

Notes

Jinliang Yang

July 23, 2015

GBS data Summary

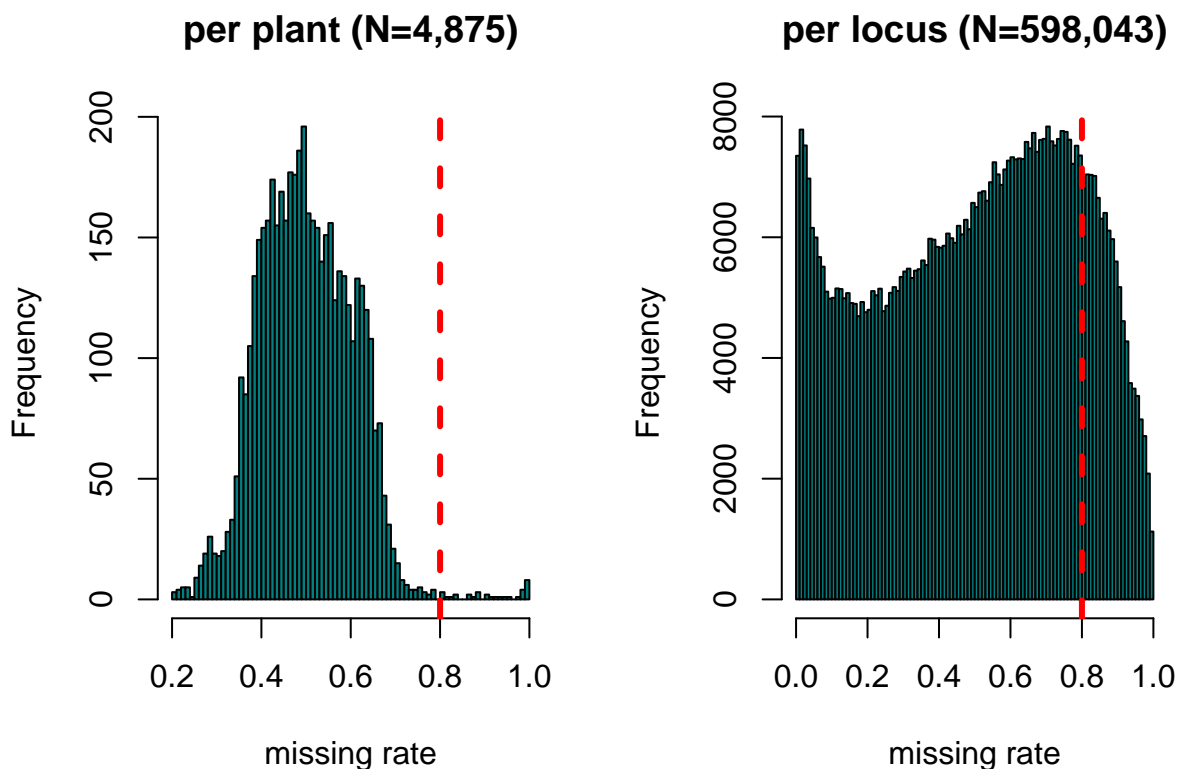
Missing rate and MAF of GBS data

Loading HDF5 format GBS raw data - loading in genotypes from HDF5 file `largedata/teo.h5` - filtering biallelic loci: Removed 357,647 non-biallelic loci. - data matrix dimension: [1:598043, 1:4875]

The missing rates were plotted as below for 598,043 SNPs of 4,875 plants (70/4,875 (1.4%) are founder lines). Note several plants have very high SNP missing rate, i.e. > 80%. Some of them even have a 100% missing rate. In addition, ~20% of them have very high (>80%) per locus missing rate.

```
# source("../profiling/2.ci_data/2.A.1_vsb_hdf5.R")
info <- read.csv("../data/teo_info.csv")
imiss <- read.csv("../data/teo_imiss.csv")

par(mfrow=c(1,2))
hist(imiss$imiss, main="per plant (N=4,875)", col="#008080", breaks=100, xlab="missing rate")
abline(v=0.8, col="red", lty=2, lwd=3)
hist(info$lmiss, main="per locus (N=598,043)", col="#008080", breaks=100, xlab="missing rate")
abline(v=0.8, col="red", lty=2, lwd=3)
```

