

BEIStack 02A Using Batch Effects Interface Assessments External
Tod Casasent
2019-02-05-1520

1 Using Batch Effects Interface Assessments

This document focuses on explaining the components of the Batch Effects Interface (BEI) involved with creating a job, loading data, and running assessments. This document will not address statistical issues or "how to spot" batch effects.

The URL for your install should be provided to you, but will likely be something like:

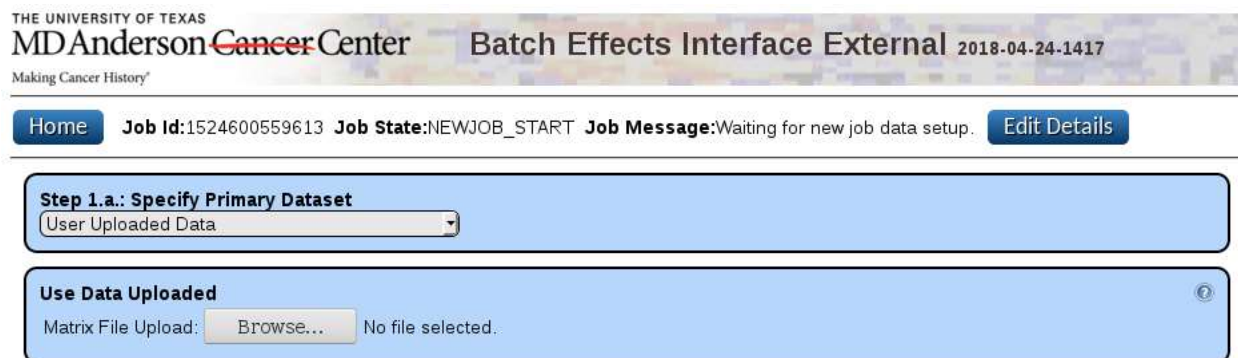
<http://your-server.your-company.com:8181/BatchEffectsInterface/>

2 Starting a Job

When the Batch Effects Interface is first entered, the user is presented with a list of jobs (if any) and the option to Start New Job. (The external/public version of BEI does not support authentication and authorization.) The initial screen is shown below.



Upon selection "Start New Job", the user is presented with a new job page and the option to Update Job Info (covered later) and Step 1, selecting or uploading data, as shown below. The job id is shown in the Update Job Info section.



Step 1.a.: Specify Primary Dataset

User Uploaded Data

Standardized Data

GDC Downloaded Data

User Uploaded Data

ected.

The next sections will follow each of these three options.

3 Downloading GDC Data

Upon selecting "Download Data from GDC", the user will be presented with the page below, asking them to agree to the TCGA and GDC release guidelines.

GDC Download Option:

Upload and use a manifest file from the GDC Data Portal.

Step 1.a.: Specify Primary Dataset

GDC Downloaded Data

State:

Use GDC Downloaded Data

Manifest File Type Selection:

Current: methylation450-bt

Data Manifest File Upload:

Browse...

No file selected.

Biospecimen Manifest File Upload:

Browse...

No file selected.

Upload Primary GDC Manifest Files

*For More Inform

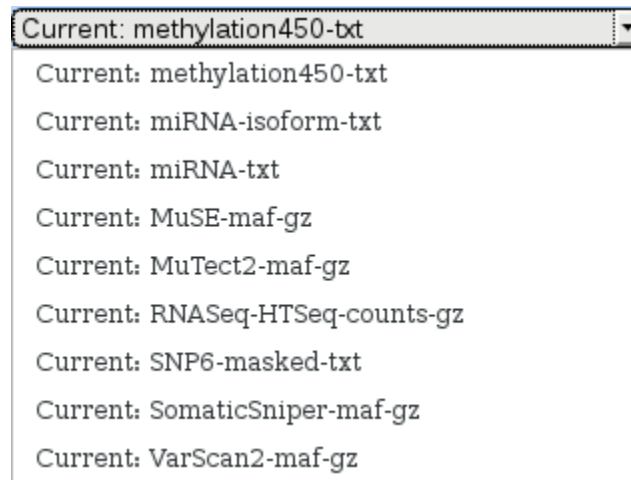
Use Primary

Agree to GDC Data Guidelines

By downloading, analyzing, and/or using TCGA data for publication purposes, the user accepts the data use restrictions and requirements as outlined in the TCGA Publication Guidelines. See https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/ for additional information.

☐ I agree

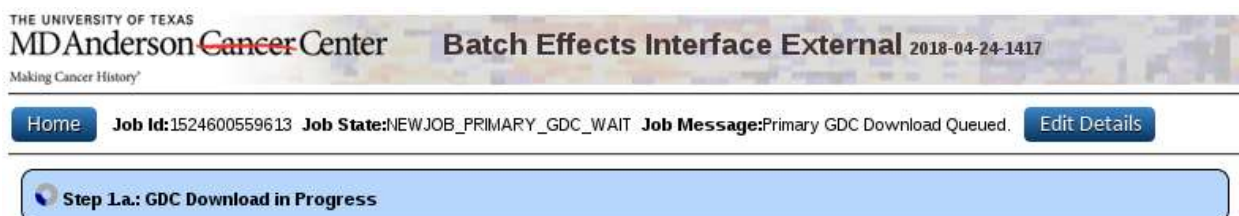
The user will need two manifest files from the GDC Data Portal <https://portal.gdc.cancer.gov/>. Data types supported for download and conversion are shown in the screenshot below. The manifest file corresponding to these data types is the "Data Manifest File".



The second manifest file is the "Biospecimen Manifest File". This corresponds to the Data Category "Biospecimen" on the GDC site.

Currently, only batch information for the TCGA projects are supported.

After uploading the Data Manifest File and Biospecimen Manifest File and pressing "Use Primary Downloaded GDC Data", the data will be downloaded and convert. The job will first go into GDC Data: Ready for download state. Depending on the setup, this could take a couple of minutes or more to process. The BEI Docker Stack is configured to run one download process at a time, due to memory constraints.



