ENVIRONMENTAL GENOMICS LABORATORY

STANDARD OPERATING PROTOCOL · EnGeL-BIOINF-SANGER

Sanger barcode sequences: quality control, sample matching and identification with BOLD and NCBI BLAST

Protocol ID	EnGeL-BIOINF-SANGER
Version	1.2.0
Date	2026-11-20
Status	draft
Area	Bioinformatics (BIOINF)
Authors	Luís Cunha (lead; pipeline scripts and manuals; review and approval; ORCID 0000-0002-5870-2537)
Reviewed and approved by	Luís Cunha
Laboratory	Environmental Genomics Laboratory (EnGeL-UC), Environmental Ecology & Ecotoxicology Group (E3G), Centre for Functional Ecology, Department of Life Sciences, University of Coimbra · Laboratório Associado TERRA
Contact	luis.cunha@uc.pt · ewormuc@gmail.com
Licence	Text: CC BY 4.0 · Code: MIT
Main hazards	None: computational protocol

1 Purpose and scope

This protocol takes the chromatogram files (AB1) returned by a Sanger sequencing provider for single-specimen barcodes and produces, for every specimen, a quality decision, a trimmed sequence, identifications from BOLD and from NCBI BLAST, a statement of how far the two agree with the identification made from morphology, and a project status report. It follows barcode PCR and Sanger sequencing in the single-specimen barcoding workflow.

It works for any marker and taxon with a reference in BOLD or GenBank. Identity thresholds depend on the marker and the group and must be set for each study (Quality control).

Out of scope: manual editing of chromatograms, assembly of forward and reverse reads into contigs, alignment and phylogenetic analysis, and submission to BOLD or GenBank.

2 Principle

Each chromatogram is read with its base-call quality scores (Phred Q). The low-quality ends are removed by Mott trimming: the script keeps the stretch of the read where the sum of (Q − cutoff) is highest. The trimmed read is then tested against minimum length, mean quality, number of bases with Q ≥ 20 and maximum percentage of ambiguous bases (N). Reads that pass all four are PASS; the others are REPEAT, with the reason recorded.

The file name of each chromatogram carries the provider's plate code and well (or tube code). These are matched to the lab's sample sheet, so every sequence gets the specimen ID, primer pair and putative species written into its FASTA header. Reads that pass but cannot be matched are kept apart, never mixed with matched reads.

The PASS sequences are identified twice, independently: against BOLD with BOLDigger3, and against GenBank nt with remote BLAST. A final table joins the sample sheet, the QC results and both identifications. Each sequence is then placed in a concordance class by comparing the BOLD and BLAST hits with the putative species: perfect, partial, genus-level, low, no match, or possible contamination (a high-identity hit to another species present in the same batch, with genus and family also disagreeing).

3 Before you start

Item	Details
Input	One folder per plate (or per tube order) with the AB1 files as delivered by the provider
Sample sheet	Excel (.xlsx) or CSV with one row per well or tube: `plateCODE`, `sequenceID` (well, e.g. A01, or tube code), `PrimersID`, `SampleIDOriginal`, `PutativeSpeciesName`, `PutativeFamilyName`; optional `species`, `primerF`, `primerR`
Taxonomy table	CSV with `Family`, `Genus`, `Species` for the group studied; used to recognise genus and family agreement and possible contamination
Computing	Internet access for BOLD and NCBI; hands-on and run time [TO CONFIRM]
Prepare in advance	Software installed (Reagent preparation); sample sheet checked against the provider's plate or tube list

CRITICAL The sample sheet is the link between a chromatogram and a specimen. Check plate codes and well or tube codes before running: a typing error in either leaves the read unmatched (it goes to `pass_unmatched.fasta`), and a swapped code assigns the read to the wrong specimen without warning.

4 Materials

4.1 Computer

- Lab workstation or a personal computer (Windows PowerShell, macOS or Linux shell) with internet access for BOLD and NCBI queries

4.2 Consumables

Not applicable.

