FIG_S2

Compiled: May 15, 2022

```
library(ggplot2)
library(reshape2)
require(ggplot2)
require(plyr)
require(gridExtra)
require(grid)
library(RColorBrewer)
library(ggpubr)
```

```
my_pal <- c("#FDBF6F", "#FF7F00", "#A6CEE3", "#1F78B4", "#FB9A99", "#E31A1C", "#B2DF8A",
"#33A02C")
my_pal_2 <- c("#FDBF6F", "#A6CEE3", "#FB9A99", "#B2DF8A")

my_comparisons_1 <- list(c("GQV", "Platypus"), c("GQV", "GATK_hc"), c("GQV", "Freebayes"
))
my_comparisons_2 <- list(c("GQV_merged", "Platypus_merged"), c("GQV_merged", "GATK_hc_me
rged"),
    c("GQV_merged", "Freebayes_merged"))
```

Fig. S2

iPSC cells

Fig. S2d - Genotype vs read depth

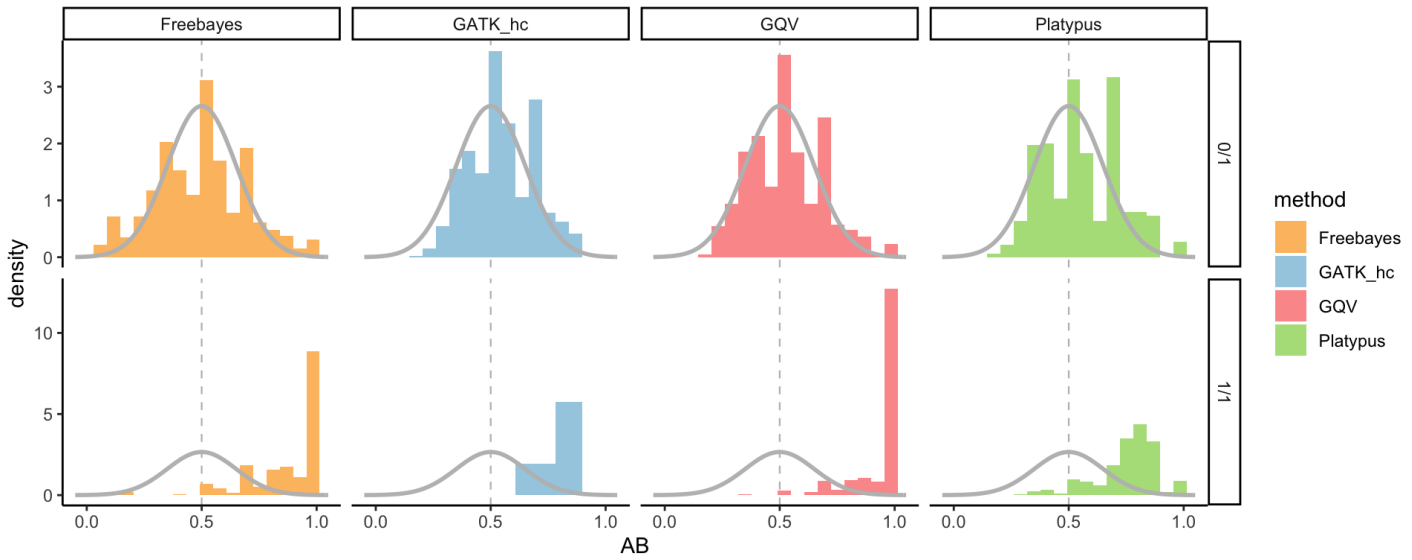
```
df6 <- read.csv("results_analysis/GSE77288.genotypes.all.txt", sep = " ", header = F)
names(df6) <- c("method", "REP", "var_type", "VAR", "GT", "AB", "AC", "DP")
head(df6)
```

method	REP	var_type	VAR	GT	AB	AC	DP
Freebayes	well_A01	indels	chr19:38836461:GA>G	0/1	0.60	6	10
Freebayes	well_A01	indels	chr20:44885754:T>TCGC	0/1	0.25	4	16
Freebayes	well_A01	indels	chr3:51385310:CGGAGGA>C	0/1	1.00	3	3
Freebayes	well_A01	indels	chr6:3224590:GA>G	1/1	1.00	5	5
Platypus	well_A01	indels	chr11:75572751:AT>A	1/1	1.00	3	3
Platypus	well_A01	indels	chr17:42998447:TG>T	1/1	0.79	15	19

```
df6 <- df6[df6$GT %in% c("0/1", "1/1"), ]
df6 <- df6[df6$DP > 5, ]

g1 <- ggplot(df6[df6$var_type == "indels", ], aes(x = AB)) + geom_vline(xintercept = 0.5
, color = "gray",
    size = 0.4, linetype = 2) + geom_histogram(bins = 20, aes(fill = method, y = ..densi
ty..)) +
    stat_function(fun = dnorm, args = list(mean = 0.5, sd = 0.15), col = "gray", size =
1) + scale_x_continuous(limits = c(-0.05,
    1.05), breaks = c(0, 0.5, 1)) + facet_grid(GT ~ method, scales = "free_y") + scale_f
ill_manual(values = my_pal_2) +
    theme_classic()

g2 <- ggplot(df6[df6$var_type == "snps", ], aes(x = AB)) + geom_vline(xintercept = 0.5,
 color = "gray",
    size = 0.4, linetype = 2) + geom_histogram(bins = 20, aes(fill = method, y = ..densi
ty..)) +
    stat_function(fun = dnorm, args = list(mean = 0.5, sd = 0.15), col = "gray", size =
1) + scale_x_continuous(limits = c(-0.05,
    1.05), breaks = c(0, 0.5, 1)) + facet_grid(GT ~ method, scales = "free_y") + scale_f
ill_manual(values = my_pal_2) +
    theme_classic()

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/S2d.pdf")
g1
g2
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1  # INDELS
```



```
g2  # SNPS
```

