Statistical report of project S1IPlocal: pairwise comparison(s) of conditions with DESeq2

Stefano

2024-01-26

The SARTools R package which generated this report has been developped at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr (mailto:hugo.varet@pasteur.fr)). Thanks to cite H. Varet, L. Brillet-Guéguen, J.-Y. Coppee and M.-A. Dillies, *SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data*, PLoS One, 2016, doi: http://dx.doi.org/10.1371/journal.pone.0157022 (https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0157022&type=printable) (http://dx.doi.org/10.1371/journal.pone.0157022) when using this tool for any analysis published.

1 Introduction

The analyses reported in this document are part of the S1IPlocal project. The aim is to find features that are differentially expressed between contr and S1IP. The statistical analysis process includes data normalization, graphical exploration of raw and normalized data, test for differential expression for each feature between the conditions, raw p-value adjustment and export of lists of features having a significant differential expression between the conditions. In this analysis, the batch effect will be taken into account in the statistical models.

The analysis is performed using the R software [1], Bioconductor [2] packages including DESeq2 [3,4] and the SARTools package developed at PF2 - Institut Pasteur. Normalization and differential analysis are carried out according to the DESeq2 model and package. This report comes with additional tab-delimited text files that contain lists of differentially expressed features.

For more details about the DESeq2 methodology, please refer to its related publications [3,4].

2 Description of raw data

The count data files and associated biological conditions are listed in the following table.

Table 1: Data files and associated biological conditions.

label	files	group	batch
contr1	coIPcontr1.tabular	contr	1
contr2	coIPcontr2.tabular	contr	2
contr3	coIPcontr3.tabular	contr	3
S1IP1	S1coIP1.tabular	S1IP	1
S1IP2	S1coIP2.tabular	S1IP	2
S1IP3	S1coIP3.tabular	S1IP	3

After loading the data we first have a look at the raw data table itself. The data table contains one row per annotated feature and one column per sequenced sample. Row names of this table are feature IDs (unique identifiers). The table contains raw count values representing the number of reads that map onto the features. For this project, there are 3543 features in the count data table.

Table 2: Partial view of the count data table.

	contr1	contr2	contr3	S1IP1	S1IP2	S1IP3
HG001_00001	592	56	7772	639	80	1900
HG001_00002	840	63	12638	999	140	2495

	contr1	contr2	contr3	S1IP1	S1IP2	S1IP3
HG001_00003	148	5	3004	191	6	537
HG001_00004	894	57	10937	1247	97	1637
HG001_00005	1937	239	35099	3173	404	6257
HG001_00006	3384	259	36528	4952	335	5337

Looking at the summary of the count table provides a basic description of these raw counts (min and max values, median, etc).

Table 3: Summary of the raw counts.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
contr1	0	5	53	16029	278	7247314
contr2	0	1	6	11542	26	6848356
contr3	0	91	285	903	643	273456
S1IP1	0	6	65	10488	340	5182337
S1IP2	0	1	10	13492	47	7973138
S1IP3	0	33	124	1583	392	1093373

Figure 1 shows the total number of mapped and counted reads for each sample. We expect total read counts to be similar within conditions, they may be different across conditions. Total counts sometimes vary widely between replicates. This may happen for several reasons, including:

- different rRNA contamination levels between samples (even between biological replicates);
- slight differences between library concentrations, since they may be difficult to measure with high precision.

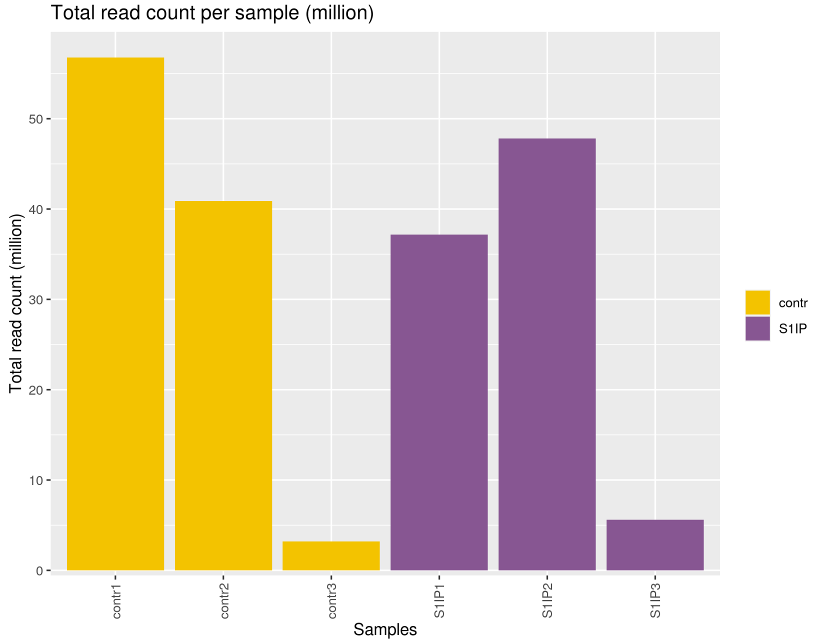


Figure 1: Number of mapped reads per sample. Colors refer to the biological condition of the sample.

Figure 2 shows the percentage of features with no read count in each sample. We expect this percentage to be similar within conditions. Features with null read counts in the 6 samples are left in the data but are not taken into account for the analysis with DESeq2. Here, 12 features (0.34%) are in this situation (dashed line). Results for those features (fold-change and p-values) are set to NA in the results files.

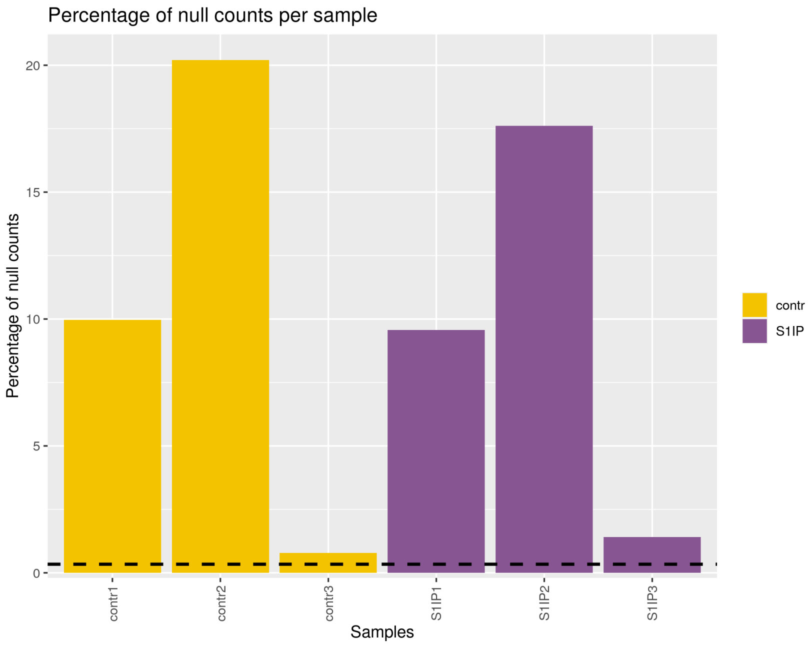


Figure 2: Percentage of features with null read counts in each sample.

Figure 3 shows the distribution of read counts for each sample (on a log scale to improve readability). Again we expect replicates to have similar distributions. In addition, this figure shows if read counts are preferably low, medium or high. This depends on the organisms as well as the biological conditions under consideration.

