BEStack 02B Using Batch Effects Interface Corrections: RBN
Tod Casasent
2018-06-06-0915

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:9999/BatchEffectsInterface/

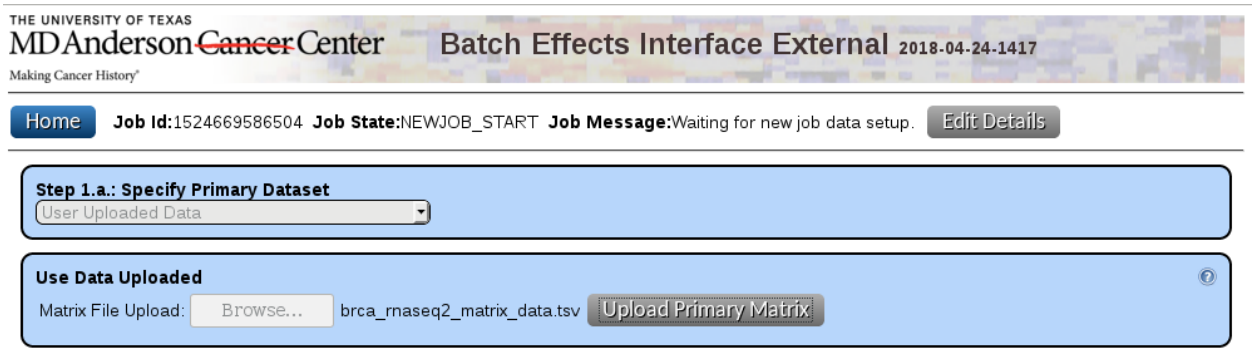
These instructions are aimed at people familiar with R and familiar with TCGA/GDC platforms and data types. They are intended to introduce the reader to producing the given assessment. These instructions will only rarely, if ever, touch on the appropriateness of the assessment algorithm or interpretation 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.

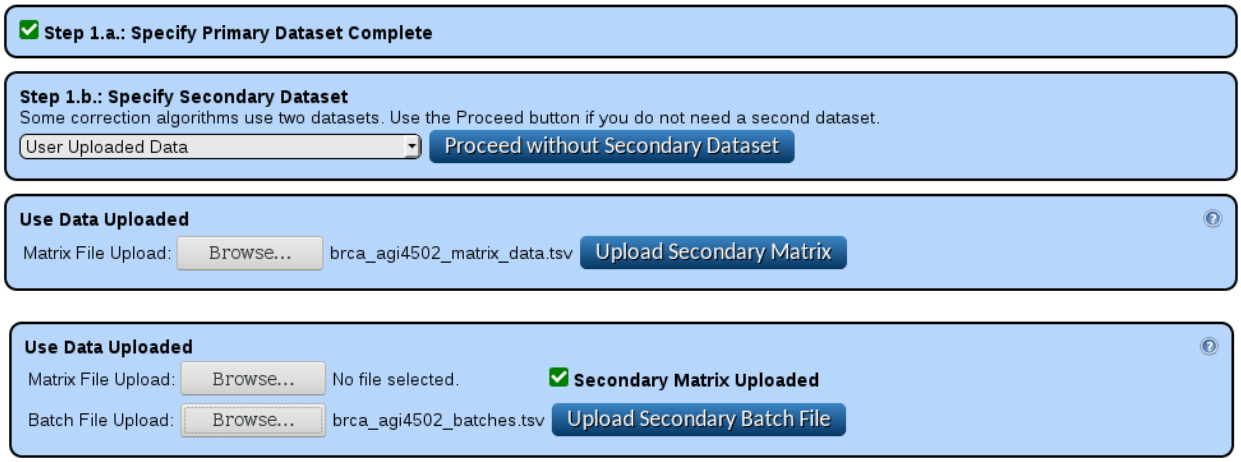
Starting a Job

See BEStack_02A_BEIUsingAssessments for more details about starting a job.

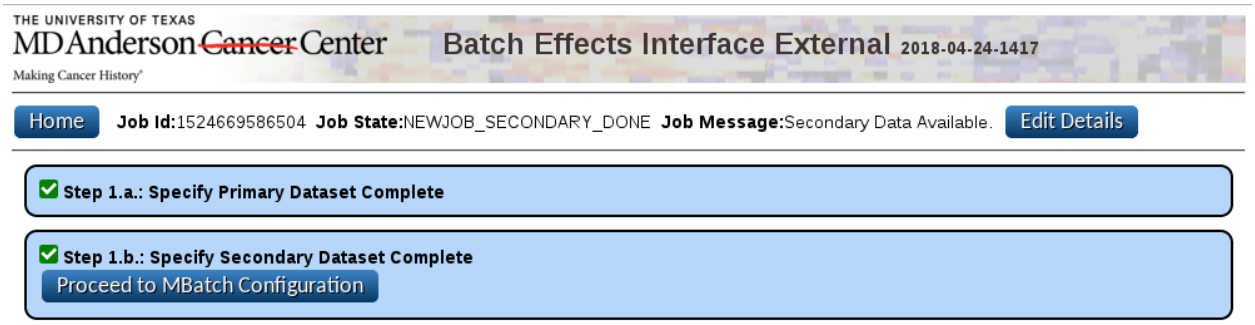
Use the "Start New Job" button and select "User Uploaded Data" for Step 1.a. From within the MATRIX_DATA.zip archive, upload brca_rnaseq2_matrix_data.tsv as the data matrix and brca_rnaseq2_batches.tsv as the batch matrix.



Then for Step 1.b. also select User Uploaded Data, and use brca_agi4502_matrix_data.tsv for the Matrix File and brca_agi4502_batches.tsv for the Batch File.



