

## 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: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.

## 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. 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.

THE UNIVERSITY OF TEXAS  
MDAnderson Cancer Center  
Making Cancer History®

Batch Effects Interface External 2018-04-24-1417

[Home](#) Job Id:1524669586504 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:  brca\_rnaseq2\_matrix\_data.tsv

**Use Data Uploaded**  
Matrix File Upload:  No file selected. ☒ **Primary Matrix Uploaded**  
Batch File Upload:  brca\_rnaseq2\_batches.tsv

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.

☒ **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

**Use Data Uploaded**  
Matrix File Upload:  brca\_agi4502\_matrix\_data.tsv

**Use Data Uploaded**  
Matrix File Upload:  No file selected. ☒ **Secondary Matrix Uploaded**  
Batch File Upload:  brca\_agi4502\_batches.tsv

Then select Proceed to MBatch Configuration.

THE UNIVERSITY OF TEXAS

MDAndersonCancerCenter

Batch Effects Interface External 2018-04-24-1417

Home Job Id:1524669586504 Job State:NEWJOB\_SECONDARY\_DONE Job Message:Secondary Data Available. Edit Details

✓ Step 1.a.: Specify Primary Dataset Complete

✓ Step 1.b.: Specify Secondary Dataset Complete

Proceed to MBatch Configuration

### 3 Configuring Assessments

See BEMStack\_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 2: Select Batches

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

Reset Defaults

Selected Types for Batch Analysis: BatchId,ShipDate

	Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
<input checked="" type="checkbox"/>		BatchId	00047:99 00056:92 00061:61 00072:50 00074:39 00080:27 00085:92 00093:61 00096:58 00103:24 00109:89 00117:46 00120:29 00124:35 00136:36 00142:49 00147:30 00155:10 00167:13 00177:17 00185:11 00202:10 00216:13 00227:18 00234:10 00239:15 00255:15 00271:22 00288:15 00296:21 00305:14 00322:12 00334:12 00338:16 00360:10 00372:10 00379:34
<input type="checkbox"/>		PlateId	A00Z:90 A034:91 A056:47 A084:46 A089:22 A109:39 A10J:27 A10U:2 A115:92 A12D:60 A12P:58 A137:23 A13Q:87 A144:46 A14D:29 A14M:35 A157:36 A169:49 A16F:30 A17B:10 A180:13 A18M:17 A19E:1 A19W:11 A213:12 A21T:13 A22K:18 A22U:10 A239:15 A24H:16 A266:21 A26B:1 A27Q:15 A28M:21 A29R:14 A31O:12 A32P:12 A32Y:2 A33J:16 A352:10 A36F:10 A41B:34 A466:2
<input checked="" type="checkbox"/>		ShipDate	2010-06-16:90 2010-07-14:91 2010-08-11:47 2010-09-29:68 2010-11-01:39 2010-12-14:27 2011-01-12:94 2011-02-08:60 2011-02-22:58 2011-03-08:23 2011-03-23:87 2011-04-06:46 2011-04-19:29 2011-04-26:35 2011-05-17:36 2011-05-31:49 2011-06-14:30 2011-07-12:10 2011-08-31:13 2011-10-04:17 2011-10-25:1 2011-11-30:11 2012-01-24:12 2012-03-28:13 2012-05-30:18 2012-06-27:10 2012-07-30:15 2012-11-07:16 2013-01-30:22 2013-03-25:15

Step 3: Select Filtering Options

Auto-Filter to Maximum Number of Values:

Max Number of Values: 4000000 ?

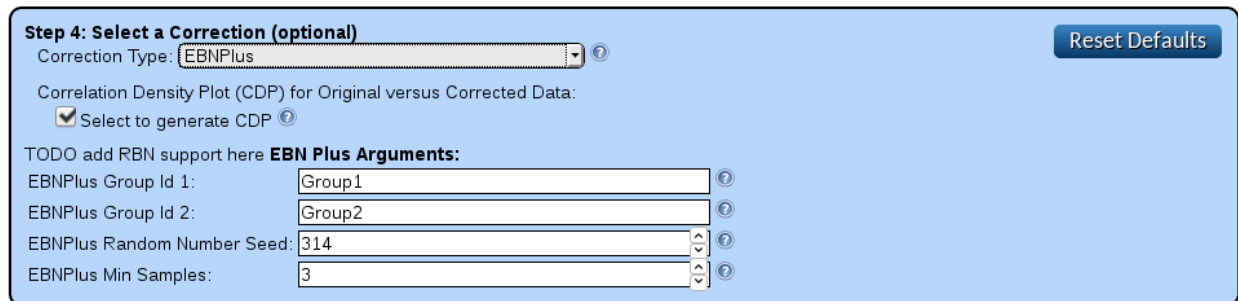
Reset Defaults

Exclude User Specified Batches:

Batch Type for Exclusion: Select Optional Batch Type for Exclusion ?

