IDENTIFY PLOIDY WITH HETEROZYGOUS VARIANS

Claudia Zirión-Martínez

Setup

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

vcfR documentation

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 all info

Extract fixed fields

```
chrom_pos_qual <- as.data.frame(vcf@fix[, c("CHROM", "POS", "QUAL", "REF", "ALT")])
chrom_pos_qual$QUAL <- as.numeric(chrom_pos_qual$QUAL)
```

Extract genotype fields

```
gt <- extract.gt(vcf, element = 'GT')
dp <- extract.gt(vcf, element = 'DP', as.numeric = TRUE)
ad <- extract.gt(vcf, element = 'AD')
```

Extract the type of variant from the INFO field

```
variant_type <- extract.info(vcf, element = "TYPE", as.numeric=FALSE)
# variant_AB <- extract.info(vcf, element = "AB", as.numeric=FALSE)
```

Combine into a dataframe

```
variant_info <- data.frame(
  CHROM = chrom_pos_qual[, "CHROM"],
  POS = chrom_pos_qual[, "POS"],
  QUAL = chrom_pos_qual[, "QUAL"],
  REF = chrom_pos_qual[, "REF"],
  ALT = chrom_pos_qual[, "ALT"],
  GT = gt[, sample],
  DP = dp[, sample],
  AD = ad[, sample],
  TYPE = variant_type
#   AB= variant_AB
  )
```

```
head(variant_info)
```

	CHROM	POS	QUAL	REF	ALT	GT	DP	AD	TYPE
CP097924.1_62800	CP097924.1	62800	679.314	CACCAGGAAGGC	TAACCCC	1/1	21	0,21	complex
CP097924.1_64483	CP097924.1	64483	253.481	ATATA	ATAA	1/1	17	1,12	del
CP097924.1_64582	CP097924.1	64582	1338.720	A	G	1/1	45	0,45	snp
CP097924.1_64914	CP097924.1	64914	4779.050	A	G	1/1	145	0,145	snp
CP097924.1_65016	CP097924.1	65016	5581.510	T	C	1/1	174	0,174	snp
CP097924.1_65307	CP097924.1	65307	465.149	A	G	0/1	126	88,38	snp

Filter variants by quality and depth

QUAL=Phred-scaled probability that the site has no variant.
DP=Total read depth in the site.

```
variant_filtered <- variant_info %>%
    filter(QUAL >= 100)%>%
    filter(DP >=20)
```

```
Number of discarded variants: 12128
```

Separate data for each allele into diferent columns

```
ad_split <- str_split_fixed(variant_filtered$AD, ",", n = max(str_count(variant_filtered$AD, ",") + 1)
type_split <- str_split_fixed(variant_filtered$TYPE, ",", n = max(str_count(variant_filtered$TYPE, ","

ad_split_df <- as.data.frame(ad_split)
type_split_df <- as.data.frame(type_split)

colnames(ad_split_df) <- paste0("AD_", seq_len(ncol(ad_split_df)))
colnames(type_split_df) <- paste0("TYPE_", seq_len(ncol(type_split_df)))

variant_split <- cbind(variant_filtered, ad_split_df, type_split_df)
```

```r
variant_split <- variant_split %>%
    rename(AD_R = AD_1,
            AD_a1 = AD_2,
            AD_a2 = AD_3,
            TYPE_a1 = TYPE_1,
            TYPE_a2 = TYPE_2)

head(variant_split[,6:14])
```

