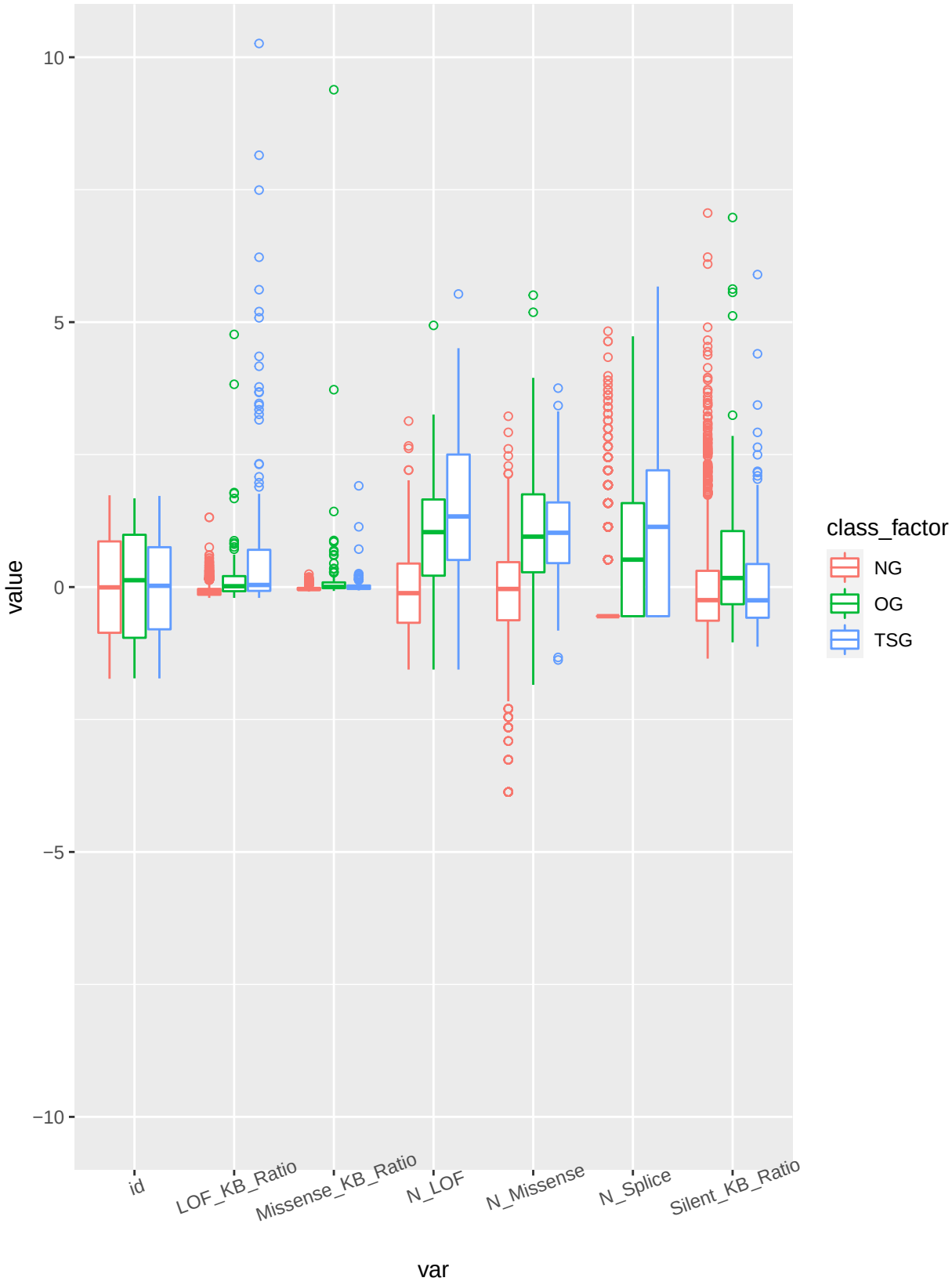
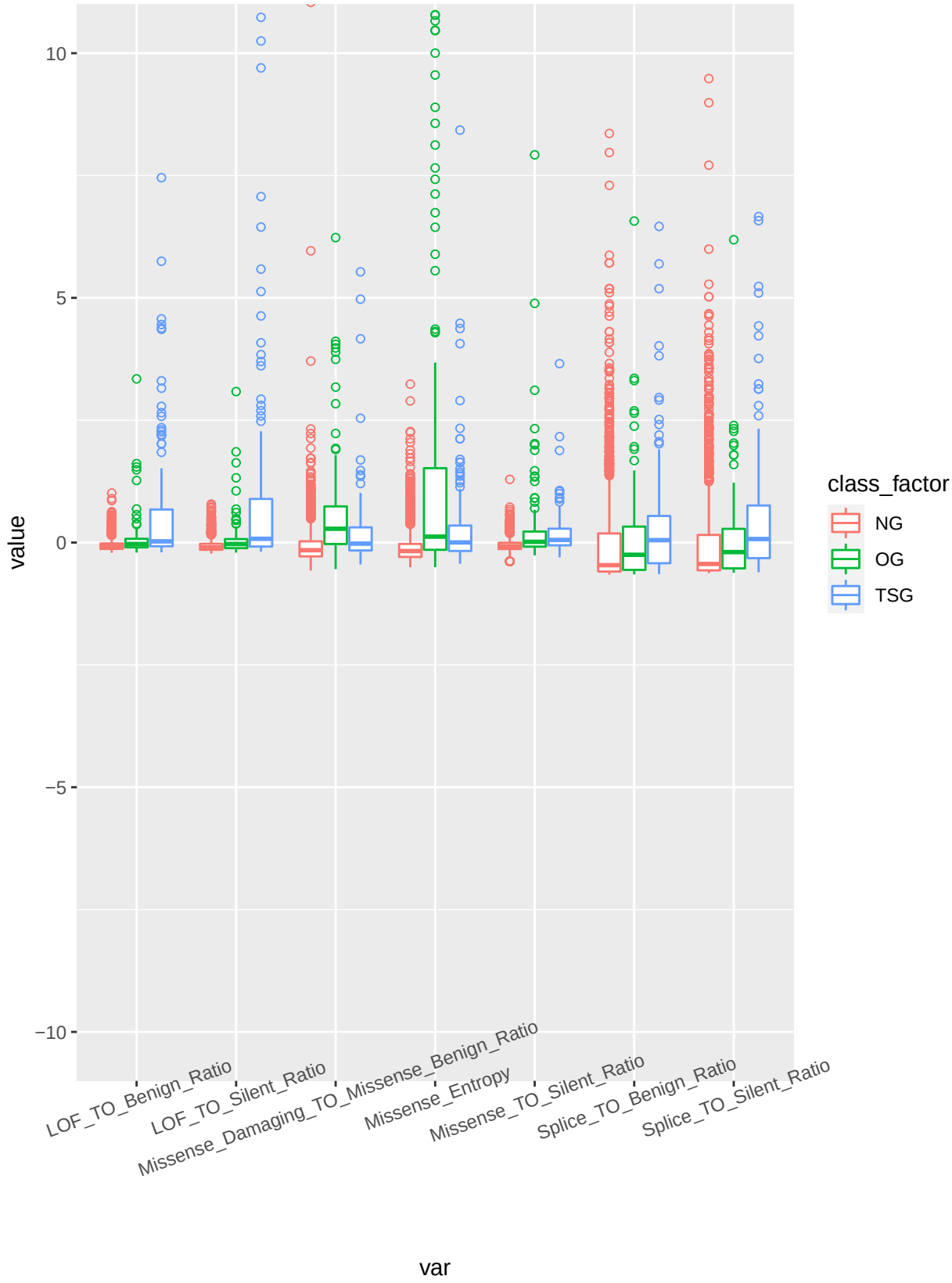
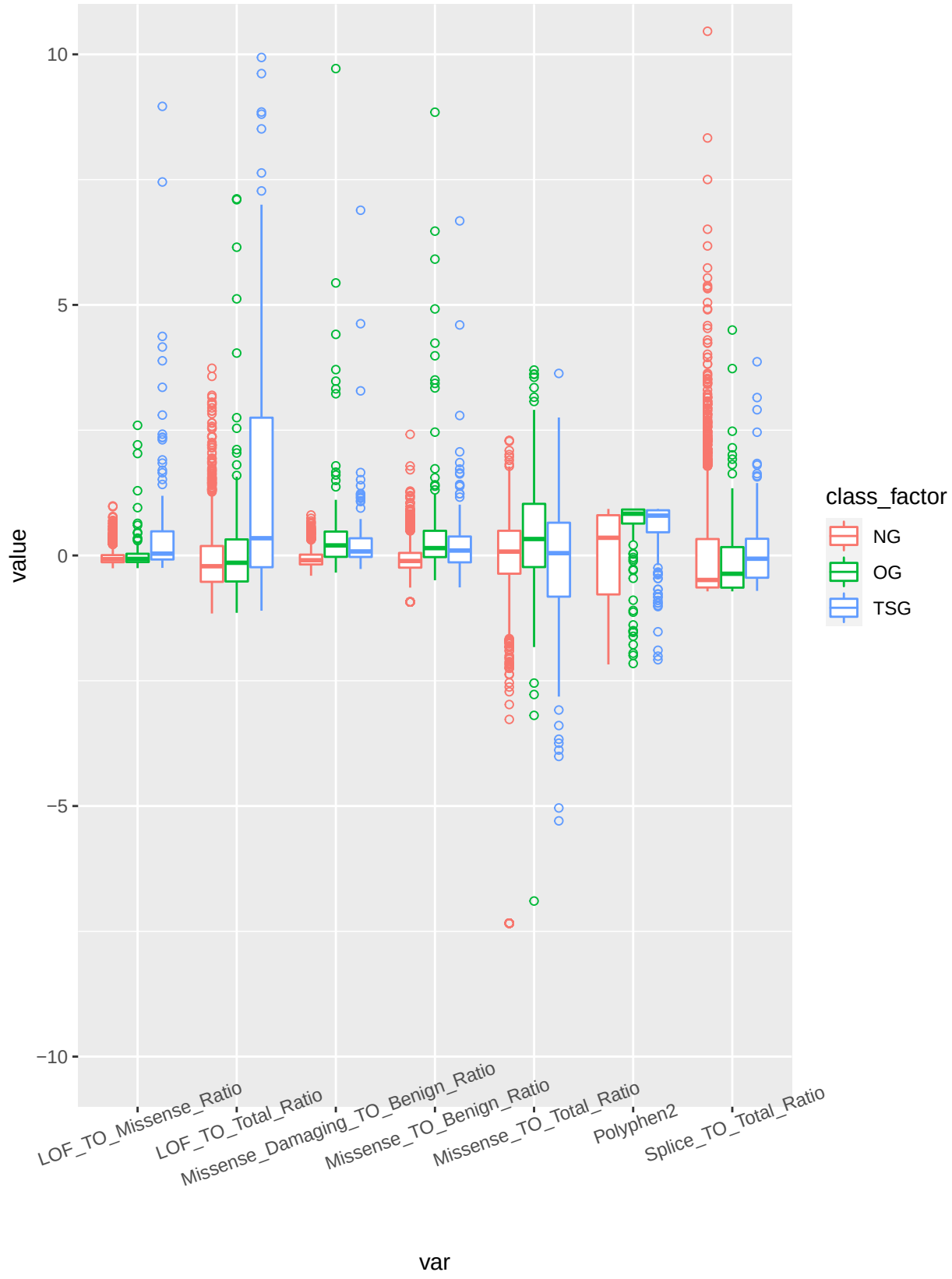
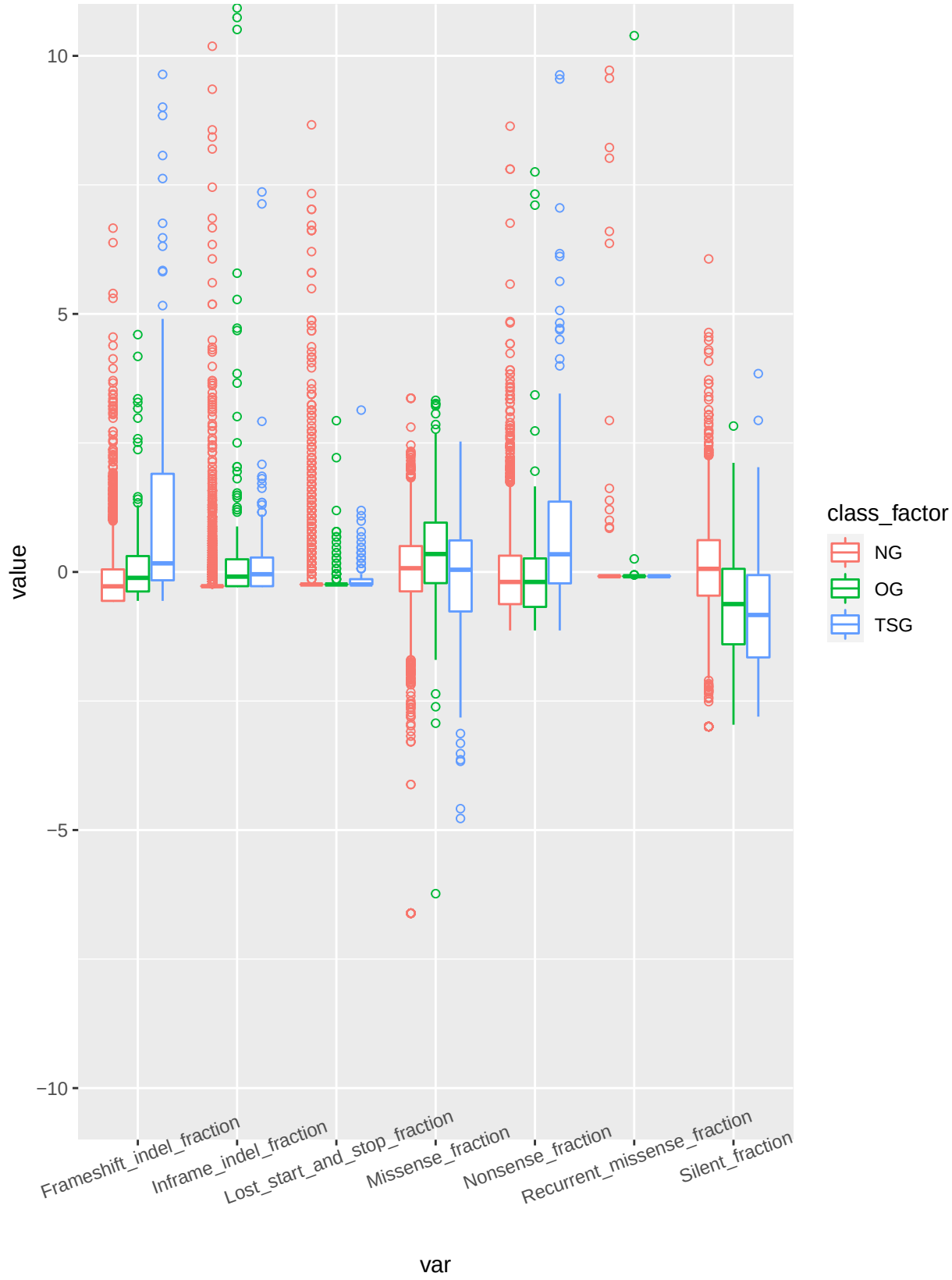
```r
source("clean_training_v2.R")
```

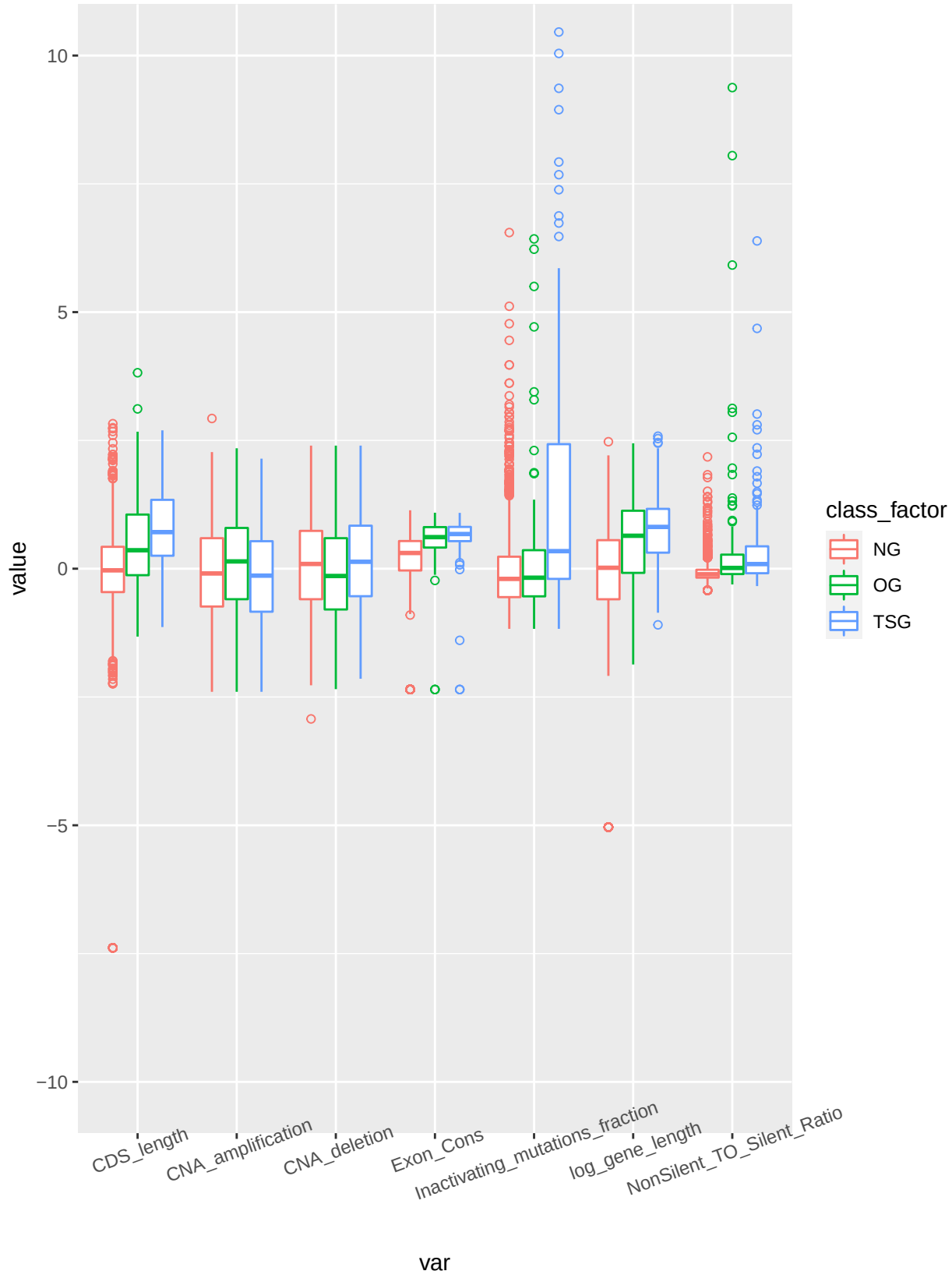
```r
library(ggplot2)

# Box plots of each variable (7 per plot) split by class
for (i in seq(1, ncol(training0) - 2, by = 7)) {
  # Select 7 columns plus class
  plot <- training0[, c(names(training0)[seq(i, i + 6)], "class_factor")] %>%
    # Scale and center numeric variables
    mutate_if(is.numeric, scale) %>%
    # Turns 7 columns into 2 columns: variable name and value
    gather(-class_factor, key = "var", value = "value") %>%
    ggplot(aes(x = var, y = value, color = class_factor)) +
      geom_boxplot(outlier.shape = 1) +
      theme(axis.text.x = element_text(angle = 20)) +
      coord_cartesian(ylim = c(-10, 10))

  print(plot)
}
```

