R Notebook

```r
suppressMessages(library("tidyverse"))
```

```
## Warning: replacing previous import by 'tidyr::%>%' when loading 'broom'

## Warning: replacing previous import by 'tidyr::gather' when loading 'broom'

## Warning: replacing previous import by 'tidyr::spread' when loading 'broom'
```

```r
library(stringi)
library(plotROC)
```

```
## Warning: replacing previous import by 'rlang::quo_name' when loading
## 'plotROC'
```

```r
variants=read_tsv("illumina_variants.tsv")
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Minimum = col_integer(),
##   Maximum = col_double(),
##   Length = col_integer(),
##   Change = col_character(),
##   Coverage = col_double(),
##   `Polymorphism Type` = col_character(),
##   `Variant Frequency` = col_character(),
##   replica = col_character(),
##   modality = col_character(),
##   freq = col_double()
## )
```

```r
barcode1v=read_tsv("BC01.variants.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```r
barcode1v$replica = 'a'
barcode2v=read_tsv("BC02.variants.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```
barcode2v$replica = 'b'
barcode3v=read_tsv("BC03.variants.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```
barcode3v$replica = 'c'
minion_variants=rbind(barcode1v, barcode2v, barcode3v)
```

```
minion_variants %>%
    filter(Qual == 0) %>%
    write_tsv(path="minion_variants.tsv")
```

```
barcode1=read_tsv("BC01.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```
barcode1$replica = 'a'
barcode2=read_tsv("BC02.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```
barcode2$replica = 'b'
barcode3=read_tsv("BC03.freqs.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer()
## )
```

```
barcode3$replica = 'c'
minion_all=rbind(barcode1, barcode2, barcode3)
```

```
minion_all %>%
    filter(Qual == 0) %>%
    write_tsv(path="minion_wt_frequencies.tsv")
```

```
expectedpositions=read_tsv("expectedpositions.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Validated = col_character()
## )
```

```
barcode1snps=read_tsv("BC01.variants.0.03.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer(),
##   VariantCov = col_integer(),
##   ForwardVariantCov = col_integer(),
##   ReverseVariantCov = col_integer()
## )
```

```
