Machine learning predicts metastatic progression using novel differentially expressed lncRNA genes as potential markers in pancreatic cancer

Analysis by *Hasan Alsharoh, M.D.*

"Iuliu Hatieganu" University of Medicine and Pharmacy

Packages and file import

In [1]:
```
import glob
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from scipy.stats import ttest_ind
import os
from pydeseq2.dds import DeseqDataSet
from pydeseq2.ds import DeseqStats
from sanbomics.plots import volcano
import gseapy as gp
from gseapy import barplot, dotplot
from gseapy.plot import gseaplot
from sanbomics.tools import id_map
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
```

Subjects

TWM

In [2]:
```
# Get all the file names in the current directory
filenames = glob.glob(r"\gdc_download_20231025_081138.443578\Included Samples\TWA\*.ts

# Create a list to store the DataFrames
dataframes = []

# Iterate over the file names and create a DataFrame for each file
```

```
for i in range(len(filenames)):
    # Create a DataFrame from the file
    df = pd.read_csv(filenames[i], sep='\t')

    # Give the DataFrame a name
    dataframe_name = f"tw{i+1}"

    # Add the DataFrame to the list
    dataframes.append((dataframe_name, df))

# Create a dictionary to store the DataFrames with their respective names
twdf = dict(dataframes)

# Print the dictionary
print(twdf)
```

```
{'tw1':                   gene_id   gene_name       gene_type  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2497
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1864
3      ENSG00000000457.14       SCYL3  protein_coding         799
4      ENSG00000000460.17    C1orf112  protein_coding         515
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         346
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           1
60659   ENSG00000288675.1  AP006621.6  protein_coding          20

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1255             1242         35.9687           8.4830
1                   0                0          0.0000           0.0000
2                 946              918        100.9062          23.7981
3                 636              638          7.5849           1.7888
4                 519              513          5.6365           1.3293
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             182              171         12.5111           2.9507
60657               0                0          0.0000           0.0000
60658               0                1          0.0068           0.0016
60659              12               16          0.7779           0.1835

       fpkm_uq_unstranded
0                 10.1327
1                  0.0000
2                 28.4261
3                  2.1367
4                  1.5879
...                   ...
60655              0.0000
60656              3.5245
60657              0.0000
60658              0.0019
60659              0.2191

[60660 rows x 9 columns], 'tw2':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2569
1       ENSG00000000005.6        TNMD  protein_coding          30
2      ENSG00000000419.13        DPM1  protein_coding        1948
3      ENSG00000000457.14       SCYL3  protein_coding         667
4      ENSG00000000460.17    C1orf112  protein_coding         239
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         190
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           3
60659   ENSG00000288675.1  AP006621.6  protein_coding          66

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1243             1326         30.3950           9.9182
1                  15               15          1.0908           0.3559
2                 982              966         86.6150          28.2634
3                 579              581          5.2007           1.6970
4                 373              400          2.1485           0.7011
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            111              90          5.6430           1.8414
60657              0               0          0.0000           0.0000
60658              3               0          0.0167           0.0055
60659             50              45          2.1084           0.6880

       fpkm_uq_unstranded
0                 11.7108
1                  0.4203
2                 33.3715
3                  2.0037
4                  0.8278
...                   ...
60655              0.0000
60656              2.1741
60657              0.0000
60658              0.0064
60659              0.8123

[60660 rows x 9 columns], 'tw3':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1013
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding        1143
3      ENSG00000000457.14       SCYL3  protein_coding        1089
4      ENSG00000000460.17    C1orf112  protein_coding         221
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         232
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           6
60659   ENSG00000288675.1  AP006621.6  protein_coding          14

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 514              500         17.6270           6.2181
1                   0                1          0.0535           0.0189
2                 561              582         74.7450          26.3671
3                 878              940         12.4880           4.4053
4                 524              462          2.9219           1.0307
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             113              128         10.1338           3.5748
60657               0                0          0.0000           0.0000
60658               4                2          0.0492           0.0174
60659              18               24          0.6578           0.2320

       fpkm_uq_unstranded
0                  6.5970
1                  0.0200
2                 27.9735
3                  4.6737
4                  1.0935
...                   ...
60655              0.0000
60656              3.7926
60657              0.0000
60658              0.0184
60659              0.2462

[60660 rows x 9 columns], 'tw4':                   gene_id   gene_name       gene_typ
```

```
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1130
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        2011
3      ENSG00000000457.14       SCYL3  protein_coding         330
4      ENSG00000000460.17    C1orf112  protein_coding         353
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         208
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           0
60659   ENSG00000288675.1  AP006621.6  protein_coding          21

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 590              540         21.3032           6.4605
1                   0                0          0.0000           0.0000
2                1003             1008        142.4768          43.2084
3                 349              345          4.0999           1.2434
4                 360              374          5.0564           1.5334
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             124               91          9.8434           2.9852
60657               0                0          0.0000           0.0000
60658               0                0          0.0000           0.0000
60659              13               23          1.0689           0.3242

       fpkm_uq_unstranded
0                  7.4173
1                  0.0000
2                 49.6071
3                  1.4275
4                  1.7605
...                   ...
60655              0.0000
60656              3.4272
60657              0.0000
60658              0.0000
60659              0.3722

[60660 rows x 9 columns], 'tw5':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1361
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1148
3      ENSG00000000457.14       SCYL3  protein_coding         711
4      ENSG00000000460.17    C1orf112  protein_coding         216
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         173
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           0
60659   ENSG00000288675.1  AP006621.6  protein_coding          10

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 706              655         33.6303          10.7114
1                   0                0          0.0000           0.0000
2                 570              578        106.6058          33.9545
3                 584              563         11.5781           3.6877
4                 354              327          4.0553           1.2916
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            108              92         10.7308           3.4178
60657              0               0          0.0000           0.0000
60658              0               0          0.0000           0.0000
60659             12              18          0.6672           0.2125

       fpkm_uq_unstranded
0                 11.4011
1                  0.0000
2                 36.1407
3                  3.9251
4                  1.3748
...                   ...
60655              0.0000
60656              3.6379
60657              0.0000
60658              0.0000
60659              0.2262

[60660 rows x 9 columns], 'tw6':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1343
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1642
3      ENSG00000000457.14       SCYL3  protein_coding         533
4      ENSG00000000460.17    C1orf112  protein_coding         200
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         155
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           1
60659   ENSG00000288675.1  AP006621.6  protein_coding          16

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 655              688         28.0375           6.8727
1                   0                0          0.0000           0.0000
2                 845              797        128.8256          31.5785
3                 371              406          7.3331           1.7975
4                 267              216          3.1724           0.7776
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              89               71          8.1229           1.9911
60657               0                0          0.0000           0.0000
60658               0                1          0.0098           0.0024
60659              17               16          0.9019           0.2211

       fpkm_uq_unstranded
0                  7.5585
1                  0.0000
2                 34.7295
3                  1.9769
4                  0.8552
...                   ...
60655              0.0000
60656              2.1898
60657              0.0000
60658              0.0027
60659              0.2431

[60660 rows x 9 columns], 'tw7':                   gene_id   gene_name       gene_typ
```

```
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        3380
1       ENSG00000000005.6        TNMD  protein_coding           7
2      ENSG00000000419.13        DPM1  protein_coding        1392
3      ENSG00000000457.14       SCYL3  protein_coding         808
4      ENSG00000000460.17    C1orf112  protein_coding         224
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         286
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           8
60659   ENSG00000288675.1  AP006621.6  protein_coding          43

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1772             1608         40.0657          13.5324
1                   3                4          0.2550           0.0861
2                 712              680         62.0098          20.9441
3                 660              667          6.3119           2.1319
4                 384              394          2.0174           0.6814
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             140              148          8.5101           2.8743
60657               0                0          0.0000           0.0000
60658               4                4          0.0447           0.0151
60659              28               43          1.3762           0.4648

       fpkm_uq_unstranded
0                 15.1690
1                  0.0965
2                 23.4771
3                  2.3897
4                  0.7638
...                   ...
60655              0.0000
60656              3.2220
60657              0.0000
60658              0.0169
60659              0.5210

[60660 rows x 9 columns], 'tw8':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1721
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding         981
3      ENSG00000000457.14       SCYL3  protein_coding         572
4      ENSG00000000460.17    C1orf112  protein_coding         135
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         124
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           4
60659   ENSG00000288675.1  AP006621.6  protein_coding          21

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 852              869         36.5572          11.4202
1                   0                1          0.0653           0.0204
2                 483              498         78.3117          24.4639
3                 453              447          8.0072           2.5014
4                 256              229          2.1788           0.6807
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656             57              71          6.6119           2.0655
60657              0               0          0.0000           0.0000
60658              2               2          0.0401           0.0125
60659             21              17          1.2044           0.3762

       fpkm_uq_unstranded
0                 11.7596
1                  0.0210
2                 25.1911
3                  2.5758
4                  0.7009
...                   ...
60655              0.0000
60656              2.1269
60657              0.0000
60658              0.0129
60659              0.3874

[60660 rows x 9 columns], 'tw9':                   gene_id   gene_name       gene_typ
e  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2870
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1813
3      ENSG00000000457.14       SCYL3  protein_coding        1254
4      ENSG00000000460.17    C1orf112  protein_coding         335
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         493
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding          15
60659   ENSG00000288675.1  AP006621.6  protein_coding          11

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1444             1426         35.9650          12.8984
1                   0                0          0.0000           0.0000
2                 894              919         85.3810          30.6210
3                1312             1240         10.3560           3.7141
4                 845              839          3.1896           1.1439
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             267              255         15.5081           5.5618
60657               0                0          0.0000           0.0000
60658               9                6          0.0886           0.0318
60659              14               20          0.3722           0.1335

       fpkm_uq_unstranded
0                 16.4931
1                  0.0000
2                 39.1547
3                  4.7491
4                  1.4627
...                   ...
60655              0.0000
60656              7.1118
60657              0.0000
60658              0.0406
60659              0.1707

[60660 rows x 9 columns], 'tw10':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding         853
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding         440
3      ENSG00000000457.14       SCYL3  protein_coding         599
4      ENSG00000000460.17    C1orf112  protein_coding         132
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         208
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           3
60659   ENSG00000288675.1  AP006621.6  protein_coding          13

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 424              429         24.0887           8.5652
1                   0                0          0.0000           0.0000
2                 228              212         46.6962          16.6037
3                 506              482         11.1477           3.9638
4                 241              290          2.8323           1.0071
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             101              116         14.7449           5.2428
60657               0                0          0.0000           0.0000
60658               1                2          0.0399           0.0142
60659              18               17          0.9912           0.3524

       fpkm_uq_unstranded
0                  8.9191
1                  0.0000
2                 17.2898
3                  4.1276
4                  1.0487
...                   ...
60655              0.0000
60656              5.4595
60657              0.0000
60658              0.0148
60659              0.3670

[60660 rows x 9 columns], 'tw11':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding         964
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding         439
3      ENSG00000000457.14       SCYL3  protein_coding         286
4      ENSG00000000460.17    C1orf112  protein_coding         121
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA          97
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           4
60659   ENSG00000288675.1  AP006621.6  protein_coding          20

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 451              513         23.0390           8.9182
1                   0                0          0.0000           0.0000
2                 222              217         39.4291          15.2627
3                 248              242          4.5045           1.7437
4                 164              165          2.1972           0.8505
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656             49              51          5.8193           2.2526
60657              0               0          0.0000           0.0000
60658              1               3          0.0451           0.0174
60659             19              12          1.2906           0.4996

       fpkm_uq_unstranded
0                 11.9065
1                  0.0000
2                 20.3769
3                  2.3279
4                  1.1355
...                   ...
60655              0.0000
60656              3.0074
60657              0.0000
60658              0.0233
60659              0.6670

[60660 rows x 9 columns], 'tw12':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2448
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding        1429
3      ENSG00000000457.14       SCYL3  protein_coding         579
4      ENSG00000000460.17    C1orf112  protein_coding         211
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         202
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          18

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1227             1221         35.1639          10.4063
1                   1                0          0.0441           0.0131
2                 720              709         77.1407          22.8287
3                 497              493          5.4810           1.6220
4                 322              333          2.3029           0.6815
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             104              104          7.2837           2.1555
60657               0                0          0.0000           0.0000
60658               0                2          0.0135           0.0040
60659              27               21          0.6981           0.2066

       fpkm_uq_unstranded
0                 11.3218
1                  0.0142
2                 24.8371
3                  1.7647
4                  0.7415
...                   ...
60655              0.0000
60656              2.3451
60657              0.0000
60658              0.0044
60659              0.2248

[60660 rows x 9 columns], 'tw13':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1714
1       ENSG00000000005.6        TNMD  protein_coding          11
2      ENSG00000000419.13        DPM1  protein_coding        1365
3      ENSG00000000457.14       SCYL3  protein_coding         471
4      ENSG00000000460.17    C1orf112  protein_coding         135
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         220
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          79

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 856              858         20.2279           8.1383
1                   6                5          0.3990           0.1605
2                 677              688         60.5395          24.3569
3                 422              434          3.6632           1.4738
4                 292              254          1.2105           0.4870
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             108              123          6.5175           2.6222
60657               0                0          0.0000           0.0000
60658               1                1          0.0111           0.0045
60659              57               52          2.5173           1.0128

       fpkm_uq_unstranded
0                 10.6818
1                  0.2107
2                 31.9691
3                  1.9344
4                  0.6392
...                   ...
60655              0.0000
60656              3.4417
60657              0.0000
60658              0.0059
60659              1.3293

[60660 rows x 9 columns], 'tw14':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1214
1       ENSG00000000005.6        TNMD  protein_coding           4
2      ENSG00000000419.13        DPM1  protein_coding        1919
3      ENSG00000000457.14       SCYL3  protein_coding        1037
4      ENSG00000000460.17    C1orf112  protein_coding         251
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           1
60656   ENSG00000288670.1  AL592295.6          lncRNA         510
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           7
60659   ENSG00000288675.1  AP006621.6  protein_coding          38

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 585              629         14.9337           4.6137
1                   1                3          0.1512           0.0467
2                 975              944         88.7136          27.4077
3                 794              836          8.4067           2.5972
4                 462              450          2.3460           0.7248
...               ...              ...             ...              ...
```

```
60655              0               1          0.0140           0.0043
60656            237             296         15.7483           4.8654
60657              0               0          0.0000           0.0000
60658              4               3          0.0406           0.0125
60659             39              39          1.2621           0.3899

       fpkm_uq_unstranded
0                  4.8768
1                  0.0494
2                 28.9705
3                  2.7453
4                  0.7661
...                   ...
60655              0.0046
60656              5.1428
60657              0.0000
60658              0.0133
60659              0.4122

[60660 rows x 9 columns], 'tw15':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1658
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        1538
3      ENSG00000000457.14       SCYL3  protein_coding         528
4      ENSG00000000460.17    C1orf112  protein_coding         190
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           1
60656   ENSG00000288670.1  AL592295.6          lncRNA         303
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          19

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 801              857         28.3285           9.2633
1                   0                2          0.1050           0.0343
2                 760              778         98.7555          32.2925
3                 453              417          5.9452           1.9441
4                 284              281          2.4666           0.8066
...               ...              ...             ...              ...
60655               0                1          0.0194           0.0064
60656             158              149         12.9956           4.2495
60657               0                0          0.0000           0.0000
60658               1                1          0.0161           0.0053
60659              19               25          0.8765           0.2866

       fpkm_uq_unstranded
0                 10.0938
1                  0.0374
2                 35.1880
3                  2.1184
4                  0.8789
...                   ...
60655              0.0069
60656              4.6305
60657              0.0000
60658              0.0057
60659              0.3123

[60660 rows x 9 columns], 'tw16':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1371
1       ENSG00000000005.6        TNMD  protein_coding           4
2      ENSG00000000419.13        DPM1  protein_coding        1005
3      ENSG00000000457.14       SCYL3  protein_coding        1012
4      ENSG00000000460.17    C1orf112  protein_coding         208
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         199
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           1
60659   ENSG00000288675.1  AP006621.6  protein_coding          22

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 685              686         25.6840           7.7835
1                   4                0          0.2303           0.0698
2                 501              505         70.7550          21.4421
3                 883              771         12.4940           3.7863
4                 415              464          2.9607           0.8972
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              99              110          9.3582           2.8360
60657               0                0          0.0000           0.0000
60658               1                0          0.0088           0.0027
60659              23               24          1.1128           0.3372

       fpkm_uq_unstranded
0                  8.0709
1                  0.0724
2                 22.2340
3                  3.9261
4                  0.9304
...                   ...
60655              0.0000
60656              2.9407
60657              0.0000
60658              0.0028
60659              0.3497

[60660 rows x 9 columns], 'tw17':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2505
1       ENSG00000000005.6        TNMD  protein_coding          12
2      ENSG00000000419.13        DPM1  protein_coding        1122
3      ENSG00000000457.14       SCYL3  protein_coding         540
4      ENSG00000000460.17    C1orf112  protein_coding         174
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         415
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding          13
60659   ENSG00000288675.1  AP006621.6  protein_coding          34

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1273             1232         40.2001          14.4699
1                   6                6          0.5918           0.2130
2                 571              551         67.6672          24.3566
3                 551              612          5.7110           2.0556
4                 430              392          2.1216           0.7637
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            202             231         16.7179           6.0176
60657              0               0          0.0000           0.0000
60658              8               5          0.0984           0.0354
60659             33              17          1.4732           0.5303

       fpkm_uq_unstranded
0                 16.8045
1                  0.2474
2                 28.2864
3                  2.3873
4                  0.8869
...                   ...
60655              0.0000
60656              6.9885
60657              0.0000
60658              0.0411
60659              0.6158

[60660 rows x 9 columns], 'tw18':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1525
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding         816
3      ENSG00000000457.14       SCYL3  protein_coding        1166
4      ENSG00000000460.17    C1orf112  protein_coding         135
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         295
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           3
60659   ENSG00000288675.1  AP006621.6  protein_coding          19

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 739              786         30.8289           9.7999
1                   0                1          0.0621           0.0197
2                 416              400         61.9932          19.7064
3                 870              890         15.5340           4.9379
4                 383              369          2.0736           0.6592
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             157              150         14.9701           4.7587
60657               0                0          0.0000           0.0000
60658               0                3          0.0286           0.0091
60659              21               26          1.0371           0.3297

       fpkm_uq_unstranded
0                 10.3320
1                  0.0208
2                 20.7764
3                  5.2061
