MBatch 03 Data for MBatch: Standardized Data
Tod Casasent
2023-10-03-0930

Introduction

These instructions are aimed at people familiar with R and familiar with TCGA/GDC platforms and data types. They are intended to introduce the reader to the format for Standardized Data, including the data feature formats (such as, gene symbols and probe ids).

Standardized Data is the results of taking TCGA data from the DCC or GDC and converting it into a format read-to-use by analysts--in other words a matrix. Particularly in the case of the DCC, converting the cryptic, oft undocumented, collections of files into an analyzable matrix can take a lot of time. The origin of MBatch was in analyzing TCGA data for batch effects, hence the format of Standardized Data became the file format for MBatch.

The website http://bioinformatics.mdanderson.org/TCGA/databrowser provides TCGA DCC Standardized Data. Within this data, the "matrix_data.tsv" files contain the actual data while the "batches.tsv" files contain the batch information.

Standardized Data

Standardized Data comes from one of two sources. GDC Standardized Data is from the GDC Data Portal. Details on the GDC project are available at https://gdc.cancer.gov/. Metabolomics Workbench (MWB) data consists of the metabolomics datasets whose data files are stored by MWB at https://www.metabolomicsworkbench.org/. In both cases, we transformed the original data into the standard data matrix described below. Standardized Data also has batch information available, also described below.

Standardized Data "Data Matrix" Format

The Standardized Data "Data Matrix" format is a tab delimited file. The first line of the file begins with a tab and contains sample identifiers. For Standardized Data, the sample identifiers are TCGA bar codes. Each subsequent row begins with a Feature Identifier and is followed by numeric data. Feature Identifiers are specific to the platform and explained later, but can be values such as Hugo Gene ids, probe ids, or microRNA identifiers.

This extract from the Data Matrix format shows four sample ids and five feature ids. Note that the first blank cell indicates the starting tab for the sample identifiers line.

	TCGA-OR-A5J2-01A-21-A39K-20	TCGA-OR-A5J3-01A-21-A39K-20	TCGA-OR-A
14-3-3_beta-R-V	0.211404	-0.14778	0.220188
14-3-3_epsilon-M-C	-0.03151	-0.12861	-0.0762
14-3-3_zeta-R-V	-0.01203	0.032791	-0.34541
4E-BP1-R-V	0.589134	0.365167	0.297887
4E-BP1_pS65-R-V	-0.13521	0.182058	-0.23654

Standardized Data Batch File Format

The Standardized Data Batch File format is also a tab delimited file. The first line of the file contains the sample id column id and batch type identifiers, none of which should contain spaces.

For GDC sourced data, the first entry should be the "aliquot_barcode" column, which contains sample ids. For TCGA data from the GDC site, the other batch type identifiers are batch_id, sample_type_name, ship_date. source_center, and tissue_source_site. Other projects/programs will have some or none or other batch types.

aliquot_barcode	batch_id	sample_type_name	ship_date	source_center	tissue_so
TCGA-18-3406-01A-01T-0981-13	39.68.0	Primary Tumor	2010-05-05	22	18
TCGA-18-3407-01A-01T-0981-13	39.68.0	Primary Tumor	2010-05-05	22	18
TCGA-18-3408-01A-01T-0981-13	39.68.0	Primary Tumor	2010-05-05	22	18
TCGA-18-3410-01A-01T-0981-13	39.68.0	Primary Tumor	2010-05-05	22	18

For MWB sourced data, the first entry should be the "Sample" column, which contains sample ids. The Batch Types from MWB vary wildly both in nature and in length. You will need to check the study (linked in both tabs in the Plot Picker) to determine the nature of the batch types.

Sample	CPAP	Gender	OSA	PCOS	Sampling
S00009859	No CPAP	Female	Negative	No PCOS	Morning fasting
S00009860	No CPAP	Female	Negative	No PCOS	Morning fasting
S00009861	No CPAP	Female	Negative	No PCOS	Morning fasting
S00009862	No CPAP	Female	Negative	No PCOS	Morning fasting

Data Features

GDC data features in matrix_data.tsv files vary based on the table below. For MWB data, the data file will have whatever metabolite information was given in their data files. These can vary wildly.

Some GDC features may be pipe delimited combinations of existing features. Some Gene Expression Quantification data will be the Gene Symbol, a pipe "|", and the Ensembl Gene ID.

GDC Features

For DCC Standardized Data, the following features are used for the following data types.

Platform	Feature	Exampl
Copy Number Segment	Gene Symbol	5_8S_r
Differential Gene Expression	Ensembl Gene ID with version	ENSG00
Gene Expression Quantification	Gene Symbol	NQO1,
Gene Level Copy Number	Gene Symbol	NQO1,
Isoform Expression Quantification	miRBase name	hsa-mir-
Masked Copy Number Segment	Gene Symbol	5_8S_r
Masked Somatic Mutation	Gene Symbol	DOCK5
Methylation Beta Value	"A unique ID for the array probe associated with a CpG site"	cg230014
Protein Expression Quantification	Protein Antibody Identifier	TUBER
miRNA Expression Quantification	miRBase name	hsa-mir-

For more details on methylation data features, see the GDC information at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_Pipeline/