Benchmarking

Daniel

May 16, 2020

Vcf files loading and preprocessing

```
#load .vcf file
#header starts with ##, column names with #. comment.char takes a single character vector but c("##") d
S288C <- read.table("~/Desktop/rDNA_analysis/S288C/SRR4074255.remap.variants.vcf",
                    "\t",
                    header = F,
                    comment.char = "#")
S288C <- as.data.frame(S288C)
#assign colnammes
vcf_col_names <- c("CHROM", "POS", "ID", "REF", "ALT", "QUAL", "FILTER", "INFO")
colnames(S288C) <- vcf_col_names
#split INFO column into several; HRUN is only for INDEL
#NB: if some future column that will not shared between different rows, try to explore the 'fill' argum
S288C <- S288C %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 20 rows [5,
## 6, 8, 9, 10, 11, 12, 13, 18, 19, 24, 30, 31, 32, 33, 34, 35, 39, 40, 44].
```

```
#make REF as char
S288C$REF<-as.character(S288C$REF)
#remove characters from the new columnus and make them appropriate class
#[A-Z] - substring starts with a letter followed by everything else * and then the = sign. replace with
S288C$DP<-as.integer(gsub("[A-Z]*=","",S288C$DP))
S288C$AF<-as.numeric(gsub("[A-Z]*=","",S288C$AF)) #set it here for numeric bc these are NOT integers
S288C$SB<-as.integer(gsub("[A-Z]*=","",S288C$SB))
S288C$DP4<-as.factor(gsub("[A-Z]*4=","",S288C$DP4)) #note the '4' in the first argument bc it starts as
S288C$HRUN<-as.integer(gsub("[A-Z]*=","",S288C$HRUN))
S288C
```

```
##    CHROM  POS ID                REF                         ALT  QUAL
## 1  S288C    3 .                   A    AGTCTTCAACTGCTTTCGCATGAA   170
## 2  S288C    6 .                   A    ATTCAACTGCTTTCGCATGAAGTA    96
## 3  S288C    6 .                   A      ACAACTGCTTTCGCATGAAGTA  3499
## 4  S288C  217 .                  GT                           G 20487
## 5  S288C  285 .                   A                           T 21543
## 6  S288C  307 .                   A                           G 21343
## 7  S288C  445 .                  CT                           C  1005
## 8  S288C  557 .                   C                           T 22525
## 9  S288C  638 .                   C                           T 26519
## 10 S288C  648 .                   A                           G 25530
## 11 S288C  817 .                   C                           A 40249
## 12 S288C 1132 .                   T                           C 27214
## 13 S288C 1450 .                   A                           T   139
## 14 S288C 1570  .                 C                          CA  2578
## 15 S288C 1570  .        CAAAAAAAAA                           C  6783
## 16 S288C 1570  .                CA                           C 11225
## 17 S288C 1570  .               CAA                           C    71
## 18 S288C 1671  .                 T                           C 24438
## 19 S288C 1720  .                 T                           G 18704
## 20 S288C 1759  . ACCGCGTCGCCGCGTCG                           A  7152
## 21 S288C 1830  .                 G                         GTA  3253
## 22 S288C 1983  .                TG                           T   189
## 23 S288C 2045  .                AT                           A  8385
## 24 S288C 2189  .                 T                           C 12466
## 25 S288C 2243  .                CG                           C    60
## 26 S288C 2244  .                 G                          GA   964
## 27 S288C 2244  .                GA                           G  8552
## 28 S288C 2244  .              GAAA                           G   179
## 29 S288C 2244  .               GAA                           G  1045
## 30 S288C 2602  .                 G                           T   275
## 31 S288C 2617  .                 A                           T   539
## 32 S288C 2620  .                 T                           C   344
## 33 S288C 2625  .                 A                           G   448
## 34 S288C 2653  .                 T                           A   569
## 35 S288C 4445  .                 A                           C    97
## 36 S288C 4521  .                 T                         TAC    90
## 37 S288C 4521  .                 T                       TCTAC   402
## 38 S288C 8471  .              AAAG                           A 17213
## 39 S288C 8474  .                 G                           A   578
## 40 S288C 8488  .                 T                           A    73
## 41 S288C 8491  .                 C                          CT  4407
## 42 S288C 8491  .             CTTTT                           C  7377
## 43 S288C 8491  .                CT                           C 13136
## 44 S288C 8934  .                 G                           A  9697
## 45 S288C 9128  .                 G GCTTTCGCATGAAGTACCTCCCAACTA   263
## 46 S288C 9131  .                 T                 TTCGCATGAAG  1197
## 47 S288C 9131  .                 T             TTCGCATGAAGTACC   579
## 48 S288C 9132  .                 T           TCGCATGAAGTACCTCC  3459
## 49 S288C 9133  .                 C                      CGCATG   330
## 50 S288C 9133  .                 C                    CGCATGAA   118
## 51 S288C 9133  .                 C                 CGCATGAAGTA   283
## 52 S288C 9133  .                 C         CGCATGAAGTACCTCCCAA   123
## 53 S288C 9133  .                 C      CGCATGAAGTACCTCCCAACTA    65
##     FILTER    DP       AF  SB                  DP4 INDEL HRUN
## 1    PASS   343 0.058309   0           243,1,20,0 INDEL    1
## 2    PASS   879 0.026166   0           571,7,23,0 INDEL    2
## 3    PASS   879 0.238908   7          571,7,210,0 INDEL    2
## 4    PASS 11307 0.081366  37    6728,3732,648,272 INDEL    2
## 5    PASS 13010 0.089008  13    5534,6275,579,579  <NA>   NA
## 6    PASS 13665 0.084815   8    7019,5451,627,532  <NA>   NA
## 7    PASS 22941 0.004315   0    12772,10082,56,43 INDEL    5
## 8    PASS 19910 0.068106   9   7351,11167,508,848  <NA>   NA
## 9    PASS 23513 0.067282   2  15613,6290,1139,443  <NA>   NA
## 10   PASS 24226 0.067778   2  15750,6786,1160,482  <NA>   NA
## 11   PASS 28393 0.077202  33 13303,12755,1205,987  <NA>   NA
## 12   PASS 14387 0.098909  50    4434,8500,573,850  <NA>   NA
## 13   PASS  8435 0.028927 181     4321,3193,206,38  <NA>   NA
## 14   PASS 17425 0.079197   9  13186,2710,1121,259 INDEL   17
## 15   PASS 17425 0.088780 216  10693,2050,1137,410 INDEL   17
## 16   PASS 17425 0.150187   6  10693,2050,2169,448 INDEL   17
## 17   PASS 17425 0.029154   8    10693,2050,414,94 INDEL   17
## 18   PASS 18863 0.122144   3   4417,8581,800,1504  <NA>   NA
## 19   PASS 19880 0.098239  74  4453,13414,379,1574  <NA>   NA
## 20   PASS 16391 0.075895 282    6107,9138,304,940 INDEL    2
## 21   PASS 13304 0.068025 229    5372,7098,546,359 INDEL    1
## 22   PASS 10315 0.132719  84    2952,6075,560,809 INDEL    1
## 23   PASS  9093 0.113384  49    3408,4673,361,670 INDEL    8
## 24   PASS  8160 0.123897  33    2514,4162,323,688  <NA>   NA
## 25   PASS  5846 0.014369 109       1505,4228,0,84 INDEL    1
## 26   PASS  5789 0.046986  94     1474,4082,29,243 INDEL   16
## 27   PASS  5789 0.162031 118    1260,3179,164,774 INDEL   16
## 28   PASS  5789 0.024011  12     1260,3179,29,110 INDEL   16
## 29   PASS  5789 0.048540  54     1260,3179,45,236 INDEL   16
## 30   PASS 22284 0.005161  62     12466,9548,91,24  <NA>   NA
## 31   PASS 22372 0.005185 105     12209,9984,98,18  <NA>   NA
## 32   PASS 22850 0.005339  64    12198,10251,94,28  <NA>   NA
## 33   PASS 22366 0.005276 115    11709,10463,99,19  <NA>   NA
## 34   PASS 23857 0.005617 112   11980,11641,107,27  <NA>   NA
## 35   PASS 20494 0.003562   8     8574,11774,37,36  <NA>   NA
## 36   PASS 19206 0.004322 298      8360,10853,83,0 INDEL    1
## 37   PASS 19206 0.003853 266      8360,10853,74,0 INDEL    1
## 38   PASS 20711 0.042538   1   14130,5917,615,266 INDEL    6
## 39   PASS 20500 0.010537 578    13786,5544,39,177  <NA>   NA
## 40   PASS 21378 0.006175 223    13621,7273,30,102  <NA>   NA
## 41   PASS 21347 0.032417   0   13258,7521,444,248 INDEL   11
## 42   PASS 21347 0.043425   2   12201,6872,583,344 INDEL   11
## 43   PASS 21347 0.064927   7   12201,6872,861,525 INDEL   11
## 44   PASS 18643 0.043502  17   3490,14299,132,679  <NA>   NA
## 45   PASS  1886 0.059915  31       127,1656,0,113 INDEL    1
## 46   PASS  1848 0.108766  58       109,1390,0,201 INDEL    3
## 47   PASS  1848 0.080628   5       109,1390,7,142 INDEL    3
## 48   PASS  1710 0.198830 114       106,1134,0,340 INDEL    1
## 49   PASS  1702 0.070505  12        17,1094,5,115 INDEL    1
## 50   PASS  1702 0.051704 802        17,1094,71,17 INDEL    1
## 51   PASS  1702 0.066980   4        17,1094,0,114 INDEL    1
## 52   PASS  1702 0.052291   0         17,1094,1,88 INDEL    1
## 53   PASS  1702 0.044653   2         17,1094,0,76 INDEL    1
```