4                  0.6949
...                   ...
60655              0.0000
60656              5.0171
60657              0.0000
60658              0.0096
60659              0.3476

[60660 rows x 9 columns], 'tw19':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1257
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        1195
3      ENSG00000000457.14       SCYL3  protein_coding         400
4      ENSG00000000460.17    C1orf112  protein_coding         113
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         143
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          27

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 638              619         25.6005           7.9655
1                   1                1          0.1252           0.0389
2                 604              591         91.4634          28.4584
3                 348              362          5.3687           1.6704
4                 231              219          1.7486           0.5441
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              75               73          7.3108           2.2747
60657               0                0          0.0000           0.0000
60658               0                2          0.0192           0.0060
60659              23               20          1.4847           0.4620

       fpkm_uq_unstranded
0                  8.2322
1                  0.0403
2                 29.4115
3                  1.7264
4                  0.5623
...                   ...
60655              0.0000
60656              2.3509
60657              0.0000
60658              0.0062
60659              0.4774

[60660 rows x 9 columns], 'tw20':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2178
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        2149
3      ENSG00000000457.14       SCYL3  protein_coding         821
4      ENSG00000000460.17    C1orf112  protein_coding         387
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         429
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding          15
60659   ENSG00000288675.1  AP006621.6  protein_coding          45

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1060             1118         25.5823           8.1724
1                   0                0          0.0000           0.0000
2                1056             1093         94.8602          30.3037
3                 759              732          6.3551           2.0302
4                 556              583          3.4538           1.1033
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            217             238         12.6489           4.0408
60657              0               0          0.0000           0.0000
60658              7               8          0.0831           0.0265
60659             59              43          1.4271           0.4559

       fpkm_uq_unstranded
0                  8.8974
1                  0.0000
2                 32.9919
3                  2.2103
4                  1.2012
...                   ...
60655              0.0000
60656              4.3992
60657              0.0000
60658              0.0289
60659              0.4963

[60660 rows x 9 columns], 'tw21':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2686
1       ENSG00000000005.6        TNMD  protein_coding           3
2      ENSG00000000419.13        DPM1  protein_coding        1123
3      ENSG00000000457.14       SCYL3  protein_coding         693
4      ENSG00000000460.17    C1orf112  protein_coding         249
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         156
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           3
60659   ENSG00000288675.1  AP006621.6  protein_coding          29

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1346             1340         48.4849          14.5921
1                   2                1          0.1664           0.0501
2                 543              580         76.1809          22.9275
3                 550              545          8.2438           2.4811
4                 334              345          3.4151           1.0278
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              91               73          7.0687           2.1274
60657               0                0          0.0000           0.0000
60658               3                0          0.0255           0.0077
60659              34               30          1.4134           0.4254

       fpkm_uq_unstranded
0                 15.2541
1                  0.0524
2                 23.9677
3                  2.5936
4                  1.0744
...                   ...
60655              0.0000
60656              2.2239
60657              0.0000
60658              0.0080
60659              0.4447

[60660 rows x 9 columns], 'tw22':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1048
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1362
3      ENSG00000000457.14       SCYL3  protein_coding         823
4      ENSG00000000460.17    C1orf112  protein_coding         183
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         283
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           5
60659   ENSG00000288675.1  AP006621.6  protein_coding          11

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 505              543         18.3531           6.3153
1                   0                0          0.0000           0.0000
2                 684              678         89.6377          30.8442
3                 668              706          9.4982           3.2683
4                 393              378          2.4350           0.8379
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             140              157         12.4408           4.2809
60657               0                0          0.0000           0.0000
60658               1                4          0.0413           0.0142
60659               8               13          0.5201           0.1790

       fpkm_uq_unstranded
0                  6.8355
1                  0.0000
2                 33.3849
3                  3.5375
4                  0.9069
...                   ...
60655              0.0000
60656              4.6335
60657              0.0000
60658              0.0154
60659              0.1937

[60660 rows x 9 columns], 'tw23':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2850
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1281
3      ENSG00000000457.14       SCYL3  protein_coding         535
4      ENSG00000000460.17    C1orf112  protein_coding         153
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         202
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           5
60659   ENSG00000288675.1  AP006621.6  protein_coding          18

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1401             1449         43.0208          14.6661
1                   0                0          0.0000           0.0000
2                 656              625         72.6689          24.7734
3                 434              443          5.3221           1.8143
4                 268              258          1.7548           0.5982
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            109             107          7.6542           2.6094
60657              0               0          0.0000           0.0000
60658              4               1          0.0356           0.0121
60659             24              19          0.7336           0.2501

       fpkm_uq_unstranded
0                 16.0928
1                  0.0000
2                 27.1833
3                  1.9908
4                  0.6564
...                   ...
60655              0.0000
60656              2.8632
60657              0.0000
60658              0.0133
60659              0.2744

[60660 rows x 9 columns], 'tw24':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1284
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding        1124
3      ENSG00000000457.14       SCYL3  protein_coding         525
4      ENSG00000000460.17    C1orf112  protein_coding         171
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         115
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           4
60659   ENSG00000288675.1  AP006621.6  protein_coding          21

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 633              651         22.7559           7.3503
1                   1                0          0.0545           0.0176
2                 558              566         74.8620          24.1808
3                 442              407          6.1317           1.9806
4                 248              266          2.3026           0.7438
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              65               58          5.1161           1.6525
60657               0                0          0.0000           0.0000
60658               1                3          0.0334           0.0108
60659              28               17          1.0049           0.3246

       fpkm_uq_unstranded
0                  8.0879
1                  0.0194
2                 26.6074
3                  2.1793
4                  0.8184
...                   ...
60655              0.0000
60656              1.8184
60657              0.0000
60658              0.0119
60659              0.3572

[60660 rows x 9 columns], 'tw25':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1457
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        1255
3      ENSG00000000457.14       SCYL3  protein_coding        1080
4      ENSG00000000460.17    C1orf112  protein_coding         227
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         342
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           8
60659   ENSG00000288675.1  AP006621.6  protein_coding          36

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 728              729         16.1380           6.0046
1                   0                2          0.0681           0.0253
2                 633              622         52.2396          19.4372
3                 841              827          7.8833           2.9332
4                 424              421          1.9104           0.7108
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             175              191          9.5089           3.5381
60657               0                0          0.0000           0.0000
60658               5                3          0.0418           0.0155
60659              29               33          1.0766           0.4006

       fpkm_uq_unstranded
0                  7.0392
1                  0.0297
2                 22.7862
3                  3.4386
4                  0.8333
...                   ...
60655              0.0000
60656              4.1477
60657              0.0000
60658              0.0182
60659              0.4696

[60660 rows x 9 columns], 'tw26':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1330
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1438
3      ENSG00000000457.14       SCYL3  protein_coding         506
4      ENSG00000000460.17    C1orf112  protein_coding         135
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           1
60656   ENSG00000288670.1  AL592295.6          lncRNA         208
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           3
60659   ENSG00000288675.1  AP006621.6  protein_coding          27

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 642              688         18.7401           6.8827
1                   0                0          0.0000           0.0000
2                 697              741         76.1455          27.9663
3                 439              391          4.6986           1.7257
4                 229              251          1.4453           0.5308
...               ...              ...             ...              ...
```

```
60655              1               0          0.0160           0.0059
60656             91             123          7.3570           2.7020
60657              0               0          0.0000           0.0000
60658              1               2          0.0199           0.0073
60659             31              36          1.0272           0.3773

       fpkm_uq_unstranded
0                  7.4872
1                  0.0000
2                 30.4222
3                  1.8772
4                  0.5774
...                   ...
60655              0.0064
60656              2.9393
60657              0.0000
60658              0.0080
60659              0.4104

[60660 rows x 9 columns], 'tw27':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2139
1       ENSG00000000005.6        TNMD  protein_coding          91
2      ENSG00000000419.13        DPM1  protein_coding        2767
3      ENSG00000000457.14       SCYL3  protein_coding         232
4      ENSG00000000460.17    C1orf112  protein_coding         351
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         107
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           7
60659   ENSG00000288675.1  AP006621.6  protein_coding          66

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1104             1035         31.7336          10.8061
1                  46               45          4.1489           1.4128
2                1447             1320        154.2707          52.5333
3                 429              420          2.2683           0.7724
4                 481              504          3.9565           1.3473
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              60               50          3.9848           1.3569
60657               0                0          0.0000           0.0000
60658               4                3          0.0490           0.0167
60659              46               39          2.6437           0.9003

       fpkm_uq_unstranded
0                 12.6531
1                  1.6543
2                 61.5124
3                  0.9044
4                  1.5776
...                   ...
60655              0.0000
60656              1.5889
60657              0.0000
60658              0.0195
60659              1.0541

[60660 rows x 9 columns], 'tw28':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2202
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        1724
3      ENSG00000000457.14       SCYL3  protein_coding        1416
4      ENSG00000000460.17    C1orf112  protein_coding         742
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         396
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          16

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1072             1130         23.8286           8.1127
1                   1                1          0.0665           0.0226
2                 889              835         70.1106          23.8698
3                1220             1155         10.0981           3.4380
4                 882              871          6.1008           2.0771
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             215              196         10.7570           3.6623
60657               0                0          0.0000           0.0000
60658               0                2          0.0102           0.0035
60659              15               22          0.4675           0.1592

       fpkm_uq_unstranded
0                  9.0177
1                  0.0252
2                 26.5327
3                  3.8215
4                  2.3088
...                   ...
60655              0.0000
60656              4.0709
60657              0.0000
60658              0.0039
60659              0.1769

[60660 rows x 9 columns], 'tw29':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        3310
1       ENSG00000000005.6        TNMD  protein_coding         180
2      ENSG00000000419.13        DPM1  protein_coding        2588
3      ENSG00000000457.14       SCYL3  protein_coding        1868
4      ENSG00000000460.17    C1orf112  protein_coding         502
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           1
60656   ENSG00000288670.1  AL592295.6          lncRNA         533
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding          17
60659   ENSG00000288675.1  AP006621.6  protein_coding          71

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1671             1639         26.8066           8.7351
1                  86               94          4.4800           1.4598
2                1290             1298         78.7670          25.6666
3                1364             1374          9.9698           3.2487
4                 700              731          3.0890           1.0066
...               ...              ...             ...              ...
```

```
60655              0               1          0.0092          0.0030
60656            291             260         10.8357          3.5309
60657              0               0          0.0000          0.0000
60658              5              12          0.0649          0.0212
60659             80              74          1.5525          0.5059

       fpkm_uq_unstranded
0                  8.5904
1                  1.4356
2                 25.2416
3                  3.1949
4                  0.9899
...                   ...
60655              0.0030
60656              3.4724
60657              0.0000
60658              0.0208
60659              0.4975

[60660 rows x 9 columns], 'tw30':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1426
1       ENSG00000000005.6        TNMD  protein_coding           3
2      ENSG00000000419.13        DPM1  protein_coding        1333
3      ENSG00000000457.14       SCYL3  protein_coding         339
4      ENSG00000000460.17    C1orf112  protein_coding         136
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         208
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           1
60659   ENSG00000288675.1  AP006621.6  protein_coding          84

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 711              715         19.3546           6.9232
1                   1                2          0.1251           0.0448
2                 681              652         67.9924          24.3212
3                 476              450          3.0322           1.0846
4                 360              374          1.4025           0.5017
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             109              113          7.0867           2.5349
60657               0                0          0.0000           0.0000
60658               1                0          0.0064           0.0023
60659              57               58          3.0783           1.1011

       fpkm_uq_unstranded
0                  8.1004
1                  0.0524
2                 28.4566
3                  1.2691
4                  0.5870
...                   ...
60655              0.0000
60656              2.9660
60657              0.0000
60658              0.0027
60659              1.2883

[60660 rows x 9 columns], 'tw31':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2300
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        1901
3      ENSG00000000457.14       SCYL3  protein_coding         650
4      ENSG00000000460.17    C1orf112  protein_coding         246
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         340
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           7
60659   ENSG00000288675.1  AP006621.6  protein_coding          21

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1174             1126         28.3311           9.0669
1                   1                1          0.0757           0.0242
2                 937              964         88.0003          28.1629
3                 517              582          5.2765           1.6886
4                 382              367          2.3023           0.7368
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656             174              178         10.5131           3.3645
60657               0                0          0.0000           0.0000
60658               4                3          0.0407           0.0130
60659              36               39          0.6984           0.2235

       fpkm_uq_unstranded
0                  9.9279
1                  0.0265
2                 30.8373
3                  1.8490
4                  0.8068
...                   ...
60655              0.0000
60656              3.6840
60657              0.0000
60658              0.0142
60659              0.2447

[60660 rows x 9 columns], 'tw32':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding         667
1       ENSG00000000005.6        TNMD  protein_coding          63
2      ENSG00000000419.13        DPM1  protein_coding        1552
3      ENSG00000000457.14       SCYL3  protein_coding         626
4      ENSG00000000460.17    C1orf112  protein_coding         241
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           2
60656   ENSG00000288670.1  AL592295.6         lncRNA         119
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           5
60659   ENSG00000288675.1  AP006621.6  protein_coding          24

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 344              323         13.1364           4.2551
1                  29               34          3.8131           1.2351
2                 786              766        114.8699          37.2082
3                 554              540          8.1249           2.6318
4                 367              366          3.6063           1.1681
...               ...              ...             ...              ...
```

```
60655              1               1          0.0448          0.0145
60656             59              63          5.8832          1.9056
60657              0               0          0.0000          0.0000
60658              3               2          0.0464          0.0150
60659             23              21          1.2762          0.4134

       fpkm_uq_unstranded
0                  4.0978
1                  1.1895
2                 35.8331
3                  2.5345
4                  1.1250
...                   ...
60655              0.0140
60656              1.8352
60657              0.0000
60658              0.0145
60659              0.3981

[60660 rows x 9 columns], 'tw33':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        1074
1       ENSG00000000005.6        TNMD  protein_coding           0
2      ENSG00000000419.13        DPM1  protein_coding        1120
3      ENSG00000000457.14       SCYL3  protein_coding         742
4      ENSG00000000460.17    C1orf112  protein_coding         150
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6          lncRNA         180
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           5
60659   ENSG00000288675.1  AP006621.6  protein_coding          27

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 515              559         20.0443           6.2496
1                   0                0          0.0000           0.0000
2                 559              561         78.5542          24.4924
3                 529              530          9.1261           2.8454
4                 251              247          2.1270           0.6632
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              94               90          8.4328           2.6293
60657               0                0          0.0000           0.0000
60658               2                3          0.0440           0.0137
60659              17               22          1.3605           0.4242

       fpkm_uq_unstranded
0                  6.4370
1                  0.0000
2                 25.2268
3                  2.9307
4                  0.6831
...                   ...
60655              0.0000
60656              2.7081
60657              0.0000
60658              0.0141
60659              0.4369

[60660 rows x 9 columns], 'tw34':                   gene_id   gene_name       gene_ty
```

```
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding         783
1       ENSG00000000005.6        TNMD  protein_coding           1
2      ENSG00000000419.13        DPM1  protein_coding        1101
3      ENSG00000000457.14       SCYL3  protein_coding         246
4      ENSG00000000460.17    C1orf112  protein_coding         108
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA          58
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           2
60659   ENSG00000288675.1  AP006621.6  protein_coding          22

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                 388              395         11.8723           4.5641
1                   1                0          0.0466           0.0179
2                 577              524         62.7374          24.1182
3                 360              354          2.4581           0.9450
4                 293              289          1.2442           0.4783
...               ...              ...             ...              ...
60655               0                0          0.0000           0.0000
60656              24               37          2.2076           0.8487
60657               0                0          0.0000           0.0000
60658               0                2          0.0143           0.0055
60659              22               24          0.9007           0.3462

       fpkm_uq_unstranded
0                  5.5053
1                  0.0216
2                 29.0919
3                  1.1399
4                  0.5770
...                   ...
60655              0.0000
60656              1.0237
60657              0.0000
60658              0.0066
60659              0.4176

[60660 rows x 9 columns], 'tw35':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        3672
1       ENSG00000000005.6        TNMD  protein_coding           2
2      ENSG00000000419.13        DPM1  protein_coding        2267
3      ENSG00000000457.14       SCYL3  protein_coding        1056
4      ENSG00000000460.17    C1orf112  protein_coding         260
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           0
60656   ENSG00000288670.1  AL592295.6         lncRNA         246
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           5
60659   ENSG00000288675.1  AP006621.6  protein_coding          28

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1881             1791         42.6760          13.5734
1                   1                1          0.0714           0.0227
2                1156             1111         99.0144          31.4922
3                 901              770          8.0880           2.5724
4                 426              475          2.2959           0.7302
...               ...              ...             ...              ...
```

```
60655              0               0          0.0000           0.0000
60656            104             145          7.1768           2.2826
60657              0               0          0.0000           0.0000
60658              2               3          0.0274           0.0087
60659             30              23          0.8786           0.2795

       fpkm_uq_unstranded
0                 13.8696
1                  0.0232
2                 32.1795
3                  2.6286
4                  0.7462
...                   ...
60655              0.0000
60656              2.3324
60657              0.0000
60658              0.0089
60659              0.2856

[60660 rows x 9 columns], 'tw36':                   gene_id   gene_name       gene_ty
pe  unstranded  \
0      ENSG00000000003.15      TSPAN6  protein_coding        2250
1       ENSG00000000005.6        TNMD  protein_coding           8
2      ENSG00000000419.13        DPM1  protein_coding        1091
3      ENSG00000000457.14       SCYL3  protein_coding         459
4      ENSG00000000460.17    C1orf112  protein_coding         266
...                   ...         ...             ...         ...
60655   ENSG00000288669.1  AC008763.4  protein_coding           1
60656   ENSG00000288670.1  AL592295.6          lncRNA         312
60657   ENSG00000288671.1  AC006486.3  protein_coding           0
60658   ENSG00000288674.1  AL391628.1  protein_coding           6
60659   ENSG00000288675.1  AP006621.6  protein_coding          42

       stranded_first  stranded_second  tpm_unstranded  fpkm_unstranded  \
0                1157             1093         37.4860          11.2519
1                   4                4          0.4096           0.1229
2                 533              558         68.3088          20.5039
3                 444              417          5.0396           1.5127
4                 339              361          3.3672           1.0107
...               ...              ...             ...              ...
60655               0                1          0.0190           0.0057
60656             154              169         13.0483           3.9166
60657               0                0          0.0000           0.0000
60658               2                4          0.0471           0.0141
60659              32               44          1.8893           0.5671

       fpkm_uq_unstranded
0                 12.2209
1                  0.1335
2                 22.2695
3                  1.6430
4                  1.0977
...                   ...
60655              0.0062
60656              4.2539
60657              0.0000
60658              0.0154
60659              0.6159

[60660 rows x 9 columns]}
```

