Using MBatch Assessments: HierarchicalClustering_Structures

Tod Casasent

2023-10-06

1 Introduction

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 for instructions on downloading test data.

2 Algorithm

HierarchicalClustering_Structures is a function used to perform batch effects assessments using hierarchical clustering.

3 Output

The primary output method for MBatch is to view results in the Batch Effects Website, described elsewhere. The PNG files are rough versions of the website output.

Graphical output is a hierarchical clustering diagram, with a dendrogram and annotations at the bottom for batches and batch types.

(See Batch Effects Viewer documentation for more details.)

FIRST IMAGES

4 Usage

HierarchicalClustering_Structures(theData, theTitle, theOutputDir, theDataVersion, theTestVersion, theBatchTypeAndValuePairsToRemove, theBatchTypeAndValuePairsToKeep)

5 Arguments

##theData An instance of BEA_DATA.

BEA_DATA objects can be created by calls of the form new("BEA_DATA", theData, theBatches, theCovariates). If you have no covariate data, use an empty data.frame created with data.frame()

mData: Object of class "matrix" A matrix where the colnames are sample ids and the rownames are gene equivalents. All names should be strings, not factors.

mBatches: Object of class "data.frame" A data.frame where the column "names" are batch types. The first batch "type" is "Sample". All names and values should be strings, not factors or numeric.

mCovariates: Object of class “data.frame” A data.frame where the column “names” are covariate types. The first covariate “type” is “Sample”. All names and values should be strings, not factors or numeric.

##theTitle A string title to use in PNG files.

##theOutputPath String giving directory in which to place output PNG files.

##theDataVersion Object of class character. Version of the data prefaced with DATA_, such as DATA_2022-09-09-1600. Use empty string to not include.

##theTestVersion Object of class character. Version of the test prefaced with TEST_, such as TEST_2022-10-10-1300. Use empty string to not include.

##theBatchTypeAndValuePairsToRemove A list of vectors containing the batch type (or * for all types) and the value to remove. list() indicates none while NULL will cause an error.

##theBatchTypeAndValuePairsToKeep A list of vectors containing the batch type (or * for all types) and a vector of the the value(s) to keep. list() indicates none while NULL will cause an error.

6 Example Call

The following code is adapted from the tests/HierarchicalClustering_Structures file. Data used is from the testing data as per the MBatch_01_InstallLinux document. In the future, we plan to make the output from MBatch more user friendly, but currently, this produces the following output at the command line.

