tidyAssayData Tutorial

Ifedayo Ojo

Background

Following the extraction of gene expression count data for either DNA microarray or RNA sequence analysis from the database or as an output of the NGS pipeline, it is required to perform a transformation task that enables us to have clean column names that represent the samples of the study. The count data usually present with the sample identified that is stacked with some other information that will not be required in the analysis. Therefore we are continuously faced with the daunting task of transforming the column name to a tidy one.

The tidyAssayData Package

With the aim to make the transformation of the count data column names easy, I present `transformMatrixcol()` function from the TidyAssayData package. This function will be useful in the transformation phase of a typical gene expression count data ETL pipeline.

Installation

You can install the tidyAssayData package using `devtools`.

```
library(devtools)
devtools::install_github("pharmlovex/tidyAssayData")
```

Import example count data

To demonstrate how the `transformMatrixCol()` function works, I am going to import count data from gotten from different studies available on the NCBI GEO database.

Import data

load RNA sequence data obtained from NGS pipeline for processing of raw data.

```
```{r}
#| warning: false
# Load library
library(tidyAssayData)
library(dplyr)

# Load count data
df <- read.delim("https://raw.githubusercontent.com/biodata-machine/tidyAssayData/main/cou
                 sep = "", header = TRUE, row.names = 1, comment.char = "#", dec = ".")

# Remove first 5 columns as it is not required for the analysis
df = df %>%
     select(-c(1:5))

head(df,5)
```
```

	mapped.SRR6176429.bam	mapped.SRR6176430.bam	mapped.SRR6176431.bam
gene-Rv0001	2657	3544	2715
gene-Rv0002	1127	1350	1172
gene-Rv0003	876	1085	1007
gene-Rv0004	537	589	549
gene-Rv0005	12152	15908	14242
	mapped.SRR6176432.bam	mapped.SRR6176433.bam	mapped.SRR6176434.bam
gene-Rv0001	1680	2444	3546
gene-Rv0002	1151	1474	1022
gene-Rv0003	960	1122	677
gene-Rv0004	499	601	432
gene-Rv0005	11901	14317	10794
	mapped.SRR6176435.bam	mapped.SRR6176436.bam	mapped.SRR6176437.bam
gene-Rv0001	3526	1339	4206
gene-Rv0002	1132	492	1528
gene-Rv0003	667	501	1392
gene-Rv0004	415	292	668
gene-Rv0005	9579	7758	20865
	mapped.SRR6176438.bam	mapped.SRR6176439.bam	mapped.SRR6176440.bam
gene-Rv0001	4596	1621	3188
gene-Rv0002	1894	517	1148

gene-Rv0003	2071	262	1272
gene-Rv0004	944	111	616
gene-Rv0005	28060	2849	23132

Transform the column names

Before applying the function, let's quickly consider the augments of the function.

- exprMat: count data frame (as a data frame)
- Sep: delimiter between the sample identifier and appendages
- nSep: count of the Delimiter
- pos: position (index) of the sample identifier

Now lets apply the function

```
transform_df = transformMatrixCol(expMat = df,
                                  Sep = ".",
                                  nSep = 2,
                                  pos = 2)

head(transform_df,5)
```

	SRR6176429	SRR6176430	SRR6176431	SRR6176432	SRR6176433	SRR6176434
gene-Rv0001	2657	3544	2715	1680	2444	3546
gene-Rv0002	1127	1350	1172	1151	1474	1022
gene-Rv0003	876	1085	1007	960	1122	677
gene-Rv0004	537	589	549	499	601	432
gene-Rv0005	12152	15908	14242	11901	14317	10794
	SRR6176435	SRR6176436	SRR6176437	SRR6176438	SRR6176439	SRR6176440
gene-Rv0001	3526	1339	4206	4596	1621	3188
gene-Rv0002	1132	492	1528	1894	517	1148
gene-Rv0003	667	501	1392	2071	262	1272
gene-Rv0004	415	292	668	944	111	616
gene-Rv0005	9579	7758	20865	28060	2849	23132

