Mesirov Lab Data Preprocessing Assignment

August 26, 2021

[1]:

```
import pandas as pd;
import os;
import sys;
!{sys.executable} -m pip install cmapPy;
import cmapPy;
from cmapPy.pandasGEXpress.parse import parse;
import matplotlib.pyplot as plt
```

```
Requirement already satisfied: cmapPy in /opt/conda/lib/python3.9/site-packages
(4.0.1)
Requirement already satisfied: h5py>=2.6.0 in /opt/conda/lib/python3.9/site-
packages (from cmapPy) (3.2.1)
Requirement already satisfied: six in /opt/conda/lib/python3.9/site-packages
(from cmapPy) (1.16.0)
Requirement already satisfied: numpy>=1.11.2 in /opt/conda/lib/python3.9/site-
packages (from cmapPy) (1.20.3)
Requirement already satisfied: pandas>=0.18 in /opt/conda/lib/python3.9/site-
packages (from cmapPy) (1.2.4)
Requirement already satisfied: requests>=2.13.0 in
/opt/conda/lib/python3.9/site-packages (from cmapPy) (2.25.1)
Requirement already satisfied: python-dateutil>=2.7.3 in
/opt/conda/lib/python3.9/site-packages (from pandas>=0.18->cmapPy) (2.8.1)
Requirement already satisfied: pytz>=2017.3 in /opt/conda/lib/python3.9/site-
packages (from pandas>=0.18->cmapPy) (2021.1)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.9/site-
packages (from requests>=2.13.0->cmapPy) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in
/opt/conda/lib/python3.9/site-packages (from requests>=2.13.0->cmapPy) (4.0.0)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.9/site-packages (from requests>=2.13.0->cmapPy)
(2021.5.30)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/opt/conda/lib/python3.9/site-packages (from requests>=2.13.0->cmapPy) (1.26.5)
```

[2]:

```
#1
def process_gct(gct_file, summary=False):
    if summary:
        gct_dataframe = cmapPy.pandasGEXpress.parse.parse(gct_file)
        print(str(gct_dataframe.data_df.shape[0]) + ' rows and ' +␣
↪str(gct_dataframe.data_df.shape[1]) + ' columns were imported.')
        return gct_dataframe
    else:
        gct_dataframe = cmapPy.pandasGEXpress.parse.parse(gct_file)
        return gct_dataframe
```

[3]:
```
#2A
gct_df = process_gct('BRCA_minimal_60x19.gct', True)
gct_df
```

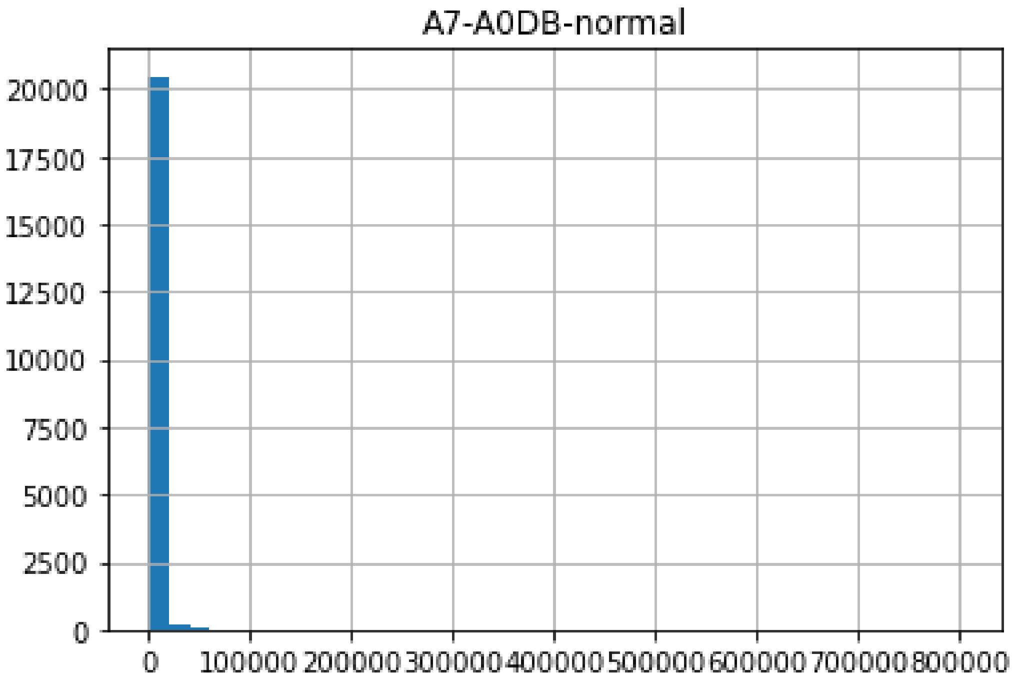
```
60 rows and 19 columns were imported.
```

