R Assignment

Trevor Nolan

February 25, 2017

Load data

```
library(ProjectTemplate)
load.project()

fang_et_al_genotypes <- read.delim("./data/fang_et_al_genotypes.txt", stringsAsFactors=FALSE)
snp_position <- read.table("./data/snp_position.txt", stringsAsFactors=FALSE, sep = "\t", header = TRUE)
#Get cols needed for analysis: SNP id (first column), chromosome location (third column), nucleotide lo
snp_position_sub <- select(snp_position, SNP_ID, Chromosome, Position)
```

Data Inspection

`fang_et_al_genotypes.txt`:

- This file has 986 columns and 2782 rows.
- The file is 11.05 MB in size.
- There are 16 Groups in the `Group` column of this file.

`snp_position.txt`:

- This file has 15 columns and 983 rows.
- The file is 82.76 KB in size.
- There are 339 candidate and 644 random SNPs in the file.

Data Processing

Select Maize or Teosinte Groups based on `Group` column

```
#select maize or teosinte groups
maize_groups <- c("ZMMIL" , "ZMMLR", "ZMMMR")
teosinte_groups <- c("ZMPBA","ZMPIL","ZMPJA")
maize <- fang_et_al_genotypes[fang_et_al_genotypes$Group %in% maize_groups, ]
teosinte <- fang_et_al_genotypes[fang_et_al_genotypes$Group %in% teosinte_groups, ]
```

Join the files and then write the .csv files for maize

- **`data/maize/asc/`** contains the files for maize seperated by chromosome in ascending order of position with missing data encoded as "?"
- **`data/maize/dsc/`** contains the files for maize seperated by chromosome in descending order with missing data encoded as "-"

```
#transpose the dfs
transposed <- t(maize)
transposed <- transposed[c(-1,-2,-3),]
#convert to data frame
transposed_df <- as.data.frame(transposed, stringsAsFactors = FALSE)
```

```
#add a column from the row names called "SNP_ID"
transposed_df <- rownames_to_column(transposed_df, "SNP_ID")

#join to SNP position file using SNP_ID column as a key
joined <- left_join(snp_position_sub, transposed_df, by = "SNP_ID")

#arrange by increasing position values
joined <- joined[mixedorder(joined$Position), ]

#for loop to write the csv files
data_in <- joined
data_out <- split(data_in,data_in[[2]])

chn <- unlist(lapply(data_out,"[",1,2))

#need to set dir that you want files written in#
wd <- getwd()
#create directories for the output files
species <- file.path(wd,"data/maize")
asc <- file.path(wd,"data/maize/asc")
dsc <- file.path(wd,"data/maize/dsc")
dir.create(species)
dir.create(asc)
dir.create(dsc)
#Changes working directory
setwd(asc)
#for loop to write the .csv files
for(i in seq_along(chn)) write.table(data_out[[i]],file=paste(chn[i],"csv",sep="."), row.names = FALSE,

setwd(dsc)

#for decreasing

#arrange the data in decreasing order according to the postion column
joined_desc <- joined[mixedorder(joined$Position, decreasing = TRUE), ]
#substitue missing values as "-" for decreasing
joined_desc[joined_desc=="?/?"] <-"-/-"

data_in <- joined_desc
data_out <- split(data_in,data_in[[2]])

chn <- unlist(lapply(data_out,"[",1,2))
for(i in seq_along(chn)) write.table(data_out[[i]],file=paste(chn[i],"csv",sep="."), row.names = FALSE,
```

