**Bioinformatic Analyses.**

**Pre-processing, mapping and filtering of RNA-seq data**

Raw sequencing reads were demultiplexed through the UPenn Functional Genomics Core and raw read fastq files from different sequencing runs were merged for each sample. FastQC (Version 0.11.3; Andrews, 2010) was used to assess sequencing read quality. Reads were mapped to the human genome (GRCh38, GENCODE release 22; Harrow et al. 2012 Genome Research) using STAR (Version 2.2.4 with option --outFilterIntronMotifs RemoveNonCanonical; Dobin et al. 2013 Bioinformatics) with default parameters and only uniquely mapping reads were selected for further analysis. Ribosomal and mitochondrial reads were removed using modified scripts from the PORT pipeline (https://github.com/itmat/Normalization).

Creation of non-overlapping gene, exon and intron annotations

Annotations were based on the comprehensive gene annotation file of the GENCODE Release 22 (GRCh38.p2). The GTF file downloaded from GENCODE (<http://www.gencodegenes.org/releases/22.html>”) was loaded into R (Version 3.2.2; R Core Team (2015): “R: A Language and Environment for Statistical Computing”, R Foundation for Statistical Computing, Vienna, Austria) and converted into a TranscriptDb object with the makeTranscriptDbFromGFF tool in the Bioconductor package GenomicFeatures (Version 1.22.8; Lawrence et al. 2013 PLoS Computational Biology). From the TranscriptDb object I extracted all annotated genes (Ensembl gene id) and their exons as a GRanges object using the tool exonsBy (with option by=”gene”). Ensembl gene ids were replaced by official gene symbols if available using the package biomaRt (Version 2.26.1) otherwise Ensembl gene ids were kept. Thereby, genes with several Ensembl gene ids were combined into one record. For each gene duplicated exons were removed and overlapping exons were combined to single, longer exons. Next, genes were defined as encompassing the entire sequence between the first base of the first exon and the last base of the last exon and the annotation was stored in a separate GRanges object. Next, regions shared by overlapping genes were removed to only count reads that map to one gene as our sequencing libraries were not strand-specific.

Counting script is included in the folder.