BEStack 02B Using Batch Effects Interface Corrections: MP
Tod Casasent
2018-04-30-1015

1 Using Batch Effects Interface Corrections: MP

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:

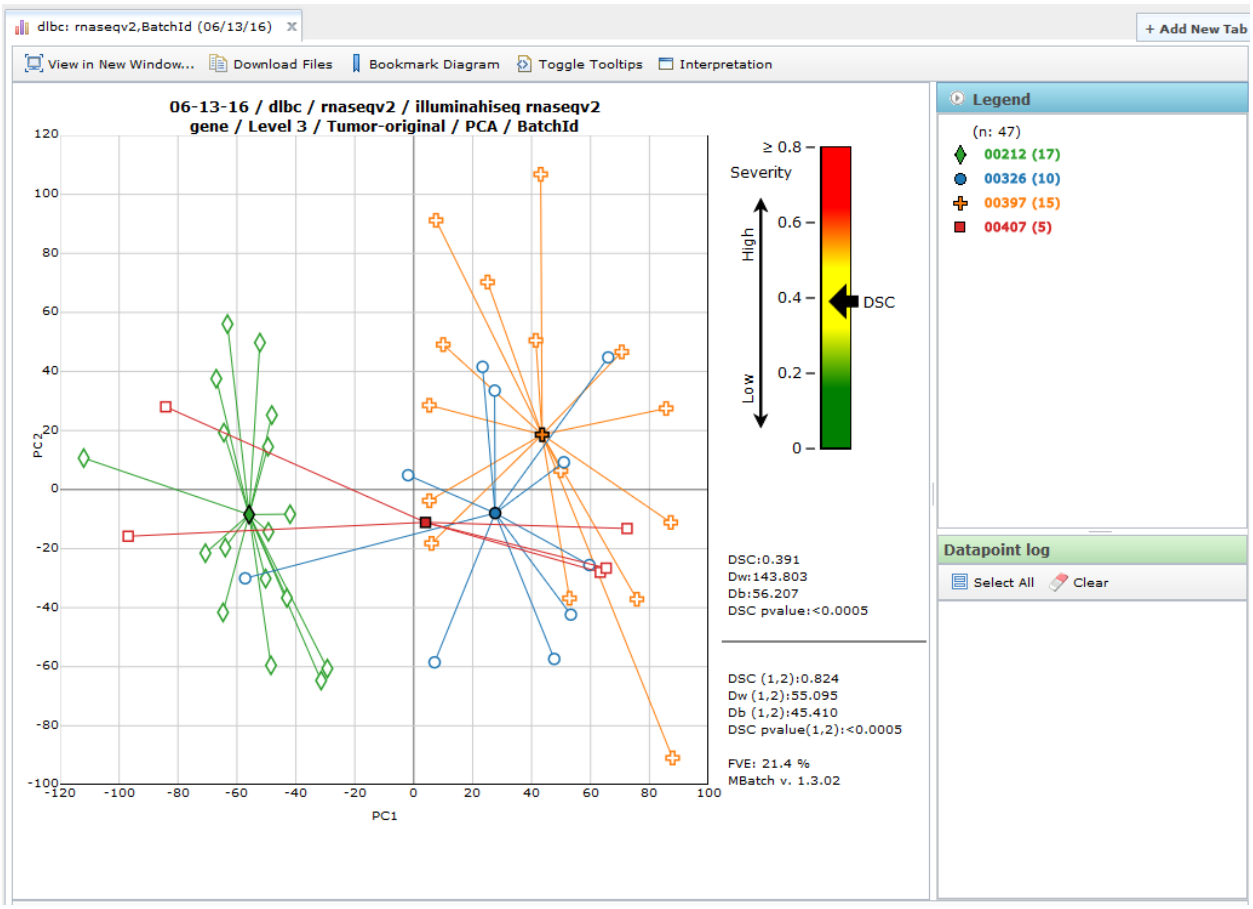
> http://your-server.your-company.com:9999/BatchEffectsInterface/

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 at https://github.com/MD-Anderson-Bioinformatics/MBatch/tree/master/pdf for instructions on downloading test data.

RBN corrections uses replicates between the two datasets to combine the two data sets based on replicates between sets.