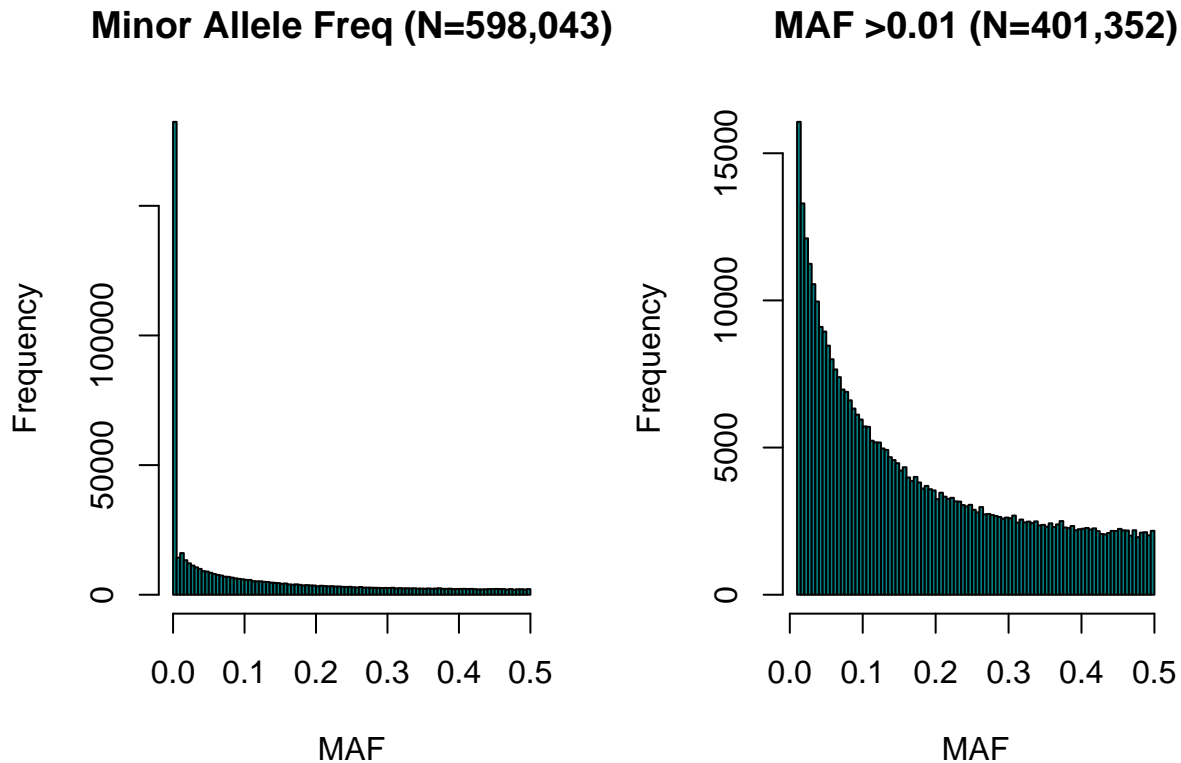


```

par(mfrow=c(1,2))
hist(info$maf, main="Minor Allele Freq (N=598,043)", col="#008080", breaks=100, xlab="MAF")
abline(v=0.8, col="red", lty=2, lwd=3)

hist(subset(info, maf>0.01)$maf, main="MAF >0.01 (N=401,352)", col="#008080", breaks=100, xlab="MAF")
abline(v=0.8, col="red", lty=2, lwd=3)

```



See above the MAF plots. Note 196665 SNPs have the MAF < 0.01.

