\log_2$ gene intensity between the two groups we are comparing, for the gene g. The coefficient v_g is the entry in the covariance matrix corresponding to g. We use $\tilde{t}_g$ rather than the usual t-statistic because in micro-array experiments a lot of genes have low variance and this causes the usual t-statistic to be extremely large. The statistics $\tilde{t}_g$ have a known distribution from which we can extract p-values. We decide to

adjust the p-values using Benjamini and Hochberg's method. Let m be the total number of p-values. This method involves sorting each p-value in ascending order, giving it a rank i. Then for each p-value, calculate $BH(p) = 0.05 \cdot i/m$. Finally, we only consider the p-values that satisfy $BH(p) < 0.05$ to be significant. The results are as follows: Out of 20206 genes, we find that 1712 are down-regulated and 1331 are up regulated. On page 8, Table 1 contains a summary of the genes with the highest absolute moderated t-statistic. On Figure 4 we show a "volcano" plot: the x-axis shows the $\log_2$ fold change of each gene, positive or negative, and the y-axis shows the $\log_2$ p-value according to the moderated t-statistic.

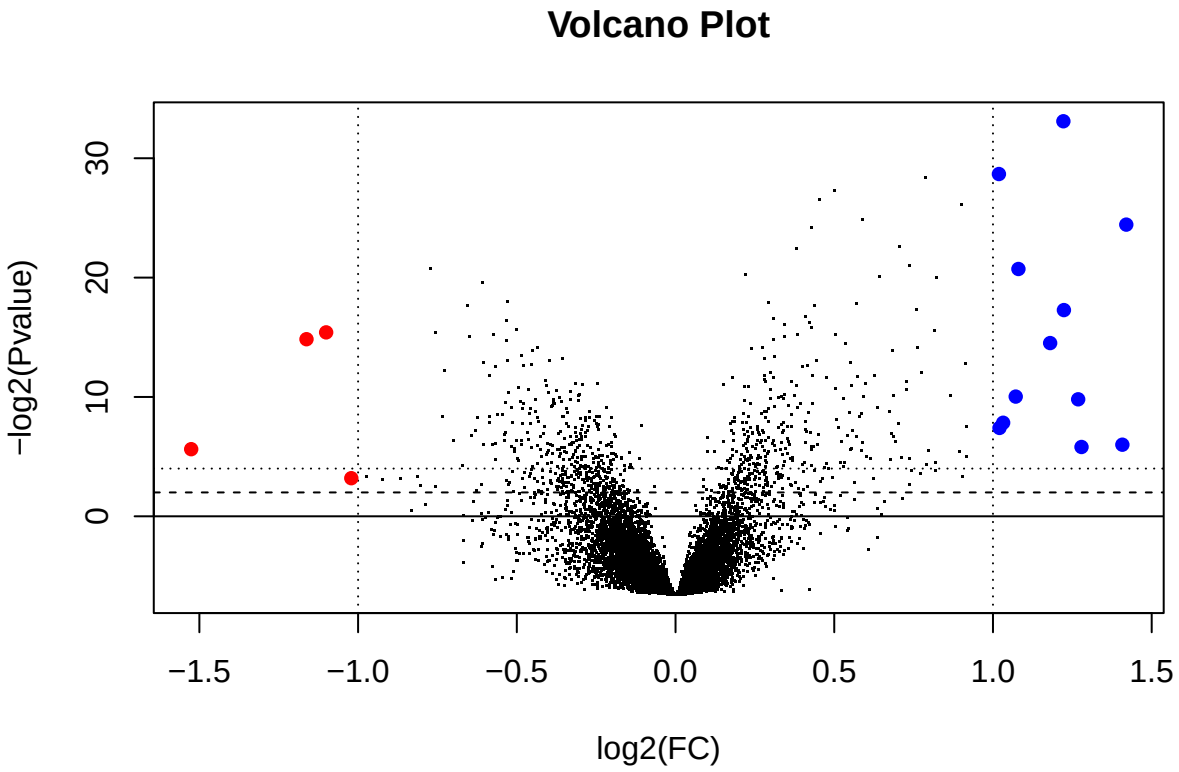

Figure 4: A plot of p-value vs. $\log_2$ fold change

5 Cluster analysis

We choose to perform a clustering analysis on the 100 genes with the highest variance using Ward's method. That is, we consider each gene to be a vector whose coordinates are given by the expression levels on each chip. Then, start with each gene being a single cluster and bit by bit we increase the "merging radius." When two clusters merge, we record at which radius it happened, the results of this procedure are shown in Figure 5. By "radius" we mean Ward's mergind cost:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} ||m_A - m_B||^2$$