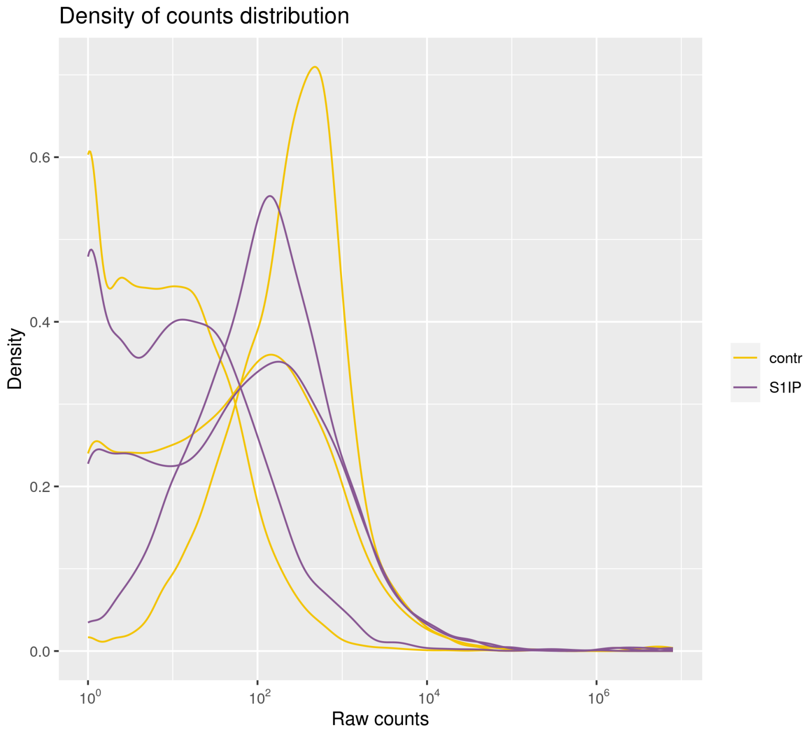


Figure 3: Density distribution of read counts.

It may happen that one or a few features capture a high proportion of reads (up to 20% or more). This phenomenon should not influence the normalization process. The DESeq2 normalization has proved to be robust to this situation [Dillies, 2012]. Anyway, we expect these high count features to be the same across replicates. They are not necessarily the same across conditions. Figure 4 and table 4 illustrate the possible presence of such high count features in the data set.

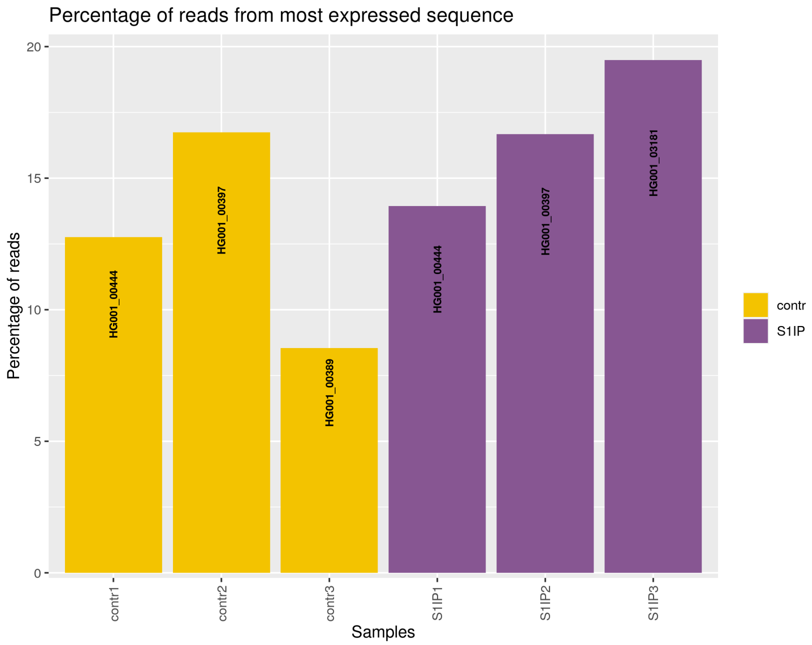


Figure 4: Percentage of reads associated with the sequence having the highest count (provided in each box on the graph) for each sample.

Table 4: Percentage of reads associated with the sequences having the highest counts.

	HG001_00444	**HG001_00396**	**HG001_02087**	**HG001_00397**	**HG001_00445**	**HG001_02200**	**HG001_00389**	**HG001_01304**	**HG001_03120**	**HG001_03181**
contr1	12.76	12.58	10.61	9.05	8.93	7.67	0.84	0.12	0.11	0.10
contr2	4.62	4.57	4.12	16.75	16.52	15.49	0.49	0.12	0.01	0.00
contr3	0.06	0.06	0.03	0.32	0.35	0.20	8.55	5.35	4.90	1.33
S1IP1	13.95	13.73	11.49	5.95	5.88	4.99	5.27	0.18	0.16	0.54
S1IP2	4.10	4.05	3.66	16.68	16.39	15.36	0.35	0.10	0.04	0.08
S1IP3	0.30	0.28	0.18	0.86	0.94	0.57	1.55	0.75	1.40	19.49

We may wish to assess the similarity between samples across conditions. A pairwise scatter plot is produced (figure 5) to show how replicates and samples from different biological conditions are similar or different (using a log scale). Moreover, as the Pearson correlation has been shown not to be relevant to measure the similarity between replicates, the SERE statistic has been proposed as a similarity index between RNA-Seq samples [5]. It measures whether the variability between samples is random Poisson variability or higher. Pairwise SERE values are printed in the lower triangle of the pairwise scatter plot. The value of the SERE statistic is:

- 0 when samples are identical (no variability at all: this may happen in the case of a sample duplication);
- 1 for technical replicates (technical variability follows a Poisson distribution);
- greater than 1 for biological replicates and samples from different biological conditions (biological variability is higher than technical one, data are over-dispersed with respect to Poisson). The higher the SERE value, the lower the similarity. It is expected to be lower between biological replicates than between samples of different biological conditions. Hence, the SERE statistic can be used to detect inversions between samples.