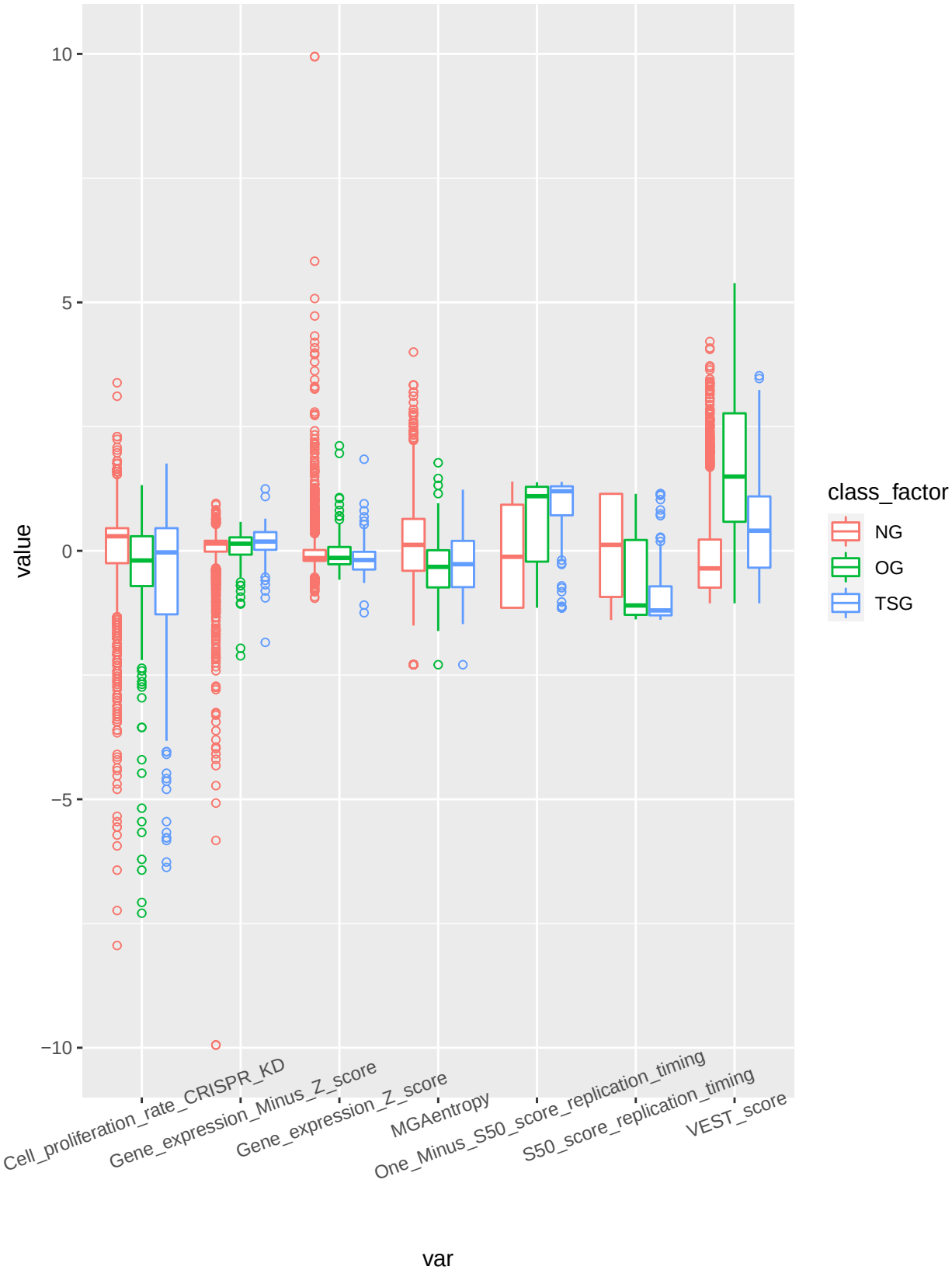

10
5
0
−5
−10
value
id
LOF_KB_Ratio
Missense_KB_Ratio
N_LOF
N_Missense
N_Splice
Silent_KB_Ratio
var
class_factor
NG
OG
TSG

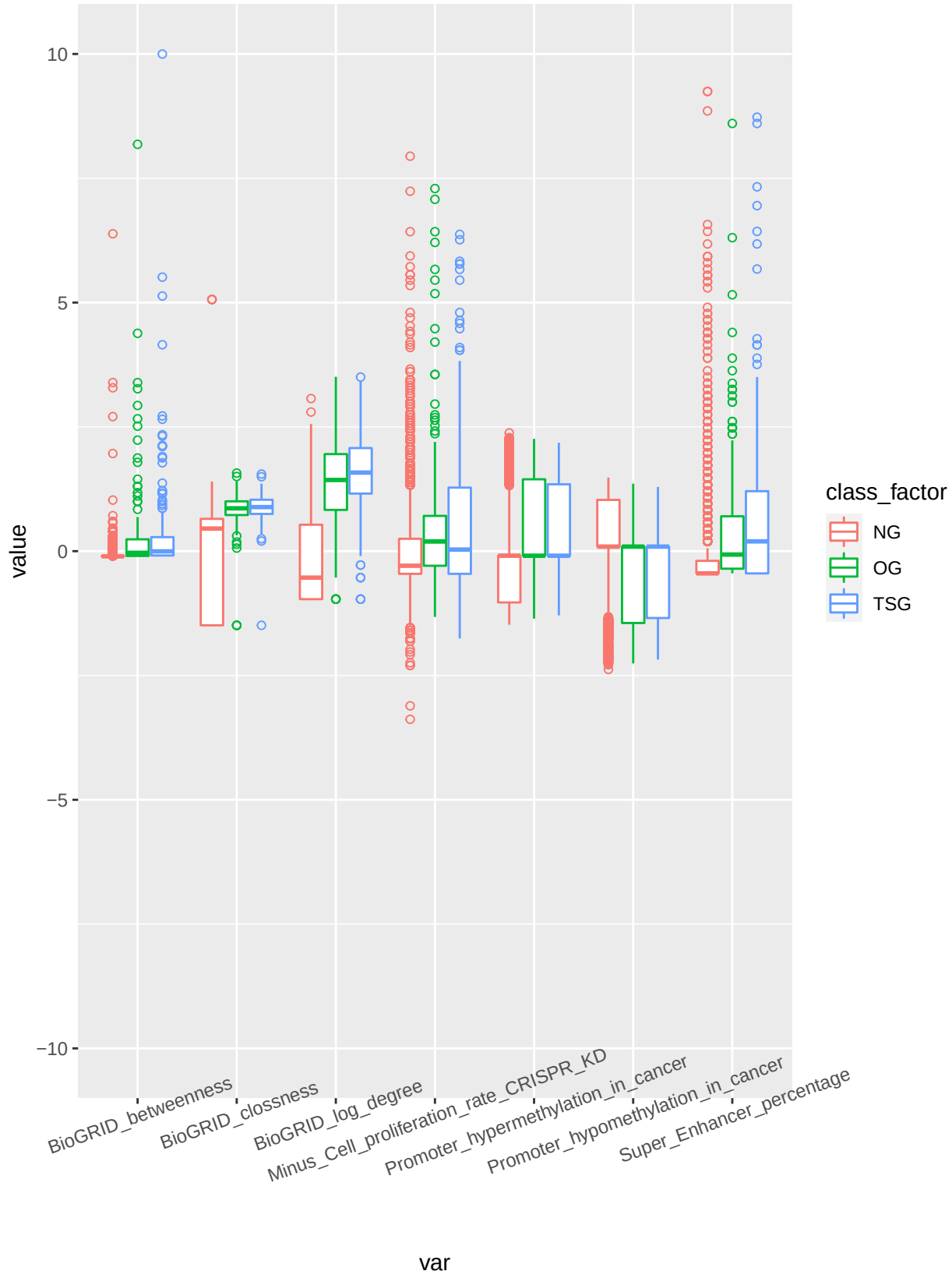

10
5
0
−5
−10
value
LOF_TO_Benign_Ratio
LOF_TO_Silent_Ratio
Missense_Damaging_TO_Missense_Benign_Ratio
Missense_Entropy
Missense_TO_Silent_Ratio
Splice_TO_Benign_Ratio
Splice_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

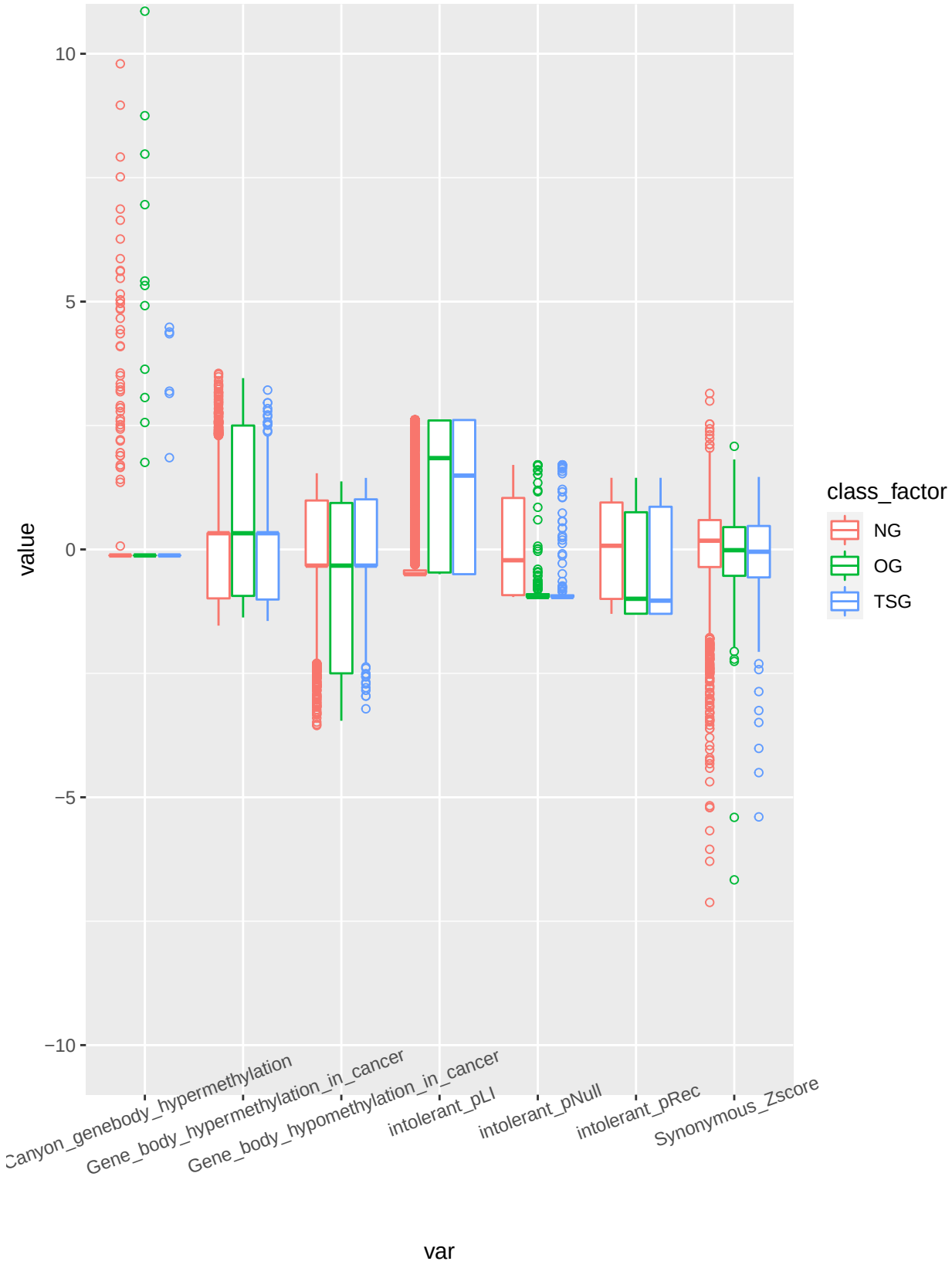

10
5
0
−5
−10
value
LOF_TO_Missense_Ratio
LOF_TO_Total_Ratio
Missense_Damaging_TO_Benign_Ratio
Missense_TO_Benign_Ratio
Missense_TO_Total_Ratio
Polyphen2
Splice_TO_Total_Ratio
var
class_factor
NG
OG
TSG

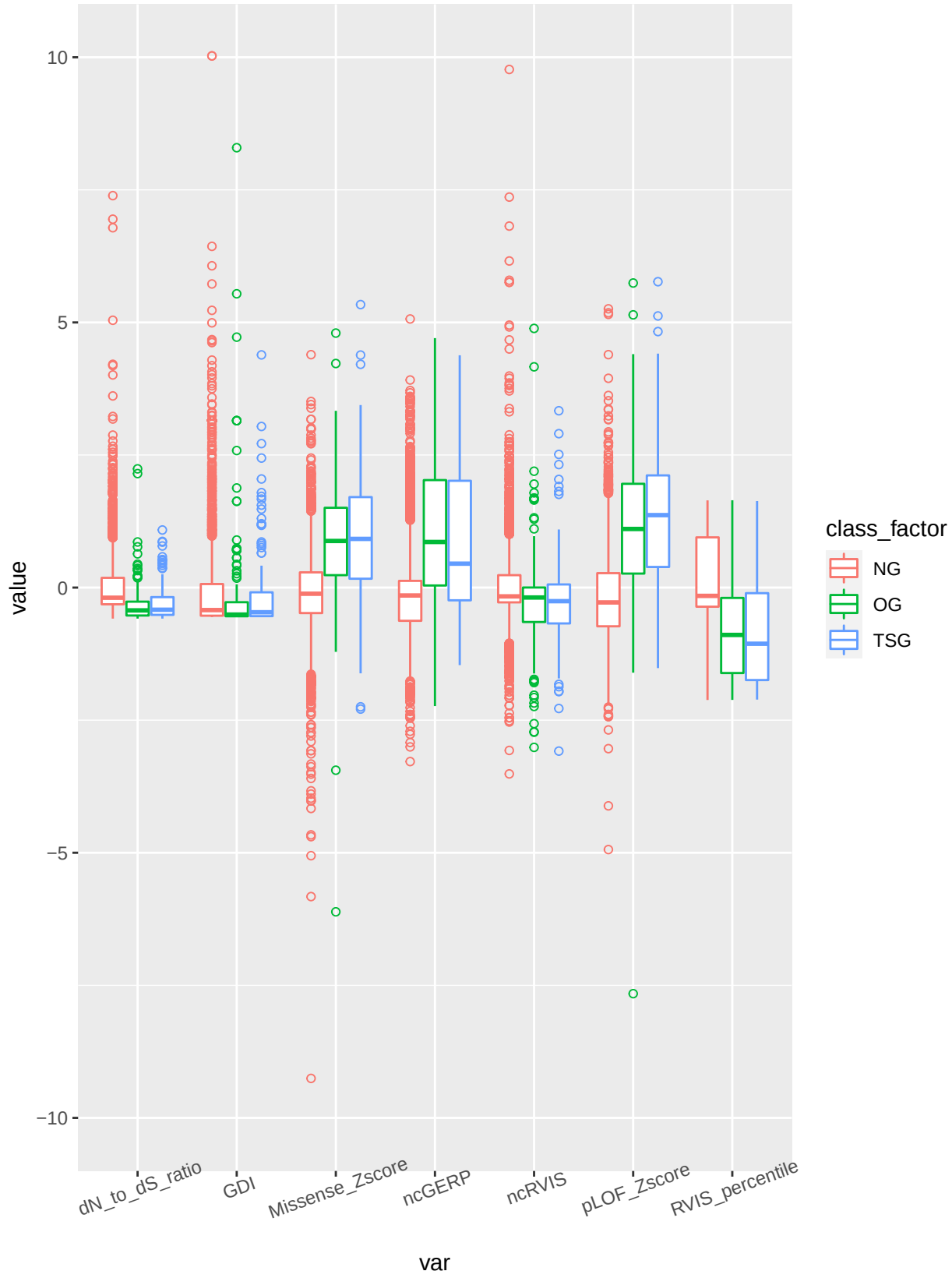

value
10
5
0
−5
−10
var
Frameshift_indel_fraction
Inframe_indel_fraction
Lost_start_and_stop_fraction
Missense_fraction
Nonsense_fraction
Recurrent_missense_fraction
Silent_fraction
class_factor
NG
OG
TSG

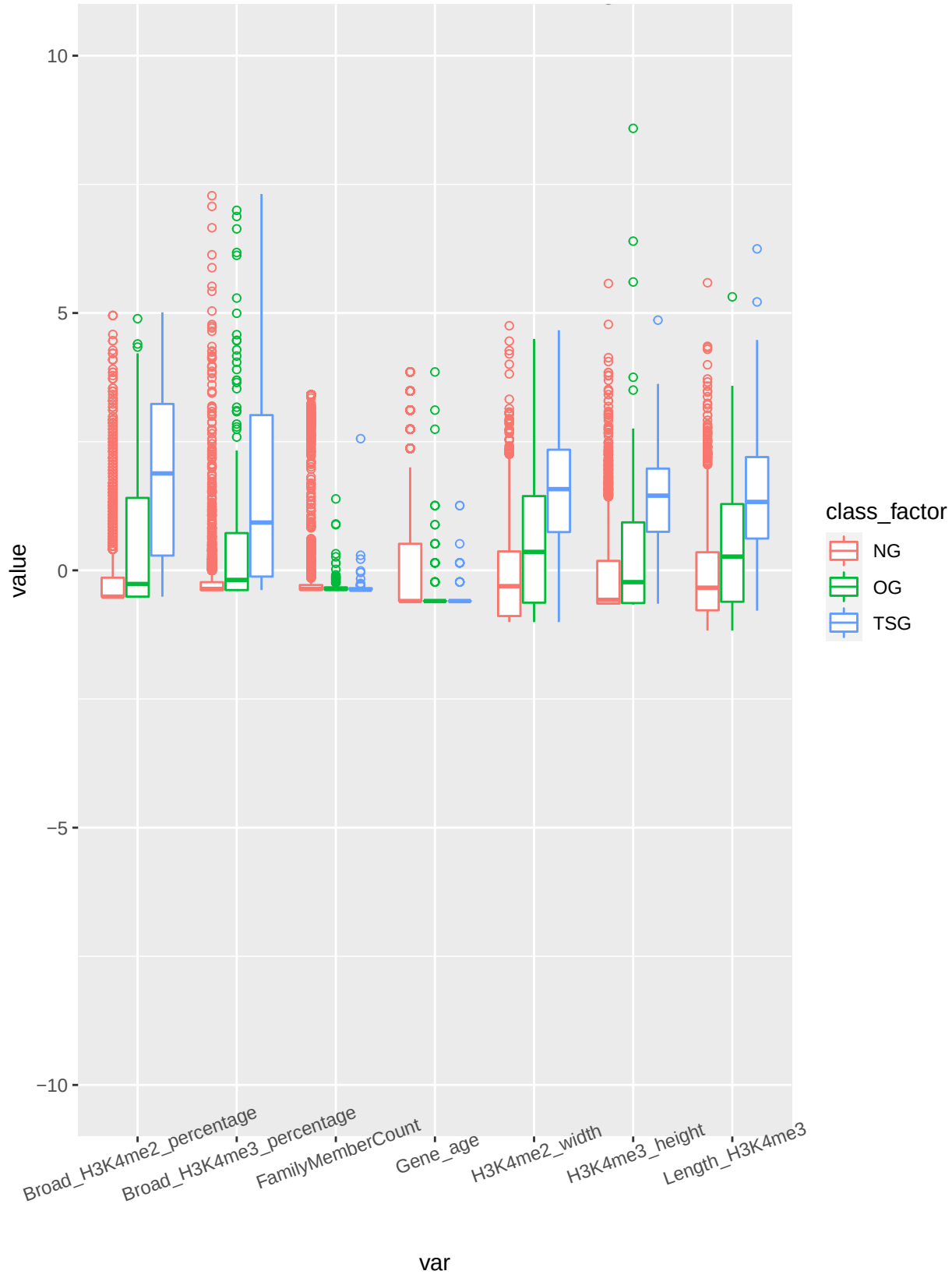

10
5
0
−5
−10
value
CDS_length
CNA_amplification
CNA_deletion
Exon_Cons
Inactivating_mutations_fraction
log_gene_length
NonSilent_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

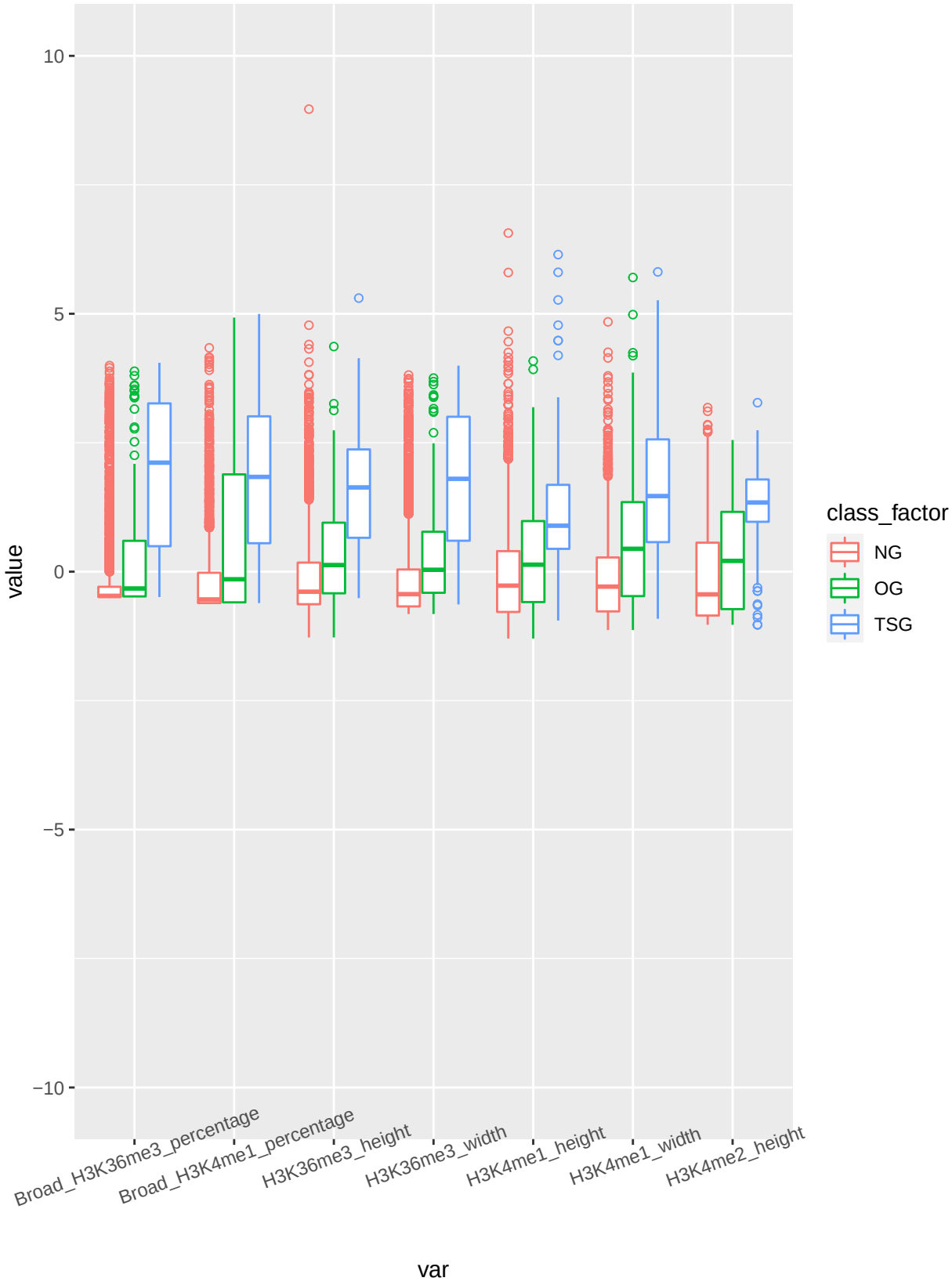

10
5
0
−5
−10
value
Cell_proliferation_rate_CRISPR_KD
Gene_expression_Minus_Z_score
Gene_expression_Z_score
MGAentropy
One_Minus_S50_score_replication_timing
S50_score_replication_timing
VEST_score
var
class_factor
NG
OG
TSG

