Clinvar mutations in clients of hsp

Load the data

```
clinvar_path <- read.delim('../../body/1raw/clinvar_patho_missense.tsv', sep = ' ')
clinvar_path$Pathogenic <- 1
clinvar_ben <- read.delim('../../body/1raw/clinvar_benign_missense.tsv', sep = ' ')
clinvar_ben$Pathogenic <- 0

clients <- read.delim('../../body/2derived/clients.with.control.nonclients.txt', sep = ' ')
clients <- clients %>%
  pivot_longer(c(client, control_gene), names_to = 'client', values_to = 'Gene')
```

Merge clinvar and clients information

```
clinvar_clients <- rbind(clinvar_ben, clinvar_path)
clinvar_clients <- merge(clinvar_clients, clients, by = 'Gene', all.x = T)

clinvar_clients <- clinvar_clients[,c(1,9,16,17,18, 19)]
```

Compare clients with all others genes - Fisher

```
clinvar_clients$hsp90client_text <- ifelse(((clinvar_clients$hsp90_client == 'TRUE') & (clinvar_clients
clinvar_clients[is.na(clinvar_clients$hsp90client_text), 'hsp90client_text'] <- 'other genes'

clinvar_clients$hsc70client_text <- ifelse(((clinvar_clients$hsc70_client == 'TRUE') & (clinvar_clients
clinvar_clients[is.na(clinvar_clients$hsc70client_text), 'hsc70client_text'] <- 'other genes'


clinvar_clients$Mutation_type_text <- ifelse(clinvar_clients$Pathogenic, 'pathogenic', 'benign')

pdf('../../body/4figures/Clinvar.mut.hsp.clients.vs.all.genes.mozaicplot.pdf')
mosaicplot(table(clinvar_clients$hsp90client_text, clinvar_clients$Mutation_type_text), ylab = 'Mutatio
           main = '', color = 'cyan3', cex.axis = 1.1)

mosaicplot(table(clinvar_clients$hsc70client_text, clinvar_clients$Mutation_type_text), ylab = 'Mutatio
           main = '', color = 'cyan3', cex.axis = 1.1)


knitr::kable(t(table(clinvar_clients$hsp90client_text, clinvar_clients$Mutation_type_text)))
```

	hsp90 clients	other genes
benign	449	45514
pathogenic	2127	41661

```
knitr::kable(t(table(clinvar_clients$hsc70client_text, clinvar_clients$Mutation_type_text)))
```

	hsc70 clients	other genes
benign	359	45604
pathogenic	1360	42428

```
ft <- fisher.test(t(table(clinvar_clients$hsp90client_text, clinvar_clients$Mutation_type_text)))
print(ft)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t(table(clinvar_clients$hsp90client_text, clinvar_clients$Mutation_type_text))
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1739875 0.2142496
## sample estimates:
## odds ratio
##  0.1932457
```

```
ft <- fisher.test(t(table(clinvar_clients$hsc70client_text, clinvar_clients$Mutation_type_text)))
print(ft)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t(table(clinvar_clients$hsc70client_text, clinvar_clients$Mutation_type_text))
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2178524 0.2762866
## sample estimates:
## odds ratio
##  0.2455618
```

```
dev.off()
```

```
## pdf
##   2
```

Compare clients with control nonclients genes - Fisher

```
clinvar_clients[clinvar_clients$hsp90client_text == 'other genes', 'hsp90client_text'] <- 'nonclients'


pdf('../../body/4figures/Clinvar.mut.hsp.clients.vs.nonclients.mozaicplot.pdf')
mosaicplot(table(clinvar_clients[!is.na(clinvar_clients$hsp90_client) & (clinvar_clients$hsp90_client),]
           main = '', color = 'cyan3', cex.axis = 1.1)




knitr::kable(t(table(clinvar_clients[!is.na(clinvar_clients$hsp90_client) & (clinvar_clients$hsp90_clie
```

	hsp90 clients	nonclients
benign	449	233
pathogenic	2127	185

```
ft <- fisher.test(t(table(clinvar_clients[!is.na(clinvar_clients$hsp90_client) & (clinvar_clients$hsp90_
print(ft)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t(table(clinvar_clients[!is.na(clinvar_clients$hsp90_client) & (clinvar_clients$hsp90_client)
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1339507 0.2097970
## sample estimates:
## odds ratio
##   0.167743
```

```
##hsc70
clinvar_clients[clinvar_clients$hsc70client_text == 'other genes', 'hsc70client_text'] <- 'nonclients'

mosaicplot(table(clinvar_clients[!is.na(clinvar_clients$hsc70_client) & (clinvar_clients$hsc70_client),]
           main = '', color = 'cyan3', cex.axis = 1.1)



knitr::kable(t(table(clinvar_clients[!is.na(clinvar_clients$hsc70_client) & (clinvar_clients$hsc70_clie
```

	hsc70 clients	nonclients
benign	359	180
pathogenic	1360	171

```
ft <- fisher.test(t(table(clinvar_clients[!is.na(clinvar_clients$hsc70_client) & (clinvar_clients$hsc70_
print(ft)
```

