FIG_2

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_merged"),
    c("GQV_merged", "Freebayes_merged"))
```

GM12878 cells

Fig. 2e - Genotype vs read depth

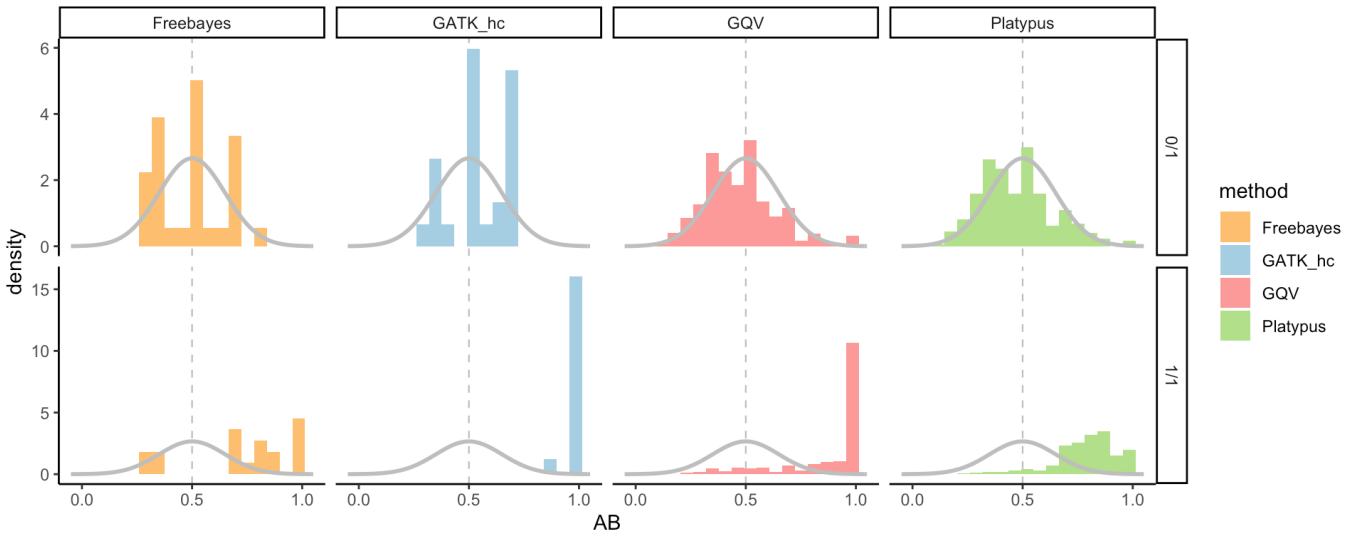
```
df6 <- read.csv("results_analysis/GM12878_scRNA_downsample.genotypes.all.txt", sep = " ", header = F)
names(df6) <- c("method", "REP", "var_type", "VAR", "GT", "AB", "AC", "DP")
head(df6)
```

method	REP	var_type	VAR	GT	AB	AC	DP
Freebayes	1	snps	chr1:631862:G>A	1/1	1	3	3
Freebayes	1	snps	chr1:944296:G>A	1/1	1	4	4
Freebayes	1	snps	chr1:1318756:G>A	1/1	1	2	2
Freebayes	1	snps	chr1:1375595:C>T	1/1	1	2	2
Freebayes	1	snps	chr1:1817973:T>G	0/1	1	2	2
Freebayes	1	snps	chr1:3780326:A>C	0/1	1	2	2

```
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 = ..density..)) +
    stat_function(fun = dnorm, args = list(mean = 0.5, sd = 0.15), col = "gray", size = 1) + scale_x_continuous(l
imits = c(-0.05,
    1.05), breaks = c(0, 0.5, 1)) + facet_grid(GT ~ method, scales = "free_y") + scale_fill_manual(values = my_pa
l_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 = ..density..)) +
    stat_function(fun = dnorm, args = list(mean = 0.5, sd = 0.15), col = "gray", size = 1) + scale_x_continuous(l
imits = c(-0.05,
    1.05), breaks = c(0, 0.5, 1)) + facet_grid(GT ~ method, scales = "free_y") + scale_fill_manual(values = my_pa
l_2) +
    theme_classic()