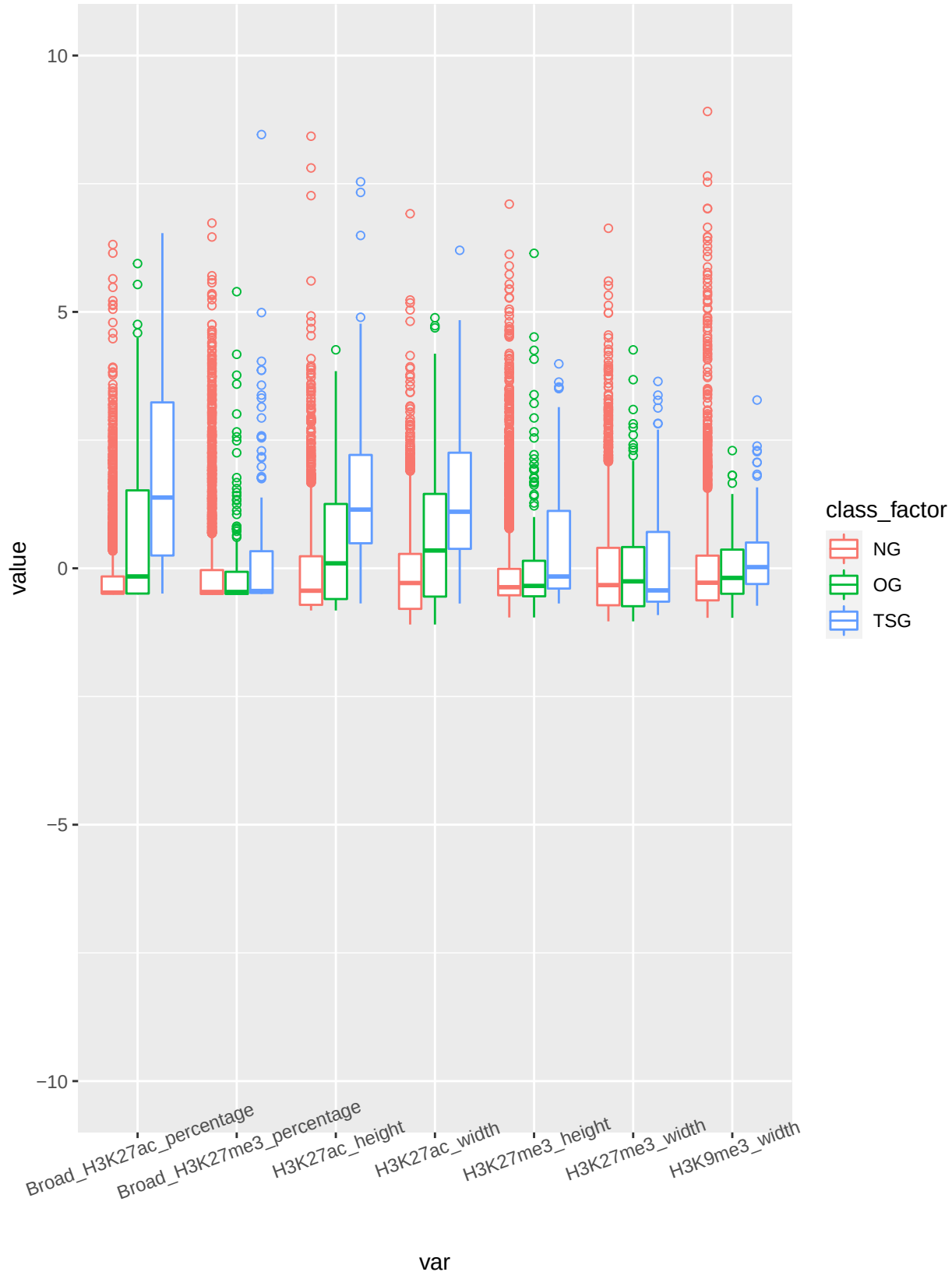

10
5
0
−5
−10
value
BioGRID_betweenness
BioGRID_clossness
BioGRID_log_degree
Minus_Cell_proliferation_rate_CRISPR_KD
Promoter_hypermethylation_in_cancer
Promoter_hypomethylation_in_cancer
Super_Enhancer_percentage
var
class_factor
NG
OG
TSG

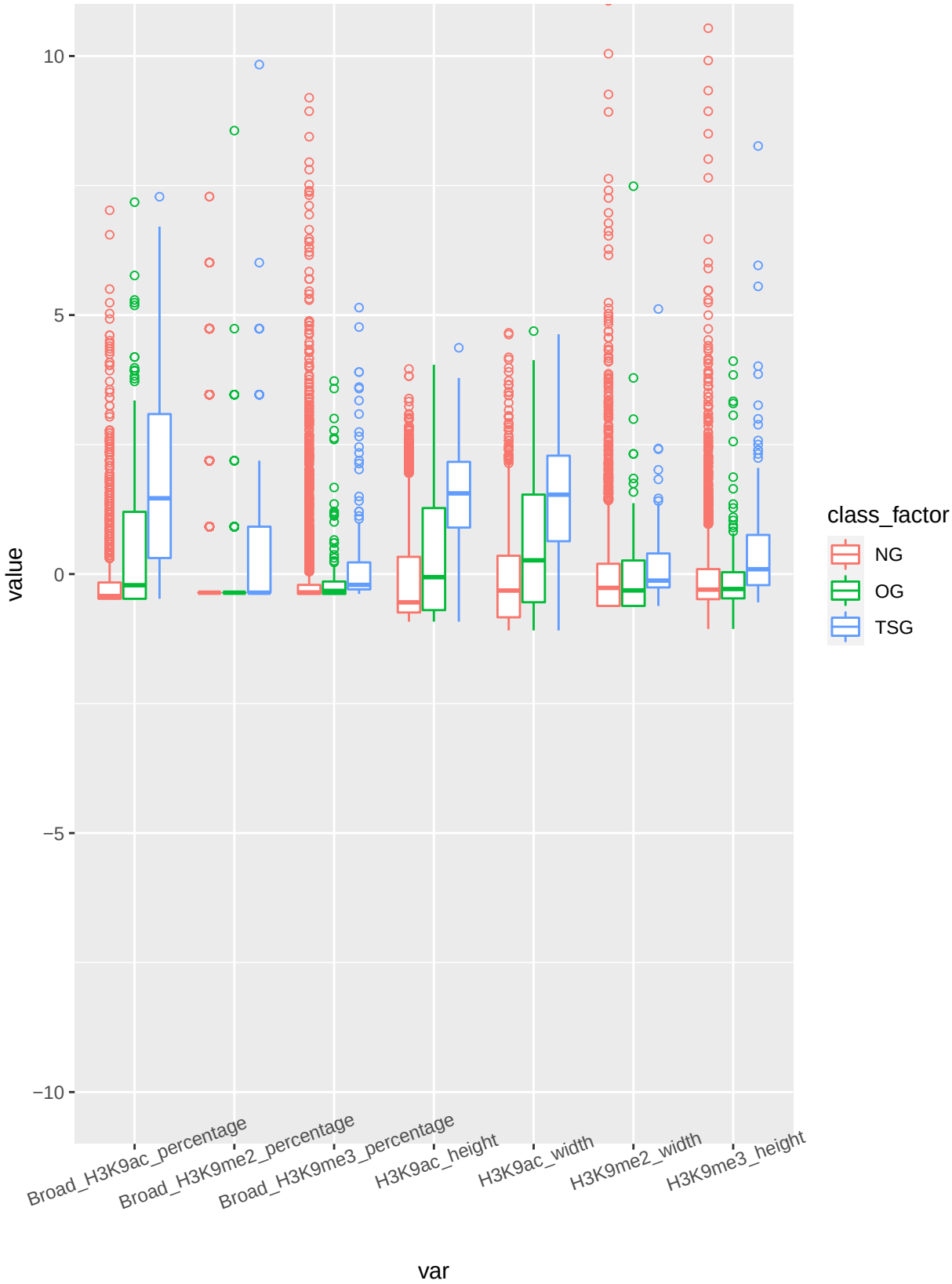

10
5
0
−5
−10
value
Canyon_genebody_hypermethylation
Gene_body_hypermethylation_in_cancer
Gene_body_hypomethylation_in_cancer
intolerant_pLI
intolerant_pNull
intolerant_pRec
Synonymous_Zscore
var
class_factor
NG
OG
TSG

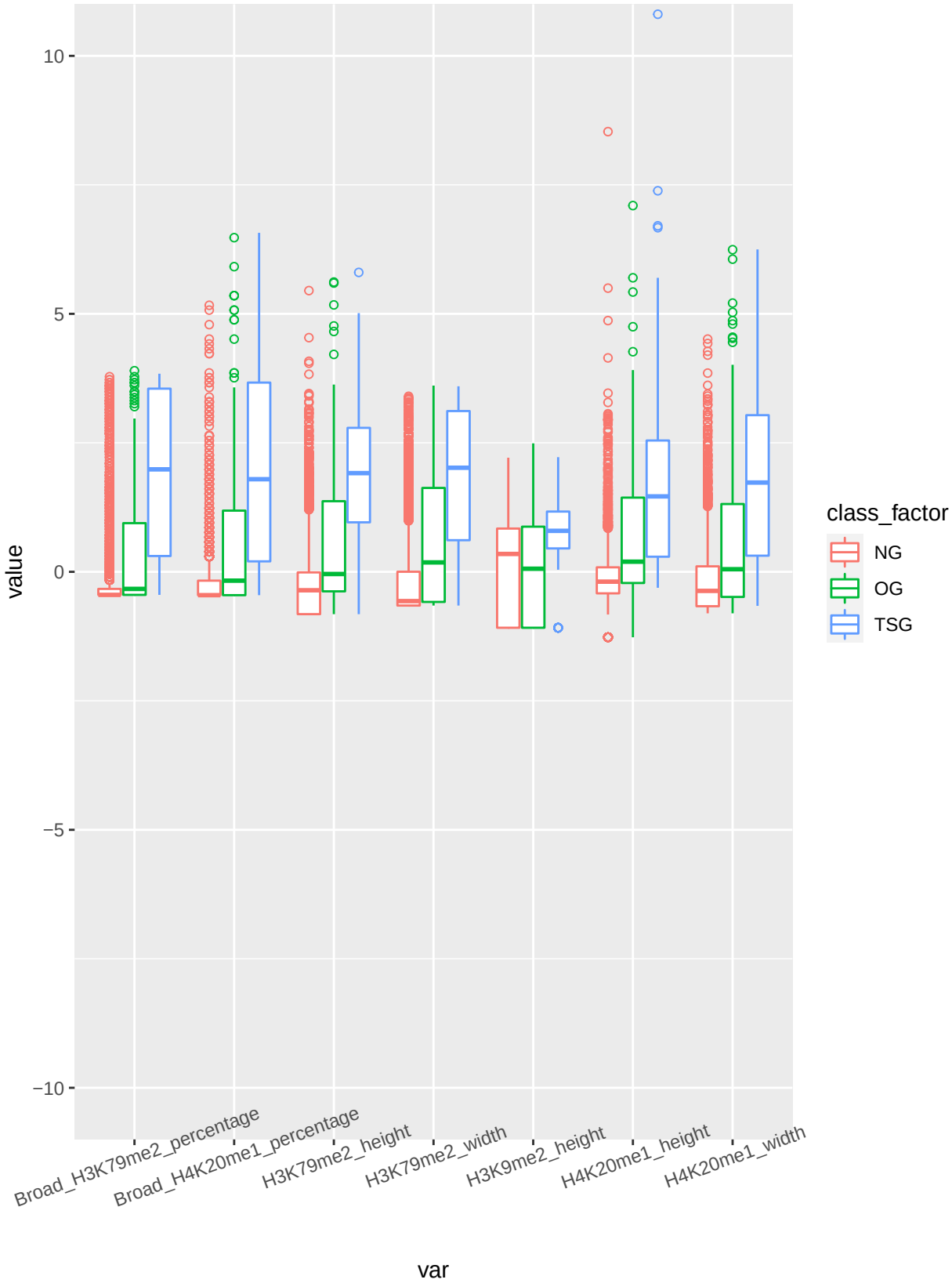

10
5
0
−5
−10
value
dN_to_dS_ratio
GDI
Missense_Zscore
ncGERP
ncRVIS
pLOF_Zscore
RVIS_percentile
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K4me2_percentage
Broad_H3K4me3_percentage
FamilyMemberCount
Gene_age
H3K4me2_width
H3K4me3_height
Length_H3K4me3
var
class_factor
NG
OG
TSG

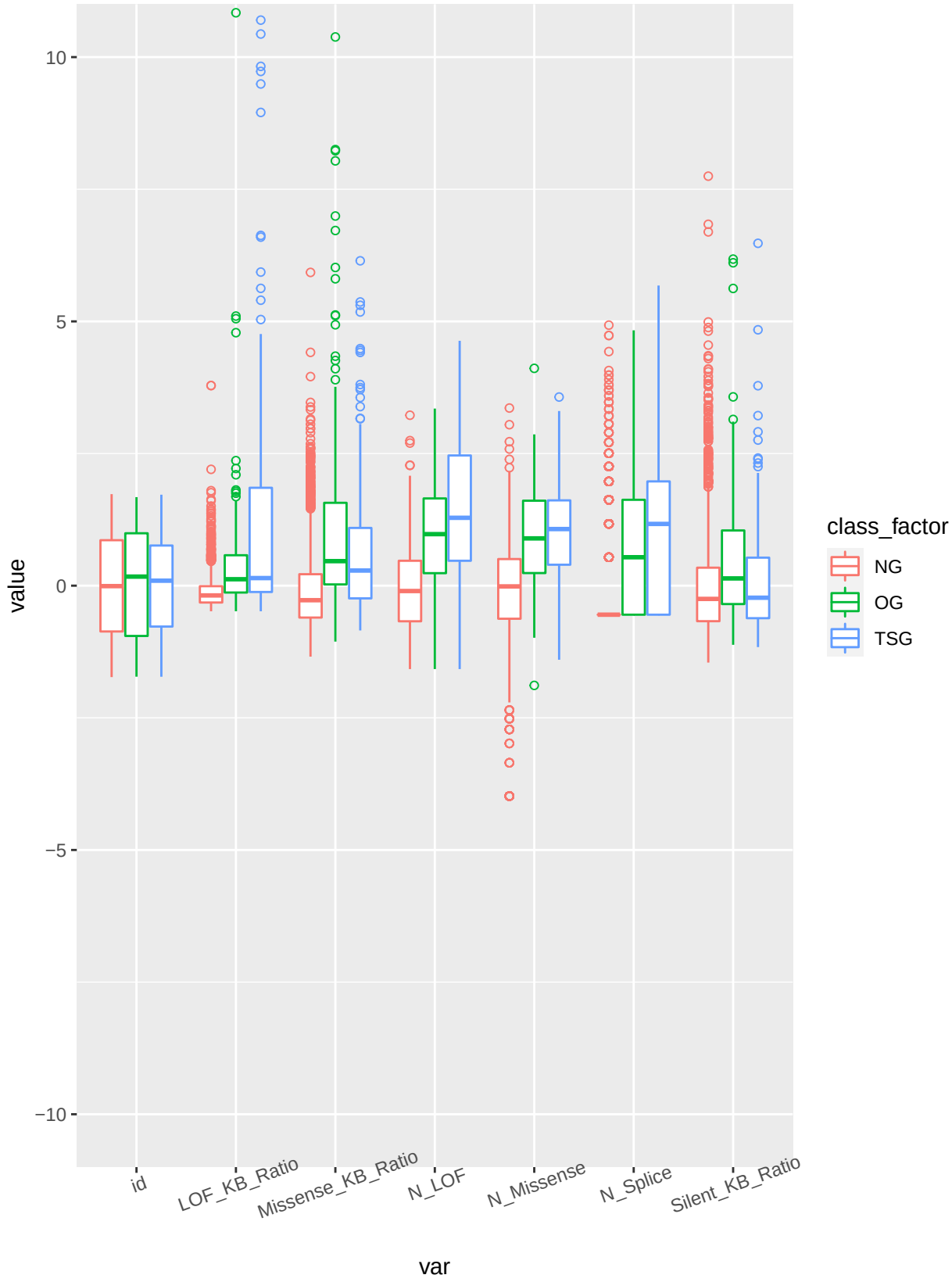

10
5
0
−5
−10
value
Broad_H3K36me3_percentage
Broad_H3K4me1_percentage
H3K36me3_height
H3K36me3_width
H3K4me1_height
H3K4me1_width
H3K4me2_height
var
class_factor
NG
OG
TSG

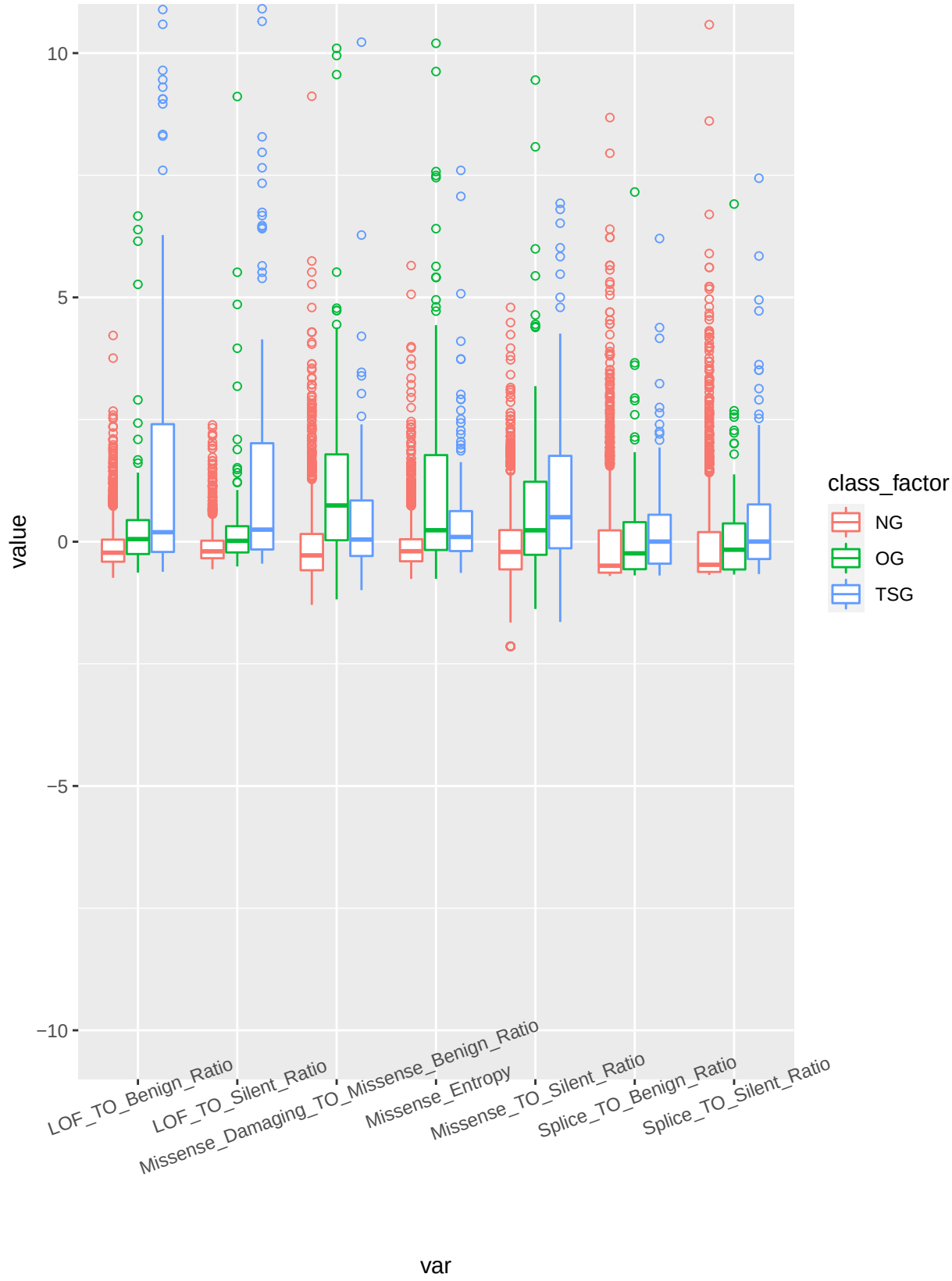

10
5
0
−5
−10
value
Broad_H3K27ac_percentage
Broad_H3K27me3_percentage
H3K27ac_height
H3K27ac_width
H3K27me3_height
H3K27me3_width
H3K9me3_width
var
class_factor
NG
OG
TSG

