Bioinformatics analysis for RNA-seq data

For each library, original reads files were quantified, trimmed and aligned to the Oar\_v3.1 reference genome using Cluster-Flow pipeline tool (vesion v0.5 dev, fastqc\_star pipeline) [1], including the following softwares built in, fastqc (version 0. 11.5) [2], trim\_galore (version 0.4.2) [3], fastq\_screen (version 0.9.3) [4], multiqc (version 0.9dev) [5] and reads alignment software STAR (version 2.5.1b\_modified) [6]. Mapped reads were sorted and indexed with samtools [7]. Since the RNA-seq data was used for genes expression quantification, duplicated reads were kept. Subread software (version 1.5.0-p2) [8] with function featureCounts was applied to the indexed bam files for counting mapped reads/fragments per annotated gene from the annotation file provided for the sheep genome (Oar\_v3.1) release.

Initial quality control included principal component analysis (PCA) and two individuals were removed as outliers (B336F2 and B388F2) for the following analysis. Differentially expressed genes were identified using R (version 3.5.3) DESeq2 package (version 1.22.2) [9], using variance stabilizing transformed expression for counts. Genes with more than one read across all samples within a contrast were retained. Additional filtering of genes with low mean read counts was automatically applied by DESeq2. For each contrast, differentially expressed genes with BH-adjusted P-values < 0.05 were identified. Log2 fold change in gene expression was plotted against the mean of read counts normalized by library size for each gene in MA plots. Different contrasts significant expressed genes were plotted in volcano plot and the intersection number of different expressed genes summary were plotted using UpSetR (version 1.4.0). For heatmap analysis, gene-level transcripts expression values were derived by normalised transformed values estimated by DESeq2.

A Bayesian method implemented in DESeq2 was used to moderate the log2 fold changes obtained for genes with low or variable expression levels. Upregulated and downregulated genes in different contrasts (BH-adjusted p < 0.01and absolute log2FoldChange > 1), were analysed for gene ontology (GO) term enrichment. Gene sets were analysed for over-representation of BP (biological process) and KEGG pathway using R package clusterProfiler (version 3.10.1). Significantly enriched terms were identified by applying the default clusterProfiler algorithm coupled with the Fisher’s exact test statistic (p ≤ 0.05, q ≤ 0.05). Corresponding gene ontology plots were using R packages enrichplot (version 1.2.0) and GOplot (version 1.0.2).

Reference:

[1] Ewels P, Krueger F, Kaler M, Andrews S. [Cluster Flow: A user-friendly bioinformatics workflow tool.](https://www.ncbi.nlm.nih.gov/pubmed/28299179) *F1000Res* 2016 Dec 6;5:2824

[2] Andrews S., Krueger F, Degonds-Pichon A., Biggins L., Krueger C. and Wingett S. FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (Jan, 2012)

[3] Krueger F. and Ewels P. Trim Galore: A wrapper tool around Cutadpt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. (Mar 2012)

[4] Wingett SW, Andrews S. [FastQ Screen: A tool for multi-genome mapping and quality control.](https://www.ncbi.nlm.nih.gov/pubmed/30254741.2) *F1000Res.* 2018 Aug 24 [revised 2018 Jan 1];7:1338. doi: 10.12688/f1000research.15931.2. eCollection

[5] Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, *Bioinformatics*, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, <https://doi.org/10.1093/bioinformatics/btw354>

[6] Dobin A., Davis CA., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. and Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.*2013 Jan 1:29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25

[7] Li H., Handsaker B., Wysoker A. Fennell T., RUan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.

[8] Yang Liao, Gordon K. Smyth and Wei Shi, The Subread aligner:fast, accurate and scalable read mapping by seed-and vote. Nucleic Acid Res. 2013 May; 41(10): e108. Published online 2013 Apr 3. doi: [10.1093/nar/gkt214](https://dx.doi.org/10.1093%2Fnar%2Fgkt214)

[9] R Core Team by R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. Vienna, Austria. 2018. <https://www.R-project.org/>

Note: All of the R packages used will be listed in Github with a summary table, with version and references. And multiqc result will be loaded to Github page with corresponding analysis Codes.