For demonstrating Median Polish (MP) corrections, we will use the DLBC RNASeq2 dataset found at: http://bioinformatics.mdanderson.org/TCGA/databrowser You can see the Batch Effects Assessment here: http://bioinformatics.mdanderson.org/BatchEffects/index.jsp?path=%252F2016_06_13_0834-2016_08_16_1052%252Fdlbc%252Frnaseqv2%252Filluminahiseq_rnaseqv2_gene%252FLevel_3%252FTumor-original%252FPCA%252FShipDate%252FManyToMany&xaxis=PC1&yaxis=PC2

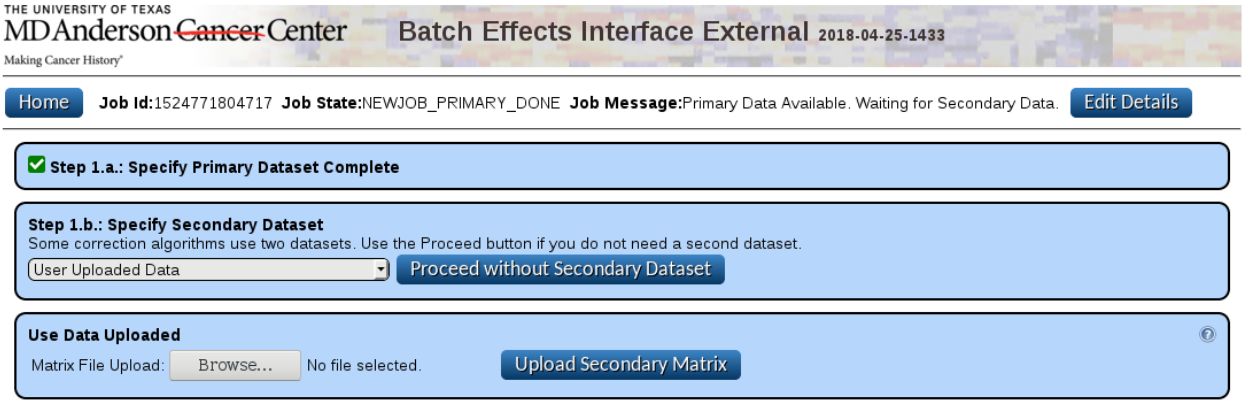
For this exercise, we have no idea if this is a real batch effect or biology, we are just using this as an example of removing a batch effects. You can see batch 00212 way to the left in green.



2 Starting a Job

See BEStack_02A_BEIUsingAssessments for more details about starting a job.

Use the "Start New Job" button and select "User Uploaded Data" for Step 1.a. Use the data matrix and batch file from the Standardized Data Browser website given above.