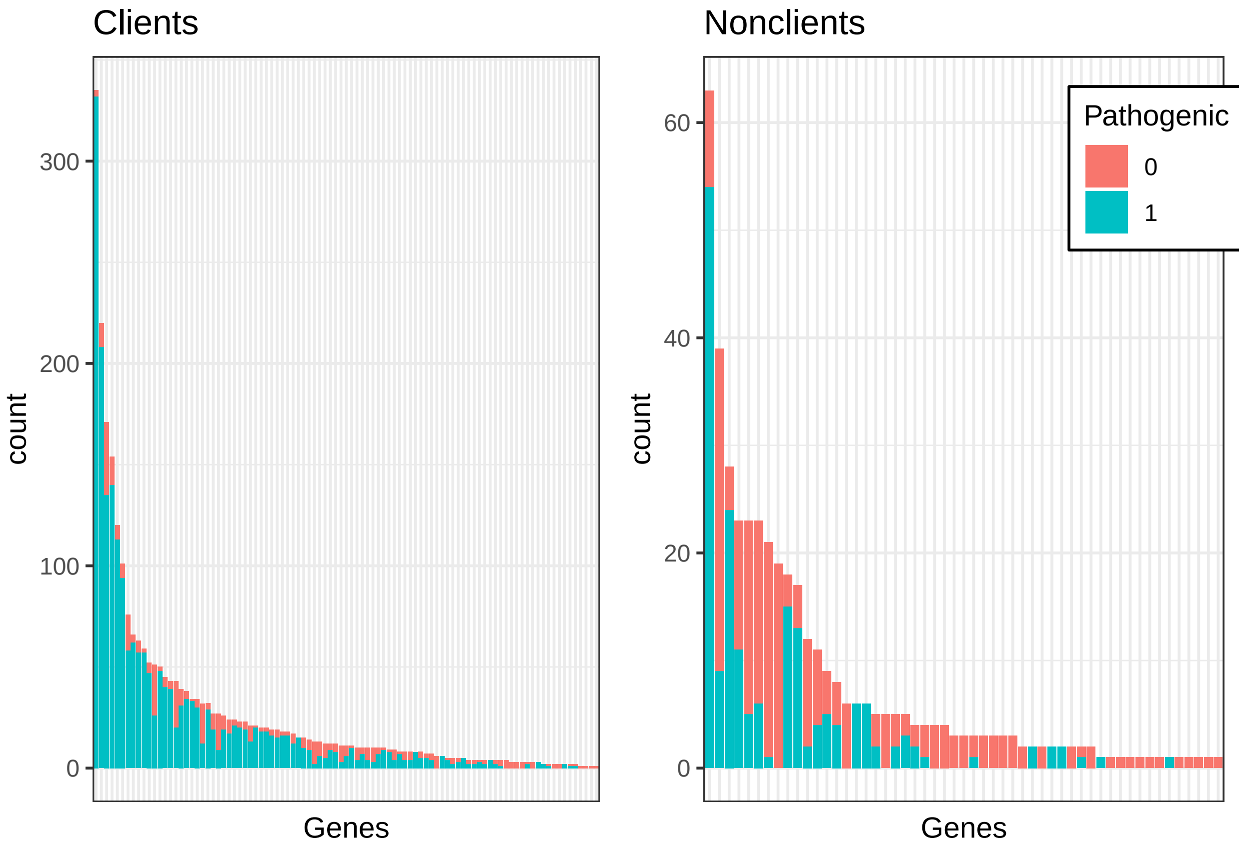
```
##
##  Fisher's Exact Test for Count Data
##
## data:  t(table(clinvar_clients[!is.na(clinvar_clients$hsc70_client) & (clinvar_clients$hsc70_client)
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1960218 0.3211073
## sample estimates:
## odds ratio
##  0.2509772
```

```
dev.off()
```

```
## pdf
##   2
```

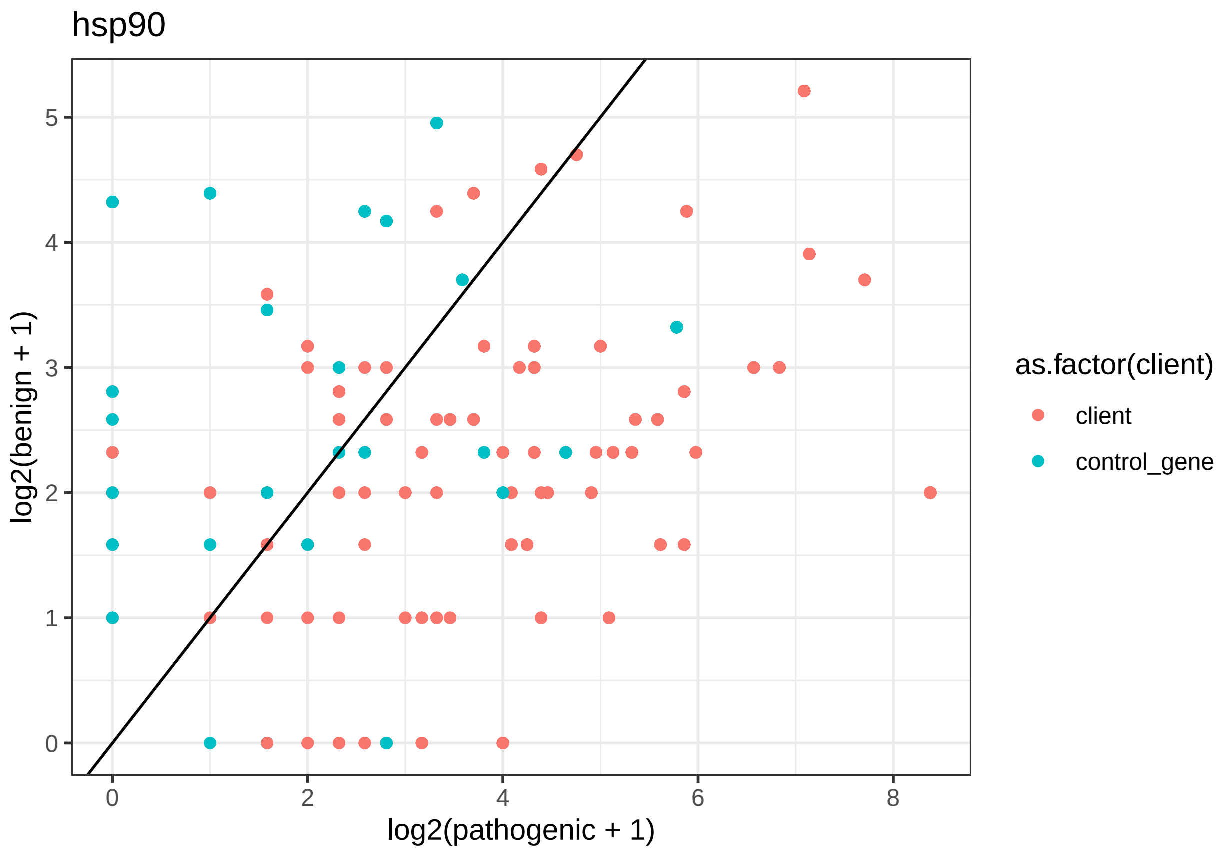
How many there are motations in one gene?

```
require(gridExtra)

clinvar_control_noncl <- clinvar_clients[!is.na(clinvar_clients$client),]

p1 <- ggplot(clinvar_control_noncl[(clinvar_control_noncl$hsp90_client) & (clinvar_control_noncl$client
  geom_bar(aes(fill = as.factor(Pathogenic)))+
  theme_bw()+
  xlab('Genes')+
  theme(axis.text.x=element_blank(), axis.ticks.x = element_blank(),
        legend.position = 'None')+
  ggtitle('Clients')

p2 <- ggplot(clinvar_control_noncl[(clinvar_control_noncl$hsp90_client) & (clinvar_control_noncl$client
  geom_bar(aes(fill = as.factor(Pathogenic)))+
  theme_bw()+
  xlab('Genes')+
  theme(axis.text.x=element_blank(), axis.ticks.x = element_blank(), legend.position = c(0.87,0.85), le
  ggtitle('Nonclients')+
  scale_fill_discrete(name = "Pathogenic")


grid.arrange(p1, p2, ncol=2)
```