TMA

In [3]:
```
# Get all the file names in the current directory
tmfilenames = glob.glob(r"\gdc_download_20231025_081138.443578\Included Samples\TMA\*.

# Create a list to store the DataFrames
tmdataframes = []

# Iterate over the file names and create a DataFrame for each file
for i in range(len(tmfilenames)):
    # Create a DataFrame from the file
    df = pd.read_csv(tmfilenames[i], sep='\t')

    # Give the DataFrame a name
    tmdataframe_name = f"tm{i+1}"

    # Add the DataFrame to the list
    tmdataframes.append((tmdataframe_name, df))

# Create a dictionary to store the DataFrames with their respective names
tmdf = dict(tmdataframes)
```

DGE of 60,660 genes

DGE preprocessing

In [815…
```
ftwun = pd.DataFrame()
for i in range(1,37):
    new_column = pd.DataFrame(zip(twdf[f'tw{i}']['unstranded']),columns = [f'tw{i}'])
    ftwun = pd.concat([ftwun, new_column], axis=1)
```

In [816…
```
ftmun = pd.DataFrame()
for i in range(1,116):
    new_column = pd.DataFrame(zip(tmdf[f'tm{i}']['unstranded']),columns = [f'tm{i}'])
    ftmun = pd.concat([ftmun, new_column], axis=1)
```

In [817…
```
ftwcounts = pd.DataFrame(zip(tmdf['tm1']['gene_id']), columns = ['geneid']) # preparir
```

In [818…
```
ffcounts = pd.concat([ftwcounts,ftmun,ftwun], axis= 1) # preparing the counts table fc
```

In [819…
```
ffcounts = ffcounts.set_index('geneid')
```

In [820…
```
ffcounts = ffcounts[ffcounts.sum(axis=1)>0]
```

In [821…
```
ffcounts = ffcounts.T
```

In [120…
```
m_list = ['m']* 115
```

In [123…
```
w_list = ['w'] * 36
```

In [124…
```
condition_list = m_list + w_list
```

In [822…
```
metadata = pd.DataFrame(zip(counts.index, condition_list), columns = ['sample','condit
```

In [824…
```
metadata = metadata.set_index('sample')
```

GSEA of 60,660 genes

We can now perform differential gene expression analysis

In [825…
```
ffdds = DeseqDataSet(counts = ffcounts,
                     metadata = metadata,
                     design_factors = 'condition')
```

In [826…
```
ffdds.deseq2()
```

```
Fitting size factors...
... done in 0.28 seconds.

Fitting dispersions...
... done in 25.67 seconds.

Fitting dispersion trend curve...
... done in 29.45 seconds.

Fitting MAP dispersions...
... done in 25.07 seconds.

Fitting LFCs...
... done in 15.77 seconds.

Refitting 2626 outliers.

Fitting dispersions...
... done in 1.79 seconds.

Fitting MAP dispersions...
... done in 1.81 seconds.

Fitting LFCs...
... done in 1.46 seconds.
```

In [827…
```
ffstat_res = DeseqStats (ffdds, n_cpus = 10, contrast=('condition','m','w'))
```

In [828…
```
ffstat_res.summary()
```

```
Running Wald tests...
... done in 10.78 seconds.
```

```
Log2 fold change & Wald test p-value: condition m vs w
```

geneid	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000000003.15	1829.309026	0.021268	0.097091	0.219050	0.826611	0.968752
ENSG00000000005.6	9.455680	0.022183	0.530114	0.041846	0.966621	0.993462
ENSG00000000419.13	1420.295880	-0.032490	0.072979	-0.445189	0.656183	0.926687
ENSG00000000457.14	716.181056	-0.013263	0.087746	-0.151153	0.879855	0.978931
ENSG00000000460.17	224.097287	-0.102700	0.093493	-1.098477	0.271996	0.761485
...	...	...	...	...	...	...
ENSG00000288667.1	0.197151	0.042768	0.959904	0.044554	0.964462	NaN
ENSG00000288669.1	0.135754	-0.289547	1.025187	-0.282433	0.777611	NaN
ENSG00000288670.1	220.750806	-0.233423	0.101494	-2.299863	0.021456	0.406075
ENSG00000288674.1	5.373051	0.203953	0.197930	1.030432	0.302807	0.782173
ENSG00000288675.1	28.956439	-0.113175	0.154417	-0.732920	0.463607	0.861426

54901 rows × 6 columns

In [829…
```
ffres = ffstat_res.results_df
```

In [142…
```
mapper = id_map(species = 'human')
```

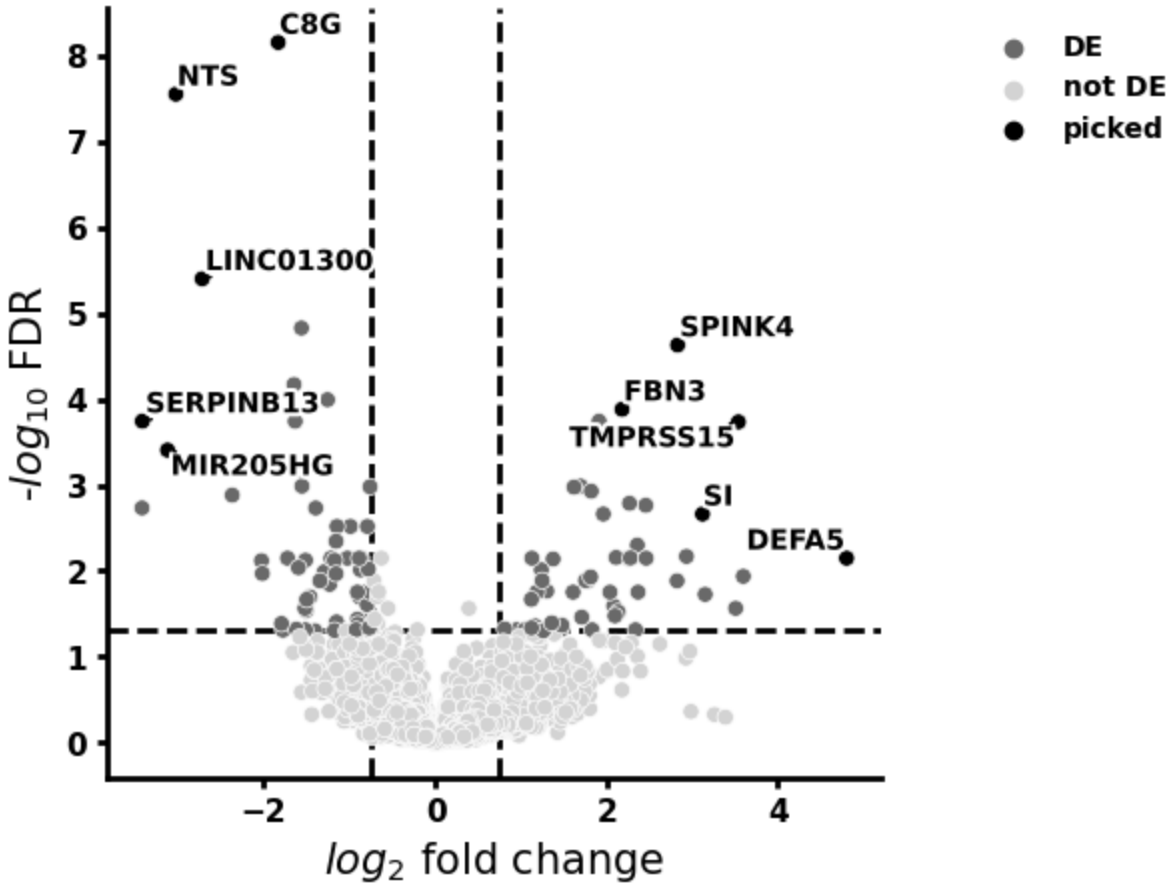
In [830…
```
# Split the column values into a list of strings
index = ffres.index.str.split('.', 1).str[0]

# Keep the first element of the list
ffres.index = index
```

In [831…
```
ffres['symbol'] = ffres.index.map(mapper.mapper)
```

DGE results

In [832…
```
fdegs = ffres[(abs(ffres.log2FoldChange) > 0.5)&(ffres.padj<0.05)]
```

In [833…
```
volcano(ffres, symbol = 'symbol')
```



In [837…
```
pfdegs = pd.DataFrame(fdegs)
pfdegs.to_csv("fdegs.csv", sep = ',')
```

In [112…
```
franking
```

Out[1120]:

geneid	symbol	stat
ENSG00000122711	SPINK4	5.854698
ENSG00000142449	FBN3	5.489955
ENSG00000070019	GUCY2C	5.381530
ENSG00000154646	TMPRSS15	5.363727
ENSG00000095627	TDRD1	4.991840
...	...	...
ENSG00000214711	CAPN14	-5.959830
ENSG00000253595	LINC01300	-6.207438
ENSG00000108786	HSD17B1	-6.327977
ENSG00000133636	NTS	-7.041705
ENSG00000176919	C8G	-7.325947

47528 rows × 2 columns

GSEA results

In [838…
```
franking = ffres[['symbol','stat']].dropna().sort_values('stat',ascending= False)
```

In [877…
```
prefres1 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout1 = []

for term in list(prefres1.results):
    fout1.append([term,
                prefres1.results[term]['fdr'],
                prefres1.results[term]['es'],
                prefres1.results[term]['nes'],
                prefres1.results[term]['pval'],
                 prefres1.results[term]['matched_genes']])

fout_df1 = pd.DataFrame(fout1, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df1
```

```
2023-10-28 21:16:22,313 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:16:22,356 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[877]:

	Term	fdr	es	nes	pval	
0	KEGG_JAK_STAT_SIGNALING_PATHWAY	0.000000	0.439582	1.698906	0.000000	
1	KEGG_HUNTINGTONS_DISEASE	0.000000	-0.479983	-1.879261	0.000000	DI
2	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.000000	0.408095	1.666393	0.000000	
3	KEGG_ALZHEIMERS_DISEASE	0.000000	-0.470835	-1.831722	0.000000	CHI
4	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.001542	0.376987	1.554841	0.000000	GRIA
5	KEGG_CHEMOKINE_SIGNALING_PATHWAY	0.011951	0.358540	1.422909	0.010040	CC
6	KEGG_CALCIUM_SIGNALING_PATHWAY	0.029557	0.331145	1.315231	0.018519	CHP2
7	KEGG_WNT_SIGNALING_PATHWAY	0.034235	0.342866	1.318097	0.030060	CH
8	KEGG_PATHWAYS_IN_CANCER	0.056398	0.301453	1.255140	0.016985	NC
9	KEGG_MAPK_SIGNALING_PATHWAY	0.108720	0.290806	1.189037	0.046218	CH
10	KEGG_FOCAL_ADHESION	0.198626	0.278984	1.107217	0.179592	RO
11	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.202190	0.277889	1.115900	0.156780	FGF
12	KEGG_OLFACTORY_TRANSDUCTION	0.260900	0.253118	1.067684	0.247357	CLCA
13	KEGG_PURINE_METABOLISM	0.568789	-0.264234	-1.006045	0.438095	GUC
14	KEGG_ENDOCYTOSIS	0.589043	-0.245664	-0.955090	0.572222	P

In [879…
```
prefres2 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout2 = []
```

```
for term in list(prefres2.results):
    fout2.append([term,
                  prefres2.results[term]['fdr'],
                  prefres2.results[term]['es'],
                  prefres2.results[term]['nes'],
                  prefres2.results[term]['pval'],
                   prefres2.results[term]['matched_genes']])

fout_df2 = pd.DataFrame(fout2, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df2
```

```
2023-10-28 21:16:50,888 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:16:50,918 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[879]:

