Get GBS data for 14 genotypes

Sept. 22th, 2016

Extract GBS Data

On `farm`, extract GBS data (v2.7 on AGPv3, imputed) for 14 genotypes.

```
library(farmeR)
#### GBS genotypes
gt0 <- read.csv("data/AllZeaGBSv2.7_publicSamples_metadata20140411.csv", header=T)
gt <- subset(gt0, Project %in% c("2010 Ames Lines", "AMES Inbreds", "Ames282") )
gt$Pedigree <- as.character(gt$Pedigree)

### 14 ids
sam <- read.csv("data/SAM_cellcount.csv")
sam <- subset(sam, !is.na(Count_Cells))
id <- as.character(unique(sam$Genotype))

sub <- subset(gt, toupper(DNASample) %in% toupper(id) | Pedigree %in% toupper(id))
as.character(unique(sub$DNASample))
length(unique(sub$DNASample, sub$Pedigree))

sub <- sub[order(sub$DNASample),]
write.table(sub, "cache/cellnum_GBS_sampleid.csv", sep=",", row.names=FALSE, quote=FALSE)
###>>> Manually curated the ids

#### extract subset of the data: biallelic SNPs only and no-update
ids <- read.csv("cache/cellnum_GBS_sampleid_curated.csv")
write.table(ids[, 1], "cache/id14.txt", sep="\t", row.names=FALSE, col.names=FALSE, quote=FALSE)

tabix <- "tabix -p vcf AllZeaGBSv2.7_publicSamples_imputedV3b_agpv3_sorted.vcf.gz"
cmd <- paste("bcftools view -S /home/jolyang/Documents/Github/zmHapMap/cache/id14.txt",
             "--no-update -m2 -M2 -v snps",
             "AllZeaGBSv2.7_publicSamples_imputedV3b_agpv3_sorted.vcf.gz",
             "-Oz -o AllZeaGBSv2.7_id14_imputedV3b.vcf.gz")
set_farm_job(slurmsh = "slurm-script/bcf2plink.sh",
             shcode = c(tabix, cmd), wd = NULL, jobid = "bcftools",
             email = "yangjl0930@gmail.com", runinfo = c(TRUE, "bigmemh", "3", "23000"))
```

Filtering

```
### extract id and convert to PLINK
cmd1 <- c("mv /home/jolyang/dbcenter/AllZeaGBS/AllZeaGBSv2.7_id14_imputedV3b.vcf.gz largedata/")
cmd2 <- "cd largedata"
cmd3 <- paste("plink -vcf AllZeaGBSv2.7_id14_imputedV3b.vcf.gz",
              "--biallelic-only --snps-only --set-missing-var-ids @_# --out GBSv2.7_id14",
              "--allow-extra-chr --make-bed --freq --missing")

set_farm_job(slurmsh = "slurm-script/bcf2plink.sh",
             shcode = c(cmd1, cmd2, cmd3), wd = NULL, jobid = "maf",
             email = "yangjl0930@gmail.com", runinfo = c(TRUE, "bigmemh", "3", "23000"))
```

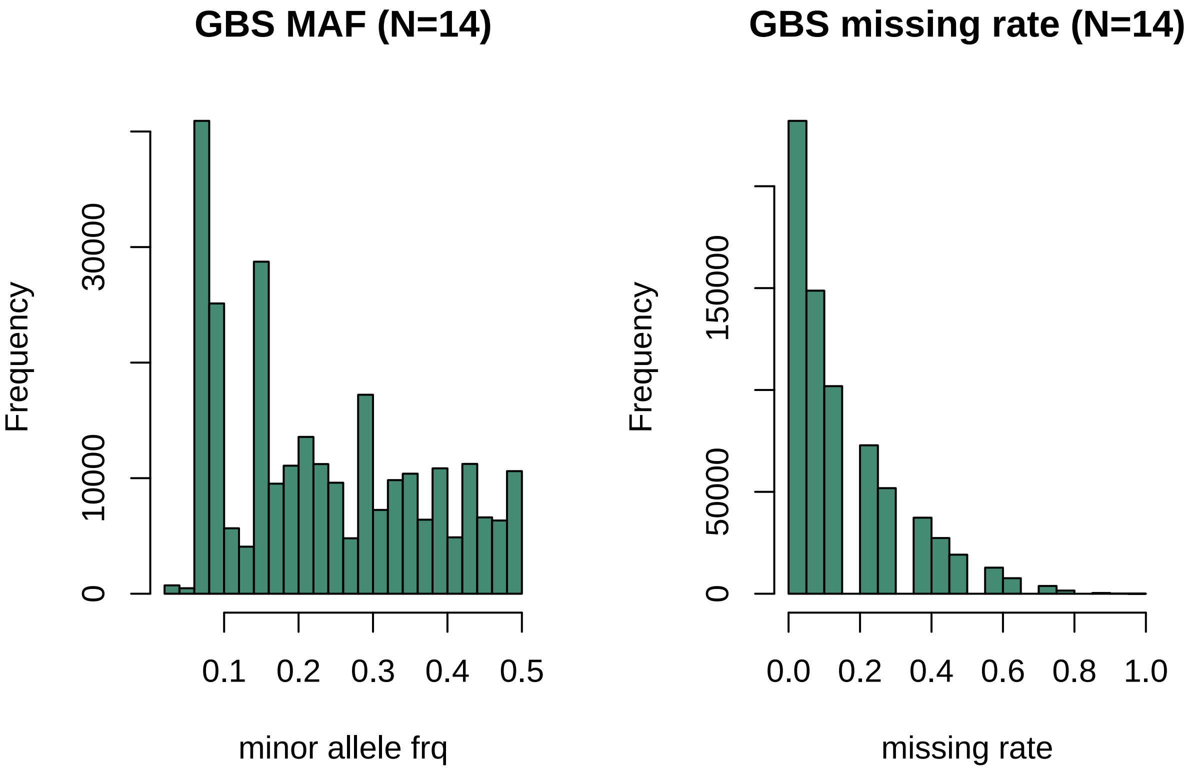
Including Plots

```
library(data.table)
frq <- fread("largedata/GBSv2.7_id14.frq")
lmiss <- fread("largedata/GBSv2.7_id14.lmiss")

d <- merge(frq[,2:5, with=FALSE], lmiss[, 2:5, with=FALSE], by="SNP")
d <- as.data.frame(d)
snps <- nrow(subset(d, MAF > 0.01 & lmiss < 0.6))
```

We kept only **717,588** biallelic SNPs. After filtering MAF > 0.01 and lmiss < 0.6 in the 14 genotypes, 349167 SNPs remain.

```
par(mfrow=c(1,2))
hist(d$MAF[d$MAF > 0], main="GBS MAF (N=14)", xlab="minor allele frq", col="#458b74")
hist(d$F_MISS, main="GBS missing rate (N=14)", xlab="missing rate", col="#458b74")
```



GEMMA for relatedness calculation

```
cmd <- "gemma -bfile GBSv2.7_id14_flt -gk 2 -o mx"
```