Once the download starts, the download log file will be tailed, as shown below.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center **Batch Effects Interface External** 2018-04-24-1417
Making Cancer History®

[Home](#) **Job Id:**1524600559613 **Job State:**NEWJOB_PRIMARY_GDCRUN_WAIT **Job Message:**Primary GDC Download in Progress. [Edit Details](#)

Log File Tail (last 100 lines):

```
/BEI/OUTPUT/1524600559613/DatasetConfig.log
1 -- downloadTextToString = http://bei_service:8080/BatchEffectsInterface/JOBUpdate?jobId=1524600559613&status=NE
1 -- main start
1 -- PARALLEL-THREADS = 10
1 -- PARALLEL-CHECK 5 = 4 (type of data usually one to 4)
1 -- myList = 0
1 -- PARALLEL-CHECK 4 = 0 (number of datasets usually in dozens)
1 -- myList = 0
```

Here is another view of the log from further in the process.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center **Batch Effects Interface External** 2018-04-24-1417
Making Cancer History®

[Home](#) **Job Id:**1524603107862 **Job State:**NEWJOB_PRIMARY_GDCRUN_WAIT **Job Message:**Primary GDC Download in Progress. [Edit Details](#)

Log File Tail (last 100 lines):

```
/BEI/OUTPUT/1524603107862/DatasetConfig.log
17 -- Download succeeded: 14e06cc7-9479-408f-9a22-aec778e5138.xml for URL=https://gdc-api.nci.nih.gov/data/14e06cc7-9479-408f-9a22-aec778e5138.xml
12 -- Download succeeded: 9e003164-365e-4dd5-9739-fcddbd174037b.xml for URL=https://gdc-api.nci.nih.gov/data/9e003164-365e-4dd5-9739-fcddbd174037b.xml
10 -- Download succeeded: fc7778d8-5931-4b20-a93e-021851fcd06b.xml for URL=https://gdc-api.nci.nih.gov/data/fc7778d8-5931-4b20-a93e-021851fcd06b.xml
10 -- downloadJsonToString = https://gdc-api.nci.nih.gov/files?filters={"op":"and","content":{"op":"in","content":{"field":"file","value":"14e06cc7-9479-408f-9a22-aec778e5138.xml"}}}}
19 -- Download succeeded: af8ff061-ded6-45f0-858b-9d9b1425ca85.xml for URL=https://gdc-api.nci.nih.gov/data/af8ff061-ded6-45f0-858b-9d9b1425ca85.xml
1 -- Download succeeded: 7316935a-51f9-4f29-bc2f-0ecc0872546b.xml for URL=https://gdc-api.nci.nih.gov/data/7316935a-51f9-4f29-bc2f-0ecc0872546b.xml
1 -- downloadJsonToString = https://gdc-api.nci.nih.gov/files?filters={"op":"and","content":{"op":"in","content":{"field":"file","value":"7316935a-51f9-4f29-bc2f-0ecc0872546b.xml"}}}}
```

Below, the downloaded files are being converted into a matrix format.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center **Batch Effects Interface External** 2018-04-24-1417
Making Cancer History®

[Home](#) **Job Id:**1524603107862 **Job State:**NEWJOB_PRIMARY_GDCRUN_WAIT **Job Message:**Primary GDC Download in Progress. [Edit Details](#)

Log File Tail (last 100 lines):

```
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
1 -- Disease Type = ACC
1 -- Converting /BEI/OUTPUT/1524603107862/downloaded/TCGA/TCGA-ACC/GeneExpressionQuantification/HTSeq-Cou
```

Once the download and convert are complete, the user will be shown the finished log, and the option to download or otherwise select a second dataset.

THE UNIVERSITY OF TEXAS

MDAndersonCancerCenter

Batch Effects Interface External

2018-04-24-1417

[Home](#)
Job Id:1524603107862
Job State:NEWJOB_PRIMARY_DONE
Job Message:Primary Data Available. Waiting for Secondary Data.
[Edit Details](#)

Step 1.a.: Specify Primary Dataset Complete

Step 1.b.: Specify Secondary Dataset

Some correction algorithms use two datasets. Use the Proceed button if you do not need a second dataset.

GDC Downloaded Data

Proceed without Secondary Dataset

Log File Tail (last 100 lines):

```

1 -- Disease Type = ACC
1 -- There are 711 rows to test.
1 -- convert done
1 -- Finish Proc downloadMe = TCGA -- TCGA-ACC -- Gene Expression Quantification -- HTSeq - Counts --
1 -- finish proc
1 -- TIMING -- GDCConvert::convertFiles -- Time=40015 -- program=TCGA -- project=TCGA-ACC -- datatype=Gene Expression Quantification -- workflow=
copy matrix_data.tsv and batches.tsv to basedir
finished, success
post job 1524603107862 for status NEWJOB_PRIMARY_DONE
1 -- downloadTextToString = http://bel_service:8080/BatchEffectsInterface/JobUpdate?jobId=1524603107862&status=NEWJOB_PRIMARY_DONE

```

Use GDC Downloaded Data

Select	Program	Project	Workflow	Datatype	Category	Platform
<input checked="" type="radio"/>	TCGA	TCGA-ACC	HTSeq - Counts	Gene Expression Quantification	RNASeq	NA
<input type="radio"/>	TCGA	TCGA-ACC	HTSeq - FPKM	Gene Expression Quantification	RNASeq	NA
<input type="radio"/>	TCGA	TCGA-ACC	HTSeq - FPKM-UQ	Gene Expression Quantification	RNASeq	NA

☒ Enable Filter Options

Filter: Program

Select Entries to Display

☐ TCGA
☐ TARGET

Filter: Project

Select Entries to Display

☒ TCGA-ACC
☐ TCGA-BLCA
☐ TCGA-BRCA
☐ TCGA-CESC

Filter: Workflow

Select Entries to Display

☐ Liftover
☐ BCGSC miRNA Profiling
☐ HTSeq - Counts
☐ HTSeq - FPKM

Filter: Datatype

Select Entries to Display

☐ Methylation Beta Value
☐ Isoform Expression Quantification
☐ miRNA Expression Quantification
☒ Gene Expression Quantification

Filter: Category

Select Entries to Display

☐ Methylation-450
☐ Methylation-27
☐ miRNA-isoform
☐ miRNA-exp

Filter: Platform

Select Entries to Display

☐ Illumina Human Methylation 450
☐ Illumina Human Methylation 27
☐ N/A

*For More Information on GDC Terminology: <https://docs.gdc.cancer.gov/Data/Introduction/>

Use Downloaded GDC Data

The secondary dataset is used for correction algorithms such as EBN-plus and RBN. For now, we will select the “Proceed without Secondary Dataset” button, to explore the basic assessment settings.

