CeGaT

Research &
Pharma Solutions

Transcriptome Sequencing

Bioinformatic Note

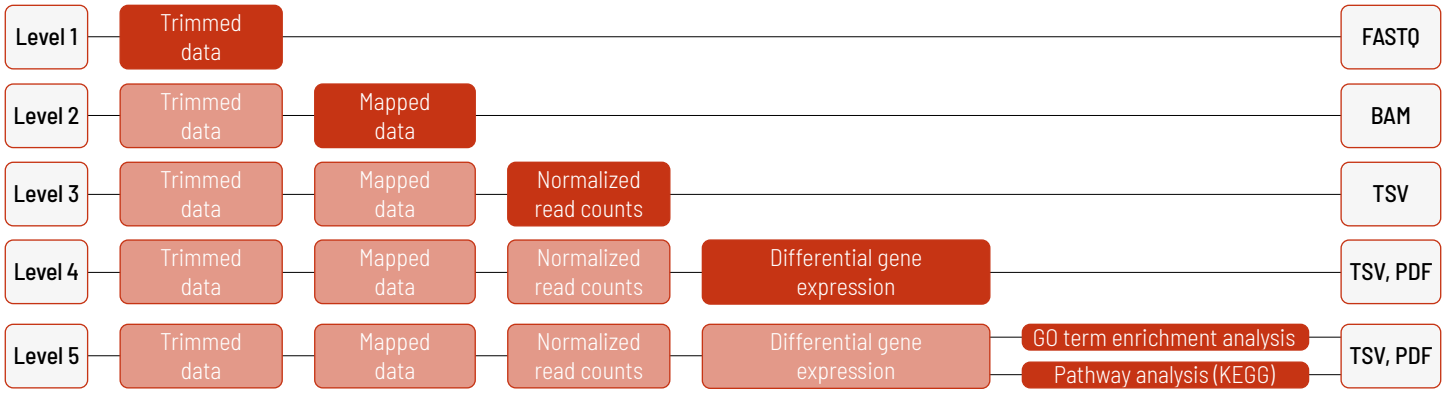
Transcriptome Sequencing

The transcriptome comprises all RNAs present in a specific cell or tissue type at a distinct time. Their presence and abundance correspond to the current metabolic state of the cells and are affected by external and internal changes. Transcriptome sequencing is a powerful method to detect and quantify RNA molecules.

Application areas and objectives for transcriptome sequencing are diverse and include

- the analysis of differential gene expression levels, and
- the detection of alternative splicing and previously unknown transcripts.

Different levels of bioinformatic data analysis are available:



With increasing bioinformatics analysis level, more data is delivered. All higher levels include the data from the lower levels. For example, in Level 2, trimmed data and mapped data are provided. In addition to the data, and independent of the analysis level, a project report is generated.

Levels 1 and 2

If you wish to analyze your data yourself, we recommend the Levels 1 or 2. The default level for raw data is Level 1, where trimmed reads in FASTQ format are delivered. In this level, the sequencing data are demultiplexed and trimmed. This level is provided for every project, regardless of additionally purchased bioinformatic analyses.

If you wish to receive Level 2, the trimmed reads are additionally mapped. In addition to the trimmed reads in FASTQ format, you will receive the mapped reads as BAM files.

The project report provides information for every sample about the laboratory protocol, including data about quality control of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing. For the trimmed data, the number of sequenced fragments and bases are reported, and the sequence length, quality of the reads, and the GC content are illustrated in bar plots for all samples.

For Level 2, the project report also includes a table with statistics of the mapped fragments including the number of mapped fragments, and the proportion of sequenced fragments.

Level 3

The raw counts derived from the mapping contain the number of reads that map to each gene. Based on these numbers, the normalized counts are calculated. We remove genes with less than two reads over all samples. This improves the detection power by making the multiple testing adjustment of the p-values less severe.

In addition to the previous files, two TSV files are delivered containing the raw counts and the normalized counts. Table 1 shows an excerpt of the raw counts file. The columns contain the samples, while the rows list the genes. The gene_id column contains the Ensembl ID, the gene column the gene name given by the HGNC (HUGO Gene Nomenclature Committee). Every project receives a unique S-number (SXXXX). In this example, the S-number is S1163. For every sample of the project, a column is included in the TSV file. The numbers in the columns indicate how many reads were mapped to the respective genes.

The normalized counts file has the same structure; it only differs in the numbers as they are normalized, as described above.

Table 1 | Excerpt of the raw counts file.

gene_id	gene	S1163-Nr100	S1163-Nr101	S1163-Nr102	S1163-Nr103	S1163-Nr104	S1163-Nr105
ENSG00000210049	MT-TF	7	7	8	7	14	12
ENSG00000211459	MT-RNR1	17271	17618	17976	19933	20499	19973
ENSG00000210077	MT-TV	0	0	0	0	0	0
ENSG00000210082	MT-RNR2	169218	166563	165531	150446	150235	149715
ENSG00000209082	MT-TL1	0	0	0	0	0	0
ENSG00000198888	MT-ND1	97863	104062	110199	92974	92951	89602
ENSG00000210100	MT-TI	5	9	6	7	12	7
ENSG00000210107	MT-TQ	20	17	21	89	94	84
ENSG00000210112	MT-TM	2	0	0	2	1	4

Level 3 provides only the raw and normalized counts. These numbers might give first insights into the metabolic state of the cells. With an additional differential expression analysis, even more questions can be answered as explained in the next section under Level 4.

To visualize the relationship between samples, we use two methods: Hierarchical clustering and principal component analysis (PCA). The hierarchical clustering is shown in figure 1 as a dendrogram. It is a one-dimensional representation of the relationship between the samples. The PCA plot in figure 2 shows the two dimensions of the data room, which explains most of the variance between the samples.

