

RNA NGS

Data Analysis Report

Project:

Sample Project 00059

Customer:

Dr. Sample Owner

Company/Institute:

QIAGEN

Date:

Thursday, February 1, 2018

Performed by:

QIAGEN GmbH

QIAGEN Genomic Services

Qiagen Str. 1

40724 Hilden

Germany

+49 2103 29 11649

QIAGEN.com/GenomicServices

Analysis reference: 00059

Contents

Summary 3

Results files overview 4

Experimental overview 5

Workflow 6

Data QC..... 7

Mapping and yields 9

Results 11

Conclusion and next steps 23

Data analysis workflow..... 25

Materials and methods 28

References 29

Frequently asked questions 30

Definitions 32

Summary

Dear Dr. Sample Owner,

We have now finalized the Next Generation Sequencing analysis of the RNAs identified in the samples you have submitted to QIAGEN Genomic Services.

Next Generation Sequencing libraries were successfully prepared, quantified and sequenced for all your samples. The collected reads were subjected to quality control and downstream analysis. The principal findings are summarized in this document. Additional information and further details on specific RNA transcripts can be found in the various documents listed on page 4.

The principal findings are summarized in this document, including the results of the unsupervised analysis, supervised differential expression analysis and GO enrichment analysis. Additional information, graphs and plots can be found in the files available in My Projects at [XploreRNA](#).

The easiest way to identify relevant targets for downstream validation and further study, is to use the Gene Sorting Wizard in My Projects at [XploreRNA](#). With the wizard you can explore each group comparison and sort the full list of differentially expressed transcripts using all relevant criteria, including expression level, fold change, and statistical significance.

For more information about QIAGEN's products for validation and functional analysis of your mRNAs or ncRNAs of interest, please go to our [RNA Universe](#).

If you have any questions, please contact your local QIAGEN representative or the lab at Genomic.Services@qiagen.com.

Kind regards,

QIAGEN Genomic Services

Results files overview

Table 1 below lists the results files of your RNA NGS data analysis.

Content	Description
file_descriptions.html	Overview of all result files.
Data QC	All information related to the Quality Control (QC) of your samples. This includes QC of individual reads as well as QC of the overall mapping results of all samples.
Mapping results	Alignment files (BAM) and alignment index files (BAI) as well as information about splice junctions, deletions, and insertions for each sample.
Assembled transcripts	Genes and isoforms identified for each sample including their raw FPKM abundance estimates. In addition a .cxb file is provided to ease further analysis.
Analysis	Result file from the analysis of your samples, including: <ul style="list-style-type: none"> • Tables with expression values • Differential expression analysis results • Unsupervised analysis results • Gene Ontology Enrichment (GOE) analysis

Table 1: List of results files. The files can be downloaded from My Projects at [XploreRNA](#). Go to "Explore Results" and click the "download all files" link. In the root folder you will find a full description of all files provided, and recommended programs to view the different files.

Experimental overview

Sample overview

Table 2 below lists all the samples in this project and their specifications according to the information provided.

Sample name	Sample groups	File name
KD_R1	knockdown	00059-00165.218.R1.fastq.gz, 00059-00165.218.R2.fastq.gz
KD_R2	knockdown	00059-00166.221.R1.fastq.gz, 00059-00166.221.R2.fastq.gz
KD_R3	knockdown	00059-00167.222.R1.fastq.gz, 00059-00167.222.R2.fastq.gz
Scramble_R1	scramble	00059-00162.225.R1.fastq.gz, 00059-00162.225.R2.fastq.gz
Scramble_R2	scramble	00059-00163.227.R1.fastq.gz, 00059-00163.227.R2.fastq.gz
Scramble_R3	scramble	00059-00164.229.R1.fastq.gz, 00059-00164.229.R2.fastq.gz

Table 2: Sample names, sample groupings and filenames.

Reference genome

Annotation of the obtained sequences was performed using the reference annotation listed below.

Organism: Homo_sapiens
Reference genome: GRCh38.p10
Annotation reference: Ensembl_90

Experimental design

The experiments were performed using the following settings:

Instrument: NextSeq500
Average number of reads: 2x30 million reads per sample
Number of sequencing cycles (read length): 75 nt. paired end reads

Library preparation

Protocol: Illumina TruSeq Stranded mRNA Library Prep Kit
Platform: Illumina

Workflow

Figure 1 below outlines the complete workflow with QIAGEN Genomic Services for RNA Next Generation Sequencing.



Figure 1: Overview of an RNA NGS project with QIAGEN Genomic Services.

Data QC

The following sections provide a summary of the QC results obtained for your dataset. Following sequencing, intensity correction and base calling (into BCL files), FASTQ files are generated using the appropriate bcl2fastq software (Illumina Inc.) which includes quality scoring of each individual base in a read. At this stage the data is separated for paired end reads, to determine whether the second read significantly differs from the first in terms of overall quality.

Average Read Quality

An overview of the average read quality is shown in Figure 2. As illustrated, we found that the vast majority of the data has a Q-score greater than 30 (>99.9% correct), indicating that high quality data was obtained for all samples. Read pairs R1 (read1) and R2 (read2) are presented separately.

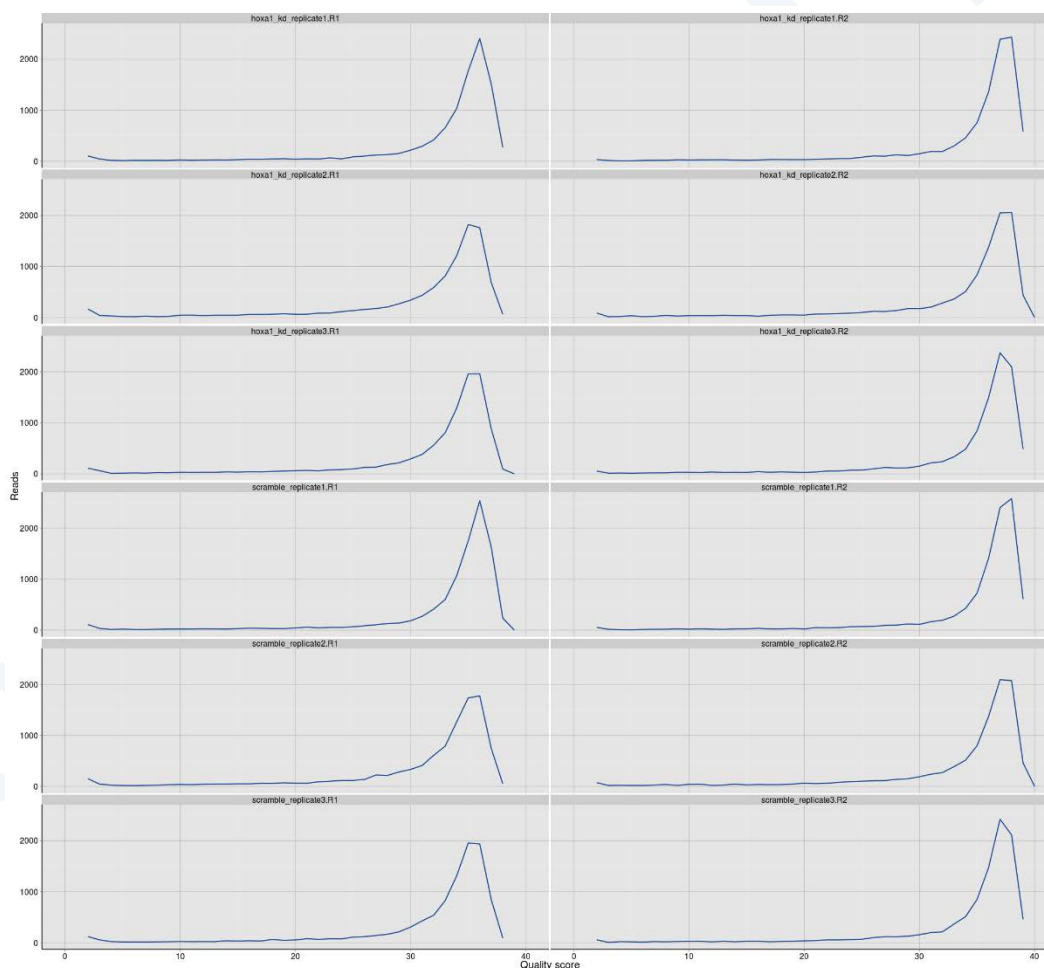


Figure 2: Average read quality of the NGS data. The average read Q-score is plotted on the x-axis and the number of reads on the y-axis. A Q-score above 30 is considered high quality data. If paired end sequencing was performed, then read pairs R1 (read1) and R2 (read2) are presented separately for each sample.

Average Base Quality

An overview of the average base quality is shown in Figure 3. As for the average read quality we found that the vast majority of the bases have a Q score greater than 30 (>99.9% correct), indicating that high quality data was obtained for all samples.

