RNA\_seq Processing

Sequencing adapters and low-quality reads were trimmed using Trimmomatic-0.39. ‘Maximum information quality filtering’ was used as the trimming algorithm. This algorithm balances read length, read coverage and error rate. A cut-off length of 40 and strictness setting 0.999 was used as described (1). The cut-off for the minimum read length was set to 36bp.

Sequences were aligned using Salmon-1.4.0 (2, 3) . The index was built by concatenating the GRCm38 mouse genome assembly to the end of the GENCODE vM25 mouse transcriptome. The genome assembly functions as a source of ‘decoy’ sequences. They sequester reads that better map to the unannotated genome rather than to the annotated transcriptome. This is a key feature in the selective alignment strategy and mitigates spurious mappings. Random hexamer priming bias and gc-bias correction were implemented. Average fragment length and SD (relevant only for single-end reads) was given as determined by TapeStation.

Tximeta-1.8.2 was used to import Salmon quant.sf files and generate raw count matrices. Reads were summarised at the gene level. Gene symbols were mapped to Ensemble IDs using org.Mm.eg.db-3.12.0 and AnnotationDbi-1.52.0. Some Ensemble IDs didn’t map to gene symbols, so a column was created containing a mix of Ensemble IDs and gene symbols. A separate csv file contains the ‘Ensembl ID”, “gene symbol” and “Ensemble IDs plus gene symbols combined’ columns. Rows (genes) with all 0 counts were removed.

1. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

2. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature methods. 2017;14(4):417-9.

3. Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, et al. Alignment and mapping methodology influence transcript abundance estimation. Genome Biology. 2020;21(1):239.