5

This will take the user Step 2: Select Batches for the MBatch Configuration, which is described later.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History®

Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process Edit Details

Step 2: Select Batches Reset Defaults

Select Sample Identifier (Select Sample Id Batch Type): Sample

Selected Types for Batch Analysis:

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/> Must specify at least one Batch Type	BatchId	304.63.0.79
<input type="checkbox"/> Must specify at least one Batch Type	PlateId	A295.79
<input type="checkbox"/> Must specify at least one Batch Type	ShipDate	2013-05-08:79
<input type="checkbox"/> Must specify at least one Batch Type	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/> Must specify at least one Batch Type	Type	01 Primary Tumor:79

Save Configuration Load Configuration

From here, hit the “Home” button, and you can see the job we just created which says “MBatch Configuration in Process”.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History®

Home Start New Job

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524603107862		4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

4 Using Your Own Data

For this example, we will use data provided by the user. From the main job list, hit “Start New Job”. The “User Uploaded Data” is selected by default for Step 1.a. Specify Primary Dataset.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History

Home Job id:1524662045619 Job State:NEWJOB_START Job Message:Waiting for new job data setup. Edit Details

Step 1.a: Specify Primary Dataset
User Uploaded Data

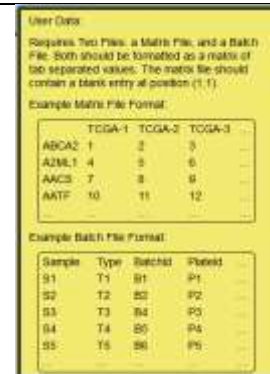
Use Data Uploaded
Matrix File Upload: Browse... No file selected.

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. The formats are described in the tooltip for User Data Uploaded section.

4.1 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. The features (left-most column) can be any set of unique strings. For proper processing, the rows and columns should be sorted.



Tooltip for Upload User Data

	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

4.2 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.

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

4.3 User File Uploads

The first part of Step 1.a. is to upload a matrix file. For this example, we downloaded the results of the GDC Download Example and extracted that file. (The results archive is described later.)

Pick Browse, select the matrix_data.tsv file (BEI will accept any file name provided and rename the file when uploaded) and click Upload Primary Matrix. The below screenshot shows the upload process.

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center **Batch Effects Interface External** 2018-04-24-1417
Making Cancer History®

[Home](#) **Job Id:**1524662045619 **Job State:**NEWJOB_START **Job Message:**Waiting for new job data setup. [Edit Details](#)

Step 1.a.: Specify Primary Dataset
User Uploaded Data

Use Data Uploaded

Matrix File Upload: matrix_data.tsv

Then the user will be presented with the controls to upload a batch file (named batches.tsv in the downloaded archive).

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524662045619 Job State:NEWJOB_PRIMARY_USER_MATRIX Job Message:Primary Matrix Data Uploaded. Waiting for Batch Data. Edit Details

Step 1.a.: User Data Upload in Progress

Use Data Uploaded

Matrix File Upload: Browse... No file selected. ☒ Primary Matrix Uploaded

Batch File Upload: Browse... No file selected.

Browse is used to select the file, after which an "Upload Primary Batch File" button allows the file to be uploaded.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524662045619 Job State:NEWJOB_PRIMARY_USER_MATRIX Job Message:Primary Matrix Data Uploaded. Waiting for Batch Data. Edit Details

Step 1.a.: User Data Upload in Progress

Use Data Uploaded

Matrix File Upload: Browse... No file selected. ☒ Primary Matrix Uploaded

Batch File Upload: Browse... batches.tsv. Upload Primary Batch File

After the batch file is uploaded, the user is given the option of providing a second dataset.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524662045619 Job State:NEWJOB_PRIMARY_DONE Job Message:Primary Data Available. Waiting for Secondary Data. Edit Details

☒ Step 1.a.: Specify Primary Dataset Complete

Step 1.b.: Specify Secondary Dataset
Some correction algorithms use two datasets. Use the Proceed button if you do not need a second dataset.

User Uploaded Data Proceed without Secondary Dataset

Use Data Uploaded

Matrix File Upload: Browse... No file selected. Upload Secondary Matrix

We will select the “Proceed without Secondary Dataset” option. This will take the user Step 2: Select Batches for the MBatch Configuration, which is described later.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History™

Home Job Id:1524662045619 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process Edit Details

Step 2: Select Batches Reset Defaults

Select Sample Identifier (Select Sample Id Batch Type): Sample

Selected Types for Batch Analysis:

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/> Must specify at least one Batch Type	BatchId	304.63.0.79
<input type="checkbox"/> Must specify at least one Batch Type	PlateId	A295.79
<input type="checkbox"/> Must specify at least one Batch Type	ShipDate	2013-05-08.79
<input type="checkbox"/> Must specify at least one Batch Type	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/> Must specify at least one Batch Type	Type	01 Primary Tumor.79

Save Configuration Load Configuration

Then, we hit Home, to return to the job list, which now has two jobs.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History™

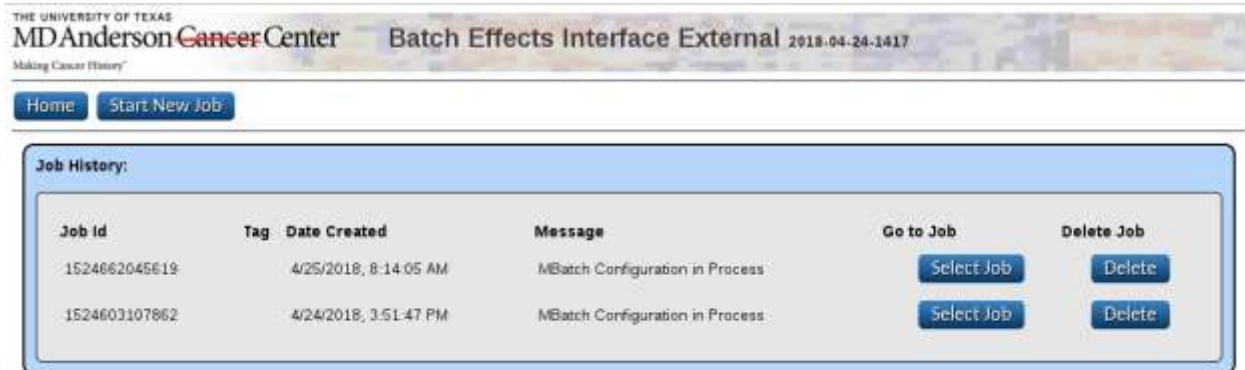
Home Start New Job

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524662045619		4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete
1524603107862		4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

5 Job Details

The current state of the job list is a little confusing. We have two jobs, in the same state, and with mostly similar Job Ids. Here, we will set the Job Details to make the jobs easier to tell apart. Let's start with the bottom job from 4.24, which is the one that used GDC Downloaded data. Click "Select Job" for that job.