	Term	fdr	es	nes	pval	mat
0	MIR3692_3P	0.000000	0.455758	1.886549	0.000000	ZNF780B;SOCS6;CASK;KIT;USP42;PUM1;ATG
1	MIR187_5P	0.000000	0.478932	1.877132	0.000000	SLC26A3;GATA5;BEND4;BRINP3;SLC4A7;GK5
2	MIR338_5P	0.000000	0.492003	2.125026	0.000000	WIF1;GDAP2;CD300LD-AS1;UNC13C;HORMAD
3	MIR221_3P	0.000000	0.478242	1.852269	0.000000	BEND4;SYBU;PHACTR4;KIT;ARHGEF38;HIPK
4	MIR205_5P	0.000000	0.545690	2.213377	0.000000	GABRA4;DSC2;NAA30;MAP3K13;TBX3;DMXL2;T
...	...	...	...	...	...	
749	MIR661	0.874765	-0.233782	-0.894893	0.736842	TOB2;CEACAM8;APOL6;CUX2;SCMH1;SHROOM4
750	MIR9500	0.876339	-0.233064	-0.889673	0.768642	GDAP2;MAGI1;DNAJB4;STOX2;ZNF106;GSPT
751	MIR6893_5P	0.895192	-0.224352	-0.872524	0.838115	MLEC;SLC4A7;ZNF592;PFKFB3;SCN4A;ZBTE
752	MIR940	0.904305	-0.225429	-0.874899	0.820408	MLEC;SLC4A7;ZNF592;PFKFB3;SCN4A;ZBTE
753	MIR6808_5P	0.930164	-0.218404	-0.849136	0.894094	MLEC;SLC4A7;ZNF592;PFKFB3;SCN4A;ZBTE

754 rows × 6 columns

In [880…

```
prefres3 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout3 = []

for term in list(prefres3.results):
    fout3.append([term,
                  prefres3.results[term]['fdr'],
                  prefres3.results[term]['es'],
                  prefres3.results[term]['nes'],
                  prefres3.results[term]['pval'],
                   prefres3.results[term]['matched_genes']])

fout_df3 = pd.DataFrame(fout3, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df3
```

```
2023-10-28 21:19:25,776 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:19:25,808 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[880]:

	Term	fdr	es	nes	pval	
0	PGM3_TARGET_GENES	0.000000	-0.457218	-1.859643	0.000000	KLF10;EIF5;MTMR12;KIF1B;THRA;G
1	MIER1_TARGET_GENES	0.000000	-0.461530	-1.925887	0.000000	ALB;TCP11L2;CNTNAP2;NPPB;UBR3
2	FREAC2_01	0.000000	0.417354	1.713671	0.000000	PTCHD1;MTTP;HIBADH;IRS4;MAB2
3	HNF1_Q6	0.000000	0.425302	1.742708	0.000000	TMPRSS15;SI;MTTP;ANXA13;CSF3;
4	MCM3_TARGET_GENES	0.000152	-0.422901	-1.698722	0.000000	COBL;MAP3K13;SPACA3;NKX6-3;ZI
...	...	...	...	...	...	
516	E2F1_Q4_01	0.763553	-0.227766	-0.913290	0.705224	ARHGAP36;CCNT2;DMD;STAG2;H2B(
517	ALPHACP1_01	0.847957	-0.218810	-0.887160	0.847195	ATOH1;MAP3K13;BRINP3;MYOCD;F(
518	HIF1_Q5	0.911616	0.214560	0.866737	0.879493	EPO;ZZZ3;KLF11;PPP1R3C;PHLPP
519	ZNF781_TARGET_GENES	0.951559	0.215069	0.844066	0.903766	AKAP9;TOP3B;MB;NAPSA;FABP5P3;
520	HIF1_Q3	0.953898	-0.212665	-0.838192	0.938462	ZZZ3;KLF11;PPP1R3C;TMTC1;FG

521 rows × 6 columns

In [881…

```
prefres4 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout4 = []

for term in list(prefres4.results):
    fout4.append([term,
                  prefres4.results[term]['fdr'],
                  prefres4.results[term]['es'],
                  prefres4.results[term]['nes'],
                  prefres4.results[term]['pval'],
                   prefres4.results[term]['matched_genes']])

fout_df4 = pd.DataFrame(fout4, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df4
```

```
2023-10-28 21:21:09,389 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:21:09,429 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[881]:

	Term	fdr	es	nes	pval	ma
0	MORF_PSMC1	0.000000	-0.501524	-1.945502	0.000000	EIF5;HSP90AA1;ZC3H14;BAG5;SETD3;YY1;P
1	MORF_DDB1	0.000000	-0.437089	-1.752514	0.000000	MLEC;TRRAP;AHCYL1;ZNF271P;HUWE1;SEC
2	MORF_DAP3	0.000000	-0.526755	-2.081573	0.000000	SETD3;AHCYL1;NIPSNAP2;HUWE1;EIF1AX;XF
3	MORF_AATF	0.000000	-0.427064	-1.698524	0.000000	ZZZ3;MLEC;ICE1;TRRAP;MTOR;HUWE1;EIF1
4	MORF_NME2	0.000000	-0.573933	-2.215518	0.000000	HSP90AA1;XBP1;PSMC1;HUWE1;EIF4G2;HSP
...	...	...	...	...	...	
81	GCM_MLL	0.497463	0.253914	0.993163	0.472574	ATG2B;ZNF644;THUMPD1;TRAPPC6B;PRKRA;
82	MORF_BCL2	0.509444	0.245052	0.986510	0.484277	PAX7;RBBP8;NOS2;ATP6V1B1;SLC6A11;AR
83	MORF_TNFRSF25	0.531306	-0.240800	-0.973220	0.552381	PAX7;RBBP8;NOS2;SOCS6;CYP2D6;SLC18A1
84	MORF_MT4	0.648992	0.230702	0.944813	0.644841	RBBP8;ROCK1;PNMT;ATP6V1B1;NEK9;FOXN
85	GCM_GSPT1	0.717528	0.234467	0.921969	0.704782	NAA30;ATG2B;RALGAPA1;ZNF644;PRKRA;D

86 rows × 6 columns

In [882…

```
prefres5 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout5 = []

for term in list(prefres5.results):
    fout5.append([term,
                prefres5.results[term]['fdr'],
                prefres5.results[term]['es'],
                prefres5.results[term]['nes'],
                prefres5.results[term]['pval'],
                 prefres5.results[term]['matched_genes']])

fout_df5 = pd.DataFrame(fout5, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df5
```

```
2023-10-28 21:21:30,907 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:21:30,943 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[882]:

	Term	fdr	es	nes	pva
0	GOBP_AEROBIC_RESPIRATION	0.000000	-0.528701	-2.062570	0.00000
1	GOBP_ELECTRON_TRANSPORT_CHAIN	0.000000	-0.478379	-1.849215	0.00000
2	GOBP_MITOCHONDRIAL_GENE_EXPRESSION	0.000000	-0.471769	-1.842552	0.00000
3	GOBP_CELLULAR_RESPIRATION	0.000000	-0.467001	-1.870414	0.00000
4	GOBP_EMBRYONIC_ORGAN_MORPHOGENESIS	0.000000	0.426873	1.770838	0.00000
...	...	...	...	...	
661	GOBP_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_MI...	0.969304	-0.216490	-0.839146	0.88223
662	GOBP_NEGATIVE_REGULATION_OF_WNT_SIGNALING_PATHWAY	0.970942	-0.216091	-0.834840	0.91907
663	GOBP_ALPHA_BETA_T_CELL_ACTIVATION	0.972769	-0.210429	-0.821531	0.93861
664	GOBP_TISSUE_REMODELING	0.981658	-0.209194	-0.804257	0.97307
665	GOBP_SPROUTING_ANGIOGENESIS	0.993659	0.202828	0.779381	0.97959

666 rows × 6 columns

In [883…

```
prefres6 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout6 = []

for term in list(prefres6.results):
    fout6.append([term,
                prefres6.results[term]['fdr'],
                prefres6.results[term]['es'],
                prefres6.results[term]['nes'],
                prefres6.results[term]['pval'],
                 prefres6.results[term]['matched_genes']])

fout_df6 = pd.DataFrame(fout6, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df6
```

```
2023-10-28 21:23:47,277 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:23:47,318 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[883]:

	Term	fdr	es	nes	pv
0	GOCC_AXONEME	0.000000	0.432339	1.666872	0.0000
1	GOCC_RIBOSOME	0.000000	-0.483104	-1.943242	0.0000
2	GOCC_INNER_MITOCHONDRIAL_MEMBRANE_PROTEIN_COMPLEX	0.000000	-0.637610	-2.438406	0.0000
3	GOCC_EXTERNAL_SIDE_OF_PLASMA_MEMBRANE	0.000000	0.445537	1.893317	0.0000
4	GOCC_RIBOSOMAL_SUBUNIT	0.000000	-0.528325	-2.099387	0.0000
...	...	...	...	...	
94	GOCC_ENDOPLASMIC_RETICULUM_LUMEN	0.781394	0.219461	0.916598	0.7987
95	GOCC_VACUOLAR_LUMEN	0.798682	-0.234012	-0.904881	0.7999
96	GOCC_CHROMOSOME_TELOMERIC_REGION	0.847155	-0.229311	-0.884346	0.7924
97	GOCC_MICROTUBULE_ASSOCIATED_COMPLEX	0.902633	0.226913	0.874333	0.8138
98	GOCC_SPINDLE_POLE	0.915731	0.221322	0.860389	0.8608

99 rows × 6 columns

In [884…

```
prefres7 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout7 = []

for term in list(prefres7.results):
    fout7.append([term,
                prefres7.results[term]['fdr'],
                prefres7.results[term]['es'],
                prefres7.results[term]['nes'],
                prefres7.results[term]['pval'],
                 prefres7.results[term]['matched_genes']])

fout_df7 = pd.DataFrame(fout7, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df7
```

```
2023-10-28 21:24:11,411 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:24:11,445 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[884]:

	Term	fdr	es	nes	pval
0	GOMF_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	0.000000	-0.516717	-2.013231	0.000000
1	GOMF_PRIMARY_ACTIVE_TRANSMEMBRANE_TRANSPORTER_...	0.001883	-0.413692	-1.588820	0.000000
2	GOMF_CARBOHYDRATE_BINDING	0.009846	0.365165	1.506738	0.000000
3	GOMF_DNA_BINDING_TRANSCRIPTION_ACTIVATOR_ACTIVITY	0.010798	0.358862	1.547459	0.000000
4	GOMF_SODIUM_ION_TRANSMEMBRANE_TRANSPORTER_ACTI...	0.013127	0.394263	1.507223	0.003929
...	...	...	...	...	...
89	GOMF_PHOSPHOPROTEIN_PHOSPHATASE_ACTIVITY	0.685275	-0.243144	-0.943397	0.618774
90	GOMF_CYTOKINE_ACTIVITY	0.690663	0.231495	0.937233	0.657732
91	GOMF_ISOMERASE_ACTIVITY	0.770842	-0.238317	-0.909602	0.736308
92	GOMF_MRNA_BINDING	0.779619	-0.212021	-0.913296	0.870968
93	GOMF_PHOSPHATIDYLINOSITOL_PHOSPHATE_BINDING	0.828809	0.227935	0.895104	0.768916

94 rows × 6 columns

In [885…

```
prefres8 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout8 = []

for term in list(prefres8.results):
    fout8.append([term,
                prefres8.results[term]['fdr'],
                prefres8.results[term]['es'],
                prefres8.results[term]['nes'],
                prefres8.results[term]['pval'],
                 prefres8.results[term]['matched_genes']])

fout_df8 = pd.DataFrame(fout8, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df8
```

```
2023-10-28 21:24:34,686 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:24:34,717 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[885]:

	Term	fdr	es	nes	pva
0	HP_INCREASED_SERUM_LACTATE	0.000000	-0.465984	-1.856808	0.00000(
1	HP_ABNORMALITY_OF_THE_MITOCHONDRION	0.000612	-0.433011	-1.724123	0.00000(
2	HP_LACTIC_ACIDOSIS	0.000816	-0.435132	-1.695912	0.00000(
3	HP_POOR_HEAD_CONTROL	0.016901	-0.385794	-1.523136	0.00000(
4	HP_HYPERTROPHIC_CARDIOMYOPATHY	0.020054	-0.371721	-1.525556	0.00000(
...	...	...	...	...	..
587	HP_ABNORMAL_METAPHYSIS_MORPHOLOGY	0.873918	-0.222314	-0.879865	0.86821
588	HP_GENU_VALGUM	0.881303	0.226229	0.877254	0.81262
589	HP_ABNORMAL_OVARIAN_MORPHOLOGY	0.925037	0.222275	0.857063	0.86587
590	HP_ABNORMAL_PERIPHERAL_NERVOUS_SYSTEM_PHYSIOLOGY	0.939802	0.219944	0.847280	0.87128
591	HP_DYSMETRIA	0.943652	-0.216367	-0.840987	0.91038

592 rows × 6 columns

In [886…

```
prefres9 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
fout9 = []

for term in list(prefres9.results):
    fout9.append([term,
                prefres9.results[term]['fdr'],
                prefres9.results[term]['es'],
                prefres9.results[term]['nes'],
                prefres9.results[term]['pval'],
                 prefres9.results[term]['matched_genes']])

fout_df9 = pd.DataFrame(fout9, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
fout_df9
```

```
2023-10-28 21:26:45,454 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:26:45,497 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[886]:

	Term	fdr	es	nes	pval	
0	TBK1.DF_DN	0.000000	0.436613	1.799579	0.000000	IMPACT;OSBPL1A;ZZZ3;AKAP9
1	MEK_UP.V1_DN	0.000000	0.442958	1.739286	0.000000	SYCP2;PRKAR2B;H2BC6;H2BC7;C
2	PRC2_EED_DN.V1_DN	0.000000	-0.474977	-1.862220	0.000000	BTBD7;TOP3B;FOSL2;CHM;DCLRE
3	PGF_UP.V1_UP	0.000000	0.532836	2.108426	0.000000	ZZZ3;AKAP9;PAPOLA;ZC3H14;T
4	CAMP_UP.V1_UP	0.000000	-0.453075	-1.781152	0.000000	ATP1A1;ETV5;AHCYL1;TGFBI;KA
...	...	...	...	...	...	
105	RPS14_DN.V1_DN	0.744680	-0.231364	-0.907662	0.746641	PCK1;AOC2;GAL;SHMT1;CA3;PS
106	SNF5_DN.V1_UP	0.801266	0.231405	0.900807	0.775258	MT3;GABRQ;ABI3BP;CXCR4;DDX
107	P53_DN.V1_UP	0.803770	0.229239	0.903875	0.731463	NKX2-1;DSC2;CAT;MMP1;CLDN3;
108	ESC_V6.5_UP_LATE.V1_DN	0.805734	0.226556	0.895515	0.795132	CTH;GBX2;MED29;GDF3;MORC1;ZN
109	KRAS.DF.V1_UP	0.853188	0.221438	0.875221	0.860082	CSF3;CAT;MMP1;DPYD;CCNT2;SL

110 rows × 6 columns

In [888…

```
prefres12 = gp.prerank (rnk= franking, gene_sets = r"C:\Users\Hasan\Desktop\Research\M
fout12 = []

for term in list(prefres12.results):
    fout12.append([term,
                prefres12.results[term]['fdr'],
                prefres12.results[term]['es'],
                prefres12.results[term]['nes'],
                prefres12.results[term]['pval'],
                 prefres12.results[term]['matched_genes']])

fout_df12 = pd.DataFrame(fout12, columns = ['Term','fdr', 'es', 'nes','pval','matched_
fout_df12
```

```
2023-10-28 21:41:12,682 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 21:41:12,713 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[888]:

	Term	fdr	es	nes	pval	
0	HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.000000	-0.525510	-2.050436	0.000000	CPT1A
1	HALLMARK_P53_PATHWAY	0.000481	-0.441726	-1.740925	0.000000	CCNK;A
2	HALLMARK_MYC_TARGETS_V1	0.022786	-0.355483	-1.408924	0.005725	EIF
3	HALLMARK_E2F_TARGETS	0.043326	-0.341172	-1.350569	0.009579	SHM
4	HALLMARK_KRAS_SIGNALING_DN	0.060331	0.355734	1.410849	0.002083	AC
5	HALLMARK_INFLAMMATORY_RESPONSE	0.081655	0.339728	1.344805	0.008457	RO
6	HALLMARK_APICAL_JUNCTION	0.092621	-0.321490	-1.275751	0.025926	ACTC
7	HALLMARK_ALLOGRAFT_REJECTION	0.114422	0.319661	1.269749	0.030738	N
8	HALLMARK_HEME_METABOLISM	0.131411	0.322447	1.281400	0.017316	S
9	HALLMARK_FATTY_ACID_METABOLISM	0.156237	0.320248	1.227141	0.054475	ADH
10	HALLMARK_KRAS_SIGNALING_UP	0.176388	0.296424	1.183908	0.103158	CIDEA
11	HALLMARK_MITOTIC_SPINDLE	0.180821	0.300029	1.196841	0.069959	CEP19
12	HALLMARK_COMPLEMENT	0.223512	0.289337	1.146399	0.120240	APC
13	HALLMARK_INTERFERON_GAMMA_RESPONSE	0.323293	0.273088	1.083918	0.242171	APO
14	HALLMARK_XENOBIOTIC_METABOLISM	0.340375	0.273575	1.091149	0.239351	ADH
15	HALLMARK_ESTROGEN_RESPONSE_EARLY	0.351326	0.264558	1.052380	0.281573	RBB
16	HALLMARK_IL2_STAT5_SIGNALING	0.361137	0.267860	1.059437	0.268750	IL1
17	HALLMARK_ESTROGEN_RESPONSE_LATE	0.400715	-0.268424	-1.060909	0.284333	RB
18	HALLMARK_APOPTOSIS	0.431867	-0.274222	-1.064862	0.265152	ROC
19	HALLMARK_UV_RESPONSE_UP	0.472853	-0.278697	-1.068326	0.286252	EIF5;G
20	HALLMARK_MYOGENESIS	0.488412	-0.274269	-1.081569	0.232604	ACTC1
21	HALLMARK_GLYCOLYSIS	0.496160	-0.279770	-1.098322	0.205273	LC
22	HALLMARK_G2M_CHECKPOINT	0.522098	0.248400	0.986416	0.492754	SLC12
23	HALLMARK_ADIPOGENESIS	0.652797	0.237568	0.934964	0.649789	G
24	HALLMARK_HYPOXIA	0.834656	-0.235306	-0.932640	0.642991	P
25	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.836076	-0.228011	-0.896133	0.740458	MMP3
26	HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.902781	-0.227409	-0.897044	0.766859	
27	HALLMARK_MTORC1_SIGNALING	0.945877	-0.210076	-0.826812	0.943888	CTH;

mirTarBase_2017

In [849…

```
prefresmir = gp.prerank (rnk= franking, gene_sets = 'miRTarBase_2017', min_size = 150)
fout1mir = []

for term in list(prefresmir.results):
    fout1mir.append([term,
```

```
                prefresmir.results[term]['fdr'],
                prefresmir.results[term]['es'],
                prefresmir.results[term]['nes'],
                prefresmir.results[term]['pval'],
                 prefresmir.results[term]['matched_genes']])

fout_df1mir = pd.DataFrame(fout1mir, columns = ['Term','fdr', 'es', 'nes','pval','matc
fout_df1mir
```

```
2023-10-28 18:54:29,561 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 18:54:29,619 [WARNING] Duplicated values found in preranked stats: 7.48% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

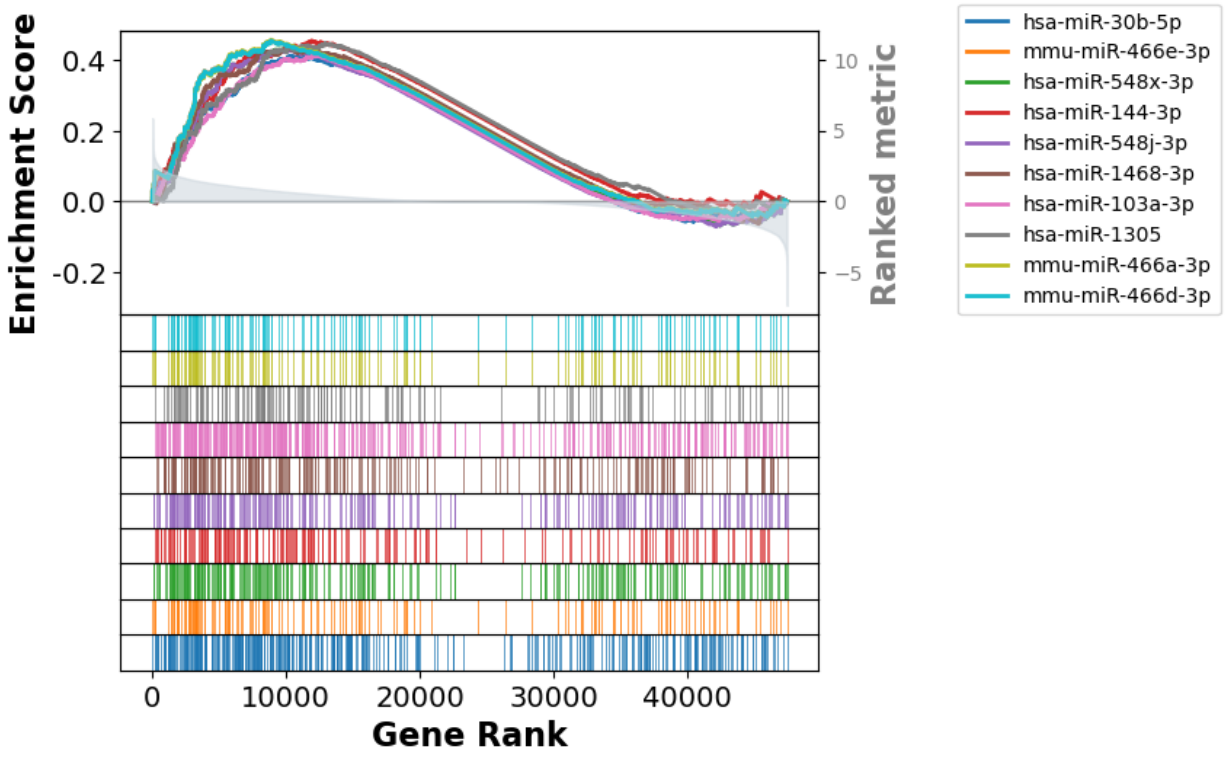
Out[849]:

	Term	fdr	es	nes	pval	matched_ge
0	hsa-miR-30b-5p	0.000000	0.410658	1.754697	0.000000	ACTC1;BTBD7;KLF10;CSNK1G1;CAT;CHAT;DDAH1;UHRF
1	mmu-miR-466e-3p	0.000000	0.455183	1.772769	0.000000	TDRD1;PADI2;TMEM241;TACR2;PIAS1;SLC16A10;ZNF4
2	hsa-miR-548x-3p	0.000000	0.432533	1.791567	0.000000	CLNK;CABLES1;EXD2;MYBPC1;COL4A3;HSPA14;PSAT1;
3	hsa-miR-144-3p	0.000000	0.454239	1.806754	0.000000	PURB;FGA;NAA30;ZCCHC2;CCNK;FGG;MLEC;IL20RB;AT
4	hsa-miR-548j-3p	0.000000	0.429322	1.774909	0.000000	CLNK;CABLES1;EXD2;MYBPC1;COL4A3;HSPA14;PSAT1;
...	...	...	...	...	...	
703	hsa-miR-3153	0.924510	-0.220060	-0.861241	0.876471	NXT2;ZBTB33;ANGEL1;LONRF3;ZNF460;RPH3A;ARSK;P
704	hsa-miR-6733-5p	0.924510	-0.220060	-0.861241	0.876471	NXT2;ZBTB33;ANGEL1;LONRF3;ZNF460;RPH3A;ARSK;P
705	hsa-miR-483-3p	0.937227	-0.217758	-0.847185	0.890335	TACR2;SPTA1;GRIN2B;SHISA9;ATP2B1;SMAD4;S1PR1;
706	hsa-miR-2276-3p	0.943847	-0.217624	-0.847872	0.897388	WWC1;NRAS;ERAP2;SEC14L3;ZNF267;YY1;INO80;STRN
707	hsa-miR-6739-5p	0.957247	-0.210087	-0.828106	0.936416	NR2E1;NXT2;ZBTB33;ANGEL1;LONRF3;ZNF460;RPH3A;

708 rows × 6 columns

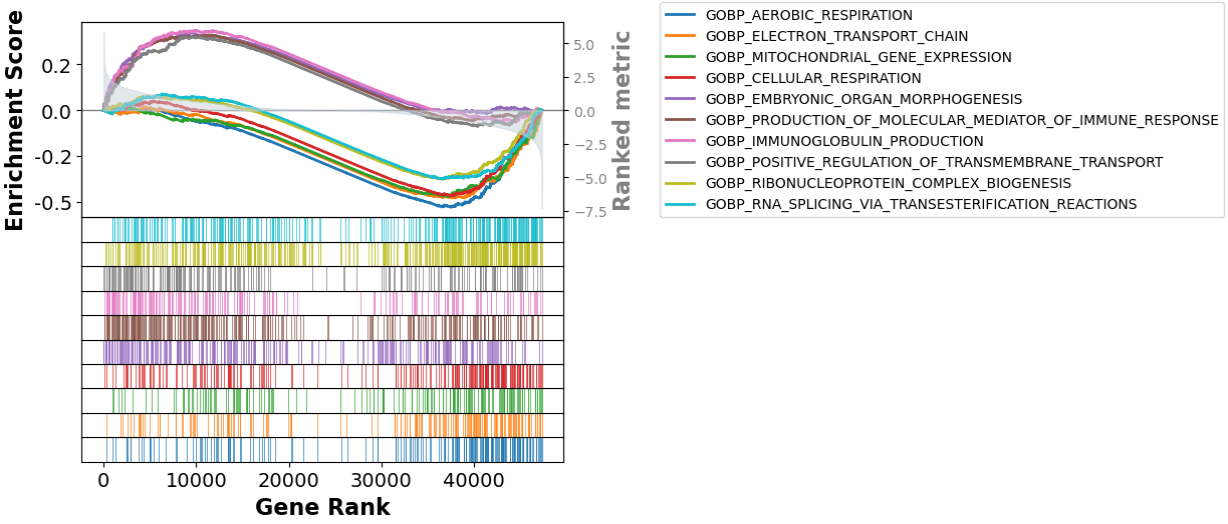
In [963…

```
axs = prefresmir.plot(terms=fout_df1mir.iloc[0:10].Term,
                      #legend_kws={'loc': (1.2, 0)}, # set the legend loc
                      show_ranking=True, # whether to show the second yaxis
                      figsize=(6,5),
                     legend_kws={'loc':(1.2,0)})
    #mirtarbase
```



In [964…

```
axs = prefres5.plot(terms=fout_df5.iloc[0:10].Term,
                    #legend_kws={'loc': (1.2, 0)}, # set the legend loc
                    show_ranking=True, # whether to show the second yaxis
                    figsize=(6,5),
                   legend_kws={'loc':(1.2,0)})
    #mirtarbase
```



lncRNA analyses

Data cleaning

TMA lncRNA

In [4]:
```
for i in range(1, 116):
    tmdf[f'tm{i}'] = tmdf[f'tm{i}'][tmdf[f'tm{i}']["gene_type"] == "lncRNA"]
```

TWM lncRNA

In [5]:
```
for i in range(1, 37):
    twdf[f'tw{i}'] = twdf[f'tw{i}'][twdf[f'tw{i}']["gene_type"] == "lncRNA"]
```

DGE set up

Get unstranded data from TWM and TMA

In [8]:
```
twun = pd.DataFrame()
for i in range(1,37):
    new_column = pd.DataFrame(zip(twdf[f'tw{i}']['unstranded']),columns = [f'tw{i}'])
    twun = pd.concat([twun, new_column], axis=1)
```

In [9]:
```
tmun = pd.DataFrame()
for i in range(1,116):
    new_column = pd.DataFrame(zip(tmdf[f'tm{i}']['unstranded']),columns = [f'tm{i}'])
    tmun = pd.concat([tmun, new_column], axis=1)
```

In [10]:
```
twcounts = pd.DataFrame(zip(tmdf['tm1']['gene_id']), columns = ['geneid']) # preparing
```

In [11]:
```
counts = pd.concat([twcounts,tmun,twun], axis= 1) # preparing the counts table for DGE
```

In [12]:
```
counts
```

Out[12]:

	geneid	tm1	tm2	tm3	tm4	tm5	tm6	tm7	tm8	tm9	...	tw27	tw28	tw29
0	ENSG00000082929.8	1	2	0	0	6	6	3	2	3	...	19	0	10
1	ENSG00000083622.8	3	14	11	4	12	4	0	8	6	...	1	19	1
2	ENSG00000093100.13	1	1	8	4	10	4	11	5	2	...	0	6	5
3	ENSG00000099869.8	41	9	14	14	6	14	260	11	13	...	76	24	41
4	ENSG00000103472.10	12	8	22	13	62	5	10	38	25	...	13	12	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16896	ENSG00000288662.1	0	0	0	0	0	0	0	0	0	...	0	0	0
16897	ENSG00000288663.1	18	16	23	24	39	20	31	39	18	...	13	31	38
16898	ENSG00000288665.1	0	0	0	0	0	0	0	0	0	...	0	0	0
16899	ENSG00000288667.1	0	0	0	0	1	0	0	0	0	...	0	0	1
16900	ENSG00000288670.1	144	186	377	285	112	236	211	139	118	...	107	396	533

16901 rows × 152 columns

Counts table and metadata

In [13]:
```
counts = counts.set_index('geneid')
```

In [14]:
```
counts = counts[counts.sum(axis=1)>0]
```

In [15]:
```
# Split the column values into a list of strings
index = counts.index.str.split('.', 1).str[0]

# Keep the first element of the list
counts.index = index
```

In [17]:
```
counts = counts.T
```

In [18]:
```
m_list = ['m']* 115
```

In [19]:
```
w_list = ['w'] * 36
```

In [20]:
```
condition_list = m_list + w_list
```

In [21]:
```
metadata = pd.DataFrame(zip(counts.index, condition_list), columns = ['sample','condit
```

In [22]:
```
metadata = metadata.set_index('sample')
```

We can now perform differential gene expression analysis

In [23]:
```
dds = DeseqDataSet(counts = counts,
                   metadata = metadata,
                   design_factors = 'condition')
```

In [24]:
```
dds.deseq2()
```

```
Fitting size factors...
... done in 0.08 seconds.

Fitting dispersions...
... done in 2.22 seconds.

Fitting dispersion trend curve...
... done in 2.57 seconds.

Fitting MAP dispersions...
... done in 2.56 seconds.

Fitting LFCs...
... done in 2.09 seconds.

Refitting 779 outliers.

Fitting dispersions...
... done in 0.57 seconds.

Fitting MAP dispersions...
... done in 0.60 seconds.