Join the files and then write the .csv files for teosinte

- **data/teosinte/asc/** contains the files for teosinte seperated by chromosome in ascending order of position with missing data encoded as "?"
- **data/teosinte/dsc/** contains the files for teosinte seperated by chromosome in descending order with missing data encoded as "-"

```
#transpose the dfs
transposed <- t(teosinte)
transposed <- transposed[c(-1,-2,-3),]
transposed_df <- as.data.frame(transposed, stringsAsFactors = FALSE)
transposed_df <- rownames_to_column(transposed_df, "SNP_ID")

#join to SNP position file using SNP_ID column as a key
joined <- left_join(snp_position_sub, transposed_df, by = "SNP_ID")

#arrange by increasing position values
joined <- joined[mixedorder(joined$Position), ]

#for loop to write the csv files
data_in <- joined
data_out <- split(data_in,data_in[[2]])

chn <- unlist(lapply(data_out,"[",1,2))
#need to set dir that you want files written in#
wd <- getwd()
species <- file.path(wd,"data/teosinte")
asc <- file.path(wd,"data/teosinte/asc")
dsc <- file.path(wd,"data/teosinte/dsc")
dir.create(species)
dir.create(asc)
dir.create(dsc)
#Changes working directory
setwd(asc)

for(i in seq_along(chn)) write.table(data_out[[i]],file=paste(chn[i],"csv",sep="."), row.names = FALSE,

setwd(dsc)

#for decreasing
# arrange by decreasing position values
joined_desc <- joined[mixedorder(joined$Position, decreasing = TRUE), ]
#substitue missing values as "-" for decreasing
joined_desc[joined_desc=="?/?"] <-"-/-"

data_in <- joined_desc
data_out <- split(data_in,data_in[[2]])

chn <- unlist(lapply(data_out,"[",1,2))
for(i in seq_along(chn)) write.table(data_out[[i]],file=paste(chn[i],"csv",sep="."), row.names = FALSE,
```

Part II Data Visualization

Process data for Graphs

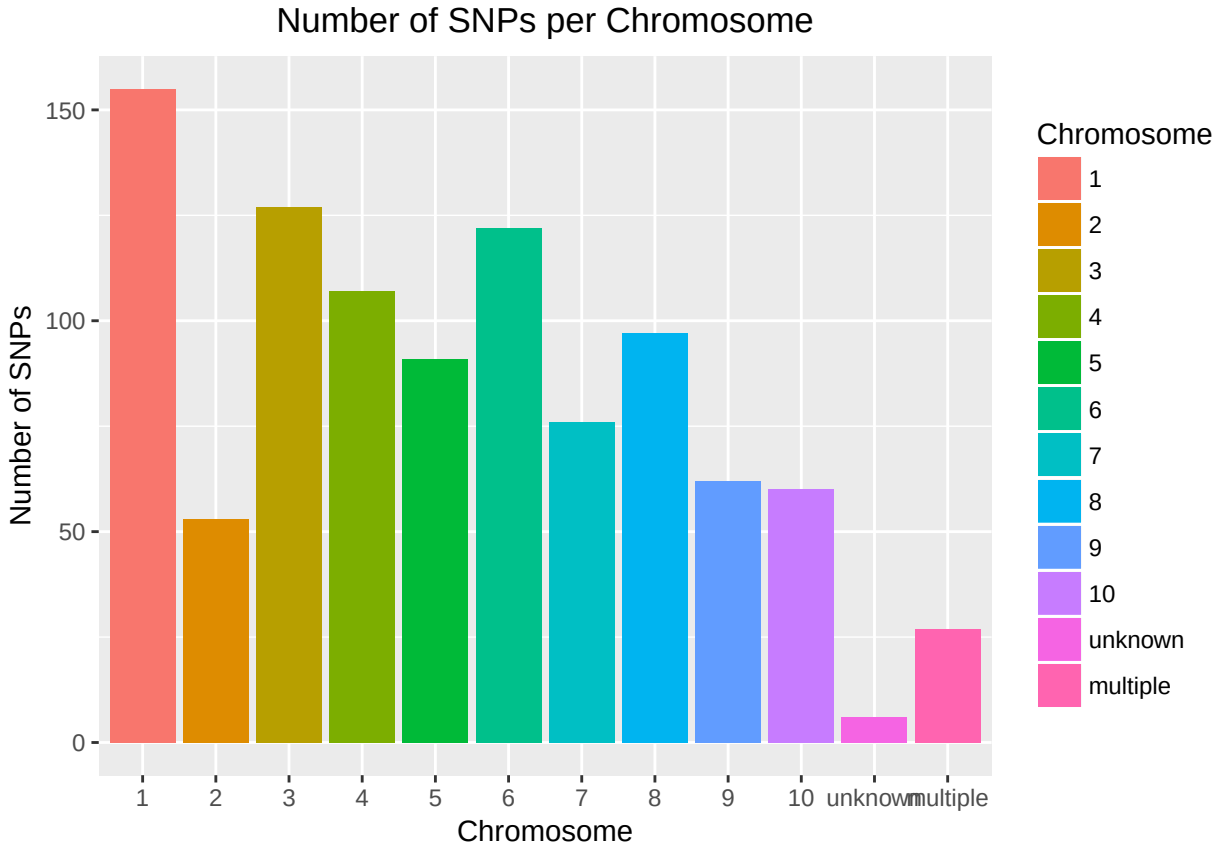
```
#make the data tidy
fang_gather <- gather(fang_et_al_genotypes, "SNP_ID", "SNP", 4:986)
#join with SNP position
```

```
fang_IDs_long <- left_join(fang_gather, snp_position, by = "SNP_ID")
```