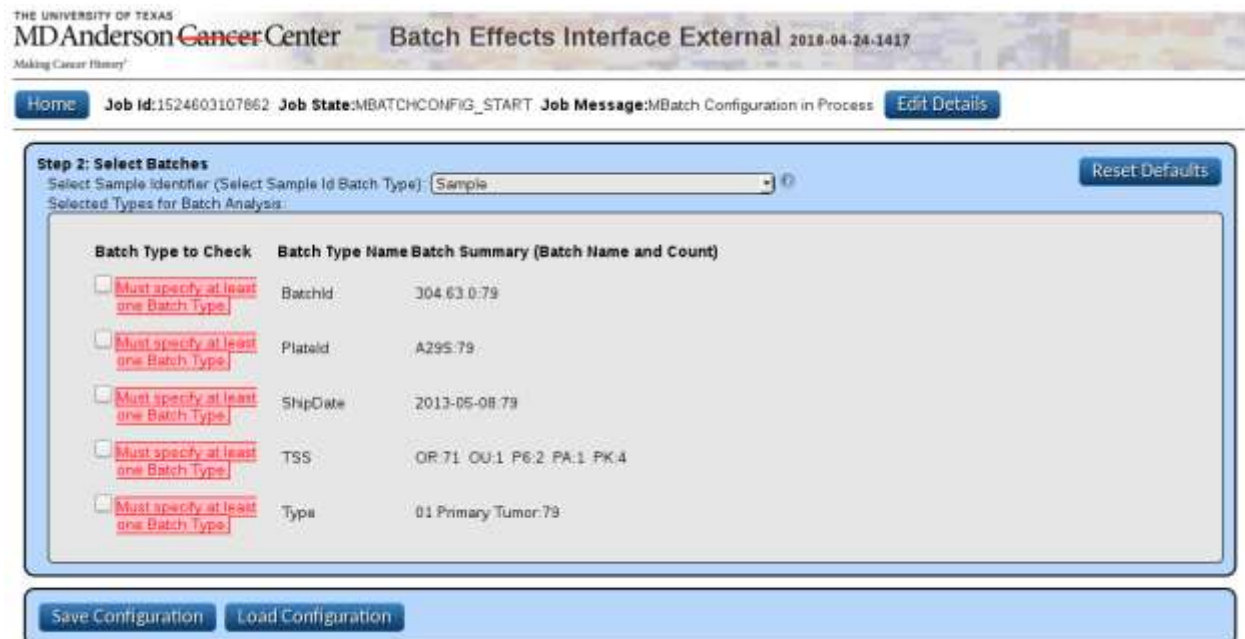
THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History

Home Start New Job

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524662045619		4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete
1524603107862		4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

This takes us to the MBatch Configuration page.



THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History

Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process Edit Details

Step 2: Select Batches Reset Defaults

Select Sample Identifier (Select Sample Id Batch Type): Sample

Selected Types for Batch Analysis:

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/> Must specify at least one Batch Type	BatchId	304.63:0.79
<input type="checkbox"/> Must specify at least one Batch Type	PlateId	A295:79
<input type="checkbox"/> Must specify at least one Batch Type	ShipDate	2013-05-08:79
<input type="checkbox"/> Must specify at least one Batch Type	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/> Must specify at least one Batch Type	Type	01 Primary Tumor:79

Save Configuration Load Configuration

But for now, just click the “Edit Details” button, which takes us to the Edit Details page.

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History®

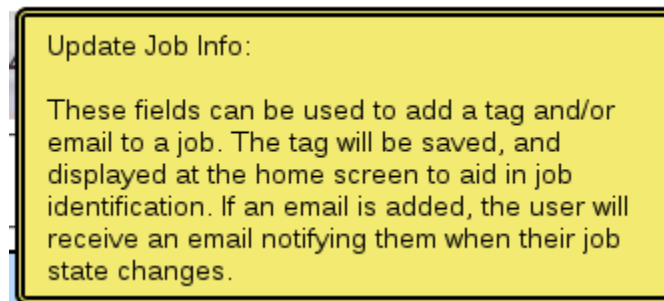
Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process

Update Job Info

Job Id: 1524603107862 Tag: Email:

Update Cancel

On this page, we hover over the tooltip question mark to see the help notice.



Since we are using the External Version, we have a Tag to describe the job, and an email field. Note, that emails may or may not be supported by your local installation.

We will leave the email blank, but for the Tag put “GDC Download Example”.

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center Batch Effects Interface External 2018-04-24-1417
Making Cancer History®

Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process

Update Job Info

Job Id: 1524603107862 Tag: GDC Download Example Email:

Update Cancel

Then we hit “Update”, which lets us know the update occurred.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process

Update Job Info

Job Id: 1524603107862 Tag: GDC Download Example Email:

Update Cancel

Updated!

OK

And returns us to the job page for that job, which right now is the MBatch Configuration page.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process Edit Details

Step 2: Select Batches

Select Sample Identifier (Select Sample Id Batch Type):

Reset Defaults

Selected Types for Batch Analysis:

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/> Must specify at least one Batch Type	BatchId	304.63.0:79
<input type="checkbox"/> Must specify at least one Batch Type	PlateId	A29S:79
<input type="checkbox"/> Must specify at least one Batch Type	ShipDate	2013-05-08:79
<input type="checkbox"/> Must specify at least one Batch Type	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/> Must specify at least one Batch Type		

Save Configuration Load Configuration

If we hit “Home”, we get the job list, which shows the updated Tag for this job.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External2018-04-24-1417

HomeStart New Job

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524662045619		4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete
1524603107862	GDC Download Example	4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

We can do the same process and add the “User Data Example” to the other job.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External2018-04-24-1417

HomeStart New Job

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524662045619	User Data Example	4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete
1524603107862	GDC Download Example	4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

6 Configuring Assessments

This is based on the GDC Download data from above, but will be similar for each data type selected. From the Job List page, hit “Select Job” for the GDC Download Example job.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

[Home](#) [Start New Job](#)

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524662045619	User Data Example	4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete
1524603107862	GDC Download Example	4/24/2018, 3:51:47 PM	MBatch Configuration in Process	Select Job	Delete

6.1 Step 2

Below is the initial step in setting up assessments, Step 2: Select Batches, where the user selects the "Sample Identifier" from the batch file. For TCGA/GDC data, this is always "Sample". If the batch file does not contain a Sample column, the user will be prompted to select the Sample Identifier column from the Batch file.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

[Home](#) **Job Id:**1524603107862 **Job State:**MBATCHCONFIG_START **Job Message:**MBatch Configuration in Process [Edit Details](#)

