Spring_project

Marianna Baranovskaia

21 05 2022

Download the data

The data are downloaded, 2 data sets are joined with *rbind()*, and the structure is corrected

```
kozak_dataset1 <-  read.csv("gnomad_hg38_af5pct_in_gencode_kozak_CDSstrands_snps_noborders_pos_sorted_ar
kozak_dataset2 <- read.csv("clinvar_in_gencode_kozak_CDSstrands_snps_F_noborders_pos_sorted_annot_combo

kozak_dataset <- rbind(kozak_dataset1, kozak_dataset2)

kozak_dataset$variant_annotation = as.factor(kozak_dataset$variant_annotation)
kozak_dataset$Change_description = as.factor(kozak_dataset$Change_description)
kozak_dataset$Kozak_type = as.factor(kozak_dataset$Kozak_type)
kozak_dataset$Clin_Sig = as.factor(kozak_dataset$Clin_Sig)
kozak_dataset$Gene = as.factor(kozak_dataset$Gene)
kozak_dataset$Ref_Kozak_efficiency = as.numeric(kozak_dataset$Ref_Kozak_efficiency)
```

```
## Warning:                      NA
```

```
kozak_dataset$Ref_Kozak_lower = as.numeric(kozak_dataset$Ref_Kozak_lower)
```

```
## Warning:                      NA
```

```
kozak_dataset$Ref_Kozak_upper = as.numeric(kozak_dataset$Ref_Kozak_upper)
```

```
## Warning:                      NA
```

```
kozak_dataset$Alt_Kozak_efficiency = as.numeric(kozak_dataset$Alt_Kozak_efficiency)
```

```
## Warning:                      NA
```

```
kozak_dataset$Alt_Kozak_lower = as.numeric(kozak_dataset$Alt_Kozak_lower)
```

```
## Warning:                      NA
```

```
kozak_dataset$Alt_Kozak_upper = as.numeric(kozak_dataset$Alt_Kozak_upper)
```

```
## Warning:                      NA
```

```
kozak_dataset$Relative_efficiency = as.numeric(kozak_dataset$Relative_efficiency)
```

```
## Warning:                      NA
```

```
summary(kozak_dataset)
```

```
##       ID             chromosome          position             Ref
##  Length:7921        Length:7921        Min.   :    49234   Length:7921
##  Class :character   Class :character   1st Qu.: 31261013   Class :character
##  Mode  :character   Mode  :character   Median : 55279554   Mode  :character
##                                        Mean   : 70451523
##                                         3rd Qu.:101407903
##                                         Max.   :244864086
##
##      Alt             Kozak_start          Kozak_end            Chain
##  Length:7921        Min.   :    49232   Min.   :    49244   Length:7921
##  Class :character   1st Qu.: 31261010   1st Qu.: 31261022   Class :character
##  Mode  :character   Median : 55279551   Median : 55279563   Mode  :character
##                     Mean   : 70451517   Mean   : 70451529
##                     3rd Qu.:101407897   3rd Qu.:101407909
##                     Max.   :244864085   Max.   :244864097
##
##  Kozak_variant_position           variant_annotation         Kozak_type
##  Min.   : 0.000         Error in annotation: 568     AUG_Kozak    :5211
##  1st Qu.: 3.000         missense           : 889     not_AUG_Kozak:2710
##  Median : 6.000         no_start           :2571
##  Mean   : 5.359         nonsense           :  45
##  3rd Qu.: 8.000         synonymous         :  15
##  Max.   :10.000         upstream           :3833
##
##  Ref_Kozak_efficiency Ref_Kozak_lower Ref_Kozak_upper  Alt_Kozak_efficiency
##  Min.   : 21.00       Min.   : 19.0   Min.   : 23.00   Min.   : 17.00
##  1st Qu.: 75.00       1st Qu.: 69.0   1st Qu.: 81.00   1st Qu.: 75.00
##  Median : 86.00       Median : 80.0   Median : 93.00   Median : 86.00
##  Mean   : 86.88       Mean   : 80.2   Mean   : 94.12   Mean   : 85.78
##  3rd Qu.:102.00       3rd Qu.: 94.0   3rd Qu.:110.00   3rd Qu.:100.00
##  Max.   :138.00       Max.   :127.0   Max.   :149.00   Max.   :144.00
##  NA's   :4945         NA's   :4945    NA's   :4945     NA's   :4945
##  Alt_Kozak_lower  Alt_Kozak_upper       Change_description Relative_efficiency
##  Min.   : 16.00   Min.   : 19.00   .             :4945     Min.   :0.373
##  1st Qu.: 69.00   1st Qu.: 81.00   equal         : 516     1st Qu.:0.929
##  Median : 79.00   Median : 93.00   getting higher:1241     Median :1.000
##  Mean   : 79.17   Mean   : 92.93   getting lower :1219     Mean   :1.001
##  3rd Qu.: 92.00   3rd Qu.:108.00                           3rd Qu.:1.060
##  Max.   :132.00   Max.   :156.00                           Max.   :3.857
##  NA's   :4945     NA's   :4945                             NA's   :4945
##                                           Clin_Sig         Gene
##  Uncertain_significance                       :3251   .      : 585
##  Likely_benign                                :1263   BRCA1  : 172
##  Benign                                       :1030   SMARCA4:  80
##  Pathogenic                                   : 871   MLH1   :  76
##  Likely_pathogenic                            : 629   RAD51C :  62
##  Conflicting_interpretations_of_pathogenicity: 435   TP53   :  58
##  (Other)                                      : 442   (Other):6888
```

```
str(kozak_dataset)
```

```
## 'data.frame':    7921 obs. of  21 variables:
##  $ ID                    : chr  "rs6660139" "rs61774959" "rs61777494" "rs1462467408" ...
##  $ chromosome            : chr  "chr1" "chr1" "chr1" "chr1" ...
##  $ position              : int  981169 1657267 1722599 13199588 13226110 16206527 16936795 23868283 2
##  $ Ref                   : chr  "A" "G" "G" "A" ...
##  $ Alt                   : chr  "G" "A" "A" "C" ...
##  $ Kozak_start           : int  981160 1657258 1722590 13199583 13226100 16206520 16936784 23868280 2
##  $ Kozak_end             : int  981172 1657270 1722602 13199595 13226112 16206532 16936796 23868292 2
##  $ Chain                 : chr  "-" "-" "-" "-" ...
##  $ Kozak_variant_position: int  2 2 2 6 1 4 10 8 3 10 ...
##  $ variant_annotation    : Factor w/ 6 levels "Error in annotation",..: 6 6 6 3 6 6 2 3 6 2 ...
##  $ Kozak_type            : Factor w/ 2 levels "AUG_Kozak","not_AUG_Kozak": 1 1 1 1 1 2 2 1 1 2 ...
##  $ Ref_Kozak_efficiency  : num  53 96 96 NA 89 NA NA NA 64 NA ...
##  $ Ref_Kozak_lower       : num  49 89 89 NA 82 NA NA NA 59 NA ...
##  $ Ref_Kozak_upper       : num  58 103 103 NA 97 NA NA NA 70 NA ...
##  $ Alt_Kozak_efficiency  : num  48 89 89 NA 88 NA NA NA 71 NA ...
##  $ Alt_Kozak_lower       : num  45 82 82 NA 80 NA NA NA 66 NA ...
##  $ Alt_Kozak_upper       : num  52 96 96 NA 95 NA NA NA 77 NA ...
##  $ Change_description    : Factor w/ 4 levels ".","equal","getting higher",..: 4 4 4 1 4 1 1 1 3 1 .
##  $ Relative_efficiency   : num  0.906 0.927 0.927 NA 0.989 ...
##  $ Clin_Sig              : Factor w/ 15 levels "","Affects","Benign",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Gene                  : Factor w/ 2252 levels ".","A2ML1","AAAS",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The preprocessed data set is written in the file.

```
write.csv(kozak_dataset, "sum_dataset_01.csv")
```

Subset of the variants with known pathogenicity

I should analyse on this step of the project the variants with clear pathogenicity: Benign and Pathogenic or Likely Pathogenic or Pathogenic/Likely Pathogenic

```
known_sign = c('Pathogenic', 'Likely_pathogenic', 'Pathogenic/Likely_pathogenic', 'Benign')
kozak_dataset_short <-subset(kozak_dataset, Clin_Sig %in% known_sign)
kozak_dataset_short$group <- ifelse(kozak_dataset_short$Clin_Sig %in% c('Pathogenic', 'Likely_pathogeni
head(kozak_dataset_short)
```

```
##             ID chromosome position Ref Alt Kozak_start Kozak_end Chain
## 1    rs6660139       chr1   981169   A   G      981160    981172     -
## 2   rs61774959       chr1  1657267   G   A     1657258   1657270     -
## 3   rs61777494       chr1  1722599   G   A     1722590   1722602     -
## 4 rs1462467408       chr1 13199588   A   C    13199583  13199595     -
## 5            .       chr1 13226110   G   C    13226100  13226112     -
## 6     rs221052       chr1 16206527   C   T    16206520  16206532     -
##   Kozak_variant_position variant_annotation    Kozak_type Ref_Kozak_efficiency
## 1                      2           upstream     AUG_Kozak                   53
## 2                      2           upstream     AUG_Kozak                   96
## 3                      2           upstream     AUG_Kozak                   96
## 4                      6           no_start     AUG_Kozak                   NA
## 5                      1           upstream     AUG_Kozak                   89
## 6                      4           upstream not_AUG_Kozak                   NA
##   Ref_Kozak_lower Ref_Kozak_upper Alt_Kozak_efficiency Alt_Kozak_lower
## 1              49              58                   48              45
## 2              89             103                   89              82
## 3              89             103                   89              82
## 4              NA              NA                   NA              NA
## 5              82              97                   88              80
## 6              NA              NA                   NA              NA
##   Alt_Kozak_upper Change_description Relative_efficiency Clin_Sig Gene  group
## 1              52      getting lower           0.9056604   Benign    . benign
## 2              96      getting lower           0.9270833   Benign    . benign
## 3              96      getting lower           0.9270833   Benign    . benign
## 4              NA                  .                  NA   Benign    . benign
## 5              95      getting lower           0.9887640   Benign    . benign
## 6              NA                  .                  NA   Benign    . benign
```

```
nrow(kozak_dataset_short[kozak_dataset_short$group == 'benign', ])
```

```
## [1] 1030
```

```
nrow(kozak_dataset_short[kozak_dataset_short$group == 'pathogenic', ])
```

```
## [1] 1627
```

```
nrow(kozak_dataset_short[kozak_dataset_short$Kozak_type == 'not_AUG_Kozak', ])
```

```
## [1] 673
```

```
nrow(kozak_dataset_short[kozak_dataset_short$Kozak_type == 'AUG_Kozak', ])
```

```
## [1] 1984
```

```
nrow(kozak_dataset_short)
```

```
## [1] 2657
```

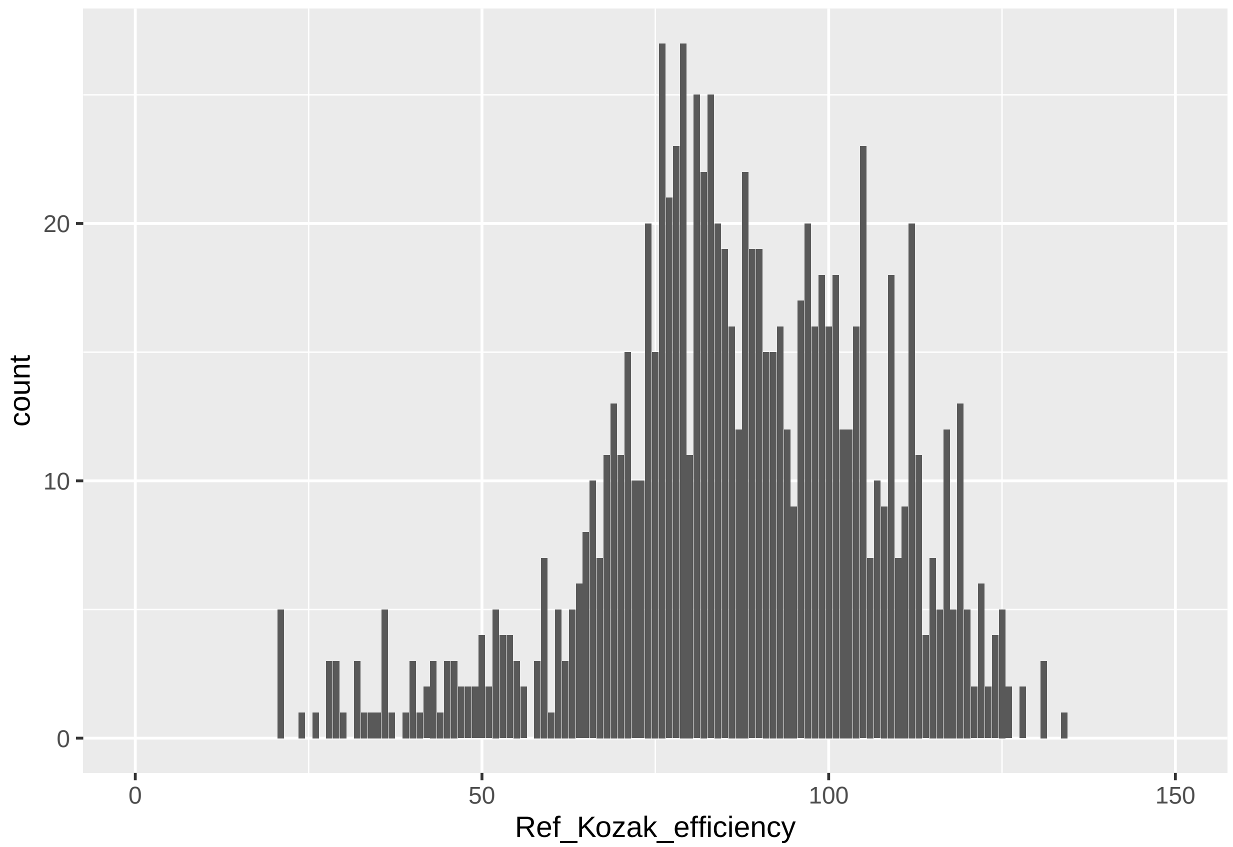
This data set is written in the file too.

