# RNA-Seq Analysis v2.0

Project / Study: EF-DEMO
Date: 27 May, 2021
RNAseq Pipeline v2.0

# 1 Data Analysis Report

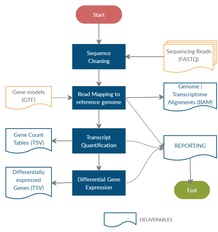
## 1.1 Samples Analysed

The list of samples in the analysis is as follows:

| ID | Sample Name |
|---|---|
| 1 | control_1 |
| 2 | control_2 |
| 3 | sample_1 |
| 4 | sample_2 |

## 1.2 Analysis Workflow

Schematic diagram showing the main steps of the analysis method followed to perform the data analysis.




## 1.3 Reference databases

- List of reference genome and gene annotations used
  - Genome: name of reference genome
  - Gene models: gene model

Source: Reference genome source link

## 1.4 Quality Control of raw sequencing data

Raw sequencing data are preprocessed to generate clean data for downstream analysis. In this step, quality of raw sequencing is checked and filtered to retain only high quality bases by performing adapter trimming, quality filtering and per-read quality pruning.

Quality is interpreted as the probability of an incorrect base call or, equivalently, the base call accuracy. The quality score is logarithmically based, so a quality score of 10 reflects a base call accuracy of 90%, but a quality score of 20 reflects a base call accuracy of 99% and a quality score of 30 reflects a base call accuracy of 99.9%. These probabilistic values are the results from the base calling algorithm and depend on how much signal was captured for the base incorporation.

Sequencing reads representing reads with quality score at least Q30 is above 90% is of very good quality. For a reasonably good sample source material, according to Illumina specifications, one could expect >75% reads with at least Q30 Phred quality.

Raw sequencing data is processed using fastp[1] software to remove poor quality bases (below Phred Quality 20) using the sliding window approach where in if the average quality of the bases drops below Q20, those bases are removed from the reads. After quality trimming, program checks for presence of any adapters in the reads and removes from the reads. Further, shorter reads which are <30bp length are also removed to retain only high quality sequencing reads for each sample in the analysis. In case of paired-end reads, both the sequencing reads which pass the QC criteria are considered for downstream analysis.

After QC processing, QC metrics such as Q30 reads and GC content can be used to assess the sequencing and sample quality across the samples.

## 1.5 Read Statistics

- Table 1: Sequence Quality Metrics overview. For each sample, the following QC metrics are provided:
  - Sample Name: name of the sample.
  - Total Raw Reads: the total number of raw sequencing reads generated for the sample.
  - Total HQ Reads: the total number of high quality reads after sequence cleaning and filtering.
  - HQ Bases (Q30): Percentage of high quality bases having at least phred quality 30.
  - GC Content: GC content in percentile of high quality sequencing reads.
  - Mean Read Length (bp): Average read length in bp of high quality sequencing reads.
  - HQ Reads %: High Quality Reads percentage

| ID | Sample Name | Total Raw Reads | Total HQ Reads | HQ Bases (Q30) | GC Content | Mean Read Length (bp) | HQ Reads % |
|---|---|---|---|---|---|---|---|
| 1 | control_1 | 68.84 M | 68.07 M | 92.7% | 50.2% | 149 | 98.9% |
| 2 | control_2 | 49.41 M | 48.89 M | 92.3% | 49.5% | 150 | 99.0% |
| 3 | sample_1 | 59.4 M | 58.74 M | 92.5% | 49.5% | 150 | 98.9% |
| 4 | sample_2 | 68.84 M | 68.07 M | 92.7% | 50.2% | 149 | 98.9% |

## 1.6 Mapping to reference genome/transcriptome

High quality sequence reads are aligned to the reference genome using STAR (Spliced Transcripts Alignment to a Reference) along with the known gene models.

STAR[2] is an aligner designed to specifically address many of the challenges of RNA-Seq data mapping using a strategy to account for spliced alignments. In general, STAR algorithm achieves highly efficient mapping by performing a two-step process – i) Seed searching, followed by ii) Clustering, stitching, and scoring.

