

BESack 02B Using Batch Effects Interface Corrections: MP
Tod Casasent
2018-04-30-1015

Using Batch Effects Interface Corrections: MP

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:

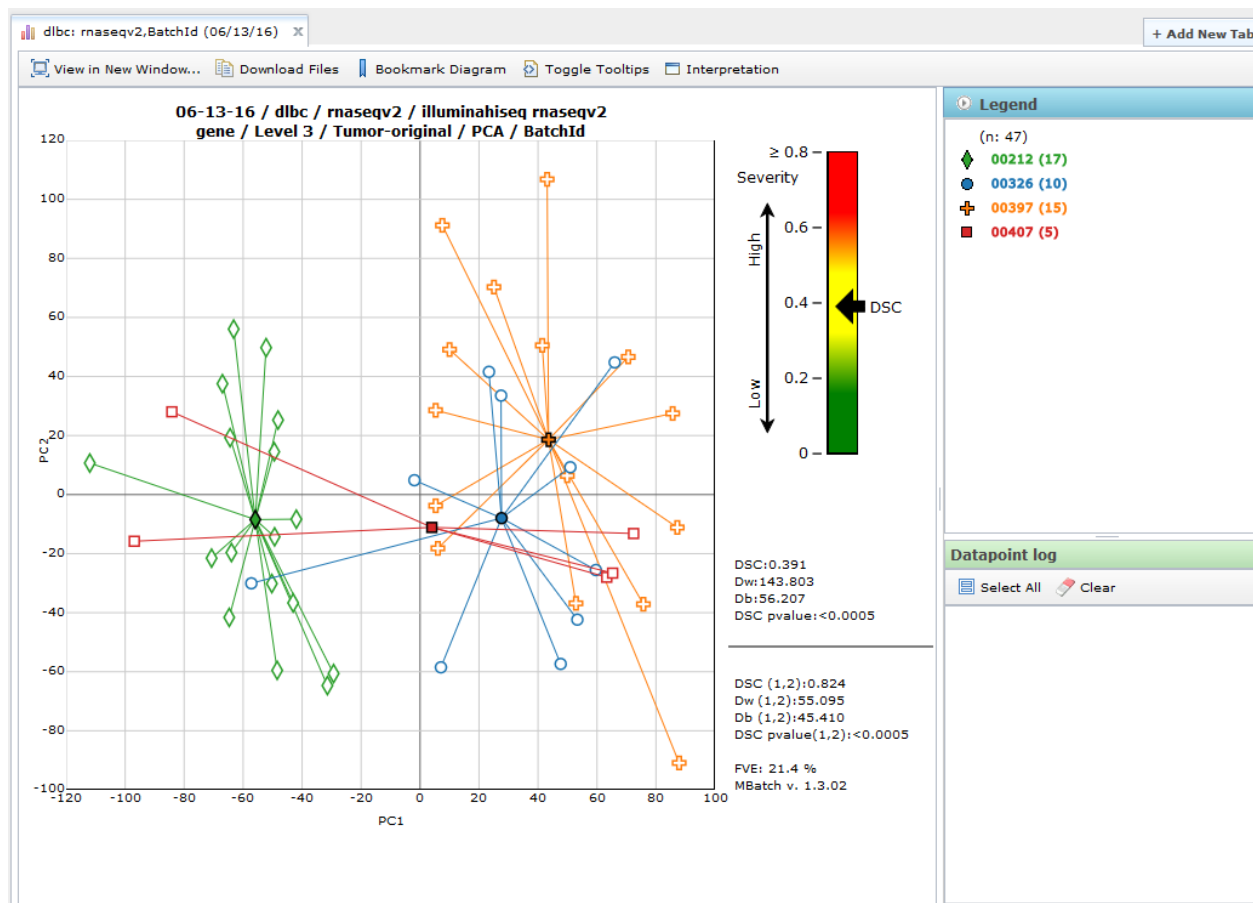
`http://your-server.your-company.com:9999/BatchEffectsInterface/`

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 Median Polish (MP) 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=%2F2016_06_13_0834-2016_08_16_1052%2Fdlbc%2Frna-seq2%2Filluminahiseq_rna-seq2_gene%2FLevel_3%2FTumor-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.



