UNIX Assignment:

Data Inspection:

Head

tail

**Fang\_et\_al\_genotypes.txt**

Command: wc fang\_et\_al\_genotypes.txt

File size: 11M **du -h**

Number of columns: 986 tail -n +2 fang\_et\_al\_genotypes.txt | awk -F "\t" '{print NF; exit}'

number of lines: 2783 **wc -l**

# of words: 2744038 **wc -w**

# of bytes: 11051939 **wc –c**

**ASCII characters**

**Snp\_position.txt**

File size: 84K **du -h**

Number of columns: 15 awk -F "\t" '{print NF; exit}' snp\_position.txt

number of lines: 984 **wc -l**

# of words: 13198 **wc -w**

# of bytes: 82763 **wc -c**

**Data processing**

Workflow:

Moved data files into my UNIX\_Assignment folder

**Saved fang\_header**

head -n 1 fang\_et\_al\_genotypes.txt > fang\_header.txt

**Split maize and teosinte genotypes:**

grep -E "ZMMIL|ZMMLR|ZMMMR" fang\_et\_al\_genotypes.txt > maize.txt

grep -E "ZMPBA|ZMPIL|ZMPJA" fang\_et\_al\_genotypes.txt > teosinte.txt

**Add the fang header to maize.txt or teosinte.txt**

Cat fang\_header.txt maize.txt > maize\_with\_header.txt

cat fang\_header.txt teosinte.txt > teosinte\_with\_header.txt

**Cut the first 2 columns:**

cut -f 3-986 maize\_with\_header.txt > maize\_with\_header\_pretranspose.txt

cut -f 3-986 teosinte\_with\_header.txt > teosinte\_with\_header\_pretranspose.txt

**Transpose data:**

awk -f transpose.awk maize\_with\_header\_pretranspose.txt > transposed\_maize.txt

awk -f transpose.awk teosinte\_with\_header\_pretranspose.txt > transposed\_teosinte.txt

**Remove Header:**

Tail –n +2 transposed\_maize.txt > transposed\_maize\_nohead.txt

tail -n +2 transposed\_teosinte.txt > transposed\_teosinte\_nohead.txt

tail -n +2 snp\_position.txt > snp\_position\_nohead.txt

**Sort:**

sort -k1 transposed\_maize\_nohead.txt > sort\_transposed\_maize.txt

sort -k1 snp\_position\_nohead.txt > sort\_snp\_position.txt

sort -k1 transposed\_teosinte\_nohead.txt > sort\_transposed\_teosinte.txt

**Join:**

Create tab delimited files with –t $’\t’

join -1 1 -2 1 -t $'\t' sort\_snp\_position.txt sort\_transposed\_maize.txt > maize\_snp\_join.txt

join -1 1 -2 1 -t $'\t' sort\_snp\_position.txt sort\_transposed\_teosinte.txt > teosinte\_snp\_join.txt

**Filter out the multiple/unknown position files:**

grep -Ev multiple maize\_snp\_join.txt | grep -Ev unknown maize\_snp\_join.txt > known\_maize\_snp.txt

grep -Ev multiple teosinte\_snp\_join.txt | grep -Ev unknown teosinte\_snp\_join.txt > known\_teosinte\_snp.txt

**Save the multiple & unknown files:**

grep -E multiple maize\_snp\_join.txt > multiple\_positions\_maize.txt

grep -E multiple teosinte\_snp\_join.txt > multiple\_positions\_teosinte.txt

grep -E unknown maize\_snp\_join.txt > unknown\_positions\_maize.txt

grep -E unknown teosinte\_snp\_join.txt > unknown\_positions\_teosinte.txt

**Save 1-10 chromosomes:**

**Increasing:**

awk '$3 == 1' known\_maize\_snp.txt | sort -k3,3n > maize\_chrom1\_increasing.txt

awk '$3 == 1' known\_teosinte\_snp.txt | sort -k3,3n > teosinte\_chrom1\_increasing.txt

**Decreasing:**

awk '$3==1' known\_maize\_snp.txt | sort -k3,3n -r > maize\_chrom1\_decreasing.txt

awk '$3==10' known\_teosinte\_snp.txt | sort -k3,3n -r > teosinte\_chrom10\_decreasing.txt

**Replace missing data with ?:**

**Replace missing data with -:**

sed 's/?/-/g' maize\_chrom10\_decreasing.txt > maize\_chrom10\_decr\_encoded-.txt

sed 's/?/-/g' teosinte\_chrom10\_decreasing.txt > teosinte\_chrom10\_decr\_encoded-.txt

**Cut the extra columns:**

cut -f 1,3,4,16-1588 maize\_chrom1\_increasing.txt > maize\_chrom1\_inc.txt

cut -f 1,3,4,16-1588 teosinte\_chrom1\_increasing.txt > teosinte\_chrom1\_inc.txt

cut -f 1,3,4,16-1588 maize\_chrom1\_decr\_encoded-.txt > maize\_chrom1\_decr.txt

cut -f 1,3,4,16-1588 teosinte\_chrom1\_decr\_encoded-.txt > teosinte\_chrom1\_decr.txt

**Create final header:**

Cut –f 1,3,4 snp\_header.txt > final\_header.txt

**Append final header to each final file:**

cat final\_header.txt teosinte\_multiple.txt > teosinte\_multiple\_final.txt

cat final\_header.txt maize\_unknown.txt > maize\_unknown\_final.txt

cat final\_header.txt teosinte\_chrom10\_decr.txt > teosinte\_chrom10\_decr\_final.txt

cat final\_header.txt maize\_chrom1\_incr.txt > maize\_chrom1\_incr\_final.txt