BEStack 02B Using Batch Effects Interface Corrections: EBNPlus
Tod Casasent
2018-04-25-1520

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/

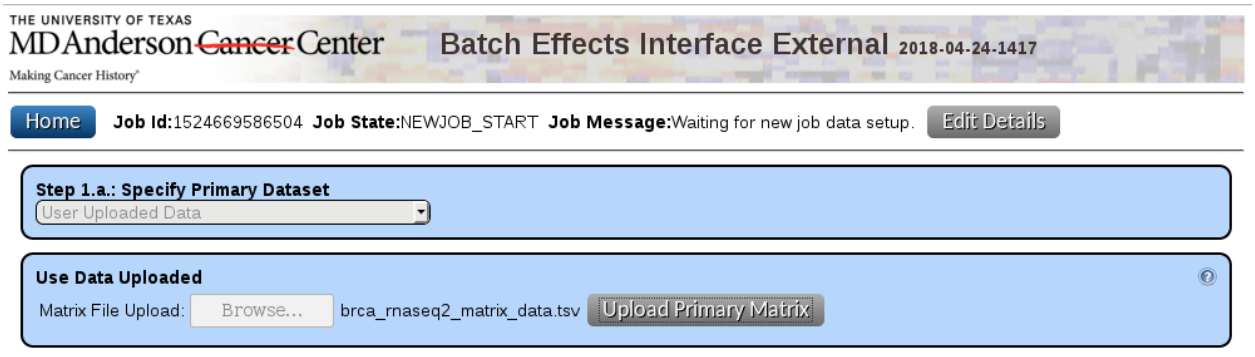
These instructions are aimed at people familiar with R and familiar with TCGA/GDC platforms and data types. They are intended to introduce the reader to producing the given assessment. These instructions will only rarely, if ever, touch on the appropriateness of the assessment algorithm or interpretation of output. See MBatch_01_InstallLinux at https://github.com/MD-Anderson-Bioinformatics/MBatch/tree/master/pdf for instructions on downloading test data.

EBNPlus corrections uses replicates between the two datasets for training and combines the two data sets based on replicates between sets.

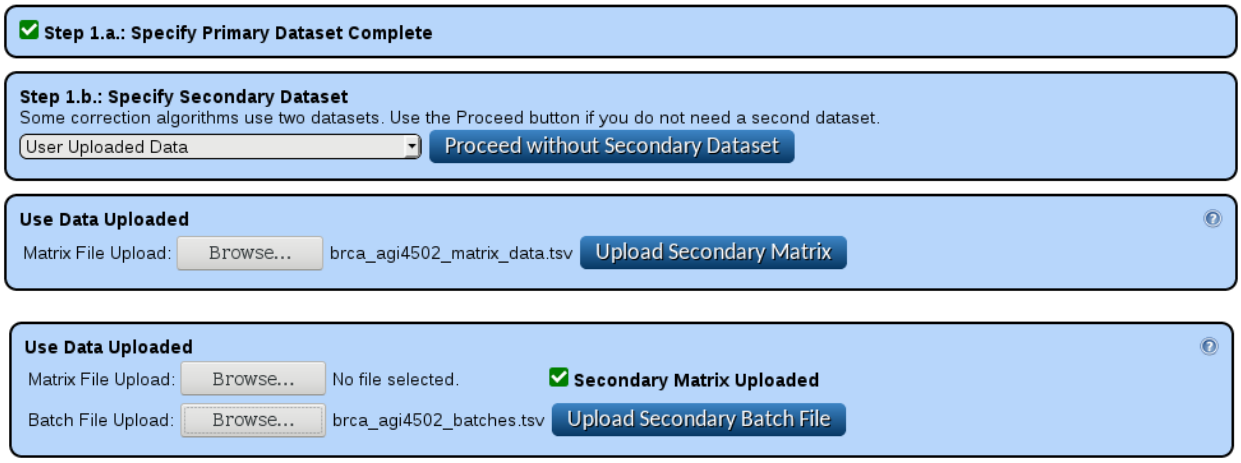
Starting a Job

See BEStack_02A_BEIUsingAssessments for more details about starting a job.

