IDENTIFY PLOIDY WITH HETEROZYGOUS VARIANS

Claudia Zirión-Martínez

```
library(tidyverse)
library(vcfR)
library(pinfsc50)
```

Example in

vcfR documentation

```
vcf_file <- system.file("extdata", "pinf_sc50.vcf.gz",
                        package = "pinfsc50")
vcf <- read.vcfR(vcf_file, verbose = FALSE)
vcf
```

One Desjardins sample with high heterozygosity Bt206

```
sample <- "SRS409075"
```

```
vcf <- read.vcfR(paste("../data/processed/haploid_freebayes/", sample, ".snps.raw.vcf", sep = ""))
```

```
vcf
```

```
***** Object of Class vcfR *****
1 samples
14 CHROMs
253,198 variants
Object size: 167.9 Mb
0 percent missing data
*****        *****         *****
```

Extract genotype

```
head(vcf@gt)
```

```
     FORMAT
[1,] "GT:DP:AD:RO:QR:AO:QA:GL"
[2,] "GT:DP:AD:RO:QR:AO:QA:GL"
[3,] "GT:DP:AD:RO:QR:AO:QA:GL"
[4,] "GT:DP:AD:RO:QR:AO:QA:GL"
[5,] "GT:DP:AD:RO:QR:AO:QA:GL"
[6,] "GT:DP:AD:RO:QR:AO:QA:GL"
     SRS409075
```

```
[1,] "1/1:21:0,21:0:0:21:824:-74.4792,-6.32163,0"
[2,] "1/1:17:1,12:1:38:12:430:-35.2729,-0.483863,0"
[3,] "1/1:45:0,45:0:0:45:1647:-148.482,-13.5464,0"
[4,] "1/1:145:0,145:0:0:145:5363:-482.666,-43.6494,0"
[5,] "1/1:174:0,174:0:0:174:6254:-562.81,-52.3792,0"
[6,] "0/1:126:88,38:88:3055:38:1280:-77.5583,0,-237.21"
```

Extract Allele Depth

```
ad <- extract.gt(vcf, element = 'AD')
head(ad)
```

```
                SRS409075
CP097924.1_62800 "0,21"
CP097924.1_64483 "1,12"
CP097924.1_64582 "0,45"
CP097924.1_64914 "0,145"
CP097924.1_65016 "0,174"
CP097924.1_65307 "88,38"
```