Parentage Infomation

```

# source("../profiling/2.ci_data/2.B.1_cj_parentage.R")
pinfo <- read.table("../data/parentage_sum.txt", header=TRUE)
#dim(pinfo) 68,5
subset(pinfo, !is.na(WGS))

```

##	sid	founder	nselfer	nox	WGS
## 5	PC_I11_ID2	PC_I11_ID2_mrg:250276264	43	126	yes
## 7	PC_I50_ID2	PC_I50_ID2_mrg:250276265	55	101	yes
## 10	PC_I55_ID2	PC_I55_ID2_mrg:250276267	81	94	yes
## 12	PC_I58_ID2	PC_I58_ID2_mrg:250276268	30	105	yes
## 16	PC_J07_ID2	PC_J07_ID2_mrg:250276269	40	92	yes
## 22	PC_J14_ID2	PC_J14_ID2_mrg:250276270	60	63	yes
## 23	PC_J48_ID2	PC_J48_ID2_mrg:250276262	46	101	yes
## 29	PC_K55_ID2	PC_K55_ID2_mrg:250276291	47	135	yes
## 31	PC_L06_ID2	PC_L06_ID2_mrg:250276271	24	98	yes

```
## 35 PC_L12_ID2 PC_L12_ID2_mrg:250276272      61  57 yes
## 38 PC_L48_ID2 PC_L48_ID2_mrg:250276273      48  78 yes
## 44 PC_N03_ID2 PC_N03_ID2_mrg:250276274      14 107 yes
## 47 PC_N07_ID2 PC_N07_ID2_mrg:250276276      38  95 yes
## 50 PC_N10_ID2 PC_N10_ID2_mrg:250276277      45  47 yes
## 54 PC_N14_ID2 PC_N14_ID2_mrg:250276278      58  85 yes
## 58 PC_N57_ID2 PC_N57_ID2_mrg:250276279      45 116 yes
## 60 PC_N58_ID2 PC_N58_ID2_mrg:250276280      46 141 yes
## 63 PC_008_ID2 PC_008_ID2_mrg:250276281      62  97 yes
## 66 PC_051_ID2 PC_051_ID2_mrg:250276282      97  13 yes
```

```
subset(pinfo, nox < 30)
```

```
##          sid          founder nselfer nox  WGS
## 4  PC_I11_ID1  PC_I11_ID1_1:250276201      NA  7 <NA>
## 9  PC_I53_ID1  PC_I53_ID1_1:250276206      NA  4 <NA>
## 13 PC_J01_ID1  PC_J01_ID1_1:250276209      NA  1 <NA>
## 15 PC_J07_ID1  PC_J07_ID1_1:250276211      NA 28 <NA>
## 21 PC_J14_ID1  PC_J14_ID1_1:250276217      NA 16 <NA>
## 28 PC_K55_ID1  PC_K55_ID1_1:250276224      NA  4 <NA>
## 33 PC_L10_ID1  PC_L10_ID1_1:250276228      NA  6 <NA>
## 37 PC_L48_ID1  PC_L48_ID1_1:250276231      NA 19 <NA>
## 49 PC_N10_ID1  PC_N10_ID1_1:250276243      NA  2 <NA>
## 53 PC_N14_ID1  PC_N14_ID1_1:250276247      NA  1 <NA>
## 56 PC_N56_ID1  PC_N56_ID1_1:250276250      NA 15 <NA>
## 57 PC_N57_ID1  PC_N57_ID1_1:250276251      NA  5 <NA>
## 61 PC_N60_ID1  PC_N60_ID1_1:250276255      NA  3 <NA>
## 62 PC_008_ID1  PC_008_ID1_1:250276256      NA 12 <NA>
## 64 PC_010_ID1  PC_010_ID1_1:250276258      NA 15 <NA>
## 65 PC_051_ID1  PC_051_ID1_1:250276259      NA  5 <NA>
## 66 PC_051_ID2 PC_051_ID2_mrg:250276282      97 13 yes
## 67 PC_059_ID1  PC_059_ID1_1:250276261      NA 29 <NA>
```

We calculated the number of selfers and the number of outcrossers associated with the founder lines. In total, 68 unique founder lines were involved in the crosses. Among them, 49 founder lines had selfing families; and the bottom ten smallest family had their family size of 3, 7, 12, 14, 15, 16, 18, 20, 22, 24, respectively.