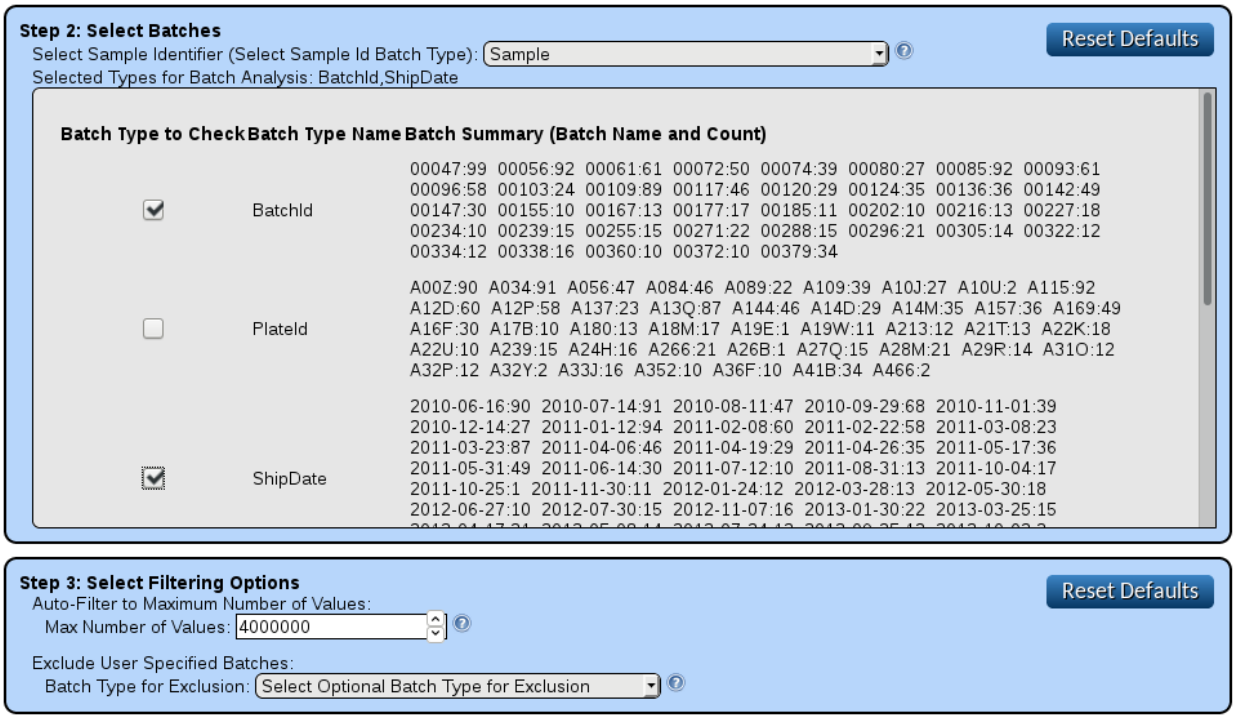
Then select Proceed to MBatch Configuration.



Configuring and Running Assessments

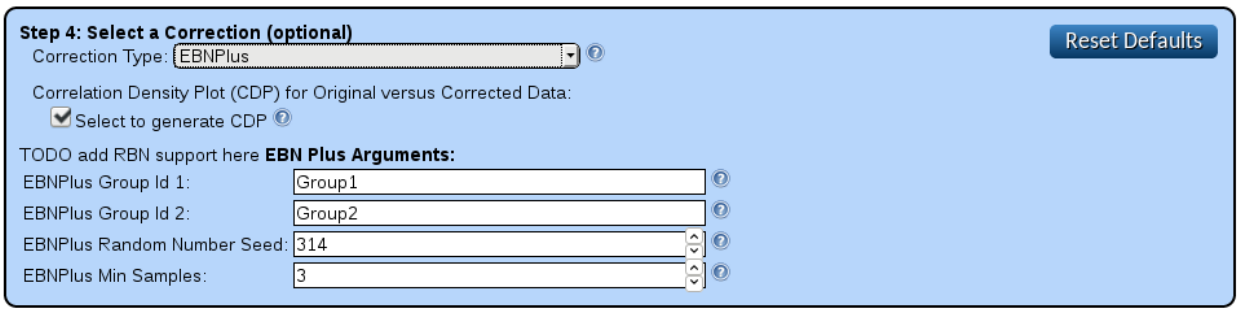
See BEStack_02A_BEIUsingAssessmentsExternal for more details about Configuring Assessments.

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.

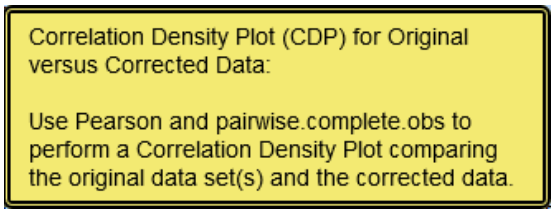


Step 4 (RBN with Replicates)

In Step 4, we begin by selecting RBN as the optional Correction Type.