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.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History®

Batch Effects Interface External 2018-04-25-1433

[Home](#) Job Id:1524771804717 Job State:NEWJOB_PRIMARY_DONE Job Message:Primary Data Available. Waiting for Secondary Data. [Edit Details](#)

✔ 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 [Proceed without Secondary Dataset](#)

Use Data Uploaded

Matrix File Upload: [Browse...](#) No file selected. [Upload Secondary Matrix](#)

Then for Step 1.b. select Proceed without Secondary Dataset.

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.

THE UNIVERSITY OF TEXAS

MDAnderson Cancer Center

Making Cancer History®

Batch Effects Interface External 2018-04-25-1433

[Home](#)
Job Id:1524771804717
Job State:MBATCHCONFIG_START
Job Message:MBatch Configuration in Process
[Edit Details](#)

Step 2: Select Batches

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

Selected Types for Batch Analysis: BatchId,ShipDate

Batch Type to Check

Batch Type Name

Batch Summary (Batch Name and Count)

☒

BatchId

00212:17 00248:1 00326:10 00397:15 00407:5

☐

PlateId

2213:17 2404:1 A31O:10 A38C:15 A39D:5

☒

ShipDate

2012-03-14:17 2012-09-10:1 2013-07-24:10 2014-03-26:15 2014-04-30:5

☐

TSS

FA - Asterand:9 FF - SingHealth:11 FM - International Genomics Consortium:1 G8 - Roswell Park:7 GR - University of Nebraska Medical Center (UNMC):6 GS - Fundacio Clinic per a la Recerca Biomedica:10 RQ - St Josephs Hospital AZ:3 VB - Global BioClinical - Georgia:1

☐

Type

01:48

Reset Defaults

Step 3: Select Filtering Options

Auto-Filter to Maximum Number of Values:

Max Number of Values: 4000000

Exclude User Specified Batches:

Batch Type for Exclusion: Select Optional Batch Type for Exclusion

Reset Defaults

Step 4 Median Polish Overall

In Step 4, we begin by selecting MedianPolish-Overall as the optional Correction Type. (The batch type to correct and minimum batch size are not appropriate for MP-Overall and will be removed.)

Step 4: Select a Correction (optional)

Correction Type: MedianPolish-Overall

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

☒ Select to generate CDP

Batch Type to Correct: BatchId

Minimum Valid Batch Size for Corrections: 1

Reset Defaults

We accept the defaults for the rest of the data.

Step 4 Median Polish By Batch

In Step 4, if we want “By Batch” instead, we select MedianPolish-ByBatch as the optional Correction Type. Since we want the problem within the Batch Id type corrected, we select that.

Step 4: Select a Correction (optional)

Correction Type: MedianPolish-ByBatch

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

Batch Type to Correct: BatchId

Minimum Valid Batch Size for Corrections: 1

Reset Defaults

We accept the defaults for the rest of the data.

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 UNIVERSITY OF TEXAS
MDAnderson Cancer Center
Making Cancer History[®]

Batch Effects Interface External
2018-04-25-1433

Home Job Id:1524771804717 Job State:MBATCHRUN_START_WAIT Job Message:MBatch Run Queued Edit Details

The MBatch run is queued and waiting for assignment to a processing node.

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

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_RUNNING_WAIT Job Message:MBatch Run in Progress Edit Details

The MBatch run is underway on a processing node.

Log File Tail (last 100 lines):

```

2018 04 25 15:44:37.058 DEBUG 1549a9e9f75d pvalueDSCwithExcerpt start
2018 04 25 15:44:37.059 DEBUG 1549a9e9f75d nrow(thePcaDataExcerpt)= 1815
2018 04 25 15:44:37.060 DEBUG 1549a9e9f75d ncol(thePcaDataExcerpt)= 1815
2018 04 25 15:44:37.061 DEBUG 1549a9e9f75d length(theBatchIdsForSamples)= 1815
2018 04 25 15:44:37.064 DEBUG 1549a9e9f75d getDSCwithExcerpt before java
2018 04 25 15:44:38.491 DEBUG 1549a9e9f75d getDSCwithExcerpt after java
2018 04 25 15:44:38.494 DEBUG 1549a9e9f75d pvalueDSCwithExcerpt after getDSCwithExcerpt
2018 04 25 15:44:38.496 DEBUG 1549a9e9f75d pvalueDSCwithExcerpt length(unique(theBatchIdsForSamples))= 38
2018 04 25 15:44:38.499 DEBUG 1549a9e9f75d pvalueDSCwithExcerpt call doDscPerms
2018 04 25 15:44:38.502 DEBUG 1549a9e9f75d doDscPerms before java

```

Finished Job

After the job has finished, use the Download option to get the corrected data. Here, for reasons to be explained soon, we look at the MP-Overall Example.

