RNA-Seq Analysis v2.0

Project / Study: NG-26005
Date: 18 May, 2021
RNAseq Pipeline v2.0

1 Data Analysis Report

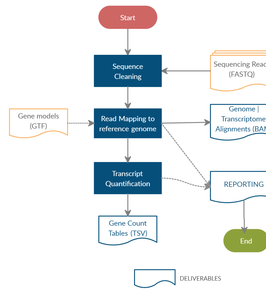
1.1 Samples Analysed

The list of samples in the analysis is as follows:

	ID	Sample Name	ID	Sample Name
1	H1_1	11	P2
2	H1_2	12	P3
3	H2	13	P4
4	H3_1	14	P5
5	H3_2		
6	H4		
7	H5		
8	N3		
9	N5		
10	P1		

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: Human (hg38 | GRCh38)
  - Gene models: gencode v27

  Source:https://www.ensembl.org/Homo_sapiens/Info/Index

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 probability 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	H1_1	59.21 M	58.4 M	89.8%	48.1%	149	98.6%
2	H1_2	66.39 M	65.68 M	91.7%	47.9%	150	98.9%
3	H2	58.15 M	57.58 M	91.7%	48.5%	149	99.0%
4	H3_1	72.09 M	71.16 M	90.0%	47.0%	149	98.7%
5	H3_2	71.63 M	70.93 M	91.7%	48.3%	149	99.0%
6	H4	107.74 M	106.72 M	91.6%	48.4%	149	99.0%
7	H5	71.11 M	70.19 M	90.3%	48.2%	149	98.7%
8	N3	64.88 M	63.89 M	89.3%	47.4%	149	98.5%
9	N5	88.32 M	87.29 M	91.4%	47.2%	149	98.8%
10	P1	62.62 M	61.73 M	89.6%	48.0%	148	98.6%
11	P2	103.54 M	102.45 M	91.6%	49.0%	150	99.0%
12	P3	85.13 M	83.83 M	89.3%	48.4%	149	98.5%
13	P4	107.96 M	106.88 M	91.7%	48.2%	149	99.0%
14	P5	64.82 M	63.78 M	89.0%	48.3%	149	98.4%

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 seed 1. 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 sequence) 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 6000 S2 sequencing platform. On an average around 60-100 million reads were produced per sample of which on an average around 99% reads could be placed onto Human 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 genome.
  - 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	H1_1	29.2 M	377.21 K(1.3%)	27.9 M(95.5%)	28.82 M(98.7%)
2	H1_2	32.84 M	361.65 K(1.1%)	31.26 M(95.2%)	32.48 M(98.9%)
3	H2	28.79 M	308.94 K(1.1%)	27.58 M(95.8%)	28.48 M(98.9%)
4	H3_1	35.58 M	479.41 K(1.3%)	33.62 M(94.5%)	35.1 M(98.7%)
5	H3_2	35.46 M	346.57 K(1.0%)	33.85 M(95.4%)	35.12 M(99.0%)
6	H4	53.36 M	505.99 K(0.9%)	51.27 M(96.1%)	52.85 M(99.1%)
7	H5	35.1 M	470.73 K(1.3%)	33.74 M(96.1%)	34.63 M(98.7%)
8	N3	31.95 M	475.01 K(1.5%)	30.56 M(95.7%)	31.47 M(98.5%)
9	N5	43.65 M	827.14 K(1.9%)	41.58 M(95.3%)	42.82 M(98.1%)
10	P1	30.86 M	447.2 K(1.4%)	29.44 M(95.4%)	30.42 M(98.6%)
11	P2	51.23 M	570.61 K(1.1%)	49.15 M(96.0%)	50.66 M(98.9%)
12	P3	41.92 M	622.88 K(1.5%)	40.02 M(95.5%)	41.29 M(98.5%)
13	P4	53.44 M	536.87 K(1.0%)	51.38 M(96.1%)	52.91 M(99.0%)
14	P5	31.89 M	506.6 K(1.6%)	30.22 M(94.8%)	31.38 M(98.4%)

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 expected counts for each gene annotated in the known gene model. Read counts are further normalized to account for sequencing depth and gene length biases, fragment per kilobase per million (FPKM) and Transcripts per million (TPM) values are generated and reported. Gene wise "expected counts" can then be used to identify differentially expressed genes.

2 Output Files and Descriptions

- NG-26005_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.
- NG-26005_gene_counts.tar.gz : Gene Count directory contains the sample wise raw counts per tissue type are found in the files listed:
  - `*.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*

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. 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.

5. 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.

6. 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.