```
write.csv(kozak_dataset_short, "sum_dataset_known.csv")
```

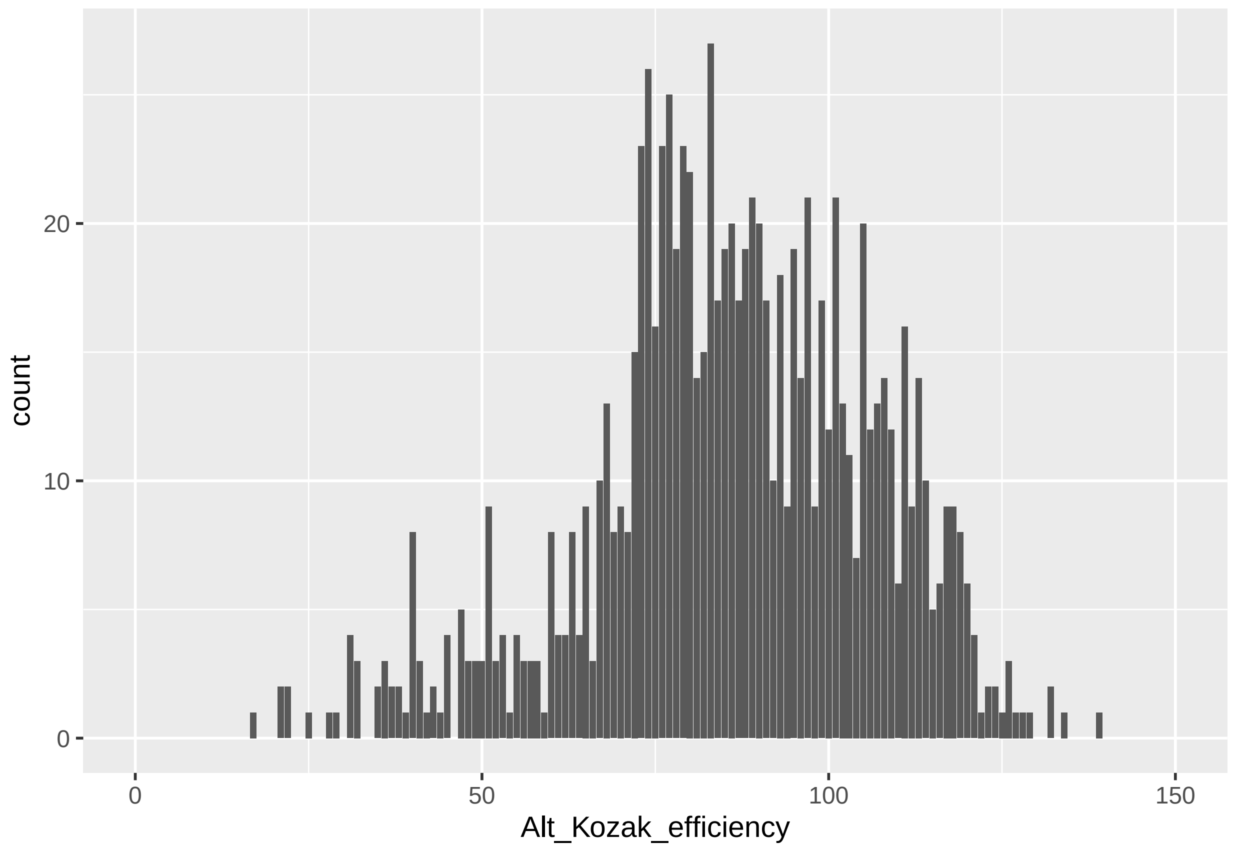
Vizualization for the big data set

Distributions of Kozak sequence efficiencies:

```
plot_F_01 <- ggplot(kozak_dataset_short)+
  geom_histogram(aes(x=Ref_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
plot_F_01
```

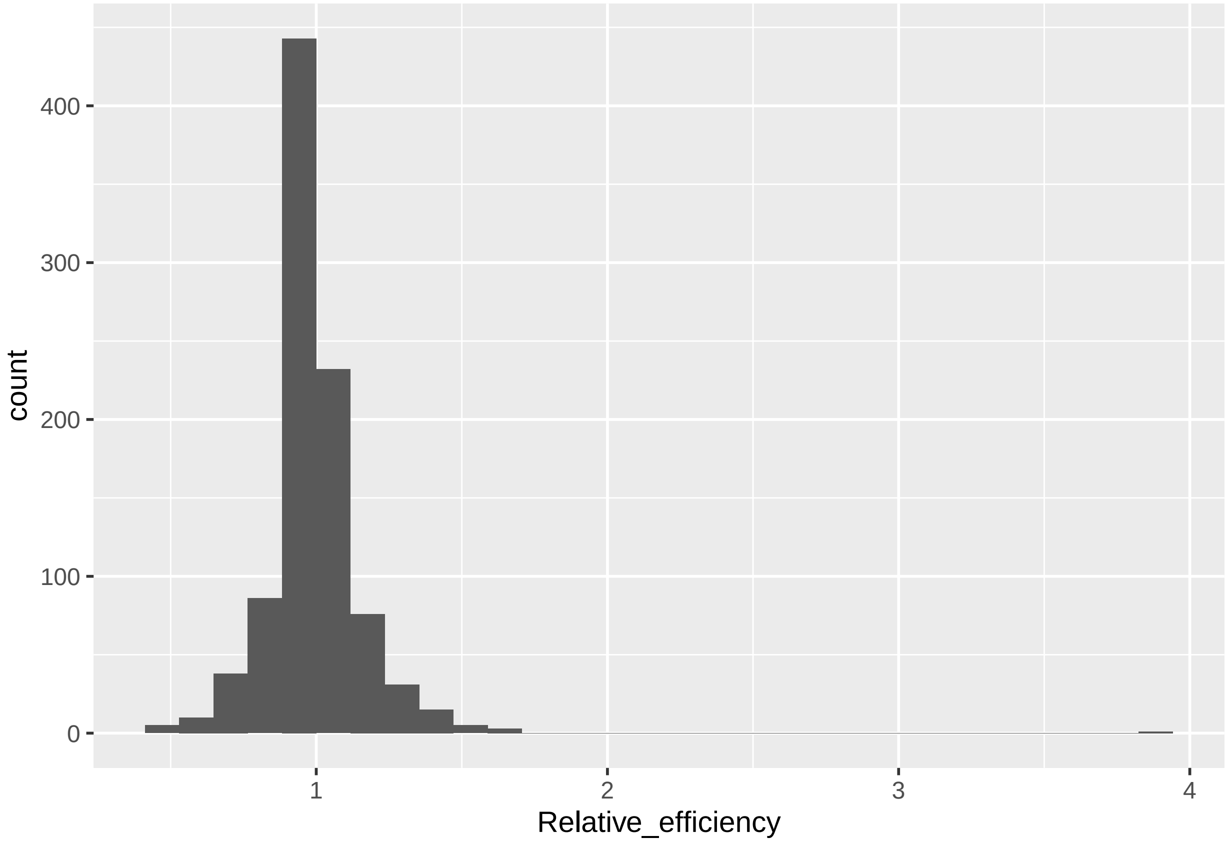


```r
plot_F_02 <- ggplot(kozak_dataset_short)+
  geom_histogram(aes(x=Alt_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
plot_F_02
```

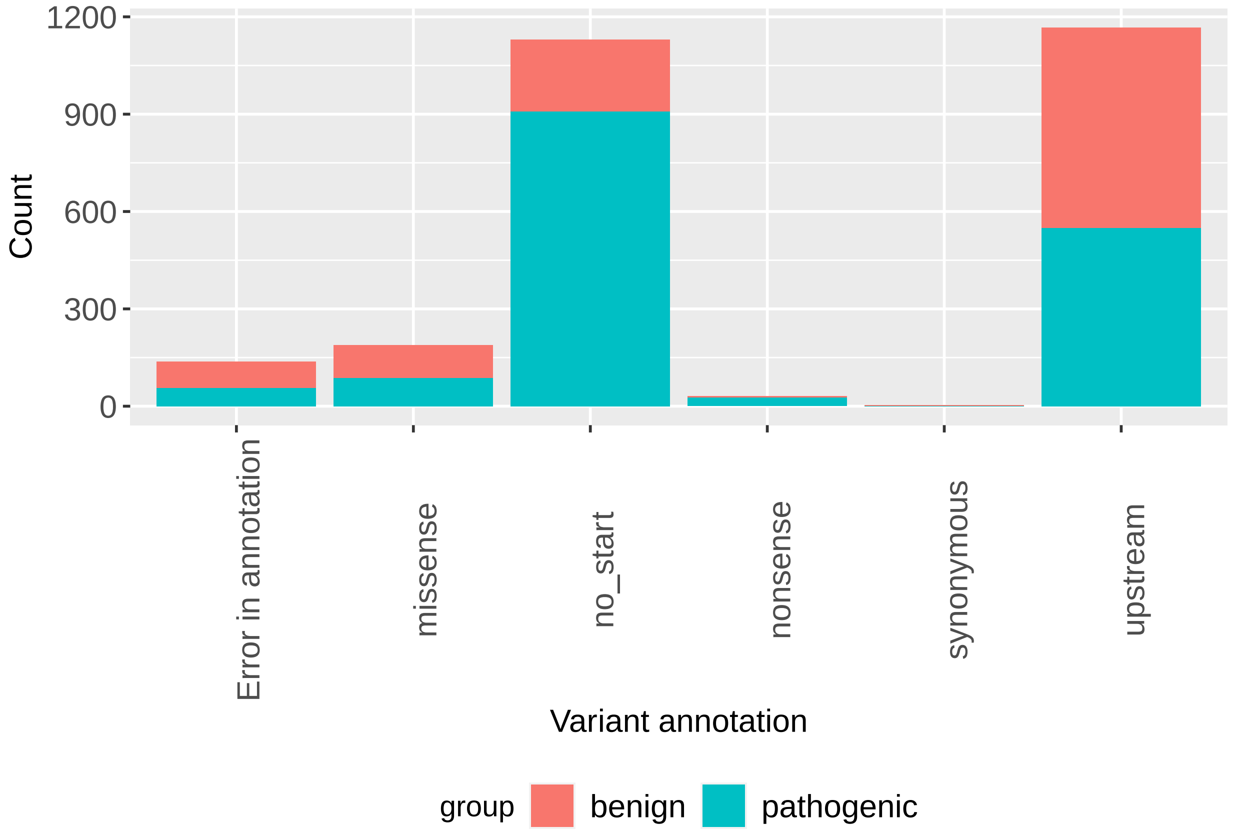