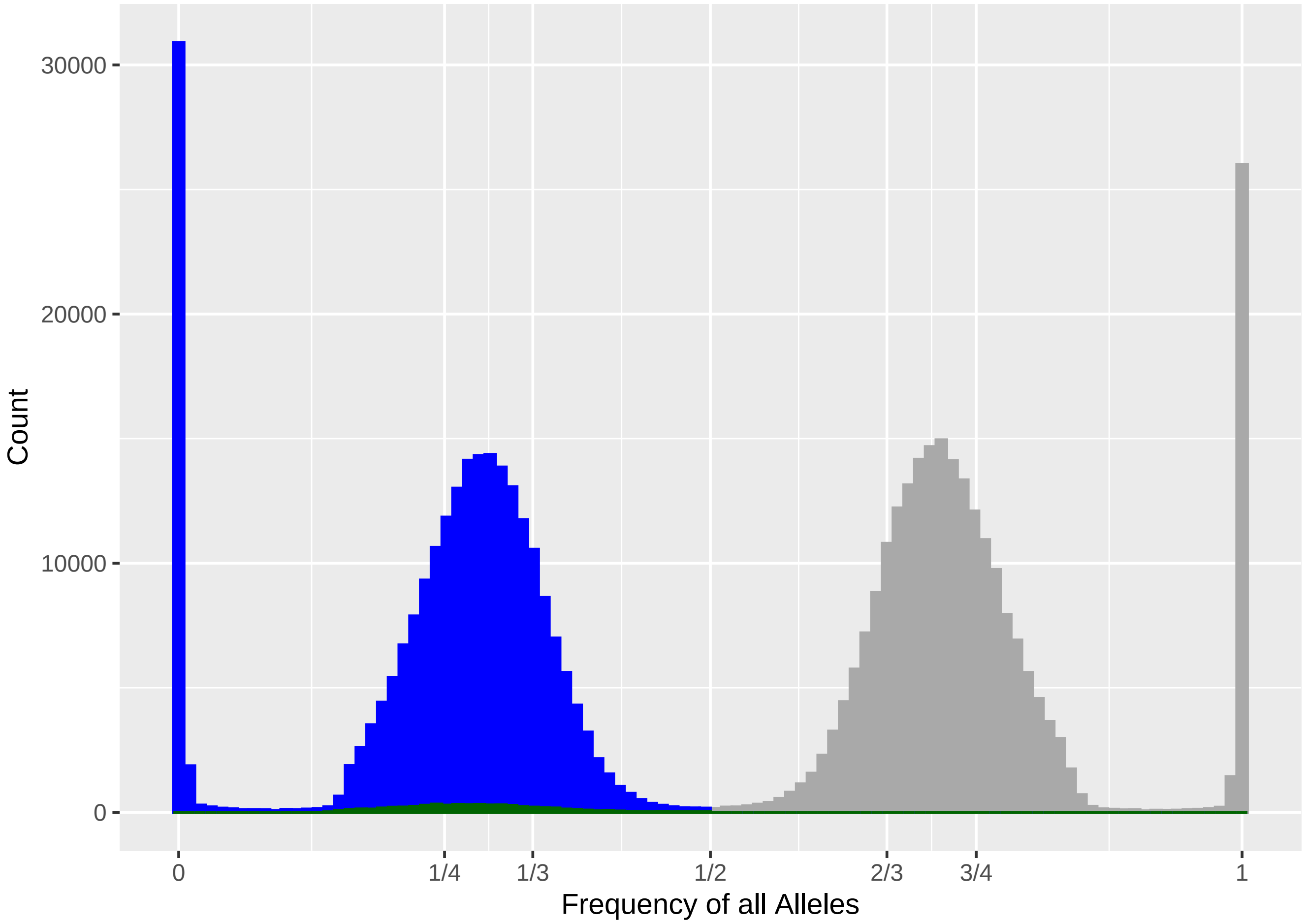
	GT	DP	AD	TYPE	AD_R	AD_a1	AD_a2	TYPE_a1	TYPE_a2
CP097924.1_62800	1/1	21	0,21	complex	0	21		complex	
CP097924.1_64582	1/1	45	0,45	snp	0	45		snp	
CP097924.1_64914	1/1	145	0,145	snp	0	145		snp	
CP097924.1_65016	1/1	174	0,174	snp	0	174		snp	
CP097924.1_65307	0/1	126	88,38	snp	88	38		snp	
CP097924.1_65581	0/1	25	16,9	snp	16	9		snp	

Allelic fractions

DP is not allways the sum of the depth of all alleles (AD_R + AD_a1 + AD_a2) because some reads are counted in the DP but "don't support any allele", so they are not included in AD.