In seed searching step (i), for every read that STAR aligns, STAR will search for the longest sequence that exactly matches one or more locations on the reference genome. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs). The different parts of the read that are mapped separately are called 'seeds'. So the first MMP that is mapped to the genome is called seed1. STAR will then search again for only the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome, or the next MMP, which will be seed2. This sequential searching of only the unmapped portions of reads underlies the efficiency of the STAR algorithm. STAR uses an uncompressed suffix array (SA) to efficiently search for the MMPs, this allows for quick searching against even the largest reference genomes. If STAR does not find an exact matching sequence for each part of the read due to mismatches or indels, the previous MMPs will be extended. If extension does not give a good alignment, then the poor quality or adapter sequence (or other contaminating sequences) will be soft clipped. In clustering, stitching, and scoring step (ii), the separate seeds are stitched together to create a complete read by first clustering the seeds together based on proximity to a set of 'anchor' seeds, or seeds that are not multi-mapping. Then the seeds are stitched together based on the best alignment for the read (scoring based on mismatches, indels, gaps, etc.).

## 1.7 Sequencing and mapping

150bp paired reads were generated from each sequencing library on NovaSeq sequencing platform. On an average around 50-60 million reads were produced per sample of which on an average around 98% reads could be placed onto reference sequence.

- Table 2: Mapping statistics overview. For each sample, the following statistics are provided:
  - Total HQ Paired Reads: the total paired end high quality reads after sequence cleaning and filtering
  - Reads Mapped: the total number of paired reads mapped to the reference.
  - Uniquely Mapped Reads: number of uniquely mapped reads, i.e. read can only be mapped to one reference locus.
  - Unmapped Reads: number of unmapped reads, i.e. read could not get mapped to the reference.

| ID | Sample Name | Total HQ Paired Reads | Unmapped Reads | Uniquely Mapped Reads | Reads Mapped |
|---|---|---|---|---|---|
| 1 | control_1 | 34.04 M | 376.13 K(1.1%) | 32.1 M(94.3%) | 33.66 M(98.9%) |
| 2 | control_2 | 24.45 M | 238.16 K(1.0%) | 23.1 M(94.5%) | 24.21 M(99.0%) |
| 3 | sample_1 | 29.38 M | 378.56 K(1.3%) | 27.66 M(94.1%) | 29 M(98.7%) |
| 4 | sample_2 | 34.04 M | 376.13 K(1.1%) | 32.1 M(94.3%) | 33.66 M(98.9%) |

## 1.8 Transcript Quantification

Gene wise quantification is achieved by inspecting transcriptome alignments using RSEM[3] tool. Using just the number of reads mapped to a transcript as a proxy for the transcript's expression level, leads to the problem that the origin of some reads cannot always be uniquely determined; if two or more distinct transcripts in a particular sample share some common sequence (for example, if they are alternatively spliced mRNAs or mRNAs derived from paralogous genes), then sequence alignment may not be sufficient to discriminate the true origin of reads mapping to these transcripts. One approach to addressing this issue involves discarding these multiple-mapped reads (multireads for short) entirely. Another involves partitioning and distributing portions of a multiread's expression value between all of the transcripts to which it maps (rescue-method). RSEM improves upon this approach, utilizing an Expectation-Maximization (EM) algorithm to estimate maximum likelihood expression levels accounting for multimapped reads generating near accurate expression levels. Using RSEM corrected p-value (adjusted p-value) <0.1 as a threshold, significantly differentially expressed genes between conditions were identified and reported.

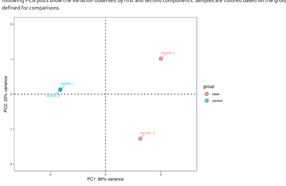
Sample groupings and the comparisons performed are listed in the following table -

| sample | condition |
|---|---|
| control_1 | control |
| control_2 | control |
| sample_1 | case |
| sample_2 | case |

| group1 | group2 |
|---|---|
| case | control |

## 1.9 PCA analysis

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. The following PCA plots show the variation observed by first and second components. Samples are colored based on the groups defined for comparisons.




## 1.10 Differential gene expression

Genes receiving less than 10 reads on an average across the compared groups were removed. The abundance counts of each gene were then used to perform differential gene expression (DGE). DGE was performed using R/Bioconductor DESeq2 package [4], which essentially normalizes the abundance counts to account for observed variance (due to differences in sequencing depths, sample groups and replicates) generating normalized gene counts. Statistical tests were performed for each gene to compare the distributions between conditions (treatment vs control) generating p-values for each gene. The final p-values were corrected by determining false discovery rates (FDR) using the Benjamini–Hochberg method. Using a FDR corrected p-value (adjusted p-value) <0.1 as a threshold, significantly differentially expressed genes between conditions were identified and reported.

