assignment1_VCFpreliminaryanalysis_Rscript_MeenaEaswaran.R

meena

2025-01-06

```
#set working directory
#setwd()

#note the system time
Sys.time()
```

```
## [1] "2025-01-06 13:10:37 PST"
```

```
#install vcfR library if required
#install.packages("vcfR")

#load vcfR library
library(vcfR)
```

```
##
##    *****       ***   vcfR   ***       *****
##    This is vcfR 1.15.0
##      browseVignettes('vcfR') # Documentation
##      citation('vcfR') # Citation
##    *****       *****      *****       *****
```

```
#read the genotype VCF file
vcf <- read.vcfR("assignment1_geno.vcf", verbose = FALSE)
vcf
```

```
## ***** Object of Class vcfR *****
## 750 samples
## 2 CHROMs
## 6,500 variants
## Object size: 39.6 Mb
## 0 percent missing data
## *****        *****         *****
```

```
# Check the first 6 lines of the VCF to get information on the meta, fixed and genotype sections
head(vcf)
```

```
## [1] "***** Object of class 'vcfR' *****"
## [1] "***** Meta section *****"
## [1] "##fileformat=VCFv4.1"
```

```
## [1] "##FILTER=<ID=PASS,Description=\"All filters passed\">"
## [1] "##fileDate=20150218"
## [1] "##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/refe [Truncated]"
## [1] "##source=1000GenomesPhase3Pipeline"
## [1] "##contig=<ID=1,assembly=b37,length=249250621>"
## [1] "First 6 rows."
## [1]
## [1] "***** Fixed section *****"
##      CHROM POS        ID            REF ALT   QUAL  FILTER
## [1,] "10"  "96419645" "rs531623987" "A" "C"   "100" "PASS"
## [2,] "10"  "96419648" "rs12414245"  "G" "A,C" "100" "PASS"
## [3,] "10"  "96419704" "rs527990601" "G" "A"   "100" "PASS"
## [4,] "10"  "96419750" "rs548052622" "G" "A"   "100" "PASS"
## [5,] "10"  "96419760" "rs567861799" "A" "G"   "100" "PASS"
## [6,] "10"  "96419766" "rs10882489"  "G" "A"   "100" "PASS"
## [1]
## [1] "***** Genotype section *****"
##      FORMAT HG00096 HG00097 HG00099 HG00100 HG00101
## [1,] "GT"   "0|0"   "0|0"   "0|0"   "0|0"   "0|0"
## [2,] "GT"   "0|0"   "0|0"   "0|0"   "0|0"   "0|0"
## [3,] "GT"   "0|0"   "0|0"   "0|0"   "0|0"   "0|0"
## [4,] "GT"   "0|0"   "0|0"   "0|0"   "0|0"   "0|0"
## [5,] "GT"   "0|0"   "0|0"   "0|0"   "0|0"   "0|0"
## [6,] "GT"   "1|1"   "1|1"   "1|1"   "1|1"   "1|1"
## [1] "First 6 columns only."
## [1]
## [1] "Unique GT formats:"
## [1] "GT"
## [1]
```

```
#Preliminary analysis with VCF:
#a. Does the file contain diploid alleles or haploid?
#Based on the above commands, the file seems to contain diploid alleles (2 chromosome sets)
#In a VCF file, diploid genotypes are typically represented as two alleles separated by either "/" or "
#Confirm with the commands below

# Create a data frame that maps the unique IDs to the original IDs
#unique ID needed for genotype extraction as ID contains non-unique names (duplicates)
id_map <- data.frame(
  UniqueID = paste0("var", seq_len(nrow(vcf@fix))),
  OriginalID = vcf@fix[,"ID"]
)
#Unique IDS Will be needed for mapping original variants at step 3

#Check for potential duplicates in the original variant IDs
duplicates_origignalvariantID <- id_map[duplicated(id_map$OriginalID) | duplicated(id_map$OriginalID, fi
duplicates_origignalvariantID
```

```
##      UniqueID  OriginalID
## 5129  var5129 rs112607901
## 5130  var5130 rs112607901
## 5131  var5131 rs112607901
```

```r
#duplicates present for one variant rs112607901

# Assign unique IDs to the variants
vcf@fix[,"ID"] <- id_map$UniqueID

# Extract genotype data with element GT
geno_data <- extract.gt(vcf, element = "GT", as.numeric = FALSE)

# Check if the genotypes are diploid or haploid
if (any(grepl("/|\\|", geno_data))) {
  print("The VCF file contains diploid alleles.")
} else {
  print("The VCF file contains haploid alleles.")
}
```

```
## [1] "The VCF file contains diploid alleles."
```

```r
#Alternatively, check for unique genotypes
unique(as.vector(geno_data))
```

```
##  [1] "0|0" "1|1" "1|0" "0|1" "2|0" "0|2" "1|2" "2|2" "1|3" "2|3" "0|3" "3|0"
## [13] "2|1" "3|2" "3|1" "3|3"
```

```r
#The outcomes give unique genotypes in terms of ploidy in the VCF file
#all are diploid and phased

# Convert geno_data to a data frame
geno_data <- as.data.frame(geno_data)

#####################################################################

#b. To check if the alleles are phased or unphased:

#Phased alleles are denoted by | and Unphased alleles are denoted by /
#Based on the above command, when checked for unique genotypes, all alleles are phased
#Confirm by using the following commands

# Check if the genotype data contains the phasing character "|"
phased_info <- grep("|", vcf@gt, value = TRUE)

# If the length of the phased_info vector is greater than 0, the alleles are phased
if(length(phased_info) > 0) {
  cat("The alleles are phased.")
} else {
  cat("The alleles are unphased.")
}
```

```
## The alleles are phased.
```

```r
#####################################################################

#c. To count how many variants have passed quality control:
```

```r
# Check unique quality scores and unique filter notations
unique_qual_scores <- unique(vcf@fix[, "QUAL"])
unique_qual_scores
```

```
## [1] "100"
```

```r
unique_filter <- unique(vcf@fix[, "FILTER"])
unique_filter
```

```
## [1] "PASS"
```

```r
#Based on above command, 100 is the only unique QUAL score and PASS is the only unique FILTER notation
#This preliminary indicates all 6500 variants passed the quality control. Rule out the role of NAs as b

# Count the number of variants that have a non-NA quality score
passed_variants <- sum(!is.na(vcf@fix[, "QUAL"]))
passed_variants
```

```
## [1] 6500
```

```r
######################################################################

#d. To count how many SNPs exist in the file

vcf2 <- extract.indels(vcf, return.indels = FALSE)
#Setting return.indels false, will return only SNPs. If TRUE, it will return only indels
vcf2
```

```
## ***** Object of Class vcfR *****
## 750 samples
## 2 CHROMs
## 6,265 variants
## Object size: 38.1 Mb
## 0 percent missing data
## *****        *****         *****
```

```r
#Extract information
info_vcf2 <- extract_info_tidy(vcf2)

#Ensure if SNP is only unique form in the variant type column
unique(info_vcf2$VT)
```

```
## [1] "SNP"
```

```r
#to know the length or count of the revised VCF
SNP_count <- length(vcf2@fix[, "ID"])
SNP_count
```

```
## [1] 6265
```

```
#####################################################################

#e. How many indels exist in the file? Can you tell how many of them are deletions?

