MBatch 04-00

Using MBatch Assessments: Parameters, Batch Types and Values

Tod Casasent

2017-10-25-0900

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 producing the given assessment. These instructions will only rarely, if ever, touch on the appropriateness of the assessment algorithm or interpretation of output. See MBatch_01_InstallLinux.docx for instructions on downloading test data.

This particular document details a pair of arguments that are common to many API functions.

2 Remove and Keep Parameters

Many of the MBatch API functions have a remove (theBatchTypeAndValuePairsToRemove) and a keep (theBatchTypeAndValuePairsToKeep) parameter. These are used to specify particular sets of Batch Types and Batches (Values) to either preserve in the case of the keep parameter (filtering samples that do not match the given values) or filter out in the case of the remove parameter (filtering samples that do match the given values).

3 Batch Types and Values Lists

The list of Batch Types and Values used for the remove and keep parameters follow the same format. Batch Types are the column names from the Batch File (or names from the Batch Data Frame). The Values in the Vector are the batches in the columns of the Batch Data Frame.

3.1 Removing Batch Types and Values

Take this list as an example. In this case, samples whose batches are not available are assigned batches of "Unknown" or "unknown".

```
list(   c("*", "unknown"),
        c("*", "Unknown"),
        c("Type", c("01", "02", "03", "04", "05", "06", "07", "08", "09", "40")))
```

This list has three pairs. If passed into the theBatchTypeAndValuePairsToRemove argument, this list causes the following to occur. The first pair c("*", "unknown") results in, for all batch types ("*") removing all elements with a batch value of "unknown". The second pair is the same, except the batch value removed is "Unknown". The third pair is c("Type", c("01", "02", "03", "04", "05", "06", "07", "08", "09", "40")). The third pair applies to only the batch type named "Type". This one removes all the tumor samples. (The given sample types are the tumor sample types as per https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes)

3.2 Keeping Batch Types and Values

Take this second list for the next example. In this case, samples whose batches are not available are assigned batches of "Unknown" or "unknown".

```
list(   c("*", c("unknown", "Unknown")),
        c("Type", c("01", "02", "03", "04", "05", "06", "07", "08", "09", "40")))
```

If the list is passed to theBatchTypeAndValuePairsToKeep, then first, for all batch types, only values of "unknown" or "Unknown" are **kept while all others are removed**. Second, for the Type batch type, only batches matching the 10 tumor types are kept. All other samples are removed. (This particular list would likely remove all or most of the samples from most data sets.)