WGS of 19 *Teosintes*

```
# source("../profiling/2.ci_data/2.B.2_cj_wgsdata.R")

source("../profiling/2.ci_data/2.B.3_cj_comp.R")

### load WGS data of 19 teosintes and recoded to `0, 1, 2` format and `3` indicates missing.
wgs <- recode()
###>>> WGS [ 396818 ] | GBS [ 597607 ] | shared [ 315514 ]
###>>> consistent SNP calling [ 301249 ]

### load GBS data of 19 teosintes
gbs <- gbsgeno(wgs)
```

```

###>>> GBS of [ 598043 ] SNPs and [ 19 ] plants
###>>> Common SNPs [ 301249 ]

### estimate the GBS SNP calling error rate
res <- comp_alleles(wgs, gbs)
###>>> Heterozygote error rate [ 49.1 ] and Homozygote error rate [ 1.7 ]
###>>> het err=[ 494582 ]; het tot=[ 1008196 ]; hom err=[ 45239 ]; hom err=[ 2714395 ]

### calculate missing rate and MAF for 19 teosintes
maf_missing(wgs, gbs)
###>>> Data write to: [ cache/teo_gbs_wgs.RData]

```

From this file `/group/jrigrp4/phasing/cj_teosinte/genotypes_teosinte_19_noScaffolds_or_organelles.geno`, a total number of 96,908,505 SNPs were called for 19 *Teosinte* lines, of which 315,514 matched with GBS SNPs according to their v2 coordinates.

After checking `ref` and `alt` SNP calls, 301,249 sites are consistent between two datasets (concordant rate 95.5%). We then recoded the SNPs to 0, 1, 2 format, where the numbers indicated the copies of alternate alleles.

The heterozygote error rate for GBS was estimated to be 49.1%; and the homozygote error rate was estimated to be 1.7%.

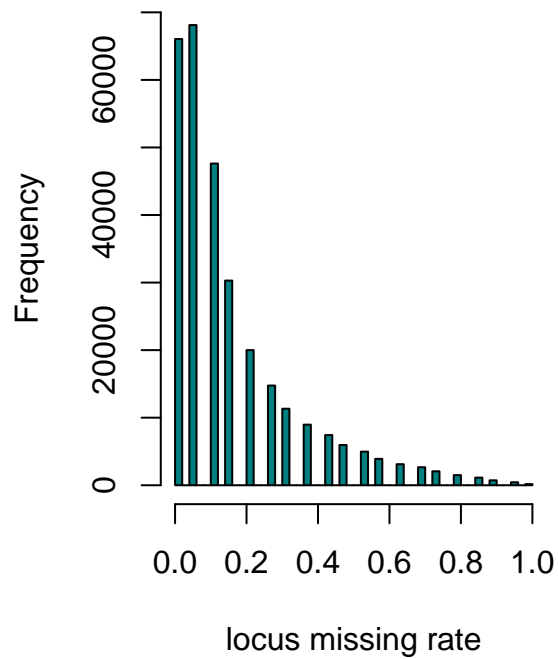
Finally, we calculate the missing rate and MAF for 19 teosintes of the two datasets, respectively. The results were stored in `cache/teo_gbs_wgs.RData`. In the following code chunk, we will plot the results.

```

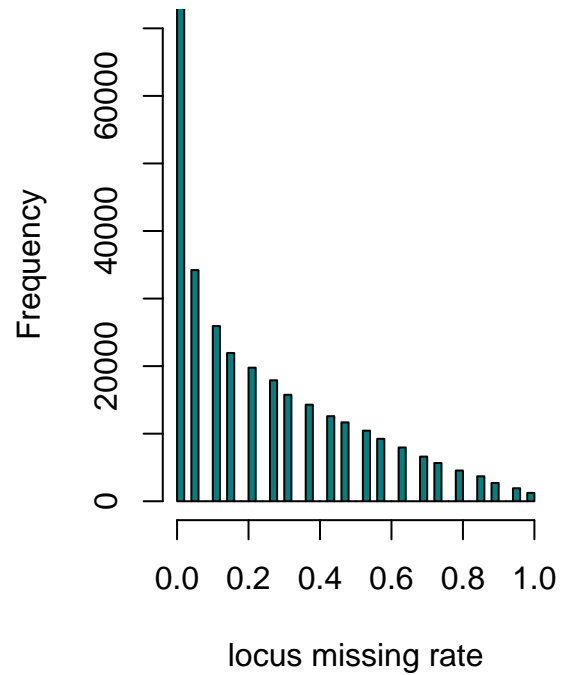
ob <- load("../cache/teo_gbs_wgs.RData")
par(mfrow=c(1,2))
hist(lmiss1, main="WGS (N=301,249)", ylim=c(0, 70000), col="#008080",
     breaks=50, xlab="locus missing rate")
#abline(v=0.8, col="red", lty=2, lwd=3)
hist(lmiss2, main="GBS (N=301,249)", ylim=c(0, 70000), col="#008080",
     breaks=50, xlab="locus missing rate")

