GDC scRNA-Seq Analysis

19 May 2025

Bill Wysocki, Ph.D – Director of User Services
Zhenyu Zhang, Ph.D – Director of Bioinformatics
Center for Translational Data Science, University of Chicago

Xin Zhou, Ph.D - Director of Data Visualization
Computational Biology Department, St. Jude Children's Research Hospital

NIH NATIONAL CANCER INSTITUTE

Webinar Logistics

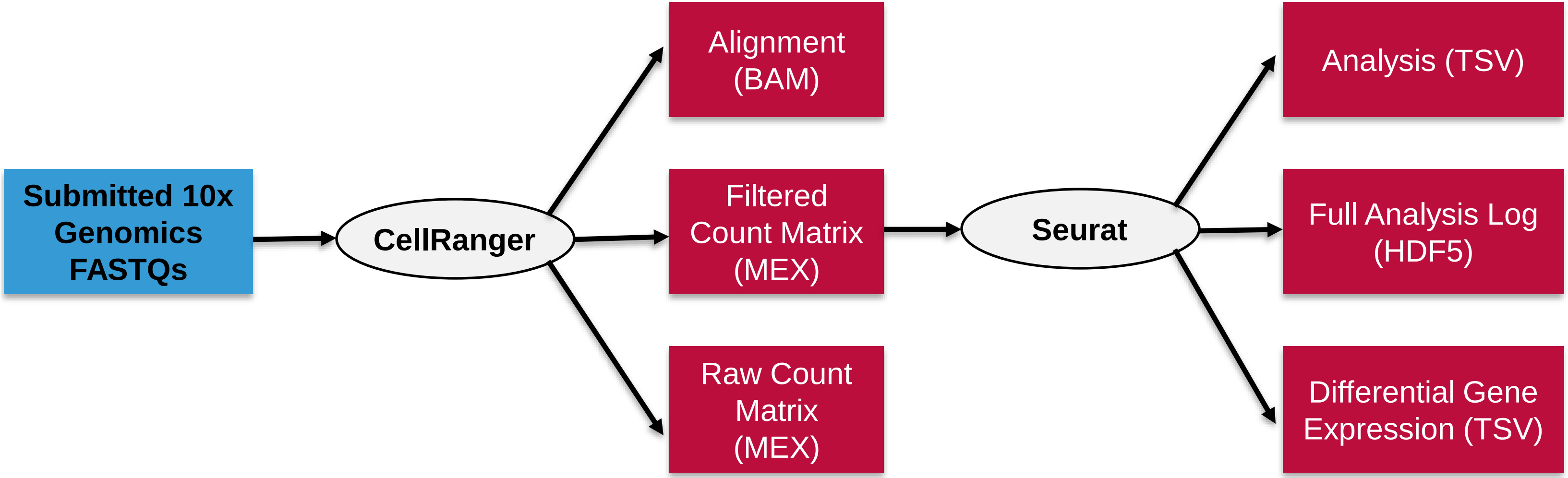
- ***Webinar will be recorded***
- ***Recording and slides will be made available soon***
- ***Type any questions in the Q&A panel – they will be addressed at the end***

Agenda

1. *Overview of GDC scRNA-Seq harmonization workflows*
2. *GDC scRNA-Seq file overview*
3. *GDC Single Cell RNA-Seq tool*
4. *GDC scRNA-Seq expression API*
5. *Questions*