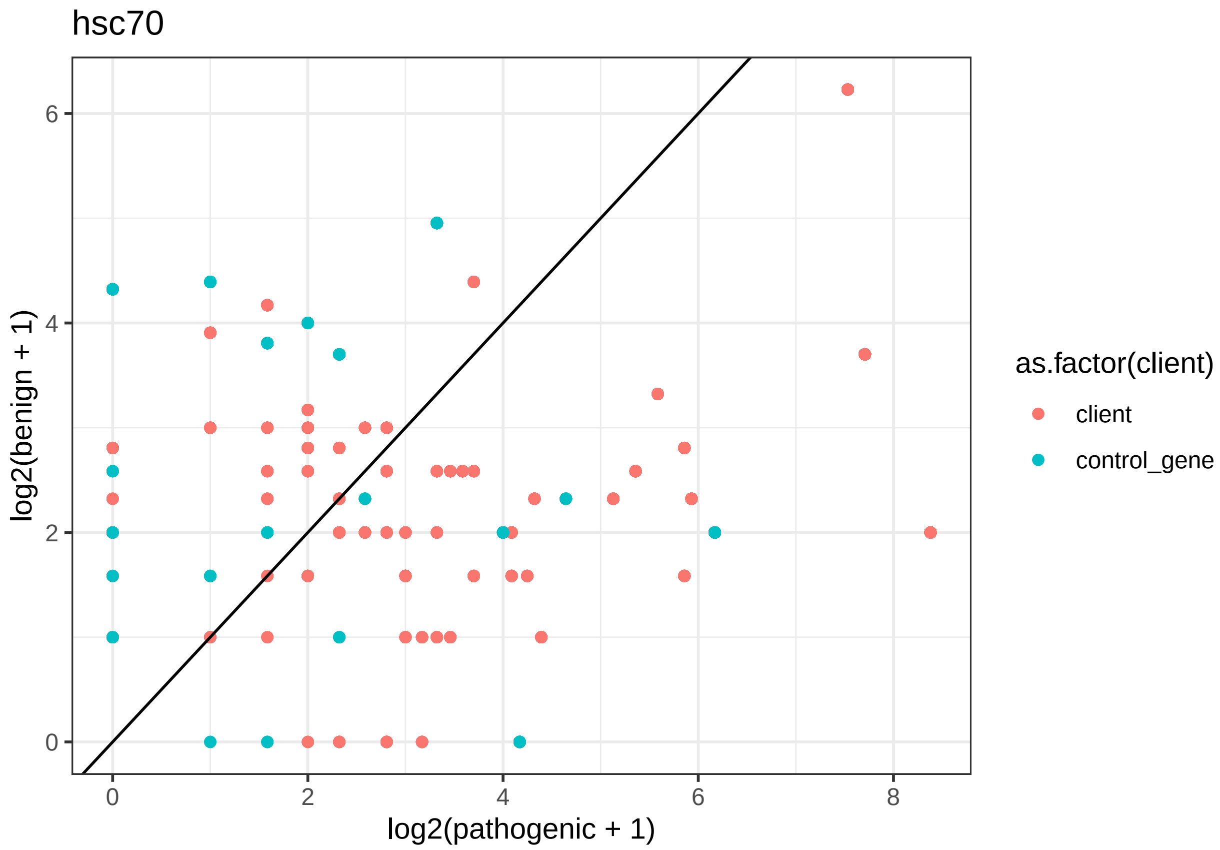


```
library('tidyverse')
```

```
count_path <- clinvar_control_noncl %>%
  count(Gene, Pathogenic)

count_path <- count_path %>%
  pivot_wider(names_from = 'Pathogenic', values_from = 'n')

count_path[is.na(count_path)] <- 0
colnames(count_path) <- c('Gene', 'pathogenic', 'benign')
count_path <- merge(count_path, clinvar_control_noncl, by = 'Gene')
count_path <- count_path[!duplicated(count_path),]

ggplot(count_path[(count_path$hsp90_client),], aes(log2(pathogenic+1), log2(benign+1), color = as.facto
  geom_point()+
  geom_abline(intercept = 0, slope = 1)+
  theme_bw()+
  ggtitle('hsp90')
```



```
ggplot(count_path[(count_path$hsc70_client),], aes(log2(pathogenic+1), log2(benign+1), color = as.facto
  geom_point()+
  geom_abline(intercept = 0, slope = 1)+
  theme_bw()+
  ggtitle('hsc70')
```



Look if the number of mutations and proportion of pathogenic per gene is different between clients and nonclients

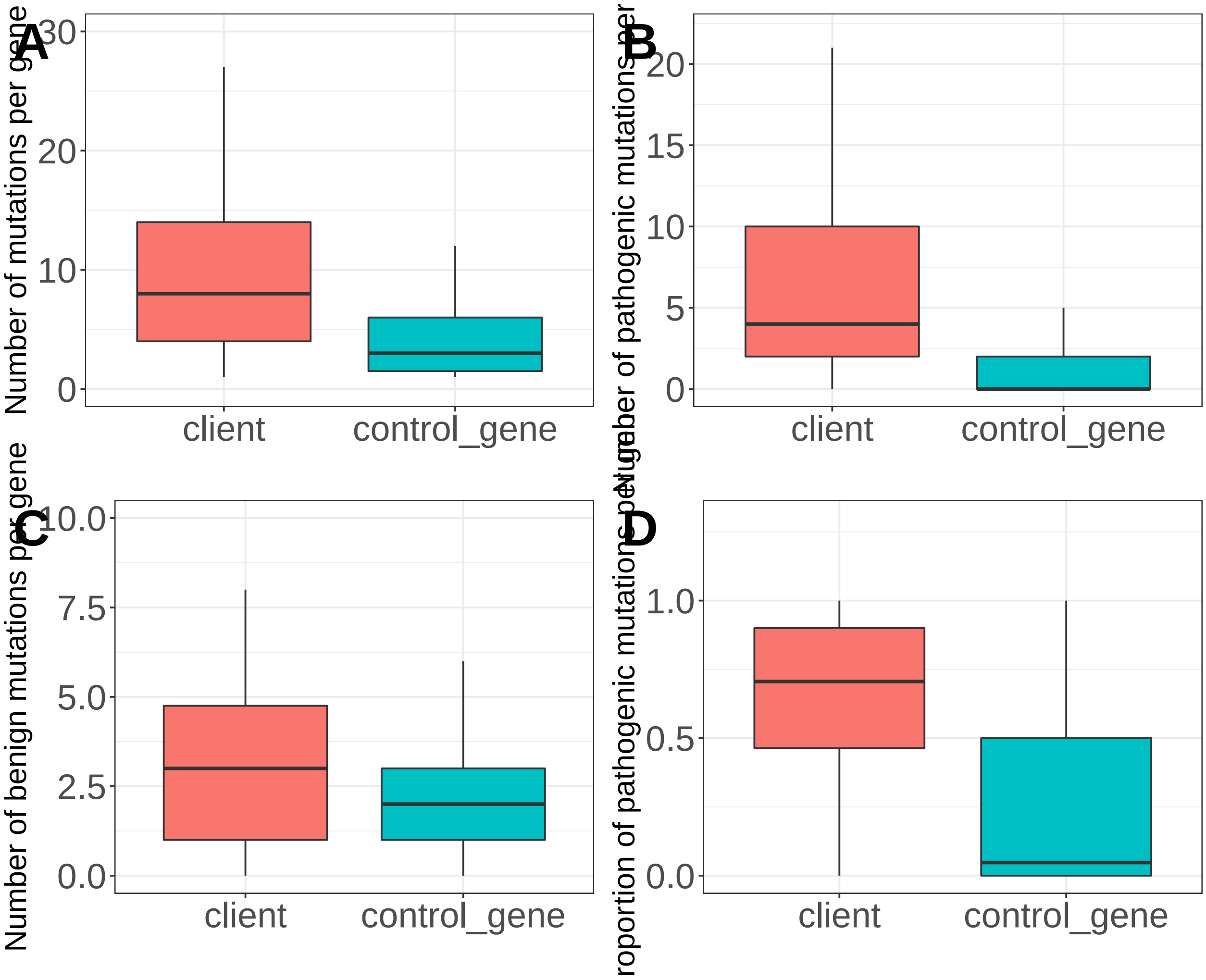
```
hsp90 <- clinvar_control_noncl[clinvar_control_noncl$hsp90_client,]

hsp90_summary <- aggregate(hsp90$Pathogenic, list(hsp90$Gene), FUN = mean)
hsp90_summary <- merge(hsp90_summary, aggregate(hsp90$Pathogenic, list(hsp90$Gene), FUN = length), by =
hsp90_summary <- merge(hsp90_summary, aggregate(hsp90$Pathogenic, list(hsp90$Gene), FUN = sum), by = 'G
colnames(hsp90_summary) <- c('Gene', 'proportion_of_pat', 'number_of_mut', 'number_of_pat' )
hsp90_summary <- merge(hsp90_summary, hsp90[,c(1,6)][!duplicated(hsp90[,c(1,6)]),], by = 'Gene', all.x =
hsp90_summary$num_of_ben <- hsp90_summary$number_of_mut - hsp90_summary$number_of_pat



p1 <- ggplot(hsp90_summary, aes(y = number_of_mut, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,30)+
  ylab('Number of mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p2 <- ggplot(hsp90_summary, aes(y = number_of_pat, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,22)+
  ylab('Number of pathogenic mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p3 <- ggplot(hsp90_summary, aes(y = num_of_ben, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,10)+
  ylab('Number of benign mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p4 <- ggplot(hsp90_summary, aes(y = proportion_of_pat, x = client, fill = client))+
  geom_boxplot()+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,1.3)+
  ylab('Proportion of pathogenic mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

pp <- plot_grid(p1,p2,p3,p4, labels = c('A','B','C','D'), ncol =2, nrow=2, label_size = 30)
```

```
## Warning: Removed 24 rows containing non-finite values (stat_boxplot).
```