Fitting LFCs...
... done in 0.50 seconds.
```

In [25]:
```
stat_res = DeseqStats (dds, n_cpus = 10, contrast=('condition','m','w'))
```

In [26]:
```
stat_res.summary()
```

```
Running Wald tests...
Log2 fold change & Wald test p-value: condition m vs w
... done in 2.26 seconds.
```

geneid	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000082929	3.618199	0.236171	0.327031	0.722166	0.470192	0.852272
ENSG00000083622	7.743589	0.348906	0.344355	1.013215	0.310957	0.761001
ENSG00000093100	5.060219	0.187632	0.212014	0.884997	0.376158	0.798384
ENSG00000099869	30.814672	-0.235906	0.291912	-0.808141	0.419009	0.827274
ENSG00000103472	11.580722	-0.106233	0.179790	-0.590877	0.554603	0.887063
...	...	...	...	...	...	...
ENSG00000288659	0.131184	-0.222847	1.730939	-0.128743	0.897561	NaN
ENSG00000288662	0.140537	0.163917	1.191795	0.137538	0.890606	NaN
ENSG00000288663	22.018264	-0.018001	0.138431	-0.130035	0.896539	0.978015
ENSG00000288667	0.185237	0.084151	0.960537	0.087608	0.930188	NaN
ENSG00000288670	218.854701	-0.197753	0.105866	-1.867966	0.061767	0.446154

15879 rows × 6 columns

In [27]:
```
res = stat_res.results_df
```

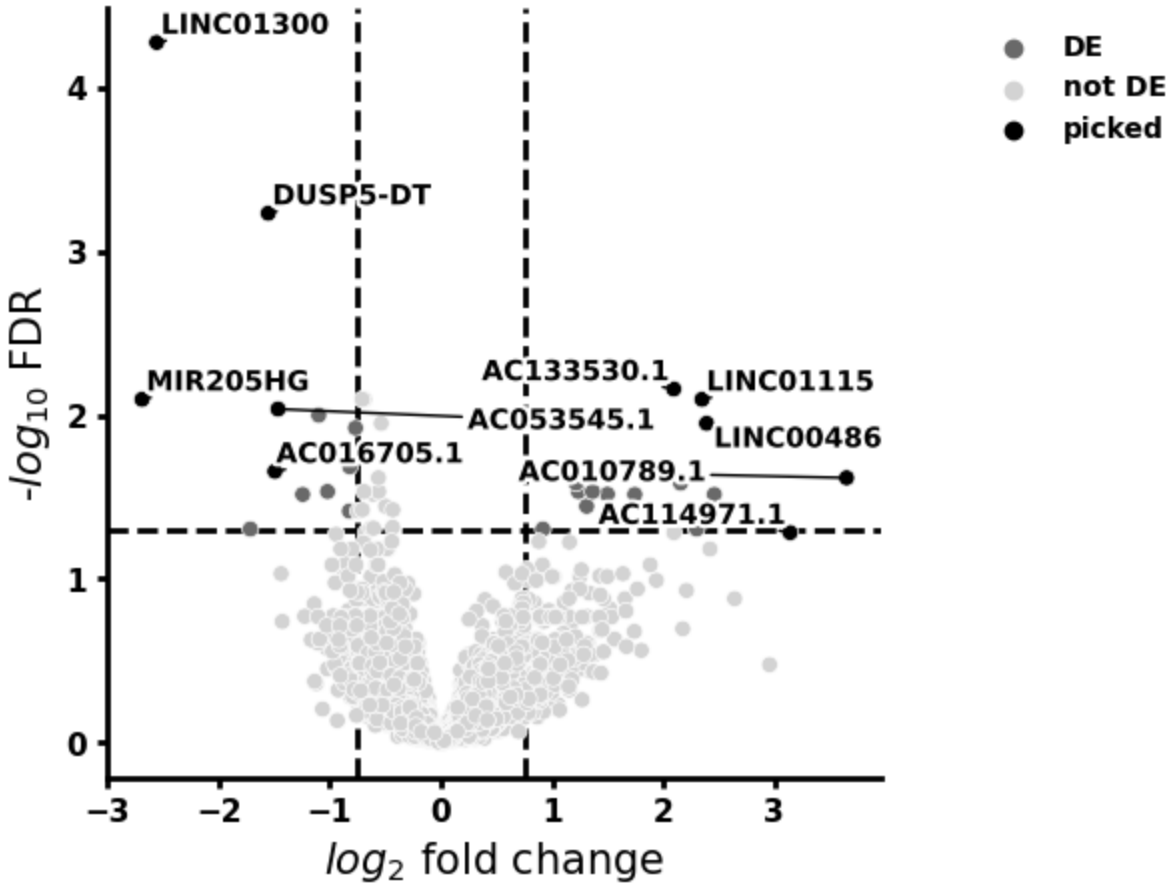
In [28]:
```
mapper = id_map(species = 'human')
```

In [29]:
```
res['symbol'] = res.index.map(mapper.mapper)
```

DGEA results

In [30]:
```
degs = res[(abs(res.log2FoldChange) > 0.5)&(res.padj<0.05)]
```

In [31]:
```
volcano(res, symbol = 'symbol')
```



In [153…
```
pdegs =pd.DataFrame(degs)
```

In [241…
```
pdegs.to_csv('DEGs.csv',sep=',')
```

Correlations

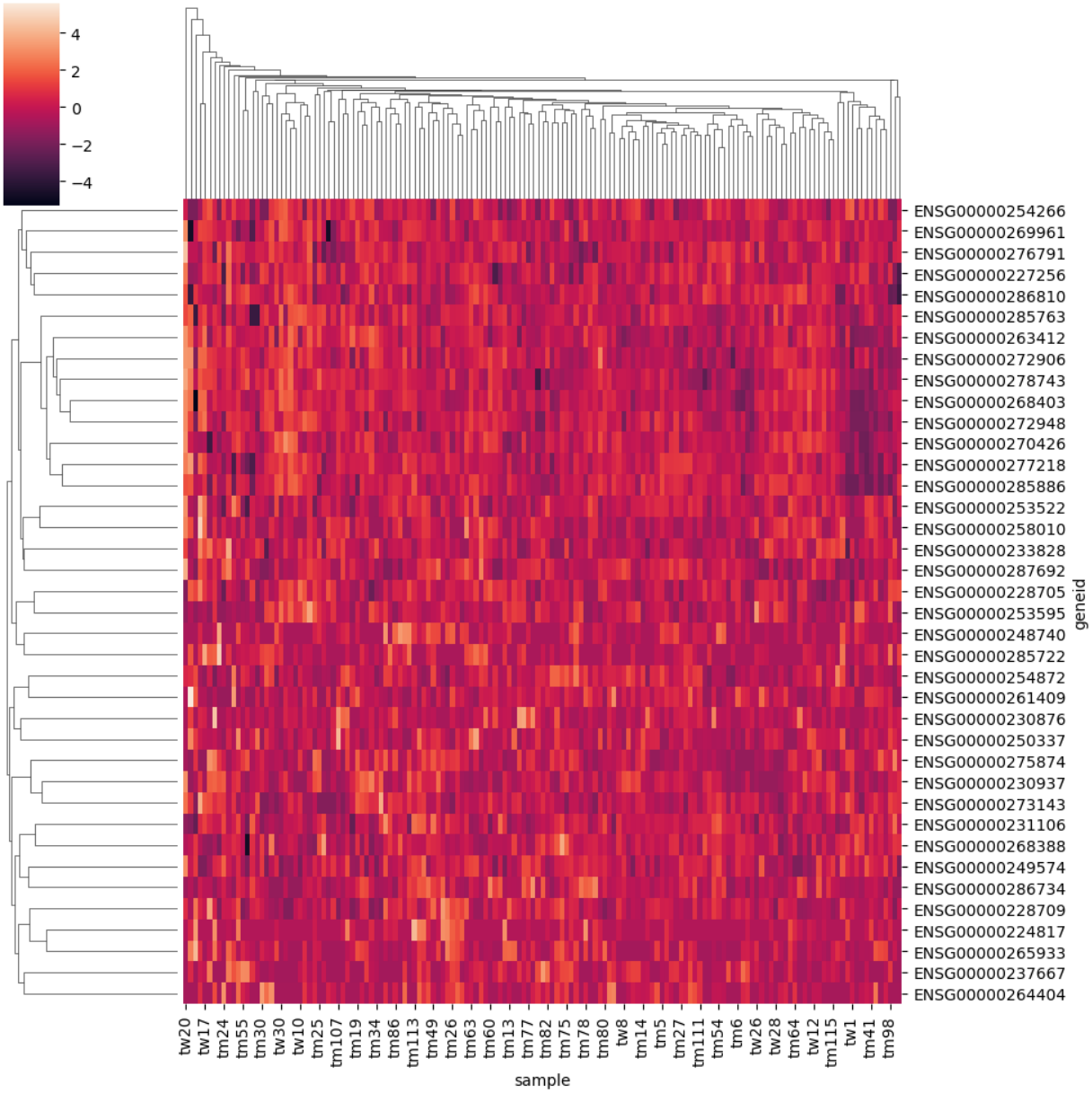
In [32]:
```
dds.layers['log1p']= np.log1p(dds.layers['normed_counts'])
```

In [34]:
```
ddsigs= dds[:,degs.index]
```

In [39]:
```
cluster = pd.DataFrame(ddsigs.layers['log1p'].T,
                       index=ddsigs.var_names,
                       columns=ddsigs.obs_names)
```

In [47]:
```
sns.clustermap(cluster, z_score=0, color= 'red')
```

Out[47]:
```
<seaborn.matrix.ClusterGrid at 0x175d50e1550>
```



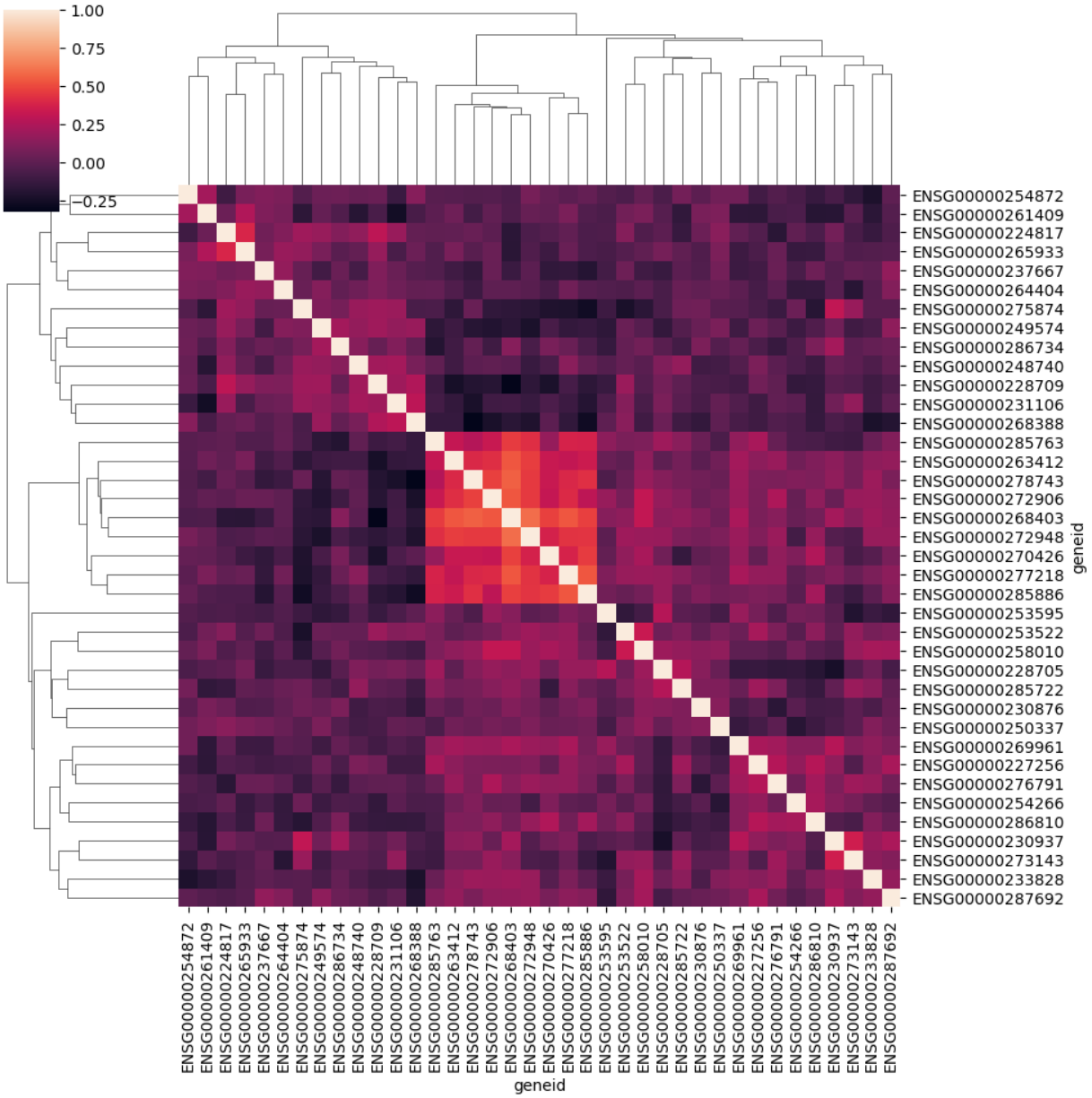
In [48]:
```
heat= cluster.T
```

In [49]:
```
correlation = heat.corr()

# Create a new dataframe with the gene names as the index and columns
heatmap_df = pd.DataFrame(correlation, index=cluster.index, columns=cluster.index)
```

In [50]:
```
correlation.to_csv('correlations.csv', sep=',')
```

In [51]:
```
sns.clustermap(correlation)
```

Out[51]:
```
<seaborn.matrix.ClusterGrid at 0x175d51c43d0>
```



We therefore proceed to gene set enrichment analysis

Set up

In [157…
```
ranking = res[['symbol','stat']].dropna().sort_values('stat',ascending= False)
```

In [876…
```
ranking
```

Out[876]:

geneid	symbol	stat
ENSG00000286734	AC133530.1	4.689599
ENSG00000237667	LINC01115	4.509832
ENSG00000230876	LINC00486	4.312419
ENSG00000224817	AC010789.1	4.045730
ENSG00000249574	AC226118.1	3.998806
...	...	...
ENSG00000268403	AC132192.2	-4.482140
ENSG00000230937	MIR205HG	-4.509800
ENSG00000286810	AL513128.3	-4.562037
ENSG00000273143	DUSP5-DT	-5.248009
ENSG00000253595	LINC01300	-5.792289

15568 rows × 2 columns

GSEA

In [904…

```
pre_res1 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out1 = []

for term in list(pre_res1.results):
    out1.append([term,
                 pre_res1.results[term]['fdr'],
                 pre_res1.results[term]['es'],
                 pre_res1.results[term]['nes'],
                 pre_res1.results[term]['pval'],
                  pre_res1.results[term]['matched_genes']])

out_df1 = pd.DataFrame(out1, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df1
```

```
2023-10-28 22:44:08,790 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:44:08,809 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
2023-10-28 22:44:08,831 [ERROR] No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.

2023-10-28 22:44:08,833 [ERROR] The first entry of your gene_sets (gmt) look like thi
s : { KEGG_N_GLYCAN_BIOSYNTHESIS: [ALG13, DOLPP1, RPN1, ALG14, MAN1B1, ALG3, B4GALT1,
MGAT5, RPN2, STT3A, MGAT3, DAD1, MGAT2, ALG12, TUSC3, MAN1C1, DPM2, DPM1, GANAB, ALG
1, MGAT4A, ALG10B, STT3B, MAN1A2, ALG10, ALG11, ALG8, ALG2, DPAGT1, RFT1, DPM3, DDOS
T, MGAT4B, ALG6, MAN2A2, MAN1A1, MAN2A1, ST6GAL1, B4GALT3, ALG5, B4GALT2, MGAT5B, ALG
9, MOGS, FUT8, MGAT1]}
2023-10-28 22:44:08,835 [ERROR] The first 5 genes look like this : [ AC133530.1, LINC
01115, LINC00486, AC010789.1, AC226118.1 ]
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
Cell In[904], line 1
----> 1 pre_res1 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAPK in NSCLC\Gene Sets\msigdb_v2023.1.Hs_GMTs\MSigDB\c2.cp.kegg.v2023.1.Hs.symbols.gmt", min_size = 3)
      2 out1 = []
      4 for term in list(pre_res1.results):

File ~\anaconda3\Lib\site-packages\gseapy\__init__.py:377, in prerank(rnk, gene_sets, outdir, pheno_pos, pheno_neg, min_size, max_size, permutation_num, weighted_score_type, ascending, threads, figsize, format, graph_num, no_plot, seed, verbose, *arg, **kwarg)
    357     threads = kwarg["processes"]
    358 pre = Prerank(
    359     rnk,
    360     gene_sets,
   (...)
    375     verbose,
    376 )
--> 377 pre.run()
    378 return pre

File ~\anaconda3\Lib\site-packages\gseapy\gsea.py:435, in Prerank.run(self)
    433 self._logger.info("Parsing data files for GSEA.............................")
    434 # filtering out gene sets and build gene sets dictionary
--> 435 gmt = self.load_gmt(gene_list=dat2.index.values, gmt=self.gene_sets)
    436 self.gmt = gmt
    437 self._logger.info(
    438     "%04d gene_sets used for further statistical testing....." % len(gmt)
    439 )

File ~\anaconda3\Lib\site-packages\gseapy\base.py:302, in GSEAbase.load_gmt(self, gene_list, gmt)
    294     self._logger.error(
    295         "The first entry of your gene_sets (gmt) look like this : %s"
    296         % dict_head
    297     )
    298     self._logger.error(
    299         "The first 5 genes look like this : [ %s ]"
    300         % (", ".join(list(gene_list)[:5]))
    301     )
--> 302     raise LookupError(msg)
    304 # self._gmtdct = genesets_dict
    305 return genesets_dict

LookupError: No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.
```

In [908…

```
pre_res3 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out3 = []

for term in list(pre_res3.results):
    out3.append([term,
                 pre_res3.results[term]['fdr'],
                 pre_res3.results[term]['es'],
                 pre_res3.results[term]['nes'],
                 pre_res3.results[term]['pval'],
```

```
                pre_res3.results[term]['matched_genes']])

out_df3 = pd.DataFrame(out3, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df3
```

```
2023-10-28 22:45:43,684 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:45:43,700 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[908]:

	Term	fdr	es	nes	pval	mat
0	MIR4283	0.644746	0.597674	0.910745	0.653430	LINC02908;LINC0295
1	MIR548AJ_5P_MIR548G_5P_MIR548X_5P	0.703898	0.654094	1.033505	0.479401	LINC02693;DOCK8
2	MIR548F_5P	0.703898	0.654094	1.033505	0.479401	LINC02693;DOCK8
3	MIR765	0.704791	-0.872317	-1.369573	0.090526	AS1;LMO7DN
4	MIR3662	0.985896	-0.288575	-0.453026	0.993952	LINC02693;GOLGA8
5	MIR373_5P	1.000000	-0.526330	-0.842586	0.676087	COLCA1;LINC0
6	MIR371B_5P	1.000000	-0.526330	-0.842586	0.676087	COLCA1;LINC0
7	MIR616_5P	1.000000	-0.526330	-0.842586	0.676087	COLCA1;LINC0
8	MIR153_5P	1.000000	-0.346589	-0.623366	0.912946	PELAT AS1;C15c DT;CC
9	MIR6844	1.000000	-0.456804	-0.718078	0.797665	LINC01517;LINC0290
10	MIR543	1.000000	-0.451660	-0.705620	0.812371	LINC02693;GOLGA8N

In [937…

```
out_df13.to_csv('mirtar.csv', sep=',')
```

In [924…

```
pre_res4 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out4 = []

for term in list(pre_res4.results):
    out1.append([term,
                pre_res4.results[term]['fdr'],
                pre_res4.results[term]['es'],
                pre_res4.results[term]['nes'],
                pre_res4.results[term]['pval'],
                 pre_res4.results[term]['matched_genes']])

out_df4 = pd.DataFrame(out1, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df4
```

```
2023-10-28 23:29:13,893 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 23:29:13,906 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[924]:

	Term	fdr	es	nes	pval	matched_genes
0	TOX4_TARGET_GENES	0.011571	-0.708107	-1.929676	0.002137	WEE2-AS1;MACC1-DT;ANXA2R-OT1;SMG7-AS1;DMXL1-DT...
1	ELF2_TARGET_GENES	0.015910	-0.533303	-1.936616	0.000000	MED28-DT;VPS33B-DT;DYNLRB2-AS1;LINC00339;WEE2-...
2	CEBPZ_TARGET_GENES	0.018803	-0.528553	-1.958051	0.000000	CD101-AS1;SRP14-DT;PEF1-AS1;RHPN1-AS1;LINC0020...
3	PRKDC_TARGET_GENES	0.030374	-0.524052	-1.813057	0.000000	ID2-AS1;RHPN1-AS1;ZFHX3-AS1;FGD5-AS1;ITFG2-AS1...
4	DBP_TARGET_GENES	0.033990	-0.495382	-1.846017	0.000000	PLBD1-AS1;IQCH-AS1;UBE2H-DT;UBL7-DT;EXOSC10-AS...
...	...	...	...	...	...	...
960	NKX25_01	1.000000	0.482623	0.793190	0.751371	LINC02880;LINC00649;NEXN-AS1;MIR9-1HG
961	NKX6_1_TARGET_GENES	1.000000	0.547742	0.833776	0.748120	LINC02136;LINC02251;LINC02564
962	PAX3_TARGET_GENES	1.000000	0.366568	1.393499	0.003317	FENDRR;POU6F2-AS2;CD300LD-AS1;ALDH1A3-AS1;LINC...
963	FBXL19-AS1-ASO_G0260852_02-DEGs Down	1.000000	0.888320	1.285561	0.114723	FSIP2-AS2;FBXL19-AS1
964	ZNF586_TARGET_GENES	1.000000	0.512544	1.058944	0.407692	LINC00926;WEE2-AS1;LOXL1-AS1;LINC02930;LINC029...

965 rows × 6 columns

In [911…

```
pre_res6 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out6 = []

for term in list(pre_res6.results):
    out6.append([term,
                 pre_res6.results[term]['fdr'],
                 pre_res6.results[term]['es'],
                 pre_res6.results[term]['nes'],
                 pre_res6.results[term]['pval'],
                  pre_res6.results[term]['matched_genes']])

out_df6 = pd.DataFrame(out6, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df6
```

```
2023-10-28 22:47:30,903 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:47:30,917 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[911]:

	Term	fdr	es	nes	pval
0	GOBP_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	0.256678	0.853607	1.338241	0.083643
1	GOBP_SPLICEOSOMAL_COMPLEX_ASSEMBLY	0.256678	0.853607	1.338241	0.083643
2	GOBP_RNA_SPLICING_VIA_TRANSESTERIFICATION_REAC...	0.256678	0.853607	1.338241	0.083643
3	GOBP_RIBONUCLEOPROTEIN_COMPLEX_SUBUNIT_ORGANIZ...	0.256678	0.853607	1.338241	0.083643
4	GOBP_MRNA_METABOLIC_PROCESS	0.258597	0.872429	1.455084	0.023985
...	...	...	...	...	...
67	GOBP_PROTEIN_CONTAINING_COMPLEX_ORGANIZATION	1.000000	-0.510659	-1.086048	0.367713
68	GOBP_RESPONSE_TO_ENDOGENOUS_STIMULUS	1.000000	-0.530254	-0.861010	0.662366
69	GOBP_RNA_PROCESSING	1.000000	-0.375356	-0.912724	0.557173
70	GOBP_INFLAMMATORY_RESPONSE	1.000000	-0.565338	-0.910632	0.605664
71	GOBP_POST_TRANSCRIPTIONAL_REGULATION_OF_GENE_E...	1.000000	-0.370725	-1.026099	0.392157

72 rows × 6 columns

In [912...

```
pre_res7 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out7= []

for term in list(pre_res7.results):
    out7.append([term,
                 pre_res7.results[term]['fdr'],
                 pre_res7.results[term]['es'],
                 pre_res7.results[term]['nes'],
                 pre_res7.results[term]['pval'],
                  pre_res7.results[term]['matched_genes']])

out_df7 = pd.DataFrame(out7, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df7
```

```
2023-10-28 22:47:43,229 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:47:43,232 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[912]:

	Term	fdr	es	nes	pval
0	GOCC_NUCLEAR_PROTEIN_CONTAINING_COMPLEX	0.148445	0.753320	1.322757	0.112963
1	GOCC_SM_LIKE_PROTEIN_FAMILY_COMPLEX	0.161614	0.853607	1.338241	0.083643
2	GOCC_SPLICEOSOMAL_SNRNP_COMPLEX	0.161614	0.853607	1.338241	0.083643
3	GOCC_ORGANELLE_SUBCOMPARTMENT	0.249223	0.832177	1.366794	0.062500
4	GOCC_NUCLEAR_OUTER_MEMBRANE_ENDOPLASMIC_RETICU...	0.268812	0.929641	1.427170	0.023346
5	GOCC_NUCLEAR_BODY	0.298432	-0.812710	-1.278934	0.161863
6	GOCC_RIBONUCLEOPROTEIN_COMPLEX	0.318786	-0.434662	-1.211342	0.201342
7	GOCC_SIGNAL_RECOGNITION_PARTICLE	0.368624	-0.761728	-1.315338	0.169565
8	GOCC_NUCLEOLUS	0.374461	-0.616052	-1.416767	0.102510
9	GOCC_RNAI_EFFECTOR_COMPLEX	0.435524	-0.422342	-1.059002	0.368539
10	GOCC_CHROMOSOME	0.711132	0.507067	0.842453	0.726606
11	GOCC_SEX_CHROMOSOME	0.718102	0.601323	0.920622	0.634981

In [929…

```
pre_res8 = gp.prerank (rnk= ranking, gene_sets = "GO_Molecular_Function_2023", min_siz
out8 = []

for term in list(pre_res8.results):
    out8.append([term,
                 pre_res8.results[term]['fdr'],
                 pre_res8.results[term]['es'],
                 pre_res8.results[term]['nes'],
                 pre_res8.results[term]['pval'],
                  pre_res8.results[term]['matched_genes']])

out_df8 = pd.DataFrame(out8, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df8
```

```
2023-10-28 23:38:33,896 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 23:38:33,910 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
2023-10-28 23:38:34,154 [ERROR] No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.

2023-10-28 23:38:34,155 [ERROR] The first entry of your gene_sets (gmt) look like thi
s : { 1-Acylglycerol-3-Phosphate O-acyltransferase Activity (GO:0003841): [LPGAT1, MB
OAT7, TAFAZZIN, LPCAT2, MBOAT2, LPCAT1, AGPAT1, AGPAT2, CRLS1, LCLAT1, AGPAT3, AGPAT
4, AGPAT5, MBOAT1, LPCAT3, GPAT2, GPAT4, GPAT3, PNPLA3]}
2023-10-28 23:38:34,157 [ERROR] The first 5 genes look like this : [ AC133530.1, LINC
01115, LINC00486, AC010789.1, AC226118.1 ]
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
Cell In[929], line 1
----> 1 pre_res8 = gp.prerank (rnk= ranking, gene_sets = "GO_Molecular_Function_202
3", min_size = 3)
      2 out8 = []
      4 for term in list(pre_res8.results):

File ~\anaconda3\Lib\site-packages\gseapy\__init__.py:377, in prerank(rnk, gene_sets,
outdir, pheno_pos, pheno_neg, min_size, max_size, permutation_num, weighted_score_typ
e, ascending, threads, figsize, format, graph_num, no_plot, seed, verbose, *arg, **kw
arg)
    357     threads = kwarg["processes"]
    358 pre = Prerank(
    359     rnk,
    360     gene_sets,
   (...)
    375     verbose,
    376 )
--> 377 pre.run()
    378 return pre

File ~\anaconda3\Lib\site-packages\gseapy\gsea.py:435, in Prerank.run(self)
    433 self._logger.info("Parsing data files for GSEA.............................")
    434 # filtering out gene sets and build gene sets dictionary
--> 435 gmt = self.load_gmt(gene_list=dat2.index.values, gmt=self.gene_sets)
    436 self.gmt = gmt
    437 self._logger.info(
    438     "%04d gene_sets used for further statistical testing....." % len(gmt)
    439 )

File ~\anaconda3\Lib\site-packages\gseapy\base.py:302, in GSEAbase.load_gmt(self, gen
e_list, gmt)
    294     self._logger.error(
    295         "The first entry of your gene_sets (gmt) look like this : %s"
    296         % dict_head
    297     )
    298     self._logger.error(
    299         "The first 5 genes look like this : [ %s ]"
    300         % (", ".join(list(gene_list)[:5]))
    301     )
--> 302     raise LookupError(msg)
    304 # self._gmtdct = genesets_dict
    305 return genesets_dict

LookupError: No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.
```

In [914…

```
pre_res9 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAP
out9= []

for term in list(pre_res9.results):
    out9.append([term,
                 pre_res9.results[term]['fdr'],
                 pre_res9.results[term]['es'],
                 pre_res9.results[term]['nes'],
                 pre_res9.results[term]['pval'],
                  pre_res9.results[term]['matched_genes']])
```

```
out_df9 = pd.DataFrame(out9, columns = ['Term','fdr', 'es', 'nes','pval','matched_gene
out_df9
```

```
2023-10-28 22:47:47,352 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:47:47,364 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[914]:

	Term	fdr	es	nes	pval
0	HP_CLINODACTYLY	0.578001	0.746706	1.152056	0.275862
1	HP_FEEDING_DIFFICULTIES	0.578001	0.746706	1.152056	0.275862
2	HP_DEVIATION_OF_THE_HAND_OR_OF_FINGERS_OF_THE_...	0.578001	0.746706	1.152056	0.275862
3	HP_FEEDING_DIFFICULTIES_IN_INFANCY	0.578001	0.746706	1.152056	0.275862
4	HP_MOTOR_DELAY	0.578001	0.746706	1.152056	0.275862
...	...	...	...	...	...
119	HP_CONSTITUTIONAL_SYMPTOM	1.000000	-0.495708	-0.836841	0.685466
120	HP_UNUSUAL_INFECTION	1.000000	-0.381058	-0.646507	0.864583
121	HP_ABNORMAL_PALATE_MORPHOLOGY	1.000000	0.868847	1.345383	0.062500
122	HP_ABNORMALITY_OF_THE_LOWER_URINARY_TRACT	1.000000	-0.784176	-1.251180	0.222717
123	HP_ABNORMAL_RENAL_MORPHOLOGY	1.000000	-0.356106	-0.562522	0.919679

124 rows × 6 columns

In [921…

```
pre_res10 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
out10 = []

for term in list(pre_res10.results):
    out10.append([term,
                pre_res10.results[term]['fdr'],
                pre_res10.results[term]['es'],
                pre_res10.results[term]['nes'],
                pre_res10.results[term]['pval'],
                 pre_res10.results[term]['matched_genes']])

out_df10 = pd.DataFrame(out10, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
out_df10
```

```
2023-10-28 22:57:20,938 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:57:20,950 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[921]:

	Term	fdr	es	nes	pval	
0	E2F3_UP.V1_DN	0.132722	-0.650737	-1.599925	0.032680	LINC00842;WT1-AS;LINC00662;R
1	IL2_UP.V1_DN	0.139238	0.930986	1.513096	0.003968	LINC00472;LINC0155
2	STK33_SKM_DN	0.149509	0.764810	1.565444	0.013208	CCDC26;LINC019
3	RAPA_EARLY_UP.V1_UP	0.315424	-0.800085	-1.461569	0.068966	ERVK-28;LINC00563;RETREG1-
4	SRC_UP.V1_UP	0.348044	-0.600672	-1.401455	0.095032	LINC00661;GAS6-DT;KCNQ
5	PRC2_SUZ12_UP.V1_DN	0.578350	0.858383	1.317926	0.100193	LINC01558
6	STK33_UP	0.589582	-0.639237	-1.121394	0.383562	MIR646HG;MIR22HG;ATF
7	SRC_UP.V1_DN	0.663471	0.751542	1.337386	0.106762	NDUFV1-DT;LINC00842;MIR1
8	IL21_UP.V1_DN	0.670688	-0.715846	-1.123210	0.401361	MIR
9	JAK2_DN.V1_DN	0.674435	-0.469281	-1.030834	0.408389	LINC00937;HYMAI;POLR2J4;LINC
10	DCA_UP.V1_UP	0.724509	-0.715846	-1.144274	0.359833	MIR6
11	PRC2_EZH2_UP.V1_DN	0.728079	-0.539491	-0.846325	0.683468	RHPN1-AS1;
12	STK33_SKM_UP	0.764333	-0.471073	-0.869964	0.644351	AS1;LINC02910;I
13	MYC_UP.V1_UP	0.770079	-0.665762	-1.227714	0.258586	SLC25A25-AS1;LINC00951;LIN
14	MTOR_UP.N4.V1_DN	0.774910	-0.536194	-0.918326	0.597753	POLR2J4;HHI
15	PDGF_UP.V1_DN	0.799858	-0.589122	-1.166574	0.270563	LINC01558;MIR124-1HG;SM
16	STK33_DN	0.847409	0.645280	1.188278	0.265152	CCDC26;LINC00926;UCA1;DOCK
17	KRAS.BREAST_UP.V1_UP	0.855195	-0.429197	-0.672531	0.840580	FA
18	NFE2L2.V2	0.934718	0.612644	1.205416	0.236036	LY86-AS1;NF
19	PTEN_DN.V1_DN	0.935916	0.393019	0.604513	0.945155	KIAA0
20	CAMP_UP.V1_DN	0.943188	0.434932	0.658713	0.903614	MIR155
21	STK33_NOMO_UP	0.954191	0.471532	0.793643	0.750484	DPP10-AS1;LINC0
22	KRAS.600_UP.V1_DN	0.971011	0.546402	0.824769	0.764595	KIAA0
23	NOTCH_DN.V1_UP	0.974663	0.706454	1.083932	0.390566	GSN-AS1
24	PRC1_BMI_UP.V1_DN	0.982399	0.383581	0.673140	0.875912	FAM106A;LINC02249;GUSB
25	STK33_NOMO_DN	1.000000	0.395064	0.827997	0.720000	LINC00926;DLX6-AS1;
26	KRAS.LUNG_UP.V1_UP	1.000000	0.716616	1.110636	0.383486	LINC00472;MIR
27	BMI1_DN_MEL18_DN.V1_DN	1.000000	0.640389	0.987835	0.529301	KLF3-A
28	PTEN_DN.V2_UP	1.000000	0.679857	1.038175	0.477064	WT1-

	Term	fdr	es	nes	pval	
29	IL15_UP.V1_UP	1.000000	0.447479	0.679701	0.893617	PLA
30	MYC_UP.V1_DN	1.000000	0.415727	0.834780	0.678571	DT;LINC00842;MEG3;LIN
31	JNK_DN.V1_DN	1.000000	0.529022	0.864703	0.692168	XIST;LINC00
32	CTIP_DN.V1_DN	1.000000	0.504475	0.847318	0.700555	GSN-AS1;FPB41L4A-I

In [916…

```
pre_res12 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
out12= []

for term in list(pre_res12.results):
    out12.append([term,
                pre_res12.results[term]['fdr'],
                pre_res12.results[term]['es'],
                pre_res12.results[term]['nes'],
                pre_res12.results[term]['pval'],
                 pre_res12.results[term]['matched_genes']])

out_df12 = pd.DataFrame(out12, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
out_df12
```

```
2023-10-28 22:48:14,014 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:48:14,027 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
2023-10-28 22:48:14,037 [ERROR] No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.

