Database SNV summary plots

mfpfox

2023-05-31

```
source("1_import_annotated_variants.R")
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
## 'data.frame':    9110589 obs. of  30 variables:
##  $ keyID37aa                : chr  "10_000093000_G/A_A" "10_000093003_C/T_V" "10_000093004_A/G_
V/A" "10_000093007_T/A_E/V" ...
##  $ CDS.position             : chr  "1332" "1329" "1328" "1325" ...
##  $ Protein.position         : chr  "444" "443" "443" "442" ...
##  $ Amino.acids              : chr  "A" "V" "V/A" "E/V" ...
##  $ Codons                   : chr  "gcC/gcT" "gtG/gtA" "gTg/gCg" "gAg/gTg" ...
##  $ SYMBOL                   : chr  "TUBB8" "TUBB8" "TUBB8" "TUBB8" ...
##  $ SYMBOL.SOURCE            : chr  "HGNC" "HGNC" "HGNC" "HGNC" ...
##  $ SIFT                     : chr  "-" "-" "tolerated_low_confidence" "deleterious_low_confidenc
e" ...
##  $ SIFT.score               : num  NA NA 0.62 0 0.15 0.05 NA NA 0.69 0.6 ...
##  $ PolyPhen                 : chr  "-" "-" "benign" "benign" ...
##  $ PolyPhen.score           : num  NA NA 0 0.013 0.557 0.305 NA NA 0.001 0.001 ...
##  $ DOMAINS                  : chr  "-" "Coiled-coils_(Ncoils):Coil" "Coiled-coils_(Ncoils):Coil"
"Coiled-coils_(Ncoils):Coil,Low_complexity_(Seg):seg" ...
##  $ Source                   : chr  "['WGS', 'WES']" "['WES']" "['WES']" "['WES']" ...
##  $ AC                       : num  10 1 2 1 1 1 2 1 1 11 ...
##  $ AN                       : num  194416 171722 174636 180624 194710 ...
##  $ nhomalt                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AF                       : num  5.14e-05 5.82e-06 1.15e-05 5.54e-06 5.14e-06 ...
##  $ nhomalt.x2               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ nhetalt                  : num  10 1 2 1 1 1 2 1 1 11 ...
##  $ ratio.nhomalt.over.nhetalt: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ keyAA                    : chr  "A" "V" "V/A" "E/V" ...
##  $ CONSEQ                   : chr  "synonymous_variant" "synonymous_variant" "missense_variant"
"missense_variant" ...
##  $ clinvarAA                : chr  NA NA NA NA ...
##  $ clinvarGeneSymbol        : chr  NA NA NA NA ...
##  $ clinvarCONSEQ            : chr  NA NA NA NA ...
##  $ HGVSp.VEP                : chr  NA NA NA NA ...
##  $ HGVSc.VEP                : chr  NA NA NA NA ...
##  $ StarReviewStatus         : chr  NA NA NA NA ...
##  $ myClinVarLabels          : chr  NA NA NA NA ...
##  $ LABEL                    : chr  NA NA NA NA ...
```

```
# used colors
unusedcolors = c("skyblue", "#FD6467","#F4B5BD","#3B9AB2" ,
                     "#DD8D29", "#E2D200", "#46ACC8",
                    "#7294D4", "#C6CDF7","#FD6467", "#5B1A18",
                    "#F2AD00","#90D4CC","#FD6467","#00A08A",
                    "#FF0000", "#08519c",
                "red2", "orange2", "pink1")



maf_colors = c("#C6CDF7", "plum", "purple3")

maf2_colors = c("#C6CDF7", "purple3")

var_colors = c( "blue", "#85D4E3", "green4")

class_colors =  c( "#DD8D29", "#E2D200", "#46ACC8")

goflof_colors = c( "#FF0000", "#08519c" ,"#D9D0D3")
```

```
wgs %>%
  dplyr::group_by(Database) %>%
  tally()
```

```
## # A tibble: 3 × 2
##   Database               n
##   <chr>              <int>
## 1 ClinVar           719911
## 2 gnomAD           7559995
## 3 gnomAD & ClinVar  830683
```

```
wgs %>%
  dplyr::group_by(Database, ProteinConsequence) %>%
  tally()
```

```
## # A tibble: 9 × 3
## # Groups:   Database [3]
##   Database         ProteinConsequence       n
##   <chr>            <fct>                <int>
## 1 ClinVar          stop_gained          41200
## 2 ClinVar          missense_variant    489791
## 3 ClinVar          synonymous_variant  188920
## 4 gnomAD           stop_gained         153147
## 5 gnomAD           missense_variant   4966854
## 6 gnomAD           synonymous_variant 2439994
## 7 gnomAD & ClinVar stop_gained          11958
## 8 gnomAD & ClinVar missense_variant    561928
## 9 gnomAD & ClinVar synonymous_variant  256797
```

```
wgs %>%
  dplyr::group_by(myClinVarLabels) %>%
  tally()
```

```
## # A tibble: 4 × 2
##   myClinVarLabels       n
##   <fct>             <int>
## 1 PATHO             97718
## 2 VUS              969721
## 3 BENIGN           483155
## 4 <NA>            7559995
```

```
wgs %>%
  dplyr::group_by(MAF) %>%
  tally()
```

```
## # A tibble: 4 × 2
##   MAF                       n
##   <fct>                 <int>
## 1 Rare (<1%)          8317499
## 2 Less common (<10%)    38139
## 3 Common (>=10%)        35040
## 4 <NA>                 719911
```

```
wgs %>%
  dplyr::group_by(MAF2) %>%
  tally()
```

```
## # A tibble: 3 × 2
##   MAF2                n
##   <fct>           <int>
## 1 Rare (<5%)    8346414
## 2 Common (>=5%)   44264
## 3 <NA>           719911
```

```
wgs %>%
  dplyr::group_by(gnomadCONSEQ) %>%
  tally()
```

```
## # A tibble: 4 × 2
##   gnomadCONSEQ             n
##   <fct>                <int>
## 1 stop_gained         165105
## 2 missense_variant   5528782
## 3 synonymous_variant 2696791
## 4 <NA>                719911
```

```
wgs %>%
  dplyr::group_by(clinvarCONSEQ) %>%
  tally()
```

```
## # A tibble: 4 × 2
##   clinvarCONSEQ            n
##   <fct>                <int>
## 1 stop_gained          53158
## 2 missense_variant   1051719
## 3 synonymous_variant  445717
## 4 <NA>               7559995
```

```
total_gnomad <- wgs %>%
  dplyr::group_by(gnomadCONSEQ) %>%
  dplyr::tally() %>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
names(total_gnomad) <- c("ProteinConsequence", "n", "Percent")
total_gnomad$Database <- "gnomAD"
sum(total_gnomad$n)
```

```
## [1] 8390678
```

```
SUM.gnomad =  "8,390,678"

total_cv <- clinvar %>%
  dplyr::group_by(clinvarCONSEQ)%>%
  dplyr::tally()%>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
names(total_cv) <- c("ProteinConsequence", "n", "Percent")
total_cv$Database <- "ClinVar"
sum(total_cv$n)
```

```
## [1] 1550594
```

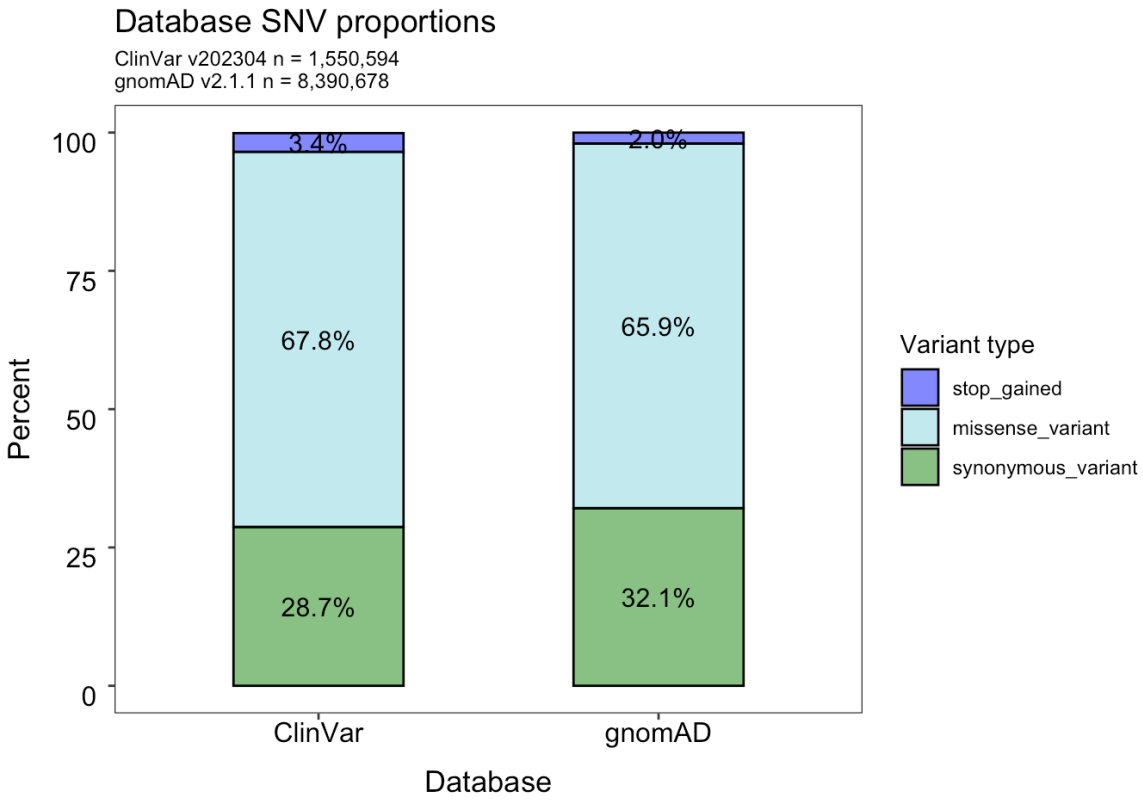
```
SUM.cv =  "1,550,594"

total_both <- bind_rows(total_cv, total_gnomad)
total_both
```

```
## # A tibble: 6 × 4
##   ProteinConsequence       n Percent Database
##   <fct>                <int>   <dbl> <chr>
## 1 stop_gained          53158     3.4 ClinVar
## 2 missense_variant   1051719    67.8 ClinVar
## 3 synonymous_variant  445717    28.7 ClinVar
## 4 stop_gained         165105     2   gnomAD
## 5 missense_variant   5528782    65.9 gnomAD
## 6 synonymous_variant 2696791    32.1 gnomAD
```

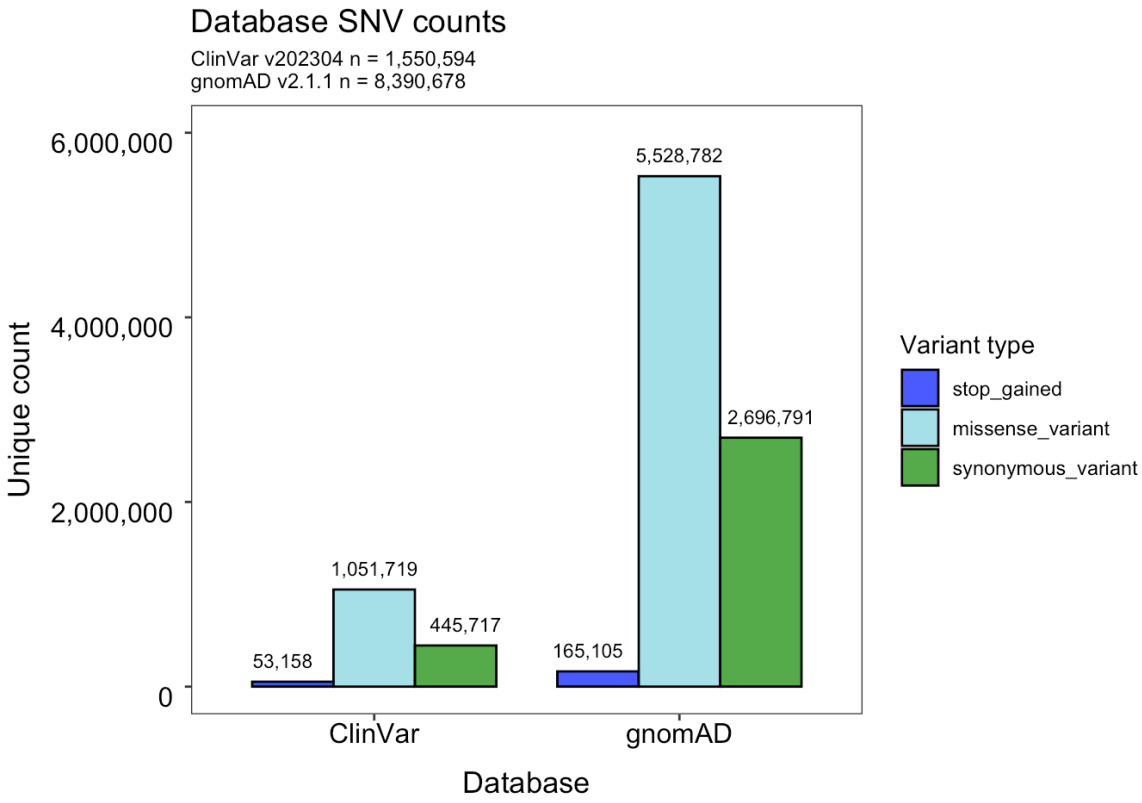
```
SUB1 = paste0("ClinVar v202304 n = ", SUM.cv)
SUB2 = paste0("gnomAD v2.1.1 n = ", SUM.gnomad)
SUB.total = paste(SUB1, SUB2, sep = "\n")

DBname <- "Exclusive ClinVar SNV n = 719,911\nExclusive gnomAD SNV n = 7,559,995\nOverlap gnomAD & C
linVar SNV n = 830,683"

SUB0 = "Overlap gnomAD & ClinVar SNV n = 830,683"
```

Database total SNV counts and %

```
stack_both <- ggplot(total_both, aes(x=Database, y=Percent, fill=ProteinConsequence)) +
  geom_bar(stat="identity", width=0.5, colour="black", alpha=0.6) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_stack(vjust=0.5),
            colour="black",
            size = 4,
            check_overlap = TRUE) +
  scale_fill_manual(values = var_colors) +
  labs(fill='Variant type') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Database",
       y= "Percent",
       title="Database SNV proportions",
       subtitle = SUB.total
         ) +
  theme(legend.position='right') +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
stack_both
```