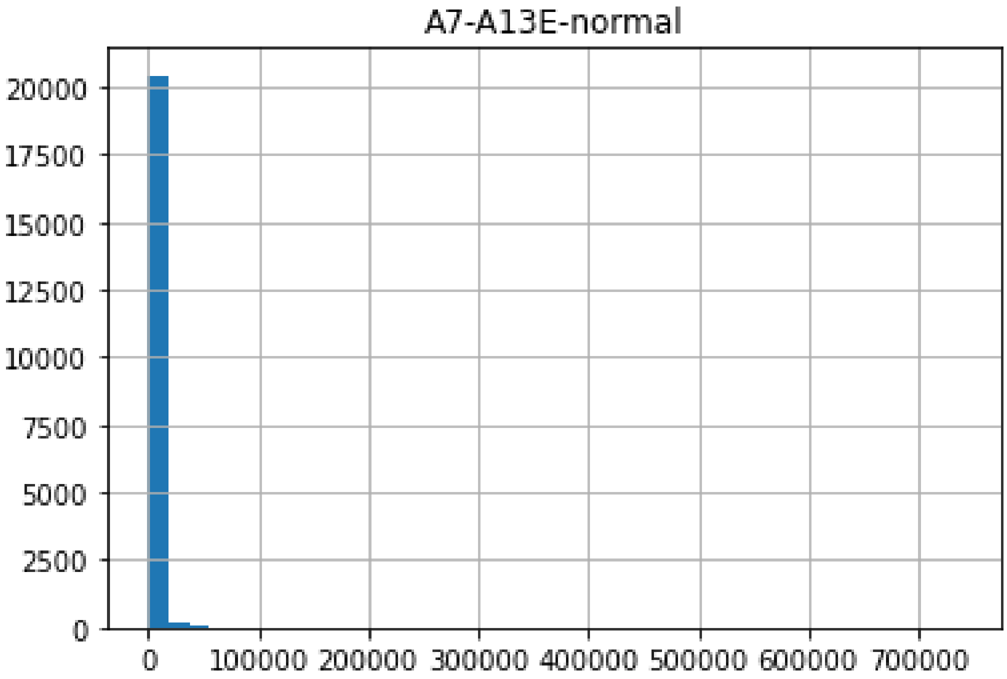
[3]:
```
<cmapPy.pandasGEXpress.GCToo.GCToo at 0x7ff14ac786d0>
```

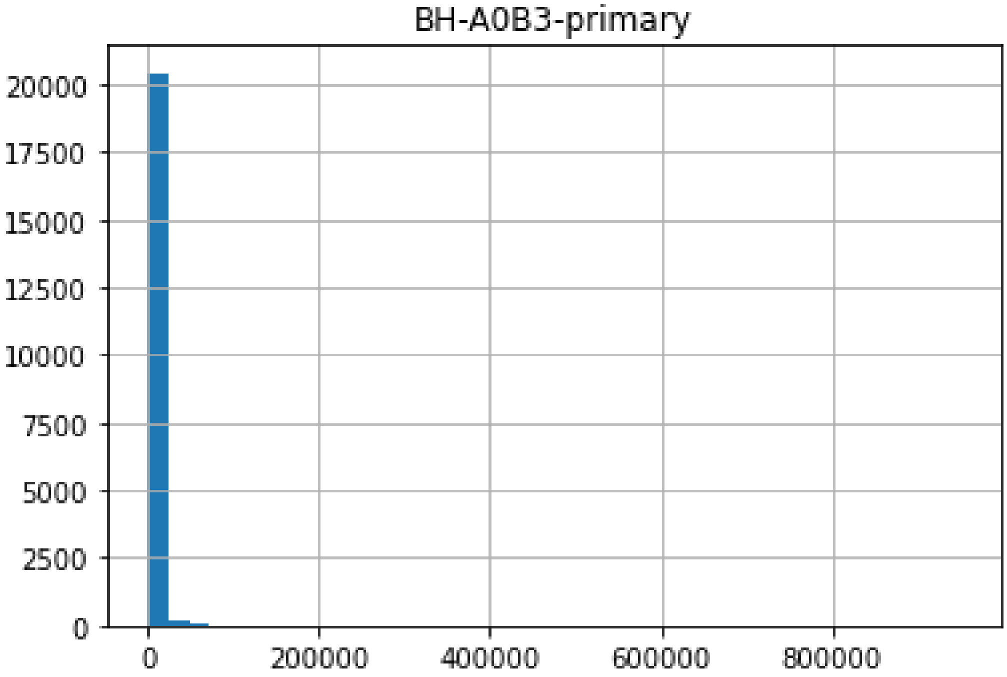
[4]:
```
#2B
gct_df = process_gct('BRCA_large_20783x40.gct')
gct_df
```