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/2e.pdf")
print(g1)
print(g2)
dev.off()
```

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

```
print(g1)
```




```
print(g2)
```

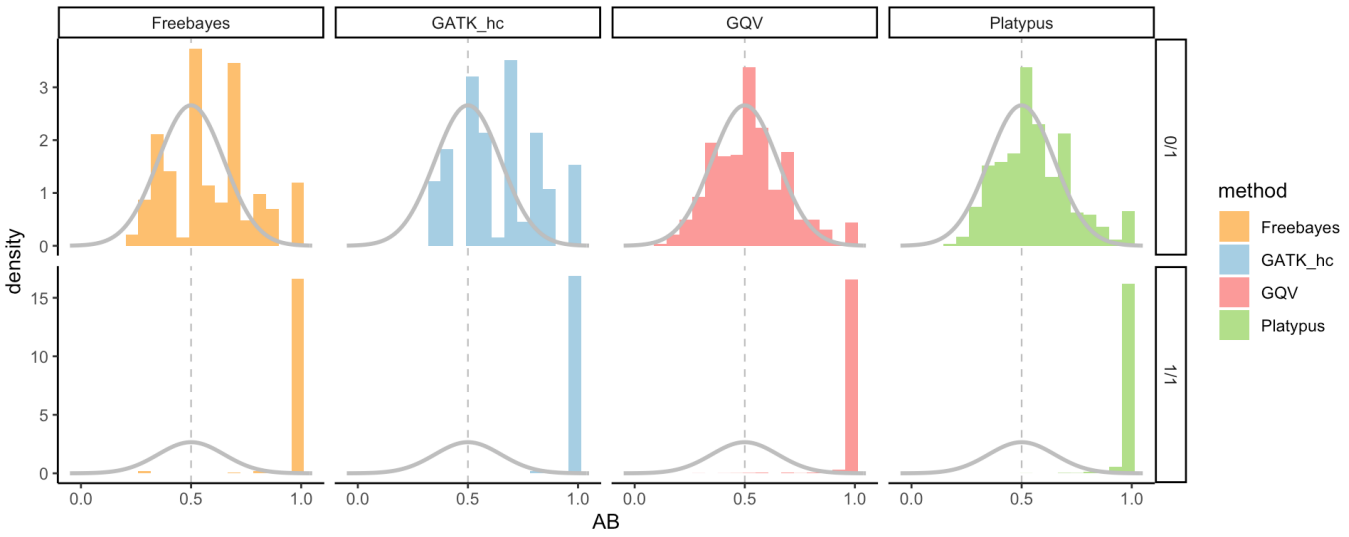


Fig. 2d - Read depth

```
df4 <- read.csv("results_analysis/GM12878_scRNA_downsample.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
109	GM12878	10	Freebayes	indels	maxF1	2
52	GM12878	10	Freebayes	indels	maxF1	3
26	GM12878	10	Freebayes	indels	maxF1	4
16	GM12878	10	Freebayes	indels	maxF1	5
13	GM12878	10	Freebayes	indels	maxF1	6
4	GM12878	10	Freebayes	indels	maxF1	7

```
df4$group_id = paste0(df4$method, df4$VAR, df4$FP_cutoff)
df4 <- df4[df4$cov >= 2, ]

df5 <- df4 %>%
    dplyr::filter(cov < 5 & method %in% c("GQV", "Platypus") & VAR == "indels") %>%
    dplyr::group_by(rep, VAR, method) %>%
    dplyr::summarise(n = sum(count))

df5 %>%
    dplyr::group_by(VAR) %>%
    dplyr::do(w = wilcox.test(n ~ method, data = ., paired = FALSE)) %>%
    dplyr::summarise(VAR, Wilcox = w$p.value)
```

VAR	Wilcox
indels	0

```
df5 <- df4 %>%
    dplyr::filter(cov < 5 & method %in% c("GQV", "Freebayes") & VAR == "snps") %>%
    dplyr::group_by(rep, VAR, method) %>%
    dplyr::summarise(n = sum(count))

df5 %>%
    dplyr::group_by(VAR) %>%
    dplyr::do(w = wilcox.test(n ~ method, data = ., paired = FALSE)) %>%
    dplyr::summarise(VAR, Wilcox = w$p.value)
```