This output can generally be skipped as very long and generally obscure. After the output is an explanation of files and directories created.

```
{
  require(MBatch)

  logTransform <- function(mymatrix)
  {
    print("****Log transform data****")
    # convert to vector
    myVector <- as.vector(mymatrix)
    # so we can remove all zero values
    myVector <- myVector[myVector>0]
    # so when we remove NAs and look for .1 quantile,
    # we get a non-zero answer
    qt <- quantile(myVector, .1, na.rm=TRUE)
    # that gives us non-zero (and non-infinite) values
    # within the transformed matrix
    mymatrix <- log2(mymatrix+qt)
    mymatrix
  }

  inputDir <- getTestInputDir()
  outputDir <- getTestOutputDir()
  compareDir <- getTestCompareDir()

  # set the paths
  # data used here has already been log transformed previously
  theGeneFile=cleanFilePath(inputDir, "AN002418_matrix_data.tsv")
  theBatchFile=cleanFilePath(inputDir, "AN002418_batches.tsv")
  theOutputDir=cleanFilePath(outputDir, "MWB_HierarchicalClustering_Structures")
  print(theGeneFile)
```

```
  print(theBatchFile)
  print(theOutputDir)
  theRandomSeed=314

  # make sure the output dir exists and is empty
  print(theOutputDir)
  unlink(theOutputDir, recursive=TRUE)
  dir.create(theOutputDir, showWarnings=FALSE, recursive=TRUE)

  # load the data and reduce the amount of data to reduce run time
  myData <- mbatchLoadFiles(theGeneFile, theBatchFile)
  myData@mData <- logTransform(myData@mData)

  # here, we take most defaults
  HierarchicalClustering_Structures(myData, "Metabolomics WorkBench Test Data", theOutputDir,
                                    theDataVersion="DATA_2022-09-09-1600",
                                    theTestVersion="TEST_2022-10-10-1300",
                                    theBatchTypeAndValuePairsToRemove=list(),
                                    theBatchTypeAndValuePairsToKeep=list())
}
```

```
## [1] "/builds/BatchEffects_clean/BatchEffectsPackage/data/testing_static/MATRIX_DATA/AN002418_matrix_
## [1] "/builds/BatchEffects_clean/BatchEffectsPackage/data/testing_static/MATRIX_DATA/AN002418_batches
## [1] "/BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/MWB_HierarchicalClustering_Structures"
## [1] "/BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/MWB_HierarchicalClustering_Structures"
## 2023 10 06 12:32:10.649 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.649 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:10.650 INFO qcprludev10 Starting mbatchLoadFiles
## 2023 10 06 12:32:10.650 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:10.651 INFO qcprludev10 read batch file= /builds/BatchEffects_clean/BatchEffectsPacl
## 2023 10 06 12:32:10.652 INFO qcprludev10 read gene file= /builds/BatchEffects_clean/BatchEffectsPacka
## 2023 10 06 12:32:10.660 INFO qcprludev10 filter samples in batches using gene samples
## 2023 10 06 12:32:10.660 INFO qcprludev10 sort batches by gene file samples
## 2023 10 06 12:32:10.661 INFO qcprludev10 Finishing mbatchLoadFiles
## 2023 10 06 12:32:10.662 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## [1] "****Log transform data****"
## 2023 10 06 12:32:10.664 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.664 INFO qcprludev10 \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/ \/
## 2023 10 06 12:32:10.665 INFO qcprludev10 mbatchFilterData Starting
## 2023 10 06 12:32:10.665 INFO qcprludev10 MBatch Version: BEA_VERSION_TIMESTAMP
## 2023 10 06 12:32:10.666 DEBUG qcprludev10 rows pre filter  1053
## 2023 10 06 12:32:10.811 DEBUG qcprludev10 rows post filter  1053
## 2023 10 06 12:32:10.811 DEBUG qcprludev10 mbatchFilterData Prefilter, gene data had  1053  while pos
## 2023 10 06 12:32:10.812 DEBUG qcprludev10 mbatchFilterData Prefilter, batch data had  15  while post
## 2023 10 06 12:32:10.813 INFO qcprludev10 mbatchFilterData Finishing
## 2023 10 06 12:32:10.813 INFO qcprludev10 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
## 2023 10 06 12:32:10.814 DEBUG qcprludev10 hierclust_outputGraphics
## 2023 10 06 12:32:10.814 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.814 DEBUG qcprludev10 hierclust_calc
## 2023 10 06 12:32:10.815 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.815 DEBUG qcprludev10 calculating HC
## 2023 10 06 12:32:10.816 DEBUG qcprludev10 hierclust_calc
## 2023 10 06 12:32:10.817 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.818 DEBUG qcprludev10 calculating HC
```

```
## 2023 10 06 12:32:10.875 DEBUG qcprludev10 checkCreateDir:  /BEA/BatchEffectsPackage_data/testing_dyna
## 2023 10 06 12:32:10.876 DEBUG qcprludev10 createBatchEffectsOutput_hierclust RData output
## 2023 10 06 12:32:10.876 INFO qcprludev10 writeHCDataTSVs start
## 2023 10 06 12:32:10.877 INFO qcprludev10 writeHCDataTSVs udend
## 2023 10 06 12:32:10.877 INFO qcprludev10 writeHCDataTSVs HCData= /BEA/BatchEffectsPackage_data/testin
## 2023 10 06 12:32:10.878 INFO qcprludev10 writeHCDataTSVs HCOrder= /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:32:10.879 INFO qcprludev10 writeHCDataTSVs rdataFile= /BEA/BatchEffectsPackage_data/tes
## 2023 10 06 12:32:10.880 INFO qcprludev10 writeHCDataTSVs done
## 2023 10 06 12:32:10.881 INFO qcprludev10 writeHCDataTSVs start
## 2023 10 06 12:32:10.881 INFO qcprludev10 writeHCDataTSVs udend
## 2023 10 06 12:32:10.881 INFO qcprludev10 writeHCDataTSVs HCData= /BEA/BatchEffectsPackage_data/testin
## 2023 10 06 12:32:10.886 INFO qcprludev10 writeHCDataTSVs HCOrder= /BEA/BatchEffectsPackage_data/testi
## 2023 10 06 12:32:10.888 INFO qcprludev10 writeHCDataTSVs rdataFile= /BEA/BatchEffectsPackage_data/tes
## 2023 10 06 12:32:10.891 INFO qcprludev10 writeHCDataTSVs done
## 2023 10 06 12:32:10.892 DEBUG qcprludev10 hierclust_draw
## 2023 10 06 12:32:10.892 DEBUG qcprludev10 Changing LC_COLLATE to C for duration of run
## 2023 10 06 12:32:10.896 DEBUG qcprludev10 writeTitleFile - pre title  Metabolomics WorkBench Test Dat
## 2023 10 06 12:32:10.897 DEBUG qcprludev10 writeTitleFile - theTitle  Metabolomics WorkBench Test Data
## 2023 10 06 12:32:10.897 DEBUG qcprludev10 writeTitleFile - titleFile  /BEA/BatchEffectsPackage_data/t
## 2023 10 06 12:32:10.898 DEBUG qcprludev10 writeClusteringImage /BEA/BatchEffectsPackage_data/testing_

## 2023 10 06 12:32:10.948 DEBUG qcprludev10 writeIndividualLegend /BEA/BatchEffectsPackage_data/testing
## 2023 10 06 12:32:10.949 DEBUG qcprludev10 mbatchStandardLegend - theTitle  Treatment.Protocol
## 2023 10 06 12:32:10.949 DEBUG qcprludev10 