Show rows with at least three alleles

```
ad[grep("[^0],[^0],[^0]", ad[,sample]),]
```

```
 CP097924.1_463018 CP097924.1_1281039 CP097924.1_1284507 CP097924.1_1518648
        "30,3,5,3"           "46,6,4"           "56,8,7"            "1,2,3"
CP097924.1_1755943  CP097925.1_102701  CP097925.1_317617  CP097925.1_979075
         "106,6,8"           "27,2,2"           "33,3,3"           "66,4,6"
CP097925.1_1089153 CP097925.1_1646995 CP097925.1_1750573  CP097926.1_474951
           "1,9,5"          "71,9,13"           "36,3,3"          "96,8,14"
 CP097926.1_546205  CP097926.1_724825  CP097926.1_735673  CP097926.1_882398
         "88,7,11"            "5,2,2"           "1,8,31"         "26,2,3,5"
  CP097927.1_90125  CP097927.1_346135  CP097927.1_525963  CP097927.1_762108
         "115,9,7"           "3,6,18"         "116,8,22"           "61,4,6"
CP097927.1_1089404  CP097928.1_103505  CP097928.1_220853  CP097928.1_241376
          "72,5,7"          "58,4,11"           "88,9,7"           "1,5,13"
 CP097928.1_500988  CP097928.1_501303 CP097928.1_1259131   CP097929.1_66460
          "28,2,2"         "111,8,18"           "43,3,6"       "19,3,2,2,5"
 CP097929.1_294808  CP097929.1_774907  CP097929.1_782246  CP097929.1_936314
        "69,7,8,9"            "8,8,4"           "42,7,5"           "34,6,4"
CP097929.1_1200336 CP097929.1_1351186 CP097929.1_1427349   CP097930.1_86114
           "9,2,3"           "1,8,24"           "37,4,4"         "136,9,23"
CP097930.1_1111664 CP097930.1_1395359  CP097931.1_294073  CP097931.1_563628
          "98,9,7"            "3,8,3"           "2,5,18"            "1,5,9"
CP097931.1_1160078 CP097931.1_1391565  CP097932.1_238222  CP097932.1_322857
        "101,9,17"           "64,7,6"           "34,2,3"            "2,4,8"
 CP097932.1_532865  CP097932.1_920577 CP097932.1_1115274 CP097932.1_1268105
          "66,4,6"           "2,5,25"           "42,5,2"            "1,2,5"
  CP097933.1_11455   CP097933.1_80445  CP097933.1_343861 CP097934.1_1086843
          "1,9,24"            "5,2,2"           "2,4,13"           "58,6,4"
CP097934.1_1091656 CP097934.1_1092055 CP097934.1_1106338 CP097934.1_1653495
```

```
          "32,2,6"           "12,3,2"           "33,2,4"           "1,4,11"
CP097934.1_1690600 CP097934.1_2002929 CP097934.1_2009793 CP097934.1_2028428
          "32,2,4"           "1,7,26"           "1,9,27"          "133,9,8"
 CP097935.1_287138  CP097936.1_206578  CP097936.1_477414  CP097936.1_524717
        "115,7,16"          "113,7,9"           "1,7,29"       "94,7,10,16"
 CP097936.1_575264  CP097936.1_685208  CP097936.1_698352  CP097936.1_832894
           "9,2,2"           "44,7,4"           "2,7,25"          "137,9,6"
 CP097937.1_538169  CP097937.1_986543
        "34,4,3,6"          "126,9,8"
```

Show rows with at least four alleles

```
ad[grep("[^0],[^0],[^0],[^0]", ad[,sample]),]
```

```
CP097926.1_882398  CP097929.1_66460 CP097929.1_294808 CP097937.1_538169
       "26,2,3,5"      "19,3,2,2,5"        "69,7,8,9"        "34,4,3,6"
```

Show rows with at least five alleles

```
ad[grep("[^0],[^0],[^0],[^0],[^0]", ad[,sample]),]
```

```
[1] "19,3,2,2,5"
```

Fraction of depth for each allele

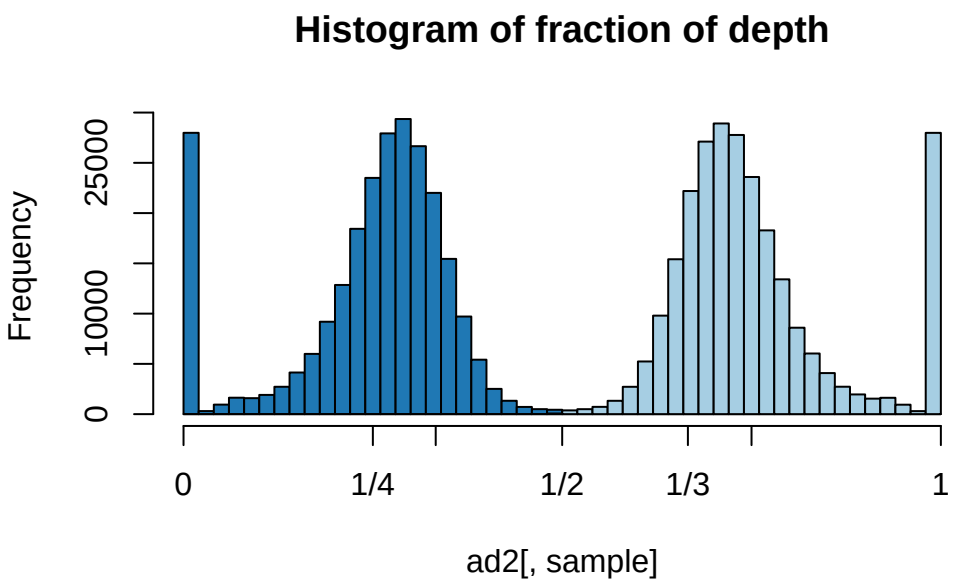
If one variant has AD "3,6" the allele with the lowst depth (*a1*) is found with a depth of 3, and the allele with the highest depth (*a2*) with a depth of 6. *a1* has a fraction of depth of 0.3333333 and *a2* has a fraction of depth of 0.6666667.