VAR	Wilcox
snps	1.8e-06

```
g1 <- ggplot(df4, aes(x = cov, y = count, group = group_id)) + geom_smooth(aes(color = method, fill = method),
    span = 0.3, level = 0.99) + labs(x = "Coverage (DP)", y = "TP count") + facet_grid(VAR ~ FP_cutoff,
    scales = "free") + scale_color_manual(values = my_pal_2) + scale_fill_manual(values = my_pal_2) +
    theme_minimal() + scale_x_log10(breaks = c(2, 5, 10, 15)) + theme(panel.border = element_rect(colour = "blac
k",
    fill = NA, size = 0.75))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/2d.pdf")
g1
dev.off()
```

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

```
g1
```

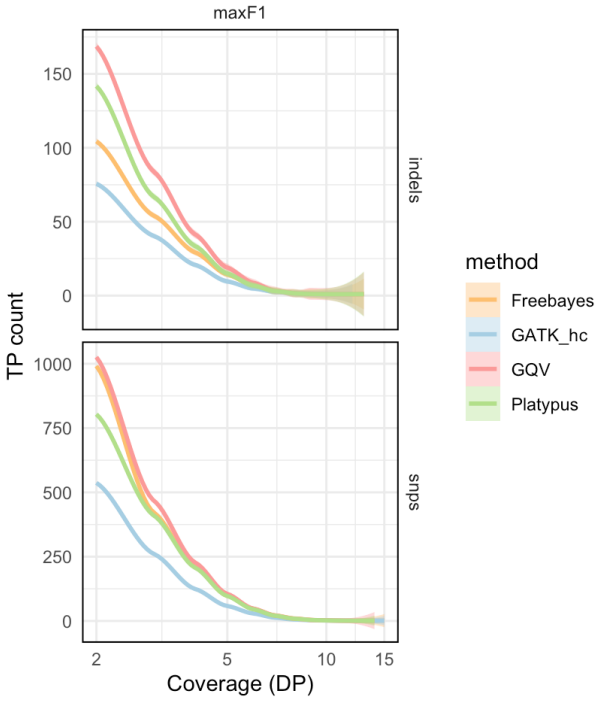


Fig. 2a - SENSPE (Non difficult regions)

```
df2 <- read.csv("results/2022/Apr/analysis/GM12878_scRNA_downsample.roc_all_methods.non_difficult_regions.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	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	10.26	146	1
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	9.82	399	5
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	9.16	788	10
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	8.31	1090	20
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	7.63	1327	30
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	6.66	1907	40

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

df3 <- df2[df2$FP %in% c(1, 5, 10, 20, 30, 50, 100), ]

df_tile <- df3 %>%
    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	ALL	INDELS	100	471	SPE_85	471.00
34	ALL	INDELS	50	471	SPE_85	283.50
55	ALL	INDELS	30	471	SPE_85	170.10
73	ALL	INDELS	20	471	SPE_85	113.40
151	ALL	INDELS	10	471	SPE_85	56.70
177	ALL	INDELS	5	471	SPE_85	28.35

```
g1 <- ggplot(df3, aes(y = TP_base, x = as.factor(FP))) + labs(x = "FP cutoff", y = "TP count") +
    geom_tile(data = df_tile, aes(x = as.factor(FP), y = 1), fill = "gray95", height = Inf, width = 0.8) +
    geom_tile(data = df_tile, aes(x = as.factor(FP), y = value, color = variable), height = 0, width = 0.8) +

geom_boxplot(aes(color = Method), outlier.shape = NA) + facet_wrap(Var_Type ~ benchmark, ncol = 3,
    scales = "free") + scale_color_manual(values = c(my_pal, c("gray", "orange"))) + scale_fill_manual(values = c
(my_pal,
    c("gray", "orange"))) + theme_classic() + theme(text = element_text(size = 10)) + theme(panel.border = elemen
t_rect(colour = "black",
    fill = NA, size = 0.75))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/2a.pdf")
g1
dev.off()
```

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

