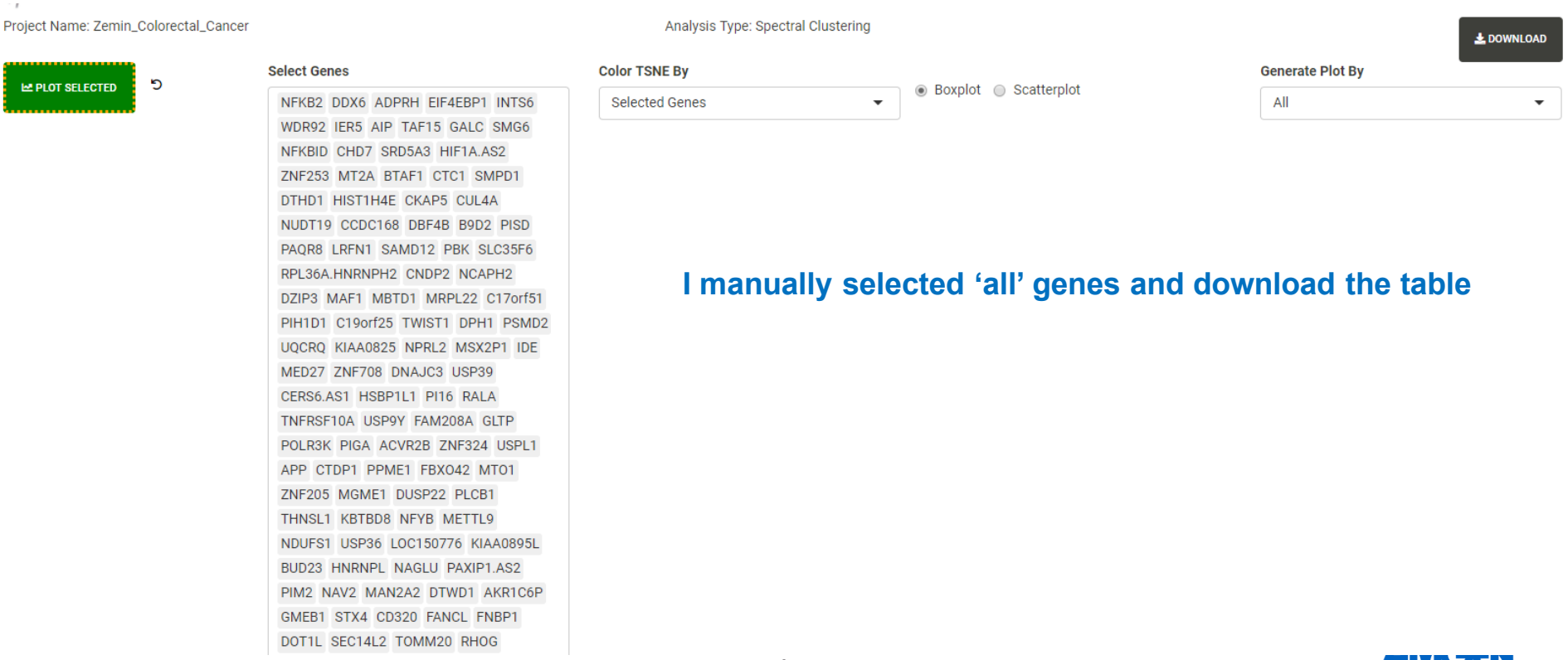
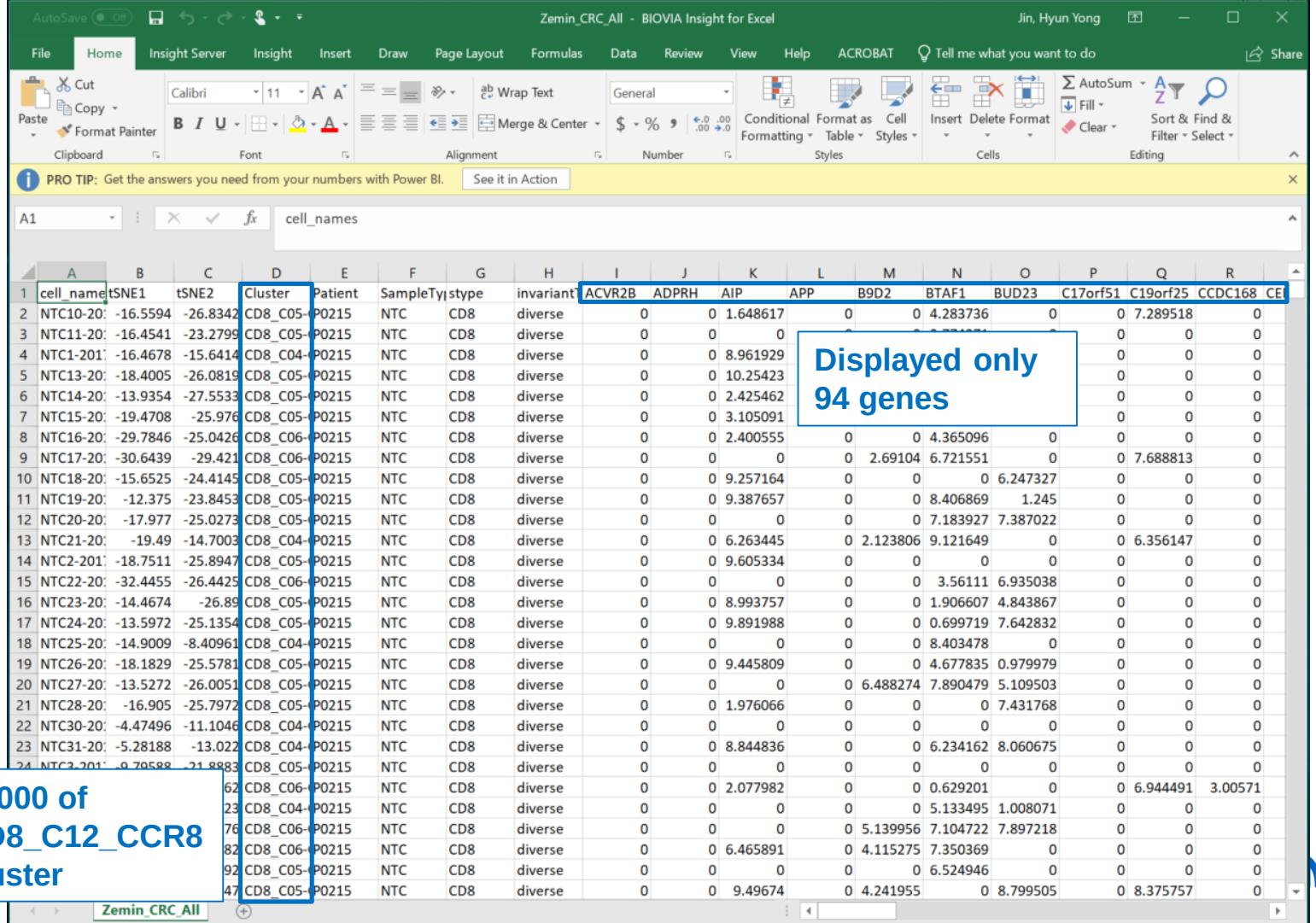
DOWNLOAD DATA FROM TAMATOA V1.4