4.3 Software

- Python 3 with biopython, numpy, pandas, matplotlib, openpyxl and xlsxwriter [TO CONFIRM: Python and package versions]
- BOLDigger3 (pip package boldigger3, with duckdb) [TO CONFIRM: version]
- NCBI BLAST+ (blastn, run in remote mode against nt) [TO CONFIRM: version]
- Quarto, to render the status report [TO CONFIRM: version]
- EnGeL Sanger pipeline scripts: sanger_folder_qc_IDmatch (v1.5; plate and tube variants), run_boldigger3_merged (v0.3.9), run_ncbi_blast_merged (v0.1.4), make_final_tables (v0.3), compare_blast_bold_status (v0.1.7.3), sanger_quarto_status_report (v0.8.2) [TO CONFIRM: name and DOI of the public repository]

5 Reagent preparation

Not applicable. Software installation (from the script manuals):

```
pip install biopython numpy pandas matplotlib openpyxl xlsxwriter   # QC and tables
pip install boldigger3 duckdb                                       # BOLD identification
```

NCBI BLAST+ is installed from NCBI (the `blastn` program must be on the PATH, or given with `--blastn-path`). Quarto is needed only to render the report.

6 Procedure

Commands are shown for a Unix shell; in Windows PowerShell use `.\` and `\` in paths and the backtick (`` ` ``) for line continuation. Replace names in angle brackets.

6.1 1. Organise the data

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.

> **Tube orders** Plate orders carry a plate code and a well (A01–H12) in the file name. For tube orders use `sanger_folder_qc_IDmatch_tubes_v1.5.py`, which accepts files named only by tube code, or give your own pattern with `--id_regex` (named groups `plateCode` and `sequenceID`). Alternatively, add a batch prefix to the tube codes in the sample sheet so that the plate version can be used.

6.2 2. Quality control and sample matching

3. Run QC on all folders at once:

```
python scripts/sanger_folder_qc_IDmatch_v1.5.py \
  --indir data/sanger/<plate1> data/sanger/<plate2> \
  --outdir qc_<date> \
  --map_file <sample_sheet>.xlsx --map_sheet Sheet1 \
  --min_length 200 --min_mean_q 20 --min_q20 200 --max_n_pct 5 \
  --min_fasta_len 50 -v
```

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.

PAUSE POINT. The QC folder can be kept indefinitely; the next steps need only `merged_pass.fasta` and `merged_results.csv`.

6.3 3. Identification against BOLD (BOLDigger3)

7. Run BOLDigger3 on the PASS sequences:

```
python scripts/run_boldigger3_merged_v0.3.9.py \
  --fasta qc_<date>/merged_pass.fasta \
  --metadata-xlsx <sample_sheet>.xlsx --meta-sheet Sheet1 --meta-id-col SampleIDOriginal \
  --db 1 --mode 2 \
  --thresholds 97 95 90 80 \
  -o boldigger3_<date> --fresh
```

The lab standard is database 1 (animal library, public records) with mode 2 (genus and species search), and thresholds of 97%, 95%, 90% and 80%: the identities BOLDigger3 requires to accept a name at species, genus, family and order level. Record any change.

Other choices, for other questions or groups: databases 2 and 3 add private records (2 keeps only records named to species); 4 is the validated Canadian arthropod library; 5 plants, 6 fungi, 7 secondary animal markers, 8 the validated animal Red List library. Mode 1 (rapid species search) is faster but only looks for close species-level matches; mode 3 (exhaustive) searches more widely and is slower. Check the BOLDigger3 documentation for the version installed, as the list changes with BOLD. Use `--resume` to continue an interrupted run without repeating finished queries.

6.4 4. Identification against GenBank (NCBI BLAST)

8. Run remote BLAST on the same file:

```
python scripts/run_ncbi_blast_merged_v0.1.4.py \
  --fasta qc_<date>/merged_pass.fasta \
  --metadata-xlsx <sample_sheet>.xlsx --meta-sheet Sheet1 --meta-id-col SampleIDOriginal \
  -o ncbi_blast_<date>
```

Defaults: megablast against nt, E-value 1e-20, 5 target sequences per query. Before the search the script writes a cleaned copy of the FASTA (`NCBI_sanitized_<input>.fasta`, with a short report): non-ASCII characters and spaces in the headers are replaced and headers broken over several lines are rejoined, because remote BLAST rejects or misreads such headers. The original file is not changed; `--no-sanitize-fasta` turns this off. `--entrez-query` limits the search to a group (for example a class or order name followed by `[ORGN]`); record it if used. The script prints the exact `blastn` command so that it can be repeated.