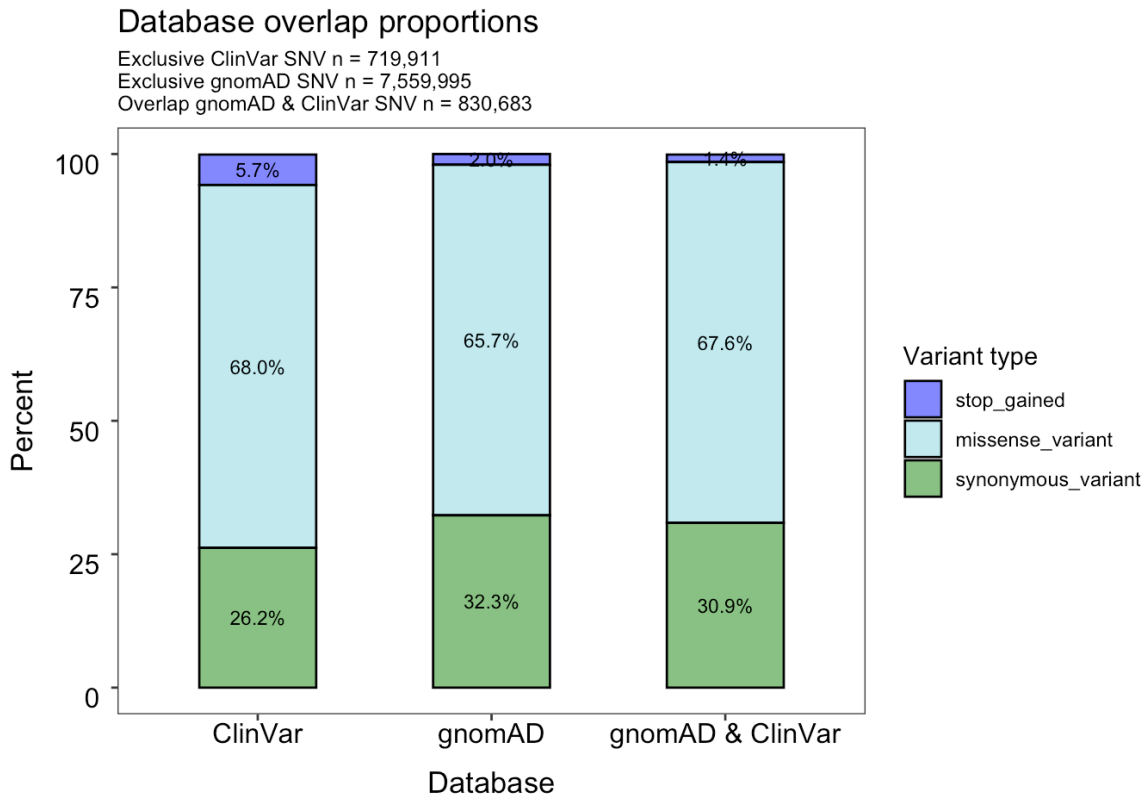


```
ggsave("Database_barstack_percent.png", width = 8, height = 6, dpi=300)



dodge_both <- ggplot(total_both, aes(x=Database, y=n, fill=ProteinConsequence)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.9),
            vjust=-1,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,6000000)) +
  scale_fill_manual(values = var_colors) +
  labs(fill='Variant type') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Database",
       y= "Unique count",
       title="Database SNV counts",
      subtitle = SUB.total
         ) +
  theme(legend.position='right') +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
dodge_both
```

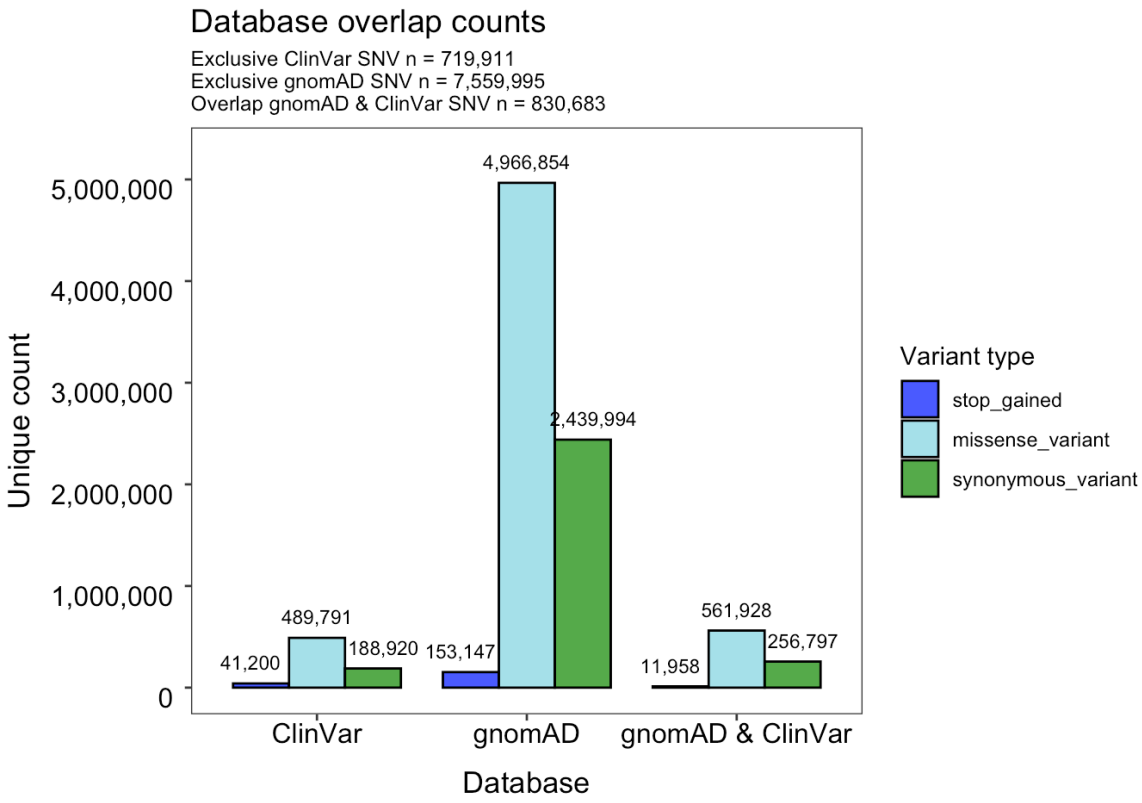



```
ggsave("Database_bardodge_count.png", width = 8.5, height = 6, dpi=300)
```

```
# database source n and percent ########################
total_db <- wgs %>%
  dplyr::group_by(Database, ProteinConsequence) %>%
  dplyr::tally() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))

stack_total_db <- ggplot(total_db, aes(x=Database, y=Percent, fill=ProteinConsequence)) +
  geom_bar(stat="identity", width=0.5, colour="black", alpha=0.6) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_stack(vjust=0.5),
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_fill_manual(values = var_colors) +
  labs(fill='Variant type') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Database",
       y= "Percent",
       title="Database overlap proportions",
       subtitle = DBname) +
  theme(legend.position='right') +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
stack_total_db
```


Database overlap proportions
Exclusive ClinVar SNV n = 719,911
Exclusive gnomAD SNV n = 7,559,995
Overlap gnomAD & ClinVar SNV n = 830,683
Percent
100
75
50
25
0
5.7%
68.0%
26.2%
2.0%
65.7%
32.3%
1.4%
67.6%
30.9%
ClinVar
gnomAD
gnomAD & ClinVar
Database
Variant type
stop_gained
missense_variant
synonymous_variant

```
ggsave("DatabaseExclusive_barstack_percent.png", width = 8, height = 6, dpi=300)

dodge_total_db <- ggplot(total_db, aes(x=Database, y=n, fill=ProteinConsequence)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.95),
            vjust=-1,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,5250000)) +
  scale_fill_manual(values = var_colors) +
  labs(fill='Variant type') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Database",
       y= "Unique count",
       title="Database overlap counts",
       subtitle = DBname) +
  theme(legend.position='right') +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
dodge_total_db
```



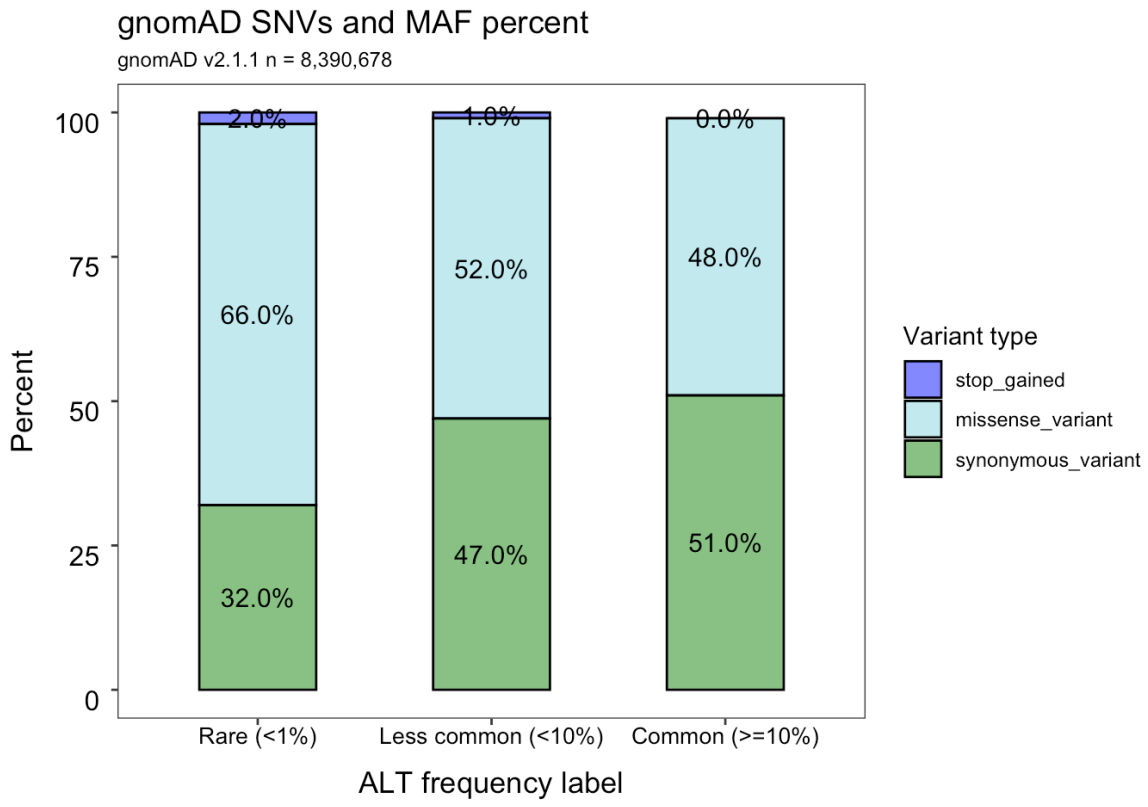

```
ggsave("DatabaseExclusive_bardodge_count.png", width = 8.5, height = 6, dpi=300)
```

Database MAF and conseq

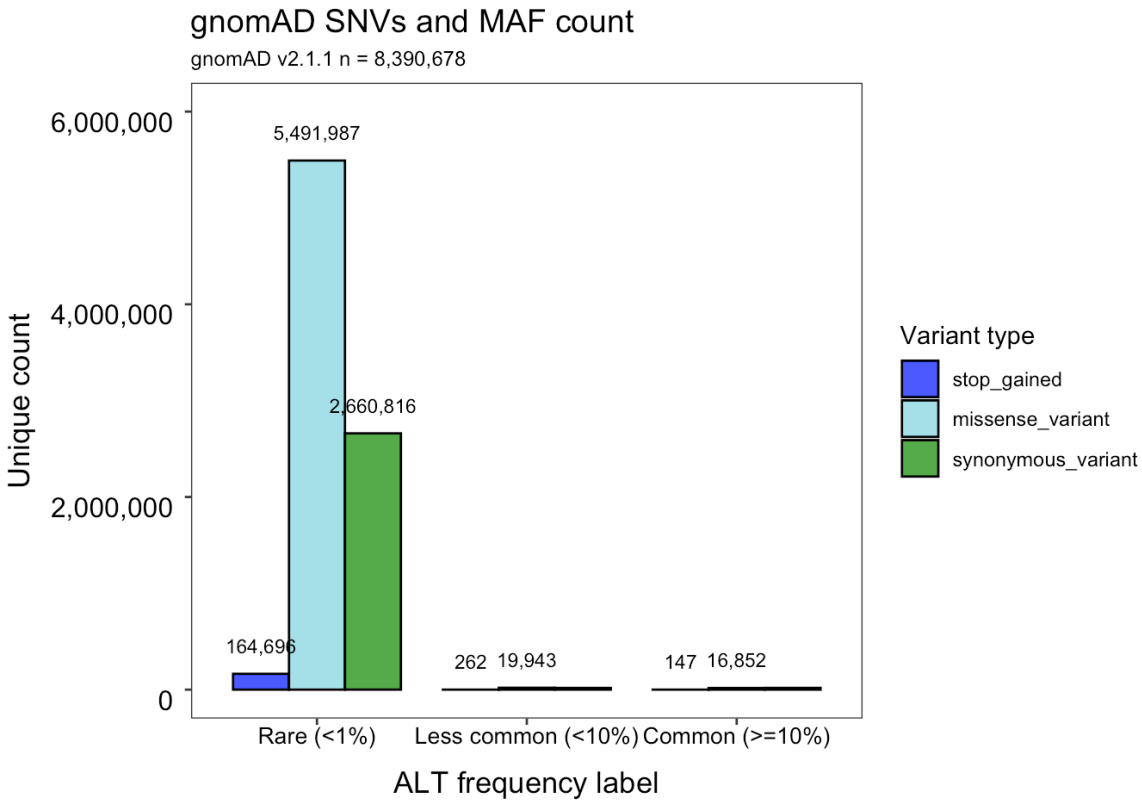
```
# gnomAD - x = MAF, y = n | Percent, fill = variant type ############################
gnomad_maf_conseq <- wgs %>%
  dplyr::group_by(MAF, gnomadCONSEQ) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100))
print(gnomad_maf_conseq)
```

```
## # A tibble: 9 × 4
## # Groups:   MAF [3]
##   MAF                gnomadCONSEQ            n Percent
##   <fct>              <fct>               <int>   <dbl>
## 1 Rare (<1%)         stop_gained         164696       2
## 2 Rare (<1%)         missense_variant   5491987      66
## 3 Rare (<1%)         synonymous_variant 2660816      32
## 4 Less common (<10%) stop_gained            262       1
## 5 Less common (<10%) missense_variant     19943      52
## 6 Less common (<10%) synonymous_variant   17934      47
## 7 Common (>=10%)     stop_gained            147       0
## 8 Common (>=10%)     missense_variant     16852      48
## 9 Common (>=10%)     synonymous_variant   18041      51
```

