Input:

 Sample data (FASTQ)

1. **QC**

- Stage 1
 - Fastq quality scores
 - Sequencing depth
 - Nucleotide composition
 - Strand bias

Input:

- · Genome reference (FASTA)
- Annotations (FASTA / GFF)

Stage 2

- · Jaccard similarity index
- Complexity analysis
- · Point to point correlation



2. Genome & annotations alignments



3. Normalization

- Identification of optimal normalization:
 - Scaling (RPKM, TMM (edgeR), DEseq2)
 - Rank based (quantile)
 - Subsampling and bootstrapping, followed by a quantile correction
- QC on normalized data:
 - Jaccard similarity index, complexity analysis, Point to point correlation between replicates
 - Point to point correlation with the riginal data (for the subsampling normalization)

4. Calculation of differential expression (DE)

- Analysis of within replicate set and between treatment variability
- Identification of hierarchy levels for the differential expression
- Calculation of offset fold change (OFC)
- · Identification of differentially expressed genes

5. Functional annotation / enrichment

- Annotation of differentially expressed genes
- Enrichment analysis

6. Validation

 Low throughput validation of differentially expressed

S1 Fig. Analysis framework for the *D. melanogaster* mRNA-seq data.

Required inputs (sequencing data in FASTQ format, and the corresponding reference genome and transcriptome in FASTA/GFF) and the six main steps of the analysis are shown in a workflow diagram, following Conesa et al. 2016 (Genome Biology, 17:13). The steps, for which additional details are included, are: Quality check (QC), alignment, normalization of gene abundances, identification of DE, functional enrichment and finally low-throughput validation