I manually selected ‘all’ genes and download the table


Zemin_CRC_All - BIOVIA Insight for Excel
Jin, Hyun Yong
File Home Insight Server Insight Insert Draw Page Layout Formulas Data Review View Help ACROBAT Tell me what you want to do Share
PRO TIP: Get the answers you need from your numbers with Power BI. See it in Action
A1
cell_names
Displayed only 94 genes
~1000 of CD8_C12_CCR8 cluster
Zemin_CRC_All

RAW TABLE FOR DOWNSTREAM ANALYSIS

Detected genes: Including CCR8

All Cluster: Including CD4_C12_CC8 cluster

Log (TPM+1) values

START FROM CALCULATED TPM: GSE108989

Series GSE108989 Query DataSets for GSE108989

Status	Public on Oct 29, 2018
Title	Lineage tracking reveals dynamic relationships of T cells in colorectal cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	T cells are central players in cancer immunotherapy1, yet some of their fundamental properties such as development and migration within tumours remain elusive. The enormous T cell receptor (TCR) repertoire, required for recognising foreign and self-antigens2,3, could serve as lineage tags to track these T cells in tumours4. Here, we obtained transcriptomes of 11,138 single T cells from 12 colorectal cancer (CRC) patients and developed STARTRAC (Single T-cell Analysis by Rna-seq and Tcr TRACking) indices to quantitatively analyse dynamic relationships among 20 identified T cell subsets with distinct functions and clonalities. While both CD8+ effector and ?exhausted? T cells exhibited high clonal expansion, they were independently connected with tumour-resident CD8+ effector memory cells, implicating a TCR-based fate decision. Of the CD4+ T cells, the majority of tumour-infiltrating Tregs showed clonal exclusivity, whereas certain Treg clones were developmentally linked to multiple TH clones. Notably, we identified two IFNG+ TH1-like clusters in tumours, the GZMK+ TEM and CXCL13+ TH1-like clusters, which were associated with distinct IFN-?-regulating transcription factors, EOMES/RUNX3 and BHLHE40, respectively. Only BHLHE40+ CXCL13+ TH1-like cells were preferentially enriched in tumours of microsatellite-instable (MSI) patients, which might explain their favourable response rates to immune-checkpoint blockade. Furthermore, we found IGFLR1 to be highly expressed in both BHLHE40+CXCL13+ TH1-like and CD8+ exhausted T cells and possessed co-stimulatory functions. Our integrated STARTRAC analyses provided a powerful avenue to comprehensively dissect the T cell properties in CRC, which could shed new insights into the dynamic relationships of T cells in other cancers.

