UNIX_Exercise

Angela Bunning

BCB 546

Spring 2017

This is my repo for the UNIX homework for BCB546

- First thing I did was to create a new repo on my github for the specific assignment and used `git clone https://github.com/abunning4/UNIX_Exercise.git` to clone my repo to my desktop

Navigating my Repo

- All the final 40 files are in the Final Files directory on github and within that directory the 40 files are divided by ?/? (question data) and -/- (dash data)
    - both the teosinte and maize are in the those two files
- All the files generated to get to the final files are in a directory named Intermediate Files

Data Inspection

1. `fang_et_al_genotypes.txt`
    - `wc` lines: 2783 / words: 2744038 / bytes: 11051939
    - file size by `du -h` : 11 M (11 megabytes)
    - number of columns using `awk` : 986
2. `snp_position.txt`
    - `wc` : lines: 984 / words: 13198 / bytes: 82763
    - file size by `du -h` : 84K (84 kilobytes)
    - number of columns using `awk` : 15

Data Processing

- First, the maize and teosinte data using
    - `grep "ZMMIL" fang_et_al_genotypes.txt >> maize_genotypes_fang.txt` for each genotype an put them into two files: `maize_genotypes_fang.txt` and `teosinte_genotypes_fang.txt`
- I then used the `awk` command Dr. Hufford wrote for the class to transform both the `maize_genotypes_fang.txt` and `teosinte_genotypes_fang.txt` files.
- `cut` was used to excise the columns needed from `snp_position.txt` --> columns for SNP ID, Chromosome, Position
    - `cut -f 1,3,4 snp_position.txt > cut_snp_position`

    - 1,3,4 = SNP ID column, Chromosome column, Position column

- The **common column** between each transposed genotype data for maize and teosinte and the cut snp position file is the SNPID/SampleID
    - These files are already sorted appropriately, and therefore can be joined without `sort` at this time
- The `join` command was used to merge each individual transposed genotype files with the cut snps file.
    - `$ join -t $'\t' -1 1 -2 1 cut_snp_position.txt transposed_teosinte_genotypes.txt > joined_teosinte_snp.txt`
    - `join -t $'\t' -1 1 -2 1 cut_snp_position.txt transposed_maize_genotypes.txt > joined_maize_snp.txt`
- The `sort` command is then used to sort each joined maize snp and teosinte snp increasing chromosome number
- `awk` was then used to pull out the each chromosome and put it in a new file
    - **Example command**: `awk -F \t '$2=="1"' sorted_joined_maize_snp.txt > chr1_maize_question.txt`
    - chromosome was in column 2, and I was pulling out chromosome 1
    - This was completed for every chromosome in both maize and teosinte to generate ten files for the ?/? data
- In each of these chromosome specific files, the `sort` function was used to order them in ascending chromosome position value
    - **Example command**: `sort -k 3,3 chr1_teosinte_question.txt > chr1_teosinte_sorted_question.txt`
    - This was completed for every chromosome for maize and teosinte
- the `sed` function was then used to generate the data files replacing the ?/? with -/-
    - **Example command**: `sed 's/\?/\-/g' chr10_teosinte_sorted_question.txt > chr10_teosinte_sorted_dash.txt`
    - This was completed for every chromosome for teosinte and maize