UNIX Assignment

Data Inspection

For data inspecton we can use several UNIX command such as cat, less, head, tail, wc etc.

Here are the codes used to inspect the fang_et_al_genotypes files:

```
1.  du -h fang_et_al_genotypes.txt
```

```
2. cat fang_et_al_genotypes.txt | wc -l
```

```
3. tail -n +3 fang_et_al_genotypes.txt | awk -F "\t" '{print NF; exit}'
```

```
4. head -n 10 fang_et_al_genotypes.txt
```

```
5. tail -n 10 fang_et_al_genotypes.txt
```

Attributes of fang_et_al_genotypes

By inspecting this file I learned that

- The size of the file is 6.1M
- There are 2783 rows, 2744038 words, and 11051939 characters.
- The total column of the file is 986.
- head command showed the first 10 line of the file.
- tail command showed the last 10 line of the file.

Here are the codes used to inspect the fang_et_al_genotypes and snp_position.txt files:

```
1.  du -h snp_position.txt
```

```
2. cat fang_et_al_genotypes.txt | wc -l
```

```
3. tail -n +3 snp_position.txt | awk -F "\t" '{print NF; exit}'
```

```
4. head -n 10 snp_position.txt
```

```
5. tail -n 10 snp_position.txt
```

Attributes of snp_position.txt

By inspecting this file I learned that

- The size of the file is 38K
- There are 984 rows, 13198 words, and 82763 characters.
- The total column of the file is 15.
- head command showed the first 10 line of the file.
- tail command showed the last 10 line of the file.

Data Processing

For maize genotypes:

1. Extract data from snp_position.txt

```
cut -f 1,3,4 snp_position.txt > snp_position_3Extracts.txt
```

2. Delet the header of the SNP files and sorting them:

```
grep -v "SNP_ID" snp_position_cut.txt | sort -k1,1 snp_position_cut.txt >
snp_position_cut_sorted.txt
```

3. Extract the maize genotypes from the fang_et_al_genotypes.txt:

```
grep -E -w "(Sample_ID|ZMMIL|ZMMLR|ZMMMR)" fang_et_al_genotypes.txt >
maize_3Genotypes.txt
```

4. Transposed the maize_3Genotypes:

```
awk -f transpose.awk maize_3Genotypes.txt >transposed_maize_3genotypes.txt
```

5.Delet the header of the transposed_maize_3genotypes files and sorting them:

```
 grep -v "Sample_ID" transposed_maize_3genotypes.txt | sort -k1,1 >
transposed_maize_3genotypes_sorted.txt
```

6.Joining

```
join -1 1 -2 1 -t $'\t' snp_position_cut_sorted.txt
transposed_maize_3genotypes_sorted.txt > maize-SNP_joined.txt
```

7. 10 files (1 for each chromosome) with SNPs ordered based on increasing position values and with missing data encoded by this symbol: ?

```
sed 's/unknown/?/g' maize-SNP_joined2.txt | sort -k3,3n > maize_q_sorted.txt
```

```
 awk '$2~ /1/' maize_q_sorted.txt > maize_increasing_chr1.txt
 awk '$2~ /2/' maize_q_sorted.txt > maize_increasing_chr2.txt
 awk '$2~ /3/' maize_q_sorted.txt > maize_increasing_chr3.txt
 awk '$2~ /4/' maize_q_sorted.txt > maize_increasing_chr4.txt
 awk '$2~ /5/' maize_q_sorted.txt > maize_increasing_chr5.txt
 awk '$2~ /6/' maize_q_sorted.txt > maize_increasing_chr6.txt
 awk '$2~ /7/' maize_q_sorted.txt > maize_increasing_chr7.txt
 awk '$2~ /8/' maize_q_sorted.txt > maize_increasing_chr8.txt
 awk '$2~ /9/' maize_q_sorted.txt > maize_increasing_chr9.txt
 awk '$2~ /10/' maize_q_sorted.txt > maize_increasing_chr10.txt
```