Step 2: Select Batches

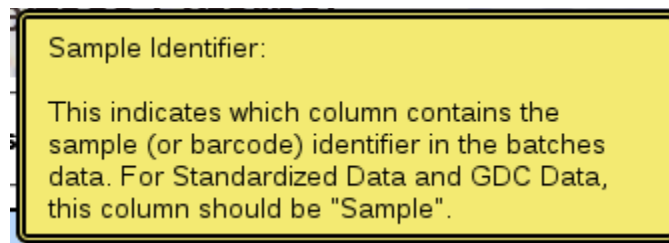
Select Sample Identifier (Select Sample Id Batch Type): Sample [Reset Defaults](#)

Selected Types for Batch Analysis:

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/> Must specify at least one Batch Type	BatchId	304.63:0:79
<input type="checkbox"/> Must specify at least one Batch Type	PlateId	A29S:79
<input type="checkbox"/> Must specify at least one Batch Type	ShipDate	2013-05-08:79
<input type="checkbox"/> Must specify at least one Batch Type	TSS	OR:71 OU:1 P:6.2 PA:1 PK:4
<input type="checkbox"/> Must specify at least one Batch Type	Type	01 Primary Tumor:79

[Save Configuration](#) [Load Configuration](#)

The tooltip text for Select Sample Identifier is shown below.



Here, we need to select the Batch Type to check for batch effects. TCGA ACC Gene Expression data does not have particularly interested batches. In this case, we will pick TSS, since it has 5 different batches, while all the others have one.

Step 2: Select Batches

Select Sample Identifier (Select Sample Id Batch Type): Sample

Selected Types for Batch Analysis: TSS

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/>	BatchId	304.63.0:79
<input type="checkbox"/>	PlateId	A29S:79
<input type="checkbox"/>	ShipDate	2013-05-08:79
<input checked="" type="checkbox"/>	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/>	Type	01 Primary Tumor:79

Reset Defaults

Once at least one Batch Type is selected, the rest of the MBatch Configuration steps are shown.

6.2 Step 3

In Step 3: Select Filtering Options, the user is asked how many data points to use and if there are batch to exclude.



The screenshot shows the 'Step 3: Select Filtering Options' dialog box. It has a 'Reset Defaults' button in the top right. The 'Auto-Filter to Maximum Number of Values:' section contains a 'Max Number of Values:' input field with the value '4000000'. The 'Exclude User Specified Batches:' section contains a 'Batch Type for Exclusion:' dropdown menu with the text 'Select Optional Batch Type for Exclusion'.

The Max Number of Values is the number of rows (features like genes) times the number of columns (samples). Features are dropped based on variance, since lower variance features are less likely to contain batch effects.

Maximum Number of Values:

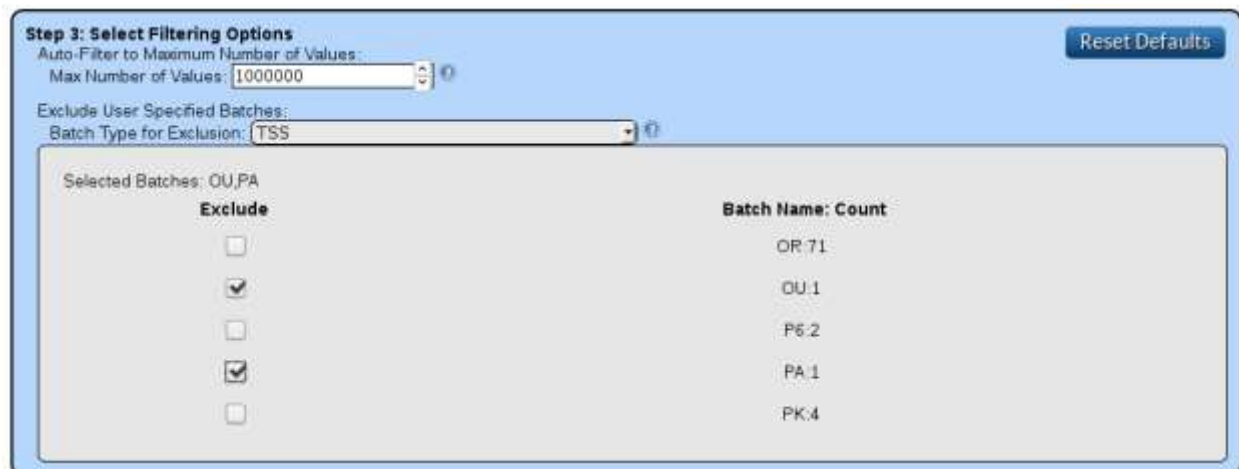
Maximum number of cells allowed in the data matrix. Matrix size is reduced by filtering features (such as, genes or probes) based on using IQR (interquartile range) as a measurement of variant, with the assumption that batch effects are more likely to be found where there is large variation. 0 means no filtering. Minimum value is . Recommended maximum value is .

Exclude User Specified Batches:

Select a batch type from the selection box to the left to be presented with a list of batches within that type. Check boxes to exclude specific batches from processing.

Tooltip Text for Step 3

Below, we reduced the Max Number of Values and selected to exclude two TSS batches, which have only one sample each. The user is not limited to selecting only the TSS batch type—any one batch type may be used for excluding data. The use-case for this is that a researcher has discovered a problem with a particular plate, so you elect to exclude it from processing.

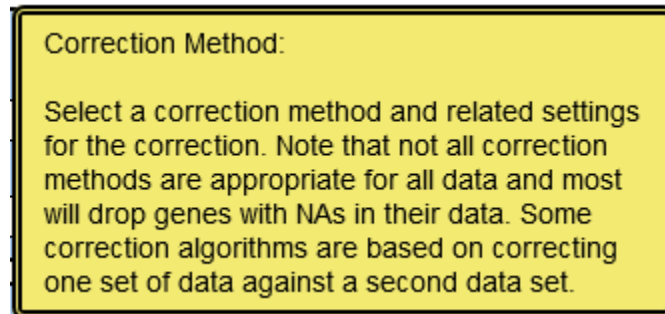


The screenshot shows the 'Step 3: Select Filtering Options' dialog box with the 'Max Number of Values' set to '1000000' and the 'Batch Type for Exclusion' dropdown set to 'TSS'. Below the dropdown, a list of selected batches is shown: 'OU,PA'. To the right, a table lists the batches and their counts.

Exclude	Batch Name: Count
<input type="checkbox"/>	OR.71
<input checked="" type="checkbox"/>	OU.1
<input type="checkbox"/>	P6.2
<input checked="" type="checkbox"/>	PA.1
<input type="checkbox"/>	PK.4

6.3 Step 4

In Step 4, a Correction is optionally selected. In this document, we will accept the default of "none" and only run assessments.



