Bulk RNA-Seq Data Processing Pipeline

Justin Shim[1], Lipei Shao[1], Alex Pelayo[1], Jinxia Ma[1], Ping Jin[1], David Stroncek[1]

[1]Center for Cellular Engineering, National Institute of Health Clinical Center

Introduction

- RNA-Seq is a powerful technique used to study gene expression patterns at the transcriptome level
- However, the data sets that are generated from these studies are very large by nature and typically require expertise in computational analysis
- Additionally, new sequencing techniques and advancements in statistical strategies pose a stiff challenge for researchers lacking in-depth coding knowledge to analyze RNA-data effectively
- We have developed a data processing pipeline using bash scripting that is comprehensive and accessible.
- In this poster presentation, we outline the key steps of the pipeline and demonstrate its application using a sample dataset

Tools

- Quality Control (fastQC)
- Read Alignment (STAR)
- Gene Level Expression Quantification (Subread > featureCounts)

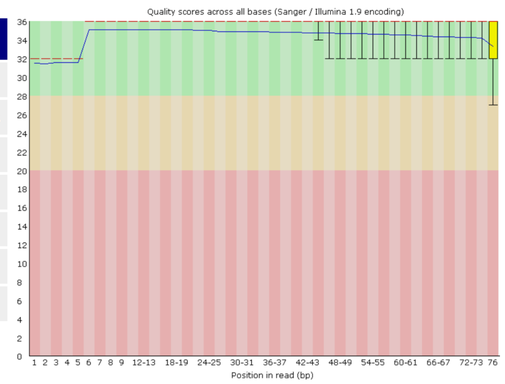
Github

Main Procedures

Measure	Value
Filename	CD22#32.fq1
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	43676996
Sequences flagged as poor quality	0
Sequence length	35-76
%GC	48

Figure 1. (Above) FastQc Report: Basic Statistics for sample #32.fq1

Figure 2. (Right) FastQC quality graph for sample #32.fq1




```
                         Started job on |       Jun 26 14
                     Started mapping on |       Jun 26 14
                            Finished on |       Jun 26 14
Mapping speed, Million of reads per hour |       304.13

                       Number of input reads |       43676996
                   Average input read length |       151
                                UNIQUE READS:
              Uniquely mapped reads number |       38415840
                   Uniquely mapped reads % |       87.95%
                     Average mapped length |       150.52
                  Number of splices: Total |       13806696
       Number of splices: Annotated (sjdb) |       13574227
                  Number of splices: GT/AG |       13668426
                  Number of splices: GC/AG |       104919
                  Number of splices: AT/AC |       11530
      Number of splices: Non-canonical |       21821
             Mismatch rate per base, % |       0.24%
                Deletion rate per base |       0.01%
               Deletion average length |       1.56
              Insertion rate per base |       0.01%
```

Figure 3. Alignment BAM output for sample #32

```
Geneid  Aligned.sortedByCoord.out.bam
ENSG00000223972.5       0
ENSG00000227232.5       103
ENSG00000278267.1       15
ENSG00000243485.3       0
ENSG00000274890.1       0
ENSG00000237613.2       0
ENSG00000268020.3       0
ENSG00000240361.1       0
ENSG00000186092.4       0
ENSG00000238009.6       0
ENSG00000239945.1       0
ENSG00000233750.3       0
```

Figure 4. Subread > featureCount output (.txt) for sample #19

GeneID	Gene Symbol
ENSG00000227232.5	WASH7P
ENSG00000279457.3	FO538757.2
ENSG00000225630.1	MTND2P28
ENSG00000276171.1	AC114498.1
ENSG00000237973.1	MTCO1P12
ENSG00000248527.1	MTATP6P1
ENSG00000198744.5	MTCO3P12
ENSG00000228327.3	RP11-206L10.2
ENSG00000237491.8	RP11-206L10.9
ENSG00000230092.7	RP11-206L10.8
ENSG00000177757.2	FAM87B

Figure 5. Bulk conversion of GeneID to GeneSymbol identifier

Gene Symbol	CD22#19	CD22#21	CD22#32	CD22#33	Extra-CD-22-7	Extra-CD-22-11
LINC00115	60	75	114	69	171	87
FAM41C	0	0	0	0	0	0
TUBB8P11	0	0	0	0	0	0
FAM166AP3	0	0	0	0	0	0
RP11-54O7.16	0	0	0	0	0	0
RP11-54O7.1	0	0	0	0	0	1
RP11-54O7.2	0	0	0	0	0	0
RP11-54O7.3	0	0	0	0	0	2
SAMD11	6	9	7	6	21	3
NOC2L	4614	2510	3555	2647	4131	2962
KLHL17	614	479	703	433	928	586
AL645608.1	0	0	0	0	0	0
PLEKHN1	58	60	78	47	111	63
PERM1	2	2	10	6	20	7
RP11-54O7.17	0	0	2	0	2	0
HES4	4	0	0	0	0	0

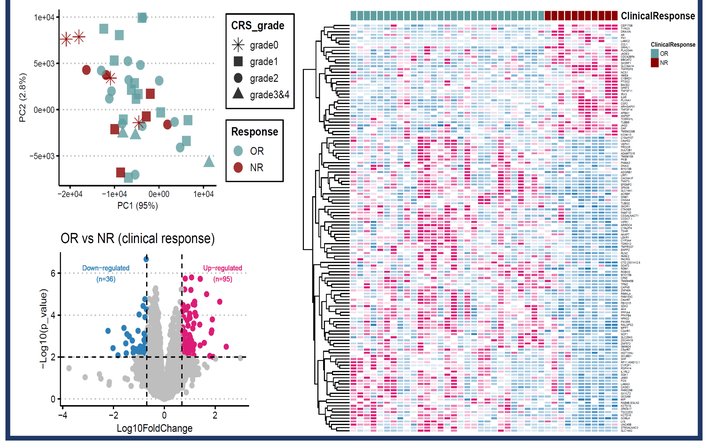
Figure 6. Bulk trimming of low read data (<15)

Conclusion

- Simplifies and automates the analysis of RNA-Seq data
- Assess data quality, align reads to the reference genome, and quantify gene expression levels accurately and seamlessly
- Gain valuable insights into gene expression profiles in various biological contexts
- Output data is organized and straightforward

Future Direction

- Understanding the transcriptional landscape of CAR T-cells before infusion (pre-infusion products)
- Identifying transcriptomic signatures from the data that are responsible for driving different clinical outcomes



Acknowledgments

- FastQC was developed by Simon Andrews, Babraham Bioinformatics
- Alexander Dobin *et al. STAR: ultrafast universal RNA-seq aligner*
- Subread on Biowulf was developed by Yang Liao, Gordon K Smyth, Wei Shi

National Institutes of Health Clinical Center