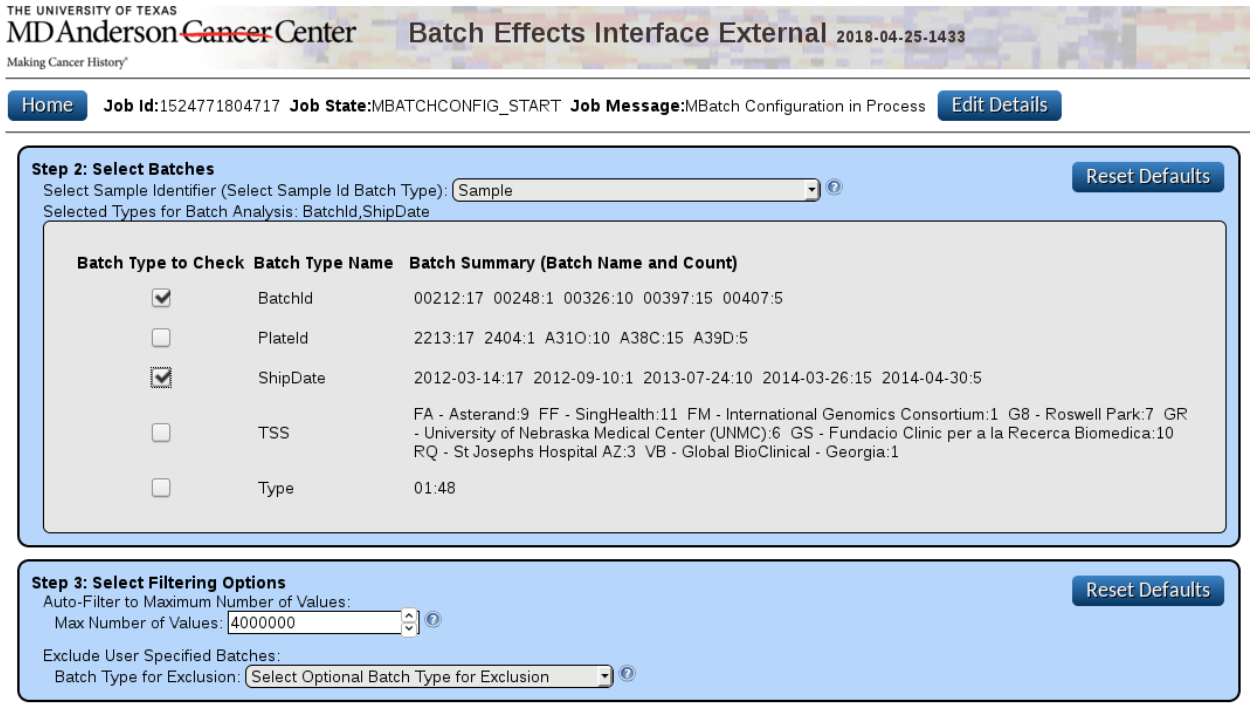
Then for Step 1.b. select Proceed without Secondary Dataset.

3 Configuring Assessments

See BEStack_02A_BEIUsingAssessmentsExternal for more details about Configuring Assessments.

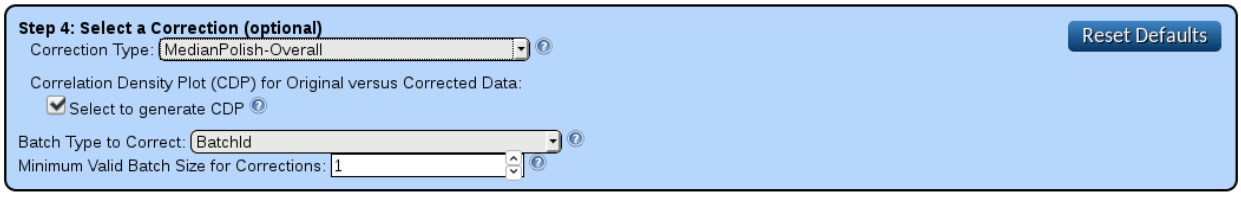
Below, we have selected Sample as the Sample Identifier, and selected BatchId as well as ShipDate as the assessment batch types. For Step 3, we have kept the defaults.

Note that the batch type you wish to correct must be one of the batch types checked.



3.1 Step 4 Median Polish Overall

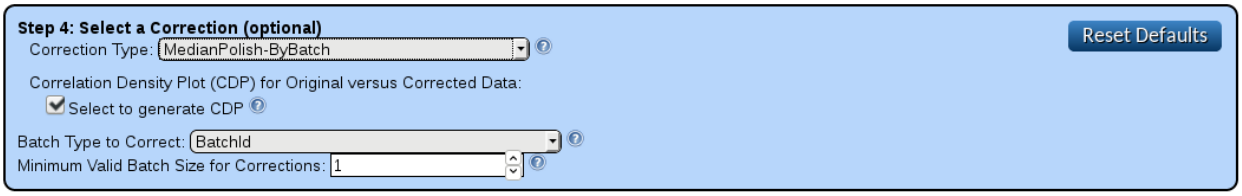
In Step 4, we begin by selecting MedianPolish-Overall as the optional Correction Type. (The batch type to correct and minimum batch size are not appropriate for MP-Overall and will be removed.)



We accept the defaults for the rest of the data.