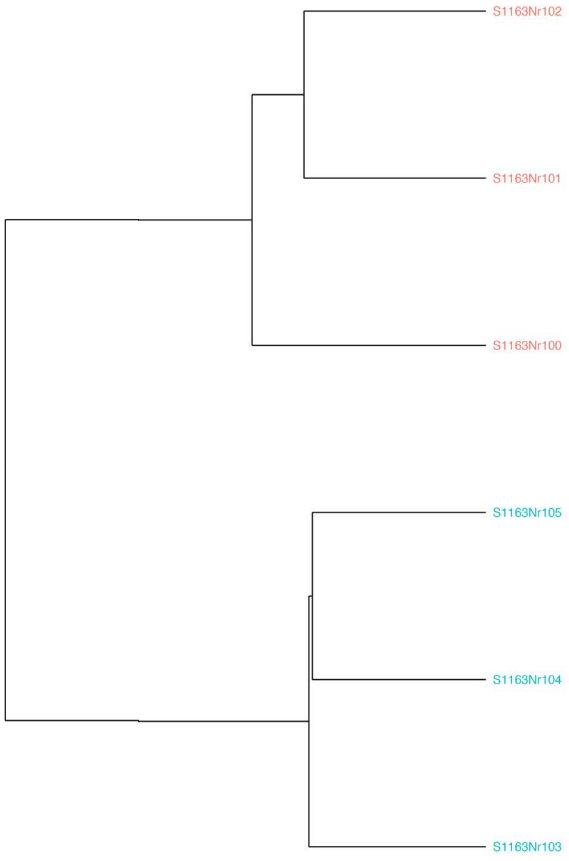


Figure 1 | Hierarchical clustering of the relationship between the samples in a dendrogram. Samples are colored according to the groups (if provided).

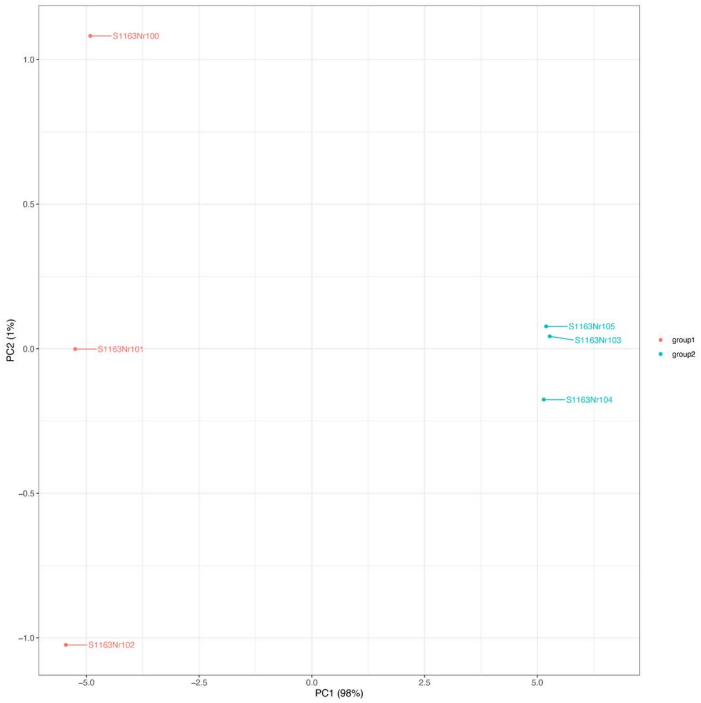


Figure 2 | Principal component analysis of expression data transformed with variance stabilizing transformation of all genes that received at least two reads. The percentage values on the axes describe how much of the variance between samples is captured in this principal component. Samples are colored according to the groups (if provided).

Level 4

In Level 4, we perform one or several group comparisons and report differentially expressed genes. For these comparisons, at least three replicates per group are required. Additionally, the groups that should be compared need to be indicated. In our example, group 1 comprises the samples S1163Nr100 – S1163Nr102, while group 2 includes the samples S1163Nr103 – S1163Nr105.

Using normalized counts, the log2 fold change is calculated. Hereby, the p-value reports the statistical significance of the result. Since we compare thousands of genes, we must correct for multiple tests to adjust the p-value (padj). Hence, the padj value should be used to determine significant differences in gene expression.

The group comparison and differential expression analysis (DEA) result is supplied in a TSV file. Table 2 shows an excerpt of this TSV file. As in table 1, the gene_id and gene columns indicate the Ensembl ID and HGNC gene name, respectively. The baseMean column returns the mean of the normalized read counts over all samples, independent of their group. The log2FoldChange reflects the difference in expression levels between the groups. It is also calculated based on the normalized read counts. Positive numbers indicate an upregulation in group 2, while negative numbers represent a downregulation in group 2 compared to group 1. The columns lfcSE and stat are the standard error of the log2 Fold Change and a statistical value for the calculation of the Wald test, respectively. The last two columns contain the p-value and the adjusted p-value. In addition to the differential expression analysis file containing statistics for all annotated genes, a TSV file with only significantly differentially expressed genes is provided.

Table 2 | Excerpt of the differential expression analysis file.

gene_id	gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000210049	MT-TF	9.10	0.59	0.64	0.93	0.35	0.55
ENSG00000211459	MT-RNR1	18807.55	0.21	0.03	7.10	1.28E-12	1.93E-11
ENSG00000210082	MT-RNR2	157951.93	-0.14	0.03	-4.99	6.03E-07	4.80E-06
ENSG00000198888	MT-ND1	97523.01	-0.17	0.04	-4.32	1.55E-05	9.96E-05
ENSG00000210100	MT-TI	7.63	0.39	0.70	0.55	0.58	0.75
ENSG00000210107	MT-TQ	54.10	2.22	0.29	7.56	3.91E-14	6.71E-13
ENSG00000210112	MT-TM	1.50	1.83	1.75	1.05	0.30	NA

In addition to visualizing the differences between the samples, we provide visualizations of the expression data and the normalized counts in form of a heatmap, an MA plot, and a volcano plot.

The heatmap (figure 3) is a color-coded representation of the normalized values. This map shows which genes and which samples have a similar expression profile. Similar expression profiles can be strengthened by clustered rows and columns. The heatmap displays genes with the smallest p-values. Thus, this graph shows the 25 up- and down-regulated genes, respectively, that have the most significant differences but not the highest (absolute) expression.