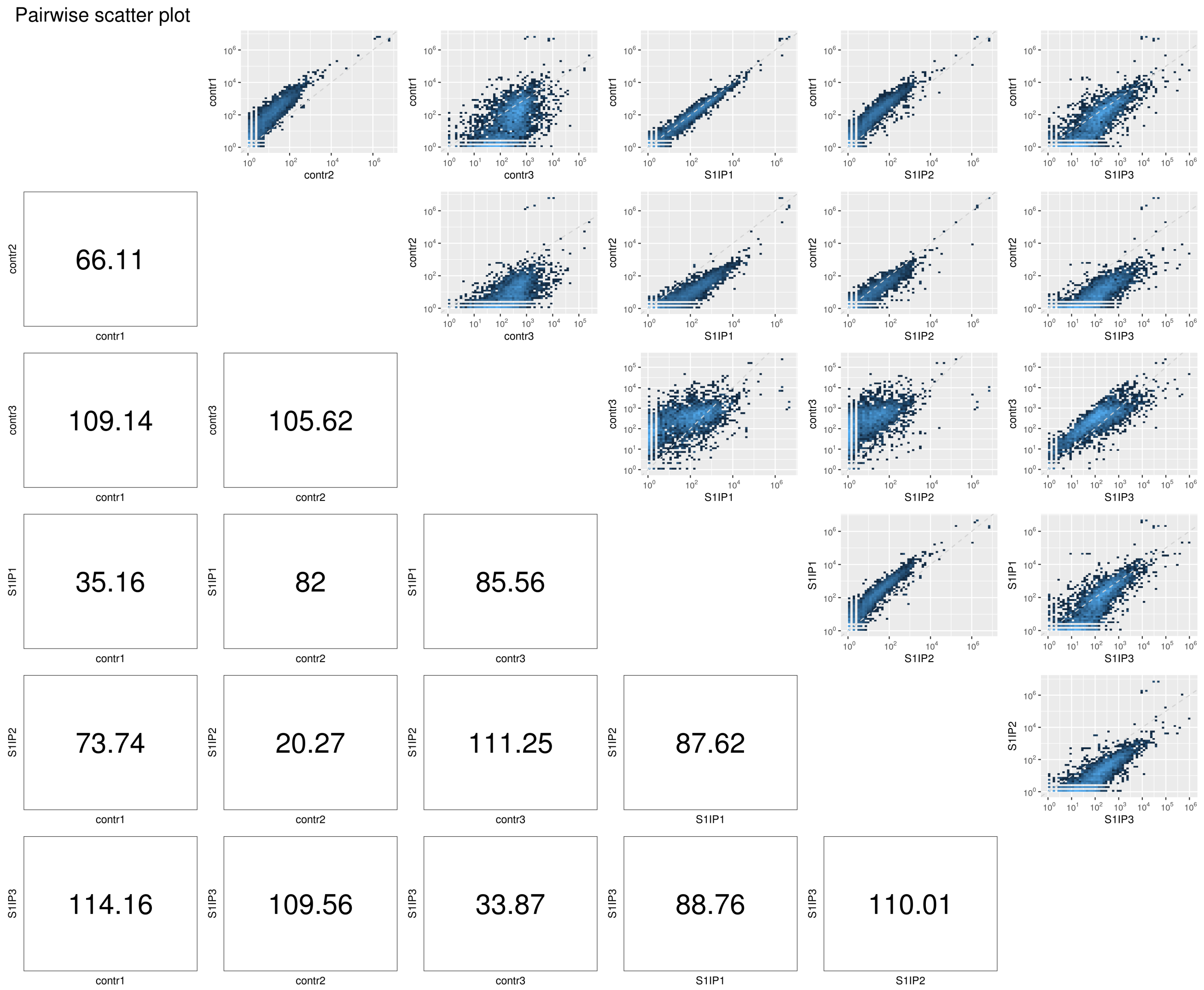


Figure 5: Pairwise comparison of samples (not produced when more than 12 samples).

3 Variability within the experiment: data exploration

The main variability within the experiment is expected to come from biological differences between the samples. This can be checked in two ways. The first one is to perform a hierarchical clustering of the whole sample set. This is performed after a transformation of the count data which can be either a Variance Stabilizing Transformation (VST) or a regularized log transformation (rlog) [3,4].

A VST is a transformation of the data that makes them homoscedastic, meaning that the variance is then independent of the mean. It is performed in two steps: (i) a mean-variance relationship is estimated from the data with the same function that is used to normalize count data and (ii) from this relationship, a transformation of the data is performed in order to get a dataset in which the variance is independent of the mean. The homoscedasticity is a prerequisite for the use of some data analysis methods, such as hierarchical clustering or Principal Component Analysis (PCA). The regularized log transformation is based on a GLM (Generalized Linear Model) on the counts and has the same goal as a VST but is more robust in the case when the size factors vary widely.

Figure 6 shows the dendrogram obtained from VST-transformed data. An euclidean distance is computed between samples, and the dendrogram is built upon the Ward criterion. We expect this dendrogram to group replicates and separate biological conditions.

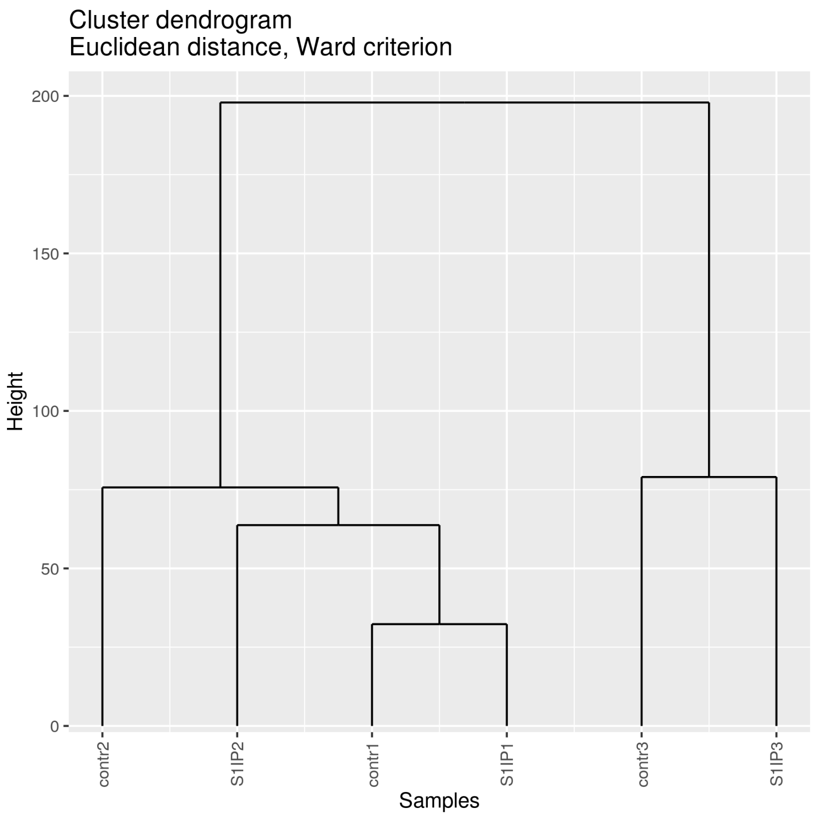


Figure 6: Sample clustering based on normalized data.

Another way of visualizing the experiment variability is to look at the first principal components of the PCA, as shown on the figure 7. On this figure, the first principal component (PC1) is expected to separate samples from the different biological conditions, meaning that the biological variability is the main source of variance in the data.

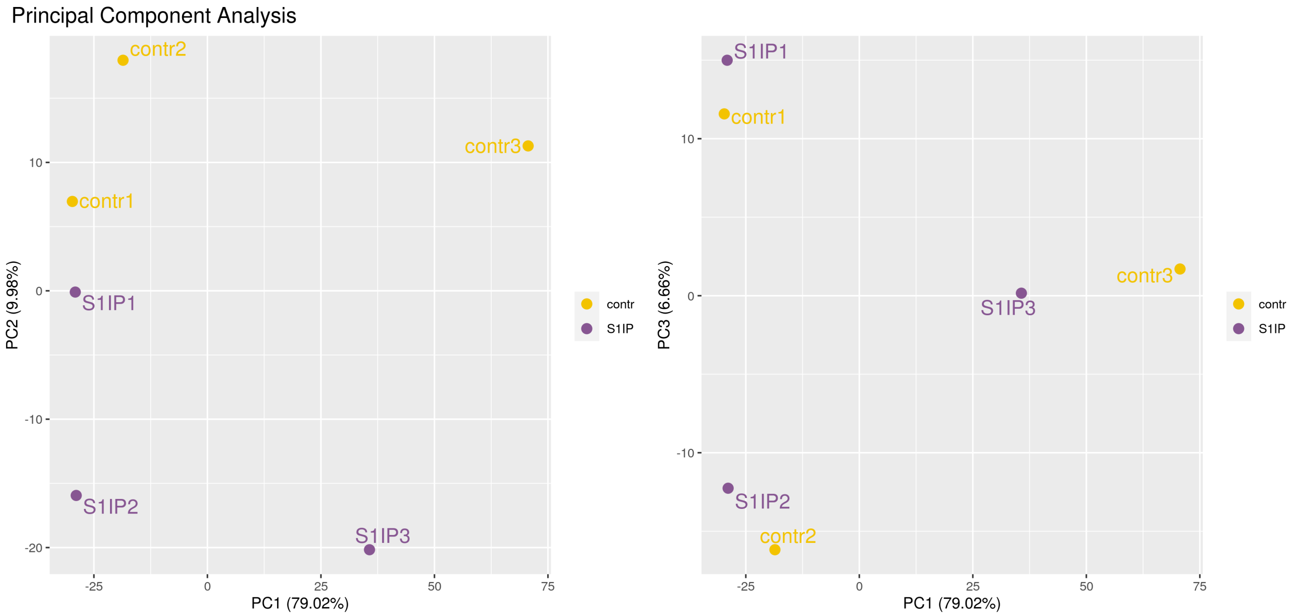


Figure 7: First two components of a Principal Component Analysis, with percentages of variance associated with each axis.