vcf3 <- extract.indels(vcf, return.indels = TRUE)
#Setting return.indels TRUE, will return only indels
vcf3
```

```
## ***** Object of Class vcfR *****
## 750 samples
## 2 CHROMs
## 235 variants
## Object size: 1.5 Mb
## 0 percent missing data
## *****        *****         *****
```

```
#Extract information
info_vcf3<- extract_info_tidy(vcf3)

#Ensure if indel is only unique form in the variant type column
#This should have no SNP
unique(info_vcf3$VT)
```

```
## [1] "INDEL"     "SNP,INDEL" "SV"
```

```
#this is not the case as "INDEL", "SNP,INDEL" "SV"  do come up

#to get only indel count
# Get unique counts for each type of variant present (, "INDEL", "SNP,INDEL", "SV")
counts_vcf3 <- table(info_vcf3$VT)
counts_vcf3
```

```
##
##     INDEL SNP,INDEL        SV
##       227         5         3
```

```
#explore metadata
vcf@meta
```

```
##   [1] "##fileformat=VCFv4.1"
##   [2] "##FILTER=<ID=PASS,Description=\"All filters passed\">"
##   [3] "##fileDate=20150218"
##   [4] "##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_ass
##   [5] "##source=1000GenomesPhase3Pipeline"
##   [6] "##contig=<ID=1,assembly=b37,length=249250621>"
##   [7] "##contig=<ID=2,assembly=b37,length=243199373>"
##   [8] "##contig=<ID=3,assembly=b37,length=198022430>"
##   [9] "##contig=<ID=4,assembly=b37,length=191154276>"
##  [10] "##contig=<ID=5,assembly=b37,length=180915260>"
##  [11] "##contig=<ID=6,assembly=b37,length=171115067>"
##  [12] "##contig=<ID=7,assembly=b37,length=159138663>"
```

```
##  [13] "##contig=<ID=8,assembly=b37,length=146364022>"
##  [14] "##contig=<ID=9,assembly=b37,length=141213431>"
##  [15] "##contig=<ID=10,assembly=b37,length=135534747>"
##  [16] "##contig=<ID=11,assembly=b37,length=135006516>"
##  [17] "##contig=<ID=12,assembly=b37,length=133851895>"
##  [18] "##contig=<ID=13,assembly=b37,length=115169878>"
##  [19] "##contig=<ID=14,assembly=b37,length=107349540>"
##  [20] "##contig=<ID=15,assembly=b37,length=102531392>"
##  [21] "##contig=<ID=16,assembly=b37,length=90354753>"
##  [22] "##contig=<ID=17,assembly=b37,length=81195210>"
##  [23] "##contig=<ID=18,assembly=b37,length=78077248>"
##  [24] "##contig=<ID=19,assembly=b37,length=59128983>"
##  [25] "##contig=<ID=20,assembly=b37,length=63025520>"
##  [26] "##contig=<ID=21,assembly=b37,length=48129895>"
##  [27] "##contig=<ID=22,assembly=b37,length=51304566>"
##  [28] "##contig=<ID=GL000191.1,assembly=b37,length=106433>"
##  [29] "##contig=<ID=GL000192.1,assembly=b37,length=547496>"
##  [30] "##contig=<ID=GL000193.1,assembly=b37,length=189789>"
##  [31] "##contig=<ID=GL000194.1,assembly=b37,length=191469>"
##  [32] "##contig=<ID=GL000195.1,assembly=b37,length=182896>"
##  [33] "##contig=<ID=GL000196.1,assembly=b37,length=38914>"
##  [34] "##contig=<ID=GL000197.1,assembly=b37,length=37175>"
##  [35] "##contig=<ID=GL000198.1,assembly=b37,length=90085>"
##  [36] "##contig=<ID=GL000199.1,assembly=b37,length=169874>"
##  [37] "##contig=<ID=GL000200.1,assembly=b37,length=187035>"
##  [38] "##contig=<ID=GL000201.1,assembly=b37,length=36148>"
##  [39] "##contig=<ID=GL000202.1,assembly=b37,length=40103>"
##  [40] "##contig=<ID=GL000203.1,assembly=b37,length=37498>"
##  [41] "##contig=<ID=GL000204.1,assembly=b37,length=81310>"
##  [42] "##contig=<ID=GL000205.1,assembly=b37,length=174588>"
##  [43] "##contig=<ID=GL000206.1,assembly=b37,length=41001>"
##  [44] "##contig=<ID=GL000207.1,assembly=b37,length=4262>"
##  [45] "##contig=<ID=GL000208.1,assembly=b37,length=92689>"
##  [46] "##contig=<ID=GL000209.1,assembly=b37,length=159169>"
##  [47] "##contig=<ID=GL000210.1,assembly=b37,length=27682>"
##  [48] "##contig=<ID=GL000211.1,assembly=b37,length=166566>"
##  [49] "##contig=<ID=GL000212.1,assembly=b37,length=186858>"
##  [50] "##contig=<ID=GL000213.1,assembly=b37,length=164239>"
##  [51] "##contig=<ID=GL000214.1,assembly=b37,length=137718>"
##  [52] "##contig=<ID=GL000215.1,assembly=b37,length=172545>"
##  [53] "##contig=<ID=GL000216.1,assembly=b37,length=172294>"
##  [54] "##contig=<ID=GL000217.1,assembly=b37,length=172149>"
##  [55] "##contig=<ID=GL000218.1,assembly=b37,length=161147>"
##  [56] "##contig=<ID=GL000219.1,assembly=b37,length=179198>"
##  [57] "##contig=<ID=GL000220.1,assembly=b37,length=161802>"
##  [58] "##contig=<ID=GL000221.1,assembly=b37,length=155397>"
##  [59] "##contig=<ID=GL000222.1,assembly=b37,length=186861>"
##  [60] "##contig=<ID=GL000223.1,assembly=b37,length=180455>"
##  [61] "##contig=<ID=GL000224.1,assembly=b37,length=179693>"
##  [62] "##contig=<ID=GL000225.1,assembly=b37,length=211173>"
##  [63] "##contig=<ID=GL000226.1,assembly=b37,length=15008>"
##  [64] "##contig=<ID=GL000227.1,assembly=b37,length=128374>"
##  [65] "##contig=<ID=GL000228.1,assembly=b37,length=129120>"
##  [66] "##contig=<ID=GL000229.1,assembly=b37,length=19913>"
```

```
