Data Query Form
Tod Casasent
2020-08-12-1430

Introduction

This document gives simple instructions for using the Data Query Form. This document will provide minimal, but some, descriptions of elements such as Platforms and data descriptions. It also includes instructions for accessing Standardized Data from the DAPIR (Data API for R) R package.

Links

Query Form / Standardized Data Browser
https://bioinformatics.mdanderson.org/StandardizedDataBrowser

Batch Effects Viewer
https://bioinformatics.mdanderson.org/BatchEffectsViewer

DAPIR on GitHub
https://github.com/MD-Anderson-Bioinformatics/DataAPI

BMB on GitHub (source code and additional documentation)
https://github.com/MD-Anderson-Bioinformatics/

BCB on Docker Hub (Docker images)
https://hub.docker.com/u/mdabcb

Categories and Category Options

The Data Query Form has 11 categories. Most categories work the same for selecting and filtering data---Versions is the one exception. For the regular categories, selecting an option within a category limits the results to elements with that option. Selecting more than one option acts as an "or" so that, for example if you select "TCGA-COAD" and "TCGA-READ" Sub-Projects, you get both COAD and READ data.

Files Category

The Files Category lists the available types of files found within each dataset archive.

Sources Category

The Source Category lists the source of the data, such as the GDC or PanCan Study Group.

Derivations Category

The Derivations Category lists the derivation of data within a Source, such as, for the GDC, "current" or "legacy" data.

Archive Type Category

The Archive Type Category lists the variation of data in the dataset--for the GDC, this is "standardized". Other datasets may provide other Archive Type Categories.

Algorithms Category

The Algorithms Category divides the data into "continuous", amenable to most standardize statistical processing, and "discrete", generally sparse matrices and not amenable to many statistical methods

Versions Category

The Versions Category are the timestamps for when the data was acquired by the Query Form. This Category works different from the rest. By default, the Query Form will show the newest version of each dataset. Selecting one or more Versions, limits the results to that particular version. Note that in Standardized Data, each Version may only contain a single dataset.

Projects Category

The Projects Category lists the higher-level project, like TCGA or TARGET, for the dataset.

Sub-Projects Category

The Sub-Projects Category lists what is generally the disease (cancer type) being processed. Some Projects do not divide data by disease, hence the more generic name for this Category.

Data Type Category

The Data Type Category divides the datasets into general type of data. Currently, some categories can be overly specific (such as for different mutation data) and some overly general or redundant (such as "Copy Number Segment" and "Copy number variation"). Filters are in place to address much of that.

Details Category

The Details Category allows filtering on detailed options for datasets, in particular the Methylation data option to include (wXY) or exclude (noXY) sex

chromosomes.

Platforms Category

The Platforms Category lists the available platforms. Currently, some may be redundant and misleading, such as the Legacy GDC data having "Illumina Human Methylation 27" and "Illumina Human Methylation 450" compared to the Current GDC data using "Liftover".

File Formats

There are five files provided within datasets.

matrix_data.tsv

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 bar codes. Each subsequent row begins with a Feature Identifier and is followed by numeric data. Feature Identifiers are specific to the platform 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. The features (left-most column) can be any set of unique strings. For proper processing, the rows and columns should be sorted.

	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

batches.tsv

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. Some non-batch types may include type and patient entries for cross-reference purposes.

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

Sample	Type	BatchId	PlateId	ShipDate	TSS
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

clinical.tsv

Clinical TSV files follow the same format as batches.tsv with different column headers rather than batch information.

index.json

This is a simple JSON file describing each Category option that defines this dataset:

{

"source": "GDC",

"variant": "current",

"project": "TCGA",

"subProject": "TCGA-CHOL",

"category": "Copy Number Segment",

"platform": "DNAcopy",

"data": "standardized",

"algorithm": "discrete",

"details": "",

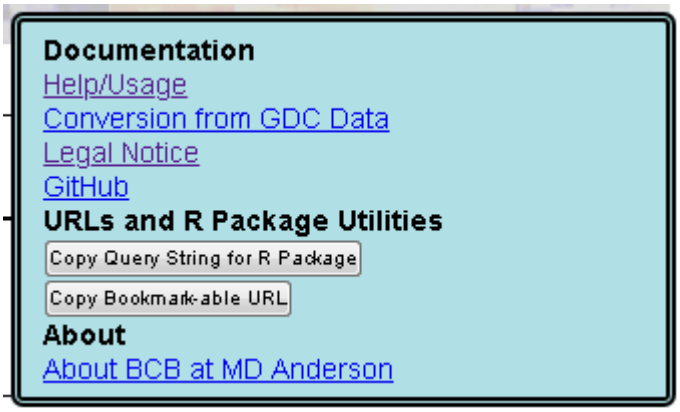
"version": "2020_01_31_0845"

}

mutations.tsv

The mutations.tsv file is a tab delimited file based on the Mutation MAF files with column headers standardized. Headers should be self-evident to users familiar with mutation datasets.

URLs and R Package Utilities



Copy Query String for R Package

Clicking the "Copy Query String for R Package" button copies a string to the clipboard based on the data query selected in the GUI. For example, this string gives current TCGA-ACC data from the 2020_01_31_0845 data run.

{\"mFiles\":[\"batches.tsv\"],\"mSources\":[],\"mVariants\":[\"current\"],\"mProjects\":[],\"mSubprojects\":[\"T
ACC\"],\"mCategories\":[],\"mPlatforms\":[],\"mData\":[],\"mAlgorithms\":[],\"mDetails\":[],\"mVersions\":[\"2

This string is used in the DAPIR R Package function checkDownloadedDataStatus to create and update a local copy of Standardized Data.

Copy Bookmark-able URL

Clicking the "Copy Bookmark-able URL" button copies a string to the clipboard based on the data query selected in the GUI. The string is a URL that links to the selected query defined in the GUI.

DAPIR R Package

The DataAPI project page (https://github.com/MD-Anderson-Bioinformatics/DataAPI) on GitHub gives instructions for installing the DAPIR R package.

And example of downloading data using the Query String is given below.

```
# TCGA-ACC data -- pasted the "Copy Query String for R Package" button into queryOne <- paste("")

queryOne <- paste("{\"mFiles\":[\"batches.tsv\"],\"mSources\":[],\"mVariants\":[\"current\"],",

"\"mProjects\":[],\"mSubprojects\":[\"TCGA-ACC\"],\"mCategories\":[],",

"\"mPlatforms\":[],\"mData\":[],\"mAlgorithms\":[],\"mDetails\":[],",
```

```
"\"mVersions\":[]}", sep="")
# temp directory
datasetDir <- file.path(tempdir(), "DAPIR")
print(datasetDir)
dir.create(datasetDir, showWarnings=FALSE, recursive=TRUE)
# get data status
results <- checkDownloadedDataStatus(queryOne, datasetDir)
print(results)
# Download initial datasets
newDatasets <- results$NEW
downloadData(newDatasets, datasetDir)
```