```
plot_F_03 <- ggplot(kozak_dataset_short)+
  geom_histogram(aes(x=Relative_efficiency))+
  theme(legend.position="bottom", legend.box = "horizontal")
plot_F_03
```

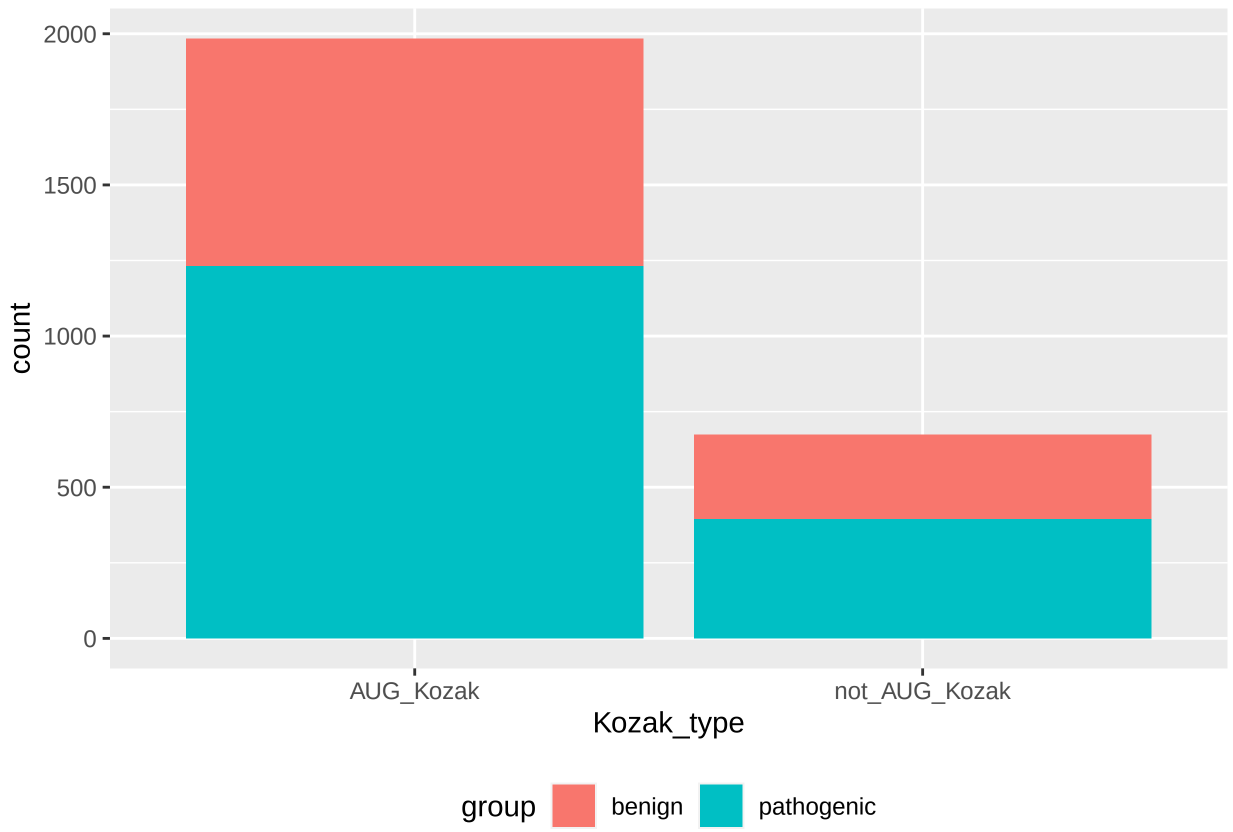
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
plot_F_04 <- ggplot(kozak_dataset_short)+
  geom_histogram(aes(x=variant_annotation, fill = group), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text=element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12, angle = 90),
        axis.text.y=element_text(size=12))+
  ylab("Count")+
  xlab("Variant annotation")
plot_F_04
```



```
plot_F_05 <- ggplot(kozak_dataset_short)+
  geom_histogram(aes(x=Kozak_type, fill = group), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")
plot_F_05
```



Saving the plots silently =)

Significant part of the data belongs to the 'not_AUG_Kozak' subset but we work now only with AUG Kozak sequences.

"AUG_Kozak" subset

```
kozak_dataset_short_AUG <- subset(kozak_dataset_short, Kozak_type == 'AUG_Kozak')
nrow(kozak_dataset_short_AUG)
```

```
## [1] 1984
```

```
nrow(kozak_dataset_short_AUG[kozak_dataset_short_AUG$group == 'pathogenic', ])
```

```
## [1] 1232
```

```
nrow(kozak_dataset_short_AUG[kozak_dataset_short_AUG$group == 'benign', ])
```

```
## [1] 752
```

This data set is written in the file too.

```
write.csv(kozak_dataset_short_AUG, "sum_dataset_known_AUG.csv")
```

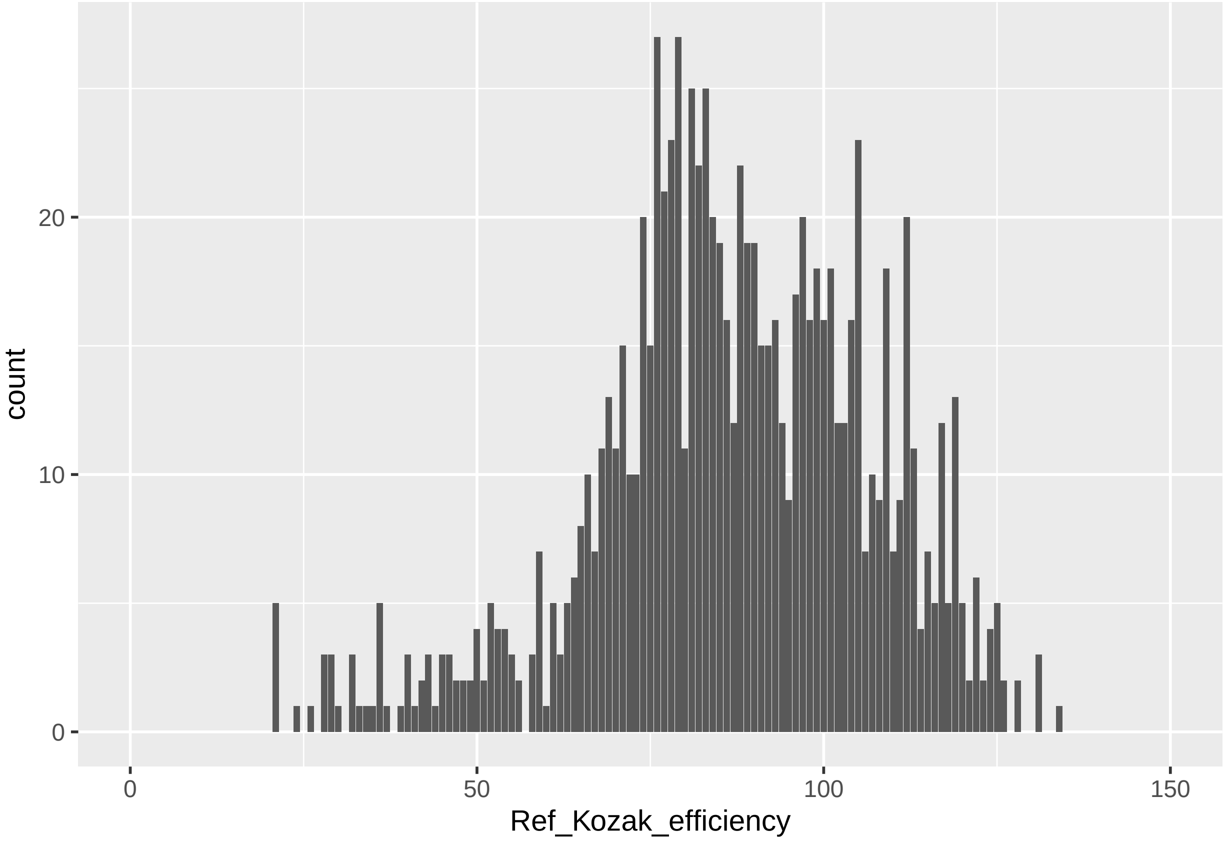
Vizualization for the AUG_Kozak data set

```
plot_F_06 <- ggplot(kozak_dataset_short_AUG)+
  geom_histogram(aes(x=Ref_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_06
```

```
## Warning: Removed 1039 rows containing non-finite values (stat_count).
```

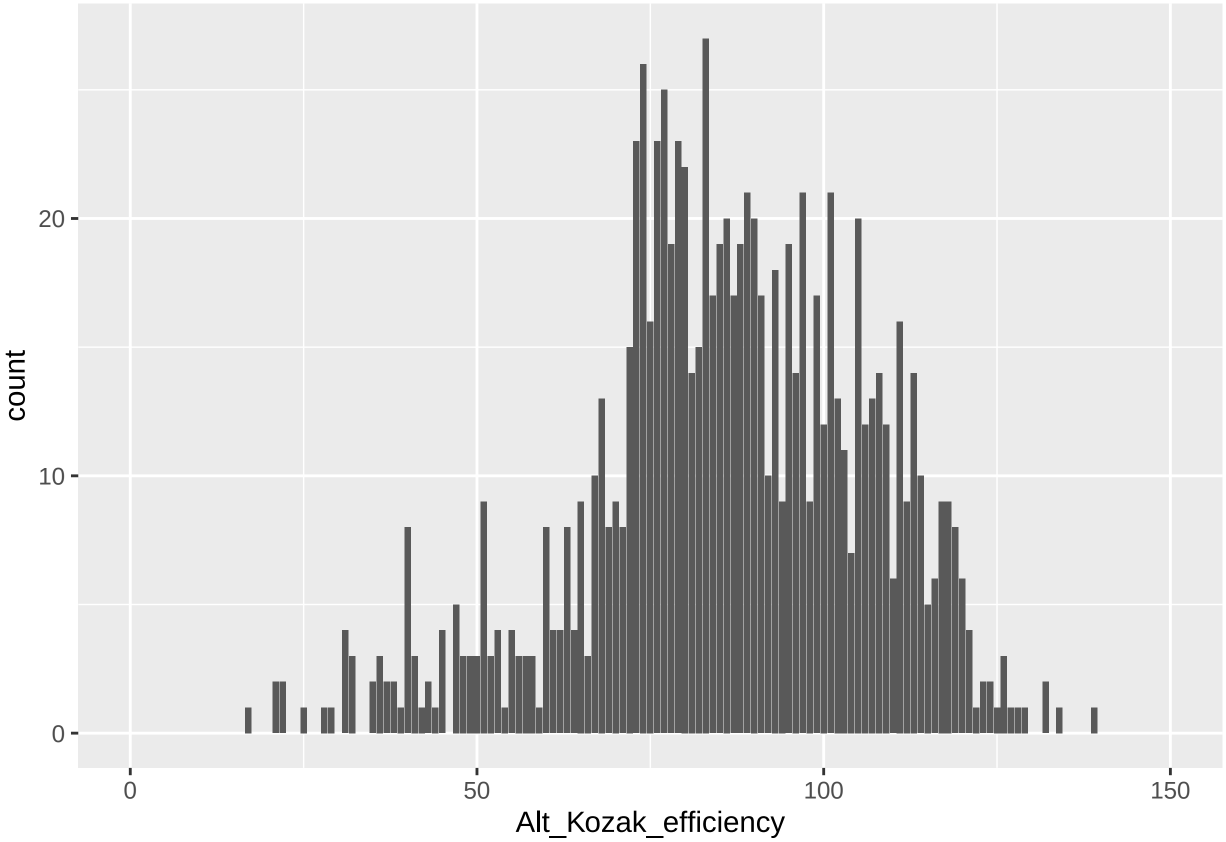


```
plot_F_07 <- ggplot(kozak_dataset_short_AUG)+
  geom_histogram(aes(x=Alt_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_07
```

