**Bioinformatics Core**

**Centre for PanorOmic Sciences (CPOS)**

LKS Faculty of Medicine

The University of Hong Kong

**Bioinformatics Report**

Transcriptome Sequencing

|  |  |
| --- | --- |
| User Information | |
| **Institution** | The University of Hong Kong |
| **PI Name** | Prof LEUNG, Yu Hung Anskar |
| **User Name** | Chee Chean Dang |

|  |  |
| --- | --- |
| General Information | |
| **Job ID** | CPOS-211019-CCD-10040a |
| **Date of Data Release** | 23 Nov 2021 |
| **Number of Samples** | 18 |
| **Sample Name** | K052-C-2N-1, K052-C-2N-2, K052-C-2N-3, K052-C-4N-1, K052-C-4N-2, K052-C-4N-3, K052-C-2N4N-1, K052-C-2N4N-2, K052-C-2N4N-3, K052-T-2N-1, K052-T-2N-2, K052-T-2N-3, K052-T-4N-1, K052-T-4N-2, K052-T-4N-3, K052-T-6N-1, K052-T-6N-2, K052-T-6N-3 |
| **Data Analysis Performed By** | CPOS, Bioinformatics Core |

**1. Data Analysis Pipeline**

The samples were analyzed with the work flow as shown in Figure 1.

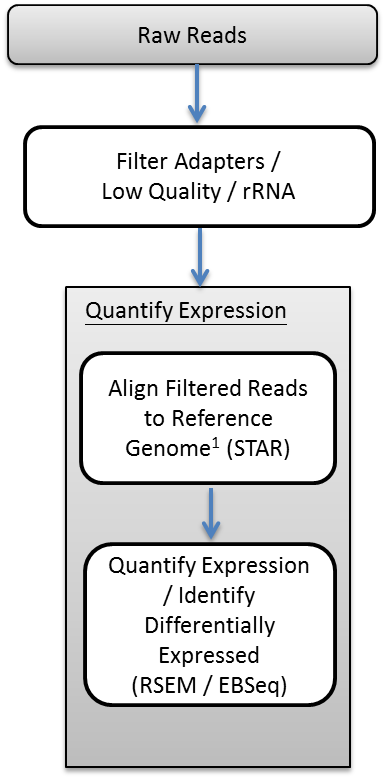


Figure 1. Bioinformatics analysis pipeline. 1**Human Genome GRCh38** (downloaded from GENCODE) was used as reference genome. Note that the annotation contains a total of 60,725 genes and the total number of transcripts is 199,348.

**2. Software**

The following software was used for data analysis:

1. STAR Version 2.7.8: Alignment

(https://github.com/alexdobin/STAR)

1. RSEM Version 1.2.31: Quantify expression

(https://deweylab.github.io/RSEM/)

1. EBSeq Version 1.18.0: Identify differentially expressed genes

(http://bioconductor.org/packages/release/bioc/html/EBSeq.html)

**3. Data Statistics**

**3.1 Pre-processing of Reads**

Sequencing reads were first filtered for adapter sequence and low quality sequence followed by retaining only reads with read length ≥ 40bp. Low Quality is defined as:

1. Reads with more than 5% unknown bases (“N”)
2. Reads having more than 50% of bases with quality value <= 11

Subsequently, sequencing reads were filtered for rRNA sequence and remaining reads were used for downstream analysis (Table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample ID** | **Number of raw reads (read1 + read2)** | **Number of filtered reads** | **% of filtered reads** |
| K052-C-2N-1 | 59,211,698 | 58,310,176 | 98.48% |
| K052-C-2N-2 | 57,541,640 | 56,568,222 | 98.31% |
| K052-C-2N-3 | 61,362,634 | 59,827,636 | 97.50% |
| K052-C-2N4N-1 | 55,801,484 | 54,806,118 | 98.22% |
| K052-C-2N4N-2 | 62,111,344 | 61,157,710 | 98.46% |
| K052-C-2N4N-3 | 59,378,392 | 58,329,932 | 98.23% |
| K052-C-4N-1 | 61,558,230 | 60,258,754 | 97.89% |
| K052-C-4N-2 | 67,624,014 | 66,511,698 | 98.36% |
| K052-C-4N-3 | 62,663,152 | 61,453,078 | 98.07% |
| K052-T-2N-1 | 64,285,866 | 63,268,000 | 98.42% |
| K052-T-2N-2 | 54,740,764 | 53,775,842 | 98.24% |
| K052-T-2N-3 | 61,124,242 | 59,876,700 | 97.96% |
| K052-T-4N-1 | 66,384,192 | 64,746,738 | 97.53% |
| K052-T-4N-2 | 55,518,488 | 53,726,790 | 96.77% |
| K052-T-4N-3 | 58,069,086 | 56,211,436 | 96.80% |
| K052-T-6N-1 | 62,628,456 | 60,555,548 | 96.69% |
| K052-T-6N-2 | 56,755,726 | 55,773,938 | 98.27% |
| K052-T-6N-3 | 53,850,110 | 52,740,280 | 97.94% |

Table 1. Summary of filtered reads for each sample.