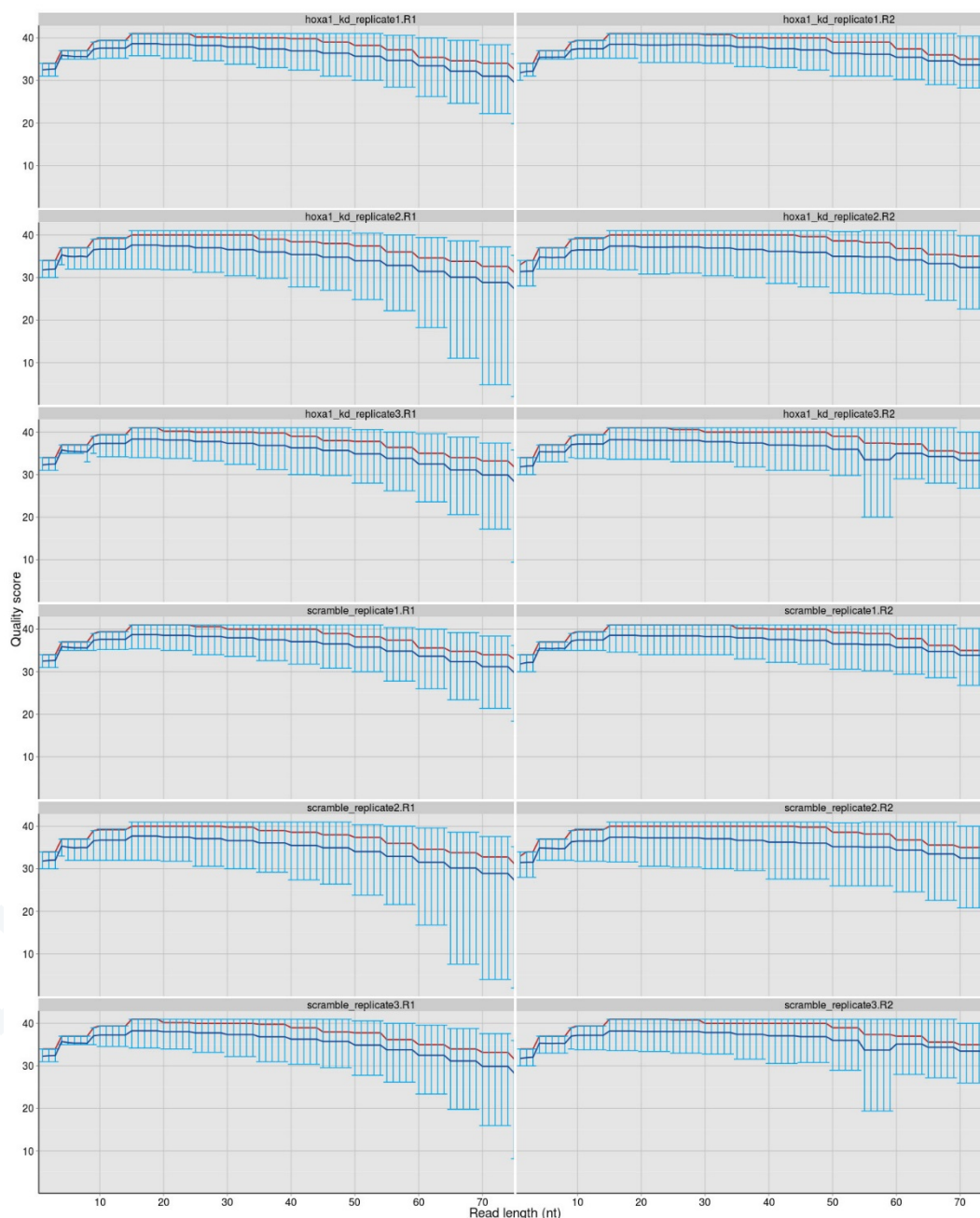


Figure 3: Base quality of the NGS data along the reads. The position in the read is plotted on the x-axis and the Q-score is plotted on the y-axis. The red line is the median value Q-score. The dark blue line is the mean value Q-score. The box-plot represents the inter-quartile range, while the whiskers represents the 10% and 90% points. A Q-score above 30 (>99.9% correct) is considered high quality data. If paired end sequencing was performed, then read pairs R1 (read1) and R2 (read2) are presented separately for each sample.

Mapping and yields

Mapping of the sequencing data is a useful quality control step in the NGS data analysis pipeline as it can help to evaluate the quality of the samples.

Reads are classified into the following classes:

Mappable reads:	Aligning to reference genome (including mRNA, pre-mRNA, poly-A tailed lncRNA and pri-miRNA)
Outmapped reads or high abundance reads:	For example; rRNA, mtRNA, poly-A and poly-C homopolymers
Unmapped reads:	No alignment possible

In a typical experiment it is possible to align 60-90% of the reads to the reference genome. However, this number depends upon multiple factors, including the quality of the sample and the coverage of the relevant reference genome; if the sample RNA was degraded, fewer reads will be mRNA or lncRNA specific and more material will be degraded rRNA.

Table 3 and Figure 4 below summarize the mapping results. In addition to the mapping results, the table below also shows the total number of reads obtained for each sample.

On average 19.3 million reads were obtained for each sample and the average genome mapping rate was 94.5%.

Sample	Reads	mtRNA	rRNA	Mapped	Unmapped
KD_R1	17,916,102	2.1	0.3	93.4	4.2
KD_R2	20,141,813	2.1	0.2	90.5	7.2
KD_R3	23,544,153	2.2	0.9	91.6	5.3
Scramble_R1	15,117,833	1.8	0.2	93.6	4.4
Scramble_R2	17,433,672	1.8	0.6	90.9	6.7
Scramble_R3	21,830,449	1.9	0.4	92.4	5.3

Table 3: Summary of the mapping results for each sample. If the alignment rate does not sum to 100% it is due to other contaminants that we have not specified in this table, e.g. PhiX.

Figure 4 summarizes the mapping results for each sample.

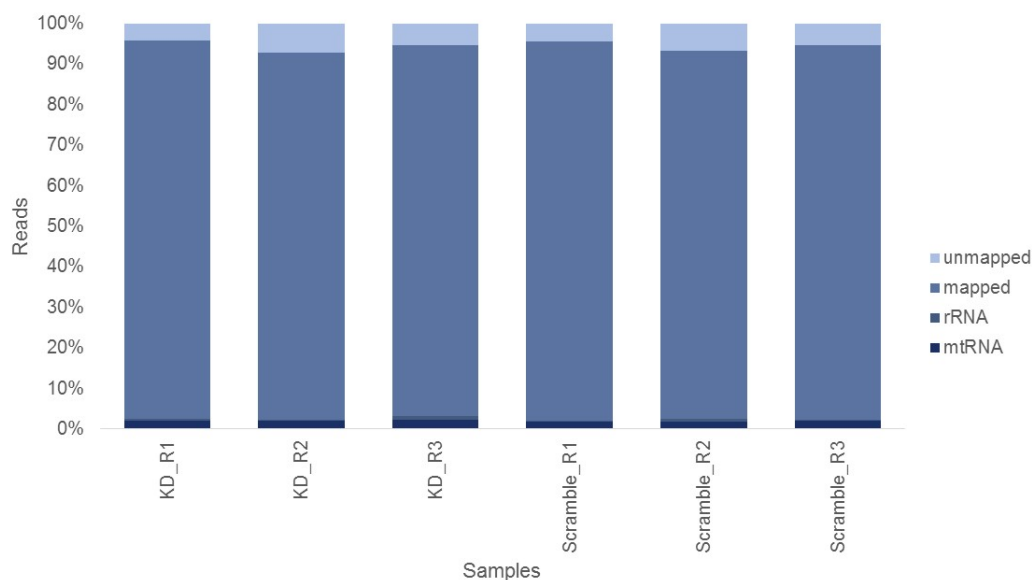


Figure 4: Summary of mapping results of the reads by sample. Each sample consists of reads that can be classified into the following categories: mapped (reads which align to reference genome), unmapped or high abundance (e.g. rRNA, miRNA and homopolymers) and reads which did not align to the reference genome (unmapped).

If you wish to inspect the mapping in detail, please download the BAM alignment files from My Projects at [XploreRNA](#). The detected deletions, insertions, and splice junctions are also available in the Mapping Results folder. Go to “Explore Results” and click the “download all files” link. In the root folder you will find a full description of all files provided as well as recommended programs to view the different files.

The BAM files can be viewed and inspected in any standard genome viewer such as the [Integrative Genomics Viewer](#) (Robinson *et al.*, 2011) and (Thorvaldsdóttir *et al.*, 2013) downloadable from [Broad Institute](#).

Results

Below you will find a summary of the principal findings for this project. The complete analysis may be found in the associated files listed on page 4. For a detailed description of the data analysis process, see the Data analysis workflow section on page 25.

Identified genes

The number of identified genes per sample was calculated based on alignment to the reference genome. When performing statistical comparisons between groups, we include all genes irrespective of abundance.

Ideally, all samples in the study should have similar call rates (similar numbers of genes identified), in order to be comparable.

Sample name	Number of Genes identified	Number of Isoforms identified
KD_R1	13,666	44,361
KD_R2	13,692	44,514
KD_R3	13,635	44,656
Scramble_R1	13,465	44,131
Scramble_R2	13,491	44,107
Scramble_R3	13,581	44,619

Table 4: Number of genes and isoforms identified in each sample which have a fragment count (FPKM) estimation of at least 10 counts per gene or isoform.

The number of genes identified for each sample based on different fragment count cut-off values is illustrated in the radar plot in Figure 5. The sample name is indicated on the outer rim of the plot. The number of genes identified which have a fragment count estimation of at least 1, 10, 100 or 1000 counts per gene are illustrated as colored rings. If one sample results in a significantly lower number of genes identified at each fragment count cut-off, this is an indication that the sample is deviating from the remaining samples. Ideally, comparable samples should show similar number of genes identified at each fragment count cut-off, resulting in concentric rings of different colors.