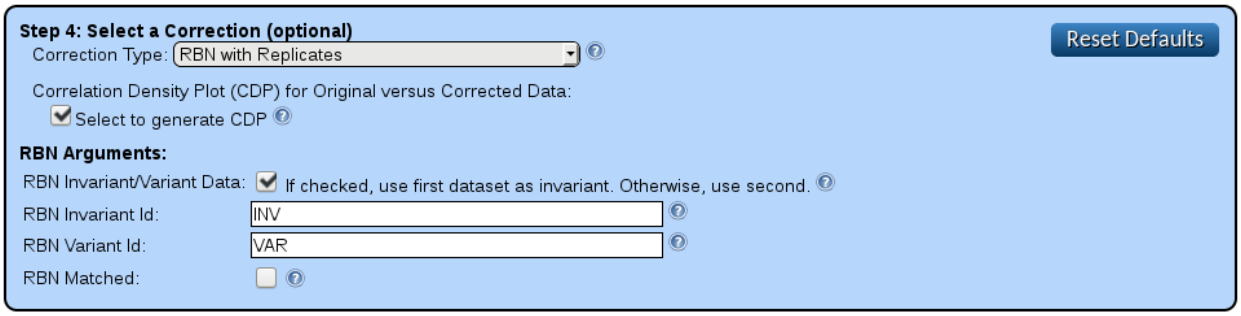
Generating a Correlation Density Plot is selected automatically. (In a future version, this option will appear for any assessment, rather than just for certain correction options.)



Tooltip Text for Correlation Density Plot

The Group Ids must be alphanumeric values without spaces.

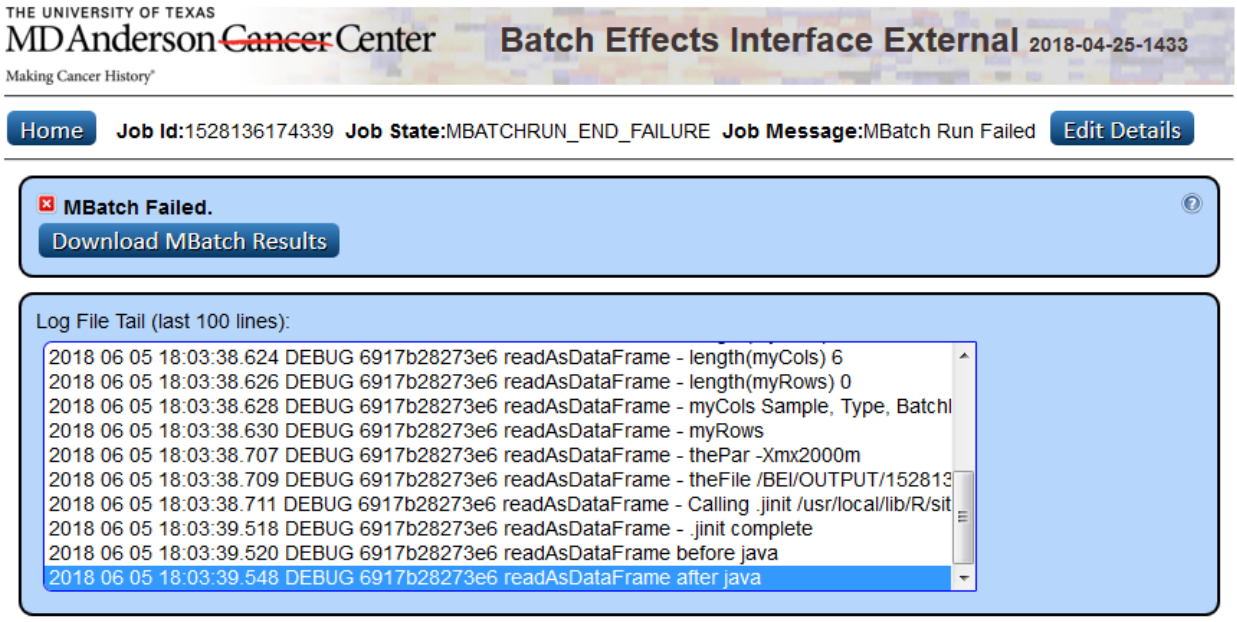
Here, we have kept all the default settings. Invariant and Variant ids are alphanumeric values without spaces but with underscores allowed.



We do not use these for this run, instead we use the other RBN option.