```
g1
```

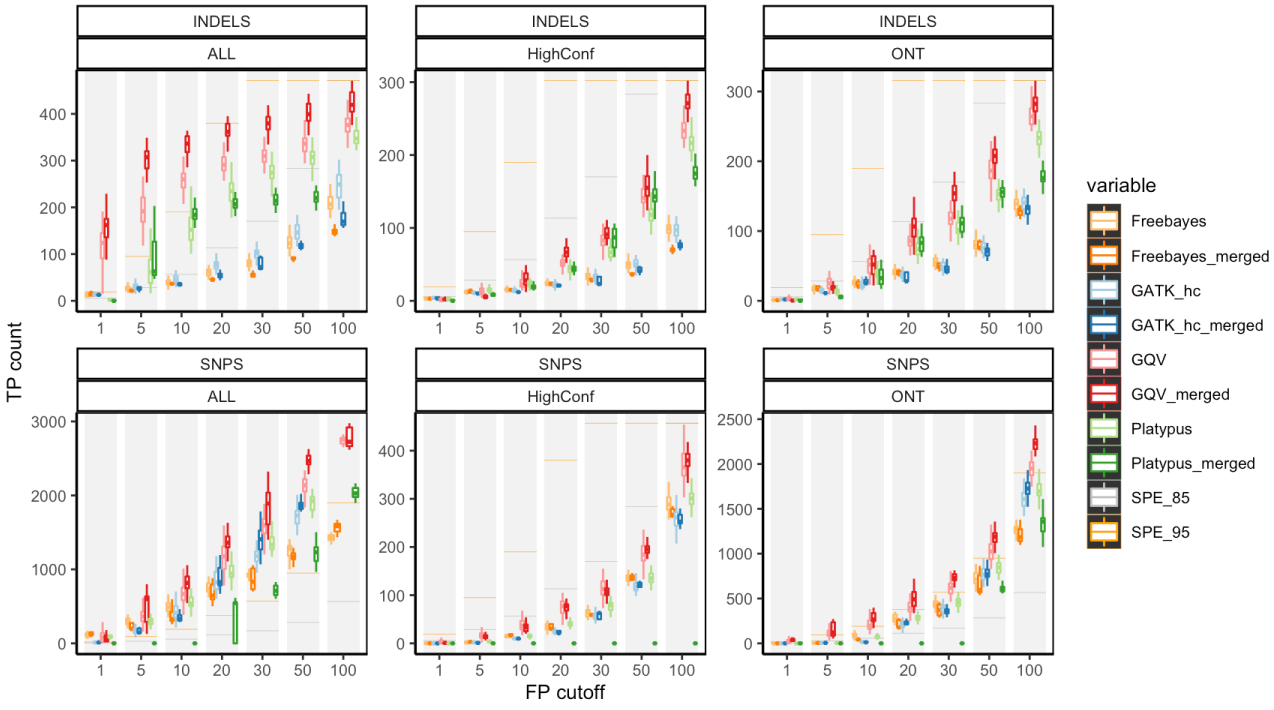
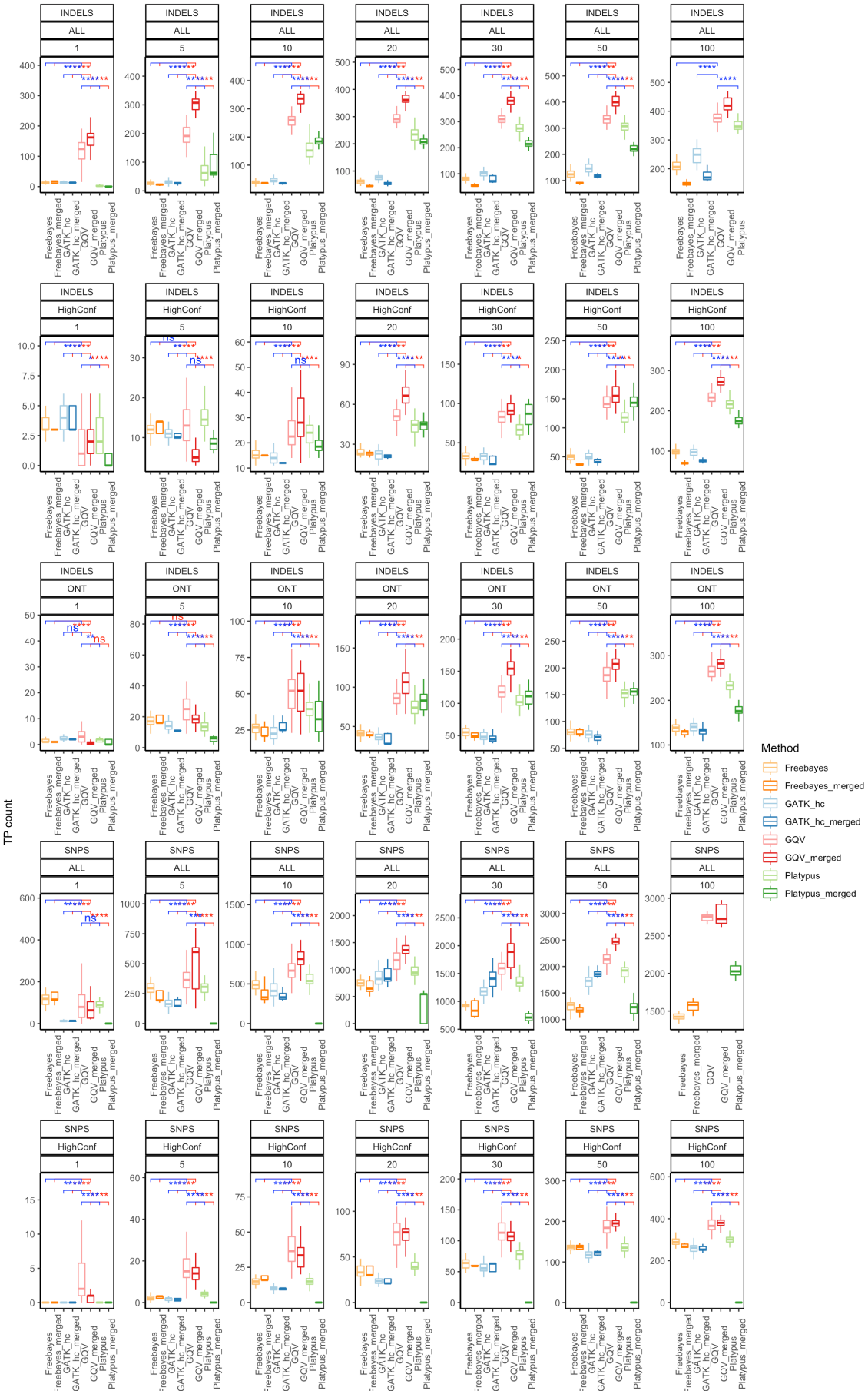


Fig. 2b - SENSPE (Difficult regions)

p-values