Calculate Total Allele Depth (TAD), get the fraction of depth of each allele. Separate the alleles into allele with major, intermediate and minimal fraction (instead of using the original order).

```r
variant_fractions <- variant_split %>%
    mutate(AD_R = as.numeric(AD_R),
           AD_a1 = as.numeric(AD_a1),
           AD_a2 = as.numeric(AD_a2)) %>%
    mutate(TAD = rowSums(across(c(AD_R, AD_a1, AD_a2), ~replace_na(., 0))),
           AF_R = AD_R / TAD,
           AF_a1 = AD_a1 / TAD,
           AF_a2 = AD_a2 / TAD)%>%
    mutate(AF_aMax = pmax(AF_R, AF_a1, AF_a2, na.rm=TRUE),
           AF_aMed = apply(cbind(AF_R, AF_a1, AF_a2), 1, median),
           AF_aMin = pmin(AF_R, AF_a1, AF_a2, na.rm=TRUE))
```



Separate Homo and Heterozygotes

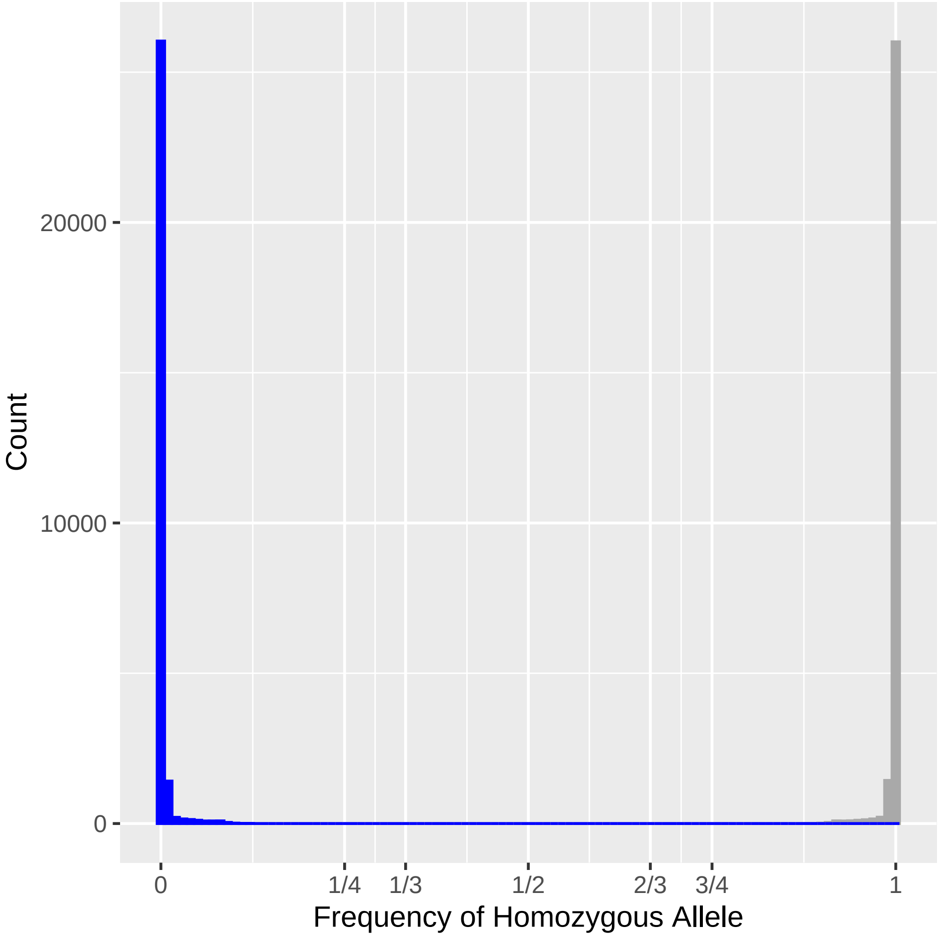
Possible genotypes

```
unique(variant_fractions$GT)
```

```
[1] "1/1" "0/1" "1/2" "0/2"
```

