EDA

Loading data

100 gibbs samples and bootstrap

```
library(tximport)
source("helper_func.R")
load("out_1/sim_counts_matrix.rda")
dir <- "out"
gsFiles <- file.path(dir, c("ERR188297_GS", "sample_01_GS"), "quant.sf")
txiInfRepGS <- tximport(gsFiles, type = "salmon", txOut = TRUE)

bootFiles <- file.path(dir, c("ERR188297_B", "sample_01_B"), "quant.sf")
txiInfRepBoot <- tximport(bootFiles, type = "salmon", txOut = TRUE)

#fileDf <- vroom::vroom("fastq/sample_01_1.fastq.gz", delim = "\n", col_names = F)
#simCounts <- countReads(fileDf)

txiInfRepGS <- computeConfInt(txiInfRepGS)
txiInfRepBoot <- computeConfInt(txiInfRepBoot)
```

Simulated Data

Obtaining the transcripts with zero counts

```
zeroBInds <- which(txiInfRepBoot$counts[,2] == 0)
zeroGSInds <- which(txiInfRepGS$counts[,2] == 0)
print(length(zeroGSInds))
```

```
## [1] 157390
```

```
print(length(zeroBInds))
```

```
## [1] 157387
```

```
print(length(setdiff(zeroGSInds, zeroBInds)))
```

```
## [1] 5
```

```
print(length(setdiff(zeroBInds, zeroGSInds)))
```

```
## [1] 2
```

So mostly same transcripts have zero counts at two independent EM runs

Obtaining transcripts that have zero means over the bootstrap runs

```
zeroMeanGS <- which(txiInfRepGS$conf[[2]][,3] <= 3)
zeroMeanB <- which(txiInfRepBoot$conf[[2]][,3] <= 3)

print(length(zeroMeanGS))
```

```
## [1] 134317
```

```
print(length(zeroMeanB))
```

```
## [1] 156142
```

```
print(length(setdiff(zeroBInds, zeroMeanB)))
```

```
## [1] 2350
```

```
print(length(setdiff(zeroGSInds, zeroMeanGS)))
```

```
## [1] 23634
```

```
print(length(intersect(zeroMeanB, zeroMeanGS)))
```

```
## [1] 134280
```

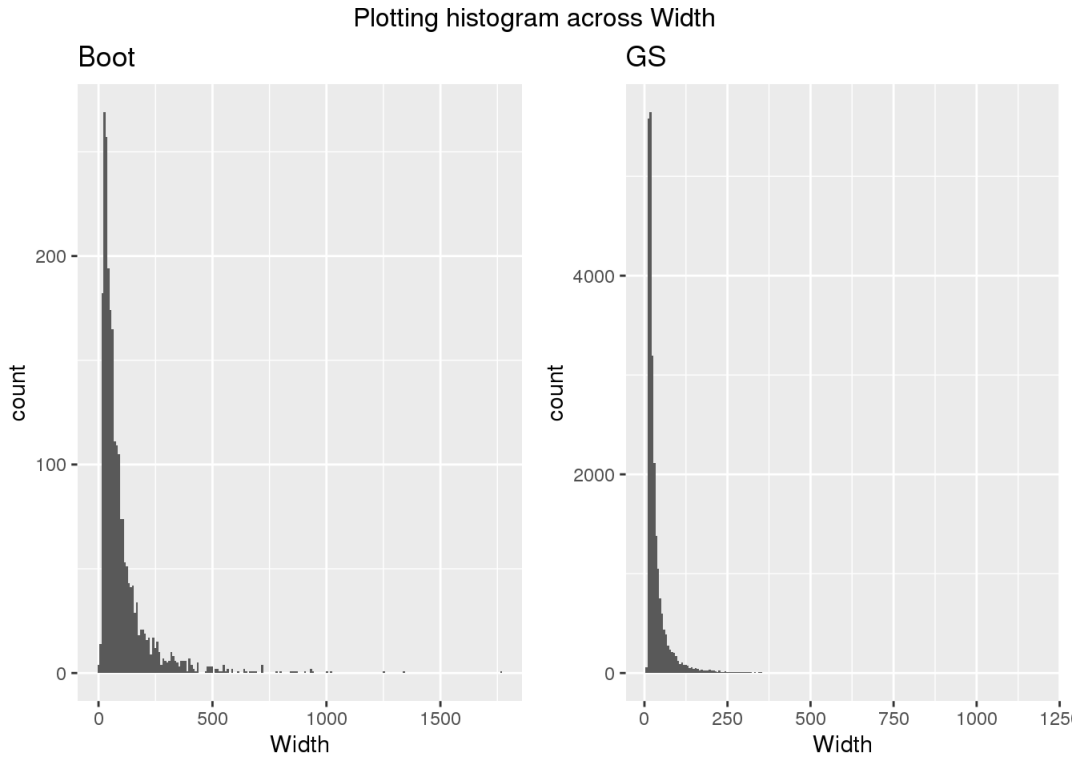
```
plotDfZeros <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[2]], "GS" = txiInfRepGS$conf[[2]]), list("Boot
" = setdiff(zeroBInds, zeroMeanB), "GS" = setdiff(zeroGSInds, zeroMeanGS)))
plotHist(plotDfZeros, "Width")
```

```
## Loading required package: ggplot2
```

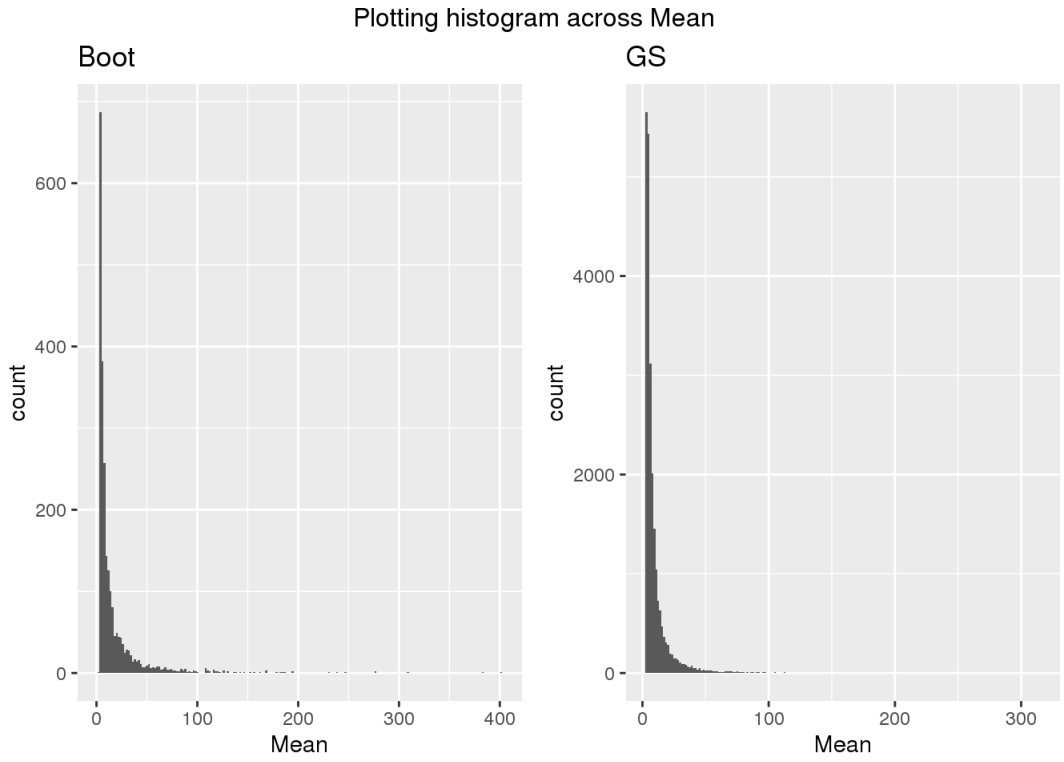
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

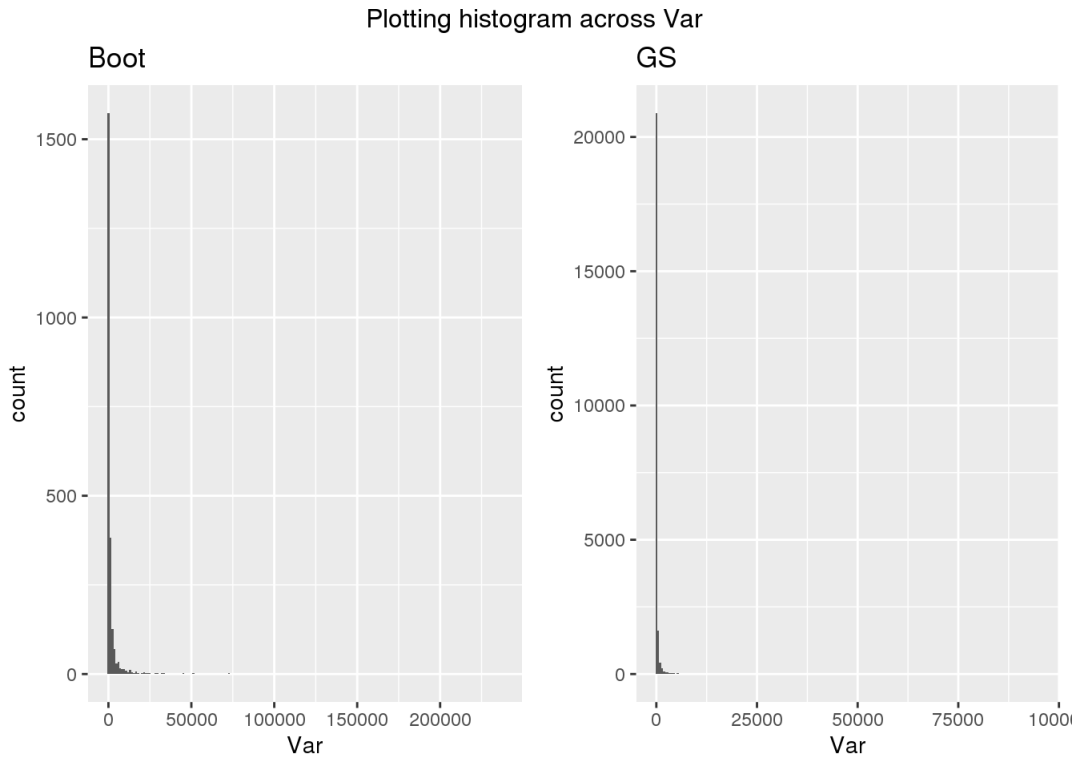
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```




```
plotHist(plotDfZeros, "Mean")
```




```
plotHist(plotDfZeros, "Var")
```



