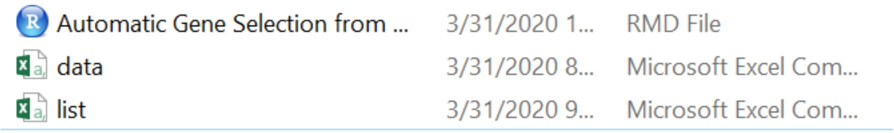
Automatic gene selection from NanoString data with R (Hyun Yong Jin)

1. ***Put a RMD file, data.csv and list.csv file in the same folder.***

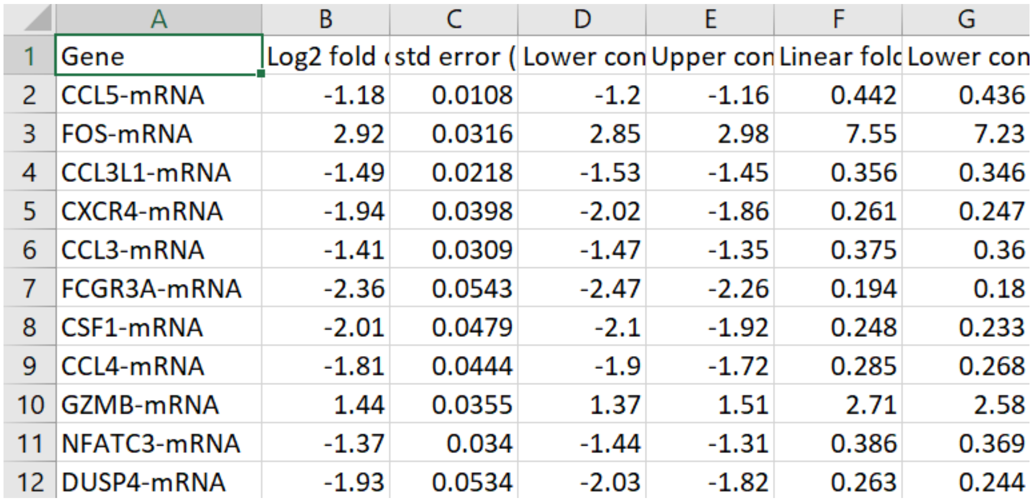


Automatic Gene Selection from ...	3/31/2020 1...	RMD File
data	3/31/2020 8...	Microsoft Excel Com...
list	3/31/2020 9...	Microsoft Excel Com...

1.1. The RMD file contains executable R code.

1.2. data.csv is a two dimensional dataframe containing NanoString data.

1.2.1. data.csv has a top row as a header, and one column MUST have “Gene” as a header.