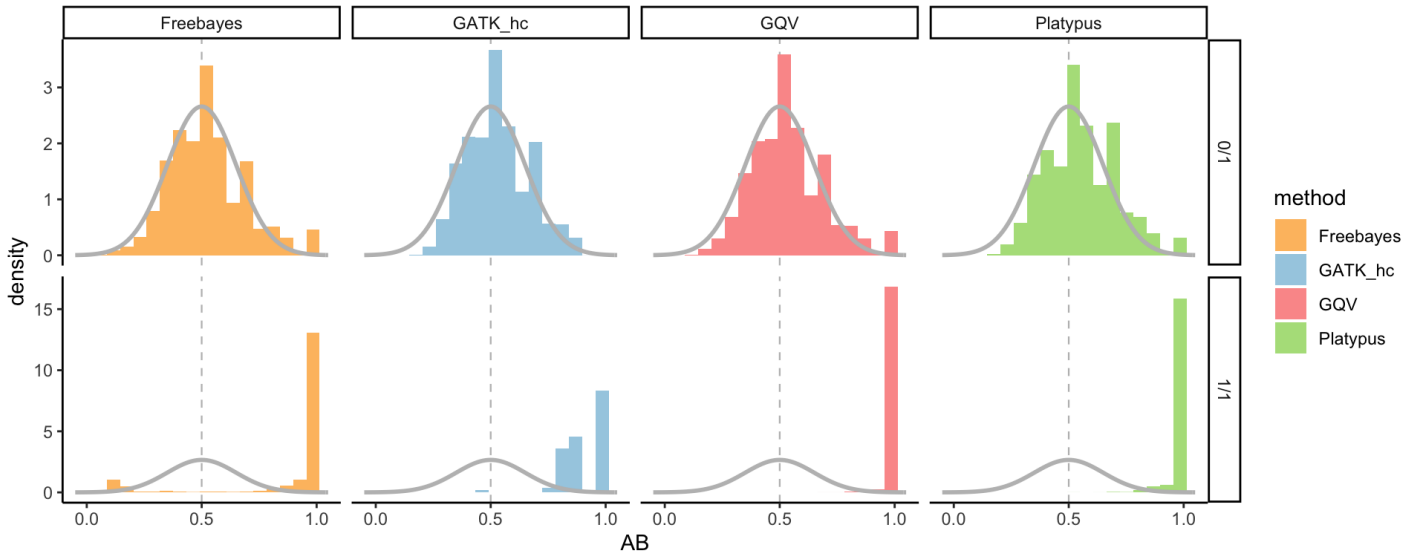


Fig. S2c - Read coverage

```
df4 <- read.csv("results_analysis/GSE77288.all_sites_cov.txt", sep = " ", header = F)
names(df4) <- c("count", "IND", "rep", "method", "VAR", "FP_cutoff", "cov")

head(df4)
```

count	IND	rep	method	VAR	FP_cutoff	cov
5	NA19098	r1	Freebayes	indels	maxF1	10
2	NA19098	r1	Freebayes	indels	maxF1	11
1	NA19098	r1	Freebayes	indels	maxF1	12
2	NA19098	r1	Freebayes	indels	maxF1	13
1	NA19098	r1	Freebayes	indels	maxF1	14
2	NA19098	r1	Freebayes	indels	maxF1	15

```
df4$group_id = paste0(df4$method, df4$VAR, df4$FP_cutoff)
df4 <- df4[df4$cov >= 3, ]

g1 <- ggplot(df4, aes(x = cov, y = count, group = group_id)) + geom_smooth(aes(color = m
ethod, fill = method),
    span = 0.3, level = 0.99) + labs(x = "Coverage (DP)", y = "TP count") + facet_grid(V
AR ~ FP_cutoff,
    scales = "free") + scale_color_manual(values = my_pal_2) + scale_fill_manual(values
 = my_pal_2) +
    theme_classic() + scale_x_log10() + theme(panel.border = element_rect(colour = "blac
k", fill = NA,
    size = 0.75))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/S2c.pdf")
g1
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```

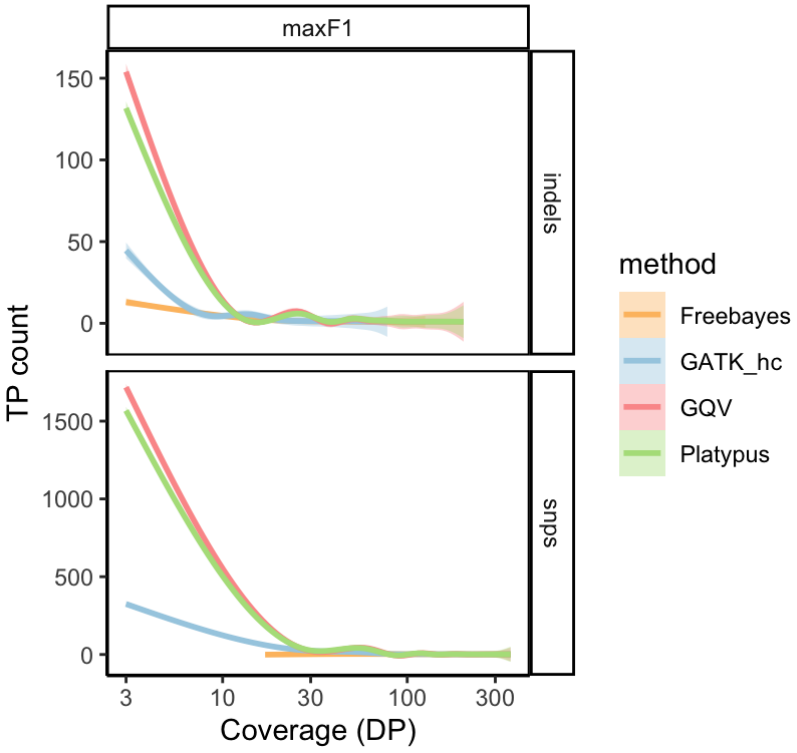


Fig. S2a - SENSPE (Non difficult regions)

```
df2 <- read.csv("results/2022/Apr/analysis/GSE77288.roc_all_methods.non_difficult_region
s.txt",
    sep = " ", header = F)

names(df2) <- c("Method", "IND", "benchmark", "sample", "rep", "Var_Type", "score", "TP_
base", "FP")

head(df2)
```

Method	IND	benchmark	sample	rep	Var_Type	score	TP_base	FP
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	10.67	159	5
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	10.21	669	20
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	9.85	1850	50
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	8.71	6440	200
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	6.91	12236	500
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	3.01	20265	2000