Let's try another count data

Import data

Load a DNA microarray count data from a study in NCBI GEO database with accession number GSE158643

```
urls = "https://raw.githubusercontent.com/pharmlovex/ETLPipeline-DNAmicroarray/main/GSE158
df <- read.delim(urls, header = TRUE, sep = ",",
                 row.names = 1)

head(df,5)
```

```
        GSM4804913_LNCAP_CO_2_HuGene.1_0.st.v1_.CEL
7892501                                    6.904828
7892502                                    5.812288
7892503                                    3.520698
7892504                                    8.764470
7892505                                    3.074535
        GSM4804914_LNCAP_CTL_1_HuGene.1_0.st.v1_.CEL
7892501                                     7.380824
7892502                                     2.765083
7892503                                     2.826773
7892504                                     9.794184
7892505                                     3.572457
        GSM4804915_LNCAP_CTL_2_HuGene.1_0.st.v1_.CEL
7892501                                     8.501313
7892502                                     4.965421
7892503                                     3.028078
7892504                                     9.236701
7892505                                     3.156453
        GSM4804916_DU145_CO_1_HuGene.1_0.st.v1_.CEL
7892501                                    5.206744
7892502                                    5.947775
7892503                                    2.828441
7892504                                    8.172379
7892505                                    3.891313
        GSM4804917_DU145_CO_2_HuGene.1_0.st.v1_.CEL
7892501                                    5.164993
7892502                                    4.858495
7892503                                    2.578302
7892504                                    8.775429
```

```
7892505                                      3.002443
        GSM4804918_DU145_CTL_1_HuGene.1_0.st.v1_.CEL
7892501                                      4.837603
7892502                                      4.629542
7892503                                      3.288168
7892504                                      9.256136
7892505                                      3.910643
        GSM4804919_DU145_CTL_2_HuGene.1_0.st.v1_.CEL
7892501                                      4.713379
7892502                                      3.980863
7892503                                      2.977510
7892504                                      8.665889
7892505                                      3.337936
        GSM4804920_LNCAP_CO_1_HuGene.1_0.st.v1_.CEL
7892501                                     7.487105
7892502                                     6.267868
7892503                                     3.349867
7892504                                     8.572407
7892505                                     3.336003
        GSM4804921_2_astro_lncap_HuGene.1_0.st.v1_.CEL
7892501                                        5.806061
7892502                                        4.742291
7892503                                        2.864958
7892504                                        8.398879
7892505                                        3.010074
        GSM4804922_2_astrocito_HuGene.1_0.st.v1_.CEL
7892501                                     6.087450
7892502                                     4.772105
7892503                                     2.823199
7892504                                     8.936036
7892505                                     3.405757
        GSM4804923_1_astro_du145_HuGene.1_0.st.v1_.CEL
7892501                                        3.835524
7892502                                        4.272858
7892503                                        3.681819
7892504                                        9.120992
7892505                                        4.170316
        GSM4804924_1_astro_lncap_HuGene.1_0.st.v1_.CEL
7892501                                        5.160502
7892502                                        4.720987
7892503                                        2.978415
7892504                                        8.696909
7892505                                        3.188487
```

```
        GSM4804925_1_astrocito_HuGene.1_0.st.v1_.CEL
7892501                                     6.286243
7892502                                     4.592906
7892503                                     3.389678
7892504                                     8.581454
7892505                                     3.255882
        GSM4804926_2_astro_du145_HuGene.1_0.st.v1_.CEL
7892501                                       4.878263
7892502                                       4.521671
7892503                                       3.146955
7892504                                       9.216972
7892505                                       2.909145
```