Tooltip Text for Step 4



6.4 Step 5

Step 5 concerns values that affect PCA-Plus and DSC calculations. Each option has explanatory tooltip text, covered in detail below. For more details on PCA-Plus and the DSC value, please see http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview#The_DSC_metric

sa

Step 5: Select PCA-Plus/DSC Arguments Reset Defaults

Number of DSC Permutations: 2000

Number of DSC Threads (1-5): 5

Minimum DSC Batch Size: 5

DSC Random Number Seed: 314

Maximum Number of Features: 5000

Tooltips for Step 5 are shown below.

The Number of Dispersion Separability Criterion (DSC) Permutations refers to the number of permutations used to calculate the p-value associated with the DSC. In general, for TCGA/GDC-like data, we recommend no less than 2000 permutations.	<p>Number of DSC Permutations:</p> <p>Permutations performed to calculate the DSC and associated p-values. Minimum of 500, recommended no less than 2000 for valid results.</p>
The number of DSC threads is related to the size of your data and the memory and cores available on your machine. For most TCGA/GDC data sets, 5-10 threads and 16-32GB of memory should be sufficient.	<p>Number of DSC Threads:</p> <p>Number of threads (usually equal to the desired number of cores) used to perform DSC calculations. Generally, use the default value.</p>
The minimum batch size to consider for inclusion in DSC processing. Small outlier batches can greatly influence the outcome. We generally recommend 5 for TCGA/GDC-like data.	<p>Minimum DSC Batch Size:</p> <p>Enter a minimum batch size acceptable as a usable batch (depends on correction algorithm).</p>
The seed used for performing DSC permutations. This is needed in order to reproduce results from other researchers.	<p>DSC Random Number Seed:</p> <p>Random number seed used in permutations and necessary for reproducible results.</p>
The Maximum Number of Features (such as genes or probes) to use for DSC calculations. Reduction in size is based on IQR (interquartile range) as a measurement of variance. (Generally, batch effects are found in areas of high variance.)	<p>Maximum Number of Features (DSC):</p> <p>Maximum number of features (such as, genes or probes) for DSC computations. Size is reduced based on using IQR (interquartile range) as a measurement of variant, with the assumption that batch effects are more likely to be found where there is large variation. Minimum recommended value is 1000. Maximum recommended value depends on your setup.</p>

Tooltip Text for Step 5

Here we have reduced the number of permutations and the number of features, to make our run finish sooner.



Step 5: Select PCA-Plus/DSC Arguments

Number of DSC Permutations: 500

Number of DSC Threads (1-5): 5

Minimum DSC Batch Size: 5

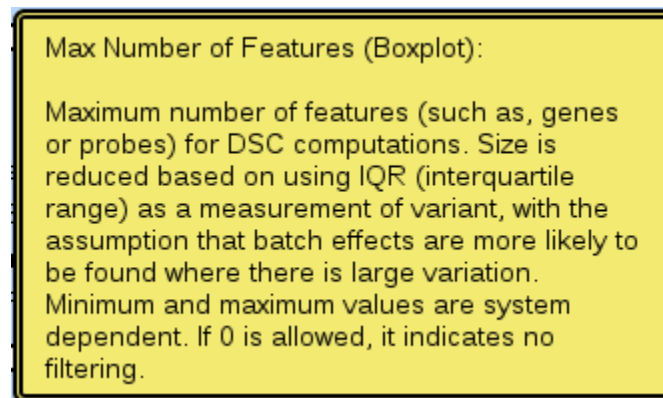
DSC Random Number Seed: 314

Maximum Number of Features: 2000

Reset Defaults

6.5 Step 6

In Step 6, the number of features (such as, genes or probes) to be used with the boxplot assessment algorithm are designated. In general, we limit this to 5000 features, as more features takes excessive amounts of memory and processing time for TCGA/GDC-like data.



Max Number of Features (Boxplot):

Maximum number of features (such as, genes or probes) for DSC computations. Size is reduced based on using IQR (interquartile range) as a measurement of variant, with the assumption that batch effects are more likely to be found where there is large variation. Minimum and maximum values are system dependent. If 0 is allowed, it indicates no filtering.

Tooltip Text for Step 6



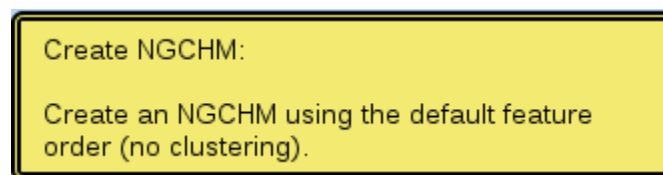
Step 6: Select Boxplot Arguments

Max Number of Features: 2500

Reset Defaults

6.6 Step 7

Step 7 is selecting the optional creation of an NGCHM (Next Generation Clustered Heatmap). Heatmap is built using default values.



Create NGCHM:

Create an NGCHM using the default feature order (no clustering).



Step 7: Select NGCHM Arguments

Perform NGCHM creation: ☒

Reset Defaults

6.7 Configuration Options and Do MBatch Run

Two different sets of buttons appear at the bottom of the page. Until Step 3 is complete, the user has only the options of Save Configuration and Load Configuration. This allows the user to save the current MBatch configuration, if they need to come back to it later (for example, after getting a list of samples to exclude). Load Configuration loads the last saved configuration.



Once Step 3 is complete, the user can also press the Do MBatch Run button, to start the run.



6.8 Complete Page View

[Home](#) Job Id:1524603107862 Job State:MBATCHCONFIG_START Job Message:MBatch Configuration in Process [Edit Details](#)

Step 2: Select Batches
Select Sample Identifier (Select Sample Id Batch Type): Sample [?](#)
Selected Types for Batch Analysis: TSS [Reset Defaults](#)

Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input type="checkbox"/>	BatchId	304.63.0:79
<input type="checkbox"/>	PlateId	A29S:79
<input type="checkbox"/>	ShipDate	2013-05-08:79
<input checked="" type="checkbox"/>	TSS	OR:71 OU:1 P6:2 PA:1 PK:4
<input type="checkbox"/>	Type	01 Primary Tumor:79

Step 3: Select Filtering Options
Auto-Filter to Maximum Number of Values:
Max Number of Values: 1000000 [?](#)
Exclude User Specified Batches:
Batch Type for Exclusion: TSS [?](#) [Reset Defaults](#)

Selected Batches: OU,PA	Exclude	Batch Name: Count
	<input type="checkbox"/>	OR:71
	<input checked="" type="checkbox"/>	OU:1
	<input type="checkbox"/>	P6:2
	<input checked="" type="checkbox"/>	PA:1
	<input type="checkbox"/>	PK:4