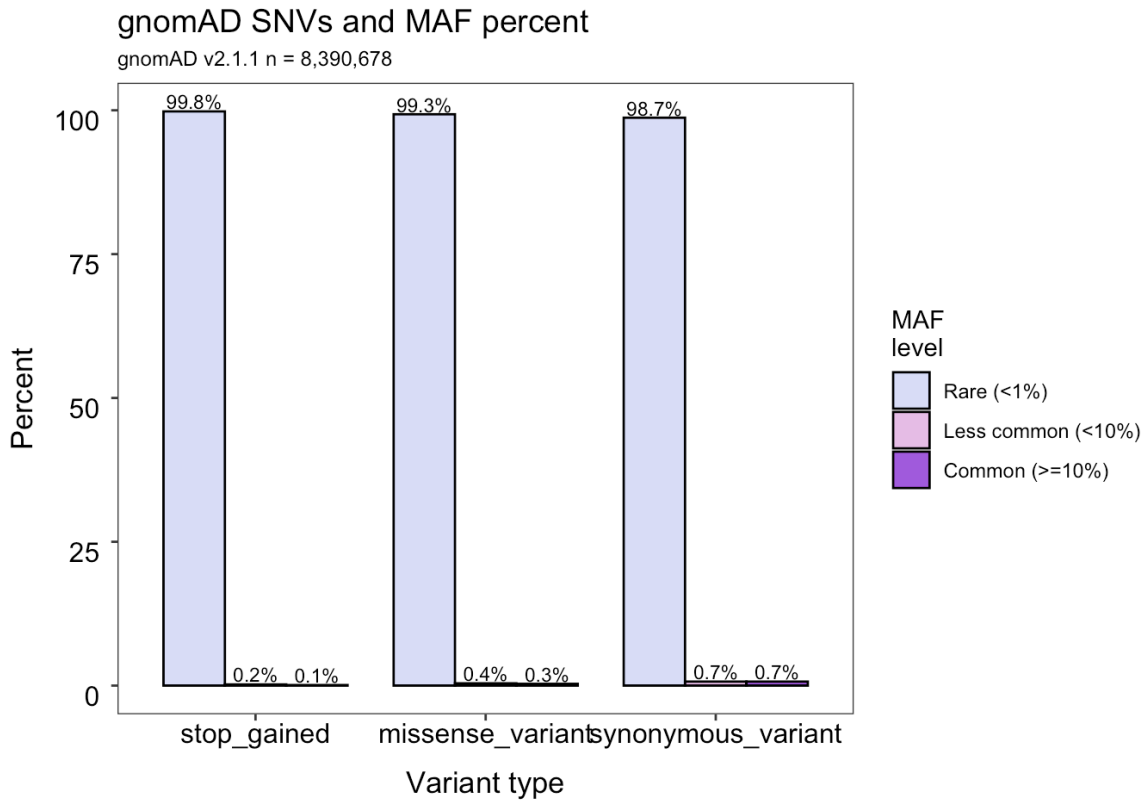
```
gnomad_maf_type_bar2 <- ggplot(gnomad_maf_conseq, aes(x=MAF,y=Percent, fill=gnomadCONSEQ)) +
  geom_bar(stat="identity", width=0.5, colour="black", alpha=0.6) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_stack(vjust=0.5),
            colour="black",
            size = 4,
            check_overlap = TRUE) +
  scale_fill_manual(values = var_colors) +
  labs(x="ALT frequency label",
       y= "Percent",
       fill='Variant type',
      title="gnomAD SNVs and MAF percent",
       subtitle = SUB2) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf_type_bar2
```



```
ggsave("gnomadDB_barstack_percent.png", width=8.5, height = 6, dpi=300)

gnomad_maf_type_bar <- ggplot(gnomad_maf_conseq, aes(x=MAF,y=n, fill=gnomadCONSEQ)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
           angle=0,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,6000000)) +
  scale_fill_manual(values = var_colors) +
  labs(x="ALT frequency label",
       y= "Unique count",
       fill='Variant type',
      title="gnomAD SNVs and MAF count",
       subtitle = SUB2) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
              axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5,
r=1, l=1)),
              axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r
=5, b = 5)),
              axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
              plot.title = element_text(size=14),
              plot.subtitle = element_text(size=9)
        )

gnomad_maf_type_bar
```

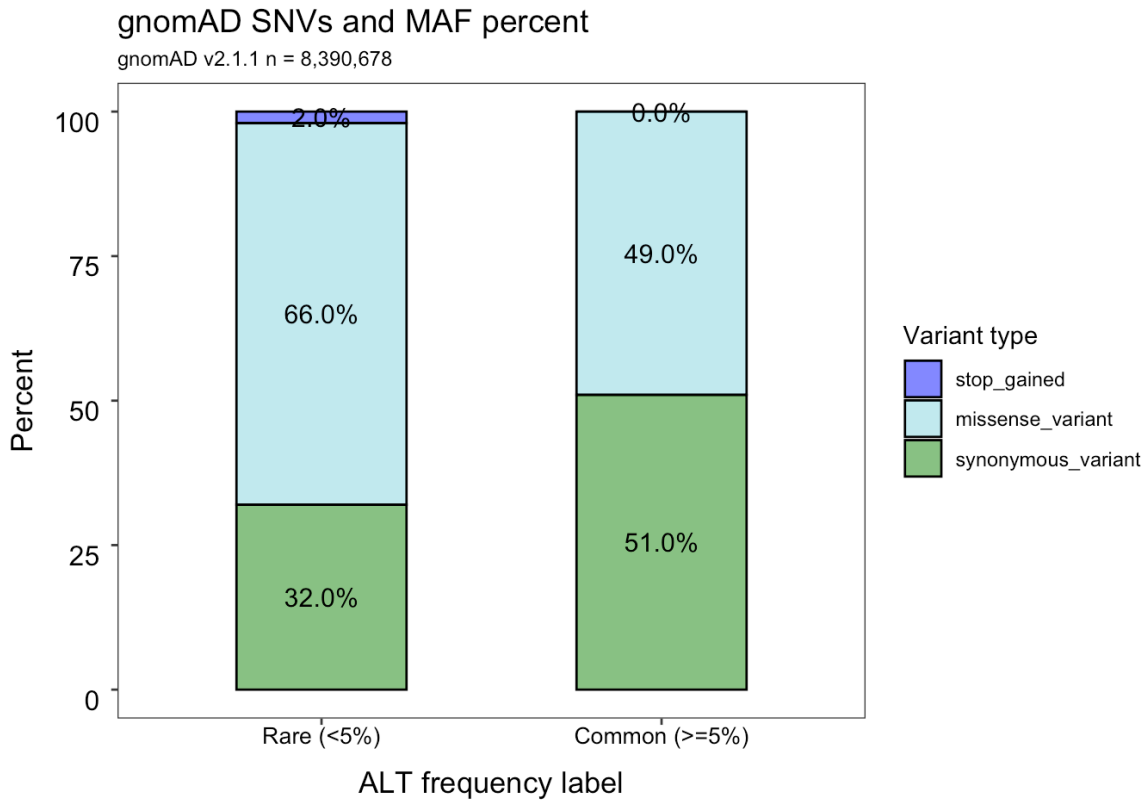

gnomAD SNVs and MAF count
gnomAD v2.1.1 n = 8,390,678
Unique count
6,000,000
4,000,000
2,000,000
0
5,491,987
2,660,816
164,696
262
19,943
147
16,852
Rare (<1%)
Less common (<10%)
Common (>=10%)
ALT frequency label
Variant type
stop_gained
missense_variant
synonymous_variant

```
ggsave("gnomadDB_bardodge_count.png", width=8.5, height = 6, dpi=300)


gnomad_maf <- wgs %>%
  dplyr::group_by(gnomadCONSEQ, MAF) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))

gnomad_maf_nonstack <- ggplot(gnomad_maf, aes(x=gnomadCONSEQ, y=Percent, fill=MAF)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-0.2,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_fill_manual(values = maf_colors) +
  labs(fill='MAF\nlevel') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Variant type",
       y= "Percent",
       title="gnomAD SNVs and MAF percent",
       subtitle = SUB2) +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf_nonstack
```

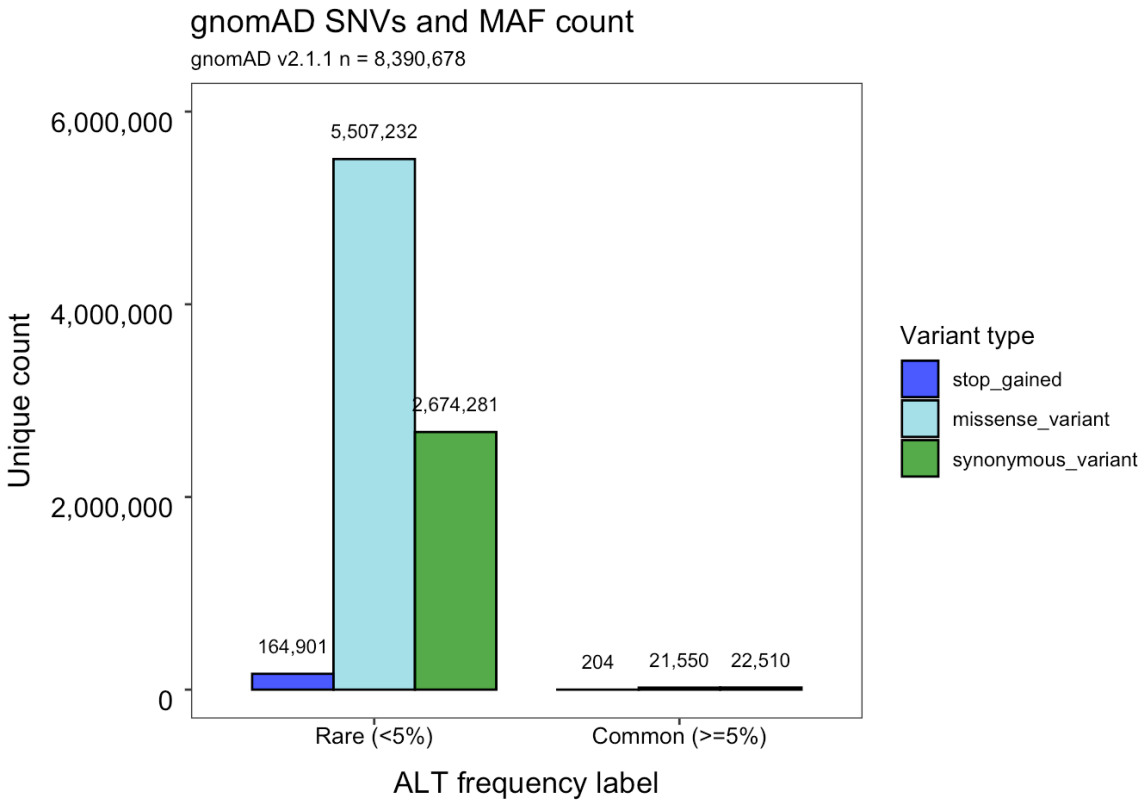



```
ggsave("gnomadDB_barstack_percent_MAFfillType.png", width=8.5, height = 6, dpi=300)
```

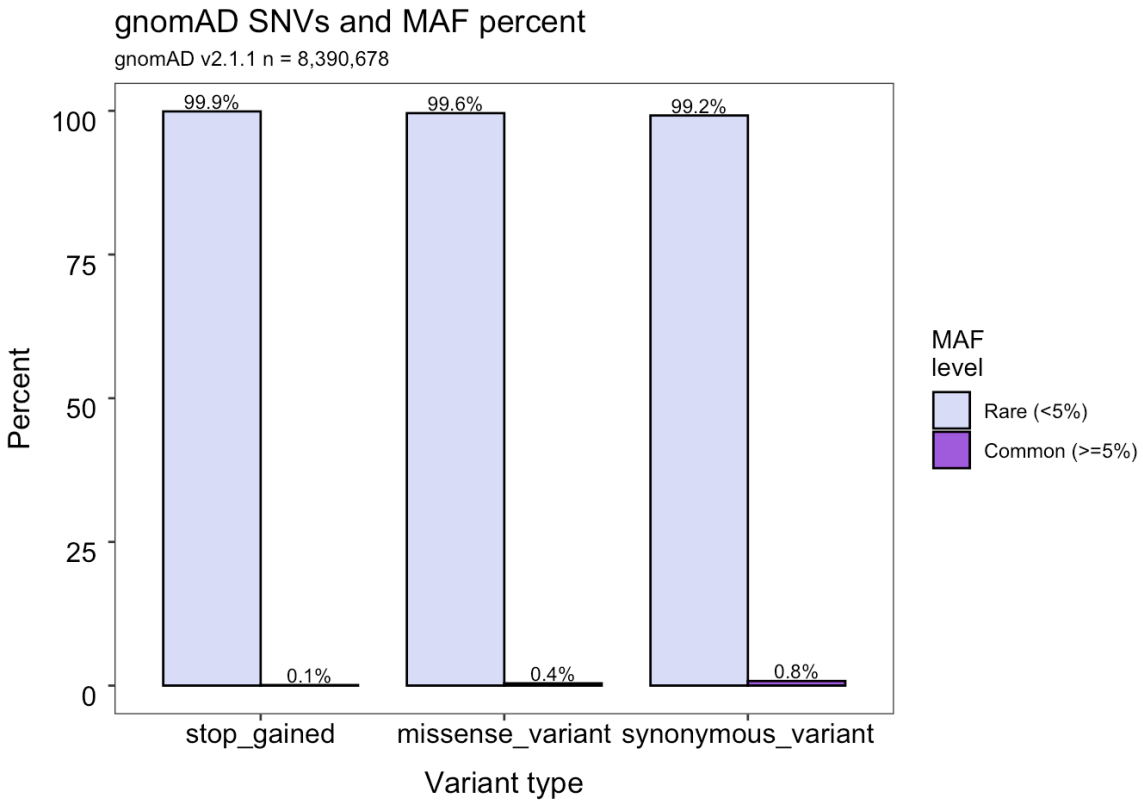
```
gnomad_maf2_conseq <- wgs %>%
  dplyr::group_by(MAF2, gnomadCONSEQ) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100))

gnomad_maf_type_bar2 <- ggplot(gnomad_maf2_conseq, aes(x=MAF2,y=Percent, fill=gnomadCONSEQ)) +
  geom_bar(stat="identity", width=0.5, colour="black", alpha=0.6) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_stack(vjust=0.5),
            colour="black",
            size = 4,
            check_overlap = TRUE) +
  scale_fill_manual(values = var_colors) +
  labs(x="ALT frequency label",
       y= "Percent",
       fill='Variant type',
      title="gnomAD SNVs and MAF percent",
       subtitle = SUB2) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf_type_bar2
```