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-25-1433
 Making Cancer History[®]

Home Job Id:1524774672723 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 26 20:49:09.204 INFO da7671a17f31 CDP_Plot theData1UnmatchedReplicates= 0
2018 04 26 20:49:09.204 INFO da7671a17f31 CDP_Plot theData2UnmatchedReplicates= 0
2018 04 26 20:49:09.286 INFO da7671a17f31 CDP_Plot pairedCorr= 48
2018 04 26 20:49:09.286 INFO da7671a17f31 CDP_Plot unmatchedCorr= 0
2018 04 26 20:49:09.290 INFO da7671a17f31 CDP_Plot pairedDensity$x= 512
2018 04 26 20:49:09.291 INFO da7671a17f31 CDP_Plot pairedDensity$y= 512
2018 04 26 20:49:09.292 INFO da7671a17f31 CDP_Plot pairedDensity$bw= 0.0239307275960845
2018 04 26 20:49:09.292 INFO da7671a17f31 CDP_Plot unmatchedDensity$x= 0
2018 04 26 20:49:09.293 INFO da7671a17f31 CDP_Plot unmatchedDensity$y= 0
2018 04 26 20:49:09.293 INFO da7671a17f31 CDP_Plot unmatchedDensity$bw=
  
```

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

THE UNIVERSITY OF TEXAS
MDAnderson Cancer Center Batch Effects Interface External 2018-04-25-1433
 Making Cancer History[®]

Home Job Id:1524774672723 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 26 20:49:09.204 INFO da7671a17f31 CDP_Plot theData1UnmatchedReplicates= 0
2018 04 26 20:49:09.204 INFO da7671a17f31 CDP_Plot theData2UnmatchedReplicates= 0
2018 04 26 20:49:09.286 INFO da7671a17f31 CDP_Plot pairedCorr= 48
2018 04 26 20:49:09.286 INFO da7671a17f31 CDP_Plot unmatchedCorr= 0
2018 04 26 20:49:09.290 INFO da7671a17f31 CDP_Plot pairedDensity$x= 512
2018 04 26 20:49:09.291 INFO da7671a17f31 CDP_Plot pairedDensity$y= 512
2018 04 26 20:49:09.292 INFO da7671a17f31 CDP_Plot pairedDensity$bw= 0.0239307275960845
2018 04 26 20:49:09.292 INFO da7671a17f31 CDP_Plot unmatchedDensity$x= 0
2018 04 26 20:49:09.293 INFO da7671a17f31 CDP_Plot unmatchedDensity$y= 0
2018 04 26 20:49:09.293 INFO da7671a17f31 CDP_Plot unmatchedDensity$bw=
  
```

Opening 1524774672723_Results.zip

You have chosen to open:
 1524774672723_Results.zip
 which is: Zip archive
 from: http://localhost:9999

What should Firefox do with this file?