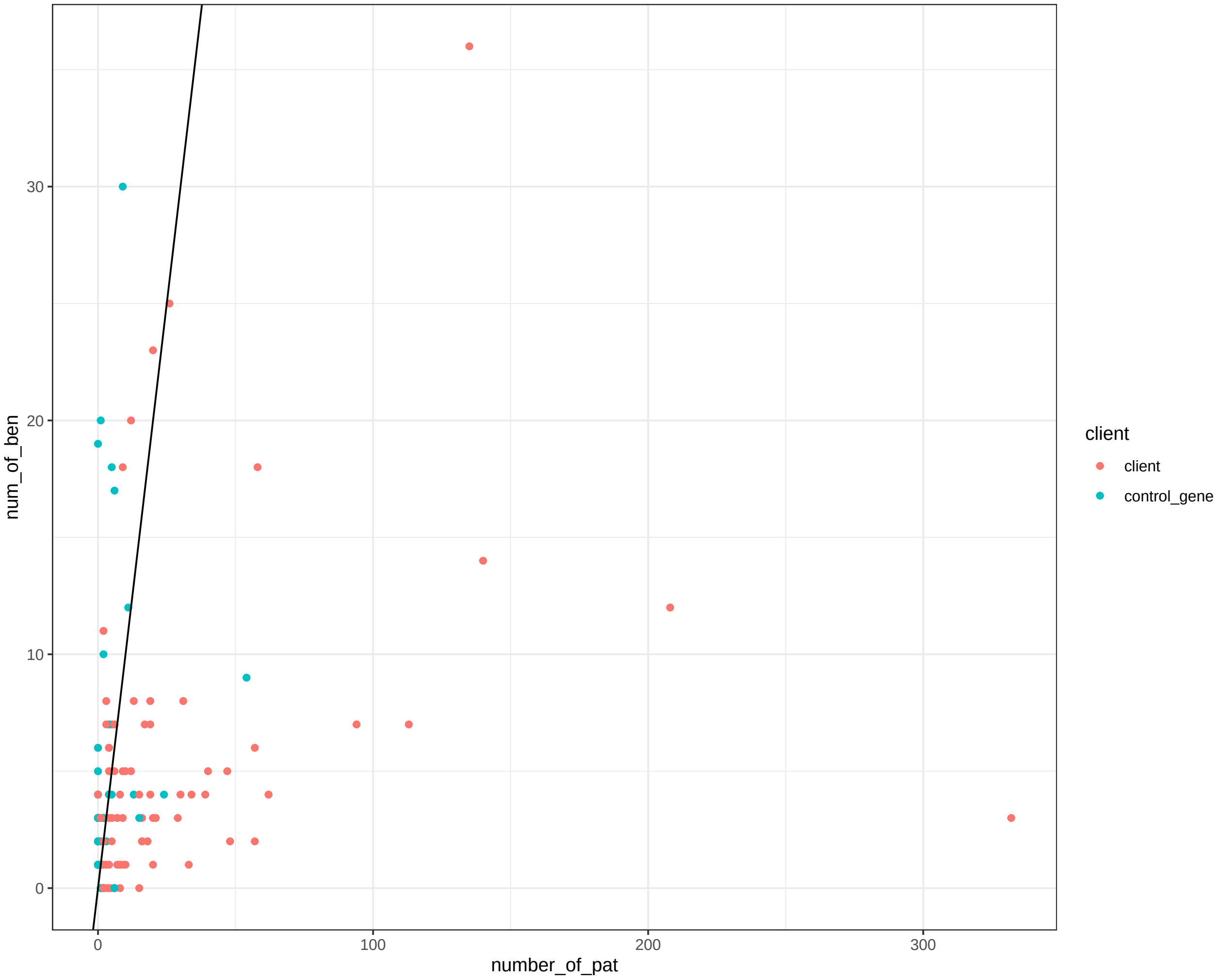
```
## Warning: Removed 22 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 15 rows containing non-finite values (stat_boxplot).
```

