1   Using Batch Effects Interface Corrections: EB

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 Empirical Bayes (EB) 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.
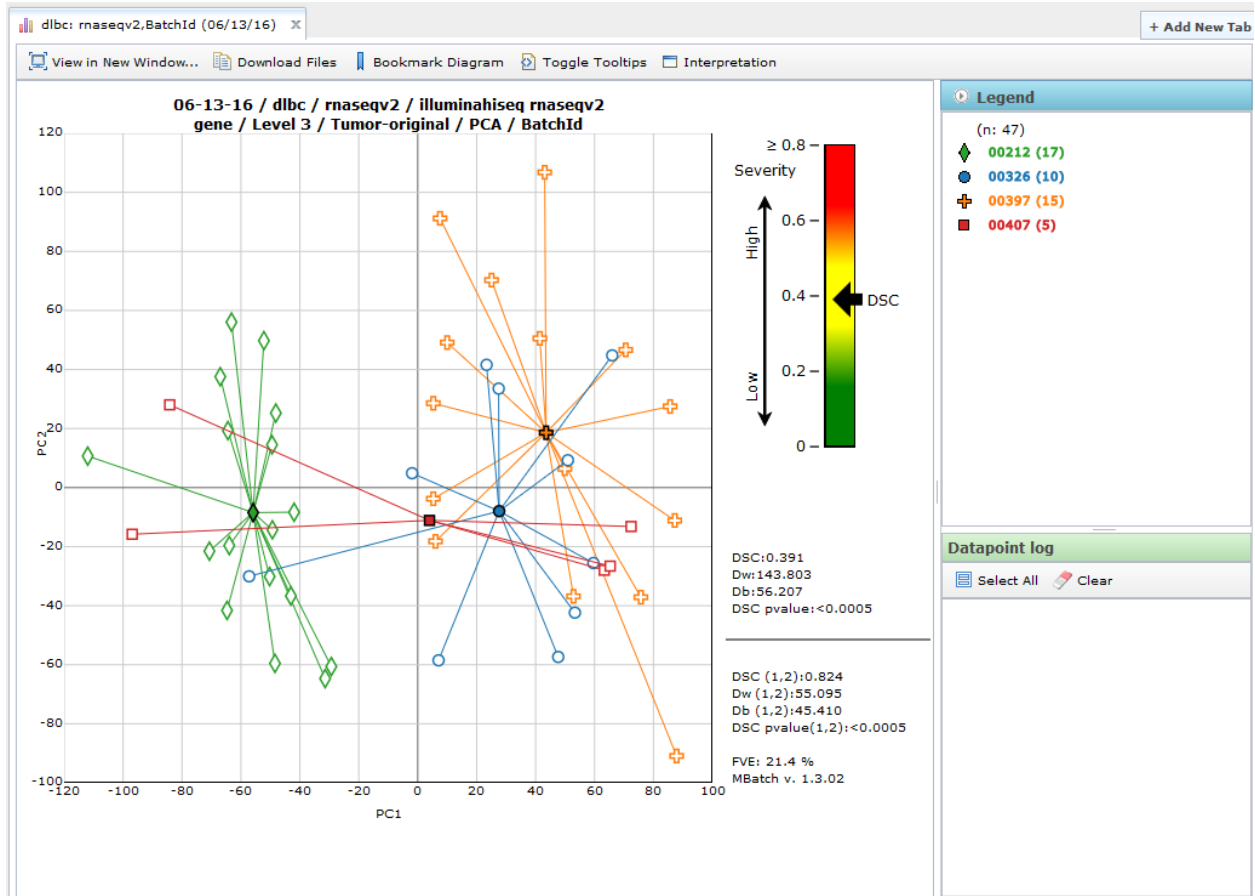
2   Starting a Job

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 Empirical Bayes

In Step 4, we begin by selecting Empirical Bayes-withNonParametricPriors as the optional Correction Type. (The setup for Empirical Bayes-withParametricPriors is the exact same.)

We accept the defaults for the rest of the data.

3.2 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.

Here you can see the nonparametric adjustments being performed, which can take 30+ minutes.

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

3.3 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.

