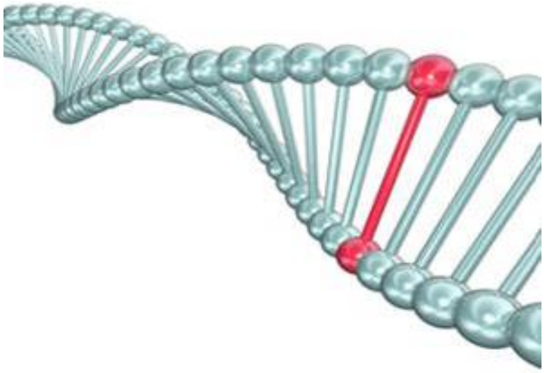
Source BioScience

Bioinformatics Report



AK

Table of Contents

Spliced mapping of reads...3

Gene-wise read counting...4

Group-wise differential gene expression analysis...5

PCA plot of samples...6

Sample-to-sample heatmap...7

MA plot...8

DEG heatmap: CONTROL vs TREATED (50 differentially expressed genes at log2fold-diff> 1.8, padj 0.05)...9

Definitions...10

SAM/BAM format...10

Differential gene expression analysis results in CSV format...11

Advanced Bioinformatics Workflow

For sequence data a customized bioinformatics pipeline has been setup and each individual sample has been processed. The following section explains the steps that have been performed and describes the generated output files.

The bioinformatics pipeline consists of three parts:

1. Spliced mapping of reads
2. Gene-wise read counting
3. Group-wise differential gene expression analysis

Spliced mapping of reads

Reference information:

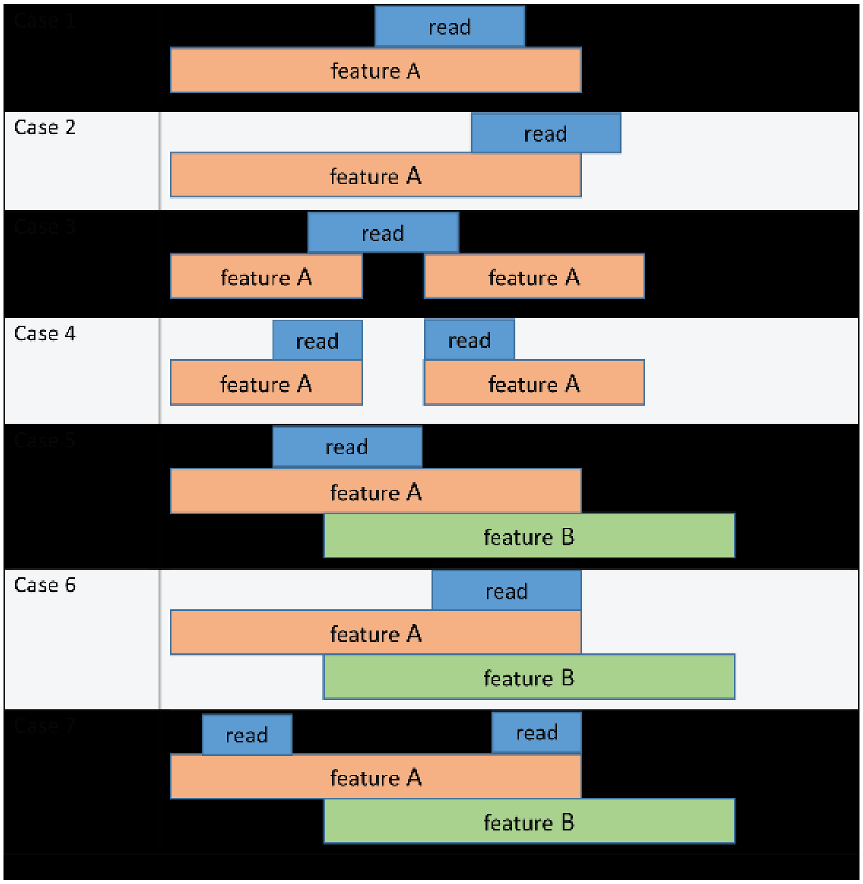
Species	Source	Size (bp)
Homo Sapiens	USCS (hg19)	3.1 Gb

Data was mapped using HISAT (version 2.0.5). Our standard parameter set for stranded RNA was used, which includes pre-alignment against annotated splice junctions, no use of discordant mapped reads and it is enforced that the right-most end (in transcript coordinates) of the fragment is the first sequenced[1]. Also reads were first mapped to known gene locations. In a second cycle all reads that could not be mapped were aligned to the entire genome.