```
df2$sample_id = paste0(df2$Method, df2$sample, df2$rep)
df2 <- df2[order(df2$score), ]

g1 <- ggplot(df2, aes(y = TP_base, x = FP, group = sample_id)) + geom_path(aes(color = M
ethod, group = sample_id),
    size = 0.8, alpha = 0.8) + facet_grid(Var_Type ~ IND, scales = "free") + scale_color
_brewer(c(my_pal,
    c("gray", "orange"))) + theme_classic() + theme(panel.border = element_rect(colour =
"black",
    fill = NA, size = 0.5))

df2 <- df2[df2$FP %in% c(5, 20, 50, 200, 500), ]

df_tile <- df2 %>%
    dplyr::group_by(benchmark, Var_Type) %>%
    dplyr::summarise(FP = FP, n = max(TP_base))

df_tile$SPE_85 = pmin(df_tile$FP * 5.67, df_tile$n)
df_tile$SPE_95 = pmin(df_tile$FP * 19, df_tile$n)

df_tile <- melt(df_tile, id.vars = c("benchmark", "Var_Type", "FP", "n"))
df_tile <- as.data.frame(unique(df_tile))
head(df_tile)
```

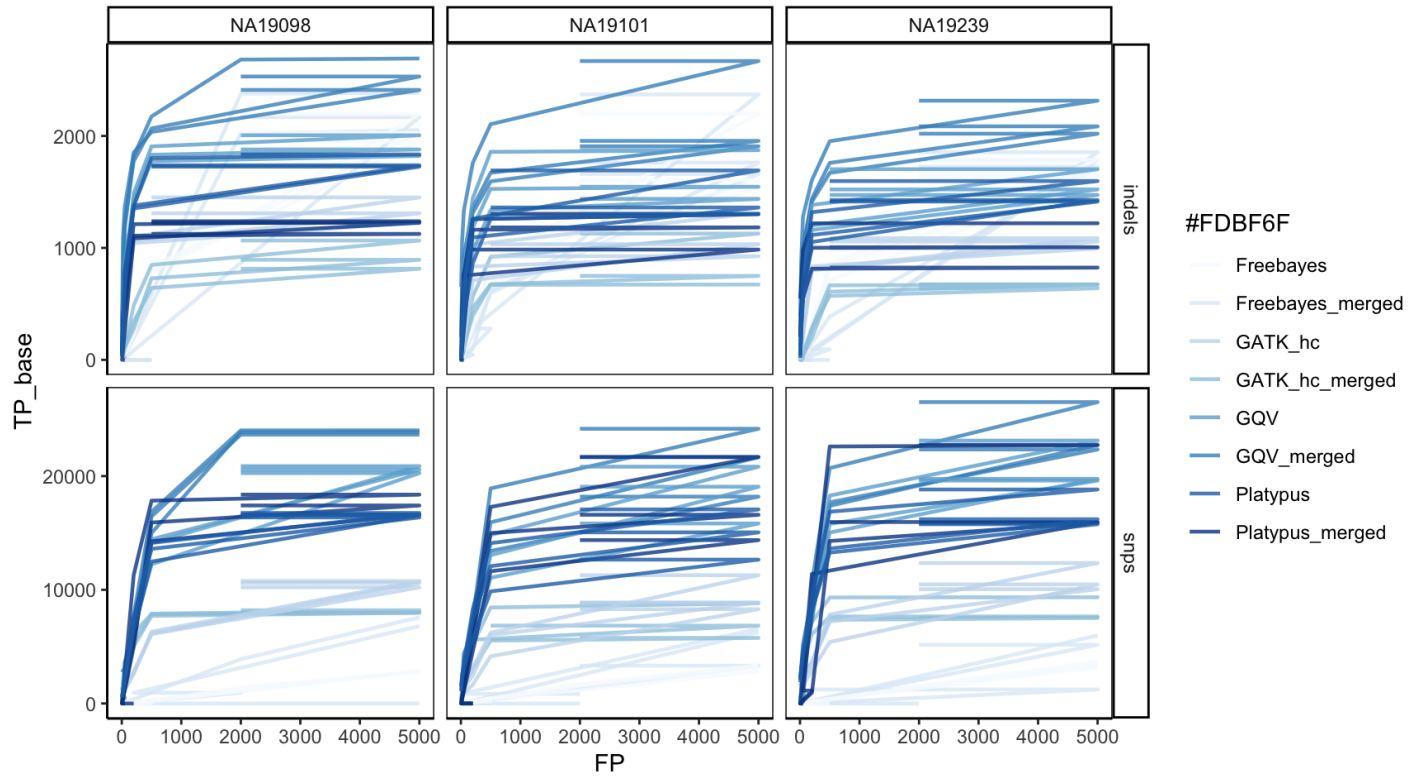
	benchmark	Var_Type	FP	n	variable	value
1	1000G	indels	500	2174	SPE_85	2174.00
10	1000G	indels	200	2174	SPE_85	1134.00

	benchmark	Var_Type	FP	n	variable	value
28	1000G	indels	50	2174	SPE_85	283.50
45	1000G	indels	20	2174	SPE_85	113.40
65	1000G	indels	5	2174	SPE_85	28.35
361	1000G	snps	500	22603	SPE_85	2835.00

```
g2 <- ggplot(df2, aes(y = TP_base, x = as.factor(FP))) + labs(x = "FP cutoff", y = "TP c
ount") +
    geom_tile(aes(x = factor(FP), y = 1, height = Inf, width = 0.8), fill = "gray80", da
ta = df_tile,
        alpha = 0.25) + geom_tile(data = df_tile, aes(x = as.factor(FP), y = value, colo
r = variable),
    height = 0, width = 0.8) + geom_boxplot(aes(color = Method), outlier.shape = NA) + f
acet_grid(Var_Type ~
    IND, scales = "free") + scale_color_manual(values = c(my_pal, c("gray", "orange")))
 + theme_classic() +
    theme(panel.border = element_rect(colour = "black", fill = NA, size = 0.75))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/S2a.pdf")
g2
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```