mbatchStandardLegend - theVersion  MBatch 2.0.3
## 2023 10 06 12:32:10.950 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath  /BEA/BatchEffectsPa
## 2023 10 06 12:32:10.950 DEBUG qcprludev10 mbatchStandardLegend - theLegendNames  allogenic (6), naive
## 2023 10 06 12:32:10.951 DEBUG qcprludev10 mbatchStandardLegend - theLegendNames length  3
## 2023 10 06 12:32:10.951 DEBUG qcprludev10 mbatchStandardLegend - theLegendColors  #0000FF, #FF0000, #
## 2023 10 06 12:32:10.951 DEBUG qcprludev10 mbatchStandardLegend - theLegendColors length  3
## 2023 10 06 12:32:10.952 DEBUG qcprludev10 mbatchStandardLegend - theLegendSymbols
## 2023 10 06 12:32:10.952 DEBUG qcprludev10 mbatchStandardLegend - theLegendSymbols length  0
## 2023 10 06 12:32:10.953 DEBUG qcprludev10 mbatchStandardLegend - myColors  #0000ff,#ff0000,#00ff00
## 2023 10 06 12:32:10.953 DEBUG qcprludev10 mbatchStandardLegend - theTitle UTF-8 =  Treatment.Protocol
## 2023 10 06 12:32:10.953 DEBUG qcprludev10 mbatchStandardLegend - theVersion UTF-8 =  MBatch 2.0.3
## 2023 10 06 12:32:10.954 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath UTF-8 =  /BEA/Batchl
## 2023 10 06 12:32:10.954 DEBUG qcprludev10 mbatchStandardLegend before Python
## 2023 10 06 12:32:10.954 DEBUG qcprludev10 mbatchStandardLegend - getGlobalMBatchEnv() =  /BEA/gendev
## 2023 10 06 12:32:10.955 DEBUG qcprludev10 mbatchStandardLegend - import(mbatch.legend.legend)
## 2023 10 06 12:32:10.955 DEBUG qcprludev10 mbatchStandardLegend - after import
## 2023 10 06 12:32:10.956 DEBUG qcprludev10 mbatchStandardLegend - after color list
## 2023 10 06 12:32:10.956 DEBUG qcprludev10 mbatchStandardLegend - after symbol list
## 2023 10 06 12:32:10.957 DEBUG qcprludev10 mbatchStandardLegend - colorList =  #0000ff, mbatchStandard
## 2023 10 06 12:32:10.957 DEBUG qcprludev10 mbatchStandardLegend - symbolList =
## 2023 10 06 12:32:10.958 DEBUG qcprludev10 mbatchStandardLegend - legendNameList =  allogenic (6), mba
## 2023 10 06 12:32:10.958 DEBUG qcprludev10 mbatchStandardLegend - myTitle =  Treatment.Protocol MBatch
## 2023 10 06 12:32:10.958 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath =  /BEA/BatchEffects
## 2023 10 06 12:32:11.025 DEBUG qcprludev10 mbatchStandardLegend after Python
## 2023 10 06 12:32:11.026 DEBUG qcprludev10 writeIndividualLegend /BEA/BatchEffectsPackage_data/testing
## 2023 10 06 12:32:11.027 DEBUG qcprludev10 mbatchStandardLegend - theTitle  time_point
## 2023 10 06 12:32:11.028 DEBUG qcprludev10 mbatchStandardLegend - theVersion  MBatch 2.0.3
## 2023 10 06 12:32:11.028 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath  /BEA/BatchEffectsPa
## 2023 10 06 12:32:11.029 DEBUG qcprludev10 mbatchStandardLegend - theLegendNames  0 (3), 42 (6), 7 (6)
## 2023 10 06 12:32:11.029 DEBUG qcprludev10 mbatchStandardLegend - theLegendNames length  3
## 2023 10 06 12:32:11.029 DEBUG qcprludev10 mbatchStandardLegend - theLegendColors  #FF0000, #00FF00, #
```

```
