McCullough T Cells Summary (All)

Stanley Yang

1/5/2018

All T cell data

Sample Summary for all T cells

```
# A tibble: 7 x 2
                       Group     N
                      <fctr> <int>
1   Aged.Female.Brain.NoStim     5
2 Aged.Female.Brain.PlusStim     5
3     Aged.Male.Brain.NoStim     6
4   Aged.Male.Brain.PlusStim     4
5    Aged.Male.Spleen.NoStim     5
6  Aged.Male.Spleen.PlusStim     5
7    Young.Male.Brain.NoStim     5
```

Normalize and Filter the data

Before normalization

```
                               lib.size norm.factors
AgedFemale.Brain.NoStim.13_S10    9655634            1
AgedFemale.Brain.NoStim.14_S1    14866050            1
AgedFemale.Brain.NoStim.15_S6     6654543            1
AgedFemale.Brain.NoStim.16_S12   10970566            1
AgedFemale.Brain.NoStim.17_S8    12591901            1
AgedFemale.Brain.PlusStim.23_S11  4920738            1
AgedFemale.Brain.PlusStim.24_S9   6062158            1
AgedFemale.Brain.PlusStim.25_S5   9203922            1
AgedFemale.Brain.PlusStim.26_S1   3954801            1
AgedFemale.Brain.PlusStim.27_S8   2216762            1
AgedMale.Brain.NoStim.10_S9      19573103            1
AgedMale.Brain.NoStim.11_S10     18771250            1
AgedMale.Brain.NoStim.12_S11     20589456            1
AgedMale.Brain.NoStim.7_S6       24973886            1
AgedMale.Brain.NoStim.8_S7       10293283            1
AgedMale.Brain.NoStim.9_S8       16088402            1
AgedMale.Brain.PlusStim.19_S4     6391537            1
AgedMale.Brain.PlusStim.20_S3     2241662            1
AgedMale.Brain.PlusStim.21_S7     3850380            1
AgedMale.Brain.PlusStim.22_S2     1156594            1
AgedMaleSpleenNoStim30_S10        8998764            1
AgedMaleSpleenNoStim31_S7         9213356            1
AgedMaleSpleenNoStim32_S12        9650685            1
AgedMaleSpleenNoStim33_S3        14709415            1
AgedMaleSpleenNoStim34_S4        10176505            1
AgedMaleSpleenPlusStim36_S2      11101911            1
```

```
AgedMaleSpleenPlusStim37_S6       9932756            1
AgedMaleSpleenPlusStim38_S5       8586075            1
AgedMaleSpleenPlusStim39_S9       6745077            1
AgedMaleSpleenPlusStim40_S11     11188159            1
YoungMale.Brain.NoStim.2_S1      15470386            1
YoungMale.Brain.NoStim.3_S2      15558582            1
YoungMale.Brain.NoStim.4_S3      12213798            1
YoungMale.Brain.NoStim.5_S4      12781536            1
YoungMale.Brain.NoStim.6_S5      13326420            1
```

before filtering low-count genes

```
                             dim.y.counts.
total number of genes detected         38924
total sample number                       35
```

keep the genes that have more than 1 count per million (cpm) in at least 2 samples

```
                                    dim.y.counts.
total number of genes after filtering         12638
total sample number                              35
```

normalize the filtered genes across all samples

```
                               lib.size norm.factors
AgedFemale.Brain.NoStim.13_S10    8450652        1.346
AgedFemale.Brain.NoStim.14_S1    12829959        1.249
AgedFemale.Brain.NoStim.15_S6     5775658        1.110
AgedFemale.Brain.NoStim.16_S12    9689366        1.200
AgedFemale.Brain.NoStim.17_S8    10768556        1.200
AgedFemale.Brain.PlusStim.23_S11  3558217        0.780
AgedFemale.Brain.PlusStim.24_S9   3879926        0.589
AgedFemale.Brain.PlusStim.25_S5   6195711        0.559
AgedFemale.Brain.PlusStim.26_S1   2247330        0.941
AgedFemale.Brain.PlusStim.27_S8   1083591        0.692
AgedMale.Brain.NoStim.10_S9      16743058        0.998
AgedMale.Brain.NoStim.11_S10     16260522        1.018
AgedMale.Brain.NoStim.12_S11     17521858        0.967
AgedMale.Brain.NoStim.7_S6       21581567        0.966
AgedMale.Brain.NoStim.8_S7        9004501        0.792
AgedMale.Brain.NoStim.9_S8       13854192        0.925
AgedMale.Brain.PlusStim.19_S4     4919751        0.797
AgedMale.Brain.PlusStim.20_S3     1531051        1.010
AgedMale.Brain.PlusStim.21_S7     2536233        0.631
AgedMale.Brain.PlusStim.22_S2      746962        0.859
AgedMaleSpleenNoStim30_S10        7317320        1.511
AgedMaleSpleenNoStim31_S7         7520542        1.461
AgedMaleSpleenNoStim32_S12        7904050        1.439
AgedMaleSpleenNoStim33_S3        12378836        1.460
AgedMaleSpleenNoStim34_S4         8315177        1.433
AgedMaleSpleenPlusStim36_S2       8927352        1.108
AgedMaleSpleenPlusStim37_S6       7522045        1.149
AgedMaleSpleenPlusStim38_S5       6625479        1.189
```