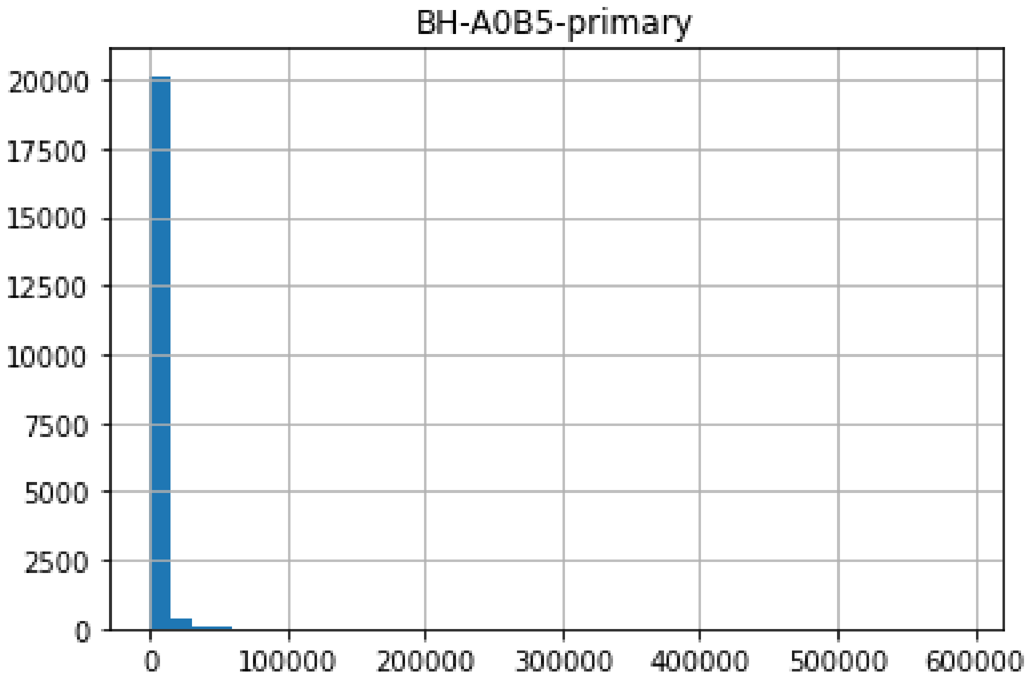
[4]:
```
<cmapPy.pandasGEXpress.GCToo.GCToo at 0x7ff14abdbb20>
```

[5]:
```
#3
hist1 = gct_df.data_df.hist(column='A7-A0DB-normal',bins=40)
hist2 = gct_df.data_df.hist(column='A7-A13E-normal', bins=40)
hist3 = gct_df.data_df.hist(column='BH-A0B3-primary', bins=40)
hist4 = gct_df.data_df.hist(column='BH-A0B5-primary', bins=40)
```




A7-A13E-normal
20000
17500
15000
12500
10000
7500
5000
2500
0
0
100000
200000
300000
400000
500000
600000
700000


BH-A0B3-primary
20000
17500
15000
12500
10000
7500
5000
2500
0
0
200000
400000
600000
800000



[15]:
```
#4
gct_df_new = gct_df.data_df.copy()
gct_df_new['Mean'] = gct_df_new.mean(numeric_only=True, axis=1)
gct_df_new['Median'] = gct_df_new.median(numeric_only=True, axis=1)
gct_df_new['Standard Deviation'] = gct_df_new.std(numeric_only=True, axis=1)
gct_df_new
```