2023-10-28 22:48:14,042 [ERROR] The first entry of your gene_sets (gmt) look like thi
s : { HALLMARK_TNFA_SIGNALING_VIA_NFKB: [JUNB, CXCL2, ATF3, NFKBIA, TNFAIP3, PTGS2, C
XCL1, IER3, CD83, CCL20, CXCL3, MAFF, NFKB2, TNFAIP2, HBEGF, KLF6, BIRC3, PLAUR, ZFP3
6, ICAM1, JUN, EGR3, IL1B, BCL2A1, PPP1R15A, ZC3H12A, SOD2, NR4A2, IL1A, RELB, TRAF1,
BTG2, DUSP1, MAP3K8, ETS2, F3, SDC4, EGR1, IL6, TNF, KDM6B, NFKB1, LIF, PTX3, FOSL1,
NR4A1, JAG1, CCL4, GCH1, CCL2, RCAN1, DUSP2, EHD1, IER2, REL, CFLAR, RIPK2, NFKBIE, N
R4A3, PHLDA1, IER5, TNFSF9, GEM, GADD45A, CXCL10, PLK2, BHLHE40, EGR2, SOCS3, SLC2A6,
PTGER4, DUSP5, SERPINB2, NFIL3, SERPINE1, TRIB1, TIPARP, RELA, BIRC2, CXCL6, LITAF, T
NFAIP6, CD44, INHBA, PLAU, MYC, TNFRSF9, SGK1, TNIP1, NAMPT, FOSL2, PNRC1, ID2, CD69,
IL7R, EFNA1, PHLDA2, PFKFB3, CCL5, YRDC, IFNGR2, SQSTM1, BTG3, GADD45B, KYNU, G0S2, B
TG1, MCL1, VEGFA, MAP2K3, CDKN1A, CCN1, TANK, IFIT2, IL18, TUBB2A, IRF1, FOS, OLR1, R
HOB, AREG, NINJ1, ZBTB10, PLPP3, KLF4, CXCL11, SAT1, CSF1, GPR183, PMEPA1, PTPRE, TLR
2, ACKR3, KLF10, MARCKS, LAMB3, CEBPB, TRIP10, F2RL1, KLF9, LDLR, TGIF1, RNF19B, DRAM
1, B4GALT1, DNAJB4, CSF2, PDE4B, SNN, PLEK, STAT5A, DENND5A, CCND1, RIGI, SPHK1, CD8
0, TNFAIP8, CCNL1, FUT4, CCRL2, SPSB1, TSC22D1, B4GALT5, SIK1, CLCF1, NFE2L2, FOSB, P
ER1, NFAT5, ATP2B1, IL12B, IL6ST, SLC16A6, ABCA1, HES1, BCL6, IRS2, SLC2A3, CEBPD, IL
23A, SMAD3, TAP1, MSC, IFIH1, IL15RA, TNIP2, BCL3, PANX1, FJX1, EDN1, EIF1, BMP2, DUS
P4, PDLIM5, ICOSLG, GFPT2, KLF2, TNC, SERPINB8, MXD1]}
2023-10-28 22:48:14,042 [ERROR] The first 5 genes look like this : [ AC133530.1, LINC
01115, LINC00486, AC010789.1, AC226118.1 ]
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
Cell In[916], line 1
----> 1 pre_res12 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MAPK in NSCLC\Gene Sets\msigdb_v2023.1.Hs_GMTs\MSigDB\h.all.v2023.1.Hs.symbols.gmt", min_size = 3)
      2 out12= []
      4 for term in list(pre_res12.results):

File ~\anaconda3\Lib\site-packages\gseapy\__init__.py:377, in prerank(rnk, gene_sets, outdir, pheno_pos, pheno_neg, min_size, max_size, permutation_num, weighted_score_type, ascending, threads, figsize, format, graph_num, no_plot, seed, verbose, *arg, **kwarg)
    357     threads = kwarg["processes"]
    358 pre = Prerank(
    359     rnk,
    360     gene_sets,
   (...)
    375     verbose,
    376 )
--> 377 pre.run()
    378 return pre

File ~\anaconda3\Lib\site-packages\gseapy\gsea.py:435, in Prerank.run(self)
    433 self._logger.info("Parsing data files for GSEA.............................")
    434 # filtering out gene sets and build gene sets dictionary
--> 435 gmt = self.load_gmt(gene_list=dat2.index.values, gmt=self.gene_sets)
    436 self.gmt = gmt
    437 self._logger.info(
    438     "%04d gene_sets used for further statistical testing....." % len(gmt)
    439 )

File ~\anaconda3\Lib\site-packages\gseapy\base.py:302, in GSEAbase.load_gmt(self, gene_list, gmt)
    294     self._logger.error(
    295         "The first entry of your gene_sets (gmt) look like this : %s"
    296         % dict_head
    297     )
    298     self._logger.error(
    299         "The first 5 genes look like this : [ %s ]"
    300         % (", ".join(list(gene_list)[:5]))
    301     )
--> 302     raise LookupError(msg)
    304 # self._gmtdct = genesets_dict
    305 return genesets_dict

LookupError: No gene sets passed through filtering condition !!!
Hint 1: Try to lower min_size or increase max_size !
Hint 2: Check gene symbols are identifiable to your gmt input.
Hint 3: Gene symbols curated in Enrichr web services are all upcases.
```

In [936…

```
pre_res13 = gp.prerank (rnk= ranking, gene_sets = 'miRTarBase_2017', min_size = 3)
out13 = []

for term in list(pre_res13.results):
    out13.append([term,
                  pre_res13.results[term]['fdr'],
                  pre_res13.results[term]['es'],
                  pre_res13.results[term]['nes'],
                  pre_res13.results[term]['pval'],
```

```
                pre_res13.results[term]['matched_genes']])

out_df13 = pd.DataFrame(out13, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
out_df13
```

```
2023-10-28 23:42:41,260 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 23:42:41,276 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[936]:

	Term	fdr	es	nes	pval	matched_genes
0	hsa-miR-186-5p	0.689203	-0.711567	-1.148775	0.343373	BDNF-AS;LINC00598;PVT1
1	hsa-miR-4279	0.706396	-0.620259	-0.999808	0.539278	LINC00598;TMEM78;LINC01560
2	hsa-miR-24-3p	0.804405	-0.635446	-1.023904	0.517467	TMEM105;LINC00632;TYMSOS
3	hsa-miR-124-3p	0.859134	-0.417551	-0.657431	0.839827	HOTAIR;DLEU1;NEAT1
4	hsa-miR-1976	0.865779	-0.744749	-1.162160	0.330454	TMEM78;LINC01560;ARIH2OS
5	hsa-miR-6812-3p	0.903404	-0.475987	-0.734761	0.785417	LINC01551;EXOC3-AS1;ARIH2OS
6	hsa-miR-4728-5p	0.929304	0.462169	0.717469	0.828846	BDNF-AS;LINC00598;DNAH10OS
7	hsa-miR-6883-5p	0.929304	0.462169	0.717469	0.828846	BDNF-AS;LINC00598;DNAH10OS
8	hsa-miR-149-3p	0.929304	0.462169	0.717469	0.828846	BDNF-AS;LINC00598;DNAH10OS
9	hsa-miR-6785-5p	0.929304	0.462169	0.717469	0.828846	BDNF-AS;LINC00598;DNAH10OS
10	hsa-miR-204-5p	0.960412	0.409809	0.628276	0.931099	UCA1;LINC00598;MALAT1
11	hsa-miR-26b-5p	0.976362	0.294564	0.516038	0.970534	RAMP2-AS1;HHLA3;CYB561D2;MIR22HG;PCOTH
12	hsa-miR-3653-5p	1.000000	-0.710344	-1.205058	0.295218	TMEM78;DLEU1;LINC01560;ARIH2OS
13	hsa-miR-6849-3p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
14	hsa-miR-759	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
15	hsa-miR-508-5p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
16	hsa-miR-8055	1.000000	0.651918	0.978381	0.555556	LINC01551;LINC00598;EXOC3-AS1
17	hsa-miR-1273g-3p	1.000000	0.683163	1.053872	0.454545	COLCA1;TSPEAR-AS2;HTR5A-AS1
18	hsa-miR-3672	1.000000	0.635983	1.180804	0.262357	LINC01556;LINC01551;TSPEAR-AS2;LINC00598;HTR5A...
19	hsa-miR-4797-5p	1.000000	0.453795	0.759469	0.797665	TSPEAR-AS2;LINC00598;HTR5A-AS1;DNAH10OS
20	hsa-miR-5586-3p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS

	Term	fdr	es	nes	pval	matched_genes
21	hsa-miR-6512-3p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
22	hsa-miR-766-3p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
23	hsa-miR-190a-3p	1.000000	-0.745064	-1.186290	0.301205	FAM182B;DNAH10OS;LINC00955
24	hsa-miR-375	1.000000	0.610068	0.932015	0.622857	SOX2-OT;MIR17HG;MALAT1
25	hsa-miR-4284	1.000000	0.565544	0.853293	0.724528	TMEM105;LINC00632;LINC00598
26	hsa-miR-6720-5p	1.000000	0.548532	0.849202	0.729730	TSPEAR-AS2;HTR5A-AS1;DNAH10OS
27	hsa-miR-129-5p	1.000000	0.651918	0.978381	0.555556	LINC01551;LINC00598;EXOC3-AS1
28	hsa-miR-6864-3p	1.000000	0.635983	1.180804	0.262357	LINC01556;LINC01551;TSPEAR-AS2;LINC00598;HTR5A...
29	hsa-miR-873-3p	1.000000	0.754026	1.176718	0.260081	LINC01556;TSPEAR-AS2;HTR5A-AS1

In [ ]:
```
pre_res13 = gp.prerank (rnk= ranking, gene_sets = r"C:\Users\Hasan\Desktop\Research\MA
out13 = []

for term in list(pre_res13.results):
    out13.append([term,
                  pre_res13.results[term]['fdr'],
                  pre_res13.results[term]['es'],
                  pre_res13.results[term]['nes'],
                  pre_res13.results[term]['pval'],
                   pre_res13.results[term]['matched_genes']])

out_df13 = pd.DataFrame(out13, columns = ['Term','fdr', 'es', 'nes','pval','matched_ge
out_df13
```

```
2023-10-28 22:51:50,703 [WARNING] Input gene rankings contains duplicated IDs, Only u
se the duplicated ID with highest value!
2023-10-28 22:51:50,719 [WARNING] Duplicated values found in preranked stats: 6.19% o
f genes
The order of those genes will be arbitrary, which may produce unexpected results.
```

Out[ ]:

	Term	fdr	es	nes	pval	
0	TOX4_TARGET_GENES	0.011584	-0.708107	-1.929676	0.002137	WEE2-AS1;MACC1-DT;ANX
1	ELF2_TARGET_GENES	0.015928	-0.533303	-1.936616	0.000000	MED28-DT;VPS3 AS1;LI
2	CEBPZ_TARGET_GENES	0.018824	-0.528553	-1.958051	0.000000	CD101-AS1;SRP14-DT;F
3	PRKDC_TARGET_GENES	0.030409	-0.524052	-1.813057	0.000000	ID2-AS1;RHPN1-AS1;ZFHX3-AS1;
4	DBP_TARGET_GENES	0.034029	-0.495382	-1.846017	0.000000	PLBD1-AS1;IQCH-AS1;UBE2H-DT;UE
...	...	...	...	...	...	
694	NKX3A_01	1.000000	0.433583	0.806295	0.722555	LINC00671;LINC01597;WT1-AS;
695	TGCCAAR_NF1_Q6	1.000000	0.508113	1.090554	0.355019	LINC00472;LINC01559;C3orf36;COLC
696	FOXJ2_02	1.000000	0.584231	1.029850	0.478088	KLF3-AS1;COLCA1;LINC01164;KIRR
697	NRF2_01	1.000000	0.627082	0.972641	0.572534	ZNF22-AS1;RP
698	CTTTGA_LEF1_Q2	1.000000	0.328484	0.791899	0.740883	LINC01559;KLF3-AS1;LINC00671;LINC

699 rows × 6 columns

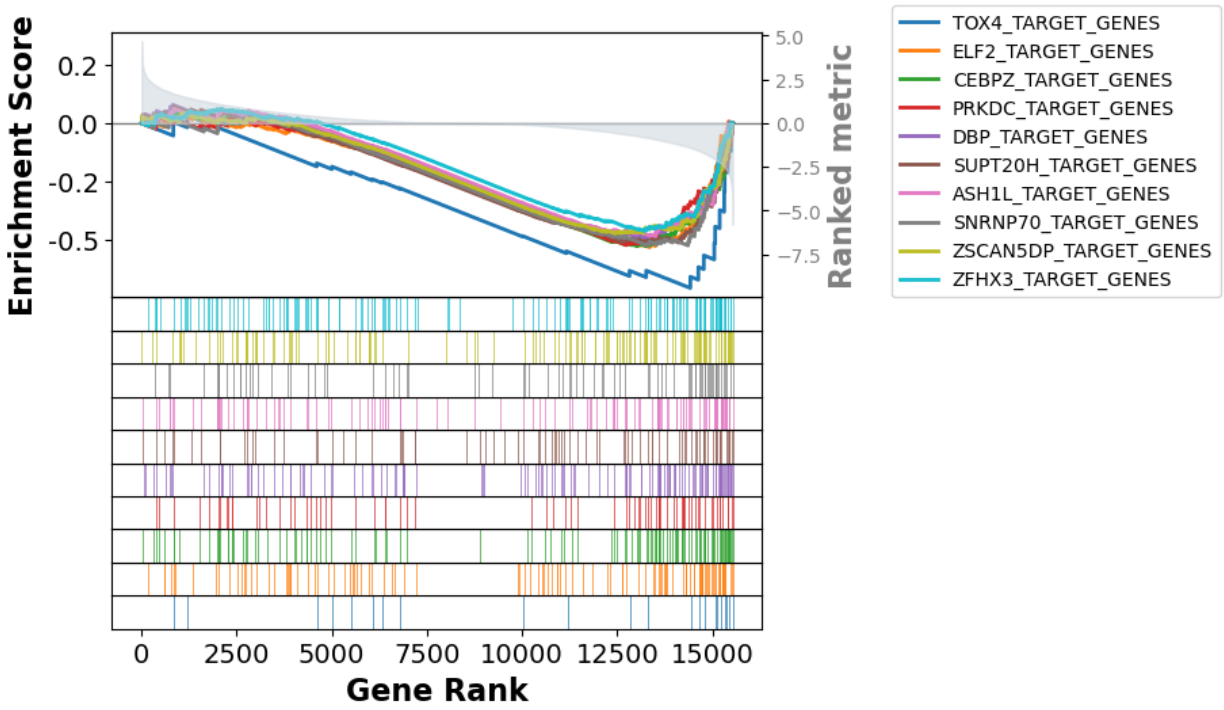
In [214…

```
out_df1.to_csv('MSigDb.csv', sep = ',')
```

GSEA visualizaiton

In [106…

```
axs = pre_res4.plot(terms=out_df4.iloc[0:10].Term,
                    #legend_kws={'loc': (1.2, 0)}, # set the legend loc
                    show_ranking=True, # whether to show the second yaxis
                    figsize=(4,5),
                   legend_kws={'loc':(1.2,0)})
    #c3.all.v2023.1.Hs.symbols.gmt
```


Enrichment Score
0.2
0.0
-0.2
-0.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Ranked metric
0
2500
5000
7500
10000
12500
15000
Gene Rank
TOX4_TARGET_GENES
ELF2_TARGET_GENES
CEBPZ_TARGET_GENES
PRKDC_TARGET_GENES
DBP_TARGET_GENES
SUPT20H_TARGET_GENES
ASH1L_TARGET_GENES
SNRNP70_TARGET_GENES
ZSCAN5DP_TARGET_GENES
ZFHX3_TARGET_GENES