## 2023 10 06 12:32:11.030 DEBUG qcprludev10 mbatchStandardLegend - theLegendColors length  3
## 2023 10 06 12:32:11.030 DEBUG qcprludev10 mbatchStandardLegend - theLegendSymbols
## 2023 10 06 12:32:11.031 DEBUG qcprludev10 mbatchStandardLegend - theLegendSymbols length  0
## 2023 10 06 12:32:11.031 DEBUG qcprludev10 mbatchStandardLegend - myColors  #ff0000,#00ff00,#0000ff
## 2023 10 06 12:32:11.032 DEBUG qcprludev10 mbatchStandardLegend - theTitle UTF-8 =  time_point
## 2023 10 06 12:32:11.032 DEBUG qcprludev10 mbatchStandardLegend - theVersion UTF-8 =  MBatch 2.0.3
## 2023 10 06 12:32:11.032 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath UTF-8 =  /BEA/Batch
## 2023 10 06 12:32:11.033 DEBUG qcprludev10 mbatchStandardLegend before Python
## 2023 10 06 12:32:11.033 DEBUG qcprludev10 mbatchStandardLegend - getGlobalMBatchEnv() =  /BEA/gendev
## 2023 10 06 12:32:11.034 DEBUG qcprludev10 mbatchStandardLegend - import(mbatch.legend.legend)
## 2023 10 06 12:32:11.034 DEBUG qcprludev10 mbatchStandardLegend - after import
## 2023 10 06 12:32:11.035 DEBUG qcprludev10 mbatchStandardLegend - after color list
## 2023 10 06 12:32:11.035 DEBUG qcprludev10 mbatchStandardLegend - after symbol list
## 2023 10 06 12:32:11.035 DEBUG qcprludev10 mbatchStandardLegend - colorList =  #ff0000, mbatchStandar
## 2023 10 06 12:32:11.036 DEBUG qcprludev10 mbatchStandardLegend - symbolList =
## 2023 10 06 12:32:11.036 DEBUG qcprludev10 mbatchStandardLegend - legendNameList =  0 (3), mbatchStan
## 2023 10 06 12:32:11.037 DEBUG qcprludev10 mbatchStandardLegend - myTitle =  time_point MBatch 2.0.3
## 2023 10 06 12:32:11.037 DEBUG qcprludev10 mbatchStandardLegend - theFilenamePath =  /BEA/BatchEffect
## 2023 10 06 12:32:11.095 DEBUG qcprludev10 mbatchStandardLegend after Python
## 2023 10 06 12:32:11.096 DEBUG qcprludev10 writeCombinedLegendHC /BEA/BatchEffectsPackage_data/testing
## 2023 10 06 12:32:11.096 DEBUG qcprludev10 mbatchStandardCombineLegends - theTitle  Metabolomics Work
## 2023 10 06 12:32:11.097 DEBUG qcprludev10 mbatchStandardCombineLegends - theFilenamePath  /BEA/Batch
## 2023 10 06 12:32:11.097 DEBUG qcprludev10 mbatchStandardCombineLegends - theListOfFiles  /BEA/BatchE
## 2023 10 06 12:32:11.097 DEBUG qcprludev10 mbatchStandardCombineLegends - theTitle UTF-8 =  Metabolom
## 2023 10 06 12:32:11.098 DEBUG qcprludev10 mbatchStandardCombineLegends - theFilenamePath UTF-8 =  /B
## 2023 10 06 12:32:11.098 DEBUG qcprludev10 mbatchStandardCombineLegends before Python
## 2023 10 06 12:32:11.099 DEBUG qcprludev10 mbatchStandardCombineLegends - import(mbatch.legend.legend
## 2023 10 06 12:32:11.121 DEBUG qcprludev10 mbatchStandardCombineLegends after Python
## 2023 10 06 12:32:11.121 INFO qcprludev10 HierarchicalClustering_Structures dirVector= /BEA/BatchEffe
## 2023 10 06 12:32:11.122 INFO qcprludev10 HierarchicalClustering_Structures rdataFileSamples= /BEA/Ba
## 2023 10 06 12:32:11.122 INFO qcprludev10 HierarchicalClustering_Structures rdataFileFeatures= /BEA/Ba

## writeBatchDataTsvForBoxplot

## /BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/MWB_HierarchicalClustering_Structures/DATA_2022

## writeAsGenericDataframe writeBatchDataTsvForBoxplot

## Writing BatchData.tsv now

## 2023 10 06 12:32:11.124 INFO qcprludev10 HierarchicalClustering_Structures done

## [[1]]
## [1] "/BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/MWB_HierarchicalClustering_Structures/DATA
##
## [[2]]
## [1] "/BEA/BatchEffectsPackage_data/testing_dynamic/MBatch/MWB_HierarchicalClustering_Structures/DATA
```