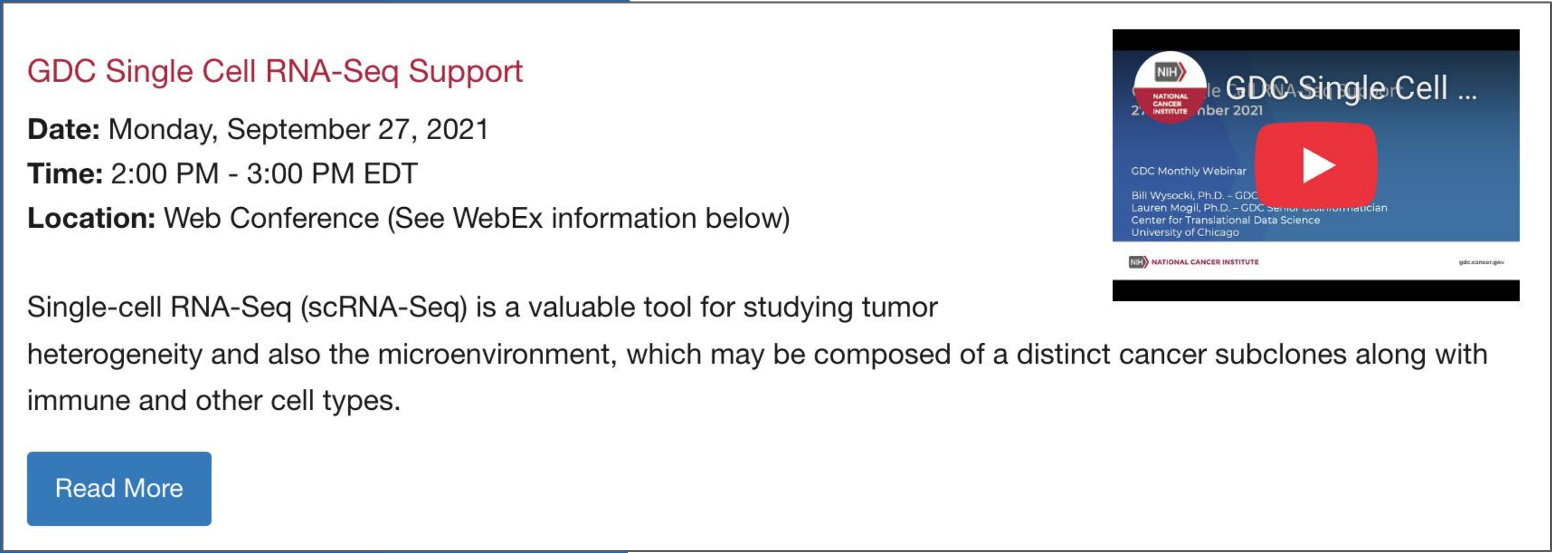
Overview of GDC scRNA-Seq Harmonization Workflows

scRNA-Seq Harmonization in the GDC



- All released scRNA-Seq files are open-access, except for the BAMs

scRNA-Seq File Overview



GDC Single Cell RNA-Seq Support

Date: Monday, September 27, 2021

Time: 2:00 PM - 3:00 PM EDT

Location: Web Conference (See WebEx information below)

Single-cell RNA-Seq (scRNA-Seq) is a valuable tool for studying tumor heterogeneity and also the microenvironment, which may be composed of a distinct cancer subclones along with immune and other cell types.

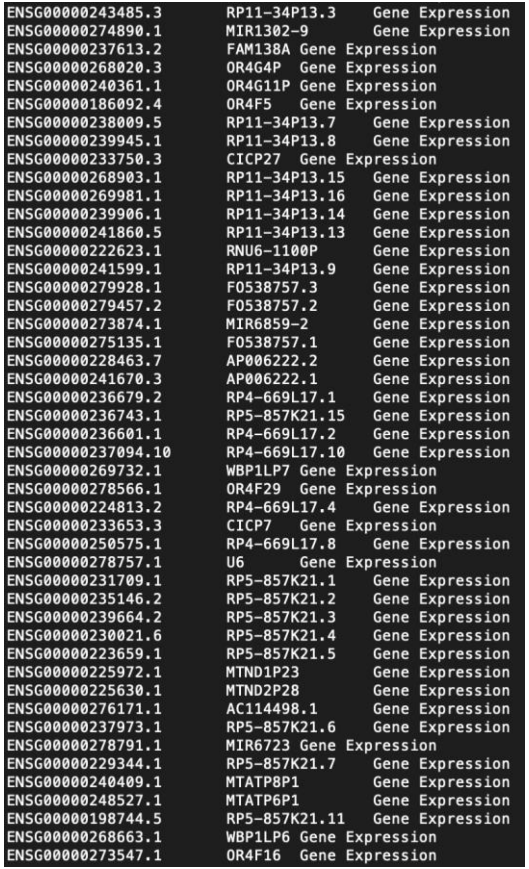
Read More

GDC Single Cell ...
nber 2021
GDC Monthly Webinar
Bill Wysocki, Ph.D. – GDC
Lauren Mogil, Ph.D. – GDC Senior Bioinformatician
Center for Translational Data Science
University of Chicago
NATIONAL CANCER INSTITUTE
gdc.cancer.gov