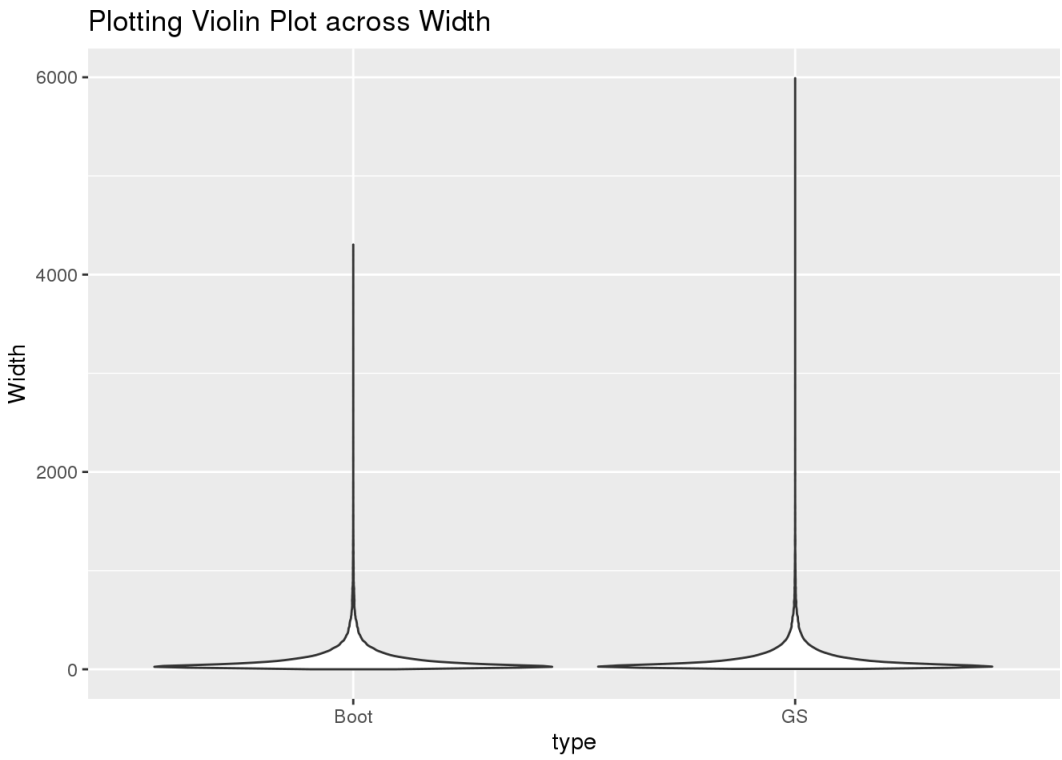
A lot of transcripts that had zero counts under Gibbs sampling have non zero means compared to bootstrap

Only observing the transcripts having non zero counts

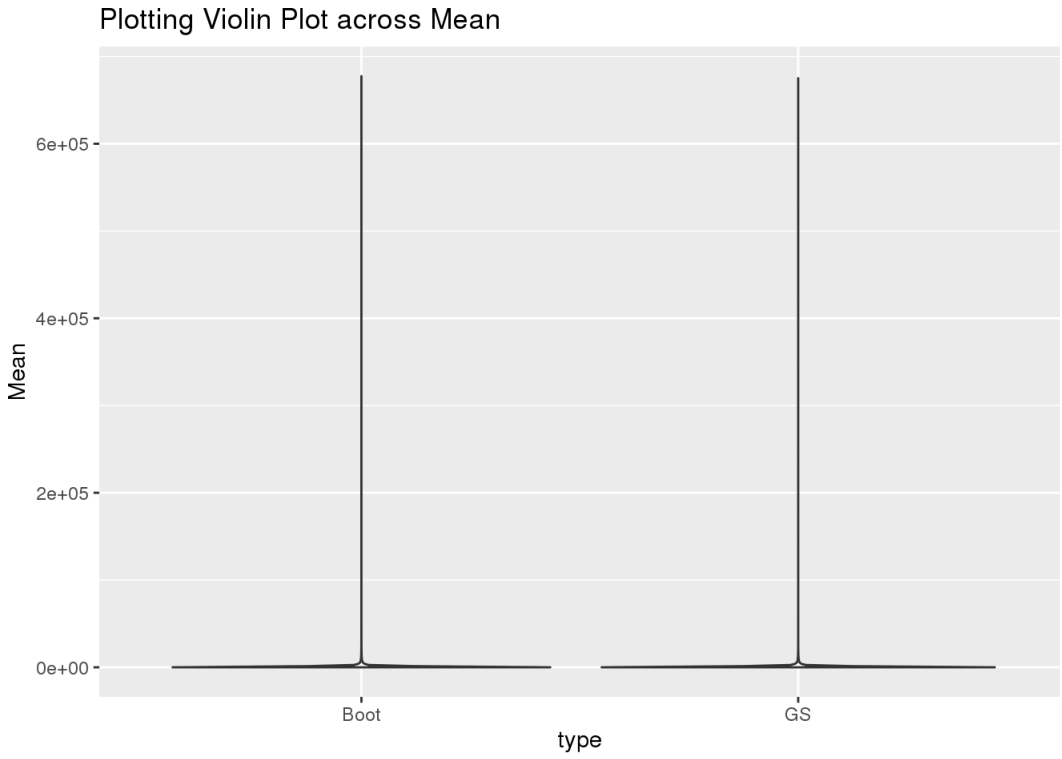
```
nZeroInds <- setdiff(1:nrow(txiInfRepBoot$conf[[2]]), union(zeroBInds, zeroGSInds))
print(length(nZeroInds))
```

```
## [1] 45635
```

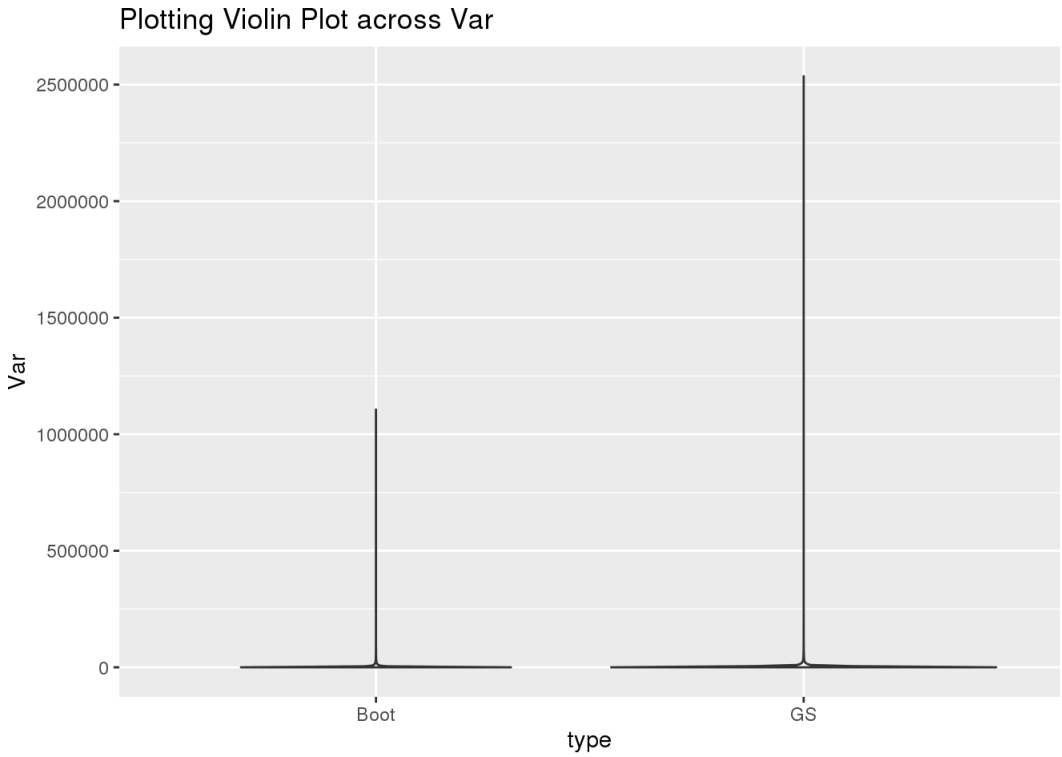
```
plotDfNonZeros <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[2]], "GS" = txiInfRepGS$conf[[2]]),
                               list("Boot" = nZeroInds, "GS" = nZeroInds))
plotViol(plotDfNonZeros, "Width")
```




```
plotViol(plotDfNonZeros, "Mean")
```




```
plotViol(plotDfNonZeros, "Var")
```



Transcripts with 95% CI of one is less than 5% CI of other (GS Dominates)