Data pre-filtering

```
#HRUN: "Homopolymer length to the right of report indel position". Remove entries that have HRUN >=4 (I
S288C_filtered <- subset(S288C, HRUN<4 | is.na(HRUN)) #since HRUN is only assigned for indels, spcifyin
#since for this alighmnet I used an rDNA prototype with no flanking regions, the first and last couple
S288C_filtered <- subset(S288C_filtered, POS>10 & POS <9100)

#for indels with high GC content
#define function first
#function requires loaded "tidyverse"
gc_indel <- function(indel) {
```

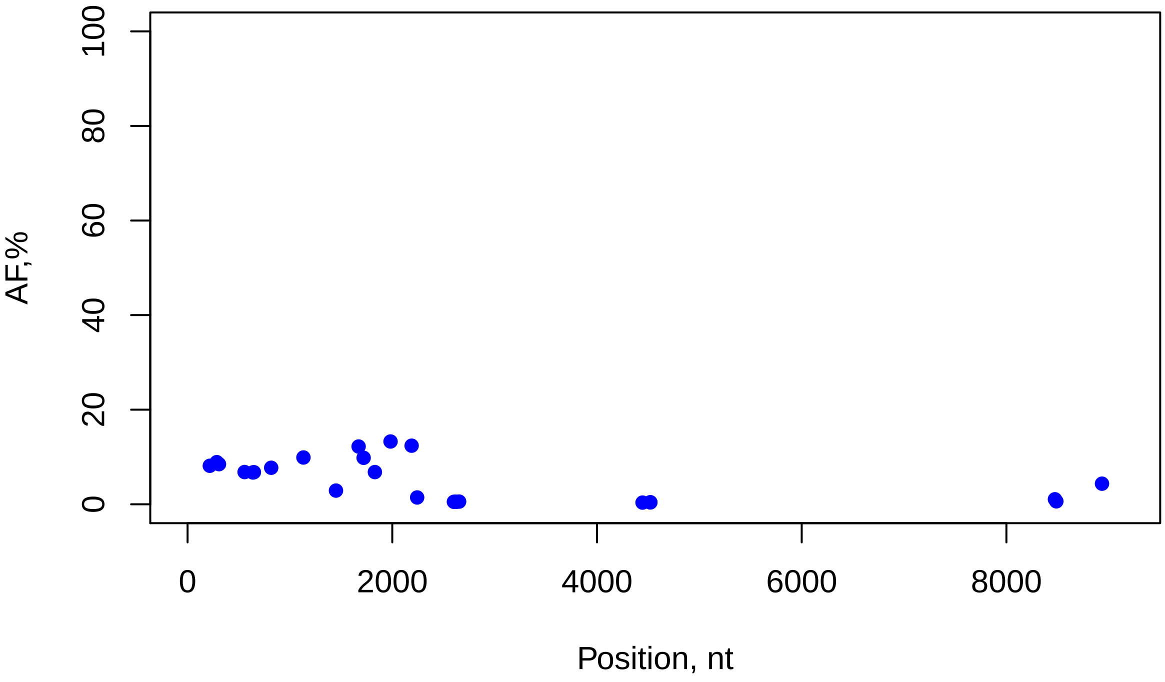
```r
  gc_content <- (
    (str_count(indel, "C") + str_count(indel, "G"))/nchar(indel)
  )
  if (nchar(indel) > 5 & gc_content > 0.6) {
      return(0)} else {return(1)}
}
GC<-as.data.frame(S288C_filtered$REF)
GC$res <- apply(GC,1,gc_indel) #calculates the gc_indel function. It returns '1' is passes the filter (
S288C_filtered <- subset(S288C_filtered,
                         GC$res==1
                         ) #subset here for GC content and homopolymer tracts
S288C_filtered
```

```
##    CHROM  POS ID REF   ALT  QUAL FILTER    DP       AF  SB
## 4  S288C  217  .  GT     G 20487   PASS 11307 0.081366  37
## 5  S288C  285  .   A     T 21543   PASS 13010 0.089008  13
## 6  S288C  307  .   A     G 21343   PASS 13665 0.084815   8
## 8  S288C  557  .   C     T 22525   PASS 19910 0.068106   9
## 9  S288C  638  .   C     T 26519   PASS 23513 0.067282   2
## 10 S288C  648  .   A     G 25530   PASS 24226 0.067778   2
## 11 S288C  817  .   C     A 40249   PASS 28393 0.077202  33
## 12 S288C 1132  .   T     C 27214   PASS 14387 0.098909  50
## 13 S288C 1450  .   A     T   139   PASS  8435 0.028927 181
## 18 S288C 1671  .   T     C 24438   PASS 18863 0.122144   3
## 19 S288C 1720  .   T     G 18704   PASS 19880 0.098239  74
## 21 S288C 1830  .   G   GTA  3253   PASS 13304 0.068025 229
## 22 S288C 1983  .  TG     T   189   PASS 10315 0.132719  84
## 24 S288C 2189  .   T     C 12466   PASS  8160 0.123897  33
## 25 S288C 2243  .  CG     C    60   PASS  5846 0.014369 109
## 30 S288C 2602  .   G     T   275   PASS 22284 0.005161  62
## 31 S288C 2617  .   A     T   539   PASS 22372 0.005185 105
## 32 S288C 2620  .   T     C   344   PASS 22850 0.005339  64
## 33 S288C 2625  .   A     G   448   PASS 22366 0.005276 115
## 34 S288C 2653  .   T     A   569   PASS 23857 0.005617 112
## 35 S288C 4445  .   A     C    97   PASS 20494 0.003562   8
## 36 S288C 4521  .   T   TAC    90   PASS 19206 0.004322 298
## 37 S288C 4521  .   T TCTAC   402   PASS 19206 0.003853 266
## 39 S288C 8474  .   G     A   578   PASS 20500 0.010537 578
## 40 S288C 8488  .   T     A    73   PASS 21378 0.006175 223
## 44 S288C 8934  .   G     A  9697   PASS 18643 0.043502  17
##                     DP4 INDEL HRUN
## 4     6728,3732,648,272 INDEL    2
## 5     5534,6275,579,579  <NA>   NA
## 6     7019,5451,627,532  <NA>   NA
## 8    7351,11167,508,848  <NA>   NA
## 9   15613,6290,1139,443  <NA>   NA
## 10  15750,6786,1160,482  <NA>   NA
## 11 13303,12755,1205,987  <NA>   NA
## 12    4434,8500,573,850  <NA>   NA
## 13     4321,3193,206,38  <NA>   NA
## 18   4417,8581,800,1504  <NA>   NA
## 19  4453,13414,379,1574  <NA>   NA
## 21    5372,7098,546,359 INDEL    1
## 22    2952,6075,560,809 INDEL    1
## 24    2514,4162,323,688  <NA>   NA
## 25       1505,4228,0,84 INDEL    1
## 30      12466,9548,91,24  <NA>   NA
## 31      12209,9984,98,18  <NA>   NA
## 32     12198,10251,94,28  <NA>   NA
## 33     11709,10463,99,19  <NA>   NA
## 34    11980,11641,107,27  <NA>   NA
## 35      8574,11774,37,36  <NA>   NA
## 36       8360,10853,83,0 INDEL    1
## 37       8360,10853,74,0 INDEL    1
## 39     13786,5544,39,177  <NA>   NA
## 40     13621,7273,30,102  <NA>   NA
## 44    3490,14299,132,679  <NA>   NA
```

```
#NB: I am keeping possible false-positives and variants with very low freq here so far, it will be expl
```