g2

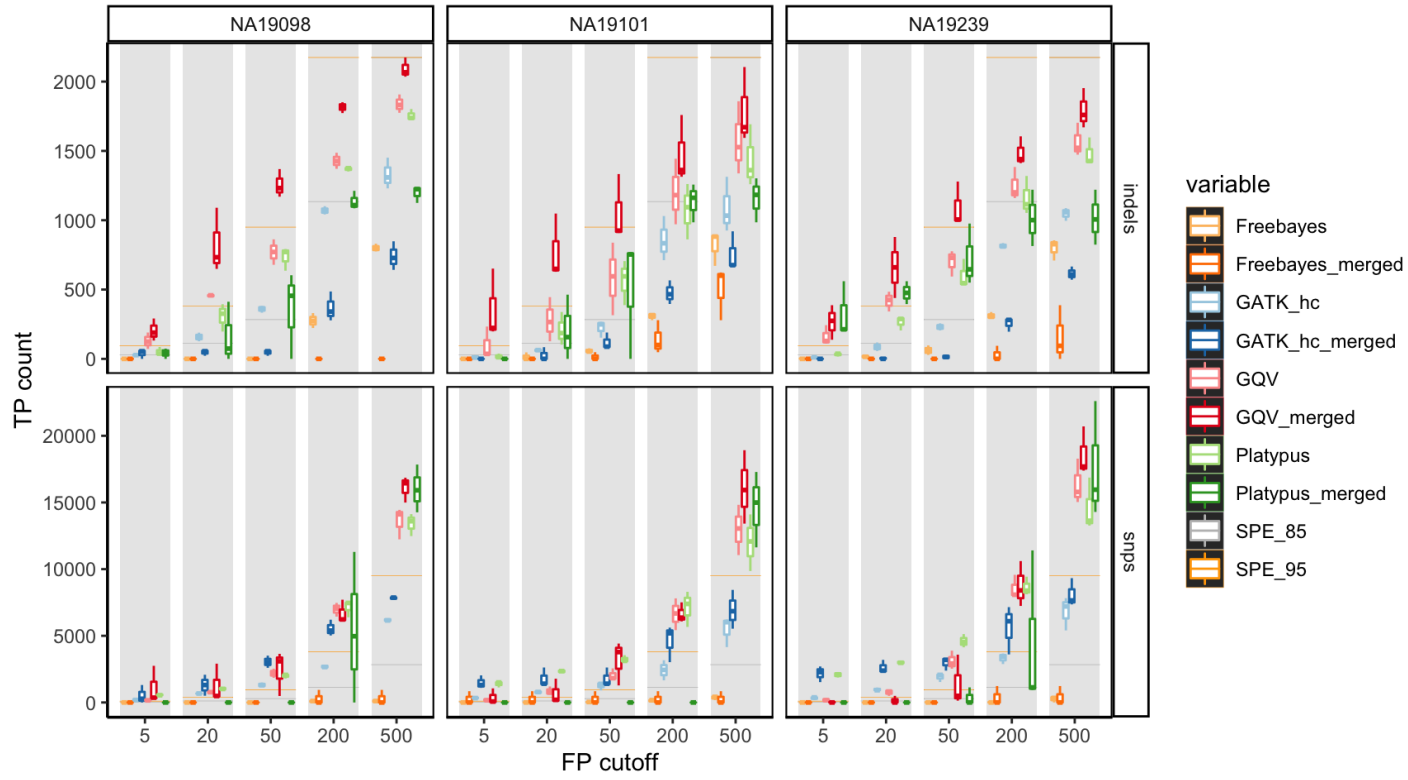
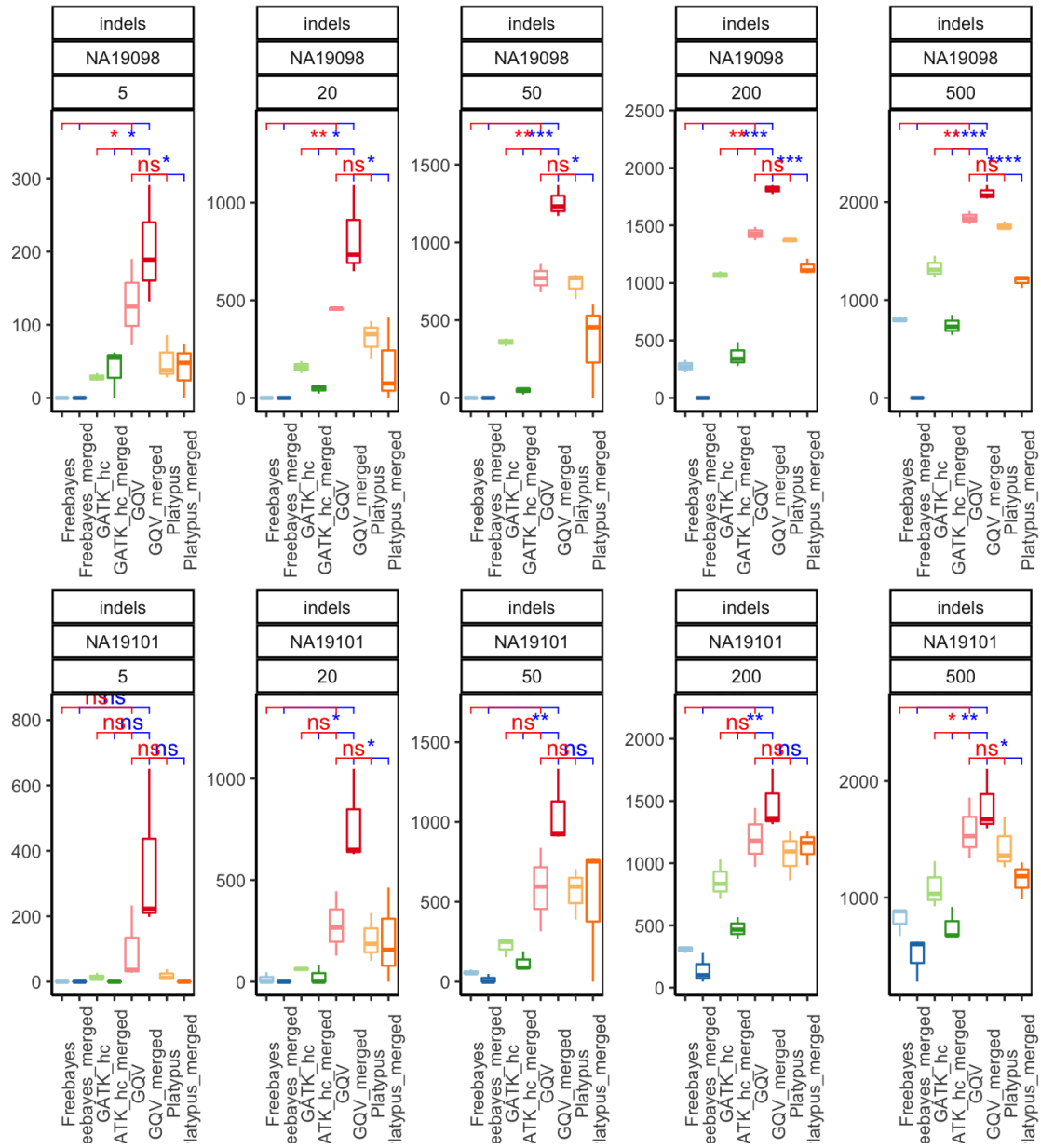
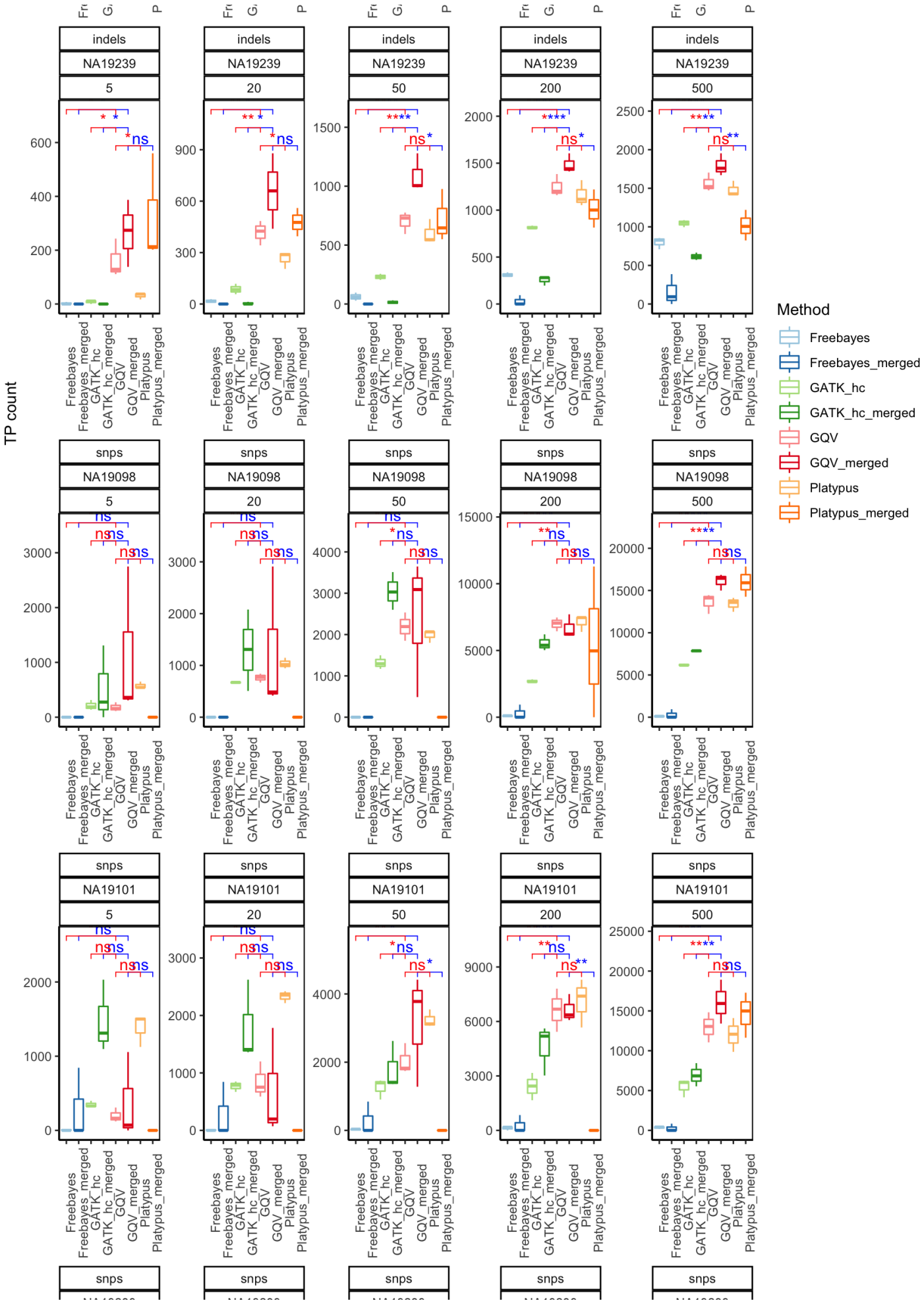