Name	GT	Description
homo_alt	“1/1”	The reference allele has 0 or very small depth
hetero_1alt	“0/1”	Heterozygous with one alternate allele
hetero_2alt	“0/2” or “1/2”	Heterozygous with two alternate allele

```
homo_alt <- variant_fractions %>%
    filter(GT == "1/1")
hetero_1alt <- variant_fractions %>%
    filter(GT == "0/1")
hetero_2alt <- variant_fractions %>%
    filter(GT %in% c("1/2","0/2"))
```



We can see that there are some variants that are not *true* homozygotes, so we will filter by the values of the frequency of alleles instead.

```
homo_alt <- variant_fractions %>%
    filter(AF_aMin == 0 & AF_aMax == 1)
hetero_1alt <- variant_fractions %>%
    filter(AF_aMin != 0 & is.na(AF_aMed))
hetero_2alt <- variant_fractions %>%
    filter(!is.na(AF_aMed))
```

```
Number of rows in homo_alt: 25080
```

```
Number of rows in hetero_1alt: 210274
```

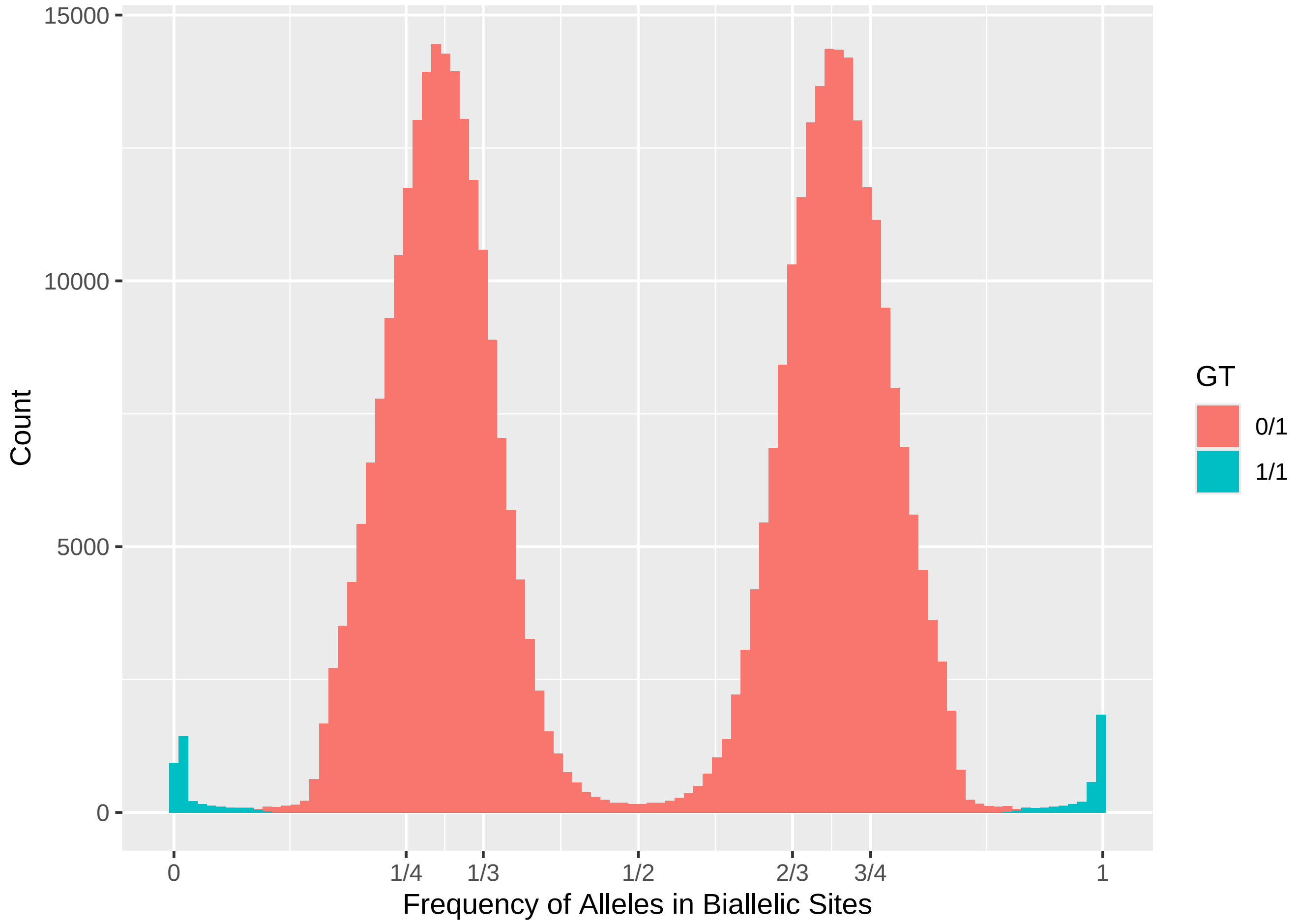
```
Number of rows in hetero_2alt: 5716
```