When we examine the PCA plot, we see the correction has been performed. (We will not comment on whether or not this has removed biological effects.)
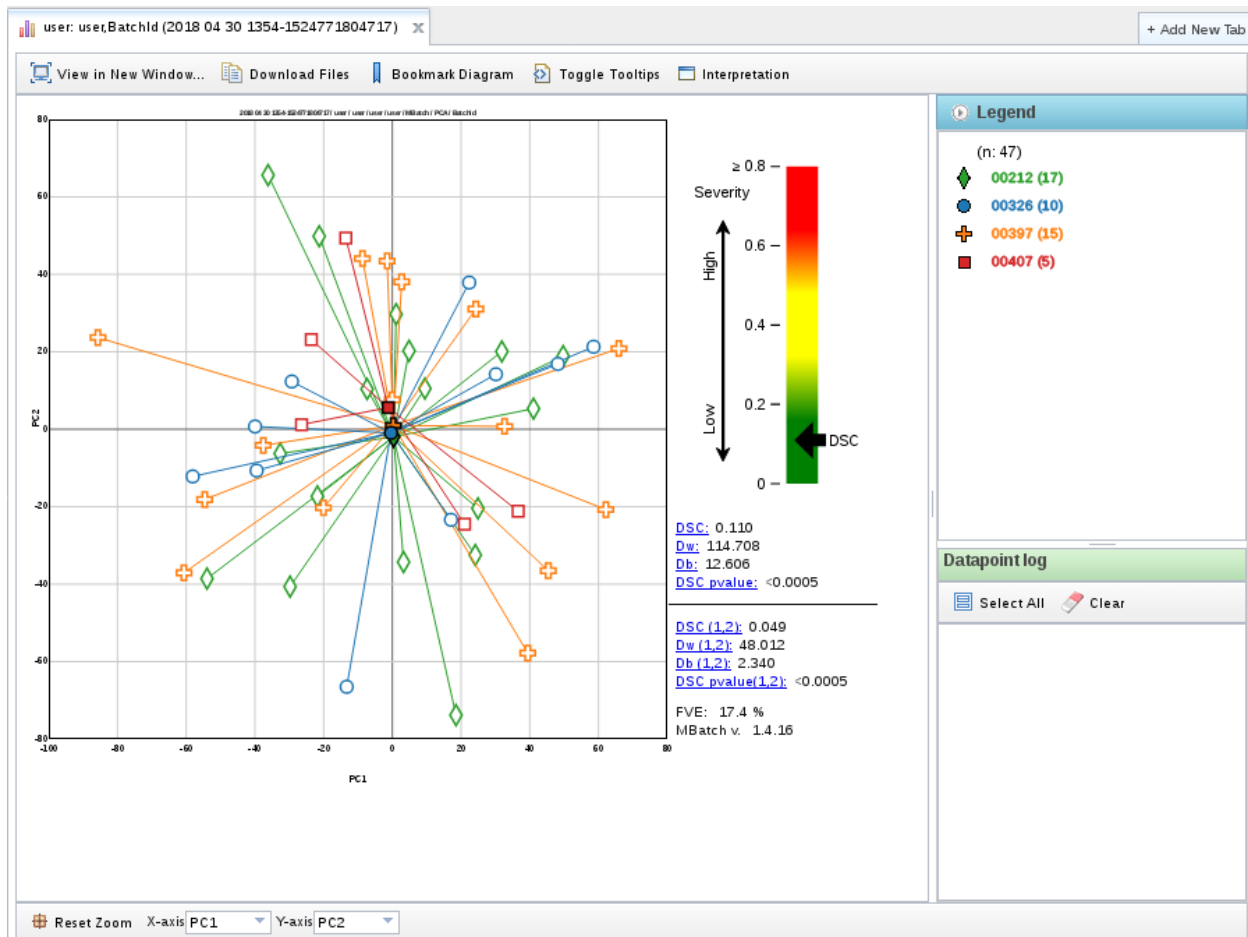
The ANY_Corrections-EBwithNonParametricPriors.tsv file contains the corrected data. Looking at an excerpt from that file below, you see the group ids have been added to the end of the sample ids (with a period to separate them).

	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	8.408407463	7.833472256	7.944558757	9.107147164
A1CF|29974	-1.428007951	-1.390367717	-1.390367717	-1.390367717
A2BP1|54715	-0.954393532	-1.312007102	-1.312007102	-1.312007102
A2LD1|87769	7.282673978	7.749438326	8.210468997	7.323894016
A2ML1|144568	-0.044995866	-1.41702626	-1.41702626	-0.204374047
A2M|2	12.2499439	11.93860208	12.31336814	11.35818912
A4GALT|53947	9.414619035	7.713706942	8.516100679	7.895134388
A4GNT|51146	0.691938264	-1.900678582	-1.900678582	-1.900678582

The BatchData.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	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