8. 10 files (1 for each chromosome) with SNPs ordered based on decreasing position values and with missing data encoded by this symbol: -

```
sed 's/unknown/-/g' maize-SNP_joined2.txt | sort -k 3 -r -n > maize_-
symbol_sorted.txt
```

```
awk '$2~ /1/' maize_-symbol_sorted.txt > maize_decreasing_chr1.txt
awk '$2~ /2/' maize_-symbol_sorted.txt > maize_decreasing_chr2.txt
awk '$2~ /3/' maize_-symbol_sorted.txt > maize_decreasing_chr3.txt
```

```
awk '$2~ /4/' maize_-symbol_sorted.txt > maize_decreasing_chr4.txt
awk '$2~ /5/' maize_-symbol_sorted.txt > maize_decreasing_chr5.txt
awk '$2~ /6/' maize_-symbol_sorted.txt > maize_decreasing_chr6.txt
awk '$2~ /7/' maize_-symbol_sorted.txt > maize_decreasing_chr7.txt
awk '$2~ /8/' maize_-symbol_sorted.txt > maize_decreasing_chr8.txt
awk '$2~ /9/' maize_-symbol_sorted.txt > maize_decreasing_chr9.txt
awk '$2~ /10/' maize_-symbol_sorted.txt > maize_decreasing_chr10.txt
```

9. SNPs with unknown positions in the genome

```
grep -w "unknown" maize-SNP_joined2.txt > maize_unknown.txt
```

10. SNPs with multiple positions in the genome

```
grep -w "multiple" maize-SNP_joined2.txt > maize_multiple.txt
```

For teosinte genotypes:

1. Extract data from snp_position.txt

```
cut -f 1,3,4 snp_position.txt > snp_position_3Extracts.txt
```

2. Delet the header of the SNP files and sorting them:

```
grep -v "SNP_ID" snp_position_cut.txt | sort -k1,1 snp_position_cut.txt >
snp_position_cut_sorted.txt
```

3. Extract the teosinte genotypes from the fang_et_al_genotypes.txt:

```
grep -E -w "(Sample_ID|ZMPBA|ZMPIL|ZMPJA)" fang_et_al_genotypes.txt >
teosinte_3genotypes.txt
```

4. Transposed the teosinte_3genotypes:

```
awk -f transpose.awk teosinte_3genotypes.txt > transposed_teosinte_3genotypes.txt
```

5. Delet the header of the transposed_teosinte_3genotypes files and sorting them:

```
 grep -v "Sample_ID" transposed_teosinte_3genotypes.txt | sort -k1,1 >
transposed_teosinte_3genotypes_sorted.txt
```

6. Joining

```
join -1 1 -2 1 -t $'\t' snp_position_cut_sorted.txt
transposed_teosinte_3genotypes_sorted.txt > teosinte-SNP_joined.txt
```

7. 10 files (1 for each chromosome) with SNPs ordered based on increasing position values and with missing data encoded by this symbol: ?

```
sed 's/unknown/?/g' teosinte-SNP_joined.txt | sort -k3,3n >
teosinte_qSymbol_sorted.txt
```

```
awk '$2~ /1/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr1.txt
awk '$2~ /2/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr2.txt
awk '$2~ /3/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr3.txt
awk '$2~ /4/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr4.txt
awk '$2~ /5/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr5.txt
awk '$2~ /6/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr6.txt
awk '$2~ /7/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr7.txt
awk '$2~ /8/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr8.txt
awk '$2~ /9/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr9.txt
awk '$2~ /10/' teosinte_qSymbol_sorted.txt > teosinte_increasing_chr10.txt
```

8. 10 files (1 for each chromosome) with SNPs ordered based on decreasing position values and with missing data encoded by this symbol: -

```
sed 's/unknown/-/g' teosinte-SNP_joined.txt | sort -k 3 -r -n > teosinte_-
symbol_sorted.txt
```

```
awk '$2~ /1/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr1.txt
awk '$2~ /2/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr2.txt
awk '$2~ /3/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr3.txt
awk '$2~ /4/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr4.txt
awk '$2~ /5/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr5.txt
awk '$2~ /6/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr6.txt
awk '$2~ /7/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr7.txt
awk '$2~ /8/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr8.txt
awk '$2~ /9/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr9.txt
awk '$2~ /10/' teosinte_-symbol_sorted.txt > teosinte_decreasing_chr10.txt
```

9. SNPs with unknown positions in the genome

```
grep -w "unknown" teosinte-SNP_joined.txt > teosinte_unknown.txt
```

10. SNPs with multiple positions in the genome

```
grep -w "multiple" teosinte-SNP_joined.txt > teosinte_multiple.txt
```