

Data preparation:

Utilizing Ubuntu windows (bash), to clean and prepare the data.

1. Folder *Original_data*

Save all the original datasets which offer by clients.

2. Folder *Clean_data*

Data:

Each patient has 8 datasets, which are activated or resting TCR record in 4 time points, a day before vaccinating and a week after vaccinating in 2016 and 2017.

After cleaning the data, each cdr3 correspond to its V-gene name, counts, and frequency. (cleaned data in sub-folder '2-digits' and '4 digits', as the client's request, the data cleaned into two versions, 2 and 4 digits which represent the digits number of V-Gene name.)

If test needed:

```
2244 Oct 4 20:45 data_clean_2d
2243 Sep 3 20:15 data_clean_4d
141 Sep 3 20:10 run_2d
141 Sep 3 20:10 run_4d
```

There are 4 bash script in folder Clean_data.

```
2. Clean data$ ./run_2d 2_digits
```

Just type (**./run_2d new_folder**), notice the first parameter is new folder name, don't use the already existing one.

The script will create a new folder, all cleaned data would save in that folder.

```
2. Clean data$ ./run_4d 4_digits
```

Same as 4 digits.

```
2. Clean data$ chmod +rwx *
```

PS: if meet the permission limited.

3. Folder *data_preparation*

2016 d0 (a day before vaccinate)	2016 d7 (a week after vaccinate)	2017 d0 (a day before vaccinate)	2017 d7 (a week after vaccinate)	ac-means activated re-means resting
Ex: acac = 2016d0 activated;2016 d7 activated				
ac ac		ac ac	L1111	ac -> activated in 7 days / ac -> activated in 7days
ac ac		re ac	L1101	ac -> activated in 7 days / re -> activated in 7 days
ac ac		re re	L1100	ac -> activated in 7 days / re -> resting
ac ac		ac re	L1110	ac -> activated in 7 days / ac -> resting
re ac		ac ac	L0111	re -> activated in 7 days / ac -> activated in 7days
re ac		re ac	L0101	re -> activated in 7 days / re -> activated in 7 days
re ac		re re	L0100	re -> activated in 7 days / re -> resting
re ac		ac re	L0110	re -> activated in 7 days / ac -> resting
re re		ac ac	L0011	re -> resting / ac -> activated in 7days
re re		re ac	L0001	re -> resting / re -> activated in 7days
re re		re re	L0000	re -> resting / re -> resting
re re		ac re	L0010	re -> resting / ac -> resting
ac re		ac ac	L1011	ac -> resting / ac -> activated in 7days
ac re		re ac	L1001	ac -> resting / re -> activated in 7days
ac re		re re	L1000	ac -> resting / re -> resting
ac re		ac re	L1010	ac -> resting / re -> resting

FigureA: Label activated TCR as '1', and resting as '0'.

Data:

Then Intersect data sets (common cdr3 as the main key), we get 16 kinds of responses (FigureA). In joined dataset (common cdr3), except cdr3 and V-gene names, we have 8 attributes columns which are 4 time points' TCR count and frequency, and a target column which is the response.

If test needed:

In folder common cdr3, for example in the sub-folder 'VGene-2digits', which include cleaned data and several bash script.

```
VGene-2digits$ ./cmcdm 06 2016 2017 2d
VGene-2digits$ ./cmcdm 08 2016 2017 2d
VGene-2digits$ ./cmcdm 10 2016 2017 2d
VGene-2digits$
```

While we just need to run script 'cmcdm', the **first parameter** is the patient number, **second** is the year 2016, **third** is the year 2017, **fourth** is 2d which represents handle 2 digits V-Gene datasets.

comcdm_2d_Sbj_05
comcdm_2d_Sbj_06
comcdm_2d_Sbj_08
comcdm_2d_Sbj_10
comcdm_2d_Sbj_11
comcdm_2d_Sbj_14

The common cdr3 dataset would appear in the same folder, Which would be used in machine learning and analysis.

```
VGene-4digits$ ./cmcdm 06 2016 2017 4d
VGene-4digits$ ./cmcdm 08 2016 2017 4d
VGene-4digits$ ./cmcdm 10 2016 2017 4d
```

Generate 4 digits datasets is the same.