```
g2 <- ggplot(df3, 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 ~ benchmark + as.factor(FP), ncol = 7, 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)) + theme(plot.margin = margin(3, 3, 3.5, 3.2, "cm")) + theme(axis.text.x = element_te
xt(angle = 90)) +
    stat_compare_means(aes(group = Method, label = ..p.signif..), method = "wilcox.test", color = "red",
        comparisons = my_comparisons_2) + stat_compare_means(aes(group = Method, label = ..p.signif..),
    method = "wilcox.test", color = "blue", comparisons = my_comparisons_1)
g2
```


INDELS
SNPS
ALL
HighConf
ONT
1
5
10
20
30
50
100
TP count
Method
Freebayes
Freebayes_merged
GATK_hc
GATK_hc_merged
GQV
GQV_merged
Platypus
Platypus_merged
ns

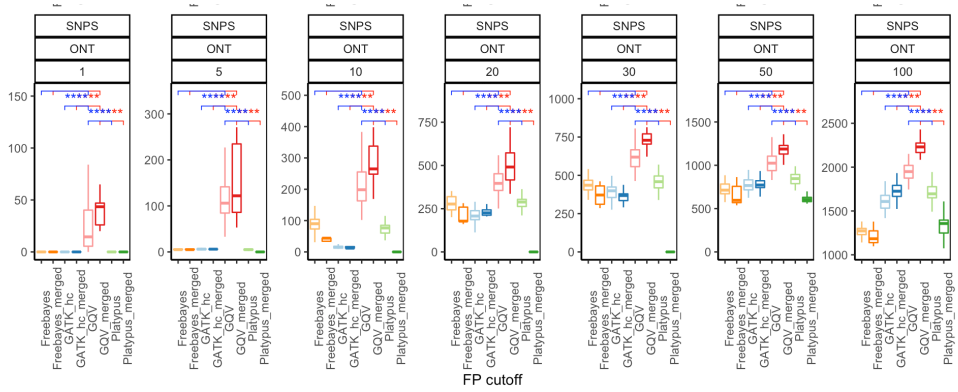


Fig. 2b - SENSPE (Difficult regions)