7 Example File Output

The above code creates the following subdirectories and files. The subdirectories correspond to the run type were requested.

```
/output/MWB_HierarchicalClustering_Structures$ ls -l DATA_2022-09-09-1600/TEST_2022-10-10-1300/

BatchData.tsv
HCData_feature.tsv
```

```
HCData.tsv
HCOrder_feature.tsv
HCOrder.tsv
HierarchicalClustering_Diagram.png
HierarchicalClustering_Legend-ALL.png
HierarchicalClustering_Legend-time_point.png
HierarchicalClustering_Legend-Treatment.Protocol.png
uDend_feature.RData
uDend.RData
```

##Files

Example data may not match output from above.

###Data TSV Files The TSV files are used by the hierarchical structure visualization code to dynamically build the diagram.

###Data RData Files RData files used internally to re-use clustering results during pipeline runs.

###Data PNG Files The PNG files are: HierarchicalClustering_Diagram.png - the actual diagram HierarchicalClustering_Legend-ALL.png - legend containing all batch types HierarchicalClustering_Legend-time_point.png - batch type specific legend HierarchicalClustering_Legend-Treatment.Protocol.png - batch type specific legend

##Diagram

Here is the diagram generated from this code.

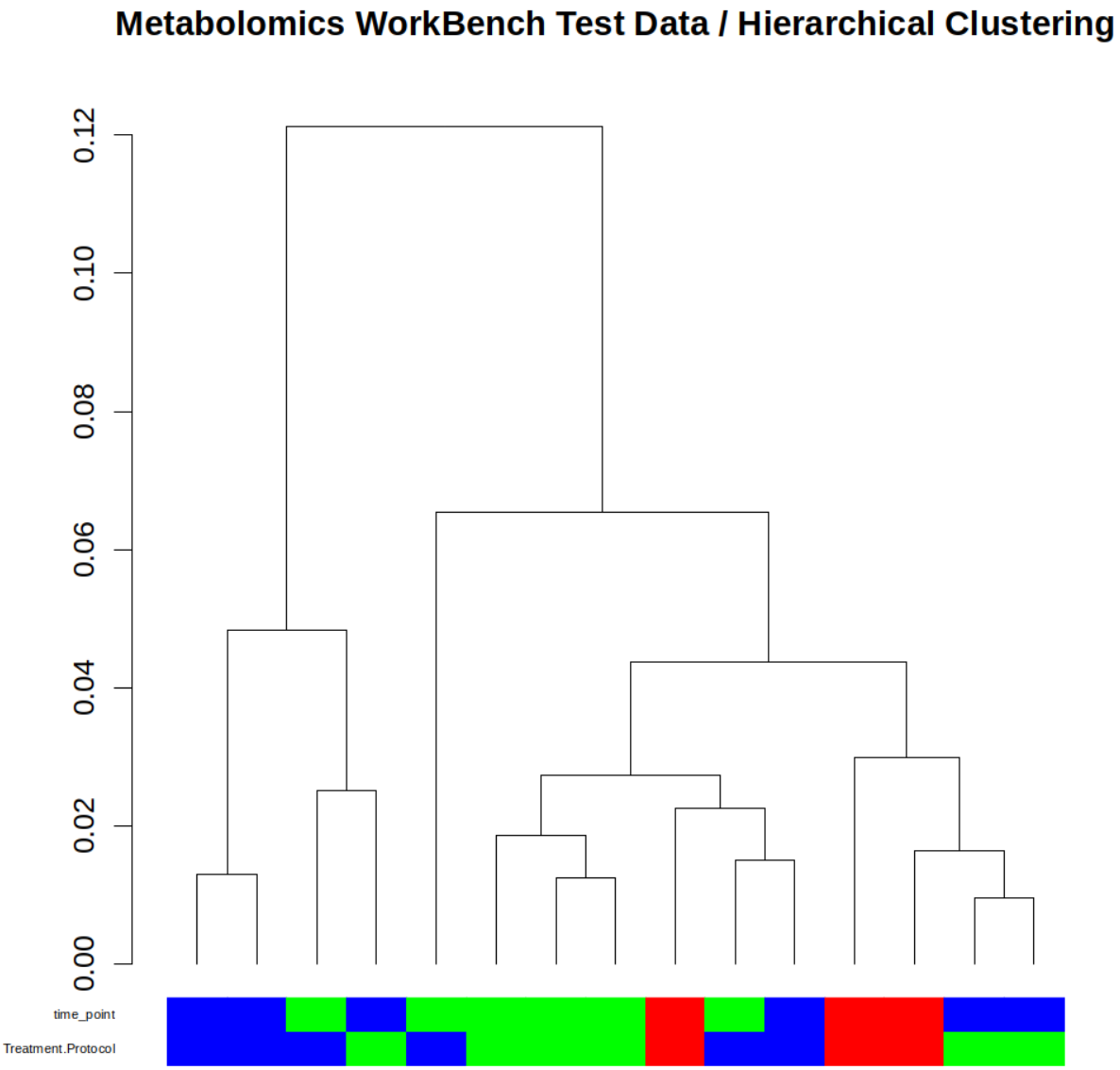

Figure 1: Hierarchical Clustering Output