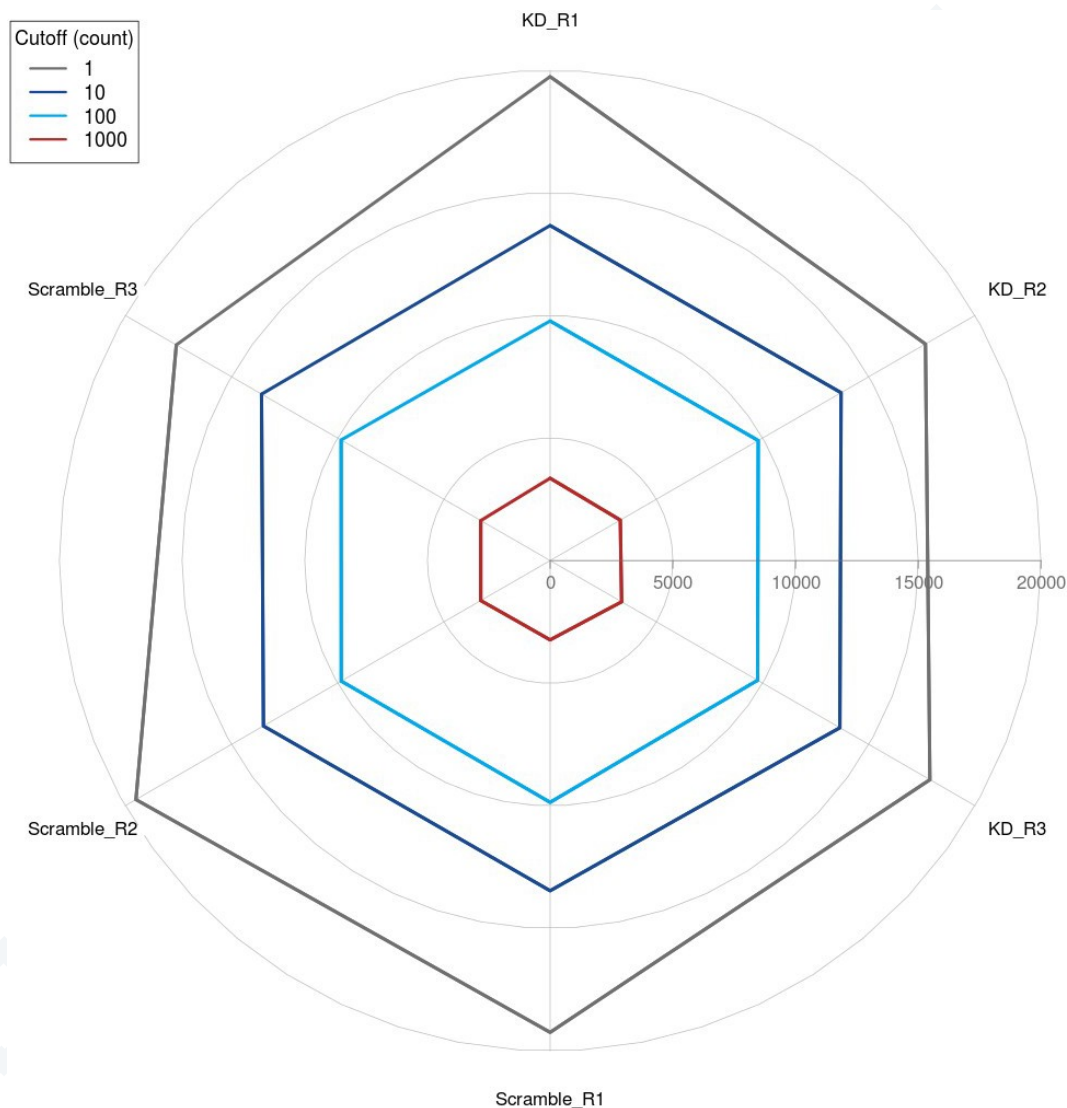


Figure 5: Radar plot showing number of genes identified for each sample at different fragment count cut-off values. See color scale at top of figure for specification of fragment count cut-off values.

FPKM is a unit of measuring gene expression used for NGS experiments. The number of reads corresponding to the particular gene is normalized to the length of the gene and the total number of mapped reads (Fragments Per Kilobase of transcript per Million mapped reads). In the analysis part the FPKM values are normalized with median of the geometric mean (Anders & Huber, 2010).

Principal Component Analysis – scramble and knockdown

Principal Component Analysis (PCA) is a method used in unsupervised analysis to reduce the dimension of large data sets and is a useful tool to explore sample classes arising naturally based on the expression profile.

The 500 genes that have the largest coefficient of variation based on FPKM abundance estimations have been included in the analysis. Figure 6 below represents an overview of how the samples cluster.

If the biological differences between the samples are pronounced, this will describe the primary components of the variation in the data. This leads to separation of samples in different regions of a PCA plot corresponding to their biology. However, if other factors, e.g. sample quality, introduce more variation in the data, the samples may not cluster according to the biology.

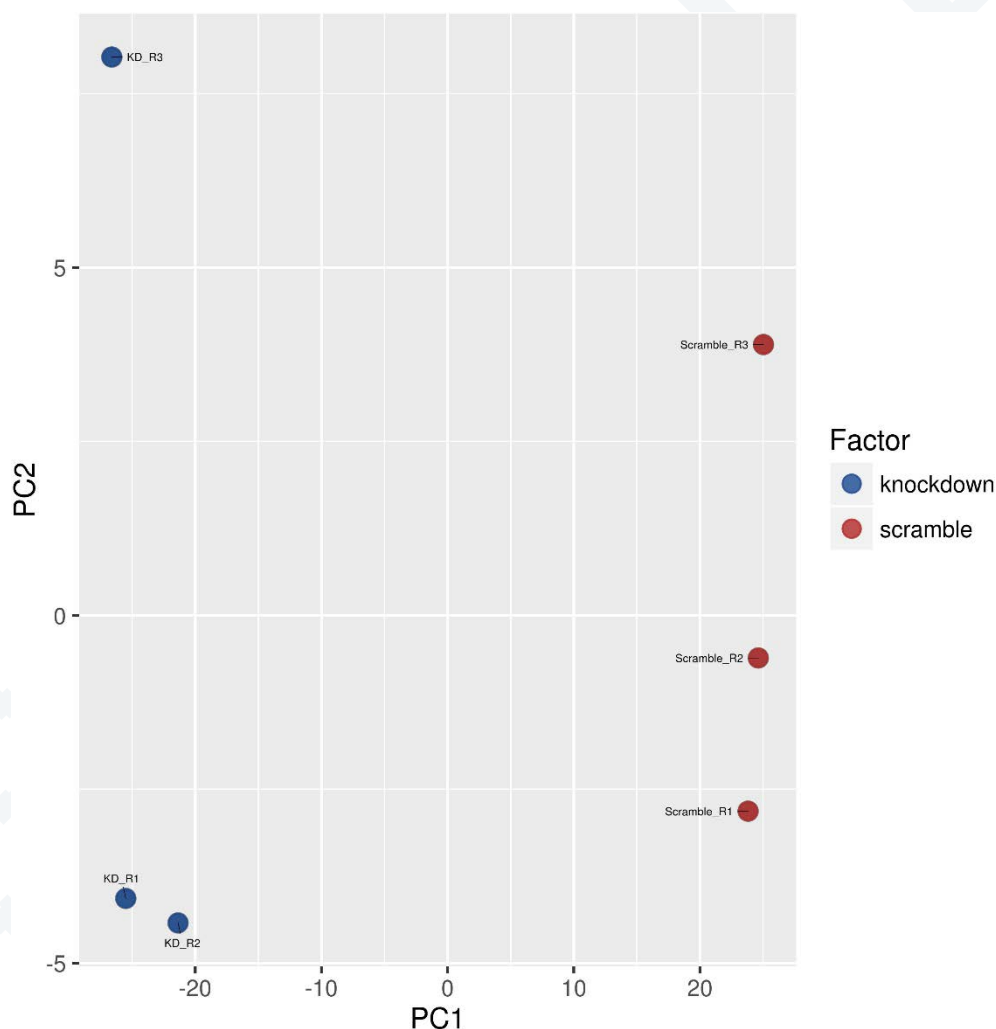


Figure 6: Principal component analysis (PCA) plot for [GROUPS]. The PCA was performed on all samples passing QC using the 500 genes that have the largest coefficient of variation based on FPKM counts. The largest component in the variation is plotted along the X-axis and the second largest is plotted on the Y-axis. Each circle represents a sample. Based on Normalized FPKM (abundance) for each gene for each sample gene. FPKM table can be downloaded from My Projects at [XploreRNA](#).

Heat map and unsupervised clustering – scramble and knockdown

The heat map diagram below shows the result of the two-way hierarchical clustering of genes and samples. It includes the 500 genes that have the largest coefficient of variation based on FPKM counts. Each row represents one gene and each column represents one sample. The color represents the relative expression level of a transcript across all samples. The color scale is shown below: red represents an expression level above the mean; green represents an expression level below the mean.