Supplementary file	Size	Download	File type/resource
GSE108989_CRC.TCell.S10805.norm.centered.txt.gz	368.5 Mb	(ftp)(http)	TXT
GSE108989_CRC.TCell.S11138.TPM.txt.gz	351.6 Mb	(ftp)(http)	TXT
GSE108989_CRC.TCell.S11138.count.txt.gz	69.7 Mb	(ftp)(http)	TXT

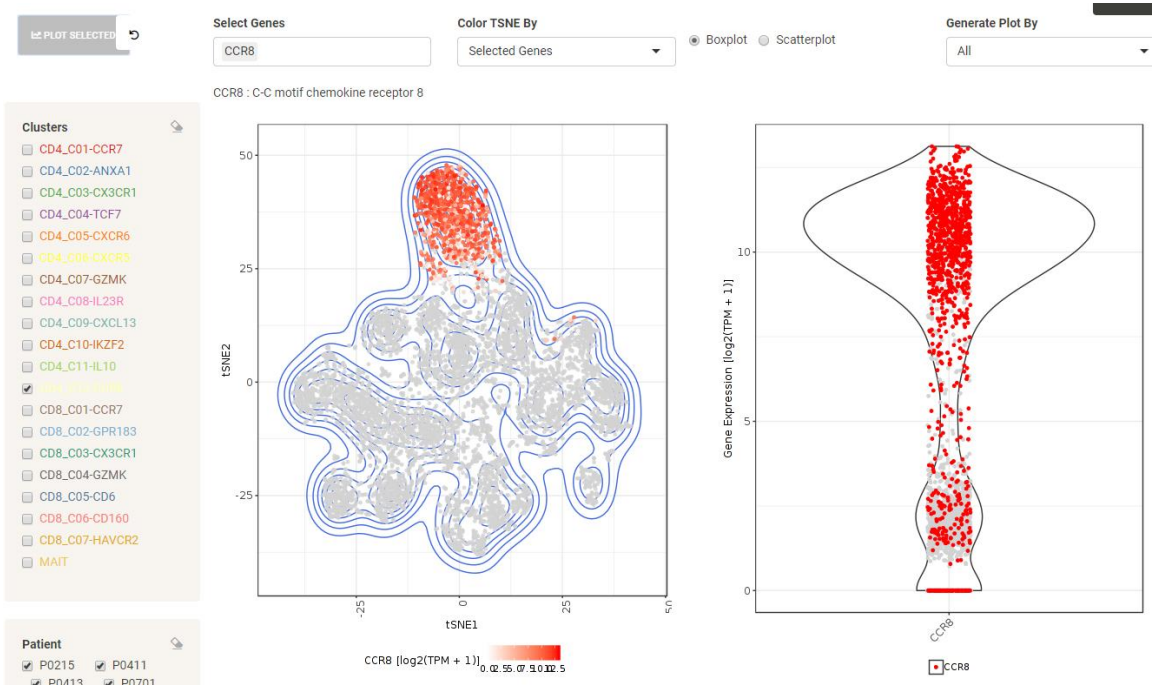
```
> tib[,1:5] #Column 2 is gene name and header is single cell identifiers
# A tibble: 12,547 x 5
      geneID geneSymbol  NP710.20180123 NP711.20180123 NP71.20180123
       <int> <fct>                <dbl>          <dbl>         <dbl>
 1         1 A1BG                -0.515         -0.515          5.52
 2       100 ADA                 -1.86          -1.86          -1.86
 3     10000 AKT3                -0.458         -0.458         -0.458
 4 100009676 ZBTB11-AS1          -0.663         -0.663         -0.663
 5     10001 MED6                -1.04          -1.04           7.16
 6     10003 NAALAD2             -0.103         -0.103         -0.103
 7 100033438 SNORD116-26         -0.0633        -0.0633        -0.0633
 8 100037417 DDTL                 3.79          -1.18          -1.18
 9     10004 NAALADL1            -0.317         -0.317         -0.317
10 100048912 CDKN2B-AS1          -0.0624        -0.0624        -0.0624
# ... with 12,537 more rows
```

```
> tib[,10802:10807]
# A tibble: 12,547 x 6
   PTC9.20161228 PTC92.20161228 PTC93.20161228 PTC94.20161228 PTC95.20161228 PTC96.20161228
           <dbl>          <dbl>          <dbl>          <dbl>          <dbl>          <dbl>
 1          2.48           2.51         -0.711           3.29         -0.711         -0.711
 2         -1.29           7.73          0.494         -1.29           6.64          -1.29
 3          7.16         -0.880           1.68         -0.880         -0.880         -0.880
 4          1.90         -0.725         -0.725         -0.725           1.14         -0.725
 5          5.76          -1.00         -0.881           5.54          -2.03           5.08
 6       -0.0687       -0.0687        -0.0687        -0.0687        -0.0687        -0.0687
 7        -0.101        -0.101         -0.101         -0.101         -0.101         -0.101
 8          2.84           1.84           2.87          -1.21           2.48           3.48
 9        -0.416           6.15         -0.416         -0.416         -0.416         -0.416
10        -0.185         -0.185         -0.185         -0.185         -0.185         -0.185
# ... with 12,537 more rows
```