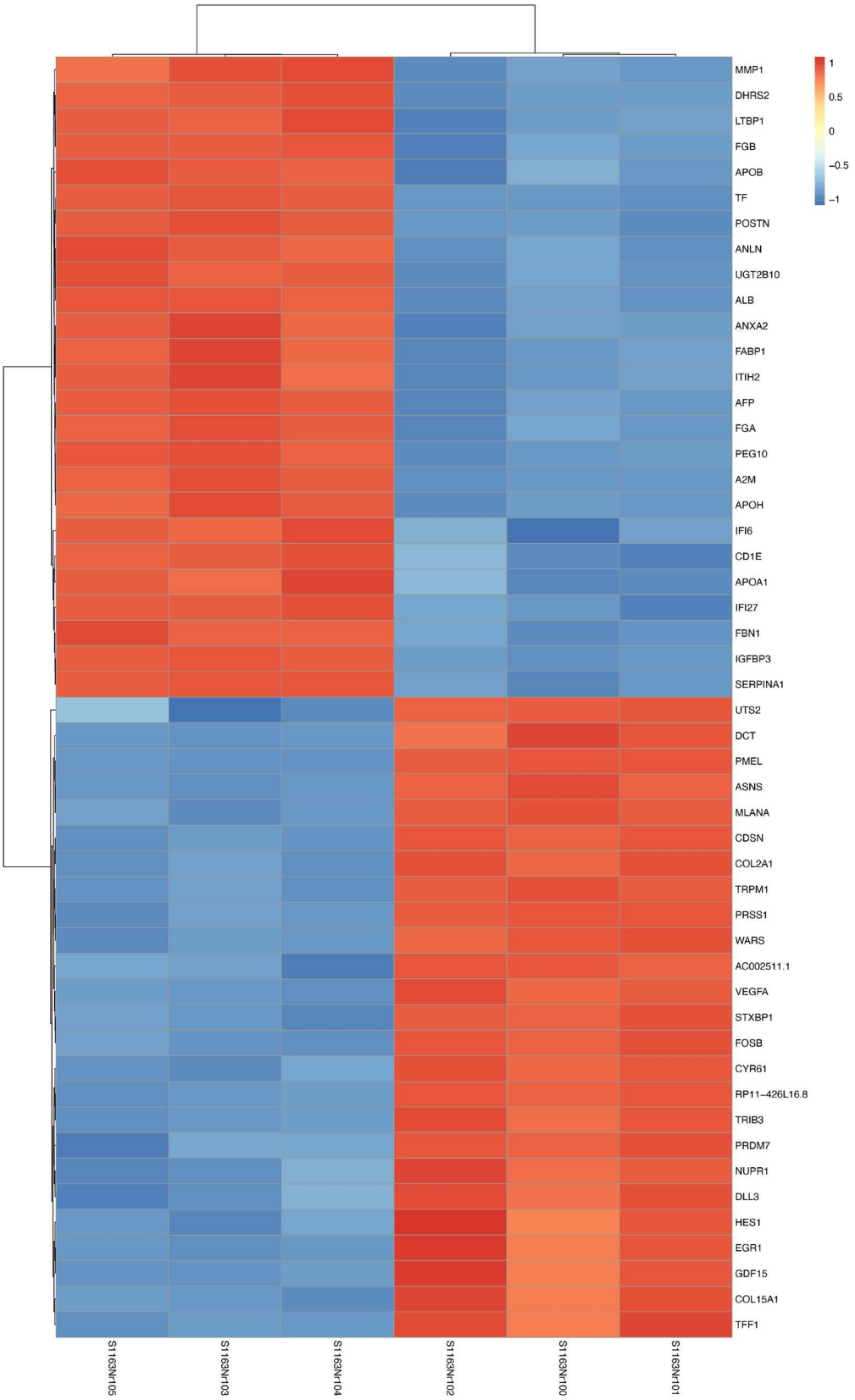


Figure 3 | Heatmap of the most significantly differentially expressed genes. Expression levels of the genes were scaled row-wise and the resulting z-scores were plotted. Red means the value is higher than the mean of all values for this gene, blue means it is lower.

The volcano plot in figure 4 is a graphical representation of the entire dataset. On the x-axis the log2 Fold Change (log2FC) is displayed, on the y-axis the negative log10 of the adjusted p-value. The vertical dashed lines represent the threshold of genes with a log2FC greater than 1.5 for upregulation or smaller than -1.5 for downregulation. This threshold can also be changed based on your needs. The horizontal dashed line represents the threshold of genes with an adjusted p-value smaller than 0.05. Thus, all red dots represent genes that are significantly differentially expressed. The blue dots have an adjusted p-value that is lower than 0.05, but the log2FC is lower than |1.5|. The green dots have a log2FC greater than |1.5| but an adjusted p-value greater than 0.05.

The grey dots have an adjusted p-value greater than 0.05 and a log2FC lower than |1.5|. Thus, the red dots might be of interest for further analyses. For that reason, gene names are indicated for those dots. If too many genes are in that area, not all genes are labeled with their gene names to facilitate the readability.