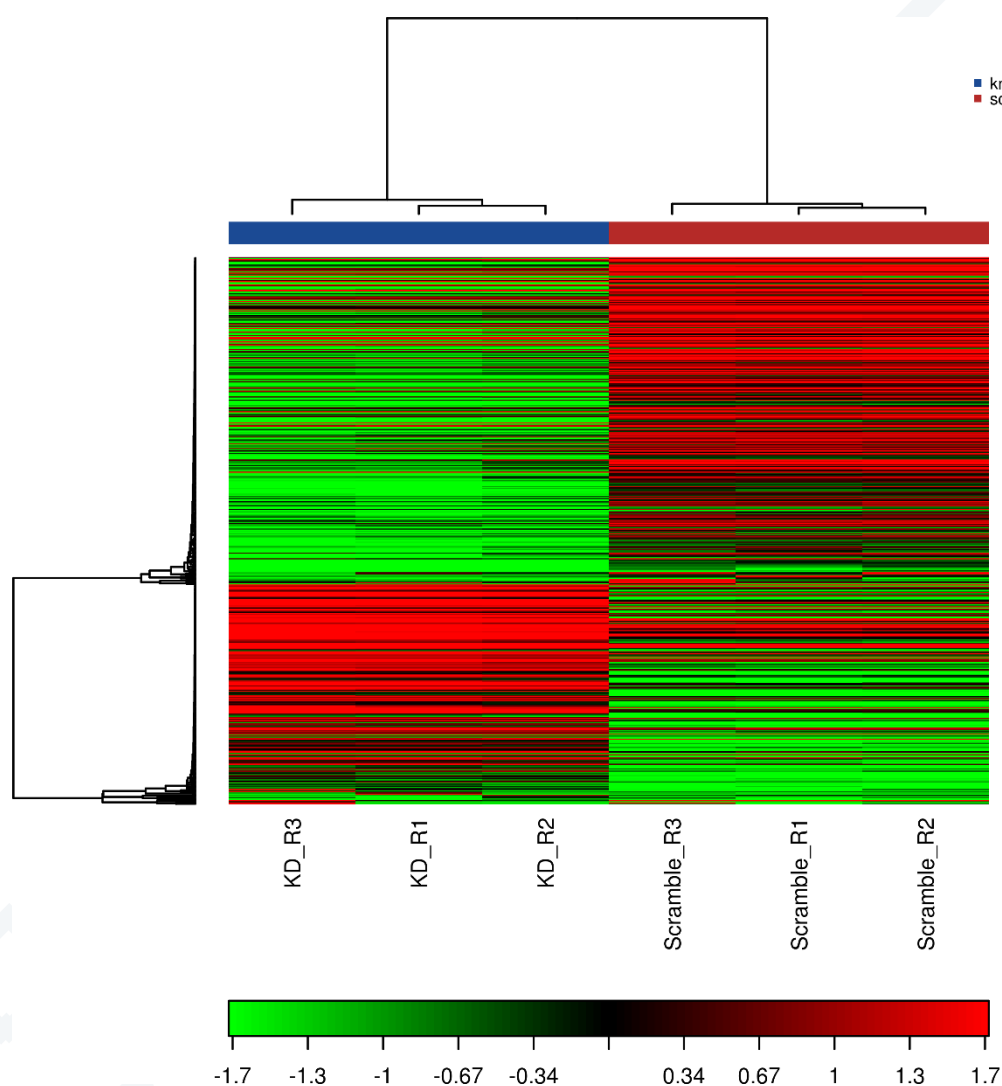


Figure 7: Heat map and unsupervised hierarchical clustering by sample and genes were performed on the listed samples using the 500 genes that have the largest coefficient of variation based on FPKM counts. Data is based on samples from the [GROUPS] groups. Based on Normalized FPKM (abundance) for each gene for each sample (raw data in genes). FPKM table can be downloaded from My Projects at [XploreRNA](#).

Identification of novel transcripts

During the transcriptome assembly process, both known and novel transcripts are identified. A novel transcript is characterized as a transcript which contains features not present in the reference annotation. Thus, a novel transcript can be both a new isoform of a known gene or a transcript without any known features. For example, a novel transcript could be the result of a previously unknown splicing event for a known gene or a previously unknown long non-coding RNA.

Identification of novel transcripts depends upon the reference annotation. Transcripts not part of the reference genome used for annotation will be classified as novel. In the results files we classify novel transcripts with known features by listing the known transcripts most closely resembling the novel transcript. For novel transcripts without any known features we provide a locally unique name as transcript identifier. In addition, we provide the genomic positions for the features of the novel transcript, e.g. the location and number of exons.

lncRNA identification and annotation is at a preliminary stage (Mattick *et al.*, 2015). Therefore we recommend referring to [RNACentral](#) or [GENCODE](#) to investigate potentially novel lncRNAs identified in your study.

A list of differentially expressed novel isoforms can be found in this report, and the full list of differentially expressed isoforms are available in My Projects at [XploreRNA](#). Go to "Explore Results" and click the "download all files" link. In the root folder you will find a full description of all files provided, and recommended programs to view the different files.

The table annotations are complex but a good reference is presented in the [Cufflinks manual](#).

Differentially expressed genes, supervised analysis

To identify differentially expressed genes, it is assumed that the number of reads produced by each transcript is proportional to both its size and abundance. QIAGEN Genomic Services has customized the analysis pipeline based on the Tuxedo suite, including the Cufflinks, Cuffmerge and Cuffdiff steps of the Tuxedo pipeline. For more details, see the Data analysis workflow on page 25.

Comparison of scramble and knockdown – genes

Table 5 below shows the individual results for the top 20 most significantly differentially expressed genes. A full list of differentially expressed genes is available as a .tsv file in My Projects at [XploreRNA](#).

Gene ID	Gene	Locus	Scramble FPKM	Knockdown FPKM	Log2 Fold Change	Q_value
XLOC_037955	AREG	4:74445133-74455009	2.39	438.9	7.52	0.000190811
XLOC_004506	PTGS2	1:186671790-186680427	0.77	43.62	5.83	0.000190811
XLOC_031249	BMP2	20:6767663-6780280	0.86	40.01	5.54	0.000190811
XLOC_049252	ANGPT2	8:6403550-6708209	0.03	1.38	5.37	0.000190811
XLOC_003928	AC241585.2	1:145164098-145216445	30.24	0.84	-5.17	0.000190811
XLOC_007664	C11orf96	11:43921058-44001157	0.58	15.09	4.71	0.000190811
XLOC_012586	SLC6A15	12:84859487-84913615	3.92	101.76	4.7	0.000190811
XLOC_014132	EDNRB	13:77895480-78659179	0.16	3.98	4.68	0.000190811
XLOC_014099	DACH1	13:71437965-71867192	0.25	6.49	4.67	0.000190811
XLOC_033151	CBR3-AS1	21:36131766-36294448	1.07	0.04	-4.66	0.000190811
XLOC_046342	KCND2	7:120273667-120752514	0.1	2.34	4.53	0.000190811
XLOC_030227	IL1B	2:112829750-112836903	1.08	22.67	4.39	0.000190811
XLOC_013943	VWA8	13:41566836-41981565	5.4	109.84	4.35	0.000190811
XLOC_008459	GRIA4	11:105609993-105982092	0.65	12.55	4.28	0.000190811
XLOC_037853	ADGRL3	4:61201257-62165554	0.49	9.41	4.28	0.000190811
XLOC_025146	GDF15	19:18374730-18389176	26.68	511.24	4.26	0.000190811
XLOC_030504	NR4A2	2:156324431-156342348	1.25	23.87	4.26	0.000190811
XLOC_007468	INSC	11:15112423-15247208	0.95	17.81	4.23	0.000190811
XLOC_008645	VWA5A	11:124115361-124147721	1.96	35.93	4.19	0.000190811
XLOC_017707	BMF	15:40087889-40108892	1.38	24.25	4.14	0.000190811

Table 5: Genes: Table of the 20 most significantly differentially expressed genes. Scramble and knockdown columns are group average FPKM values. Transcripts with the highest fold change between groups are shown at the top of the table. Fold change is the log2 fold change of the FPKM between groups scramble and knockdown. Q-values shown are p-values that have been adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) approach to correct for multiple testing. As a general guide, fold changes with q-values below 0.05 may be considered significant. The full list of differentially expressed genes is available as a .tsv file in My Projects at [XploreRNA](#).

Comparison of scramble and knockdown – isoforms

Table 6 below shows the individual results for the top 20 most significantly differentially expressed isoforms. A full list of differentially expressed novel transcripts is available as a .tsv file in My Projects at [XploreRNA](#).