> **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.

> Remote BLAST is shared with all NCBI users and can be slow for large batches. Run it outside working hours or in smaller files if it times out.

6.5 5. Final table

9. Join the sample sheet, QC results and both identifications:

```
python scripts/make_final_tables_v0.3.py \
  --base-xlsx <sample_sheet>.xlsx \
  --qc-table qc_<date>/merged_results.csv \
  --bold-file boldigger3_<date>/boldigger3_output.xlsx \
  --blast-file ncbi_blast_<date>/blast_best_hit_per_query.csv \
  --outdir final_table
```

The join is made in three steps (sample sheet + QC, + BOLD, + BLAST) on plate code and well. After each step the script writes the unmatched rows and duplicate keys (`step*_unmatched_*.csv`, `step*_duplicate_keys_*.csv`) and a metrics file. The final table is `final_table/step3_final_with_qc_bold_blast.xlsx`.

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).

6.6 6. BOLD–BLAST concordance

11. Classify each sequence:

```
python scripts/compare_blast_bold_status_v0.1.7.3.py \
  --final-xlsx final_table/step3_final_with_qc_bold_blast.xlsx \
  --outdir homology_<date> \
  --species-hi 97 --genus-hi 90 --family-hi 85 --low-thresh 85 \
  --id-col SampleIDOriginal --primer-col PrimersID \
  --putative-col PutativeSpeciesName --putative-family-col PutativeFamilyName \
  --bold-species-col species --bold-ident-col pct_identity \
  --blast-species-col blast_stitle --blast-ident-col blast_pident \
  --taxonomy-map-csv <taxonomy_table>.csv \
  --require-both-tools-for-contam --single-tool-perfect-as-perfect
```

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
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.)

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.

6.7 7. Status report

13. Write and render the report:

```
python scripts/sanger_quarto_status_report_v0.8.2.py \
  --qc_outdir qc_<date> --report_name "<Project> Sanger status report" \
  --prefer_putative_species \
  --metadata_xlsx <sample_sheet>.xlsx --meta_sheet Sheet1 \
  --meta_species_col PutativeSpeciesName --meta_family_col PutativeFamilyName \
  --meta_original_id_col SampleIDOriginal \
  --homology-dir homology_<date>
quarto render qc_<date>/sanger_status_report.qmd --to html
```

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.

7 Calculations

Mott trimming keeps the sub-sequence $[i, j)$ of the read that maximises

$$S = \sum_{k=i}^{j-1} (Q_k - Q_{\text{cut}})$$

where Q_k is the Phred quality of base k (dimensionless) and Q_{cut} the trimming cutoff (default 20). On the trimmed read the script reports length (bp), mean Q, the number of bases with $Q \geq 20$ and the percentage of N.

Worked example: a read with qualities 10, 15, 30, 35, 32, 18, 12 and $Q_{cut} = 20$ gives scores −10, −5, 10, 15, 12, −2, −8. The best stretch is bases 3 to 5 (S = 37), so 3 bases are kept, with mean Q 32.3 and 3 bases with $Q \geq 20$.

8 Quality control

Check	Where	Acceptance criterion
Read quality (PASS/REPEAT)	`merged_results.csv`	PASS: trimmed length ≥ 200 bp, ≥ 200 bases with Q ≥ 20, mean Q ≥ 20, ≤ 5% N
Every specimen has a read; every read has a specimen	`recovered_report.csv`, `merged_pass_unmatched.fasta`	No unexplained rows
Joins complete	`final_table/step*_unmatched_*.csv`, `step*_duplicate_keys_*.csv`	Empty, or explained
Identification thresholds	Steps 3 and 6	BOLD: 97/95/90/80 (species/genus/family/order); concordance: species 97, genus 90, family 85, low 85; any change recorded
Possible contamination	`contamination_table.csv`	Every case reviewed and the decision recorded

9 Data and metadata

- **Outputs:** QC folder (per-read metrics, PASS/REPEAT FASTAs, figures), BOLDigger3 and BLAST output, final table (`step3_final_with_qc_bold_blast.xlsx`), concordance tables and the HTML status report.
- **Where:** project data folder [TO CONFIRM] AB1 files kept with the raw data. Accepted barcodes are submitted to BOLD or GenBank with the voucher information [TO CONFIRM: lab practice].
- **Code:** the scripts are published in their own public repository (MIT licence) with an archived, citable version [TO CONFIRM: repository name and DOI] cite the version used.