1. Downloaded raw table from GSE108989. This table contained normalized gene expression (12547 genes) of single cells (10807 cells).
2. Downloaded table from Tamatoa. This table includes individual cell ID and cluster information, but less number of cells (7172). I assume they removed some cells with less confident analysis.
3. Merged two tables. Now I have individual cell (7172) with gene expression profile (12546, removed one un-assigned gene).
4. Select cells assigned to “CD4_C12-CCR8”. Down to 1042 cells.
5. Starting from the CCR8 cluster, I separated the individual cells into two groups. Cells belong to CCR8hi (log2 >8, 330 cells) and CCR8 low (1<log2<4, 47 cells). About half of the cells does not even have significant CCR8 but still clustered as same cluster because other gene expression patterns contributed to the clustering. I focused on cells with significant CCR8 expression.
6. From this point on, I treated the individual cells from CCR8hi group (330 cells) and low group (47 cells) as biological replicates for calculating statistics.
7. I calculated mean, FC, SD, p values and other statistics per individual genes.
8. From this stat(stat_all.csv), I selected FC > 5 and p values <0.01 genes. This table is attached, showing upregulated gene list in CCR8hi cells. CCL22 was the top hit and CCR8 was the third hit.
9. From this stat(stat_all.csv), I selected FC<0.2 genes and p values <0.2 . This table shows downregulated gene list in CCR8hi cells. Stat is very loosened because lowly detected genes have very poor statistics. SIRT1 was downregulated.

- Comments: Many cells with low abundant TPM contains exactly same values. Statistics from these values may not represent true statistics. Removing cells with low abundant values are not feasible because essentially all the cells contain some levels of low abundant mRNAs. In these case, fold-change could be more reliable values.
- Following is additional script I wrote this morning. Again, analysis log will be generated after everything is done. Due to the file size, generating analysis log takes significant computation time.