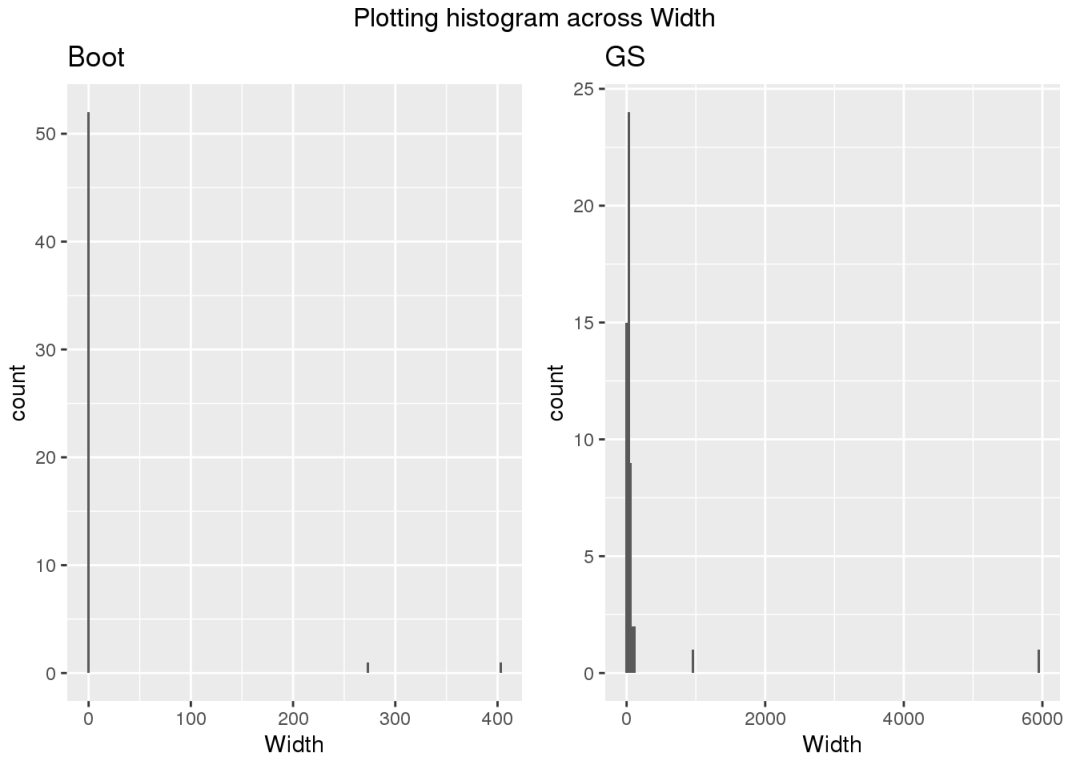
```
GSLargeInds <- which(txiInfRepBoot$conf[[2]][nZeroInds,2] < txiInfRepGS$conf[[2]][nZeroInds,1])
bootLargeInds <- which(txiInfRepGS$conf[[2]][nZeroInds,2] < txiInfRepBoot$conf[[2]][nZeroInds,1])

print(length(GSLargeInds))
```

```
## [1] 54
```

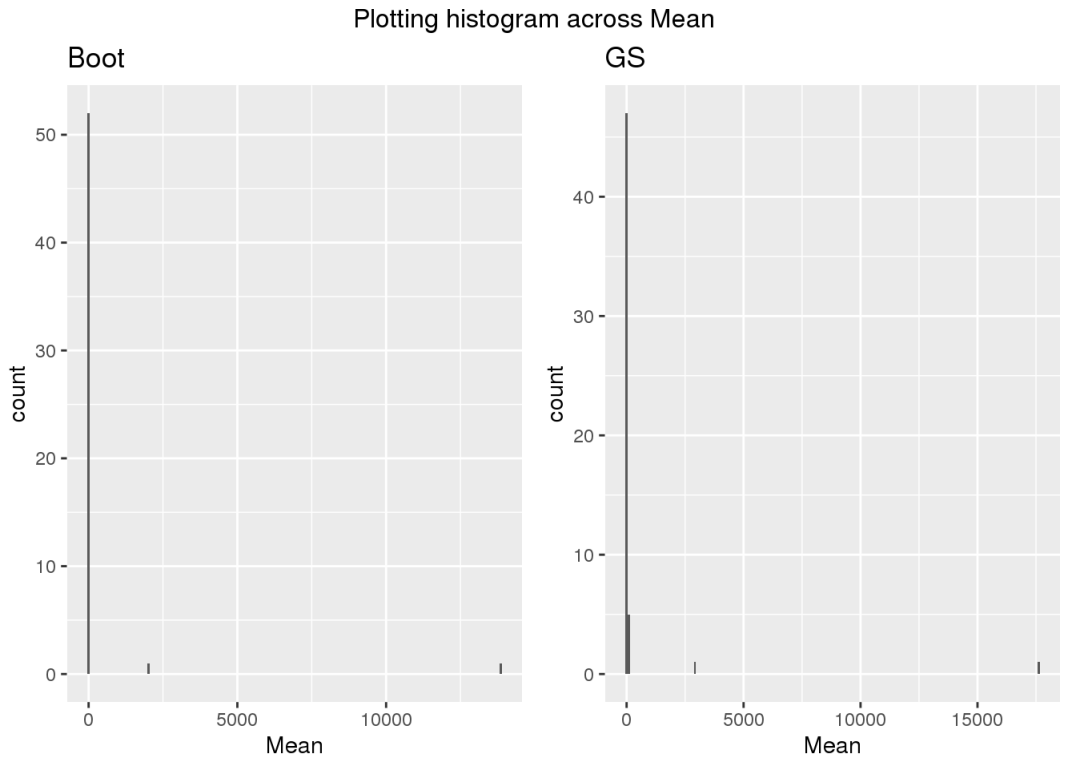
```
print(length(bootLargeInds))
```

```
## [1] 5
```

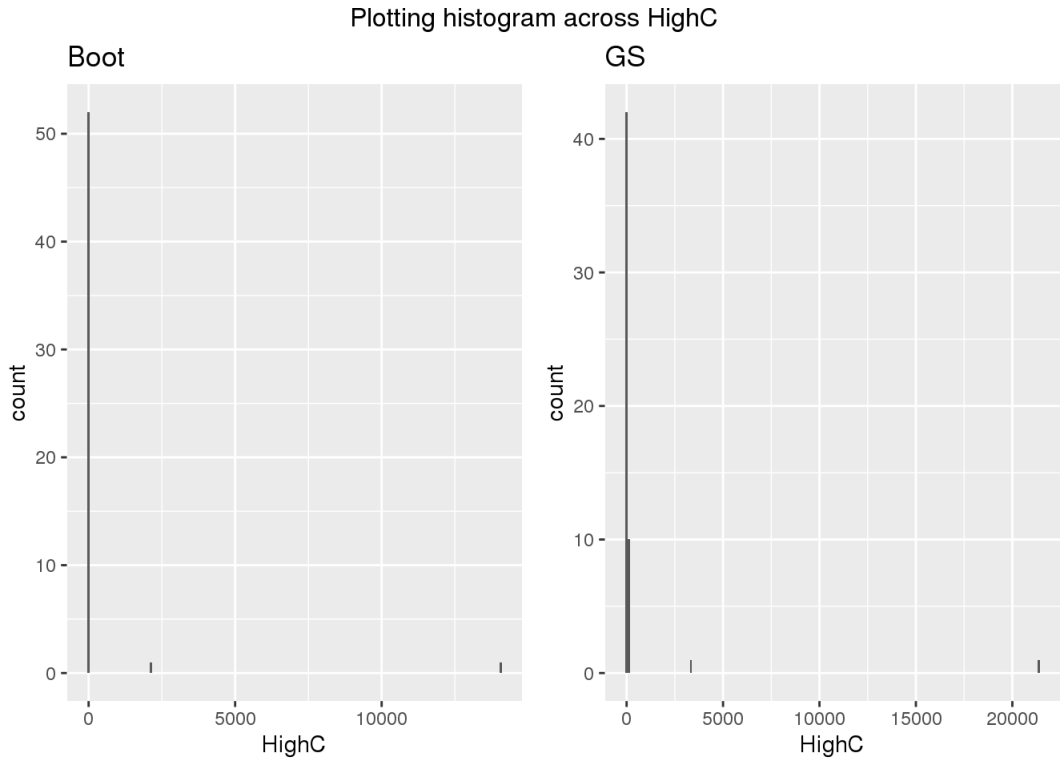
```
pGSLargeDf <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[2]], "GS" = txiInfRepGS$conf[[2]]),
                           list("Boot" = nZeroInds[GSLargeInds], "GS" = nZeroInds[GSLargeInds]))
plotHist(pGSLargeDf, "Width")
```