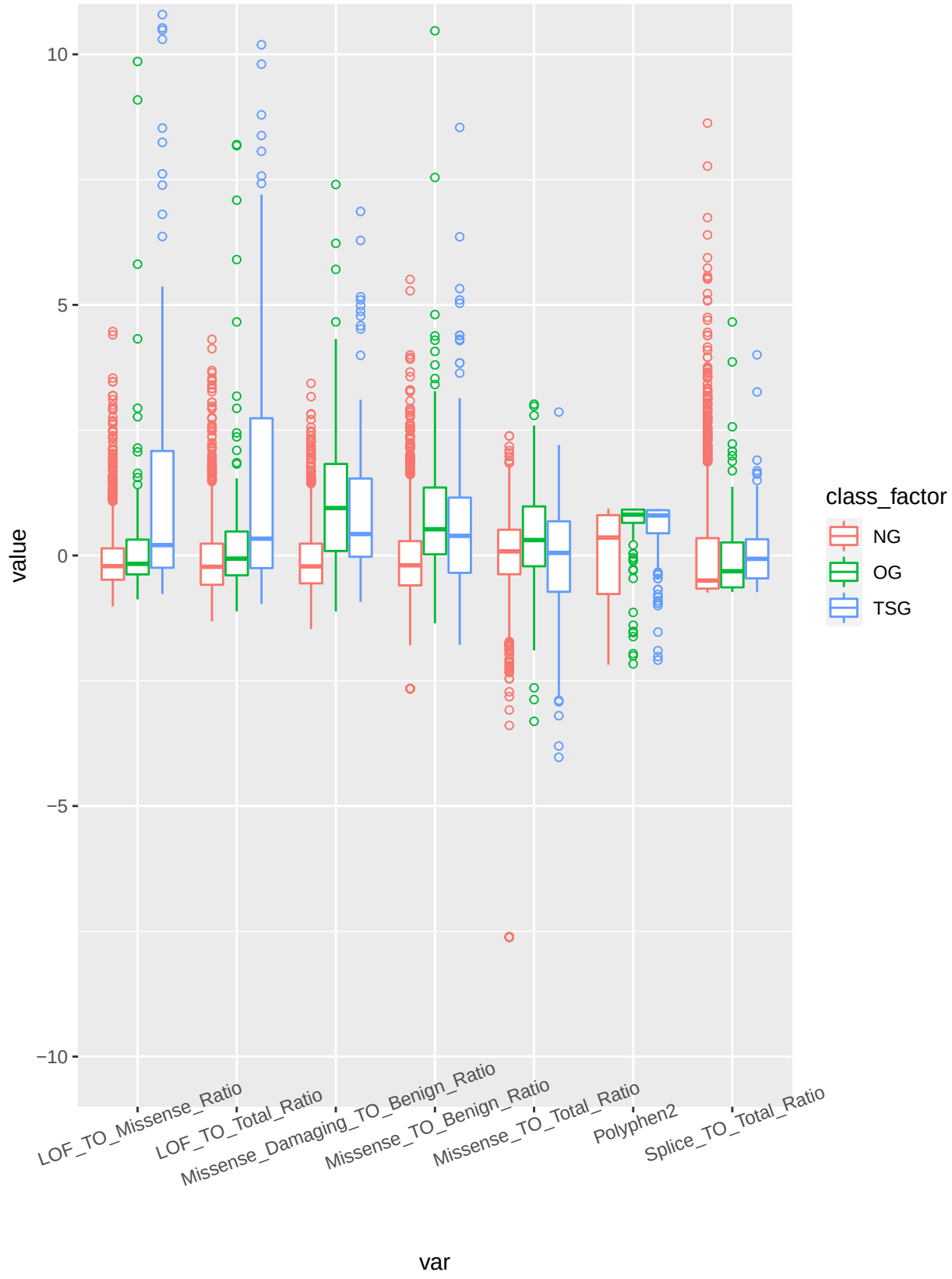

10
5
0
−5
−10
value
Broad_H3K9ac_percentage
Broad_H3K9me2_percentage
Broad_H3K9me3_percentage
H3K9ac_height
H3K9ac_width
H3K9me2_width
H3K9me3_height
var
class_factor
NG
OG
TSG

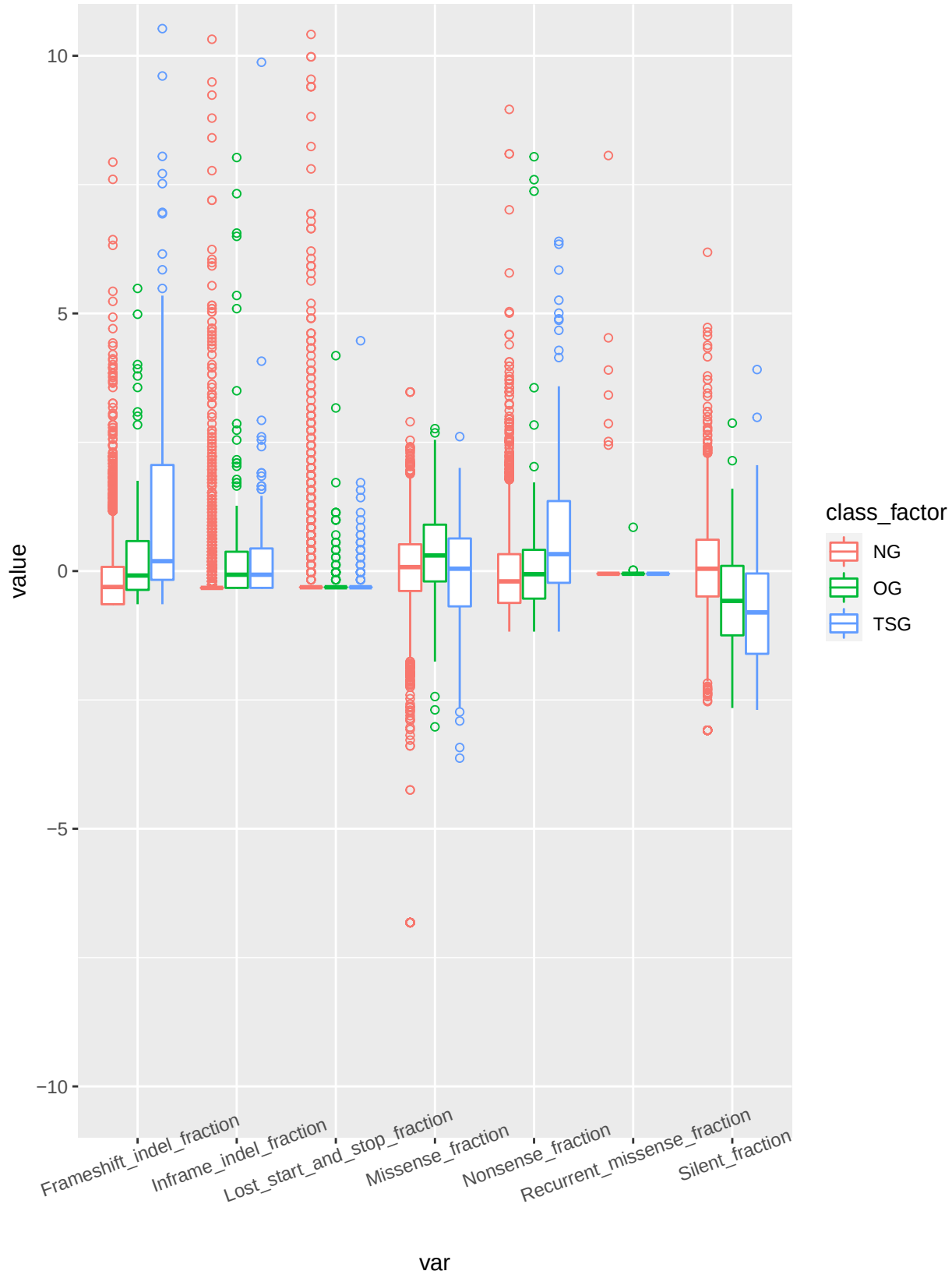

10
5
0
−5
−10
value
Broad_H3K79me2_percentage
Broad_H4K20me1_percentage
H3K79me2_height
H3K79me2_width
H3K9me2_height
H4K20me1_height
H4K20me1_width
var
class_factor
NG
OG
TSG

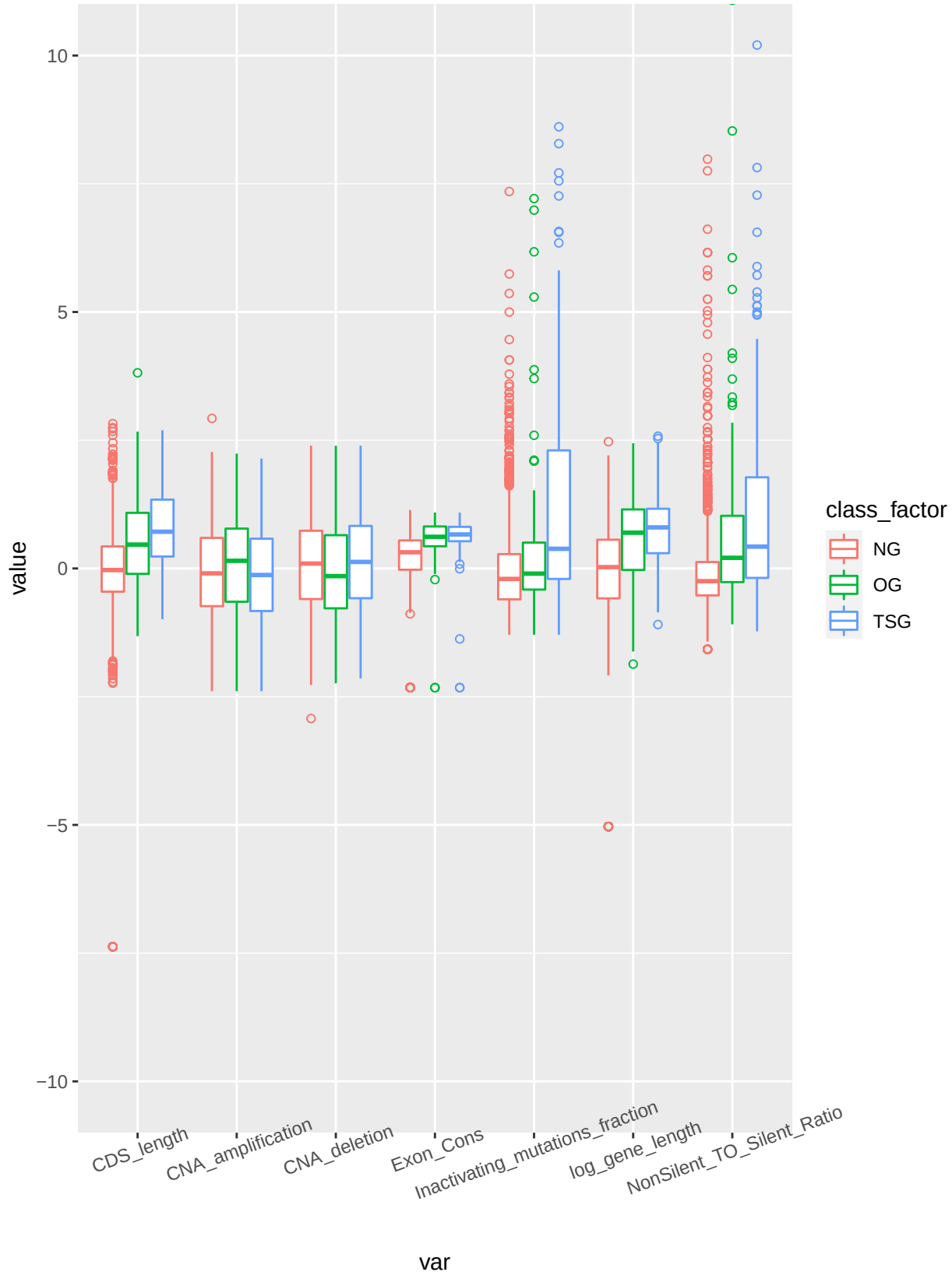
```r
# Box plots of each variable (7 per plot) split by class
for (i in seq(1, ncol(training1) - 2, by = 7)) {
  # Select 7 columns plus class
  plot <- training1[, c(names(training1)[seq(i, i + 6)], "class_factor")] %>%
    # Scale and center numeric variables
    mutate_if(is.numeric, scale) %>%
    # Turns 7 columns into 2 columns: variable name and value
    gather(-class_factor, key = "var", value = "value") %>%
    ggplot(aes(x = var, y = value, color = class_factor)) +
      geom_boxplot(outlier.shape = 1) +
      theme(axis.text.x = element_text(angle = 20)) +
      coord_cartesian(ylim = c(-10, 10))

  print(plot)
}
```

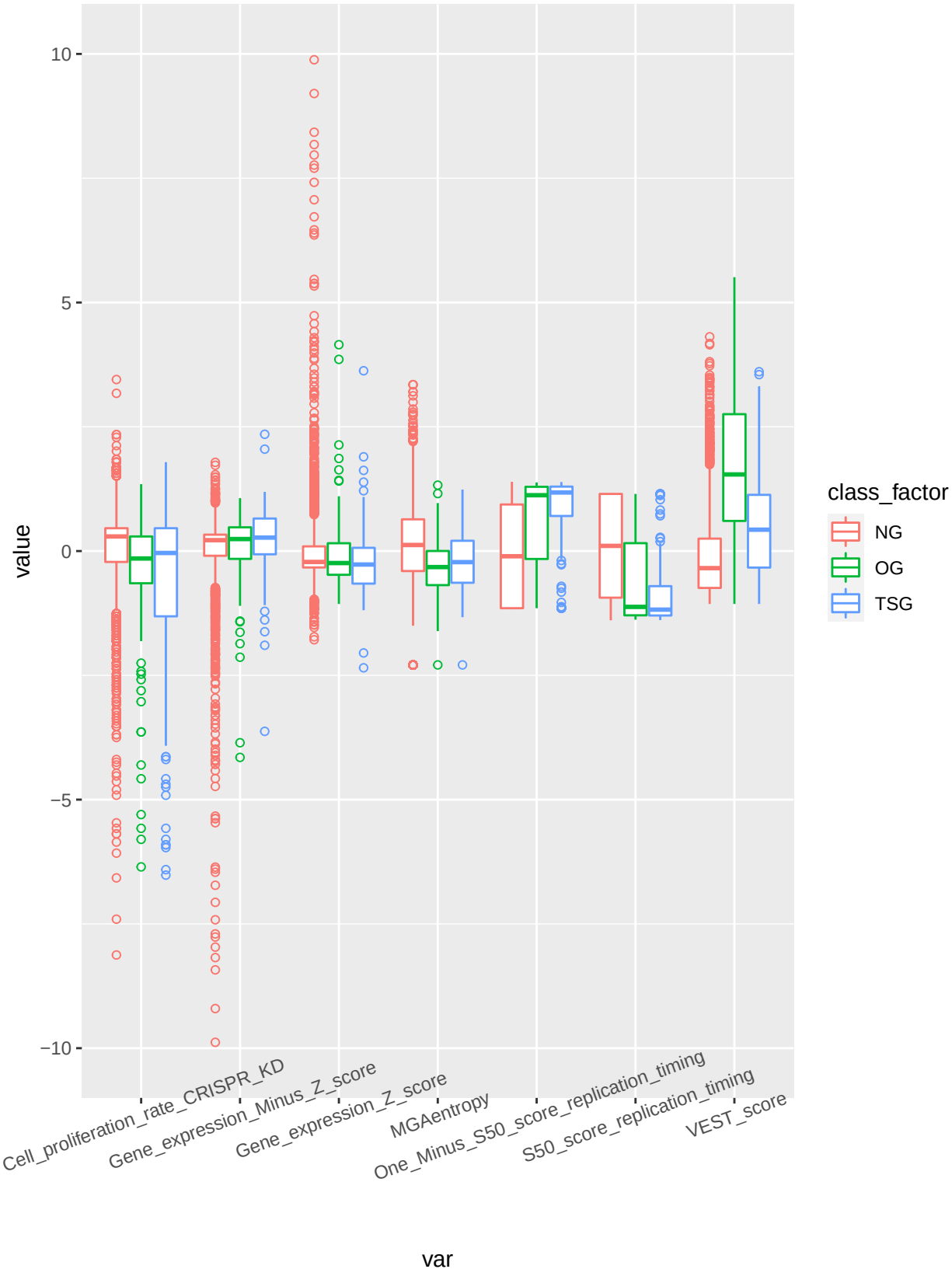

10
5
0
−5
−10
value
id
LOF_KB_Ratio
Missense_KB_Ratio
N_LOF
N_Missense
N_Splice
Silent_KB_Ratio
var
class_factor
NG
OG
TSG

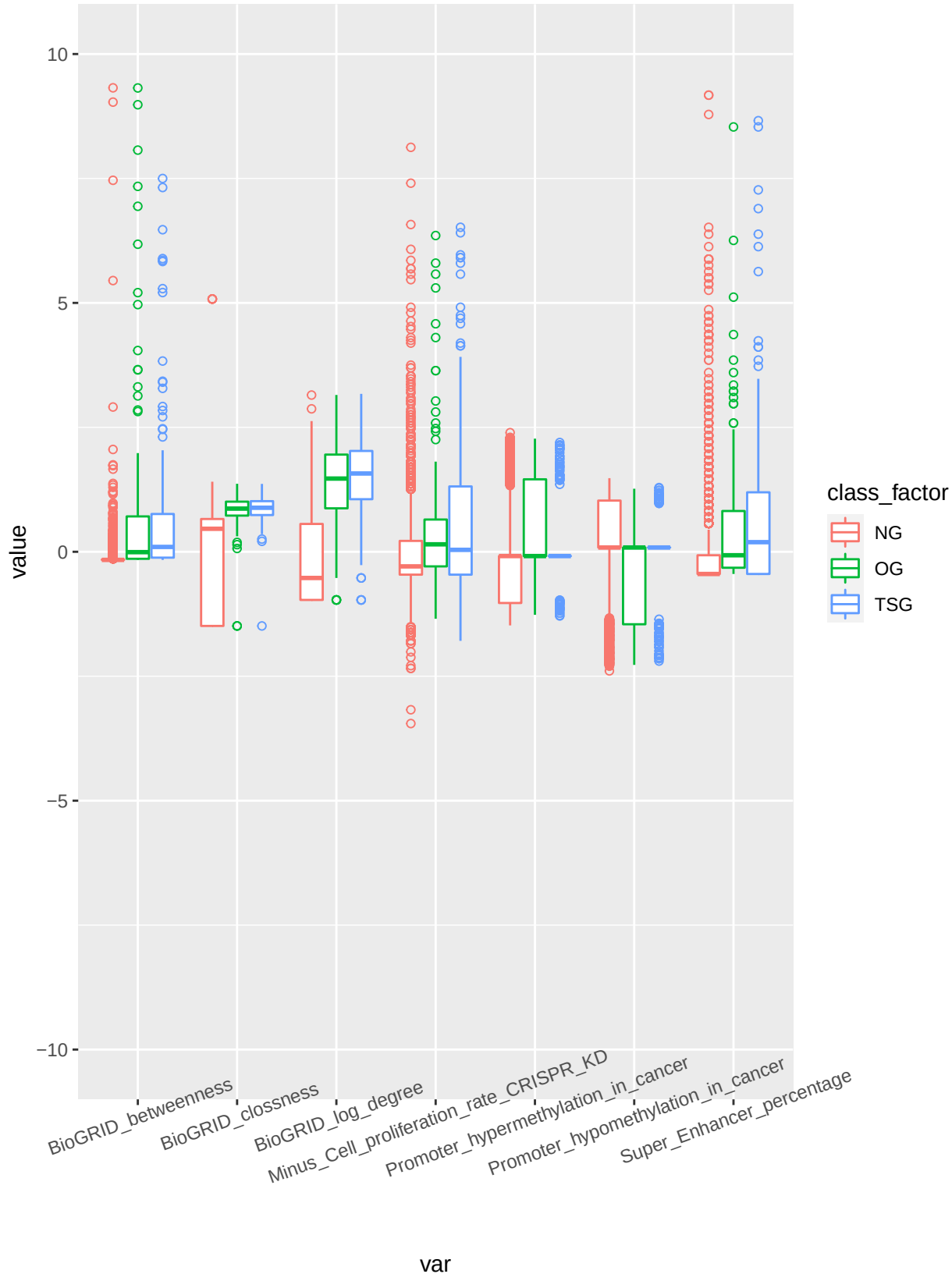

value
10
5
0
−5
−10
LOF_TO_Benign_Ratio
LOF_TO_Silent_Ratio
Missense_Damaging_TO_Missense_Benign_Ratio
Missense_Entropy
Missense_TO_Silent_Ratio
Splice_TO_Benign_Ratio
Splice_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