Fig. S2a - SENSPE (Non difficult regions)

p-value

```
args <- list(var.equal = F, alternative = "greater")

g3 <- ggplot(df2, aes(y = TP_base, x = Method)) + labs(x = "FP cutoff", y = "TP count")
 + geom_boxplot(aes(color = Method),
    outlier.shape = NA) + facet_wrap(Var_Type ~ IND + as.factor(FP), ncol = 5, scales =
"free") +
    scale_color_brewer(palette = "Paired") + theme_classic() + theme(panel.border = elem
ent_rect(colour = "black",
    fill = NA, size = 0.75)) + theme(axis.text.x = element_text(angle = 90)) + stat_comp
are_means(aes(group = Method,
    label = ..p.signif..), color = "blue", method = "t.test", method.args = args, compar
isons = my_comparisons_2) +
    stat_compare_means(aes(group = Method, label = ..p.signif..), color = "red", method
 = "t.test",
        method.args = args, comparisons = my_comparisons_1)


g3
```




TP count
indels
snps
NA19239
NA19098
NA19101
5
20
50
200
500
Method
Freebayes
Freebayes_merged
GATK_hc
GATK_hc_merged
GQV
GQV_merged
Platypus
Platypus_merged

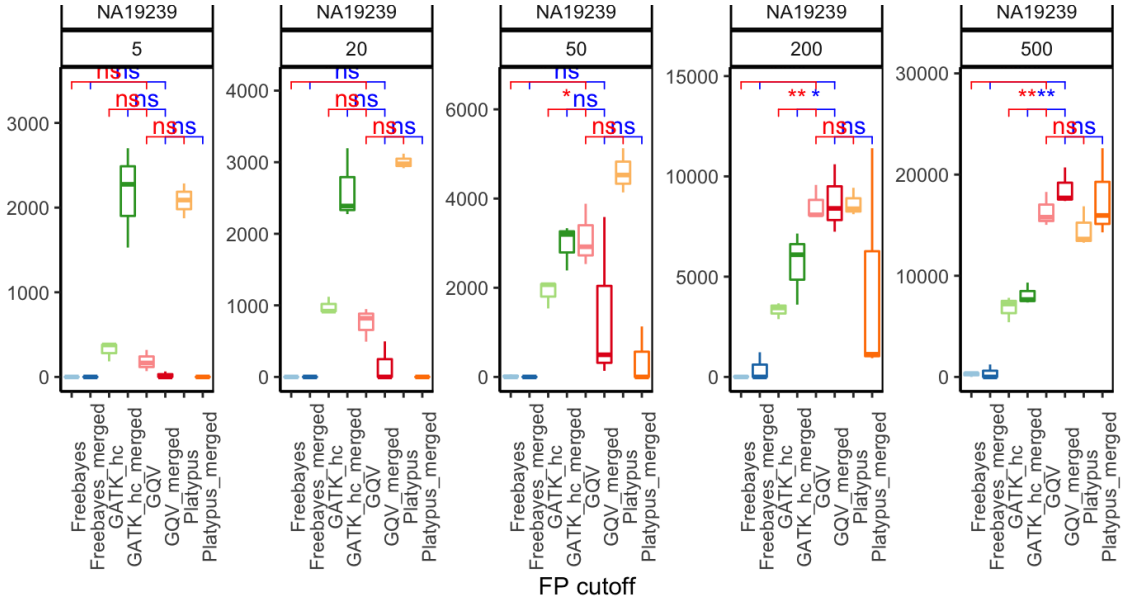


Fig. S2b - SENSPE (Difficult regions)

```
df2 <- read.csv("results/2022/Apr/analysis/GSE77288.roc_all_methods.difficult_regions.tx
t", sep = " ",
    header = F)

names(df2) <- c("Method", "IND", "benchmark", "sample", "rep", "Var_Type", "score", "TP_
base", "FP")

head(df2)
```