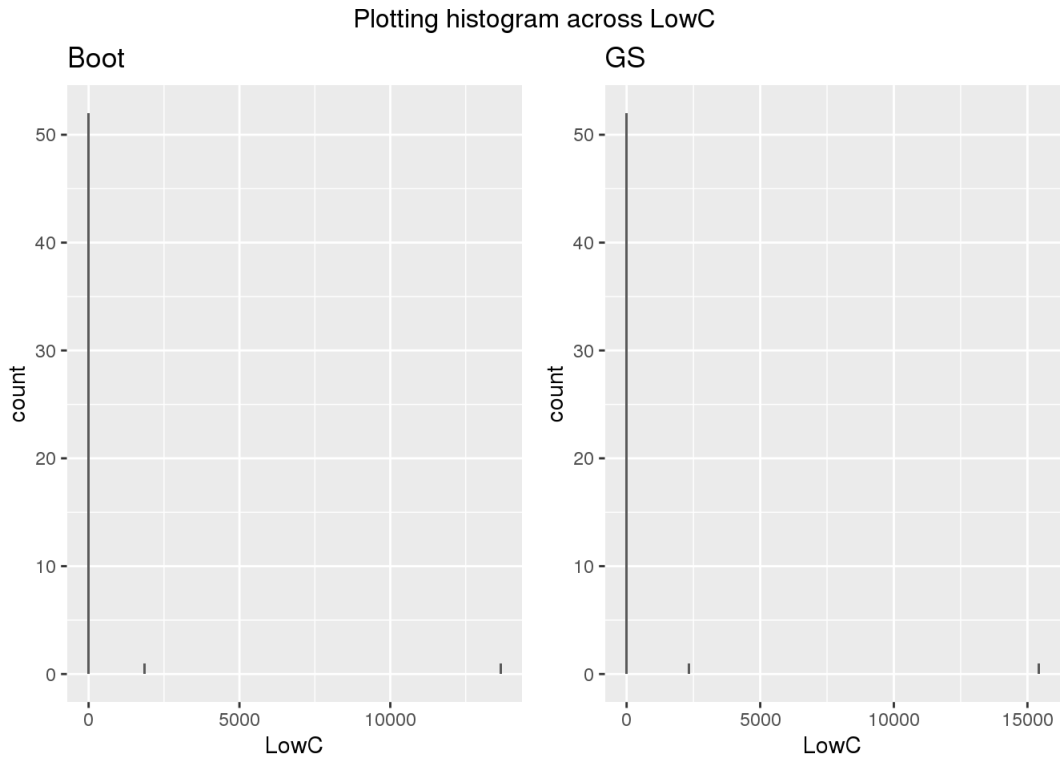

```
plotHist(pGSLargeDf, "Mean")
```




```
plotHist(pGSLargeDf, "HighC")
```

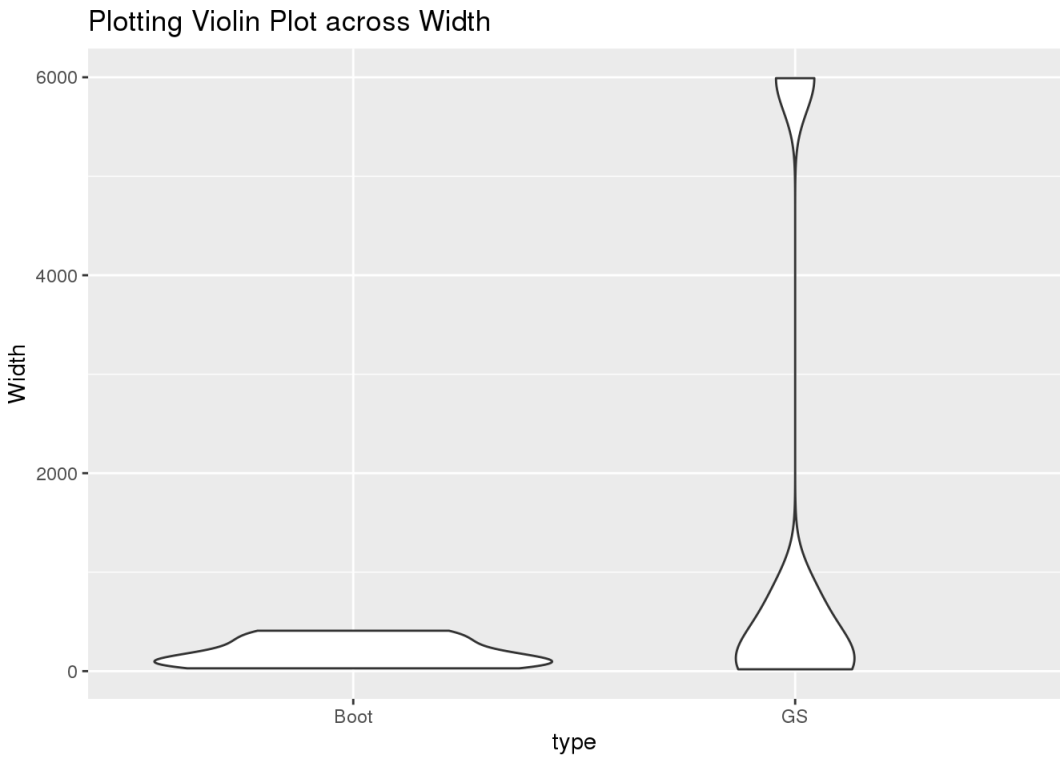



```
plotHist(pGSLargeDf, "LowC")
```

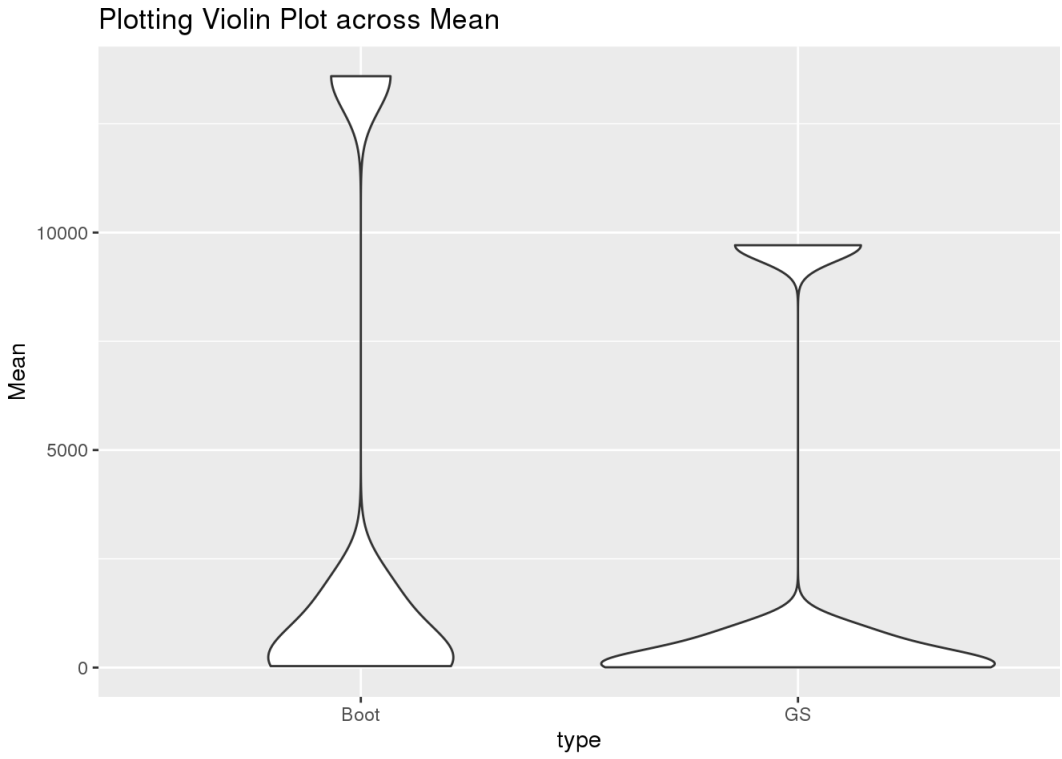


Transcripts with 95% CI of one is less than 5% CI of other (Bootstrap Dominates)

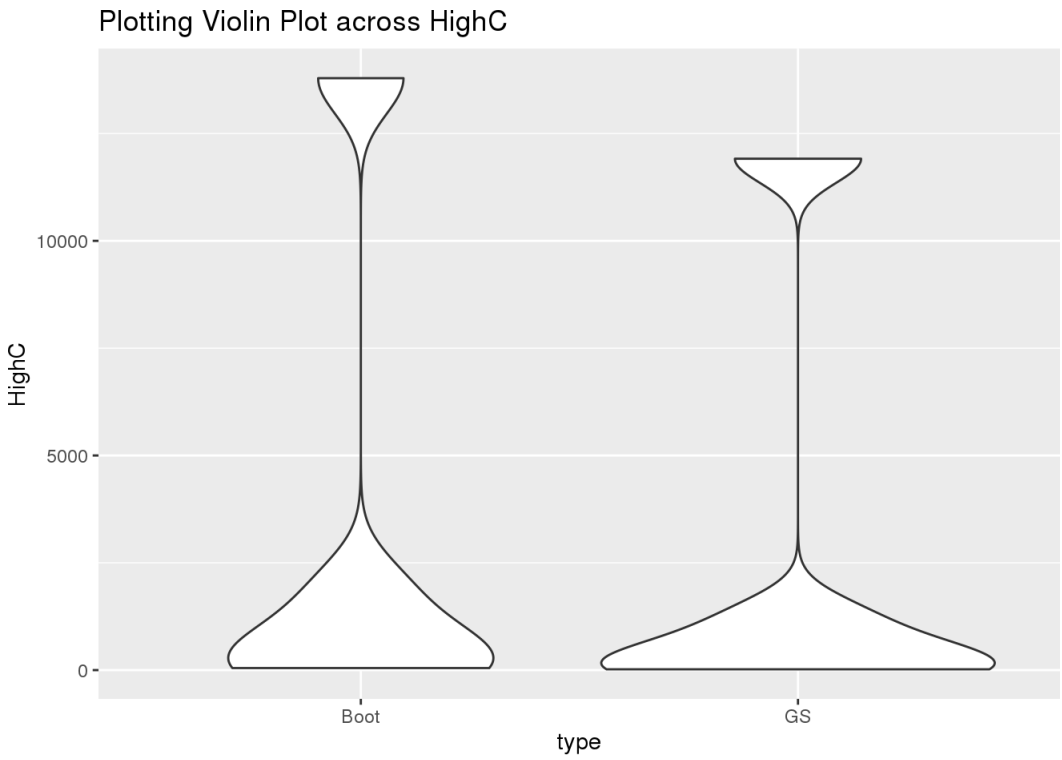
```
pBootLargeDf <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[2]], "GS" = txiInfRepGS$conf[[2]]),
                             list("Boot" = nZeroInds[bootLargeInds], "GS" = nZeroInds[bootLargeInds]))
plotViol(pBootLargeDf, "Width")
```



