#UNIX Assignment

##Data Inspection

###Attributes of `fang_et_al_genotypes`

here is my snippet of code used for data inspection:

ls -lh fang_et_al_genotypes.txt

file fang_et_al_genotypes.txt

head -n 10 fang_et_al_genotypes.txt

tail -n 10 fang_et_al_genotypes.txt

less fang_et_al_genotypes.txt

wc fang_et_al_genotypes.txt

wc -l fang_et_al_genotypes.txt

wc -w fang_et_al_genotypes.txt

wc -c fang_et_al_genotypes.txt

awk -F'\t' '{print NF; exit}' fang_et_al_genotypes.txt

awk -F'\t' '' fang_et_al_genotypes.txt | sort -u

head -n 1 fang_et_al_genotypes.txt

cut -f1 fang_et_al_genotypes.txt | sort | uniq | head -n 20

cut -f1 fang_et_al_genotypes.txt | sort | uniq | wc -l

grep -E "\bNA\b|\b.\b|\b?\b" fang_et_al_genotypes.txt | head -n 10

cut -f1 fang_et_al_genotypes.txt | sort | uniq -d

By inspecting this file I learned that:

1. The file size is 11 MB
2. It is an ACII text with very long lines
3. it has 2783 Lines
4. It has 2744038 words
5. It has 11051939 Bytes
6. It has 986 columns

###Attributes of `snp_position.txt`

here is my snippet of code used for data inspection:

ls -lh snp_position.txt

file snp_position.txt

head -n 10 snp_position.txt

tail -n 10 snp_position.txt

less snp_position.txt

wc snp_position.txt

wc -l snp_position.txt

wc -w snp_position.txt

wc -c snp_position.txt

awk -F'\t' '{print NF; exit}' snp_position.txt

awk -F'\t' '' snp_position.txt | sort -u

head -n 1 snp_position.txt

cut -f1 snp_position.txt | sort | uniq | head -n 20

cut -f1 snp_position.txt | sort | uniq | wc -l

cut -f1 snp_position.txt | sort | uniq -d

By inspecting this file I learned that:

1. The file size is 81 KB
2. It is an ACII text
3. it has 984 Lines
4. It has 13198 words
5. It has 82763 Bytes
6. It has 15 columns

##Data Processing

###Maize Data

Step 1: Extract ZMMIL, ZMMLR, and ZMMMR from the Group column awk '$3 ~ /Group|ZMMIL|ZMMLR|ZMMMR/' fang_et_al_genotypes.txt > maize_data.txt

Step 2: Transpose the extracted data awk -f transpose.awk maize_data.txt > maize_transposed.txt

Step 3: Sort the transposed file and add the header

head -n 1 maize_transposed.txt > header.txt tail -n +4 maize_transposed.txt | sort -k1,1 > sorted_maize.txt cat header.txt sorted_maize.txt > sorted_maize_with_header.txt

Step 4: Process SNP position data head -n 1 snp_position.txt > snp_header.txt tail -n +2 snp_position.txt | sort -k1,1 > sorted_snp.txt cat snp_header.txt sorted_snp.txt > sorted_snp_with_header.txt cut -f 1,3,4 sorted_snp_with_header.txt > snp_trimmed.txt

Step 5: Join SNP data with maize data sed 's/Sample_ID/SNP_ID/' sorted_maize_with_header.txt > maize_final.txt join -1 1 -2 1 -t $'\t' snp_trimmed.txt maize_final.txt > maize_joined.txt

Step 6: Extract unknown and multiple positions grep -E "(Chromosome|unknown)" maize_joined.txt > maize_unknown.txt grep -E "(Chromosome|multiple)" maize_joined.txt > maize_multiple.txt

Step 7: Sort by increasing position and replace missing data head -n 1 maize_joined.txt > maize_header.txt tail -n +2 maize_joined.txt | sort -k3,3n > maize_sorted_asc.txt cat maize_header.txt maize_sorted_asc.txt | sed 's!?/?!?!g' > maize_final_asc.txt

Step 8: Sort by decreasing position and replace missing data tail -n +2 maize_joined.txt | sort -k3,3nr > maize_sorted_desc.txt cat maize_header.txt maize_sorted_desc.txt | sed 's!?/?!-!g' > maize_final_desc.txt