Get median and standard deviation of Allele frequencies

```
hetero_a1_AF_stats <- hetero_1alt %>%
    summarise(
        median_AF_aMax = median(AF_aMax, na.rm = TRUE),
        mean_AF_aMax = mean(AF_aMax, na.rm = TRUE),
        sd_AF_aMax = sd(AF_aMax, na.rm = TRUE),
        median_AF_aMin = median(AF_aMin, na.rm = TRUE),
        mean_AF_aMin = mean(AF_aMin, na.rm = TRUE),
        sd_AF_aMin = sd(AF_aMin, na.rm = TRUE)
    )
hetero_a1_AF_stats
```

median_AF_aMax	mean_AF_aMax	sd_AF_aMax	median_AF_aMin	mean_AF_aMin	sd_AF_aMin
0.715736	0.71888	0.0664512	0.284264	0.28112	0.0664512

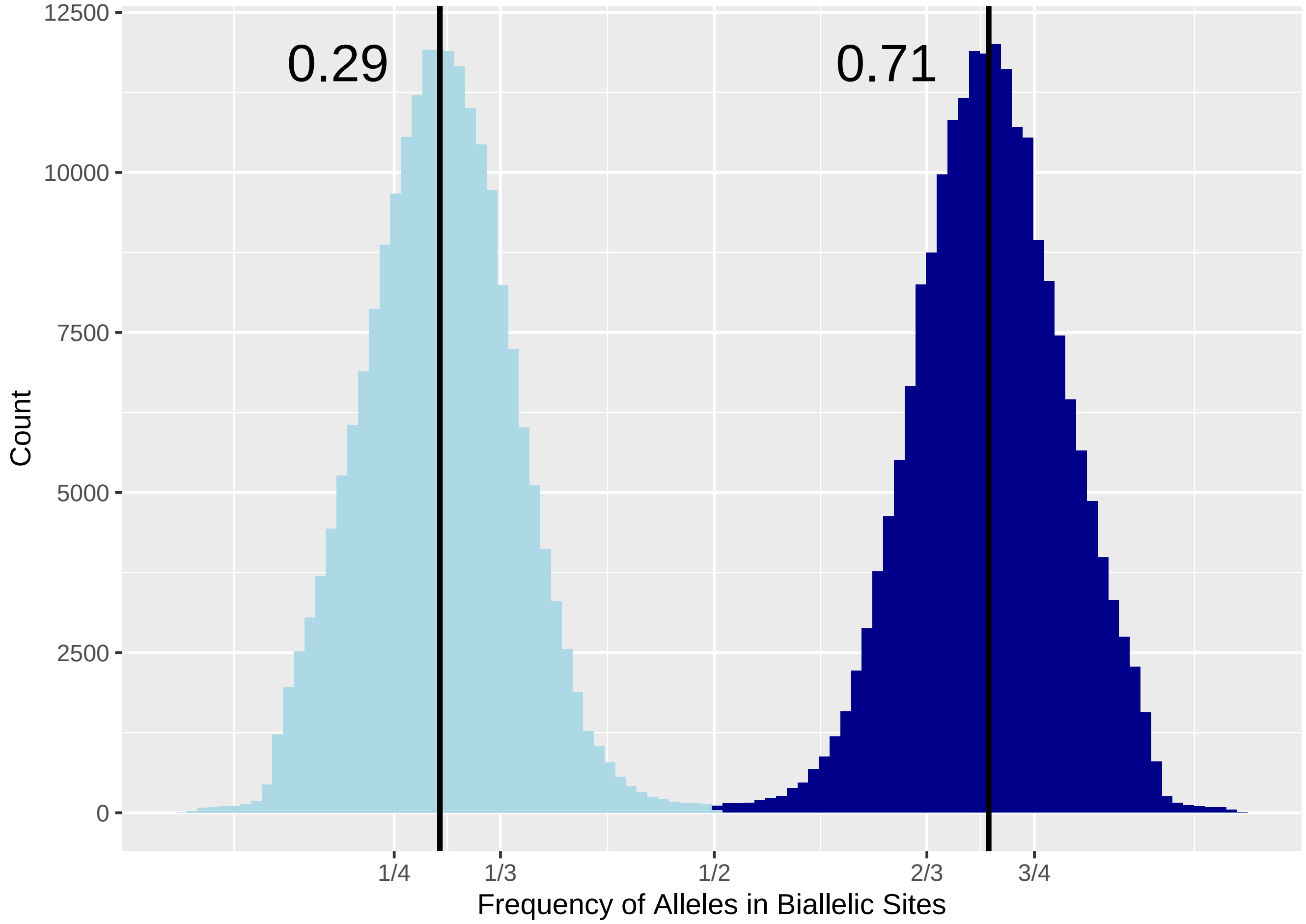
We will color by genotype because we have mixed *called* genotypes within the heterozygotes.