For the statistical analysis, we need to take into account the effect of the batch parameter. Statistical models and tests will thus be adjusted on it.

4 Normalization

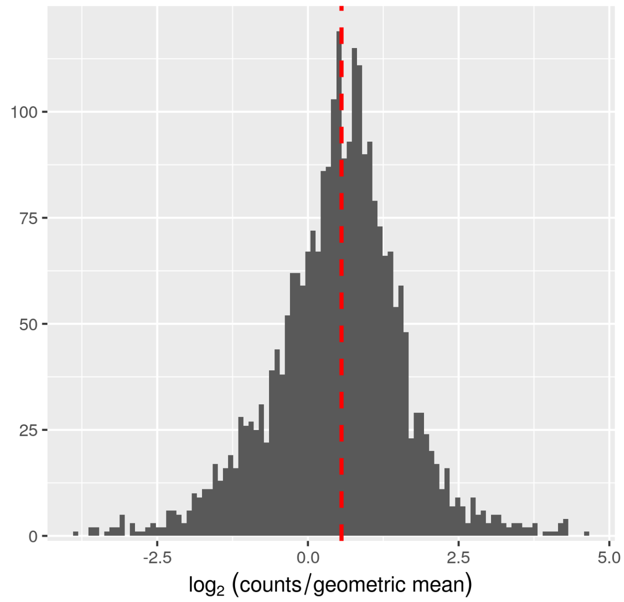
Normalization aims at correcting systematic technical biases in the data, in order to make read counts comparable across samples. The normalization proposed by DESeq2 relies on the hypothesis that most features are not differentially expressed. It computes a scaling factor for each sample. Normalized read counts are obtained by dividing raw read counts by the scaling factor associated with the sample they belong to. Scaling factors around 1 mean (almost) no normalization is performed. Scaling factors lower than 1 will produce normalized counts higher than raw ones, and the other way around. Two options are available to compute scaling factors: locfunc="median" (default) or locfunc="shorth." Here, the normalization was performed with locfunc="median."

Table 5: Normalization factors.

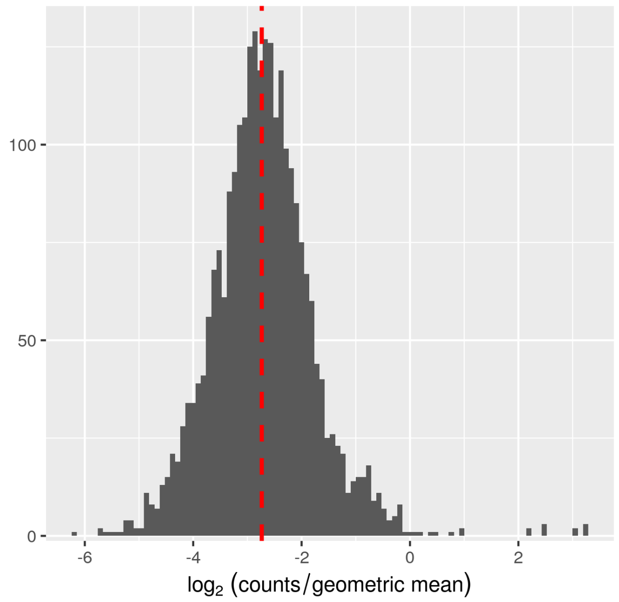
	contr1	contr2	contr3	S1IP1	S1IP2	S1IP3
Size factor	1.47	0.15	4	1.85	0.26	2.53

The histograms (figure 8) can help to validate the choice of the normalization parameter (“median” or “shorth”). Under the hypothesis that most features are not differentially expressed, each size factor represented by a red line is expected to be close to the mode of the distribution of the counts divided by their geometric means across samples.