```

WGS (N=301,249)

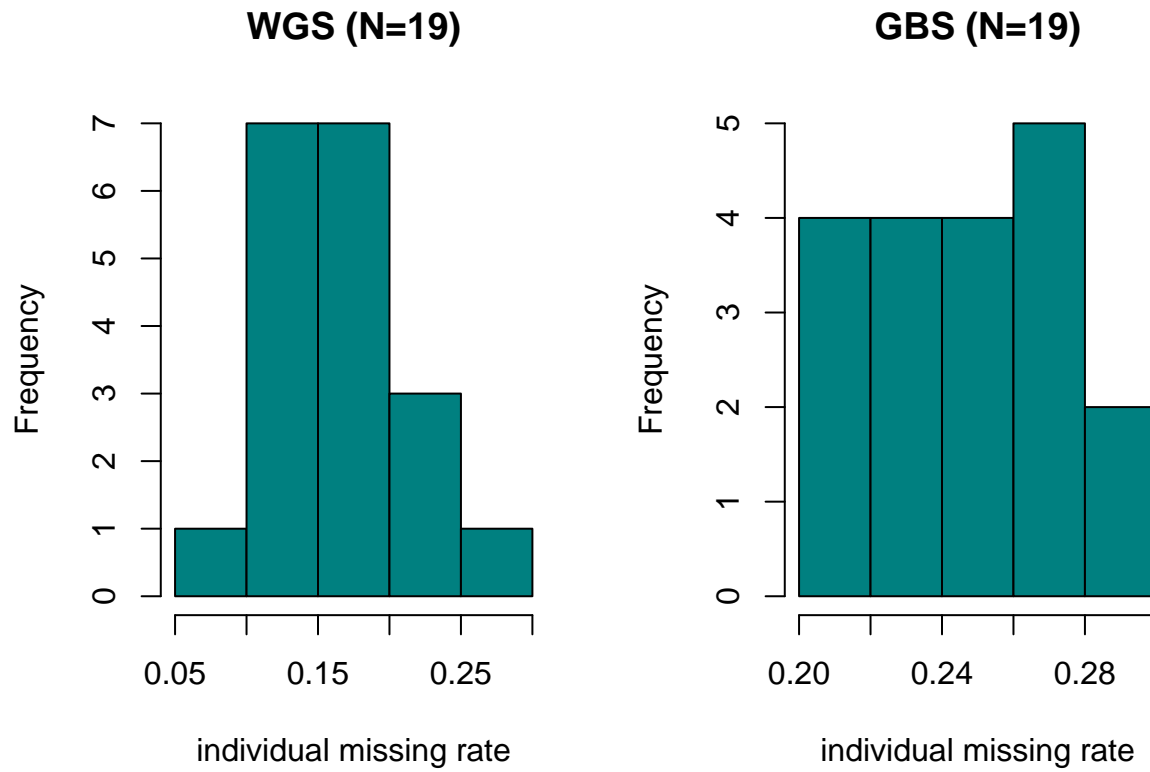


GBS (N=301,249)