3.2 Step 4 Median Polish By Batch

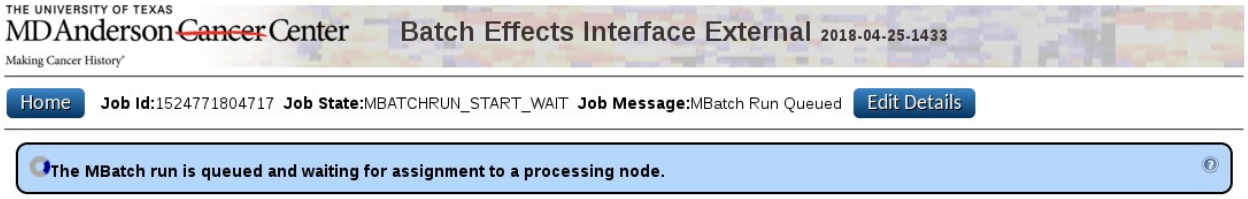
In Step 4, if we want “By Batch” instead, we select MedianPolish-ByBatch as the optional Correction Type. Since we want the problem within the Batch Id type corrected, we select that.



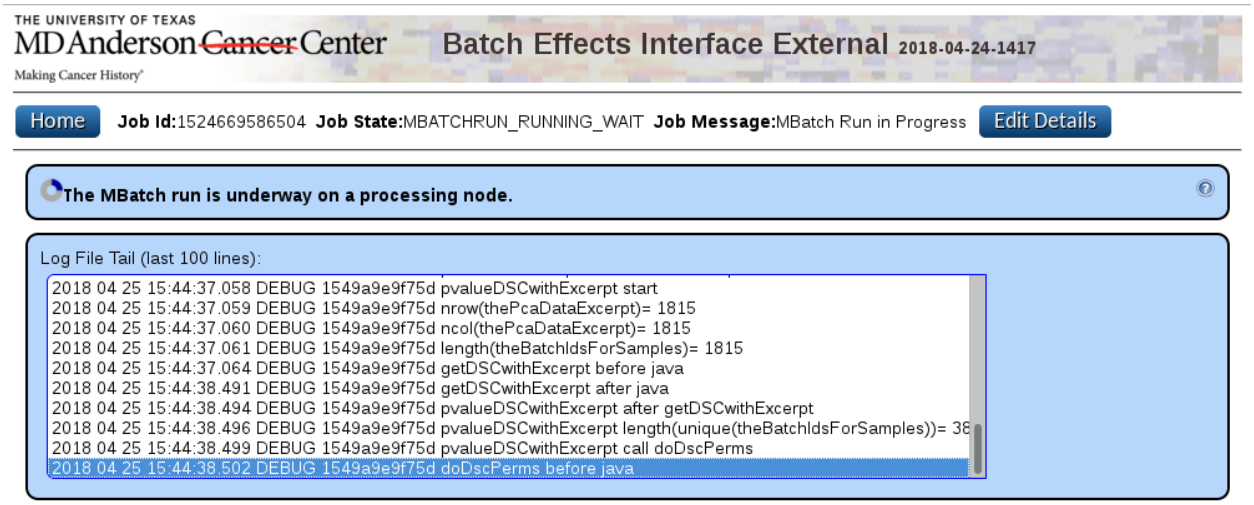
We accept the defaults for the rest of the data.

3.3 Do MBatch Run

See BEStack_02A_BEIUsingAssessments for more details about running and monitoring a run. We press Do MBatch Run from the configuration page.