gnomAD SNVs and MAF percent
gnomAD v2.1.1 n = 8,390,678
100
75
50
25
0
Percent
2.0%
66.0%
32.0%
0.0%
49.0%
51.0%
Rare (<5%)
Common (>=5%)
ALT frequency label
Variant type
stop_gained
missense_variant
synonymous_variant

```
ggsave("gnomadDB_barstack_percent_MAF2.png", width=8.5, height = 6, dpi=300)

gnomad_maf_type_bar <- ggplot(gnomad_maf2_conseq, aes(x=MAF2,y=n, fill=gnomadCONSEQ)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
           angle=0,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,6000000)) +
  scale_fill_manual(values = var_colors) +
  labs(x="ALT frequency label",
       y= "Unique count",
       fill='Variant type',
      title="gnomAD SNVs and MAF count",
       subtitle = SUB2) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
              axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5,
r=1, l=1)),
              axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r
=5, b = 5)),
              axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
              plot.title = element_text(size=14),
              plot.subtitle = element_text(size=9)
        )

gnomad_maf_type_bar
```


gnomAD SNVs and MAF count
gnomAD v2.1.1 n = 8,390,678
Unique count
6,000,000
4,000,000
2,000,000
0
5,507,232
2,674,281
164,901
204
21,550
22,510
Rare (<5%)
Common (>=5%)
ALT frequency label
Variant type
stop_gained
missense_variant
synonymous_variant

```
ggsave("gnomadDB_bardodge_count_MAF2.png", width=8, height = 6, dpi=300)




gnomad_maf2 <- wgs %>%
  dplyr::group_by(gnomadCONSEQ, MAF2) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))

gnomad_maf2_nonstack <- ggplot(gnomad_maf2, aes(x=gnomadCONSEQ, y=Percent, fill=MAF2)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-0.2,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_fill_manual(values = maf2_colors) +
  labs(fill='MAF\nlevel') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Variant type",
       y= "Percent",
      title="gnomAD SNVs and MAF percent",
       subtitle = SUB2) +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf2_nonstack
```



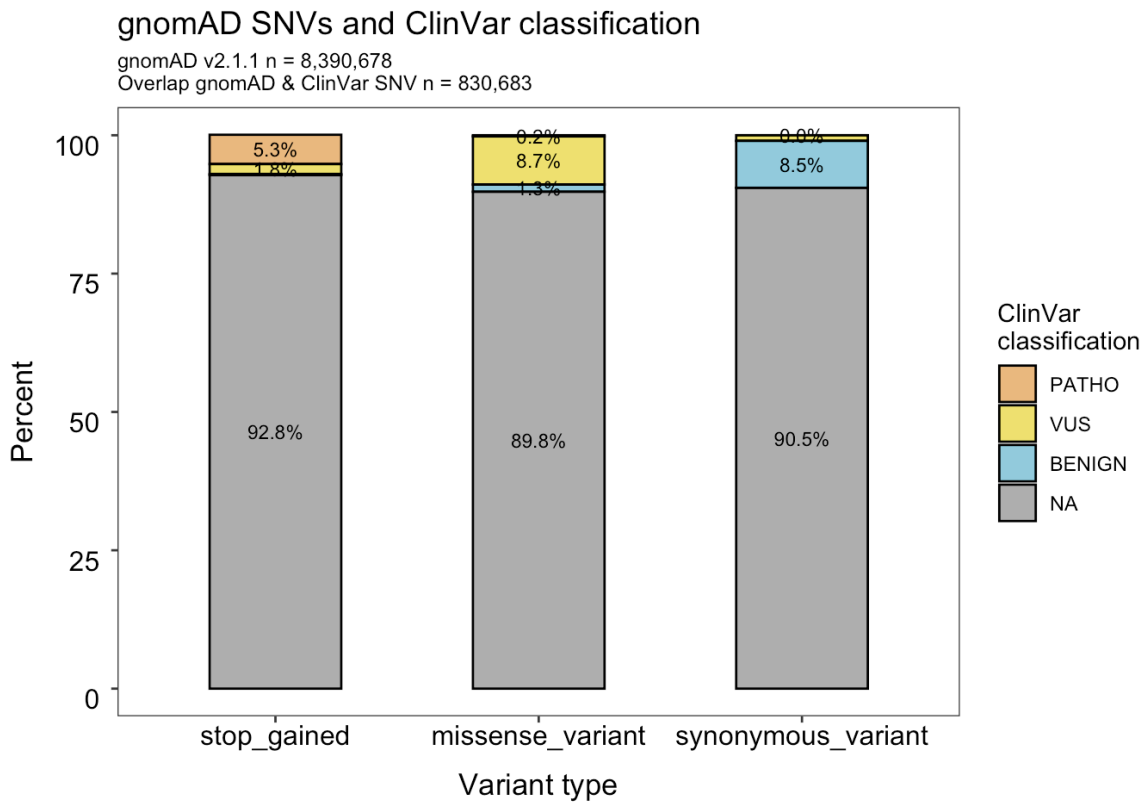

```
ggsave("gnomadDB_barstack_percent_MAF2filltype.png", width=8.5, height = 6, dpi=300)
```

Database OVERLAP MAF/classification and conseq

```
gnomad_class <- wgs %>%
  dplyr::group_by(gnomadCONSEQ, myClinVarLabels) %>%
  dplyr::tally()%>%
  drop_na(gnomadCONSEQ) %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
print(gnomad_class)
```

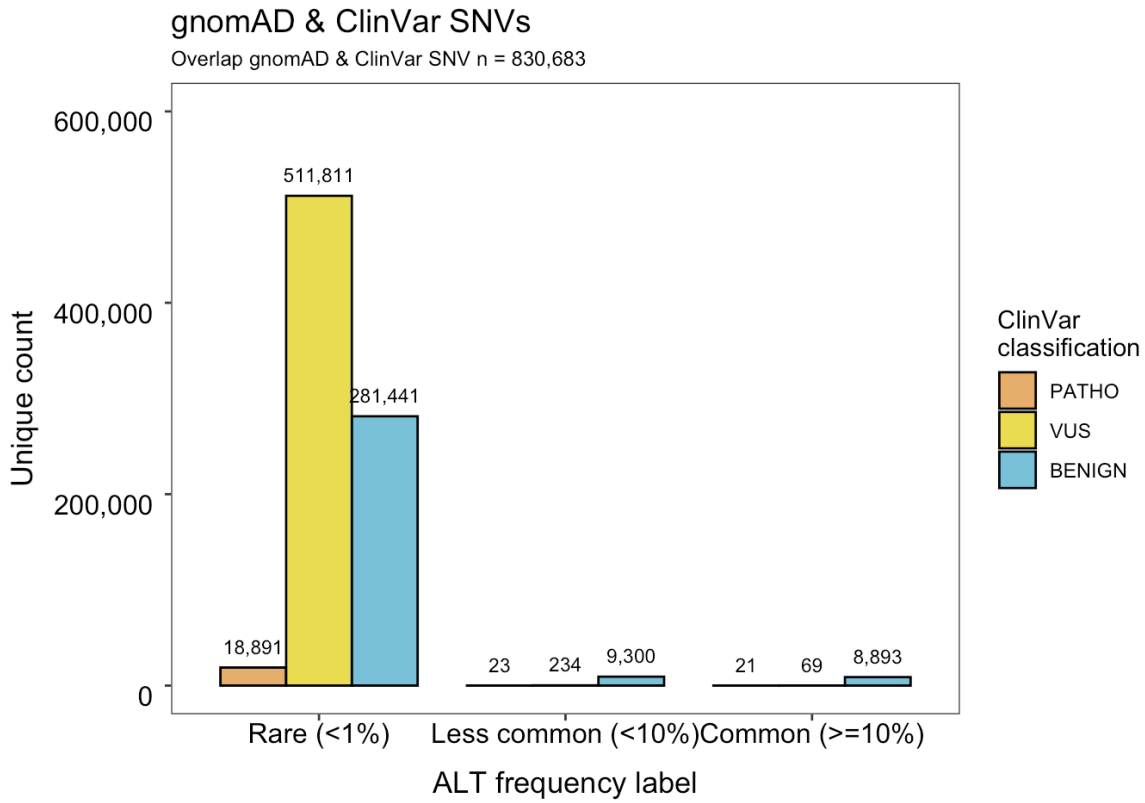
```
## # A tibble: 12 × 4
## # Groups:   gnomadCONSEQ [3]
##    gnomadCONSEQ       myClinVarLabels       n Percent
##    <fct>              <fct>             <int>   <dbl>
##  1 stop_gained        PATHO              8756     5.3
##  2 stop_gained        VUS                2904     1.8
##  3 stop_gained        BENIGN              298     0.2
##  4 stop_gained        <NA>             153147    92.8
##  5 missense_variant   PATHO             10103     0.2
##  6 missense_variant   VUS              482426     8.7
##  7 missense_variant   BENIGN            69399     1.3
##  8 missense_variant   <NA>            4966854    89.8
##  9 synonymous_variant PATHO               76     0
## 10 synonymous_variant VUS               26784     1
## 11 synonymous_variant BENIGN           229937     8.5
## 12 synonymous_variant <NA>            2439994    90.5
```

```
gnomad_stack_class <- ggplot(gnomad_class, aes(x=gnomadCONSEQ, y=Percent, fill=myClinVarLabels)) +
  geom_bar(stat="identity", width=0.5, colour="black", alpha=0.7) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_stack(vjust=0.5),
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_fill_manual(values = class_colors) +
  labs(fill='ClinVar\nclassification') +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  labs(x="Variant type",
       y= "Percent",
       title="gnomAD SNVs and ClinVar classification",
       subtitle = paste(SUB2, SUB0, sep="\n")) +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_stack_class
```

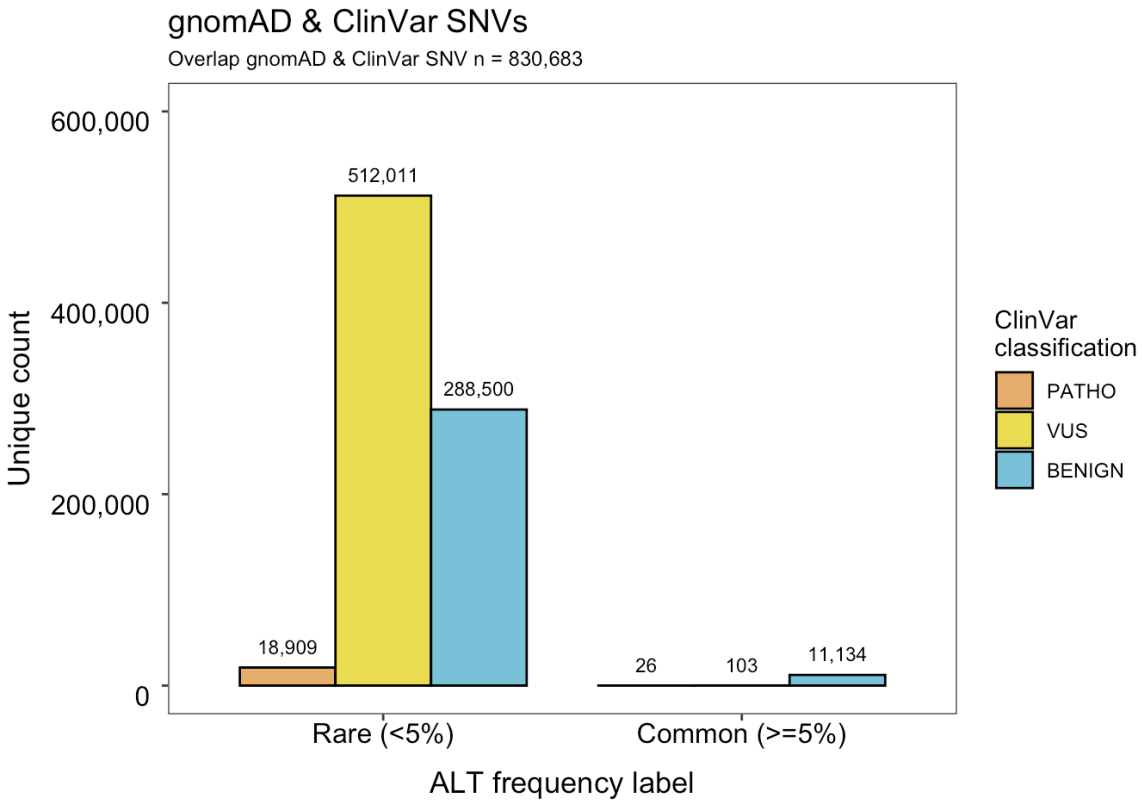



```
ggsave("gnomadDB_barstack_percent_CVCLASSfillType.png", width=8.5, height = 6, dpi=300)
```

```
# gnomAD & clinvar variants- x = MAF, y = counts, fill = class ############################
gnomad_maf_class <- wgs %>%
  dplyr::group_by(MAF, myClinVarLabels) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(percent=n/sum(n))

gnomad_maf_class_bar <- ggplot(gnomad_maf_class, aes(x=MAF,y=n, fill=myClinVarLabels)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,600000)) +
  scale_fill_manual(values = class_colors) +
  labs(x ="ALT frequency label",
       y = "Unique count",
       fill ='ClinVar\nclassification',
       title="gnomAD & ClinVar SNVs",
       subtitle = SUB0) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf_class_bar
```


gnomAD & ClinVar SNVs
Overlap gnomAD & ClinVar SNV n = 830,683
600,000
400,000
200,000
0
Unique count
511,811
281,441
18,891
23
234
9,300
21
69
8,893
Rare (<1%)
Less common (<10%)
Common (>=10%)
ALT frequency label
ClinVar classification
PATHO
VUS
BENIGN

```
ggsave("gnomadClinvarDB_bardodge_count_CVCLASSfillType.png", width=8, height = 6, dpi=300)


# gnomAD & clinvar variants- x = MAF2, y = counts, fill = class ############################
gnomad_maf2_class <- wgs %>%
  dplyr::group_by(MAF2, myClinVarLabels) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(percent=n/sum(n))

gnomad_maf2_class_bar <- ggplot(gnomad_maf2_class, aes(x=MAF2,y=n, fill=myClinVarLabels)) +
  geom_bar(position="dodge",stat="identity", width=0.8, colour="black", alpha=0.8, na.rm=TRUE) +
  geom_text(aes(label=  scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,600000)) +
  scale_fill_manual(values = class_colors) +
  labs(x ="ALT frequency label",
       y = "Unique count",
       fill ='ClinVar\nclassification',
      title="gnomAD & ClinVar SNVs",
       subtitle = SUB0) +
  theme_bw() +
  theme(legend.position='right',
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9)
  )
gnomad_maf2_class_bar
```