Plotting some data

```
#plot allele frequencies in %
plot(S288C_filtered$POS, (S288C_filtered$AF)*100, col="blue", xlim = c(1,9137), ylim = c(0,100), pch=16
     xlab="Position, nt",
     ylab="AF,%")
```



For SK1

```
#SK1:S288C(%) 100:0
SK1_S288C_100_0 <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_S288C_100_0.vcf",
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_100_0 <- as.data.frame(SK1_S288C_100_0)
colnames(SK1_S288C_100_0) <- vcf_col_names
SK1_S288C_100_0 <- SK1_S288C_100_0 %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

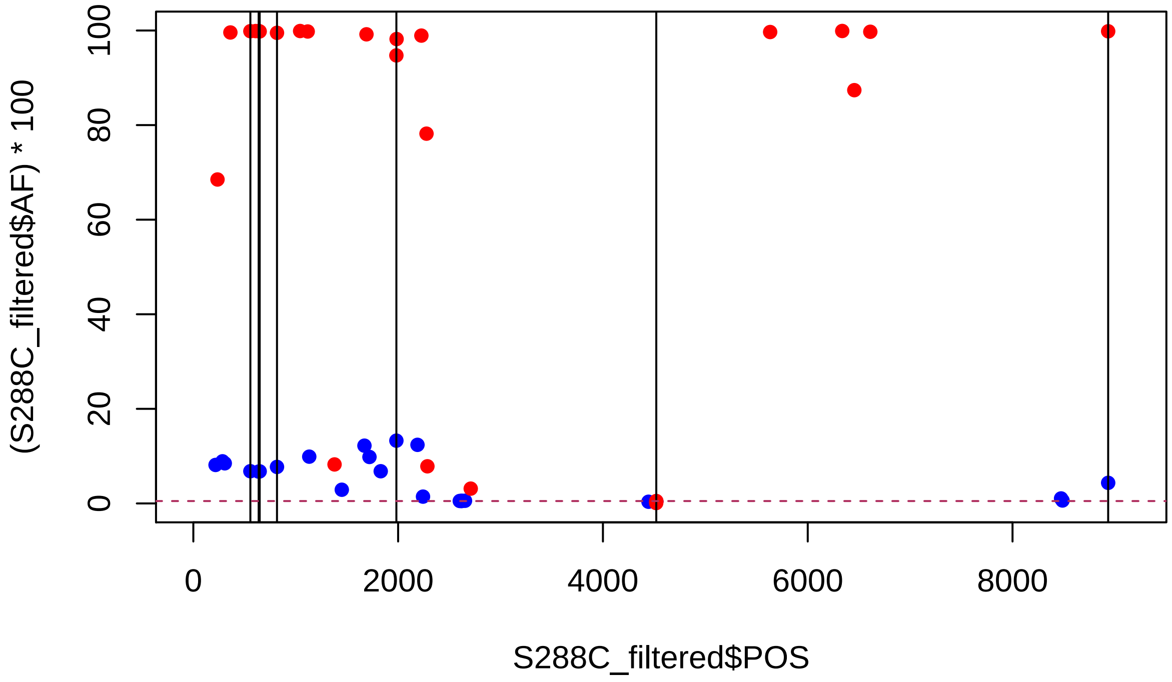
```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 18 rows [4,
## 5, 6, 7, 8, 9, 10, 11, 12, 14, 17, 19, 24, 32, 34, 35, 36, 41].
```

```
SK1_S288C_100_0$REF<-as.character(SK1_S288C_100_0$REF)
SK1_S288C_100_0$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0$DP))
SK1_S288C_100_0$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_100_0$AF))
SK1_S288C_100_0$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0$SB))
SK1_S288C_100_0$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_100_0$DP4))
SK1_S288C_100_0$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0$HRUN))

SK1_S288C_100_0_filtered <- subset(SK1_S288C_100_0, HRUN<4 | is.na(HRUN))
SK1_S288C_100_0_filtered <- subset(SK1_S288C_100_0_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_100_0_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
SK1_S288C_100_0_filtered <- subset(SK1_S288C_100_0_filtered,
                        GC1$res==1
                        )
SK1_S288C_100_0_filtered
```

```
##    CHROM  POS ID      REF     ALT  QUAL FILTER    DP       AF   SB
## 2  S288C  236  .        T TGCGGAA 49314   PASS 14285 0.684984 2321
## 4  S288C  362  .        A       G 49314   PASS 16393 0.995852   10
## 5  S288C  557  .        C       T 49314   PASS 14370 0.998608    3
## 6  S288C  609  .        C       A 49314   PASS 13646 0.998828    0
## 7  S288C  638  .        C       T 49314   PASS 16849 0.998872    7
## 8  S288C  648  .        A       G 49314   PASS 17525 0.998117    4
## 9  S288C  817  .        C       A 49314   PASS 24485 0.995140    6
## 10 S288C 1043  .        G       C 49314   PASS 25135 0.998966    4
## 11 S288C 1117  .        G       C 49314   PASS 17382 0.997929    2
## 12 S288C 1379  .        T       A  1079   PASS 12900 0.082403 1564
## 14 S288C 1691  .        C       G 49314   PASS 23654 0.991968   65
## 16 S288C 1983  .       TG       T 11349   PASS 17453 0.947344  962
## 17 S288C 1985  .        T       A  3410   PASS 17758 0.981980   89
## 19 S288C 2227  .        C       T 49314   PASS 11057 0.989238    0
## 22 S288C 2277  . CAAATAGT       C 49314   PASS  7340 0.781880  445
## 23 S288C 2286  .        A    ACTT   111   PASS  6854 0.078494  170
## 24 S288C 2709  .        T       C  6855   PASS 18767 0.031278    0
## 28 S288C 4520  .       CT       C    62   PASS 20053 0.000798   38
## 29 S288C 4521  .        T     TAC    64   PASS 20490 0.003514  253
## 30 S288C 4521  .        T   TCTAC   570   PASS 20490 0.005076  345
## 32 S288C 5633  .        T       C 49314   PASS 23875 0.996817    4
## 34 S288C 6337  .        G       A 49314   PASS  7160 0.999022    0
## 35 S288C 6455  .        T       C 49314   PASS 11717 0.873859    0
## 36 S288C 6611  .        T       C 49314   PASS 14054 0.997367    0
## 41 S288C 8934  .        G       A 49314   PASS 16121 0.998387    0
##                    DP4 INDEL HRUN
## 2  3540,873,5112,4673 INDEL    1
## 4      13,4,8681,7644  <NA>   NA
## 5       3,3,5017,9333  <NA>   NA
## 6       2,3,6747,6883  <NA>   NA
## 7      1,2,12222,4608  <NA>   NA
## 8      7,5,12550,4942  <NA>   NA
## 9     9,3,13432,10934  <NA>   NA
## 10     3,2,8240,16869  <NA>   NA
## 11     0,3,4896,12450  <NA>   NA
## 12  3430,1956,211,852  <NA>   NA
## 14    17,2,7447,16017  <NA>   NA
## 16  723,200,7227,9307 INDEL    1
## 17    57,12,8080,9358  <NA>   NA
## 19      1,5,2368,8570  <NA>   NA
## 22  314,1305,390,5349 INDEL    3
## 23     538,5652,2,536 INDEL    3
## 24  9395,8700,305,282  <NA>   NA
## 28    9281,10942,15,1 INDEL    1
## 29    9077,11415,72,0 INDEL    1
## 30   9077,11415,103,1 INDEL    1
## 32   12,9,11230,12569  <NA>   NA
## 34      1,2,2027,5126  <NA>   NA
## 35 1048,416,7293,2946  <NA>   NA
## 36      3,2,8474,5543  <NA>   NA
## 41     1,5,3177,12918  <NA>   NA
```

plotting