Split the fraction of depth of each allele in two tables.

```
allele1 <- masplit(ad, record = 1)
allele2 <- masplit(ad, record = 2)

ad1 <- allele1 / (allele1 + allele2)
ad2 <- allele2 / (allele1 + allele2)
```

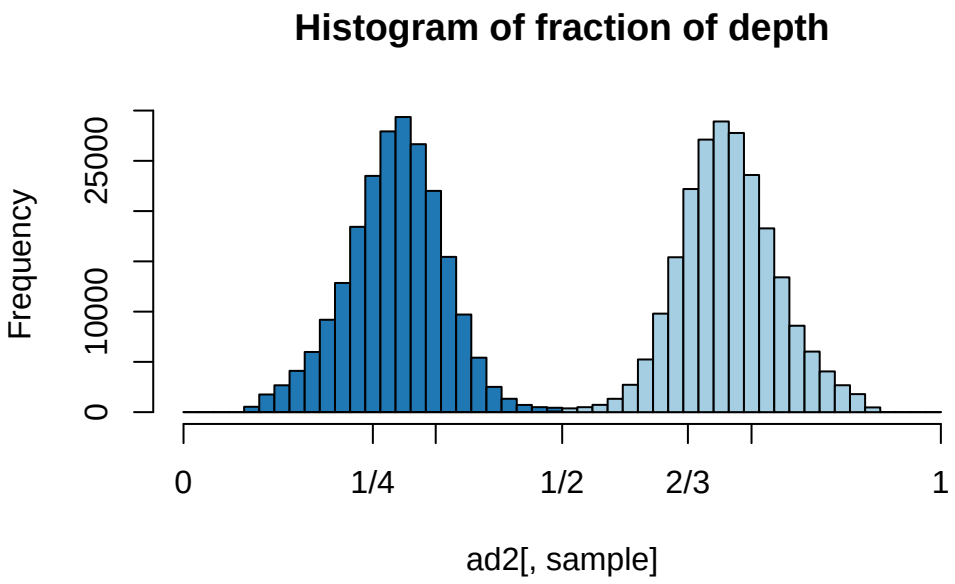
Histogram of frequency of each value of fraction of allele depth. The low frequency alleles in dark blue and the high frequency alleles in ligth blue.

```
hist(ad2[,sample], breaks = seq(0,1,by=0.02), col = "#1f78b4", xaxt="n", main = "Histogram of fraction
hist(ad1[,sample], breaks = seq(0,1,by=0.02), col = "#a6cee3", add = TRUE)
axis(side=1, at=c(0,0.25,0.333,0.5,0.666,0.75,1), labels=c(0,"1/4","1/3","1/2","1/3","3/4",1))
```



Remove homozygote variants and get the frequency of the fraction of the depth again.

```
gt <- extract.gt(vcf, element = 'GT')
hets <- is_het(gt)

is.na( ad[ !hets ] ) <- TRUE

allele1 <- masplit(ad, record = 1)
allele2 <- masplit(ad, record = 2)

ad1 <- allele1 / (allele1 + allele2)
ad2 <- allele2 / (allele1 + allele2)

hist(ad2[,sample], breaks = seq(0,1,by=0.02), col = "#1f78b4", xaxt="n", main = "Histogram of fraction
hist(ad1[,sample], breaks = seq(0,1,by=0.02), col = "#a6cee3", add = TRUE)
axis(side=1, at=c(0,0.25,0.333,0.5,0.666,0.75,1), labels=c(0,"1/4","1/3","1/2","2/3","3/4",1))
```



