Zebrafish retinal mRNA RNA-seq data processing

Nicholas Owen
UCL Institute of Ophthalmology, London, UK https://orcid.org/0000-0001-5598-6274
Mariya Moosajee (m.moosajee@ucl.ac.uk)
UCL Institute of Ophthalmology, London, UK https://orcid.org/0000-0003-1688-5360

Method Article

Keywords: RNA-seq, zebrafish, spatio-temporal transcriptome, embryo development, retina, optic fissure

Posted Date: May 19th, 2020

DOI: https://doi.org/10.21203/rs.3.pex-946/v1

Abstract

This protocol details the step-by-step procedures followed to process zebrafish retinal mRNA sequencing data generated by the SMARTSeq2 library preparation protocols in the manuscript Richardson et al 2019 [1].

Introduction

This protocol describes the bioinformatic data processing pipeline for the associated manuscript.

Reagents

Equipment

Access to a high-performance computer cluster is recommended. Alternatively, a desktop workstation running either Linux or Mac OS with 32 GB RAM and high-end CPU(s). Cloud computing platforms such as Amazon Web Services (AWS) or Elastic Cloud (EC1) are also appropriate. Requisite software and packages include R [2]/R Studio [3] (version 3.5 or higher), FASTQC [4], multiQC [5], Trim Galore [6], STAR [7], HTSEQ [8], DESeq2 [9], GOseq [10].

Procedure

1. Raw sequencing data from mRNA SMARTSeq2 libraries (100bp paired-end [PE]) was converted from bcl to FASTQ format in BaseSpace per sample.

2. All FASTQ reads were assessed for quality control using FASTQC and FastQ Screen. FastQ Screen is highly recommended as it assesses the library for sequence origin, ensuring the data match expectations.

3. Sequencing adapters (Illumina) and low quality read bases were trimmed using Trim Galore, removing reads with a quality Phred score under 6.

4. To align the read sequences to the zebrafish genome, the complete genomic sequence (build GCRz10) file (FASTA) and GTF (general transfer format) / GFF (general feature format) (version 95) were obtained from Ensembl (downloaded: FA : ftp://ftp.ensembl.org/pub/release-95/fasta/danio_rerio/dna/ GTF: ftp://ftp.ensembl.org/pub/release-95/gtf/danio_rerio/)

5. Appropriate index files were created using STAR (v2.7.1a):

STAR –runMode genomeGenerate –runThreadN 4 -genomeSAindexNbases 8 –genomeChrBinNbits 14 –genomeDir ./STARIndex/ –genomeFastaFiles ./GRCz11/Danio_rerio.GRCz11.dna.toplevel.fa –sjdbGTFfile ./GRCz11/Danio_rerio.GRCz11.95.gtf –sjdbOverhang 99

6. PE FASTQ reads were aligned using STAR (v2.7.1a) in two-pass mode and guided by the GTF/GFF annotation:

STAR –readFilesIn ${fq1} ${fq2} –readFilesCommand zcat –genomeDir ./STARIndex/ –runThreadN 4 –twopassMode Basic –outFileNamePrefix ${sample_id} –outSAMtype BAM SortedByCoordinate –outSAMunmapped Within –outSAMheaderHD @HD VN:1.4 –outSAMattrRGline ID:${sample_id} CN:UCL DS:RNAseq LB:Truseq PL:${library.platform} SM:${sample_id}

7. Resulting SAM files were sorted by coordinate, converted to BAM format, and indexed using samtools (v1.9):

samtools sort ${input.sam} -o ${output.sorted.bam} -O bam / samtools index ${output.sorted.bam}

8. After mapping, read duplicates were marked using the Picard (v2.20.4) MarkDuplicates command and gene (or transcript) level counts calculated with a module of HTSeq (v0.11.3), htseq-count keeping duplicate reads:

htseq-count –order=pos –stranded=no - ./GRCz11/Danio_rerio.GRCz11.95.gtf > ./counts/${BAM_file_name}_htseq_counts_keepdups.tsv

9. Quality of mapping was carried out, summarizing the number of mapped, multimapped, unmapped, as well as mapped to exons, intergenic, or intragenic regions using RNA-SeQC/picard metrics/Qualimap.

10. Summary reports of all metrics were created using MultiQC (v1.8) [5].

11. The R computing environment (>3.5) and Bioconductor packages, DESeq2 (v1.28.0) was used for statistical modelling of the count data, carrying out pairwise comparisons of optic fissure (OF) tissue and dorsal retina (DR) tissue at each specific developmental time point using WALD testing to generate p values. Metadata of the samples should include; sample id, condition, timepoint. Per sample count data was imported using tximport package, creating a single count dataframe. Low counts were filtered and removed (basemean <10) before modelling. Pairwise comparisons were carried out using specific contrasts called in the DESeq2 command. This ensured the complete dataset was preprocessed and consistent for all comparisons. For the time course analysis, DESeq2 was used with a likelihood-ratio test (LRT) to generate significance values on the complete data set using the interaction of time with tissue origin factors (time:tissue) .

12. The data was explored using several graphical plots using ggplot2 in R [11], including the MA plot – this shows the log2 fold change, per feature, plotted against the mean of normalized counts, for all the samples. An overview of the level of similarity was created using a sample-to-sample distance heatmap, using hierarchical clustering. Principle component analysis (PCA plot) was used to show the samples in a 2D space, separated by the first two principle components, key to identification of outliers and batch effects. All plots can be saved as resolution independent SVG images.

13. Differentially expressed genes (DEGs) were defined in this dataset as those having an absolute log2 fold change >= 1 and an FDR <= 0.05.

14. (optional) Genes can be further annotated using the biomaRt package [12]; providing common gene name, chromosomal location, biotype and strand specificity as well as associated protein identifiers and homologues in a number of species.

15. Gene ontology over representation analysis was carried out using GOseq, which has the benefit of considering any length bias in the data. Alternative tools include GOrilla, DAVID and Enrichr (accessible through web interfaces and APIs).

16. Hierarchical clustering of the identified DEGs can be used to identify clades/groups of differentially expressed genes, either up or down regulated in samples represented together colour coded in the heatmap output. The rows of the heatmap (representing an individual gene/transcript) are reordered according to the clustering result, putting similar observations close to one another. Clustering used the rlog transformed assay data from the DESeq2 object.

17. For inter sample comparisons of specific gene(s) expression, transcripts per million (TPM) values were calculated [13] and appropriate plots created.

Troubleshooting

Time Taken

Anticipated Results

Please see associated Publication

References

1 Richardson, R. *et al.* Transcriptome profiling of zebrafish optic fissure fusion. *Sci Rep* **9**, 1541, doi:10.1038/s41598-018-38379-5 (2019).

2 R Core Team. (Vienna, Austria, 2017).

3 RStudio Team. (Boston, MA, 2015).

4 Andrews, S. *FASTQC - A Quality Control tool for High Throughput Sequence Data*, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2014).

5 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

6 Krueger, F. *Trim Galore!*, <https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/> (2012).

7 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

8 Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

9 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

10 Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14, doi:10.1186/gb-2010-11-2-r14 (2010).

11 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

12 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).

13 Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281-285, doi:10.1007/s12064-012-0162-3 (2012).

Acknowledgements

This research was funded by the Wellcome Trust (205174/Z/16/Z).