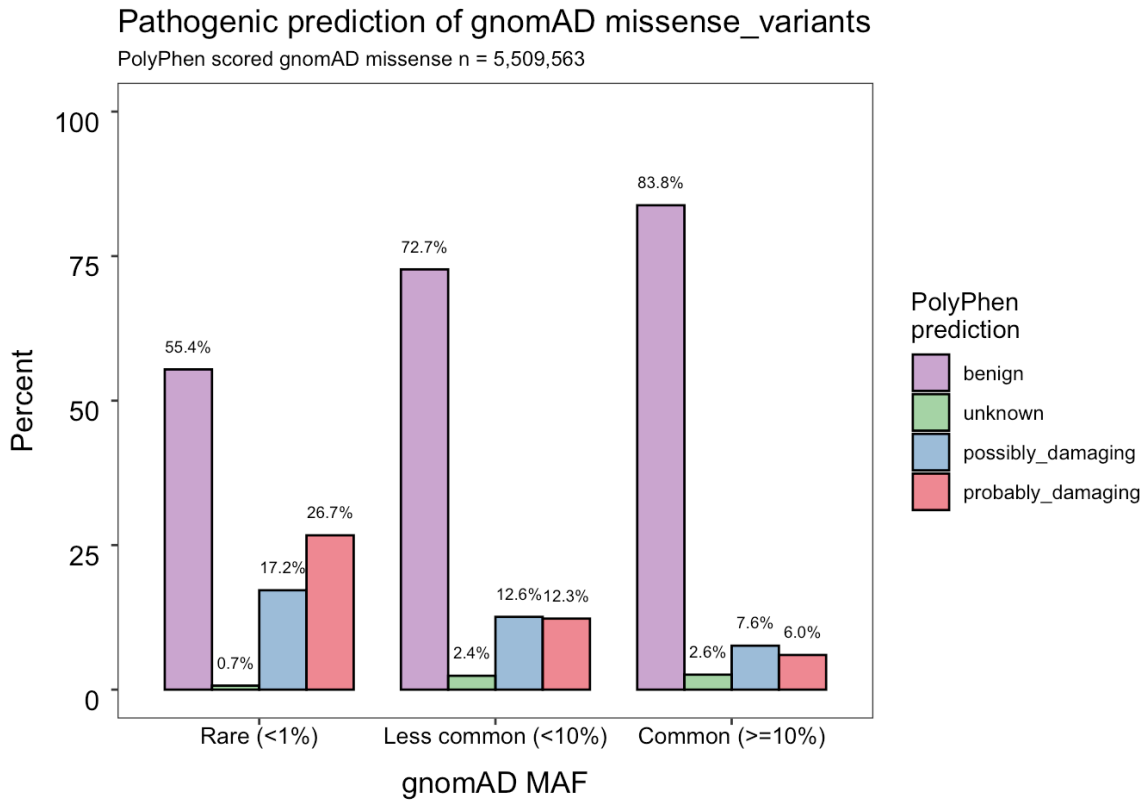

```
ggsave("gnomadClinvarDB_bardodge_count_CVCLASSfillType_MAF2.png", width=8, height = 6, dpi=300)
```

gnomAD missense predictions

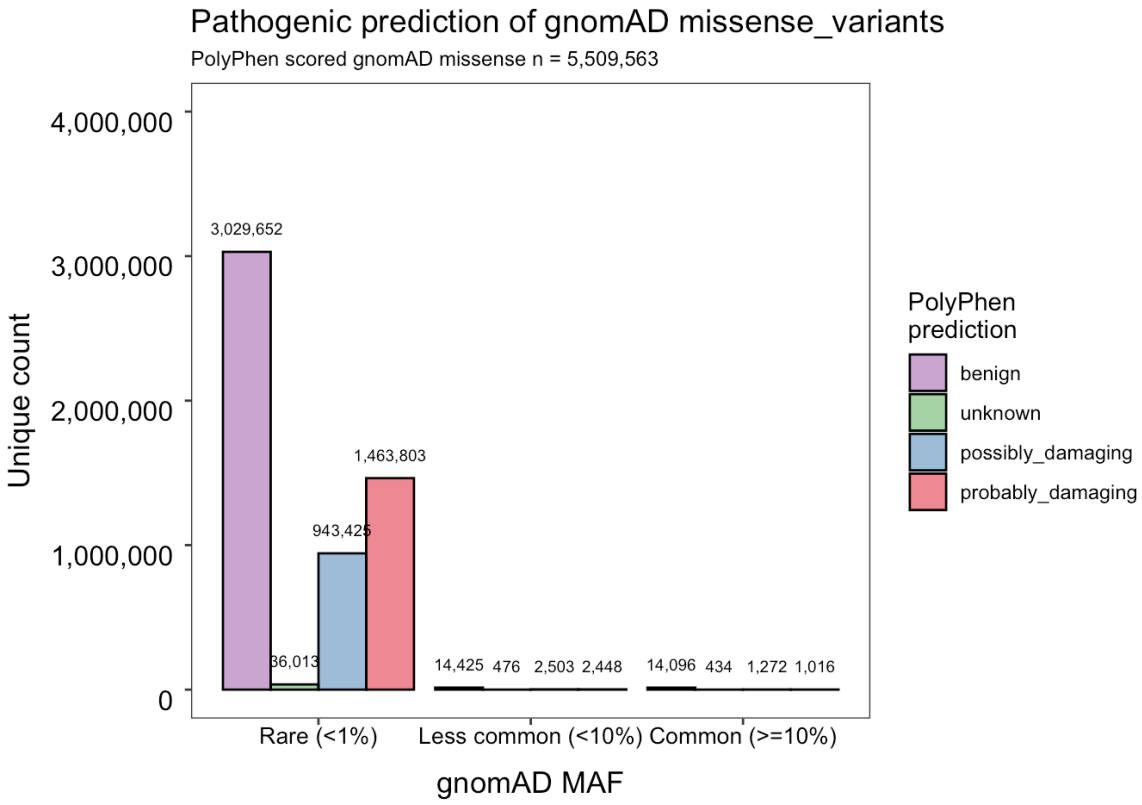
```
## polyphen gnonmad ######################
polyphen_gnomad <- wgs %>%
  dplyr::group_by(MAF, PolyPhen) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
sum(polyphen_gnomad$n) # 5,509,563
```

```
## [1] 5509563
```

```
## polyphen gnonmad plots ######################
polyphen_gnomad_percent <- ggplot(polyphen_gnomad, aes(x=MAF,
                                          y=Percent,
                                          fill=PolyPhen)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.6, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(limit=c(0,100)) +
   scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="gnomAD MAF",
       y= "Percent",
       fill='PolyPhen\nprediction',
       title="Pathogenic prediction of gnomAD missense_variants",
       subtitle = "PolyPhen scored gnomAD missense n = 5,509,563") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
polyphen_gnomad_percent
```


Pathogenic prediction of gnomAD missense_variants
PolyPhen scored gnomAD missense n = 5,509,563
Percent
100
75
50
25
0
55.4%
0.7%
17.2%
26.7%
72.7%
2.4%
12.6%
12.3%
83.8%
2.6%
7.6%
6.0%
Rare (<1%)
Less common (<10%)
Common (>=10%)
gnomAD MAF
PolyPhen prediction
benign
unknown
possibly_damaging
probably_damaging

```
ggsave("gnomadDB_bardodge_percent_PolyPhen.png", width = 8, height=6, dpi= 300)

polyphen_gnomad_plot2 <- ggplot(polyphen_gnomad, aes(x=MAF,
                                                     y=n,
                                                     fill=PolyPhen)) +
  geom_bar(position="dodge", stat="identity", width=0.9, colour="black", alpha=0.6, na.rm=TRUE) +
  geom_text(aes(label= scales::comma(n)),
            position=position_dodge2(width=0.9),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,4000000)) +
   scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="gnomAD MAF",
       y= "Unique count",
       fill='PolyPhen\nprediction',
       title="Pathogenic prediction of gnomAD missense_variants",
       subtitle = "PolyPhen scored gnomAD missense n = 5,509,563") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
polyphen_gnomad_plot2
```

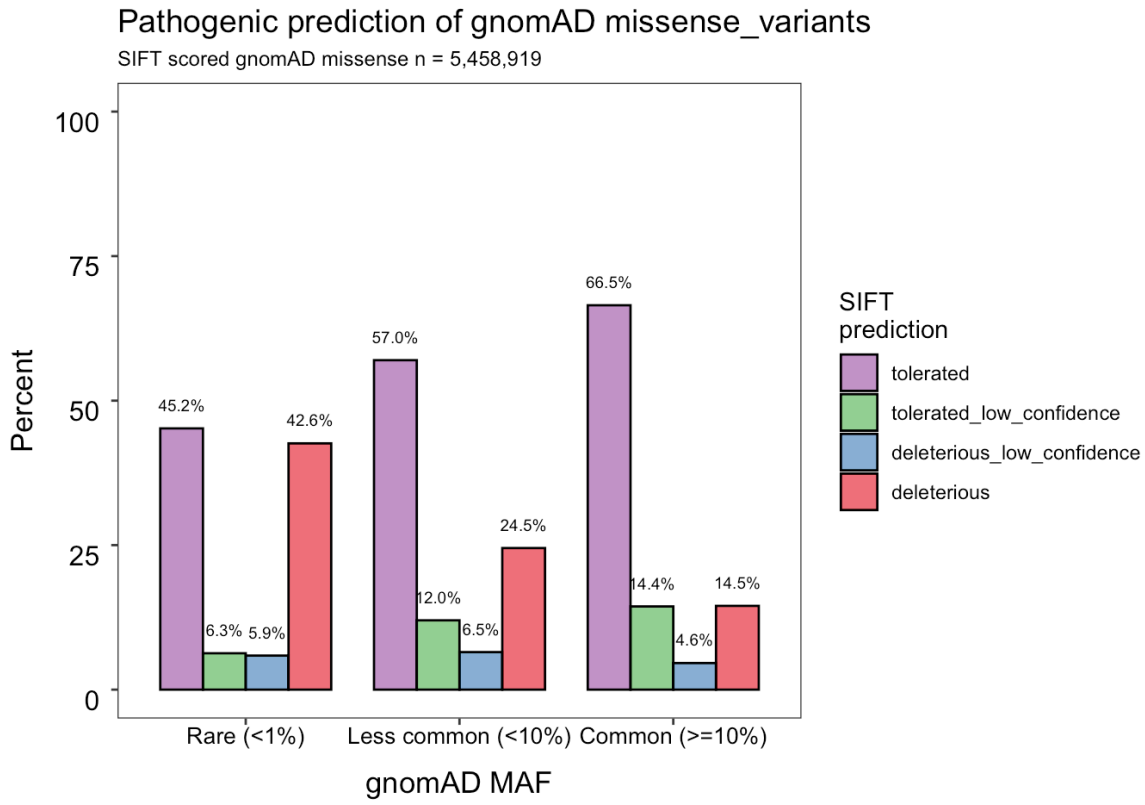



```
ggsave("gnomadDB_bardodge_count_PolyPhen.png", width = 8, height=6, dpi= 300)
```

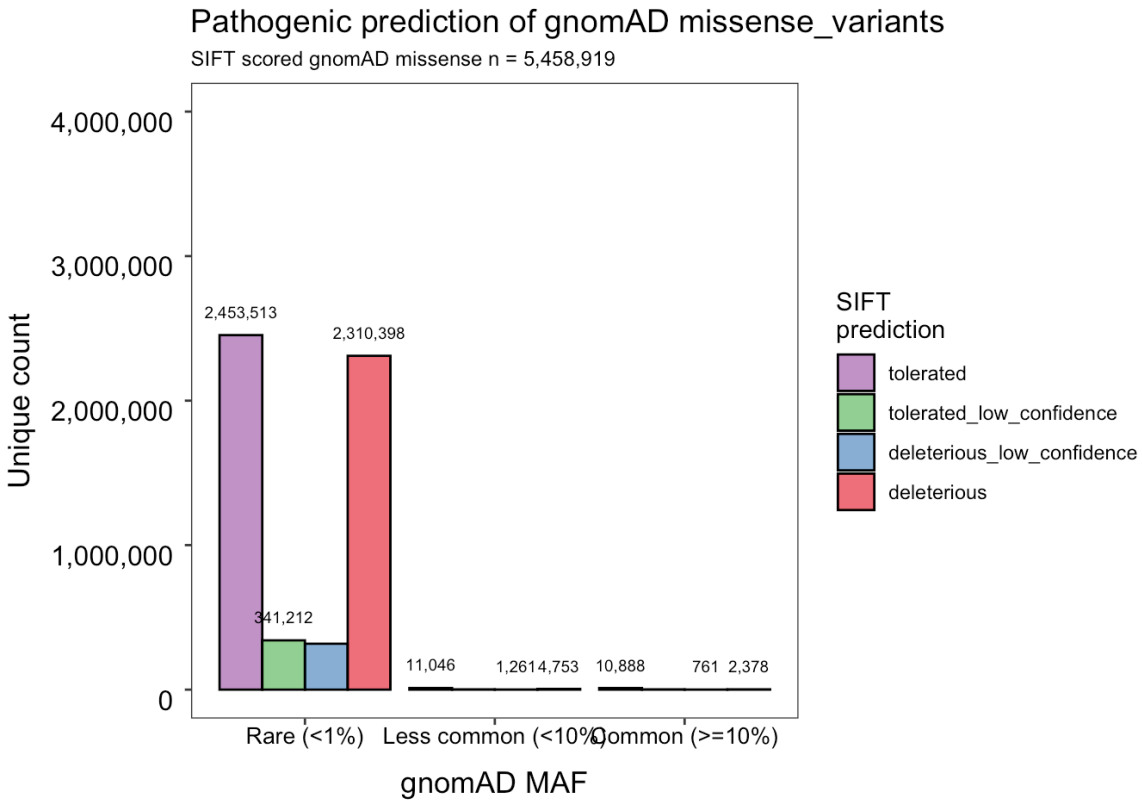
```
## sift gnomad ############################
sift_gnomad <- wgs %>%
  dplyr::group_by(MAF, SIFT) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
sum(sift_gnomad$n) # 5,458,919
```

```
## [1] 5458919
```

```
## sift gnomad plots ############################
sift_gnomad_percent <- ggplot(sift_gnomad, aes(x=MAF,
                                               y=Percent,
                                               fill=SIFT)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.7, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(limit=c(0,100)) +
  scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="gnomAD MAF",
       y= "Percent",
       fill='SIFT\nprediction',
       title="Pathogenic prediction of gnomAD missense_variants",
       subtitle = "SIFT scored gnomAD missense n = 5,458,919") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
sift_gnomad_percent
```