scRNA-Seq File Formats – Gene Expression

File.tar.gz - MEX Format

- **barcodes.tsv.gz** - a list of barcodes for this sample
- **features.tsv.gz** - a list of each gene id and name
- **matrix.mtx.gz** - the gene expression matrix
- Note: If you are using a Mac, unzipping the three internal files may not work in the browser. Use the following command in terminal: `gunzip barcodes.tsv.gz`



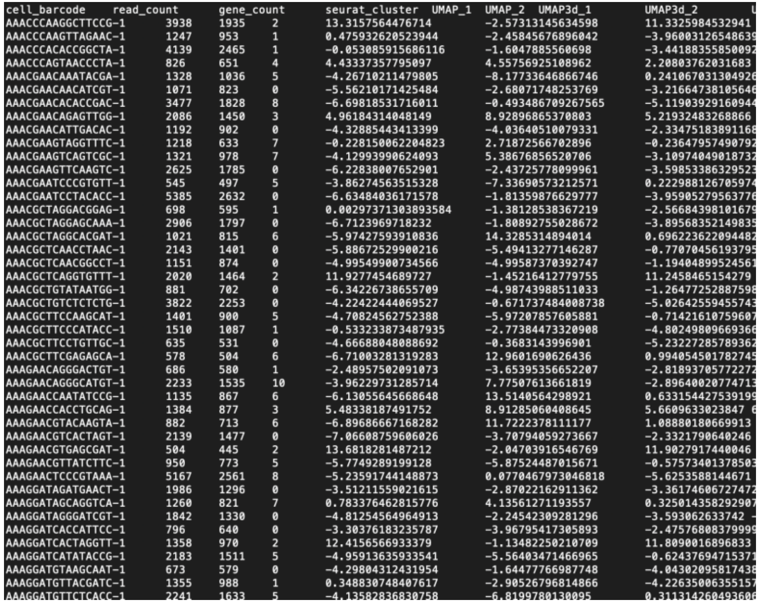
```
ENSG00000243485.3       RP11-34P13.3    Gene Expression
ENSG00000274890.1       MIR1302-9       Gene Expression
ENSG00000237613.2       FAM138A Gene Expression
ENSG00000268020.3       OR4G4P  Gene Expression
ENSG00000240361.1       OR4G11P Gene Expression
ENSG00000186092.4       OR4F5   Gene Expression
ENSG00000238009.5       RP11-34P13.7    Gene Expression
ENSG00000239945.1       RP11-34P13.8    Gene Expression
ENSG00000233750.3       CICP27  Gene Expression
ENSG00000268903.1       RP11-34P13.15   Gene Expression
ENSG00000269981.1       RP11-34P13.16   Gene Expression
ENSG00000239906.1       RP11-34P13.14   Gene Expression
ENSG00000241860.5       RP11-34P13.13   Gene Expression
ENSG00000222623.1       RNU6-1100P      Gene Expression
ENSG00000241599.1       RP11-34P13.9    Gene Expression
ENSG00000279928.1       FO538757.3      Gene Expression
ENSG00000279457.2       FO538757.2      Gene Expression
ENSG00000273874.1       MIR6859-2       Gene Expression
ENSG00000275135.1       FO538757.1      Gene Expression
ENSG00000228463.7       AP006222.2      Gene Expression
ENSG00000241670.3       AP006222.1      Gene Expression
ENSG00000236679.2       RP4-669L17.1    Gene Expression
ENSG00000236743.1       RP5-857K21.15   Gene Expression
ENSG00000236601.1       RP4-669L17.2    Gene Expression
ENSG00000237094.10      RP4-669L17.10   Gene Expression
ENSG00000269732.1       WBP1LP7 Gene Expression
ENSG00000278566.1       OR4F29  Gene Expression
ENSG00000224813.2       RP4-669L17.4    Gene Expression
ENSG00000233653.3       CICP7   Gene Expression
ENSG00000250575.1       RP4-669L17.8    Gene Expression
ENSG00000278757.1       U6      Gene Expression
ENSG00000231709.1       RP5-857K21.1    Gene Expression
ENSG00000235146.2       RP5-857K21.2    Gene Expression
ENSG00000239664.2       RP5-857K21.3    Gene Expression
ENSG00000230021.6       RP5-857K21.4    Gene Expression
ENSG00000223659.1       RP5-857K21.5    Gene Expression
ENSG00000225972.1       MTND1P23        Gene Expression
ENSG00000225630.1       MTND2P28        Gene Expression
ENSG00000276171.1       AC114498.1      Gene Expression
ENSG00000237973.1       RP5-857K21.6    Gene Expression
ENSG00000278791.1       MIR6723 Gene Expression
ENSG00000229344.1       RP5-857K21.7    Gene Expression
ENSG00000240409.1       MTATP8P1        Gene Expression
ENSG00000248527.1       MTATP6P1        Gene Expression
ENSG00000198744.5       RP5-857K21.11   Gene Expression
ENSG00000268663.1       WBP1LP6 Gene Expression
ENSG00000273547.1       OR4F16  Gene Expression
```