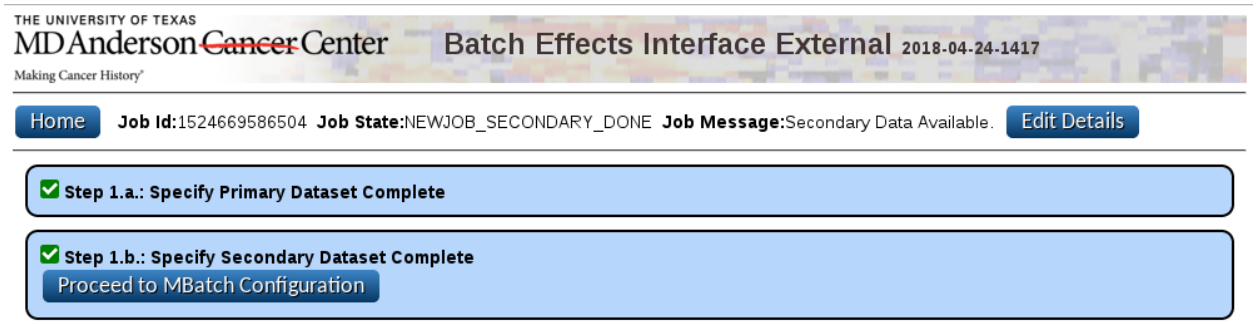
Use the "Start New Job" button and select "User Uploaded Data" for Step 1.a. From within the MATRIX_DATA.zip archive, upload brca_rnaseq2_matrix_data.tsv as the data matrix and brca_rnaseq2_batches.tsv as the batch matrix.



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.



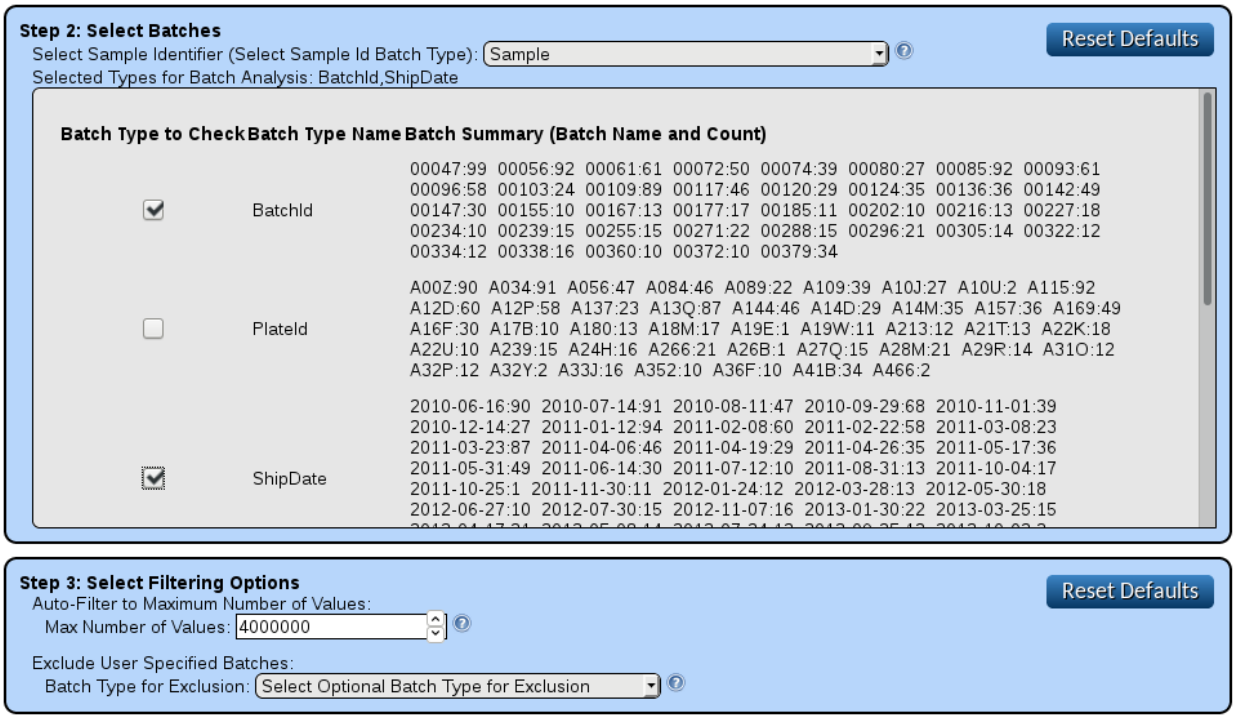
Then select Proceed to MBatch Configuration.