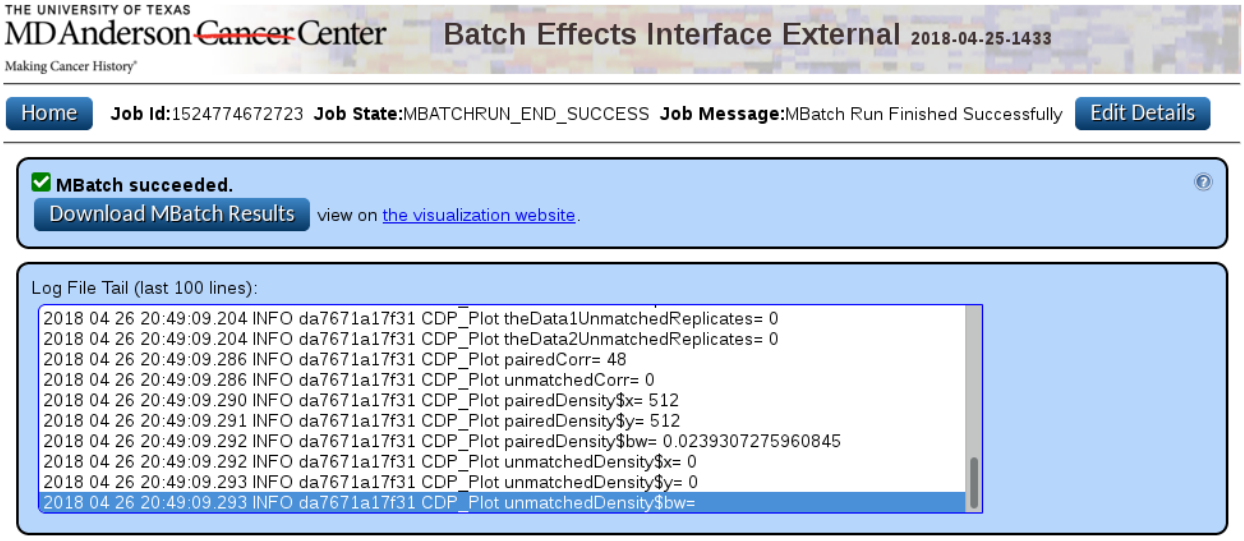


The DSC Permutations step will take some time, which can take 30+ minutes.

