ST790-003, Homework 1

Due Friday, Jan 23, 2015 @ 11:59PM

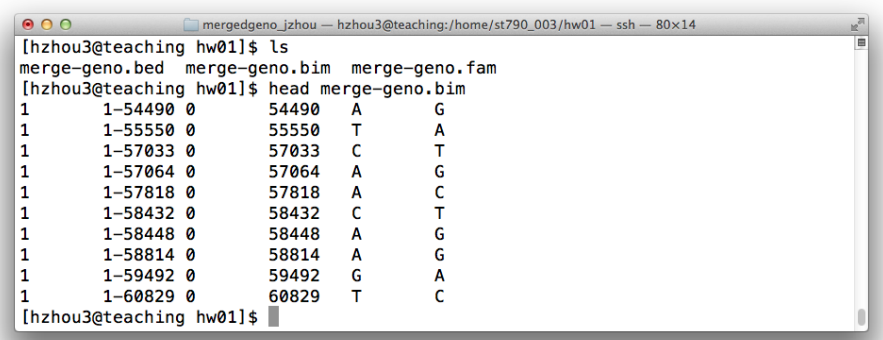
This is a *solo* homework. No group work is allowed.

1. Create a private repository `st790-2015spr` on `github.ncsu.edu`. Add teaching assistant (`tzhang9`) and instructor (`hzhou3`) as your collaborators. Top directories should be `hw1`, `hw2`, ... Create two branches `master` and `develop`. The `develop` branch will be your own playground, the place where you develop solution (code) to homework problems and write up report. The `master` branch will be your presentation area. Put your homework submission files (pdf, html from R Markdown, code to reproduce results, ...) in this branch. *No handwritten homework reports are accepted for this course.* After each homework due date, teaching assistant and instructor will check out your `master` branch for grading. Tag each of your homework submissions with tag names `st790hw1`, `st790hw2`, ...

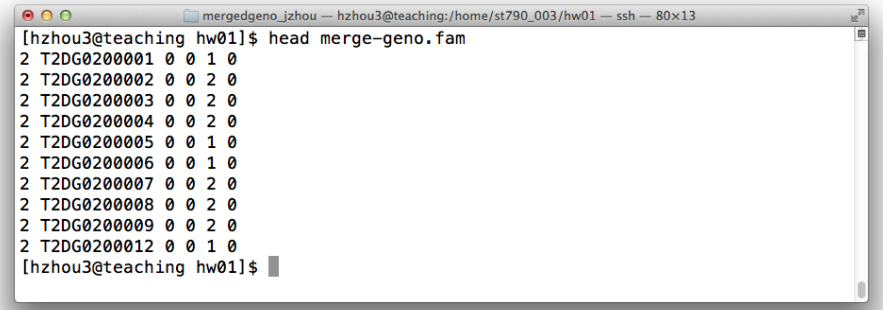
   Efficient and abundant use of Git, e.g., frequent and well-documented commits, is an important criterion for evaluating your homework.

2. The `teaching.stat.ncsu.edu:/home/st790_003/hw01` folder contains a typical genetic data set in plink format. If interested, you can read plink documentation at `http://pngu.mgh.harvard.edu/~purcell/plink/`. But it's definitely not necessary for this homework.

   - `merge-geno.bed` contains genotypes of each individual in binary format.
   - `merge-geno.bim` contains information of each genetic marker (SNP). Each line is a SNP and has fields: Chromosome, SNP ID, Genetic Distance (morgan), Base Pair Position (bp), Allele 1, Allele 2.



   - `merge-geno.fam` contains individual information. Each line is one individual and has fields: Family ID, Person ID, Father ID, Mother ID, Sex coded as 1 (male) or 2 (female), Affection Status. Father ID = 0 means that person's father is not in this data set. Similarly Mother ID = 0 means that person's mother is not in this data set.



Please, do *not* put these data files into Git. They are *huge*. You even don't need to copy them into your directory. Just read from the data folder `teaching.stat.ncsu.edu:/home/st790_003/hw01` directly.

