MBatch 03 Data for MBatch: User Data
Tod Casasent
2017-11-02-0850

1 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 as it impacts user supplied data.

2 MBatch Data Files

MBatch uses two different files (available from Standardized Data or to be provided by the user), for which the package provides code to read the files.

2.1 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-A5J6-01A-41-A39K-20	TCGA-OR-A5J7-01A-21-A39K-20
14-3-3_beta-R-V	0.211404	-0.14778	0.220188	-0.02738
14-3-3_epsilon-M-C	-0.03151	-0.12861	-0.0762	-0.02275
14-3-3_zeta-R-V	-0.01203	0.032791	-0.34541	0.136629
4E-BP1-R-V	0.589134	0.365167	0.297887	7.34E-05
4E-BP1_pS65-R-V	-0.13521	0.182058	-0.23654	-0.0974

The features (left-most column) can be any set of unique strings. For proper processing, the rows and columns should be sorted. The MBatch function readAsGenericMatrix will read a matrix TSV and sort the rows and columns.

2.2 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. The first entry should be the "Sample" column, which contains sample ids. For TCGA data (from the DCC and the GDC), the other batch type identifiers are Type, BatchId, PlateId, ShipDate, and TSS. The MBatch function readAsGenericDataframe will read a batch information data frame TSV.

Sample	Type	BatchId	PlateId	ShipDate	TSS
TCGA-OR-A5J2-01A-21-A39K-20	1	304	A39K	5/7/2014	OR - University of Michigan
TCGA-OR-A5J3-01A-21-A39K-20	1	304	A39K	5/7/2014	OR - University of Michigan
TCGA-OR-A5J6-01A-41-A39K-20	1	304	A39K	5/7/2014	OR - University of Michigan
TCGA-OR-A5J7-01A-21-A39K-20	1	304	A39K	5/7/2014	OR - University of Michigan

3 MBatch Data Structures

MBatch uses two different data structures, for which the package provides APIs to handle in functions containing the name "structures". The function mbatchLoadStructures turns the two data structures into an R object and calls the matrix theGeneMatrix and the batch information theBatchDataframe.

3.1 Matrix Data

The actual data is in a matrix formatted with the "Data Matrix" format, named by convention matrix_data.tsv. When reading a matrix file manually in R or creating a data matrix, row names (samples) and column names (features) should be strings, not factors or numbers, and should be sorted.

3.2 Batch Data

The batch information is in a data frame, formatted in the Batch Data format, and named by convention batches.tsv. When reading a batch file manually in R or creating batch data frame, names and values should be strings, not factors or numbers. The first column of the data frame should be "Sample" and contain sample ids. None of the column names should contain spaces.

3.3 MBatch Data Object

The MBatch Data Object (of class BEA_DATA) can be generated from files by calling mbatchLoadFiles or from structures (a matrix and a dataframe) by calling mbatchLoadStructures. Both functions will filter the batch information to only contain information relevant to the matrix data.