barcode1snps$replica = 'a'
barcode2snps=read_tsv("BC02.variants.0.03.txt")
```

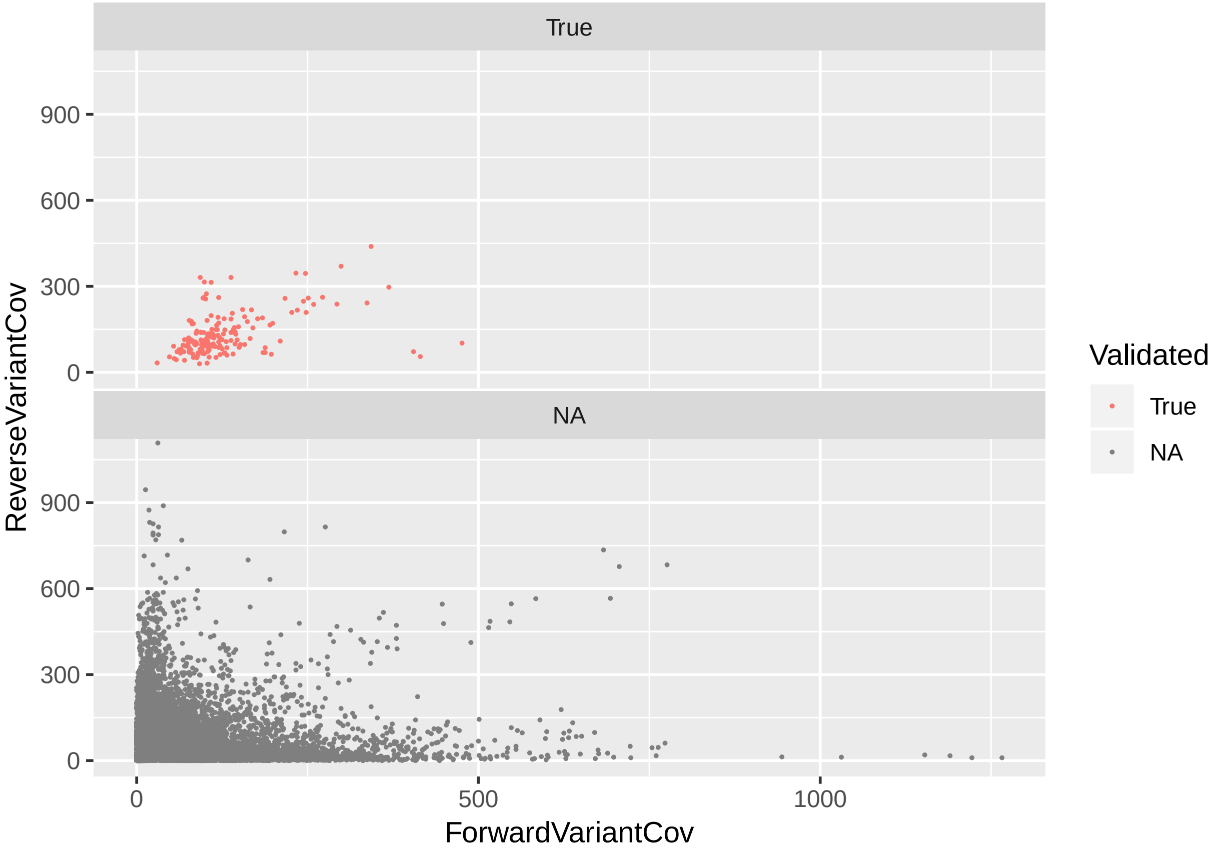
```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer(),
##   VariantCov = col_integer(),
##   ForwardVariantCov = col_integer(),
##   ReverseVariantCov = col_integer()
## )
```

```
barcode2snps$replica = 'b'
barcode3snps=read_tsv("BC03.variants.0.03.txt")
```

```
## Parsed with column specification:
## cols(
##   Pos = col_integer(),
##   Qual = col_integer(),
##   Freq = col_double(),
##   Ref = col_character(),
##   Base = col_character(),
##   UngappedCoverage = col_integer(),
##   TotalCoverage = col_integer(),
##   VariantCov = col_integer(),
##   ForwardVariantCov = col_integer(),
##   ReverseVariantCov = col_integer()
## )
```

```
barcode3snps$replica = 'c'
minion_all_variants=rbind(barcode1snps, barcode2snps, barcode3snps)
minion_all_variants_positions=left_join(minion_all_variants, expectedpositions, by=c("Pos"))
```

```
minion_all_variants_positions %>%
  ggplot(aes(x=ForwardVariantCov, y=ReverseVariantCov, color=Validated)) + geom_point(size=0.2) + facet_
```



```