Pathogenic prediction of gnomAD missense_variants
SIFT scored gnomAD missense n = 5,458,919
Percent
100
75
50
25
0
45.2%
6.3%
5.9%
42.6%
57.0%
12.0%
6.5%
24.5%
66.5%
14.4%
4.6%
14.5%
Rare (<1%)
Less common (<10%)
Common (>=10%)
gnomAD MAF
SIFT prediction
tolerated
tolerated_low_confidence
deleterious_low_confidence
deleterious

```
ggsave("gnomadDB_bardodge_percent_SIFT.png", width = 8, height=6, dpi= 300)

sift_gnomad_plot2 <- ggplot(sift_gnomad, aes(x=MAF,
                                             y=n,
                                             fill=SIFT)) +
  geom_bar(position="dodge", stat="identity", width=0.9, colour="black", alpha=0.7, na.rm=TRUE) +
   geom_text(aes(label= scales::comma(n)),
             position=position_dodge2(width=0.9),
             vjust=-1.5,
             colour="black",
             angle=0,
             size = 2.5,
             check_overlap = TRUE) +
 scale_y_continuous(label=comma, limits = c(0,4000000)) +
  scale_fill_brewer(palette = "Set1", direction = -1) +
 labs(x="gnomAD MAF",
       y= "Unique count",
       fill='SIFT\nprediction',
       title="Pathogenic prediction of gnomAD missense_variants",
       subtitle = "SIFT scored gnomAD missense n = 5,458,919") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=10, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
sift_gnomad_plot2
```




```
ggsave("gnomadDB_bardodge_count_SIFT.png", width = 8, height=6, dpi= 300)
```

ClinVar missense predictions

```
## polyphen_clinvar

# ISSUE with missing score data for clinvar variants- onnly the clinvar variants overlapping gnomad variant keyID37aa have sift or polyphen scores because i only used the vep output fro the gnomad var iants before merging clinvar data from variant_summary.txt pipeline

# TODO get score data from vep output. of clinvar variants

path="/Users/mariapalafox/Desktop/BOLDcaution/7_vep_annotations/data/"

clinvar_scored = read.csv(paste0(path, "Rfigures_clinvar20230430_vep_annotated_META_merged_codingSNV s.tsv"),
                          sep="\t")

# order of factors
pp_order <- c("benign",  "unknown", "possibly_damaging",  "probably_damaging")
clinvar_scored[['PolyPhen']] <- factor(clinvar_scored[['PolyPhen']], levels = pp_order)

sift_order <- c("tolerated", "tolerated_low_confidence", "deleterious_low_confidence", "deleteriou s")
clinvar_scored[['SIFT']] <- factor(clinvar_scored[['SIFT']],levels = sift_order)

clinvar_scored[['myClinVarLabels']] <- factor(clinvar_scored[['myClinVarLabels']],
                                              levels = c("PATHO", "VUS", "BENIGN"))

clinvar_scored[['LABEL']] <- factor(clinvar_scored[['LABEL']], levels = c("LOF", "GOF"))

order_variants <- c( "stop_gained", "missense_variant", "synonymous_variant")

clinvar_scored[['clinvarCONSEQ']] <- factor(clinvar_scored[['clinvarCONSEQ']], levels = order_varian ts)

star <- clinvar_scored %>% filter(StarReviewStatus == "Star")

polyphen_clinvar_star <- star %>%
  dplyr::group_by(myClinVarLabels, PolyPhen) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
sum(polyphen_clinvar_star$n)
```

```
## [1] 946756
```

```
print(polyphen_clinvar_star)
```

```
## # A tibble: 12 × 4
## # Groups:   myClinVarLabels [3]
##    myClinVarLabels PolyPhen              n Percent
##    <fct>           <fct>             <int>   <dbl>
##  1 PATHO           benign             4388    11.2
##  2 PATHO           unknown               1     0
##  3 PATHO           possibly_damaging   5497    14
##  4 PATHO           probably_damaging  29280    74.8
##  5 VUS             benign            434351    52.4
##  6 VUS             unknown              417     0.1
##  7 VUS             possibly_damaging 146280    17.6
##  8 VUS             probably_damaging 248585    30
##  9 BENIGN          benign             57360    73.6
## 10 BENIGN          unknown               91     0.1
## 11 BENIGN          possibly_damaging   9406    12.1
## 12 BENIGN          probably_damaging  11100    14.2
```

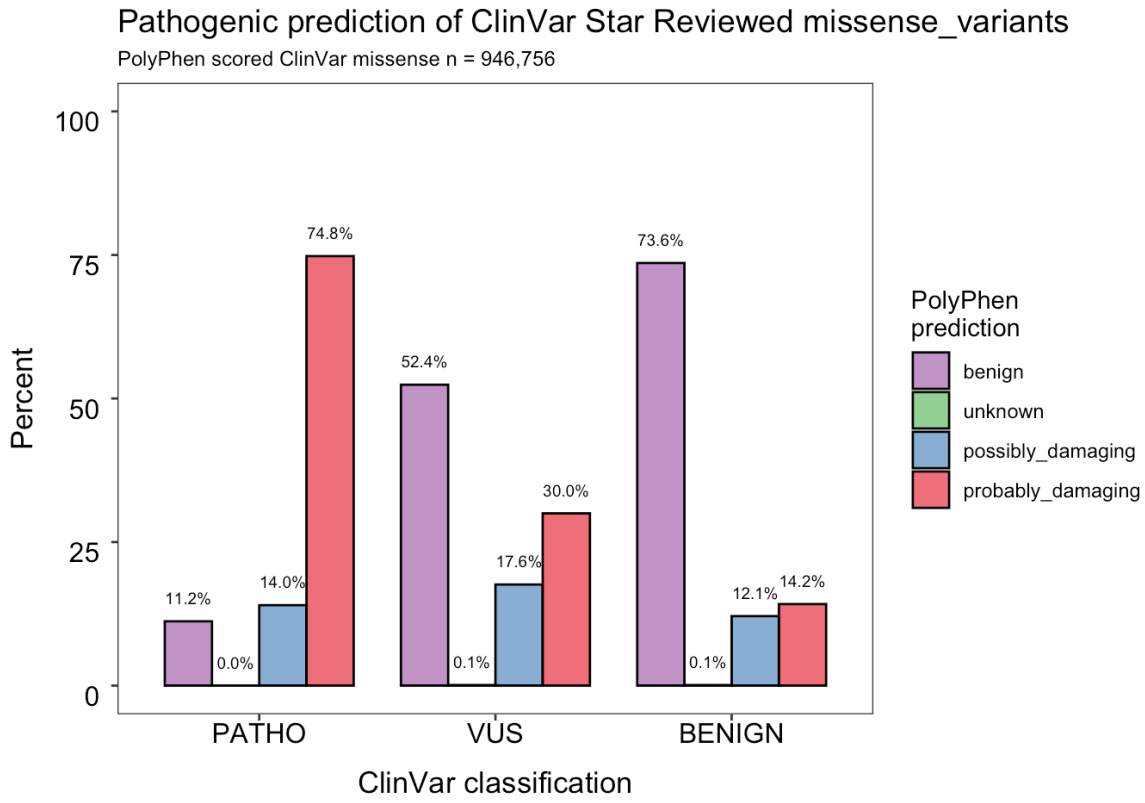
```
# difference bettween non star filtered and only. star reviewed= v1 82k
## sift clinvnar ############################
sift_clinvar <- star %>%
  dplyr::group_by(myClinVarLabels, SIFT) %>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(Percent = round(n/sum(n) * 100, 1))
sum(sift_clinvar$n)
```

```
## [1] 945568
```

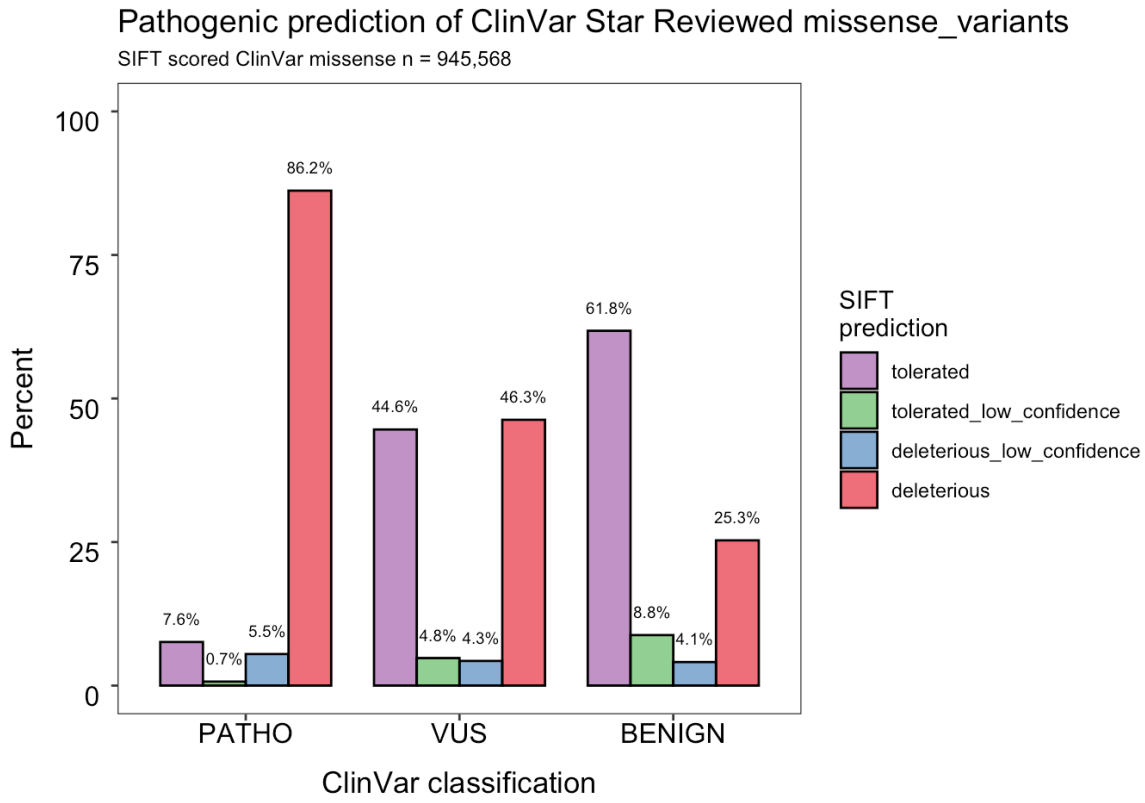
```
print(sift_clinvar)
```

```
## # A tibble: 12 × 4
## # Groups:   myClinVarLabels [3]
##    myClinVarLabels SIFT                          n Percent
##    <fct>           <fct>                     <int>   <dbl>
##  1 PATHO           tolerated                  2972     7.6
##  2 PATHO           tolerated_low_confidence    280     0.7
##  3 PATHO           deleterious_low_confidence 2153     5.5
##  4 PATHO           deleterious               33730    86.2
##  5 VUS             tolerated                369584    44.6
##  6 VUS             tolerated_low_confidence  40045     4.8
##  7 VUS             deleterious_low_confidence 35252    4.3
##  8 VUS             deleterious              383675    46.3
##  9 BENIGN          tolerated                 48163    61.8
## 10 BENIGN          tolerated_low_confidence   6835     8.8
## 11 BENIGN          deleterious_low_confidence 3165     4.1
## 12 BENIGN          deleterious               19714    25.3
```

```
polyphen_clinvar_star_bar <- ggplot(polyphen_clinvar_star, aes(x=myClinVarLabels,
                                            y=Percent,
                                            fill=PolyPhen)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.7, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(limit=c(0,100)) +
   scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="ClinVar classification",
       y= "Percent",
       fill='PolyPhen\nprediction',
       title="Pathogenic prediction of ClinVar Star Reviewed missense_variants",
       subtitle = "PolyPhen scored ClinVar missense n = 946,756") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
polyphen_clinvar_star_bar
```

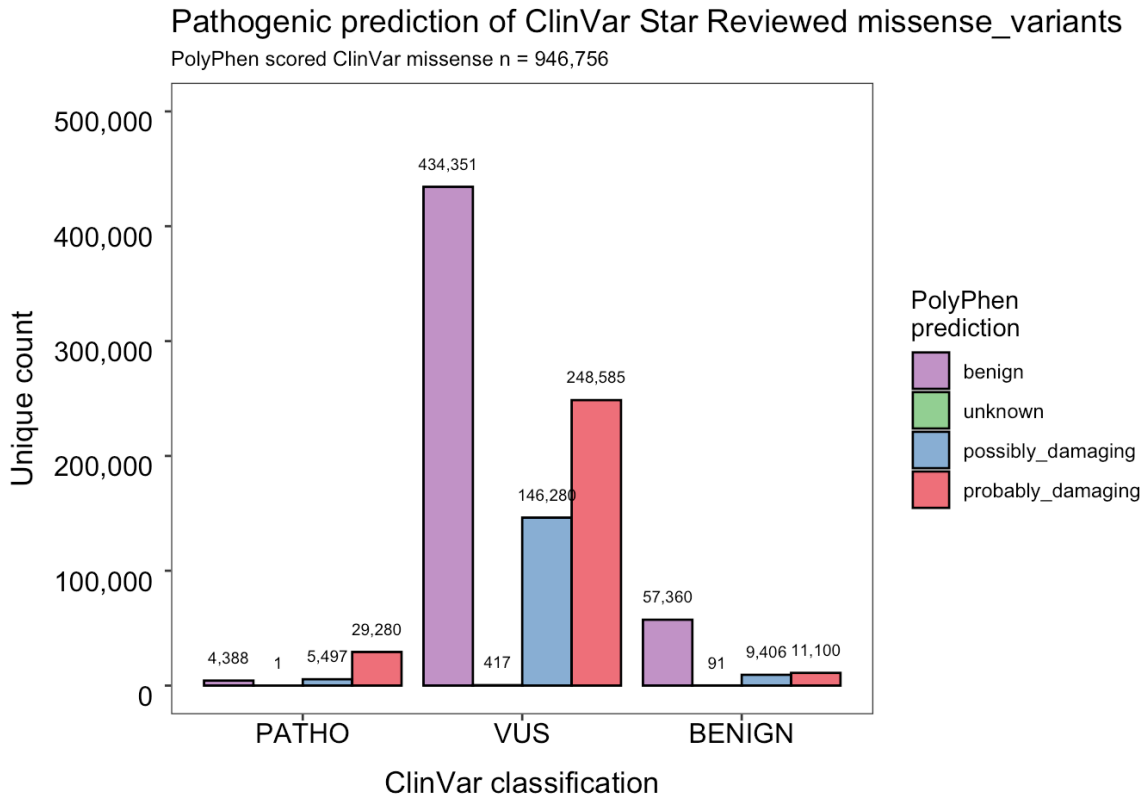

Pathogenic prediction of ClinVar Star Reviewed missense_variants
PolyPhen scored ClinVar missense n = 946,756
Percent
100
75
50
25
0
11.2%
0.0%
14.0%
74.8%
52.4%
0.1%
17.6%
30.0%
73.6%
0.1%
12.1%
14.2%
PATHO
VUS
BENIGN
ClinVar classification
PolyPhen prediction
benign
unknown
possibly_damaging
probably_damaging

```
ggsave("clinvarDB_bardodge_percent_PolyPhen.png", width = 8, height=6, dpi= 300)



clinvar_sift_bar <- ggplot(sift_clinvar, aes(x=myClinVarLabels,
                                             y=Percent,
                                             fill=SIFT)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.7, na.rm=TRUE) +
  geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
            position=position_dodge2(width=0.8),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(limit=c(0,100)) +
  scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="ClinVar classification",
       y= "Percent",
       fill='SIFT\nprediction',
       title="Pathogenic prediction of ClinVar Star Reviewed missense_variants",
       subtitle = "SIFT scored ClinVar missense n = 945,568") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
clinvar_sift_bar
```