```
#i am plotting AF in percentage here
plot(S288C_filtered$POS, (S288C_filtered$AF)*100, col="blue", xlim = c(1,9137), ylim = c(0,100), pch=16
points(SK1_S288C_100_0_filtered$POS, (SK1_S288C_100_0_filtered$AF)*100, col="red", pch=16)
#find intersection between positions
shared_pos <- intersect(S288C_filtered$POS,SK1_S288C_100_0_filtered$POS)
abline(v=shared_pos) #plot where positions of variants are the same
abline(h=0.5, lty=2, col="maroon") #I got this treshold for sensitivity from titration (see below)
```



Titration

```
#these files contain only unique for SK1 variants
#SK1 reads : S288C reads(%) 100:0
SK1_S288C_100_0_u <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_unique/SK1_S288C_10
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_100_0_u <- as.data.frame(SK1_S288C_100_0_u)
colnames(SK1_S288C_100_0_u) <- vcf_col_names
SK1_S288C_100_0_u <- SK1_S288C_100_0_u %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 13 rows [3,
## 4, 5, 6, 7, 8, 9, 10, 15, 23, 25, 26, 27].
```

```
SK1_S288C_100_0_u$REF<-as.character(SK1_S288C_100_0_u$REF)
SK1_S288C_100_0_u$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0_u$DP))
SK1_S288C_100_0_u$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_100_0_u$AF))
SK1_S288C_100_0_u$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0_u$SB))
SK1_S288C_100_0_u$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_100_0_u$DP4))
SK1_S288C_100_0_u$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_100_0_u$HRUN))

SK1_S288C_100_0_u_filtered <- subset(SK1_S288C_100_0_u, HRUN<4 | is.na(HRUN))
SK1_S288C_100_0_u_filtered <- subset(SK1_S288C_100_0_u_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_100_0_u_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
```

```
SK1_S288C_100_0_u_filtered <- subset(SK1_S288C_100_0_u_filtered,
                        GC1$res==1
                        )

#SK1 reads : S288C reads(%) 50:50
SK1_S288C_50_50_u <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_unique/SK1_S288C_50_
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_50_50_u <- as.data.frame(SK1_S288C_50_50_u)
colnames(SK1_S288C_50_50_u) <- vcf_col_names
SK1_S288C_50_50_u <- SK1_S288C_50_50_u %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 13 rows [3,
## 4, 5, 6, 7, 8, 9, 13, 14, 18, 20, 21, 22].
```

```
SK1_S288C_50_50_u$REF<-as.character(SK1_S288C_50_50_u$REF)
SK1_S288C_50_50_u$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_50_50_u$DP))
SK1_S288C_50_50_u$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_50_50_u$AF))
SK1_S288C_50_50_u$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_50_50_u$SB))
SK1_S288C_50_50_u$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_50_50_u$DP4))
SK1_S288C_50_50_u$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_50_50_u$HRUN))

SK1_S288C_50_50_u_filtered <- subset(SK1_S288C_50_50_u, HRUN<4 | is.na(HRUN))
SK1_S288C_50_50_u_filtered <- subset(SK1_S288C_50_50_u_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_50_50_u_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
SK1_S288C_50_50_u_filtered <- subset(SK1_S288C_50_50_u_filtered,
                        GC1$res==1
                        )

#SK1 reads : S288C reads(%) 10:90
SK1_S288C_10_90_u <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_unique/SK1_S288C_10_
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_10_90_u <- as.data.frame(SK1_S288C_10_90_u)
colnames(SK1_S288C_10_90_u) <- vcf_col_names
SK1_S288C_10_90_u <- SK1_S288C_10_90_u %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 12 rows [2,
## 3, 4, 5, 6, 7, 10, 11, 13, 15, 16, 17].
```

```
SK1_S288C_10_90_u$REF<-as.character(SK1_S288C_10_90_u$REF)
SK1_S288C_10_90_u$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_10_90_u$DP))
SK1_S288C_10_90_u$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_10_90_u$AF))
SK1_S288C_10_90_u$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_10_90_u$SB))
SK1_S288C_10_90_u$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_10_90_u$DP4))
```

```
SK1_S288C_10_90_u$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_10_90_u$HRUN))

SK1_S288C_10_90_u_filtered <- subset(SK1_S288C_10_90_u, HRUN<4 | is.na(HRUN))
SK1_S288C_10_90_u_filtered <- subset(SK1_S288C_10_90_u_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_10_90_u_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
SK1_S288C_10_90_u_filtered <- subset(SK1_S288C_10_90_u_filtered,
                        GC1$res==1
                        )

#SK1 reads : S288C reads(%) 1:99
SK1_S288C_1_99_u <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_unique/SK1_S288C_1_99
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_1_99_u <- as.data.frame(SK1_S288C_1_99_u)
colnames(SK1_S288C_1_99_u) <- vcf_col_names
SK1_S288C_1_99_u <- SK1_S288C_1_99_u %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