Step 9: Extract chromosomes (using a loop for efficiency) for i in {1..10}; do awk -v chr="$i" '\(2 == chr' maize_final_asc.txt > maize_chr\)_asc.txt awk -v chr="$i" '\(2 == chr' maize_final_desc.txt > maize_chr\)_desc.txt done

Here is my brief description of what this code does: I wrote brief description of each line in previous section

###Teosinte Data

here is my snippet of code used for data processing Step 1: Extract ZMPBA, ZMPIL, and ZMPJA from the Group column awk '$3 ~ /Group|ZMPBA|ZMPIL|ZMPJA/' fang_et_al_genotypes.txt > teosinte_data.txt

Step 2: Transpose the extracted data awk -f transpose.awk teosinte_data.txt > teosinte_transposed.txt

Step 3: Sort the transposed file and add the header head -n 1 teosinte_transposed.txt > header.txt tail -n +4 teosinte_transposed.txt | sort -k1,1 > sorted_teosinte.txt cat header.txt sorted_teosinte.txt > sorted_teosinte_with_header.txt

Step 4: Join SNP data with teosinte data sed 's/Sample_ID/SNP_ID/' sorted_teosinte_with_header.txt > teosinte_final.txt join -1 1 -2 1 -t $'\t' snp_trimmed.txt teosinte_final.txt > teosinte_joined.txt

Step 5: Extract unknown and multiple positions grep -E "(Chromosome|unknown)" teosinte_joined.txt > teosinte_unknown.txt grep -E "(Chromosome|multiple)" teosinte_joined.txt > teosinte_multiple.txt

Step 6: Sort by increasing position and replace missing data head -n 1 teosinte_joined.txt > teosinte_header.txt tail -n +2 teosinte_joined.txt | sort -k3,3n > teosinte_sorted_asc.txt cat teosinte_header.txt teosinte_sorted_asc.txt | sed 's!?/?!?!g' > teosinte_final_asc.txt

Step 7: Sort by decreasing position and replace missing data tail -n +2 teosinte_joined.txt | sort -k3,3nr > teosinte_sorted_desc.txt cat teosinte_header.txt teosinte_sorted_desc.txt | sed 's!?/?!-!g' > teosinte_final_desc.txt

Step 8: Extract chromosomes (using a loop for efficiency)

for i in {1..10}; do awk -v chr="$i" '\(2 == chr' teosinte_final_asc.txt > teosinte_chr\)_asc.txt awk -v chr="$i" '\(2 == chr' teosinte_final_desc.txt > teosinte_chr\)_desc.txt done

Here is my brief description of what this code does: You can Find Breif Description of each line above that.

Following steps are for making folders for better undeerstanding:

Step 1: Create Folder Structure

Create Maize and Teosinte folders

mkdir -p Maize/increasing Maize/decreasing

mkdir -p Teosinte/increasing Teosinte/decreasing

Step 2: Move Maize Files

Move increasing position files for Maize

mv maize_chr*_asc.txt Maize/increasing/

Move decreasing position files for Maize

mv maize_chr*_desc.txt Maize/decreasing/

Move unknown and multiple files for Maize

mv maize_unknown.txt maize_multiple.txt Maize/

Step 3: Move Teosinte Files

Move increasing position files for Teosinte

mv teosinte_chr*_asc.txt Teosinte/increasing/

Move decreasing position files for Teosinte

mv teosinte_chr*_desc.txt Teosinte/decreasing/

Move unknown and multiple files for Teosinte

mv teosinte_unknown.txt teosinte_multiple.txt Teosinte/

Step 4: Create a Folder for Temporary Files

Create a folder named "temp_files" in the current directory

mkdir -p temp_files

Step 5: Move Temporary Files into the Folder

Move all intermediate maize-related files

mv maize_data.txt maize_transposed.txt sorted_maize.txt header.txt maize_final.txt maize_joined.txt maize_sorted_asc.txt

maize_sorted_desc.txt maize_header.txt temp_files/

Move all intermediate teosinte-related files

mv teosinte_data.txt teosinte_transposed.txt sorted_teosinte.txt teosinte_final.txt teosinte_joined.txt teosinte_sorted_asc.txt

teosinte_sorted_desc.txt teosinte_header.txt temp_files/

Move SNP-related intermediate files

mv snp_header.txt sorted_snp.txt sorted_snp_with_header.txt snp_trimmed.txt temp_files/