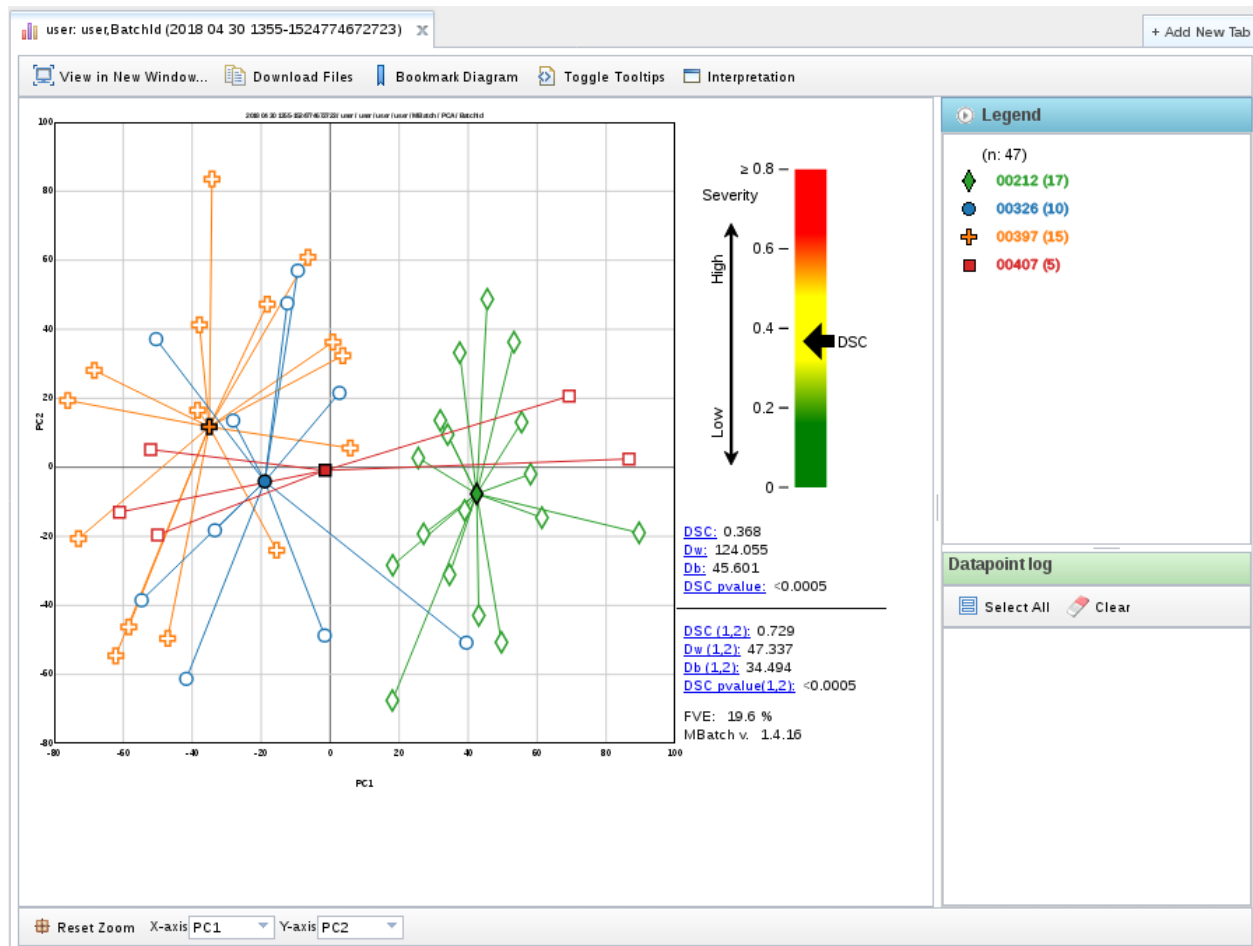
☐ Open with Xarchiver (default)

☒ Save File

☐ Do this automatically for files like this from now on.

Cancel OK

When we examine the PCA plot, we see the correction has been performed with minimal effectiveness. (We will not comment on whether or not this has removed biological effects.)



The ANY_Corrections-MPOverall.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-A4X
A1BG 1	7.576146043	6.531136709	6.694689663
A1CF 29974	6.98560411	6.945400497	6.98560411
A2BP1 54715	6.98560411	6.945400497	6.98560411
A2LD1 87769	6.371952521	7.069288154	7.663499093
A2ML1 144568	7.703779796	6.386542207	6.42674582
A2M 2	6.858983434	6.663633291	7.031545702
A4GALT 53947	8.252804603	5.776492658	6.638236223
A4GNT 51146	10.61079289	6.93980883	6.980012443

The BatchData.tsv contains the batch files. Looking at an excerpt from that file below, you see the batch data used.

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

Failed Job

Looking at the jobs, we notice the MB-ByBatch Example failed. Here, we will look at the log files.

THE UNIVERSITY OF TEXAS

MDAnderson Cancer Center

Making Cancer History®

Batch Effects Interface External 2018-04-25-1433

[Home](#)
[Start New Job](#)

Job History:

Job Id	Tag	Date Created	Message	Go to Job	Delete Job
1524775346620	ANOVA Example	4/26/2018, 3:42:26 PM	MBatch Run Finished Successfully	Select Job	Delete
1524775093543	MB-ByBatch Example	4/26/2018, 3:38:13 PM	MBatch Run Failed	Select Job	Delete
1524774672723	MP-Overall Example	4/26/2018, 3:31:12 PM	MBatch Run Finished Successfully	Select Job	Delete
1524771804717	EB Example	4/26/2018, 2:43:24 PM	MBatch Run Finished Successfully	Select Job	Delete
1524687939743	RBN Example	4/25/2018, 3:25:39 PM	MBatch Run Finished Successfully	Select Job	Delete
1524669586504	EBNPlus Example	4/25/2018, 10:19:46 AM	MBatch Run Finished Successfully	Select Job	Delete
1524662045619	User Data Example	4/25/2018, 8:14:05 AM	MBatch Configuration in Process	Select Job	Delete

The job status shows failed.