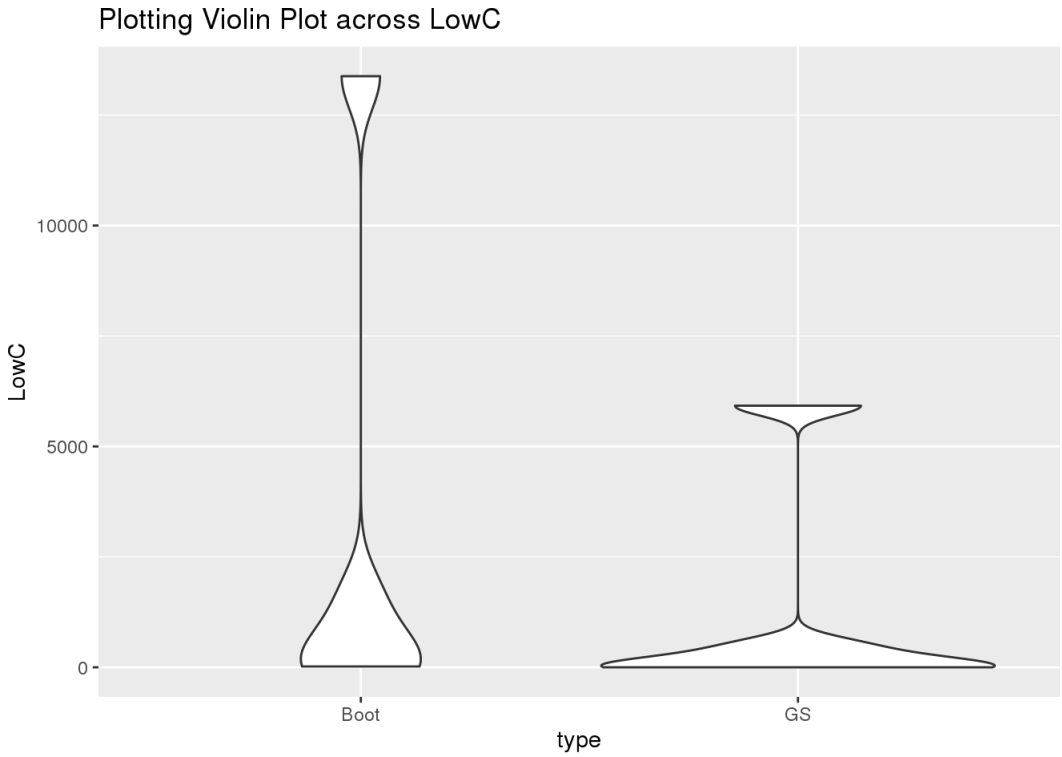

```
plotViol(pBootLargeDf, "Mean")
```




```
plotViol(pBootLargeDf, "HighC")
```




```
plotViol(pBootLargeDf, "LowC")
```



Computing the difference between magnitude of difference b/w the two

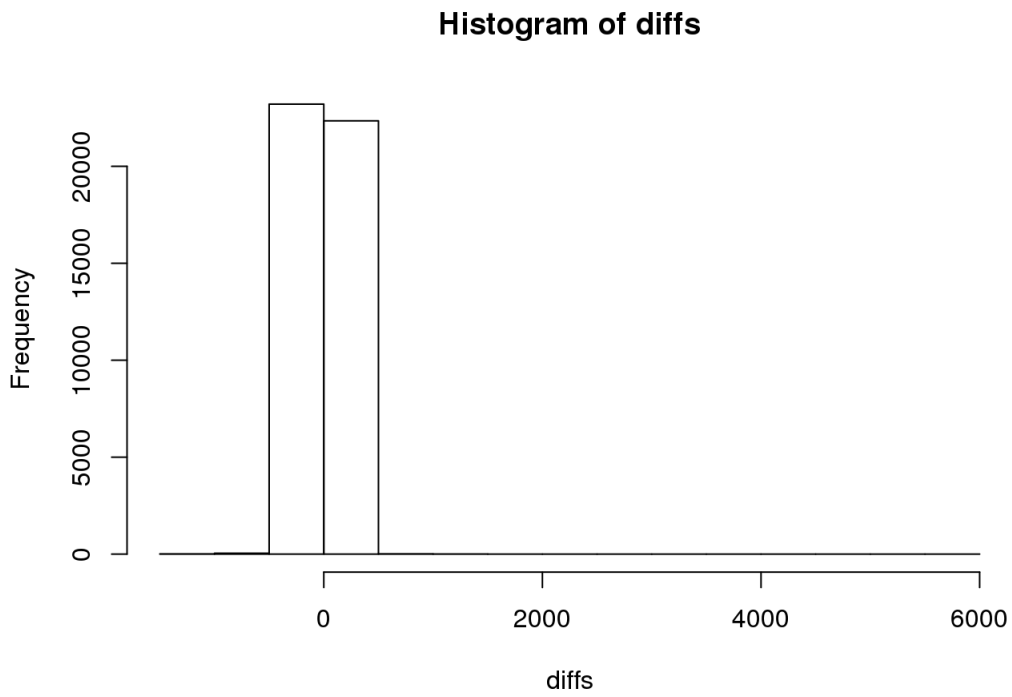
```
diffs <- txiInfRepGS$conf[[2]][nZeroInds,"Width"] - txiInfRepBoot$conf[[2]][nZeroInds,"Width"]
print(sum(diffs < 0))
```

```
## [1] 23266
```

```
print(sum(diffs > 0))
```

```
## [1] 22369
```

```
hist(diffs)
```




```
print(sum(abs(diffs) < 20))
```

```
## [1] 35372
```

```
print(sum(diffs < -20))
```

```
## [1] 5961
```

```
print(sum(diffs > 20))
```

```
## [1] 4302
```

Of the total 46K transcripts that have non zero counts, around 35K have a difference of less than 20

Real Data

Obtaining the transcripts with zero counts

```
i=1
zeroBInds <- which(txiInfRepBoot$counts[,i] == 0)
zeroGSInds <- which(txiInfRepGS$counts[,i] == 0)
print(length(zeroGSInds))
```

```
## [1] 121302
```

```
print(length(zeroBInds))
```

```
## [1] 121274
```

```
print(length(setdiff(zeroGSInds, zeroBInds)))
```

```
## [1] 49
```

```
print(length(setdiff(zeroBInds, zeroGSInds)))
```

```
## [1] 21
```

Obtaining transcripts that have zero means over the bootstrap runs

```
zeroMeanGS <- which(txiInfRepGS$conf[[i]][,3] <= 3)
zeroMeanB <- which(txiInfRepBoot$conf[[i]][,3] <= 3)

print(length(zeroMeanGS))
```

```
## [1] 110085
```

```
print(length(zeroMeanB))
```

```
## [1] 136226
```

```
print(length(setdiff(zeroBInds, zeroMeanB)))
```

```
## [1] 3305
```

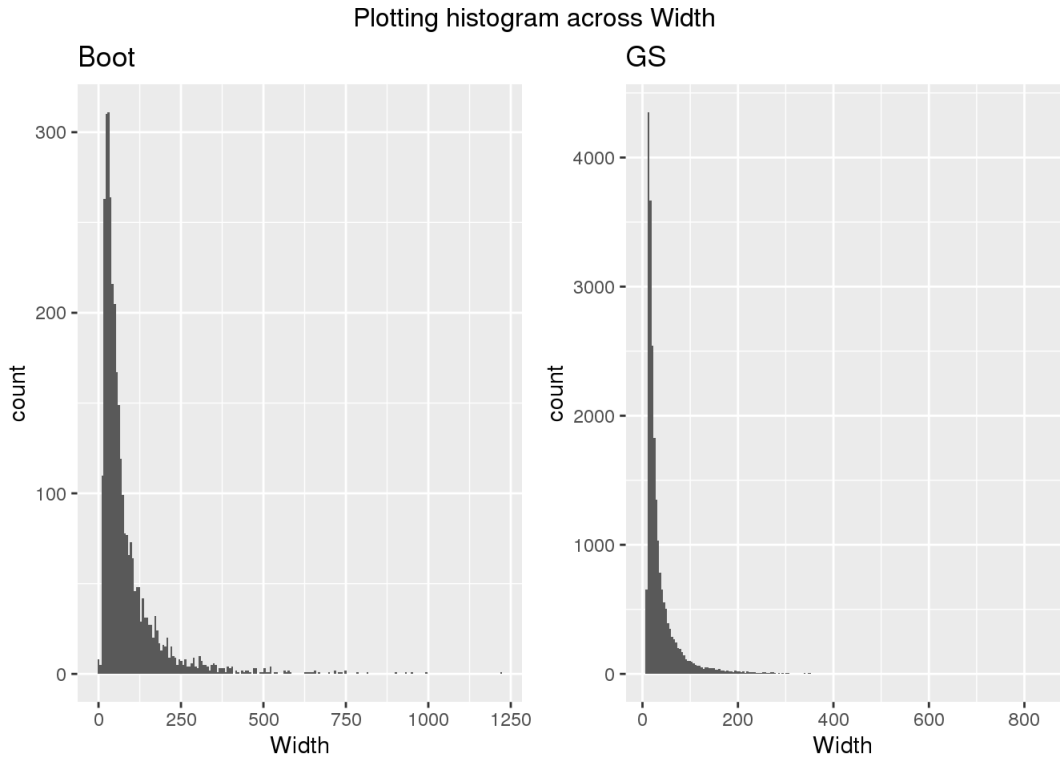
```
print(length(setdiff(zeroGSInds, zeroMeanGS)))
```

```
## [1] 21868
```