[15]:
```
cid         A7-A0CE-normal  A7-A0CH-normal  A7-A0D9-normal  A7-A0DB-normal  \
rid
TSPAN6              5404.0          5030.0          3616.0          2425.0
TNMD                 320.0          2116.0          3616.0           304.0
DPM1                2472.0          1611.0          1254.0          1137.0
SCYL3               1483.0          1154.0           820.0           687.0
C1orf112             312.0           252.0           225.0           241.0
…                        …               …               …               …
HCP5B                 20.0            20.0            20.0            20.0
SPRY4-IT1             20.0            20.0            27.0            24.0
AC018638.8            27.0            20.0            37.0            78.0
LINC02246             20.0            20.0            20.0            20.0
LINC01144            147.0            79.0            33.0            20.0

cid         A7-A13E-normal  A7-A13F-normal  A7-A13G-normal  AC-A23H-normal  \
rid
TSPAN6              3400.0          3276.0          4611.0          7362.0
```

```
TNMD                 992.0          2159.0           869.0           234.0
DPM1                1242.0          1295.0          1896.0          1813.0
SCYL3                931.0          1178.0          1262.0          1684.0
C1orf112             259.0           277.0           256.0           390.0
…                    …             …             …             …
HCP5B                 20.0            25.0            22.0            25.0
SPRY4-IT1             30.0            32.0            41.0            20.0
AC018638.8            44.0            75.0           139.0            99.0
LINC02246             20.0            21.0            73.0            27.0
LINC01144             36.0            64.0            39.0            77.0

cid         AC-A2FB-normal  AC-A2FF-normal  …  BH-A0AZ-primary  \
rid                                         …
TSPAN6              5389.0          4686.0  …           1946.0
TNMD                1218.0           103.0  …             54.0
DPM1                1930.0          2143.0  …           1235.0
SCYL3               1589.0          1829.0  …           1705.0
C1orf112             331.0           524.0  …            354.0
…                    …             …  …              …
HCP5B                 20.0            20.0  …             20.0
SPRY4-IT1             20.0            42.0  …             20.0
AC018638.8            60.0           104.0  …             23.0
LINC02246             20.0            26.0  …             20.0
LINC01144            112.0           113.0  …             47.0

cid         BH-A0B3-primary  BH-A0B5-primary  BH-A0B7-primary  \
rid
TSPAN6               2498.0           2709.0           3701.0
TNMD                   20.0             20.0             88.0
DPM1                 1853.0           1739.0           2172.0
SCYL3                1168.0           3469.0           2544.0
C1orf112             1166.0           2086.0            325.0
…                     …              …              …
HCP5B                  20.0             20.0             20.0
SPRY4-IT1              20.0             20.0             36.0
AC018638.8             20.0            104.0             28.0
LINC02246              20.0             30.0             20.0
LINC01144              75.0             60.0             48.0

cid         BH-A0B8-primary  BH-A0BA-primary  BH-A0BC-primary         Mean  \
rid
TSPAN6               2390.0           6725.0           1173.0  3703.475098
TNMD                   38.0            113.0             92.0   489.750000
DPM1                 1391.0           3203.0           1709.0  1990.775024
SCYL3                1274.0           4205.0           1687.0  1822.000000
C1orf112              462.0           2162.0           1015.0   628.875000
…                     …              …              …          …
```

```
HCP5B              20.0          20.0          20.0    23.400000
SPRY4-IT1          20.0          20.0          20.0    25.000000
AC018638.8         25.0         103.0          21.0    67.949997
LINC02246          20.0          89.0          91.0    33.099998
LINC01144         123.0          42.0         124.0    76.050003

cid          Median  Standard Deviation
rid
TSPAN6       3701.0         1772.747192
TNMD          107.0          774.445862
DPM1         1846.0          935.971130
SCYL3        1597.0          937.156921
C1orf112      416.0          504.518280
…              …                …
HCP5B          20.0           17.599112
SPRY4-IT1      20.0           16.635548
AC018638.8     61.0           41.067184
LINC02246      20.0           27.743214
LINC01144      64.0           44.973068

[20783 rows x 43 columns]
```

```
[19]: #5A
      rows_to_keep = int(0.9*len(gct_df_new))
      gct_df_new_filtered = gct_df_new.nsmallest(rows_to_keep, 'Standard Deviation')
      gct_df_new_filtered
```

```
[19]: cid          A7-A0CE-normal  A7-A0CH-normal  A7-A0D9-normal  A7-A0DB-normal  \
      rid
      MIR196A2               20.0            20.0            20.0            20.0
      AC007279.1             20.0            20.0            20.0            20.0
      GNA14-AS1              20.0            20.0            20.0            20.0
      RPEP4                  20.0            20.0            20.0            20.0
      KRT18P63               20.0            20.0            20.0            20.0
      …                       …               …               …               …
      NFIA                 7126.0          7274.0          7949.0          5941.0
      DBN1                 5361.0          5511.0          4991.0          8619.0
      SEMA3F               5125.0          3239.0          1133.0          1729.0
      COLGALT1             6159.0          6753.0          4989.0          9277.0
      POGK                 4509.0          3497.0          2309.0          2269.0

      cid          A7-A13E-normal  A7-A13F-normal  A7-A13G-normal  AC-A23H-normal  \
      rid
      MIR196A2               20.0            20.0            20.0            20.0
      AC007279.1             20.0            20.0            20.0            20.0
      GNA14-AS1              20.0            20.0            20.0            20.0
      RPEP4                  20.0            20.0            20.0            20.0
```

```
KRT18P63              20.0            20.0            20.0            20.0
…                    …             …             …             …
NFIA                6735.0          5911.0         11638.0         10207.0
DBN1                3754.0          3336.0          3123.0          5697.0
SEMA3F              2573.0          2118.0          1666.0          1877.0
COLGALT1            5731.0          5908.0          5789.0          4852.0
POGK                2607.0          3002.0          3094.0          5240.0

cid         AC-A2FB-normal  AC-A2FF-normal  …  BH-A0AZ-primary  \
rid                                         …
MIR196A2              20.0            20.0  …             20.0
AC007279.1            20.0            20.0  …             20.0
GNA14-AS1             20.0            20.0  …             20.0
RPEP4                 20.0            20.0  …             20.0
KRT18P63              20.0            20.0  …             20.0
…                    …             …  …              …
NFIA               10953.0          9169.0  …           3426.0
DBN1                8959.0          7181.0  …           4002.0
SEMA3F              3101.0          4372.0  …           2383.0
COLGALT1            6433.0          8612.0  …           6931.0
POGK                4757.0          6798.0  …           5888.0

cid         BH-A0B3-primary  BH-A0B5-primary  BH-A0B7-primary  \
rid
MIR196A2               20.0             20.0             20.0
AC007279.1             20.0             20.0             20.0
GNA14-AS1              20.0             20.0             20.0
RPEP4                  20.0             20.0             20.0
KRT18P63               20.0             20.0             20.0
…                     …              …              …
NFIA                 1914.0           8089.0          16430.0
DBN1                 6108.0           9862.0          17701.0
SEMA3F               4823.0          11359.0           6093.0
COLGALT1            15085.0           6833.0           6353.0
POGK                 8405.0          16259.0           8269.0

cid         BH-A0B8-primary  BH-A0BA-primary  BH-A0BC-primary         Mean  \
rid
MIR196A2               20.0             20.0             20.0    22.500000
AC007279.1             20.0             20.0             20.0    22.500000
GNA14-AS1              20.0             20.0             20.0    22.500000
RPEP4                  20.0             20.0             20.0    22.500000
KRT18P63               20.0             20.0             20.0    22.500000
…                     …              …              …          …
NFIA                 3870.0           6793.0           7789.0  6869.299805
DBN1                 4169.0           2219.0          13547.0  5457.674805
SEMA3F               5689.0           8673.0           7336.0  4920.299805
```