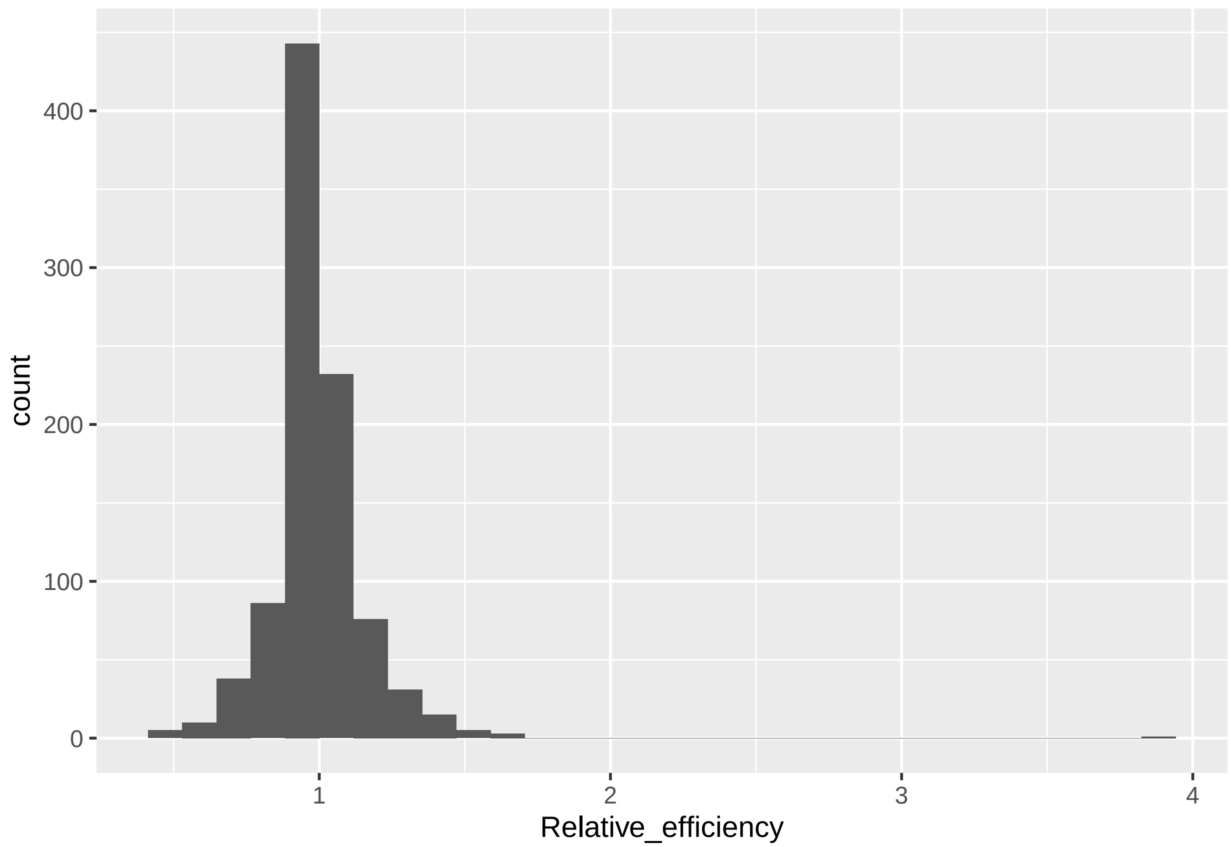
```
## Warning: Removed 1039 rows containing non-finite values (stat_count).
```



```
plot_F_08 <- ggplot(kozak_dataset_short_AUG)+
  geom_histogram(aes(x=Relative_efficiency))+
  theme(legend.position="bottom", legend.box = "horizontal")
plot_F_08
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

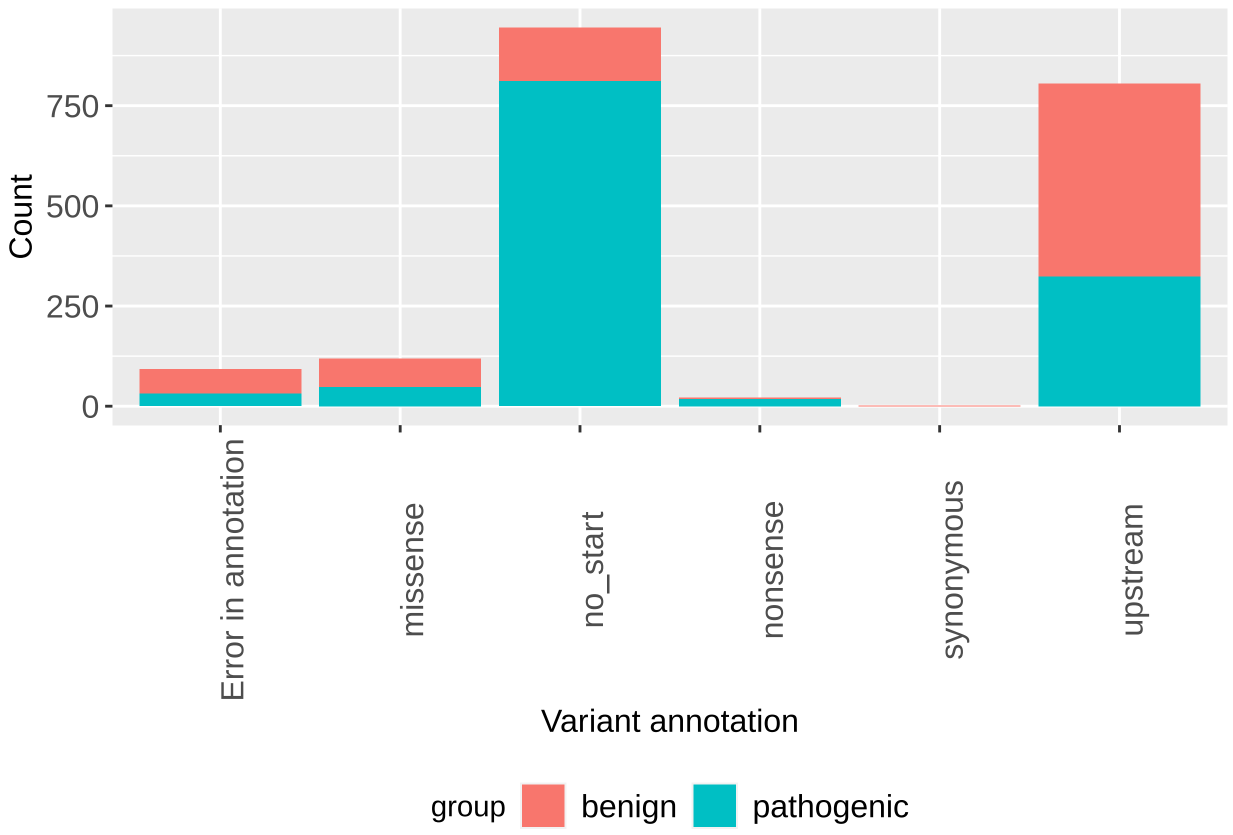
```
## Warning: Removed 1039 rows containing non-finite values (stat_bin).
```



```
plot_F_09 <- ggplot(kozak_dataset_short_AUG)+
  geom_histogram(aes(x=variant_annotation, fill = group), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text=element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12, angle = 90),
        axis.text.y=element_text(size=12))+
  ylab("Count")+
  xlab("Variant annotation")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_09
```

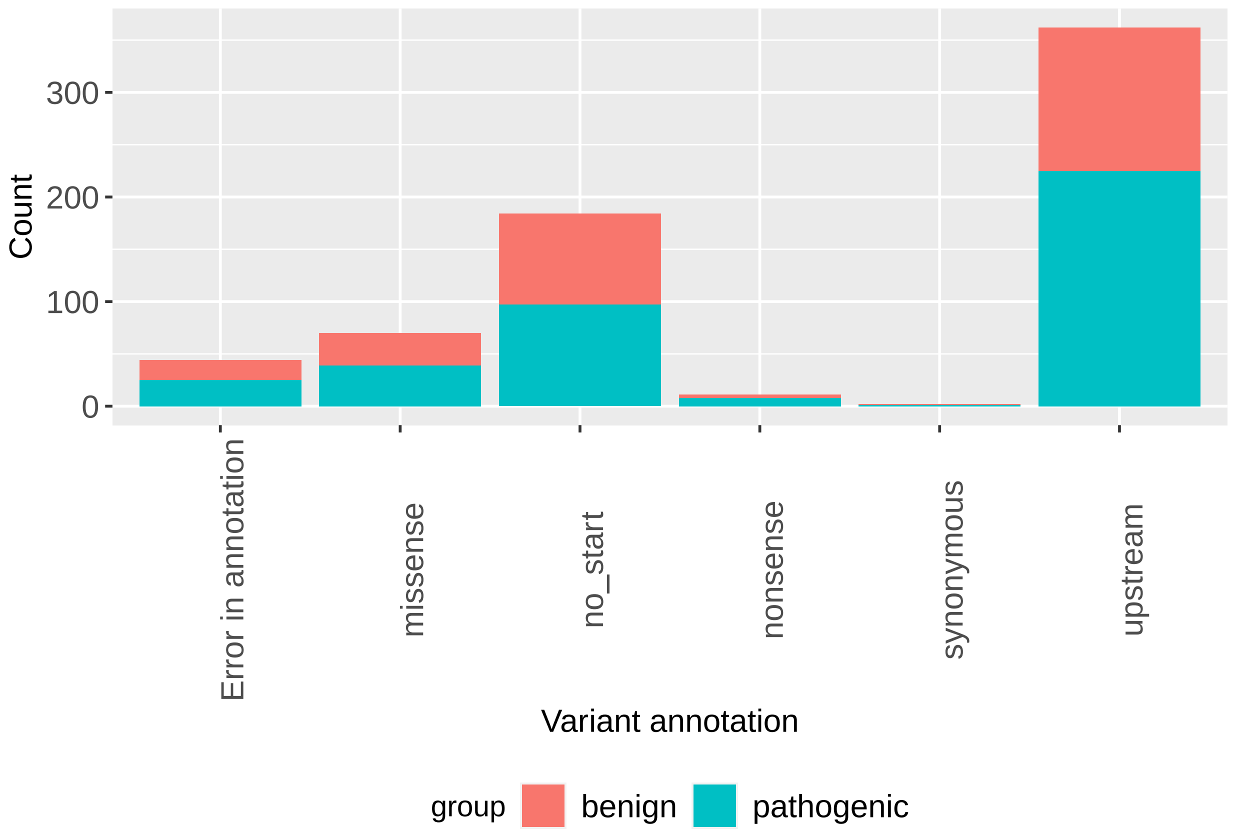


```
# additional plot for 'not_AUG_Kozak'

plot_F_09_notAUG <- ggplot(kozak_dataset_short[kozak_dataset_short$Kozak_type == 'not_AUG_Kozak', ])+
  geom_histogram(aes(x=variant_annotation, fill = group), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text=element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12, angle = 90),
        axis.text.y=element_text(size=12))+
  ylab("Count")+
  xlab("Variant annotation")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_09_notAUG
```



Saving the plots silently =)

"Upstream+synonymous" subset

We decided to do this subset because 'no_start', 'missense' and 'nonsense' variants change the protein primary structure and can be pathogenic just because of this effect and not because of up/downregulation with Kozak sequence.

```
locations = c('upstream', 'synonymous')
kozak_dataset_short_AUG_2 <- subset(kozak_dataset_short_AUG, variant_annotation %in% locations)
nrow(kozak_dataset_short_AUG_2)
```

```
## [1] 807
```

```
nrow(kozak_dataset_short_AUG_2[kozak_dataset_short_AUG_2$group == 'pathogenic', ])
```

```
## [1] 324
```

```
nrow(kozak_dataset_short_AUG_2[kozak_dataset_short_AUG_2$group == 'benign', ])
```

```
## [1] 483
```

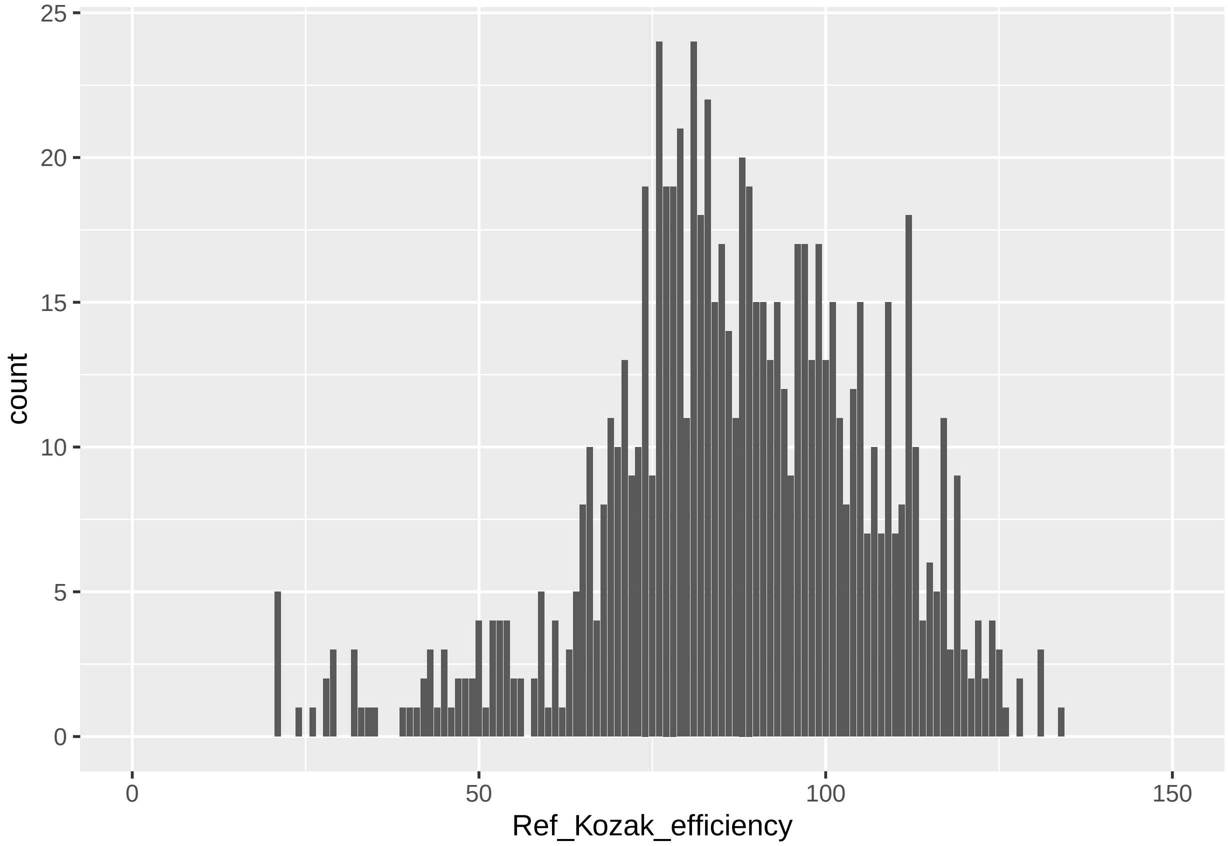
Vizualization for "Upstream+synonymous" subset

```
plot_F_10 <- ggplot(kozak_dataset_short_AUG_2)+
  geom_histogram(aes(x=Ref_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_10
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```