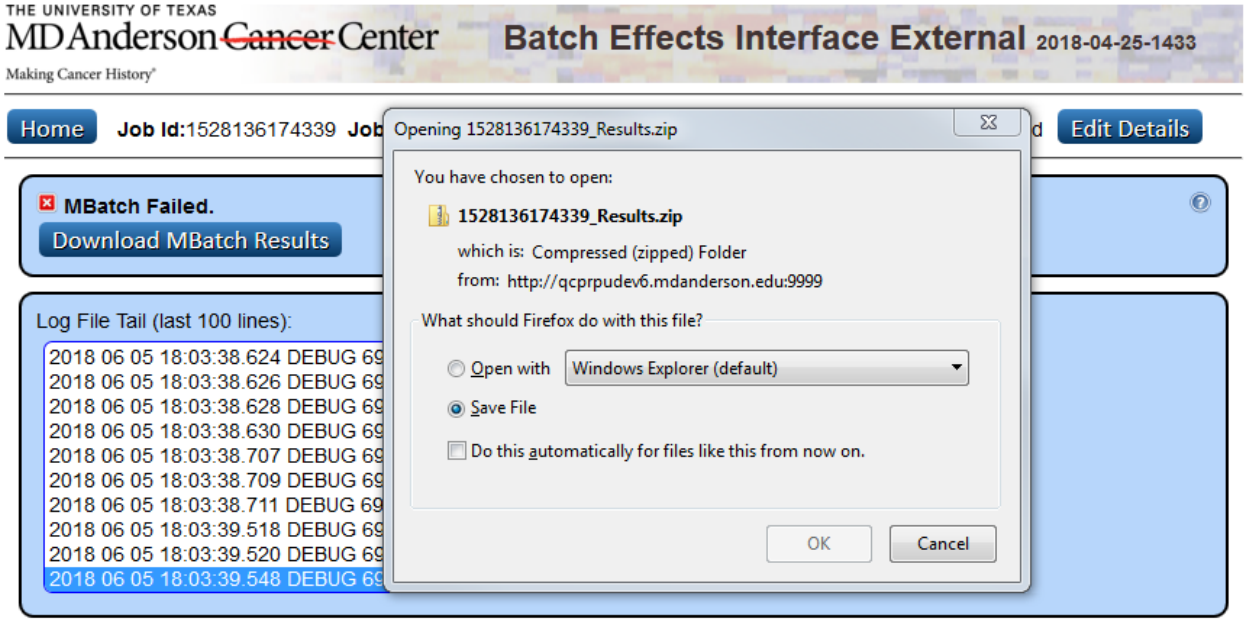
Do MBatch Run

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

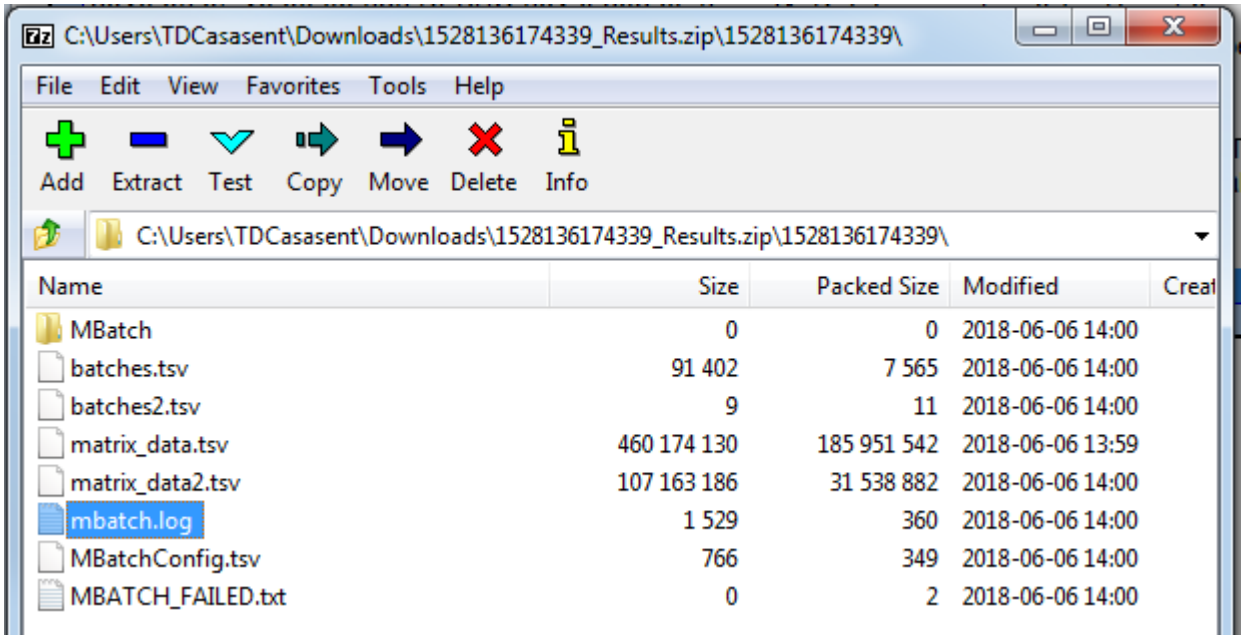


Failed Job

After the job has finished, use the Download MBatch Results button to get the results, which include a log file.



Find the mbatch.log file inside the one directory (with the job id) in the ZIP file.



Extract and view the log file. It will look much like the tail from the web page.

2018 06 05 18:03:37.535 DEBUG 6917b28273e6 readAsDataFrame - thePar -Xmx2000m

2018 06 05 18:03:37.572 DEBUG 6917b28273e6 readAsDataFrame - theFile /BEI/OUTPUT/1528136174339/batches.tsv

2018 06 05 18:03:37.572 DEBUG 6917b28273e6 readAsDataFrame - Calling .jinit /usr/local/lib/R/site-library/MBatch/ReadRJava/ReadRJava.jar

2018 06 05 18:03:38.355 DEBUG 6917b28273e6 readAsDataFrame - .jinit complete

2018 06 05 18:03:38.357 DEBUG 6917b28273e6 readAsDataFrame before java

2018 06 05 18:03:38.616 DEBUG 6917b28273e6 readAsDataFrame after java

2018 06 05 18:03:38.622 DEBUG 6917b28273e6 readAsDataFrame - length(myData) 7290

2018 06 05 18:03:38.624 DEBUG 6917b28273e6 readAsDataFrame - length(myCols) 6

2018 06 05 18:03:38.626 DEBUG 6917b28273e6 readAsDataFrame - length(myRows) 0

2018 06 05 18:03:38.628 DEBUG 6917b28273e6 readAsDataFrame - myCols Sample, Type, BatchId, PlateId, ShipDate, TSS

2018 06 05 18:03:38.630 DEBUG 6917b28273e6 readAsDataFrame - myRows

2018 06 05 18:03:38.707 DEBUG 6917b28273e6 readAsDataFrame - thePar -Xmx2000m

2018 06 05 18:03:38.709 DEBUG 6917b28273e6 readAsDataFrame - theFile /BEI/OUTPUT/1528136174339/batches2.tsv

2018 06 05 18:03:38.711 DEBUG 6917b28273e6 readAsDataFrame - Calling .jinit /usr/local/lib/R/site-library/MBatch/ReadRJava/ReadRJava.jar