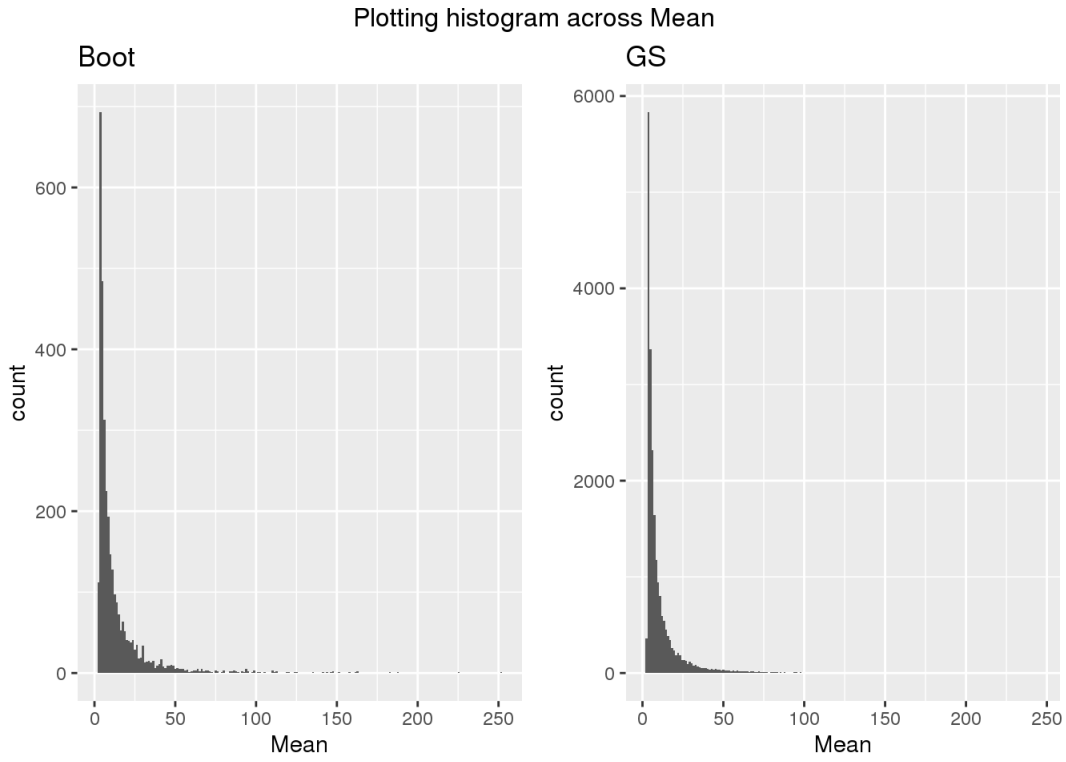
```
print(length(intersect(zeroMeanB, zeroMeanGS)))
```

```
## [1] 109848
```

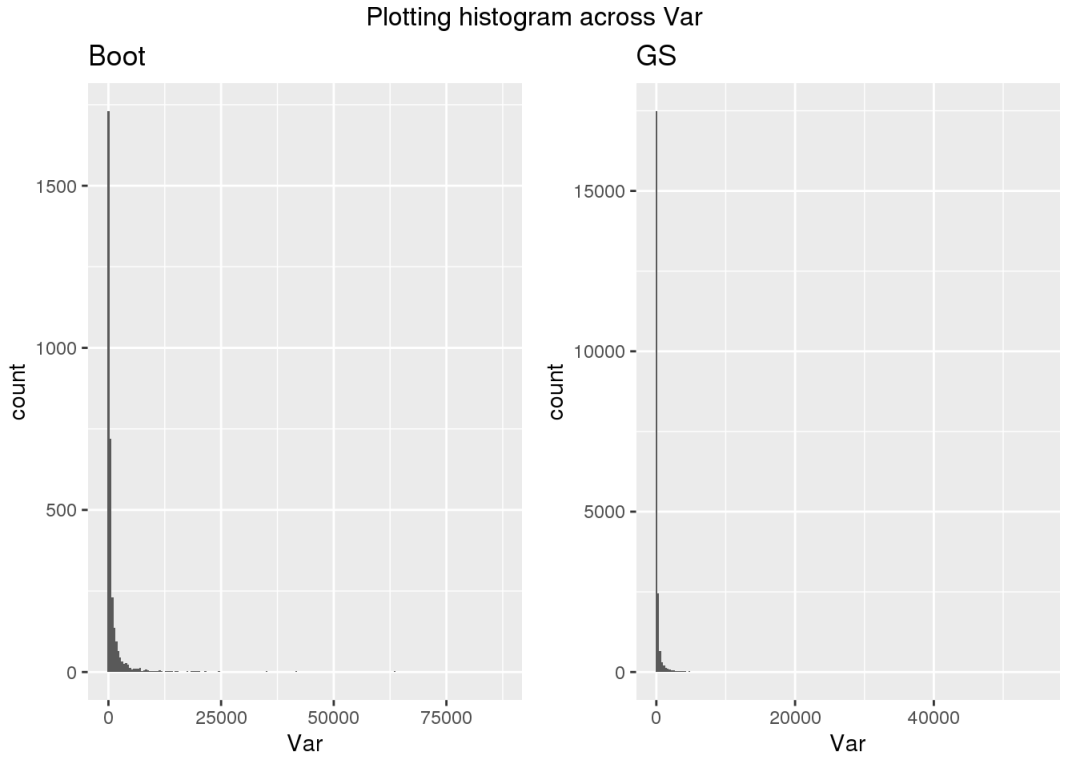
```
plotDfZeros <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[i]], "GS" = txiInfRepGS$conf[[i]]), list("Boot
" = setdiff(zeroBInds, zeroMeanB), "GS" = setdiff(zeroGSInds, zeroMeanGS)))
plotHist(plotDfZeros, "Width")
```




```
plotHist(plotDfZeros, "Mean")
```




```
plotHist(plotDfZeros, "Var")
```