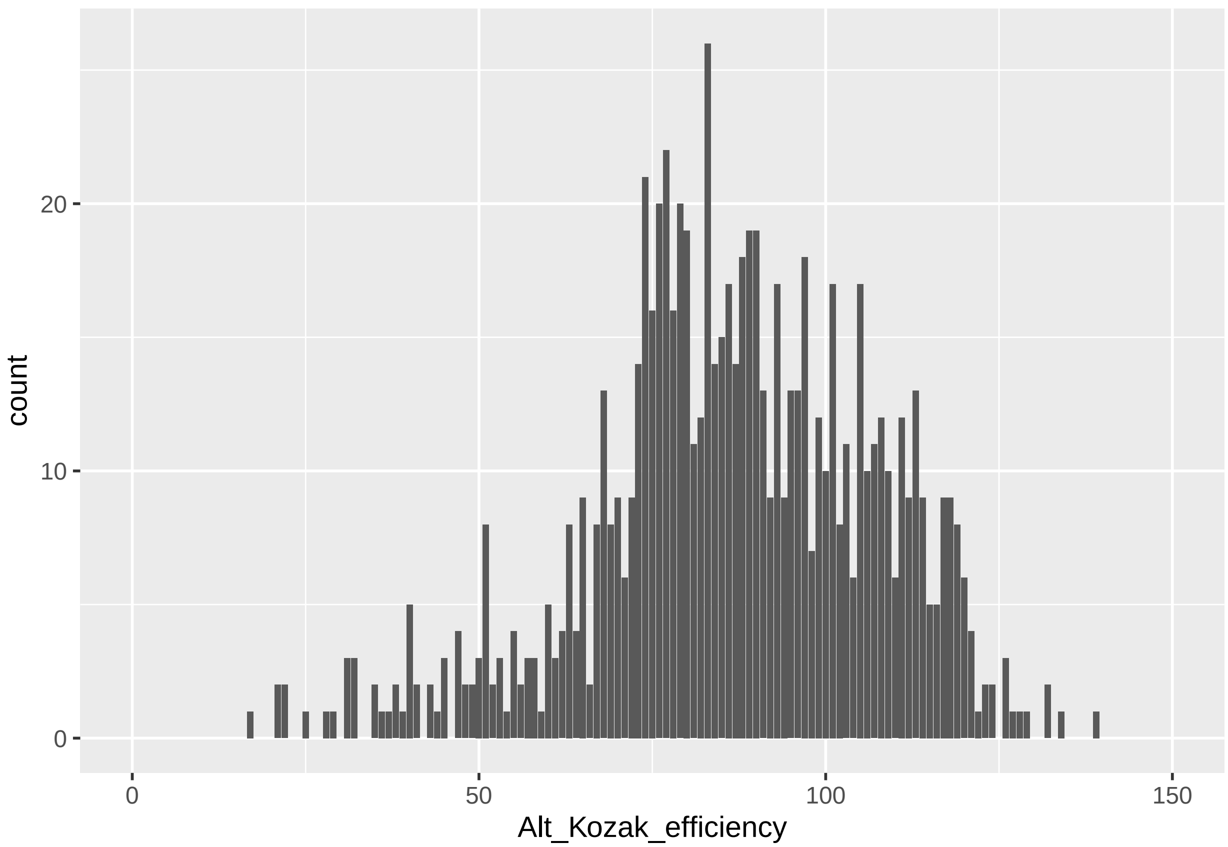


```
plot_F_11 <- ggplot(kozak_dataset_short_AUG_2)+
  geom_histogram(aes(x=Alt_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_11
```

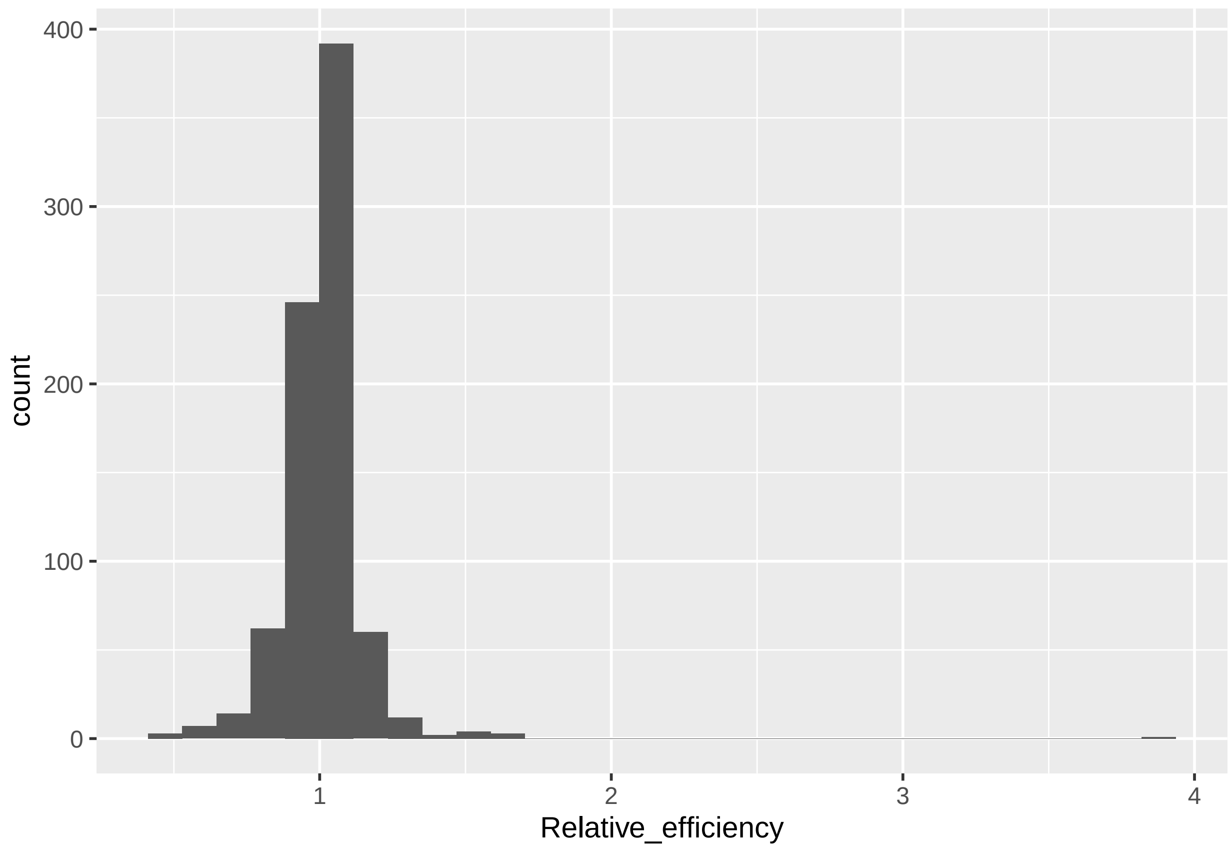
```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```



```
plot_F_12 <- ggplot(kozak_dataset_short_AUG_2)+
  geom_histogram(aes(x=Relative_efficiency))+
  theme(legend.position="bottom", legend.box = "horizontal")
plot_F_12
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

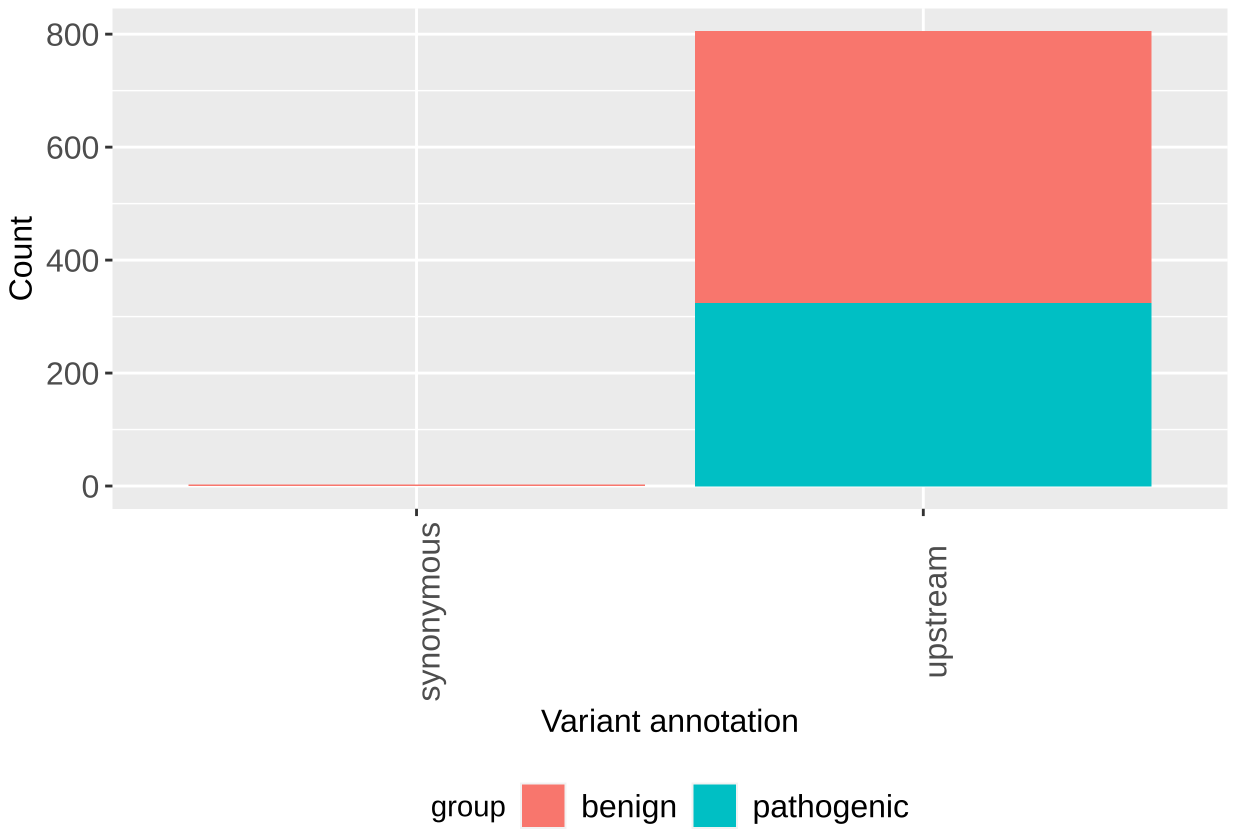
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```
plot_F_13 <- ggplot(kozak_dataset_short_AUG_2)+
  geom_histogram(aes(x=variant_annotation, fill = group), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text=element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12, angle = 90),
        axis.text.y=element_text(size=12))+
  ylab("Count")+
  xlab("Variant annotation")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_13
```



Saving the plots silently =)

Which variants have the effect more than 50%?

```
subset(kozak_dataset_short_AUG_2, Relative_efficiency>1.5)
```

```
##              ID chromosome position Ref Alt Kozak_start Kozak_end Chain
## 155   rs4457918      chr14 23953786   T   G    23953782  23953794     +
## 298  rs11878547      chr19 35902284   T   C    35902277  35902289     -
## 2448      18142      chr14 73170984   G   C    73170979  73170991     +
## 3449     381605      chr17 31163208   T   G    31163204  31163216     +
## 3684      91553      chr17 43106527   G   T    43106520  43106532     -
## 3686      54247      chr17 43106528   C   A    43106520  43106532     -
##      Kozak_variant_position variant_annotation Kozak_type Ref_Kozak_efficiency
## 155                       3           upstream  AUG_Kozak                   68
## 298                       4           upstream  AUG_Kozak                   48
## 2448                      4           upstream  AUG_Kozak                   33
## 3449                      3           upstream  AUG_Kozak                   59
## 3684                      4           upstream  AUG_Kozak                   21
## 3686                      3           upstream  AUG_Kozak                   21
##      Ref_Kozak_lower Ref_Kozak_upper Alt_Kozak_efficiency Alt_Kozak_lower
## 155               63              73                  108             100
## 298               44              53                   78              72
## 2448              31              36                   55              51
## 3449              55              64                   93              85
## 3684              19              23                   32              30
## 3686              19              23                   81              75
##      Alt_Kozak_upper Change_description Relative_efficiency
## 155              117     getting higher            1.588235
## 298               86     getting higher            1.625000
## 2448              59     getting higher            1.666667
## 3449             101     getting higher            1.576271
## 3684              35     getting higher            1.523810
## 3686              88     getting higher            3.857143
##                            Clin_Sig  Gene      group
## 155                          Benign     .     benign
## 298                          Benign     .     benign
## 2448                     Pathogenic PSEN1 pathogenic
## 3449                     Pathogenic   NF1 pathogenic
## 3684                     Pathogenic BRCA1 pathogenic
## 3686 Pathogenic/Likely_pathogenic BRCA1 pathogenic
```