Isoform ID	Gene	Locus	Scramble FPKM	Knockdown FPKM	Log2 Fold Change	Q_value
XLOC_037955	AREG	4:74445133-74455009	2.39	438.2	7.52	0.0006618
XLOC_004506	PTGS2	1:186671790-186680427	0.37	34.33	6.52	0.0006618
XLOC_008459	GRIA4	11:105609993-105982092	0.04	3.16	6.32	0.00488304
XLOC_012586	SLC6A15	12:84859487-84913615	0.71	38.08	5.74	0.0006618
XLOC_041840	MTX3	5:79976730-79991262	1.11	0.02	-5.7	0.0382217
XLOC_031249	BMP2	20:6767663-6780280	0.86	40.01	5.54	0.0006618
XLOC_002687	PLEKHG5	1:6424787-6520061	0.05	2.34	5.54	0.0418422
XLOC_007664	C11orf96	11:43921058-44001157	0.37	14.93	5.34	0.0006618
XLOC_019503	HSD17B2	16:82035003-82139631	0.28	10.34	5.22	0.0006618
XLOC_014099	DACH1	13:71437965-71867192	0.11	4.21	5.22	0.0006618
XLOC_046342	KCND2	7:120273667-120752514	0.06	1.83	4.99	0.0006618
XLOC_012586	SLC6A15	12:84859487-84913615	1.27	38.5	4.92	0.0006618
XLOC_008645	VWA5A	11:124115361-124147721	0.9	25.32	4.81	0.0006618
XLOC_014132	EDNRB	13:77895480-78659179	0.08	2.19	4.78	0.0006618
XLOC_032355	RIPOR3	20:50586107-50691546	0.37	9.63	4.69	0.0006618
XLOC_030227	IL1B	2:112829750-112836903	0.91	22.19	4.61	0.0006618
XLOC_000661	CDC20	1:43358954-43363203	4.94	0.21	-4.54	0.0498505
XLOC_039890	SLC9A3-AS1	5:473195-524332	0.15	3.3	4.42	0.0382217
XLOC_020870	PIMREG	17:6444414-6556494	6.18	0.29	-4.39	0.0006618
XLOC_014823	CDKN3	14:54396848-54420218	9.21	0.44	-4.38	0.0128823

Table 6: Isoforms: Table of the 20 most significantly differentially expressed isoforms (known and novel). Scramble and knockdown columns are group average FPKM values. Isoforms with the highest fold change between groups are shown at the top of the table. Fold change is the log2 fold change of the FPKM between groups scramble and knockdown. Q-values shown are p-values that have been adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) approach to correct for multiple testing. As a general guide, fold changes with q-values below 0.05 may be considered significant. The full list of differentially expressed isoforms is available as a .tsv file in My Projects at [XploreRNA](#).

Comparison of scramble and knockdown – novel isoforms

Table 7 below lists the top 20 most significantly differentially expressed novel isoforms identified in this project. In the second column in the table below, the known transcripts most closely resembling the novel transcript are listed. A full list of differentially expressed novel transcripts is available as a .tsv file in My Projects at [XploreRNA](#).

Isoform ID	Closest Known Transcript	Locus	Scramble FPKM	Knockdown FPKM	Log2 Fold Change	Q_value
XLOC_039563	RNF150	4:140859806-141212877	0.76	0.04	-4.14	0.0153222
XLOC_001316	HSD3BP5	1:119597701-119614892	0.06	0.89	3.85	0.0006618
XLOC_008966	NLRP10	11:7957549-7965469	1.2	0.09	-3.81	0.0006618
XLOC_035702	CLDN11	3:170418864-170860380	15.41	1.17	-3.72	0.0006618
XLOC_018101	ADPGK	15:72751366-72798199	8.71	0.7	-3.64	0.00125174
XLOC_016773	DLL4	15:40929339-40939074	1.26	15.22	3.59	0.0006618
XLOC_010820	ACVRL1	12:51906907-51923458	0.31	3.42	3.48	0.0006618
XLOC_033248	ICOSLG	21:44222990-44240966	0.12	1.33	3.43	0.0006618
XLOC_013903	POSTN	13:37562581-37598844	9.82	104.58	3.41	0.0006618
XLOC_020389	CSNK2A2	16:58129528-58217805	5.92	0.56	-3.41	0.0006618
XLOC_047901	DGKI	7:137380883-137847092	0.92	9.75	3.4	0.0006618
XLOC_049852	TRPA1	8:71828166-72118393	7.21	76.07	3.4	0.0006618
XLOC_028495	KYNU	2:142877497-143055832	0.38	3.98	3.39	0.0006618
XLOC_020670	ZDHHC7	16:84974180-85011535	4.63	0.51	-3.18	0.0355261
XLOC_002625	SLC35E2B	1:1659324-1692728	0.23	1.96	3.07	0.0033916
XLOC_032330	SULF2	20:47501808-47786616	0.39	3.1	2.98	0.0424642
XLOC_010094	PCSK7	11:117199320-117232525	66.26	9.14	-2.86	0.0006618
XLOC_030575	SPC25	2:168455861-168913371	2.91	0.43	-2.75	0.0193018
XLOC_044615	PTCHD4	6:47877458-48111181	0.33	2.18	2.73	0.0402038
XLOC_019676	UNKL	16:1351922-1414751	0.41	2.69	2.71	0.0006618

Table 7: Novel isoforms: Table of the 20 most significantly differentially expressed novel isoforms. Scramble and knockdown columns are group average FPKM values. Novel transcripts with the highest fold change between groups are shown at the top of the table. Fold change is the log2 fold change of the FPKM between groups scramble and knockdown. Q-values shown are p-values that have been adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) approach to correct for multiple testing. As a general guide, fold changes with q-values below 0.05 may be considered significant. The full list of differentially expressed novel isoforms is available as a .tsv file in My Projects at [XploreRNA](#).

Volcano Plot – introduction

The Volcano plot provides a way to perform a quick visual identification of the genes displaying large-magnitude changes which are also statistically significant. The plot is constructed by plotting $-\log_{10}(\text{p-value})$ on the y-axis, and the expression fold change between the two experimental groups on the x-axis. There are two regions of interest in the plot: those points that are found towards the top of the plot (high statistical significance) and at the extreme left or right (strongly down and up-regulated respectively).

Volcano Plot – scramble and knockdown

Genes that pass the filtering of $q\text{-value} < 0.05$ are indicated on the plot below (red). For the present study, 6699 genes pass this filtering.

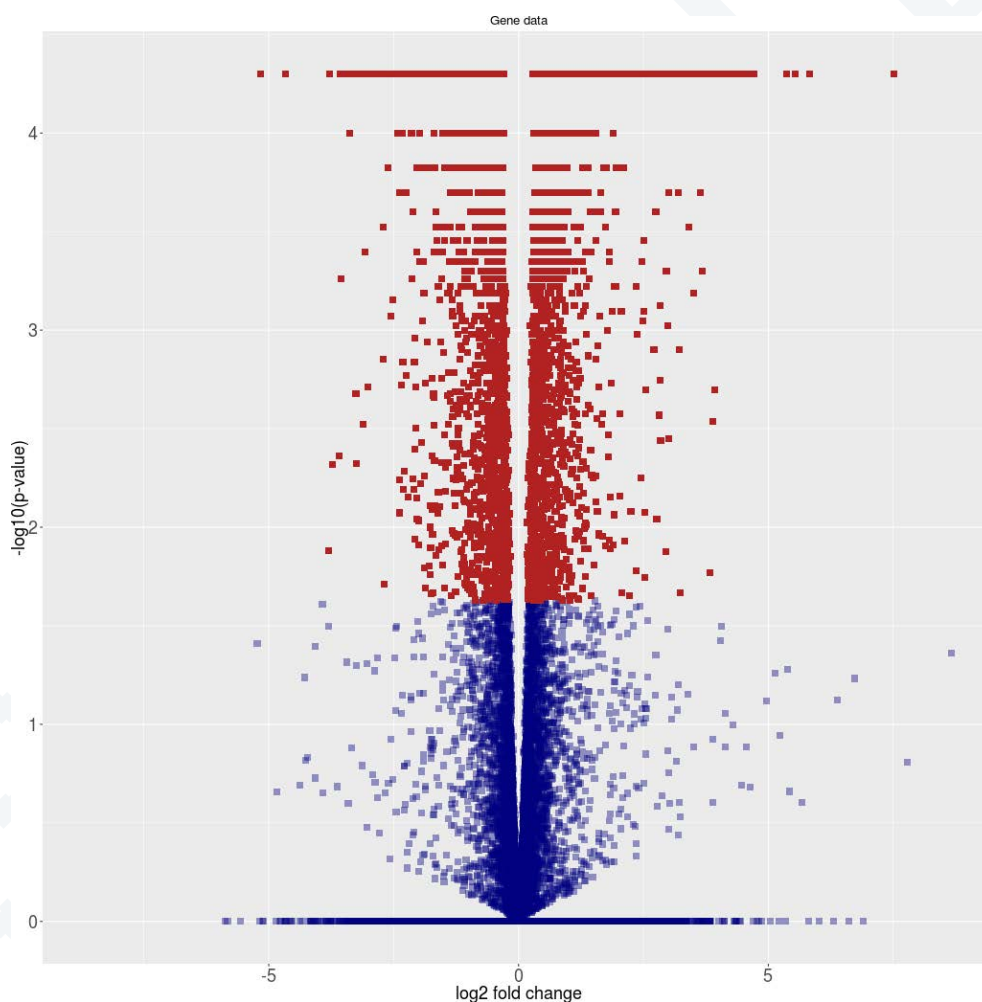


Figure 8: Volcano plot showing the relationship between the raw p-values and the log2 fold change in normalized expression (FPKM) between scramble and knockdown. Data is based on Normalized FPKM (abundance) for each gene for each sample (gene_exp.diff available in My Projects at [XploreRNA](#)).

Gene Ontology Enrichment Analysis – introduction

Gene Ontology (GO – Gene Ontology Consortium, 2000) is an initiative to describe genes, gene products and their attributes using vocabulary (GO terms) which is unified and controlled across all species. This enables functional interpretation of experimental data using GO terms, for example via enrichment analysis.