Minimum to record

- Provider, order, plate or tube codes and sequencing date
- Specimen ID (`SampleIDOriginal`), primer pair (`PrimersID`), putative species and the person who identified it
- QC thresholds, BOLD database, mode and thresholds, BLAST settings (including any `--entrez-query`), concordance thresholds
- Script versions, software versions and the date of the BOLD and BLAST queries (both databases change over time)
- Decision for every REPEAT, LOW, NO_MATCH and POSSIBLE_CONTAMINATION specimen

10 Safety and waste

Not applicable (computational protocol).

11 Troubleshooting

Problem	Likely cause	Solution
Many sequences in `merged_pass_unmatched.fasta`	Plate or well codes in the sample sheet differ from the file names	Compare file names with the sample sheet; for tube orders use the tube script or `--id_regex`.

Problem	Likely cause	Solution
Sample sheet rows without a read	Typing error, or the provider did not return the file	Check the codes; ask the provider for the missing files.
Extra codes in the provider's folder	Wells or tubes sent that are not in the sample sheet	Add them to the sample sheet or record them as not part of the study.
Many REPEAT reads with `trimmed_length` as reason	Poor PCR product, primer problem, or thresholds too strict for a short marker	Look at the chromatograms; check the marker length against `--min_length`.
BOLDigger3 stops part way	Connection to BOLD lost	Run again with `--resume`.
BLAST very slow or times out	Remote BLAST queue	Split the FASTA, or run at a quieter time.
Report classes do not match the concordance table	Report built from an older homology folder	Rebuild with `--homology-dir` pointing to the current run.

12 Remarks

Most errors in Sanger barcoding come from the link between tube and file, not from the sequence: keep the sample sheet as the single source of truth and correct it, never the file names. BOLD and GenBank differ in coverage and in the quality of their names; agreement between the two, and with the morphological identification, is what the concordance classes report. Identity thresholds are a convention that varies between groups and markers, so they are recorded with the results.

Parts of the scripts were written with the help of a generative AI assistant, as stated in their headers.

13 References

Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13(4), 477–478. [TO CONFIRM: citation used for Mott trimming]

[TO CONFIRM: references for BOLDigger3, BOLD Systems, BLAST+ and Biopython.]

14 Version history and citation

Version	Date	Change	By
1.0.0	2026-11-20	First EnGeL-UC version, from the Sanger pipeline scripts and Quarto manuals written by Luís Cunha (2025); project-specific settings removed.	Luís Cunha
1.1.0	2026-11-20	NCBI BLAST step uses run_ncbi_blast_merged v0.1.4 (FASTA headers cleaned for remote BLAST by default; default sample-ID pattern corrected); code to be published in its own public repository with a citable version; use of a generative AI assistant in writing the scripts stated in Remarks.	Luís Cunha

Version	Date	Change	By
1.2.0	2026-11-20	Values confirmed by Luís Cunha: read QC at minimum trimmed length 200 bp, at least 200 bases with Q ≥ 20, mean Q 20, at most 5% N; BOLDigger3 database 1 and mode 2 as standard (other databases and modes explained), thresholds 97/95/90/80; concordance thresholds kept (species 97, genus 90, family 85, low 85); no PCR blank is sequenced.	Luís Cunha

How to cite

Cunha L (2026). Sanger barcode sequences: quality control, sample matching and identification with BOLD and NCBI BLAST. EnGeL-UC protocol EnGeL-BIOINF-SANGER, version 1.2.0. Zenodo. DOI [TO CONFIRM].

Licence

Text: CC BY 4.0. Code: MIT. Logos are the property of their institutions and are not covered by the licence.

15 Open questions

1. Software and package versions; hands-on and run time.
2. Submission of accepted barcodes to BOLD or GenBank: lab practice.
3. Name of the public code repository and its DOI.
4. Where outputs are kept.
5. Citations for the tools; DOI of the protocol.