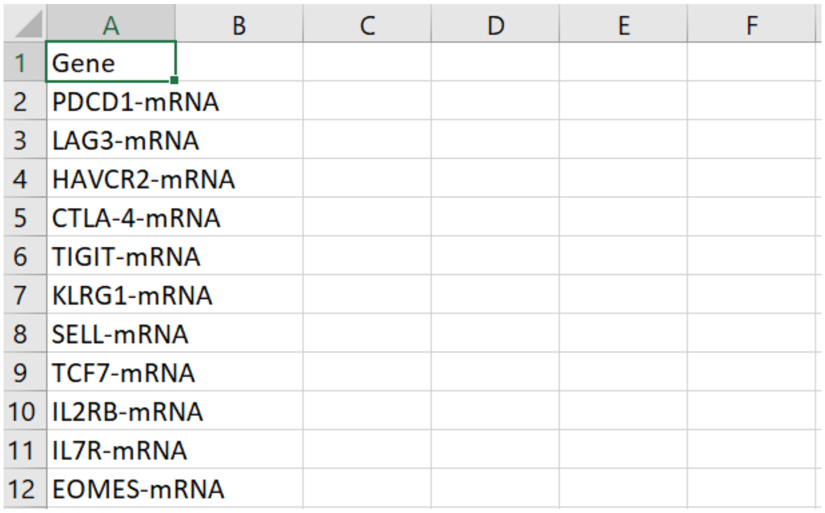


	A	B	C	D	E	F	G
1	Gene	Log2 fold (	std error (	Lower con	Upper con	Linear folc	Lower con
2	CCL5-mRNA	-1.18	0.0108	-1.2	-1.16	0.442	0.436
3	FOS-mRNA	2.92	0.0316	2.85	2.98	7.55	7.23
4	CCL3L1-mRNA	-1.49	0.0218	-1.53	-1.45	0.356	0.346
5	CXCR4-mRNA	-1.94	0.0398	-2.02	-1.86	0.261	0.247
6	CCL3-mRNA	-1.41	0.0309	-1.47	-1.35	0.375	0.36
7	FCGR3A-mRNA	-2.36	0.0543	-2.47	-2.26	0.194	0.18
8	CSF1-mRNA	-2.01	0.0479	-2.1	-1.92	0.248	0.233
9	CCL4-mRNA	-1.81	0.0444	-1.9	-1.72	0.285	0.268
10	GZMB-mRNA	1.44	0.0355	1.37	1.51	2.71	2.58
11	NFATC3-mRNA	-1.37	0.034	-1.44	-1.31	0.386	0.369
12	DUSP4-mRNA	-1.93	0.0534	-2.03	-1.82	0.263	0.244

1.3. list.csv contains list of genes you want to look up

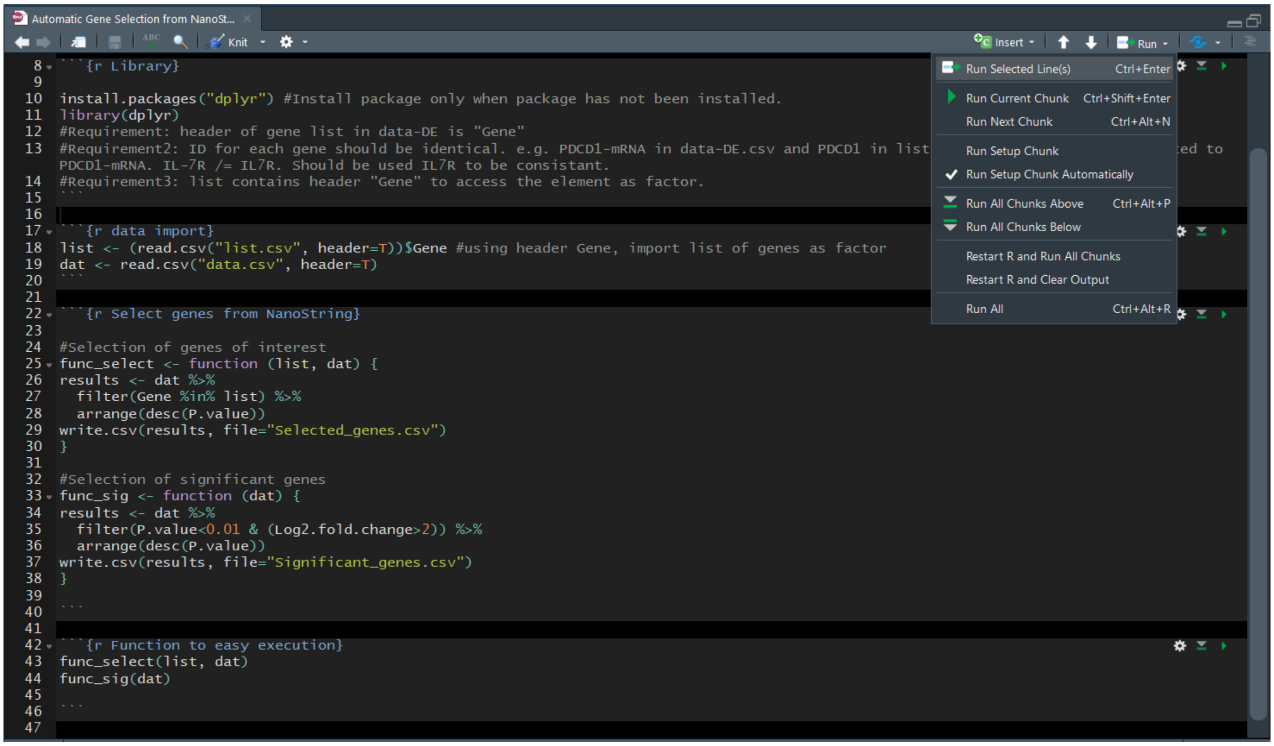
1.3.1. list.csv also has a top row as a header, and must have “Gene” as a header.