Apply function

```
df_transform <- transformMatrixCol(expMat = df,
                                   Sep = "_",
                                   nSep = 6,
                                   pos = 1)

head(df_transform)
```

```
        GSM4804913 GSM4804914 GSM4804915 GSM4804916 GSM4804917 GSM4804918
7892501   6.904828   7.380824   8.501313   5.206744   5.164993   4.837603
7892502   5.812288   2.765083   4.965421   5.947775   4.858495   4.629542
7892503   3.520698   2.826773   3.028078   2.828441   2.578302   3.288168
7892504   8.764470   9.794184   9.236701   8.172379   8.775429   9.256136
7892505   3.074535   3.572457   3.156453   3.891313   3.002443   3.910643
7892506   4.097385   5.219914   4.313771   4.771196   4.355183   4.213146
        GSM4804919 GSM4804920 GSM4804921 GSM4804922 GSM4804923 GSM4804924
7892501   4.713379   7.487105   5.806061   6.087450   3.835524   5.160502
7892502   3.980863   6.267868   4.742291   4.772105   4.272858   4.720987
7892503   2.977510   3.349867   2.864958   2.823199   3.681819   2.978415
7892504   8.665889   8.572407   8.398879   8.936036   9.120992   8.696909
7892505   3.337936   3.336003   3.010074   3.405757   4.170316   3.188487
7892506   4.081464   5.761013   3.976960   4.049248   4.316477   4.887078
        GSM4804925 GSM4804926
7892501   6.286243   4.878263
7892502   4.592906   4.521671
7892503   3.389678   3.146955
7892504   8.581454   9.216972
```

```
7892505    3.255882    2.909145
7892506    4.445348    3.505868
```

Finally the third demostration of application of the function

Import data

Load a DNA micro array count data of a study with accession number GSE55945

```
urls = "https://raw.githubusercontent.com/pharmlovex/ETLPipeline-DNAmicroarray/main/GSE559
df <- read.delim(urls, header = TRUE, sep = ",",
                 row.names = 1)

head(df,5)
```

```
          GSM1348933_011508_HGU133_PLUS_2.0_MS_36D6.CEL
1007_s_at                                      9.991477
1053_at                                        7.691397
117_at                                         5.983557
121_at                                         5.817680
1255_g_at                                      2.490914
          GSM1348934_110607_HGU133_PLUS_2.0_MS_36C1.CEL
1007_s_at                                      9.884199
1053_at                                        7.157873
117_at                                         5.329473
121_at                                         5.902699
1255_g_at                                      2.465105
          GSM1348935_011508_HGU133_PLUS_2.0_MS_36D7.CEL
1007_s_at                                     10.001607
1053_at                                        7.006980
117_at                                         4.844578
121_at                                         5.734552
1255_g_at                                      2.737927
          GSM1348936_011508_HGU133_PLUS_2.0_MS_36D8.CEL
1007_s_at                                     10.228229
1053_at                                        7.226520
117_at                                         6.202345
121_at                                         5.850842
1255_g_at                                      2.464717
          GSM1348938_110607_HGU133_PLUS_2.0_MS_36C4.CEL
1007_s_at                                     10.241452
```

```
1053_at                                              7.383506
117_at                                               5.810763
121_at                                               5.653722
1255_g_at                                            2.341199
          GSM1348939_092707_HGU133_PLUS_2.0_NUGEN_TEST07.CEL
1007_s_at                                                9.444288
1053_at                                                  6.924715
117_at                                                   5.957557
121_at                                                   5.901353
1255_g_at                                                3.068366
          GSM1348940_110607_HGU133_PLUS_2.0_MS_36C8.CEL
1007_s_at                                           10.615160
1053_at                                              7.246976
117_at                                               6.123938
121_at                                               5.629536
1255_g_at                                            2.437844
          GSM1348941_110807_HGU133_PLUS_2.0_MS_36A1.CEL
1007_s_at                                           10.213961
1053_at                                              7.289172
117_at                                               5.459237
121_at                                               5.568359
1255_g_at                                            2.279069
          GSM1348942_110607_HGU133_PLUS_2.0_MS_36C2.CEL
1007_s_at                                           10.289259
1053_at                                              7.038109
117_at                                               5.225692
121_at                                               5.388865
1255_g_at                                            2.348351
          GSM1348943_110607_HGU133_PLUS_2.0_MS_36C3.CEL
1007_s_at                                           10.909372
1053_at                                              7.180767
117_at                                               6.615794
121_at                                               5.757363
1255_g_at                                            2.234805
          GSM1348944_011508_HGU133_PLUS_2.0_MS_36C9.CEL
1007_s_at                                           10.364307
1053_at                                              7.465574
117_at                                               5.693298
121_at                                               5.649767
1255_g_at                                            2.438624
          GSM1348945_092707_HGU133_PLUS_2.0_NUGEN_TEST05.CEL
1007_s_at                                               10.003841
1053_at                                                  7.075762
```

```
117_at                                                 5.959141
121_at                                                 6.281441
1255_g_at                                              2.406546
          GSM1348946_110607_HGU133_PLUS_2.0_MS_36C6.CEL
1007_s_at                                           9.796519
1053_at                                             7.097667
117_at                                              6.362296
121_at                                              5.360538
1255_g_at                                           2.331987
          GSM1348947_110607_HGU133_PLUS_2.0_MS_36C7.CEL
1007_s_at                                          10.461698
1053_at                                             7.472731
117_at                                              6.788785
121_at                                              6.031708
1255_g_at                                           2.529830
          GSM1348949_011508_HGU133_PLUS_2.0_MS_36D3.CEL
1007_s_at                                          10.148979
1053_at                                             7.097682
117_at                                              5.263395
121_at                                              5.515413
1255_g_at                                           2.636593
          GSM1348950_011508_HGU133_PLUS_2.0_MS_36D4.CEL
1007_s_at                                          10.308416
1053_at                                             7.500263
117_at                                              5.631823
121_at                                              5.826804
1255_g_at                                           2.224224
          GSM1348951_011508_HGU133_PLUS_2.0_MS_36D5.CEL
1007_s_at                                          10.164578
1053_at                                             7.132205
117_at                                              5.654953
121_at                                              5.215922
1255_g_at                                           2.382408
          GSM1348952_110807_HGU133_PLUS_2.0_MS_36A4.CEL
1007_s_at                                          10.758033
1053_at                                             6.966189
117_at                                              6.773504
121_at                                              5.550213
1255_g_at                                           2.235432
          GSM1348953_110807_HGU133_PLUS_2.0_MS_36A5.CEL
1007_s_at                                          10.306449
1053_at                                             7.190906
117_at                                              6.243322
```

```
121_at                                          5.325679
1255_g_at                                       2.441954
```