Allele depth instead of frequency of fraction of depth

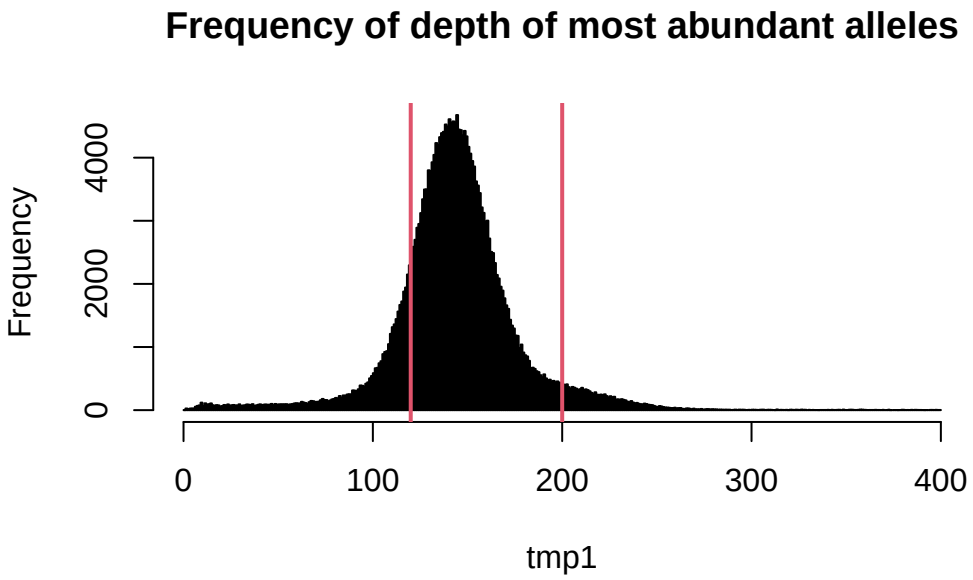
```
ad <- extract.gt(vcf, element = 'AD')

allele1 <- masplit(ad, record = 1)
allele2 <- masplit(ad, record = 2)
```

```
tmp1 <- allele1[,sample]
tmp1 <- tmp1[tmp1 <= 400]
```

```
sums <- apply(allele1, MARGIN=2, quantile, probs=c(0.15, 0.95), na.rm=TRUE)
sums[,sample]
```

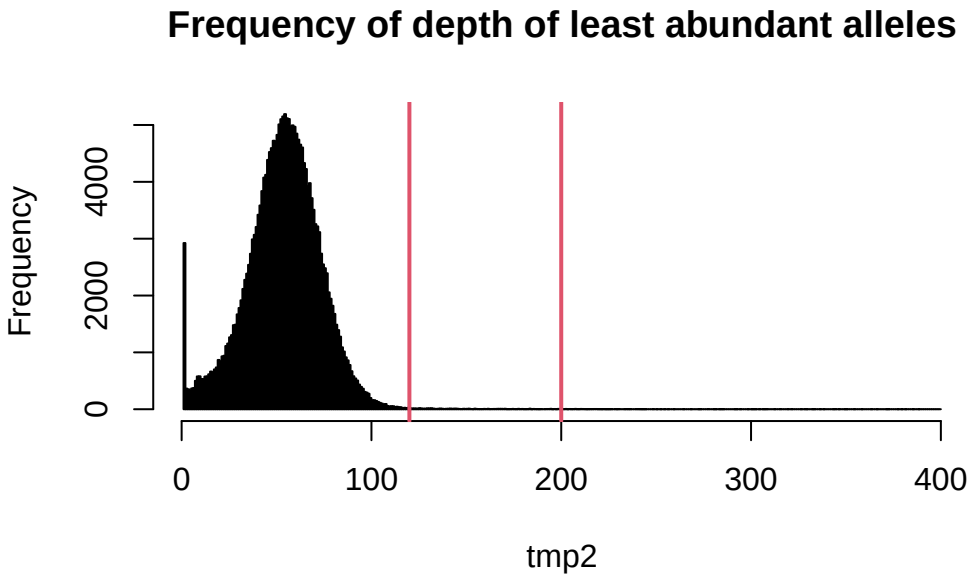
```
15% 95%
120 200
```

```
hist(tmp1, breaks=seq(0,400,by=1), col="#808080", main ="Frequency of depth of most abundant alleles")
abline(v=sums[,sample], col=2, lwd=2)
```




```
tmp2 <- allele2[,sample]
tmp2 <- tmp2[tmp2>0]
tmp2 <- tmp2[tmp2<=400]
```

```
hist(tmp2, breaks=seq(1,400,by=1), col="#808080", main="Frequency of depth of least abundant alleles")
abline(v=sums[,sample], col=2, lwd=2)
```




```
head(allele1)
```