2018 06 05 18:03:39.518 DEBUG 6917b28273e6 readAsDataFrame - .jinit complete

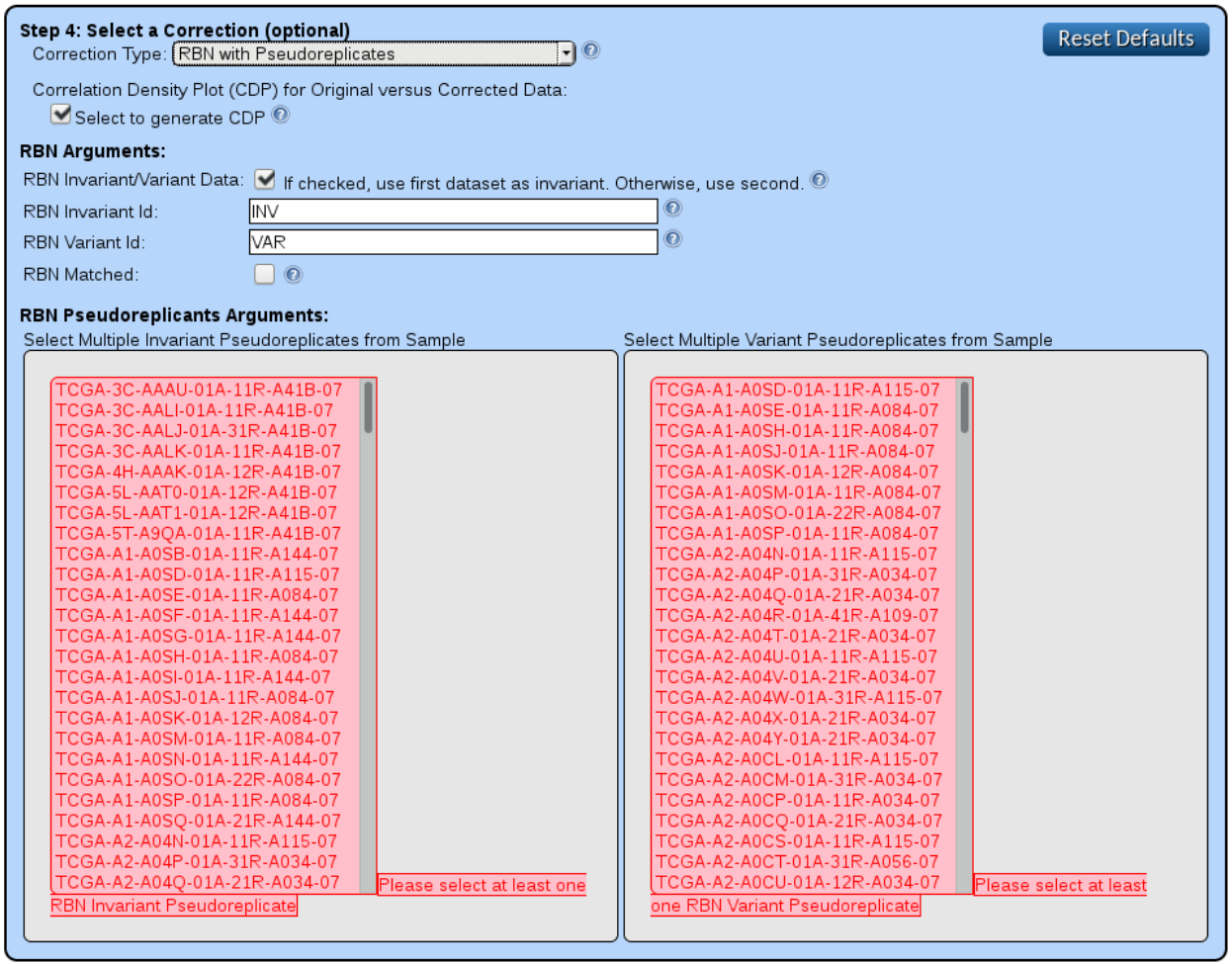
2018 06 05 18:03:39.520 DEBUG 6917b28273e6 readAsDataFrame before java