Configuring Assessments

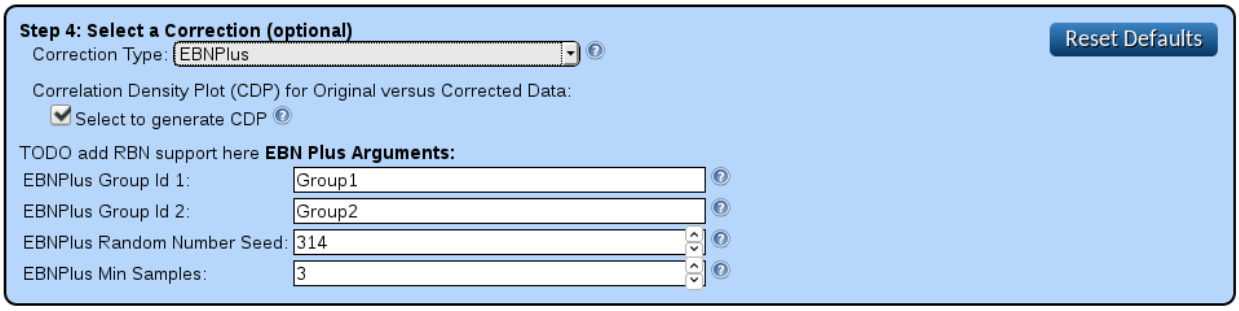
See BEStack_02A_BEIUsingAssessmentsExternal for more details about Configuring Assessments.

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.

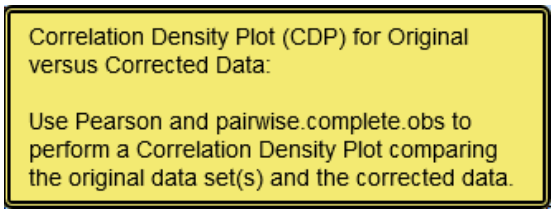


Step 4

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



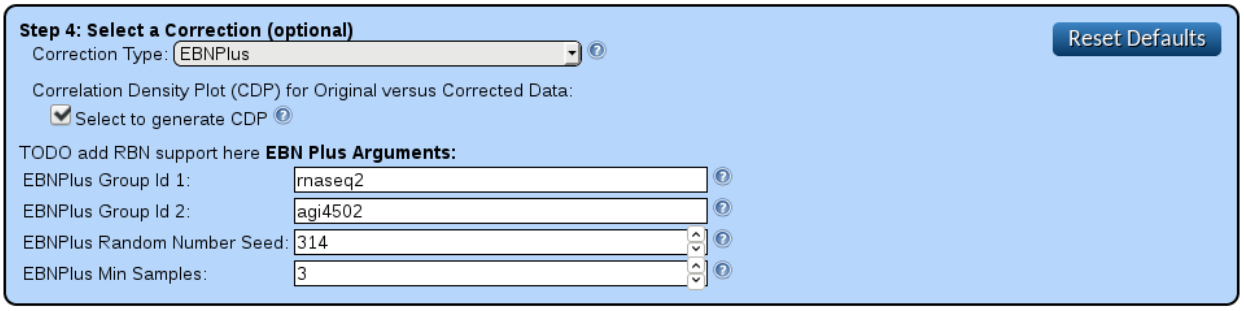
Generating a Correlation Density Plot is selected automatically. (In a future version, this option will appear for any assessment, rather than just for certain correction options.)



