Data arrangement

Christine Chang

Dec. 29th, 2023

christine.chang.c@gmail.com

Preparation

1. Prepare a raw dataset of **RNA-Seq**

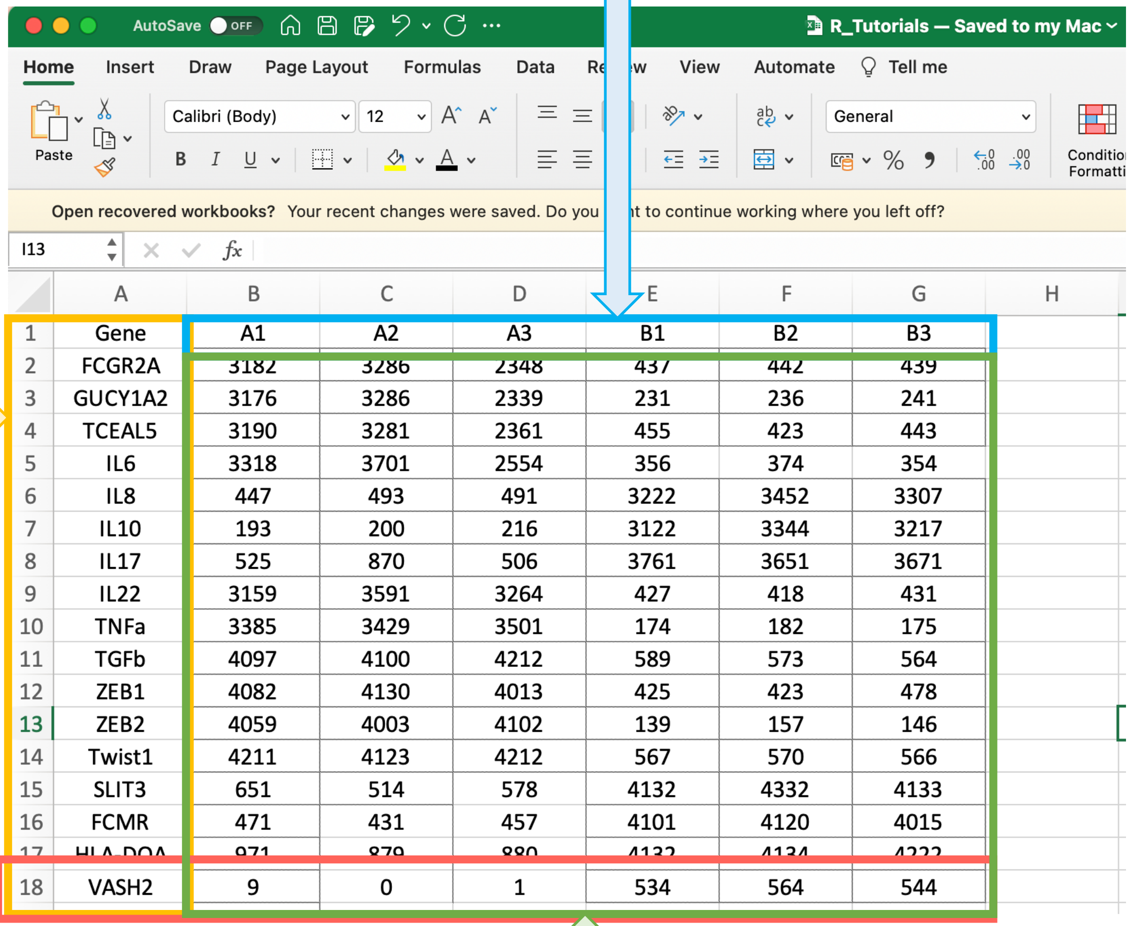
- Both RNA-Seq and small RNA-Seq are okay
- You can download dataset from https://www.ncbi.nlm.nih.gov/gds

Arrange the raw data

1. Select **Gene, (Identifier), Total count**

2. Arrange as a new dataframe (clean dataset)

Sample_Name



Open recovered workbooks? Your recent changes were saved. Do you nt to continue working where you left off?

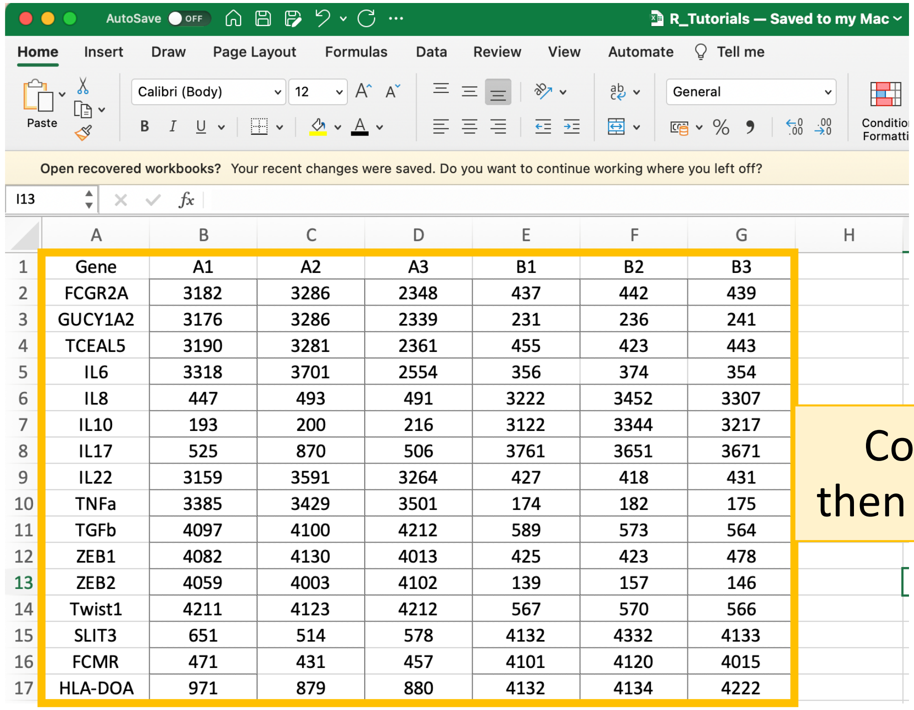
	A	B	C	D	E	F	G	H
1	Gene	A1	A2	A3	B1	B2	B3	
2	FCGR2A	3182	3286	2348	437	442	439	
3	GUCY1A2	3176	3286	2339	231	236	241	
4	TCEAL5	3190	3281	2361	455	423	443	
5	IL6	3318	3701	2554	356	374	354	
6	IL8	447	493	491	3222	3452	3307	
7	IL10	193	200	216	3122	3344	3217	
8	IL17	525	870	506	3761	3651	3671	
9	IL22	3159	3591	3264	427	418	431	
10	TNFa	3385	3429	3501	174	182	175	
11	TGFb	4097	4100	4212	589	573	564	
12	ZEB1	4082	4130	4013	425	423	478	
13	ZEB2	4059	4003	4102	139	157	146	
14	Twist1	4211	4123	4212	567	570	566	
15	SLIT3	651	514	578	4132	4332	4133	
16	FCMR	471	431	457	4101	4120	4015	
17	HLA-DOA	971	879	880	4132	4134	4222	
18	VASH2	9	0	1	534	564	544	

Gene_Name

The expression 0 will cause the “Valuerror” when computing matrix element in some algorithm, therefore, those gene will be excluded in the clean dataset

Total count, count per million (CPM) of each sample

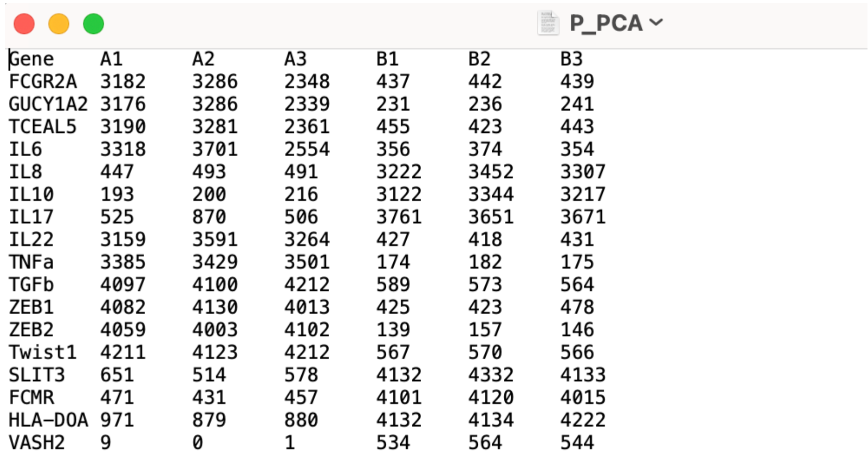
Create the .txt file for the dataset



Open recovered workbooks? Your recent changes were saved. Do you want to continue working where you left off?

	A	B	C	D	E	F	G	H
1	Gene	A1	A2	A3	B1	B2	B3	
2	FCGR2A	3182	3286	2348	437	442	439	
3	GUCY1A2	3176	3286	2339	231	236	241	
4	TCEAL5	3190	3281	2361	455	423	443	
5	IL6	3318	3701	2554	356	374	354	
6	IL8	447	493	491	3222	3452	3307	
7	IL10	193	200	216	3122	3344	3217	
8	IL17	525	870	506	3761	3651	3671	
9	IL22	3159	3591	3264	427	418	431	
10	TNFa	3385	3429	3501	174	182	175	
11	TGFb	4097	4100	4212	589	573	564	
12	ZEB1	4082	4130	4013	425	423	478	
13	ZEB2	4059	4003	4102	139	157	146	
14	Twist1	4211	4123	4212	567	570	566	
15	SLIT3	651	514	578	4132	4332	4133	
16	FCMR	471	431	457	4101	4120	4015	
17	HLA-DOA	971	879	880	4132	4134	4222	

Copy the clean data then paste on the .txt file



```
Gene    A1      A2      A3      B1      B2      B3
FCGR2A  3182    3286    2348    437     442     439
GUCY1A2 3176    3286    2339    231     236     241
TCEAL5  3190    3281    2361    455     423     443
IL6     3318    3701    2554    356     374     354
IL8     447     493     491     3222    3452    3307
IL10    193     200     216     3122    3344    3217
IL17    525     870     506     3761    3651    3671
IL22    3159    3591    3264    427     418     431
TNFa    3385    3429    3501    174     182     175
TGFb    4097    4100    4212    589     573     564
ZEB1    4082    4130    4013    425     423     478
ZEB2    4059    4003    4102    139     157     146
Twist1  4211    4123    4212    567     570     566
SLIT3   651     514     578     4132    4332    4133
FCMR    471     431     457     4101    4120    4015
HLA-DOA 971     879     880     4132    4134    4222
VASH2   9       0       1       534     564     544
```

* The .txt file will be used in the following analysis