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:

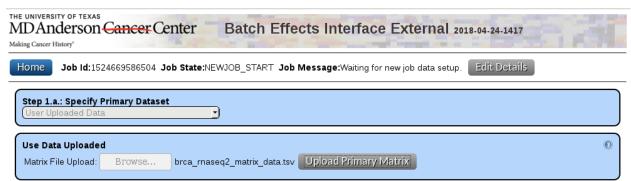> http://your-server.your-company.com:9999/BatchEffectsInterface/

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.
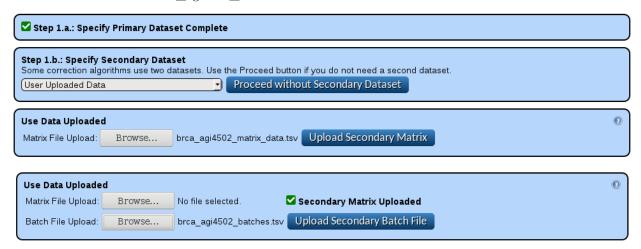
Starting a Job

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 2: Select Batches
Select Sample Identifier (Select Sample Id Batch Type): [Sample ▾] ⊘
Selected Types for Batch Analysis: BatchId,ShipDate

Reset Defaults

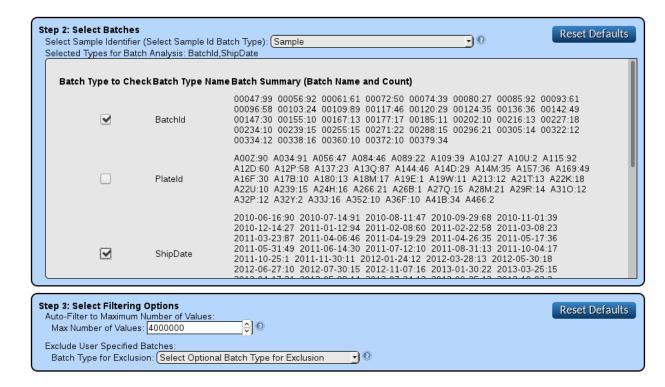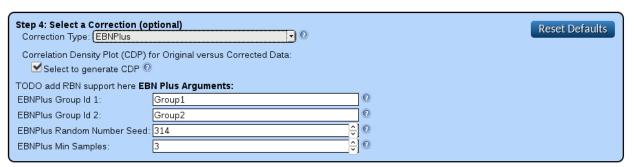| Batch Type to Check	Batch Type Name	Batch Summary (Batch Name and Count)
☑	BatchId	00047:99 00056:92 00061:61 00072:50 00074:39 00080:27 00085:92 00093:61 00096:58 00103:24 00109:89 00117:46 00120:29 00124:35 00136:36 00142:49 00147:30 00155:10 00167:13 00177:17 00185:11 00202:10 00216:13 00227:18 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
☐	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
☑	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 ▾] ⊘

Step 4

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

Step 4: Select a Correction (optional)
Correction Type: [EBNPlus ▾] ⊘

Reset Defaults

Correlation Density Plot (CDP) for Original versus Corrected Data:
    ☑ Select to generate CDP ⊘

TODO add RBN support here **EBN Plus Arguments:**

EBNPlus Group Id 1:	[Group1] ⊘
EBNPlus Group Id 2:	[Group2] ⊘
EBNPlus Random Number Seed:	[314 ⬍] ⊘
EBNPlus Min Samples:	[3 ⬍] ⊘

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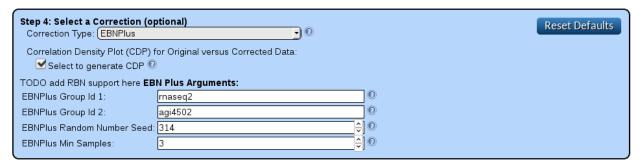
Correlation Density Plot (CDP) for Original
versus Corrected Data:

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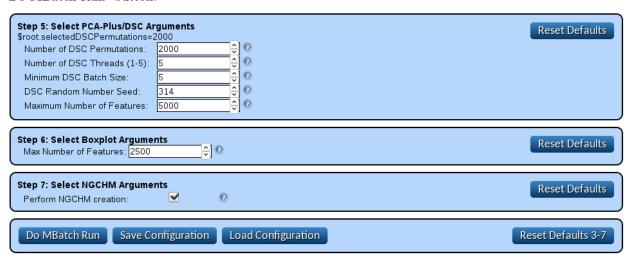the original data set(s) and the corrected data.