```
df2 <- read.csv("results/2022/Apr/analysis/GM12878_scRNA_downsample.roc_all_methods.difficult_regions.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	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	9.82	80	1
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	8.65	189	5
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	8.20	220	10
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	6.63	390	20
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	5.67	482	30
GQV	GM12878_smartseq2	ALL	1.DS_seed_11	ALL_GQV_SNPS_r1.DS_seed_11	SNPS	4.23	546	40

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

df3 <- df2[df2$FP %in% c(1, 5, 10, 20, 30, 50, 100), ]

df_tile <- df3 %>%
    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	ALL	INDELS	100	250.5	SPE_85	250.50
21	ALL	INDELS	50	250.5	SPE_85	250.50
53	ALL	INDELS	30	250.5	SPE_85	170.10
88	ALL	INDELS	20	250.5	SPE_85	113.40
162	ALL	INDELS	10	250.5	SPE_85	56.70

	benchmark	Var_Type	FP	n	variable	value
200	ALL	INDELS	5	250.5	SPE_85	28.35

```
# df3 <- df3[df3$Method %in% c('GQV', 'Platypus', 'Freebayes', 'GATK_hc'), ]
g1 <- ggplot(df3, aes(y = TP_base, x = as.factor(FP))) + labs(x = "FP cutoff", y = "TP count") +
    geom_tile(data = df_tile, aes(x = as.factor(FP), y = 1), fill = "gray95", height = Inf, width = 0.8) +
    geom_tile(data = df_tile, aes(x = as.factor(FP), y = value, color = variable), height = 0, width = 0.8) +

geom_boxplot(aes(color = Method), outlier.shape = NA) + facet_wrap(Var_Type ~ benchmark, ncol = 3,
    scales = "free") + scale_color_manual(values = c(my_pal, c("gray", "orange"))) + theme_classic() +
    theme(text = element_text(size = 10)) + theme(panel.border = element_rect(colour = "black",
    fill = NA, size = 0.75))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/2b.pdf")
g1
dev.off()
```