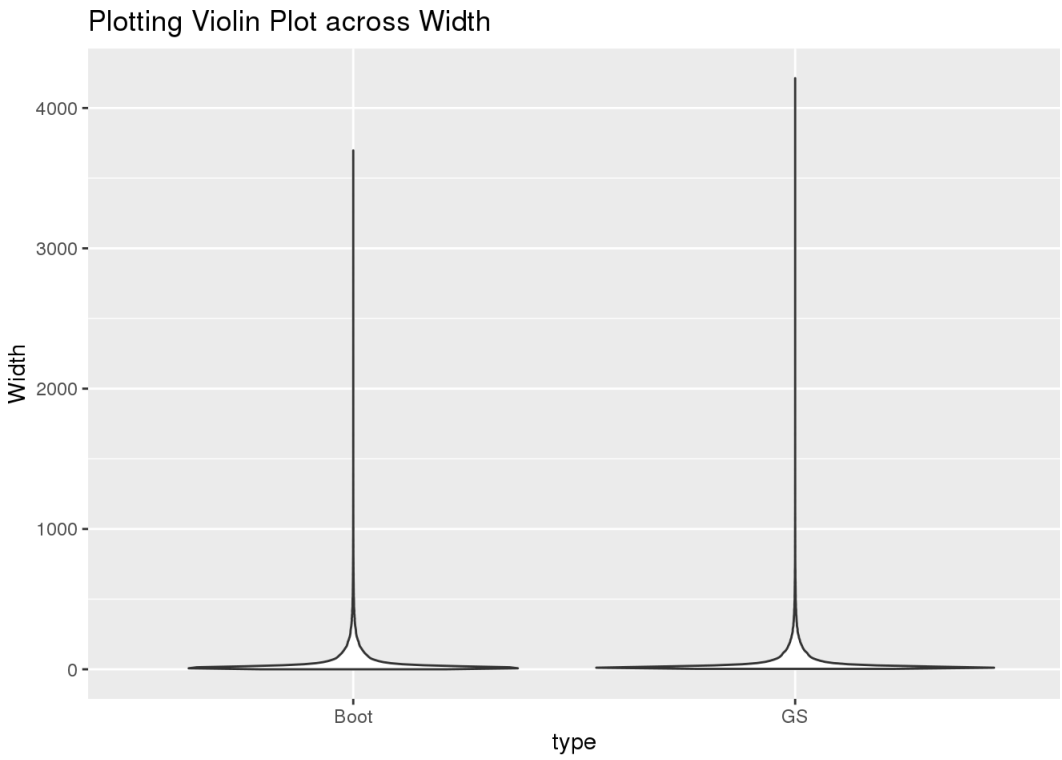


Only observing the transcripts having non zero counts

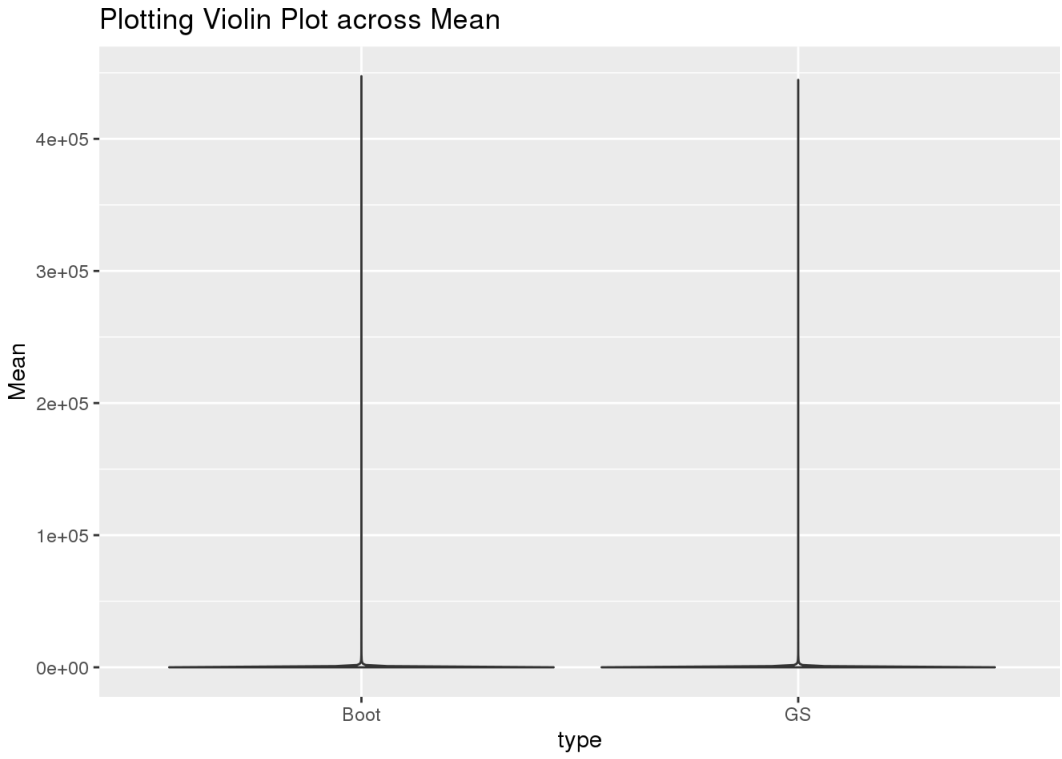
```
nZeroInds <- setdiff(1:nrow(txiInfRepBoot$conf[[i]]), union(zeroBInds, zeroGSInds))
print(length(nZeroInds))
```

```
## [1] 81704
```

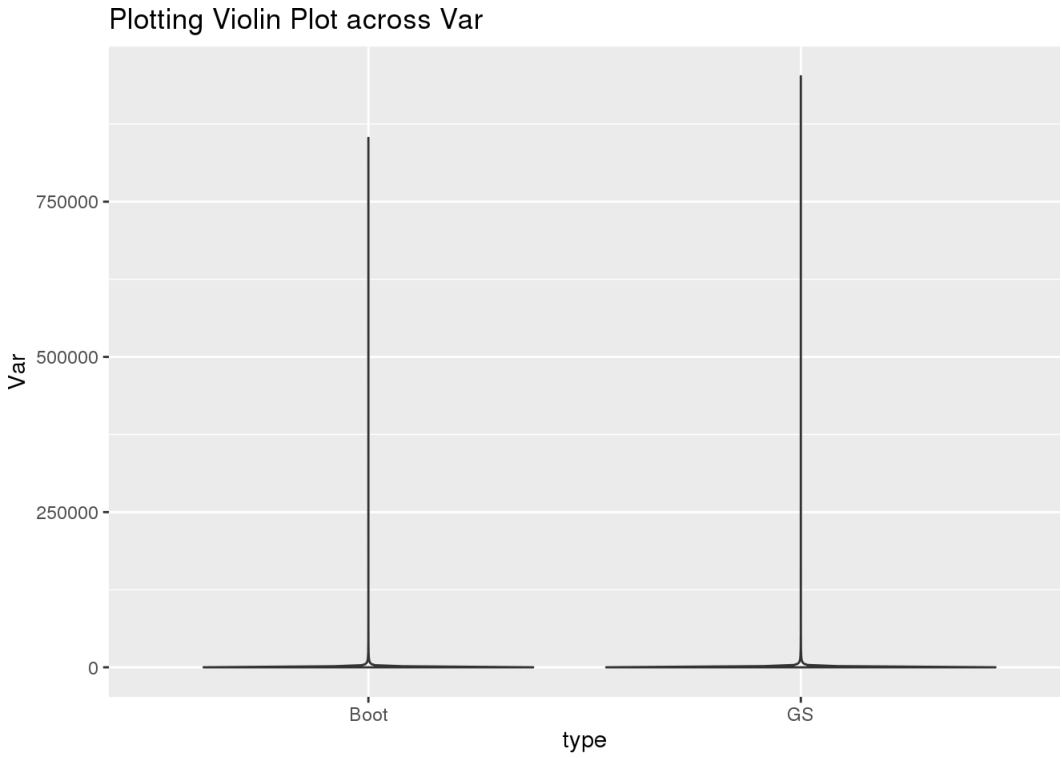
```
plotDfNonZeros <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[i]], "GS" = txiInfRepGS$conf[[i]]),
                               list("Boot" = nZeroInds, "GS" = nZeroInds))
plotViol(plotDfNonZeros, "Width")
```




```
plotViol(plotDfNonZeros, "Mean")
```




```
plotViol(plotDfNonZeros, "Var")
```



Transcripts with 95% CI of one is less than 5% CI of other (GS Dominates)

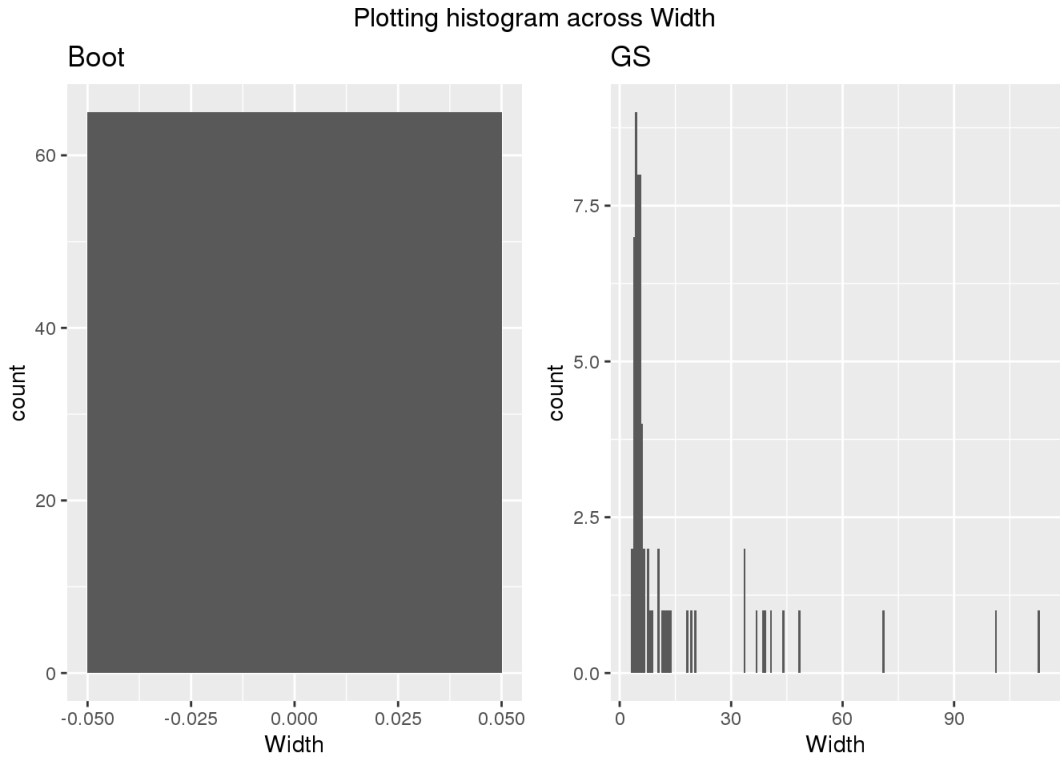
```
GSLargeInds <- which(txiInfRepBoot$conf[[i]][nZeroInds,2] < txiInfRepGS$conf[[i]][nZeroInds,1])
bootLargeInds <- which(txiInfRepGS$conf[[i]][nZeroInds,2] < txiInfRepBoot$conf[[i]][nZeroInds,1])

print(length(GSLargeInds))
```

```
## [1] 65
```

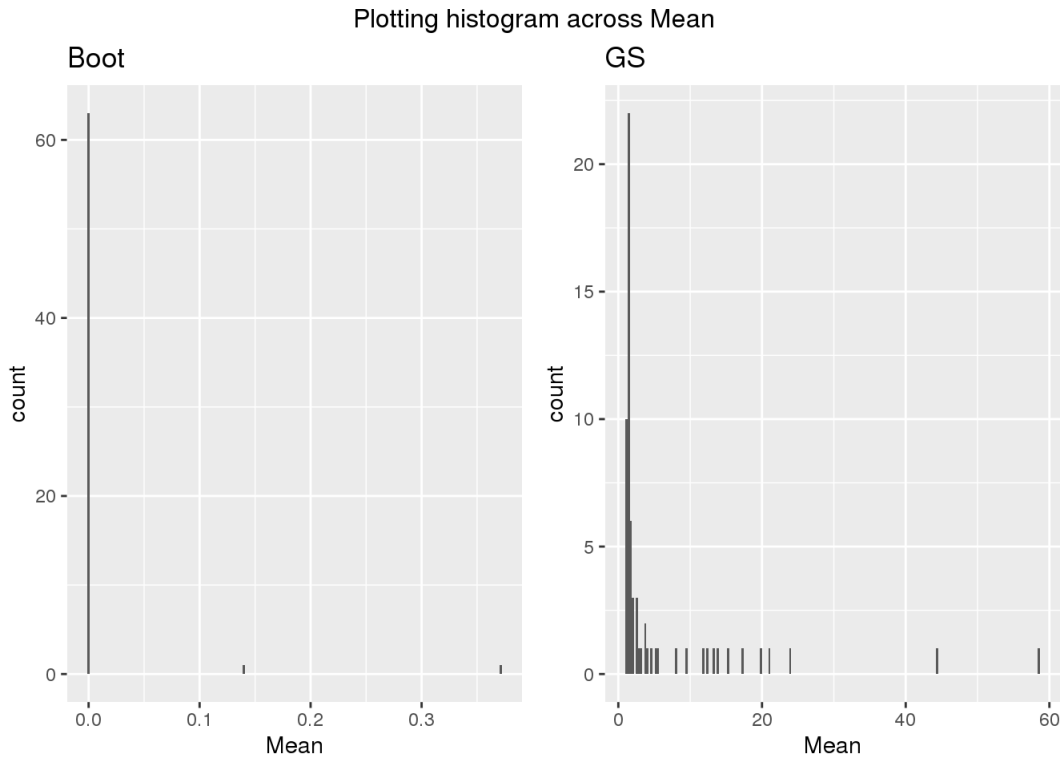
```
print(length(bootLargeInds))
```

```
## [1] 1
```