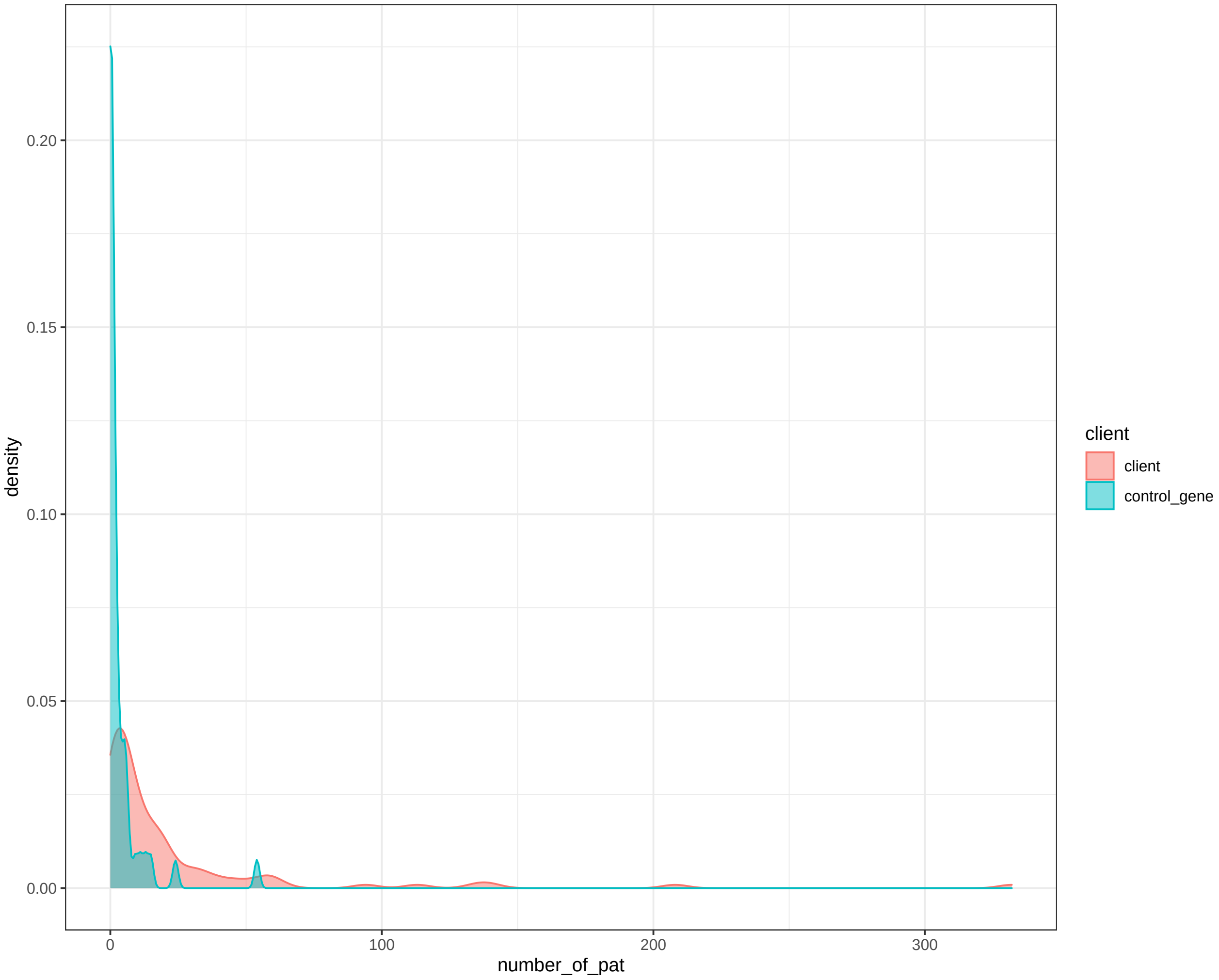
```
ggsave(pp, filename = '../../body/4figures/ClinVer.number.of.mut.per.gene.clients.vs.nonclients.pdf', wi

print(pp)
```



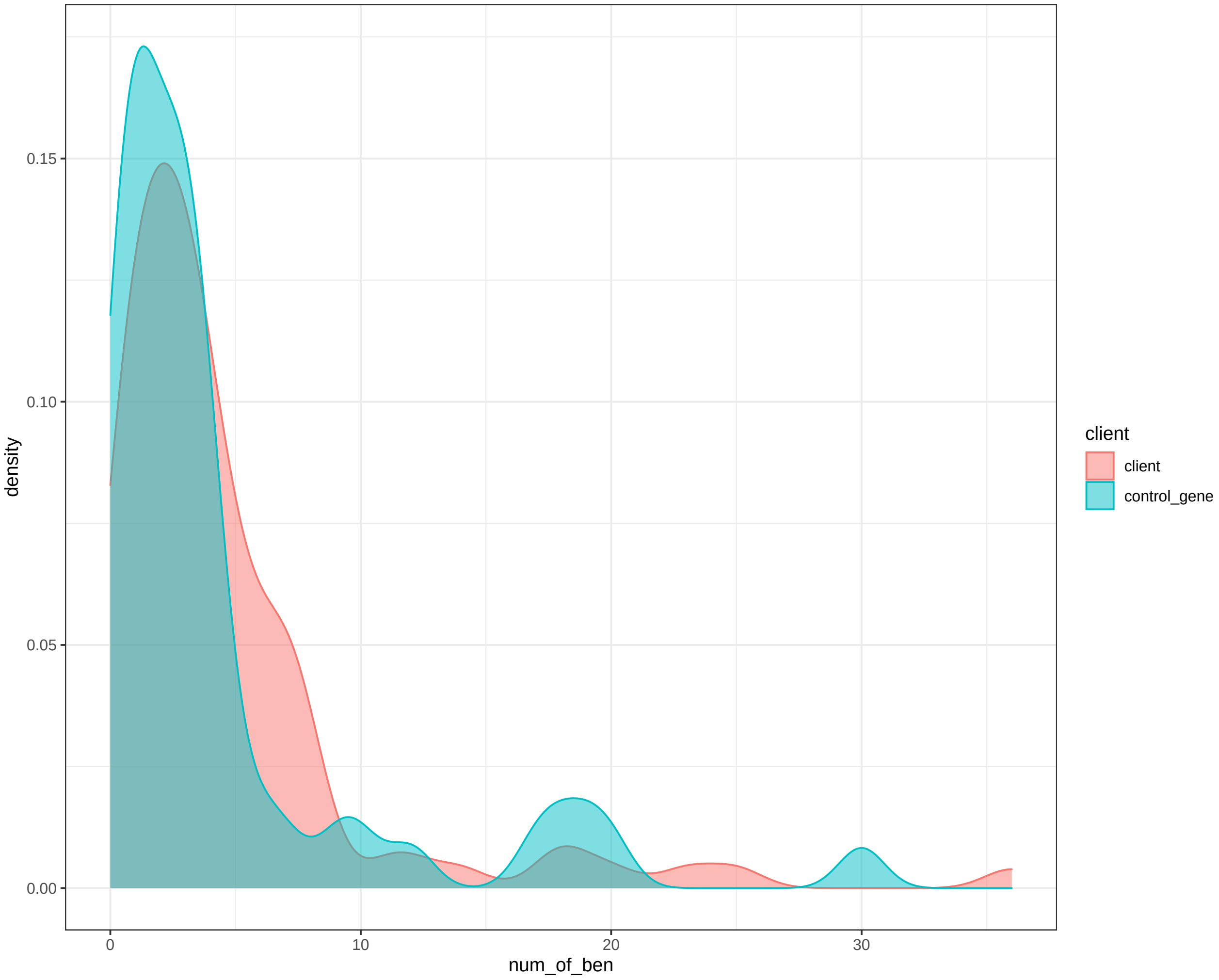
```
ggplot(hsp90_summary, aes(x = number_of_pat, y = num_of_ben, color = client))+
  geom_point()+
  geom_abline(intercept = 0, slope = 1)+
  theme_bw()
```



```
ggplot(hsp90_summary, aes(x = number_of_pat, fill = client, color = client))+
  geom_density(alpha = 0.5)+
  theme_bw()
```



```
ggplot(hsp90_summary, aes(x = num_of_ben, fill = client, color = client))+
  geom_density(alpha = 0.5)+
  theme_bw()
```



```
wilcox.test(hsp90_summary$number_of_mut ~ hsp90_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsp90_summary$number_of_mut by hsp90_summary$client
## W = 3715, p-value = 1.607e-06
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsp90_summary$number_of_pat ~ hsp90_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsp90_summary$number_of_pat by hsp90_summary$client
## W = 3955, p-value = 6.523e-09
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsp90_summary$num_of_ben ~ hsp90_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  hsp90_summary$num_of_ben by hsp90_summary$client
## W = 2850, p-value = 0.1803
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsp90_summary$proportion_of_pat ~ hsp90_summary$client)
```