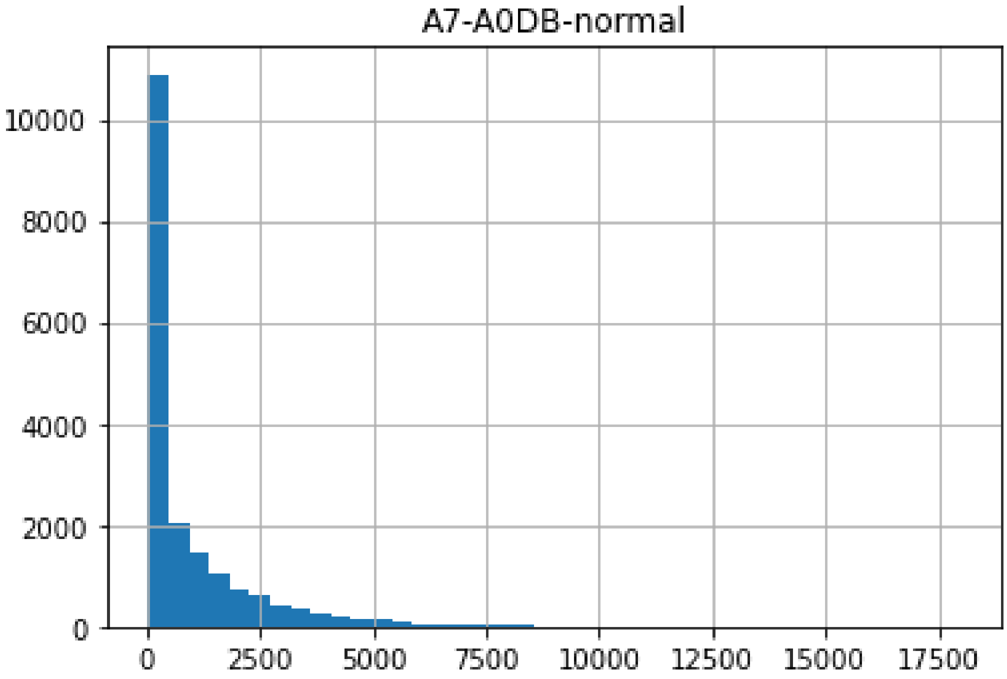
```
COLGALT1              4626.0           5211.0           9484.0  6843.549805
POGK                  5199.0           7859.0           6309.0  6007.575195

cid          Median  Standard Deviation
rid
MIR196A2       20.0           15.425749
AC007279.1     20.0           15.425749
GNA14-AS1      20.0           15.425749
RPEP4          20.0           15.425749
KRT18P63       20.0           15.425749
…               …                   …
NFIA         6793.0         3196.553711
DBN1         4391.0         3196.833008
SEMA3F       3911.0         3204.032471
COLGALT1     5908.0         3205.637451
POGK         5671.0         3207.962402

[18704 rows x 43 columns]
```