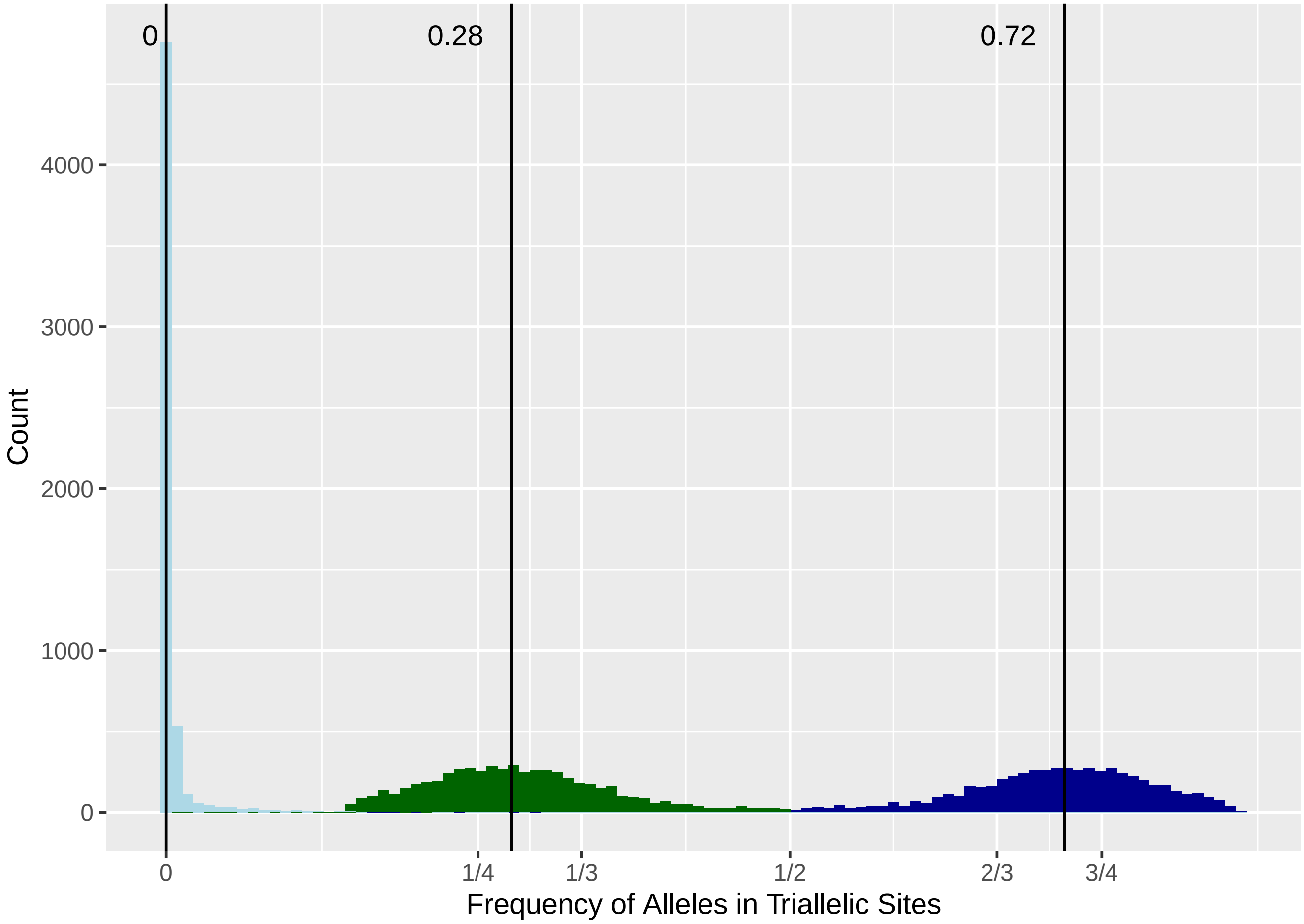
Filter out the *called* homozygotes because they have a different distribution due to very low depth of the reference allele (the alternative allele is not in all reads).

```
hetero_1alt <- hetero_1alt %>%
    filter(GT != "1/1")
```

```
hetero_a1_AF_stats <- hetero_1alt %>%
    summarise(
        median_AF_aMax = median(AF_aMax, na.rm = TRUE),
        mean_AF_aMax = mean(AF_aMax, na.rm = TRUE),
        sd_AF_aMax = sd(AF_aMax, na.rm = TRUE),
        median_AF_aMin = median(AF_aMin, na.rm = TRUE),
        mean_AF_aMin = mean(AF_aMin, na.rm = TRUE),
        sd_AF_aMin = sd(AF_aMin, na.rm = TRUE)
    )
```