where n_X and m_X are respectively the number of elements in cluster X and the barycenter of X. Looking at Figure 5 it seems reasonable to conclude that there are three main clusters being formed by taking for radius $\Delta \approx 100$. But we would need to consult with a biologist to

see if such a categorization has theoretical merit. On page 7, Figure 6 illustrates a heat map. It is essentially a color representation of the matrix that has a row for each gene, a column for each chip, and each entry is equal to the average $\log_2$ intensity of that gene on that chip. If the gene names are too small to read on paper, note that they appear in the same order as in Figure 5. From the heatmap, we find that there are group of genes which appear mutually exclusively (if one is there the other one is not).

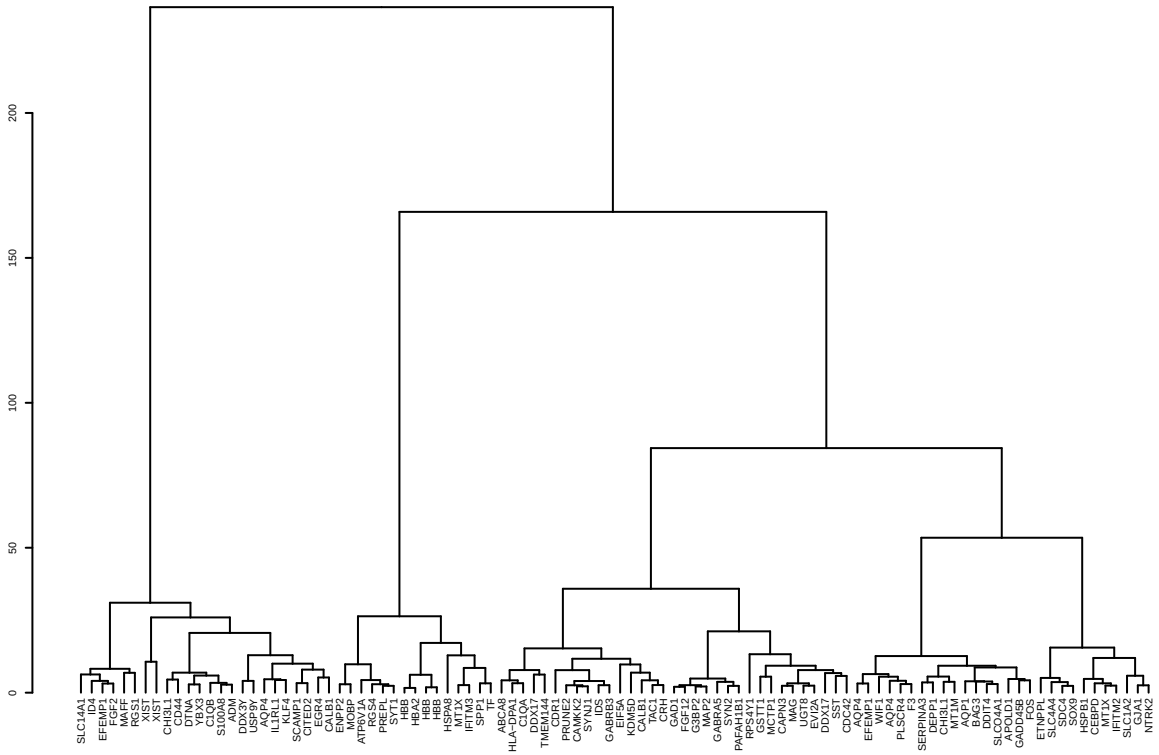

Figure 5: A dendrogram of the genes with highest variance using Ward's method

6 Conclusion

After performing a robust linear model fitting, we discarded one low-quality chip. Then, we normalized the samples so that they were as similar as possible to each other. Then, we fit a linear model to express the difference between the two groups of interest. We found that 1712 genes are down-regulated and 1331 are up-regulated. Finally we performed a clustering analysis of the 100 genes with highest variance among Huntington's disease patients. We found that it may be interesting to perform a study into a possible relationship of the genes in a classification in three clusters. Although we found a grouping while performing clustering analysis it remains unclear if it is not simply magical thinking. Consulting with a biologist and performing more experiments would be required.