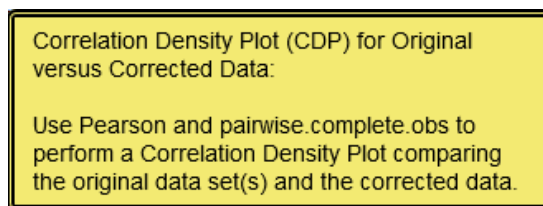
### 3.1 Step 4

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



The screenshot shows a web interface for Step 4: Select a Correction (optional). At the top right is a 'Reset Defaults' button. Below the title, 'Correction Type' is set to 'EBNPlus' with a help icon. A section for 'Correlation Density Plot (CDP) for Original versus Corrected Data:' contains a checked checkbox 'Select to generate CDP' with a help icon. Below this, a note says 'TODO add RBN support here'. The 'EBN Plus Arguments:' section includes four input fields: 'EBNPlus Group Id 1:' (Group1), 'EBNPlus Group Id 2:' (Group2), 'EBNPlus Random Number Seed:' (314), and 'EBNPlus Min Samples:' (3). Each field has a help icon.

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.)



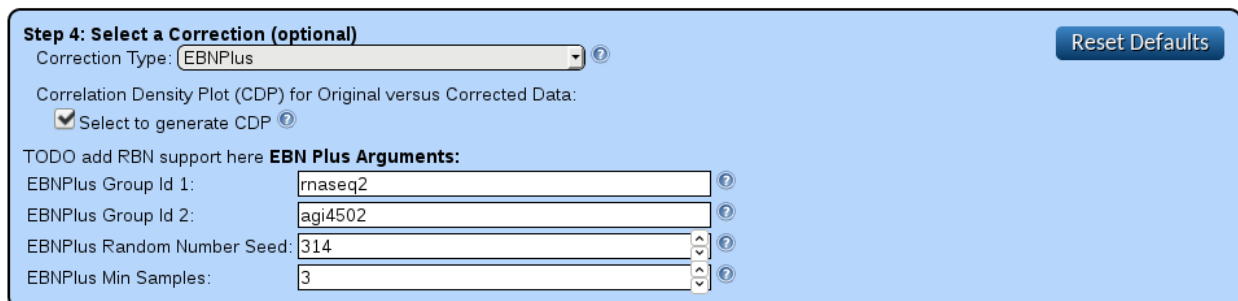
Correlation Density Plot (CDP) for Original versus Corrected Data:

Use Pearson and pairwise.complete.obs to perform a Correlation Density Plot comparing the original data set(s) and the corrected data.

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.



This screenshot shows the same Step 4 interface as before, but with updated values in the 'EBN Plus Arguments:' section. 'EBNPlus Group Id 1:' is now 'rnaseq2' and 'EBNPlus Group Id 2:' is now 'agi4502'. The other fields remain the same: 'EBNPlus Random Number Seed:' is 314 and 'EBNPlus Min Samples:' is 3. All help icons are still present.

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

**Step 5: Select PCA-Plus/DSC Arguments**Reset Defaults  
\$root.selectedDSCPermutations=2000  
Number of DSC Permutations:   
Number of DSC Threads (1-5):   
Minimum DSC Batch Size:   
DSC Random Number Seed:   
Maximum Number of Features:

**Step 6: Select Boxplot Arguments**Reset Defaults  
Max Number of Features:

**Step 7: Select NGCHM Arguments**Reset Defaults  
Perform NGCHM creation: ☒

Reset Defaults 3-7

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.

6

### 3.3 Finished Job

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

THE UNIVERSITY OF TEXAS

MDAnderson Cancer Center

Making Cancer History®

Batch Effects Interface External

2018-04-24-1417

[Home](#) **Job Id:**1524669586504 **Job State:**MBATCHRUN\_END\_SUCCESS **Job Message:**MBatch Run Finished Successfully [Edit Details](#)

✓ **MBatch succeeded.**

[Download MBatch Results](#) view on [the visualization website](#).