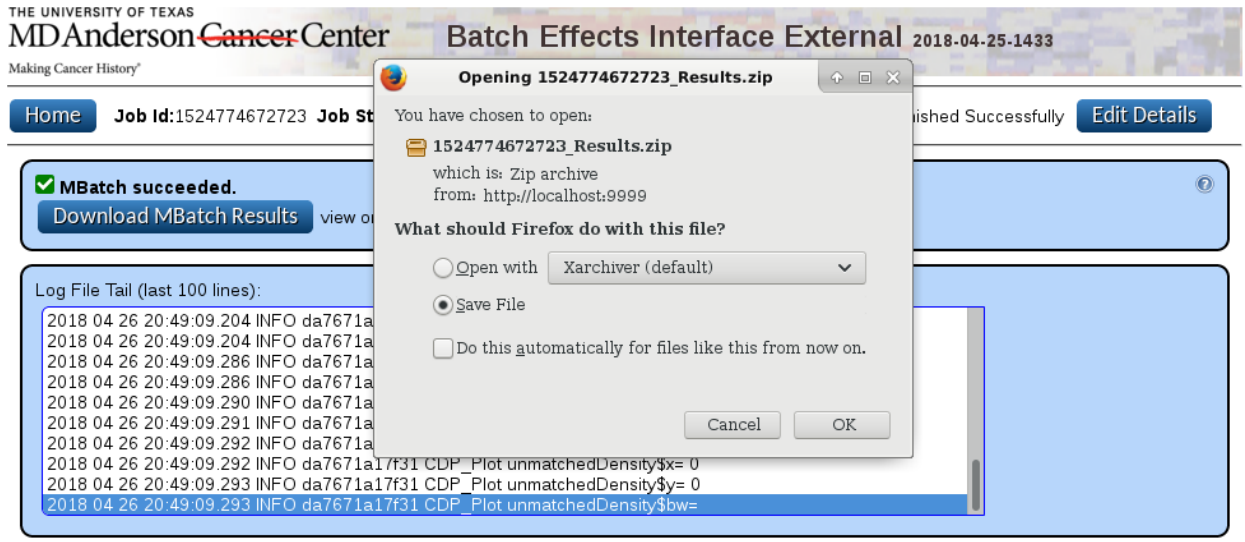


3.4 Finished Job

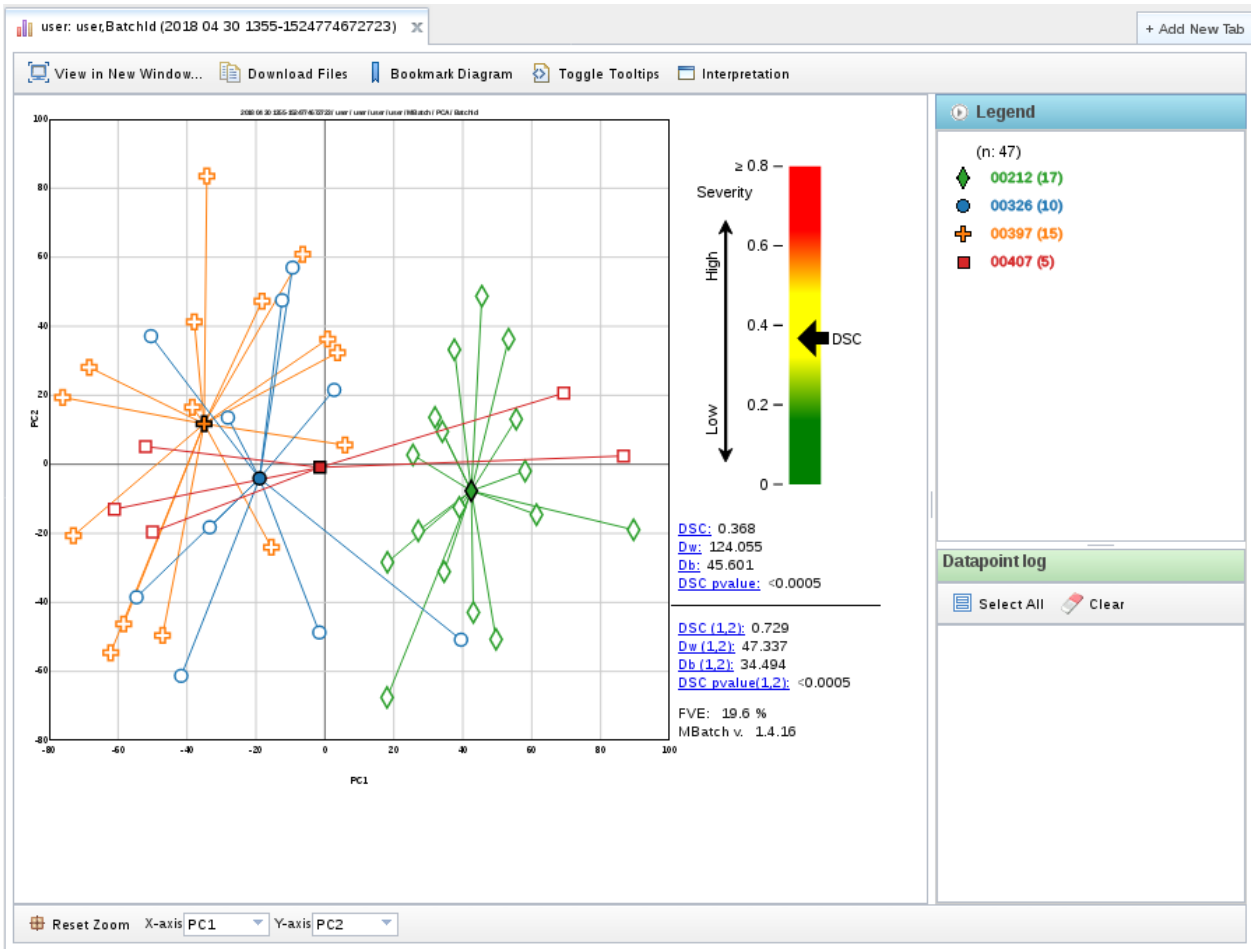
After the job has finished, use the Download option to get the corrected data. Here, for reasons to be explained soon, we look at the MP-Overall Example.



Click the Download MBatch Results button. Open or unzip the archive and enter the MBatch directory.



When we examine the PCA plot, we see the correction has been performed with minimal effectiveness. (We will not comment on whether or not this has removed biological effects.)



The ANY_Corrections-MPOverall.tsv file contains the corrected data. Looking at an excerpt from that file below, you see the group ids have been added to the end of the sample ids (with a period to separate them).