We use GO enrichment analysis to investigate whether specific GO terms are more likely to be associated with the differentially expressed transcripts. Two different statistical tests are used and compared. Firstly, a standard Fisher's test is used to investigate enrichment of terms between the two groups. Secondly, the 'Elim' method takes a more conservative approach by incorporating the topology of the GO network to compensate for local dependencies between GO which can mask significant GO terms. Comparisons of the predictions from these two methods can highlight truly relevant GO terms.

Gene Ontology Enrichment Analysis – knockdown vs scramble

The figure below shows a comparison of the results for the GO (Biological process) terms associated with the significantly differentially expressed transcripts that were identified between the two groups. Complete GO enrichment analysis for all of the comparisons, including Cellular component (CC) and Molecular functions (MF) analysis, is available in My Projects at [XploreRNA](#).

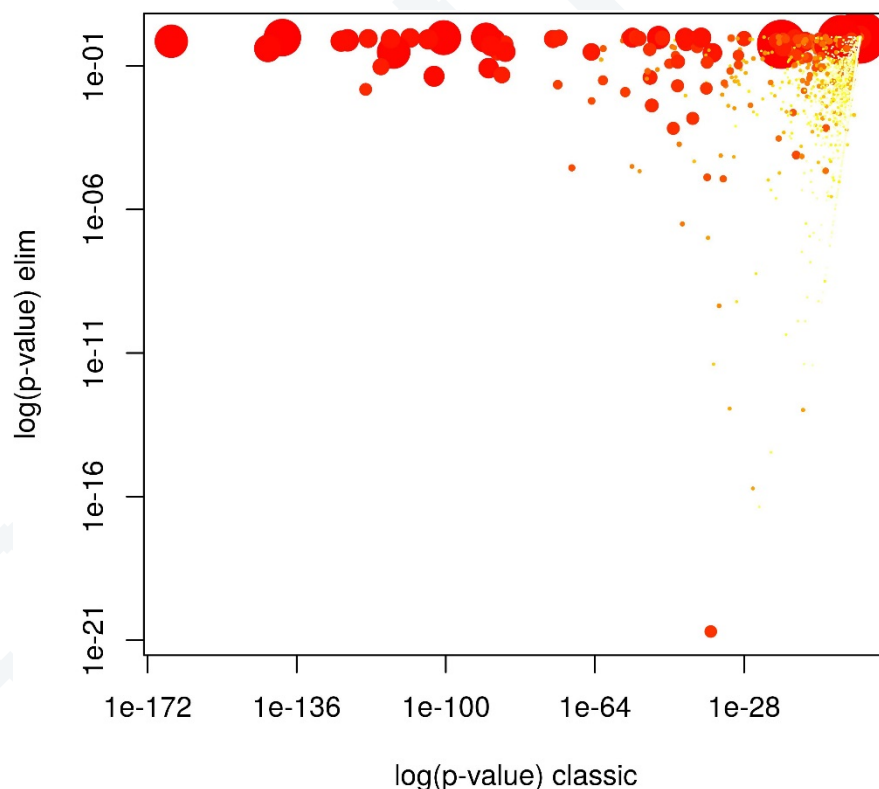


Figure 9: Scatter plot for significantly enriched GO terms associated with genes differentially expressed between scramble and knockdown. Plot shows a comparison of the results obtained by the two statistical tests used. Values along diagonal are consistent between both methods. Values in the bottom left of the plot correspond to the terms with most reliable estimates from both methods. Size of dot is proportional to number of genes mapping to that GO term and coloring represents number of significantly differentially expressed transcripts corresponding to that term with dark red representing more terms and yellow representing fewer.

Table 8 below lists the top 20 most significant GO (Biological process) terms in this project. A full list of GO (Biological process) terms is available through My Projects at [XploreRNA](#).

GO_ID	Term	Annotated	Significant	Expected	P-value
GO:0006355	regulation of transcription, DNA-templated	3335	1279	1135.97	2.0e-21
GO:0006364	rRNA processing	234	93	79.7	4.4e-17
GO:0051301	cell division	657	347	223.79	1.9e-16
GO:0000398	mRNA splicing, via spliceosome	291	109	99.12	3.5e-15
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	760	343	258.87	1.0e-13
GO:0016032	viral process	702	323	239.11	1.2e-13
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	109	51	37.13	3.8e-12
GO:0006281	DNA repair	483	235	164.52	4.1e-12
GO:0007062	sister chromatid cohesion	125	81	42.58	4.1e-12
GO:0000209	protein polyubiquitination	268	123	91.29	4.4e-11
GO:0019083	viral transcription	159	79	54.16	1.3e-10
GO:0010389	regulation of G2/M transition of mitotic cell cycle	176	97	59.95	1.3e-10
GO:0006886	intracellular protein transport	944	423	321.54	4.4e-10
GO:0070126	mitochondrial translational termination	77	26	26.23	5.8e-10
GO:0043161	proteasome-mediated ubiquitindependent protein catabolic process	395	189	134.54	6.1e-10
GO:1901796	regulation of signal transduction by p53 class mediator	158	78	53.82	6.3e-10
GO:0006413	translational initiation	182	81	61.99	8.7e-10
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	82	36	27.93	1.3e-09
GO:0070125	mitochondrial translational elongation	74	25	25.21	1.3e-09
GO:0001701	in utero embryonic development	314	149	106.95	3.2e-09

Table 8: The top 20 significant GO (Biological Process) terms associated with transcripts found to be differentially expressed between scramble and knockdown. This table aims to highlight the most relevant GO terms associated with the differentially expressed transcripts in your comparison to see if certain biological functions are enriched among these transcripts compared to the reference background.

The GO term analysis is a type of gene enrichment test and does not ensure that the transcripts that belong to a significant GO term are up or down regulated. It ensures however that a group of differentially expressed genes with similar functionality are significantly over-represented. The Expected values represent an estimate of the number of transcripts associated with the given GO term that are significant by random among all the annotated differentially expressed genes. The Significant (Observed) values represents the number of differentially expressed transcripts associated with that particular GO term in the sample (real) dataset. Annotations represents the total number of genes associated with that GO term in the sample dataset, which means that the reference background could potentially have higher number of annotations. The q value represents the test statistics for the given GO term whether it is significantly enriched or not.

To illustrate how the different GO terms are linked, a GO network has been created. The network is shown in Figure 10 below and networks of varying complexity are available through My Projects at [XploreRNA](#).

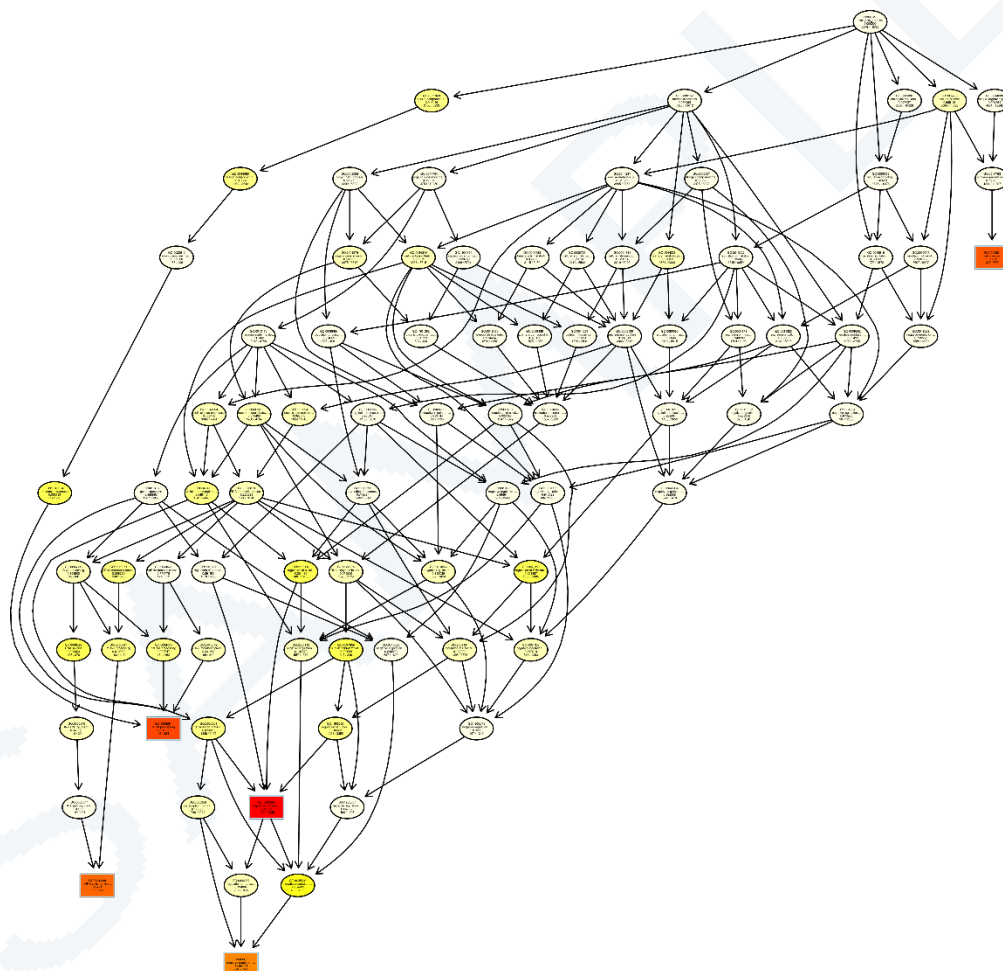


Figure 10: GO network generated for the enriched GO terms (Biological Process) associated with genes differentially expressed between scramble and knockdown. Nodes are colored from red to yellow with the node with the strongest support colored red and nodes with no significant enrichment colored yellow. The five nodes with strongest support are marked with rectangular nodes. A high-resolution version of this figure is available in My Projects at [XploreRNA](#).