Plot number of SNPs by Chromosome

- since all the SNPs seem to have at least some variation across the dataset (checked with the code below), I just plotted the total number of unique SNPs on each chromosome.

```
#all of the SNPs appear to have some variants because all 983 have more than one SNP type
SNPs_only <- group_by(fang_IDs_long, SNP_ID) %>% filter(SNP!="?/?") %>% summarize(SNPs_uniq = length(uni
snp_position_sub_plot <- snp_position_sub
snp_position_sub_plot$Chromosome <- as.factor(snp_position_sub_plot$Chromosome)
levels(snp_position_sub_plot$Chromosome) <- c(1:10, "unknown", "multiple")
ggplot(snp_position_sub_plot, aes(x=Chromosome, fill=Chromosome)) + geom_bar() + ylab("Number of SNPs")
```

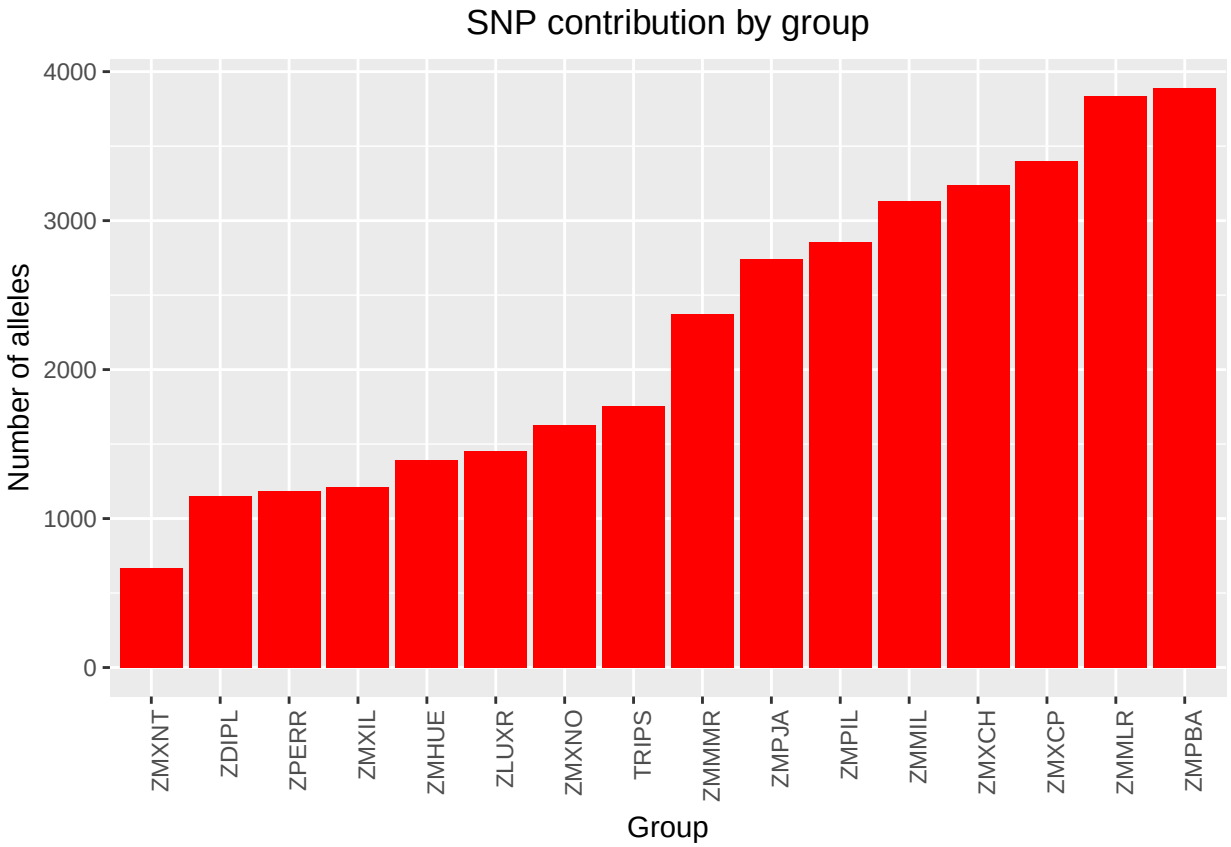


Contribution of each group to SNPs in the dataset

- The groups ZMMLR (a maize group) and ZMPBA (a teosinte group) contribute most to the SNPs in our dataset. See plot below that shows that total number of alleles (across all SNPs) for each group.