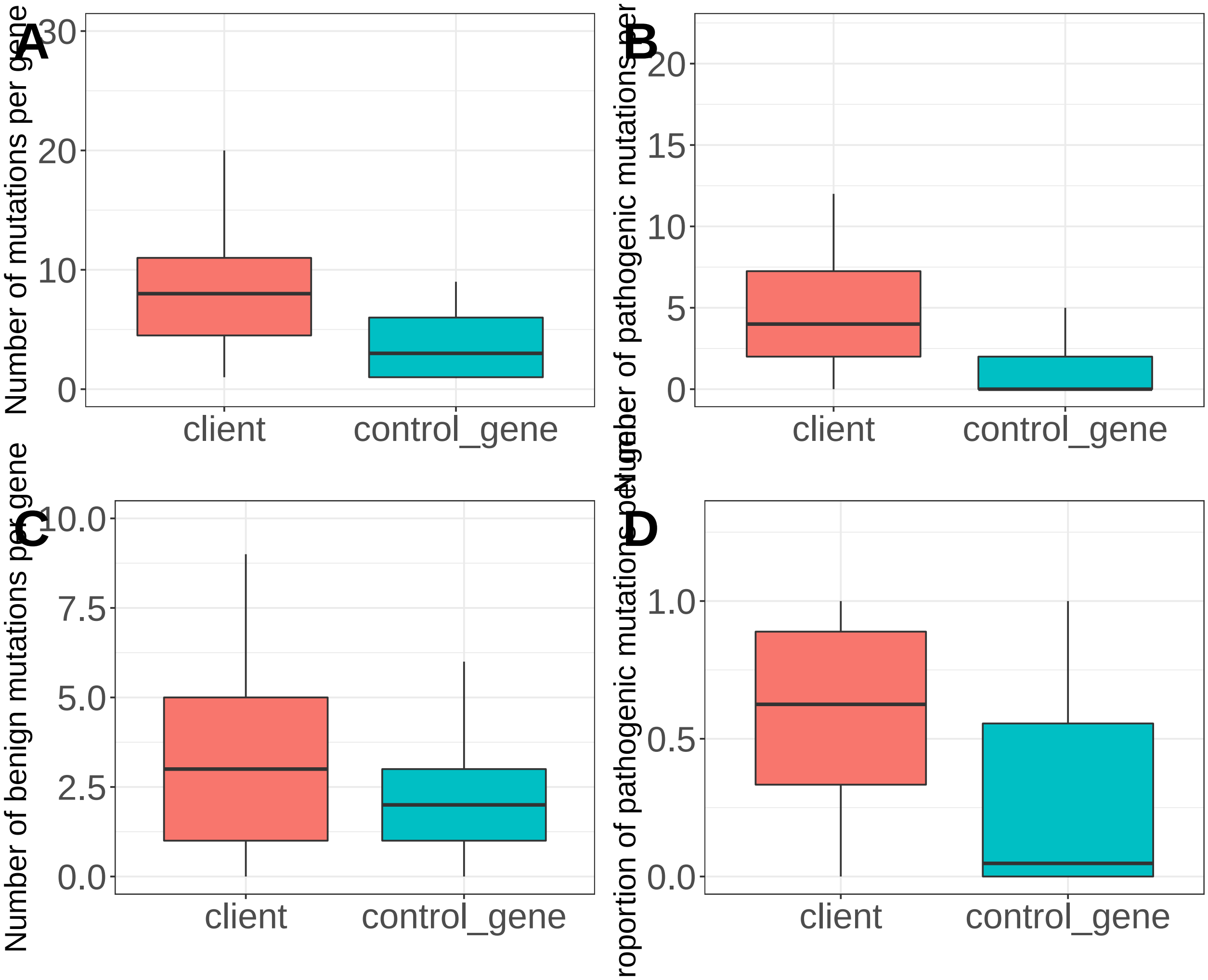
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsp90_summary$proportion_of_pat by hsp90_summary$client
## W = 3671, p-value = 3.244e-06
## alternative hypothesis: true location shift is not equal to 0
```

All the same for hsc70

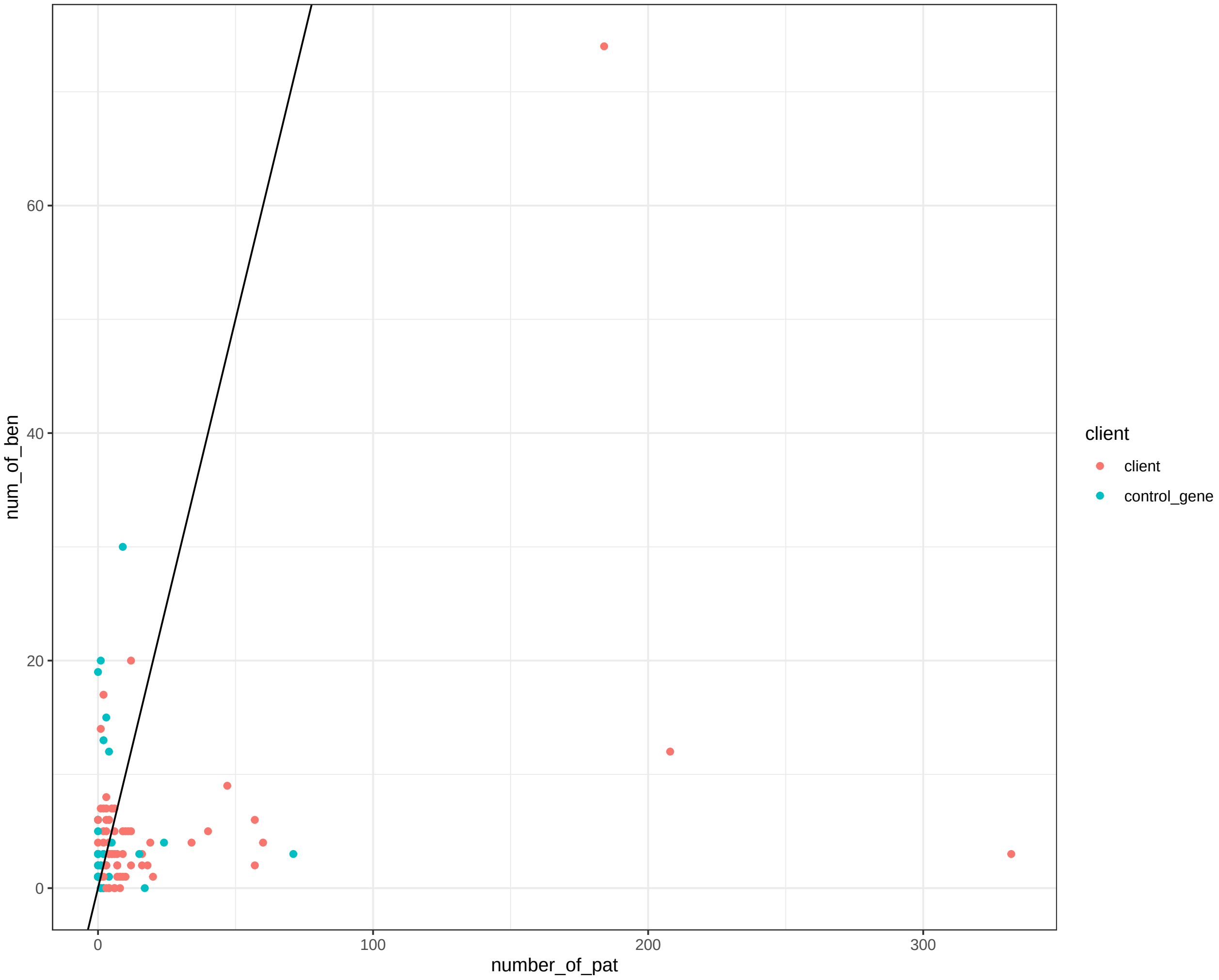
```
hsc70 <- clinvar_control_noncl[clinvar_control_noncl$hsc70_client,]

