Here is the deal: The file that we downloaded from the omnibus, GSE103322\_HNSCC\_all\_data.txt.gz from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103322>, had to be reformatted to make it nice.

**Step 1**.  We reformatted the downloaded file with a script, ***createInputFile***. After this, the gene names are in single quotes, which we removed in the cell using sed.

**$ perl createInputFileML GSE103322\_HNSCC\_all\_data.txt.gz | gzip > HNSCC\_all\_data\_all\_cell\_types.gz**

**Step 2**.  We then eliminated any column with more than 80% zeroes (we went from 2475 columns to 1415) by running **step2PreprocessingHNSCCdata.R** where takes in output of step 1 (HNSCC\_all\_data\_all\_cell\_types.gz) and user specifies percentage when deciding what cells to eliminate from analysis by eliminating cells with more than XX% percentage of its genes with zero expression in eliminateColsWithMoreThanPercentGenesWithZeroExpression() function in script → output is geneExpression\_MatrixAllCellTypes.gz where XX% is percentage used to eliminate columns with

Example with 80%: What we did for step 2 was to first prepare a table where each row corresponds to a rough summary of the “distribution” of values for the ~20K genes in each single-cell column of the original table.  The first few rows of this table are these:

#column mean median max third fifth

2 42.0769 0 76054 (0 0) (0 7)

3 42.0663 0 109669 (0 0) (0 2)

4 42.0628 0 28396 (0 0) (0 13)

5 42.1329 0 41120 (0 0) (0 0)

So for each cell, we have three quantiles: the median, the third quantile (values shown correspond to the ~33% and the ~67% values) and the fifth (values shown correspond to 20% and 80%).

After a quick look at the entire table (2475 rows), we saw that most of the ~67% values are zero (2420 out of 2475). So we went for the 80% values.  Here, 1060 rows have zero as their 80% value; we eliminated these ones (using the shell on the table).

**Step 3**.  The output from 2 was then used as input to another script, ***pruneTable***, which takes two arguments (in addition to the input file name): 1) proportion of columns (in a category) that must meet a minimum expression level and 2) minimum expression level.

**$ perl pruneTable geneExpression85MatrixAllCellTypes.gz 0.8 4 | gzip > fileForAnalysis85AllCellTypes.gz**

**Step 4**.  The output from Step 3 goes into the R script named ***analyzeExpression.R*** for differential expression analysis, UMAP, PCA, etc..