[Home](#) **Job Id:**1524775093543 **Job State:**MBATCHRUN_END_FAILURE **Job Message:**MBatch Run Failed [Edit Details](#)

MBatch Failed.

[Download MBatch Results](#)

Log File Tail (last 100 lines):

```
2018 04 26 20:51:43.053 DEBUG da7671a17f31 rows post filter 20531
2018 04 26 20:51:43.053 DEBUG da7671a17f31 mbatchFilterData Prefilter, gene data had 20531 while post filter 20531
2018 04 26 20:51:43.054 DEBUG da7671a17f31 mbatchFilterData Prefilter, batch data had 48 while post filter 48
2018 04 26 20:51:43.055 INFO da7671a17f31 mbatchFilterData Finishing
2018 04 26 20:51:43.055 INFO da7671a17f31 ~~~~~
2018 04 26 20:51:43.056 DEBUG da7671a17f31 hierclust_outputGraphics
2018 04 26 20:51:43.057 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.057 DEBUG da7671a17f31 hierclust_calc
2018 04 26 20:51:43.058 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.088 DEBUG da7671a17f31 calculating HC
```

We download the results as before.

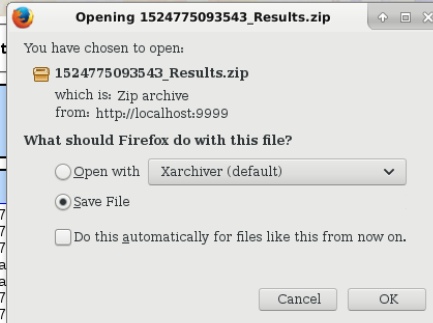
[Home](#) **Job Id:**1524775093543 **Job State:**MBATCHRUN_END_FAILURE **Job Message:**MBatch Run Failed [Edit Details](#)

MBatch Failed.

[Download MBatch Results](#)

Log File Tail (last 100 lines):

```
2018 04 26 20:51:43.053 DEBUG da7671a17f31 rows post filter 20531
2018 04 26 20:51:43.053 DEBUG da7671a17f31 mbatchFilterData Prefilter, gene data had 20531 while post filter 20531
2018 04 26 20:51:43.054 DEBUG da7671a17f31 mbatchFilterData Prefilter, batch data had 48 while post filter 48
2018 04 26 20:51:43.055 INFO da7671a17f31 mbatchFilterData Finishing
2018 04 26 20:51:43.055 INFO da7671a17f31 ~~~~~
2018 04 26 20:51:43.056 DEBUG da7671a17f31 hierclust_outputGraphics
2018 04 26 20:51:43.057 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.057 DEBUG da7671a17f31 hierclust_calc
2018 04 26 20:51:43.058 DEBUG da7671a17f31 Changing LC_COLLATE to C for duration of run
2018 04 26 20:51:43.088 DEBUG da7671a17f31 calculating HC
```



We open the archive and look at the mbatch.log file at the top level. At the end of that file we see these five lines:

```
2018 04 26 20:51:43.056 DEBUG da7671a17f31 hierclust_outputGraphics
```

```
2018 04 26 20:51:43.057 DEBUG da7671a17f31 Changing LC_COLLATE to C
for duration of run
```

```
2018 04 26 20:51:43.057 DEBUG da7671a17f31 hierclust_calc
```

```
2018 04 26 20:51:43.058 DEBUG da7671a17f31 Changing LC_COLLATE to C
for duration of run
```

```
2018 04 26 20:51:43.088 DEBUG da7671a17f31 calculating HC
```

So, we know that something happened within the hierarchical clustering calculations for this data set. This is generally a limitation of the data set itself (unsolvable solutions) or of the third party clustering algorithm. Generally, the

“limitation” type errors are caught within the code and only cause a problem for the one diagram type. In this case, since there is no error message, the problem is most likely that that relatively unlimited parameters used to process the data used more than the 24GB available on the machine on which this was run and cause R to stop without notice.