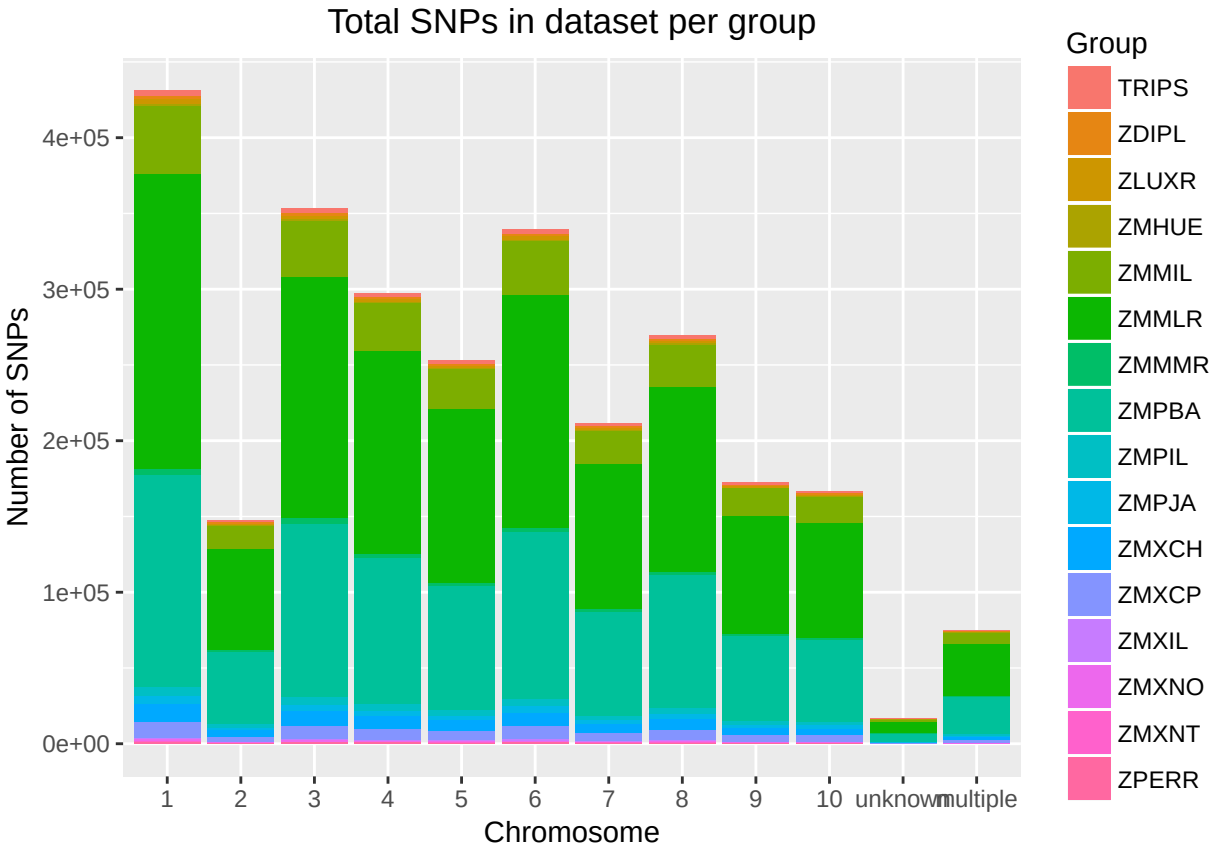
```
#calculate number of unique genotypes for each group at each SNP
SNP_summary <- group_by(fang_IDs_long, Group, SNP_ID) %>% summarize(SNPs_uniq = length(unique(SNP))) %>%
#summarize total number of alleles per group
Group_SNP_summary <- group_by(SNP_summary, Group) %>% summarize(Number_SNPs=sum(SNPs_uniq))
#change factor level for group according to number of alleles
Group_SNP_summary$Group <- factor(Group_SNP_summary$Group, levels = Group_SNP_summary$Group[order(Group_
```

```
#plot
ggplot(Group_SNP_summary, aes(x=Group, y=Number_SNPs)) + geom_bar(stat = "identity", fill="red") + theme
```



- I also plotted the total number of SNPs in the dataset (without summarizing) which appears to show a similar trend - perhaps those groups that were sampled most often contribute the most diversity to the dataset.