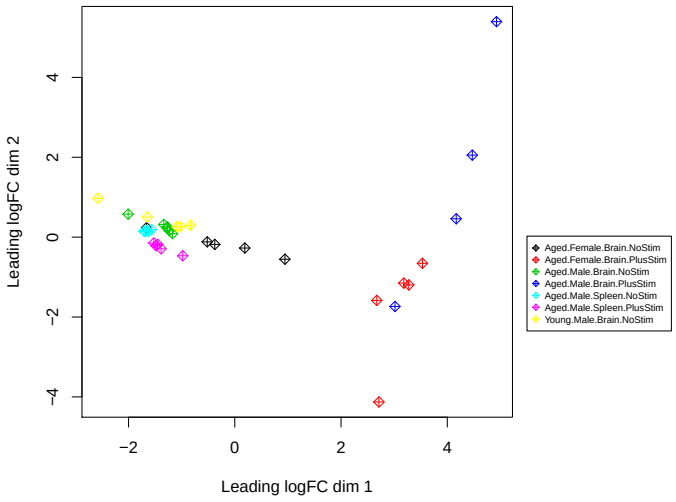

Figure 1: MDS plot

AgedMaleSpleenPlusStim39_S9	5092337	1.252
AgedMaleSpleenPlusStim40_S11	8254639	1.057
YoungMale.Brain.NoStim.2_S1	13277300	0.864
YoungMale.Brain.NoStim.3_S2	13244514	0.679
YoungMale.Brain.NoStim.4_S3	10078524	0.959
YoungMale.Brain.NoStim.5_S4	10953635	1.009
YoungMale.Brain.NoStim.6_S5	11336934	0.969

Explore the data samples

MDS plot

to check the distance among samples (Figure 1)

Principal Component Analysis (PCA)

1. scree plot to show all possible components for variance explained

- proportion of variance explained for each individual components (Figure 2)
- cumulative proportion of variance explained (Figure 3)

2. PCA plot to show the first, second and third components

- Tissue difference between brain and spleen explains the variance for the first component, as most of sample from brain (circle) is larger than -25 on PC1 axis whereas samples from spleen tissue (triangle) is less -25 on PC1 axis (Figure 4).

- Stimulation may explain the variance as the second component, as most of sample without stimulation (circle) is above 0 on PC2 axis whereas samples with stimuation (triangle) is below 0 on PC1 axis (Figure 5).
- Dimention PC2 and PC3 of PCA plot are constructed in Figure 6 and Figure 7. There is no obvious factor that contribute to PC3.