scRNA-Seq File Formats – Analysis

TSV Format – Columns:

- Contains information on clustering and visualization.

- cell_barcode
- read_count
- gene_count
- seurat_cluster
- Coordinates for UMAP, tSNE and PCA



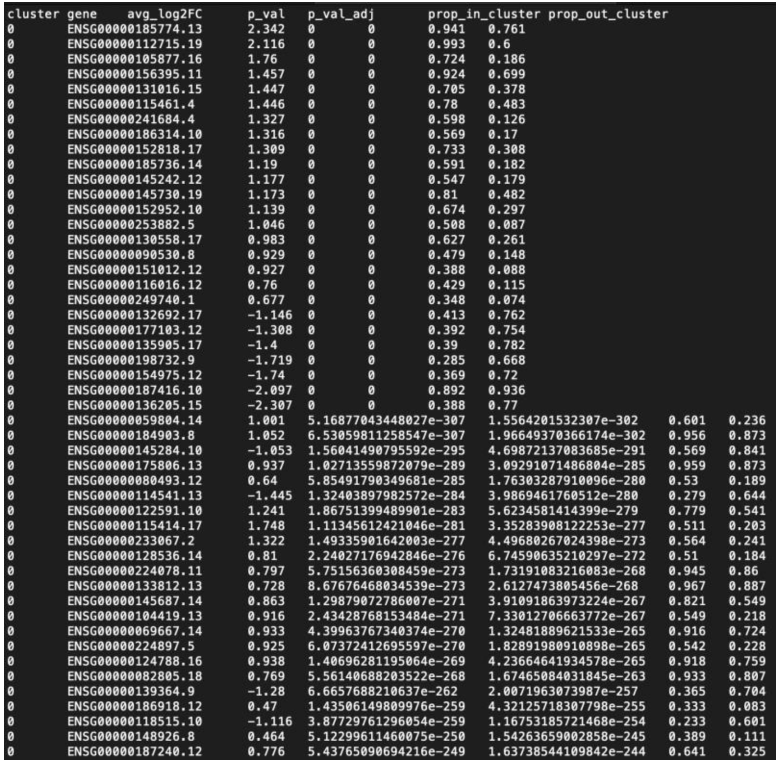
```
cell_barcode     read_count      gene_count      seurat_cluster  UMAP_1  UMAP_2  UMAP3d_1        UMAP3d_2        U
AAACCCAAGGCTTCCG-1      3938    1935    2       13.3157564476714        -2.57313145634598       11.3325984532941
AAACCCAAGTTAGAAC-1      1247    953     1       0.475932620523944       -2.45845676896042       -3.96003126548639
AAACCCACACCGGCTA-1      4139    2465    1       -0.053085915686116      -1.6047885560698        -3.44188355850092
AAACCCAGTAACCCTA-1      826     651     4       4.43337357795097        4.55756925108962        2.20803762031683
AAACGAACAAATACGA-1      1328    1036    5       -4.26710211479805       -8.17733646866746       0.241067031304926
AAACGAACAACATCGT-1      1071    823     0       -5.56210171425484       -2.68071748253769       -3.21664738105646
AAACGAACACACCGAC-1      3477    1828    8       -6.69818531716011       -0.493486709267565      -5.11903929160944
AAACGAACAGAGTTGG-1      2086    1450    3       4.96184314048149        8.92896865370803        5.21932483268866
AAACGAACATTGACAC-1      1192    902     0       -4.32885443413399       -4.03640510079331       -2.33475183891168
AAACGAAGTAGGTTTC-1      1218    633     7       -0.228150062204823      2.71872566702896        -0.23647957490792
AAACGAAGTCAGTCGC-1      1321    978     7       -4.12993990624093       5.38676856520706        -3.10974049018732
AAACGAAGTTCAAGTC-1      2625    1785    0       -6.22838007652901       -2.43725778099961       -3.59853386329523
AAACGAATCCCGTGTT-1      545     497     5       -3.86274563515328       -7.33690573212571       0.222988126705974
AAACGAATCCTACACC-1      5385    2632    0       -6.63484036171578       -1.81359876629777       -3.95905279563776
AAACGCTAGGACGGAG-1      698     595     1       0.00297371303893584     -1.38128538367219       -2.56684398101679
AAACGCTAGGAGCAAA-1      2906    1797    0       -6.7123969718232        -1.80892755028672       -3.89568352149835
AAACGCTAGGCACGAT-1      1021    815     6       -5.97427593910836       14.3285314894014        0.696223622094482
AAACGCTCAACCTAAC-1      2143    1401    0       -5.88672529900216       -5.49413277146287       -0.77070456193795
AAACGCTCAACGGCCT-1      1151    874     0       -4.99549900734566       -4.99587370392747       -1.19404899524561
AAACGCTCAGGTGTTT-1      2020    1464    2       11.9277454689727        -1.45216412779755       11.2458465154279