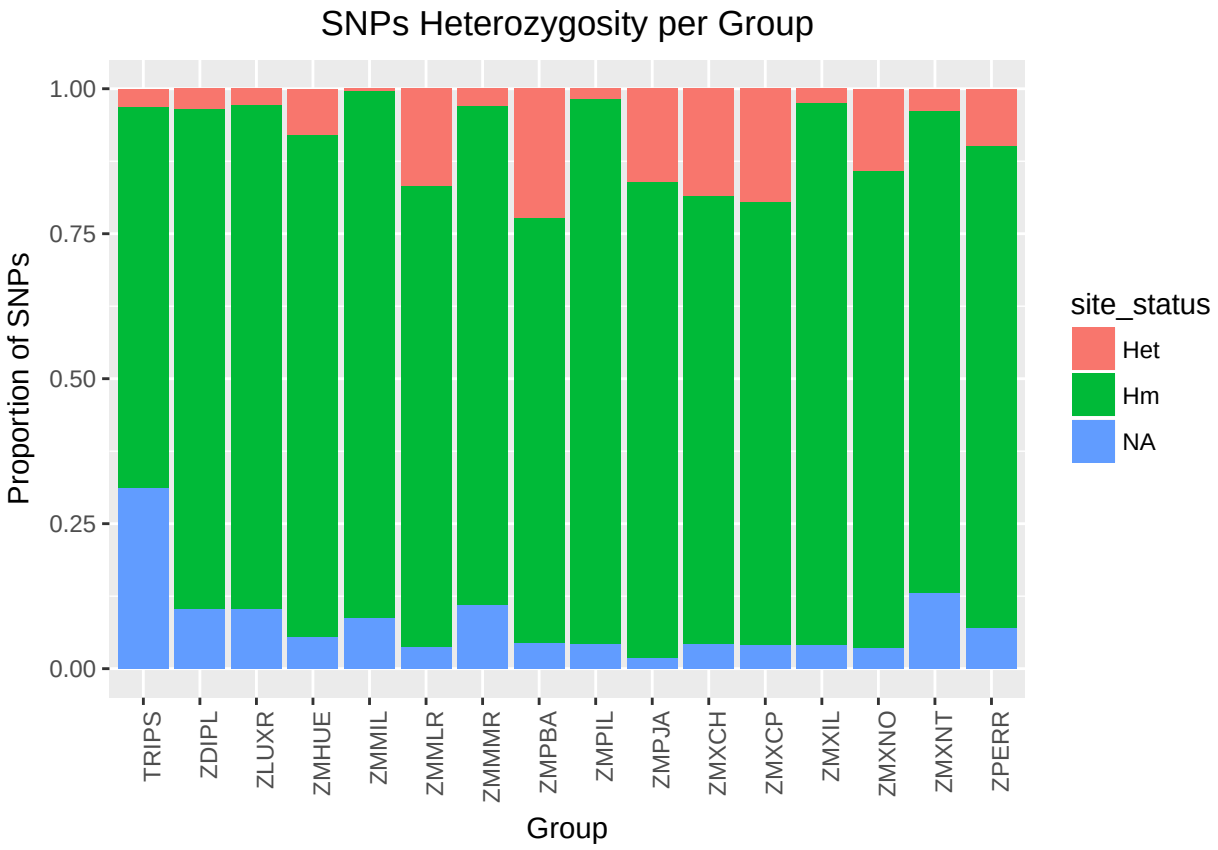
```
fang_IDs_long$Chromosome <- as.factor(fang_IDs_long$Chromosome)
levels(fang_IDs_long$Chromosome) <- c(1:10, "unknown", "multiple")
ggplot(fang_IDs_long, aes(x=Chromosome, fill=Group)) + geom_bar() + ylab("Number of SNPs") + ggtitle  (
```



Missing data and amount of heterozygosity

- `fang_IDs_long$site_status`: change missing data to "NA", heterozygous SNPs to "Het", and homozygous SNPs to "Hm"

```
fang_IDs_long$site_status <- fang_IDs_long$SNP
fang_IDs_long$site_status[fang_IDs_long$SNP=="?/?"] <-"NA"
hm <- c("A/A", "C/C", "G/G", "T/T")
fang_IDs_long$site_status[fang_IDs_long$SNP %in% hm] <-"Hm"
fang_IDs_long$site_status[fang_IDs_long$site_status!="Hm" & fang_IDs_long$site_status!="NA"] <-"Het"

#sort by group and species ID
fang_IDs_long_sorted <- arrange(fang_IDs_long, Group, JG_OTU)

#plot graph showing proportion of homozygous and heterozygous sites as well as missing data in each spe

#fist generate summaries of data

fang_summary <- group_by(fang_IDs_long_sorted, JG_OTU, Group, site_status) %>% summarise(site_n = lengt
fang_summary <- arrange(fang_summary, Group, JG_OTU, site_status)

#plot with normalized height of bars
ggplot(fang_summary, aes(x=Group, y=site_n, fill=site_status)) + geom_bar(stat = "identity", position =
```