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.

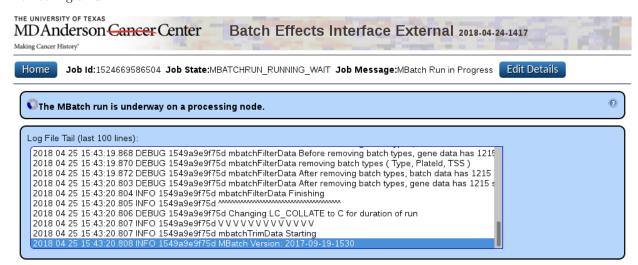Here, we have selected rnaseq2 as Group Id 1 and agi4502 as Group Id 2.

Step 4: Select a Correction (optional)

Correction Type: EBNPlus

Correlation Density Plot (CDP) for Original versus Corrected Data:
☑ Select to generate CDP

TODO add RBN support here **EBN Plus Arguments:**

EBNPlus Group Id 1: rnaseq2

EBNPlus Group Id 2: agi4502

EBNPlus Random Number Seed: 314

EBNPlus Min Samples: 3

Reset Defaults

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
$root.selectedDSCPermutations=2000

Number of DSC Permutations: 2000

Number of DSC Threads (1-5): 5

Minimum DSC Batch Size: 5

DSC Random Number Seed: 314

Maximum Number of Features: 5000

Reset Defaults

Step 6: Select Boxplot Arguments
Max Number of Features: 2500

Reset Defaults

Step 7: Select NGCHM Arguments
Perform NGCHM creation: ☑

Reset Defaults

Do MBatch Run   Save Configuration   Load Configuration
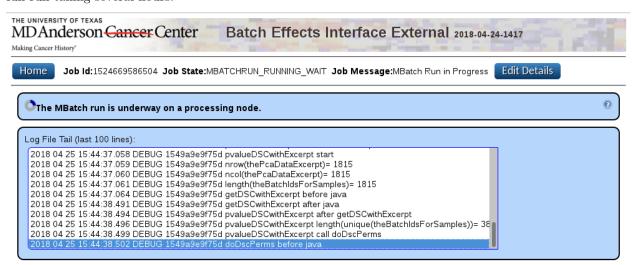
Reset Defaults 3-7

Do MBatch Run

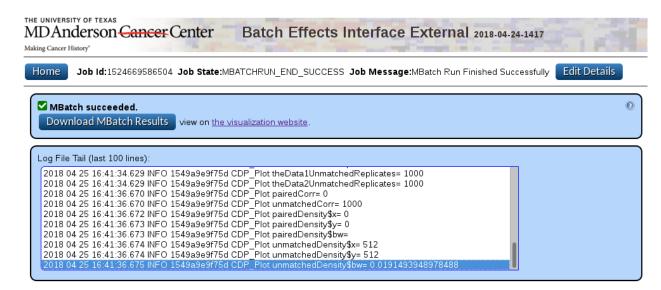See BEStack_02A_BEIUsingAssessments for more details about running and monitoring a run.

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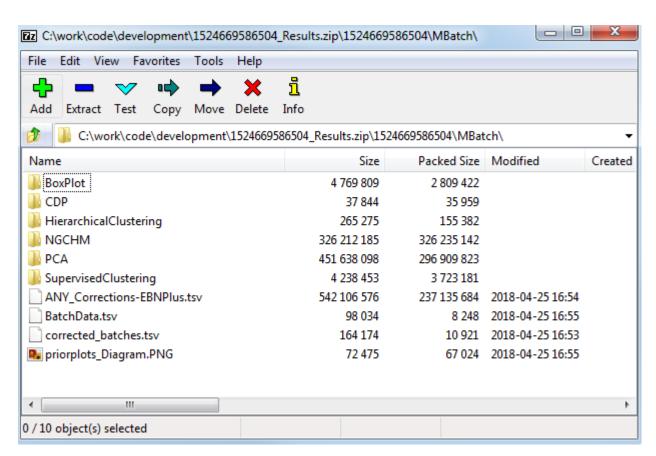

Finished Job

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 Heal
TCGA-XX-A89A-01A-11R-A36F-07.rnaseq2	1	372	A36F	1/29/2014	XX - Spectrum Heal
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