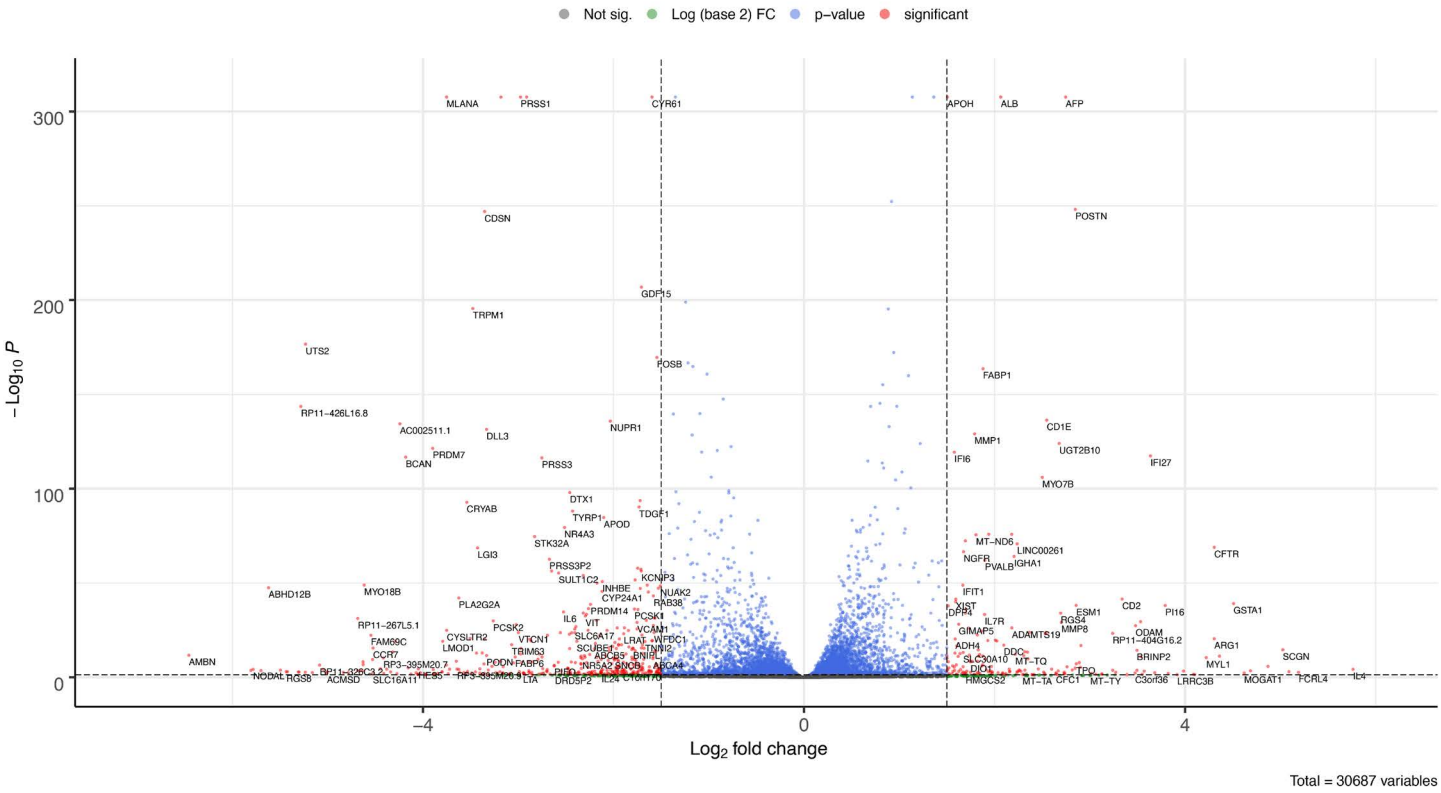


Figure 4 | Volcano plot of the entire dataset. Significantly expressed genes with an adjusted p-value < 0.05 and a log2FC greater than |1.5| are illustrated as red dots.

The MA plot in figure 5 is another possibility to visualize the entire dataset. On the x-axis, the log2 mean expression is shown, on the y-axis, the log2FC. The dotted line represents the significance threshold for the log2FC of |1.5|. Red dots are significantly upregulated genes with a log2FC greater than 1.5 and an adjusted p-value smaller than 0.05. Blue dots are significantly downregulated genes with a log2FC lower than -1.5 and an adjusted p-value smaller than 0.05. All grey dots are not considered as significantly differentially expressed. As the x-axis shows the log2 mean expression, the dots more right have a higher mean expression level compared to the dots more left. This information can be used to identify highly expressed genes.

For your convenience, we also create an overview figure that comprises all described graphs.

With this differential expression analysis, the metabolic states of different groups can be compared. Depending on your research question, the data might reveal

- effects of treatment strategies,
- effects of different nutrition environments,
- differences between sexes,
- differences between species, and much more.