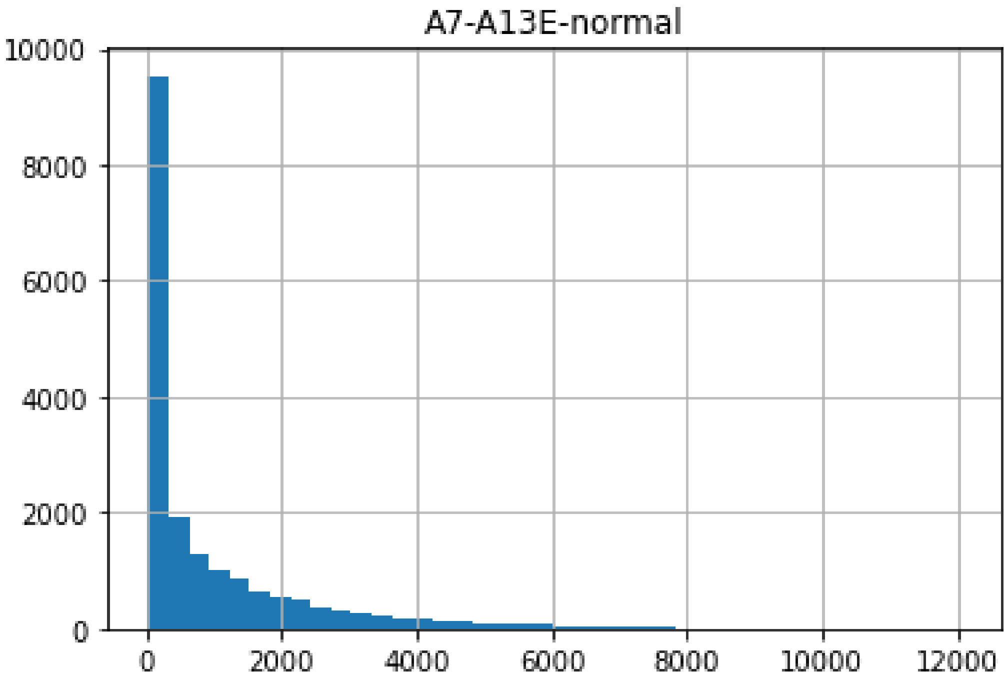
```
[31]: #5B
print(gct_df_new_filtered.loc[:, ~gct_df_new_filtered.columns.isin(['Mean',␣
 ↪'Median', 'Standard Deviation'])].mean())
print(gct_df_new_filtered.loc[:, ~gct_df_new_filtered.columns.isin(['Mean',␣
 ↪'Median', 'Standard Deviation'])].median())
new_hist1 = gct_df_new_filtered.hist(column='A7-A0DB-normal',bins=40)
new_hist2 = gct_df_new_filtered.hist(column='A7-A13E-normal', bins=40)
new_hist3 = gct_df_new_filtered.hist(column='BH-A0B3-primary', bins=40)
new_hist4 = gct_df_new_filtered.hist(column='BH-A0B5-primary', bins=40)
```