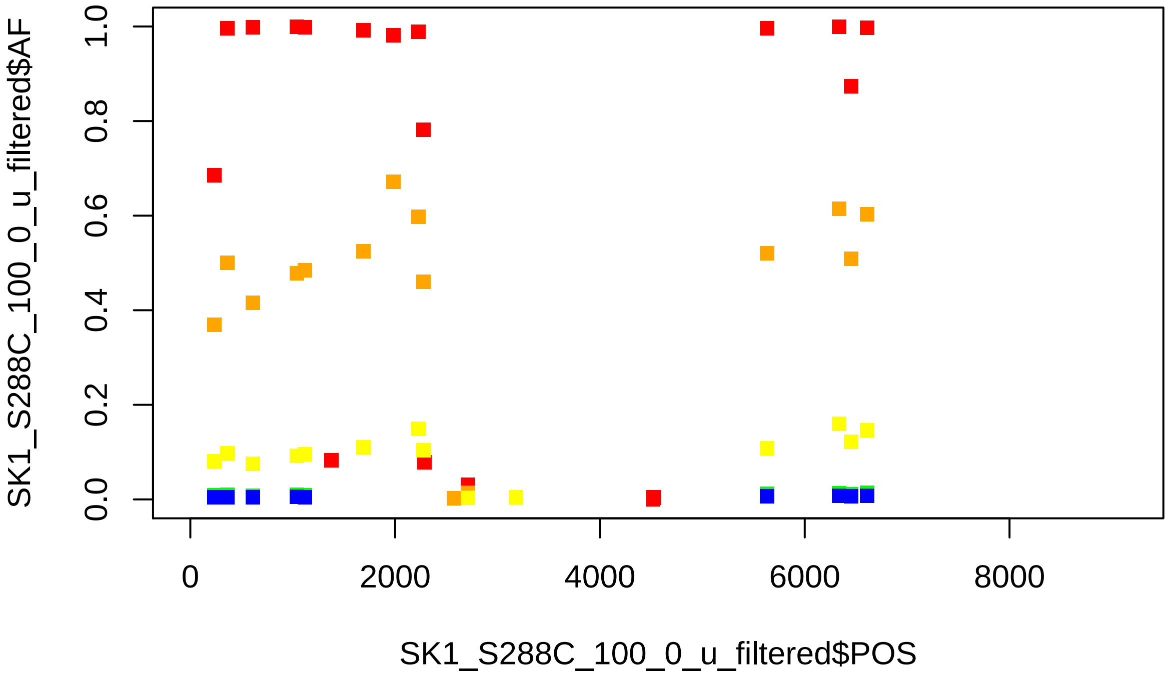
```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 8 rows [2,
## 3, 4, 5, 7, 8, 9, 10].
```

```
SK1_S288C_1_99_u$REF<-as.character(SK1_S288C_1_99_u$REF)
SK1_S288C_1_99_u$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_1_99_u$DP))
SK1_S288C_1_99_u$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_1_99_u$AF))
SK1_S288C_1_99_u$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_1_99_u$SB))
SK1_S288C_1_99_u$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_1_99_u$DP4))
SK1_S288C_1_99_u$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_1_99_u$HRUN))

SK1_S288C_1_99_u_filtered <- subset(SK1_S288C_1_99_u, HRUN<4 | is.na(HRUN))
SK1_S288C_1_99_u_filtered <- subset(SK1_S288C_1_99_u_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_1_99_u_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
SK1_S288C_1_99_u_filtered <- subset(SK1_S288C_1_99_u_filtered,
                        GC1$res==1
                        )

#SK1 reads : S288C reads(%) 0.5:99.5
SK1_S288C_05_99_u <- read.table("~/Desktop/rDNA_analysis/benchmarking/titration/SK1_unique/SK1_S288C_05_
                    "\t",
                    header = F,
                    comment.char = "#")
SK1_S288C_05_99_u <- as.data.frame(SK1_S288C_05_99_u)
colnames(SK1_S288C_05_99_u) <- vcf_col_names
SK1_S288C_05_99_u <- SK1_S288C_05_99_u %>% separate("INFO",
                   c("DP","AF","SB","DP4","INDEL","HRUN"),
                   sep=";"
                   )
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 8 rows [2,
## 3, 4, 5, 6, 7, 8, 9].
```

```
SK1_S288C_05_99_u$REF<-as.character(SK1_S288C_05_99_u$REF)
SK1_S288C_05_99_u$DP<-as.integer(gsub("[A-Z]*=","",SK1_S288C_05_99_u$DP))
SK1_S288C_05_99_u$AF<-as.numeric(gsub("[A-Z]*=","",SK1_S288C_05_99_u$AF))
SK1_S288C_05_99_u$SB<-as.integer(gsub("[A-Z]*=","",SK1_S288C_05_99_u$SB))
SK1_S288C_05_99_u$DP4<-as.factor(gsub("[A-Z]*4=","",SK1_S288C_05_99_u$DP4))
SK1_S288C_05_99_u$HRUN<-as.integer(gsub("[A-Z]*=","",SK1_S288C_05_99_u$HRUN))

SK1_S288C_05_99_u_filtered <- subset(SK1_S288C_05_99_u, HRUN<4 | is.na(HRUN))
SK1_S288C_05_99_u_filtered <- subset(SK1_S288C_05_99_u_filtered, POS>10 & POS <9100)

GC1<-as.data.frame(SK1_S288C_05_99_u_filtered$REF)
GC1$res <- apply(GC1,1,gc_indel)
SK1_S288C_05_99_u_filtered <- subset(SK1_S288C_05_99_u_filtered,
                        GC1$res==1
                        )
#lower titrations led to undetectable SK1-unique variants:
#SK1 reads : S288C reads(%) 0.1:99.9
#SK1 reads : S288C reads(%) 0.05:99.95
#SK1 reads : S288C reads(%) 0.01:99.99
```

```
plot(SK1_S288C_100_0_u_filtered$POS, SK1_S288C_100_0_u_filtered$AF, xlim = c(1,9137), ylim = c(0,1), col
points(SK1_S288C_50_50_u_filtered$POS, SK1_S288C_50_50_u_filtered$AF, col="orange", pch=15)
points(SK1_S288C_10_90_u_filtered$POS, SK1_S288C_10_90_u_filtered$AF, col="yellow", pch=15)
points(SK1_S288C_1_99_u_filtered$POS, SK1_S288C_1_99_u_filtered$AF, col="green", pch=15)
points(SK1_S288C_05_99_u_filtered$POS, SK1_S288C_05_99_u_filtered$AF, col="blue", pch=15)
```



#Plot titrations

```
SK1_S288C_100_0_AF_98 <- subset(SK1_S288C_100_0_u_filtered, AF>=0.98)
SK1_S288C_100_0_AF_98
```

```
##    CHROM  POS ID REF ALT  QUAL FILTER    DP       AF SB              DP4
## 3  S288C  362  .   A   G 49314   PASS 16393 0.995852 10   13,4,8681,7644
## 4  S288C  609  .   C   A 49314   PASS 13646 0.998828  0    2,3,6747,6883
## 5  S288C 1043  .   G   C 49314   PASS 25135 0.998966  4   3,2,8240,16869
## 6  S288C 1117  .   G   C 49314   PASS 17382 0.997929  2   0,3,4896,12450
## 8  S288C 1691  .   C   G 49314   PASS 23654 0.991968 65  17,2,7447,16017
## 9  S288C 1985  .   T   A  3410   PASS 17758 0.981980 89  57,12,8080,9358
## 10 S288C 2227  .   C   T 49314   PASS 11057 0.989238  0    1,5,2368,8570
## 23 S288C 5633  .   T   C 49314   PASS 23875 0.996817  4 12,9,11230,12569
## 25 S288C 6337  .   G   A 49314   PASS  7160 0.999022  0    1,2,2027,5126
## 27 S288C 6611  .   T   C 49314   PASS 14054 0.997367  0    3,2,8474,5543
##    INDEL HRUN
## 3   <NA>   NA
## 4   <NA>   NA
## 5   <NA>   NA
## 6   <NA>   NA
## 8   <NA>   NA
## 9   <NA>   NA
## 10  <NA>   NA
## 23  <NA>   NA
## 25  <NA>   NA
## 27  <NA>   NA
```

```r
SK1_S288C_100_0_titr <- data.frame(
  POS = c(SK1_S288C_100_0_AF_98$POS),
  AF = c(SK1_S288C_100_0_AF_98$AF),
  SK1_percent = rep(c(100), nrow(SK1_S288C_100_0_AF_98)))

SK1_S288C_50_50_titr <- subset(SK1_S288C_50_50_u_filtered, SK1_S288C_50_50_u_filtered$POS %in% SK1_S288C
#create a dataframe with POS, AF, and SK1_percent read
SK1_S288C_50_50_titr <- data.frame(
  POS = c(SK1_S288C_50_50_titr$POS),
  AF = c(SK1_S288C_50_50_titr$AF),
  SK1_percent = rep(c(50), nrow(SK1_S288C_50_50_titr))
)
#do the same for the rest of titrations
#NB:some titrations have new variants emerged, (false positives?), BUT they will be excluded from the a

#
SK1_S288C_10_90_titr <- subset(SK1_S288C_10_90_u_filtered, SK1_S288C_10_90_u_filtered$POS %in% SK1_S288C
SK1_S288C_10_90_titr <- data.frame(
  POS = c(SK1_S288C_10_90_titr$POS),
  AF = c(SK1_S288C_10_90_titr$AF),
  SK1_percent = rep(c(10), nrow(SK1_S288C_10_90_titr))
)
#
SK1_S288C_1_99_titr <- subset(SK1_S288C_1_99_u_filtered, SK1_S288C_1_99_u_filtered$POS %in% SK1_S288C_1
SK1_S288C_1_99_titr <- data.frame(
  POS = c(SK1_S288C_1_99_titr$POS),
  AF = c(SK1_S288C_1_99_titr$AF),
  SK1_percent = rep(c(1), nrow(SK1_S288C_1_99_titr))
)

#
SK1_S288C_05_99_titr <- subset(SK1_S288C_05_99_u_filtered, SK1_S288C_05_99_u_filtered$POS %in% SK1_S288C
SK1_S288C_05_99_titr <- data.frame(
  POS = c(SK1_S288C_05_99_titr$POS),
  AF = c(SK1_S288C_05_99_titr$AF),
  SK1_percent = rep(c(0.5), nrow(SK1_S288C_05_99_titr))
)

SK1_S288C_50_50_titr
```

```
##      POS       AF SK1_percent
## 1    362 0.500155          50
## 2    609 0.415583          50
## 3   1043 0.478403          50
## 4   1117 0.484677          50
## 5   1691 0.524551          50
## 6   1985 0.671668          50
## 7   2227 0.598094          50
## 8   5633 0.520365          50
## 9   6337 0.614701          50
## 10  6611 0.603252          50
```

```
SK1_S288C_10_90_titr
```

```
##    POS       AF SK1_percent
## 1  362 0.097785          10
## 2  609 0.075143          10
## 3 1043 0.092489          10
## 4 1117 0.095388          10
## 5 1691 0.110678          10
## 6 2227 0.149091          10
## 7 5633 0.108147          10
## 8 6337 0.159935          10
## 9 6611 0.146262          10
```

```
SK1_S288C_1_99_titr
```

```
##    POS       AF SK1_percent
## 1  362 0.009399           1
## 2  609 0.008011           1
## 3 1043 0.010208           1
## 4 1117 0.009170           1
## 5 5633 0.011639           1
## 6 6337 0.013155           1
## 7 6611 0.013999           1
```

```
SK1_S288C_05_99_titr
```

```
##    POS       AF SK1_percent
## 1  362 0.004909         0.5
## 2  609 0.004279         0.5
## 3 1043 0.005206         0.5
## 4 1117 0.004228         0.5
## 5 5633 0.006493         0.5
## 6 6337 0.007313         0.5
## 7 6611 0.007932         0.5
```

```
titration <- rbind(SK1_S288C_100_0_titr, SK1_S288C_50_50_titr, SK1_S288C_10_90_titr, SK1_S288C_1_99_tit
titration$POS <-as.factor(titration$POS)
titration
```