Method	IND	benchmark	sample	rep	Var_Type	score	TP_base	FP
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	10.53	192	5
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	10.01	742	20
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	9.71	1431	50
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	8.27	4268	200
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	6.15	8397	500
GQV	NA19098	1000G	r1	NA19098_GQV_indels_rr3	snps	3.01	10835	2000

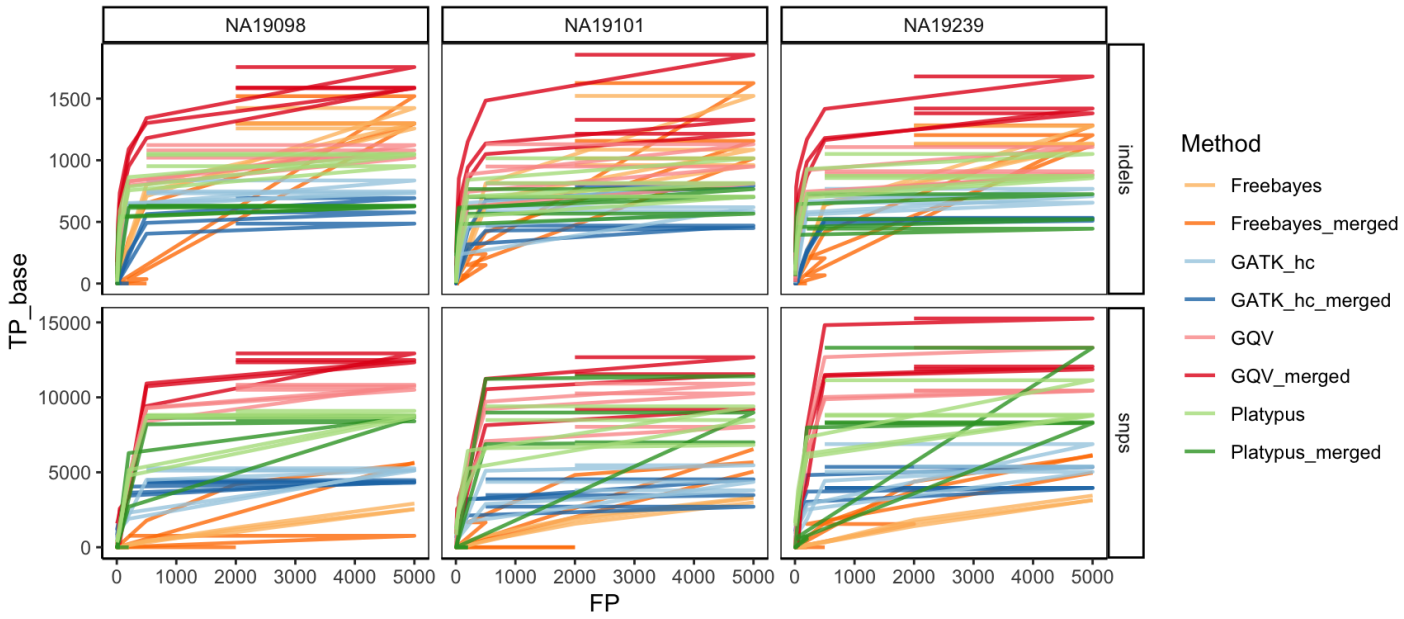
```
df2$sample_id = paste0(df2$Method, df2$sample, df2$rep)
df2 <- df2[order(df2$score), ]

g1 <- ggplot(df2, aes(y = TP_base, x = FP, group = sample_id)) + geom_path(aes(color = M
ethod, group = sample_id),
    size = 0.8, alpha = 0.8) + facet_grid(Var_Type ~ IND, scales = "free") + scale_color
_manual(values = c(my_pal,
    c("gray", "orange"))) + theme_classic() + theme(panel.border = element_rect(colour =
"black",
    fill = NA, size = 0.5))

df2 <- df2[df2$FP %in% c(5, 20, 50, 200, 500), ]

df_tile <- df2 %>%
    dplyr::group_by(benchmark, Var_Type) %>%
    dplyr::summarise(FP = FP, n = max(TP_base))

df_tile$SPE_85 = pmin(df_tile$FP * 5.67, df_tile$n)
df_tile$SPE_95 = pmin(df_tile$FP * 19, df_tile$n)

df_tile <- melt(df_tile, id.vars = c("benchmark", "Var_Type", "FP", "n"))
df_tile <- as.data.frame(unique(df_tile))
head(df_tile)
```

	benchmark	Var_Type	FP	n	variable	value
1	1000G	indels	500	1484	SPE_85	1484.00
10	1000G	indels	200	1484	SPE_85	1134.00
28	1000G	indels	50	1484	SPE_85	283.50
45	1000G	indels	20	1484	SPE_85	113.40
69	1000G	indels	5	1484	SPE_85	28.35
361	1000G	snps	500	14818	SPE_85	2835.00

```
g2 <- ggplot(df2, aes(y = TP_base, x = as.factor(FP))) + labs(x = "FP cutoff", y = "TP c
ount") +
    geom_tile(aes(x = factor(FP), y = 1, height = Inf, width = 0.8), fill = "gray80", da
ta = df_tile,
        alpha = 0.25) + geom_tile(data = df_tile, aes(x = as.factor(FP), y = value, colo
r = variable),
    height = 0, width = 0.8) + geom_boxplot(aes(color = Method), outlier.shape = NA) + f
acet_grid(Var_Type ~
    IND, scales = "free") + scale_color_manual(values = c(my_pal, c("gray", "orange")))
 + theme_classic() +
    theme(panel.border = element_rect(colour = "black", fill = NA, size = 0.75))


pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/S2b.pdf")
g2
dev.off()
```

```
## quartz_off_screen
##                   2
```

```
g1
```




```
g2
```

