Processing and Analysis of RNA sequencing data of total RNA (ribodepleted)

Leibniz Institute for Farm Animal Biology (FBN)

Institute Genome Biology, Genome Physiology Unit

Written by Rosemarie Weikard, Wietje Nolte, Christa Kühn (kuehn@fbn-dummerstorf.de)

16th September, 2019

Quality Control and Trimming

- After quality control of raw sequencing reads with FastQC (Andrew 2010), adapter and quality trimming were performed with Cutadapt v. 1.16 (Martin 2011) and Quality Trim v. 1.6.0 (Robinson 2015), respectively. In Quality Trim the start of sequences was also trimmed (option -s) and the maximum number of N bases was set to 3, while the minimum base quality was set to 15.

Alignment

- Reads were then mapped in a guided alignment with HISAT2 v.2.1.0 (Kim et al. 2015) to the bovine reference genome UMD.3.1 (Ensembl annotation release 92). Because of paired-end sequencing the RNA strandness option 'RF' was applied. Multithreading was invoked with --p 12 for 12 processors

Assembly

- After converting from SAM to BAM format, the BAM files were sorted and indexed with samtools v.1.6 (Li et al. 2009). Samples were individually assembled with Stringtie v.1.3.4d (Pertea et al. 2015) based on the reference genome and annotation used for alignment. To account for the stranded library, the --rf option was included.
- Using the individually assembled samples from all four tissues and the bovine reference genome, a new merged annotation was built in Stringtie (--merge) across tissues, while specifying for minimal transcript coverage across samples of 15 read alignments per exonic base (-c 15). Multithreading was invoked with -p 6 to use 6 processors. Additionally, to account for the stranded library, the --rf option was included.

Fragment Counting

- The new merged annotation was used for fragment counting with featureCounts (subread v.1.6.1) (Liao et al. 2014), while allowing for fractional counting (--fraction), counting multimapping reads (-M) and specifying for reverse strandedness (-s 2). Multithreading was invoked with -T12 to use 12 processors.

Code

```
#quality control
fastqc [INPUT] –outdir=[OUTPUT DIRECTORY]

# adapter trimming
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -o
[OUTPUT_read1.fastq.gz] -p [OUTPUT_read2.fastq.gz] [INPUT_read1.fastq.gz]
[INPUT_read2.fastq.gz]

#quality trimming
quality-trim -s -N3 -q15 [INPUT_read1.fastq.gz] [INPUT_read2.fastq.gz] >> [OUTPUT_REPORT.txt]

# alignment
hisat2 -p 20 --rna-strandness RF --dta --new-summary --summary-file [OUTPUT_SUMMARY.txt] -x
[REFERENCE_ANNOTATION] -1 [INPUT_read1.fastq.gz] -2 [INPUT_read2.fastq.gz] -U
[INPUT_singleton.fastq.gz] -S [OUTPUT.sam]

# conversion, sorting, indexing
samtools sort -o [OUTPUT.bam] -O bam -T [TMP.nnn.bam] -@ 12 [INPUT.sam]
samtools index [INPUT.bam]

# assembly
stringtie [INPUT.bam] -p 12 --rf -o [OUTPUT.gtf] -G [REFERENCE_ANNOTATION] -A [OUTPUT.tab] -B
>> [OUTPUT_REPORT.txt]

stringtie -p 6 --merge --rf -c 10 -G [REFERENCE_ANNOTATION] -o [OUTPUT.gtf] [LIST_SAMPLES.txt]

# fragment counting
featureCounts -T 12 -t exon -g gene_id -p -s 2 -M --fraction -a [REFERENCE_ANNOTATION] -o
[OUTPUT] [INPUT.bam] 2>> [OUTPUT_REPORT.txt]
```

Literature

- Andrew, S. 2010. 'FastQC: a quality control tool for high throughput sequence data'. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
- Kim, D., B. Langmead, and S.L. Salzberg. 2015. 'HISAT: a fast spliced aligner with low memory requirements', *Nat Meth*, 12: 357.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al. 2009. 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25: 2078-9.
- Liao, Y., G. K. Smyth, and W. Shi. 2014. 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30: 923-30.
- Martin, M. 2011. 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17: 10-12.
- Pertea, M., G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, and S.L. Salzberg. 2015. 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nat Biotechnol*, 33: 290-95.
- Robinson, A. 2015. 'Quality Trim version 1.6.0', Accessed 29 March 2018. https://bitbucket.org/arobinson/qualitytrim/downloads/.