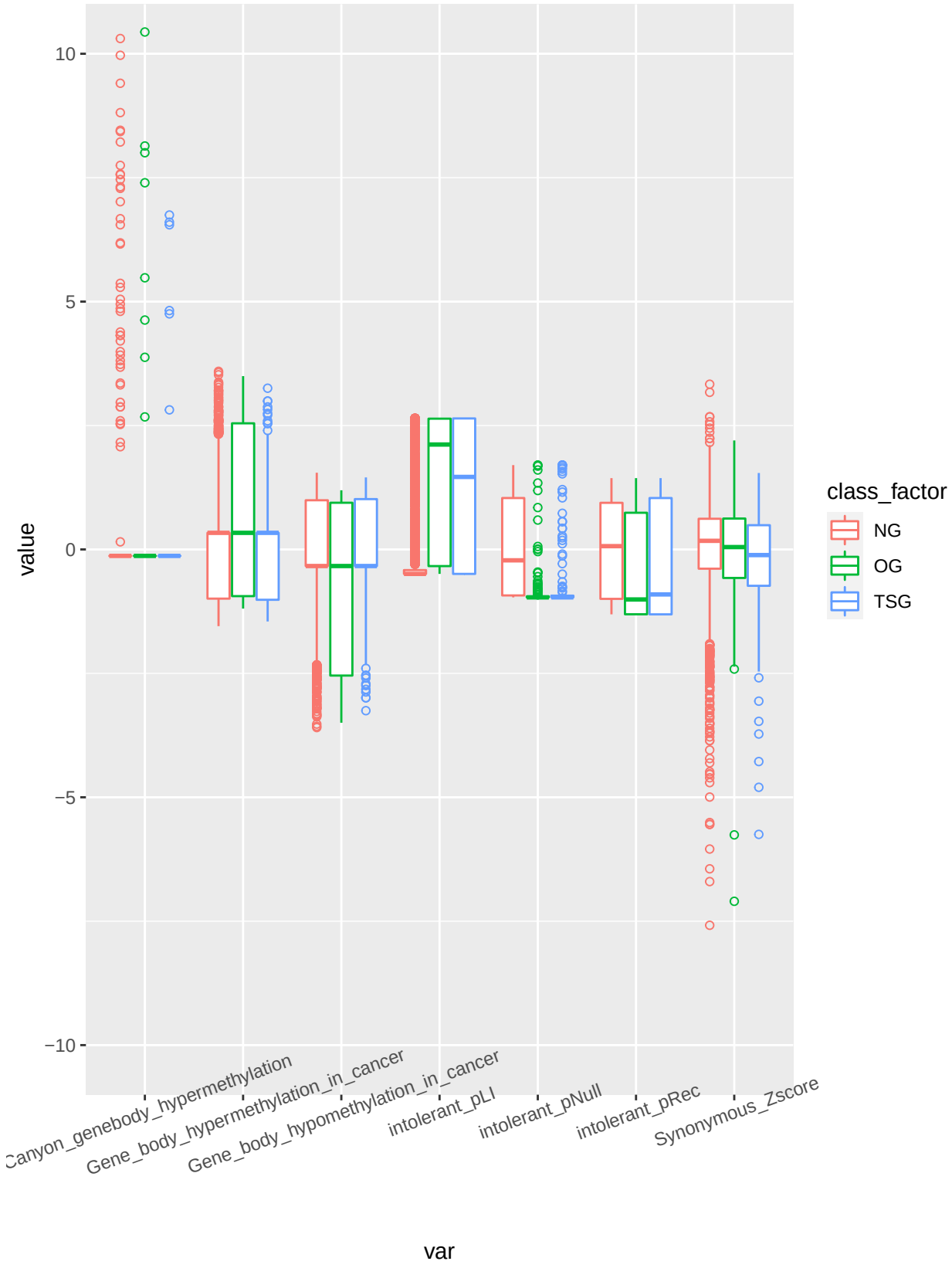

10
5
0
−5
−10
value
LOF_TO_Missense_Ratio
LOF_TO_Total_Ratio
Missense_Damaging_TO_Benign_Ratio
Missense_TO_Benign_Ratio
Missense_TO_Total_Ratio
Polyphen2
Splice_TO_Total_Ratio
var
class_factor
NG
OG
TSG

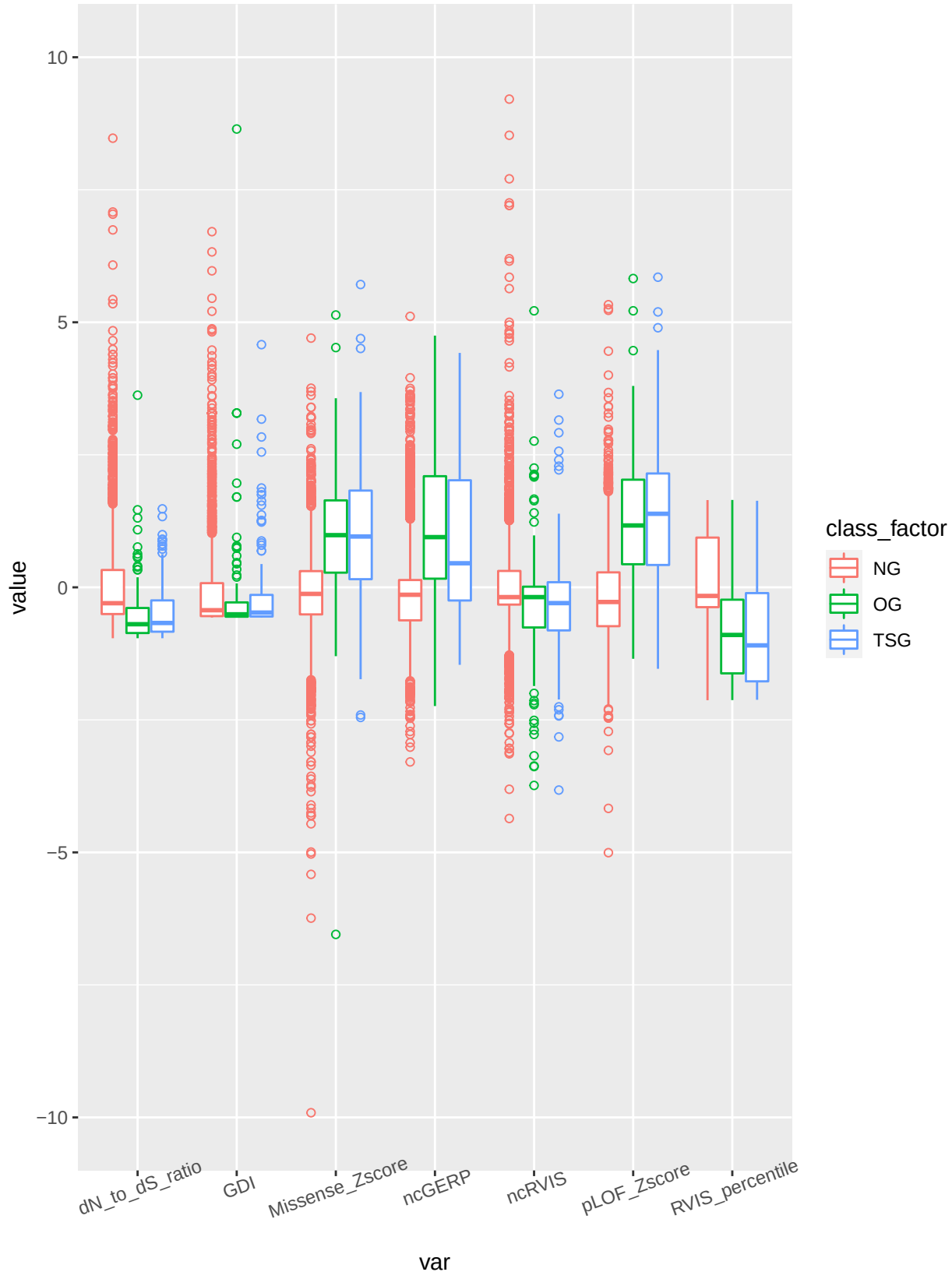

10
5
0
−5
−10
value
Frameshift_indel_fraction
Inframe_indel_fraction
Lost_start_and_stop_fraction
Missense_fraction
Nonsense_fraction
Recurrent_missense_fraction
Silent_fraction
var
class_factor
NG
OG
TSG

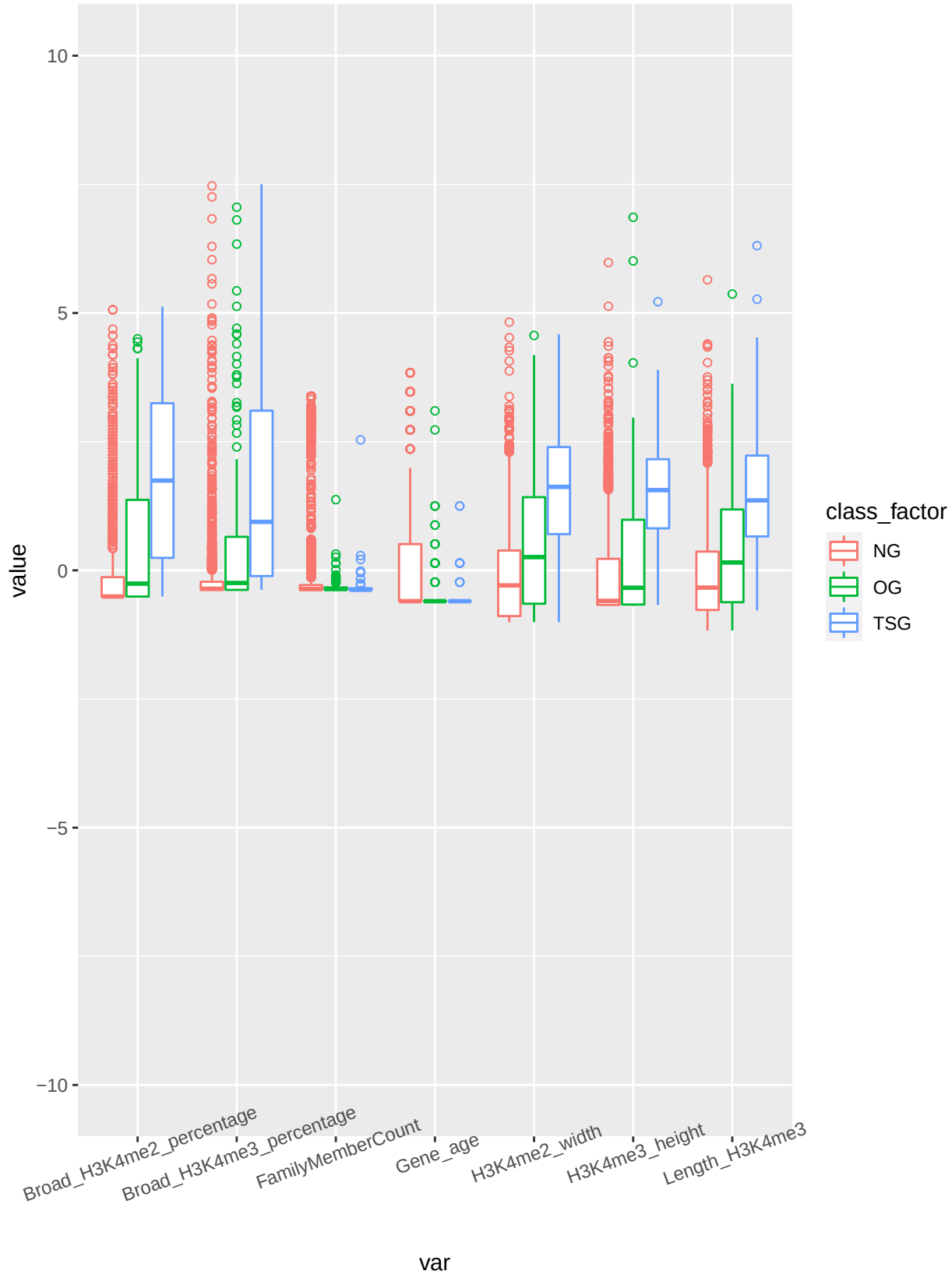

10
5
0
−5
−10
value
CDS_length
CNA_amplification
CNA_deletion
Exon_Cons
Inactivating_mutations_fraction
log_gene_length
NonSilent_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

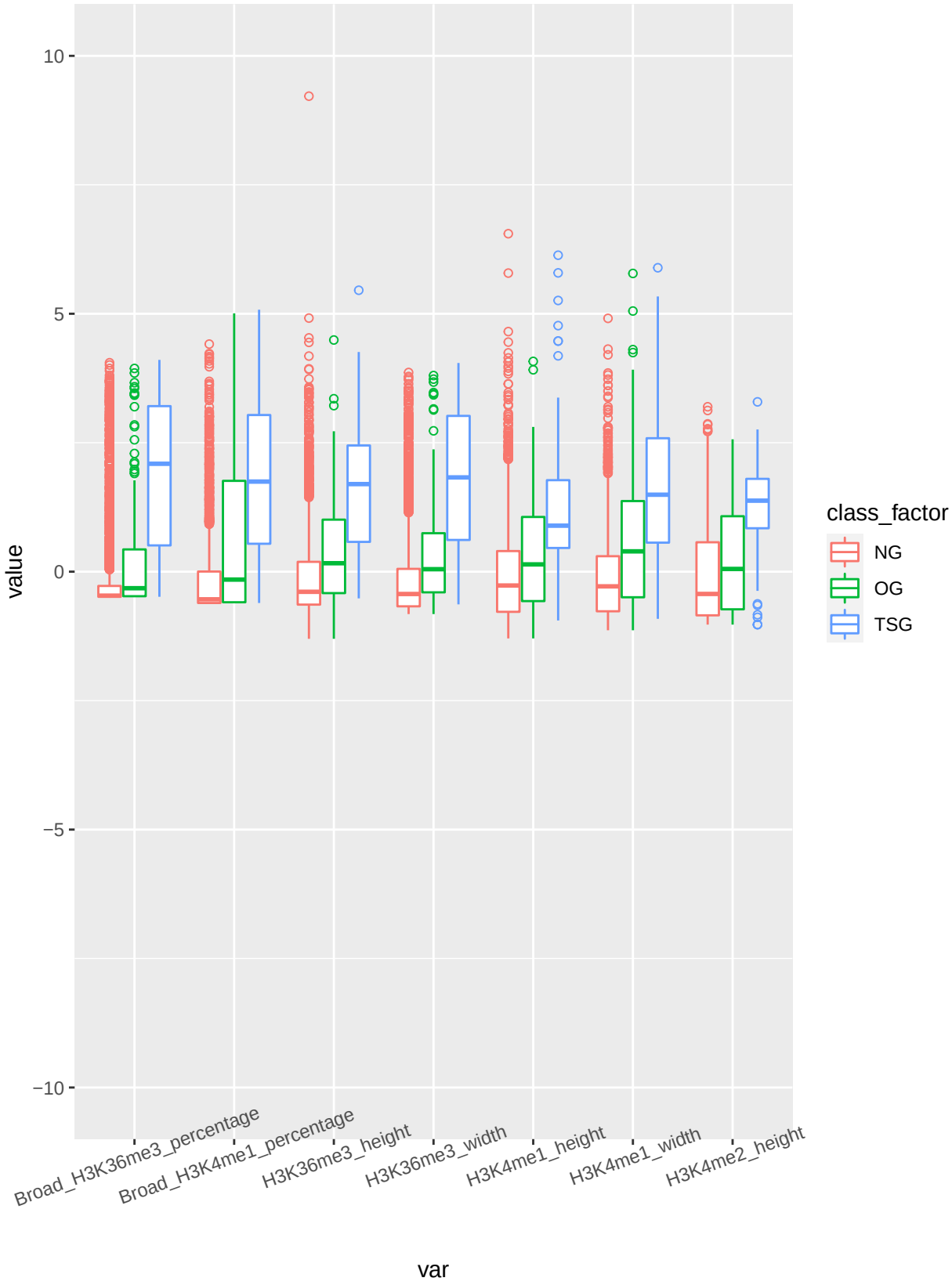

10
5
0
−5
−10
value
Cell_proliferation_rate_CRISPR_KD
Gene_expression_Minus_Z_score
Gene_expression_Z_score
MGAentropy
One_Minus_S50_score_replication_timing
S50_score_replication_timing
VEST_score
var
class_factor
NG
OG
TSG

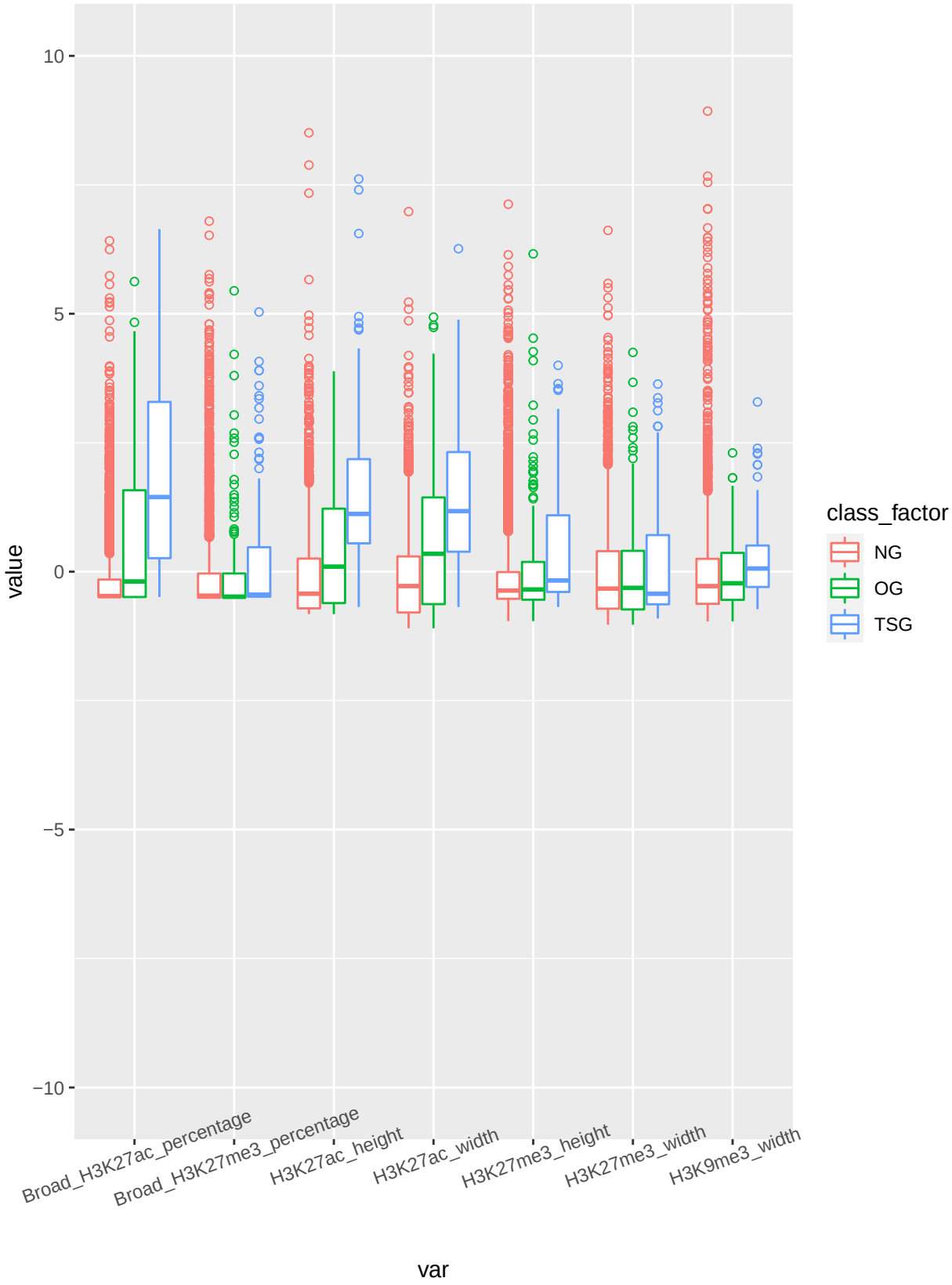

10
5
0
−5
−10
value
BioGRID_betweenness
BioGRID_clossness
BioGRID_log_degree
Minus_Cell_proliferation_rate_CRISPR_KD
Promoter_hypermethylation_in_cancer
Promoter_hypomethylation_in_cancer
Super_Enhancer_percentage
var
class_factor
NG
OG
TSG