hsc70_summary <- aggregate(hsc70$Pathogenic, list(hsc70$Gene), FUN = mean)
hsc70_summary <- merge(hsc70_summary, aggregate(hsc70$Pathogenic, list(hsc70$Gene), FUN = length), by =
hsc70_summary <- merge(hsc70_summary, aggregate(hsc70$Pathogenic, list(hsc70$Gene), FUN = sum), by = 'G
colnames(hsc70_summary) <- c('Gene', 'proportion_of_pat', 'number_of_mut', 'number_of_pat' )
hsc70_summary <- merge(hsc70_summary, hsc70[,c(1,6)][!duplicated(hsc70[,c(1,6)]),], by = 'Gene', all.x =
hsc70_summary$num_of_ben <- hsc70_summary$number_of_mut - hsc70_summary$number_of_pat




p1 <- ggplot(hsc70_summary, aes(y = number_of_mut, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,30)+
  ylab('Number of mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p2 <- ggplot(hsc70_summary, aes(y = number_of_pat, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,22)+
  ylab('Number of pathogenic mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p3 <- ggplot(hsc70_summary, aes(y = num_of_ben, x = client, fill = client))+
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,10)+
  ylab('Number of benign mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

p4 <- ggplot(hsc70_summary, aes(y = proportion_of_pat, x = client, fill = client))+
  geom_boxplot()+
  theme_bw()+
  theme(legend.position = 'None')+
  ylim(0,1.3)+
  ylab('Proportion of pathogenic mutations per gene')+xlab('')+
  theme(axis.text = element_text(size=21), axis.title = element_text(size=18))

pp <- plot_grid(p1,p2,p3,p4, labels = c('A','B','C','D'), ncol =2, nrow=2, label_size = 30)
```

```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).

## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
## Removed 11 rows containing non-finite values (stat_boxplot).
```

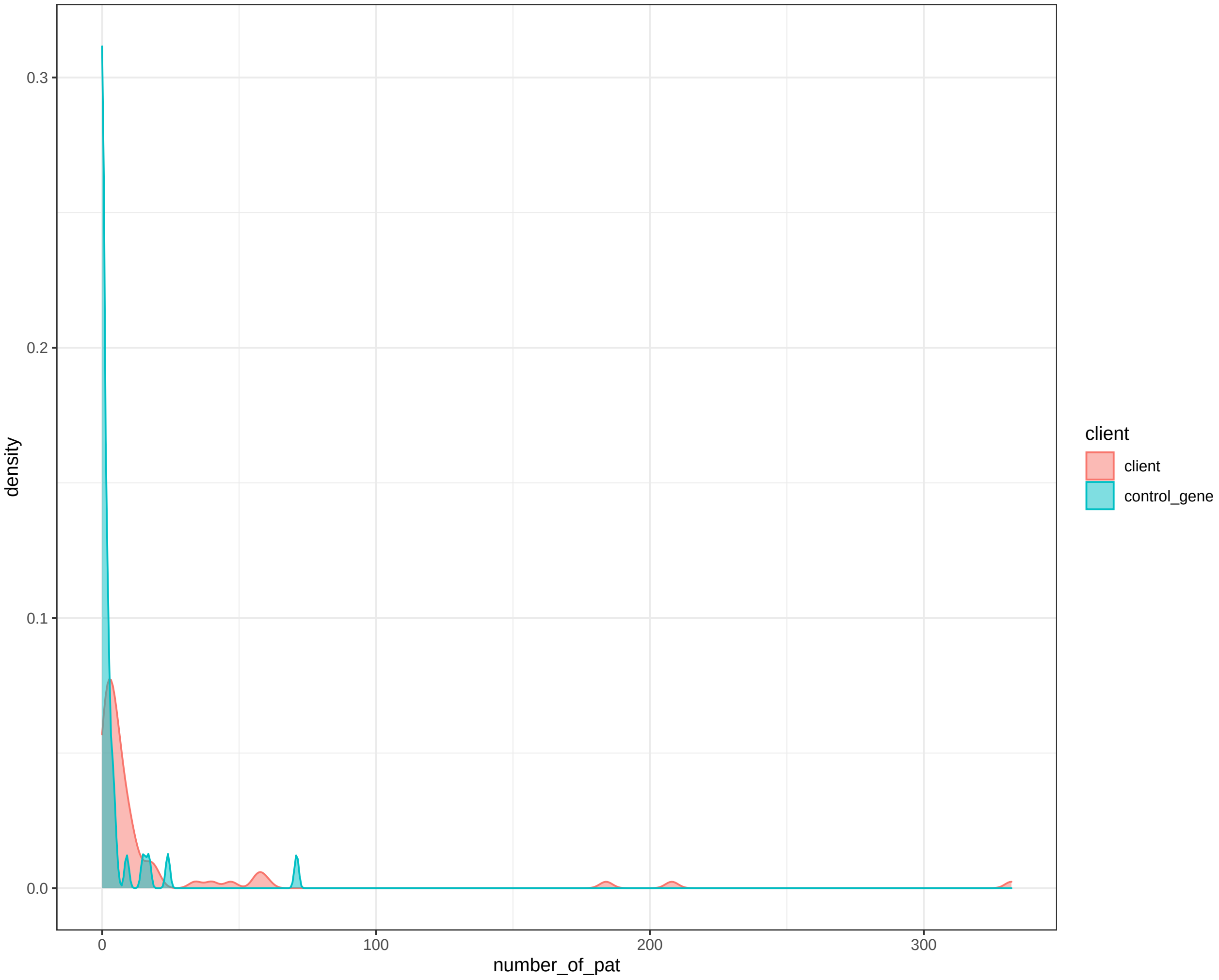
```
ggsave(pp, filename = '../../body/4figures/ClinVer.number.of.mut.per.gene.clients.vs.nonclients.pdf', wi

