Script guide for RNA-seq of lncRNA

Keep in mind to change directories according to your need. You may want to add an unzip command depending on your files.

If you are a IBU cluster user, you can find all scripts and data here used on the cluster: /data/courses/rnaseq/lncRNAs/Project2/lbrun

01read_quality (on cluster)

1. Run read_count.sh: Returns number of reads in input FASTQ file, output in read_count folder
2. Run quality_control_fastqc.sh: FastQC (v0.11.9) (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data) performs a fast quality control on input FASTQ file, output in fastqc folder
3. Run quality_control_multiqc.sh: MultiQC (v1.8) [1] compares all fastqc files, output in multiqc folder

02read_mapping (on cluster)

4. Run read_mapping.sh: Map reads with STAR (v2.7.9a) [2] to CRCh38.p13 primary assembly (PRI regions) in FASTA format (GENCODE - Human Release 39 (gencodegenes.org)) which was indexed by STAR (v2.7.9a) {index}
   Give arguments to script in command line: $1=forward read, $2=reverse read, $3=name of output file, outputs BAM files in STAR_ouput folder

03transcriptome_assembly (on cluster)

5. Run transcriptome_assembly.sh: Assembles transcriptome with StringTie (v1.3.3b) [3], use BAM files from step 4 as input and GRCh38.p13 Genecode.v38.annotation.gtf (CHR regions) in GTF format (GENCODE - Human Release 39 (gencodegenes.org) as reference, outputs GTF files in stringtie_ouput folder
6. Run stringtie_merged.sh: Merges all GTF files, outputs merged GTF file into stringtie_output
7. Use commands from stats_transcriptome_assembly.txt in the command line (with srun) to get the number of genes, exons, transcripts, novel genes and the number of genes and exons that are composed of more than one gene

04quantification (on cluster)

8. Run gtf2fasta.sh: Cufflinks (v2.2.1) (Cufflinks (cole-trapnell-lab.github.io) makes FASTA file out of merged GTF file created in step 6, GRCh38.p13 GRCh38.primary_assembly.genome.fa (PRI regions) in FASTA format (GENCODE - Human Release 39 (gencodegenes.org) was given as a reference file
9. Run indexing_kallisto.sh: kallisto (v0.46.0) [4] indexes the FASTA file and outputs a .fa.fai file
10. Run quantification_kallisto.sh: kallisto (v0.46.0) uses the index file to quantify each FASTQ file (same files as in step 1) and outputs a folder for each sample into output_quantification containing an abundance.h5, an abundance.tsv and run_info.json file. Download output_quantification folder locally to R working directory

05differential_expression (locally on RStudio (v.1.1.4)

11. Run differential_expression_sleuth.R: Sleuth (v0.30.0) [5] creates differential expression tables for transcript level and for gene level. The data can be visualised by creating a volcano plot

06integrative_analysis_prep (on cluster)

12. Run gtf2bed.sh option 1: BEDOPS (v2.4.40) [6] converts merged GTF file from step 6 to BED file
13. Run TSS_bedtool_window.sh: BEDTool window (v2.29.2) [7] returns all entries on which BED file from step 12 matches with transcript start sites listed in TSS_human.bed (Index of /data/fantom5/datafiles/phase1.3/extra/TSS_classifier (biosciencedbc.jp)
14. Run polyAsite_bedtool_window.sh: Analogous to step 12, BEDTool window (v2.29.2) returns all entries where the BED file and polyA sites listed in atlas.clusters.2.0.GRCh38.96.bed (PolyASite - Exploring 3' end processing (unibas.ch)) overlap
    Tip: Check if you have to add “chr” in column 1, otherwise it cannot recognize overlapping transcripts
15. Run coding_potential_CPAT.sh: CPAT (v1.2.4) [8] uses the merged FASTA file from step 8 as input as well as Human_Hexamer.tsv and Human_logitModel.RData (CPAT - Browse /v1.2.4 at SourceForge.net) that have to be downloaded to the working folder. CPAT outputs a R code and a dat file which can be downloaded locally to the R working directory
16. Run gtf2bed.sh option 2: BEDOPS (v2.4.40) converts the GTF annotation file GRCh38.p13 Genecode.v38.annotation.gtf (CHR regions) (GENCODE - Human Release 39 (gencodegenes.org) to a BED file
17. Run intergenic_bedtools_window: BEDTools window (v2.29.2) returns all entries which have NO overlap with the reference genome from step 16
18. Use the command from stats_intergenic_transcripts.txt in the command line (with srun) to get the number of unique intergenic transcripts

06integrative_analysis (locally on RStudio (v1.1.4)):

19. Run integrative_analysis_CPAT.R: R code from step 15 was used for this script
20. Run make_tables.R: Returns all kind of tables

References

[1] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/BIOINFORMATICS/BTW354.

[2] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, p. 15, Jan. 2013, doi: 10.1093/BIOINFORMATICS/BTS635.

[3] M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nat. Biotechnol. 2015 333*, vol. 33, no. 3, pp. 290–295, Feb. 2015, doi: 10.1038/nbt.3122.

[4] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nat. Biotechnol. 2016 345*, vol. 34, no. 5, pp. 525–527, Apr. 2016, doi: 10.1038/nbt.3519.

[5] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, "Differential analysis of RNA-seq incorporating quantification uncertainty," *Nat. Methods 2017 147*, vol. 14, no. 7, pp. 687–690, Jun. 2017, doi: 10.1038/nmeth.4324.

[6] S. Neph *et al.*, "BEDOPS: high-performance genomic feature operations," *Bioinformatics*, vol. 28, no. 14, p. 1919, Jul. 2012, doi: 10.1093/BIOINFORMATICS/BTS277.

[7] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/BIOINFORMATICS/BTQ033.

[8] L. Wang, H. J. Park, S. Dasari, S. Wang, J. P. Kocher, and W. Li, "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model," *Nucleic Acids Res.*, vol. 41, no. 6, p. e74, Apr. 2013, doi: 10.1093/NAR/GKT006.