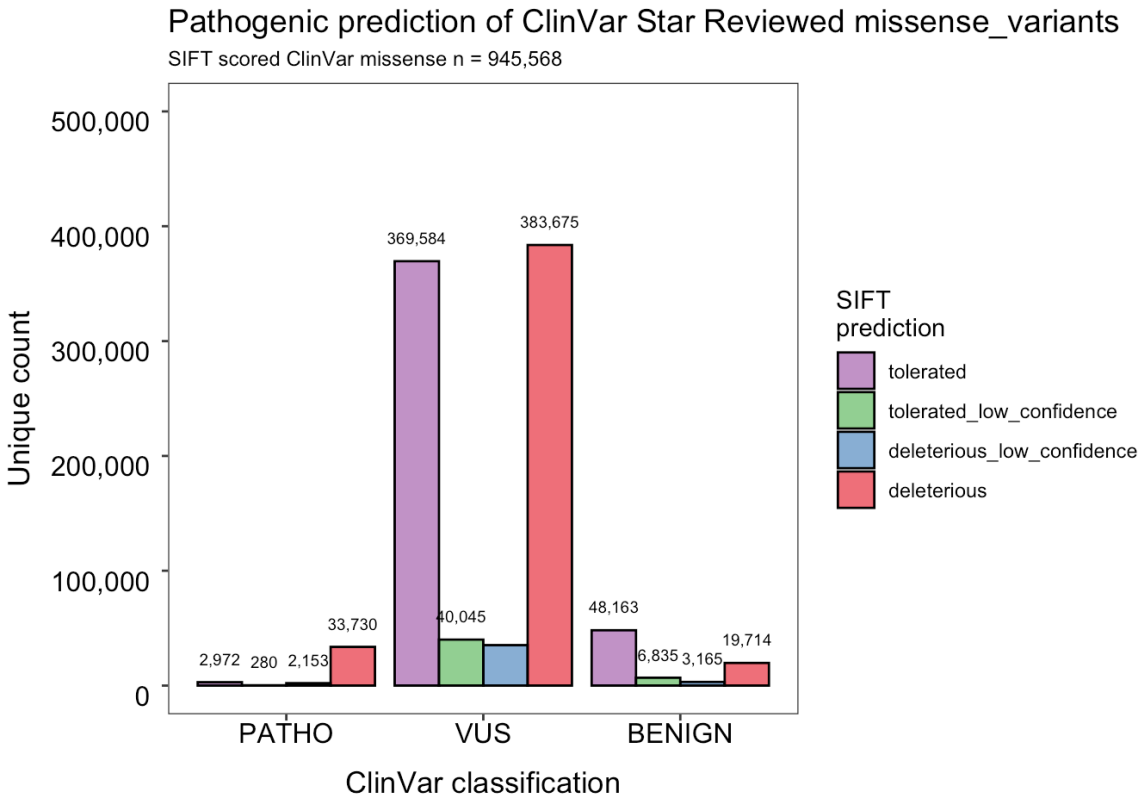



```
ggsave("clinvarDB_bardodge_percent_SIFT.png", width = 8, height=6, dpi= 300)
```

```
polyphen_clinvar_star_bar <- ggplot(polyphen_clinvar_star, aes(x=myClinVarLabels,
                                            y=n,
                                            fill=PolyPhen)) +
  geom_bar(position="dodge", stat="identity", width=0.9, colour="black", alpha=0.7, na.rm=TRUE) +
  geom_text(aes(label= scales::comma(n)),
            position=position_dodge2(width=0.9),
            vjust=-1.5,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,500000)) +
   scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="ClinVar classification",
      y= "Unique count",
       fill='PolyPhen\nprediction',
       title="Pathogenic prediction of ClinVar Star Reviewed missense_variants",
       subtitle = "PolyPhen scored ClinVar missense n = 946,756") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
polyphen_clinvar_star_bar
```


Pathogenic prediction of ClinVar Star Reviewed missense_variants
PolyPhen scored ClinVar missense n = 946,756
Unique count
500,000
400,000
300,000
200,000
100,000
0
4,388
1
5,497
29,280
434,351
417
146,280
248,585
57,360
91
9,406
11,100
PATHO
VUS
BENIGN
ClinVar classification
PolyPhen prediction
benign
unknown
possibly_damaging
probably_damaging

```
ggsave("clinvarDB_bardodge_count_PolyPhen.png", width = 8, height=6, dpi= 300)


clinvar_sift_bar <- ggplot(sift_clinvar, aes(x=myClinVarLabels,
                                             y=n,
                                             fill=SIFT)) +
  geom_bar(position="dodge", stat="identity", width=0.9, colour="black", alpha=0.7, na.rm=TRUE) +
   geom_text(aes(label= scales::comma(n)),
             position=position_dodge2(width=0.9),
             vjust=-1.5,
             colour="black",
             angle=0,
             size = 2.5,
             check_overlap = TRUE) +
 scale_y_continuous(label=comma, limits = c(0,500000)) +
  scale_fill_brewer(palette = "Set1", direction = -1) +
  labs(x="ClinVar classification",
       y= "Unique count",
       fill='SIFT\nprediction',
       title="Pathogenic prediction of ClinVar Star Reviewed missense_variants",
       subtitle = "SIFT scored ClinVar missense n = 945,568") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=12, color="black", margin = margin(t = 5, r=5, b
= 5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=14),
        plot.subtitle = element_text(size=9))
clinvar_sift_bar
```



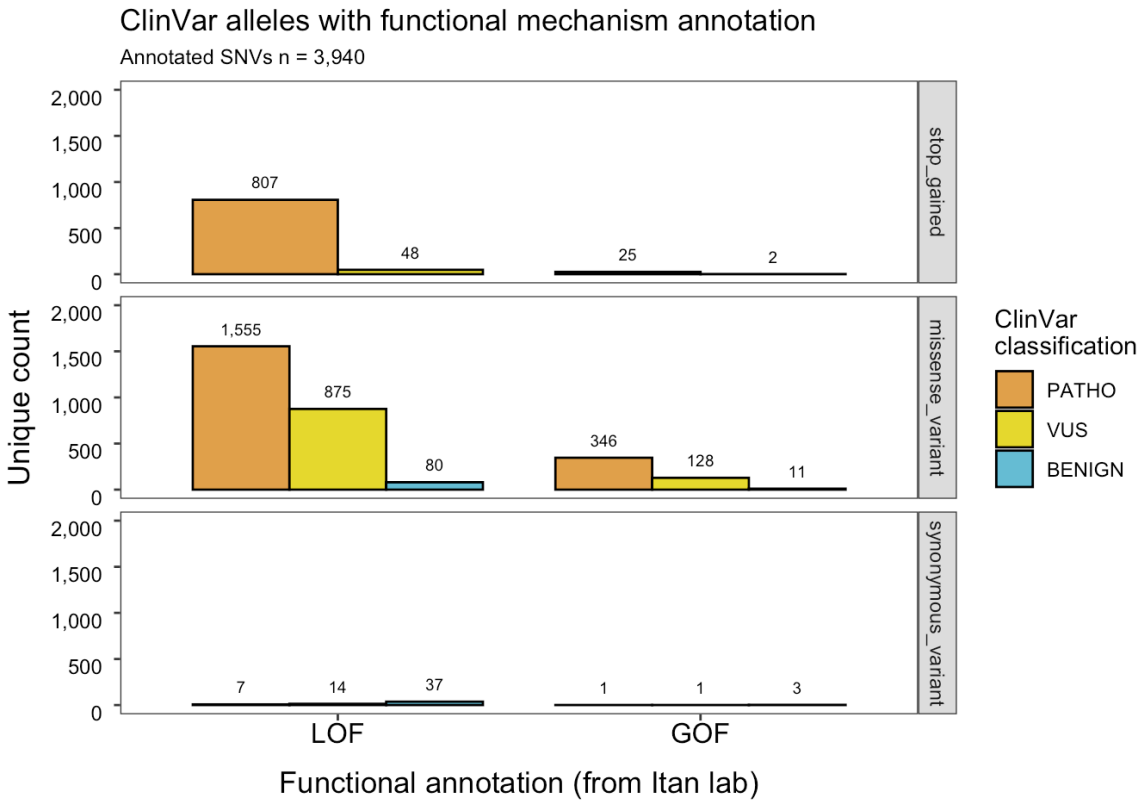

```
ggsave("clinvarDB_bardodge_count_SIFT.png", width = 8, height=6, dpi= 300)
```

ClinVar mechanism annotations from Itan lab- Bayrak et al

```
# clinvar GOF LOF ############################
gof_var1 <- clinvar %>%
  dplyr::group_by(clinvarCONSEQ, LABEL, myClinVarLabels)%>%
  dplyr::tally()%>%
 drop_na(LABEL) %>%
  dplyr::mutate(percent=n/sum(n))
print(sum(gof_var1$n))
```

```
## [1] 3940
```

```
gof_var_bar <- ggplot(gof_var1, aes(x=LABEL,
                                    y=n,
                                  fill=myClinVarLabels)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.9, na.rm=TRUE) +
 # geom_bar(stat="identity", width=0.5, colour="black", alpha=0.8) +
  #geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
  geom_text(aes(label=scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,2000), n.breaks = 5) +
  scale_fill_manual(values = class_colors) +
  labs(x="Functional annotation (from Itan lab)",
       y= "Unique count",
       fill="ClinVar\nclassification",
       title="ClinVar alleles with functional mechanism annotation",
       subtitle = "Annotated SNVs n = 3,940") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=9, color="black", margin = margin(t = 5, r=5, b =
5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=13),
        plot.subtitle = element_text(size=9)) +
  facet_grid(clinvarCONSEQ ~ . )
gof_var_bar
```

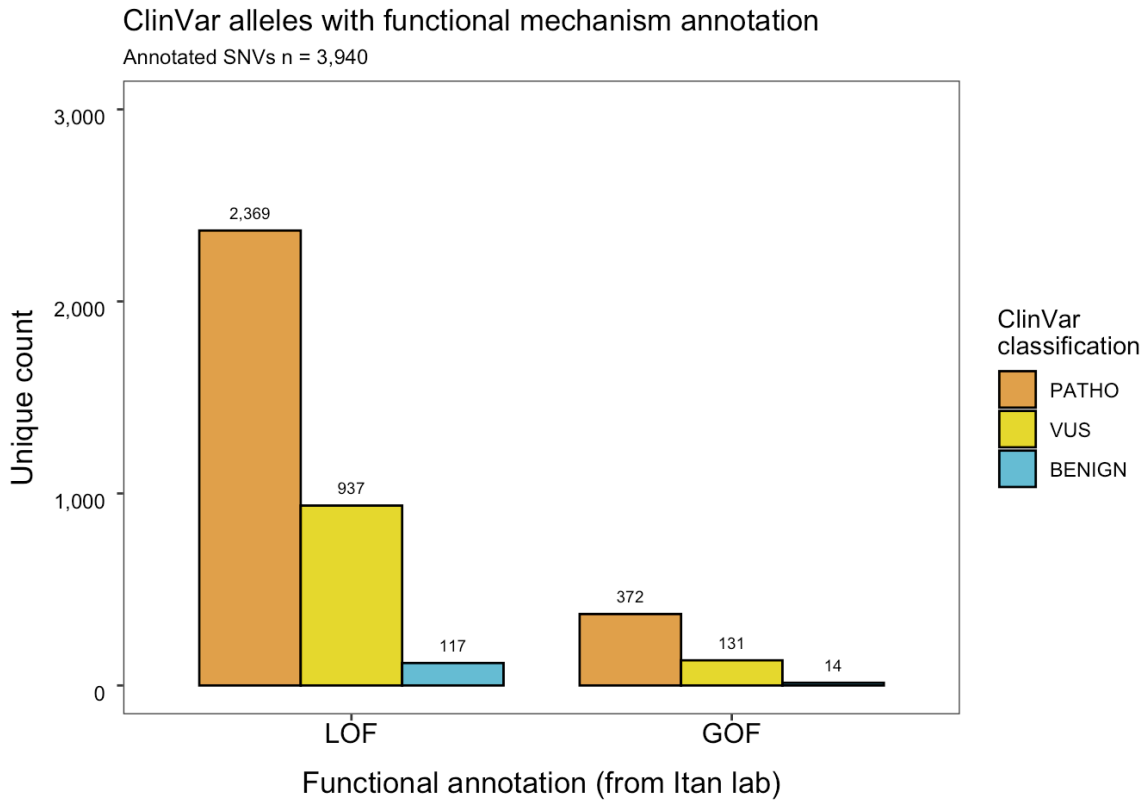



```
ggsave("clinvarDB_bardodge_count_GOF_LOF_dropna_facet.png", width = 6, height=6, dpi= 300)
```

```
# clinvar GOF LOF2 ############################
gof_var2 <- clinvar %>%
  dplyr::group_by(LABEL, myClinVarLabels)%>%
  dplyr::tally()%>%
 drop_na(LABEL) %>%
  dplyr::mutate(percent=n/sum(n))
print(sum(gof_var2$n))
```

```
## [1] 3940
```

```
gof_var_bar2 <- ggplot(gof_var2, aes(x=LABEL,
                                     y=n,
                                fill=myClinVarLabels)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.9, na.rm=TRUE) +
 # geom_bar(stat="identity", width=0.5, colour="black", alpha=0.8) +
  #geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
  geom_text(aes(label=scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            angle=0,
            size = 2.5,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,3000), n.breaks = 5) +
  scale_fill_manual(values = class_colors) +
    labs(x="Functional annotation (from Itan lab)",
       y= "Unique count",
       fill="ClinVar\nclassification",
       title="ClinVar alleles with functional mechanism annotation",
       subtitle = "Annotated SNVs n = 3,940") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=9, color="black", margin = margin(t = 5, r=5, b =
5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=13),
        plot.subtitle = element_text(size=9))
gof_var_bar2
```