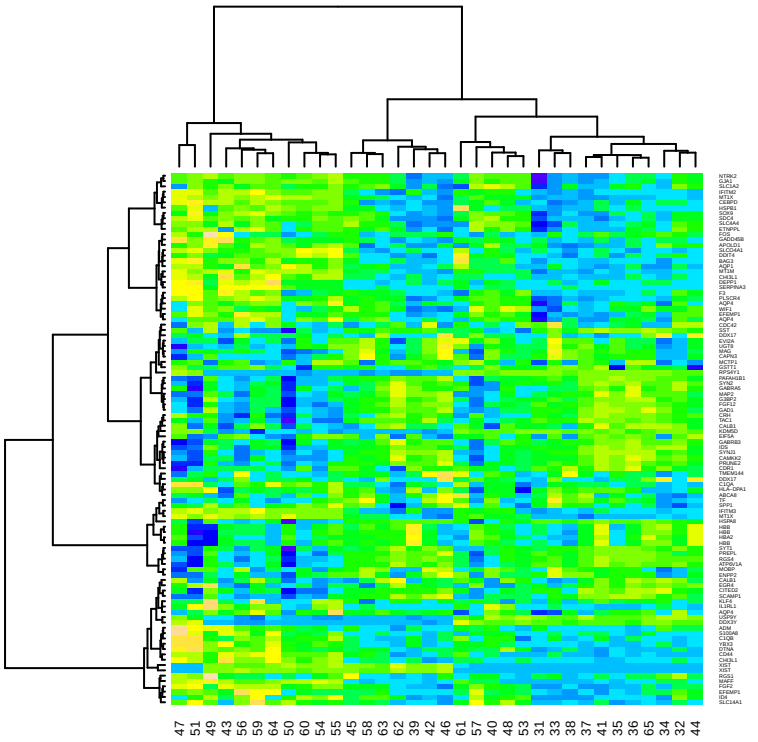

Figure 6: A heatmap of the 100 genes with higest variance, with respect to dendrograms using Ward's method, the gene names are the same as in Figure 5 so that left $\rightarrow$ right appears as down $\rightarrow$ up

Table 1: The 50 most differentially expressed genes

	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
33767_at	NEFH	1.22	10.90	12.63	2.48e-19	5.01e-15	33.10
204412_s_at	NEFH	1.02	10.74	11.39	2.69e-17	2.41e-13	28.68
214903_at	SYT2	0.79	9.81	11.29	3.58e-17	2.41e-13	28.41
202224_at	CRK	0.50	8.64	10.99	1.20e-16	6.07e-13	27.26
202817_s_at	SS18	0.45	5.24	10.83	2.44e-16	9.85e-13	26.59
220551_at	SLC17A6	0.90	8.28	10.73	3.91e-16	1.32e-12	26.14
212103_at	KPNA6	0.59	8.03	10.35	1.53e-15	4.40e-12	24.84
216086_at	SV2C	1.42	9.41	10.32	2.33e-15	5.89e-12	24.44
218903_s_at	NABP2	0.43	7.63	10.19	2.96e-15	6.65e-12	24.21
205508_at	SCN1B	0.70	10.25	9.78	1.57e-14	3.16e-11	22.62
221090_s_at	OGFOD1	0.38	8.59	9.74	1.81e-14	3.33e-11	22.48
219267_at	GLTP	0.74	6.70	9.40	8.52e-14	1.43e-10	21.00
214375_at	PPFIBP1	-0.77	7.18	-9.34	1.07e-13	1.65e-10	20.79
205336_at	PVALB	1.08	9.11	9.34	1.15e-13	1.65e-10	20.72
214714_at	ZNF394	0.22	7.80	9.18	1.82e-13	2.45e-10	20.28
207981_s_at	ESRRG	0.82	8.20	9.14	2.43e-13	2.89e-10	20.00
219628_at	ZMAT3	0.64	9.91	9.13	2.24e-13	2.83e-10	20.08
217565_at	GRIA3	-0.61	8.20	-9.00	3.82e-13	4.29e-10	19.57
210281_s_at	ZMYM2	-0.53	6.18	-8.62	1.96e-12	2.08e-09	18.00
212018_s_at	RSL1D1	0.29	7.92	8.58	2.18e-12	2.21e-09	17.90
200706_s_at	LITAF	0.57	8.90	8.55	2.40e-12	2.31e-09	17.80
213517_at	PCBP2	-0.66	8.10	-8.54	2.74e-12	2.42e-09	17.68
204082_at	PBX3	0.44	8.06	8.52	2.75e-12	2.42e-09	17.67
214650_x_at	MOG	1.22	8.97	8.46	4.18e-12	3.38e-09	17.28
213326_at	VAMP1	0.76	11.13	8.44	3.97e-12	3.34e-09	17.32
217893_s_at	AKIRIN1	0.41	8.46	8.28	7.51e-12	5.84e-09	16.71
219001_s_at	DCAF10	0.31	7.21	8.25	8.58e-12	6.42e-09	16.58
210282_at	ZMYM2	-0.53	5.12	-8.23	1.08e-11	7.81e-09	16.37
205481_at	ADORA1	0.42	9.50	8.16	1.24e-11	8.66e-09	16.23
208353_x_at	ANK1	0.34	7.80	8.11	1.50e-11	1.01e-08	16.05
213230_at	CDR2L	0.43	8.51	8.04	1.98e-11	1.29e-08	15.78
204185_x_at	PPID	0.82	8.75	8.02	2.48e-11	1.52e-08	15.57
212758_s_at	ZEB1	-0.50	8.15	-8.00	2.34e-11	1.48e-08	15.62
220940_at	ANKRD36B	-1.10	7.67	-7.99	2.93e-11	1.69e-08	15.41
213649_at	SRSF7	-0.75	9.33	-7.97	2.88e-11	1.69e-08	15.42
206339_at	CARTPT	-0.57	7.22	-7.93	3.40e-11	1.90e-08	15.26
215489_x_at	HOMER3	0.38	8.03	7.91	3.47e-11	1.90e-08	15.24
209817_at	PPP3CB	0.50	7.97	7.90	3.63e-11	1.93e-08	15.20
210794_s_at	MEG3	-0.65	9.65	-7.88	4.06e-11	2.05e-08	15.09
208756_at	EIF3I	0.34	10.01	7.87	4.05e-11	2.05e-08	15.09
212732_at	MEG3	-1.16	8.86	-7.84	5.32e-11	2.56e-08	14.84