1.3.2. Gene name (i.e.PDCD1-mRNA) must be EXACTLY the same in the gene name of data.csv. Case-sensitive.



	A	B	C	D	E	F
1	Gene					
2	PDCD1-mRNA					
3	LAG3-mRNA					
4	HAVCR2-mRNA					
5	CTLA-4-mRNA					
6	TIGIT-mRNA					
7	KLRG1-mRNA					
8	SELL-mRNA					
9	TCF7-mRNA					
10	IL2RB-mRNA					
11	IL7R-mRNA					
12	EOMES-mRNA					

2. ***Open "Automatic Gene Selection from NanoString_20200331.rmd" file in R studio and Click on "Run" button on your upper right hand corner and select "Run All"***


Automatic Gene Selection from NanoSt...
Knit
Insert
Run
Run Selected Line(s) Ctrl+Enter
Run Current Chunk Ctrl+Shift+Enter
Run Next Chunk Ctrl+Alt+N
Run Setup Chunk
✓ Run Setup Chunk Automatically
Run All Chunks Above Ctrl+Alt+P
Run All Chunks Below
Restart R and Run All Chunks
Restart R and Clear Output
Run All Ctrl+Alt+R

~~~
```{r Library}

install.packages("dplyr") #Install package only when package has not been installed.
library(dplyr)
#Requirement: header of gene list in data-DE is "Gene"
#Requirement2: ID for each gene should be identical. e.g. PDCD1-mRNA in data-DE.csv and PDCD1 in list          ted to
PDCD1-mRNA. IL-7R /= IL7R. Should be used IL7R to be consistant.
#Requirement3: list contains header "Gene" to access the element as factor.
~~~

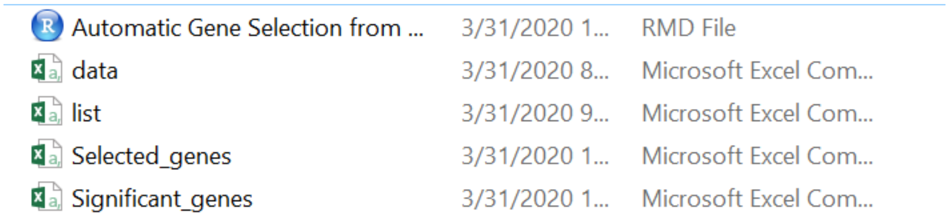
```{r data import}
list <- (read.csv("list.csv", header=T))$Gene #using header Gene, import list of genes as factor
dat <- read.csv("data.csv", header=T)
```

```{r Select genes from NanoString}

#Selection of genes of interest
func_select <- function (list, dat) {
results <- dat %>%
  filter(Gene %in% list) %>%
  arrange(desc(P.value))
write.csv(results, file="Selected_genes.csv")
}

#Selection of significant genes
func_sig <- function (dat) {
results <- dat %>%
  filter(P.value<0.01 & (Log2.fold.change>2)) %>%
  arrange(desc(P.value))
write.csv(results, file="Significant_genes.csv")
}

```