##  [67] "##contig=<ID=GL000230.1,assembly=b37,length=43691>"
##  [68] "##contig=<ID=GL000231.1,assembly=b37,length=27386>"
##  [69] "##contig=<ID=GL000232.1,assembly=b37,length=40652>"
##  [70] "##contig=<ID=GL000233.1,assembly=b37,length=45941>"
##  [71] "##contig=<ID=GL000234.1,assembly=b37,length=40531>"
##  [72] "##contig=<ID=GL000235.1,assembly=b37,length=34474>"
##  [73] "##contig=<ID=GL000236.1,assembly=b37,length=41934>"
##  [74] "##contig=<ID=GL000237.1,assembly=b37,length=45867>"
##  [75] "##contig=<ID=GL000238.1,assembly=b37,length=39939>"
##  [76] "##contig=<ID=GL000239.1,assembly=b37,length=33824>"
##  [77] "##contig=<ID=GL000240.1,assembly=b37,length=41933>"
##  [78] "##contig=<ID=GL000241.1,assembly=b37,length=42152>"
##  [79] "##contig=<ID=GL000242.1,assembly=b37,length=43523>"
##  [80] "##contig=<ID=GL000243.1,assembly=b37,length=43341>"
##  [81] "##contig=<ID=GL000244.1,assembly=b37,length=39929>"
##  [82] "##contig=<ID=GL000245.1,assembly=b37,length=36651>"
##  [83] "##contig=<ID=GL000246.1,assembly=b37,length=38154>"
##  [84] "##contig=<ID=GL000247.1,assembly=b37,length=36422>"
##  [85] "##contig=<ID=GL000248.1,assembly=b37,length=39786>"
##  [86] "##contig=<ID=GL000249.1,assembly=b37,length=38502>"
##  [87] "##contig=<ID=MT,assembly=b37,length=16569>"
##  [88] "##contig=<ID=NC_007605,assembly=b37,length=171823>"
##  [89] "##contig=<ID=X,assembly=b37,length=155270560>"
##  [90] "##contig=<ID=Y,assembly=b37,length=59373566>"
##  [91] "##contig=<ID=hs37d5,assembly=b37,length=35477943>"
##  [92] "##ALT=<ID=CNV,Description=\"Copy Number Polymorphism\">"
##  [93] "##ALT=<ID=DEL,Description=\"Deletion\">"
##  [94] "##ALT=<ID=DUP,Description=\"Duplication\">"
##  [95] "##ALT=<ID=INS:ME:ALU,Description=\"Insertion of ALU element\">"
##  [96] "##ALT=<ID=INS:ME:LINE1,Description=\"Insertion of LINE1 element\">"
##  [97] "##ALT=<ID=INS:ME:SVA,Description=\"Insertion of SVA element\">"
##  [98] "##ALT=<ID=INS:MT,Description=\"Nuclear Mitochondrial Insertion\">"
##  [99] "##ALT=<ID=INV,Description=\"Inversion\">"
## [100] "##ALT=<ID=CN0,Description=\"Copy number allele: 0 copies\">"
## [101] "##ALT=<ID=CN1,Description=\"Copy number allele: 1 copy\">"
## [102] "##ALT=<ID=CN2,Description=\"Copy number allele: 2 copies\">"
## [103] "##ALT=<ID=CN3,Description=\"Copy number allele: 3 copies\">"
## [104] "##ALT=<ID=CN4,Description=\"Copy number allele: 4 copies\">"
## [105] "##ALT=<ID=CN5,Description=\"Copy number allele: 5 copies\">"
## [106] "##ALT=<ID=CN6,Description=\"Copy number allele: 6 copies\">"
## [107] "##ALT=<ID=CN7,Description=\"Copy number allele: 7 copies\">"
## [108] "##ALT=<ID=CN8,Description=\"Copy number allele: 8 copies\">"
## [109] "##ALT=<ID=CN9,Description=\"Copy number allele: 9 copies\">"
## [110] "##ALT=<ID=CN10,Description=\"Copy number allele: 10 copies\">"
## [111] "##ALT=<ID=CN11,Description=\"Copy number allele: 11 copies\">"
## [112] "##ALT=<ID=CN12,Description=\"Copy number allele: 12 copies\">"
## [113] "##ALT=<ID=CN13,Description=\"Copy number allele: 13 copies\">"
## [114] "##ALT=<ID=CN14,Description=\"Copy number allele: 14 copies\">"
## [115] "##ALT=<ID=CN15,Description=\"Copy number allele: 15 copies\">"
## [116] "##ALT=<ID=CN16,Description=\"Copy number allele: 16 copies\">"
## [117] "##ALT=<ID=CN17,Description=\"Copy number allele: 17 copies\">"
## [118] "##ALT=<ID=CN18,Description=\"Copy number allele: 18 copies\">"
## [119] "##ALT=<ID=CN19,Description=\"Copy number allele: 19 copies\">"
## [120] "##ALT=<ID=CN20,Description=\"Copy number allele: 20 copies\">"
```

```
## [121] "##ALT=<ID=CN21,Description=\"Copy number allele: 21 copies\">"
## [122] "##ALT=<ID=CN22,Description=\"Copy number allele: 22 copies\">"
## [123] "##ALT=<ID=CN23,Description=\"Copy number allele: 23 copies\">"
## [124] "##ALT=<ID=CN24,Description=\"Copy number allele: 24 copies\">"
## [125] "##ALT=<ID=CN25,Description=\"Copy number allele: 25 copies\">"
## [126] "##ALT=<ID=CN26,Description=\"Copy number allele: 26 copies\">"
## [127] "##ALT=<ID=CN27,Description=\"Copy number allele: 27 copies\">"
## [128] "##ALT=<ID=CN28,Description=\"Copy number allele: 28 copies\">"
## [129] "##ALT=<ID=CN29,Description=\"Copy number allele: 29 copies\">"
## [130] "##ALT=<ID=CN30,Description=\"Copy number allele: 30 copies\">"
## [131] "##ALT=<ID=CN31,Description=\"Copy number allele: 31 copies\">"
## [132] "##ALT=<ID=CN32,Description=\"Copy number allele: 32 copies\">"
## [133] "##ALT=<ID=CN33,Description=\"Copy number allele: 33 copies\">"
## [134] "##ALT=<ID=CN34,Description=\"Copy number allele: 34 copies\">"
## [135] "##ALT=<ID=CN35,Description=\"Copy number allele: 35 copies\">"
## [136] "##ALT=<ID=CN36,Description=\"Copy number allele: 36 copies\">"
## [137] "##ALT=<ID=CN37,Description=\"Copy number allele: 37 copies\">"
## [138] "##ALT=<ID=CN38,Description=\"Copy number allele: 38 copies\">"
## [139] "##ALT=<ID=CN39,Description=\"Copy number allele: 39 copies\">"
## [140] "##ALT=<ID=CN40,Description=\"Copy number allele: 40 copies\">"
## [141] "##ALT=<ID=CN41,Description=\"Copy number allele: 41 copies\">"
## [142] "##ALT=<ID=CN42,Description=\"Copy number allele: 42 copies\">"
## [143] "##ALT=<ID=CN43,Description=\"Copy number allele: 43 copies\">"
## [144] "##ALT=<ID=CN44,Description=\"Copy number allele: 44 copies\">"
## [145] "##ALT=<ID=CN45,Description=\"Copy number allele: 45 copies\">"
## [146] "##ALT=<ID=CN46,Description=\"Copy number allele: 46 copies\">"
## [147] "##ALT=<ID=CN47,Description=\"Copy number allele: 47 copies\">"
## [148] "##ALT=<ID=CN48,Description=\"Copy number allele: 48 copies\">"
## [149] "##ALT=<ID=CN49,Description=\"Copy number allele: 49 copies\">"
## [150] "##ALT=<ID=CN50,Description=\"Copy number allele: 50 copies\">"
## [151] "##ALT=<ID=CN51,Description=\"Copy number allele: 51 copies\">"
## [152] "##ALT=<ID=CN52,Description=\"Copy number allele: 52 copies\">"
## [153] "##ALT=<ID=CN53,Description=\"Copy number allele: 53 copies\">"
## [154] "##ALT=<ID=CN54,Description=\"Copy number allele: 54 copies\">"
## [155] "##ALT=<ID=CN55,Description=\"Copy number allele: 55 copies\">"
## [156] "##ALT=<ID=CN56,Description=\"Copy number allele: 56 copies\">"
## [157] "##ALT=<ID=CN57,Description=\"Copy number allele: 57 copies\">"
## [158] "##ALT=<ID=CN58,Description=\"Copy number allele: 58 copies\">"