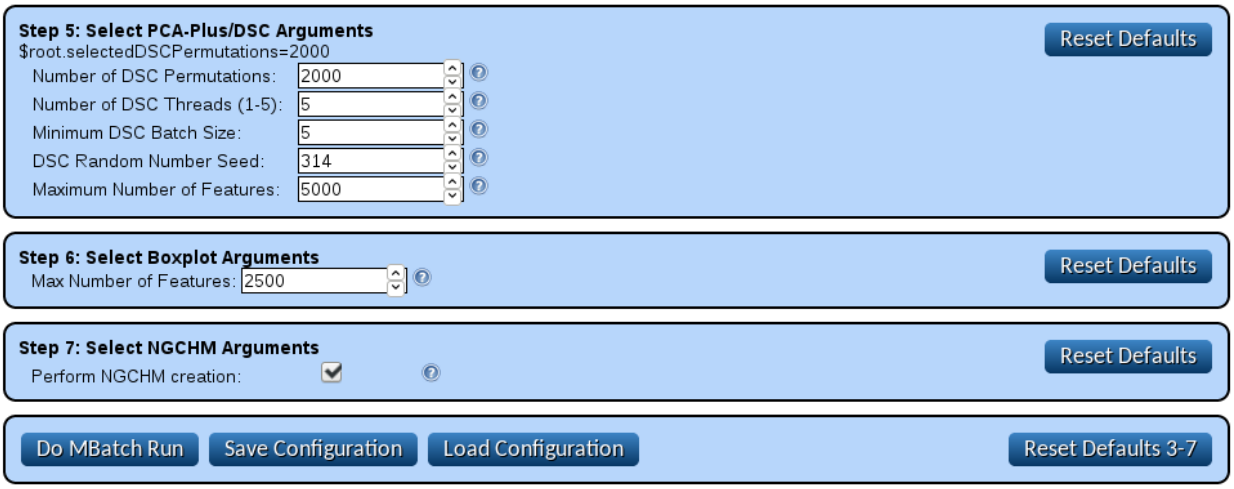
Tooltip Text for Correlation Density Plot

The Group Ids must be alphanumeric values without spaces but with underscores allowed. The random number seed is an integer used as a seed. The minimum number of samples is an integer and means any row (gene) with less than this number of samples is dropped.

Here, we have selected rnaseq2 as Group Id 1 and agi4502 as Group Id 2.

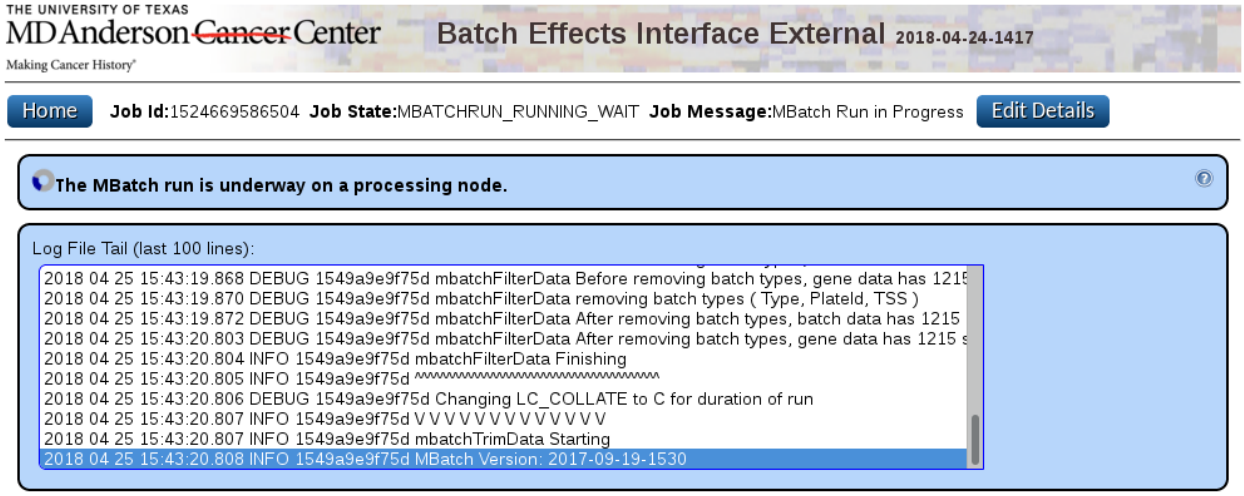


We accept the default values for the rest of the assessment settings, and hit the "Do MBatch Run" button.