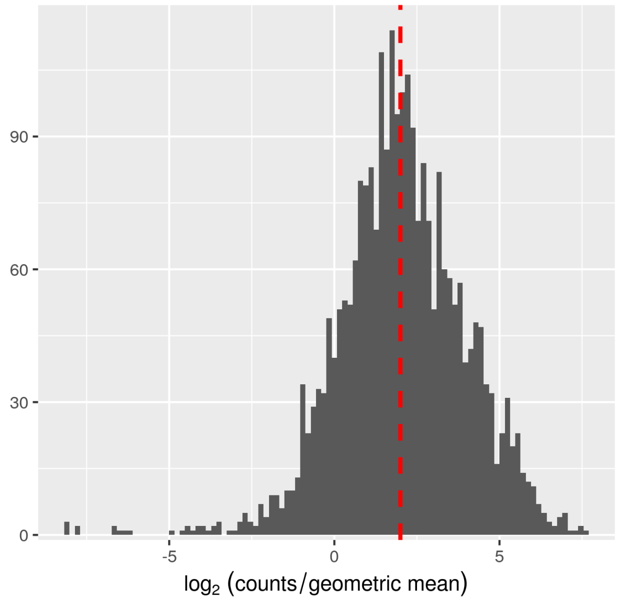
Size factor diagnostic - contr1



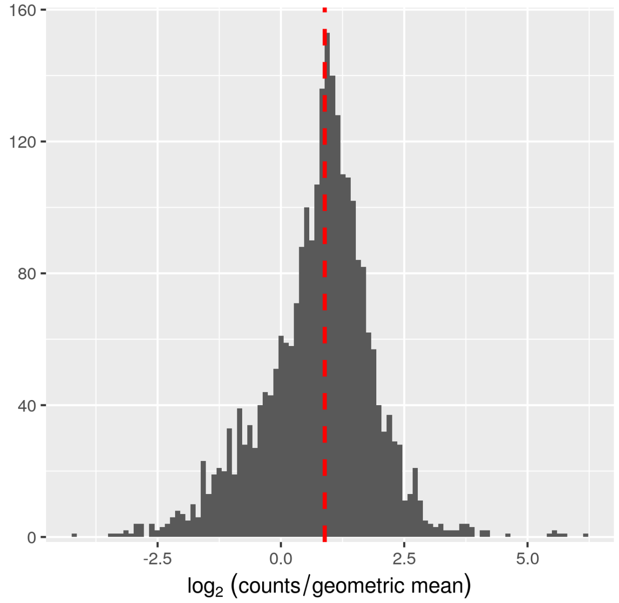
Size factor diagnostic - contr2



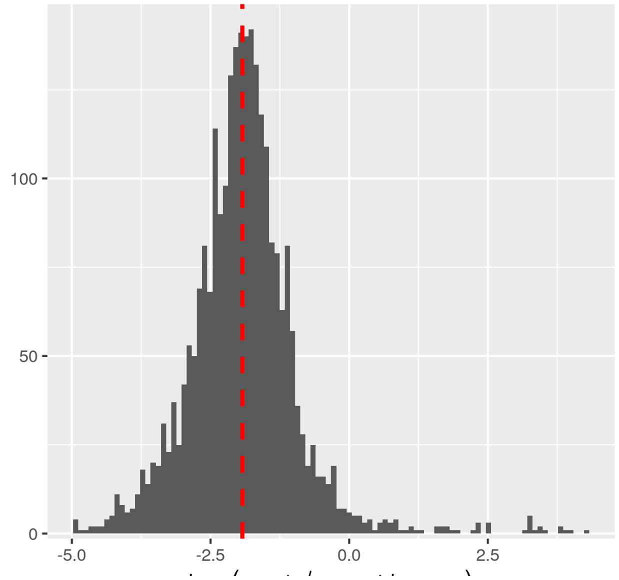
Size factor diagnostic - contr3



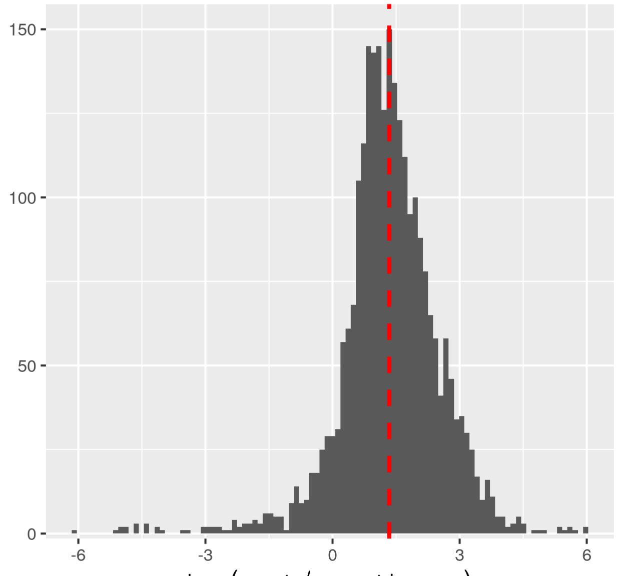
Size factor diagnostic - S1IP1



Size factor diagnostic - S1IP2



Size factor diagnostic - S1IP3



$\log_2$ (counts/geometric mean)
$\log_2$ (counts/geometric mean)

Figure 8: Diagnostic of the estimation of the size factors.

The figure 9 shows that the scaling factors of DESeq2 and the total count normalization factors may not perform similarly.

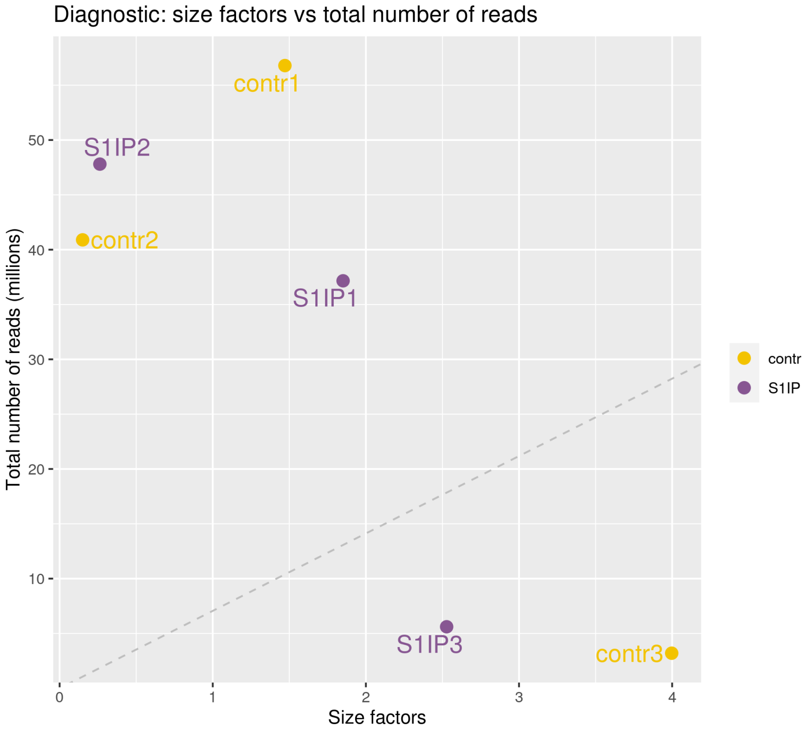


Figure 9: Plot of the estimated size factors and the total number of reads per sample.

Boxplots are often used as a qualitative measure of the quality of the normalization process, as they show how distributions are globally affected during this process. We expect normalization to stabilize distributions across samples. Figure 10 shows boxplots of raw (left) and normalized (right) data respectively.

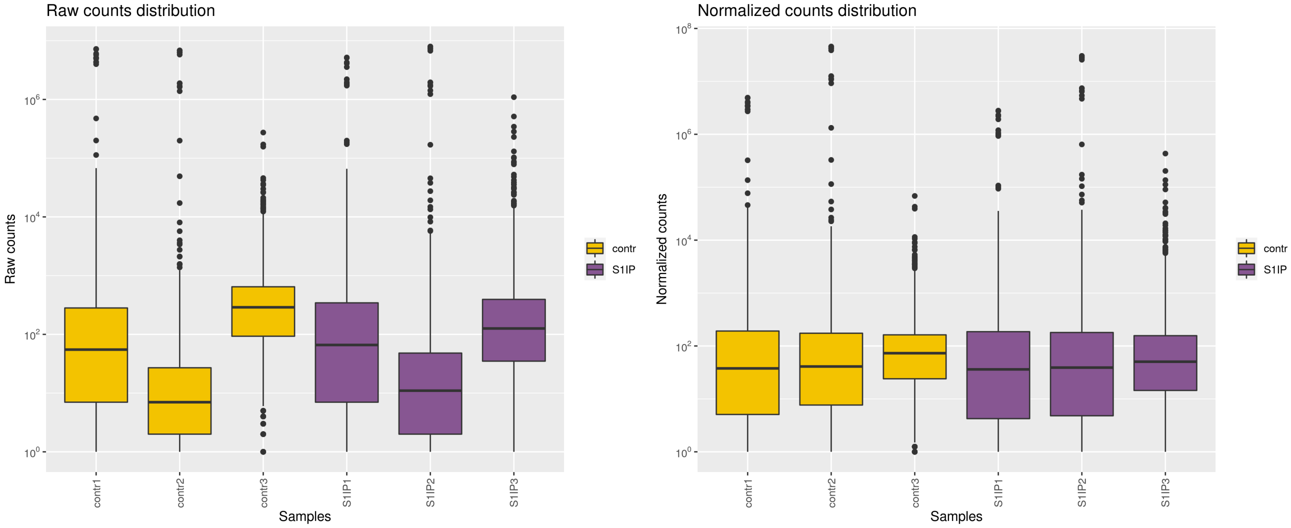


Figure 10: Boxplots of raw (left) and normalized (right) read counts.

5 Differential analysis

5.1 Modelisation

DESeq2 aims at fitting one linear model per feature. For this project, the design used is counts ~ batch + group and the goal is to estimate the models' coefficients which can be interpreted as $\log_2(\mathtt{FC})$. These coefficients will then be tested to get p-values and adjusted p-values.

5.2 Outlier detection

Model outliers are features for which at least one sample seems unrelated to the experimental or study design. For every feature and for every sample, the Cook's distance [6] reflects how the sample matches the model. A large value of the Cook's distance indicates an outlier count and p-values are not computed for the corresponding feature.

5.3 Dispersions estimation

The DESeq2 model assumes that the count data follow a negative binomial distribution which is a robust alternative to the Poisson law when data are over-dispersed (the variance is higher than the mean). The first step of the statistical procedure is to estimate the dispersion of the data. Its purpose is to determine the shape of the mean-variance relationship. The default is to apply a GLM (Generalized Linear Model) based method (fitType="parametric"), which can handle complex designs but may not converge in some cases. The alternative is to use fitType="local" as described in the original paper [3] or fitType="mean." The parameter used for this project is fitType="local." Then, DESeq2 imposes a Cox Reid-adjusted profile likelihood maximization [7 and McCarthy, 2012] and uses the maximum *a posteriori* (MAP) of the dispersion [Wu, 2013].