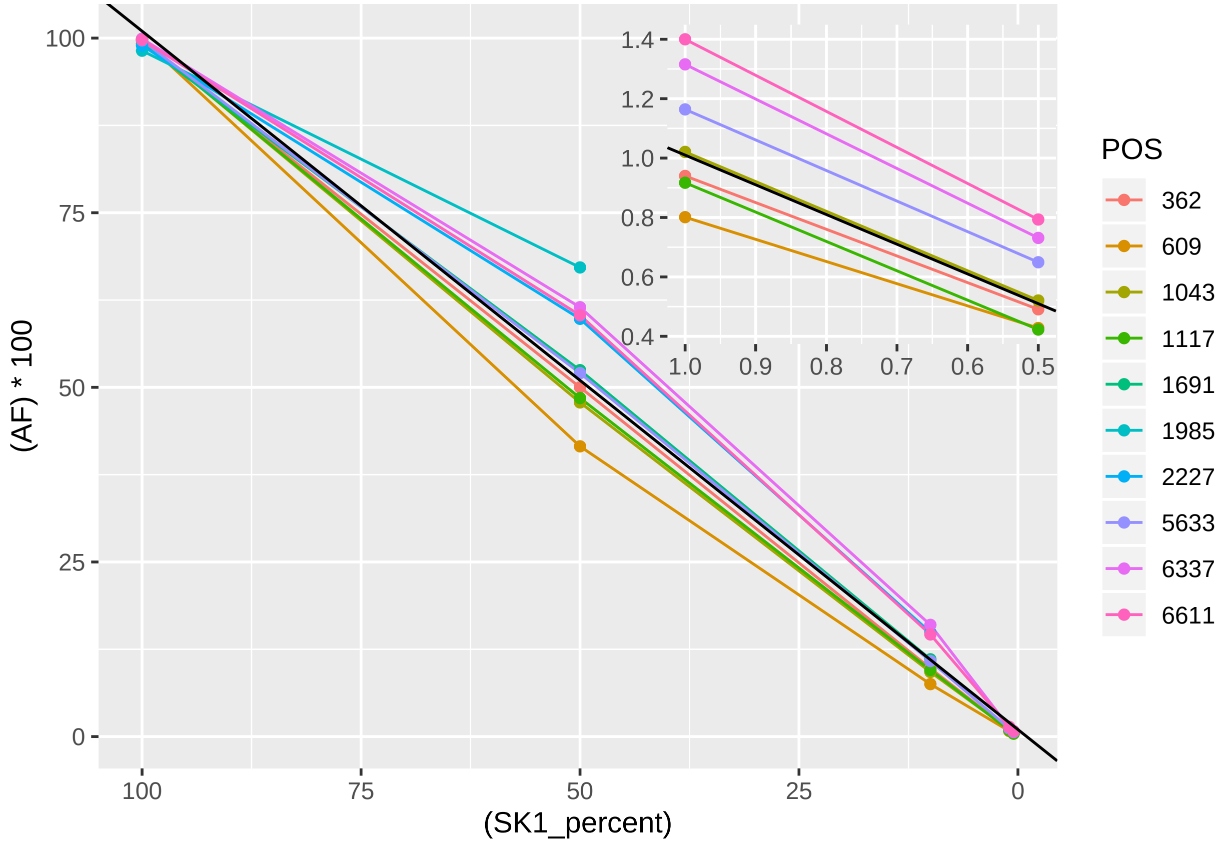
```
##     POS       AF SK1_percent
## 1   362 0.995852       100.0
## 2   609 0.998828       100.0
## 3  1043 0.998966       100.0
## 4  1117 0.997929       100.0
## 5  1691 0.991968       100.0
## 6  1985 0.981980       100.0
## 7  2227 0.989238       100.0
## 8  5633 0.996817       100.0
## 9  6337 0.999022       100.0
## 10 6611 0.997367       100.0
## 11  362 0.500155        50.0
## 12  609 0.415583        50.0
## 13 1043 0.478403        50.0
## 14 1117 0.484677        50.0
## 15 1691 0.524551        50.0
## 16 1985 0.671668        50.0
## 17 2227 0.598094        50.0
## 18 5633 0.520365        50.0
## 19 6337 0.614701        50.0
## 20 6611 0.603252        50.0
## 21  362 0.097785        10.0
## 22  609 0.075143        10.0
## 23 1043 0.092489        10.0
## 24 1117 0.095388        10.0
## 25 1691 0.110678        10.0
## 26 2227 0.149091        10.0
## 27 5633 0.108147        10.0
## 28 6337 0.159935        10.0
## 29 6611 0.146262        10.0
## 30  362 0.009399         1.0
## 31  609 0.008011         1.0
## 32 1043 0.010208         1.0
## 33 1117 0.009170         1.0
## 34 5633 0.011639         1.0
## 35 6337 0.013155         1.0
## 36 6611 0.013999         1.0
## 37  362 0.004909         0.5
## 38  609 0.004279         0.5
## 39 1043 0.005206         0.5
## 40 1117 0.004228         0.5
## 41 5633 0.006493         0.5
## 42 6337 0.007313         0.5
## 43 6611 0.007932         0.5
```

```
#can use coef(lm()) to calculate slope and intersept BUT be careful!
p1<- ggplot(data=titration, aes(x=(SK1_percent), y=(AF)*100))+
  geom_line(aes(color=POS))+
  geom_point(aes(color=POS))+
  geom_abline(slope=-1, intercept=1)+ #here intercept = 1 because AF is 1!!!!! (100 is displayed in per
  scale_x_reverse()
#plot 1%-0.5% range
#lines need to be recolorored to match the ones on the full plot
#to retrieve used colors:
```

```
ggplot_build(p1)$data[[1]]$colour
```

```
##  [1] "#F8766D" "#F8766D" "#F8766D" "#F8766D" "#F8766D" "#D89000" "#D89000"
##  [8] "#D89000" "#D89000" "#D89000" "#A3A500" "#A3A500" "#A3A500" "#A3A500"
## [15] "#A3A500" "#39B600" "#39B600" "#39B600" "#39B600" "#39B600" "#00BF7D"
## [22] "#00BF7D" "#00BF7D" "#00BFC4" "#00BFC4" "#00B0F6" "#00B0F6" "#00B0F6"
## [29] "#9590FF" "#9590FF" "#9590FF" "#9590FF" "#9590FF" "#E76BF3" "#E76BF3"
## [36] "#E76BF3" "#E76BF3" "#E76BF3" "#FF62BC" "#FF62BC" "#FF62BC" "#FF62BC"
## [43] "#FF62BC"
```

```
colgraph<-c("#F8766D","#D89000","#A3A500","#39B600", "#9590FF","#E76BF3","#FF62BC","#F8766D","#D89000",
p2<- ggplot(data=titration[30:nrow(titration),], aes(x=(SK1_percent), y=(AF)*100))+
  geom_line(aes(col=POS))+
  geom_point(aes(col=POS))+
  geom_abline(slope=-1, intercept=0.01)+
  scale_x_reverse()+
  scale_color_manual(values=colgraph)+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        legend.position = "none",
        plot.background = element_blank())
# also can play with ggplotly(p1)
titr<-p1 + annotation_custom(ggplotGrob(p2),
                        xmin=6, xmax=-47, ymin=49, ymax=104)
titr
```



```
#ggsave(file="titration.svg", plot=titr, width=10, height = 8) #can use .pdf too
pdf("titration.pdf")
```

```
print(titr)
dev.off()
```