Step 4: Select a Correction (optional)
Correction Type: none [?](#) [Reset Defaults](#)

Step 5: Select PCA-Plus/DSC Arguments
Root.selectedDSCPermutations=500
Number of DSC Permutations: 500 [?](#)
Number of DSC Threads (1-5): 5 [?](#)
Minimum DSC Batch Size: 5 [?](#)
DSC Random Number Seed: 914 [?](#)
Maximum Number of Features: 2000 [?](#) [Reset Defaults](#)

Step 6: Select Boxplot Arguments
Max Number of Features: 2500 [?](#) [Reset Defaults](#)

Step 7: Select NGCHM Arguments
Perform NGCHM creation: ☒ [?](#) [Reset Defaults](#)

[Do MBatch Run](#) [Save Configuration](#) [Load Configuration](#) [Reset Defaults 3-7](#)

6.9 Do MBatch Run

After finishing setting up the assessments, the "Do MBatch Run" button is pressed. The page will update and go into the "MBatch Run Queued" state, seen at the bottom of the screenshot below. The page will automatically update as needed.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History™

Batch Effects Interface External 2018-04-24-1417

[Home](#) Job Id:1524603107862 Job State:MBATCHRUN_START_WAIT Job Message:MBatch Run Queued [Edit Details](#)

The MBatch run is queued and waiting for assignment to a processing node. [?](#)

6.10 During the MBatch Run

The MBatch run starts, which could take up to 2 minutes in an unused system or until after other users in the queue are finished (which depends on your system load). After starting, the last 100 lines from the log file are displayed. Below shows the start of the process, with the two data files from the job being loaded. (Updating log files to have more user friendly output is also on the future list of enhancements.)

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

[Home](#) **Job Id:**1524603107862 **Job State:**MBATCHRUN_RUNNING_WAIT **Job Message:**MBatch Run in Progress [Edit Details](#)

The MBatch run is underway on a processing node.

Log File Tail (last 100 lines):

```
2018 04 25 14:01:23.314 INFO 1549a9e9f75d read gene file= /BEI/OUTPUT/1524603107862/matrix_data.tsv
2018 04 25 14:01:26.842 INFO 1549a9e9f75d filter samples in batches using gene samples
2018 04 25 14:01:26.843 INFO 1549a9e9f75d sort batches by gene file samples
2018 04 25 14:01:26.943 INFO 1549a9e9f75d Finishing mbatchLoadFiles
2018 04 25 14:01:26.943 INFO 1549a9e9f75d ~~~~~
2018 04 25 14:01:26.944 DEBUG 1549a9e9f75d Changing LC_COLLATE to C for duration of run
2018 04 25 14:01:26.945 INFO 1549a9e9f75d V V V V V V V V V V V V V V V V
2018 04 25 14:01:26.945 INFO 1549a9e9f75d mbatchFilterData Starting
2018 04 25 14:01:26.946 INFO 1549a9e9f75d MBatch Version: 2017-09-19-1530
2018 04 25 14:01:26.947 DEBUG 1549a9e9f75d rows pre filter 60483
```

The assessment portion of MBatch finishes with the line marked below saying "mbatchAssess Finishing". The NGCHM creation follows this. The NGCHM creation log is not currently shown during processing, but is on the list to be added.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

[Home](#) **Job Id:**1524603107862 **Job State:**MBATCHRUN_RUNNING_WAIT **Job Message:**MBatch Run in Progress [Edit Details](#)

The MBatch run is underway on a processing node.

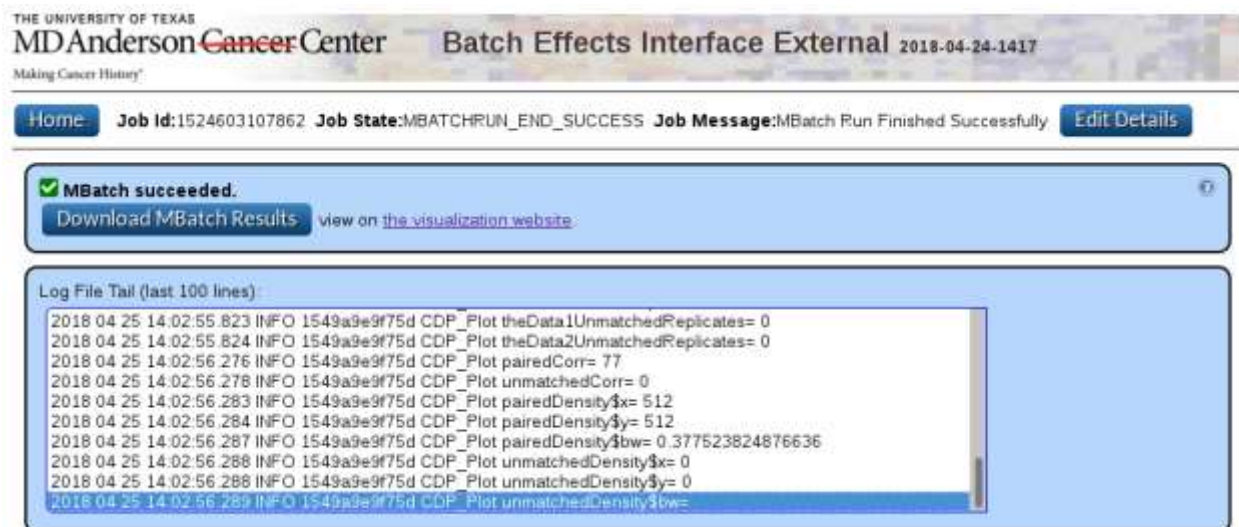
Log File Tail (last 100 lines):

```
2018 04 25 14:02:34.419 DEBUG 1549a9e9f75d writeAsDataframe - theFile /BEI/OUTPUT/1524603107862/MB
2018 04 25 14:02:34.421 DEBUG 1549a9e9f75d writeAsDataframe - length(myData) 154
2018 04 25 14:02:34.421 DEBUG 1549a9e9f75d writeAsDataframe - length(myCols) 2
2018 04 25 14:02:34.422 DEBUG 1549a9e9f75d writeAsDataframe - length(myRows) 0
2018 04 25 14:02:34.422 DEBUG 1549a9e9f75d writeAsDataframe - Calling .jinit /usr/local/lib/R/site-library/MBa
2018 04 25 14:02:34.523 DEBUG 1549a9e9f75d writeAsDataframe - .jinit complete
2018 04 25 14:02:34.524 DEBUG 1549a9e9f75d writeAsDataframe before java
2018 04 25 14:02:34.545 DEBUG 1549a9e9f75d writeAsDataframe after java
2018 04 25 14:02:34.546 DEBUG 1549a9e9f75d writeAsDataframe success= TRUE
2018 04 25 14:02:34.548 INFO 1549a9e9f75d mbatchAssess Finishing
```