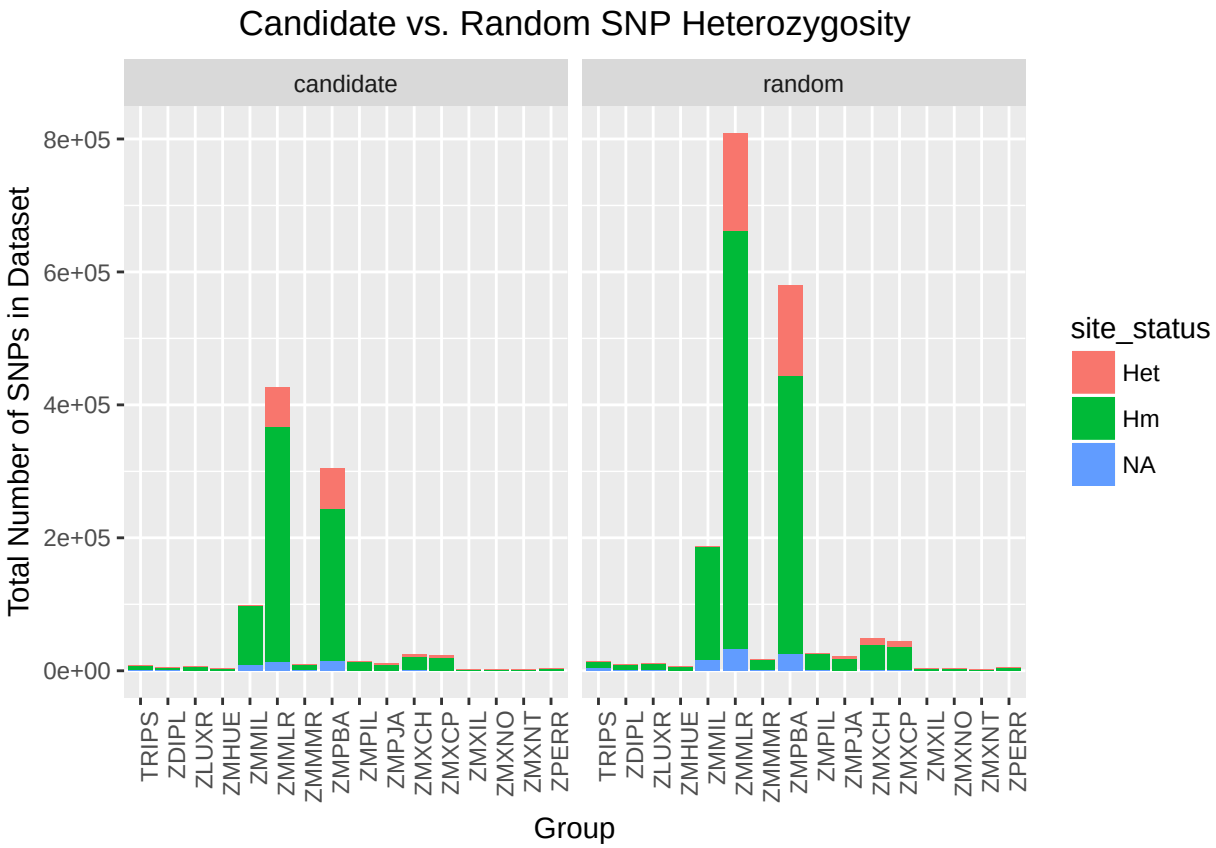


My Own Visualization

Look at differences in Heterozygosity for Candidate vs. Random SNPs from the `snp_position.txt` file

