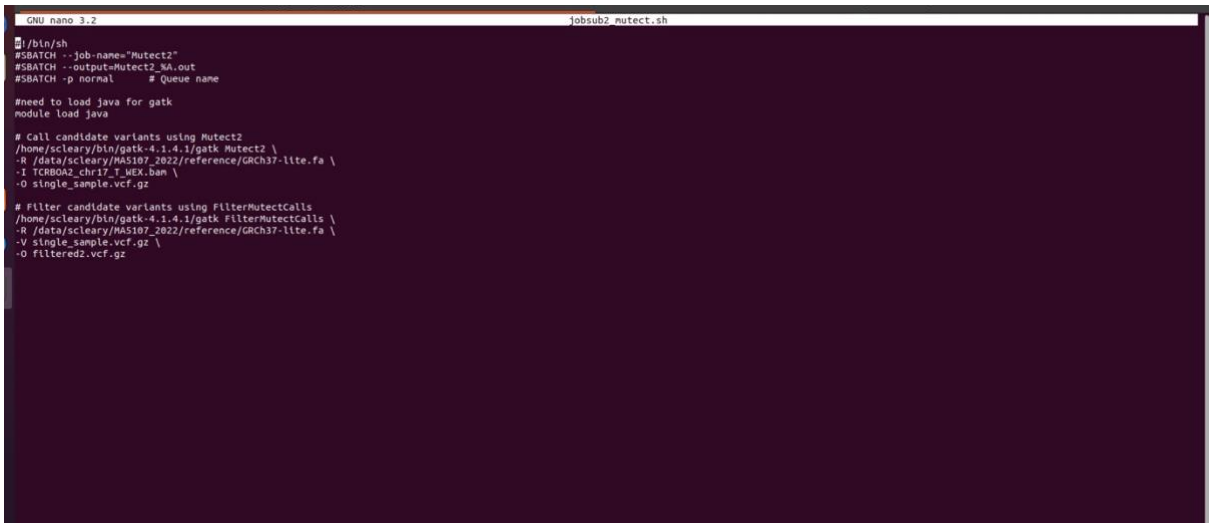
Practical 1
Aung Myat Phyo Student ID 22221184

3.

4. Germline variants calling is direct. They vary against the reference. For somatic variants calling, it compares between two samples against the reference. Germline calling assumes fixed ploidy and calling consists of genotyping sites while somatic calling allows for varying ploidy. Somatic variants calling are both (a) different from control sample and (ii) different from the reference. Somatic variants calling(Mutect2) does not offer for the calculation of reference confidence which is a feature in Germline calling (HaplotypeCaller).

5. Mutect2 offers a joint analysis of multiple samples. SNVs and small indels can be detetcted via Mutect2. Mutect2 applies Bayesian somatic genotyping model while Mutect uses Bayesian classifier.