## [159] "##ALT=<ID=CN59,Description=\"Copy number allele: 59 copies\">"
## [160] "##ALT=<ID=CN60,Description=\"Copy number allele: 60 copies\">"
## [161] "##ALT=<ID=CN61,Description=\"Copy number allele: 61 copies\">"
## [162] "##ALT=<ID=CN62,Description=\"Copy number allele: 62 copies\">"
## [163] "##ALT=<ID=CN63,Description=\"Copy number allele: 63 copies\">"
## [164] "##ALT=<ID=CN64,Description=\"Copy number allele: 64 copies\">"
## [165] "##ALT=<ID=CN65,Description=\"Copy number allele: 65 copies\">"
## [166] "##ALT=<ID=CN66,Description=\"Copy number allele: 66 copies\">"
## [167] "##ALT=<ID=CN67,Description=\"Copy number allele: 67 copies\">"
## [168] "##ALT=<ID=CN68,Description=\"Copy number allele: 68 copies\">"
## [169] "##ALT=<ID=CN69,Description=\"Copy number allele: 69 copies\">"
## [170] "##ALT=<ID=CN70,Description=\"Copy number allele: 70 copies\">"
## [171] "##ALT=<ID=CN71,Description=\"Copy number allele: 71 copies\">"
## [172] "##ALT=<ID=CN72,Description=\"Copy number allele: 72 copies\">"
## [173] "##ALT=<ID=CN73,Description=\"Copy number allele: 73 copies\">"
## [174] "##ALT=<ID=CN74,Description=\"Copy number allele: 74 copies\">"
```

```
## [175] "##ALT=<ID=CN75,Description=\"Copy number allele: 75 copies\">"
## [176] "##ALT=<ID=CN76,Description=\"Copy number allele: 76 copies\">"
## [177] "##ALT=<ID=CN77,Description=\"Copy number allele: 77 copies\">"
## [178] "##ALT=<ID=CN78,Description=\"Copy number allele: 78 copies\">"
## [179] "##ALT=<ID=CN79,Description=\"Copy number allele: 79 copies\">"
## [180] "##ALT=<ID=CN80,Description=\"Copy number allele: 80 copies\">"
## [181] "##ALT=<ID=CN81,Description=\"Copy number allele: 81 copies\">"
## [182] "##ALT=<ID=CN82,Description=\"Copy number allele: 82 copies\">"
## [183] "##ALT=<ID=CN83,Description=\"Copy number allele: 83 copies\">"
## [184] "##ALT=<ID=CN84,Description=\"Copy number allele: 84 copies\">"
## [185] "##ALT=<ID=CN85,Description=\"Copy number allele: 85 copies\">"
## [186] "##ALT=<ID=CN86,Description=\"Copy number allele: 86 copies\">"
## [187] "##ALT=<ID=CN87,Description=\"Copy number allele: 87 copies\">"
## [188] "##ALT=<ID=CN88,Description=\"Copy number allele: 88 copies\">"
## [189] "##ALT=<ID=CN89,Description=\"Copy number allele: 89 copies\">"
## [190] "##ALT=<ID=CN90,Description=\"Copy number allele: 90 copies\">"
## [191] "##ALT=<ID=CN91,Description=\"Copy number allele: 91 copies\">"
## [192] "##ALT=<ID=CN92,Description=\"Copy number allele: 92 copies\">"
## [193] "##ALT=<ID=CN93,Description=\"Copy number allele: 93 copies\">"
## [194] "##ALT=<ID=CN94,Description=\"Copy number allele: 94 copies\">"
## [195] "##ALT=<ID=CN95,Description=\"Copy number allele: 95 copies\">"
## [196] "##ALT=<ID=CN96,Description=\"Copy number allele: 96 copies\">"
## [197] "##ALT=<ID=CN97,Description=\"Copy number allele: 97 copies\">"
## [198] "##ALT=<ID=CN98,Description=\"Copy number allele: 98 copies\">"
## [199] "##ALT=<ID=CN99,Description=\"Copy number allele: 99 copies\">"
## [200] "##ALT=<ID=CN100,Description=\"Copy number allele: 100 copies\">"
## [201] "##ALT=<ID=CN101,Description=\"Copy number allele: 101 copies\">"
## [202] "##ALT=<ID=CN102,Description=\"Copy number allele: 102 copies\">"
## [203] "##ALT=<ID=CN103,Description=\"Copy number allele: 103 copies\">"
## [204] "##ALT=<ID=CN104,Description=\"Copy number allele: 104 copies\">"
## [205] "##ALT=<ID=CN105,Description=\"Copy number allele: 105 copies\">"
## [206] "##ALT=<ID=CN106,Description=\"Copy number allele: 106 copies\">"
## [207] "##ALT=<ID=CN107,Description=\"Copy number allele: 107 copies\">"
## [208] "##ALT=<ID=CN108,Description=\"Copy number allele: 108 copies\">"
## [209] "##ALT=<ID=CN109,Description=\"Copy number allele: 109 copies\">"
## [210] "##ALT=<ID=CN110,Description=\"Copy number allele: 110 copies\">"
## [211] "##ALT=<ID=CN111,Description=\"Copy number allele: 111 copies\">"
## [212] "##ALT=<ID=CN112,Description=\"Copy number allele: 112 copies\">"
## [213] "##ALT=<ID=CN113,Description=\"Copy number allele: 113 copies\">"
## [214] "##ALT=<ID=CN114,Description=\"Copy number allele: 114 copies\">"
## [215] "##ALT=<ID=CN115,Description=\"Copy number allele: 115 copies\">"
## [216] "##ALT=<ID=CN116,Description=\"Copy number allele: 116 copies\">"
## [217] "##ALT=<ID=CN117,Description=\"Copy number allele: 117 copies\">"
## [218] "##ALT=<ID=CN118,Description=\"Copy number allele: 118 copies\">"
## [219] "##ALT=<ID=CN119,Description=\"Copy number allele: 119 copies\">"
## [220] "##ALT=<ID=CN120,Description=\"Copy number allele: 120 copies\">"
## [221] "##ALT=<ID=CN121,Description=\"Copy number allele: 121 copies\">"
## [222] "##ALT=<ID=CN122,Description=\"Copy number allele: 122 copies\">"
## [223] "##ALT=<ID=CN123,Description=\"Copy number allele: 123 copies\">"
## [224] "##ALT=<ID=CN124,Description=\"Copy number allele: 124 copies\">"
## [225] "##FORMAT=<ID=GT,Number=1,Type=String,Description=\"Genotype\">"
## [226] "##INFO=<ID=CIEND,Number=2,Type=Integer,Description=\"Confidence interval around END for impre
## [227] "##INFO=<ID=CIPOS,Number=2,Type=Integer,Description=\"Confidence interval around POS for impre
## [228] "##INFO=<ID=CS,Number=1,Type=String,Description=\"Source call set.\">"
```

```