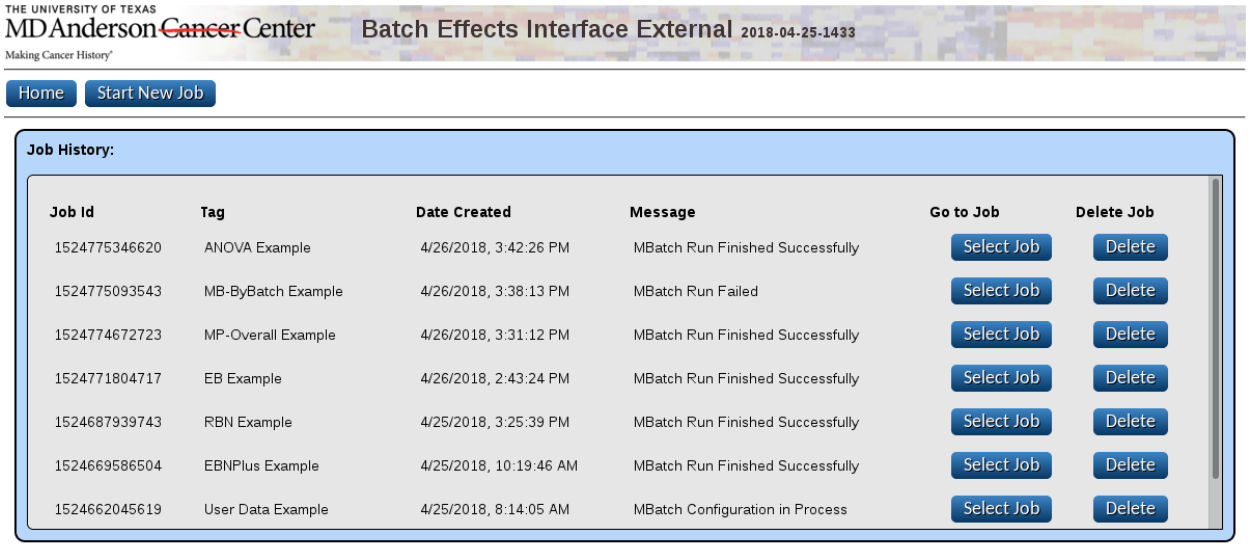
	TCGA-FA-8693-01A-11R-2404-07	TCGA-FA-A4BB-01A-11R-A31O-07	TCGA-FA-A4XK-01A-11R-A31O-07	TCGA-FA-A6HN-01A-11R-A31O-07
A1BG|1	7.576146043	6.531136709	6.694689663	7.980712792
A1CF|29974	6.98560411	6.945400497	6.98560411	6.980700672
A2BP1|54715	6.98560411	6.945400497	6.98560411	6.980700672
A2LD1|87769	6.371952521	7.069288154	7.663499093	6.593223945
A2ML1|144568	7.703779796	6.386542207	6.42674582	7.705116548
A2M|2	6.858983434	6.663633291	7.031545702	6.191399686
A4GALT|53947	8.252804603	5.776492658	6.638236223	5.997549387
A4GNT|51146	10.61079289	6.93980883	6.980012443	6.975109005

The BatchData.tsv contains the batch files. Looking at an excerpt from that file below, you see the batch data used.

