

Today is December 5, 2012.

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

> make <- FALSE
> locus <- "BMP4"
> 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"))
}
```

The following show method eats up a lot of RAM!

```
> show(vcf)

> 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	0.0831500	0.0234600	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	14	54382651	0.0016611296
2	14	54383433	0.0008305648
3	14	54383470	0.0141196013
4	14	54383819	0.0041528239
5	14	54383945	0.2956810631
6	14	54384030	0.0357142857

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

```
> save(geno.mat, file = paste0("/home/sgy/jhsph/targeted-sequencing/repo/vcf2R/data/",
  locus, "-european.geno.RData"))
```

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

```
> (n.snps <- ncol(geno.common.mat))
```

```
[1] 179
```

There are 179 SNPS with $\text{maf} \geq 0.01$ in BMP4 European parents.