## [229] "##INFO=<ID=END,Number=1,Type=Integer,Description=\"End coordinate of this variant\">"
## [230] "##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description=\"Imprecise structural variation\">"
## [231] "##INFO=<ID=MC,Number=.,Type=String,Description=\"Merged calls.\">"
## [232] "##INFO=<ID=MEINFO,Number=4,Type=String,Description=\"Mobile element info of the form NAME,STA
## [233] "##INFO=<ID=MEND,Number=1,Type=Integer,Description=\"Mitochondrial end coordinate of inserted
## [234] "##INFO=<ID=MLEN,Number=1,Type=Integer,Description=\"Estimated length of mitochondrial insert\
## [235] "##INFO=<ID=MSTART,Number=1,Type=Integer,Description=\"Mitochondrial start coordinate of inser
## [236] "##INFO=<ID=SVLEN,Number=.,Type=Integer,Description=\"SV length. It is only calculated for str
## [237] "##INFO=<ID=SVTYPE,Number=1,Type=String,Description=\"Type of structural variant\">"
## [238] "##INFO=<ID=TSD,Number=1,Type=String,Description=\"Precise Target Site Duplication for bases,
## [239] "##INFO=<ID=AC,Number=A,Type=Integer,Description=\"Total number of alternate alleles in called
## [240] "##INFO=<ID=AF,Number=A,Type=Float,Description=\"Estimated allele frequency in the range (0,1)
## [241] "##INFO=<ID=NS,Number=1,Type=Integer,Description=\"Number of samples with data\">"
## [242] "##INFO=<ID=AN,Number=1,Type=Integer,Description=\"Total number of alleles in called genotypes
## [243] "##INFO=<ID=EAS_AF,Number=A,Type=Float,Description=\"Allele frequency in the EAS populations ca
## [244] "##INFO=<ID=EUR_AF,Number=A,Type=Float,Description=\"Allele frequency in the EUR populations ca
## [245] "##INFO=<ID=AFR_AF,Number=A,Type=Float,Description=\"Allele frequency in the AFR populations ca
## [246] "##INFO=<ID=AMR_AF,Number=A,Type=Float,Description=\"Allele frequency in the AMR populations ca
## [247] "##INFO=<ID=SAS_AF,Number=A,Type=Float,Description=\"Allele frequency in the SAS populations ca
## [248] "##INFO=<ID=DP,Number=1,Type=Integer,Description=\"Total read depth; only low coverage data wer
## [249] "##INFO=<ID=AA,Number=1,Type=String,Description=\"Ancestral Allele. Format: AA|REF|ALT|IndelTyp
## [250] "##INFO=<ID=VT,Number=.,Type=String,Description=\"indicates what type of variant the line repre
## [251] "##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description=\"indicates whether a variant is within th
## [252] "##INFO=<ID=MULTI_ALLELIC,Number=0,Type=Flag,Description=\"indicates whether a site is multi-al
```

```
#line 249,  have annotations in Ancestral Allele (AA) column about indels; specific insertions or delet

# Install dplyr if required
#install.packages("dplyr")

# Load dplyr package
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Extract items with deletion mentioned in the AA column
deletion_rows <- info_vcf3 %>%
  filter(grepl("deletion", AA, ignore.case = TRUE))
nrow(deletion_rows)
```

```
## [1] 47
```

```r
#confirm if deletion_rows are part of only indel category and not SNP, INDEL
unique(deletion_rows$VT)
```

```
## [1] "INDEL"
```

```r
######################################################################

#e.For structural variations (copy number variations)

#explore metadata
vcf@meta
```

```
##   [1] "##fileformat=VCFv4.1"
##   [2] "##FILTER=<ID=PASS,Description=\"All filters passed\">"
##   [3] "##fileDate=20150218"
##   [4] "##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_ass
##   [5] "##source=1000GenomesPhase3Pipeline"
##   [6] "##contig=<ID=1,assembly=b37,length=249250621>"
##   [7] "##contig=<ID=2,assembly=b37,length=243199373>"
##   [8] "##contig=<ID=3,assembly=b37,length=198022430>"
##   [9] "##contig=<ID=4,assembly=b37,length=191154276>"
##  [10] "##contig=<ID=5,assembly=b37,length=180915260>"
##  [11] "##contig=<ID=6,assembly=b37,length=171115067>"
##  [12] "##contig=<ID=7,assembly=b37,length=159138663>"
##  [13] "##contig=<ID=8,assembly=b37,length=146364022>"
##  [14] "##contig=<ID=9,assembly=b37,length=141213431>"
##  [15] "##contig=<ID=10,assembly=b37,length=135534747>"
##  [16] "##contig=<ID=11,assembly=b37,length=135006516>"
##  [17] "##contig=<ID=12,assembly=b37,length=133851895>"
##  [18] "##contig=<ID=13,assembly=b37,length=115169878>"
##  [19] "##contig=<ID=14,assembly=b37,length=107349540>"
##  [20] "##contig=<ID=15,assembly=b37,length=102531392>"
##  [21] "##contig=<ID=16,assembly=b37,length=90354753>"
##  [22] "##contig=<ID=17,assembly=b37,length=81195210>"
##  [23] "##contig=<ID=18,assembly=b37,length=78077248>"
##  [24] "##contig=<ID=19,assembly=b37,length=59128983>"
##  [25] "##contig=<ID=20,assembly=b37,length=63025520>"
##  [26] "##contig=<ID=21,assembly=b37,length=48129895>"
##  [27] "##contig=<ID=22,assembly=b37,length=51304566>"
##  [28] "##contig=<ID=GL000191.1,assembly=b37,length=106433>"
##  [29] "##contig=<ID=GL000192.1,assembly=b37,length=547496>"
##  [30] "##contig=<ID=GL000193.1,assembly=b37,length=189789>"
##  [31] "##contig=<ID=GL000194.1,assembly=b37,length=191469>"
##  [32] "##contig=<ID=GL000195.1,assembly=b37,length=182896>"
##  [33] "##contig=<ID=GL000196.1,assembly=b37,length=38914>"
##  [34] "##contig=<ID=GL000197.1,assembly=b37,length=37175>"
##  [35] "##contig=<ID=GL000198.1,assembly=b37,length=90085>"
##  [36] "##contig=<ID=GL000199.1,assembly=b37,length=169874>"
##  [37] "##contig=<ID=GL000200.1,assembly=b37,length=187035>"
##  [38] "##contig=<ID=GL000201.1,assembly=b37,length=36148>"
##  [39] "##contig=<ID=GL000202.1,assembly=b37,length=40103>"
##  [40] "##contig=<ID=GL000203.1,assembly=b37,length=37498>"
##  [41] "##contig=<ID=GL000204.1,assembly=b37,length=81310>"
```

```