minion_all_variants_positions %>%
  mutate(StrandAF = pmin(ForwardVariantCov, ReverseVariantCov) / pmax(ForwardVariantCov, ReverseVariant
```

```
## # A tibble: 12,414 x 13
##      Pos  Qual   Freq Ref   Base  UngappedCoverage TotalCoverage
##    <int> <int>  <dbl> <chr> <chr>            <int>         <int>
##  1    53     0 0.0438 C     T                 1987          2012
##  2    54     0 0.117  T     C                 1921          2012
##  3    55     0 0.0514 A     G                 1926          2012
##  4    56     0 0.0942 G     A                 1922          2012
##  5    57     0 0.0367 C     T                 1937          2012
##  6    59     0 0.04   A     G                 1975          2012
##  7    62     0 0.0312 G     A                 1957          2012
##  8    70     0 0.0344 A     G                 1974          2012
##  9    82     0 0.0553 G     T                 1719          2012
## 10    89     0 0.0336 G     A                 1937          2012
## # ... with 12,404 more rows, and 6 more variables: VariantCov <int>,
## #   ForwardVariantCov <int>, ReverseVariantCov <int>, replica <chr>,
## #   Validated <chr>, StrandAF <dbl>
```

```
nrow(minion_all_variants_positions %>% filter(Validated == 'True'))
```

```
## [1] 168
```

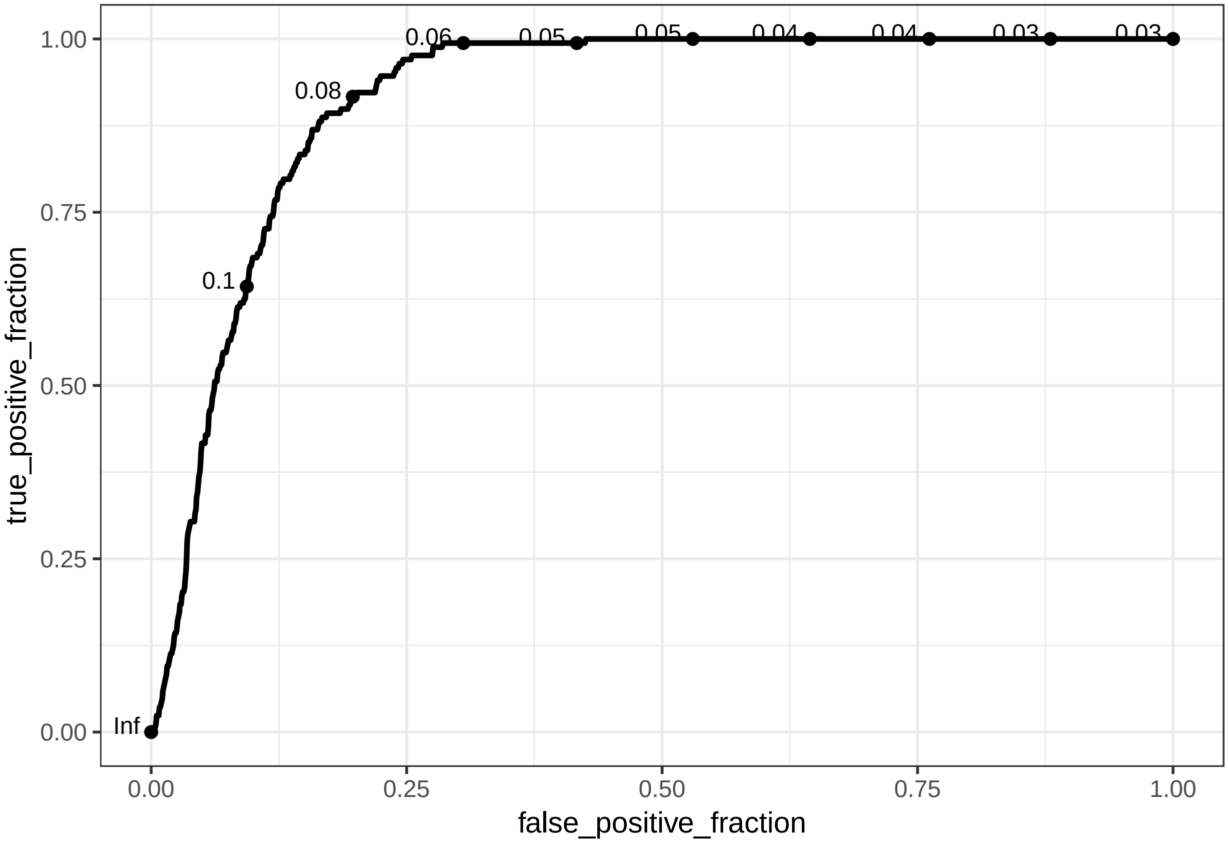
```
minion_all_variants_positions %>%
  mutate(StrandAF = pmin(ForwardVariantCov, ReverseVariantCov) / pmax(ForwardVariantCov, ReverseVariantC
  filter(ForwardVariantCov > 10) %>%
  filter(ReverseVariantCov > 10) %>%
  write_tsv("minion_variants_3pc_all.tsv")
```

```
forroc = minion_all_variants_positions %>%
  mutate(StrandAF = pmin(ForwardVariantCov, ReverseVariantCov) / pmax(ForwardVariantCov, ReverseVariantC
  filter(ForwardVariantCov > 10) %>%
  filter(ReverseVariantCov > 10) %>%
  mutate(D = ifelse(grepl("True", Validated), 1, 0))
```

```
nrow(forroc %>% filter(D==1))
```

```
## [1] 168
```

```
ggplot(forroc, aes(d = D, m = Freq)) + geom_roc(labelsize=3, labelround=2) + theme_bw()
```

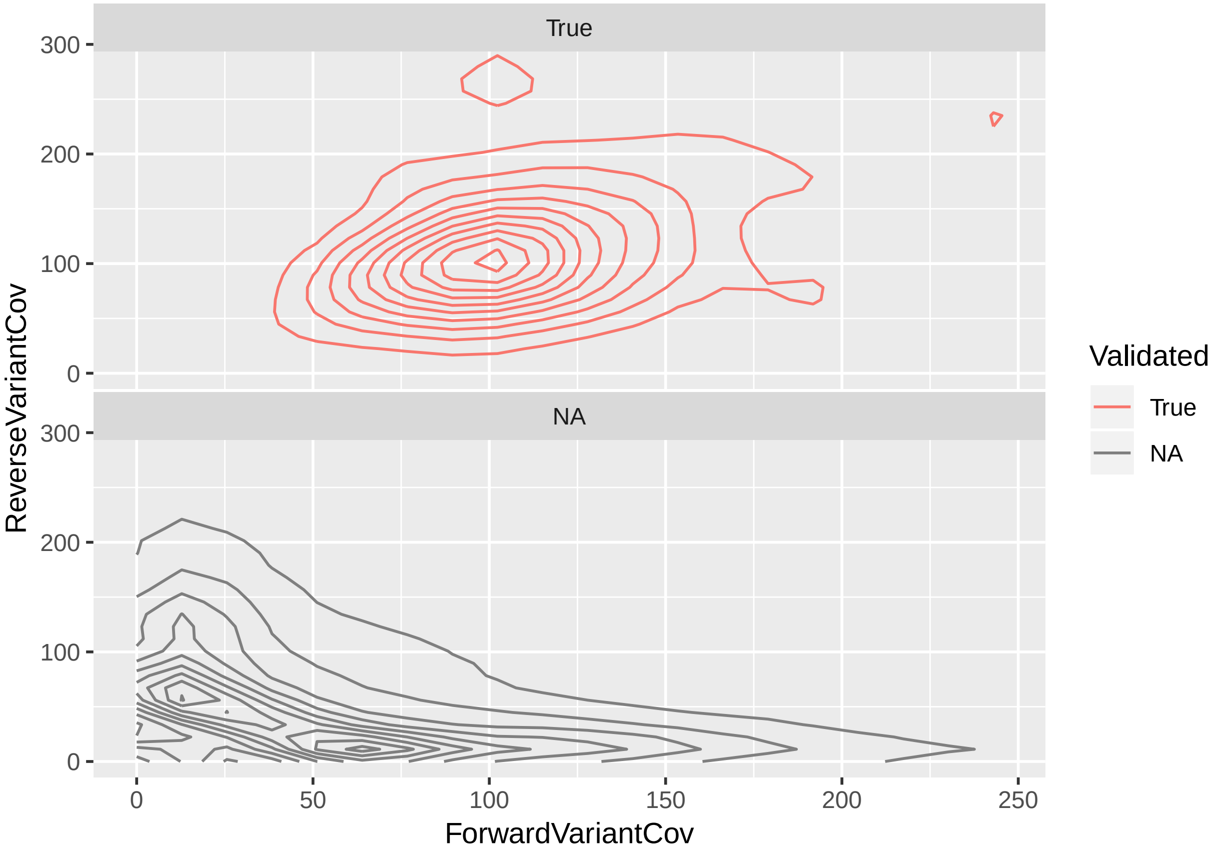


```
minion_all_variants_positions %>%
  mutate(StrandAF = pmin(ForwardVariantCov, ReverseVariantCov) / pmax(ForwardVariantCov, ReverseVariantC
  filter(ForwardVariantCov > 10) %>%
  filter(ReverseVariantCov > 10) %>%