```
subset(kozak_dataset_short_AUG_2, Relative_efficiency<0.5)
```

```
##           ID chromosome position Ref Alt Kozak_start Kozak_end Chain
## 3059  552136      chr16 56511158   C   T    56511150  56511162     -
## 6236 1067945       chr5 60922156   C   T    60922148  60922160     -
##      Kozak_variant_position variant_annotation Kozak_type Ref_Kozak_efficiency
## 3059                      3           upstream  AUG_Kozak                   89
## 6236                      3           upstream  AUG_Kozak                  108
##      Ref_Kozak_lower Ref_Kozak_upper Alt_Kozak_efficiency Alt_Kozak_lower
## 3059              82              97                   40              37
## 6236             100             117                   49              45
##      Alt_Kozak_upper Change_description Relative_efficiency          Clin_Sig
## 3059              44      getting lower           0.4494382 Likely_pathogenic
## 6236              53     getting higher           0.4537037 Likely_pathogenic
##       Gene      group
## 3059  BBS2 pathogenic
## 6236 ERCC8 pathogenic
```

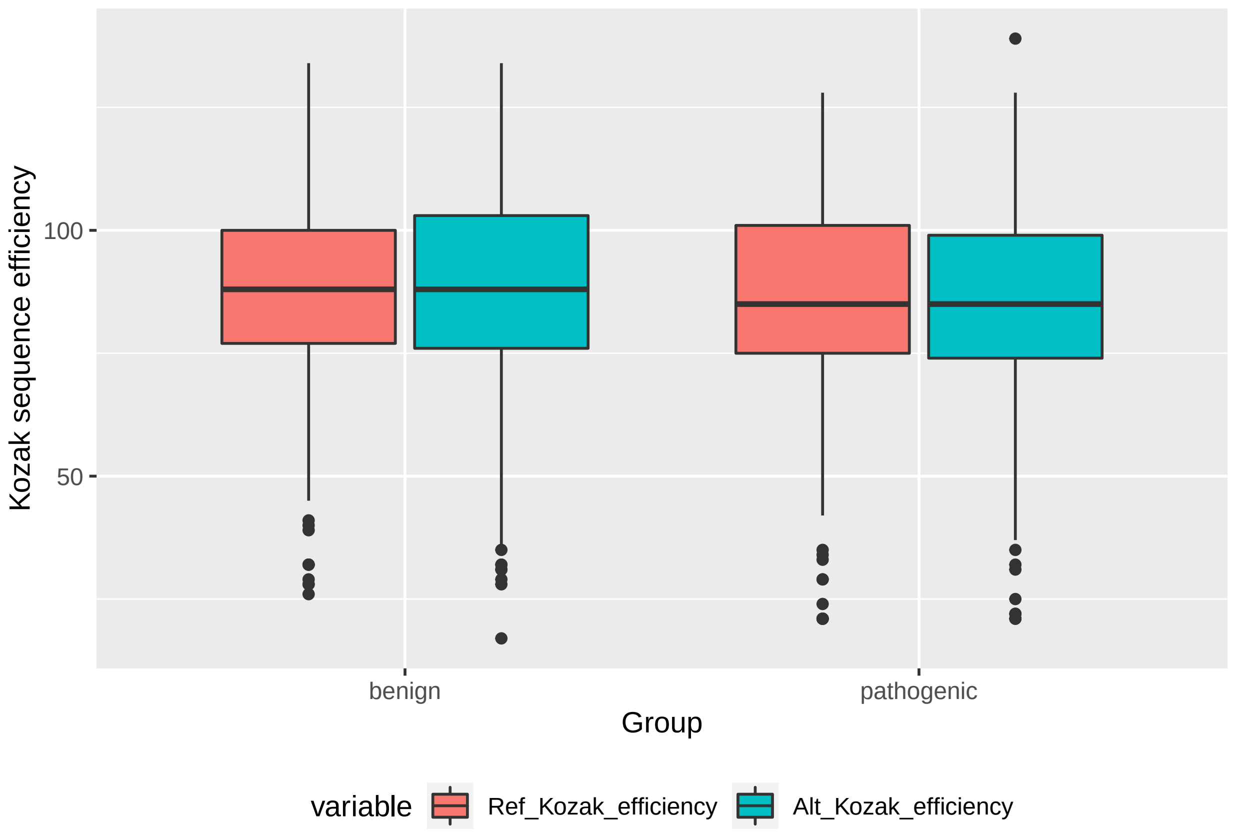
Vizualization for “Upstream+synonymous” subset (melted)

```
kozak_dataset_short_AUG_2_part_melted <- melt(kozak_dataset_short_AUG_2[, c('ID',  'group', 'Ref_Kozak_

kozak_dataset_short_AUG_2_part_melted2 <- melt(kozak_dataset_short_AUG_2[, c('ID',  'group', 'Kozak_var
kozak_dataset_short_AUG_2_part_melted2$Kozak_variant_position <- as.factor(kozak_dataset_short_AUG_2_pa
```

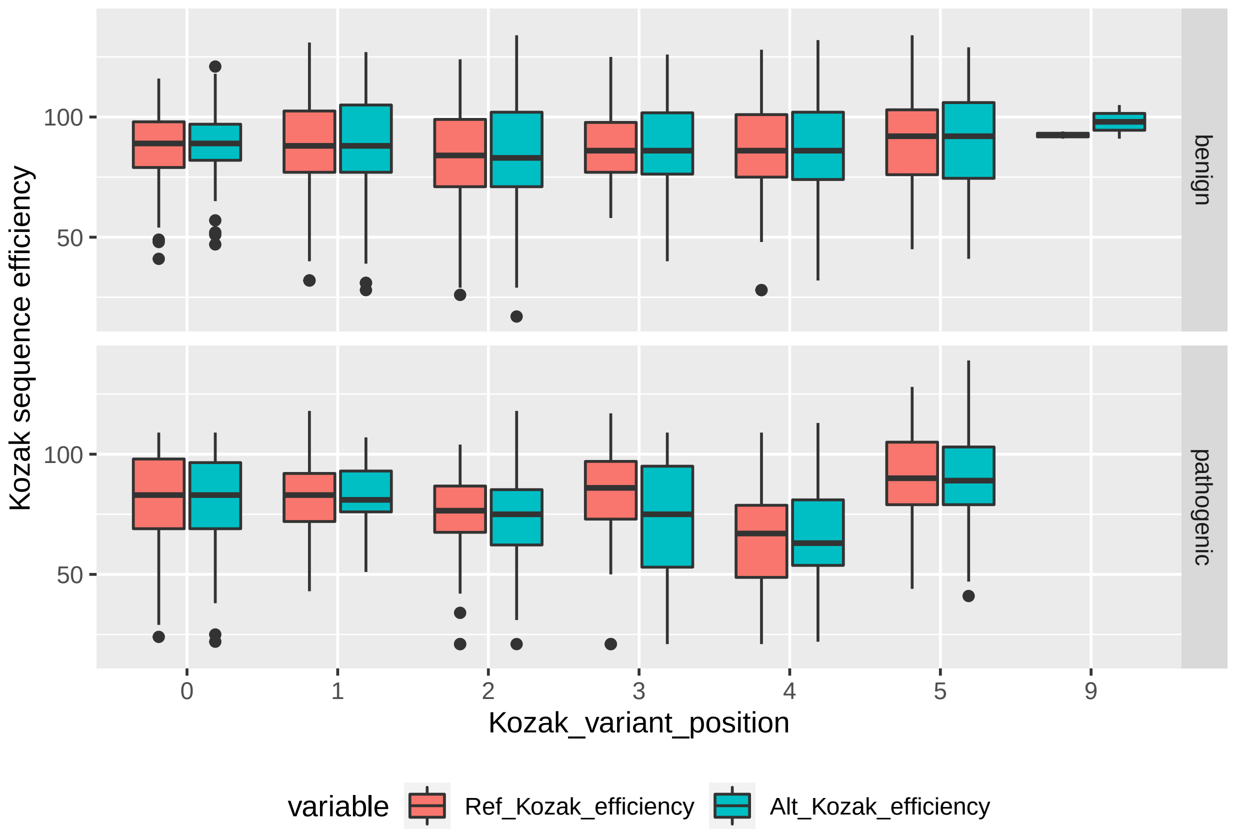
```
plot_F_14 <- ggplot(kozak_dataset_short_AUG_2_part_melted, aes(x=group, fill=variable))+
  geom_boxplot(aes(y=value))+
  theme(legend.position="bottom", legend.box = "horizontal")+
  ylab('Kozak sequence efficiency')+
  xlab('Group')
plot_F_14
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



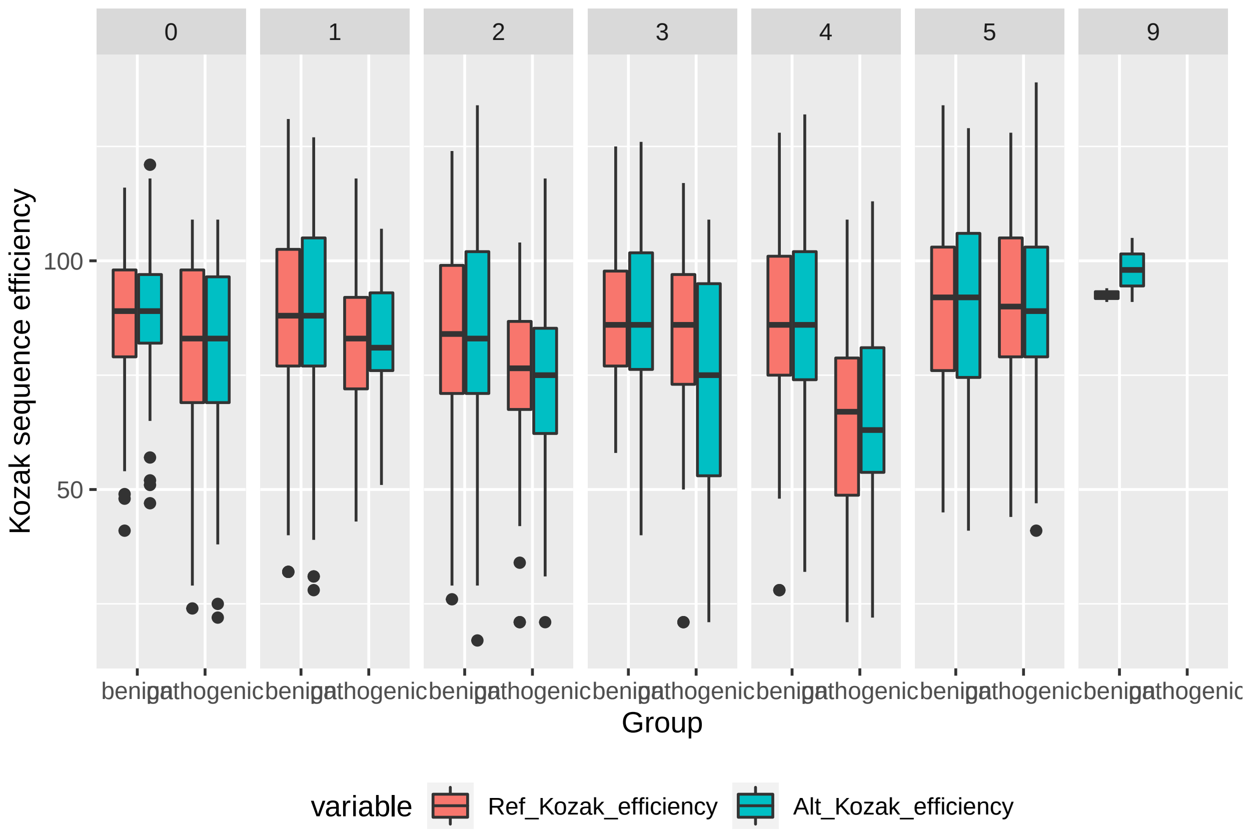
```
plot_F_15 <- ggplot(kozak_dataset_short_AUG_2_part_melted2, aes(x=Kozak_variant_position, fill=variable)
  geom_boxplot(aes(y=value))+
  theme(legend.position="bottom", legend.box = "horizontal")+
  ylab('Kozak sequence efficiency')+
  xlab('Kozak_variant_position')+
  facet_grid(rows = vars(group))
plot_F_15
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



```
plot_F_16 <- ggplot(kozak_dataset_short_AUG_2_part_melted2, aes(x=group, fill=variable))+
  geom_boxplot(aes(y=value))+
  theme(legend.position="bottom", legend.box = "horizontal")+
  ylab('Kozak sequence efficiency')+
  xlab('Group')+
  facet_grid(cols = vars(Kozak_variant_position))
plot_F_16
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



Saving the plots silently =)

There seems to be no significant differences in the distributions here.

"Only significant" subset

The last subset is the variants which have non-intersected confidence intervals (i.e. with significant change in the translation efficiency)

```
kozak_dataset_short_AUG_3 <- subset(kozak_dataset_short_AUG_2, Relative_efficiency != 'NA')

kozak_dataset_short_AUG_3$is_significant <- c(NA * nrow(kozak_dataset_short_AUG_3))

kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency > 1, ]$is_significant <-
  ifelse(kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency > 1, ]$Alt_Kozak_lower
           kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency > 1, ]$Ref_Kozak_uppe

kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency <= 1, ]$is_significant <-
  ifelse(kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency <= 1, ]$Alt_Kozak_uppe
           kozak_dataset_short_AUG_3[kozak_dataset_short_AUG_3$Relative_efficiency <= 1, ]$Ref_Kozak_lo

sign_diff_Kozaks_AUG <- subset(kozak_dataset_short_AUG_3, is_significant == TRUE)
nrow(kozak_dataset_short_AUG_3)
```

```
## [1] 806
```

```
nrow(sign_diff_Kozaks_AUG)
```

```
## [1] 92
```

```
nrow(sign_diff_Kozaks_AUG[sign_diff_Kozaks_AUG$group == 'pathogenic',])
```

```
## [1] 43
```

```
nrow(sign_diff_Kozaks_AUG[sign_diff_Kozaks_AUG$group == 'benign',])
```