##  [42] "##contig=<ID=GL000205.1,assembly=b37,length=174588>"
##  [43] "##contig=<ID=GL000206.1,assembly=b37,length=41001>"
##  [44] "##contig=<ID=GL000207.1,assembly=b37,length=4262>"
##  [45] "##contig=<ID=GL000208.1,assembly=b37,length=92689>"
##  [46] "##contig=<ID=GL000209.1,assembly=b37,length=159169>"
##  [47] "##contig=<ID=GL000210.1,assembly=b37,length=27682>"
##  [48] "##contig=<ID=GL000211.1,assembly=b37,length=166566>"
##  [49] "##contig=<ID=GL000212.1,assembly=b37,length=186858>"
##  [50] "##contig=<ID=GL000213.1,assembly=b37,length=164239>"
##  [51] "##contig=<ID=GL000214.1,assembly=b37,length=137718>"
##  [52] "##contig=<ID=GL000215.1,assembly=b37,length=172545>"
##  [53] "##contig=<ID=GL000216.1,assembly=b37,length=172294>"
##  [54] "##contig=<ID=GL000217.1,assembly=b37,length=172149>"
##  [55] "##contig=<ID=GL000218.1,assembly=b37,length=161147>"
##  [56] "##contig=<ID=GL000219.1,assembly=b37,length=179198>"
##  [57] "##contig=<ID=GL000220.1,assembly=b37,length=161802>"
##  [58] "##contig=<ID=GL000221.1,assembly=b37,length=155397>"
##  [59] "##contig=<ID=GL000222.1,assembly=b37,length=186861>"
##  [60] "##contig=<ID=GL000223.1,assembly=b37,length=180455>"
##  [61] "##contig=<ID=GL000224.1,assembly=b37,length=179693>"
##  [62] "##contig=<ID=GL000225.1,assembly=b37,length=211173>"
##  [63] "##contig=<ID=GL000226.1,assembly=b37,length=15008>"
##  [64] "##contig=<ID=GL000227.1,assembly=b37,length=128374>"
##  [65] "##contig=<ID=GL000228.1,assembly=b37,length=129120>"
##  [66] "##contig=<ID=GL000229.1,assembly=b37,length=19913>"
##  [67] "##contig=<ID=GL000230.1,assembly=b37,length=43691>"
##  [68] "##contig=<ID=GL000231.1,assembly=b37,length=27386>"
##  [69] "##contig=<ID=GL000232.1,assembly=b37,length=40652>"
##  [70] "##contig=<ID=GL000233.1,assembly=b37,length=45941>"
##  [71] "##contig=<ID=GL000234.1,assembly=b37,length=40531>"
##  [72] "##contig=<ID=GL000235.1,assembly=b37,length=34474>"
##  [73] "##contig=<ID=GL000236.1,assembly=b37,length=41934>"
##  [74] "##contig=<ID=GL000237.1,assembly=b37,length=45867>"
##  [75] "##contig=<ID=GL000238.1,assembly=b37,length=39939>"
##  [76] "##contig=<ID=GL000239.1,assembly=b37,length=33824>"
##  [77] "##contig=<ID=GL000240.1,assembly=b37,length=41933>"
##  [78] "##contig=<ID=GL000241.1,assembly=b37,length=42152>"
##  [79] "##contig=<ID=GL000242.1,assembly=b37,length=43523>"
##  [80] "##contig=<ID=GL000243.1,assembly=b37,length=43341>"
##  [81] "##contig=<ID=GL000244.1,assembly=b37,length=39929>"
##  [82] "##contig=<ID=GL000245.1,assembly=b37,length=36651>"
##  [83] "##contig=<ID=GL000246.1,assembly=b37,length=38154>"
##  [84] "##contig=<ID=GL000247.1,assembly=b37,length=36422>"
##  [85] "##contig=<ID=GL000248.1,assembly=b37,length=39786>"
##  [86] "##contig=<ID=GL000249.1,assembly=b37,length=38502>"
##  [87] "##contig=<ID=MT,assembly=b37,length=16569>"
##  [88] "##contig=<ID=NC_007605,assembly=b37,length=171823>"
##  [89] "##contig=<ID=X,assembly=b37,length=155270560>"
##  [90] "##contig=<ID=Y,assembly=b37,length=59373566>"
##  [91] "##contig=<ID=hs37d5,assembly=b37,length=35477943>"
##  [92] "##ALT=<ID=CNV,Description=\"Copy Number Polymorphism\">"
##  [93] "##ALT=<ID=DEL,Description=\"Deletion\">"
##  [94] "##ALT=<ID=DUP,Description=\"Duplication\">"
##  [95] "##ALT=<ID=INS:ME:ALU,Description=\"Insertion of ALU element\">"
```

```
##  [96] "##ALT=<ID=INS:ME:LINE1,Description=\"Insertion of LINE1 element\">"
##  [97] "##ALT=<ID=INS:ME:SVA,Description=\"Insertion of SVA element\">"
##  [98] "##ALT=<ID=INS:MT,Description=\"Nuclear Mitochondrial Insertion\">"
##  [99] "##ALT=<ID=INV,Description=\"Inversion\">"
## [100] "##ALT=<ID=CN0,Description=\"Copy number allele: 0 copies\">"
## [101] "##ALT=<ID=CN1,Description=\"Copy number allele: 1 copy\">"
## [102] "##ALT=<ID=CN2,Description=\"Copy number allele: 2 copies\">"
## [103] "##ALT=<ID=CN3,Description=\"Copy number allele: 3 copies\">"
## [104] "##ALT=<ID=CN4,Description=\"Copy number allele: 4 copies\">"
## [105] "##ALT=<ID=CN5,Description=\"Copy number allele: 5 copies\">"
## [106] "##ALT=<ID=CN6,Description=\"Copy number allele: 6 copies\">"
## [107] "##ALT=<ID=CN7,Description=\"Copy number allele: 7 copies\">"
## [108] "##ALT=<ID=CN8,Description=\"Copy number allele: 8 copies\">"
## [109] "##ALT=<ID=CN9,Description=\"Copy number allele: 9 copies\">"
## [110] "##ALT=<ID=CN10,Description=\"Copy number allele: 10 copies\">"
## [111] "##ALT=<ID=CN11,Description=\"Copy number allele: 11 copies\">"
## [112] "##ALT=<ID=CN12,Description=\"Copy number allele: 12 copies\">"
## [113] "##ALT=<ID=CN13,Description=\"Copy number allele: 13 copies\">"
## [114] "##ALT=<ID=CN14,Description=\"Copy number allele: 14 copies\">"
## [115] "##ALT=<ID=CN15,Description=\"Copy number allele: 15 copies\">"
## [116] "##ALT=<ID=CN16,Description=\"Copy number allele: 16 copies\">"
## [117] "##ALT=<ID=CN17,Description=\"Copy number allele: 17 copies\">"
## [118] "##ALT=<ID=CN18,Description=\"Copy number allele: 18 copies\">"
## [119] "##ALT=<ID=CN19,Description=\"Copy number allele: 19 copies\">"
## [120] "##ALT=<ID=CN20,Description=\"Copy number allele: 20 copies\">"
## [121] "##ALT=<ID=CN21,Description=\"Copy number allele: 21 copies\">"
## [122] "##ALT=<ID=CN22,Description=\"Copy number allele: 22 copies\">"
## [123] "##ALT=<ID=CN23,Description=\"Copy number allele: 23 copies\">"
## [124] "##ALT=<ID=CN24,Description=\"Copy number allele: 24 copies\">"
## [125] "##ALT=<ID=CN25,Description=\"Copy number allele: 25 copies\">"
## [126] "##ALT=<ID=CN26,Description=\"Copy number allele: 26 copies\">"
## [127] "##ALT=<ID=CN27,Description=\"Copy number allele: 27 copies\">"
## [128] "##ALT=<ID=CN28,Description=\"Copy number allele: 28 copies\">"
## [129] "##ALT=<ID=CN29,Description=\"Copy number allele: 29 copies\">"
## [130] "##ALT=<ID=CN30,Description=\"Copy number allele: 30 copies\">"
## [131] "##ALT=<ID=CN31,Description=\"Copy number allele: 31 