  filter(Freq > 0.1) %>%
  group_by(Validated) %>%
  summarise(n=n())
```

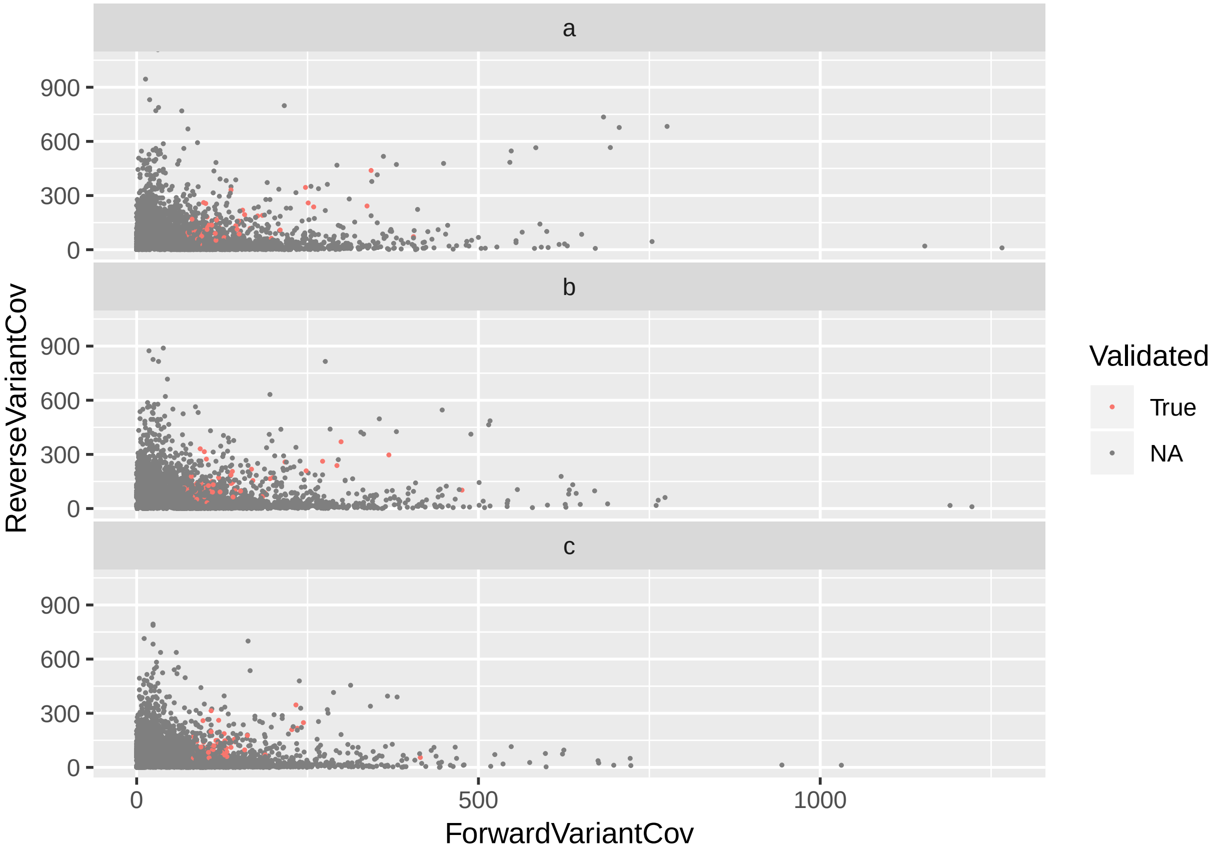
```
## # A tibble: 2 x 2
##   Validated     n
##   <chr>     <int>
## 1 True        109
## 2 <NA>        777
```

```
minion_all_variants_positions %>%
  mutate(StrandAF = pmin(ForwardVariantCov, ReverseVariantCov) / pmax(ForwardVariantCov, ReverseVariantC
  filter(ForwardVariantCov > 10) %>%
  filter(ReverseVariantCov > 10) %>%
  filter(StrandAF > 0.70) %>%
  write_tsv("minion_variants_3pc_0.7strandaf.tsv")
```

```
minion_all_variants_positions %>%
  ggplot(aes(x=ForwardVariantCov, y=ReverseVariantCov, color=Validated)) + geom_density2d() + facet_wrap
```



```
minion_all_variants_positions %>%
  ggplot(aes(x=ForwardVariantCov, y=ReverseVariantCov, color=Validated)) + geom_point(size=0.2) + facet_
```

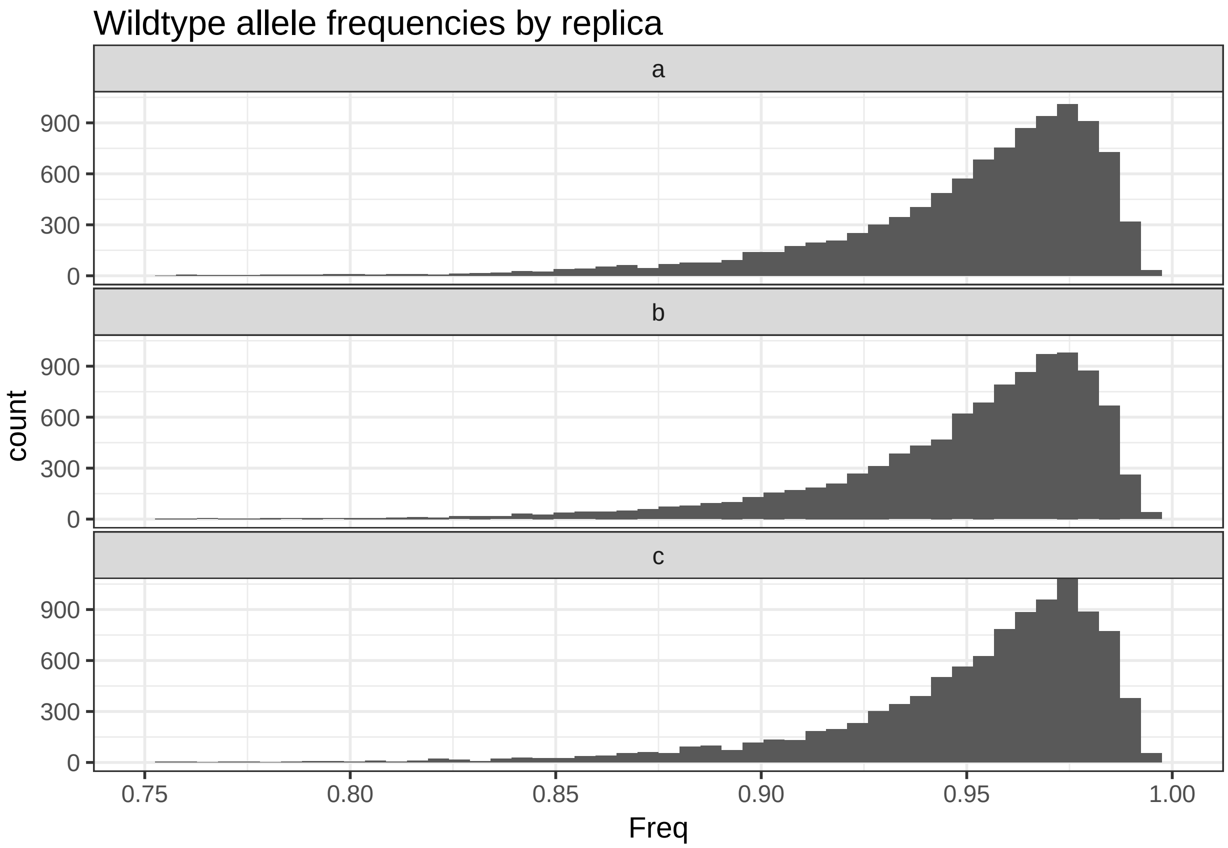


```
variant_positions=inner_join(minion_all, expectedpositions, by=c("Pos"), copy=T)
wildtype_positions=anti_join(minion_all, expectedpositions, by=c("Pos"), copy=T)
```

```
wildtype_positions %>%
  filter(UngappedCoverage >= 50) %>%