Apply function to transform the column name

```
df_transform <- transformMatrixCol(expMat = df,
                                   Sep = "_",
                                   nSep = 6,
                                   pos = 1)

head(df_transform)
```

```
          GSM1348933 GSM1348934 GSM1348935 GSM1348936 GSM1348938 GSM1348939
1007_s_at   9.991477   9.884199  10.001607  10.228229  10.241452   9.444288
1053_at     7.691397   7.157873   7.006980   7.226520   7.383506   6.924715
117_at      5.983557   5.329473   4.844578   6.202345   5.810763   5.957557
121_at      5.817680   5.902699   5.734552   5.850842   5.653722   5.901353
1255_g_at   2.490914   2.465105   2.737927   2.464717   2.341199   3.068366
1294_at     8.908288   7.000133   8.338459   8.637678   8.047593   6.840393
          GSM1348940 GSM1348941 GSM1348942 GSM1348943 GSM1348944 GSM1348945
1007_s_at  10.615160  10.213961  10.289259  10.909372  10.364307  10.003841
1053_at     7.246976   7.289172   7.038109   7.180767   7.465574   7.075762
117_at      6.123938   5.459237   5.225692   6.615794   5.693298   5.959141
121_at      5.629536   5.568359   5.388865   5.757363   5.649767   6.281441
1255_g_at   2.437844   2.279069   2.348351   2.234805   2.438624   2.406546
1294_at     8.721120   7.376658   8.016912   8.962868   8.989413   7.302721
          GSM1348946 GSM1348947 GSM1348949 GSM1348950 GSM1348951 GSM1348952
1007_s_at   9.796519  10.461698  10.148979  10.308416  10.164578  10.758033
1053_at     7.097667   7.472731   7.097682   7.500263   7.132205   6.966189
117_at      6.362296   6.788785   5.263395   5.631823   5.654953   6.773504
121_at      5.360538   6.031708   5.515413   5.826804   5.215922   5.550213
1255_g_at   2.331987   2.529830   2.636593   2.224224   2.382408   2.235432
1294_at     8.776370   9.475852   7.691538   8.230438   8.208712   8.262031
          GSM1348953
1007_s_at  10.306449
1053_at     7.190906
117_at      6.243322
121_at      5.325679
1255_g_at   2.441954
1294_at     8.144334
```

Conclusion

tidyAssayData package provides a `transformMatrixCol()` function that allow easy way to tidy up the column names of count data. Therefore it will be worth adding to your expression gene analysis toolkit.