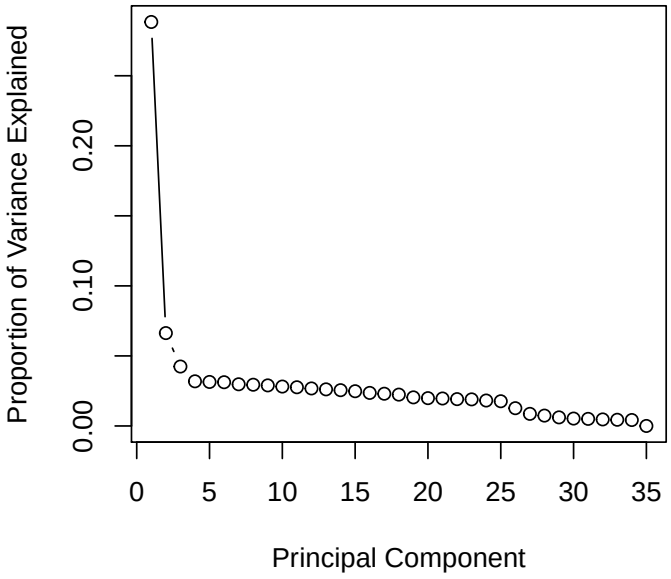

Figure 2: Scree Plot

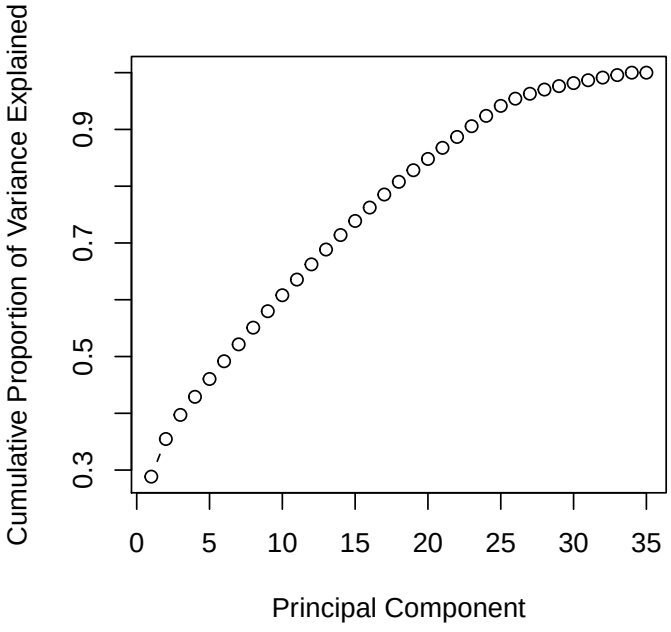

Figure 3: Scree Plot

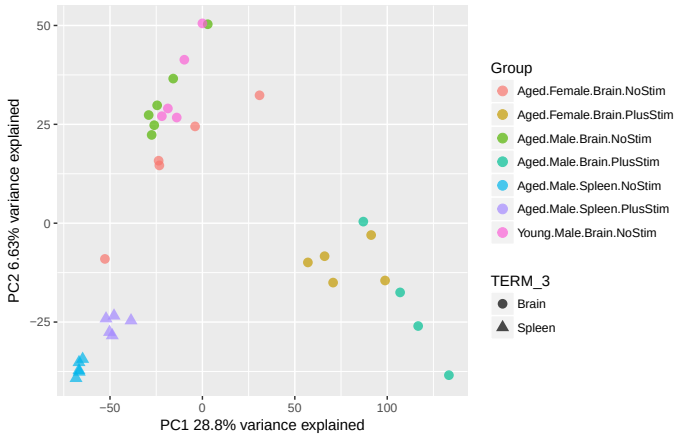

Figure 4: PCA Plot

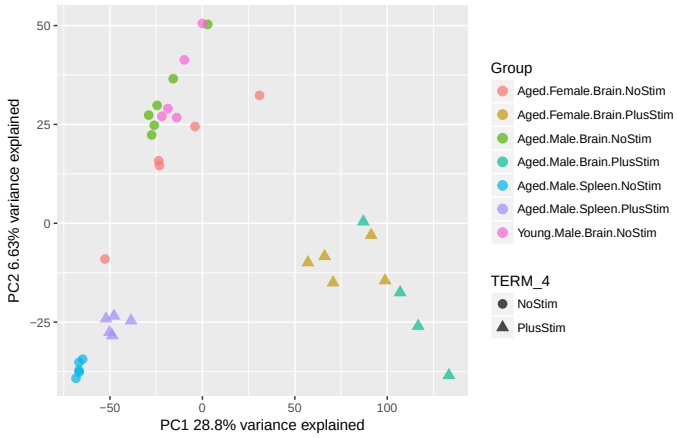

Figure 5: PCA Plot

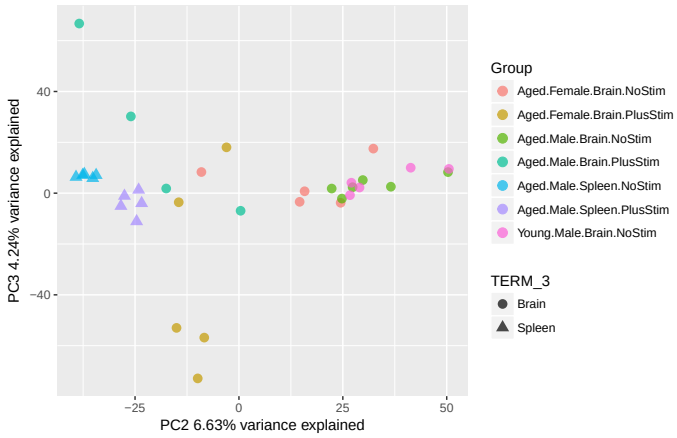

Figure 6: PCA Plot

Figure 7: PCA Plot

Differential expression analysis

testing for differential expression

1. Methods for DE gene analysis (edgeR package)

- I use generalized linear model based quasi-likelihood (QL) F-tests (glmQLFtest) instead of likelihood ratio test (LRT) for find DE genes as they give stricter error rate control by accounting for the uncertainty in dispersion estimation. (The old DE gene lists were made by LRT methods)
- There are two kinds of QL F-tests used in DE gene analysis, they are marked in the output files
  - glmTreat_1 : identifies differential expression based on statistical significance ($FDR < 0.05$ as a cutoff) regardless of how small the difference might be. (1 means the FC = 1)
  - glmTreat_1.5 : identifies the differential expression fold changes are significantly greater than a specified fold change which is 1.5 in this case. (1.5 means FC=1.5, can be changed)

2. list of all the comparisions and their meanings:

- a1_BrvsSp.A.M.NS = Aged.Male.Brain.NoStim - Aged.Male.Spleen.NoStim
  - The difference between unstimulated brain and spleen T-cells in aged male mice.
  - Positive FC indicates the gene expression is higher in Aged.Male.Brain.NoStim than in Aged.Male.Spleen.NoStim, negative FC indicates the gene expression is lower in Aged.Male.Brain.NoStim than in Aged.Male.Spleen.NoStim. The same rule applies to the rest of the comparisons.
  - a1_BrvsSp.A.M.NS will be attached to the output file names indicating this comparison type.The same rule for the rest of the comparsions.
- a2_BrvsSp.A.M.PS = Aged.Male.Brain.PlusStim - Aged.Male.Spleen.PlusStim
  - The difference between stimulated brain and spleen T-cells in aged male mice.
- a3_Stim_BrvsSp.A.M = (Aged.Male.Brain.PlusStim - Aged.Male.Brain.NoStim) - (Aged.Male.Spleen.PlusStim - Aged.Male.Spleen.NoStim)