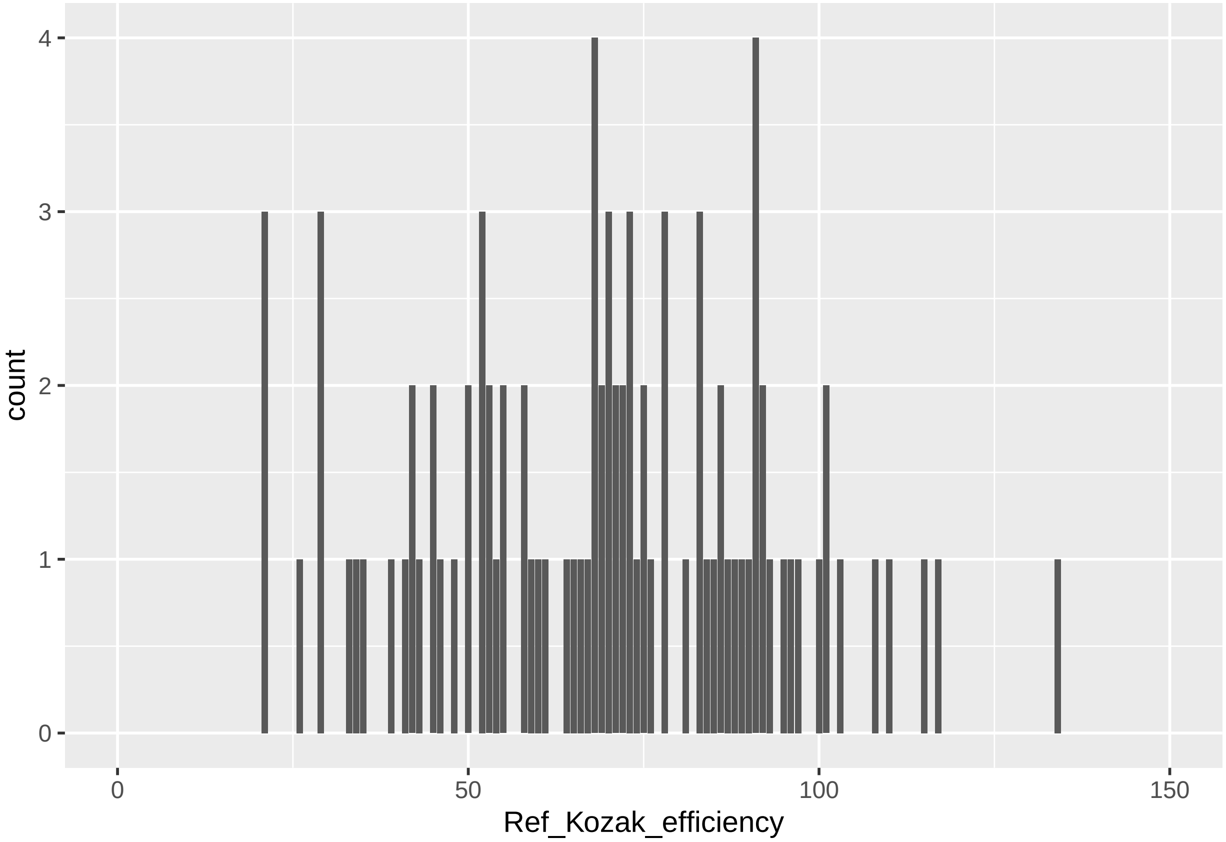
```
## [1] 49
```

Vizualization of “Only significant” subset

```
plot_F_16 <- ggplot(sign_diff_Kozaks_AUG)+
  geom_histogram(aes(x=Ref_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

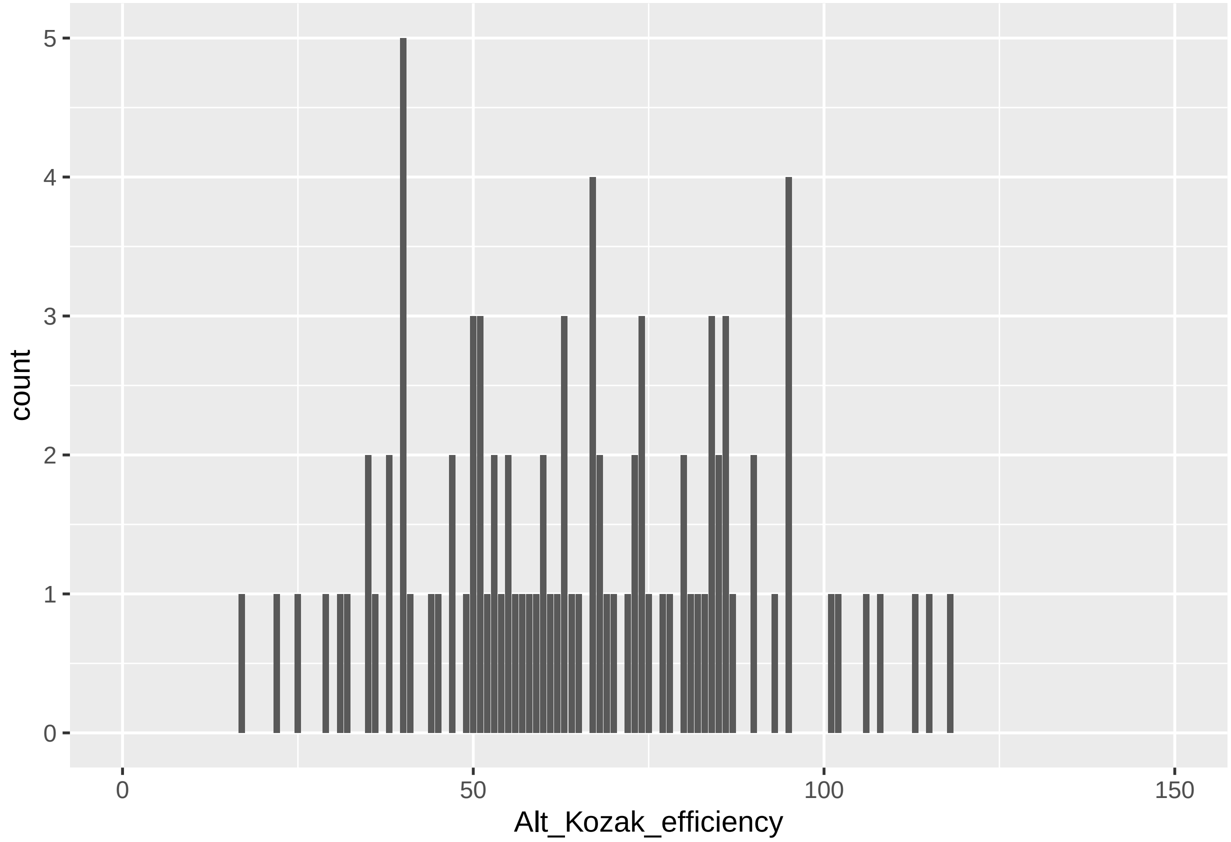
```
plot_F_16
```



```
plot_F_17 <- ggplot(sign_diff_Kozaks_AUG)+
  geom_histogram(aes(x=Alt_Kozak_efficiency), stat="count")+
  theme(legend.position="bottom", legend.box = "horizontal")+
  xlim(c(0,150))
```

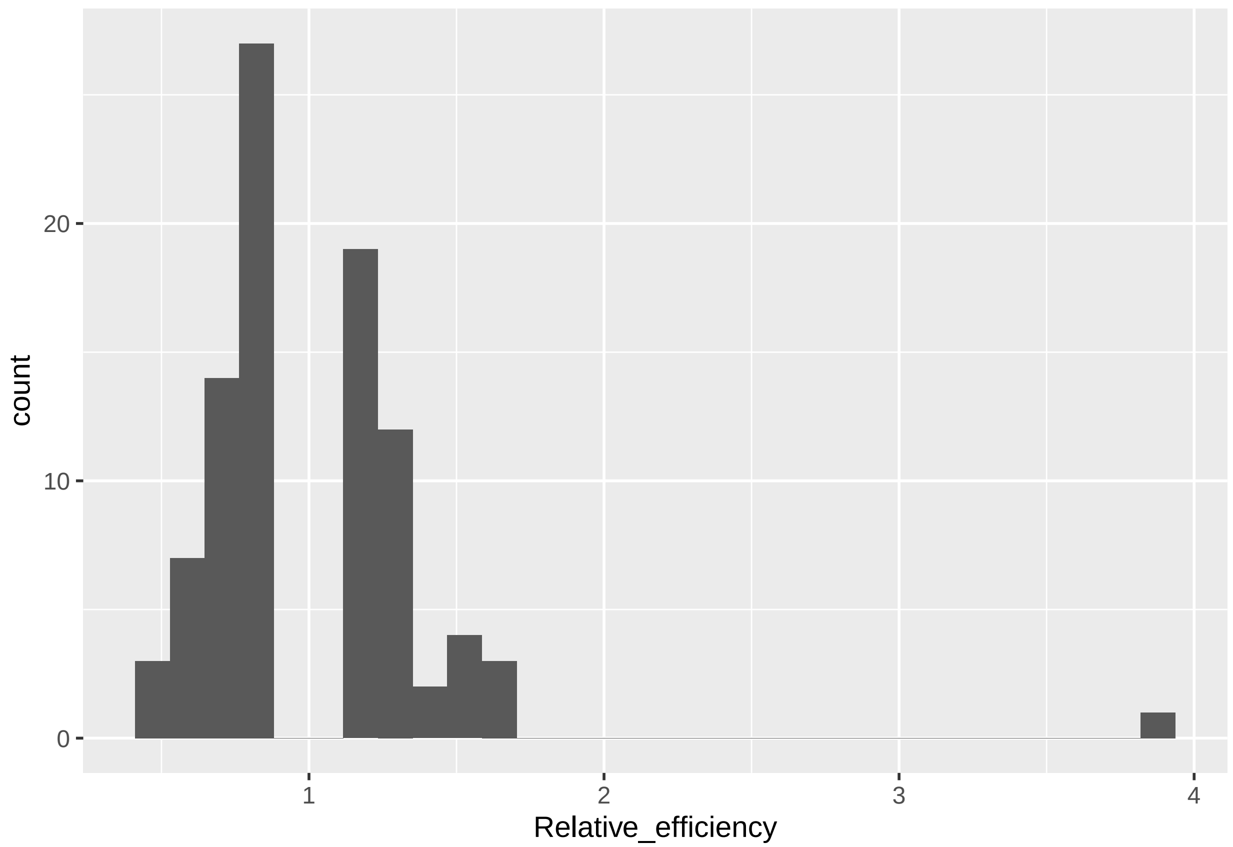
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
plot_F_17
```



```
plot_F_18 <- ggplot(sign_diff_Kozaks_AUG)+
  geom_histogram(aes(x=Relative_efficiency))+
  theme(legend.position="bottom", legend.box = "horizontal")
plot_F_18
```

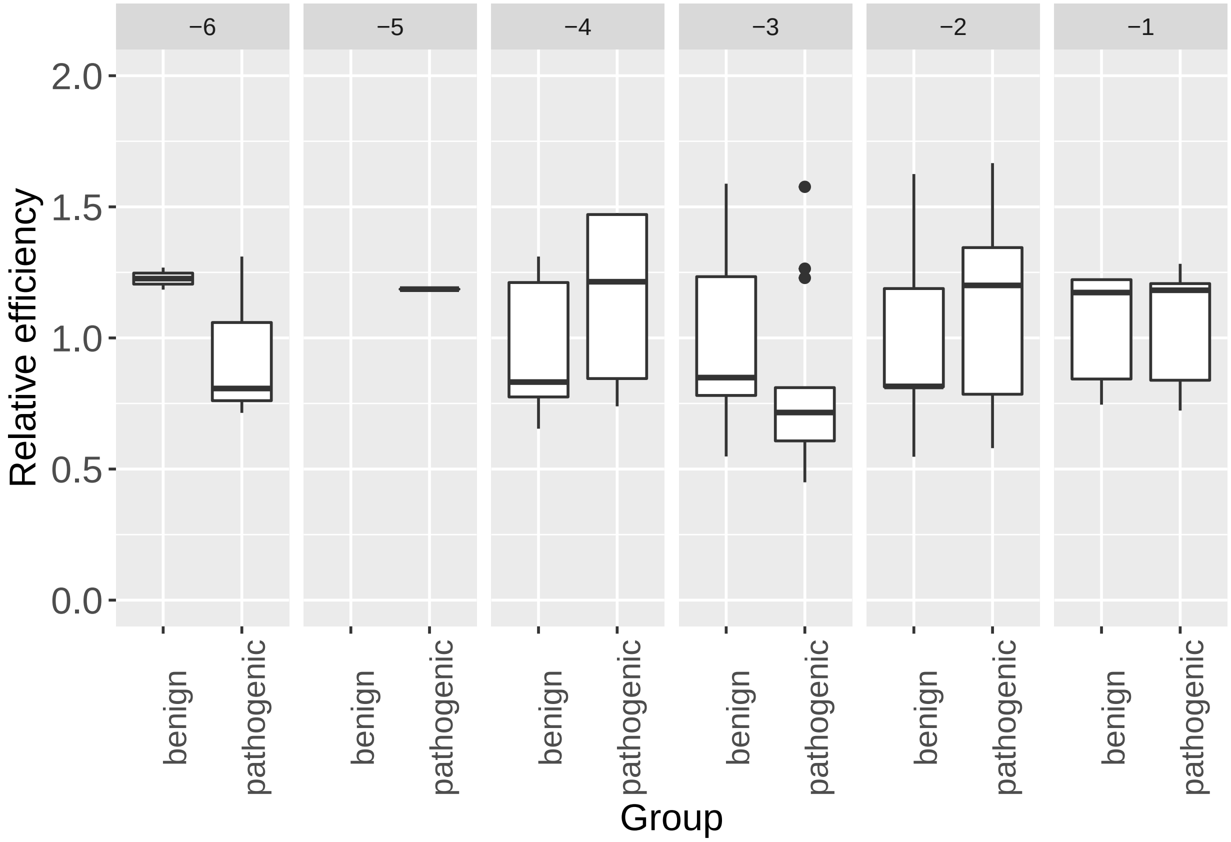
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# additional
new_labels = c('-6', '-5', '-4', '-3', '-2', '-1')
names(new_labels) = c(0, 1, 2, 3, 4, 5)

plot_F_22 <- ggplot(sign_diff_Kozaks_AUG, aes(x=group, y=Relative_efficiency))+
  geom_boxplot()+
  theme(legend.position="bottom", legend.box = "horizontal")+
  ylab('Relative efficiency')+
  ylim(c(0, 2))+
  xlab('Group')+
  facet_grid(cols = vars(Kozak_variant_position),
             labeller = labeller(Kozak_variant_position = new_labels))+
  scale_fill_manual(values = c("cyan", "gray"))+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text = element_text(size=14),
        axis.title.x=element_text(size=14),
        axis.title.y=element_text(size=14),
        axis.text.y=element_text(size=14),
        axis.text.x=element_text(size=12, angle = 90),)
plot_F_22
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