AAACGCTGTATAATGG-1      881     702     0       -6.34226738655709       -4.98743988511033       -1.26477252887598
AAACGCTGTCTCTCTG-1      3822    2253    0       -4.22422444069527       -0.671737484008738      -5.02642559455743
AAACGCTTCCAAGCAT-1      1401    900     5       -4.70824562752388       -5.97207857605881       -0.71421610759607
AAACGCTTCCCATACC-1      1510    1087    1       -0.533233873487935      -2.77384473320908       -4.80249809669366
AAACGCTTCCTGTTGC-1      635     531     0       -4.66688048088692       -0.3683143996901        -5.23227285789362
AAACGCTTCGAGAGCA-1      578     504     6       -6.71003281319283       12.9601690626436        0.994054501782745
AAAGAACAGGGACTGT-1      686     580     1       -2.48957502091073       -3.65395356652207       -2.81893705772272
AAAGAACAGGGCATGT-1      2233    1535    10      -3.96229731285714       7.77507613661819        -2.89640020774713
AAAGAACCAATATCCG-1      1135    867     6       -6.13055645668648       13.5140564298921        0.633154427539199
AAAGAACCACCTGCAG-1      1384    877     3       5.48338187491752        8.91285060408645        5.6609633023847 6
AAAGAACGTACAAGTA-1      882     713     6       -6.89686667168282       11.7222378111177        1.08880180669913
AAAGAACGTCACTAGT-1      2139    1477    0       -7.06608759606026       -3.70794059273667       -2.3321790640246
AAAGAACGTGAGCGAT-1      504     445     2       13.6818281487212        -2.04703916546769       11.9027917440046
AAAGAACGTTATCTTC-1      950     773     5       -5.7749289199128        -5.87524487015671       -0.57573401378503
AAAGAACTCCCGTAAA-1      5167    2561    8       -5.23591744148873       0.0770467973046818      -5.6253588144671
AAAGGATAGATGAACT-1      1986    1296    0       -3.51211559021615       -2.87022162911362       -3.36174606727472
AAAGGATAGCAGGTCA-1      1260    821     7       0.783376462815776       4.13561271193557        0.325014358292907
AAAGGATAGGGATCGT-1      1842    1330    0       -4.81254564964913       -2.24542309281296       -3.593062633742 -
AAAGGATCACCATTCC-1      796     640     0       -3.30376183235787       -3.96795417305893       -2.47576808379999
AAAGGATCACTAGGTT-1      1358    970     2       12.4156566933379        -1.13482250210709       11.8090016896833
AAAGGATCATATACCG-1      2183    1511    5       -4.95913635933541       -5.56403471466965       -0.62437694715371
AAAGGATGTAAGCAAT-1      673     579     0       -4.29804312431954       -1.64477766987748       -4.04302095817438
AAAGGATGTTACGATC-1      1355    988     1       0.348830748407617       -2.90526796814866       -4.22635006355157
AAAGGATGTTCTCACC-1      2241    1633    5       -4.13582836830758       -6.8199780130095        0.311314260493606
```