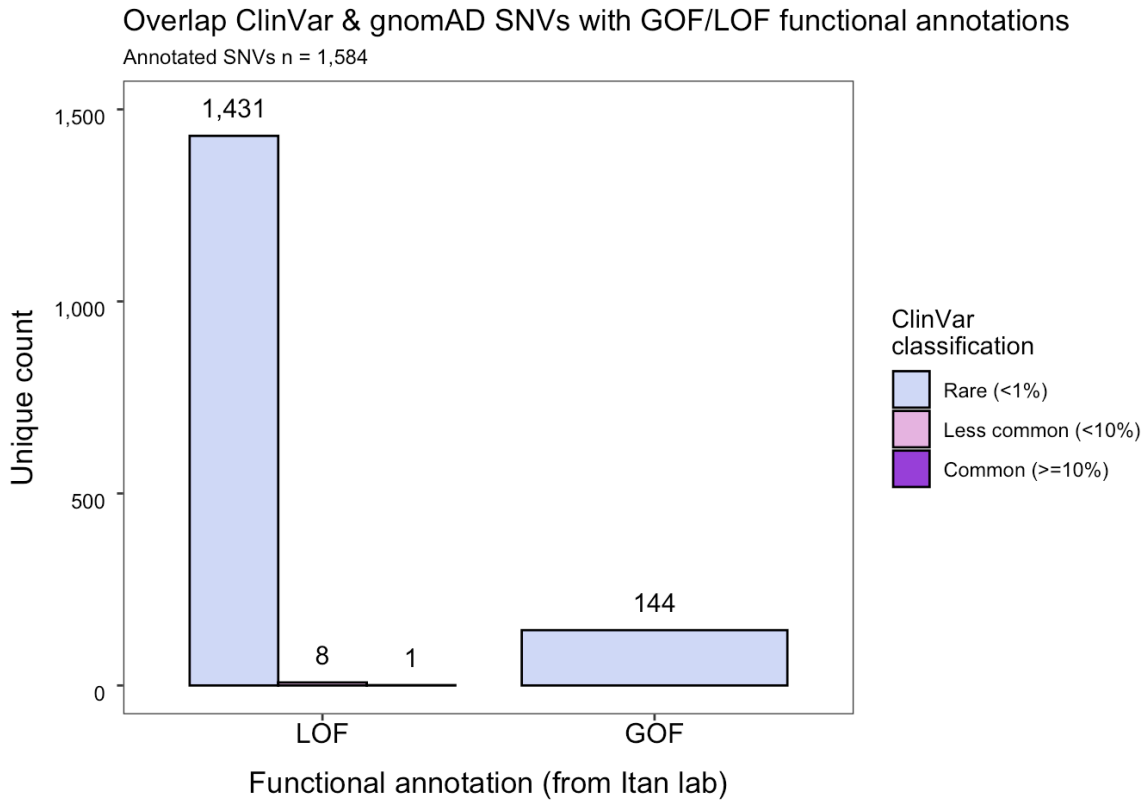



```
ggsave("clinvarDB_bardodge_count_GOF_LOF_dropna.png", width = 6, height=6, dpi= 300)
```

```
gof_var_maf <- clinvar %>%
  dplyr::group_by(LABEL, MAF)%>%
  dplyr::tally()%>%
  drop_na() %>%
  dplyr::mutate(percent=n/sum(n))
print(sum(gof_var_maf$n))
```

```
## [1] 1584
```

```
gof_maf_bar <- ggplot(gof_var_maf, aes(x=LABEL,
                                   y=n,
                                   fill=MAF)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.9, na.rm=TRUE) +
  # geom_bar(stat="identity", width=0.5, colour="black", alpha=0.8) +
  #geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
  geom_text(aes(label=scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            angle=0,
            size = 4,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,1500), n.breaks = 5) +
  scale_fill_manual(values = maf_colors) +
  labs(x="Functional annotation (from Itan lab)",
       y= "Unique count",
       fill="ClinVar\nclassification",
      title="Overlap ClinVar & gnomAD SNVs with GOF/LOF functional annotations ",
       subtitle = "Annotated SNVs n = 1,584") +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=9, color="black", margin = margin(t = 5, r=5, b =
5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=13),
        plot.subtitle = element_text(size=9))
gof_maf_bar
```

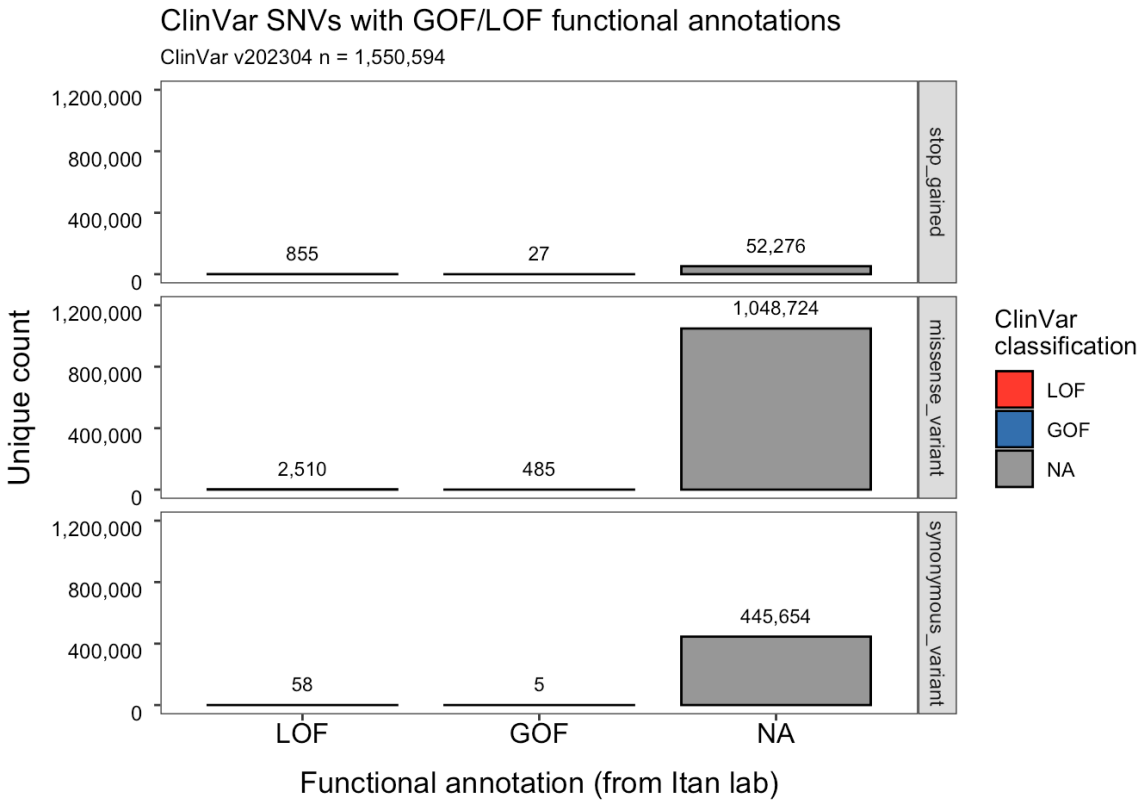



```
ggsave("clinvarDB_bardodge_count_GOF_LOF_dropna_MAF.png", width = 6, height=6, dpi= 300)
```

```
gof_var2 <- clinvar %>%
  dplyr::group_by(clinvarCONSEQ, LABEL) %>%
  dplyr::tally()%>%
  dplyr::mutate(Percent= round(n/sum(n), 2))
print(sum(gof_var2$n))
```

```
## [1] 1550594
```

```
gof_var_bar2 <- ggplot(gof_var2, aes(x=LABEL,
                                     y= n,
                                     fill=LABEL)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.9, na.rm=TRUE) +
  #geom_bar(stat="identity", width=0.5, colour="black", alpha=0.8) +
  #geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
 geom_text(aes(label=scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            angle=0,
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,1200000), n.breaks = 4) +
  scale_fill_manual(values = goflof_colors) +
  labs(x="Functional annotation (from Itan lab)",
       y= "Unique count",
       fill="ClinVar\nclassification",
       title="ClinVar SNVs with GOF/LOF functional annotations ",
       subtitle = SUB1) +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=9, color="black", margin = margin(t = 5, r=5, b =
5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=13),
        plot.subtitle = element_text(size=9)) +
  facet_grid(clinvarCONSEQ ~ . )
gof_var_bar2
```

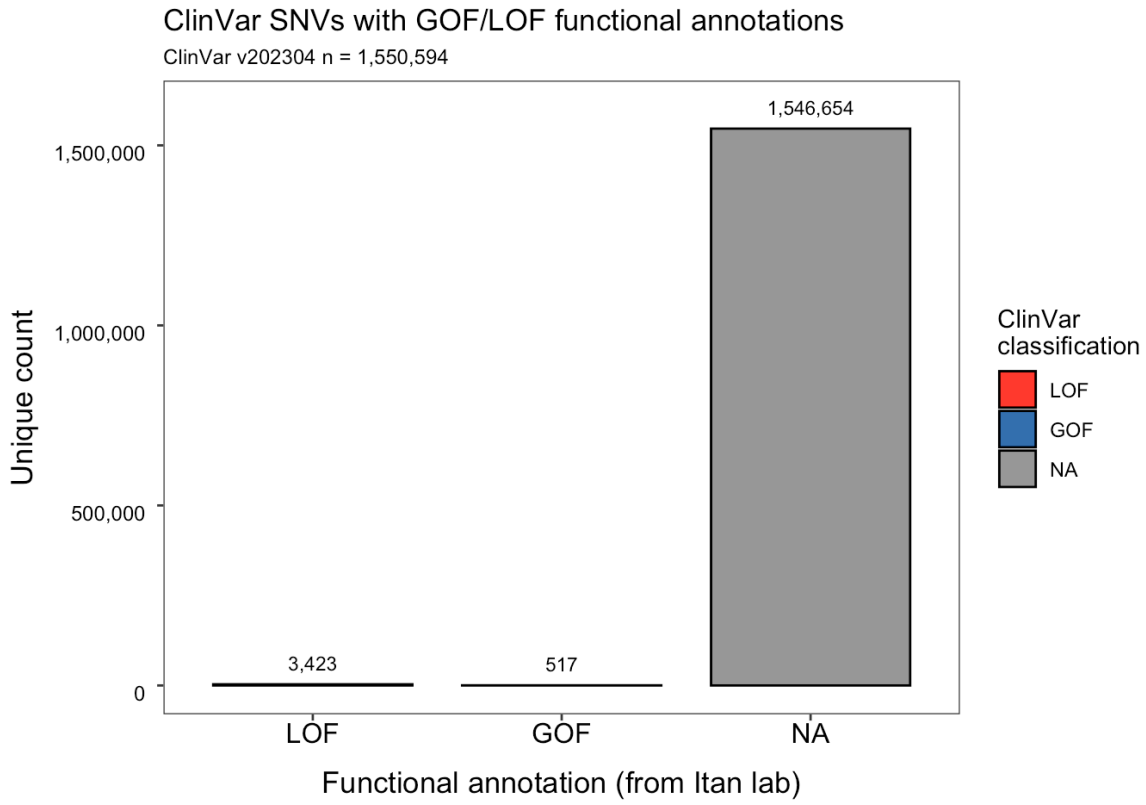



```
ggsave("clinvarDB_bardodge_count_GOF_LOF_keepna_facet.png", width = 6, height=6, dpi= 300)
```

```
gof_var3 <- clinvar %>%
  dplyr::group_by(LABEL) %>%
  dplyr::tally()%>%
  dplyr::mutate(Percent= round(n/sum(n), 2))
print(sum(gof_var2$n))
```

```
## [1] 1550594
```

```
gof_var_bar3 <- ggplot(gof_var3, aes(x=LABEL,
                                     y= n,
                                     fill=LABEL)) +
  geom_bar(position="dodge", stat="identity", width=0.8, colour="black", alpha=0.9, na.rm=TRUE) +
  #geom_bar(stat="identity", width=0.5, colour="black", alpha=0.8) +
  #geom_text(aes(label=paste0(sprintf("%1.1f", Percent),"%")),
 geom_text(aes(label=scales::comma(n)),
            position=position_dodge2(width=0.8),
            vjust=-1,
            colour="black",
            angle=0,
            size = 3,
            check_overlap = TRUE) +
  scale_y_continuous(label=comma, limits = c(0,1600000), n.breaks = 5) +
  scale_fill_manual(values = goflof_colors) +
  labs(x="Functional annotation (from Itan lab)",
       y= "Unique count",
       fill="ClinVar\nclassification",
       title="ClinVar SNVs with GOF/LOF functional annotations ",
       subtitle = SUB1) +
  theme_bw() +
  theme(legend.position='right') +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_blank()) +
  theme(axis.title.x = element_text(size=13, color="black", margin=margin(t=10, b=5)),
        axis.title.y = element_text(vjust = 3, size=13, color="black", margin=margin(t=5, b=5, r=1,
l=1)),
        axis.text.y = element_text(vjust = 1, size=9, color="black", margin = margin(t = 5, r=5, b =
5)),
        axis.text.x = element_text(size=12, color="black", margin = margin(t = 1)),
        plot.title = element_text(size=13),
        plot.subtitle = element_text(size=9))
gof_var_bar3
```




```
ggsave("clinvarDB_bardodge_count_GOF_LOF_keepna.png", width = 6, height=6, dpi= 300)
```