**Data Query Form**
**Tod Casasent**
**2021-04-30-1000**

# Introduction

This document gives simple instructions for using the Data Query Form -- the Query Form is part of the MetaBatch Omic Browser combined GUI, which allows the user to sort and select datasets. This document will provide minimal, but some, descriptions of elements such as Platforms and data descriptions.

## Links

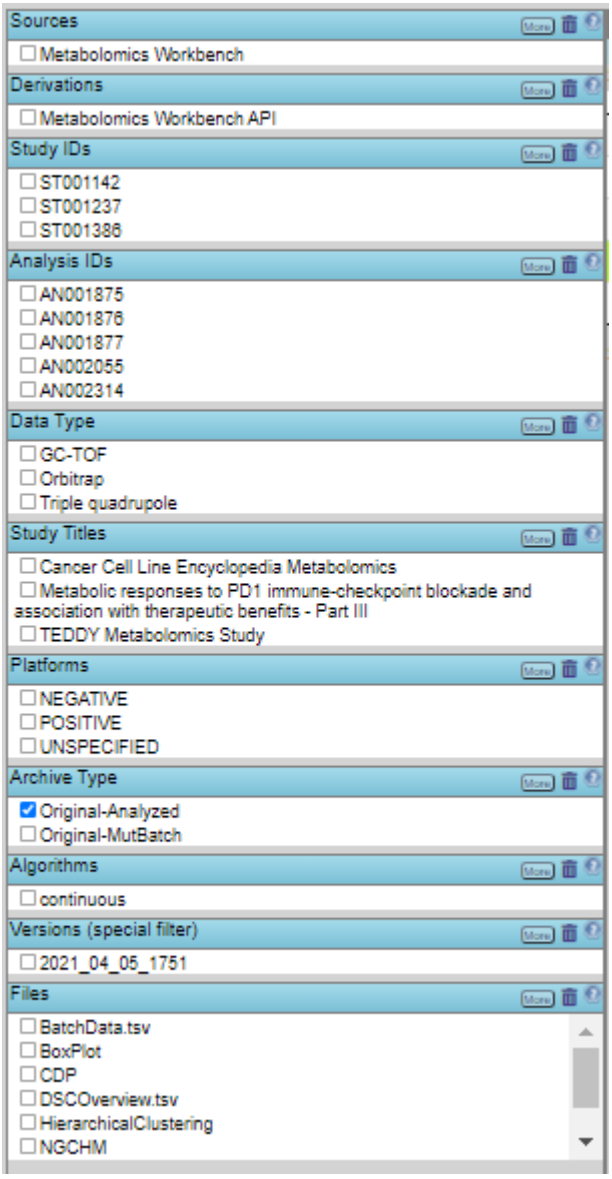
DAPIR on GitHub
https://github.com/MD-Anderson-Bioinformatics/DataAPI

BCB 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 several categories. Most categories work the same for selecting and filtering data---Versions is the one exception. For the regular categories, we emulate the GDC Portal search algorithm with selections within categories operating as OR and between categories as AND. Category options are narrowed based on selections from other categories.




## DSC P-Value

This is only available in the MetaBatch Omic Browser Query Form when DSC values are selected. DSC P-Values are not corrected for multiple testing. The DSC P-Value Category lists common p-value cut-offs.

### Min DSC Value

This is only available in the MetaBatch Omic Browser Query Form when DSC values are selected. The Min DSC Value indicates the DSC of the results should be less than or equal to this value. Empty means accept any value.

### Max DSC Value

This is only available in the MetaBatch Omic Browser Query Form when DSC values are selected. The Max DSC Value indicates the DSC of the results should be less than or equal to this value. Empty means accept any value.

### Sources

The Source Category lists the source of the data, such as the Metabolomics Workbench.

### Derivations

The Derivations Category lists the derivation of data within a Source, such as, for the Metabolomics Workbench, "API" indicating data was accessed via the Metabolomics Workbench API.

### Study IDs

The Study IDs Category lists the higher-level project for the dataset -- for Metabolomics Workbench datasets, this is the Study ID.

### Analysis IDs

The Analysis IDs Category lists the lower-level project for the dataset -- for Metabolomics Workbench datasets, this is the Analysis ID.

### Data Type

The Data Type Category divides the datasets into general type of data. Some names have been converted to generic version for usability.

### Study Title

The Study Title Category allows filtering or searching on study titles.

### Platforms

The Platforms Category lists the available platforms. For metabolomics data, this is negative, positive, and unspecified.

### Archive Type

The Archive Type Category lists the variation of data in the dataset--this is Original-Analyzed for a traditional batch effect analysis and Original-MutBatch for a Kuskal-Wallis non-parametric analysis.

### Algorithms

The Algorithms Category divides the data into "continuous", amenable to most standard statistical processing, and "discrete", generally sparse matrices and not amenable to many statistical methods. Metabolomics data is classified as "continuous" but run as both.

### Versions (special filter)

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.

### Files

The Files Category lists the available types of files found within each dataset archive. You can use this to look for data with batch information or mutation files.

## 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 study but use the metabolomics names from Metabolomics Workbench.

This extract from the Data Matrix format shows six sample ids and eight 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.

| | 1000087 | 1002105 | 1002571 | 1003070 | 1003867 |
|---|---|---|---|---|---|
| 1,2-dihydroxycylohexane NIST | 7.208088008 | 7.060479781 | 8.089582893 | 7.251624424 | 7.426516366 |

| | 1000087 | 1002105 | 1002571 | 1003070 | 1003867 |
|---|---|---|---|---|---|
| 1,5-anhydroglucitol | 15.37626786 | 14.88150738 | 15.67871979 | 15.51111176 | 17.61022522 |
| 1-monoolein | 7.111031312 | 12.77566789 | 9.103602544 | 11.15600089 | 7.449148645 |
| 1-monopalmitin | 9.095317897 | 8.325215495 | 8.034358798 | 9.002955622 | 8.175175262 |
| 1-monostearin | 10.14030576 | 8.247120518 | 7.976592762 | 7.968609199 | 9.619889819 |
| 2,3-dihydroxybutanoic acid NIST | 8.055228198 | 7.605109074 | 10.55252501 | 7.265286858 | 7.417430613 |
| 2-deoxytetronic acid | 8.575803937 | 8.247025547 | 9.26352743 | 8.577013332 | 9.009100789 |
| 2-deoxytetronic acid NIST | 9.480244923 | 10.09477708 | 9.448446735 | 9.937462101 | 9.960914192 |

## 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. While labeled "batches", these are the study "factors" from Metabolomics Workbench which may or may not be appropriate or interesting for batch effects.

| Sample | Sex | cc |
|---|---|---|
| 1000087 | Female | 4 |
| 1002105 | Male | 4 |
| 1002571 | Female | 4 |
| 1003070 | Male | 5 |
| 1003867 | Female | 1 |
| 1004016 | Female | 1 |
| 1004091 | Male | 4 |
| 1004917 | Female | 1 |
| 1005692 | Female | 4 |
| 1006191 | Male | 1 |
| 1007755 | Female | 3 |

## index.json

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

{

"source": "MW",

"variant": "API",

"project": "ST001386",

"subProject": "AN002314",

"category": "GC-TOF",
"platform": "UNSPECIFIED",
"data": "standardized",
"algorithm": "continuous",
"details": "TEDDY Metabolomics Study",
"version": "2021_04_05_1751"
}