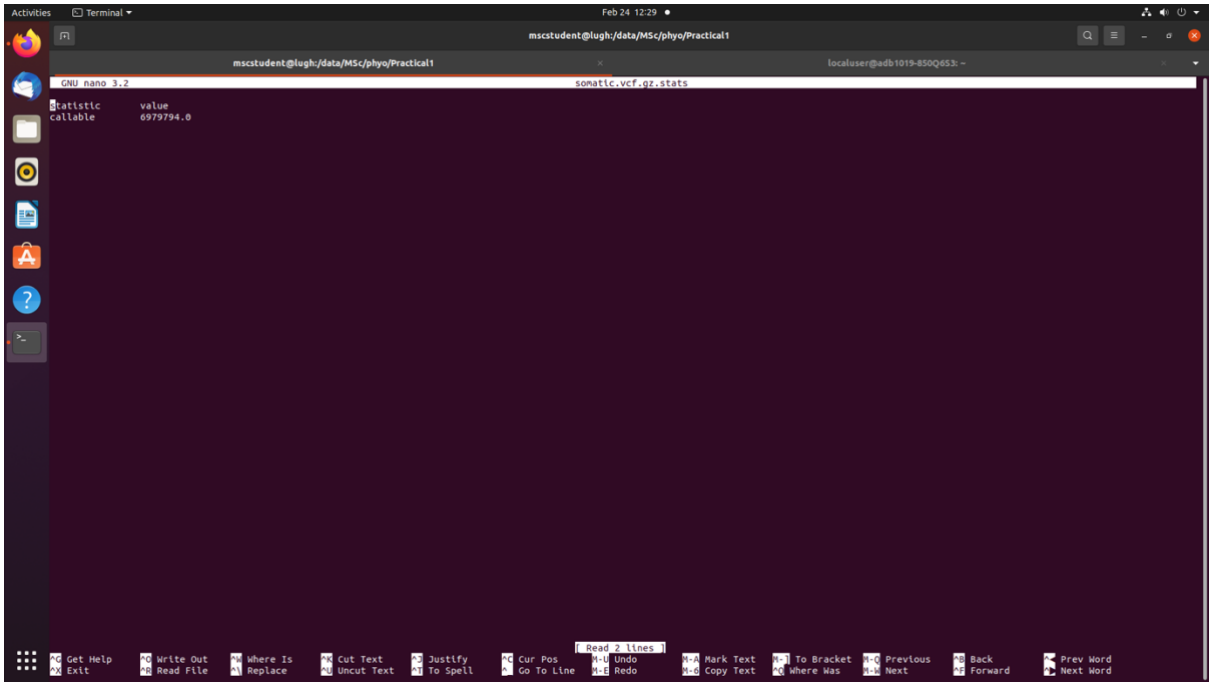
6.

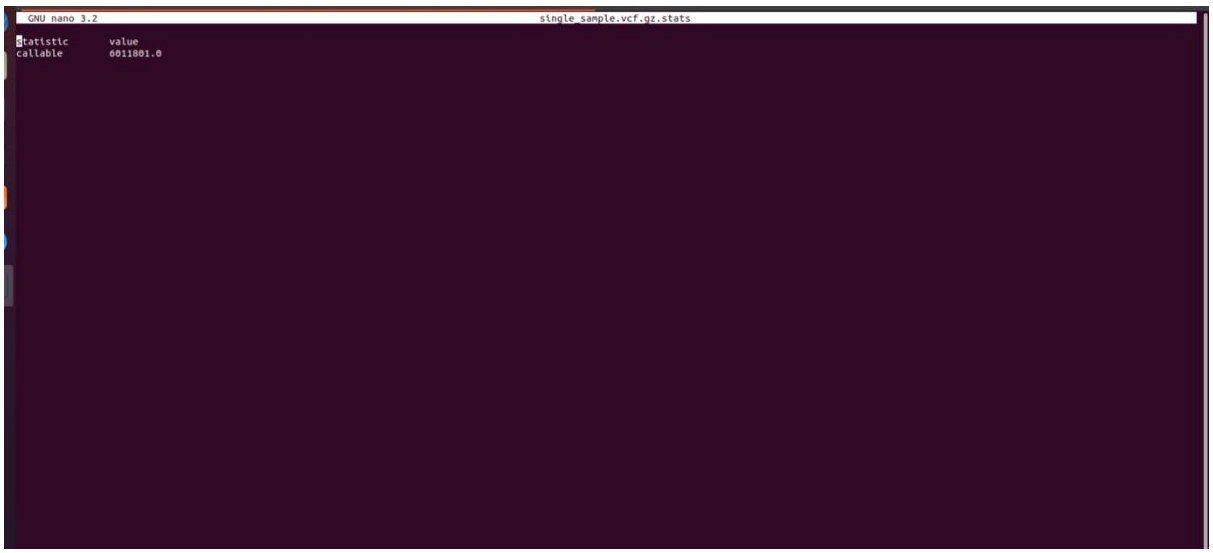


7.

```
 GNU nano 3.2                                                    filtered.vcf.gz.filteringStats.tsv

#<METADATA>Ln prior of deletion of length 10=-20.72326583694641
#<METADATA>Ln prior of deletion of length 9=-20.72326583694641
#<METADATA>Ln prior of deletion of length 8=-20.72326583694641
#<METADATA>Ln prior of deletion of length 7=-20.72326583694641
#<METADATA>Ln prior of deletion of length 6=-20.72326583694641
#<METADATA>Ln prior of deletion of length 5=-20.72326583694641
#<METADATA>Ln prior of deletion of length 4=-20.72326583694641
#<METADATA>Ln prior of deletion of length 3=-20.72326583694641
#<METADATA>Ln prior of deletion of length 2=-20.72326583694641
#<METADATA>Ln prior of deletion of length 1=-20.72326583694641
#<METADATA>Ln prior of SNV=-15.065382788849113
#<METADATA>Ln prior of insertion of length 1=-20.72326583694641
#<METADATA>Ln prior of insertion of length 2=-20.72326583694641
#<METADATA>Ln prior of insertion of length 3=-20.72326583694641
#<METADATA>Ln prior of insertion of length 4=-20.72326583694641
#<METADATA>Ln prior of insertion of length 5=-20.72326583694641
#<METADATA>Ln prior of insertion of length 6=-20.72326583694641
#<METADATA>Ln prior of insertion of length 7=-20.72326583694641
#<METADATA>Ln prior of insertion of length 8=-20.72326583694641
#<METADATA>Ln prior of insertion of length 9=-20.72326583694641
#<METADATA>Ln prior of insertion of length 10=-20.72326583694641
#<METADATA>High-AF beta-binomial cluster=weight = 0.2000, alpha = 10.00, beta = 1.00
#<METADATA>Background beta-binomial cluster=weight = 0.4000, alpha = 0.60, beta = 1.22
#<METADATA>Binomial cluster 1=weight = 0.6667, mean = 0.375
#<METADATA>threshold=0.503
#<METADATA>fdr=0.125
#<METADATA>sensitivity=0.656
filter  FP      FDR     FN      FNR
weak_evidence   0.02    0.01    0.35    0.13
strand_bias     0.02    0.01    0.0     0.0
normal_artifact 0.0     0.0     0.07    0.03
slippage        0.0     0.0     0.0     0.0
haplotype       0.0     0.0     0.0     0.0
germline        0.22    0.11    0.0     0.0
```

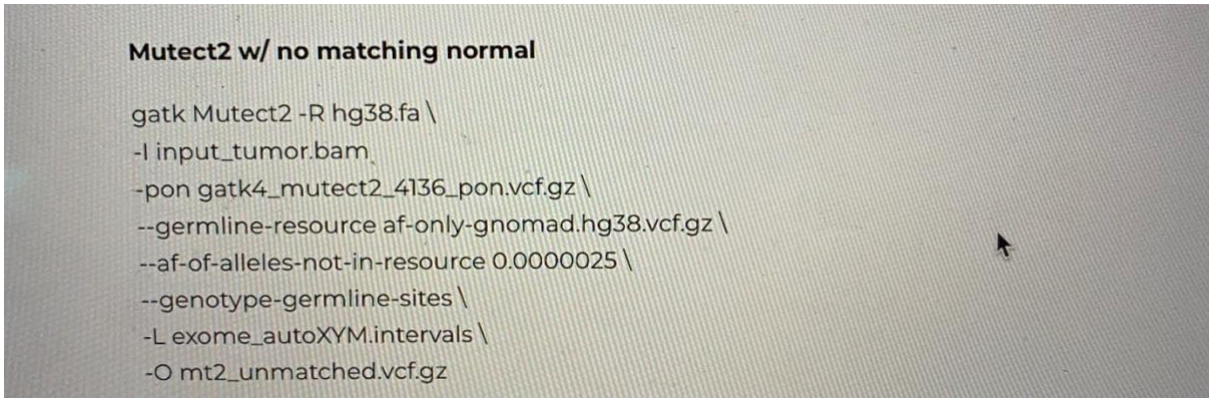
```
msc-student@lugh:/data/MSc/phyo/Practical1                    localuser@adb1019-850Q6S3: ~
 GNU nano 3.2                                                   filtered2.vcf.gz.filteringStats.tsv

#<METADATA>Ln prior of deletion of length 10=-14.510622640173326
#<METADATA>Ln prior of deletion of length 9=-14.510622640173326
#<METADATA>Ln prior of deletion of length 8=-14.510622640173326
#<METADATA>Ln prior of deletion of length 7=-13.999797016407333
#<METADATA>Ln prior of deletion of length 6=-13.999797016407333
#<METADATA>Ln prior of deletion of length 5=-14.91608774828149
#<METADATA>Ln prior of deletion of length 4=-12.518192475483119
#<METADATA>Ln prior of deletion of length 3=-12.112727367374955
#<METADATA>Ln prior of deletion of length 2=-11.41958018681501
#<METADATA>Ln prior of deletion of length 1=-10.485270949438176
#<METADATA>Ln prior of SNV=-5.706898038062702
#<METADATA>Ln prior of insertion of length 1=-10.24794276313201
#<METADATA>Ln prior of insertion of length 2=-11.99831701619721
#<METADATA>Ln prior of insertion of length 3=-12.564712491118012
#<METADATA>Ln prior of insertion of length 4=-12.564712491118012
#<METADATA>Ln prior of insertion of length 5=-13.999797016407333
#<METADATA>Ln prior of insertion of length 6=-13.817475459613338
#<METADATA>Ln prior of insertion of length 7=-14.510622640173326
#<METADATA>Ln prior of insertion of length 8=-13.817475459613338
#<METADATA>Ln prior of insertion of length 9=-15.609234928841435
#<METADATA>Ln prior of insertion of length 10=-20.72326583694641
#<METADATA>High-AF beta-binomial cluster=weight = 0.0115, alpha = 10.29, beta = 0.52
#<METADATA>Background beta-binomial cluster=weight = 0.1020, alpha = 9.97, beta = 9.15
#<METADATA>Binomial cluster 1=weight = 0.7406, mean = 0.990
#<METADATA>Binomial cluster 1=weight = 0.1475, mean = 0.499
#<METADATA>Binomial cluster 1=weight = 0.0754, mean = 0.540
#<METADATA>Binomial cluster 1=weight = 0.0097, mean = 0.460
#<METADATA>Binomial cluster 1=weight = 0.0073, mean = 0.535
#<METADATA>Binomial cluster 1=weight = 0.0069, mean = 0.486
#<METADATA>Binomial cluster 1=weight = 0.0060, mean = 0.592
#<METADATA>Binomial cluster 1=weight = 0.0055, mean = 0.451
#<METADATA>Binomial cluster 1=weight = 0.0006, mean = 0.718
#<METADATA>Binomial cluster 1=weight = 0.0004, mean = 0.089
#<METADATA>threshold=0.524
#<METADATA>fdr=0.081
#<METADATA>sensitivity=0.985
filter  FP      FDR     FN      FNR
weak_evidence   1874.82 0.08    301.51  0.01
strand_bias     2.83    0.0     0.0     0.0
slippage        4.95    0.0     8.01    0.0
haplotype       13.0    0.0     6.99    0.0
germline        21.79   0.0     4.1     0.0
```





SNVs are different with 15.06 and 5.70 for steps 3 & 6 respectively. Moreover, fdr is also different with 0.503 & 0.081. Matched tumour normal variant calling is better for downstream analysis.

8.



We can remove possible germline mutations following the above code. Moreover, we can use population-based data if there is no matched normal sample.