[1] More information can be found on https://ccb.jhu.edu/software/hisat2/manual.shtml

Gene-wise read counting

Using the provided reference annotation stated in the previous section, the number of sequence fragments that have been assigned to each gene were counted for each sample, using the featureCounts model illustrated below:



featureCounts model: overview of different counting scenarios[2]

To quantify differences between conditions and measure statistical significance, the DeSeq2 package was used to estimate the dispersion and logarithmic fold change between comparative groups for each gene.

[2] Source: https://uk.mathworks.com/help/bioinfo/ref/featurecount.html?s_tid=gn_loc_drop

Group-wise differential gene expression analysis

For the analysis of differentially expressed genes (DEG), 4 samples were grouped into the TEST group. For the CONTROL group, the data from 3 samples provided by the customer was used. A single comparison was performed, between the TEST and CONTROL groups.

Groups:

Control	Treated
D4_S10	V4_S12
D5_S13	V5_S15
D6_S16	V6_S18

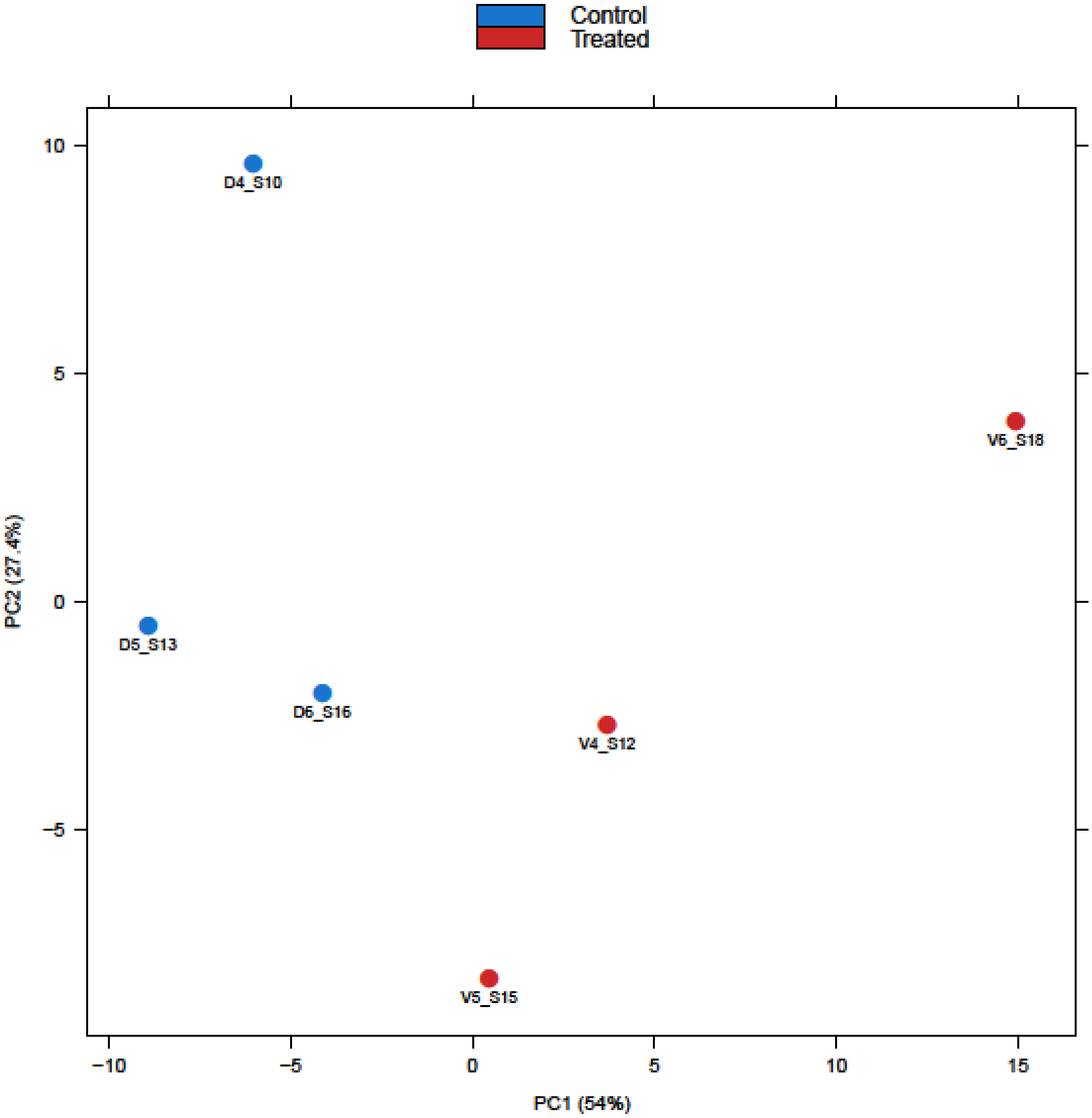
Comparisons:

1. CONTROL vs TREATED

Results were written to a comma-separated output file with the prefix “Results_” and have also been converted to Microsoft Excel format. More information about the file format as well as a description of the actual content can be found in section “Definitions”.
Using the DeSeq2 tool, we have generated the following plots to help visualize your data. These plots are shown below and are also available as separate files:

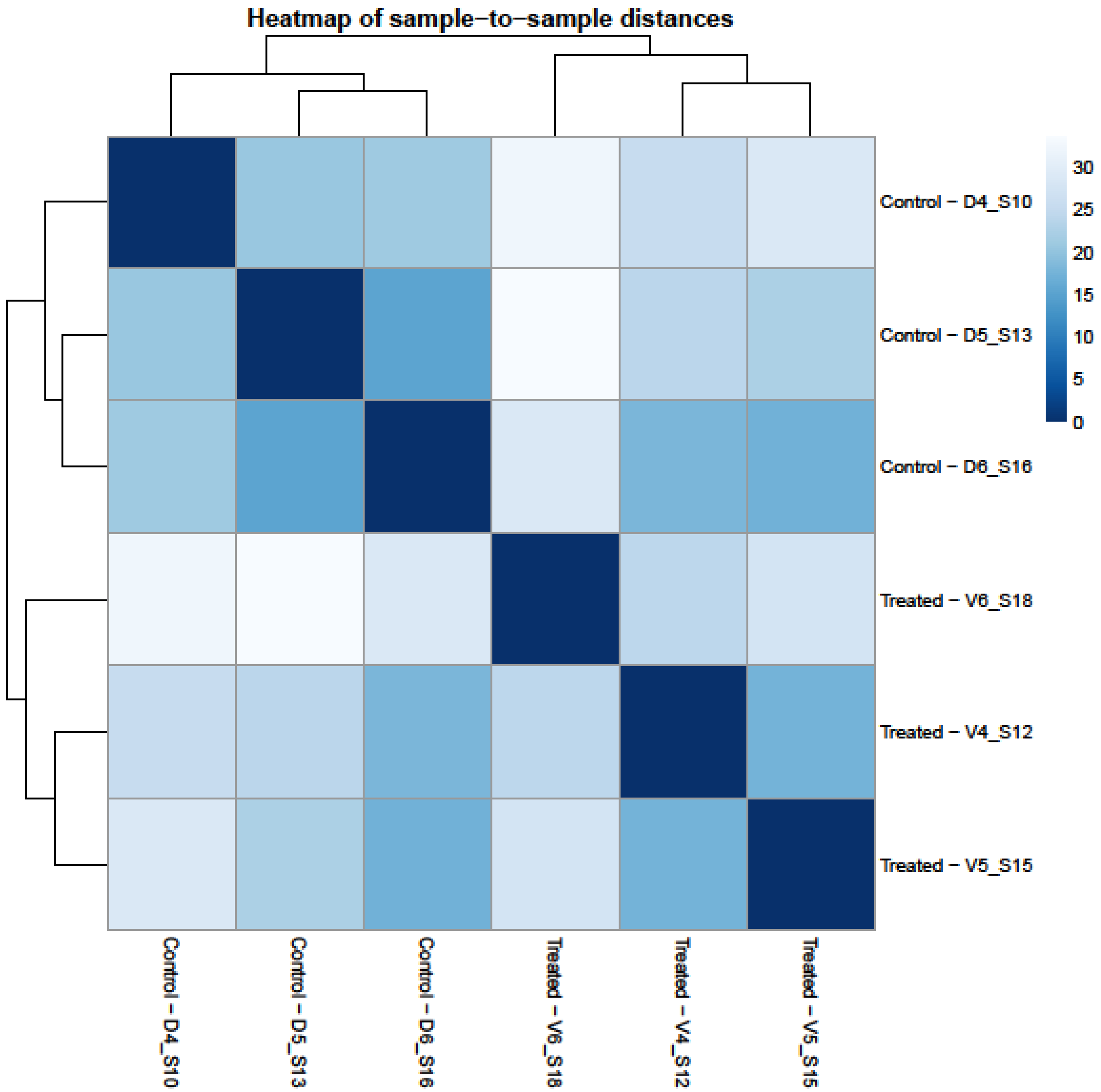
- a PCA plot of the first two principal components between all samples
- a sample-to-sample distance heatmap showing Euclidian distances between samples
- an MA plot for the CONTROL vs TEST comparison, where the log$_{2}$-fold change (lfc) for each gene is plotted against the average expression level (normalised count)
- a heatmap of counts for all differentially expressed genes with lfc > 1.8, adj $p < 0.05$ for the CONTROL vs TREATED comparison.

