Today is December 4, 2012.

```
> rm(list = ls())
> source("~/jhsph/R/reload.R")
> library("VariantAnnotation")
> library("vcf2R")

> make <- FALSE
> locus <- "8q24"
> if (make) {
     fl <- paste0("/thumper/ctsa/beaty/targeted_seq/regional/",
         locus, "-european.recode.vcf.gz")
     vcf <- readVcf(fl, "hg19")
     save(vcf, file = paste0("/home/bst/student/syounkin/R/packages/vcf2R/data/",
         locus, "-european.vcf.RData"))
 } else {
     data(list = paste0(locus, "-european.vcf"))
 }

> show(vcf)

class: CollapsedVCF
dim: 27639 602
rowData(vcf):
  GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
info(vcf):
  DataFrame with 9 columns: DQ, DA, NS, PS, DP, MQ, BA, AB, AF
info(header(vcf)):
      Number Type    Description
   DQ 1      Float   De Novo Mutation Quality
   DA 1      Integer De Novo Mutation Allele
   NS 1      Integer Number of Samples With Data
   PS 1      Float   Percentage of Samples With Data
   DP 1      Integer Total Read Depth
   MQ 1      Float   Average Map Quality
   BA 1      String  Best Alternative Allele
   AB 1      Float   Allelic Balance
   AF .      Float   Reference Allele Frequency
geno(vcf):
  SimpleList of length 12: DNGL, DNGT, DNGQ, GT, GQ, DP, DS, GL, FT, DNFT, PL, BA
geno(header(vcf)):
        Number Type    Description
   DNGL .      Integer Denovo Genotype Likelihoods
   DNGT 1      String  Genotype
   DNGQ 1      Integer Genotype Quality
   GT   1      String  Genotype
   GQ   1      Integer Genotype Quality
   DP   1      Integer Read Depth
   DS   1      Float   Dosage: Defined As the Expected Alter...
   GL   .      Integer Genotype Likelihoods
   FT   1      String  Per Sample Filter Status
   DNFT 1      String  Denovo Filter Status
   PL   .      Integer Phred-scaled Genotype Likelihood
   BA   1      String  Best Alterantive Allele
```

```
> geno.mat <- geno(vcf)$GT
> geno.mat <- t(ifelse(geno.mat == "0/0", 0L, ifelse(geno.mat ==
     "0/1", 1L, ifelse(geno.mat == "1/1", 2L, NA))))
> geno.mat <- geno.mat[, colSums(geno.mat, na.rm = TRUE) !=
     0]
> maf <- colSums(geno.mat, na.rm = TRUE)/2/nrow(geno.mat)
> maf <- ifelse(maf > 0.5, 1 - 0.5, maf)
> summary(maf)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0008306 0.0008306 0.0016610 0.0625100 0.0240900 0.5000000
```

```
> ll <- strsplit(x = colnames(geno.mat), split = ":")
> map.df <- data.frame((do.call("rbind", ll)))
> names(map.df) <- c("chr", "pos")
> maf.df <- data.frame(map.df, maf = maf)
> rownames(maf.df) <- NULL
> head(maf.df)
```

```
  chr       pos          maf
1   8 129295457 0.0033222591
2   8 129295502 0.0008305648
3   8 129295625 0.0033222591
4   8 129295789 0.0024916944
5   8 129295998 0.0191029900
6   8 129296000 0.1877076412
```

```
> id.vec <- rownames(geno.mat)
> lll <- strsplit(x = id.vec, split = "-")
> id.vec <- data.frame((do.call("rbind", lll)))[, 3]
> head(id.vec)
```

```
[1] DS10776_2 DS10776_3 DS10777_2 DS10777_3 DS10778_2 DS10778_3
602 Levels: DS10776_2 DS10776_3 DS10777_2 DS10777_3 ... DS11418_3
```

```
> colnames(geno.mat) <- NULL
> rownames(geno.mat) <- NULL
> head(geno.mat[, 1:5], 10)
```

```
      [,1] [,2] [,3] [,4] [,5]
 [1,]    0    0    0    0    0
 [2,]    0    0    0    0    0
 [3,]   NA    0    0    0    0
 [4,]   NA    0    0    0    0
 [5,]   NA    0    0    0    1
 [6,]   NA    0    0    0    0
 [7,]    0    0    0    0    1
 [8,]    0    0    0    0    0
 [9,]   NA    0    0    0    0
[10,]   NA    0    0    0    0
```

```
> geno.common.mat <- subset(geno.mat, select = (maf >=
      0.01))
> (n.snps <- ncol(geno.common.mat))
```

```
[1] 3130
```

There are 3130 SNPS with maf ≥ 0.01 in BMP4 European parents.