Bench sheet · EnGeL-BIOINF-SANGER

Sanger barcode sequences: quality control, sample matching and identification with BOLD and NCBI BLAST · version 1.2.0 (2026-11-20). Full protocol: EnGeL-BIOINF-SANGER. Tick each step when done; write deviations in Notes.

Date	
Operator	
Plate ID / samples	
Sequencing provider order or plate codes	
Script versions and software versions; date of BOLD and BLAST queries	

1. Organise the data

☐	No.	Step
☐	1	Make a project folder with `scripts/`, the sample sheet and the taxonomy table, and copy each plate (or tube order) into its own folder under `data/sanger/`, keeping the provider's folder name (the plate code).
☐	2	Keep the original AB1 files read-only. The pipeline never changes them.

2. Quality control and sample matching

☐	No.	Step
☐	3	Run QC on all folders at once:
☐	4	Open `qc_<date>/merged_results.csv` and the figures (`fig_lengths.png`, `fig_meanQ_vs_len.png`, `fig_pass_rate.png`). Check the reasons given for REPEAT reads.
☐	5	Check `recovered_report.csv`: every row of the sample sheet should have a read. Rows without a read and reads without a row (in `merged_pass_unmatched.fasta`) usually come from a typing error in the sample sheet or a missing file from the provider. Correct the sample sheet, not the file names, and run again.
!		**CRITICAL** Use the same QC thresholds for all plates of a study and record them. The lab values are: minimum trimmed length 200 bp, at least 200 bases with Q ≥ 20, mean Q ≥ 20, at most 5% N (trimming cutoff Q 20). A read of 200 good-quality bases is enough for a reliable hit. The script defaults are stricter (400 bp and 300 bases with Q ≥ 20) and must be overridden as above.
☐	6	Send REPEAT specimens back for re-sequencing (or a new PCR) and record the decision.

3. Identification against BOLD (BOLDigger3)

☐	No.	Step
☐	7	Run BOLDigger3 on the PASS sequences:

4. Identification against GenBank (NCBI BLAST)

☐	No.	Step
☐	8	Run remote BLAST on the same file:
!		**CRITICAL** Use version 0.1.4 or later. In version 0.1.3 the default pattern that reads the sample ID from the FASTA headers was wrongly escaped and never matched, so, unless `--id-regex` was given, the script's own table of BLAST hits with metadata was not joined to the sample sheet. The final table (step 5) joins on plate and well and was not affected.

5. Final table

☐	No.	Step
☐	9	Join the sample sheet, QC results and both identifications:
☐	10	Read the unmatched and duplicate files for each step. They should be empty, or explained (for example REPEAT reads, which are not sent to BOLD or BLAST).

6. BOLD–BLAST concordance

☐	No.	Step
☐	11	Classify each sequence:

Class	Meaning
PERFECT_CONCORDANCE	BOLD and BLAST both give the putative species at or above the species threshold (with `--single-tool-perfect-as-perfect`: one tool does and the other has only a weak, low or no hit)
PARTIAL_CONCORDANCE	One tool gives the putative species at or above the species threshold; or genus or family agree at or above their thresholds
GENUS_CONCORDANCE	Genus agreement only (shown as a separate class only with `--genus-as-own-category`)
LOW_CONCORDANCE	Hits exist but identity is below the low threshold
NO_MATCH	Neither tool gives a usable hit

Class	Meaning
POSSIBLE_CONTAMINATION	A hit at or above the species threshold to another species present in the batch, with genus (and family, when known) also disagreeing; with `--require-both-tools-for-contam`, both tools must show this. Only applied when the putative name is a full binomial (not sp., spp. or cf.)

☐	No.	Step
☐	12	Check `contamination_table.csv` and the LOW and NO_MATCH lists with the person who identified the specimens. A discordant result is a question about the specimen, the sequence or the reference database; it is not a re-identification in itself.

7. Status report

☐	No.	Step
☐	13	Write and render the report:
!		**CRITICAL** `--homology-dir` must point to the output of step 6 from the same run. A report built from an older concordance folder will show out-of-date classes without any warning.

Notes / deviations