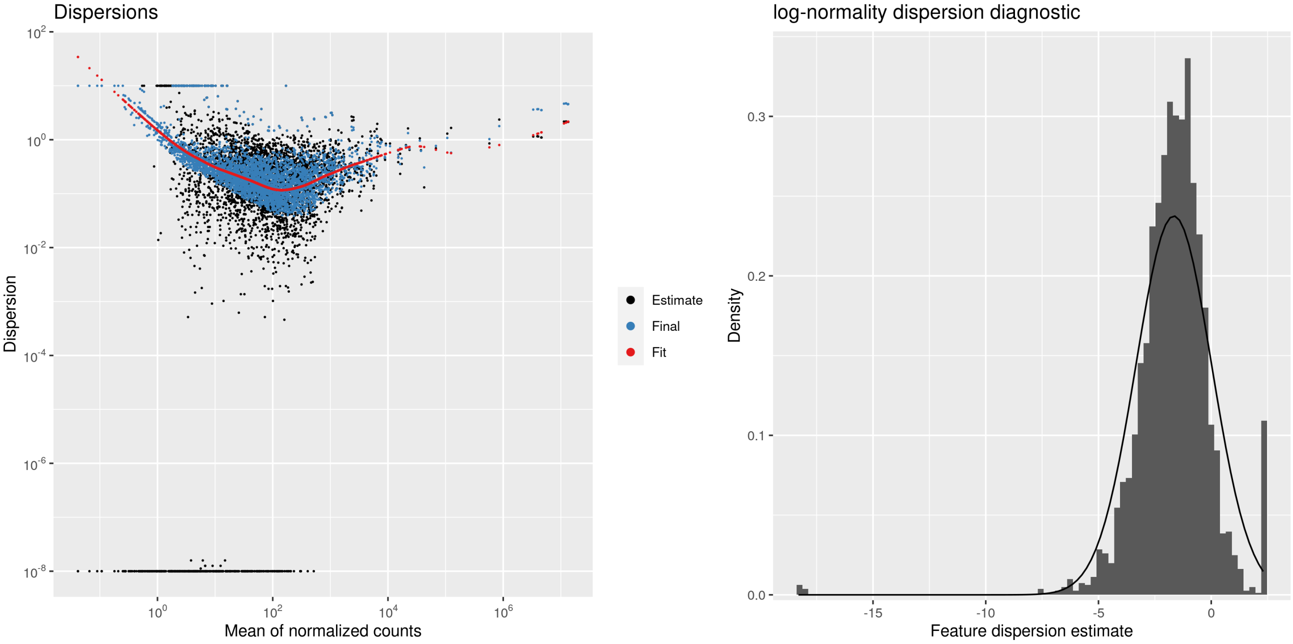


Figure 11: Dispersion estimates (left) and diagnostic of log-normality (right).

The left panel on figure 11 shows the result of the dispersion estimation step. The x- and y-axes represent the mean count value and the estimated dispersion respectively. Black dots represent empirical dispersion estimates for each feature (from the observed counts). The red dots show the mean-variance relationship function (fitted dispersion value) as estimated by the model. The blue dots are the final estimates from the maximum *a posteriori* and are used to perform the statistical test. Blue circles (if any) point out dispersion outliers. These are features with a very high empirical variance (computed from observed counts). These high dispersion values fall far from the model estimation. For these features, the statistical test is based on the empirical variance in order to be more conservative than with the MAP dispersion. These features will have low chance to be declared significant. The figure on the right panel allows to check the hypothesis of log-normality of the dispersions.

5.4 Statistical test for differential expression

Once the dispersion estimation and the model fitting have been done, DESeq2 can perform the statistical testing. Figure 12 shows the distributions of raw p-values computed by the statistical test for the comparison(s) done. This distribution is expected to be a mixture of a uniform distribution on $[0, 1]$ and a peak around 0 corresponding to the differentially expressed features.

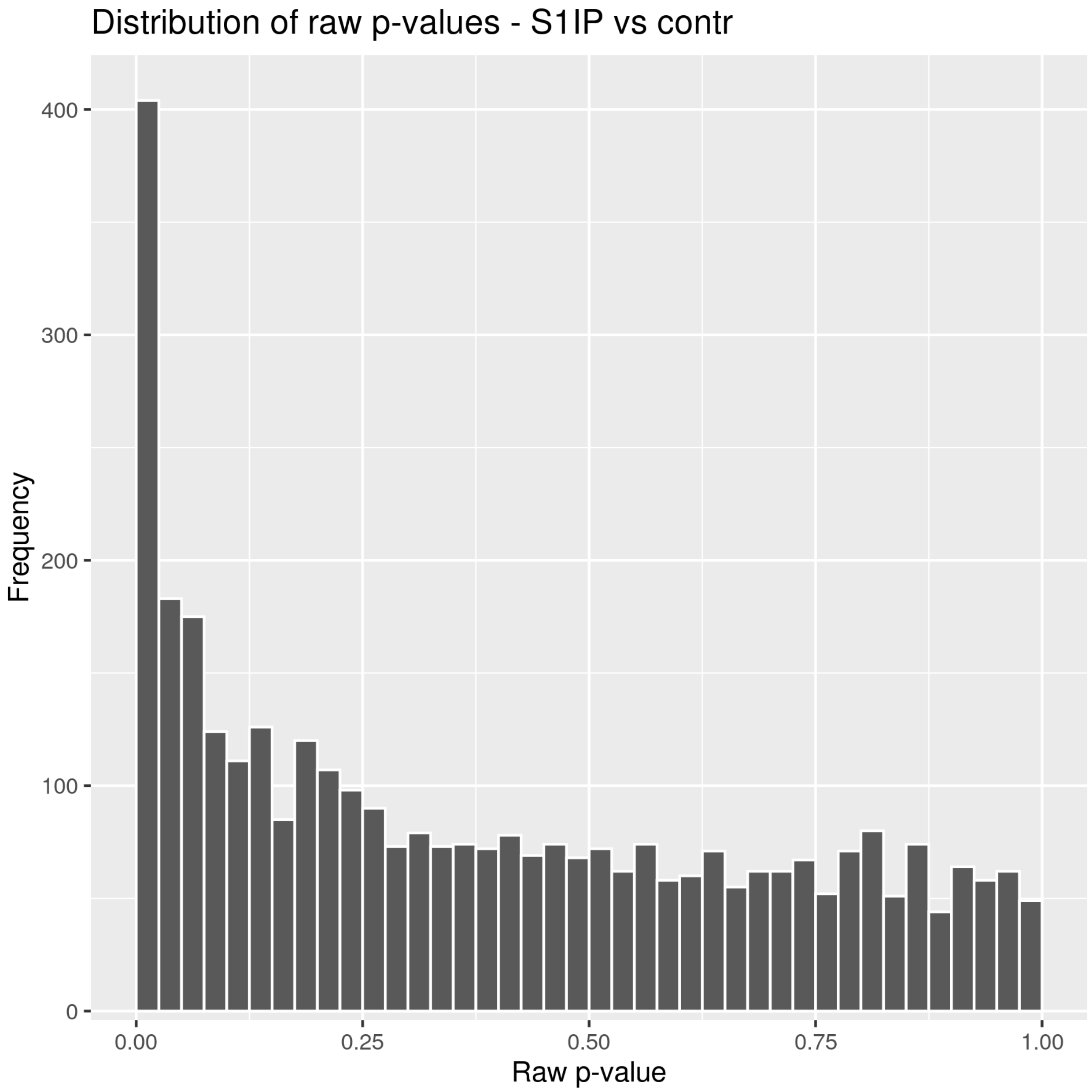


Figure 12: Distribution(s) of raw p-values.

5.5 Independent filtering

DESeq2 can perform an independent filtering to increase the detection power of differentially expressed features at the same experiment-wide type I error. Since features with very low counts are not likely to see significant differences typically due to high dispersion, it defines a threshold on the mean of the normalized counts irrespective of the biological condition. This procedure is independent because the information about the variables in the design formula is not used [4].

Table 6 reports the thresholds used for each comparison and the number of features discarded by the independent filtering. Adjusted p-values of discarded features are then set to NA.