CCR8 HI AND LOW CUT-OFF



```
> merged[c(1:8), c(1:12)]
      cell_names     tSNE1     tSNE2       Cluster Patient SampleType stype invariantTCR
1 NTC10.20170215 -16.55942 -26.83424   CD8_C05-CD6   P0215        NTC   CD8      diverse
2 NTC11.20170215 -16.45410 -23.27989   CD8_C05-CD6   P0215        NTC   CD8      diverse
3  NTC1.20170215 -16.46778 -15.64138  CD8_C04-GZMK   P0215        NTC   CD8      diverse
4 NTC13.20170215 -18.40049 -26.08195   CD8_C05-CD6   P0215        NTC   CD8      diverse
5 NTC14.20170215 -13.93536 -27.55328   CD8_C05-CD6   P0215        NTC   CD8      diverse
6 NTC15.20170215 -19.47082 -25.97597   CD8_C05-CD6   P0215        NTC   CD8      diverse
7 NTC16.20170215 -29.78458 -25.04256 CD8_C06-CD160   P0215        NTC   CD8      diverse
8 NTC17.20170215 -30.64388 -29.42102 CD8_C06-CD160   P0215        NTC   CD8      diverse
          Units      A1BG       ADA         AKT3
1 log2(TPM + 1)  3.904416 -1.429880 -0.003865219
2 log2(TPM + 1) -1.077304  6.728823 -0.793387736
3 log2(TPM + 1) -1.077304 -2.219403  4.671928970
4 log2(TPM + 1) -1.077304 -2.219403 -0.793387736
5 log2(TPM + 1) -1.077304 -2.219403  0.636907032
6 log2(TPM + 1) -1.077304 -2.219403 -0.793387736
7 log2(TPM + 1) -1.077304 -2.219403  7.966519913
8 log2(TPM + 1) -1.077304  2.386216 -0.793387736
>
```

DETAILED ANALYSIS LOG

<Knitter 1>

Zemin_CRC_GSE108989-CCR8_Analysis

Hyun Yong Jin

August 6, 2019

Ver1.1 as of 20190808 Code readability has been improved.

1. Downloaded raw table from GSE108989. This table contained normalized gene expression (12547 genes) of single cells (10807 cells).
2. Downloaded table from Tamatoa. This table includes individual cell ID and cluster information, but less number of cells (7172). I assume they removed some cells with less confident analysis.
3. Merged two tables. Now I have individual cell (7172) with gene expression profile (12546, removed one un-assigned gene).
4. Select cells assigned to CD4_C12-CCR8. Down to 1042 cells.
5. Starting from the CCR8 cluster, I separated the individual cells into two groups. Cells belong to CCR8hi (log2 >8, 330 cells) and CCR8 low (1<log2<4, 47 cells). About half of the cells does not even have significant CCR8 but still clustered as same cluster because other gene expression patterns contributed to the clustering. I focused on cells with significant CCR8 expression.
6. From this point on, I treated the individual cells from CCR8hi group (330 cells) and low group (47 cells) as biological replicates for calculating statistics.
7. I calculated mean, FC, SD, p values and other statistics per individual genes.
8. From this stat(stat_all.csv), I selected FC > 5 and p values <0.01 genes. This table is attached, showing upregulated gene list in CCR8hi cells. CCL22 was the top hit and CCR8 was the third hit.
9. From this stat(stat_all.csv), I selected FC<0.2 genes and p values <0.2 . This table shows downregulated gene list in CCR8hi cells. Stat is very loosened because lowly detected genes have very poor statistics. SIRT1 was downregulated.

```
knitr::opts_chunk$set(fig.width=12, fig.height=8, fig.path='Output/',
                      warning=FALSE)
```

Download Data

```
#GSE108989
library(GEOquery)
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

<Knitter 2>

Zemin_CRC_GSE108989_visualization

Hyun Yong Jin

August 8, 2019

```
knitr::opts_chunk$set(echo = TRUE)
```

Visualization of from stat_all table

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 3.5.3
```

```
a <- read.csv("stat_all.csv", header=T, stringsAsFactors = FALSE)

a$Significant <- ifelse((a$FC>18&a$p.value<0.01)|a$p.value<1/10^15, "Significant", "Not Sig")

pl<-ggplot(a, aes(x = log2(FC), y = -log10(p.value))) +
  geom_point(aes(color = Significant)) +
  scale_color_manual(values = c("grey", "red")) +
  theme_bw(base_size = 12) + theme(legend.position = "bottom") +
  geom_vline(xintercept=0, linetype="dashed", color = "blue", size=1)+
  geom_vline(xintercept=log2(10), linetype="dashed", color = "blue")+
  geom_vline(xintercept=log2(0.1), linetype="dashed", color = "blue")+
  geom_hline(yintercept=-log10(1/10^15), linetype="dashed", color = "blue")+
  geom_text_repel(
    data = subset(a, (FC>18&p.value<0.01)|p.value<1/10^15),
    aes(label = gene_name),
    size = 5,
    box.padding = unit(0.35, "lines"),
    point.padding = unit(0.3, "lines")
  )
```