Bench sheet · EnGeL-BIOINF-RNASEQ

RNA sequencing: from reads to differentially expressed genes and gene ontology (FastQC, Trimmomatic, STAR, featureCounts, DESeq2, DAVID) · version 1.1.2 (2026-11-20). Full protocol: EnGeL-BIOINF-RNASEQ. Tick each step when done; write deviations in Notes.

Date	
Operator	
Plate ID / samples	
Sequencing provider, order and run IDs	
Genome assembly and annotation versions; software versions	

1. Quality check of the raw reads (FastQC)

☐	No.	Step
☐	1	Run FastQC on all raw files and look at the reports (per-base quality, adapter content, overrepresented sequences):

2. Trimming (Trimmomatic)

☐	No.	Step
☐	2	Trim each pair. Adapters are removed with the lab's adapter file (`ILLUMINACLIP` 2:30:10); then bases below quality 3 at either end, windows of 4 bases with mean quality below 15, and reads shorter than 36 bases:
☐	3	Run FastQC again on the paired outputs (`fastqc -t $THREADS -o 1.qc/trimmed 2.trimmed/*_paired.fq.gz`) and check that adapter content has gone. Only the `_paired` files are used from here on.

3. Genome index (STAR)

☐	No.	Step
☐	4	Build the index once per genome and annotation version. `--sjdbOverhang` is the read length minus 1 (149 for 2 × 150 bp reads). `--genomeSAindexNbases` is set from the genome size, as the STAR manual advises: min(14, log2(genome length)/2 − 1), rounded down (about 12 for a 250 Mb genome, 14 for genomes of about 1 Gb or more):

4. Alignment (STAR)

☐	No.	Step
☐	5	Align each sample and write a coordinate-sorted BAM file directly, named after the sample:
☐	6	For each sample, read `4.mapping/<sample>_Log.final.out` and record the input reads and the percentage of uniquely mapped reads.

5. Counting reads per gene (featureCounts)

☐	No.	Step
☐	7	Count read pairs (fragments) per gene for all samples in one table. The libraries are not strand-specific (`-s 0`):
☐	8	Open `5.counts/featureCounts_output.txt.summary` and record the percentage of assigned reads per sample.

6. Differential expression (DESeq2)

☐	No.	Step
☐	9	Import the counts and the sample table:
☐	10	Fit the model and check the dispersion plot: the fitted line should follow the cloud of gene estimates, which should fall towards it as mean counts increase:
☐	11	Transform the counts for plots and look at how the samples group. Use rlog for usual experiments; for large data sets (more than about 30 samples) use `vst(dds, blind = TRUE)` instead, which is much faster and gives practically the same plots. The transformation is used only for plots, never for the test. Look at how the samples group, with a sample-distance heatmap and a principal component analysis:
!		**CRITICAL** Replicates are never removed only because they look different. A sample is excluded only for a documented laboratory reason (for example a known extraction, library or sample-handling problem, or a sample swap confirmed by other evidence), the decision and reason are recorded, and the analysis with all samples is kept and reported. A sample that groups away from its replicates is a reason to check the laboratory records, not a reason to drop it.
☐	12	List the coefficients and get the results for each planned contrast, with α = 0.05 for independent filtering:
☐	13	Shrink the fold changes (apeglm) for ranking, plots and the fold-change threshold. apeglm works on the coefficients listed by `resultsNames`, i.e. each condition against the reference; for other comparisons (e.g. high vs low) either set that level as the reference and refit, or use `type = "ashr"` with `contrast = c("condition", "high", "low")`:
☐	14	Call genes differentially expressed with the lab's criteria, **padj < 0.05 and |log2 fold change| ≥ 0.5**, taking padj from `results()` and the fold change from the shrunken estimates, so that a gene passes only if its change is both large and well supported, and write the tables:
☐	15	Draw a volcano plot from the same table:
☐	16	Repeat the last four steps for each planned contrast.
☐	17	If a functional annotation table exists for the genome (gene ID, gene name, description, GO, KEGG), join it to the results by gene ID before writing the final tables, and record its source and version.

7. Gene ontology and pathway enrichment (DAVID)

☐	No.	Step
☐	18	For each contrast, write three gene lists as plain text, one identifier per line: genes up, genes down, and the **background** of all genes tested (genes with a padj value, i.e. not removed by independent filtering):
☐	19	Check that the identifiers are of a type DAVID recognises for the species (for example NCBI gene IDs or gene symbols). Convert them if needed and record the conversion.
☐	20	In DAVID, upload the up list as a gene list and the background list as background, select the species, and run the functional annotation chart for GO biological process, molecular function and cellular component, and KEGG pathways. Repeat for the down list. Export the charts.
☐	21	Report terms with a Benjamini-adjusted p-value below 0.05. Use both the functional annotation chart (individual terms) and the functional annotation clustering (groups of related, redundant terms, ranked by enrichment score) to see which processes stand out. Record the DAVID version and knowledgebase date.

8. Branch: species without a reference genome

[TO CONFIRM: protocol to be written.] Without a genome the reads are assembled into a transcriptome (de novo assembly), transcripts are quantified and grouped into genes, and the transcriptome is annotated by similarity to protein databases before the same DESeq2 and enrichment steps (6 and 7). Tools, settings and quality checks for this branch are to be decided by the lab.

Notes / deviations