Use Linux shell commands or scripts to answer following questions.

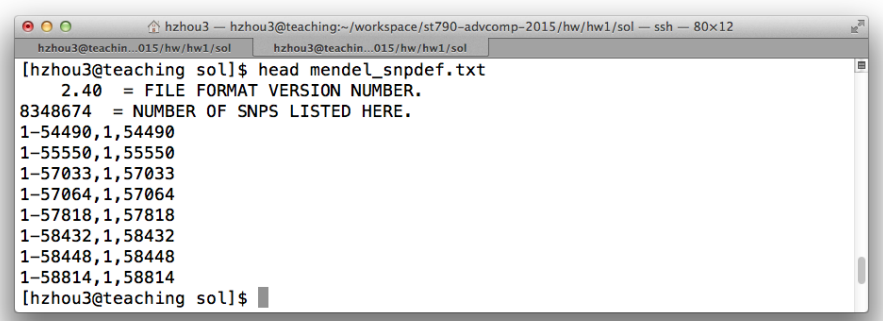
(a) How many persons are in the data set (statisticians call this n)? How many SNPs are in the data set (statisticians call this p)?

(b) Which chromosomes does this data set contain? How many SNPs are in each chromosome?

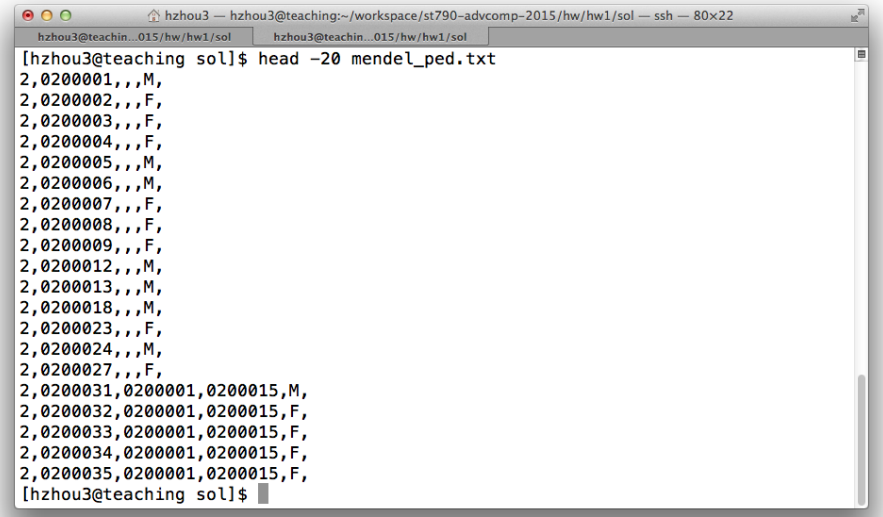
(c) MAP4 (microtubule-associated protein 4) is a gene on chromosome 3 spanning positions 47,892,180 bp – 48,130,769 bp. How many SNPs are located within MAP4 gene?

(d) Statistical geneticists often have to reformat a data set to feed into various analysis programs. For example, to use the Mendel software (`http://www.genetics.ucla.edu/software/mendel`), we have to reformat the data set to be read by Mendel.

   i. Mendel's SNP definition file is similar to the plink `bim` file but has format
   SNP ID, Chromosome, Base Pair Position
   with each field separated by a comma. Write a Linux shell command to convert `merge-geno.bim` to Mendel SNP definition file. The first few lines of the Mendel SNP definition file should look like



ii. Mendel's pedigree file is similar to the plink `fam` file but has format
Family ID, Person ID, Father ID, Mother ID, Sex coded as M or F, Twin Status
with each field separated by a comma. Write a Linux shell command to convert `merge-geno.fam` to Mendel pedigree file. Since twin status is not available in plink format, we put nothing for that field. Also Mendel limits Person ID to have length less or equal to 8 characters, so we have to strip the string `T2DG` from the IDs. First few lines of the Mendel pedigree should look like



Again, do *not* put output files into Git. They are huge.