PCA plot of samples



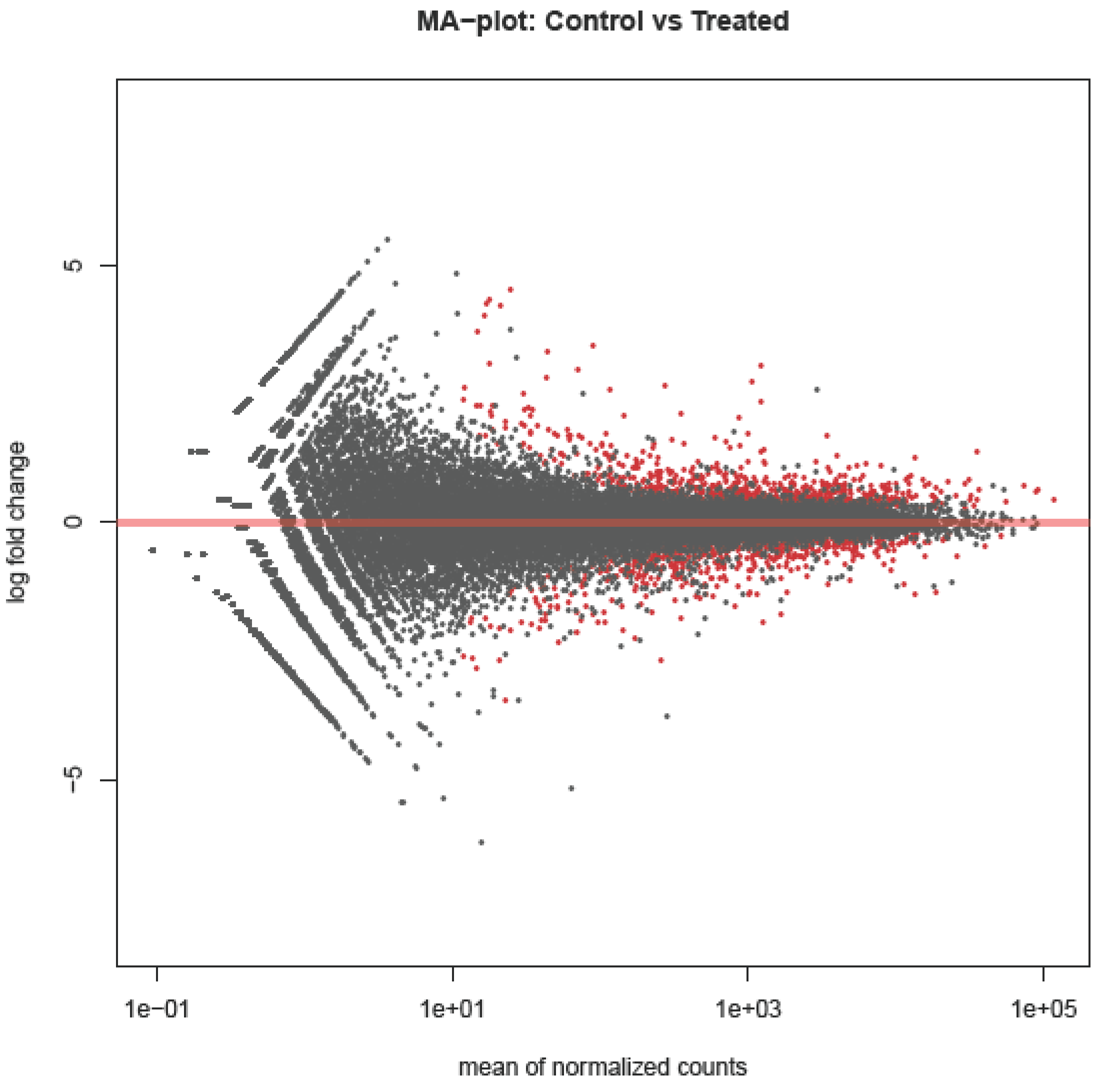
PCA plot showing all samples in a two dimensional plane spanned by their first two principle components. Closely related samples are clustering.

