AGED Vignette on TCGA_PAAD Data

Michael Sweeney, Luke Torre-Healy

11/23/2021

Load TCGA_PAAD data

```
load("../../data/TCGA_PAAD.RData")
```

Load the expression matrix & prep for FaStaNMF

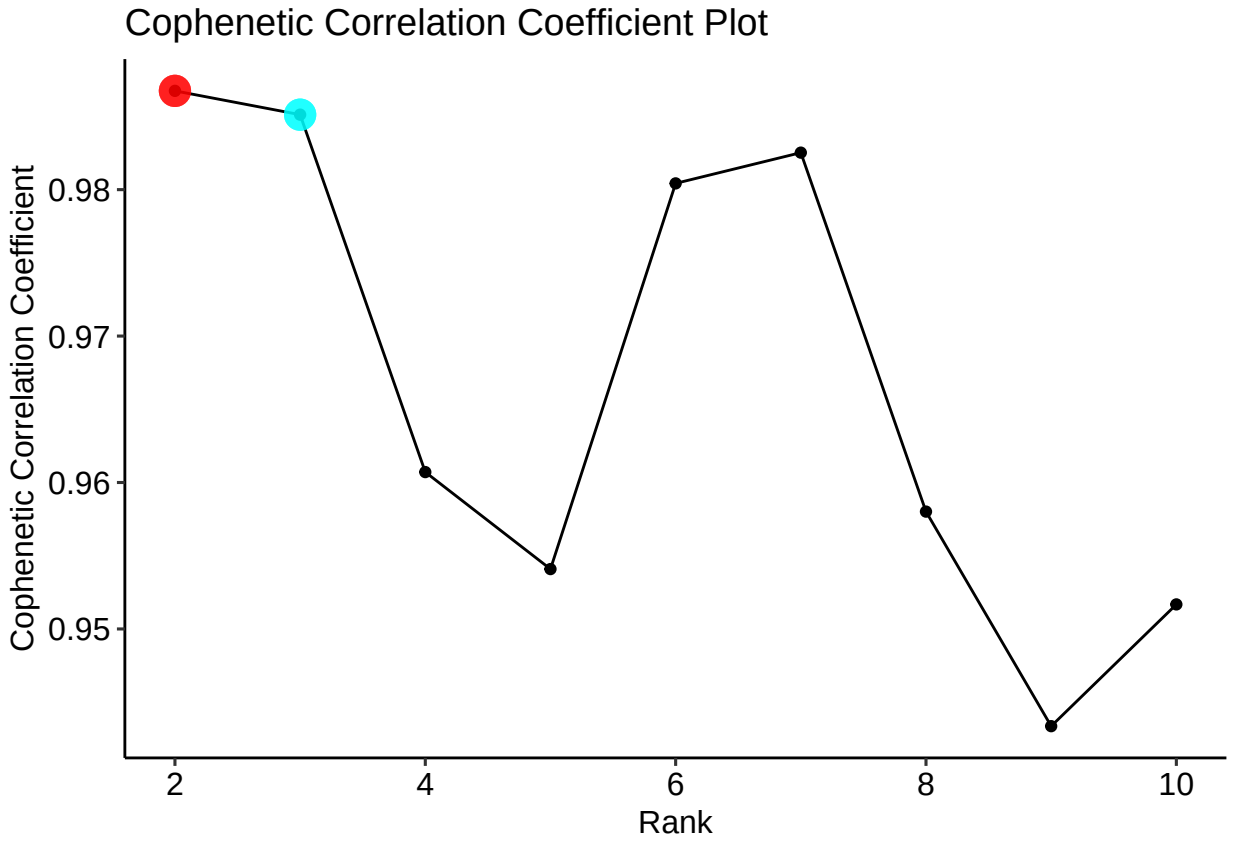
Before running the Automatic Gene Expression Deconvolution (AGED) pipeline, consider if you should remove rows with low variance, and what the threshhold should be. If you are working with unnormalized data, such as raw counts, then also consider normalizing your data using VST or log. In this case, the LungLiverBrain data is pre-normalized.

```
df = as.matrix(TCGA_PAAD$ex)
```

Consider the cophenetic correlation coefficient to pick a rank

Next, consider estimating a possible rank for FaStaNMF using the cophenetic correlation coefficient plot generator, if desired. The generator will highlight up to two key ranks: the maximum rank will be highlighted in red, and the rank directly before the steepest slope to the next rank (before the first positive slope is witnessed) will be highlighted in cyan. If both of these ranks are the same rank, it will be highlighted in purple. If the range of ranks you selected is not a consecutive set of numbers, only the maximum rank will be highlighted. This step is intended to mirror commonly used heuristics for selecting optimal ranks for NMF. The highlighted values are intended to be light suggestions. The rank selected is ultimately up to the user and their pre-existing knowledge of the data.

```
aged::cophenetic_generator(df, rank_range = 2:10, nrun = 200, .options = "p4")
```