  ggplot(aes(x=Freq)) + geom_histogram(bins=50) + facet_wrap(~replica, ncol=1) + theme_bw() + xlim(0.75
```

```
## Warning: Removed 96 rows containing non-finite values (stat_bin).
```

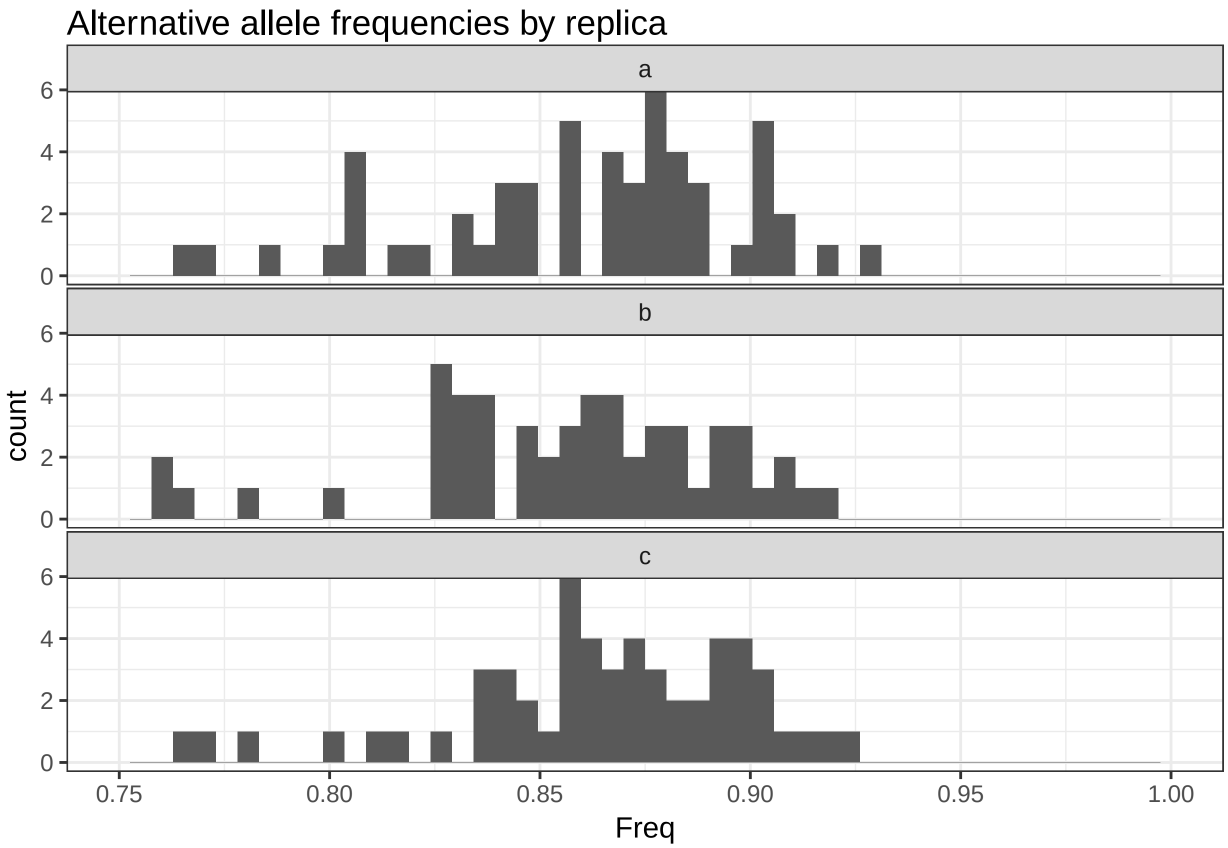
```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



```
variant_positions %>%
  filter(UngappedCoverage >= 50) %>%
  ggplot(aes(x=Freq)) + geom_histogram(bins=50) + facet_wrap(~replica, ncol=1) + theme_bw() + xlim(0.75
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```



```
snps = variants %>%
 filter(stri_startswith_fixed(`Polymorphism Type`, 'SNP'))
joined = inner_join(snps, minion_variants, by=c("Minimum" = "Pos"))
```

```
filtered = joined %>%
  filter(Qual == 0) %>%
  filter(replica.x == replica.y) %>%
  filter(modality == 'amplicon')

fit = lm(filtered$Freq ~ filtered$freq)
summary(fit)
```

```
##
## Call:
## lm(formula = filtered$Freq ~ filtered$freq)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24185 -0.02864 -0.00710  0.01820  0.47913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.041244   0.003927   10.50   <2e-16 ***
## filtered$freq 0.642182   0.040000   16.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

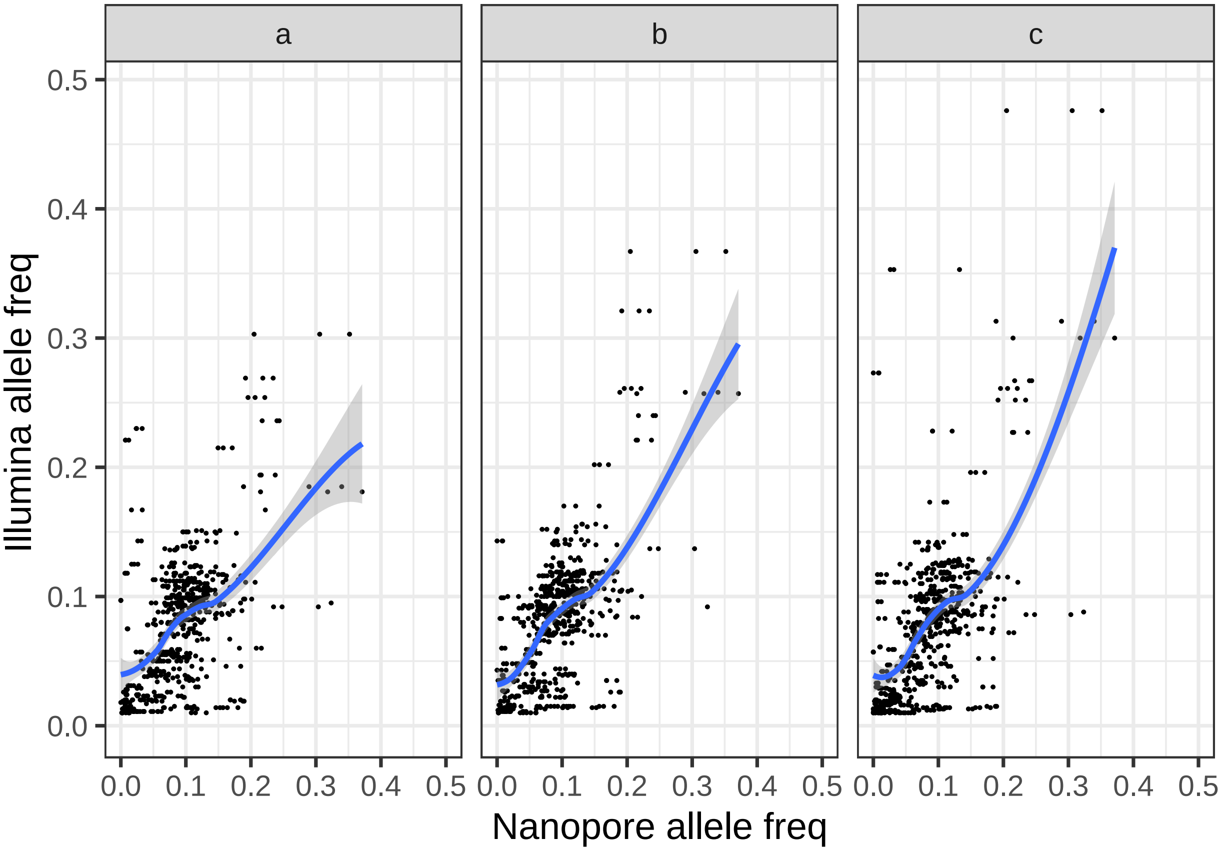