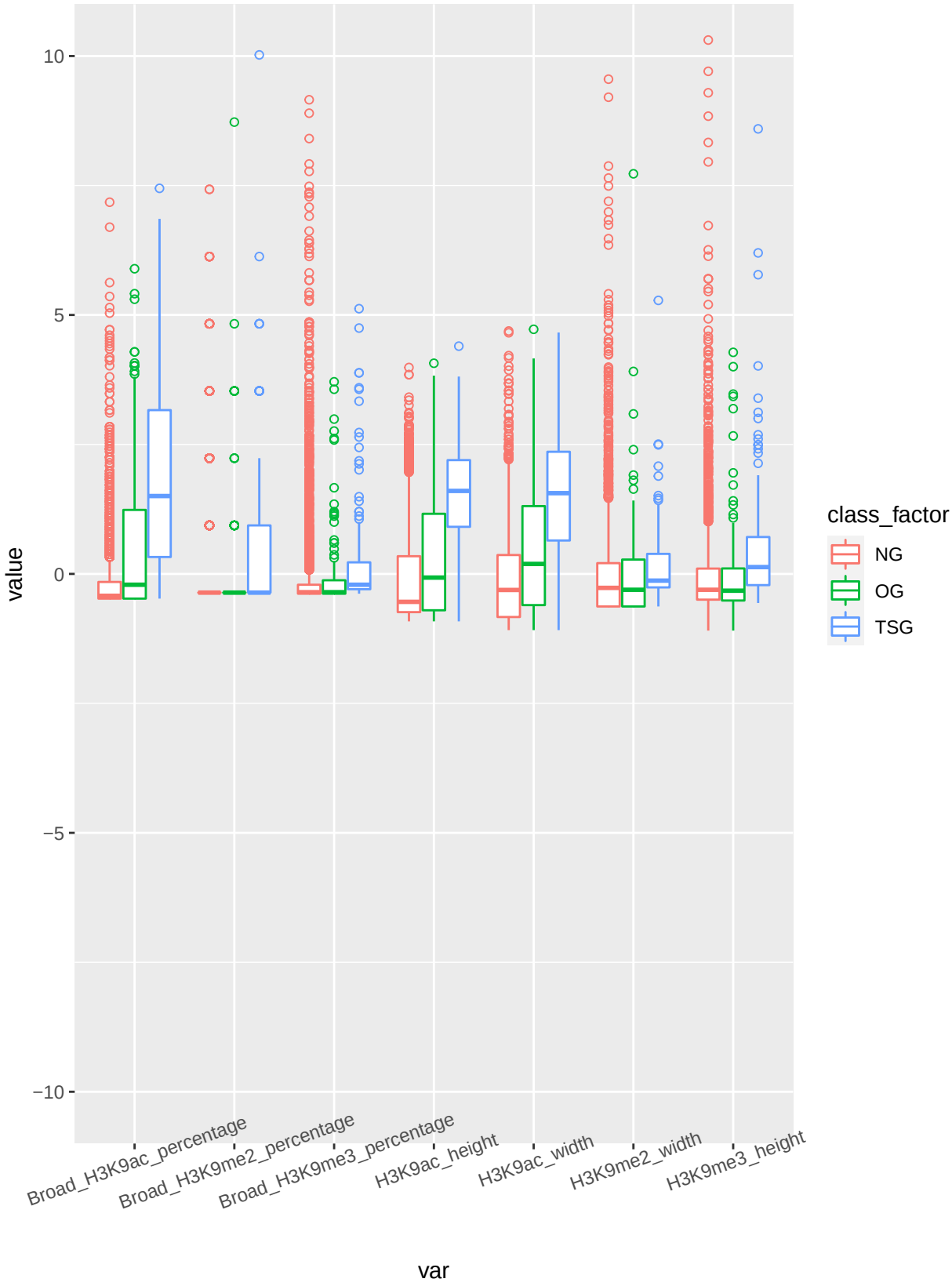

10
5
0
−5
−10
value
Canyon_genebody_hypermethylation
Gene_body_hypermethylation_in_cancer
Gene_body_hypomethylation_in_cancer
intolerant_pLI
intolerant_pNull
intolerant_pRec
Synonymous_Zscore
var
class_factor
NG
OG
TSG

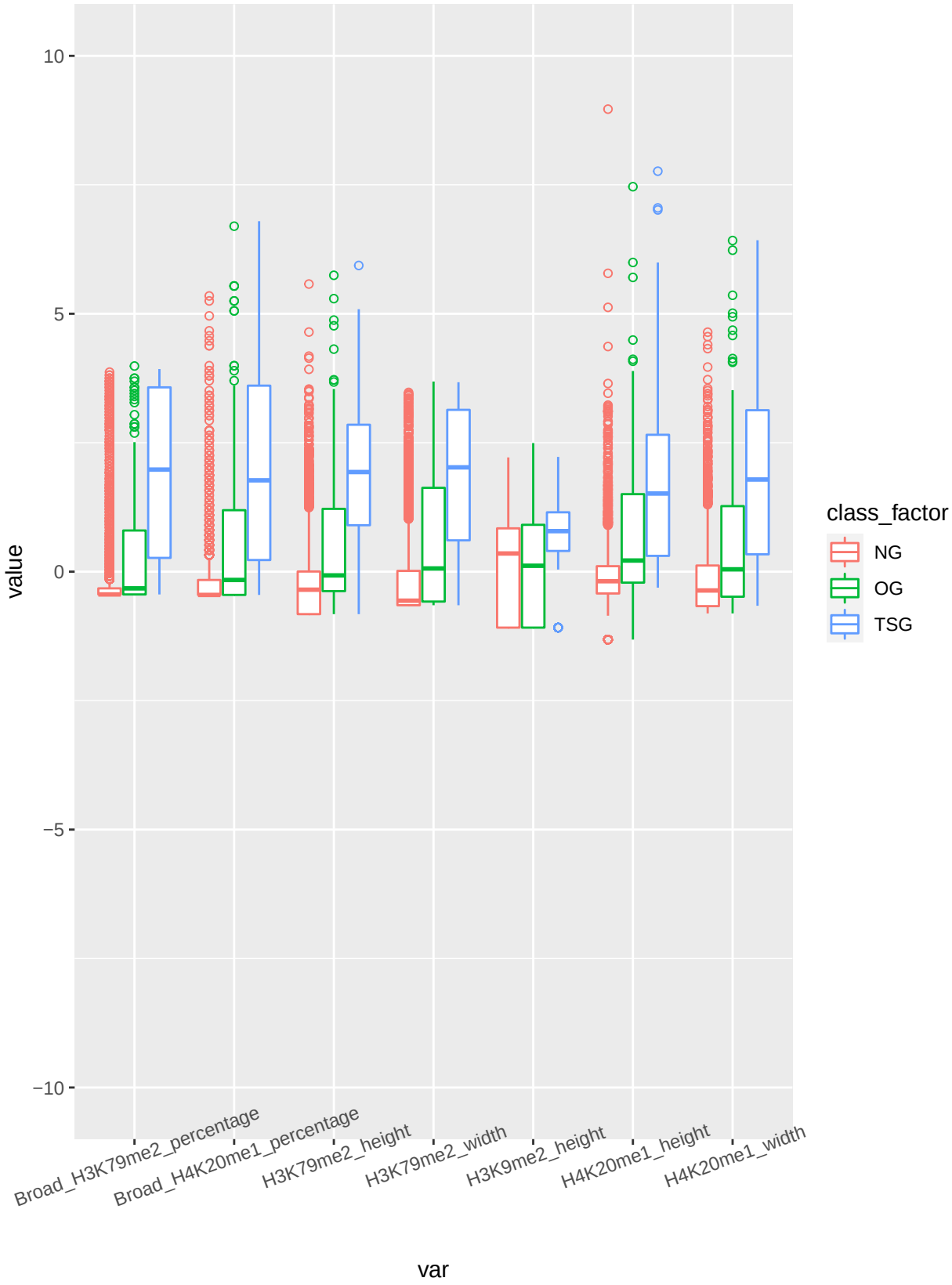

10
5
0
−5
−10
value
dN_to_dS_ratio
GDI
Missense_Zscore
ncGERP
ncRVIS
pLOF_Zscore
RVIS_percentile
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K4me2_percentage
Broad_H3K4me3_percentage
FamilyMemberCount
Gene_age
H3K4me2_width
H3K4me3_height
Length_H3K4me3
var
class_factor
NG
OG
TSG

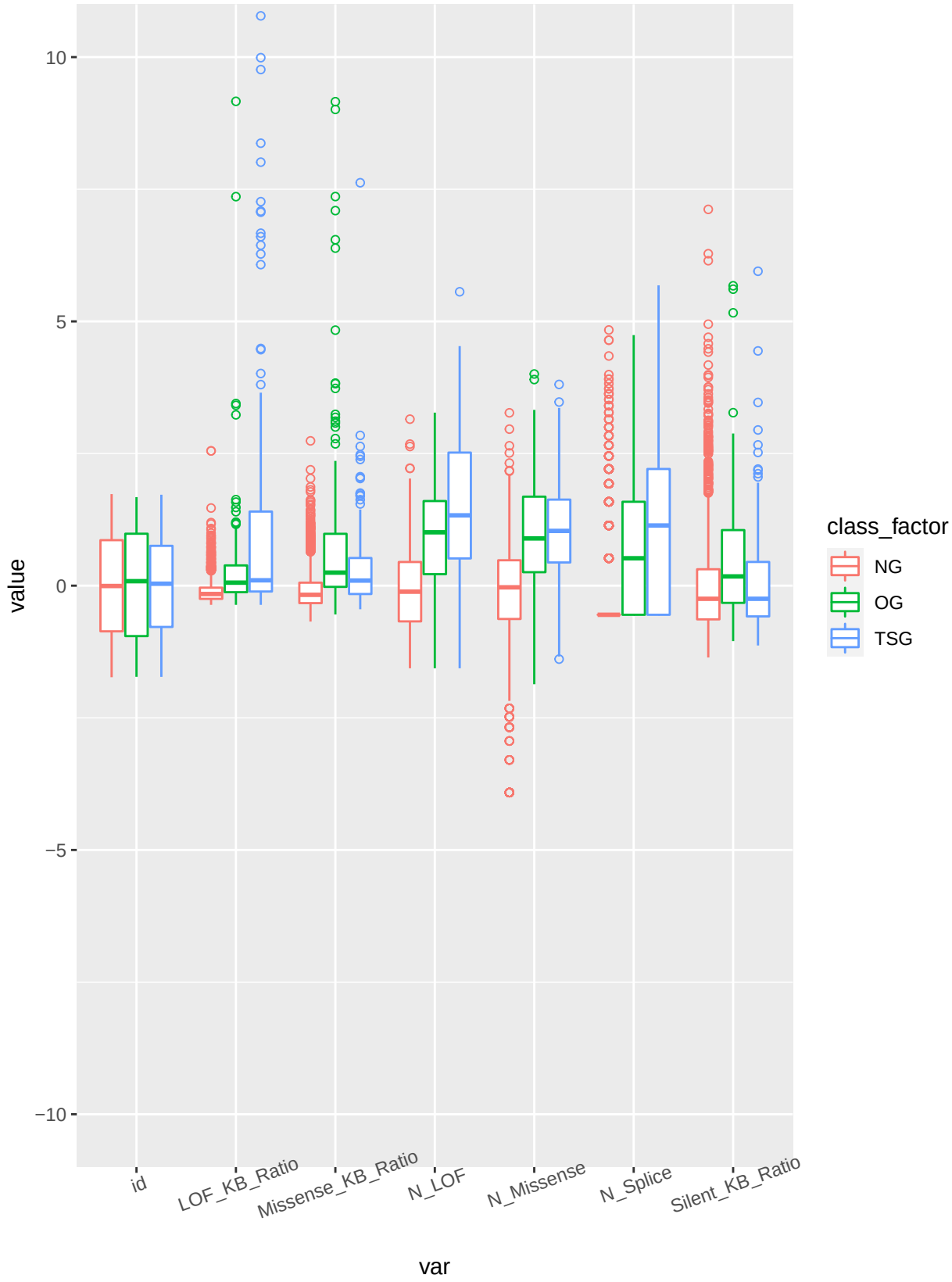

10
5
0
−5
−10
value
Broad_H3K36me3_percentage
Broad_H3K4me1_percentage
H3K36me3_height
H3K36me3_width
H3K4me1_height
H3K4me1_width
H3K4me2_height
var
class_factor
NG
OG
TSG

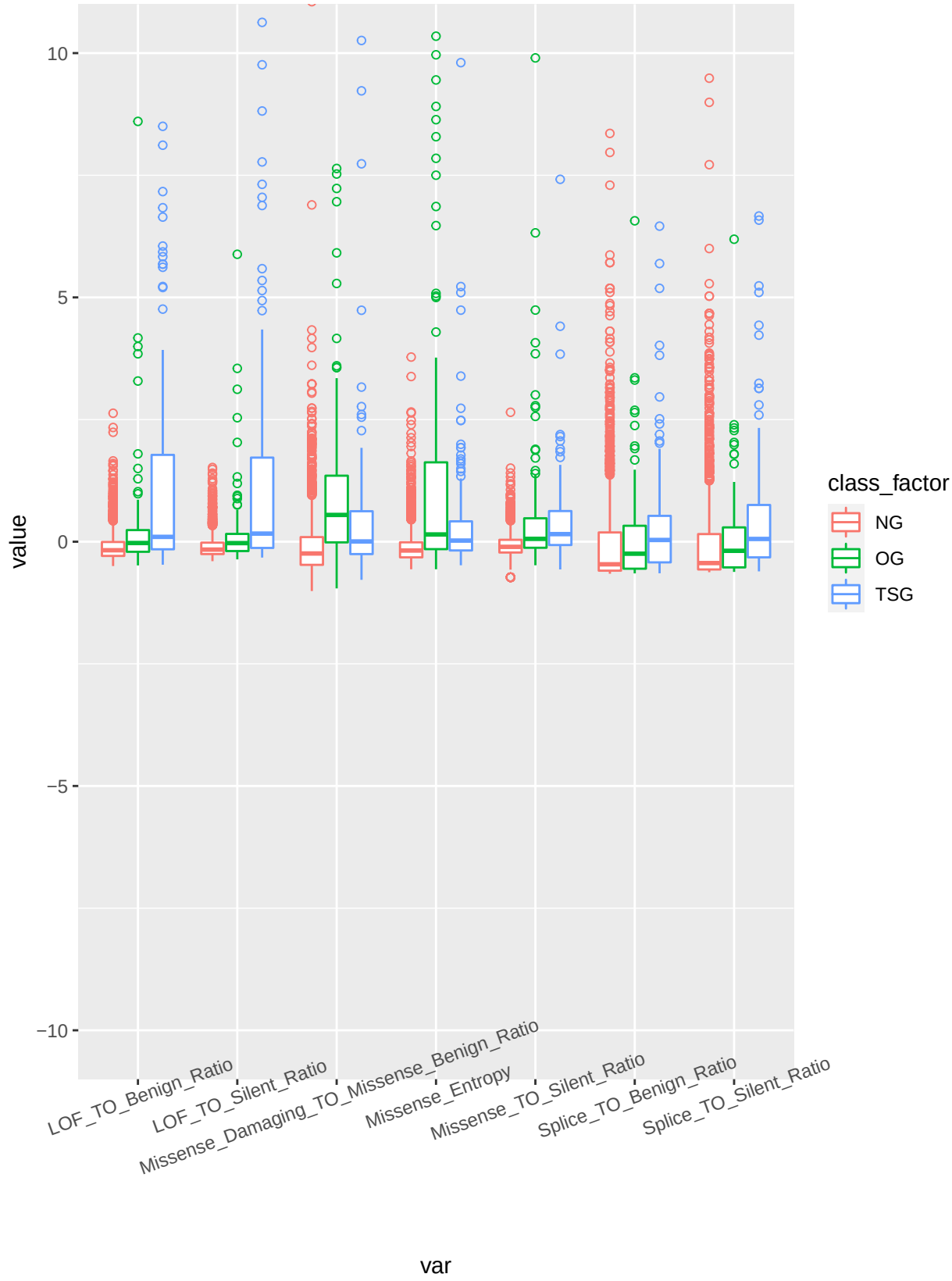

10
5
0
−5
−10
value
Broad_H3K27ac_percentage
Broad_H3K27me3_percentage
H3K27ac_height
H3K27ac_width
H3K27me3_height
H3K27me3_width
H3K9me3_width
var
class_factor
NG
OG
TSG

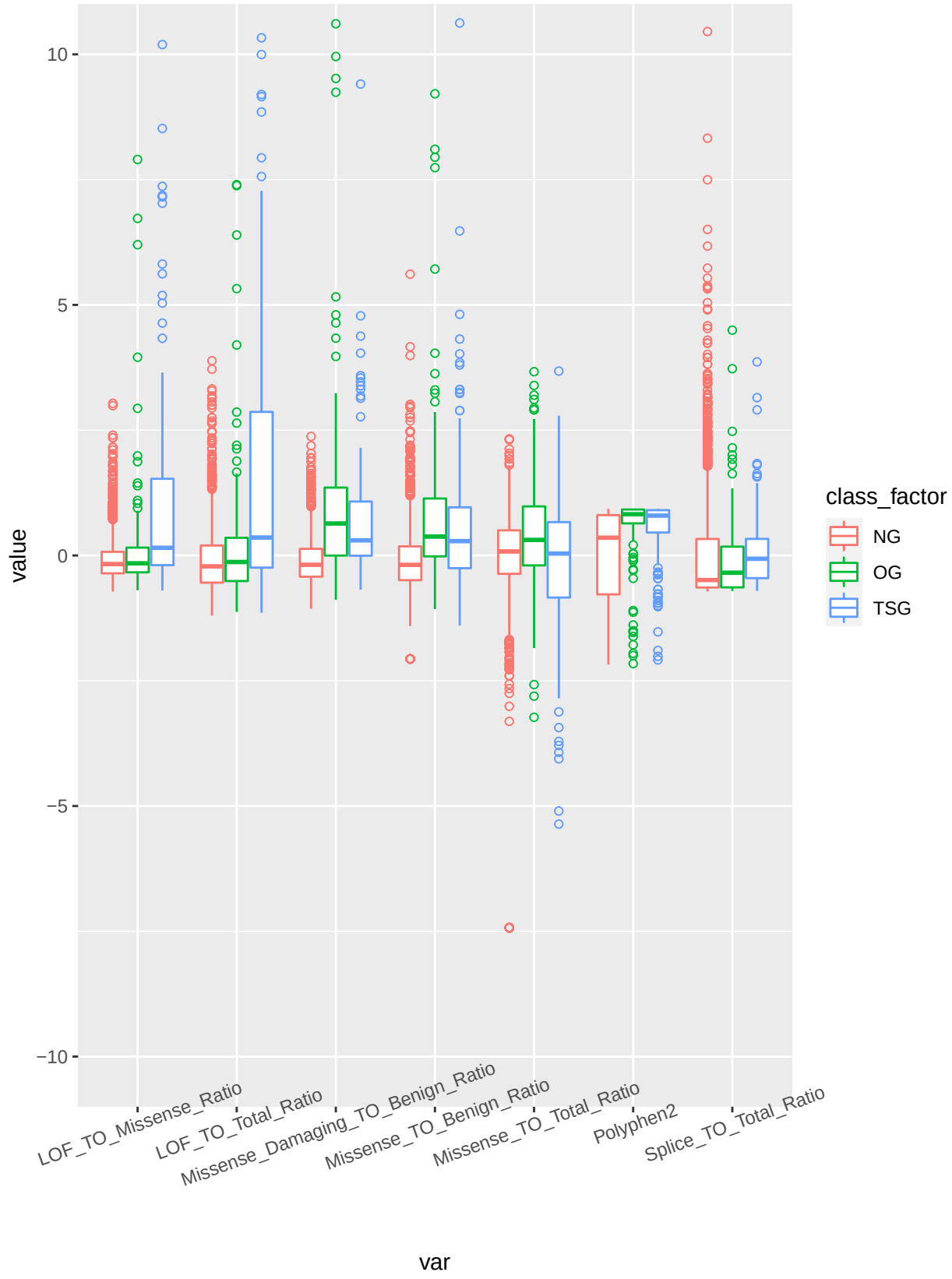

10
5
0
−5
−10
value
Broad_H3K9ac_percentage
Broad_H3K9me2_percentage
Broad_H3K9me3_percentage
H3K9ac_height
H3K9ac_width
H3K9me2_width
H3K9me3_height
var
class_factor
NG
OG
TSG

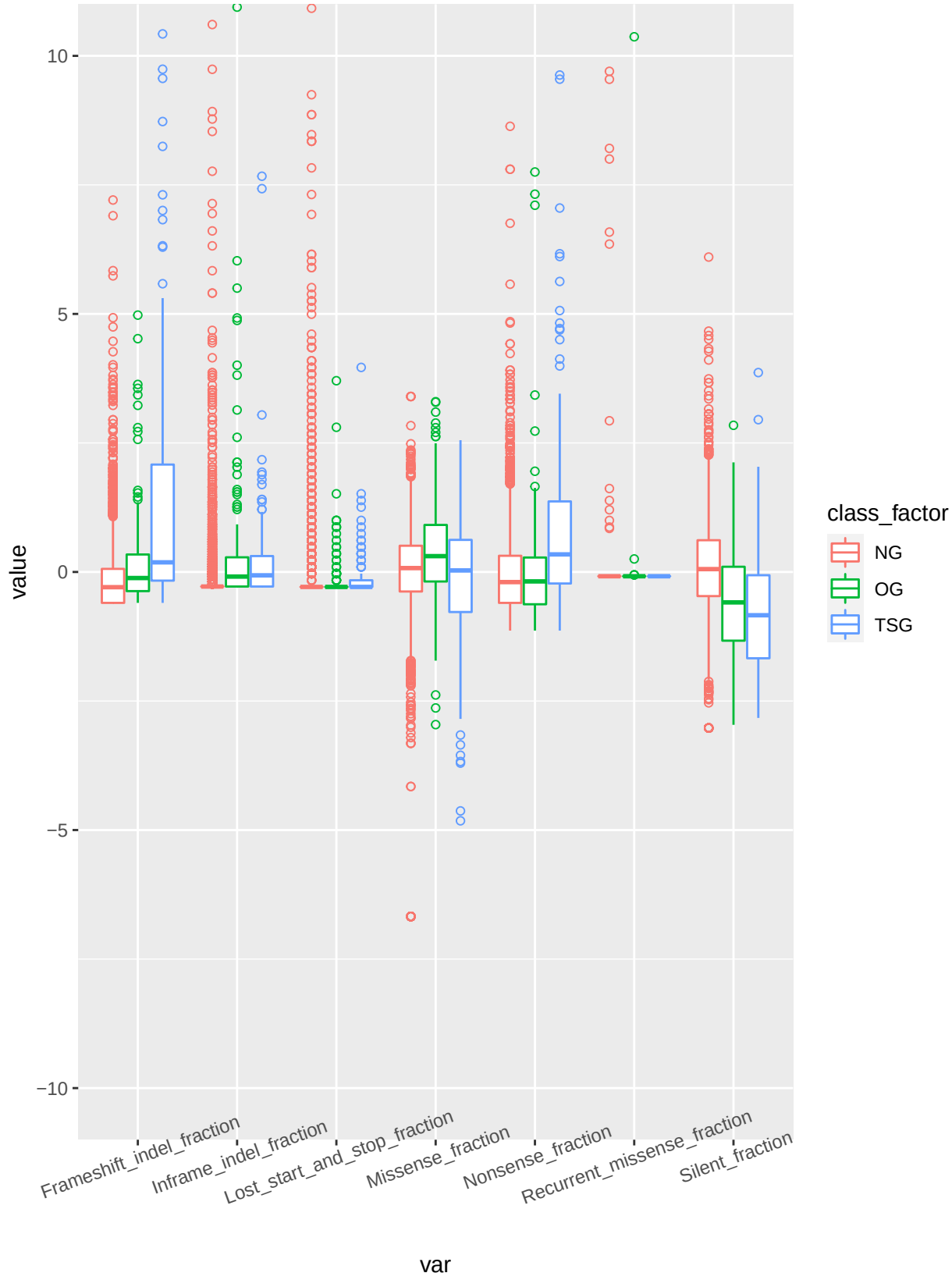

10
5
0
−5
−10
value
Broad_H3K79me2_percentage
Broad_H4K20me1_percentage
H3K79me2_height
H3K79me2_width
H3K9me2_height
H4K20me1_height
H4K20me1_width
var
class_factor
NG
OG
TSG

