Single oocyte/embryo RNASeq data processing

ADITYA SANKAR (✉ adisankara2000@gmail.com)

University of Copenhagen <https://orcid.org/0000-0002-1840-3356>

Jens Vilstrup Johansen

University of Copenhagen <https://orcid.org/0000-0001-7094-6801>

Method Article

Keywords: developmental epigenetics, single cell RNASeq, SMARTSeq2

Posted Date: January 27th, 2020

DOI: <https://doi.org/10.21203/rs.2.21804/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

This protocol details the step-by-step procedures followed to process the single oocyte/embryo mRNASeq and 2-cell embryo Total RNASeq data generated using the SMARTSeq2 technology in the associated publication.

Introduction

This protocol is a bioinformatic data processing protocol that is linked to the associated publication.

Equipment

Windows/Mac OS workstation with 12 GB RAM and high end processing power. Alternatively, there can be efficient and faster processing of data if connecting to a computing core using VPN

Procedure

1. For both the single-cell mRNA (151bp PE) and the totalRNA (76bp PE) sequencing the raw data was converted from bcl to fastq format and reads trimmed in BaseSpace.
2. After download from BaseSpace the raw reads were quality assessed with FastQC¹ and Fastq Screen².
3. Afterwards trimmed using Trimmomatic v0.32³ (mRNA-seq settings: PE ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:1:true HEADCROP:15 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:25; totalRNA-seq settings: PE ILLUMINACLIP:TruSeq2-PE.fa:2:30:10:1:true HEADCROP:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:25).
4. For mRNASeq libraries, the trimmed reads were aligned to the mm10 genome assembly using STAR⁴(v2.5.1a) in two-pass mode and guided by a RefSeq (UCSC,2018/08/05) gene annotation (settings: -sjdbOverhang 135 -twopassMode Basic -outSAMtype BAM SortedByCoordinate -outSAMattributes All -outSAMunmapped Within -outFilterMismatchNoverLmax 0.1 -outFilterMatchNmin 16 -outFilterMismatchNmax 5).
5. After mapping, the reads were assigned to genes with featureCounts⁵ (v1.5.1, settings: -primary -p -B -O -M -fraction -s 0 -J) generating a count table.
6. Using R (v3.5.1) (<https://www.r-project.org>), the quality of samples were again estimated using various quantitative and qualitative methods available in the Scater package⁶.
7. One 8-cell MZ mutant sample was excluded from the dataset due to extremely low total gene count (404). the remaining samples had total gene counts in the range 0,5-4,6 mio (mean:2,4 mio).

8. The DESeq2 (v1.22.1) package⁷ was used for statistical analysis of the count data comparing the knockout and wild-type samples within each cell stage. The clusterProfiler package⁸ was used to test for under/overrepresentation of genes in various gene sets.
9. The DBTMEE⁹ gene-to-cluster annotation (cluster_gene_v2.tsv) was downloaded from <http://dbtmee.hgc.jp/download/download.phpm>, and non/low expressed genes removed by imposing the filter criteria FPKM>3 for cell stages Oocyte, 1C, 2C and 4C.
10. Differentially expressed genes from the DeSeq2 analysis were defined as having an absolute log2 fold change ≥ 1 and FDR $\leq 5\%$.
11. Using the compareCluster function from the clusterProfiler R-package we looked for over- or underrepresented DBTMEE gene sets in our list of DE genes. From the compareCluster results we derived the observed/expected ratio based on the values of 'GeneRatio' (Obs) and 'BgRatio' (Exp).
12. The compareCluster results were filtered to only included gene sets with FDR<10% and (DE Gene) Count>3. Furthermore, to simplify the plot we limited the color scale to +1/-1.
13. For total-RNASeq libraries, we tested for differential expression of repeat elements we first used RepEnrich2 (v0.1)¹⁰ to map our totalRNA reads against the RepeatMasker database (mm10,4.0.5,2014013) followed by statistical analysis in R using the edgeR package (v3.24.0)¹¹ as per the authors suggested analysis pipeline.

References

1. S, Andrews. FastQC: a quality control tool for high throughput sequence data from <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.(2010).
2. Andrews, S. FastQ Screen (2011).
3. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
4. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
5. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
6. McCarthy, D.J., Campbell, K.R., Lun, A.T. & Wills, Q.F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186 (2017).
7. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

8. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
9. Park, S.J., Shirahige, K., Ohsugi, M. & Nakai, K. DBTMEE: a database of transcriptome in mouse early embryos. *Nucleic Acids Res* **43**, D771-776 (2015).
10. Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 583 (2014).
11. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

Acknowledgements

Mads Lerdrup, University of Copenhagen