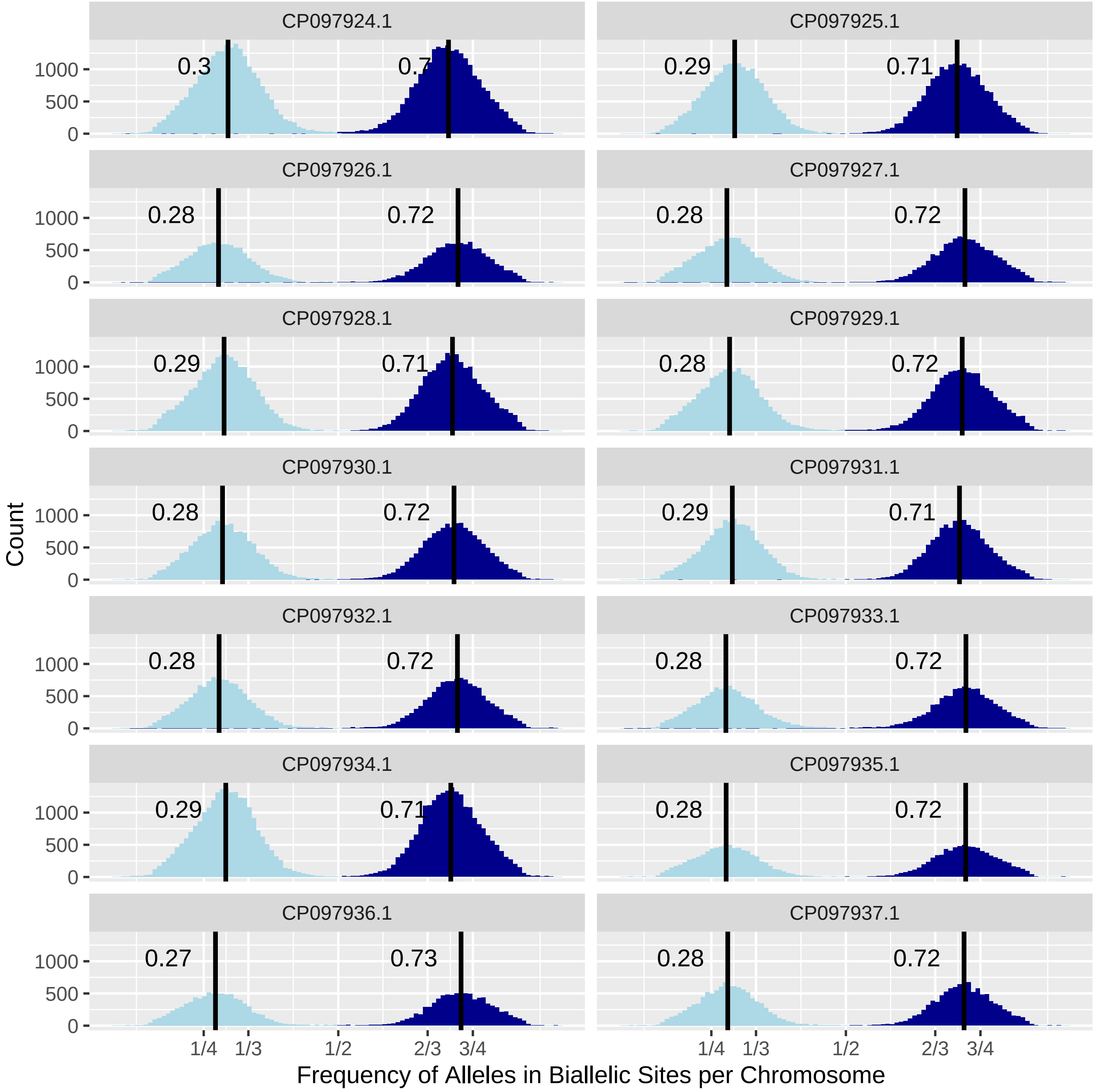
```
hetero_a2_AF_stats <- hetero_2alt %>%
    summarise(
        median_AF_aMax = median(AF_aMax, na.rm = TRUE),
        mean_AF_aMax = mean(AF_aMax, na.rm = TRUE),
        sd_AF_aMax = sd(AF_aMax, na.rm = TRUE),
        median_AF_aMed = median(AF_aMed, na.rm = TRUE),
        mean_AF_aMed = mean(AF_aMed, na.rm = TRUE),
        sd_AF_aMed = sd(AF_aMed, na.rm = TRUE),
        median_AF_aMin = median(AF_aMin, na.rm = TRUE),
        mean_AF_aMin = mean(AF_aMin, na.rm = TRUE),
        sd_AF_aMin = sd(AF_aMin, na.rm = TRUE)
    )
```



