Single oocyte/embryo RNASeq data processing

CURRENT STATUS: POSTED

ADITYA SANKAR
University of Copenhagen

adisankara2000@gmail.com*Corresponding Author*
ORCiD: https://orcid.org/0000-0002-1840-3356

Jens Vilstrup Johansen
University of Copenhagen

ORCiD: https://orcid.org/0000-0001-7094-6801

Rehannah Borup
University of Copenhagen

ORCiD: https://orcid.org/0000-0002-6947-9094

DOI:
10.21203/rs.2.21804/v2

SUBJECT AREAS
Computational biology and bioinformatics

KEYWORDS
developmental epigenetics, single cell RNASeq, SMARTSeq2

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. A separate section highlights processing of human MII oocytes.

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 FastQC[1] and Fastq Screen[2].

3. Afterwards trimmed using Trimmomatic v0.32[3] (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 STAR[4](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 featureCounts[5] (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 package[6].

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) package[7] was used for statistical analysis of the count data comparing the knockout and wild-type samples within each cell stage. The clusterProfiler package[8] was used to test for under/overrepresentation of genes in various gene sets.

9. The DBTMEE[9] 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<=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.

Single Cell RNA seq Data Analysis on 15 Human MII oocytes

- Human oocytes were processed same as mouse oocytes in order to prepare cDNA using the SMART-Seq2[12] protocol with subsequent library preparation using the Nextera XT DNA library preparation kit as described previously.

- Single cell libraries were paired-end sequenced using the Illumina NextSeq500 Instrument to generate 75 base pair reads of the human oocyte libraries.

- We transformed the per-cycle base call (BCL) file output from the sequencing run of 15 human MII oocytes into per-read FASTQ files using the bcl2fastq2 Conversion Software v2.19 from Illumina.

- The samples libraries were multiplexed across four sequencing lanes and the FastQ files from each of the four lanes were concatenated to generate one set of paired fastq files per sample.

- We performed sample QC and filtering of reads to remove low quality (Q < 20) reads and adaptor sequences with AfterQC[13] on the human samples. One sample, PT979-4, was excluded from downstream analysis due to inadequate library generation.

- Subsequent to filtering, we used the remaining paired reads for alignment by hisat23 to the human genome GeneCode v.27 release with the paired GenCode v.27 gtf file containing gene annotations using 'HISAT2 -p 22 --dta -x .gencode.v27 -1 R1.fastq -2 R2.fastq -S sample.sam'.

- The resulting sam files were sorted, indexed and transformed to bam files using samtools[14]. - We filtered the bam files for mitochondrial reads and Stringtie (using the settings '-G $gtffile -e -A $tsvfile $bamfile') was applied to merge and assemble reference guided transcripts for gene level quantifications of fragments per million per kilobasepair (FPKM)[15].

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).

12. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9, 171-81 (2014).

13. Chen, S. et al. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18, 80 (2017).

14. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-9 (2009).

15. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11, 1650-67 (2016).

Acknowledgements

Rehannah Borup and Mads Lerdrup, University of Copenhagen