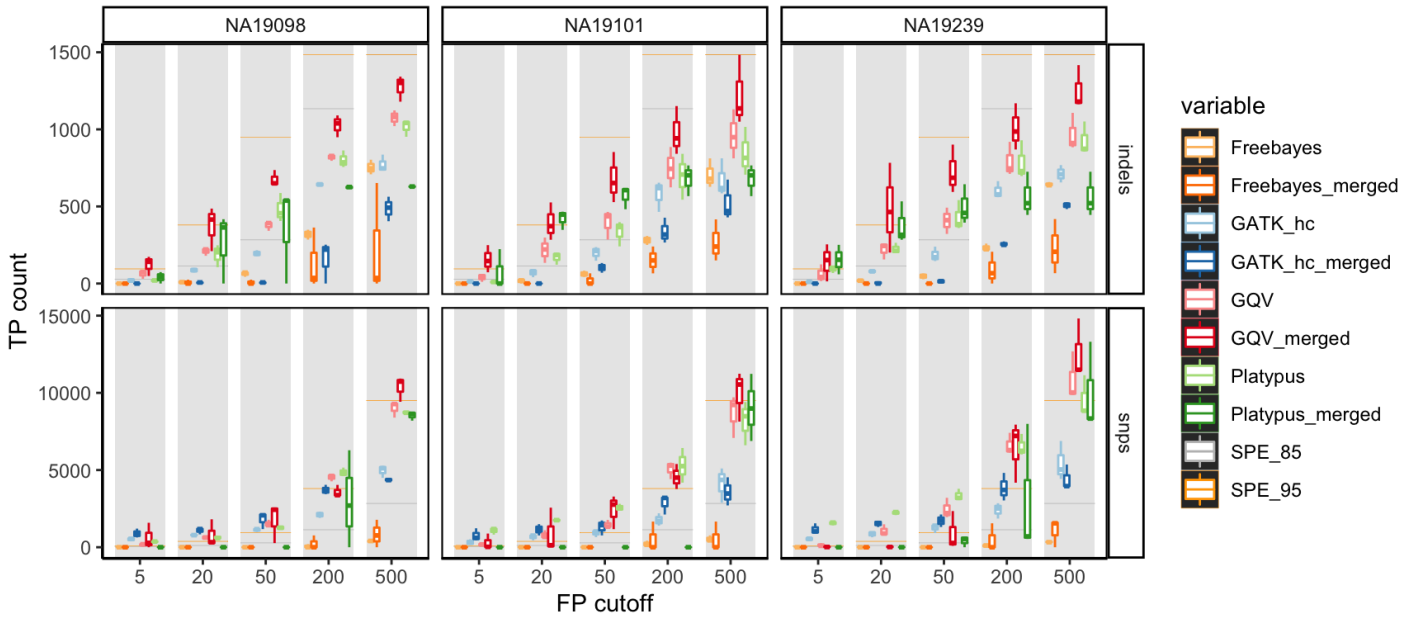
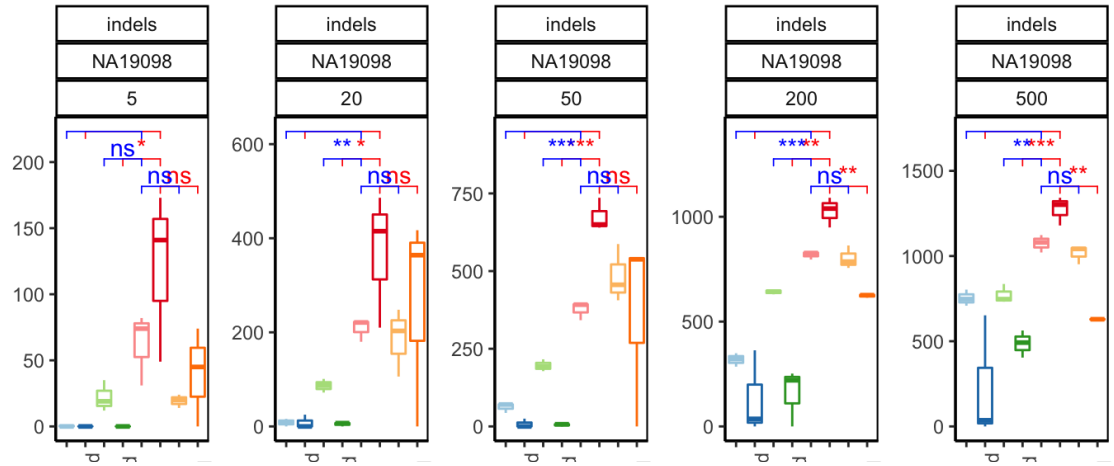
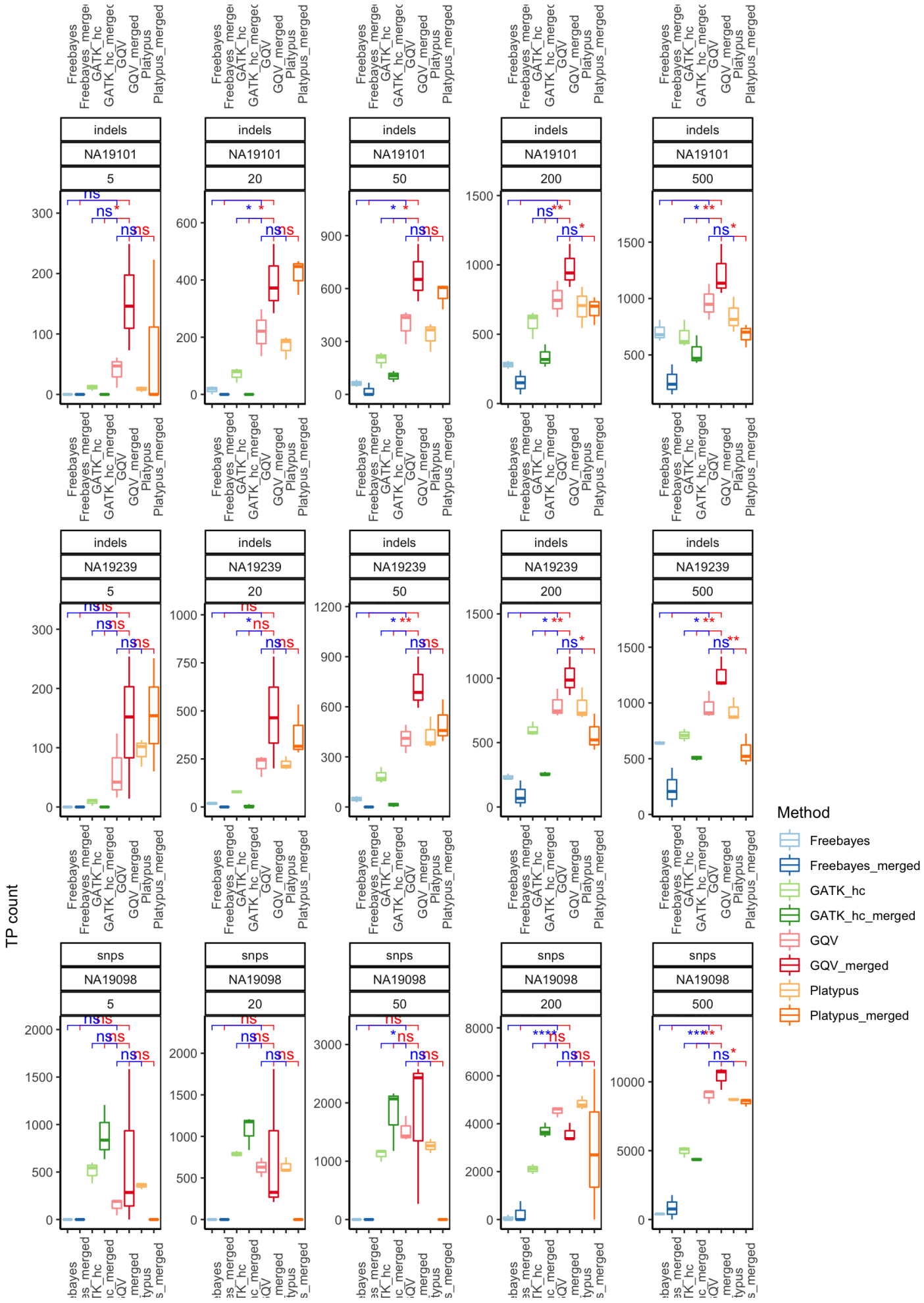