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

A glimpse on some false-positives (just preliminary look, more thorough analysis will be in another protocol)

```
#some titrations have new variants emerged, (false positives?):
SK1_S288C_50_50_new_arised <- subset(SK1_S288C_50_50_u_filtered, !(SK1_S288C_50_50_u_filtered$POS %in% S
#do the same for the rest
SK1_S288C_10_90_new_arised <- subset(SK1_S288C_10_90_u_filtered, !(SK1_S288C_10_90_u_filtered$POS %in% S
SK1_S288C_1_99_new_arised <- subset(SK1_S288C_1_99_u_filtered, !(SK1_S288C_1_99_u_filtered$POS %in% SK1_
SK1_S288C_05_99_new_arised <- subset(SK1_S288C_05_99_u_filtered, !(SK1_S288C_05_99_u_filtered$POS %in% S

SK1_S288C_50_50_new_arised
```

```
##    CHROM  POS ID REF ALT QUAL FILTER    DP       AF SB              DP4
## 13 S288C 2573  .   T   C  100   PASS 21509 0.002464 39 12665,8681,45,8
##    INDEL HRUN
## 13  <NA>   NA
```

```
SK1_S288C_10_90_new_arised
```

```
##    CHROM  POS ID REF ALT QUAL FILTER    DP       AF SB               DP4
## 11 S288C 3180  .   G   A  122   PASS 18566 0.004363 45 6947,11448,13,68
##    INDEL HRUN
## 11  <NA>   NA
```

```
SK1_S288C_1_99_new_arised
```

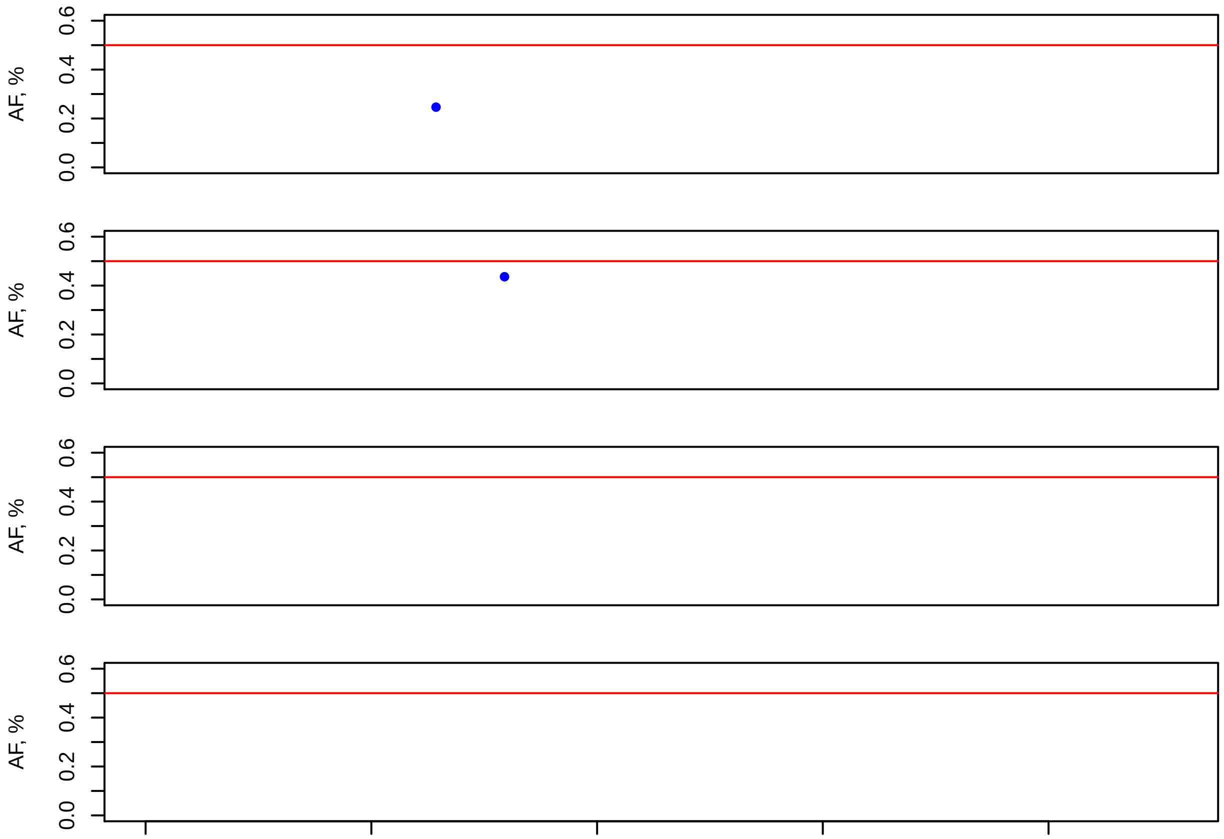
```
##  [1] CHROM  POS    ID     REF    ALT    QUAL   FILTER DP     AF     SB
## [11] DP4    INDEL  HRUN
## <0 rows> (or 0-length row.names)
```

```
SK1_S288C_05_99_new_arised
```

```
##  [1] CHROM  POS    ID     REF    ALT    QUAL   FILTER DP     AF     SB
## [11] DP4    INDEL  HRUN
## <0 rows> (or 0-length row.names)
```

```
#HERE I AM PLOTTING PERCENTAGE TOO!
#mai=c(bottom,left,)
par(mfrow=c(4,1), mai=c(0.1,0.5,0.2,0.2))
plot(SK1_S288C_50_50_new_arised$POS, (SK1_S288C_50_50_new_arised$AF)*100, col="blue", pch=16, xlim = c(
     xlab = "", ylab = "AF, %",
     xaxt='n')
#mtext("100% SK1 reads", side = 4)
abline(h=0.5, col="red")

plot(SK1_S288C_10_90_new_arised$POS, (SK1_S288C_10_90_new_arised$AF)*100, col="blue", pch=16, xlim = c(
     xlab = "", ylab = "AF, %",
     xaxt='n')
abline(h=0.5, col="red")
```

```
plot(SK1_S288C_1_99_new_arised$POS, (SK1_S288C_1_99_new_arised$AF)*100, col="blue", pch=16, xlim = c(1,
     xlab = "", ylab = "AF, %",
     xaxt='n')
abline(h=0.5, col="red")

plot(SK1_S288C_05_99_new_arised$POS, (SK1_S288C_05_99_new_arised$AF)*100, col="blue", pch=16, xlim = c(
     xlab = "Position, nt", ylab = "AF, %",
     )
abline(h=0.5, col="red")
```



distribution of SNPs and INDELs

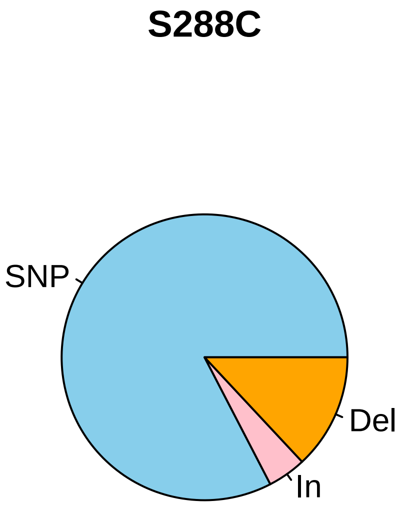
```
#this is a really basic look at the data. I counted snps, in and dels 'manually' now, but it can be cod

#first, apply the calculated threshold (AF<0.5% same as AF<0.005)
s288c_threshold_pass <- subset(S288C_filtered, AF*100 > 0.5) #here do not forget to AF*100 bc they are 
sk1_threshold_pass <- subset(SK1_S288C_100_0_filtered, AF*100 > 0.5)
s288c_threshold_pass
```

