BioR — SUPER Challenge (Ecology): Intertidal data integration with tricks

Tidy + join + QC + hybrid objects (nested data + one model per site)

Scenario

You sampled intertidal sites with a foundation-species canopy (CANOPY) and nearby open patches (OPEN). Biomass was recorded in a spreadsheet (wide format). Microclimate (temperature and humidity) was logged at high frequency with separate loggers. Microbiome swabs were sequenced and delivered as an ASV table. Your task is to integrate everything into an analysis-ready dataset, create explicit QC flags, and build a hybrid table with one linear model per site.

Dataset folder

Use this folder (contains all input files):
/mnt/Hard_intertidal_challange

Input files

- biomass_wide.csv (wide biomass; includes ND/dead/blanks; sample_id separators vary)
- quadrat_meta.xlsx (Excel; canopy_cover_pct may be '45%' or impossible values; sample_id duplicates)
- logger_raw.csv (mixed timestamp formats; duplicated lines; drift)
- logger_map.csv (logger_id has trailing spaces and case mismatches)
- asv_table.csv (microbiome long table; samples can appear in multiple runs; some low depth)

Deliverables

- outputs/clean_master.csv (sample_id × day with biomass + meta + microclimate + microbiome proxies)
- outputs/qc_report.csv (one row per sample_id with QC flags)
- outputs/site_models.csv (one row per site: slope, intercept, r2, n for biomass ~ microclimate)
- outputs/plots/01_biomass_timeseries.png
- outputs/plots/02_biomass_vs_temp.png (or vs humidity)
- outputs/plots/03_qc_flags.png
- scripts/super_challenge_intertidal.R (your script)

Tasks

- A) Tidy biomass: pivot_longer day_* -> (day, biomass). Day must be numeric.
- - Convert biomass to numeric; treat ND/dead/blanks as NA; create qc_non_numeric_biomass flag.
- - Parse sample_id into site, habitat, rep. Handle '_' and '-' separators, mixed case, and R01 vs R1.
- B) Quadrat metadata: read the Excel file; parse canopy_cover_pct into numeric; flag impossible values (>100).

- - Detect duplicated sample_id rows in meta; decide a strategy and document it (qc_meta_duplicate).
- C) Microclimate: join logger_map to logger_raw; parse timestamps robustly; aggregate mean temp and mean RH per site×habitat×day (days 0,7,14).
- - Join microclimate into biomass using left_join; also try inner_join and report row counts and what changes.
- D) Microbiome proxies: combine runs by summing reads per sample_id×asv_id; compute library_size and richness per sample_id.
- - Flag low depth samples: qc_low_depth = library_size < 1000.
- E) Join all sources into one master table (sample_id×day). Make missing combos explicit with complete() for days {0,7,14}.
- F) Make a QC report (one row per sample_id) that summarizes key flags.
- G) Hybrid models: one row per site with nested data and fit = lm(biomass ~ mean_temp) (or mean_rh). Extract slope, intercept, r2, n.

Tricks to watch for (common failure modes)

- Joins fail if you do not standardize keys first (sample_id/logger_id).
- Timestamp parsing will fail unless you handle multiple formats.
- Meta has percent strings and duplicates; decide how to resolve duplicates (and flag).
- ASV table may contain multiple runs per sample; combine before computing proxies.
- lm() drops NA rows: filter before fitting and report n.