scRNA-Seq File Formats – Differential Gene Expression

TSV Format – Columns:

- Contains information on the DGE between clusters for differentially expressed genes.

- cluster
- gene
- avg_logFC – average in fold-change
- p_val – p-value
- p_val_adj – adjusted p-value
- prop_in_cluster – proportion of cells in the cluster
- prop_out_cluster

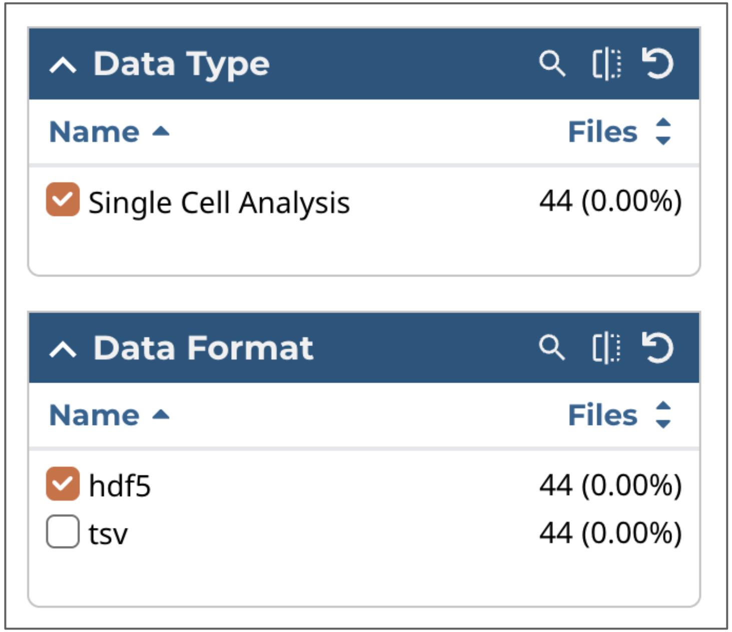


```
cluster gene    avg_log2FC      p_val   p_val_adj       prop_in_cluster prop_out_cluster
0       ENSG00000185774.13      2.342   0       0       0.941   0.761
0       ENSG00000112715.19      2.116   0       0       0.993   0.6
0       ENSG00000105877.16      1.76    0       0       0.724   0.186
0       ENSG00000156395.11      1.457   0       0       0.924   0.699
0       ENSG00000131016.15      1.447   0       0       0.705   0.378
0       ENSG00000115461.4       1.446   0       0       0.78    0.483
0       ENSG00000241684.4       1.327   0       0       0.598   0.126
0       ENSG00000186314.10      1.316   0       0       0.569   0.17
0       ENSG00000152818.17      1.309   0       0       0.733   0.308
0       ENSG00000185736.14      1.19    0       0       0.591   0.182
0       ENSG00000145242.12      1.177   0       0       0.547   0.179
0       ENSG00000145730.19      1.173   0       0       0.81    0.482
0       ENSG00000152952.10      1.139   0       0       0.674   0.297
0       ENSG00000253882.5       1.046   0       0       0.508   0.087
0       ENSG00000130558.17      0.983   0       0       0.627   0.261
0       ENSG00000090530.8       0.929   0       0       0.479   0.148
0       ENSG00000151012.12      0.927   0       0       0.388   0.088
0       ENSG00000116016.12      0.76    0       0       0.429   0.115
0       ENSG00000249740.1       0.677   0       0       0.348   0.074
0       ENSG00000132692.17      -1.146  0       0       0.413   0.762
0       ENSG00000177103.12      -1.308  0       0       0.392   0.754
0       ENSG00000135905.17      -1.4    0       0       0.39    0.782
0       ENSG00000198732.9       -1.719  0       0       0.285   0.668
0       ENSG00000154975.12      -1.74   0       0       0.369   0.72
0       ENSG00000187416.10      -2.097  0       0       0.892   0.936
0       ENSG00000136205.15      -2.307  0       0       0.388   0.77
0       ENSG00000059804.14      1.001   5.16877043448027e-307   1.5564201532307e-302    0.601   0.236
0       ENSG00000184903.8       1.052   6.53059811258547e-307   1.96649370366174e-302   0.956   0.873
0       ENSG00000145284.10      -1.053  1.56041490795592e-295   4.69872137083685e-291   0.569   0.841