Table 6: Number of features discarded by the independent filtering for each comparison.

Test vs Ref	BaseMean Threshold	# discarded
S1IP vs contr	19.82	1039

5.6 Final results

A p-value adjustment is performed to take into account multiple testing and control the false positive rate to a chosen level α. For this analysis, a BH p-value adjustment was performed [8 and BY2001] and the level of controlled false positive rate was set to 0.05.

Table 7: Number of up-, down- and total number of differentially expressed features for each comparison.

Test vs Ref	# down	# up	# total
S1IP vs contr	58	72	130

Figure 13 represents the MA-plot of the data for the comparisons done, where differentially expressed features are highlighted in red. A MA-plot represents the log ratio of differential expression as a function of the mean intensity for each feature. Triangles correspond to features having a too low/high $\log_2(\mathrm{FC})$ to be displayed on the plot.

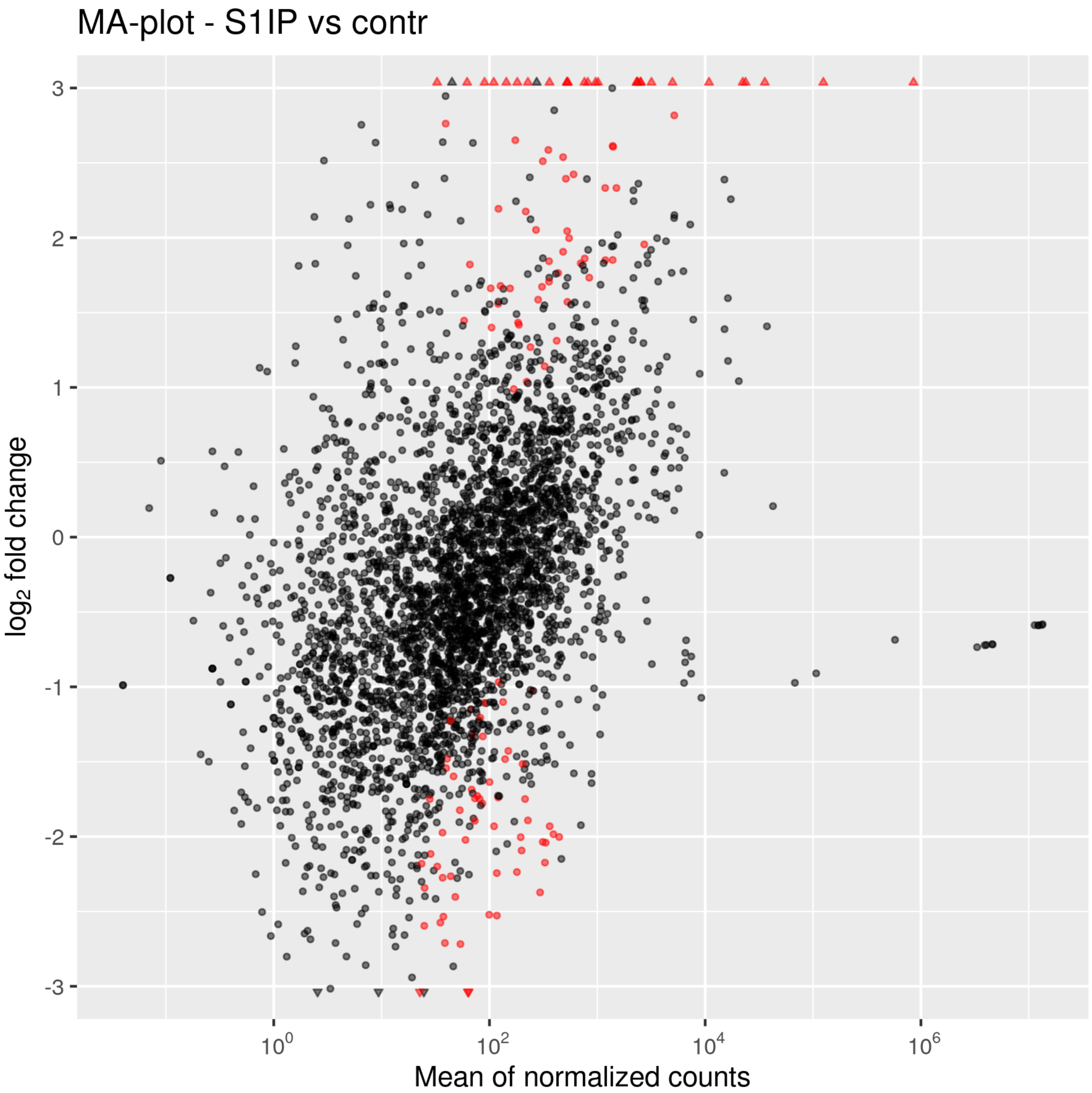


Figure 13: MA-plot(s) of each comparison. Red dots represent significantly differentially expressed features.

Figure 14 shows the volcano plots for the comparisons performed and differentially expressed features are still highlighted in red. A volcano plot represents the log of the adjusted P value as a function of the log ratio of differential expression.

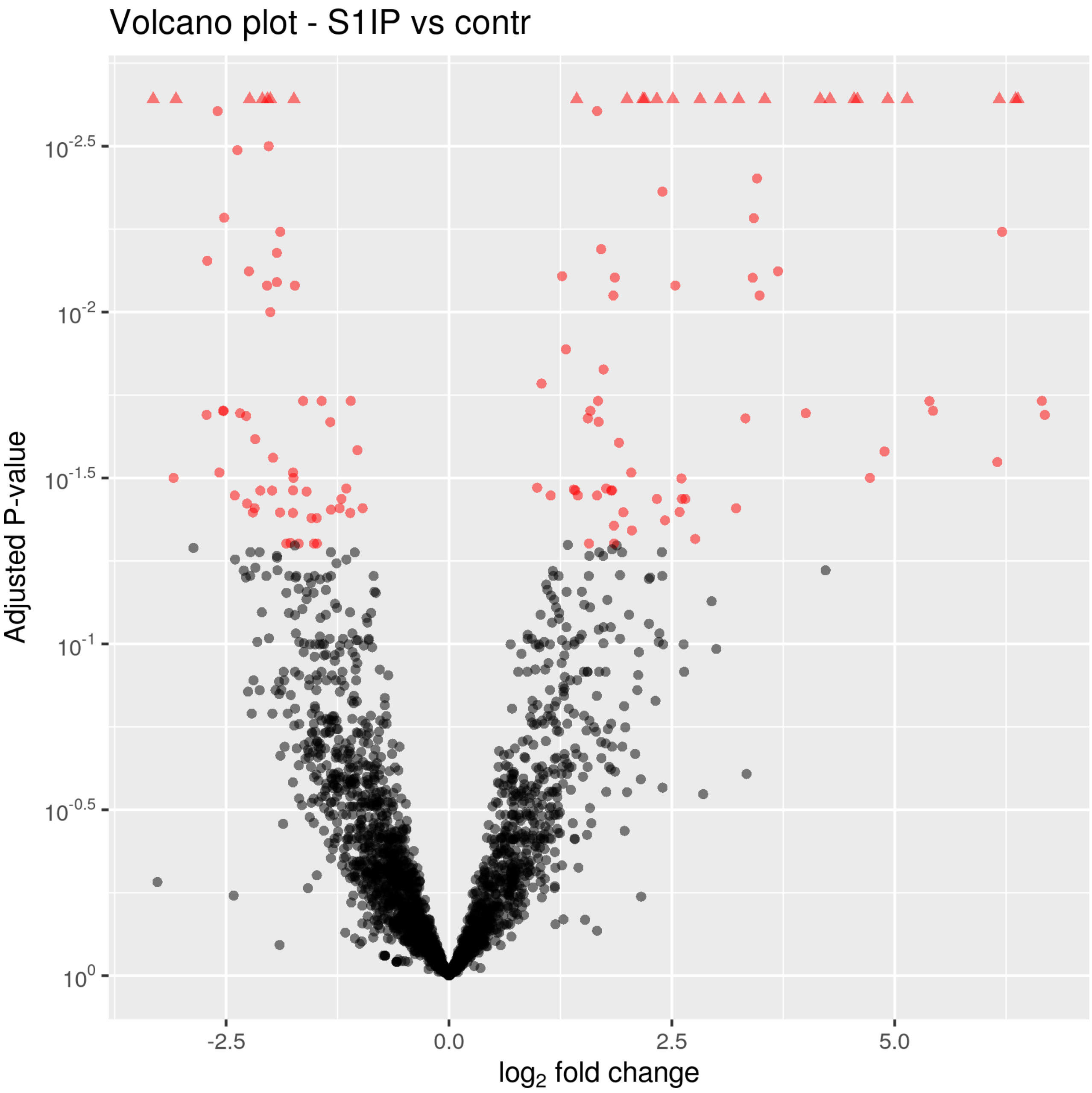


Figure 14: Volcano plot(s) of each comparison. Red dots represent significantly differentially expressed features.