Sample-to-sample heatmap



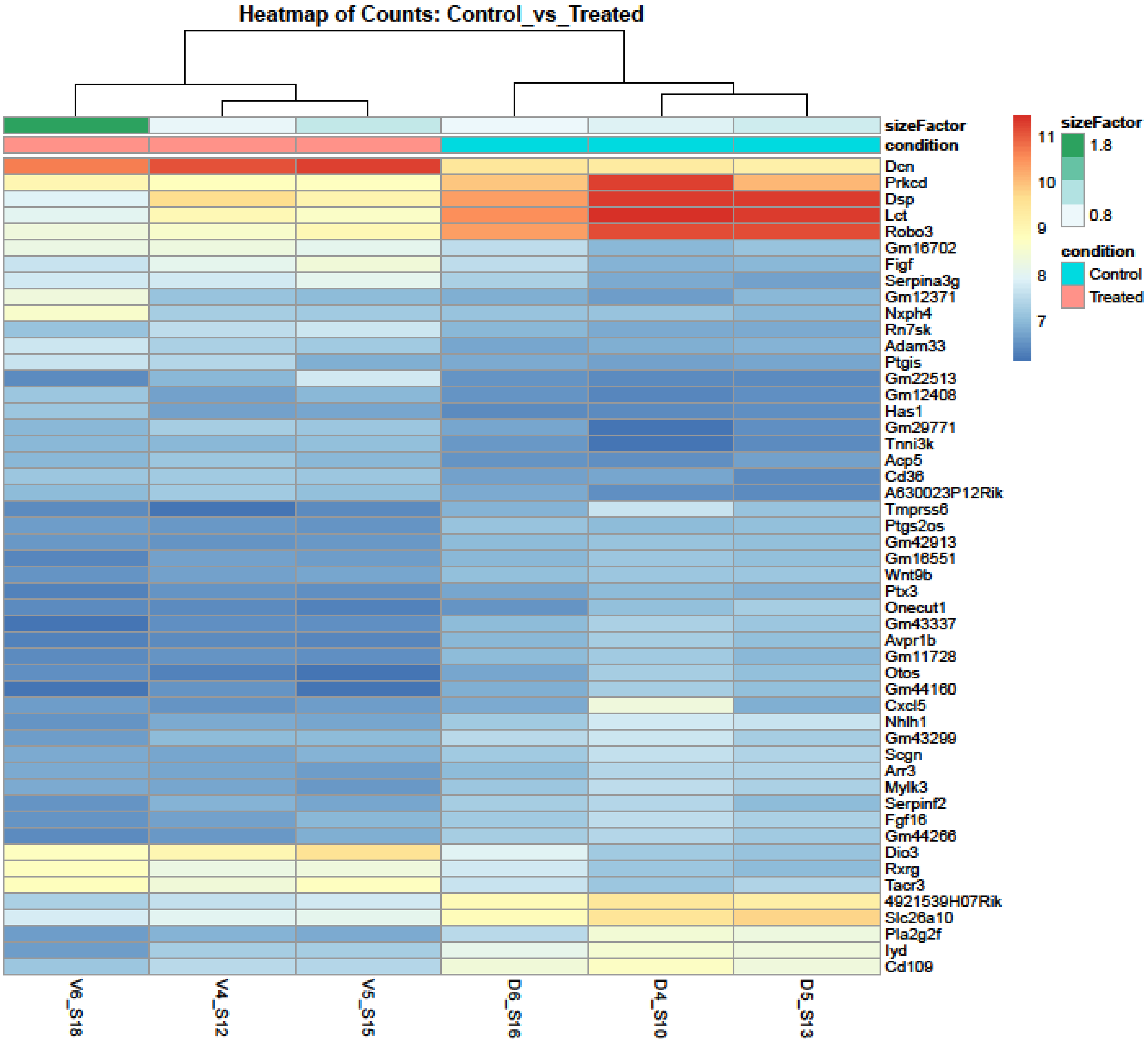
Genome-wide gene expression correlation heatmap between all samples. Samples are clustered by the Euclidean distance between rows/columns and single linkage clustering.

MA plot



MA-plot for the comparison between CONTROL and TREATED group samples. The $\log_2$-fold change (M = y axis) of each differentially expressed gene is plotted against the average expression level (A = x axis), for each comparison. Points are coloured in red if the adjusted *p* value is less than 0.1.

DEG heatmap: CONTROL vs TREATED (50 differentially expressed genes at log2fold-diff> 1.8, padj 0.05)



Expression data of significantly differentially expressed genes (adjusted *p* < 0.05) with a $\log_2$ fold difference > 1.8 as a heatmap, for comparison between the CONTROL and TREATED group samples. The data has been transformed with variance stabilizing transformation prior to clustering.

Definitions

SAM/BAM format

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with `@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information. BAM is a compressed SAM file in BGZF format.

Every row of the output of a SAM/BAM file contains the following information:

Col	**Field**	**Description**
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSition
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Where flag can be one of the following:

Char	Flag	Description
P	0x0001	the read is paired in sequencing
P	0x0002	the read is mapped in a proper pair
U	0x0004	the query sequence itself is unmapped
U	0x0008	the mate is unmapped
R	0x0010	strand of the query (1 for reverse)
R	0x0020	strand of the mate
1	0x0040	the read is the first read in a pair
2	0x0080	the read is the second read in a pair
S	0x0100	the alignment is not primary
F	0x0200	QC failure

Differential gene expression analysis results in CSV format

CSV is a platform independent format that can be read by common spreadsheet software. In this case, every row represents one gene with the following seven attributes:

Name	Description
1. gene_id	Entrez ID (if annotated) or Cufflinks ID (de-novo)
2. baseMean	the base mean over all rows
3. log2FoldChange	log2 fold change (MAP): condition treated vs untreated
4. lfcSE	standard error: condition treated vs untreated
5. stat	Wald statistic: condition treated vs untreated
6. pvalue	Wald test p-value: condition treated vs untreated
7. padj	Benjamini-Hochberg adjusted p-values