The maximum cophenetic correlation coefficient witnessed here is rank = 2, and there is also a significant drop from rank = 3 to rank = 4. Therefore, I will use rank = 3 to run the deconvolution. This decision mirrors the elbow method heuristic for number of clusters selection and will let us look at a basic deconvolution of separating the data into three metagenes. Further runs of AGED with higher ranks may be beneficial after looking at results from primitive runs of AGED with lower ranks.

Run NMF using aged()

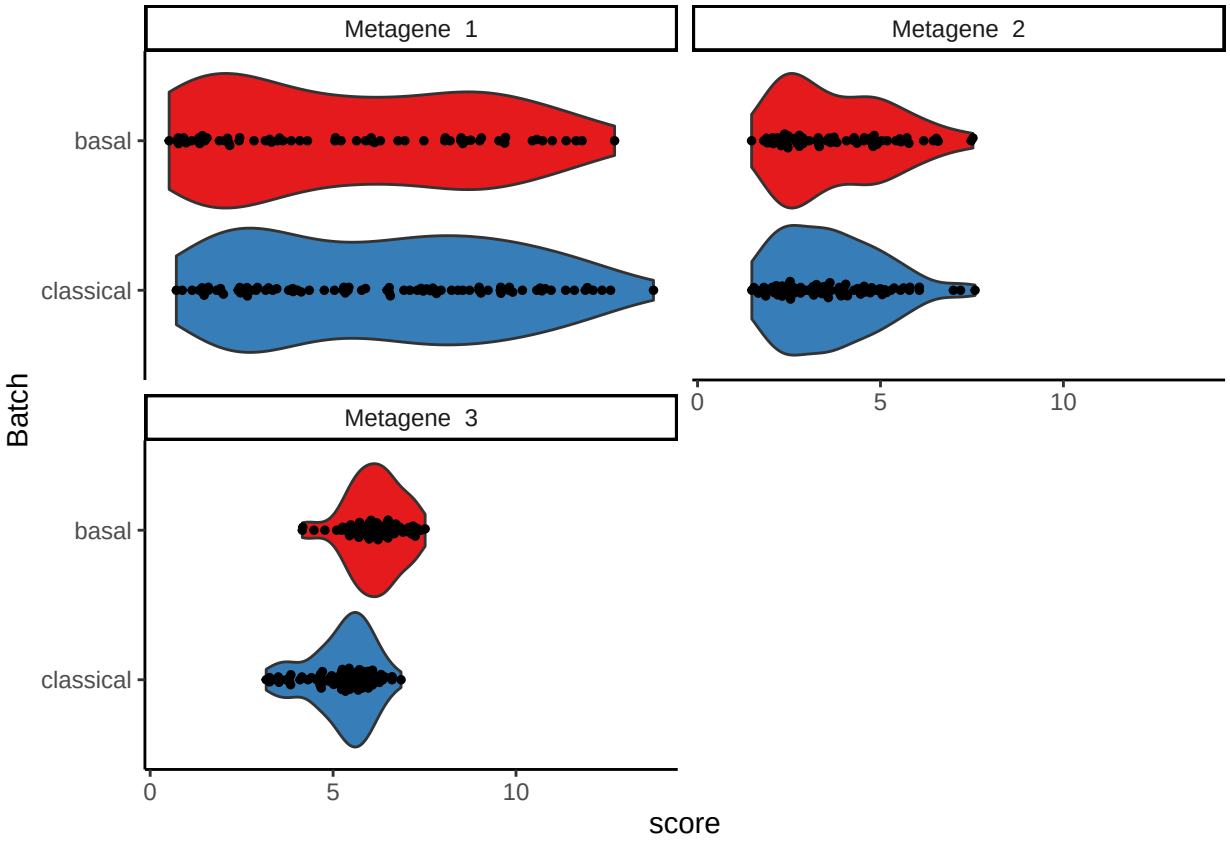
```
aged_results = aged::aged(data = df,
               rank = 3,
               .options = "p4")
```

```
## [1] "Starting FaStaNMF with rank 3..."
## [1] "Reducing the dataset to the most variable genes..."
## [1] "Performing the initial NMF run..."
## [1] "Creating the initial synthetic seed..."
## [1] "Running seeded NMF..."
## [1] "Expanding back to the full dataset..."
## [1] "Normalizing NMF results..."
## [1] "Calculating barcode genes for each metagene..."
```

The gene expression deconvolution will return an unnormalized, raw object from FaStaNMF. In that object, the H matrix will give a proportion of each metagene per sample, and the W matrix will provide the most prominent barcode genes per metagene. AGED also returns lists of the most prominent barcode genes per each metagene after normalization.