6.11 Finished Job

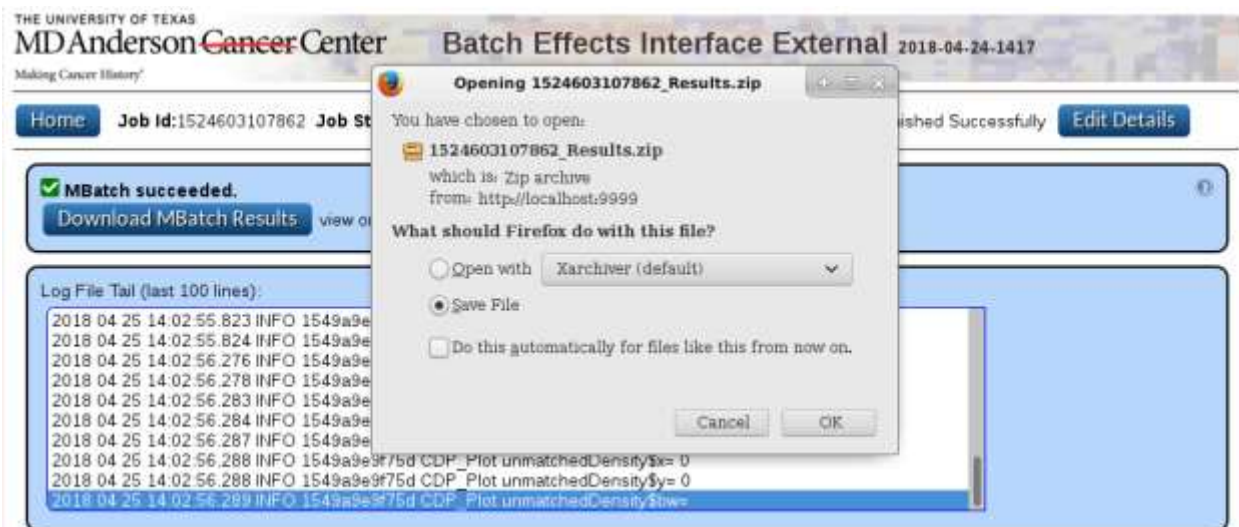
After a job has finished, the user is presented with a "Download MBatch Results" button that offers the ability to download the complete job directory, with data used, downloaded, configuration files, and results. (The Download option is described in more detail later.)

If successful, the user will also be presented with a link to the Batch Effects Viewer website, to view the results.

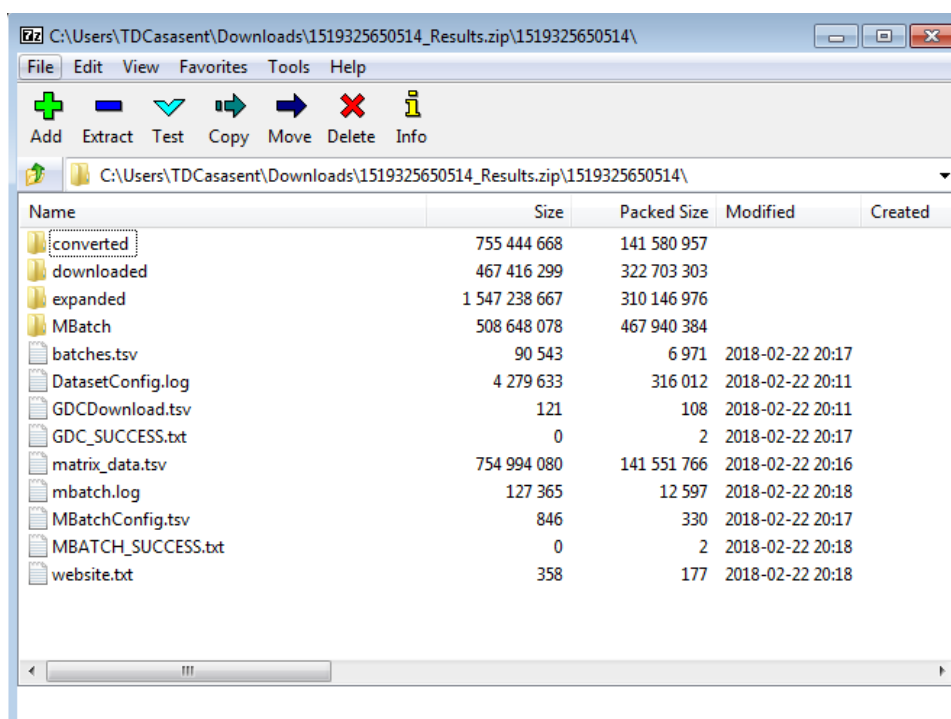


7 Download MBatch Results

If the user presses the Download MBatch Results button, the job directory will be compressed and sent to the browser. (The exact download process will vary by browser. This screenshot is Firefox setup to ask the user to Open or Save.)



Within the ZIP archive are the following directories and files.



The directories can mostly be ignored by users. The MBatch directory has results that are used for display by the Batch Effects Viewer. To save space, the converted, downloaded, and expanded directories can be deleted if desired. The downloaded directory contains the files downloaded from the GDC. The expanded directory contains the uncompressed version of the GDC files. The converted directory has the converted data, batch file, covariates file with information like demographics, and annotations file with information like the value added before log normalization, if performed.

Filename	Contents
batches.tsv	Batch file for data. See user data documentation for file format.
DatasetConfig.log	Log from downloading GDC data. This is not particularly user-friendly at present.
GDCDownload.tsv	Contains a description of the data downloaded from the GDC.
GDC_SUCCESS.txt	An empty "flag" file indicating successful completion of the GDC Download process.
matrix_data.tsv	Data file for data. See user data documentation for file format.
mbatch.log	Log from MBatch run. This is not particularly user-friendly at present.
MBatchConfig.tsv	Configuration file for MBatch assessments and corrections.
MBATCH_SUCCESS.txt	An empty "flag" file indicating successful completion of the MBatch run.
website.txt	Timestamp and links to the original job and output.

8 View MBatch Results

For details on the Batch Effects Viewer website, please see

<http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview>

This document will be updated when a new version of the Batch Effects Viewer, currently in development, becomes available.

9 Job List

Pressing the Home button takes the user to the main page, with a list of jobs.

Clicking on the Select Job button takes the user to the results page, which also has the Update Job Info. Jobs can also be deleted. The external version of Batch Effects Interface has not security on viewing, running, or deleting jobs.