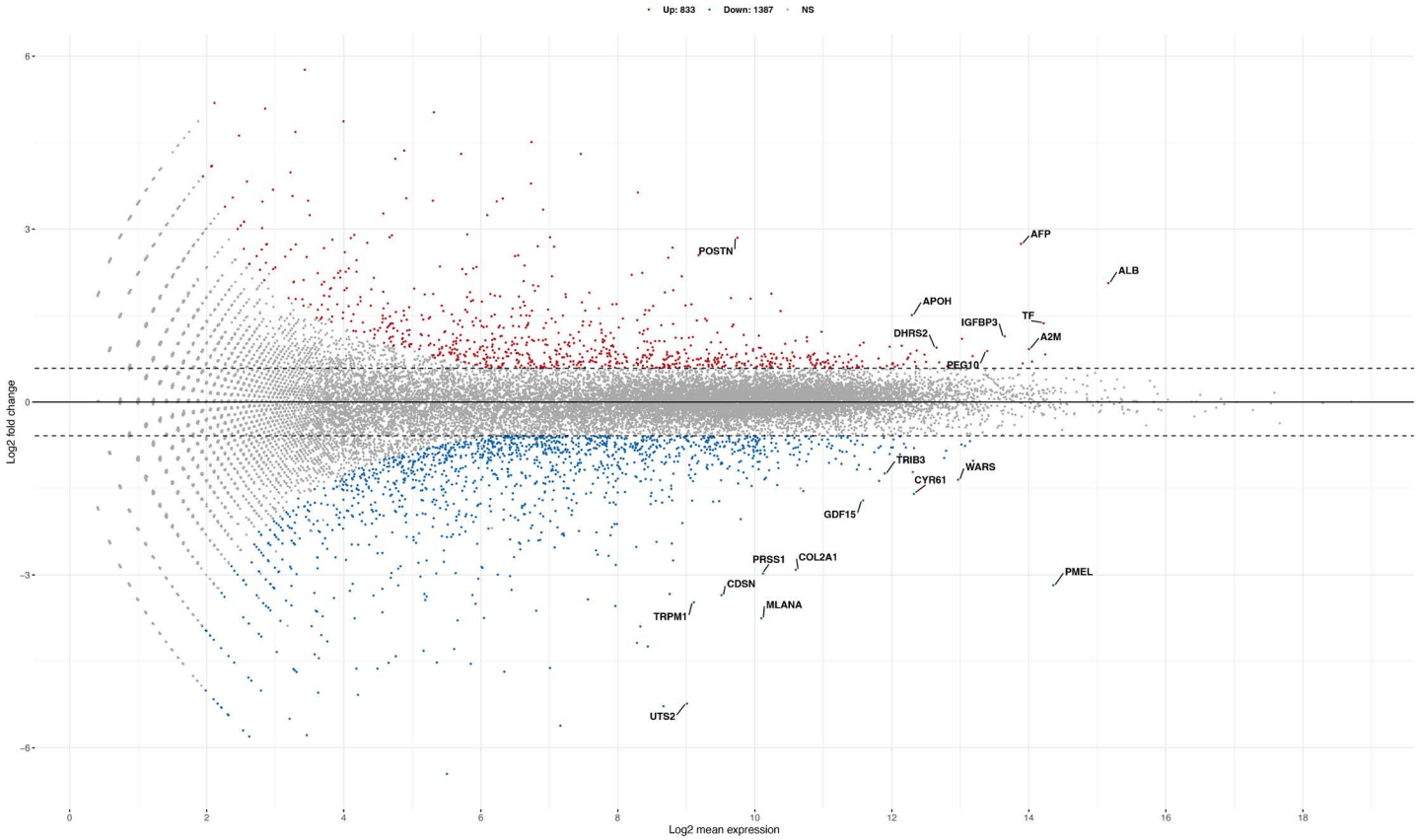


Figure 5 | MA-plot of the entire dataset. Significantly upregulated genes with an adjusted p-value < 0.05 and a log2FC greater than 1.5 are illustrated as red dots. Significantly downregulated genes with an adjusted p-value < 0.05 and a log2FC lower than -1.5 are illustrated as blue dots. All grey dots are not considered as significantly differentially expressed.

Level 5

If you are interested in further investigations of the significantly differentially expressed genes, we offer the following enrichment analyses: GO term enrichment analysis or KEGG pathway enrichment analysis. Please note that Level 5 only includes one of the two enrichment analyses. However, if you wish to receive both analyses, we can send you a respective offer.

Go Term Enrichment Analysis

The Gene Ontology (GO) knowledgebase comprises information on the functions of genes. It is divided into three large areas, so-called ontologies: Molecular function (MF), biological process (BP), and cellular component (CC). Each ontology has GO terms consisting of a name and a unique accession identifier. The GO terms have a tree-like structure and become more precise with increasing depth within the tree structure. The GO term enrichment analysis aims to identify overrepresented GO terms in a gene list, e.g., of differentially expressed genes (Ashburner et al. 2000), compared to a gene universe (e.g., the human transcriptome).

The list of the identified differentially expressed genes and an annotated gene list of the organism of interest is used to map and annotate all genes to the available GO terms. As this annotation file can become significant for a large number of differentially expressed genes assigned to many GO terms, a statistical test is conducted to identify significantly overrepresented GO terms in the list of differentially expressed genes in contrast to the annotated gene list. As we run this statistical test for every GO term, we correct for multiple tests.

The result of this analysis is a TSV file for each ontology containing the significantly enriched GO terms. Table 3 shows an excerpt of the result of the GO term enrichment analysis for the BPs ontology. The first two columns indicate the identified GO term with its ID and descriptive name, respectively. The gene ratio states the proportion of genes from the differentially expressed gene list that are assigned to the GO term. In the first row, for example, 202 genes from the 5976 differentially expressed genes were annotated with the GO term GO:0009123. The Bg Ratio column states how many genes from that organism are assigned to this GO term. In our example, 384 out of the 19,786 human genes in that database were annotated with this GO term. Using these ratios, the hypergeometric test is conducted, and the p-value, adjusted p-value (due to multiple testing), and the q-value from the statistical test are returned. The q-value provides information about the false discovery rate. The geneID column lists all genes from the differential expression gene list assigned to the GO term. The IDs are Ensembl IDs. Please note that in table 3 only the first ten genes are listed for the sake of brevity. The provided TSV file contains, of course, all gene IDs. The last column again provides the count of genes from the differential gene list annotated with the respective GO term.

Table 3 | Excerpt of significant GO terms (BP).

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0009123	nucleoside monophosphate metabolic process	202/ 5976	384/ 19786	2,22E-06	1,06E-02	7,00E-03	ND1/ND2/COX2/ATP8/ND3/ND4L/ND6/NADK/ PARK7/PINK1/...	202
GO:0043062	extracellular structure organization	223/ 5976	437/ 19786	3,25E-07	1,06E-02	7,00E-03	VWA1/PDPN/MFAP2/PLA2G2A/HSPG2/COL16A1/ COL8A2/COL9A2/CCN1/VCAM1/...	223
GO:0030198	extracellular matrix organization	198/ 5976	377/ 19786	6,72E-06	1,36E-02	9,02E-03	VWA1/PDPN/MFAP2/HSPG2/COL16A1/COL8A2/ COL9A2/CCN1/VCAM1/COL11A1/...	198
GO:0140014	mitotic nuclear division	148/ 5976	258/ 19786	8,38E-06	1,36E-02	9,02E-03	AURKAIP1/PHF13/MAD2L2/CEP85/RCC1/CDCA8/ CDC20/KIF2C/PSRC1/POGZ/ ...	148
GO:0009161	ribonucleoside monophosphate metabolic process	191/ 5976	365/ 19786	5,13E-06	6,21E-03	4,12E-02	ND1/ND2/COX2/ATP8/ND3/ND4L/ND6/NADK/ PARK7/PINK1/...	191

To reduce the redundancy of similar GO terms, we calculate a so-called redundancy factor. The redundancy factor is calculated based on the semantic similarity of GO terms by including the GO terms' positions in the tree and the parent nodes. It ranges from 0 to 1, where higher values correspond to similar GO terms. A high redundancy factor is used for a rough overview of the GO terms. In contrast, a small redundancy factor is used for a more detailed view.

The reduced significant GO term list is also provided as TSV file for every ontology. It has the same structure as described in table 3.

The 20 most significant GO terms are visualized in both a bar plot and a dot plot, and the 30 most significant GO terms are additionally visualized in an enrichment map. These plots are created for each of the three ontologies (MF, BP, CC), respectively.

Figure 6 shows the bar plot of the GO term enrichment analysis for the BP ontology. The color indicates the adjusted p-value, and the length of the bar plot the frequency of occurrence of this GO term in the differentially expressed gene dataset.