  - The difference between Brain and Spleen T cells in response to stimulation (interaction effect between tissue and stimulation).
- a4_AvsY.M.Br.NS = Aged.Male.Brain.NoStim - Young.Male.Brain.NoStim
  - The difference between unstimulated T-cells in aged and young male mice.
- a5_FvsM.A.Br.NS = Aged.Female.Brain.NoStim - Aged.Male.Brain.NoStim
  - The difference between unstimulated T-cells in female and male mice.
- a6_FvsM.A.Br.PS = Aged.Female.Brain.PlusStim - Aged.Male.Brain.PlusStim
  - The difference between stimulated T-cells in female and male mice.
- a7_Stim_FvsM.A.Br = (Aged.Female.Brain.PlusStim - Aged.Female.Brain.NoStim) - (Aged.Male.Brain.PlusStim - Aged.Male.Brain.NoStim),

  - The difference between brain T-cells of female and male in response to stimulation (interaction effect between sex and stimulation).
- a8_Stim_A.M.Br = Aged.Male.Brain.PlusStim - Aged.Male.Brain.NoStim
  - The difference between stimulated and unstimulated T-cells in aged male brain
- a9_PSvsNS_A.F.Br = Aged.Female.Brain.PlusStim - Aged.Female.Brain.NoStim
  - The difference between stimulated and unstimulated T-cells in aged female brain
- a10_PSvsNS_A.M.Sp = Aged.Male.Spleen.PlusStim - Aged.Male.Spleen.NoStim
  - The difference between stimulated and unstimulated T-cells in aged male spleen

3. Downstream analysis of DE gene set

- Gene ontology analysis:
  - all the genes from **glmTreat_1** or **glmTreat_1.5** lists with FDR<0.05 were put into gene ontology analysis.
  - The **Up** and **Down** columns indicate the number of genes within the GO terms that are sigificantly up- and down-regulated in this differential expression comparison, respectively. The **P.Up** and **P.Down** columns contain the p-values for over-representation of the GO term in the up- and down-regulated genes, respectively.
  - GO terms with p-value less than $\mathbf{10^{-5}}$ were kept.
- KEGG pathway analysis
  - all the genes from **glmTreat_1** or **glmTreat_1.5** lists with FDR<0.05 were put into KEGG analysis.
  - same meaning for **Up** and **Down** ,and **P.Up** and **P.Down** columns as in GO terms
  - I kept **p-value < 0.05** for KEGG analysis. May need $\mathbf{p < 10^{-5}}$ for more stringent threthold.

