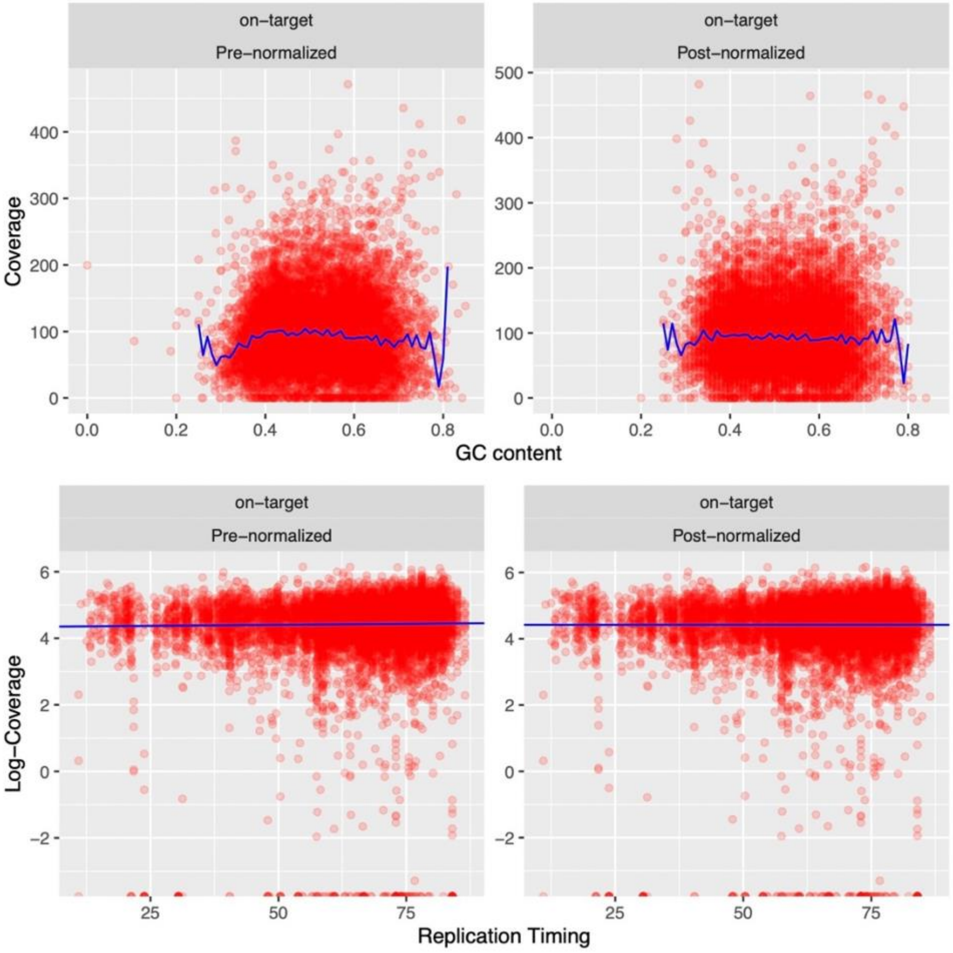
Aung Myat Phyo 22221184
Practical 3

3.



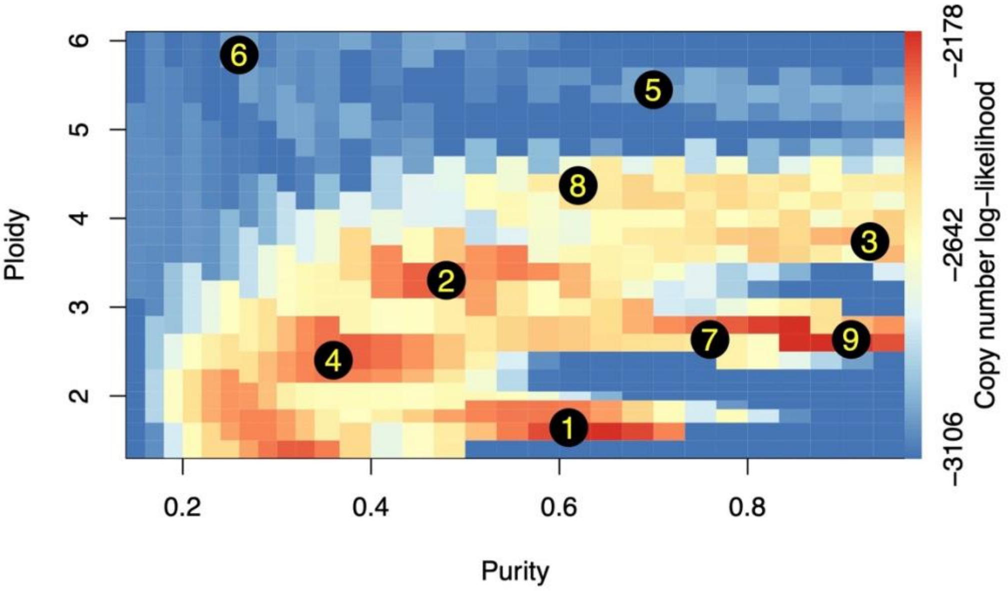
GC-content is one of the most important library biases that means regions of high GC or AT content are not exactly produced with the same efficiency in tumour and normal. Especially, high GC content regions result in PCR amplification biases and sequencing errors and low GC regions probably have lower sequencing depth and coverage. GC normalization is important to remove these biases by adjusting the read counts on its GC content. Finally, GC normalization is important in the high throughput sequencing data downstream analysis. In the above plot, there is a difference between pre normalized and post normalized GC content.

4. Normal database gives a baseline to identify and remove germline CNVs of the general population. By comparing CNVs in tumour samples and CNVs in normal database, it can distinguish between somatic CNVs of specific tumour and germline CNVs in both tumour and normal samples. As a result, it can reduce false positive rate and increase specificity. Moreover, normal database identify and filter technical artifacts and sequences biases of the data.

5. To get high accuracy in CNV detection, PureCN needs high quality annotations such as dbSNP annotation (common genetic variations in human). dbSNP information identify and

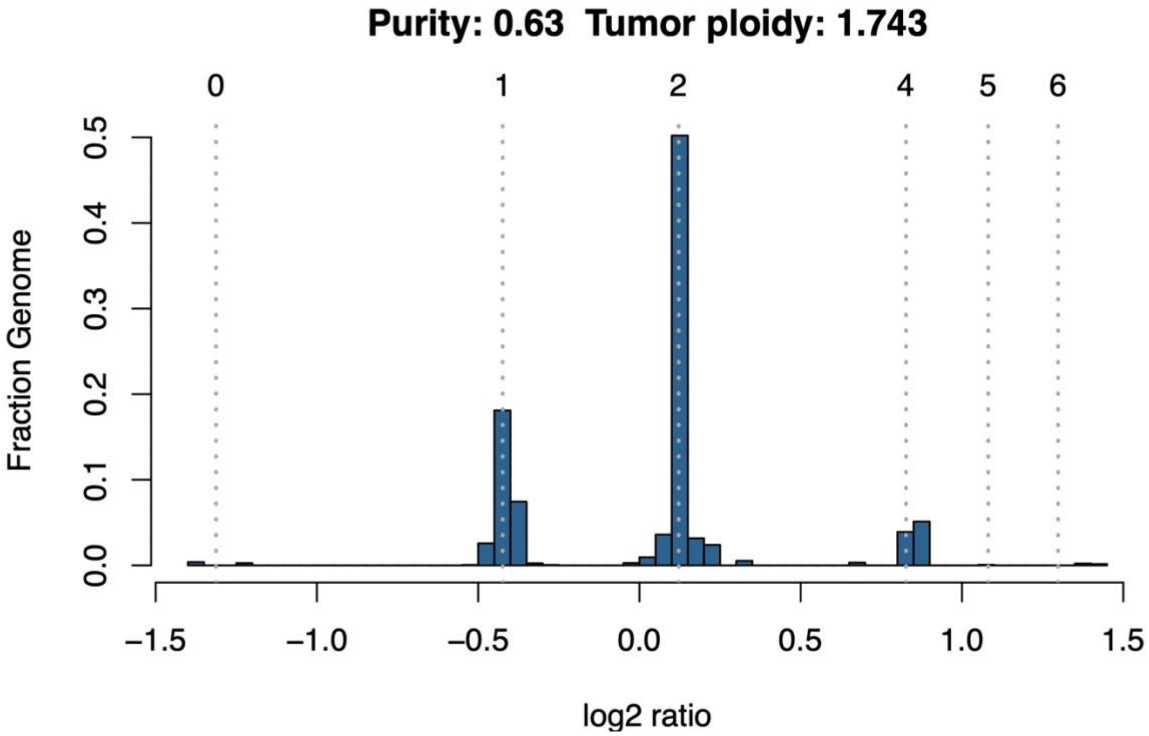
exclude genome regions with frequent genome variations and genomic polymorphisms which disturb CNV detection. As a result, PureCN decrease false positive CNV calls and increase accuracy and sensitivity of CNV detection (for example, cancer genomics).

6.



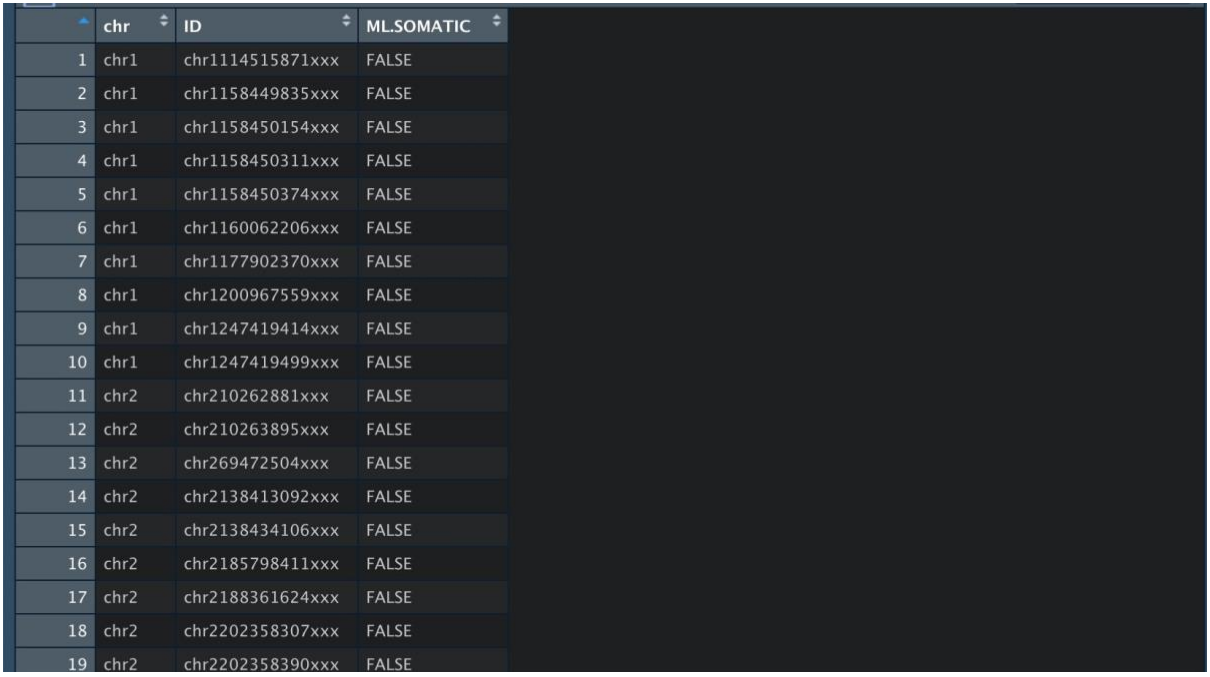
The most likely values for purity and ploidy for this sample is 0.62 and 1.6 respectively.

7.



a. 0.5 (50%) of genome has a diploid copy number.
b. 0.18 (18%) of genome has copy number of 1.
c. 0 of genome has copy number of 3.

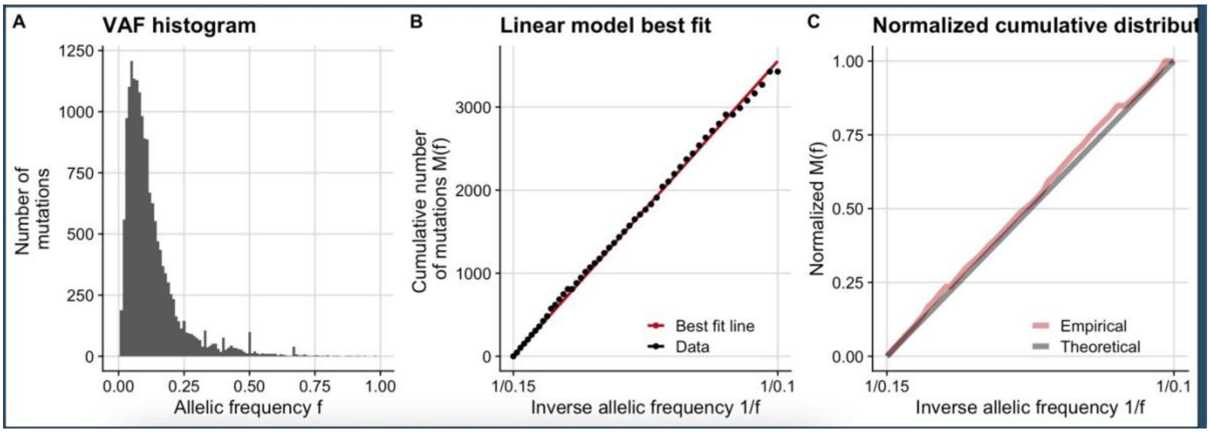
9. There is no any somatic mutations present.
In ML.SOMATIC column, all data are false.



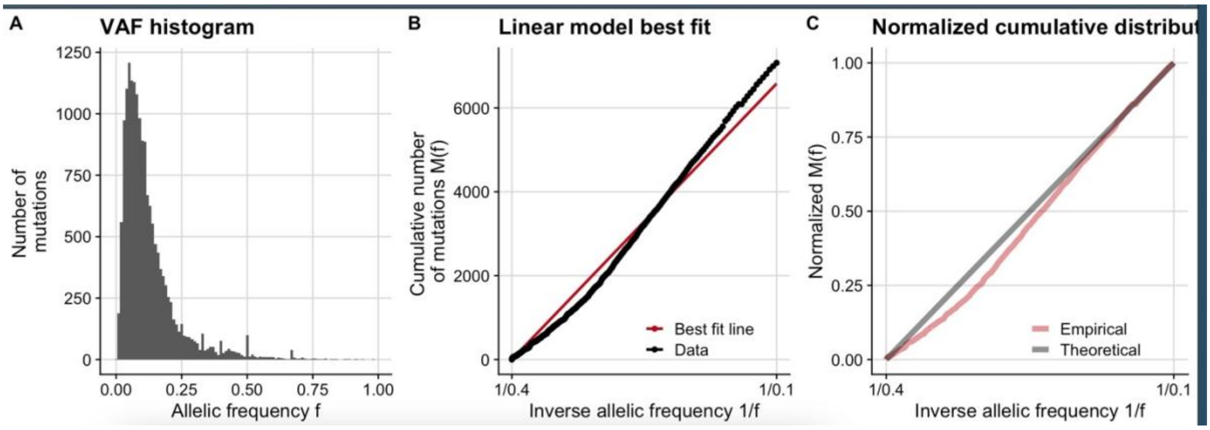
	chr	ID	ML.SOMATIC
1	chr1	chr1114515871xxx	FALSE
2	chr1	chr1158449835xxx	FALSE
3	chr1	chr1158450154xxx	FALSE
4	chr1	chr1158450311xxx	FALSE
5	chr1	chr1158450374xxx	FALSE
6	chr1	chr1160062206xxx	FALSE
7	chr1	chr1177902370xxx	FALSE
8	chr1	chr1200967559xxx	FALSE
9	chr1	chr1247419414xxx	FALSE
10	chr1	chr1247419499xxx	FALSE
11	chr2	chr210262881xxx	FALSE
12	chr2	chr210263895xxx	FALSE
13	chr2	chr269472504xxx	FALSE
14	chr2	chr2138413092xxx	FALSE
15	chr2	chr2138434106xxx	FALSE
16	chr2	chr2185798411xxx	FALSE
17	chr2	chr2188361624xxx	FALSE
18	chr2	chr2202358307xxx	FALSE
19	chr2	chr2202358390xxx	FALSE

Neutral Evolutionary Model

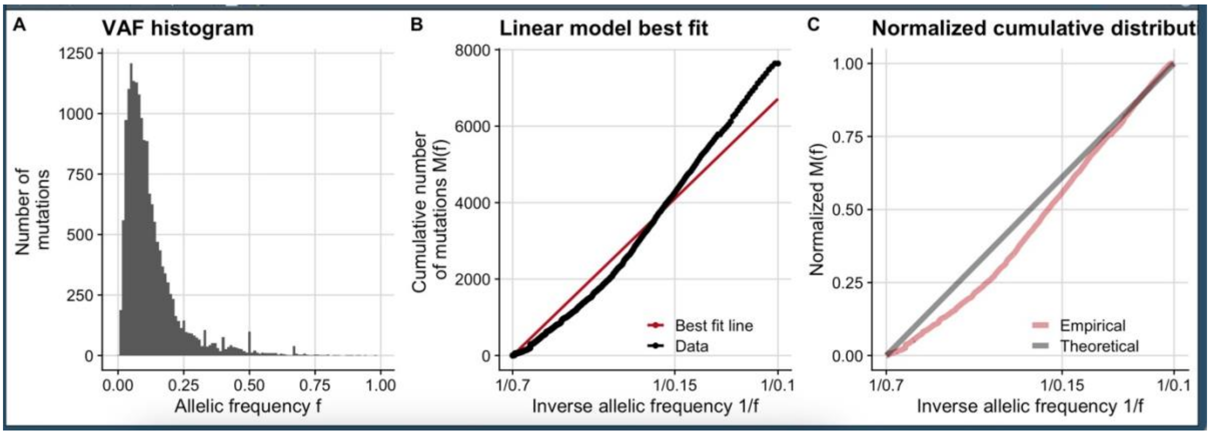
fmax=0.15



fmax=0.4



fmax=0.7



4.a According to the histogram, most number of mutations are under 0.25. Therefore, we should choose around 0.25 for fmax.

B There is an evidence of positive selection acting on subclonal mutation because of linear model best fit line.