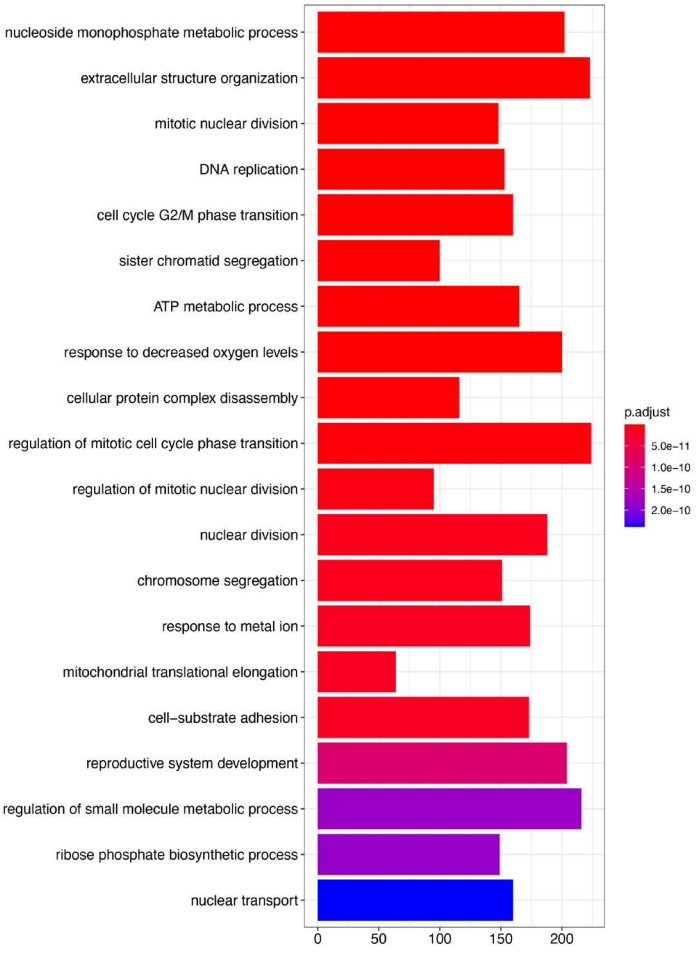


Figure 6 | Bar plot of the results of the GO term enrichment analysis. The 20 most significantly enriched GO terms for the biological process (BP) ontology are illustrated and their frequency of occurrence is indicated.

Similarly, figure 7 shows the GO term enrichment analysis for the BP ontology as a dot plot. Again, the adjusted p-value is indicated by the color coding. The occurrence frequency of the GO term is, however, indicated by the size of the dot. The position on the x-axis displays the gene ratio that is also given in the TSV file (see table 3).

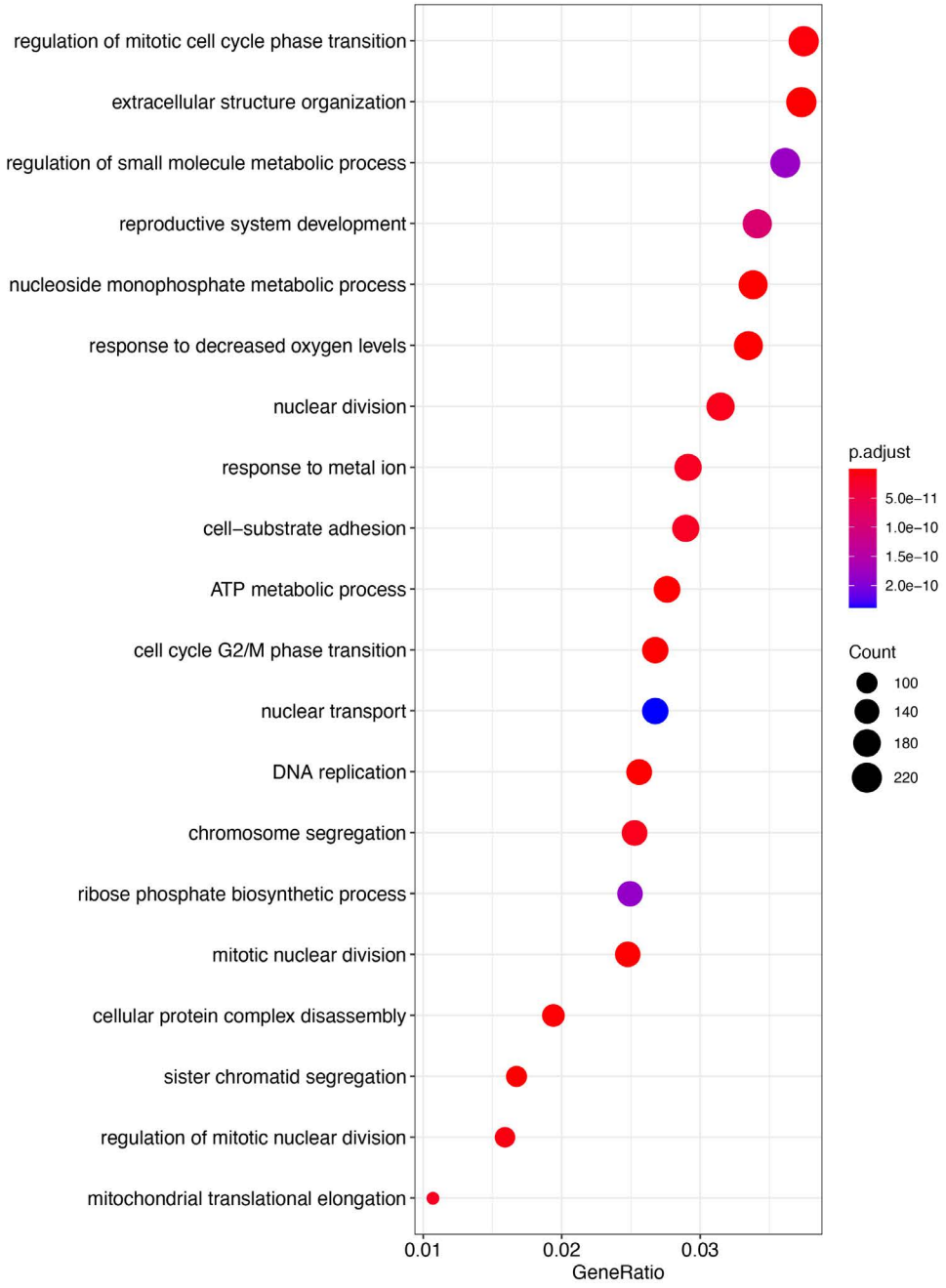


Figure 7 | Dot plot of the results of the GO term enrichment analysis. The 20 most significantly enriched GO terms for the biological process (BP) ontology are illustrated and their frequency of occurrence is indicated.