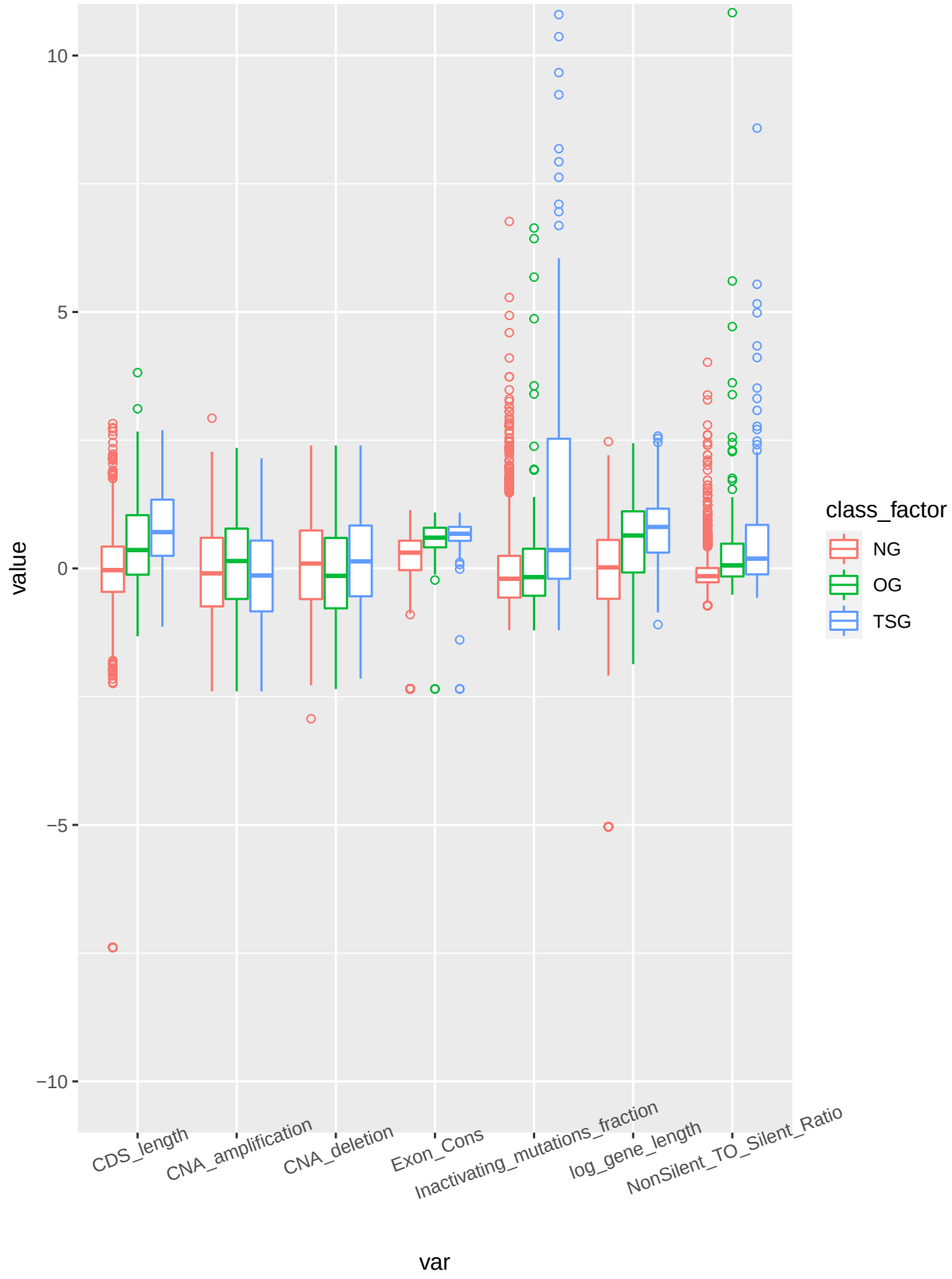
```r
# Box plots of each variable (7 per plot) split by class
for (i in seq(1, ncol(training2) - 2, by = 7)) {
  # Select 7 columns plus class
  plot <- training2[, c(names(training2)[seq(i, i + 6)], "class_factor")] %>%
    # Scale and center numeric variables
    mutate_if(is.numeric, scale) %>%
    # Turns 7 columns into 2 columns: variable name and value
    gather(-class_factor, key = "var", value = "value") %>%
    ggplot(aes(x = var, y = value, color = class_factor)) +
      geom_boxplot(outlier.shape = 1) +
      theme(axis.text.x = element_text(angle = 20)) +
      coord_cartesian(ylim = c(-10, 10))

  print(plot)
}
```

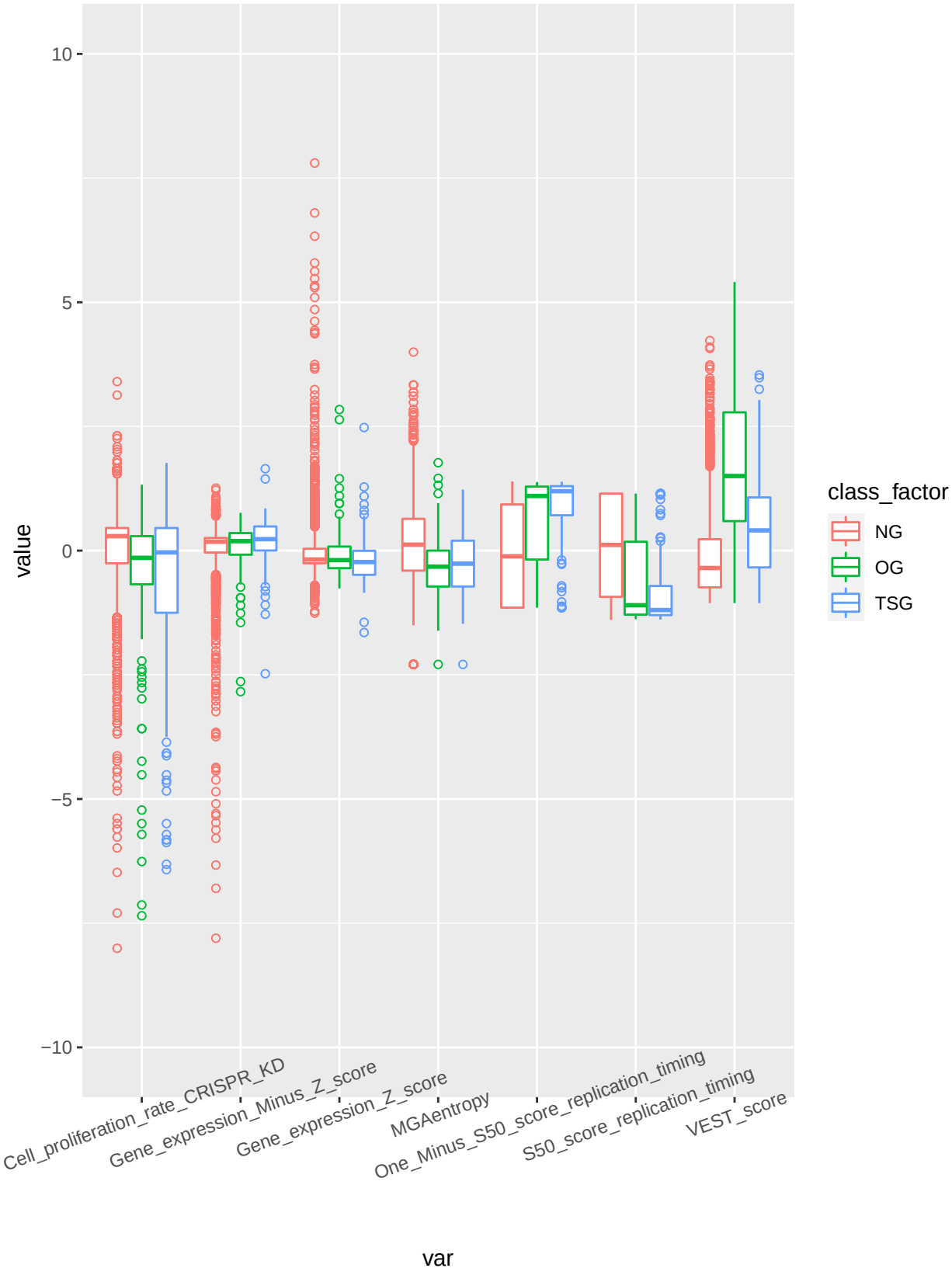

10
5
0
−5
−10
value
id
LOF_KB_Ratio
Missense_KB_Ratio
N_LOF
N_Missense
N_Splice
Silent_KB_Ratio
var
class_factor
NG
OG
TSG

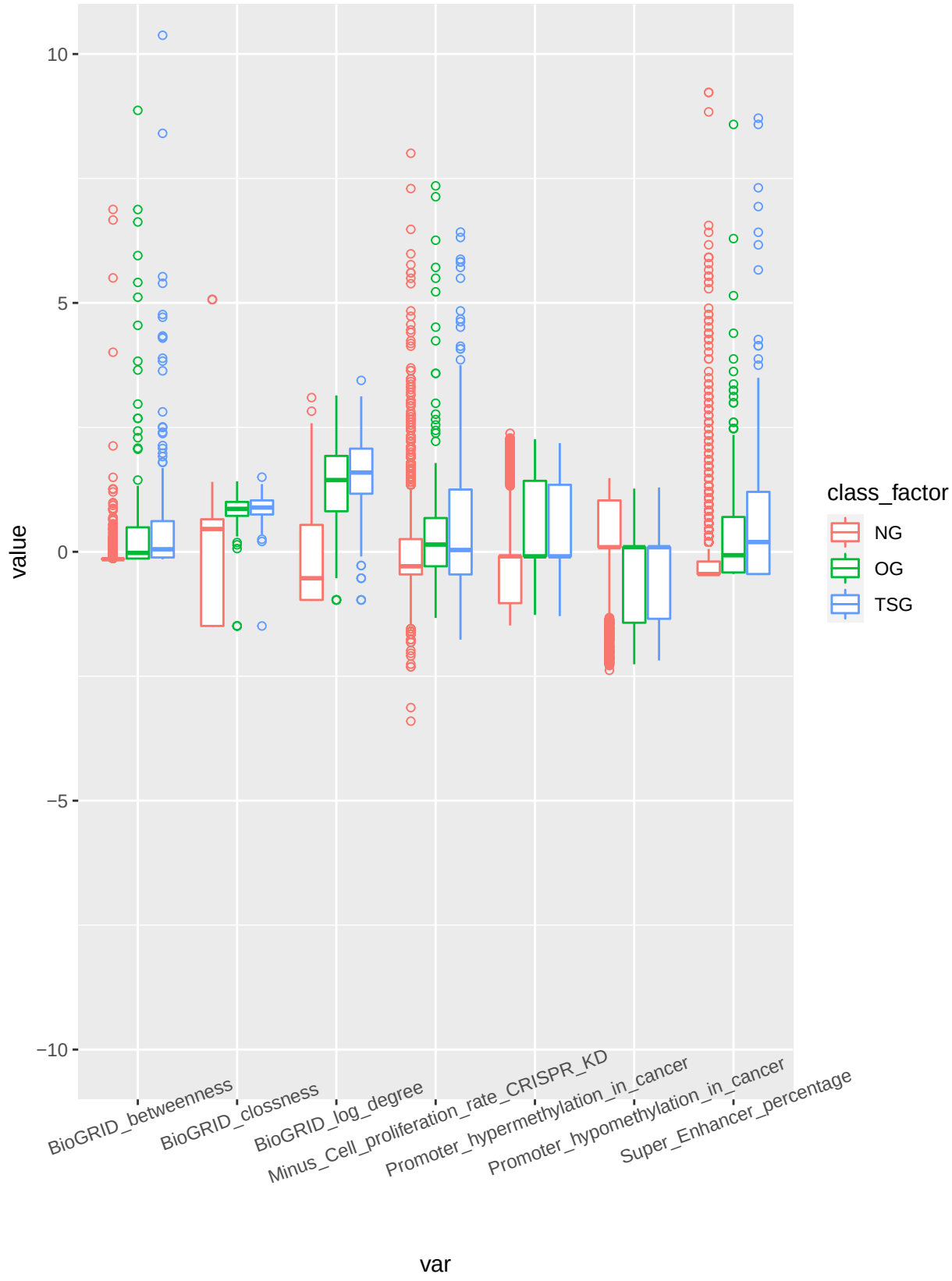

10
5
0
−5
−10
value
LOF_TO_Benign_Ratio
LOF_TO_Silent_Ratio
Missense_Damaging_TO_Missense_Benign_Ratio
Missense_Entropy
Missense_TO_Silent_Ratio
Splice_TO_Benign_Ratio
Splice_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

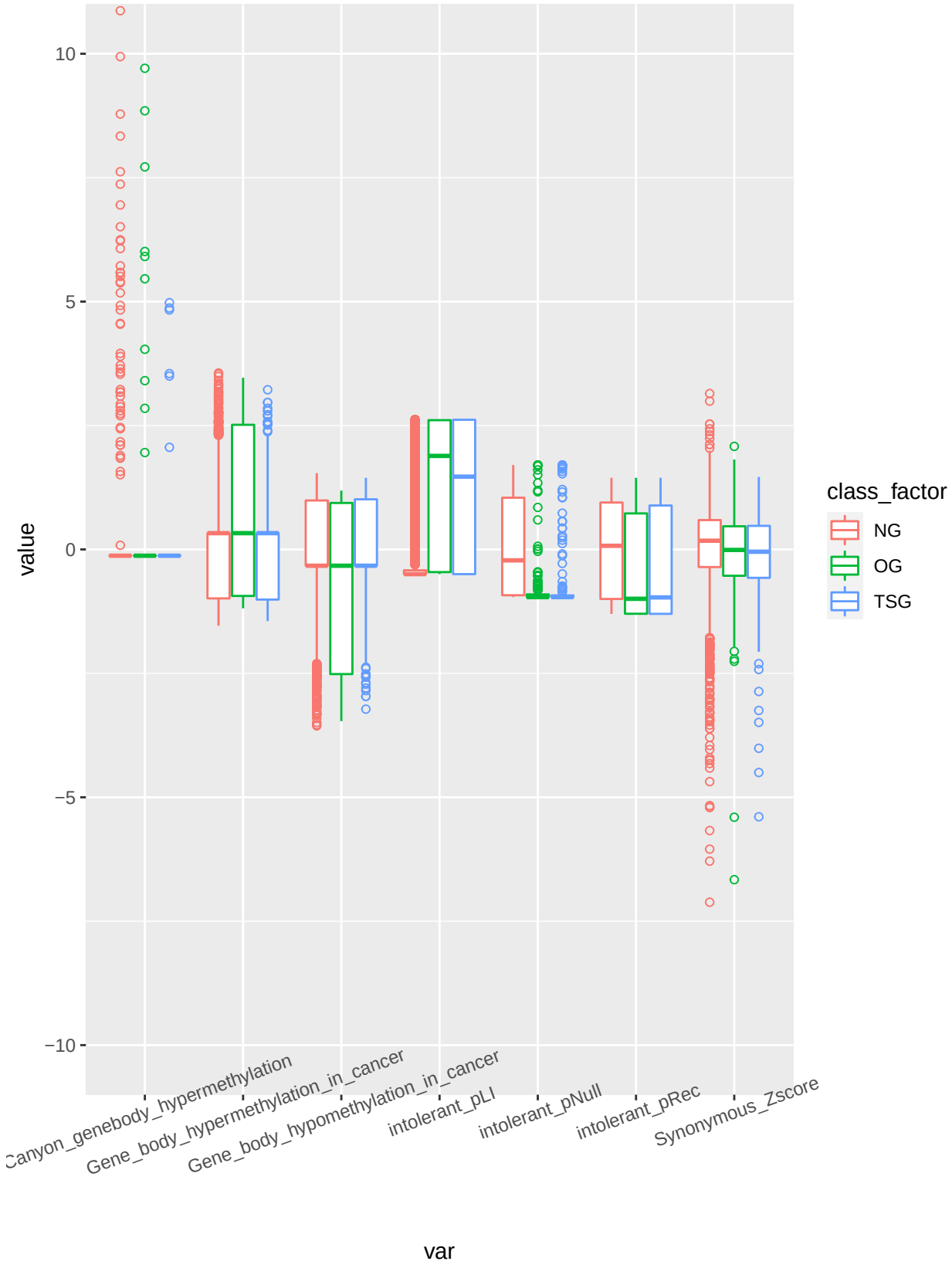

10
5
0
−5
−10
value
LOF_TO_Missense_Ratio
LOF_TO_Total_Ratio
Missense_Damaging_TO_Benign_Ratio
Missense_TO_Benign_Ratio
Missense_TO_Total_Ratio
Polyphen2
Splice_TO_Total_Ratio
var
class_factor
NG
OG
TSG

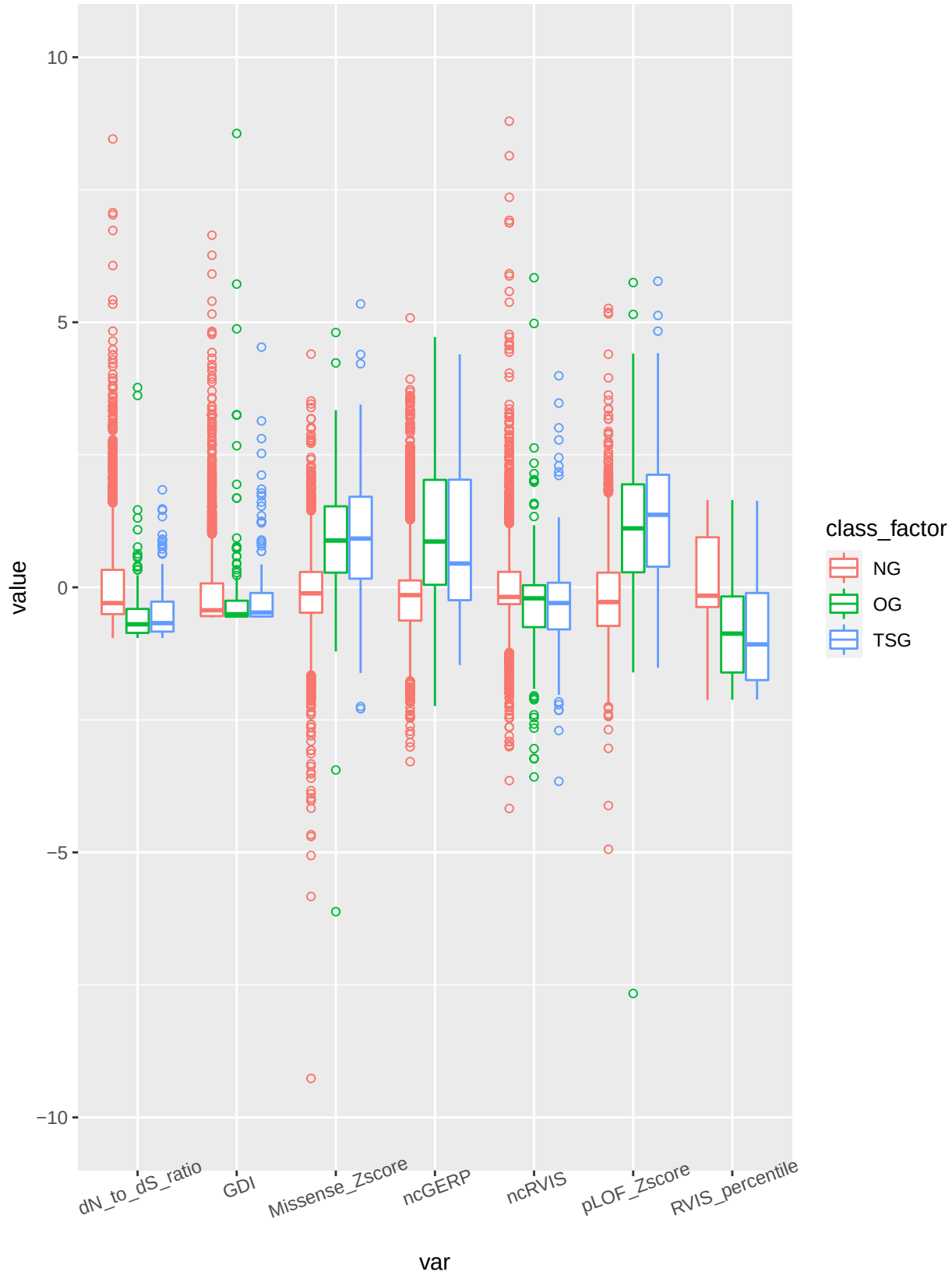

10
5
0
−5
−10
value
Frameshift_indel_fraction
Inframe_indel_fraction
Lost_start_and_stop_fraction
Missense_fraction
Nonsense_fraction
Recurrent_missense_fraction
Silent_fraction
var
class_factor
NG
OG
TSG