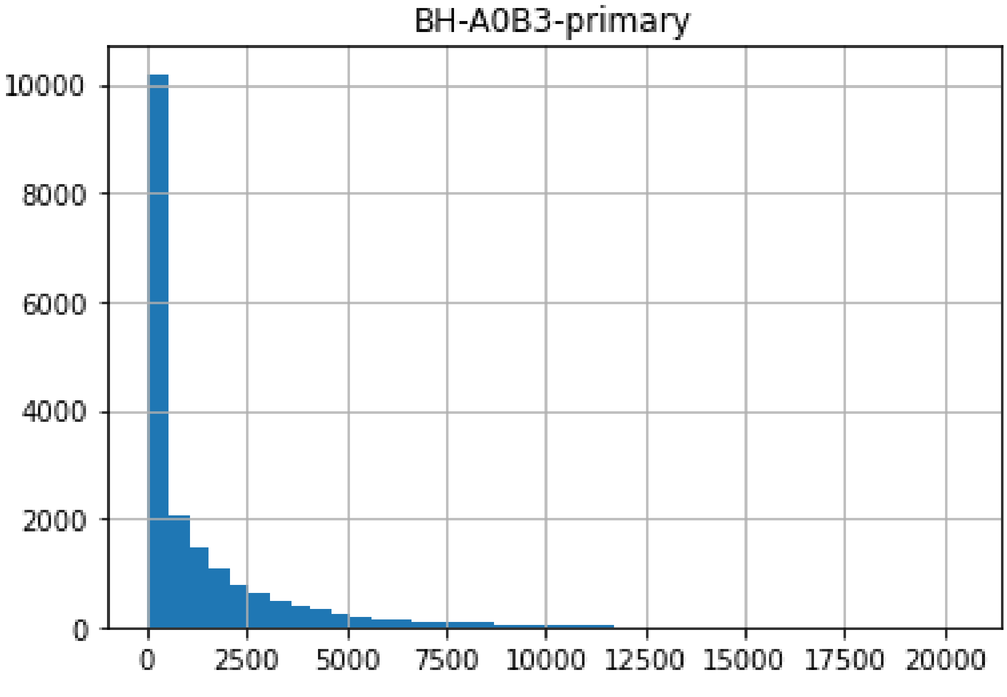
```
cid
A7-A0CE-normal     1483.951050
A7-A0CH-normal     1226.841553
A7-A0D9-normal      958.922791
A7-A0DB-normal      915.061829
A7-A13E-normal      951.364197
A7-A13F-normal     1041.000610
A7-A13G-normal     1281.722046
AC-A23H-normal     1381.501343
AC-A2FB-normal     1550.661621
AC-A2FF-normal     1737.991333
AC-A2FM-normal     1189.578613
BH-A0AU-normal     1223.027100
BH-A0AY-normal     1074.947876
BH-A0AZ-normal     1621.079590
BH-A0B3-normal     1434.162109
BH-A0B5-normal     1091.059448
BH-A0B7-normal     1184.854614
```