```{r Function to easy execution}
func_select(list, dat)
func_sig(dat)

```

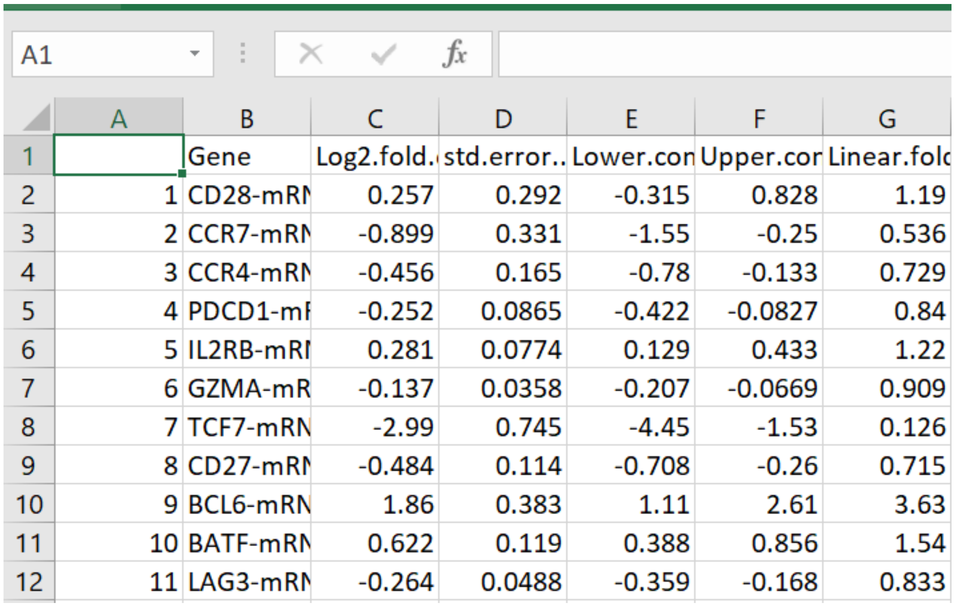
```

2.1. The first code block install and initiate necessary package in R (dplyr). You may delete or comment out (#) "install.packages(dplyr)" once you install dplyr once.

2.2. The second cod block import list.csv and data.csv in R in desired format.

2.3. The third code bock contains two functions. One function select the row from data.csv using gene names derived from list.csv, the other function select rows from data.csv based on calculated statistics.