2018 06 05 18:03:39.548 DEBUG 6917b28273e6 readAsDataFrame after java

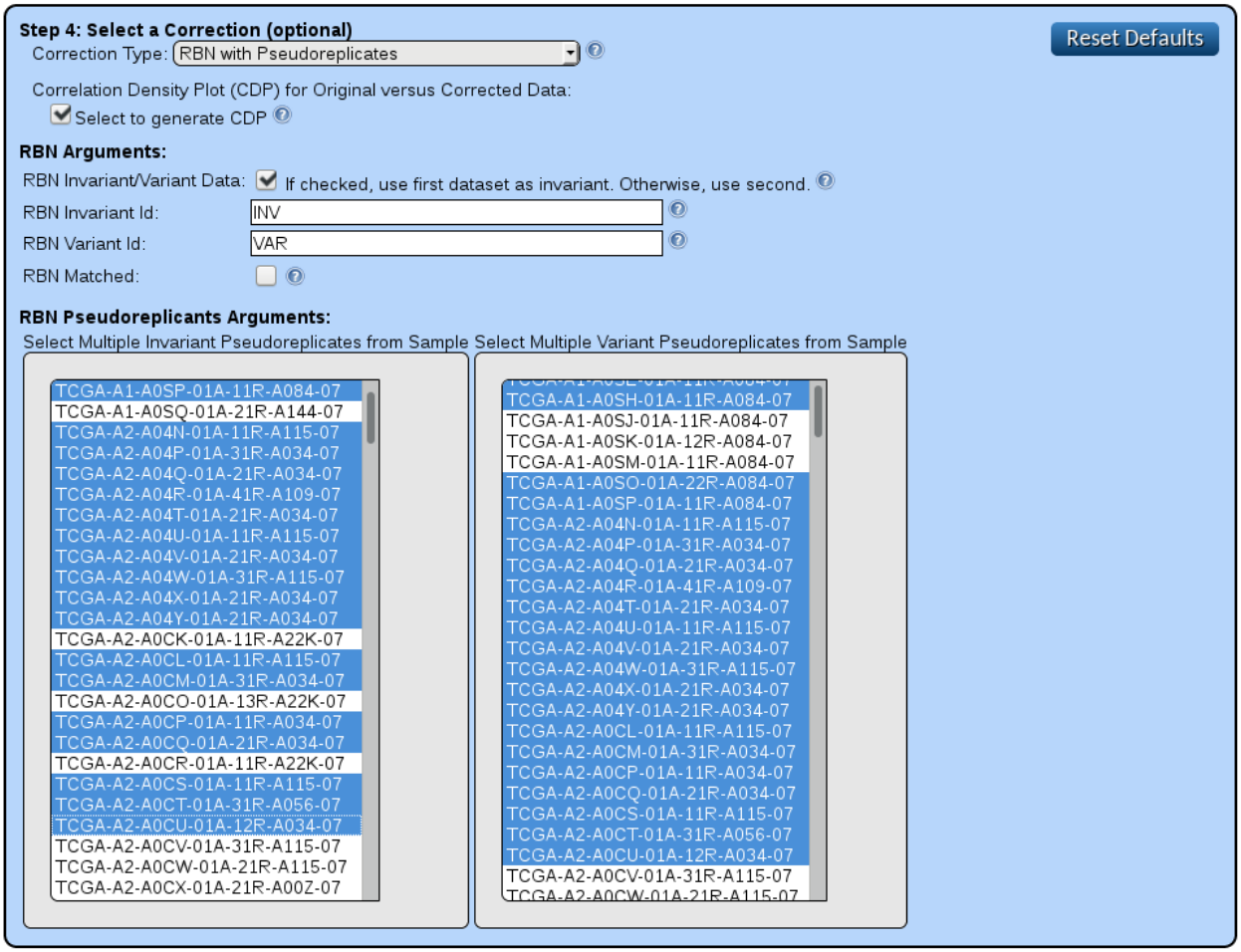
Notice that there is no error or any sort of exception in the log file. This generally means that the memory requirements for processing the data exceeded the memory available, and that R crashed. The system detected this crash, and marked the job as failed.

Step 4 (RBN with Pseudoreplicates)

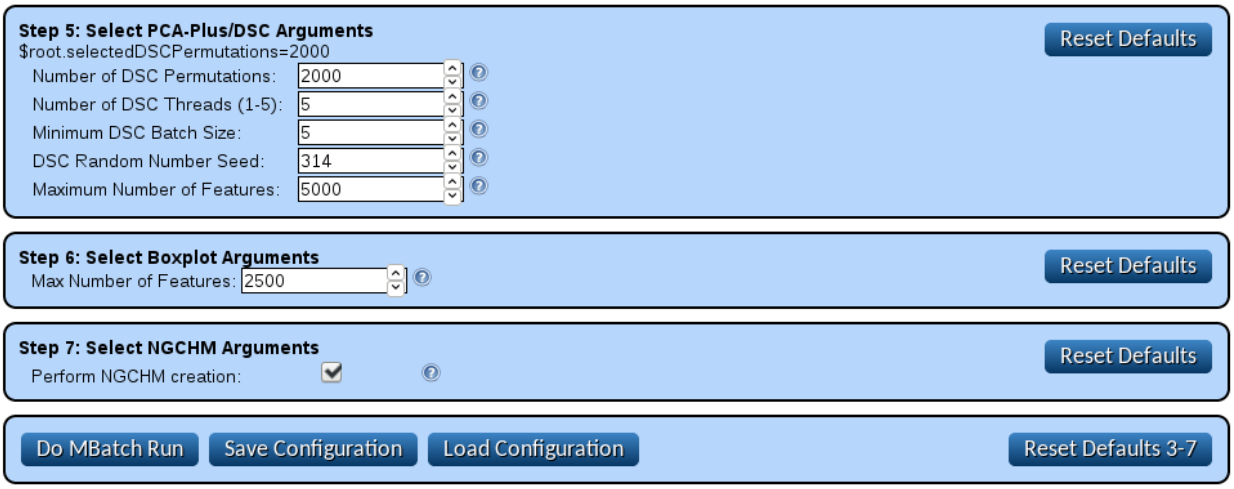
Pseudoreplicates are used when you have replicates, or something you consider "close enough", in the dataset, but they have different sample ids. When the Pseudoreplicates option comes up, it looks similar to the Replicates option, except for the lists from which to select Pseudoreplicates. We recommend at least 30 pseudoreplicates in each group.



Here, we have selected pseudoreplicates from each group.