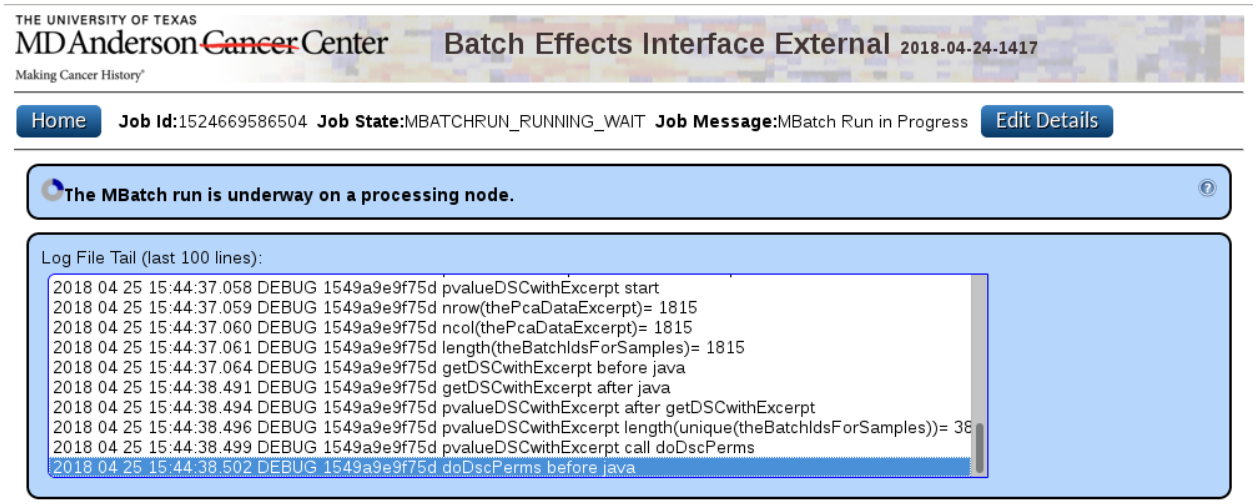


Do MBatch Run

See BEStack_02A_BEIUsingAssessments for more details about running and monitoring a run.

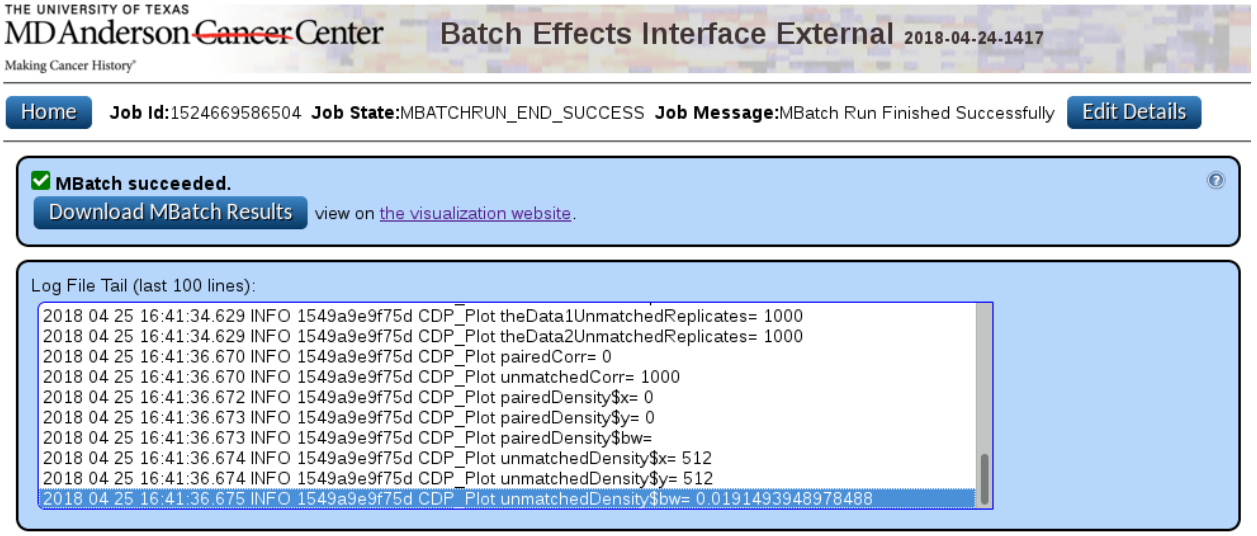


The DSC Permutations step will take some time—30 minutes or more with the full run taking several hours.