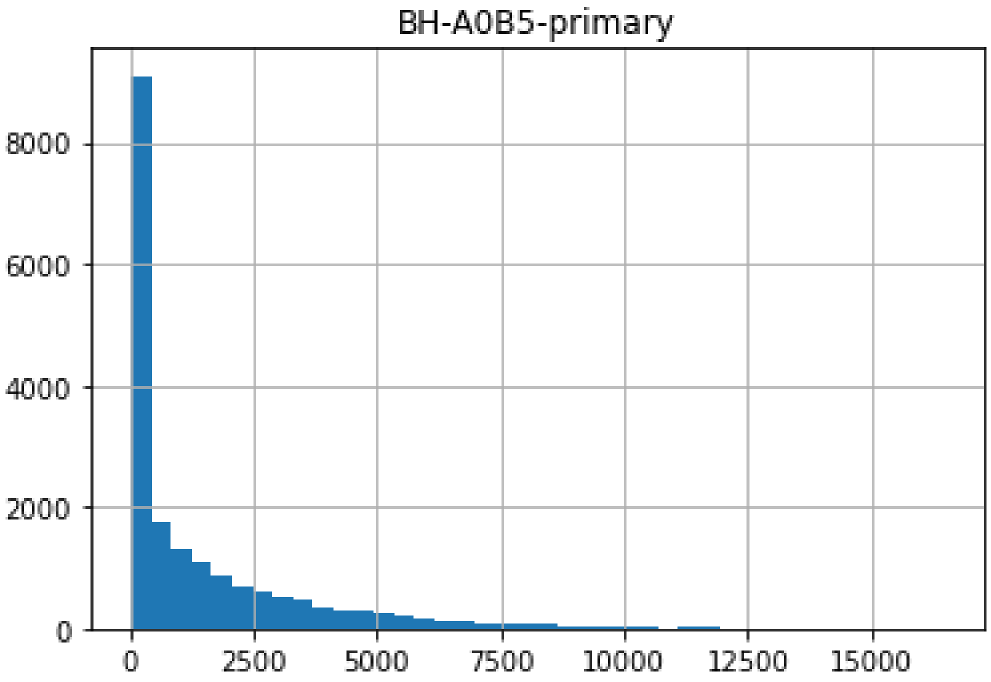
```
BH-A0B8-normal     1040.873779
BH-A0BA-normal      986.037964
BH-A0Bc-normal     1024.977051
A7-A0CE-primary    1458.019531
A7-A0CH-primary    1075.178833
A7-A0D9-primary    1265.842041
A7-A0DB-primary    1238.317749
A7-A13E-primary    1211.962769
A7-A13F-primary    1176.750122
A7-A13G-primary    1037.985718
AC-A23H-primary    1657.331421
AC-A2FB-primary    1484.233154
AC-A2FF-primary    1482.230713
AC-A2FM-primary    1512.691772
BH-A0AU-primary    1134.330078
BH-A0AY-primary    1035.985718
BH-A0AZ-primary     957.272461
BH-A0B3-primary    1276.469238
BH-A0B5-primary    1433.535889
BH-A0B7-primary    1324.126587
BH-A0B8-primary    1024.696167
BH-A0BA-primary    1367.004028
BH-A0BC-primary    1523.738037
dtype: float32
cid
A7-A0CE-normal     566.5
A7-A0CH-normal     440.5
A7-A0D9-normal     245.0
A7-A0DB-normal     256.0
A7-A13E-normal     296.0
A7-A13F-normal     372.5
A7-A13G-normal     320.0
AC-A23H-normal     460.0
AC-A2FB-normal     554.0
AC-A2FF-normal     667.0
AC-A2FM-normal     410.0
BH-A0AU-normal     460.0
BH-A0AY-normal     399.0
BH-A0AZ-normal     595.0
BH-A0B3-normal     551.0
BH-A0B5-normal     238.0
BH-A0B7-normal     424.5
BH-A0B8-normal     259.0
BH-A0BA-normal     367.5
BH-A0Bc-normal     378.0
A7-A0CE-primary    396.0
A7-A0CH-primary    306.0
A7-A0D9-primary    350.0
```

```
A7-A0DB-primary    358.0
A7-A13E-primary    307.0
A7-A13F-primary    294.5
A7-A13G-primary    396.0
AC-A23H-primary    382.0
AC-A2FB-primary    494.5
AC-A2FF-primary    535.0
AC-A2FM-primary    390.0
BH-A0AU-primary    315.0
BH-A0AY-primary    299.5
BH-A0AZ-primary    311.0
BH-A0B3-primary    366.5
BH-A0B5-primary    484.0
BH-A0B7-primary    416.0
BH-A0B8-primary    280.0
BH-A0BA-primary    435.0
BH-A0BC-primary    466.5
dtype: float32
```




A7-A13E-normal
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
12000


BH-A0B3-primary
10000
8000
6000
4000
2000
0
0
2500
5000
7500
10000
12500
15000
17500
20000


BH-A0B5-primary
8000
6000
4000
2000
0
0
2500
5000
7500
10000
12500
15000

```
[ ]:
```