Description of Data for Harmonization Based on Zethoven et al.

Here we provide a set of data files for developing and assessing data harmonization tools. The challenge is to harmonize sample metadata from ten sources. Unlike in Use Case Exercise 1, we do not specify a format to harmonize into. Instead, you may choose a format, such as GDC format or a format used in one of the source papers, or treat creating a format to capture the data as part of the challenge.

This is based on data gathered in the paper Zethoven et al. 2022 [1], which was the inspiration for a use case in Use Case Exercise 4 and corresponds to Data Request Prompt #1 in the file 'Research Data Request Prompts.xlsx'. Zethoven et al. assembled gene expression data for Pheochromocytoma and Paraganglioma (PCPG) samples from prior projects. This paper's results section briefly describes the purpose of this dataset:

> *Despite the utility of snRNA-seq to identify genotype associations among PCPG, the limited number of biological replicates among genotypes limited the generalizability to the broader PCPG population. We therefore sought to integrate our snRNA-seq data with a large compendium of published microarray and RNA-seq data (n = 735 samples) (Supplementary Data 3).*

The integrated sample metadata are in Zethoven et al.'s Supplementary Data 3. We do not recommend treating this file as an example harmonized data set to compare against because it includes some data that were calculated via molecular data analyses as part of the Zethoven et al. study, has some redundant variables, and has a few variables that were specific to single source datasets. Furthermore, there is no data dictionary associated with this file and the values for at least two variables, "PCPG" and "purity" (the percent or fraction of a sample that is tumor cells rather than normal cells), are not normalized and standardized. However, there are a few variables with names that include "raw" that also have corresponding variable versions with normalized values; among these, the variable "Genotype" appears to make use of HGNC gene symbols. We have created an abbreviated version of S3, in a file called "integrated_dataset-abbreviated_version_of_Zethoven_et_al_S3.csv", with some variables removed and value standardization for 'purity' (but not normalization for PCPG), which could potentially be used for comparison.

Source datasets

The data sources used in Zethoven et al. include:

- Seven GEO accessions: GSE2841, GSE19422, GSE19987, GSE39716, GSE50442, GSE51081, and GSE67066. These GEO files do not all include the same variables, and the values for some variables are not normalized across files.
- ArrayExpress accession E-MTAB-733

- Data associated with the paper Flynn et al. 2015 [2]. Sample metadata appears in Flynn et al. Table S1.
- TCGA PCPG cohort data, some of which appears to be in the Genomic Data Commons (GDC) and some of which appears to be from Table S2 in Fishbein et al. 2017 [3].

For the datasets in repository sources, here is information about how to obtain them:

- The data for the GEO accessions (e.g., GSE2841) can be obtained using the R script in this folder, ‘script_to_get_GEO_datasets.R’.
- E-MTAB-733 is a project dataset posted in the European Bioinformatics Institute (EBI) repository. The metadata can be copied from the table, with all rows displayed, from here: Samples and Data < E-MTAB-733 < ArrayExpress < BioStudies < EMBL-EBI.
- The GDC TCGA-PCPG datasets can be obtained using the following faceted search options in the GDC repository user interface:
  - Cohort Builder Query: PROJECT IS TCGA-PCPG and EXPERIMENTAL STRATEGY IS RNA-Seq and TISSUE TYPE IS tumor
  - Repository Query: EXPERIMENTAL STRATEGY IS RNA-Seq and TISSUE TYPE IS tumor

For the two datasets from papers, here is information about where you can obtain them:

- Flynn et al. 2015 The genomic landscape of phaeochromocytoma - The Journal of Pathology [2]
  - Table S1. Clinical information. https://pathsocjournals.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fpath.4503&file=path4503-sup-0011-TableS1.xls
  - Accessing this paper requires a journal subscription, and its license might not be consistent with machine processing. However, the Supporting Information files are freely available.

- Fishbein et al. 2017 Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma: Cancer Cell [3]
  - Table S2. Individual Sample Level Data https://www.cell.com/cms/10.1016/j.ccell.2017.01.001/attachment/3036e7bb-acf1-42b0-abd3-10366e71fb08/mmc3.xls
  - This paper was identified as a source of data via a citation in the methods section in Zethoven et al. This covers the same project as the data from GDC. Note, while this paper is open access for human reading, its license type does NOT appear to be consistent with open machine processing.

Some tips about the GDC TCGA-PCPG data: There are multiple metadata files from the GDC for tumor samples with RNA-Seq data from the TCGA-PCPG project. The values for the Zethoven et al. S3 variable “Sample.raw” can be used to find associated file metadata in the “metadata.repository” json file using the variable there “entity_submitter_id”, and “Sample.raw” values can be used to find sample information in the “biospecimen.cohort” json file using the variable there “submitter_id”. There are also additional files with clinical metadata, biospecimen metadata, and file metadata that may be useful.

References

[1] M. Zethoven *et al.*, “Single-nuclei and bulk-tissue gene-expression analysis of pheochromocytoma and paraganglioma links disease subtypes with tumor microenvironment,” *Nat Commun*, vol. 13, no. 1, p. 6262, Oct. 2022, doi: 10.1038/s41467-022-34011-3.

- This paper is available with a CC-BY-4.0 license

[2] A. Flynn *et al.*, “The genomic landscape of phaeochromocytoma,” *The Journal of Pathology*, vol. 236, no. 1, pp. 78–89, 2015, doi: 10.1002/path.4503.

[3] L. Fishbein *et al.*, “Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma,” *Cancer Cell*, vol. 31, no. 2, pp. 181–193, Feb. 2017, doi: 10.1016/j.ccell.2017.01.001.

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.