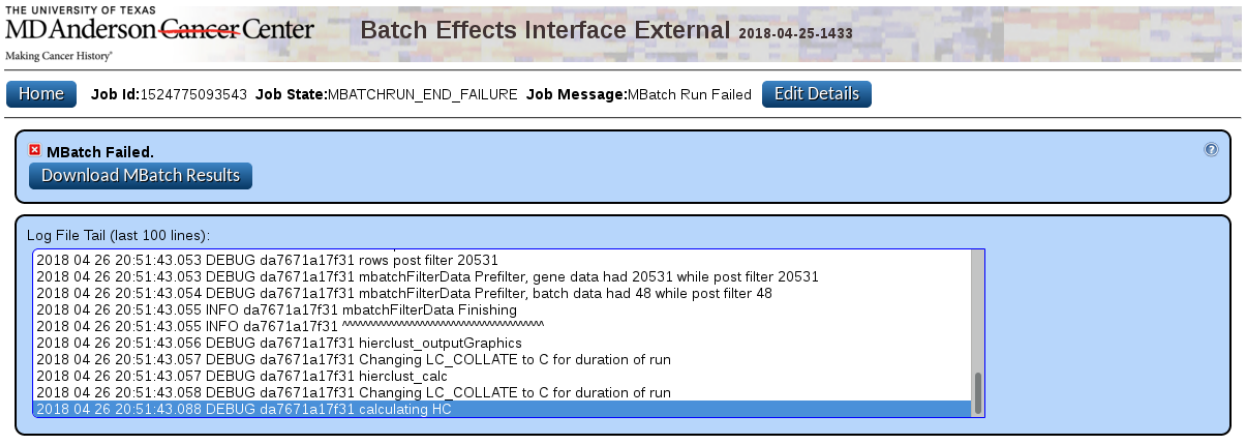
Sample	BatchId	ShipDate
TCGA-FA-8693-01A-11R-2404-07	00248	2012-09-10
TCGA-FA-A4BB-01A-11R-A31O-07	00326	2013-07-24
TCGA-FA-A4XK-01A-11R-A31O-07	00326	2013-07-24
TCGA-FA-A6HN-01A-11R-A31O-07	00326	2013-07-24
TCGA-FA-A6HO-01A-11R-A31O-07	00326	2013-07-24
TCGA-FA-A7DS-01A-11R-A38C-07	00397	2014-03-26
TCGA-FA-A7Q1-01A-11R-A38C-07	00397	2014-03-26
TCGA-FA-A82F-01A-11R-A38C-07	00397	2014-03-26

3.5 Failed Job

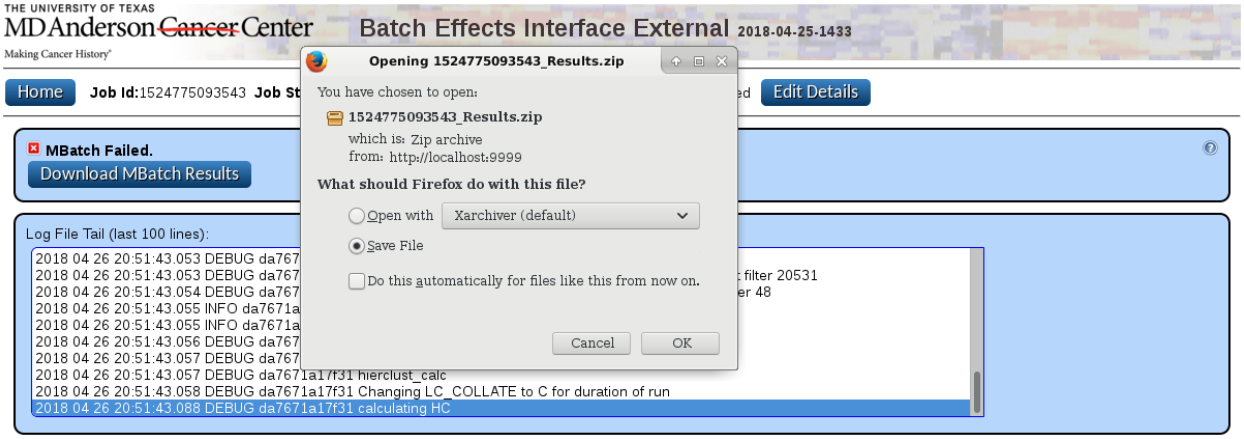
Looking at the jobs, we notice the MB-ByBatch Example failed. Here, we will look at the log files.



The job status shows failed.



We download the results as before.



We open the archive and look at the mbatch.log file at the top level. At the end of that file we see these five lines:

2018 04 26 20:51:43.056 DEBUG da7671a17f31 hierclust_outputGraphics
2018 04 26 20:51:43.057 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.057 DEBUG da7671a17f31 hierclust_calc
2018 04 26 20:51:43.058 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.088 DEBUG da7671a17f31 calculating HC

So, we know that something happened within the hierarchical clustering calculations for this data set. This is generally a limitation of the data set itself (unsolvable solutions) or of the third party clustering algorithm. Generally, the “limitation” type errors are caught within the code and only cause a problem for the one diagram type. In this case, since there is no error message, the problem is most likely that that relatively unlimited parameters used to process the data used more than the 24GB available on the machine on which this was run and cause R to stop without notice.