2.4. The final code block execute functions in the third code block.
```

3. ***Go to the original folder you put a rmd file and two csv file. The resulting two csv file must be generated (Selected_genes.csv and Significant_genes.csv).***



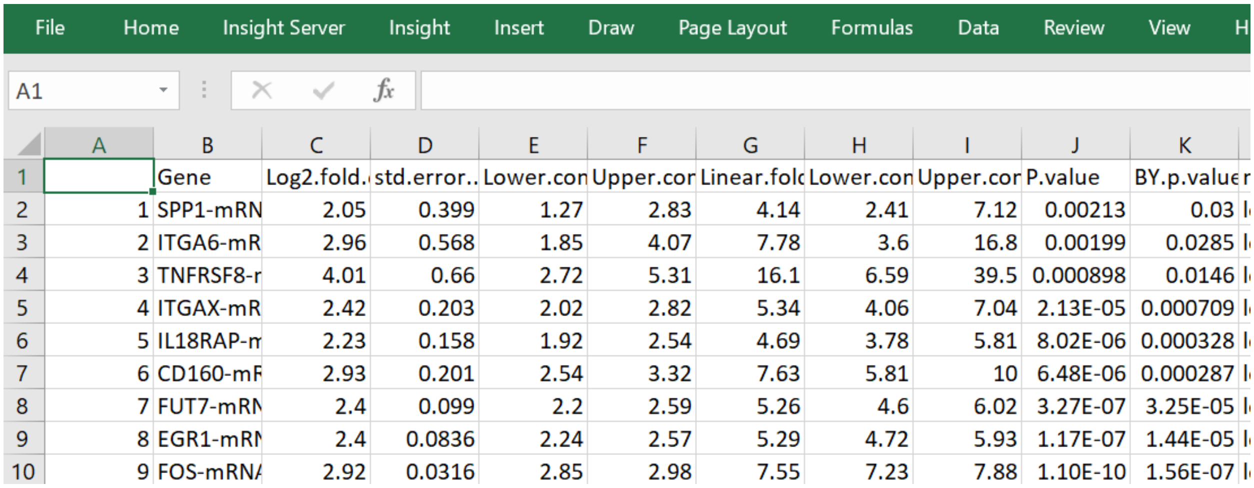
Automatic Gene Selection from ...	3/31/2020 1...	RMD File
data	3/31/2020 8...	Microsoft Excel Com...
list	3/31/2020 9...	Microsoft Excel Com...
Selected_genes	3/31/2020 1...	Microsoft Excel Com...
Significant_genes	3/31/2020 1...	Microsoft Excel Com...

3.1. **Selected_genes.csv** contains results from your list.csv. Note that gene names that was not found in data.csv is automatically ignored.



	A	B	C	D	E	F	G
1		Gene	Log2.fold.(	std.error..	Lower.con	Upper.cor	Linear.fold
2	1	CD28-mRN	0.257	0.292	-0.315	0.828	1.19
3	2	CCR7-mRN	-0.899	0.331	-1.55	-0.25	0.536
4	3	CCR4-mRN	-0.456	0.165	-0.78	-0.133	0.729
5	4	PDCD1-mF	-0.252	0.0865	-0.422	-0.0827	0.84
6	5	IL2RB-mRI	0.281	0.0774	0.129	0.433	1.22
7	6	GZMA-mR	-0.137	0.0358	-0.207	-0.0669	0.909
8	7	TCF7-mRN	-2.99	0.745	-4.45	-1.53	0.126
9	8	CD27-mRN	-0.484	0.114	-0.708	-0.26	0.715
10	9	BCL6-mRN	1.86	0.383	1.11	2.61	3.63
11	10	BATF-mRN	0.622	0.119	0.388	0.856	1.54
12	11	LAG3-mRN	-0.264	0.0488	-0.359	-0.168	0.833

3.2. **Significant_genes.csv** contains list of genes that are significant from data.csv. In the default setting, the list of genes are selected based on fold-change more than 2-fold, and p values less than 0.01. In this tutorial, 9 genes were selected.



	A	B	C	D	E	F	G	H	I	J	K	
1		Gene	Log2.fold.(	std.error..	Lower.con	Upper.cor	Linear.fold	Lower.con	Upper.cor	P.value	BY.p.value	r
2	1	SPP1-mRN	2.05	0.399	1.27	2.83	4.14	2.41	7.12	0.00213	0.03	l
3	2	ITGA6-mR	2.96	0.568	1.85	4.07	7.78	3.6	16.8	0.00199	0.0285	l
4	3	TNFRSF8-r	4.01	0.66	2.72	5.31	16.1	6.59	39.5	0.000898	0.0146	l
5	4	ITGAX-mR	2.42	0.203	2.02	2.82	5.34	4.06	7.04	2.13E-05	0.000709	l
6	5	IL18RAP-m	2.23	0.158	1.92	2.54	4.69	3.78	5.81	8.02E-06	0.000328	l
7	6	CD160-mR	2.93	0.201	2.54	3.32	7.63	5.81	10	6.48E-06	0.000287	l
8	7	FUT7-mRN	2.4	0.099	2.2	2.59	5.26	4.6	6.02	3.27E-07	3.25E-05	l
9	8	EGR1-mRN	2.4	0.0836	2.24	2.57	5.29	4.72	5.93	1.17E-07	1.44E-05	l
10	9	FOS-mRNA	2.92	0.0316	2.85	2.98	7.55	7.23	7.88	1.10E-10	1.56E-07	l