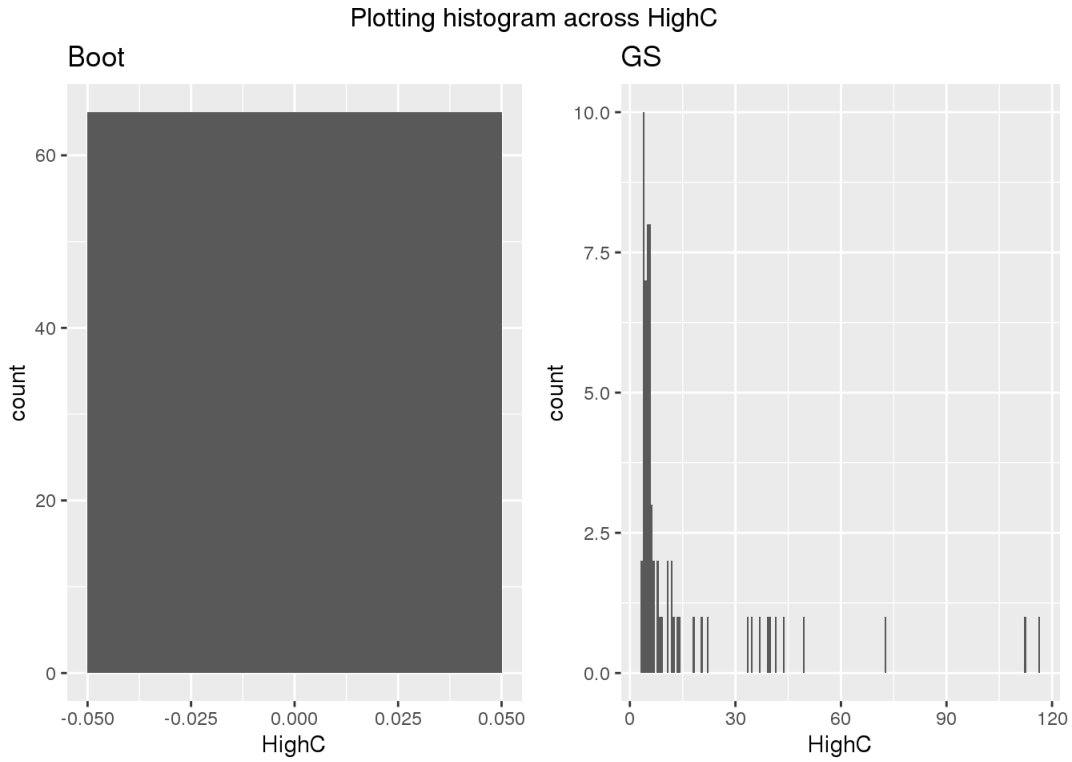
```
pGSLargeDf <- createPlotDf(list("Boot" = txiInfRepBoot$conf[[i]], "GS" = txiInfRepGS$conf[[i]]),
                           list("Boot" = nZeroInds[GSLargeInds], "GS" = nZeroInds[GSLargeInds]))
plotHist(pGSLargeDf, "Width")
```



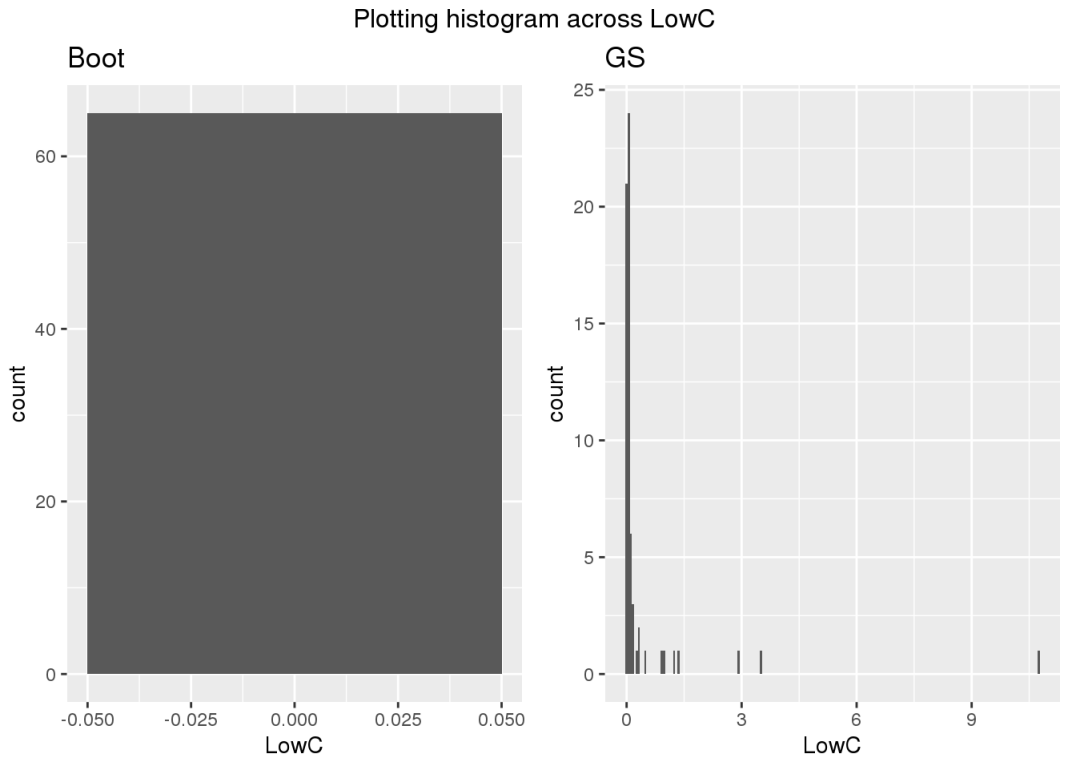

```
plotHist(pGSLargeDf, "Mean")
```




```
plotHist(pGSLargeDf, "HighC")
```




```
plotHist(pGSLargeDf, "LowC")
```



Computing the difference between magnitude of difference b/w the two

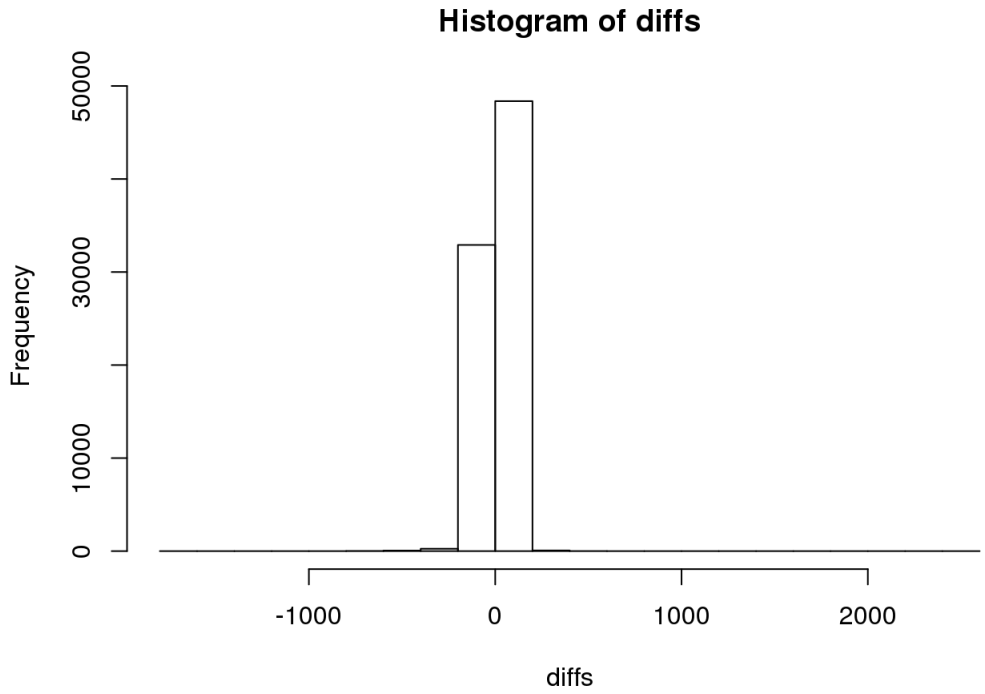
```
diffs <- txiInfRepGS$conf[[i]][nZeroInds,"Width"] - txiInfRepBoot$conf[[i]][nZeroInds,"Width"]
print(sum(diffs < 0))
```

```
## [1] 33248
```

```
print(sum(diffs > 0))
```

```
## [1] 48456
```

```
hist(diffs)
```




```
print(sum(abs(diffs) < 20))
```

```
## [1] 70378
```

```
print(sum(diffs < -20))
```

```
## [1] 7304
```

```
print(sum(diffs > 20))
```

```
## [1] 4022
```