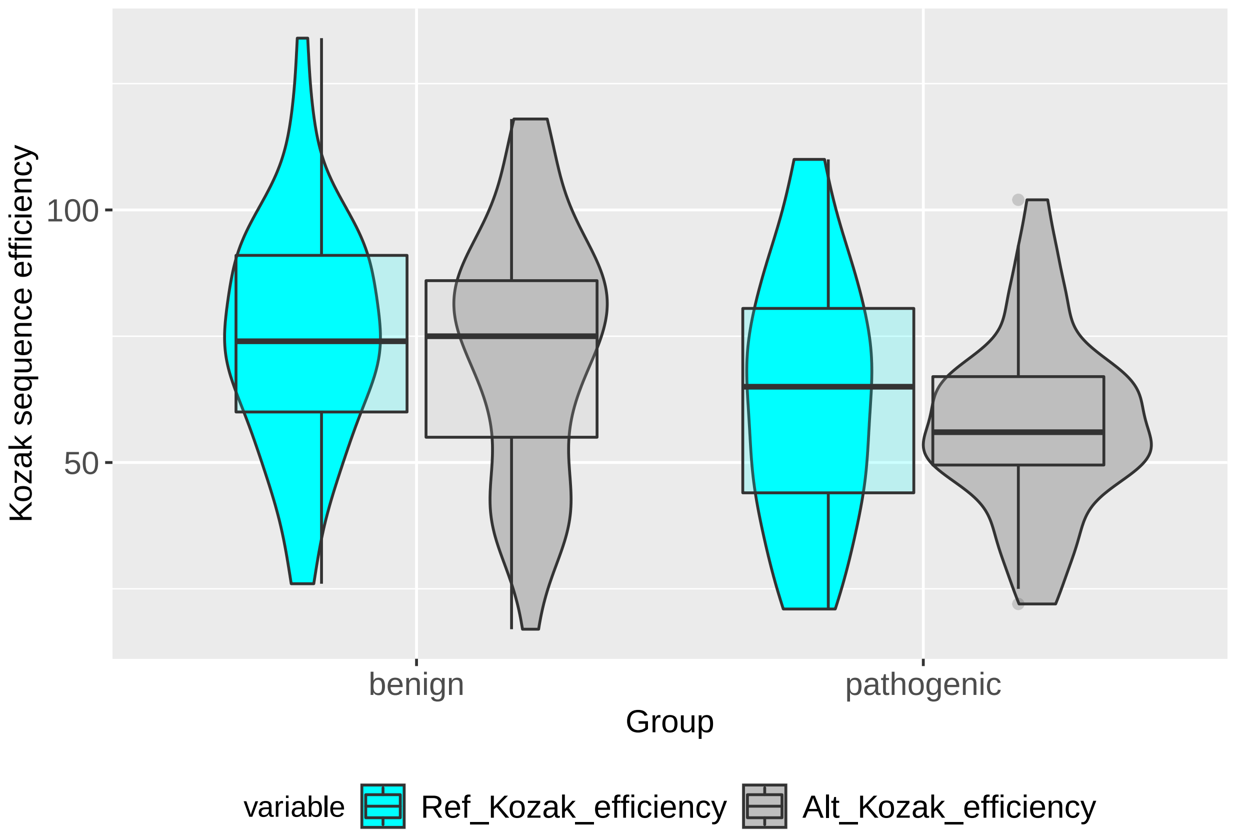
Saving the plots silently =)

Vizualization of "Only significant" subset (melted)

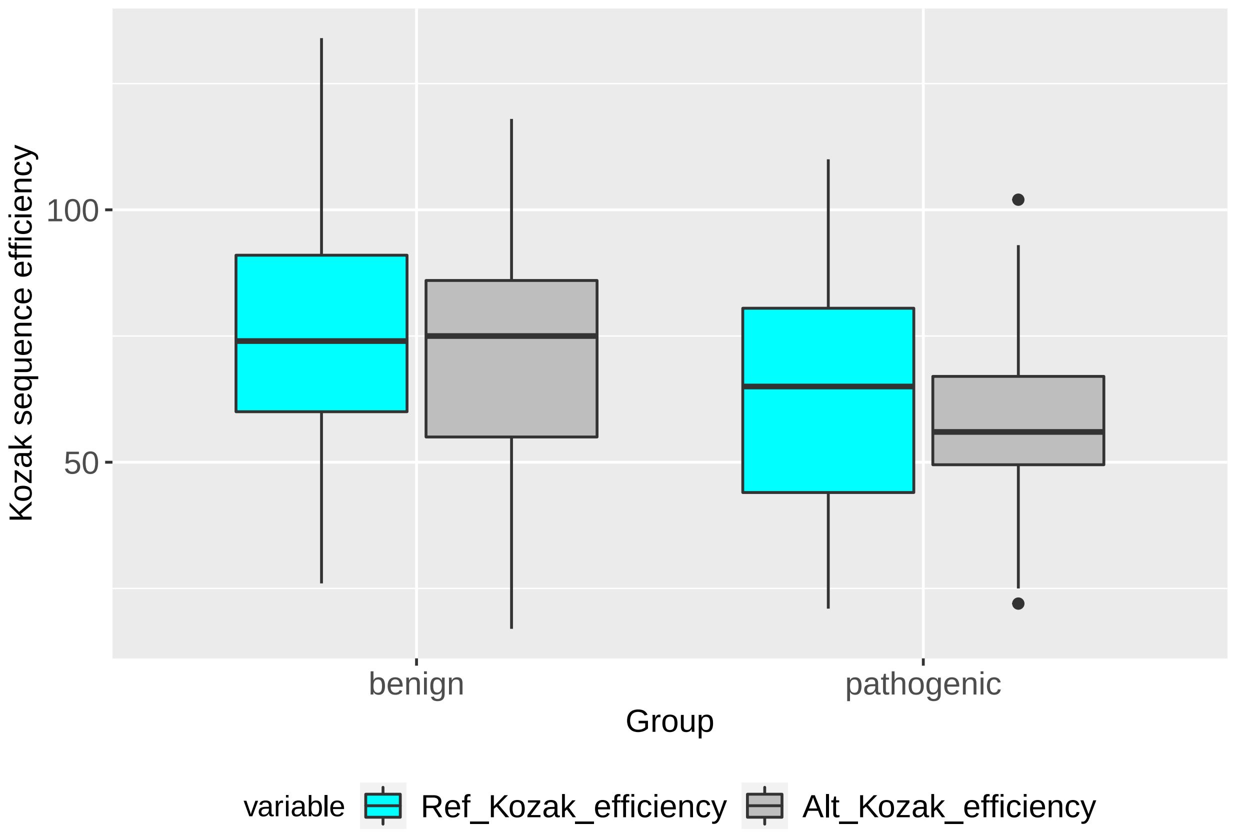
```
sign_diff_Kozaks_AUG_part_melted <- melt(sign_diff_Kozaks_AUG[, c('ID',  'group', 'Ref_Kozak_efficiency

sign_diff_Kozaks_AUG_part_melted2 <- melt(sign_diff_Kozaks_AUG[, c('ID',  'group', 'Kozak_variant_posit
sign_diff_Kozaks_AUG_part_melted2$Kozak_variant_position <- as.factor(sign_diff_Kozaks_AUG_part_melted2
```

```
plot_F_19 <- ggplot(sign_diff_Kozaks_AUG_part_melted, aes(x=group, fill=variable))+
  geom_violin(aes(y=value))+
  geom_boxplot(aes(y=value), alpha=0.2)+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text = element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12),
        axis.text.y=element_text(size=12))+
  ylab('Kozak sequence efficiency')+
  xlab('Group')+
  scale_fill_manual(values = c("cyan", "gray"))
plot_F_19
```



```
plot_F_20 <- ggplot(sign_diff_Kozaks_AUG_part_melted, aes(x=group, fill=variable))+
  geom_boxplot(aes(y=value))+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text = element_text(size=12),
        axis.title.x=element_text(size=12),
        axis.title.y=element_text(size=12),
        axis.text.x=element_text(size=12),
        axis.text.y=element_text(size=12))+
  ylab('Kozak sequence efficiency')+
  xlab('Group')+
  scale_fill_manual(values = c("cyan", "gray"))
plot_F_20
```

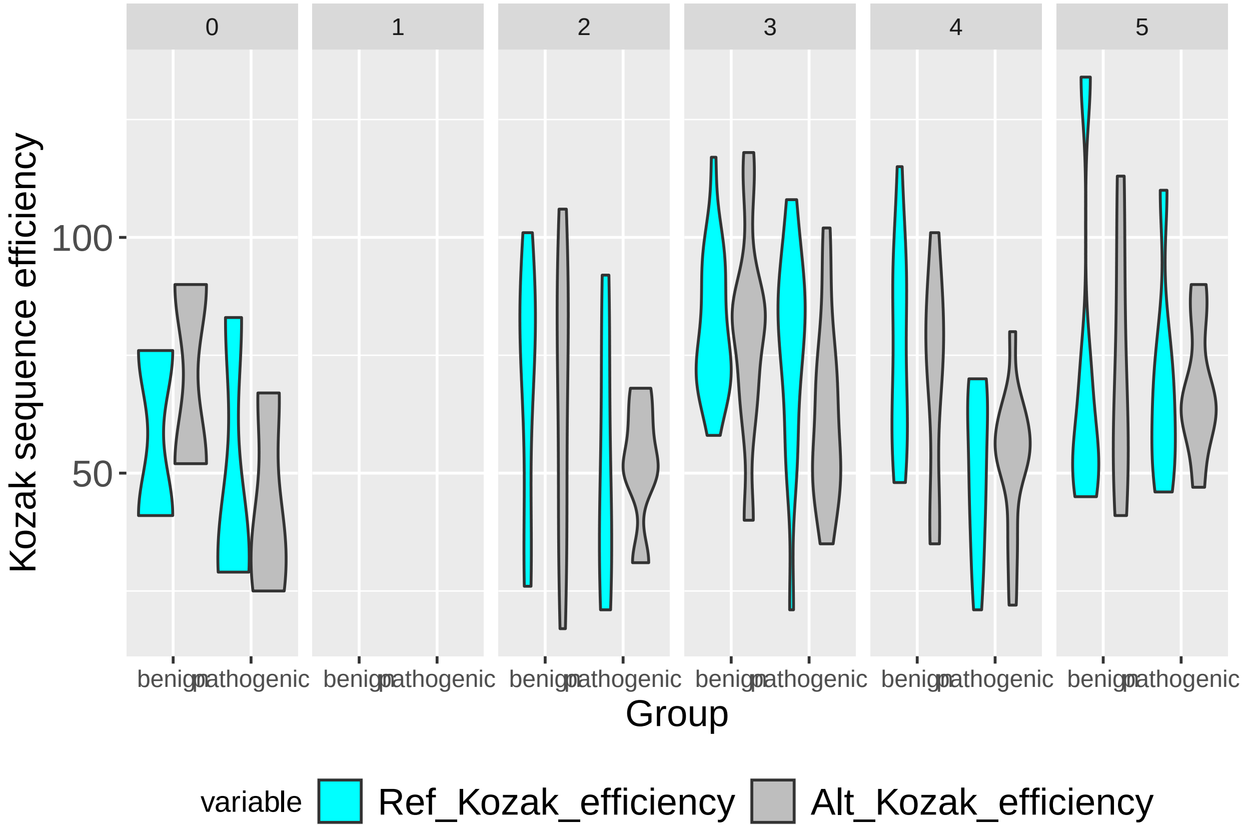


```r
plot_F_21 <- ggplot(sign_diff_Kozaks_AUG_part_melted2, aes(x=group, y=value))+
  geom_violin(aes(fill=variable))+
  theme(legend.position="bottom", legend.box = "horizontal")+
  ylab('Kozak sequence efficiency')+
  xlab('Group')+
  facet_grid(cols = vars(Kozak_variant_position))+
  scale_fill_manual(values = c("cyan", "gray"))+
  theme(legend.position="bottom", legend.box = "horizontal",
        legend.text = element_text(size=14),
        axis.title.x=element_text(size=14),
        axis.title.y=element_text(size=14),
        axis.text.y=element_text(size=14))
plot_F_21
```

```
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
```

```
## Warning in max(data$density):   'max'                  ;
## -Inf
```

```
## Warning: Computation failed in `stat_ydensity()`:
##       1    ,      -- 0
```



Saving the plots silently =)