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

```
g1
```

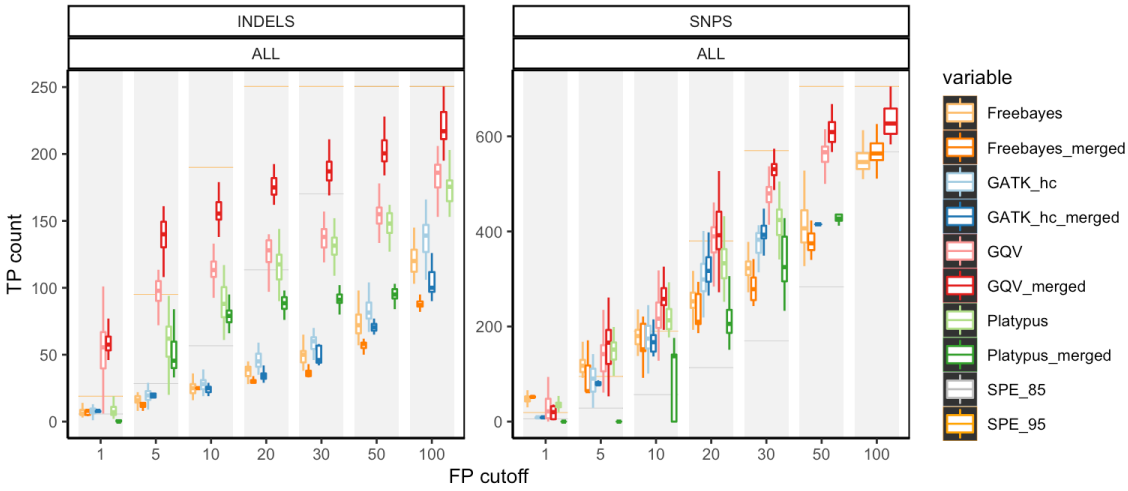
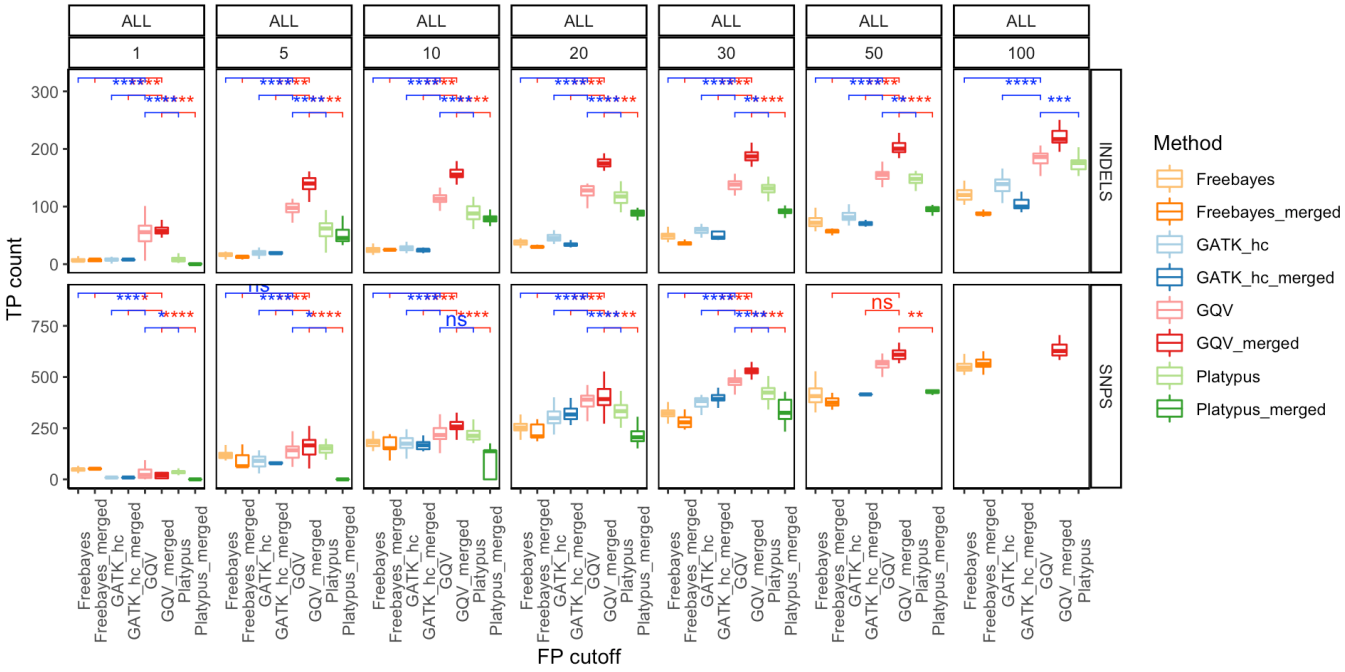


Fig. 2b - SENSPE (Difficult regions)

p-values

```
g2 <- ggplot(df3, aes(y = TP_base, x = Method)) + labs(x = "FP cutoff", y = "TP count") + geom_boxplot(aes(color
= Method),
    outlier.shape = NA) + facet_grid(Var_Type ~ benchmark + as.factor(FP), 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)) + stat_compare_means(aes(group = Method, label = ..p.signif..), method = "wilcox.tes
t",
    color = "red", comparisons = my_comparisons_2) + stat_compare_means(aes(group = Method, label = ..p.signi
f..),
    method = "wilcox.test", color = "blue", comparisons = my_comparisons_1) + theme(axis.text.x = element_text(an
gle = 90))


g2
```


ALL
1
5
10
20
30
50
100
INDELS
SNPS
TP count
FP cutoff
Method
Freebayes
Freebayes_merged
GATK_hc
GATK_hc_merged
GQV
GQV_merged
Platypus
Platypus_merged
ns