The enrichment map in figure 8 organizes enriched terms in a network that connects overlapping gene sets with edges. Thus, overlapping gene sets tend to cluster together. This clustering facilitates the identification of functional modules. As seen in figure 7, the size of the dots indicates the frequency of the GO term, and the color of the adjusted p-value.

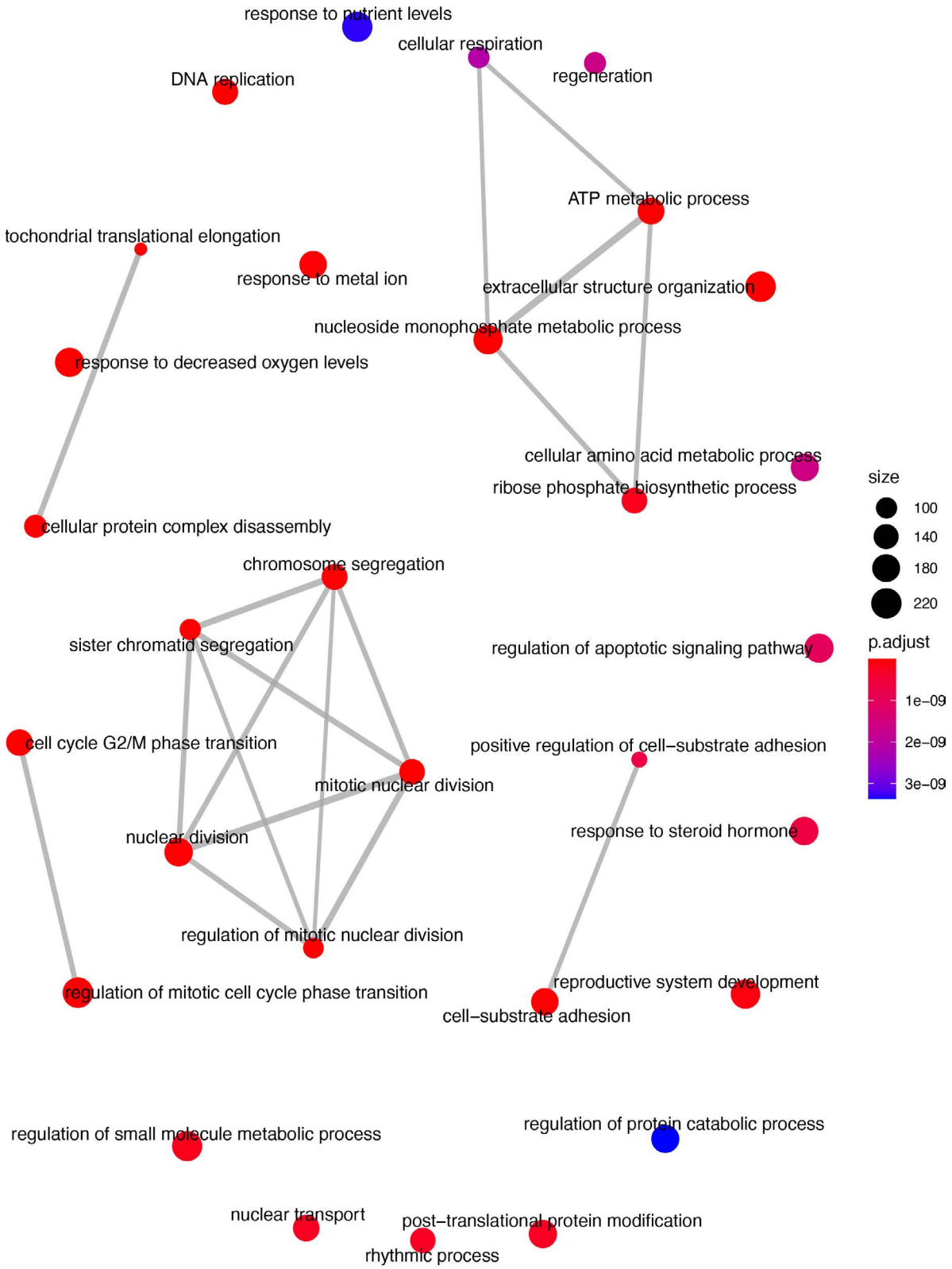


Figure 8 | Enrichment map of the results of the GO term enrichment analysis. The 30 most significantly enriched GO terms for the biological process (BP) ontology are illustrated and the edges connect overlapping gene sets.

KEGG Analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database represents biological systems in terms of molecular networks, especially in the form of manually curated KEGG pathway maps (Kanehisa, Goto 2000).

As for the GO term enrichment analysis, the differentially expressed gene list and an annotated gene list of the organism of interest are used to map and annotate all genes to the KEGG pathways. As this annotation file can become significant for a large number of differentially expressed genes with many KEGG pathways, a statistical test is conducted to identify significantly overrepresented KEGG pathways in the list of differentially expressed genes in contrast to the annotated gene list. As we run this statistical test for every KEGG pathway, we correct for multiple testing.

The resulting TSV file has the same structure as the TSV file for the GO term enrichment analysis displayed in table 3. Only the ID does not refer to a GO term, but to a KEGG pathway.

The 20 most significant KEGG pathways are visualized in a bar plot and the 5 most significant pathways additionally in a dot plot. The dot plot contains all comparisons, which is especially interesting when more than two groups are compared. The bar plot and the dot plot have the same structure as figure 6 and figure 7, respectively, and are therefore not explained again.

For the three most significant KEGG pathways, the KEGG pathway maps are provided, where the differentially expressed genes are marked. One example is given in figure 9. In this pathway map, the gene expression values are mapped to a gradient color scale, highlighting differentially expressed genes.