copies\">"
## [132] "##ALT=<ID=CN32,Description=\"Copy number allele: 32 copies\">"
## [133] "##ALT=<ID=CN33,Description=\"Copy number allele: 33 copies\">"
## [134] "##ALT=<ID=CN34,Description=\"Copy number allele: 34 copies\">"
## [135] "##ALT=<ID=CN35,Description=\"Copy number allele: 35 copies\">"
## [136] "##ALT=<ID=CN36,Description=\"Copy number allele: 36 copies\">"
## [137] "##ALT=<ID=CN37,Description=\"Copy number allele: 37 copies\">"
## [138] "##ALT=<ID=CN38,Description=\"Copy number allele: 38 copies\">"
## [139] "##ALT=<ID=CN39,Description=\"Copy number allele: 39 copies\">"
## [140] "##ALT=<ID=CN40,Description=\"Copy number allele: 40 copies\">"
## [141] "##ALT=<ID=CN41,Description=\"Copy number allele: 41 copies\">"
## [142] "##ALT=<ID=CN42,Description=\"Copy number allele: 42 copies\">"
## [143] "##ALT=<ID=CN43,Description=\"Copy number allele: 43 copies\">"
## [144] "##ALT=<ID=CN44,Description=\"Copy number allele: 44 copies\">"
## [145] "##ALT=<ID=CN45,Description=\"Copy number allele: 45 copies\">"
## [146] "##ALT=<ID=CN46,Description=\"Copy number allele: 46 copies\">"
## [147] "##ALT=<ID=CN47,Description=\"Copy number allele: 47 copies\">"
## [148] "##ALT=<ID=CN48,Description=\"Copy number allele: 48 copies\">"
## [149] "##ALT=<ID=CN49,Description=\"Copy number allele: 49 copies\">"
```

```
## [150] "##ALT=<ID=CN50,Description=\"Copy number allele: 50 copies\">"
## [151] "##ALT=<ID=CN51,Description=\"Copy number allele: 51 copies\">"
## [152] "##ALT=<ID=CN52,Description=\"Copy number allele: 52 copies\">"
## [153] "##ALT=<ID=CN53,Description=\"Copy number allele: 53 copies\">"
## [154] "##ALT=<ID=CN54,Description=\"Copy number allele: 54 copies\">"
## [155] "##ALT=<ID=CN55,Description=\"Copy number allele: 55 copies\">"
## [156] "##ALT=<ID=CN56,Description=\"Copy number allele: 56 copies\">"
## [157] "##ALT=<ID=CN57,Description=\"Copy number allele: 57 copies\">"
## [158] "##ALT=<ID=CN58,Description=\"Copy number allele: 58 copies\">"
## [159] "##ALT=<ID=CN59,Description=\"Copy number allele: 59 copies\">"
## [160] "##ALT=<ID=CN60,Description=\"Copy number allele: 60 copies\">"
## [161] "##ALT=<ID=CN61,Description=\"Copy number allele: 61 copies\">"
## [162] "##ALT=<ID=CN62,Description=\"Copy number allele: 62 copies\">"
## [163] "##ALT=<ID=CN63,Description=\"Copy number allele: 63 copies\">"
## [164] "##ALT=<ID=CN64,Description=\"Copy number allele: 64 copies\">"
## [165] "##ALT=<ID=CN65,Description=\"Copy number allele: 65 copies\">"
## [166] "##ALT=<ID=CN66,Description=\"Copy number allele: 66 copies\">"
## [167] "##ALT=<ID=CN67,Description=\"Copy number allele: 67 copies\">"
## [168] "##ALT=<ID=CN68,Description=\"Copy number allele: 68 copies\">"
## [169] "##ALT=<ID=CN69,Description=\"Copy number allele: 69 copies\">"
## [170] "##ALT=<ID=CN70,Description=\"Copy number allele: 70 copies\">"
## [171] "##ALT=<ID=CN71,Description=\"Copy number allele: 71 copies\">"
## [172] "##ALT=<ID=CN72,Description=\"Copy number allele: 72 copies\">"
## [173] "##ALT=<ID=CN73,Description=\"Copy number allele: 73 copies\">"
## [174] "##ALT=<ID=CN74,Description=\"Copy number allele: 74 copies\">"
## [175] "##ALT=<ID=CN75,Description=\"Copy number allele: 75 copies\">"
## [176] "##ALT=<ID=CN76,Description=\"Copy number allele: 76 copies\">"
## [177] "##ALT=<ID=CN77,Description=\"Copy number allele: 77 copies\">"
## [178] "##ALT=<ID=CN78,Description=\"Copy number allele: 78 copies\">"
## [179] "##ALT=<ID=CN79,Description=\"Copy number allele: 79 copies\">"
## [180] "##ALT=<ID=CN80,Description=\"Copy number allele: 80 copies\">"
## [181] "##ALT=<ID=CN81,Description=\"Copy number allele: 81 copies\">"
## [182] "##ALT=<ID=CN82,Description=\"Copy number allele: 82 copies\">"
## [183] "##ALT=<ID=CN83,Description=\"Copy number allele: 83 copies\">"
## [184] "##ALT=<ID=CN84,Description=\"Copy number allele: 84 copies\">"
## [185] "##ALT=<ID=CN85,Description=\"Copy number allele: 85 copies\">"
## [186] "##ALT=<ID=CN86,Description=\"Copy number allele: 86 copies\">"
## [187] "##ALT=<ID=CN87,Description=\"Copy number allele: 87 copies\">"
## [188] "##ALT=<ID=CN88,Description=\"Copy number allele: 88 copies\">"
## [189] "##ALT=<ID=CN89,Description=\"Copy number allele: 89 copies\">"
## [190] "##ALT=<ID=CN90,Description=\"Copy number allele: 90 copies\">"
## [191] "##ALT=<ID=CN91,Description=\"Copy number allele: 91 copies\">"
## [192] "##ALT=<ID=CN92,Description=\"Copy number allele: 92 copies\">"
## [193] "##ALT=<ID=CN93,Description=\"Copy number allele: 93 copies\">"
## [194] "##ALT=<ID=CN94,Description=\"Copy number allele: 94 copies\">"
## [195] "##ALT=<ID=CN95,Description=\"Copy number allele: 95 copies\">"
## [196] "##ALT=<ID=CN96,Description=\"Copy number allele: 96 copies\">"
## [197] "##ALT=<ID=CN97,Description=\"Copy number allele: 97 copies\">"
## [198] "##ALT=<ID=CN98,Description=\"Copy number allele: 98 copies\">"
## [199] "##ALT=<ID=CN99,Description=\"Copy number allele: 99 copies\">"
## [200] "##ALT=<ID=CN100,Description=\"Copy number allele: 100 copies\">"
## [201] "##ALT=<ID=CN101,Description=\"Copy number allele: 101 copies\">"
## [202] "##ALT=<ID=CN102,Description=\"Copy number allele: 102 copies\">"
## [203] "##ALT=<ID=CN103,Description=\"Copy number allele: 103 copies\">"
```

```