Conclusion and next steps

The analysis of your RNA Next Generation Sequencing data has been completed.

mRNA Next Generation Sequencing libraries were successfully prepared, quantified and sequenced for all your samples. The profiling of your samples was successfully completed. The data passed all QC metrics; the NGS data had high Q-Score, indicating good technical performance of the NGS experiment.

A large number of novel transcripts were identified. Note, however, that many of these will be novel isoforms or start sites of known genes and transcripts.

It is clear from the unsupervised analysis that the six samples cluster according to their biological groups, indicating that the sample groups are causing the largest variation on the samples.

The supervised analysis showed large numbers of significantly differentially expressed mRNA at the CDS (Coding DNA Sequence) and gene level as well as at the isomer level. Note: when navigating through these data, counts lower than 1-5 FPKM (on average) per group might be difficult to validate in a qPCR experiment.

As part of your service project with QIAGEN Genomic Services, we are now ready to setup a teleconference with our service scientists to go through your data, assist with interpretation of your results and discuss next steps such as how to validate your findings.

Please contact your local QIAGEN representative or the lab at Genomic.Services@qiagen.com, so we can schedule the teleconference.

We are also happy suggest an appropriate validation scheme for your results. QIAGEN offers a wide range tools for validating potentially regulated mRNAs, microRNAs, lncRNAs as well as other non-coding RNAs by qPCR, *in situ* hybridization, Northern blot, or gene silencing.

Using the Gene Sorting Wizard to select candidates for downstream validation

The easiest way to select your candidates for validation and further study is using the Gene Sorting Wizard in My Projects at [XploreRNA](#). There you can explore each group comparison and sort the full list of differentially expressed genes using all relevant criteria, including expression level, fold change, and statistical significance. Below are some general considerations when selecting candidates for validation and further study.

- **Fold change.** Smaller fold changes tend to be more affected by technical variance, and hence may be at greater risk of false-positive signals. To study transcripts with small fold changes, considerably more technical or biological replicates should be included in the validation.
- **Expression level.** When navigating through these data, counts lower than 1-5 FPKM (on average) per group might be difficult to validate in a qPCR experiment.
- **Statistical significance.** Fold changes with adjusted p-values (q-values) below 0.05 may be considered significant, and have a greater chance of being validated by qPCR.
- **Novel transcripts.** Many of the novel transcripts identified may be novel isoforms of known transcripts, or alternative start sites of known transcripts. Potentially novel lncRNAs identified may be further investigated using specialised lncRNA resources e.g. [RNACentral](#) or [GENCODE](#).
- **Reference genes.** NGS data can also be used to identify stably expressed transcripts that may be used as reference genes for normalization in qPCR validation experiments. Genes that are stably expressed across all samples can be identified using e.g. NormFinder or geNorm (Vandesompele *et al.*, 2002).

Data analysis workflow

Figure 11 below outlines the QIAGEN Genomic Services data analysis pipeline for mRNA Next Generation Sequencing.



Figure 11: Overview of the mRNA NGS data analysis pipeline.

Software tools used for the analysis

Our NGS data analysis pipeline is based on the Tuxedo software package, which is a combination of open-source software, and implements peer-reviewed statistical methods. In addition we employ specialized software developed internally at QIAGEN Genomic Services to interpret and improve the readability of the final results.

The components of our NGS data analysis pipeline for RNA-seq include Bowtie2 (v. 2.2.2, see Langmead B and Salzberg S. (2012)), Tophat (v2.0.11, see Trapnell, C., *et al.* (2009)) and Cufflinks (v2.2.1, see Trapnell, C., *et al.* (2010) and Trapnell, C., *et al.* (2012)), and are described in detail below. See Figure 12 for an illustrated overview of the analysis pipeline.

Tophat is a fast splice junction mapper for RNA-seq reads. It aligns the sequencing reads to the reference genome using the sequence aligner Bowtie2. Tophat also uses the sequence alignments to identify splice junctions for both known and novel transcripts as well as identification of insertions and deletions.

Cufflinks takes the alignment results from Tophat and assembles the aligned sequences into transcripts, thereby constructing a map or a snapshot of the transcriptome. To guide the assembly process, an existing transcript annotation is used (RABT assembly). In addition, we perform fragment bias correction which seeks to correct for sequence bias during library preparation (see Kasper *et al.*, 2010 and Adam *et al.*, 2011). The Cufflinks assembles aligned reads into different transcript isoforms based on exon usage and also determines the transcriptional start sites (TSSs).

When comparing groups, Cuffdiff is used to calculate the FPKM (number of fragments per kilobase of transcript per million mapped fragments) and test for differential expression and regulation among the assembled transcripts across the submitted samples using the Cufflinks output. Cuffdiff can be used to test differential expression at different levels, from CDS and gene specific, down to the isoform and TSS transcript level. For more information on the Cuffdiff module, see Trapnell *et al.*, (2013).

As a final step custom software is used for post processing of Cuffnorm and Cuffdiff results. We use these tools to generate a visual representation of your sequencing results to aid the interpretation of the sequencing data and the analysis results.

mRNA / whole transcriptome workflow

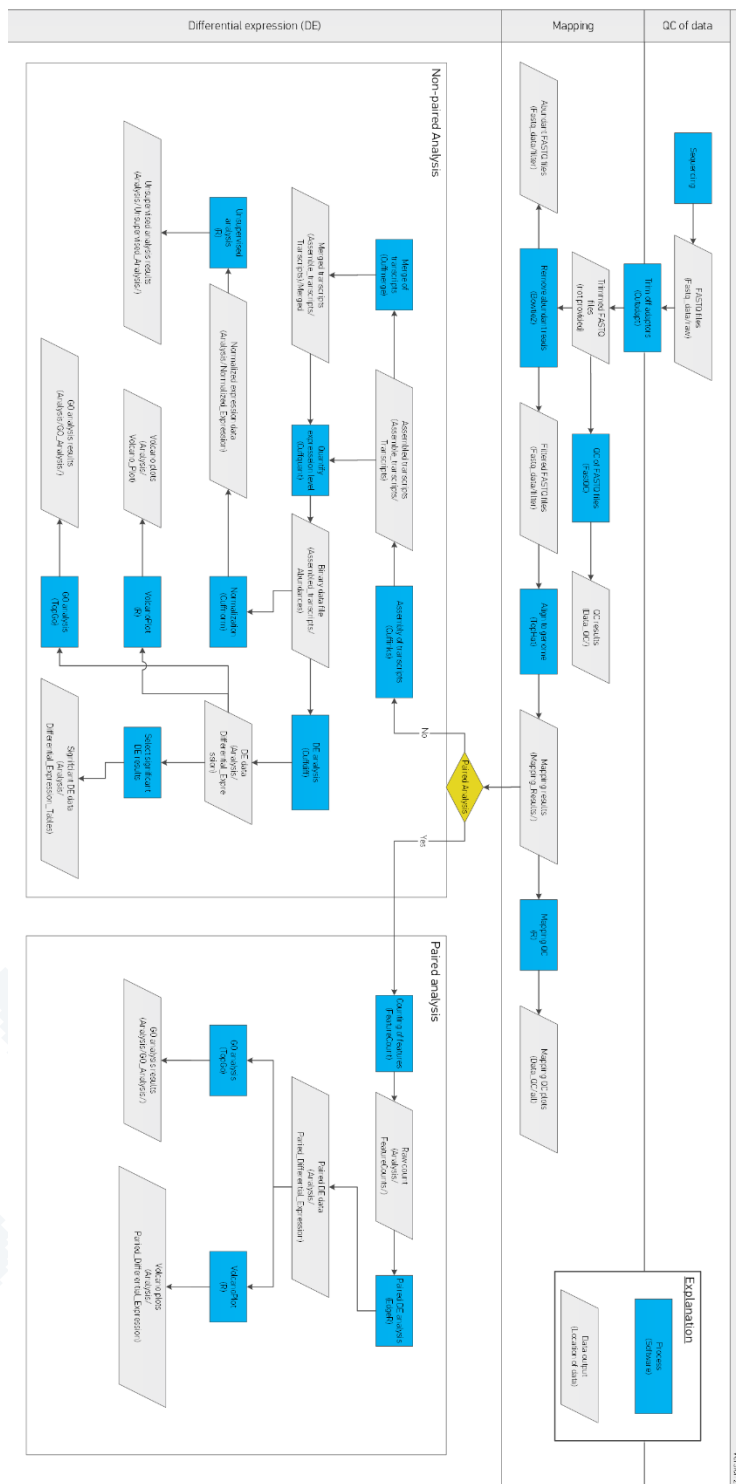


Figure 12: Overview of the analysis pipeline for mRNA and whole transcriptome projects. Blue square boxes indicates a process, the software/tool performing the process is specified in parenthesis. Grey parallelogram boxes indicates a data output and the output location is specified in parenthesis. Note that the workflow is different depending on if the samples are paired. Refer to the summary report (including references) for further explanation of each step. Also refer to the file_description.html file for details about the data outputs, how to browse the files and how to interpret the results.

Materials and methods

All experiments were conducted by QIAGEN Genomic Services.