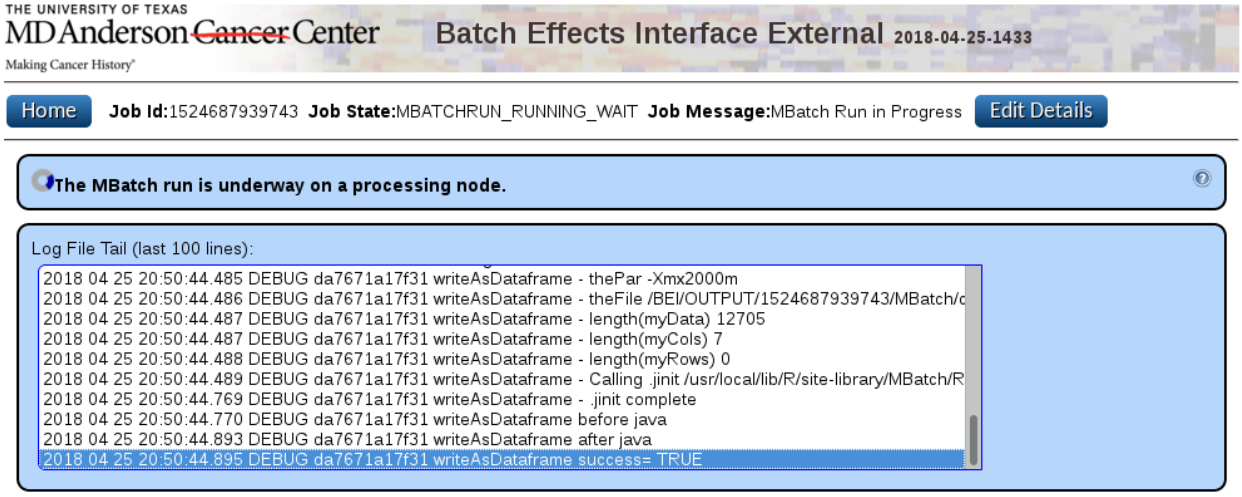


And we take the defaults for the assessments.

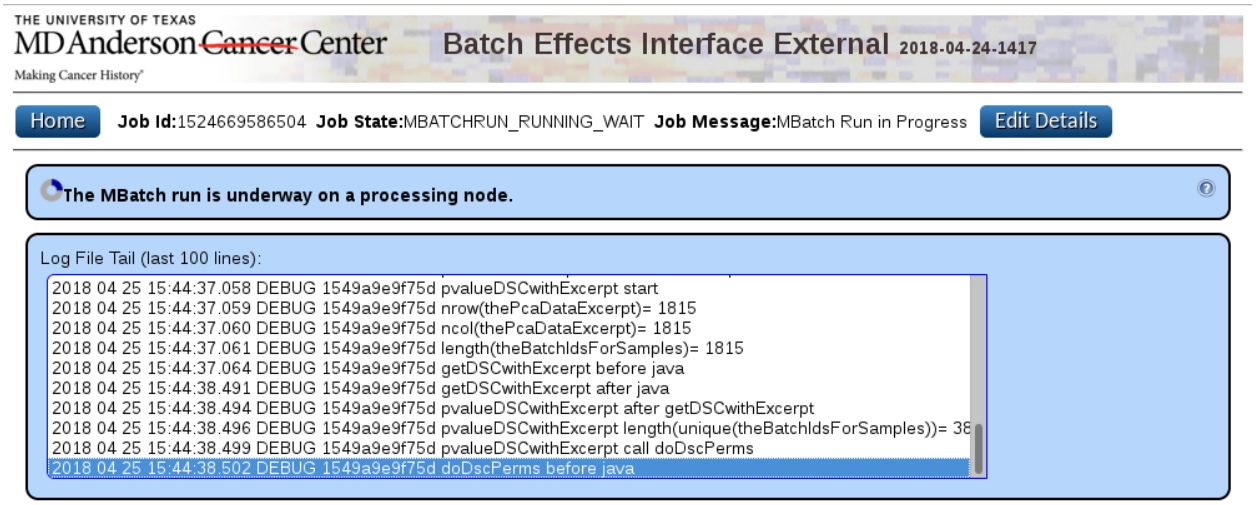


Do MBatch Run

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

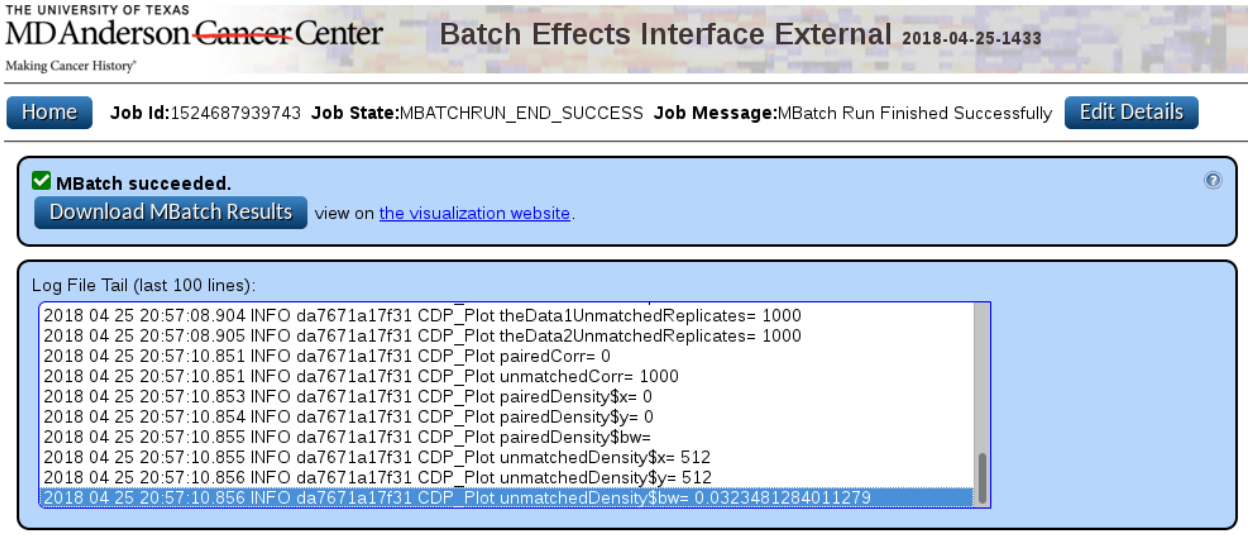


The DSC Permutations step will take some time—30 minutes or more with the full run taking several hours.