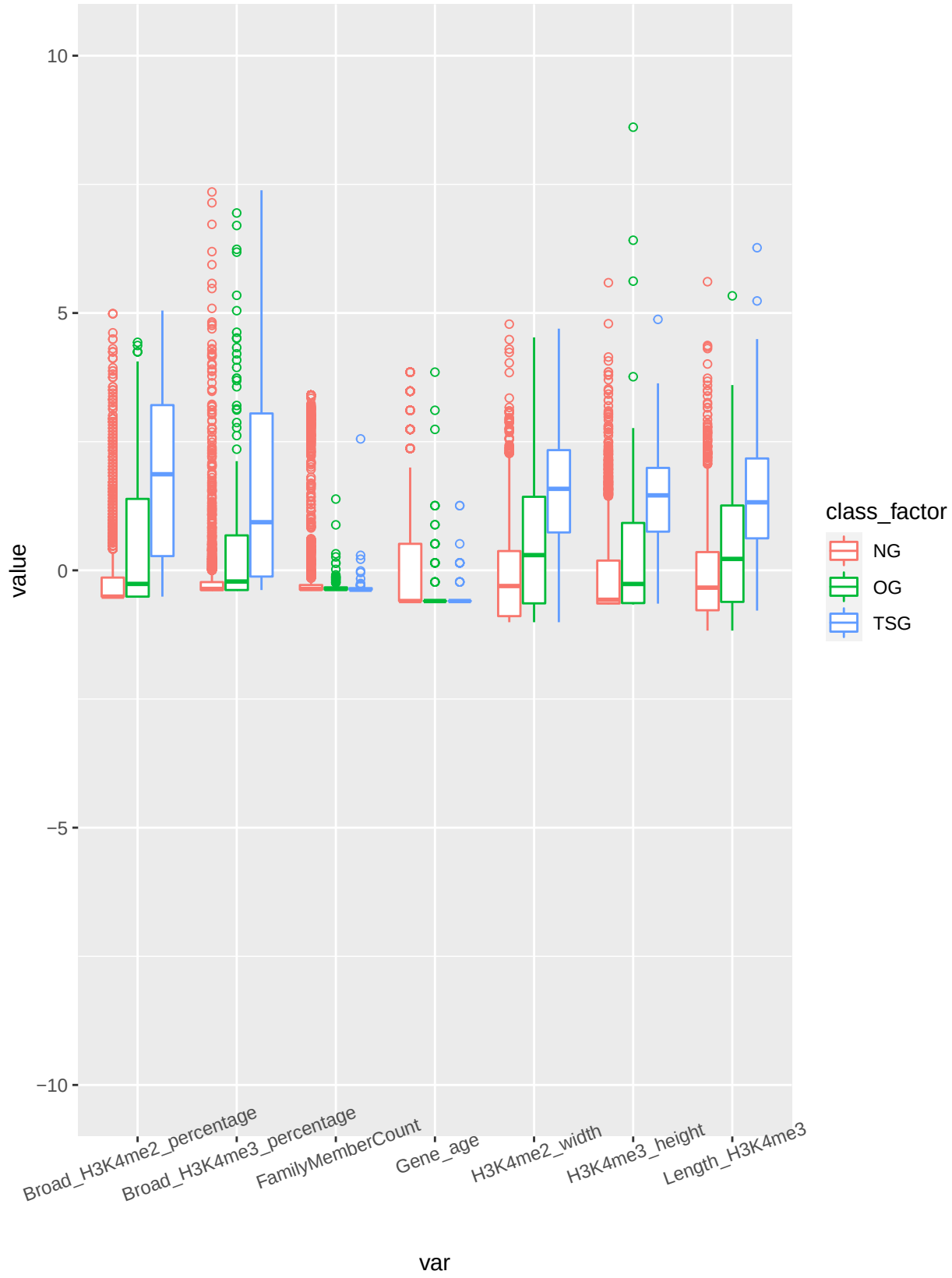

10
5
0
−5
−10
value
CDS_length
CNA_amplification
CNA_deletion
Exon_Cons
Inactivating_mutations_fraction
log_gene_length
NonSilent_TO_Silent_Ratio
var
class_factor
NG
OG
TSG

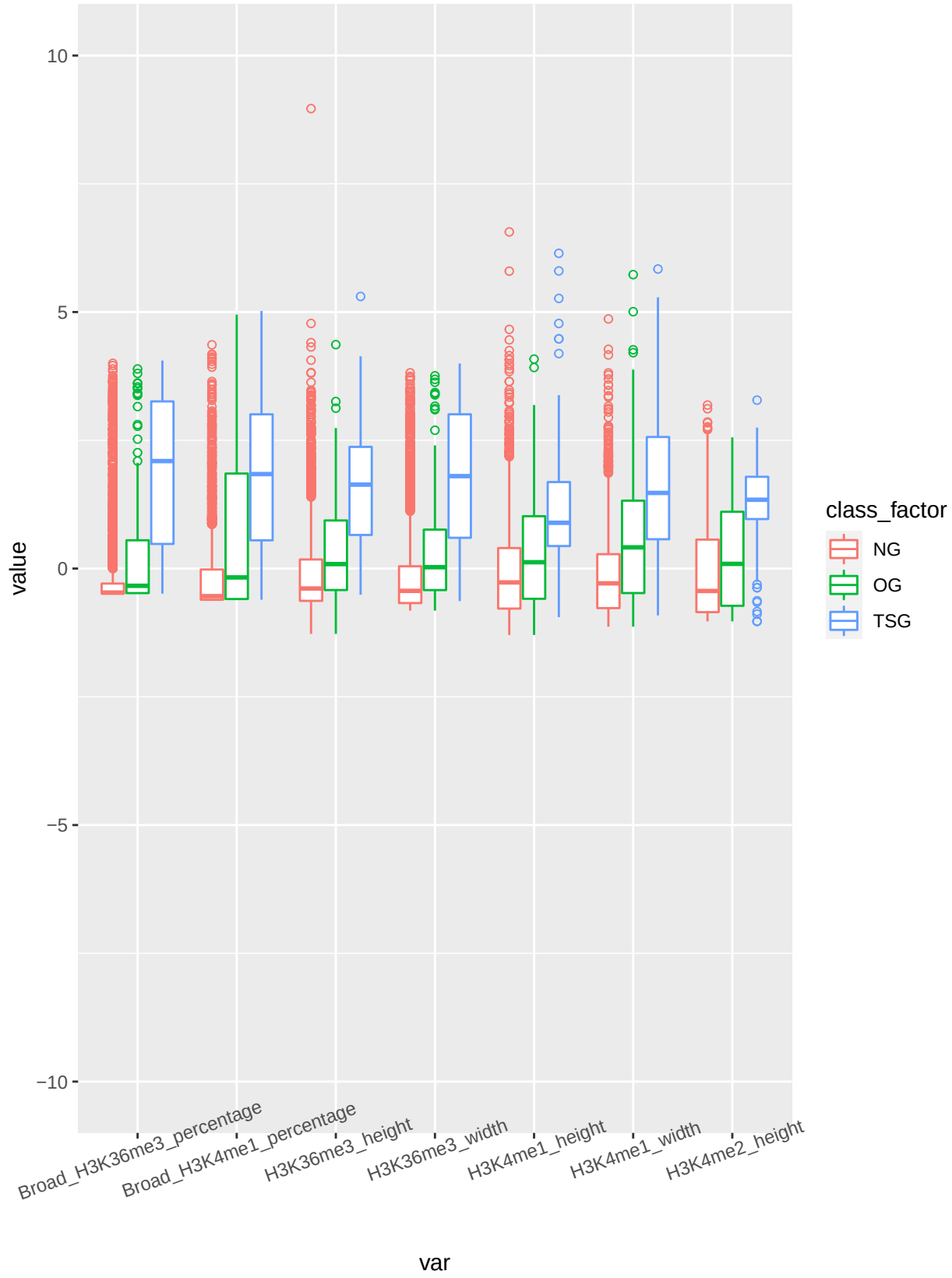

10
5
0
−5
−10
value
Cell_proliferation_rate_CRISPR_KD
Gene_expression_Minus_Z_score
Gene_expression_Z_score
MGAentropy
One_Minus_S50_score_replication_timing
S50_score_replication_timing
VEST_score
var
class_factor
NG
OG
TSG

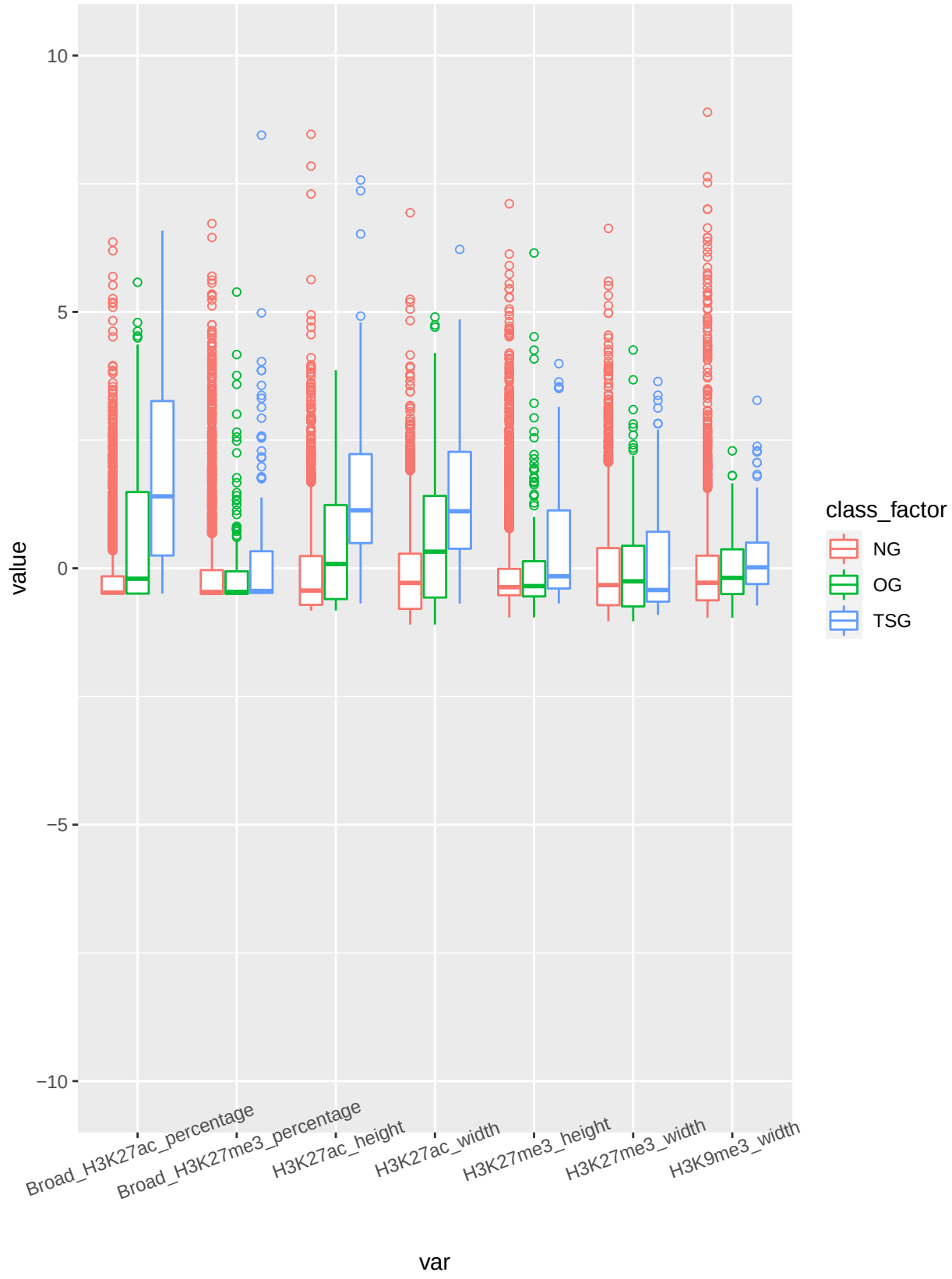

10
5
0
−5
−10
value
BioGRID_betweenness
BioGRID_clossness
BioGRID_log_degree
Minus_Cell_proliferation_rate_CRISPR_KD
Promoter_hypermethylation_in_cancer
Promoter_hypomethylation_in_cancer
Super_Enhancer_percentage
var
class_factor
NG
OG
TSG

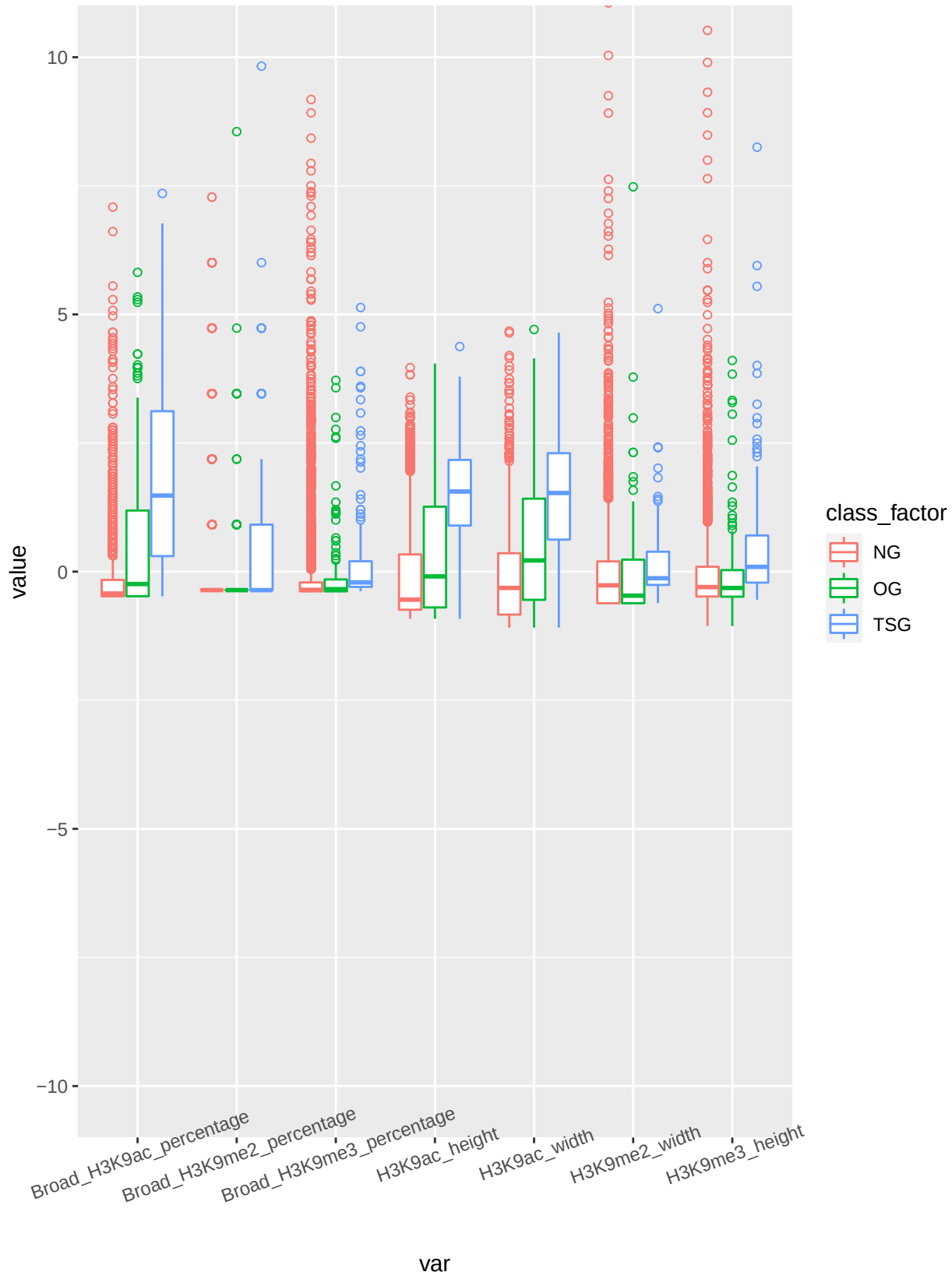

10
5
0
−5
−10
value
Canyon_genebody_hypermethylation
Gene_body_hypermethylation_in_cancer
Gene_body_hypomethylation_in_cancer
intolerant_pLI
intolerant_pNull
intolerant_pRec
Synonymous_Zscore
var
class_factor
NG
OG
TSG

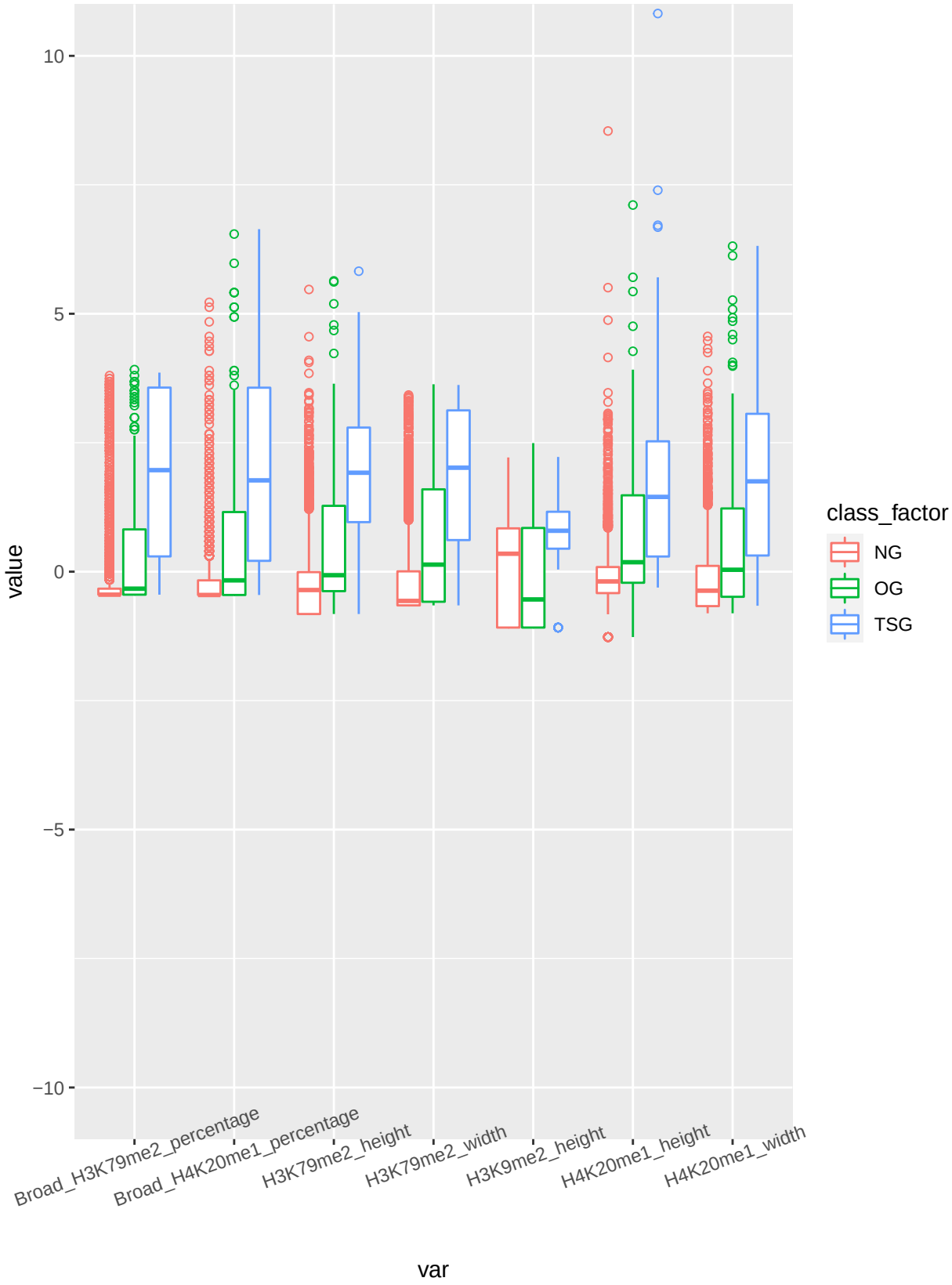

10
5
0
−5
−10
value
dN_to_dS_ratio
GDI
Missense_Zscore
ncGERP
ncRVIS
pLOF_Zscore
RVIS_percentile
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K4me2_percentage
Broad_H3K4me3_percentage
FamilyMemberCount
Gene_age
H3K4me2_width
H3K4me3_height
Length_H3K4me3
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K36me3_percentage
Broad_H3K4me1_percentage
H3K36me3_height
H3K36me3_width
H3K4me1_height
H3K4me1_width
H3K4me2_height
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K27ac_percentage
Broad_H3K27me3_percentage
H3K27ac_height
H3K27ac_width
H3K27me3_height
H3K27me3_width
H3K9me3_width
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K9ac_percentage
Broad_H3K9me2_percentage
Broad_H3K9me3_percentage
H3K9ac_height
H3K9ac_width
H3K9me2_width
H3K9me3_height
var
class_factor
NG
OG
TSG


10
5
0
−5
−10
value
Broad_H3K79me2_percentage
Broad_H4K20me1_percentage
H3K79me2_height
H3K79me2_width
H3K9me2_height
H4K20me1_height
H4K20me1_width
var
class_factor
NG
OG
TSG