**3.2 Alignment**

Reads were mapped to the reference genome using STAR (default parameters) with the following exception:

1. twopassMode: Basic

The results of the alignment are summarized in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample ID** | **Number of filtered reads** | **Number of reads mapped to reference genome** | **% of reads mapped to reference genome** |
| K052-C-2N-1 | 58,310,176 | 56,191,858 | 96.37% |
| K052-C-2N-2 | 56,568,222 | 54,587,434 | 96.50% |
| K052-C-2N-3 | 59,827,636 | 57,746,758 | 96.52% |
| K052-C-2N4N-1 | 54,806,118 | 52,815,044 | 96.37% |
| K052-C-2N4N-2 | 61,157,710 | 58,892,928 | 96.30% |
| K052-C-2N4N-3 | 58,329,932 | 56,128,948 | 96.23% |
| K052-C-4N-1 | 60,258,754 | 58,070,496 | 96.37% |
| K052-C-4N-2 | 66,511,698 | 64,239,044 | 96.58% |
| K052-C-4N-3 | 61,453,078 | 59,278,198 | 96.46% |
| K052-T-2N-1 | 63,268,000 | 61,138,158 | 96.63% |
| K052-T-2N-2 | 53,775,842 | 51,841,138 | 96.40% |
| K052-T-2N-3 | 59,876,700 | 57,840,042 | 96.60% |
| K052-T-4N-1 | 64,746,738 | 62,308,212 | 96.23% |
| K052-T-4N-2 | 53,726,790 | 51,878,686 | 96.56% |
| K052-T-4N-3 | 56,211,436 | 54,291,522 | 96.58% |
| K052-T-6N-1 | 60,555,548 | 58,298,126 | 96.27% |
| K052-T-6N-2 | 55,773,938 | 53,782,202 | 96.43% |
| K052-T-6N-3 | 52,740,280 | 50,990,300 | 96.68% |

Table 2. Summary of reads mapped to reference genome for each sample.

**3.3 Quantify Expression / Differentially Expression Analysis**

Differentially expression analysis was done using EBSeq (Table 3).

|  |  |  |
| --- | --- | --- |
| **Sample pair** | **Number of differentially expressed genes**1 | **Number of differentially expressed transcripts**1 |
| K052-C-2N vs K052-T-2N | 4,819 | 8,038 |
| K052-C-4N vs K052-T-4N | 4,088 | 6,598 |
| K052-C-2N4N vs K052-T-6N | 5,191 | 8,859 |

Table 3. Number of differentially expressed genes and transcripts for each sample pair.1Differentially expressed genes/transcripts were defined as FDR < 0.05.

**4. Data Deliverables**

**4.1 List of Analysis Results Files**

|  |  |
| --- | --- |
| **File** | **Description** |
| \*\_Gene\_DiffExpressed.xlsx | Results of differentially expressed gene analysis |
| \*\_Transcript\_DiffExpressed.xlsx | Results of differentially expressed transcript analysis |
| \*\_Analysis\_Report.docx | Analysis report |
| Various files in Partek folder | Import to Partek for pathway analysis |

Table 4. List of analysis results files.

**4.2 List of Analysis Data Folders**

|  |  |
| --- | --- |
| **Folder** | **Description** |
| RAW | Raw sequencing files |
| FastQC | Data QC of sequencing files |
| Clean | Sequencing files after filtering |
| STAR\_RSEM\_GeneCounts | Alignment files |
| RSEM | Result files of quantify expression |
| EBSeq | Result files of differentially expression analysis |
| Partek | Files for importing into Partek |

Table 5. List of analysis data folders.

**4.3 Explanation of Results**

**4.3.1 Results of Differentially Expressed Genes/Transcripts**

The excel worksheets contain the following columns:

1. **transcript\_id:** Ensembl Transcript ID (Only for Isoforms)
2. **gene\_id:** Ensembl Gene ID
3. **gene\_name:** Ensembl Gene Name
4. **gene\_biotype:** Gene Classification including: protein\_coding, pseudogene, processed\_pseudogene, miRNA, rRNA, scRNA, snoRNA, snRNA. Please visit <https://www.gencodegenes.org/pages/biotypes.html> for details
5. **chr:** Chromosome Number
6. **start:** Start of Gene/Transcript
7. **stop:** Stop of Gene/Transcript
8. **strand:** Strand of Gene/Transcript
9. **Sample 1 TPM:** Sample 1 Transcripts Per Million (TPM)
10. **Sample 2 TPM:** Sample 2 Transcripts Per Million (TPM)
11. **Sample 1 Normalised Read Counts:** Sample 1 Normalised Read Counts. This number has been used for conducting differential analysis
12. **Sample 2 Normalised Read Counts:** Sample 2 Normalised Read Counts. This number has been used for conducting differential analysis
13. **Posterior Fold Change:** The posterior fold change (condition 1 over condition2) for a gene/transcript. It is the ratio between posterior mean expression estimates of the gene/transcript for each condition. The posterior fold change tends to shrink the fold change estimates of low expressers to values close to 1
14. **Real Fold Change:** The real fold change (condition 1 over condition2) for a gene/transcript. It is the ratio of the normalized within condition 1 mean count over normalized within condition 2 mean count for the gene/transcript
15. **PPDE:** The posterior probability that a gene/transcript is differentially expressed
16. **FDR:** False Discovery Rate. Calculated as (1 - PPDE). Therefore with a FDR of 5%, using the calculation and choose those with FDR < 0.05

**Questions?**

Further questions about laboratory work, contact Illumina Sequencing Platform at:

E: nextgenseq.cpos@hku.hk

P: +852 2831-5477 or +852 2831-5478

Questions about data transfer and analysis, contact Bioinformatics team at:

E: bioinfo.cpos@hku.hk

P: +852 2831-5418

For more information about the types of bioinformatics analysis available and deliverables, visit us at:

W: https://cpos.hku.hk