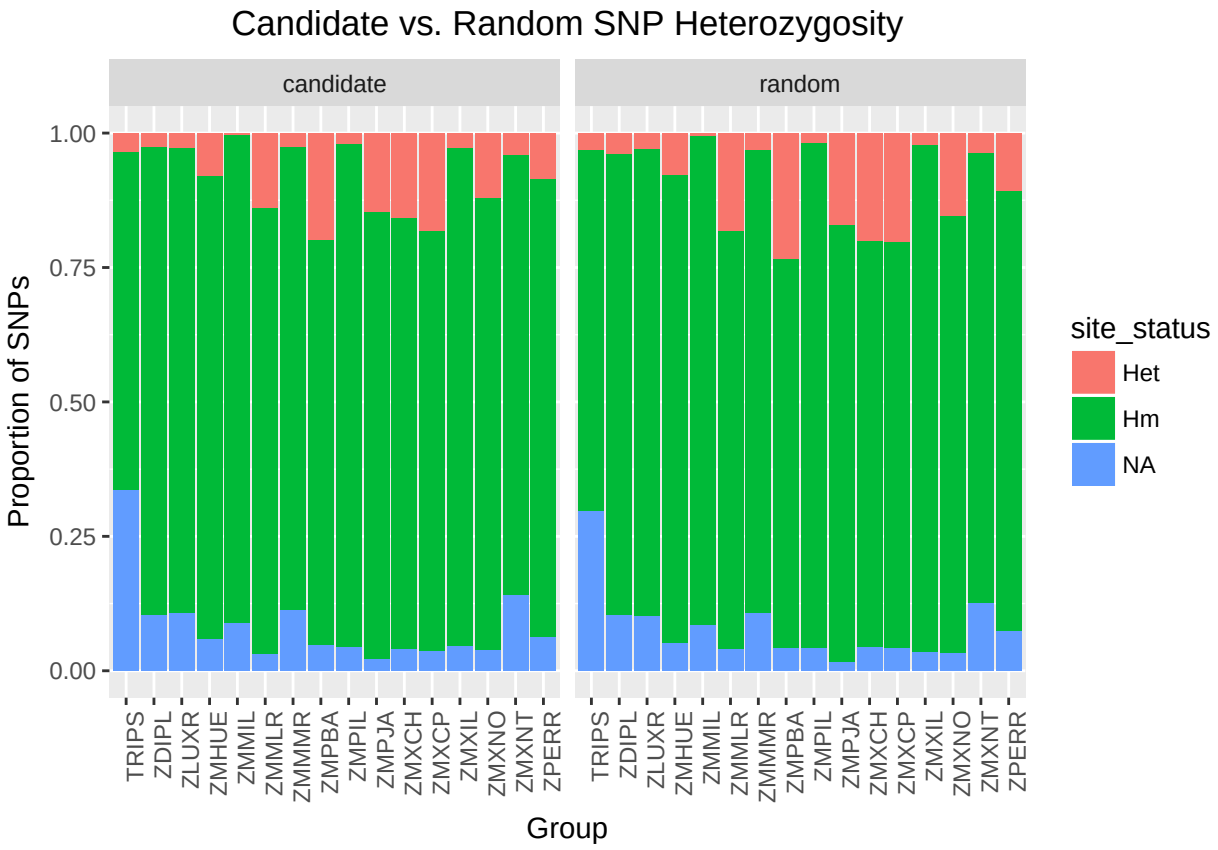
- There seem to be more random that candidate SNPs, but there are also more markers for random than candidate.
- There are 339 candidate and 644 random SNPs in the file.

```
#spread data
candidate_summary <- group_by(fang_IDs_long_sorted, Group, candidate.random, site_status) %>% summarise

ggplot(candidate_summary, aes(x=Group, y=candidate.random_n, fill=site_status, position="dodge")) + geo
```



- To account for this I generated a second plot with the bars normalized by groups and faceted by candidate or random. Now the heterozygousity looks fairly similar in cadidate vs. random SNPs.

```
ggplot(candidate_summary, aes(x=Group, y=candidate.random_n, fill=site_status)) + geom_bar(stat="identi
```


Candidate vs. Random SNP Heterozygosity
candidate
random
Proportion of SNPs
1.00
0.75
0.50
0.25
0.00
TRIPS
ZDIPL
ZLUXR
ZMHUE
ZMMIL
ZMMLR
ZMMMR
ZMPBA
ZMPIL
ZMPJA
ZMXCH
ZMXCP
ZMXIL
ZMXNO
ZMXNT
ZPERR
Group
site_status
Het
Hm
NA