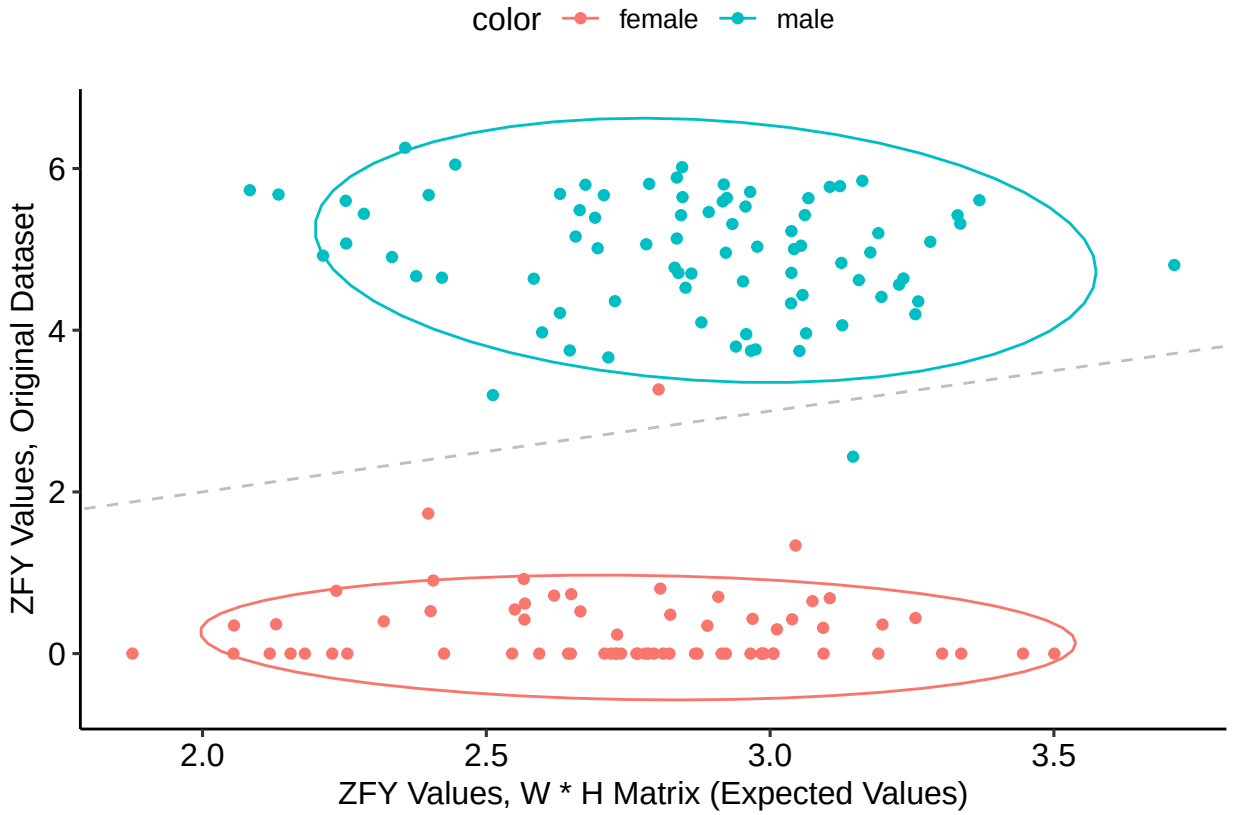
Visualize difference in metagene expression by Moffitt subtype using violin_generator

```
aged::violin_generator(aged_results = aged_results,
                  data = df,
                  batch = TCGA_PAAD$sampInfo$MoffittTumor,
                  var_axis = "Batch")
```



Visualize expected vs. observed expression of a specific gene by gender using scatterplot_generator

```
aged::scatterplot_generator(aged_results = aged_results,
                        data = df,
                        gene = "ZFY",
                        x_axis = "wh",
                        color = TCGA_PAAD$sampInfo$Gender,
                        ellipse = TRUE,
                        reg = F)
```



Visualize metagene expression by Moffitt subtype and gender using heatmap_generator

```
aged::heatmap_generator(aged_results = aged_results,
                        data = df,
                        samp_info = TCGA_PAAD$sampInfo,
                        batches = c("MoffittTumor", "Gender"),
                        hmap_color = c("blue","white","red"),
                        dendrogram = "column",
                        key = T)
```