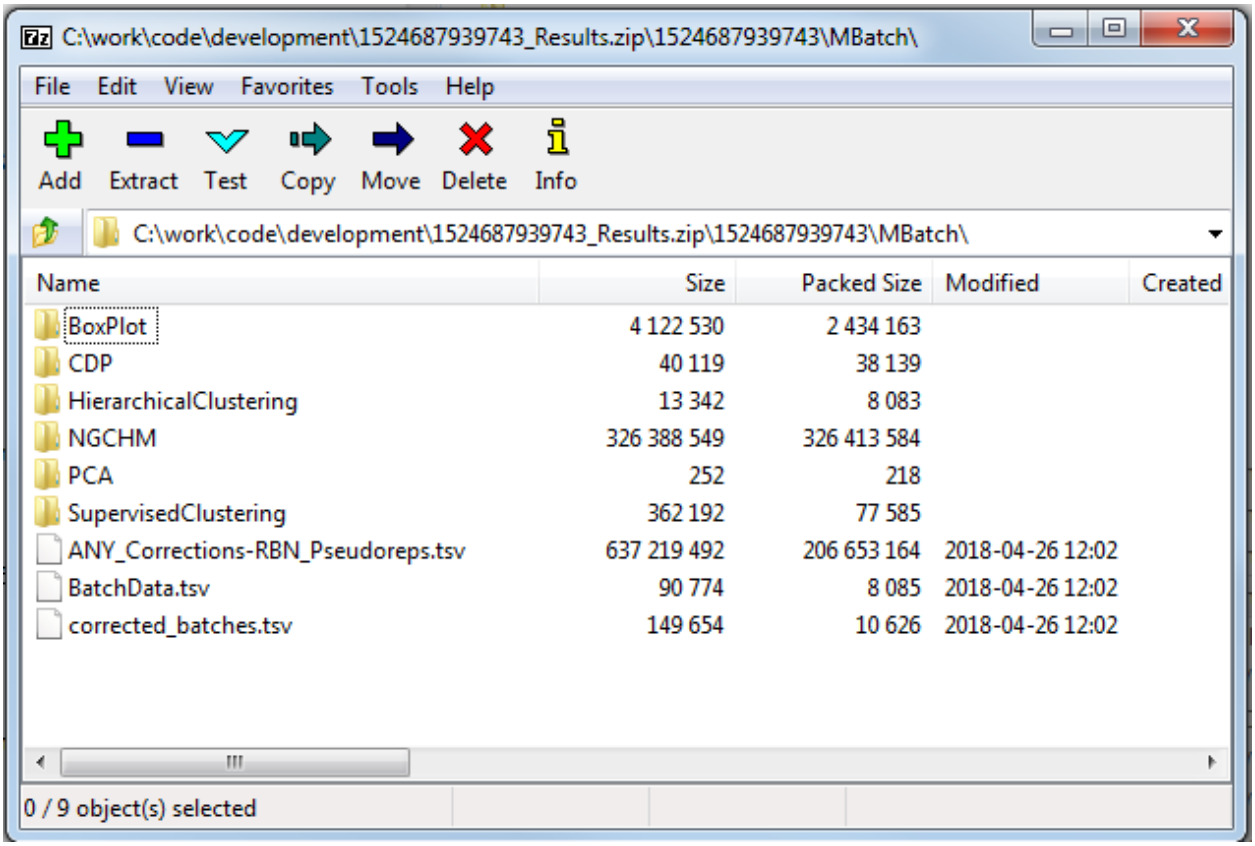


Finished Job

After the job has finished, use the Download option to get the corrected data.



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



The ANY_Corrections-RBN_Pseudoreps.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-Z7-A8R5-01A-42R-A41B-07-INV	TCGA-Z7-A8R6-01A-11R-A41B-07-INV	TCGA-
A1BG|1	8.780670006	7.957533334	NA
A1CF|29974	-2.016249352	-2.016249352	NA
A2BP1|54715	-0.243830672	-2.016249352	NA
A2LD1|87769	6.349581128	4.668675122	NA
A1BG	NA	NA	0.94933
A2BP1	NA	NA	0.53916
A2M	NA	NA	0.242
A2ML1	NA	NA	0.4235

The corrected_batches.tsv contains the combined batch files. 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).

Sample	Type	BatchId	PlateId	ShipDate	TSS
TCGA-XX-A899-01A-11R-A36F-07-INV	01	00372	A36F	2014-01-29	XX - Spectrum Health
TCGA-XX-A89A-01A-11R-A36F-07-INV	01	00372	A36F	2014-01-29	XX - Spectrum Health
TCGA-Z7-A8R5-01A-42R-A41B-07-INV	01	00379	A41B	2014-05-28	Z7 - John Wayne Cance
TCGA-Z7-A8R6-01A-11R-A41B-07-INV	01	00379	A41B	2014-05-28	Z7 - John Wayne Cance
TCGA-A1-A0SD-01A-11R-A115-07-VAR	01	00085	A115	2011-01-12	A1 - UCSF
TCGA-A1-A0SE-01A-11R-A084-07-VAR	01	00072	A084	2010-09-29	A1 - UCSF
TCGA-A1-A0SH-01A-11R-A084-07-VAR	01	00072	A084	2010-09-29	A1 - UCSF
TCGA-A1-A0SJ-01A-11R-A084-07-VAR	01	00072	A084	2010-09-29	A1 - UCSF