RNA-seq Data Delivery Specifications

Output data info

- FASTQ.GZ files containing raw sequences

Depending on the library application kit, reads will be provided either with or without masking of adapter sequences. No quality clipping is provided with raw reads delivery. Information content of each read and strandness are dependent on preparation type. See Table below.

Library preparation protocol	Adapter masked	First read sequence	Second read sequence	UMI sequence	cDNA strandness
Illumina TruSeq Stranded Total RNA Gold	Yes	R1	R2	No	fr-firststrand
Illumina TruSeq Stranded Total RNA Plant	Yes	R1	R2	No	fr-firststrand
Illumina TruSeq Stranded mRNA	Yes	R1	R2	No	fr-firststrand
Illumina TruSeq smallRNA	No	R1	No	No	unstrand
NuGEN Ovation Universal Plus mRNA-seq	Yes	R1	R2	No	fr-secondstrand
NuGEN Ovation SoLo RNA-Seq	No	R1	R3	R2	fr-secondstrand
NuGEN Ovation RNA-Seq v2	Yes	R1	R2	No	unstrand

RNA-Seq standard bioinformatics analysis

- Alignments in BAM format
- Estimated gene-level expression with the FPKM and TPM values per sample
- Metrics describing overall quality metrics computed from the BAM file (gene body coverage, strandness, etc..).
- BedGraph format files allowing for data visualization in a Genome Browser such as UCSC
- A REPORT describing library preparation and analysis flow

- Differential expression analysis output (only when required):
    - a list of significantly differentially expressed genes
    - genes found to be differentially expressed in any of the tests
    - a table with all the statistical testing for a given pairwise test
    - gene lists filtered with adjusted p-value < 0.01
    - MA-Plot for each comparison
    - PCA analysis of all samples/groups with VST-normalized data
    - A hierarchical clustering of samples using VST-normalized data
    - Files for hierarchical and k-means cluster analysis visualization using JavaTreeView (not for smallRNA)

A description of each file is provided in the REPORT.

For multifactorial experiments, analyses can be carried by means of general linear models (interaction of several factors or uncontrolled "batch effect"): please inquire to define parameters.

smallRNA-seq standard bioinformatics analysis

- Raw counts of tags mapping against all miRNAs available in mirBASE and piRNAs (only for human through piRNABase).
- A histogram of size distribution and the classification of small RNA sequences
- A REPORT describing library preparation and analysis flow (with sequence adapters)

- Differential expression outputs (only when required):
    - A list of miRNAs with tag-counts for each sample along with fold changes, p-values and q-values
    - MA-plots for each comparison
    - PCA analysis of all samples/groups
    - A heatmap of expression levels for all miRNAs

IMPORTANT: for RNA-seq/smallRNA-Seq analyses, we strongly advocate the presence of a minimum of three biological replicates for each condition. With no replicates, detection of differentially expressed genes is severely impaired.

FAQ

Are reads quality trimmed?

Delivered raw data are not quality trimmed. However, our internal analysis pipelines always rely on a quality trimming step which is described in the delivery REPORT.

Are PCR duplicates removed during RNA-seq analyses?

Duplicates in RNA-seq are not necessarily an artifact. In fact, observing high rates of read duplicates in RNA-seq libraries is common. It may not be an indication of poor library complexity as long as the input material is recognized to be sufficient. For low input samples, specific protocols integrate UMIs (unique molecular identifiers that act as original cDNA fragment barcodes) permitting to faithfully remove PCR duplicates (also from single reads dataset).

Are raw reads depleted from duplicates?

With standard delivery of raw data only, deduplicated raw reads are NOT provided. In case RNA-Seq analysis has been requested, only for UMI-bearing libraries deduplicated FASTQ files are generated and can be delivered upon request.

Does mRNA-seq detect long non-coding RNAs?

LncRNAs are 50/50 polyA+ and polyA-. If the RNA-seq library is polyA+ enriched there will be a bias in analysis for those lncRNAs that are polyA+.