Log File Tail (last 100 lines):

```
2018 04 25 16:41:34.629 INFO 1549a9e9f75d CDP_Plot theData1UnmatchedReplicates= 1000
2018 04 25 16:41:34.629 INFO 1549a9e9f75d CDP_Plot theData2UnmatchedReplicates= 1000
2018 04 25 16:41:36.670 INFO 1549a9e9f75d CDP_Plot pairedCorr= 0
2018 04 25 16:41:36.670 INFO 1549a9e9f75d CDP_Plot unmatchedCorr= 1000
2018 04 25 16:41:36.672 INFO 1549a9e9f75d CDP_Plot pairedDensity$x= 0
2018 04 25 16:41:36.673 INFO 1549a9e9f75d CDP_Plot pairedDensity$y= 0
2018 04 25 16:41:36.673 INFO 1549a9e9f75d CDP_Plot pairedDensity$bw=
2018 04 25 16:41:36.674 INFO 1549a9e9f75d CDP_Plot unmatchedDensity$x= 512
2018 04 25 16:41:36.674 INFO 1549a9e9f75d CDP_Plot unmatchedDensity$y= 512
2018 04 25 16:41:36.675 INFO 1549a9e9f75d CDP_Plot unmatchedDensity$bw= 0.0191493948978488
```

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

C:\work\code\development\1524669586504\_Results.zip\1524669586504\MBatch\

File Edit View Favorites Tools Help

+

—

✓

→

→

✗

i

Add Extract Test Copy Move Delete Info

📁

 C:\work\code\development\1524669586504\_Results.zip\1524669586504\MBatch\

Name	Size	Packed Size	Modified	Created
BoxPlot	4 769 809	2 809 422		
CDP	37 844	35 959		
HierarchicalClustering	265 275	155 382		
NGCHM	326 212 185	326 235 142		
PCA	451 638 098	296 909 823		
SupervisedClustering	4 238 453	3 723 181		
ANY_Corrections-EBNPlus.tsv	542 106 576	237 135 684	2018-04-25 16:54	
BatchData.tsv	98 034	8 248	2018-04-25 16:55	
corrected_batches.tsv	164 174	10 921	2018-04-25 16:53	
priorplots_Diagram.PNG	72 475	67 024	2018-04-25 16:55	

0 / 10 object(s) selected



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	TCGA-A1-A0SD-01A-11R-A115-07.rnaseq2	TCGA-A1-A0SE-01A-11R-A084-07.agi4502
A1BG	2.4365111687448295	4.565452148506589	3.7312694095439234	3.708911067971168
A2BP1	0.7439255863822336	2.4835734270886842	-1.3208886816920007	-0.6581895449438544
A2M	8.545488308256441	6.629607302478645	8.137917647366036	7.434465827947289
A2ML1	0.787232177552283	1.6477352958044498	1.6447986673920711	1.4887545672394458
A4GALT	4.748263228165531	4.799944822217981	4.886308815099057	4.550275954972794
A4GNT	0.1884566850556865	0.3153918401926371	-0.06336625921993128	0.17228602035088397
AAAS	4.452459243976453	4.3829614276156645	3.995812362559631	4.096645218343809
AACS	6.252403781207319	4.691106228333074	4.818314878643724	5.068015002849296
AADAC	-2.3617543340468186	0.2346797540826212	-2.3617543340468186	0.1719909368074798

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	EBNPlus
TCGA-XX-A899-01A-11R-A36F-07.rnaseq2	1	372	A36F	1/29/2014	XX - Spectrum Health	rnaseq2
TCGA-XX-A89A-01A-11R-A36F-07.rnaseq2	1	372	A36F	1/29/2014	XX - Spectrum Health	rnaseq2
TCGA-Z7-A8R5-01A-42R-A41B-07.rnaseq2	1	379	A41B	5/28/2014	Z7 - John Wayne Cancer Center	rnaseq2
TCGA-Z7-A8R6-01A-11R-A41B-07.rnaseq2	1	379	A41B	5/28/2014	Z7 - John Wayne Cancer Center	rnaseq2
TCGA-A1-A0SD-01A-11R-A115-07.agi4502	1	85	A115	1/12/2011	A1 - UCSF	agi4502
TCGA-A1-A0SE-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF	agi4502
TCGA-A1-A0SH-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF	agi4502
TCGA-A1-A0SJ-01A-11R-A084-07.agi4502	1	72	A084	9/29/2010	A1 - UCSF	agi4502