Metadata Management Strategy for RNA-seq Projects

Goal: Robust, reproducible handling of large clinical + molecular metadata (~250 columns × 700+ samples), allowing easy updates and new sample integration.

1. Initial Cleaning & Consolidation

- Load raw metadata (read_csv / readRDS).
- Explore dataset: skimr::skim(), DataExplorer::create_report().
- Sanity check sample IDs (unique, no missing, match with expression matrix).
- Standardize column names: janitor::clean_names(), optional manual renaming.
- Fix data types: convert characters, factors, numerics; trim whitespace.
- Handle missing values: visualize (naniar::vis_miss()), impute / label unknown / drop column.
- Check redundancy: caret::nearZeroVar(), cross-check for duplicate info.

2. Handling Inclusion Flags / Binary Columns

- Detect In* columns using regex.
- Determine exclusivity: rowSums() <=1 -> collapse into single categorical variable.
- Collapse exclusive sets: create RNA_set.
- Retain non-exclusive sets as logical flags (is_in_*).
- Validate consistency: no sample assigned to all sets unless expected.

3. Domain-Aware Metadata Object

- Structure: meta_obj = list(data = meta_tidy, domains = domain_df).
- domain_df example: variable | domain (Age -> clinical, Stage -> clinical, etc.).
- Helper function get_domain(meta_obj, "clinical") returns subset (always includes sample_id).
- Supports multiple domains, optionally returns just variable names.

4. Repeatable Metadata Update Workflow

- Column renaming: maintain column_map.csv, old -> new names.
- Schema-driven validation: define schema tibble (variable, type, allowed, action).
- check_metadata() validates new data; auto-fix or warn for mismatches.
- extend_metadata(meta_obj, new_data, schema): validate, fix, append, check duplicates.
- Optional logging: record renamed columns, fixed values, warnings.

5. Suggested Logic Flow (Checklist)

Incoming new metadata:

1. Load new metadata.
2. Apply column renaming map.
3. Standardize column names (janitor::clean_names()).
4. Validate against schema (type, allowed values, missing fields).
5. Auto-fix common issues (e.g., gender abbreviations, factor levels).
6. Append new data; check for duplicate sample IDs.
7. Update domain-aware object (domain_df) if new columns appear.

8. Save updated metadata object (saveRDS).
9. Generate optional report / summary (skimr::skim() or custom log).

6. Best Practices

- Keep raw metadata separate from cleaned data.
- Version control: column map, schema, domain definitions.
- Maintain data dictionary (column definitions, units, expected values).
- Modular functions: renaming, validation, extension.
- Log warnings and auto-fixes for reproducibility.

Deliverables / Outputs

- Cleaned metadata object (meta_obj) with data + domains.
- Schema for validation.
- Column renaming map.
- Domain helper functions (get_domain(), list_domains()).
- Version-controlled, repeatable update workflow.