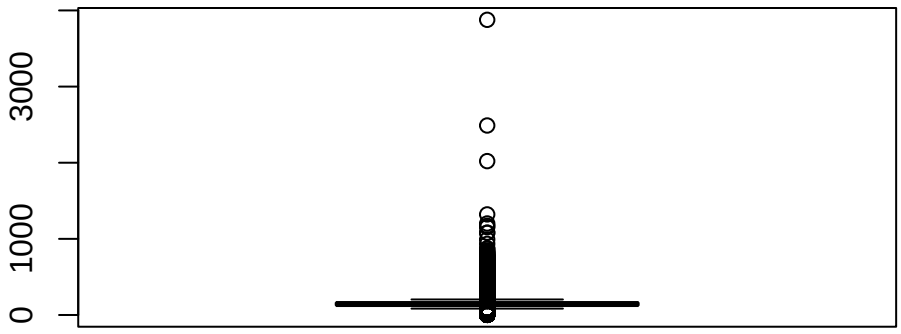
```
                SRS409075
CP097924.1_62800        21
CP097924.1_64483        12
CP097924.1_64582        45
```

```
CP097924.1_64914       145
CP097924.1_65016       174
CP097924.1_65307        88
```

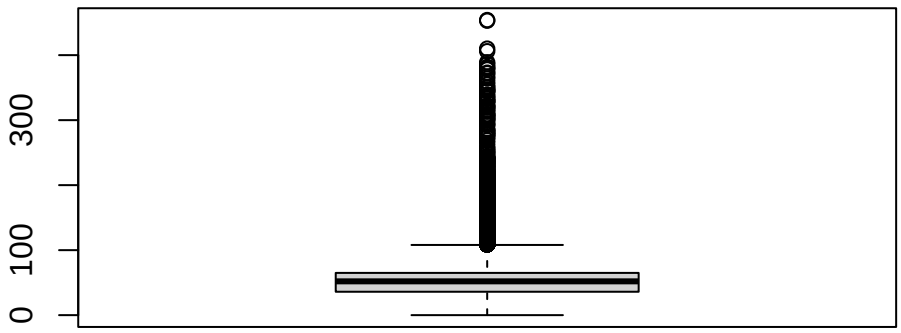
```
head(allele2)
```

```
                SRS409075
CP097924.1_62800         0
CP097924.1_64483         1
CP097924.1_64582         0
CP097924.1_64914         0
CP097924.1_65016         0
CP097924.1_65307        38
```

```
boxplot(allele1)
```




```
boxplot(allele2)
```



Set as NA the variants that are outside the 15 and 95 quantiles.

```
sums <- apply(allele1, MARGIN=2, quantile, probs=c(0.15, 0.95), na.rm=TRUE)
```

Allele 1 (high frequency allele)

Substract the 15 quartile value from the allele1 depth

```
dp2 <- sweep(allele1, MARGIN=2, FUN = "-", sums[1,])
```

In the VCF genotype informations, set to NA the rows that ended up being negative in the dp2 table

```
vcf@gt[,-1][dp2 < 0 & !is.na(vcf@gt[,-1]) ] <- NA
```

Substract the 95 quartile value from the allele1 depth In the VCF genotype informations, set to NA the rows that ended up being negative in the dp2 table

```
dp2 <- sweep(allele1, MARGIN=2, FUN = "-", sums[2,])
vcf@gt[,-1][dp2 > 0] <- NA
```