The triallelic variants have very similar distributions as the biallelic ones, so it looks like there are biallelic sites where the reference allele was detected in a very low proportion of reads.

By chromosome

```
hetero_a1_AF_stats_chrom <- hetero_1alt %>%
    group_by(CHROM)%>%
    summarise(
        median_AF_aMax = median(AF_aMax, na.rm = TRUE),
        mean_AF_aMax = mean(AF_aMax, na.rm = TRUE),
        sd_AF_aMax = sd(AF_aMax, na.rm = TRUE),
        median_AF_aMin = median(AF_aMin, na.rm = TRUE),
        mean_AF_aMin = mean(AF_aMin, na.rm = TRUE),
        sd_AF_aMin = sd(AF_aMin, na.rm = TRUE)
    )
```


CP097924.1
0.3
0.7
CP097925.1
0.29
0.71
CP097926.1
0.28
0.72
CP097927.1
0.28
0.72
CP097928.1
0.29
0.71
CP097929.1
0.28
0.72
CP097930.1
0.28
0.72
CP097931.1
0.29
0.71
CP097932.1
0.28
0.72
CP097933.1
0.28
0.72
CP097934.1
0.29
0.71
CP097935.1
0.28
0.72
CP097936.1
0.27
0.73
CP097937.1
0.28
0.72
Count
1000
500
0
1/4
1/3
1/2
2/3
3/4
Frequency of Alleles in Biallelic Sites per Chromosome