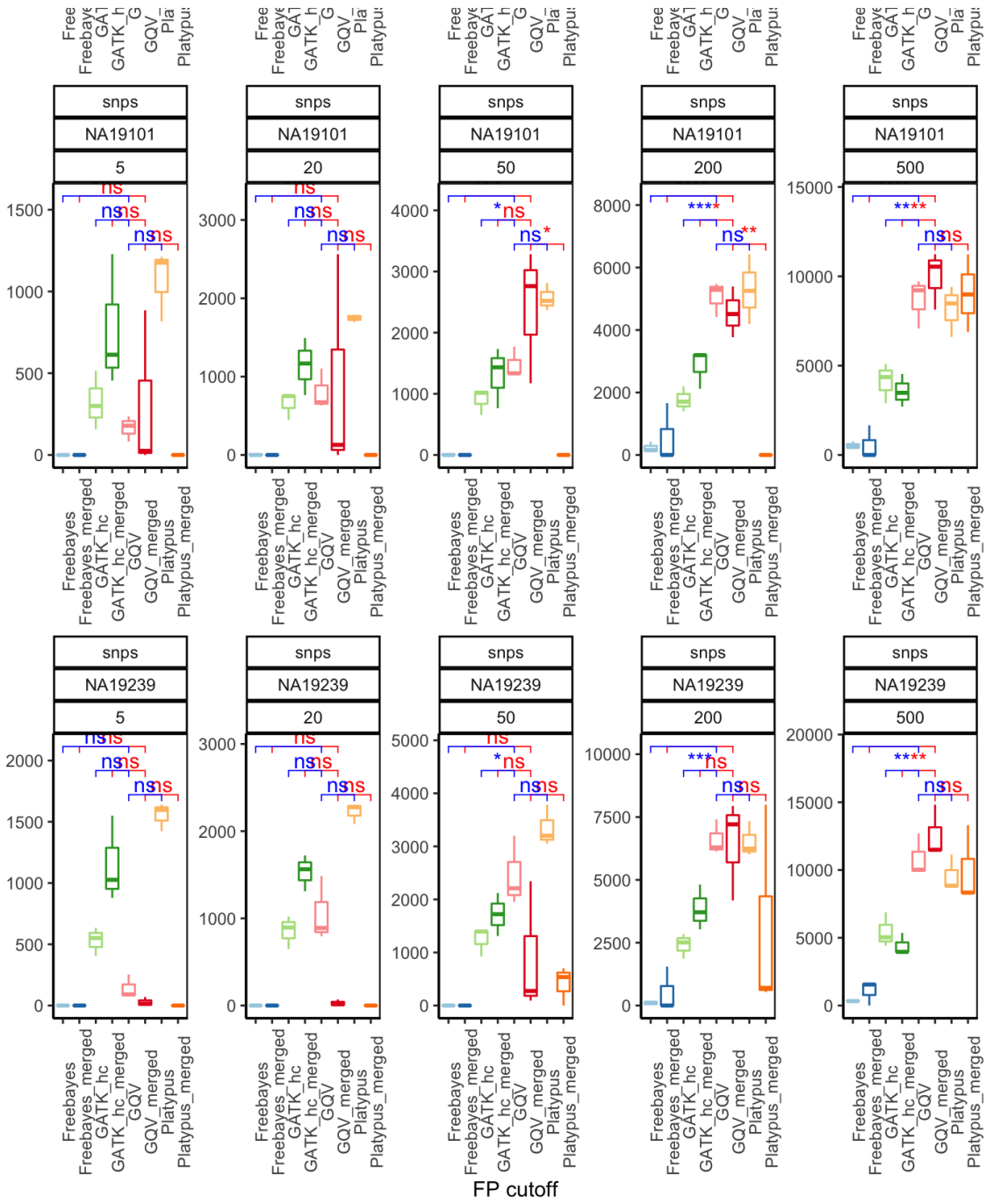
Fig. S2b - SENSPE (Difficult regions)

p-value

```
g3 <- ggplot(df2, aes(y = TP_base, x = Method)) + labs(x = "FP cutoff", y = "TP count")
 + geom_boxplot(aes(color = Method),
    outlier.shape = NA) + facet_wrap(Var_Type ~ IND + as.factor(FP), ncol = 5, scales =
"free") +
    scale_color_brewer(palette = "Paired") + theme_classic() + theme(panel.border = elem
ent_rect(colour = "black",
    fill = NA, size = 0.75)) + theme(axis.text.x = element_text(angle = 90)) + stat_comp
are_means(aes(group = Method,
    label = ..p.signif..), method = "t.test", method.args = args, color = "red", compari
sons = my_comparisons_2) +
    stat_compare_means(aes(group = Method, label = ..p.signif..), method = "t.test", met
hod.args = args,
        color = "blue", comparisons = my_comparisons_1)


g3
```




TP count
indels
snps
NA19101
NA19239
NA19098
5
20
50
200
500
Freebayes
Freebayes_merged
GATK_hc
GATK_hc_merged
GQV
GQV_merged
Platypus
Platypus_merged
Method
ns
*
**
***
****


snps
NA19101
5
20
50
200
500
NA19239
Freebayes
Freebayes_merged
GATK_hc
GATK_hc_merged
GQV
GQV_merged
Platypus
Platypus_merged
ns
*
**
***
****
FP cutoff