Full results as well as lists of differentially expressed features are provided in the following text files which can be easily read in a spreadsheet. For each comparison:

- TestVsRef.complete.txt contains results for all the features;
- TestVsRef.up.txt contains results for significantly up-regulated features. Features are ordered from the most significant adjusted p-value to the less significant one;
- TestVsRef.down.txt contains results for significantly down-regulated features. Features are ordered from the most significant adjusted p-value to the less significant one.

These files contain the following columns:

- Id: unique feature identifier;
- sampleName: raw counts per sample;
- norm.sampleName: rounded normalized counts per sample;
- baseMean: base mean over all samples;
- contr and S1IP: means (rounded) of normalized counts of the biological conditions;
- FoldChange: fold change of expression, calculated as $2^{\log_2(\mathrm{FC})}$;
- log2FoldChange: $\log_2(\mathrm{FC})$ as estimated by the GLM model. It reflects the differential expression between Test and Ref and can be interpreted as $\log_2(\frac{\mathrm{Test}}{\mathrm{Ref}})$. If this value is:
  - around 0: the feature expression is similar in both conditions;
  - positive: the feature is up-regulated ($\mathrm{Test} > \mathrm{Ref}$);

  - negative: the feature is down-regulated ($\text{Test} < \text{Ref}$);
- stat: Wald statistic for the coefficient tested;
- pvalue: raw p-value from the statistical test;
- padj: adjusted p-value on which the cut-off α is applied;
- dispGeneEst: dispersion parameter estimated from feature counts (i.e. black dots on figure 11);
- dispFit: dispersion parameter estimated from the model (i.e. red dots on figure 11);
- dispMAP: dispersion parameter estimated from the Maximum *A Posteriori* model;
- dispersion: final dispersion parameter used to perform the test (i.e. blue dots and circles on figure 11);
- betaConv: convergence of the coefficients of the model (TRUE or FALSE);
- maxCooks: maximum Cook's distance of the feature.

6 R session information and parameters

The versions of the R software and Bioconductor packages used for this analysis are listed below. It is important to save them if one wants to re-perform the analysis in the same conditions.

- R version 4.0.2 (2020-06-22), x86_64-conda_cos6-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=en_US.UTF-8, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Running under: CentOS Linux 7 (Core)
- Matrix products: default
- BLAS/LAPACK: /shared/ifbstor1/galaxy/mutable-data/dependencies/_conda/envs/__r-sartools@1.7.3/lib/libopenblasp-r0.3.10.so (mailto:shared/ifbstor1/galaxy/mutable-data/dependencies/_conda/envs/__r-sartools@1.7.3/lib/libopenblasp-r0.3.10.so)
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.50.0, BiocGenerics 0.36.0, DESeq2 1.30.0, edgeR 3.32.0, GenomeInfoDb 1.26.0, GenomicRanges 1.42.0, ggplot2 3.3.3, IRanges 2.24.0, kableExtra 1.3.1, limma 3.46.0, MatrixGenerics 1.2.0, matrixStats 0.57.0, optparse 1.6.6, S4Vectors 0.28.0, SARTools 1.7.3, SummarizedExperiment 1.20.0
- Loaded via a namespace (and not attached): annotate 1.68.0, AnnotationDbi 1.52.0, BiocParallel 1.24.0, bit 4.0.4, bit64 4.0.5, bitops 1.0-6, blob 1.2.1, colorspace 2.0-0, compiler 4.0.2, crayon 1.3.4, DBI 1.1.0, DelayedArray 0.16.0, digest 0.6.27, dplyr 1.0.2, ellipsis 0.3.1, evaluate 0.14, farver 2.0.3, genefilter 1.72.0, geneplotter 1.68.0, generics 0.1.0, GenomeInfoDbData 1.2.4, getopt 1.20.3, GGally 2.1.0, ggdendro 0.1.22, ggrepel 0.9.0, glue 1.4.2, grid 4.0.2, gridExtra 2.3, gtable 0.3.0, highr 0.8, htmltools 0.5.0, httr 1.4.2, knitr 1.30, labeling 0.4.2, lattice 0.20-41, lifecycle 0.2.0, locfit 1.5-9.4, magrittr 2.0.1, MASS 7.3-53, Matrix 1.3-2, memoise 1.1.0, munsell 0.5.0, pillar 1.4.7, pkgconfig 2.0.3, plyr 1.8.6, purrr 0.3.4, R6 2.5.0, RColorBrewer 1.1-2, Rcpp 1.0.5, RCurl 1.98-1.2, reshape 0.8.8, rlang 0.4.10, rmarkdown 2.6, RSQLite 2.2.1, rstudioapi 0.13, rvest 0.3.6, scales 1.1.1, splines 4.0.2, stringi 1.5.3, stringr 1.4.0, survival 3.2-7, tibble 3.0.4, tidyselect 1.1.0, tools 4.0.2, vctrs 0.3.6, viridisLite 0.3.0, webshot 0.5.2, withr 2.3.0, xfun 0.20, XML 3.99-0.5, xml2 1.3.2, xtable 1.8-4, XVector 0.30.0, yaml 2.2.1, zlibbioc 1.36.0

Parameter values used for this analysis are:

- workDir: /shared/ifbstor1/galaxy/jobs/003/859/3859009/working
- projectName: S1IPlocal
- author: Stefano
- targetFile: /shared/ifbstor1/galaxy/datasets/005/771/dataset_5771673.dat
- rawDir: /shared/ifbstor1/galaxy/jobs/003/859/3859009/working/rawDir_unzipped
- featuresToRemove: alignment_not_unique, ambiguous, no_feature, not_aligned, too_low_aQual
- varInt: group
- condRef: contr
- batch: batch
- fitType: local
- cooksCutoff: TRUE
- independentFiltering: TRUE
- alpha: 0.05
- pAdjustMethod: BH
- typeTrans: VST
- locfunc: median
- colors: #f3c300, #875692, #f38400, #a1caf1, #be0032, #c2b280, #848482, #008856, #e68fac, #0067a5

Bibliography

1. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria : R Foundation for Statistical Computing, 2017 :
2. Gentleman RC, Carey VJ, Bates DM, *et al.* Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 2004 ; 5 : R80.
3. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology* 2010 ; 11 : R106.
4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014 ; 15 : 550.
5. Schulze SK, Kanwar R, Gölzenleuchter M, *et al.* SERE: Single-parameter quality control and sample comparison for RNA-seq. *BMC Genomics* 2012 ; 13 : 524.
6. Cook RD. Detection of influential observation in linear regression. *Technometrics* 1977 ; 19 : 15–18.
7. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* 1987 ; 49 : 1–39.

8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995 ; 57 : 289–300.