Library preparation and Next Generation Sequencing

The library preparation was done using TruSeq® Stranded mRNA Sample preparation kit (Illumina Inc.).

The starting material (500 ng) of total RNA was mRNA enriched using the oligodT bead system. The isolated mRNA was subsequently fragmented using enzymatic fragmentation. Then first strand synthesis and second strand synthesis were performed and the double stranded cDNA was purified (AMPure XP, Beckman Coulter).

The cDNA was end repaired, 3' adenylated and Illumina sequencing adaptors ligated onto the fragments ends, and the library was purified (AMPure XP). The mRNA stranded libraries were pre-amplified with PCR and purified (AMPure XP). The libraries size distribution was validated and quality inspected on a Bioanalyzer 2100 or BioAnalyzer 4200 tapeStation (Agilent Technologies).

High quality libraries were pooled in equimolar concentrations based on the Bioanalyzer Smear Analysis tool (Agilent Technologies). The library pool(s) were quantified using qPCR and optimal concentration of the library pool used to generate the clusters on the surface of a flowcell before sequencing on a NextSeq500 instrument (75 cycles) according to the manufacturer instructions (Illumina Inc.).

References

- Anders S. and Huber W. (2010) Differential expression analysis for sequence count data. *Genome Biology* 11: R106.
- Benjamini and Hochberg (1995). *Journal of the Royal Statistical Society Series B*, *57*, 289-300.
- Kasper D., *et al.*, (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming *Nucleic Acids Research*, Volume 38, Issue 12.
- Kellis, M., *et al.*, (2013) Defining functional DNA elements in the human genome. *PNAS*, Vol. 111:6131-6138.
- Langmead B, Salzberg S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357-359.
- Marinov, G. K., *et al.*, (2014) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res*. 24: 496-510.
- Roberts, A., *et al.*, (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17): 2325-2329.
- Roberts, A., *et al.*, (2011) Improving RNA-Seq expression estimates by correcting for fragment bias *Genome Biology*, Volume 12, R22.
- Robinson, J.T., *et al.*, (2011) Integrative Genomics Viewer. *Nature Biotechnology* 29, 24–26.
- Thorvaldsdóttir, H., *et al.*, (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*.
- Trapnell, C., *et al.*, (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5): 511-515.
- Trapnell, C., *et al.*, (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7, 562–578.
- Trapnell, C., *et al.*, (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* (Oxford, England), 25(9):1105-1111.
- Vandesompele *et al.*, (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7):research0034.1–0034.11.

Frequently asked questions

Question: What is Q-score?

Answer: A quality score (or Q-score) is an estimation of the probability of an incorrect base call. $Q\text{-score} = -10 \log_{10}(p)$ where p is the estimated probability of the base call being wrong. A quality score of 10 indicates an error probability of 0.1, a quality score of 20 indicates an error probability of 0.01, a quality score of 30 indicates an error probability of 0.001, and so on. A Q-score above 30 (>99.9% correct) is considered high data quality. In order to pass the Data QC, all samples must have an average Q-score across both the first half and the second half of the reads of at least 16 (>97.5% correct).

Question: What is the difference between FPKM and RPKM?

Answer: RPKM stands for Reads per Kilobase of transcript per Million mapped reads.

FPKM stands for Fragments per Kilobase of transcript per Million mapped fragments. The term “fragments” refers to the cDNA fragments present during library preparation.

Both RPKM and FPKM are normalized numbers which tell you something about the relative abundance of, for example, an assembled transcript.

In paired-end sequencing, two reads are produced per cDNA fragment during library preparation, whereas only one read is produced per cDNA fragment in single-end sequencing. Thus, single-end versus paired-end sequencing will affect the value of RPKM but not FPKM. Consequently, FPKM is preferred over RPKM as it will provide values comparable between single-end sequencing and paired-end sequencing.

Question: What does 1 FPKM mean in terms of abundance?

Answer: This is difficult to estimate and highly variable according to cell type and the total number of mRNAs in a given cell. For example, it was estimated that in a single cell analysis of the cell line GM12878 that one transcript copy corresponds to 10 FPKM (Marinov 2014). Others find that “FPKM are not directly comparable among different subcellular fractions, as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell (Kellis 2013). Depending on the total amount of RNA in a cell, one transcript copy per cell corresponds to between 0.5 and 5 FPKM in poly-A+ whole-cell samples according to current estimates with the upper end of that range corresponding to small cells with little RNA and vice versa”.

Question: What is a “novel” RNA transcript?

Answer: A novel transcript is characterized as a transcript which contains features not present in the reference annotation. Thus, a novel transcript can be both a new isoform of a known gene or a transcript without any known features. For example, a novel transcript could be the result of a previously unknown splicing event for a known gene or a previously unknown long noncoding RNA.

Question: A novel transcript identified seems to be a known gene when I look it up in the gene browser, why is that?

Answer: Most novel transcripts are not new “genes” but different isoforms of previously annotated genes. A novel transcript is most commonly a novel combination of exons or a different start site.

Question: Where can I find the most up to date information on lncRNA annotations?

Answer: lncRNA identification and annotation is at a preliminary stage (Mattick *et al.*, 2015). [RNACentral](#) or [GENCODE](#) are recommended lncRNA resources to investigate potentially novel lncRNAs identified in your study.

Question: Where do I find the result files?

Answer: The result files are located in [XploreRNA](#) on the results page. Go to MyProjects and click on the relevant project and analysis. For a description of the result files, please open the file “file_descriptions.html” (using an internet browser) located in the result folder in [XploreRNA](#).

Definitions

CDS	Coding DNA Sequence
Exon	Sequence that remains present within the final mature RNA product of that gene after introns have been removed by splicing.
FPKM	Fragments Per Kilobase of transcript per Million mapped fragments.
Fragment Count	The number of fragments originating from a feature.
Gene	The standard definition of a gene is a high level feature on the genome that codes for a protein or RNA with a function in the organism. The same gene may encode multiple different RNA transcripts (or isoforms) through alternative splicing or different transcriptional start sites. In the context of RNA-seq, the term "gene" is used to refer to all RNA transcripts (or isoforms) encoded by the same gene. For example, when analyzing differentially expressed genes, the reads from all transcripts derived from the same gene are included.
Gene_ID	Gene identifier. For known genes this will be the gene id from the annotation source, e.g. an Ensembl gene ID. For novel genes, this will be a unique generic identifier, e.g. "CUFF.2".
GO	Gene Ontology (GO – Gene Ontology Consortium, 2000) is an initiative to describe genes, gene products and their attributes using vocabulary (GO terms) which is unified and controlled across all species. The GO terms are categorized into three GO domains: molecular function, biological process and cellular component.
GO_ID	Unique identifier for a Gene Ontology (GO) term.
Intron	A sequence within a gene that is removed by splicing during maturation of the final RNA product.
Isoforms	Different closely related transcripts arising from the same primary transcript (and same gene or DNA sequence) by alternative splicing of exons for example. Isoforms may also be referred to as transcripts.
lncRNA	long non-coding RNA
Mappable reads	Sequences which can be aligned to the reference genome.
mtRNA	mitochondrial RNA

Novel transcript	A transcript which contains features not present in the reference annotation. A novel transcript can be both a new isoform of a known gene or a transcript without any known features. A novel transcript is most commonly a novel combination of exons or a different start site.
Outmapped reads or high abundance reads	For example; rRNA, mtRNA, poly-A and poly-C homopolymers.
P_ID	Promoter ID. This value is extracted from reference annotations, which contain CDS information.
phiX	Libraries generated from the phiX virus used as a control in sequencing runs.
Primary transcript (pre-mRNA)	RNA sequence transcribed from DNA. The primary transcript is then processed (e.g. by addition of 5' cap, 3'-polyadenylation, alternative splicing) to yield various mature RNA products such as mRNAs, ncRNAs, tRNAs, and rRNAs. Multiple primary transcripts may be transcribed from the same gene by use of different transcriptional start sites.
pri-miRNA	primary microRNA transcript
Promoter	A region of DNA that initiates transcription of a particular gene. Promoters are located near the transcriptional start sites of genes.
Q-score	Quality score used to assess the quality of the bases or reads in NGS data. See FAQs for further explanation.
q-value	P-values that have been adjusted to correct for multiple testing.
Reads	DNA Sequence generated by the sequencing machine (for paired end sequencing the same strand is sequenced in both directions – forward and reverse)
rRNA	ribosomal RNA
Transcript	RNA sequence (e.g. mRNA, ncRNA, tRNA or rRNA) transcribed from DNA. Transcripts may also be referred to as isoforms.
Transcript ID	Transcript identifier. For known genes this will be the transcript id from the annotation source, e.g. an Ensembl gene ID. For novel transcripts, this will be a generic identifier, e.g. "CUFF.2.1".

tRNA	transfer RNA
TSS	Transcriptional Start Site
TSS_ID	A unique identifier for the inferred Transcriptional Start Site (TSS). Note: this identifier is unique only within a single analysis.
Unmapped reads	No alignment to the reference genome is possible. Explanations for this include the read being too short, low read quality or contaminations.

SAMPLE



USA, Canada and Mexico

QIAGEN Genomic Services
6951 Executive Way
Frederick, Maryland 21703
USA
+1 301 673 5045

Europe and all other countries

QIAGEN Genomic Services
Qiagen Str. 1
40724 Hilden
Germany
+49 2103 29 11649