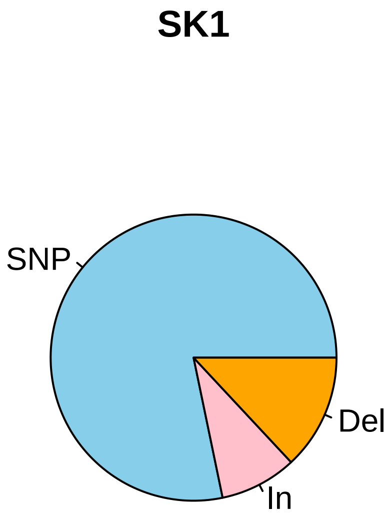
```
##     CHROM  POS ID REF ALT  QUAL FILTER    DP       AF  SB
## 4   S288C  217  .  GT   G 20487   PASS 11307 0.081366  37
## 5   S288C  285  .   A   T 21543   PASS 13010 0.089008  13
## 6   S288C  307  .   A   G 21343   PASS 13665 0.084815   8
## 8   S288C  557  .   C   T 22525   PASS 19910 0.068106   9
## 9  S288C  638  .   C   T 26519   PASS 23513 0.067282   2
## 10 S288C  648  .   A   G 25530   PASS 24226 0.067778   2
## 11 S288C  817  .   C   A 40249   PASS 28393 0.077202  33
## 12 S288C 1132  .   T   C 27214   PASS 14387 0.098909  50
## 13 S288C 1450  .   A   T   139   PASS  8435 0.028927 181
## 18 S288C 1671  .   T   C 24438   PASS 18863 0.122144   3
## 19 S288C 1720  .   T   G 18704   PASS 19880 0.098239  74
## 21 S288C 1830  .   G GTA  3253   PASS 13304 0.068025 229
## 22 S288C 1983  .  TG   T   189   PASS 10315 0.132719  84
## 24 S288C 2189  .   T   C 12466   PASS  8160 0.123897  33
## 25 S288C 2243  .  CG   C    60   PASS  5846 0.014369 109
## 30 S288C 2602  .   G   T   275   PASS 22284 0.005161  62
## 31 S288C 2617  .   A   T   539   PASS 22372 0.005185 105
## 32 S288C 2620  .   T   C   344   PASS 22850 0.005339  64
## 33 S288C 2625  .   A   G   448   PASS 22366 0.005276 115
## 34 S288C 2653  .   T   A   569   PASS 23857 0.005617 112
## 39 S288C 8474  .   G   A   578   PASS 20500 0.010537 578
## 40 S288C 8488  .   T   A    73   PASS 21378 0.006175 223
## 44 S288C 8934  .   G   A  9697   PASS 18643 0.043502  17
##                      DP4 INDEL HRUN
## 4     6728,3732,648,272 INDEL    2
## 5     5534,6275,579,579  <NA>   NA
## 6     7019,5451,627,532  <NA>   NA
## 8    7351,11167,508,848  <NA>   NA
## 9   15613,6290,1139,443  <NA>   NA
## 10  15750,6786,1160,482  <NA>   NA
## 11 13303,12755,1205,987  <NA>   NA
## 12    4434,8500,573,850  <NA>   NA
## 13     4321,3193,206,38  <NA>   NA
## 18   4417,8581,800,1504  <NA>   NA
## 19  4453,13414,379,1574  <NA>   NA
## 21    5372,7098,546,359 INDEL    1
## 22    2952,6075,560,809 INDEL    1
## 24    2514,4162,323,688  <NA>   NA
## 25        1505,4228,0,84 INDEL    1
## 30      12466,9548,91,24  <NA>   NA
## 31      12209,9984,98,18  <NA>   NA
## 32    12198,10251,94,28  <NA>   NA
## 33    11709,10463,99,19  <NA>   NA
## 34   11980,11641,107,27  <NA>   NA
## 39    13786,5544,39,177  <NA>   NA
## 40    13621,7273,30,102  <NA>   NA
## 44   3490,14299,132,679  <NA>   NA
```

```
sk1_threshold_pass
```

```
##    CHROM  POS ID      REF     ALT  QUAL FILTER    DP       AF   SB
## 2  S288C  236  .        T TGCGGAA 49314   PASS 14285 0.684984 2321
## 4  S288C  362  .        A       G 49314   PASS 16393 0.995852   10
## 5  S288C  557  .        C       T 49314   PASS 14370 0.998608    3
## 6  S288C  609  .        C       A 49314   PASS 13646 0.998828    0
## 7  S288C  638  .        C       T 49314   PASS 16849 0.998872    7
## 8  S288C  648  .        A       G 49314   PASS 17525 0.998117    4
## 9  S288C  817  .        C       A 49314   PASS 24485 0.995140    6
## 10 S288C 1043  .        G       C 49314   PASS 25135 0.998966    4
## 11 S288C 1117   .        G       C 49314   PASS 17382 0.997929    2
## 12 S288C 1379   .        T       A  1079   PASS 12900 0.082403 1564
## 14 S288C 1691   .        C       G 49314   PASS 23654 0.991968   65
## 16 S288C 1983   .       TG       T 11349   PASS 17453 0.947344  962
## 17 S288C 1985   .        T       A  3410   PASS 17758 0.981980   89
## 19 S288C 2227   .        C       T 49314   PASS 11057 0.989238    0
## 22 S288C 2277   . CAAAATGT       C 49314   PASS  7340 0.781880  445
## 23 S288C 2286   .        A    ACTT   111   PASS  6854 0.078494  170
## 24 S288C 2709   .        T       C  6855   PASS 18767 0.031278    0
## 30 S288C 4521   .        T   TCTAC   570   PASS 20490 0.005076  345
## 32 S288C 5633   .        T       C 49314   PASS 23875 0.996817    4
## 34 S288C 6337   .        G       A 49314   PASS  7160 0.999022    0
## 35 S288C 6455   .        T       C 49314   PASS 11717 0.873859    0
## 36 S288C 6611   .        T       C 49314   PASS 14054 0.997367    0
## 41 S288C 8934   .        G       A 49314   PASS 16121 0.998387    0
##                    DP4 INDEL HRUN
## 2  3540,873,5112,4673 INDEL    1
## 4      13,4,8681,7644  <NA>   NA
## 5       3,3,5017,9333  <NA>   NA
## 6       2,3,6747,6883  <NA>   NA
## 7      1,2,12222,4608  <NA>   NA
## 8      7,5,12550,4942  <NA>   NA
## 9     9,3,13432,10934  <NA>   NA
## 10     3,2,8240,16869  <NA>   NA
## 11     0,3,4896,12450  <NA>   NA
## 12  3430,1956,211,852  <NA>   NA
## 14    17,2,7447,16017  <NA>   NA
## 16  723,200,7227,9307 INDEL    1
## 17    57,12,8080,9358  <NA>   NA
## 19      1,5,2368,8570  <NA>   NA
## 22  314,1305,390,5349 INDEL    3
## 23     538,5652,2,536 INDEL    3
## 24  9395,8700,305,282  <NA>   NA
## 30   9077,11415,103,1 INDEL    1
## 32   12,9,11230,12569  <NA>   NA
## 34      1,2,2027,5126  <NA>   NA
## 35 1048,416,7293,2946  <NA>   NA
## 36      3,2,8474,5543  <NA>   NA
## 41     1,5,3177,12918  <NA>   NA
```

```r
#snp, in, del
par(mfrow=c(1,2))
s288c_slices <- c(19,1,3)
sk1_slices <- c(18,2,3)
labels_pie_char <- c("SNP", "In", "Del")

pie(s288c_slices, labels = labels_pie_char, main = "S288C", col = c("skyblue", "pink", "orange"))
pie(sk1_slices, labels = labels_pie_char, main = "SK1", col = c("skyblue", "pink", "orange"))
```





```
#also in the final filtered data, 6 positios are shared between s288c and sk1. they also have the same
```