```
## Residual standard error: 0.05774 on 641 degrees of freedom
## Multiple R-squared:  0.2868, Adjusted R-squared:  0.2857
## F-statistic: 257.8 on 1 and 641 DF,  p-value: < 2.2e-16
```

```
p = joined %>%
  filter(Qual == 0) %>%
  filter(modality == 'amplicon') %>%
  ggplot(aes(x=Freq, y=freq)) + geom_point(size=0.2) + stat_smooth() + xlim(0, 0.5) + ylim(0, 0.5) + xla
p
```

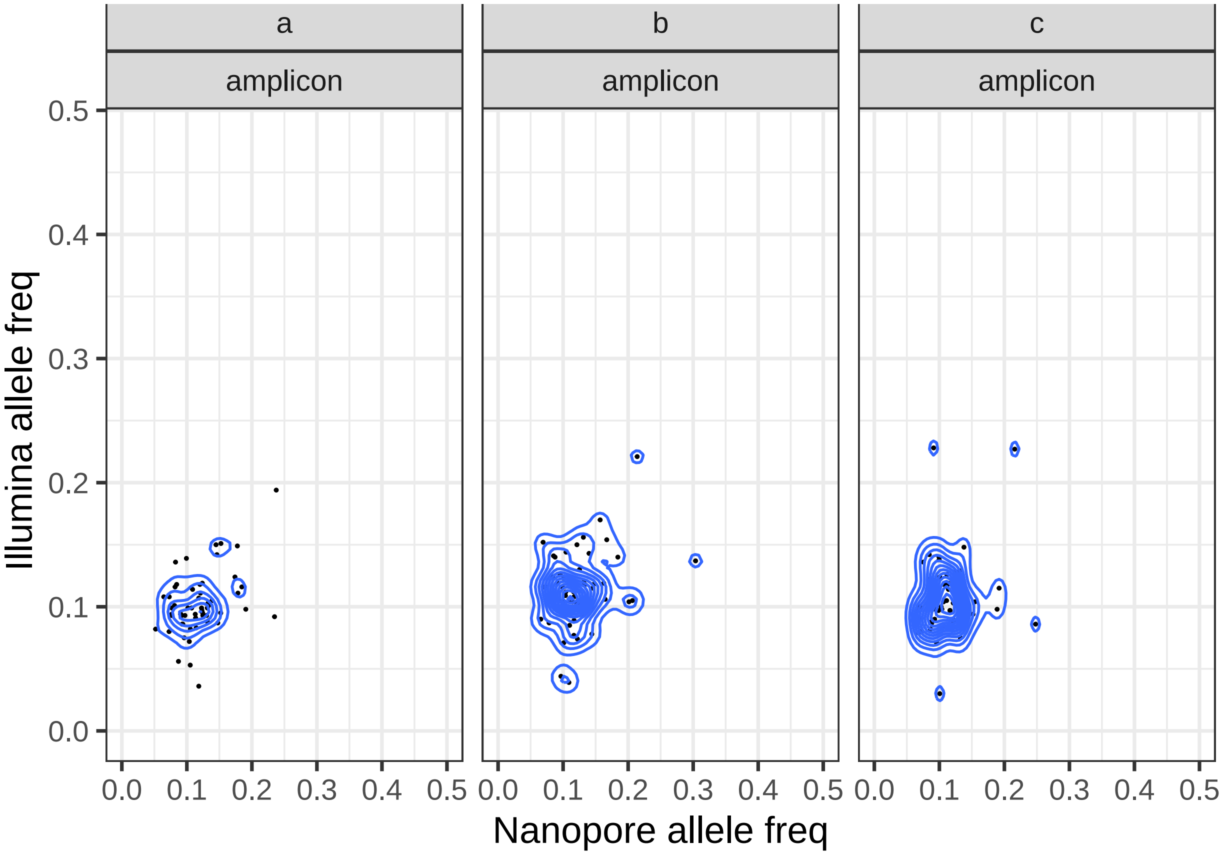
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

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

## Warning: Removed 12 rows containing missing values (geom_point).
```



```
a=inner_join(joined, expectedpositions, by=c("Minimum" = "Pos")) %>%
  filter(Qual == 0) %>%
  filter(modality == 'amplicon') %>%
  filter(replica.x == replica.y) %>%
  ggplot(aes(x=Freq, y=freq)) + geom_point(size=0.2) + geom_density2d() + xlim(0, 0.5) + ylim(0, 0.5) +
a
```


a
b
c
amplicon
amplicon
amplicon
Illumina allele freq
0.0
0.1
0.2
0.3
0.4
0.5
Nanopore allele freq
0.0
0.1
0.2
0.3
0.4
0.5