0       ENSG00000175806.13      0.937   1.02713559872079e-289   3.09291071486804e-285   0.959   0.873
0       ENSG00000080493.12      0.64    5.85491790349681e-285   1.76303287910096e-280   0.53    0.189
0       ENSG00000114541.13      -1.445  1.32403897982572e-284   3.9869461760512e-280    0.279   0.644
0       ENSG00000122591.10      1.241   1.86751399489901e-283   5.6234581414399e-279    0.779   0.541
0       ENSG00000115414.17      1.748   1.11345612421046e-281   3.35283908122253e-277   0.511   0.203
0       ENSG00000233067.2       1.322   1.49335901642003e-277   4.49680267024398e-273   0.564   0.241
0       ENSG00000128536.14      0.81    2.24027176942846e-276   6.74590635210297e-272   0.51    0.184
0       ENSG00000224078.11      0.797   5.75156360308459e-273   1.73191083216083e-268   0.945   0.86
0       ENSG00000133812.13      0.728   8.67676468034539e-273   2.6127473805456e-268    0.967   0.887
0       ENSG00000145687.14      0.863   1.29879072786007e-271   3.91091863973224e-267   0.821   0.549
0       ENSG00000104419.13      0.916   2.43428768153484e-271   7.33012706663772e-267   0.549   0.218
0       ENSG00000069667.14      0.933   4.39963767340374e-270   1.32481889621533e-265   0.916   0.724
0       ENSG00000224897.5       0.925   6.07372412695597e-270   1.82891980910898e-265   0.542   0.228
0       ENSG00000124788.16      0.938   1.40696281195064e-269   4.23664641934578e-265   0.918   0.759
0       ENSG00000082805.18      0.769   5.56140688203522e-268   1.67465084031845e-263   0.933   0.807
0       ENSG00000139364.9       -1.28   6.6657688210637e-262    2.0071963073987e-257    0.365   0.704
0       ENSG00000186918.12      0.47    1.43506149809976e-259   4.32125718307798e-255   0.333   0.083
0       ENSG00000118515.10      -1.116  3.87729761296054e-259   1.16753185721468e-254   0.233   0.601
0       ENSG00000148926.8       0.464   5.12299611460075e-250   1.54263659002858e-245   0.389   0.111
0       ENSG00000187240.12      0.776   5.43765090694216e-249   1.63738544109842e-244   0.641   0.325
```

scRNA-Seq File Formats – Full Seurat Analysis Log

HDF5 Format – Loom File:

- Contains the full analysis from Seurat
- Stores on-disk rather than in-memory.
- Can be processed with several R or Python packages (loomR or loompy)



Data Type

Name	Files
☑ Single Cell Analysis	44 (0.00%)

Data Format

Name	Files
☑ hdf5	44 (0.00%)
□ tsv	44 (0.00%)

Tool Demo

API Demo

Questions?

U.S. Department of Health & Human Services

National Institutes of Health | National Cancer Institute

https://www.cancer.gov/

1-800-4-CANCER

Produced May 2025