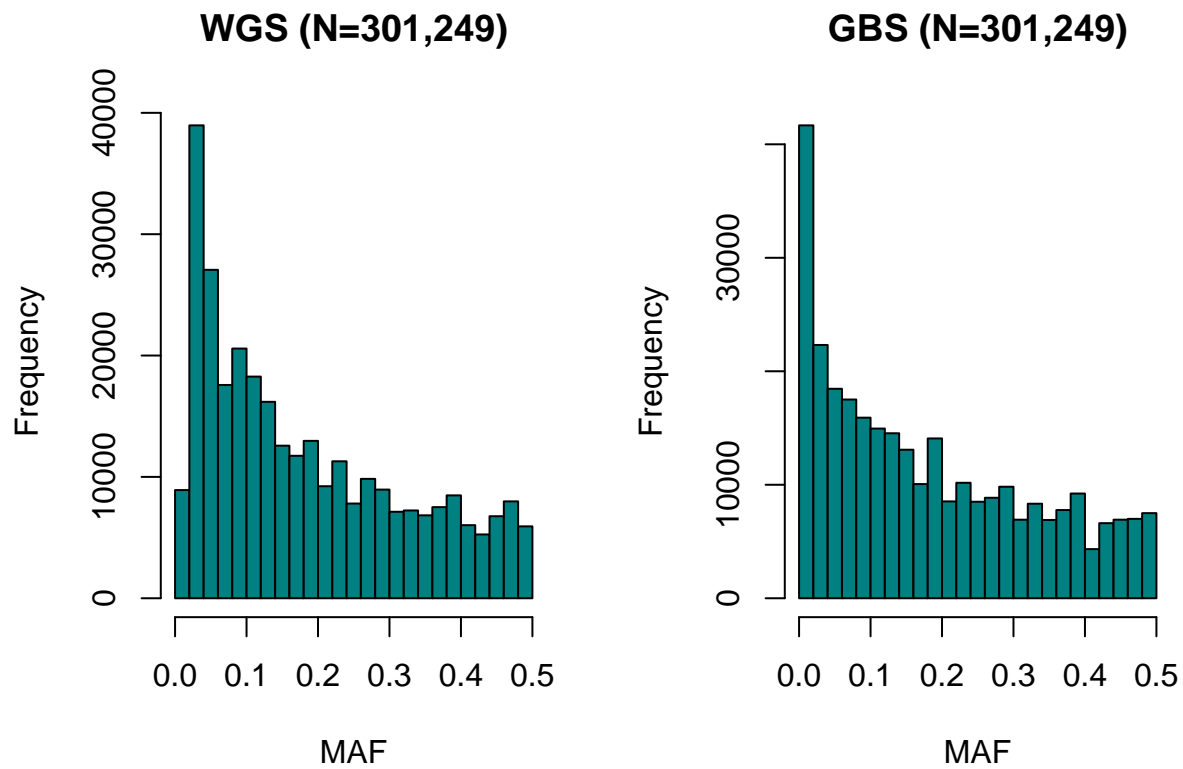
```
#abline(v=0.8, col="red", lty=2, lwd=3)

par(mfrow=c(1,2))
hist(imiss1, main="WGS (N=19)", col="#008080", xlab="individual missing rate")
#abline(v=0.8, col="red", lty=2, lwd=3)
hist(imiss2, main="GBS (N=19)", col="#008080", xlab="individual missing rate")
```



```
#abline(v=0.8, col="red", lty=2, lwd=3)

par(mfrow=c(1,2))
hist(maf1, main="WGS (N=301,249)", col="#008080", xlab="MAF")
#abline(v=0.8, col="red", lty=2, lwd=3)
hist(maf2, main="GBS (N=301,249)", col="#008080", xlab="MAF")
```



```
#abline(v=0.8, col="red", lty=2, lwd=3)
```

Comparing GBS vs. WGS