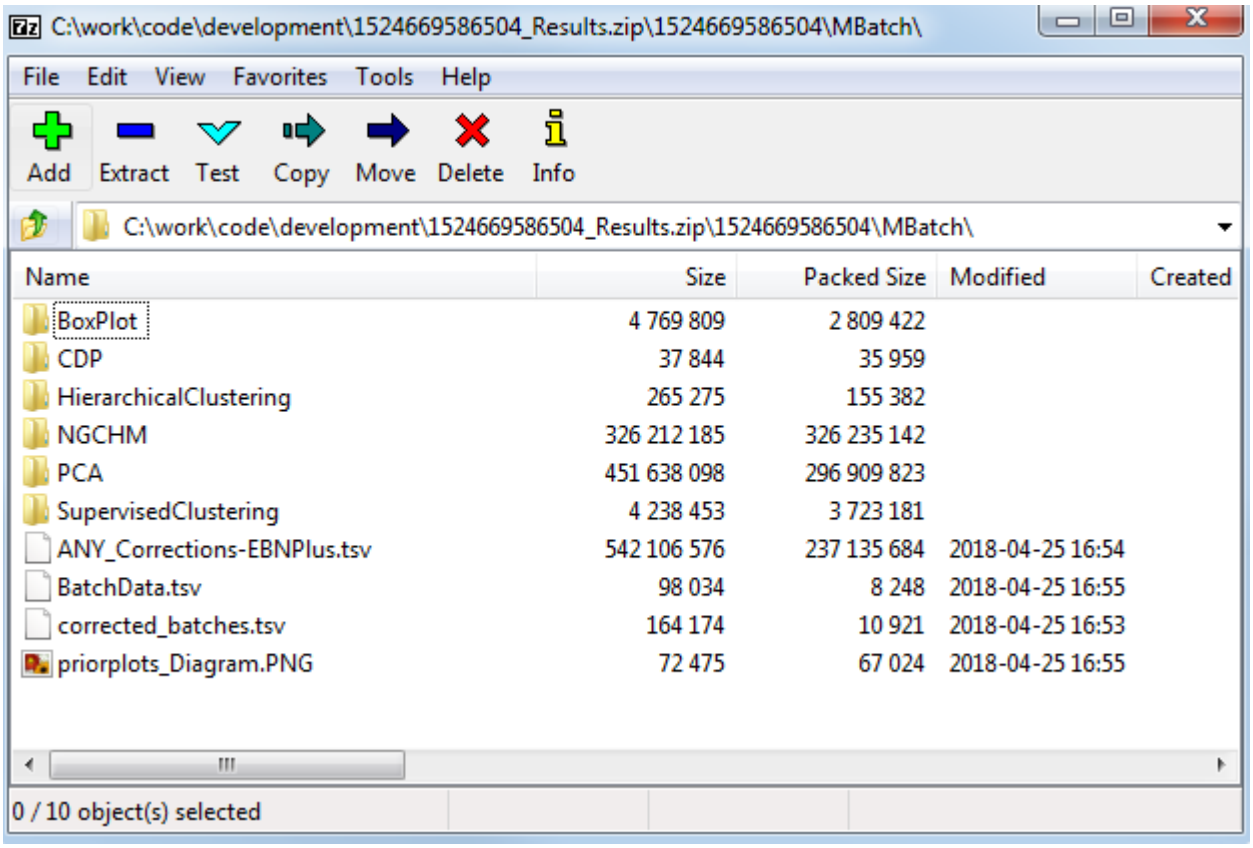


Finished Job

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-EBNPlus.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-A1-A0SB-01A-11R-A144-07.rnaseq2	TCGA-A1-A0SD-01A-11R-A115-07.agi4502	TCG
A1BG	2.4365111687448295	4.565452148506589	3.731
A2BP1	0.7439255863822336	2.4835734270886842	-1.32
A2M	8.545488308256441	6.629607302478645	8.137
A2ML1	0.787232177552283	1.6477352958044498	1.644
A4GALT	4.748263228165531	4.799944822217981	4.886
A4GNT	0.1884566850556865	0.3153918401926371	-0.06
AAAS	4.452459243976453	4.3829614276156645	3.995
AACS	6.252403781207319	4.691106228333074	4.818
AADAC	-2.3617543340468186	0.2346797540826212	-2.36

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.rnaseq2	1	372	A36F	1/29/2014	XX - Spectrum Healt
TCGA-XX-A89A-01A-11R-A36F-07.rnaseq2	1	372	A36F	1/29/2014	XX - Spectrum Healt
TCGA-Z7-A8R5-01A-42R-A41B-07.rnaseq2	1	379	A41B	5/28/2014	Z7 - John Wayne Ca
TCGA-Z7-A8R6-01A-11R-A41B-07.rnaseq2	1	379	A41B	5/28/2014	Z7 - John Wayne Ca
TCGA-A1-A0SD-01A-11R-A115-07.agi4502	1	85	A115	1/12/2011	A1 - UCSF
TCGA-A1-A0SE-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF
TCGA-A1-A0SH-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF
TCGA-A1-A0SJ-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF