COVID Project

Data summary

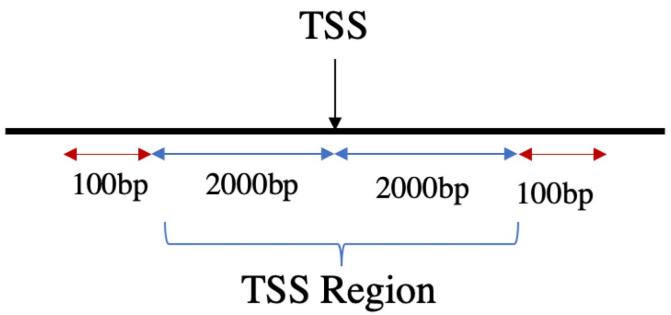
	Healthy	Mild	Moderate	Severe	Fatal
Number	7	7	3	7	1

The severity status is the status of patients when they were tested. (Current severity)

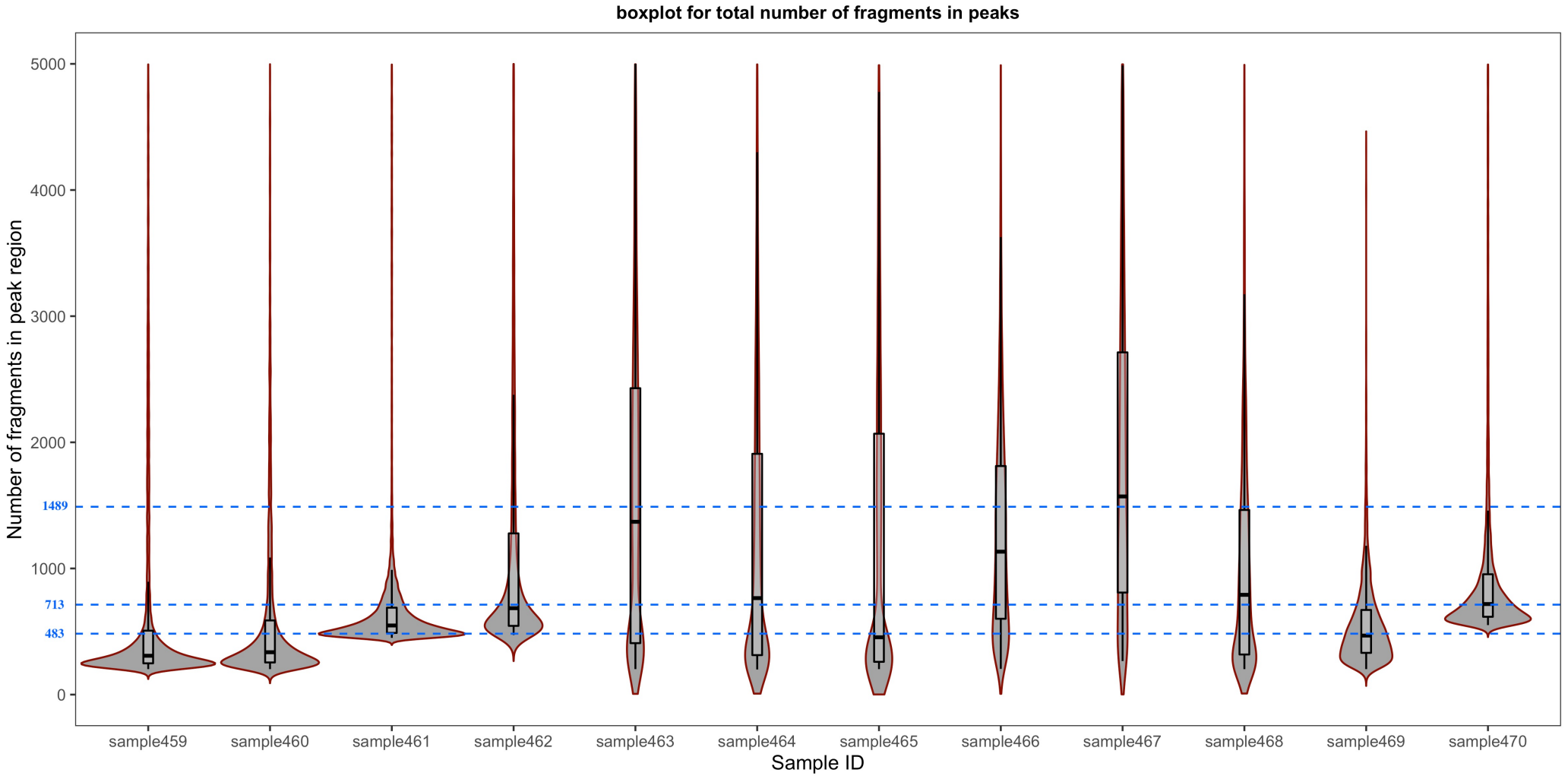
QC for different samples

1. Total number of fragments in peaks
2. The fraction of all fragments that fall within peak regions
3. TSS enrichment score

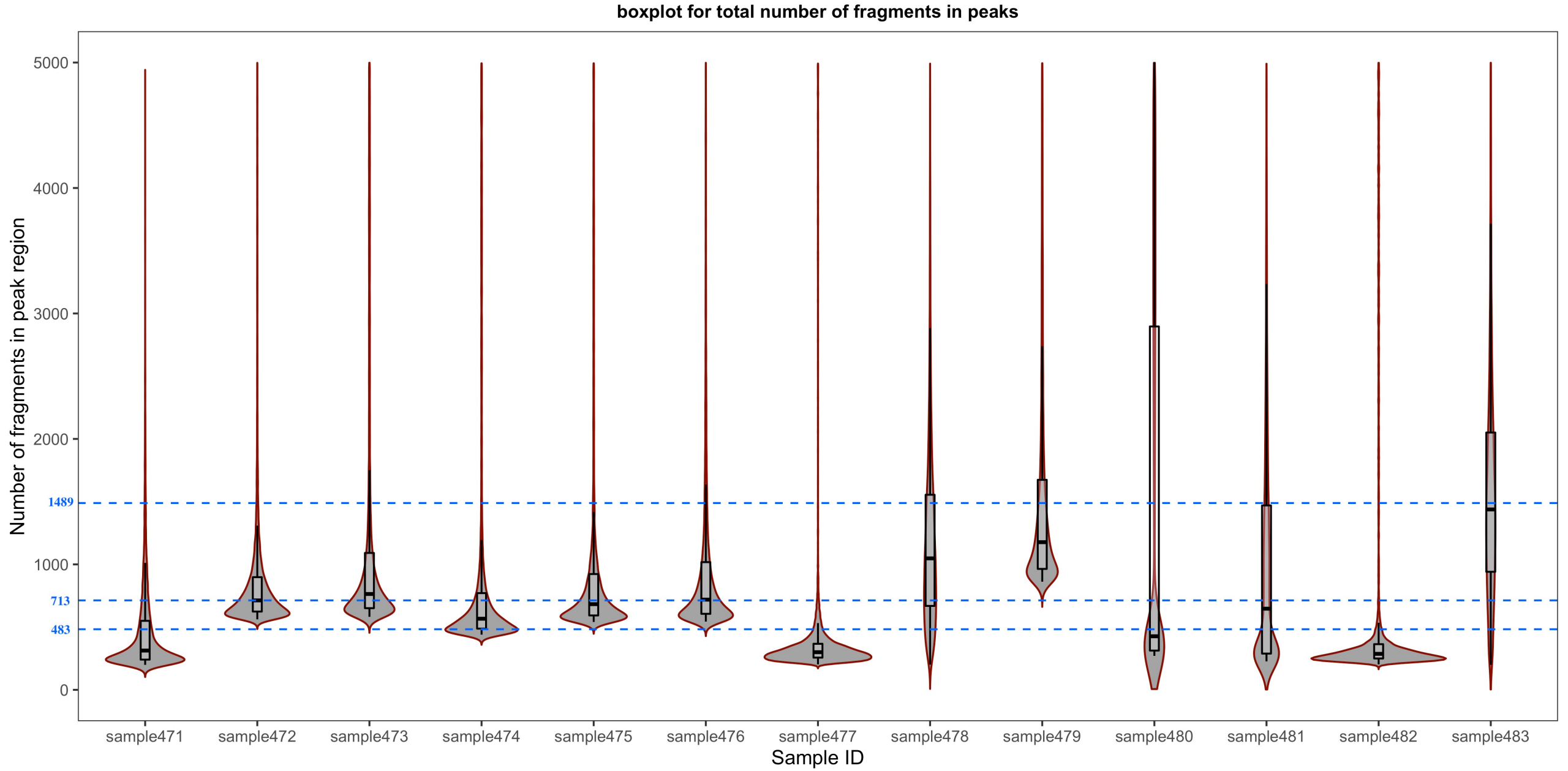
Transcription Start Site (TSS) Enrichment Score - The TSS enrichment calculation is a signal to noise calculation. The reads around a reference set of TSSs are collected to form an aggregate distribution of reads centered on the TSSs and extending to 2000 bp in either direction (for a total of 4000bp). This distribution is then normalized by taking the average read depth in the 100 bps at each of the end flanks of the distribution (for a total of 200bp of averaged data) and calculating a fold change at each position over that average read depth. This means that the flanks should start at 1, and if there is high read signal at transcription start sites (highly open regions of the genome) there should be an increase in signal up to a peak in the middle. We take the signal value at the center of the distribution after this normalization as our TSS enrichment metric. **Used to evaluate ATAC-seq.**



Number of fragments in peak region

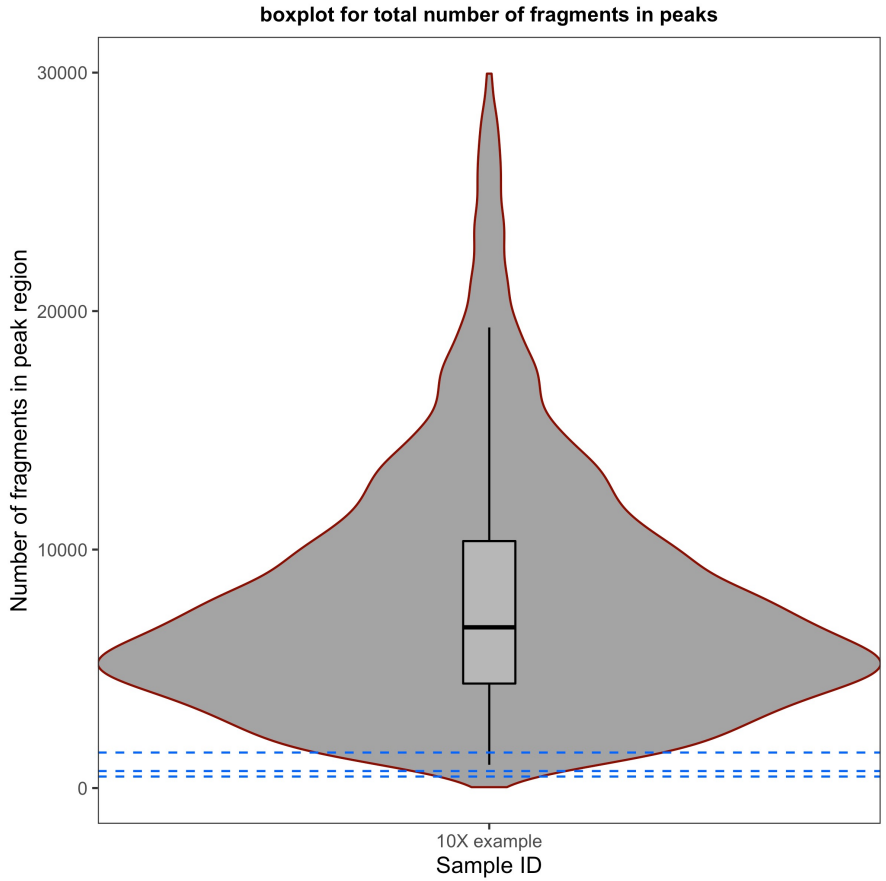


The 3 dashed bule lines indicated 25%, 50%, 75% quantile after pooling all samples.

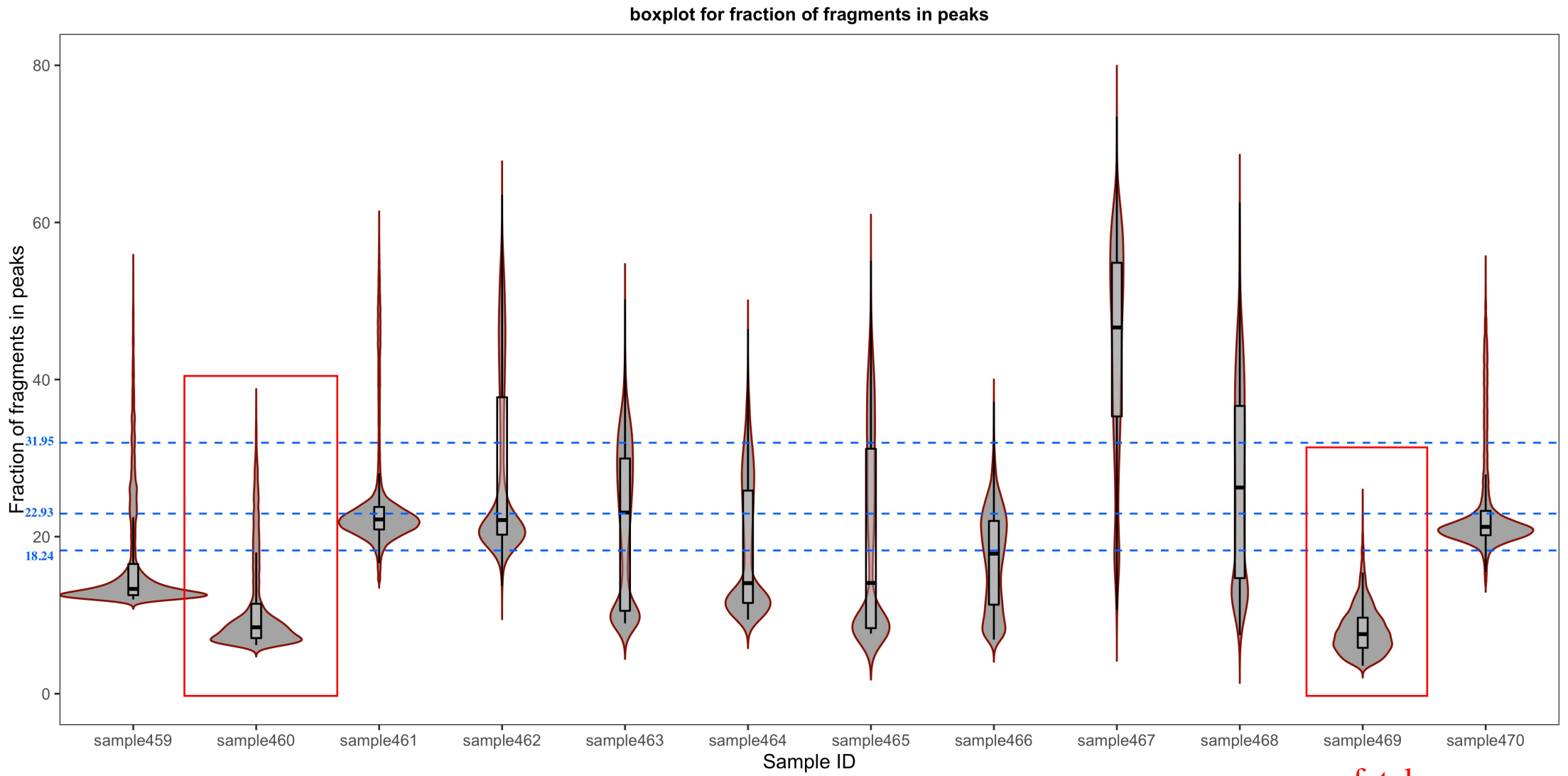

boxplot for total number of fragments in peaks
Number of fragments in peak region
5000
4000
3000
2000
1489
1000
713
483
0
sample471
sample472
sample473
sample474
sample475
sample476
sample477
sample478
sample479
sample480
sample481
sample482
sample483
Sample ID

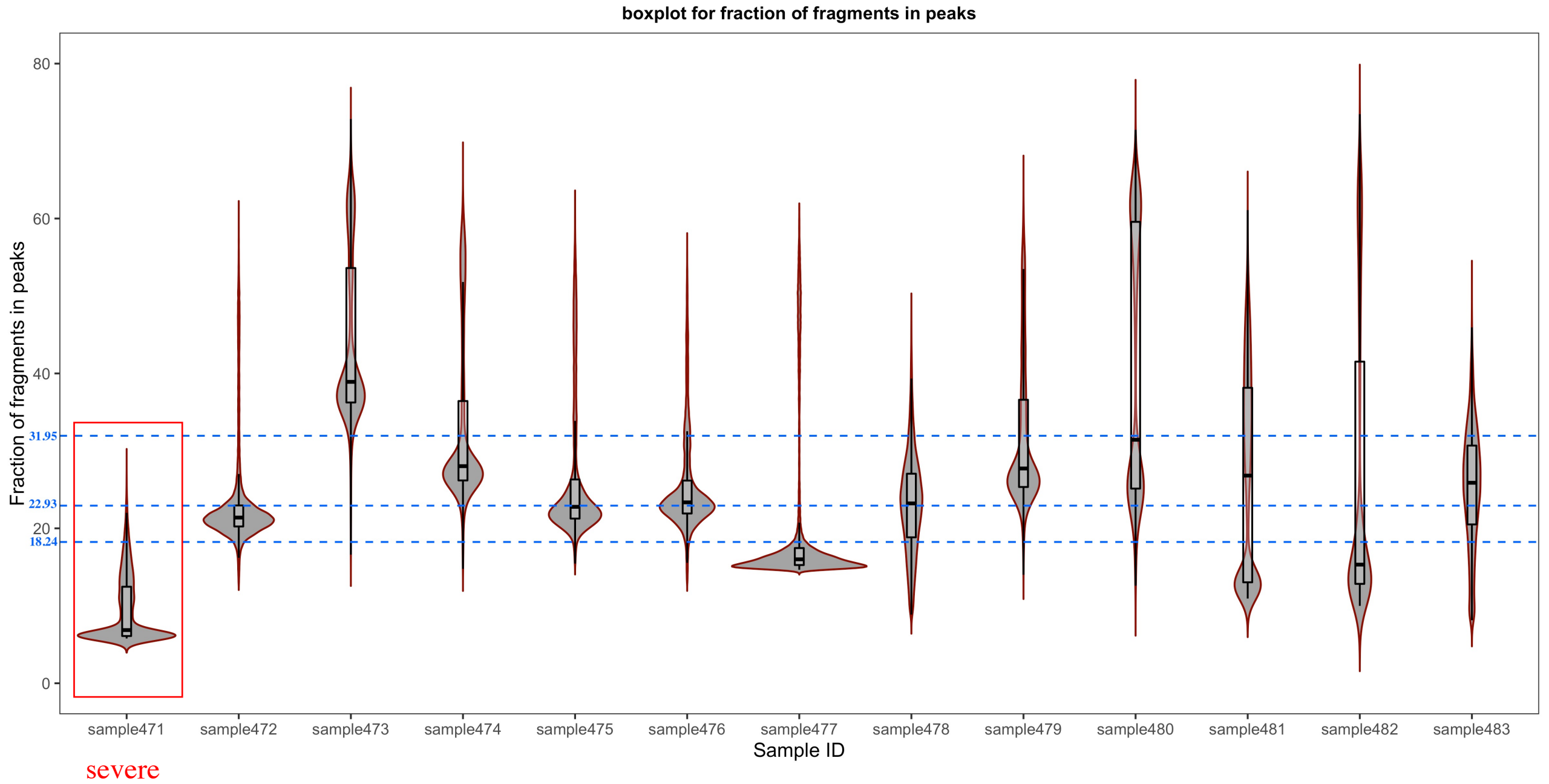
One example from 10X website

- I download one example from 10X website using the same protocol with the covid scATAC-seq data.
- Compared with data from 10X, the covid data seems too sparse.



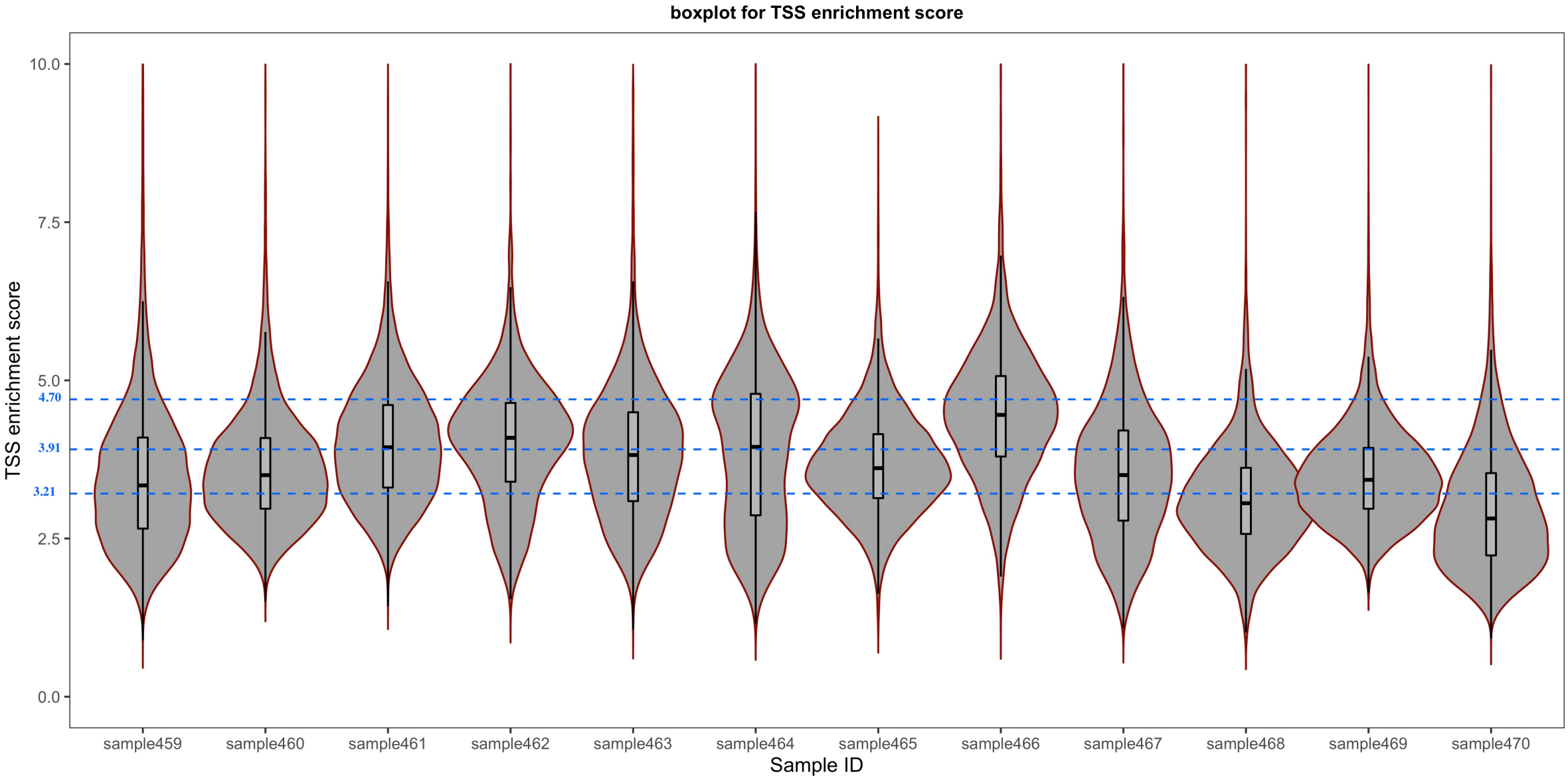
Fraction of fragments in peaks

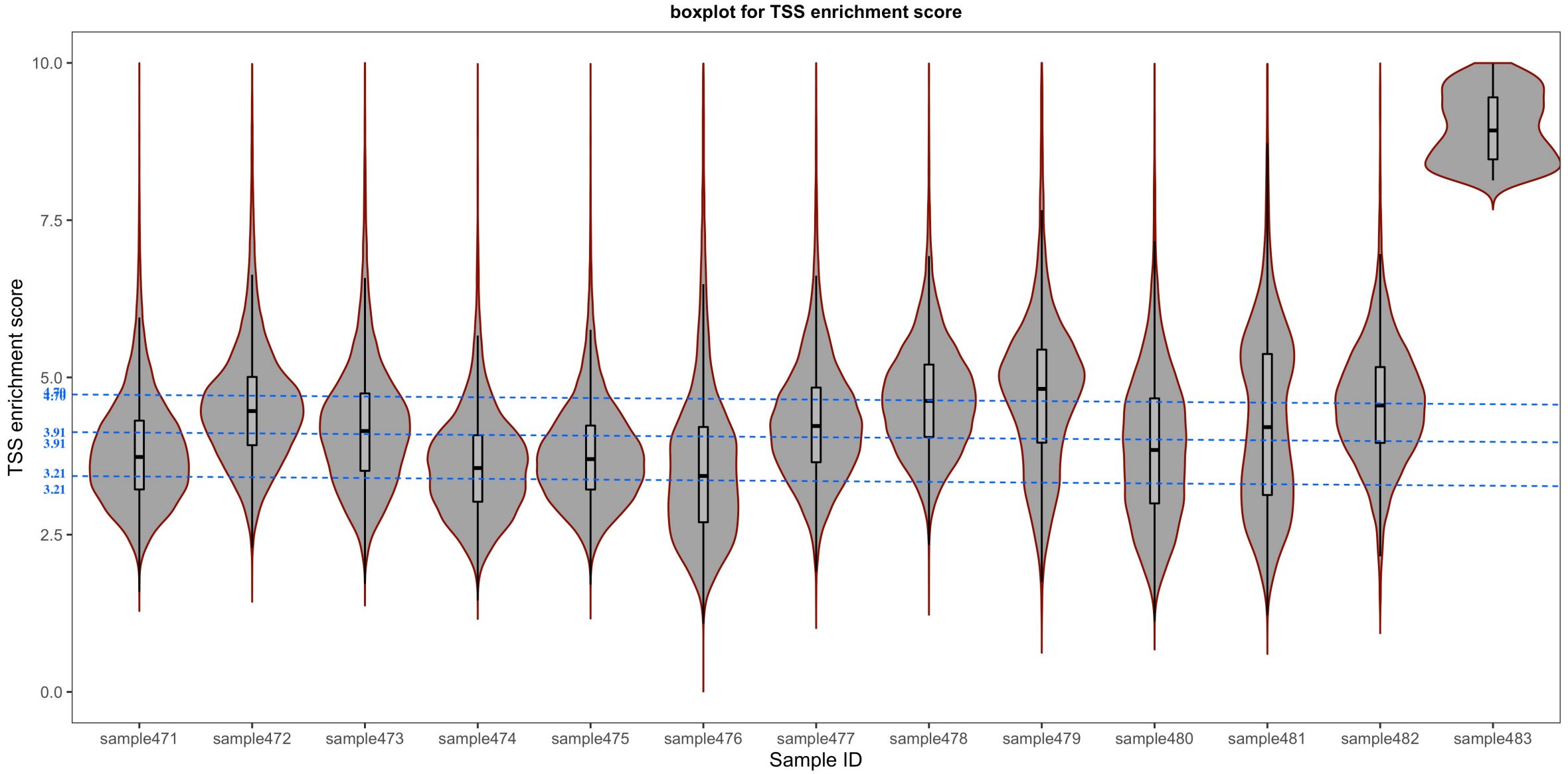



boxplot for fraction of fragments in peaks
Fraction of fragments in peaks
80
60
40
31.95
22.93
20
18.24
0
sample471
sample472
sample473
sample474
sample475
sample476
sample477
sample478
sample479
sample480
sample481
sample482
sample483
Sample ID
severe

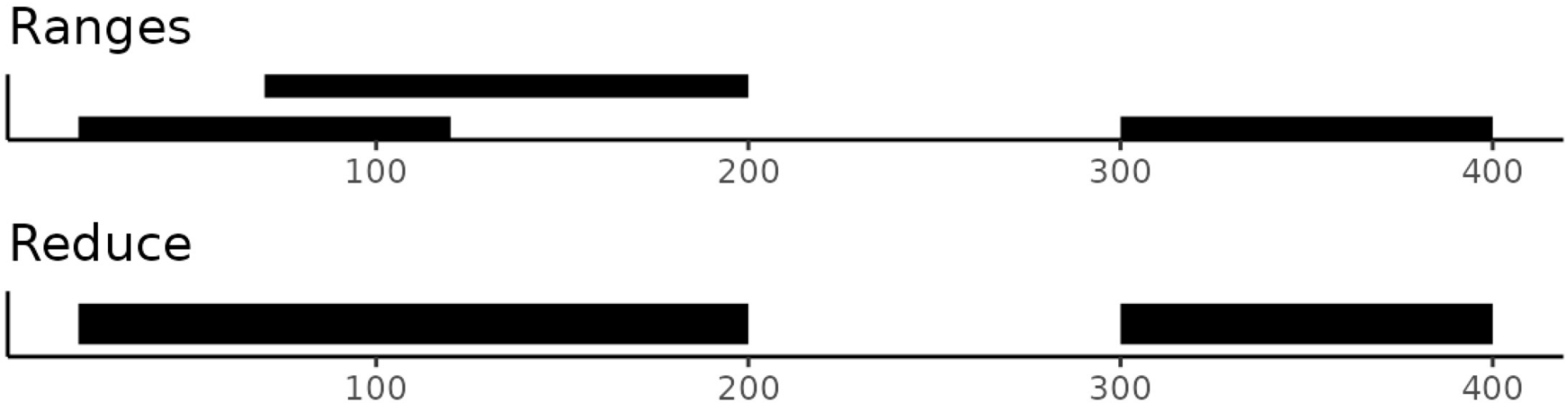
- Based on Fraction of fragments in peaks, we found sample 460, sample469 and sample471 are much worse than others.
- We decide to filter out those 3 samples. After that, we have 7 healthy samples, 6 moderate samples, 3 mild samples and 7 severe samples.

TSS enrichment score




boxplot for TSS enrichment score
TSS enrichment score
10.0
7.5
5.0
4.70
4.70
3.91
3.91
3.21
3.21
2.5
0.0
sample471
sample472
sample473
sample474
sample475
sample476
sample477
sample478
sample479
sample480
sample481
sample482
sample483
Sample ID

Get Union Peaks

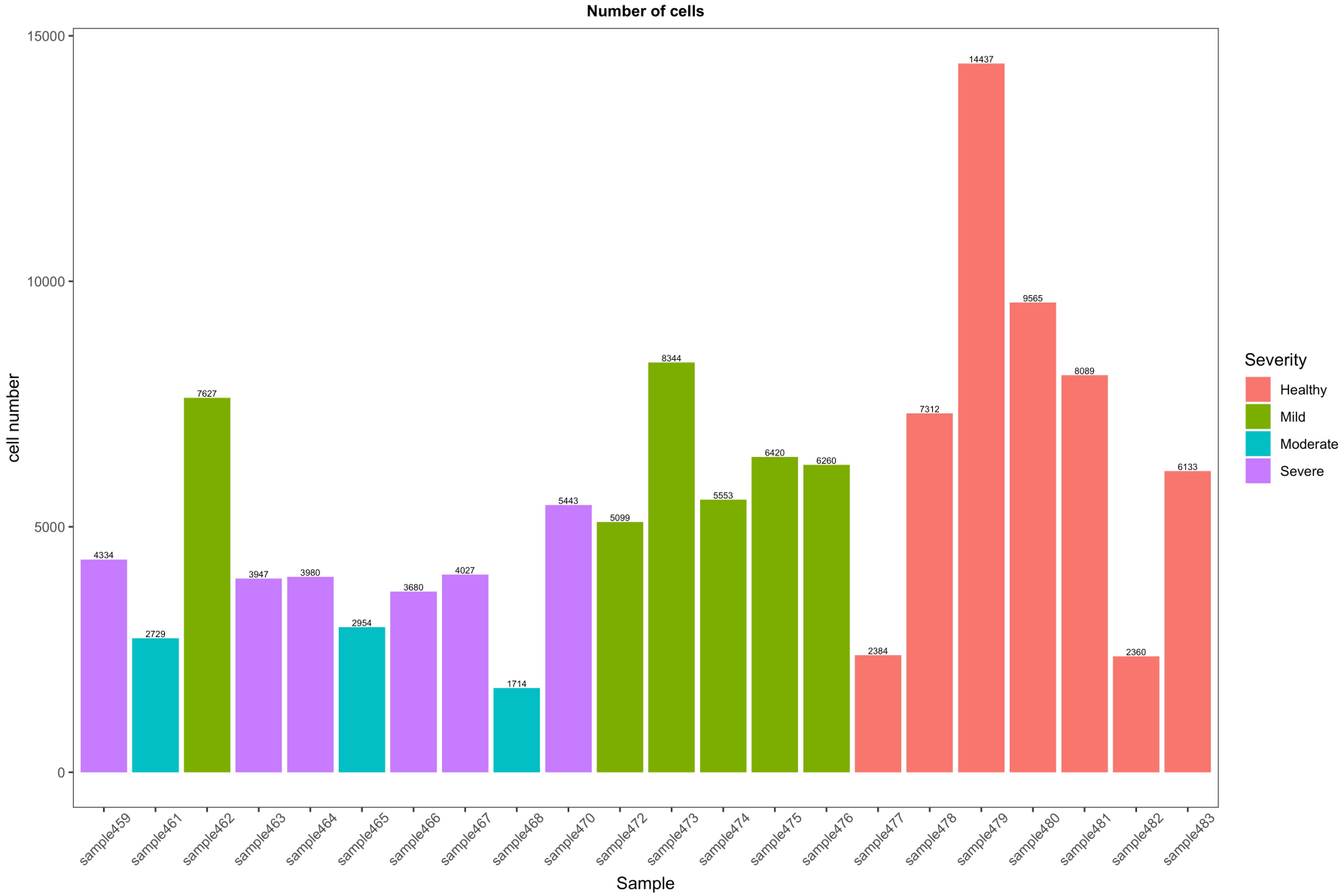


- Get common peak set for remaining 22 samples (There are 229, 457 peaks in total)
- Quantify peaks in each dataset
- Generate peak cell matrix for each sample

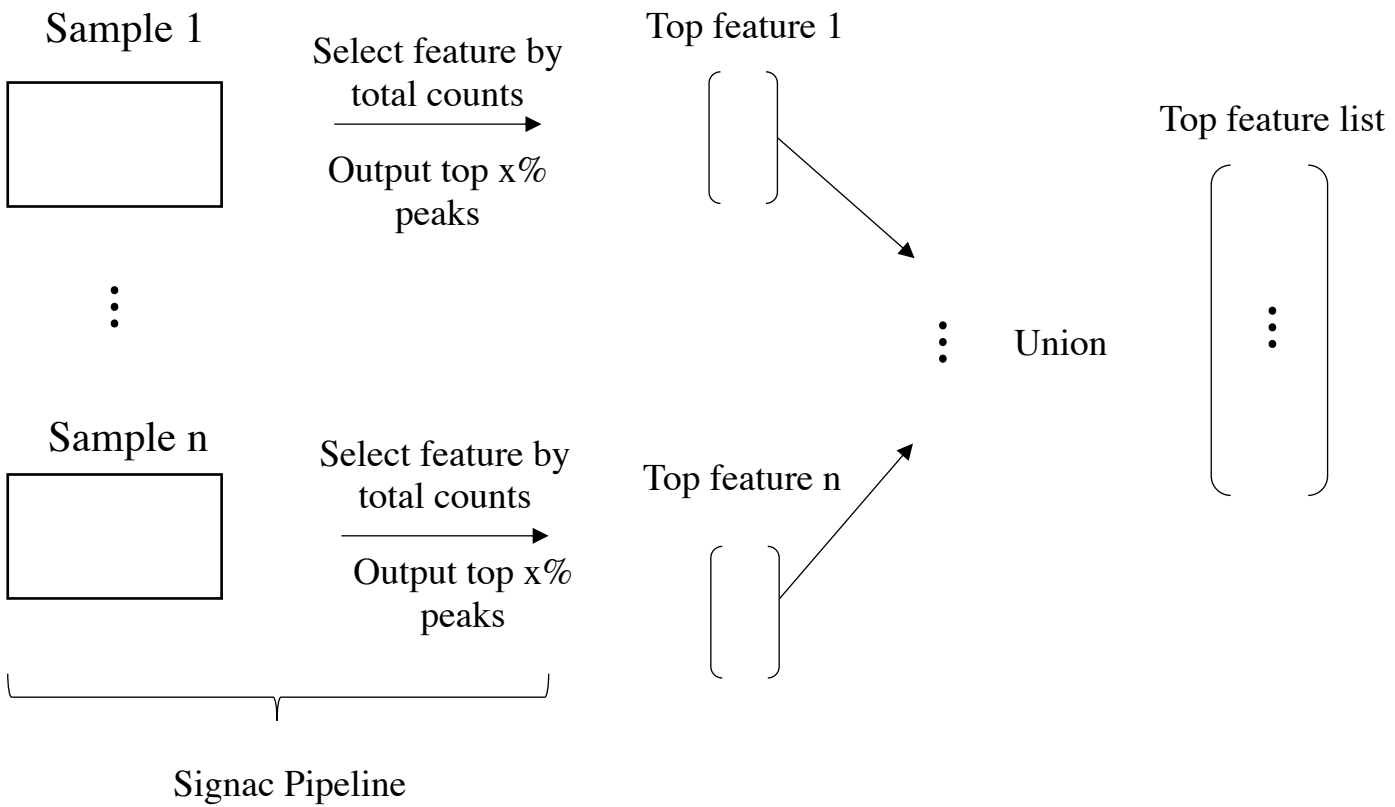
Filter low quality cells for each sample

- Number of fragments > 1000(1500 is recommended by the author) and $< 20,000$
- Percentage of reads in peak region > 13.67 (10% quantile after pooling all samples except for 3 filtered samples)
- TSS enrichment score > 2.64 (10% quantile after pooling all samples except for 3 filtered samples)

Sample information after filtering



Select top features: Method1

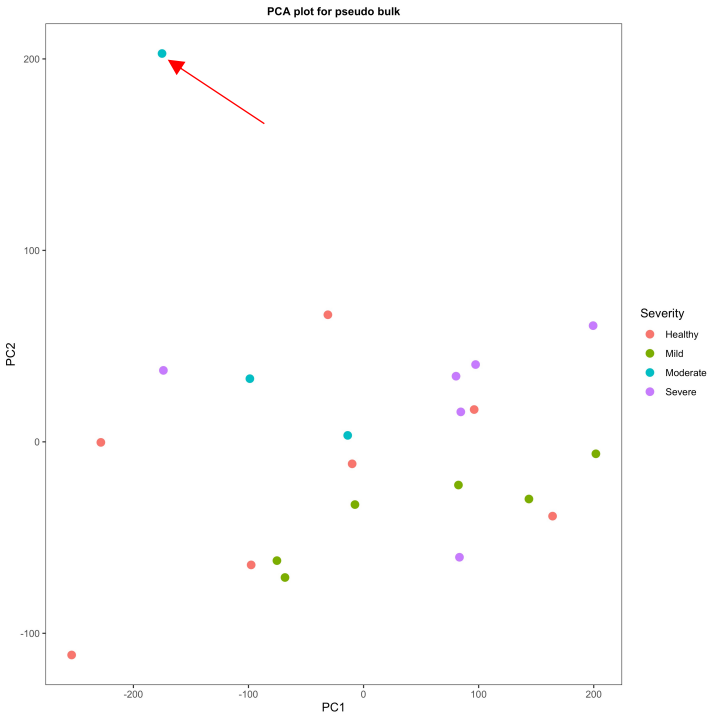


- I tried with different x.

Pseudo bulk analysis

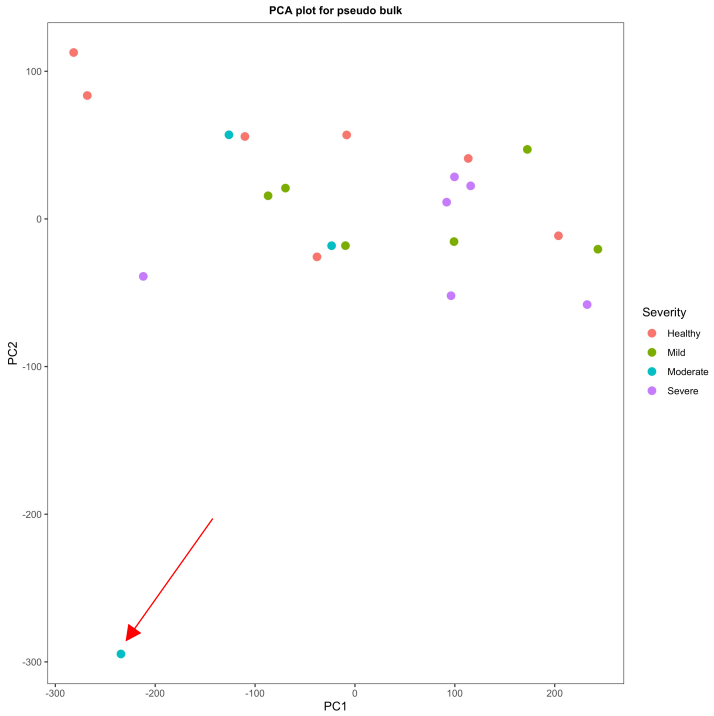
- Compute pseudo bulk for each sample
- Normalized by library size
- PCA using top features

Results for different choice of x



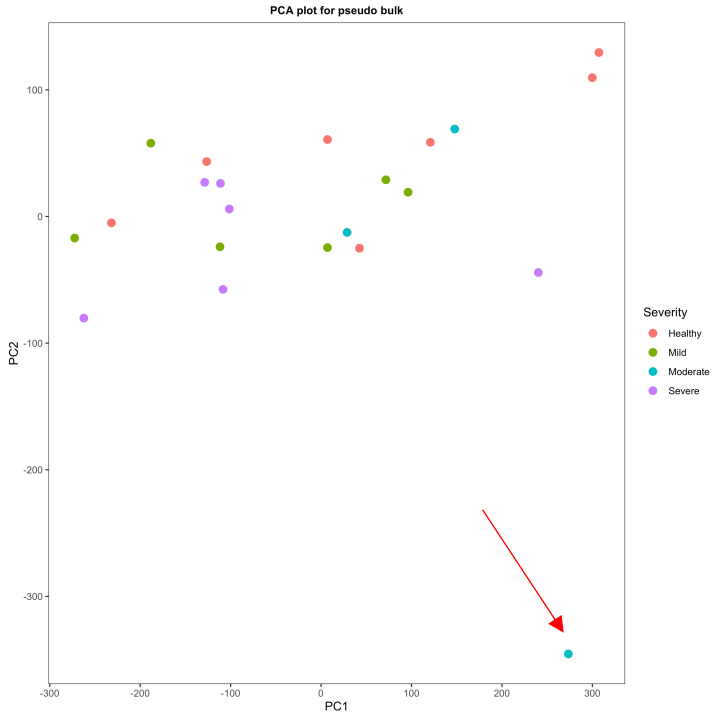
X = 10%

39, 157 peaks



X = 15%

59, 086 peaks

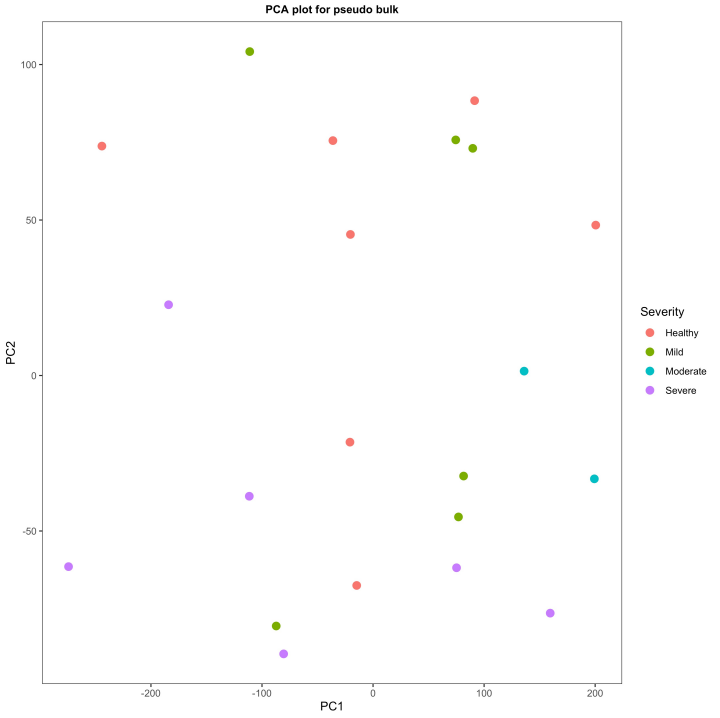


X = 20%

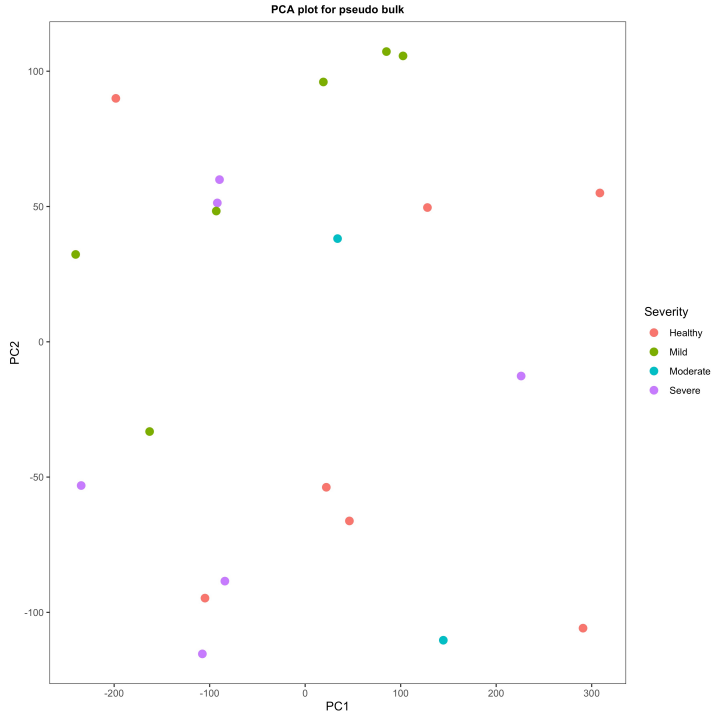
78, 305 peaks

Outliers in 3 plots are sample468. sample468 has the least cell number.

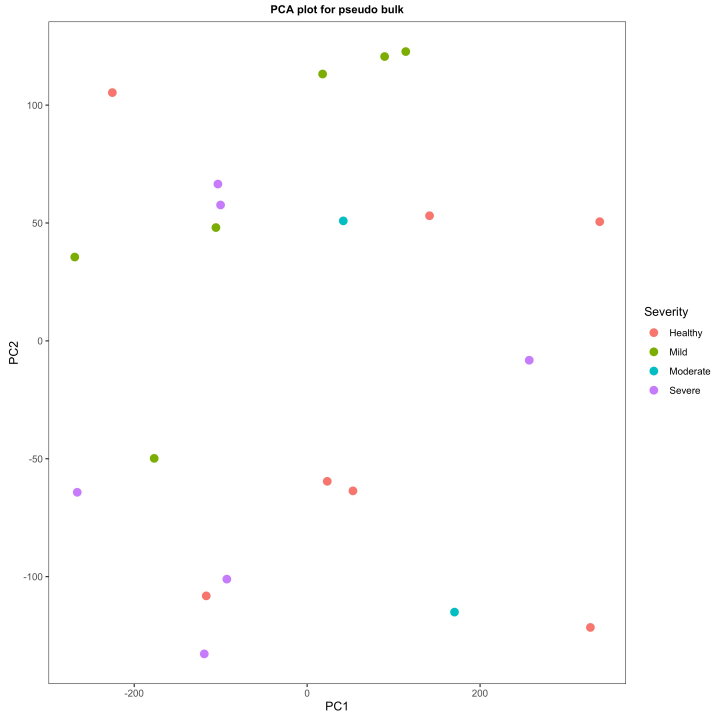
Results without sample468



X = 10%

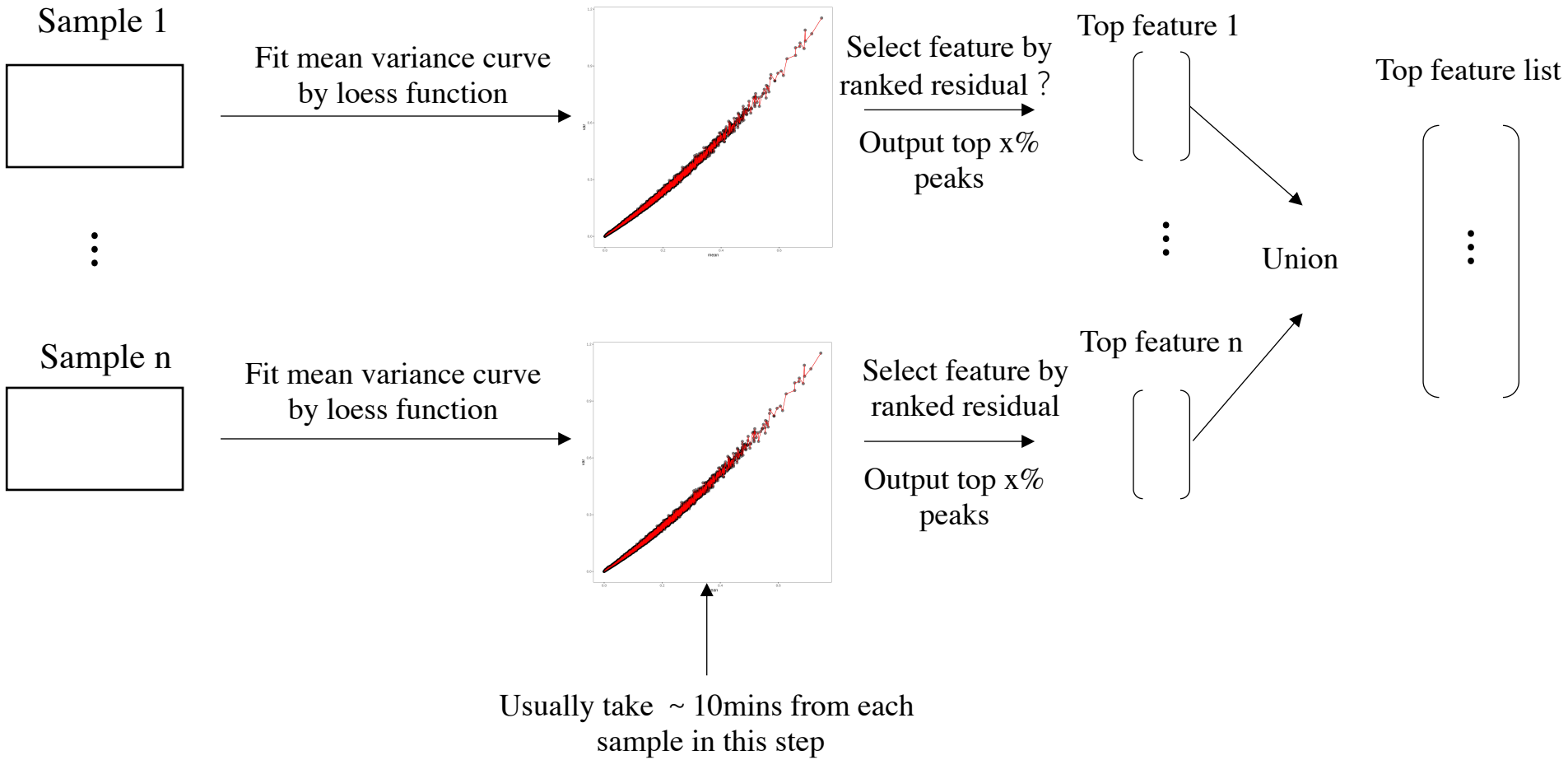


X = 15%

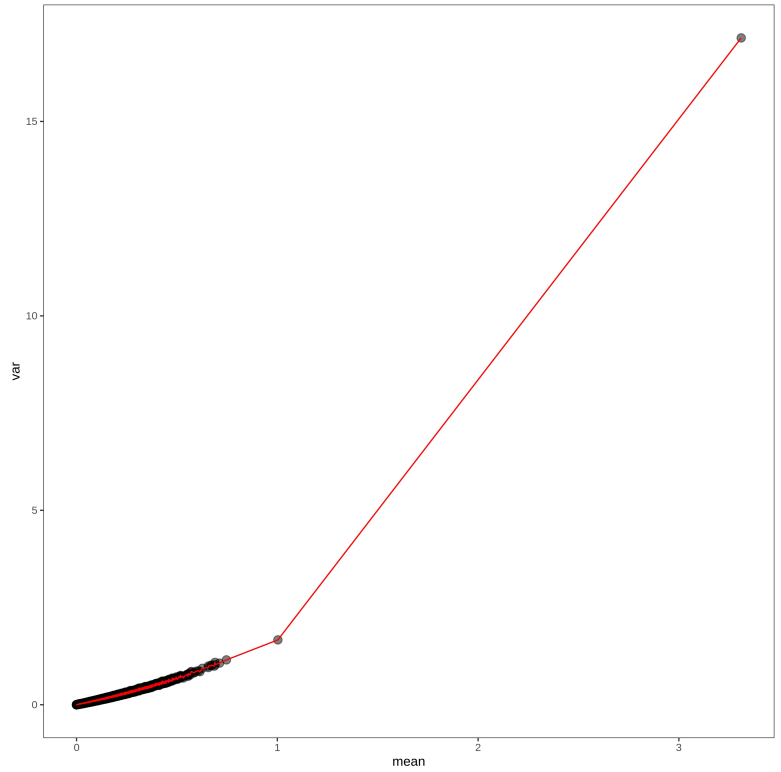


X = 20%

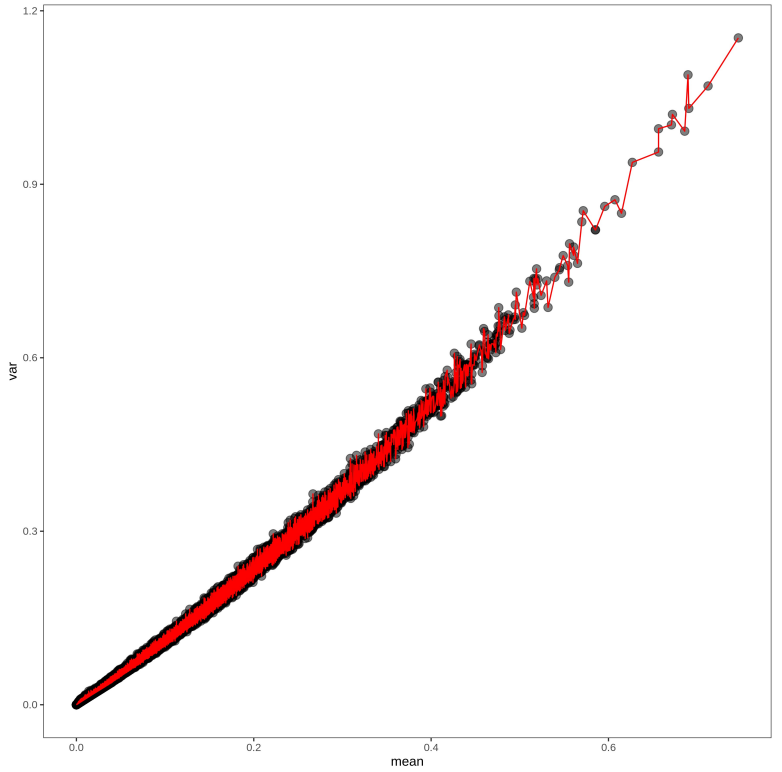
Select top features: Method2



Outlier when fitting the line



Fitted with all features



Fitted after filtering outliers

Use Harmony to integrate different samples

COVID severity may not be dominated biological difference

- Find differentiable features between covid and healthy sample after controlling for gender and age.

Use Seurat to integrate different samples

Use Cell type proportion to define distance