print(pp)
```



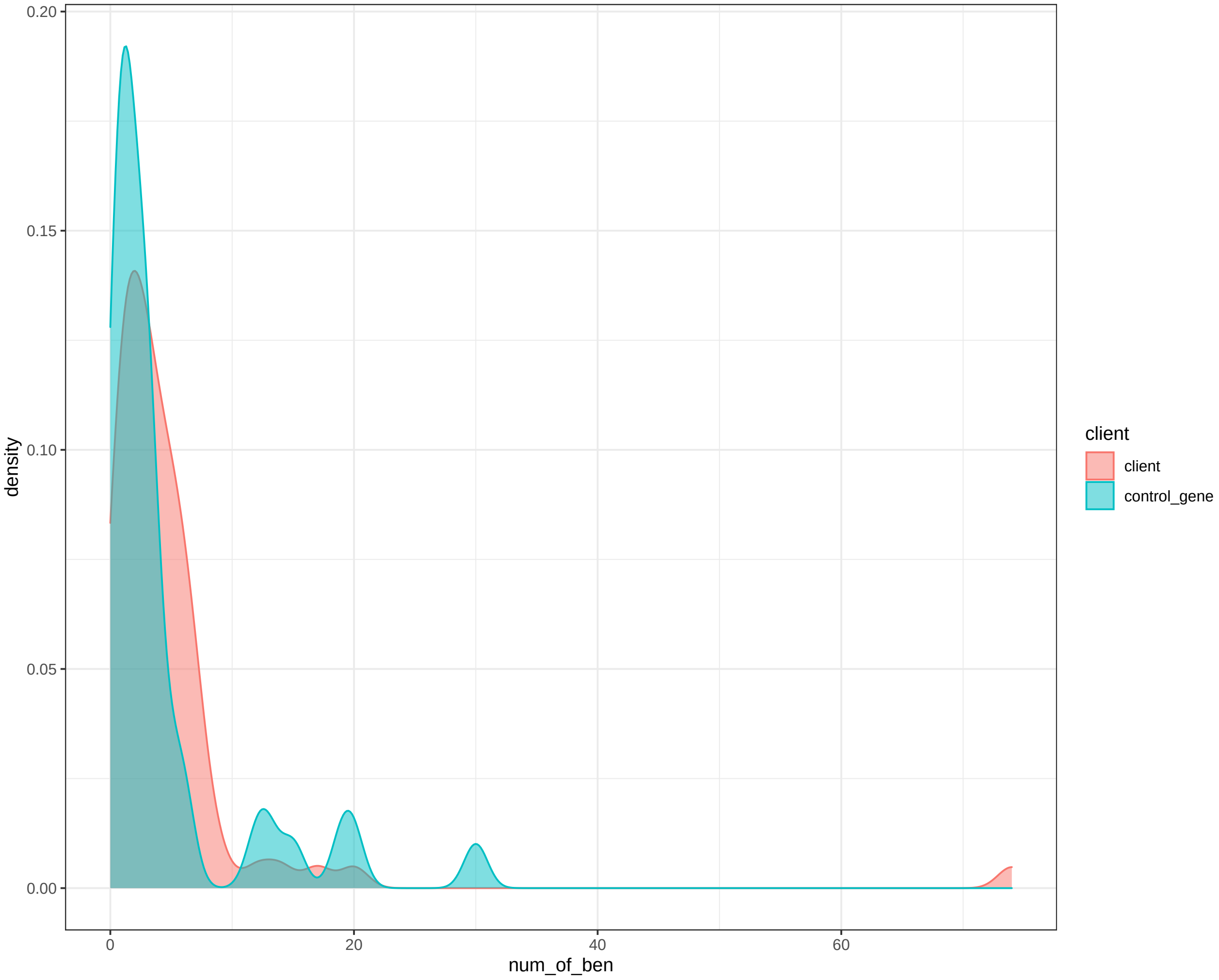
```
ggplot(hsc70_summary, aes(x = number_of_pat, y = num_of_ben, color = client))+
  geom_point()+
  geom_abline(intercept = 0, slope = 1)+
  theme_bw()
```



```
ggplot(hsc70_summary, aes(x = number_of_pat, fill = client, color = client))+
  geom_density(alpha = 0.5)+
  theme_bw()
```



```
ggplot(hsc70_summary, aes(x = num_of_ben, fill = client, color = client))+
  geom_density(alpha = 0.5)+
  theme_bw()
```



```
wilcox.test(hsc70_summary$number_of_mut ~ hsc70_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsc70_summary$number_of_mut by hsc70_summary$client
## W = 2108, p-value = 0.0002972
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsc70_summary$number_of_pat ~ hsc70_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsc70_summary$number_of_pat by hsc70_summary$client
## W = 2259, p-value = 5.313e-06
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsc70_summary$num_of_ben ~ hsc70_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  hsc70_summary$num_of_ben by hsc70_summary$client
## W = 1756, p-value = 0.1224
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(hsc70_summary$proportion_of_pat ~ hsc70_summary$client)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hsc70_summary$proportion_of_pat by hsc70_summary$client
## W = 2068, p-value = 0.0006545
## alternative hypothesis: true location shift is not equal to 0
```