## [204] "##ALT=<ID=CN104,Description=\"Copy number allele: 104 copies\">"
## [205] "##ALT=<ID=CN105,Description=\"Copy number allele: 105 copies\">"
## [206] "##ALT=<ID=CN106,Description=\"Copy number allele: 106 copies\">"
## [207] "##ALT=<ID=CN107,Description=\"Copy number allele: 107 copies\">"
## [208] "##ALT=<ID=CN108,Description=\"Copy number allele: 108 copies\">"
## [209] "##ALT=<ID=CN109,Description=\"Copy number allele: 109 copies\">"
## [210] "##ALT=<ID=CN110,Description=\"Copy number allele: 110 copies\">"
## [211] "##ALT=<ID=CN111,Description=\"Copy number allele: 111 copies\">"
## [212] "##ALT=<ID=CN112,Description=\"Copy number allele: 112 copies\">"
## [213] "##ALT=<ID=CN113,Description=\"Copy number allele: 113 copies\">"
## [214] "##ALT=<ID=CN114,Description=\"Copy number allele: 114 copies\">"
## [215] "##ALT=<ID=CN115,Description=\"Copy number allele: 115 copies\">"
## [216] "##ALT=<ID=CN116,Description=\"Copy number allele: 116 copies\">"
## [217] "##ALT=<ID=CN117,Description=\"Copy number allele: 117 copies\">"
## [218] "##ALT=<ID=CN118,Description=\"Copy number allele: 118 copies\">"
## [219] "##ALT=<ID=CN119,Description=\"Copy number allele: 119 copies\">"
## [220] "##ALT=<ID=CN120,Description=\"Copy number allele: 120 copies\">"
## [221] "##ALT=<ID=CN121,Description=\"Copy number allele: 121 copies\">"
## [222] "##ALT=<ID=CN122,Description=\"Copy number allele: 122 copies\">"
## [223] "##ALT=<ID=CN123,Description=\"Copy number allele: 123 copies\">"
## [224] "##ALT=<ID=CN124,Description=\"Copy number allele: 124 copies\">"
## [225] "##FORMAT=<ID=GT,Number=1,Type=String,Description=\"Genotype\">"
## [226] "##INFO=<ID=CIEND,Number=2,Type=Integer,Description=\"Confidence interval around END for impre
## [227] "##INFO=<ID=CIPOS,Number=2,Type=Integer,Description=\"Confidence interval around POS for impre
## [228] "##INFO=<ID=CS,Number=1,Type=String,Description=\"Source call set.\">"
## [229] "##INFO=<ID=END,Number=1,Type=Integer,Description=\"End coordinate of this variant\">"
## [230] "##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description=\"Imprecise structural variation\">"
## [231] "##INFO=<ID=MC,Number=.,Type=String,Description=\"Merged calls.\">"
## [232] "##INFO=<ID=MEINFO,Number=4,Type=String,Description=\"Mobile element info of the form NAME,STA
## [233] "##INFO=<ID=MEND,Number=1,Type=Integer,Description=\"Mitochondrial end coordinate of inserted s
## [234] "##INFO=<ID=MLEN,Number=1,Type=Integer,Description=\"Estimated length of mitochondrial insert\
## [235] "##INFO=<ID=MSTART,Number=1,Type=Integer,Description=\"Mitochondrial start coordinate of inser
## [236] "##INFO=<ID=SVLEN,Number=.,Type=Integer,Description=\"SV length. It is only calculated for stru
## [237] "##INFO=<ID=SVTYPE,Number=1,Type=String,Description=\"Type of structural variant\">"
## [238] "##INFO=<ID=TSD,Number=1,Type=String,Description=\"Precise Target Site Duplication for bases, i
## [239] "##INFO=<ID=AC,Number=A,Type=Integer,Description=\"Total number of alternate alleles in called
## [240] "##INFO=<ID=AF,Number=A,Type=Float,Description=\"Estimated allele frequency in the range (0,1)\
## [241] "##INFO=<ID=NS,Number=1,Type=Integer,Description=\"Number of samples with data\">"
## [242] "##INFO=<ID=AN,Number=1,Type=Integer,Description=\"Total number of alleles in called genotypes\
## [243] "##INFO=<ID=EAS_AF,Number=A,Type=Float,Description=\"Allele frequency in the EAS populations ca
## [244] "##INFO=<ID=EUR_AF,Number=A,Type=Float,Description=\"Allele frequency in the EUR populations ca
## [245] "##INFO=<ID=AFR_AF,Number=A,Type=Float,Description=\"Allele frequency in the AFR populations ca
## [246] "##INFO=<ID=AMR_AF,Number=A,Type=Float,Description=\"Allele frequency in the AMR populations ca
## [247] "##INFO=<ID=SAS_AF,Number=A,Type=Float,Description=\"Allele frequency in the SAS populations ca
## [248] "##INFO=<ID=DP,Number=1,Type=Integer,Description=\"Total read depth; only low coverage data wer
## [249] "##INFO=<ID=AA,Number=1,Type=String,Description=\"Ancestral Allele. Format: AA|REF|ALT|IndelTyp
## [250] "##INFO=<ID=VT,Number=.,Type=String,Description=\"indicates what type of variant the line repre
## [251] "##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description=\"indicates whether a variant is within th
## [252] "##INFO=<ID=MULTI_ALLELIC,Number=0,Type=Flag,Description=\"indicates whether a site is multi-al
```

```
#VCF file has copy number variations as seen in metadata

#CNV information to be obtained from INFO section
info_vcf <- extract_info_tidy(vcf)
```

```r
unique(info_vcf$VT)
```

```
## [1] "SNP"       "INDEL"     "SNP,INDEL" "SV"
```

```r
#confirms presence of structural variants (SV)

# Get unique counts for each type of variant present ("SNP", "INDEL", "SNP,INDEL", "SV")
counts_vcf <- table(info_vcf$VT)
counts_vcf
```

```
##
##     INDEL       SNP SNP,INDEL        SV
##       227      6265         5         3
```

```r
# Get the count specifically for "SV"
sv_count <- counts_vcf["SV"]
sv_count
```

```
## SV
##  3
```

```r
# Alternatively confirm  number of structural variant type DEL/DUP associated to CNV in file
#Deletions and duplications with a size larger than 1,000 bases (>1 kb) are often referred to as copy-n
c <- table(info_vcf$SVTYPE)
c
```

```
##
## DEL DUP
##   1   2
```

```r
######################################################################

#g. To count variants with more than one alternative allele
# based on the VCF meta data, total number of alternate alleles in called genotypes is in line 239 unde

# Install stringr if required
#install.packages("stringr")

# Load stringr packages
library(stringr)

# Filter the 'info' data frame to create 'multi_allele_rows', which includes only rows where the 'AC' c
multi_allele_rows <- info_vcf %>%
  filter(str_count(AC, ",") > 0)

nrow(multi_allele_rows)
```

```
## [1] 33
```