	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
216247_at	RPS20	0.31	7.05	7.81	5.33e-11	2.56e-08	14.83
220313_at	GPR88	-0.53	7.61	-7.79	5.85e-11	2.75e-08	14.74
205989_s_at	MOG	1.18	7.77	7.76	7.46e-11	3.42e-08	14.51
207592_s_at	HCN2	0.53	7.97	7.72	7.70e-11	3.46e-08	14.48
207100_s_at	VAMP1	0.76	9.47	7.66	1.08e-10	4.53e-08	14.16
220269_at	ZBBX	-0.44	6.06	-7.65	1.07e-10	4.53e-08	14.16
219953_s_at	AKIP1	0.27	7.82	7.64	1.06e-10	4.53e-08	14.17
208350_at	CSN1S1	0.24	4.99	7.62	1.14e-10	4.69e-08	14.10
203083_at	THBS2	0.68	9.34	7.58	1.46e-10	5.81e-08	13.86

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.

Bolstad, Ben. 2021. *affyPLM: Methods for Fitting Probe-Level Models*. https://github.com/bmbolstad/affyPLM.

Bolstad, Benjamin M. 2004. "Low Level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization and Summarization." PhD thesis, University of California, Berkeley.

Bolstad, Benjamin M, Francois Collin, Julia Brettschneider, Ken Simpson, Leslie Cope, Rafael A Irizarry, and Terence P Speed. 2005. "Quality Assessment of Affymetrix GeneChip Data." In *Bioinformatics and Computational Biology Solutions Using r and Bioconductor*, edited by R. Gentleman, V. Carey, W. Huber, R Irizarry, and S Dudoit, 33–47. New York: Springer.

Brettschneider, Julia, Francois Collin, Benjamin M Bolstad, and Terence P Speed. 2007. "Quality Assessment for Short Oligonucleotide Arrays." *Technometrics* In press.

Carlson, Marc. 2021. *Hgu133a.db: Affymetrix Affymetrix HG-U133a Array Annotation Data (Chip Hgu133a)*.

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.

Gentleman, R. 2021. *Annotate: Annotation for Microarrays*.

Irizarry, Rafael A., Laurent Gautier, Benjamin Milo Bolstad, Crispin Miller with contributions from Magnus Astrand <Magnus.Astrand@astrazeneca.com>, Leslie M. Cope, Robert Gentleman, Jeff Gentry, et al. 2021. *Affy: Methods for Affymetrix Oligonucleotide Arrays*.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Smyth, Gordon, Yifang Hu, Matthew Ritchie, Jeremy Silver, James Wettenhall, Davis

McCarthy, Di Wu, et al. 2021. *Limma: Linear Models for Microarray Data*. http://bioinf.wehi.edu.au/limma.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown.

———. 2021a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://CRAN.R-project.org/package=bookdown.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.