Sample groupings and the comparisons performed are listed in the following table -

| sample | condition |
|---|---|
| control_1 | control |
| control_2 | control |
| sample_1 | case |
| sample_2 | case |

| group1 | group2 |
|---|---|
| case | control |

## 1.11 Volcano Plot

The volcano plot shows the fold differences in gene expression levels between the compared groups. Foldchange values are plotted on x-axis and -log10 FDR corrected p-values are on y-axis. Down regulated genes have negative FoldChange values where as up regulated genes have positive FoldChange values. FoldChange differences observed in genes below 0.1 FDR corrected p-value are considered significantly differ between compared groups. Significantly expressed genes having at least 1-fold change are colored green.




# 2 Output Files and Descriptions

- EF-DEMO_alignment_index_files.tar.gz : The alignment directory contains the result files of the read mapping:
  - *_bam.bai : The index files of * bam files are needed e.g. for the visualization with IGV.
- *_bam : These files contain the results of the read mapping in BAM format. The * bam files are sorted by alignment positions and contain mapped as well as unmapped reads. Use "samtools view -F 4 in.bam > out.sam" to extract mapped reads. Use "samtools view -f 4 in.bam > out.sam" to extract unmapped reads.
- EF-DEMO_gene_counts.tar.gz : Gene Count directory contains the sample wise raw and normalized counts per tissue type are found in the files listed.
  - *sample_wise_gene_read_counts_normalized.csv* : These files contain the gene wise normalized reads counts in TSV format.
  - *sample_wise_genes_counts.tsv* : These files contain the gene wise reads counts in TSV format. The content of these files have the following information -

| Column_Description | |
|---|---|
| gene_id | *Gene identifier from gene models used in the analysis* |
| gene_name | *Corresponding gene name from refseq* |
| [sample]_counts | *Sample wise read counts observed* |

- EF-DEMO_gene_foldchanges.tar.gz : Gene Foldchanges directory contains the sample wise comparison foldchange and significant foldchange values.
  - */group1VSgroup2/group1.foldchange.csv*
  - */group1VSgroup2/foldchange.significant.csv*
  - */group1VSgroup2/group1.foldchange.volcanoplot.png*

| Column_Description | |
|---|---|
| gene_id | *Gene identifier from gene models used in the analysis* |
| gene_name | *Corresponding gene name from refseq* |
| mean_counts | *Average of the normalized counts between the compared pair* |
| log2_foldChange | *FoldChange observed in treatment compared to control and is reported on a logarithmic scale* |
| | *to base 2. Positive values indicate upregulated genes in treatment and negative values indicate* |
| | *downregulated in the treatment compared to control sample.* |
| p_value | *Statistical test performed for each gene to valid uff any experimental variability* |
| | *False discovery rate (FDR) corrected P-value and is determined using a permutation test on* |
| | *treated versus control samples and corrected for multiple testing using the Benjamini–* |
| | *Hochberg method. FoldChange differences observed in genes below 0.1 FDR corrected p-value* |
| adj_p_value | *are considered significantly differ between treatment and control sample.* |

# 3 Additional Information

The reads and their associated alignment positions are provided as BAM formatted files. BAM files are binary, so they cannot be opened or edited with a text editor. To extract information or to manipulate BAM files, please refer to samtools (http://www.htslib.org/) or to the Picard software package (http://broadinstitute.github.io/picard/). Further documentation on data in BAM or SAM format can be found in the SAM Format Specification (http://samtools.github.io/hts-specs/SAMv1.pdf). A practical tool to view alignments, expression profiles, or variant data is the Integrative Genomics Viewer (IGV) for Unix, MS Windows, and MacOS X (http://broadinstitute.org/igv).

## 3.1 Software

*List of programs and their versions used

- Fastp-v0.20.0
  - STAR-v2.7.3
  - RSEM-v1.3.3
  - multiqc-v1.8
  - R-v3.2.4

# 4 Bibliography


1. Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560.

2. Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras; STAR: ultrafast universal RNA-seq aligner, Bioinformatics, Volume 29, Issue 1, 1 January 2013, Pages 15–21.

3. Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). https://doi.org/10.1186/1471-2105-12-323

4. Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550.

5. Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6):1767–1771, 2010.

6. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–2079, 2009.

7. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.