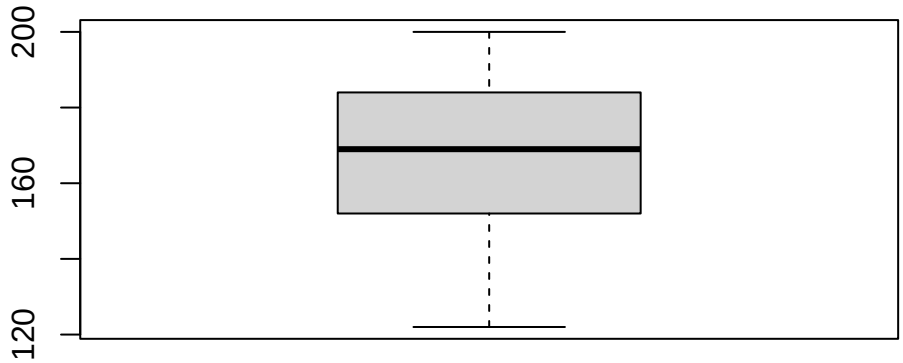
Allele 2 (low frequency allele)

```
dp2 <- sweep(allele2, MARGIN=2, FUN = "-", sums[1,])
vcf@gt[,-1][ dp2 < 0 & !is.na(vcf@gt[,-1]) ] <- NA
```

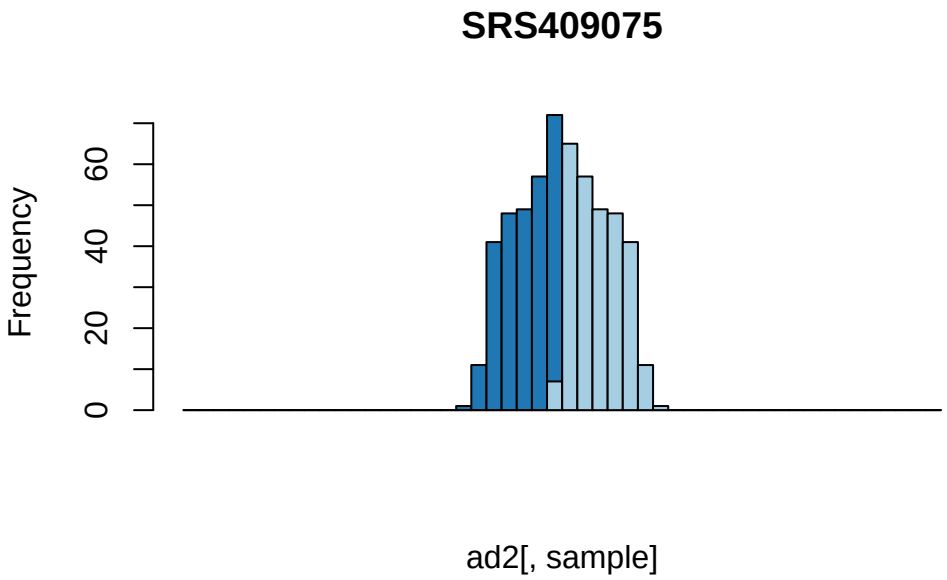
```
dp2 <- sweep(allele2, MARGIN=2, FUN = "-", sums[2,])
vcf@gt[,-1][dp2 > 0] <- NA
```

```
ad <- extract.gt(vcf, element = 'AD')
allele1 <- masplit(ad, record = 1)
```

```
allele2 <- masplit(ad, record = 2)
boxplot(allele1, las=3)
```




```
gt <- extract.gt(vcf, element = 'GT')
hets <- is_het(gt)
is.na( ad[ !hets ] ) <- TRUE

allele1 <- masplit(ad, record = 1)
allele2 <- masplit(ad, record = 2)

ad1 <- allele1 / (allele1 + allele2)
ad2 <- allele2 / (allele1 + allele2)

hist(ad2[,sample], breaks = seq(0,1,by=0.02), col = "#1f78b4", xaxt="n", main=sample)
hist(ad1[,sample], breaks = seq(0,1,by=0.02), col = "#a6cee3", add = TRUE)
```


SRS409075
Frequency
0
20
40
60
ad2[, sample]