4. Decode output files of DE gene lists and down stream analysis

Take a2_BrvsSp.A.M.PS comparison as an example: first refer to comparison table to find out this comparison means: The difference between stimulated brain and spleen T-cells in aged male mice. It contains four files starting with all_Tcells followed by the name of this comparison:

- all_Tcells_a2_BrvsSp.A.M.PS_glmTreat_1.txt
  - glmTreat using glm QL F-tests for significant DE genes (FDR<0.05) no matter how small the change is.
  - DE gene list can be further maually put into KEGG or GO online tool for detailed analysis
- all_Tcells_a2_BrvsSp.A.M.PS_glmTreat_1.5.txt
  - glmTreat using glm QL F-tests for significant DE genes (FDR<0.05) that has a fold change greater than 1.5 in either direction
- all_Tcells_a2_BrvsSp.A.M.PS_KEGG_1.txt
  - KEGG analysis of DE genes (FDR<0.05) from glmTreat_1 file
  - not all the comparison has an output file of KEGG analysis, only the gene sets meet KEGG analysis p-value standard will have this output file. Same rule applies to gene ontology analysis.
- all_Tcells_a2_BrvsSp.A.M.PS_KEGG_1.5.txt
  - KEGG analysis of DE genes (FDR<0.05) from glmTreat_1.5 file
- all_Tcells_a2_BrvsSp.A.M.PS_Ont_1.5.txt
  - Gene Ontology analysis of DE genes (FDR<0.05) from glmTreat_1.5 file
  - There is no output of Gene Ontology analysis of DE genes (FDR<0.05) from glmTreat_1 file

heatmap visualization of sample clustering

Heatmaps are a popular way to display DE results for publicaiton pruposes. Here I generated a sample heatmap based on top 100 DE genes (**glmTreat_1.5**) between unstimulated brain and spleen T-cells in aged male mice (comparison code: a1_BrvsSp.A.M.NS) (Figure 8).

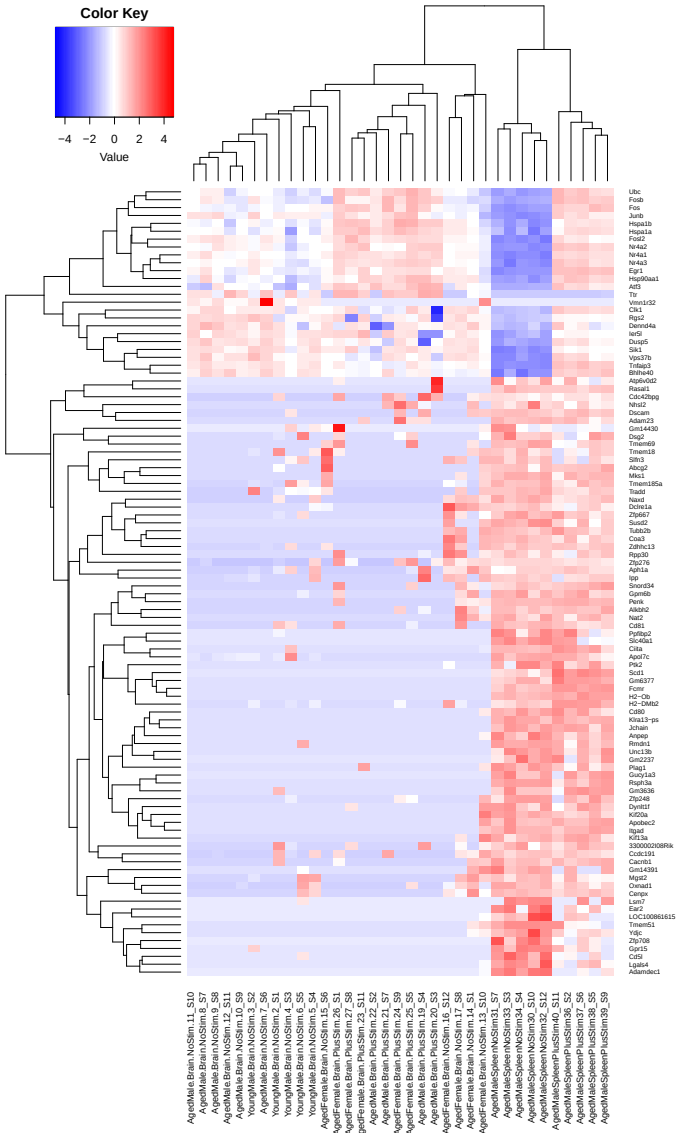

Figure 8: heat map across all the samples using the top 100 most DE genes between brain and spleen samples (in aged male mice)