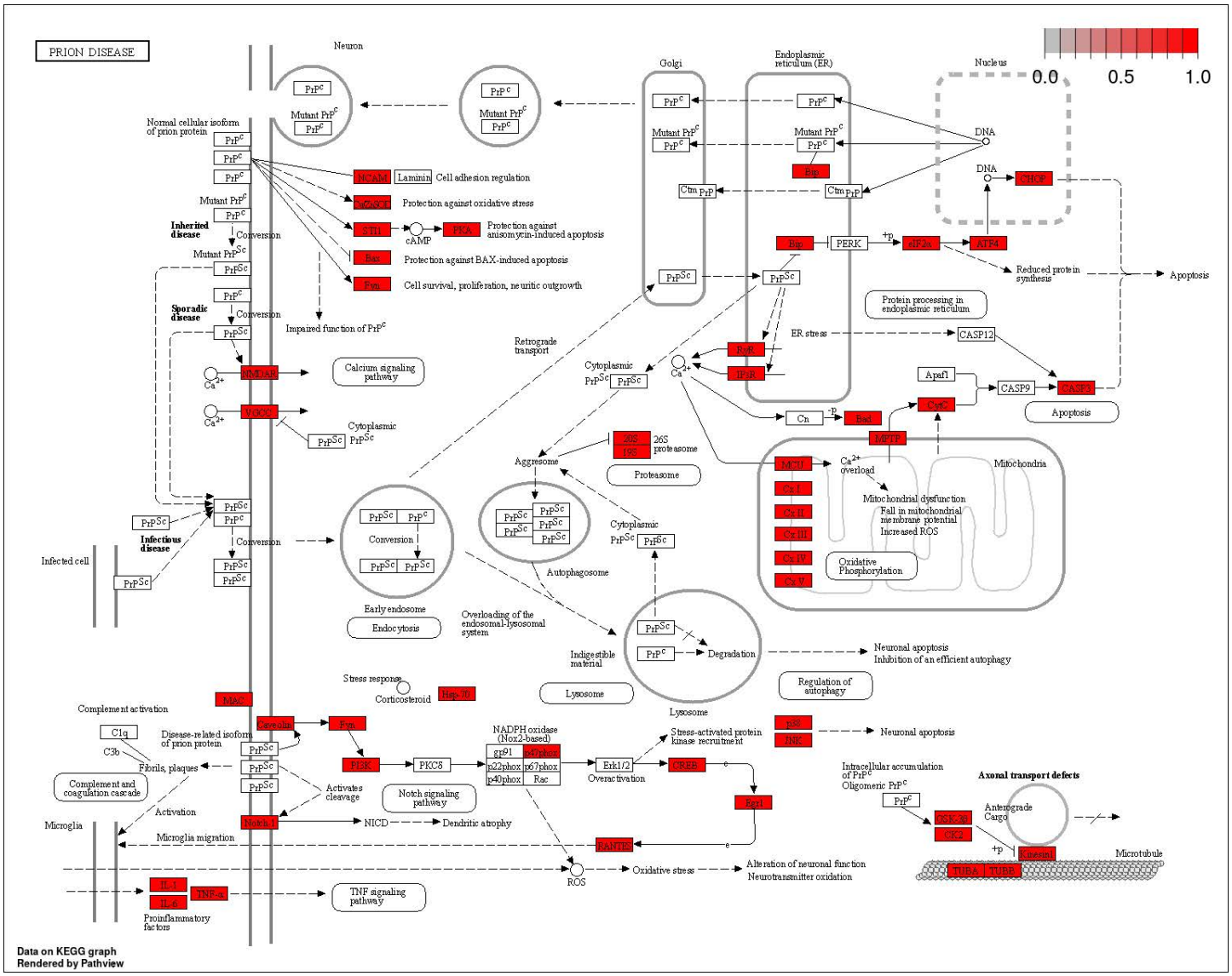


Figure 9 | Pathway map of the most significantly enriched KEGG pathway. The differentially expressed genes are highlighted in the map.

References

Ashburner, Michael; Ball, Catherine A.; Blake, Judith A.; Botstein, David; Butler, Heather; Cherry, J. Michael; Davis, Allan P.; Dolinski, Kara; Dwight, Selina S.; Eppig, Janan T.; Harris, Midori A.; Hill, David P.; Issel-Tarver, Laurie; Kasarskis, Andrew; Lewis, Suzanna; Matese, John C.; Richardson, Joel E.; Ringwald, Martin; Rubin, Gerald M.; Sherlock, Gavin (2000): Gene Ontology: tool for the unification of biology. In Nat. Genet., (1):25-9. DOI: 10.1038/75556.

Kanehisa, Minoru; Goto, Susumu (2000): KEGG: Kyoto Encyclopedia of Genes and Genomes. In Nucleic Acids Research 28(1), pp. 27-30. DOI: 10.1093/nar/28.1.27

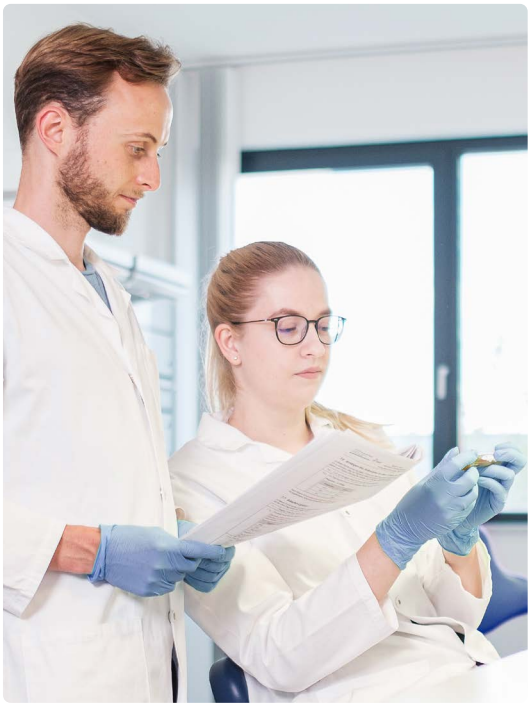
About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.

For more details please visit
www.cegat.com/rps



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00
Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

2023/12