

ST790-003, Homework 1

Due Monday, Jan 19, 2015 @ 11:59PM

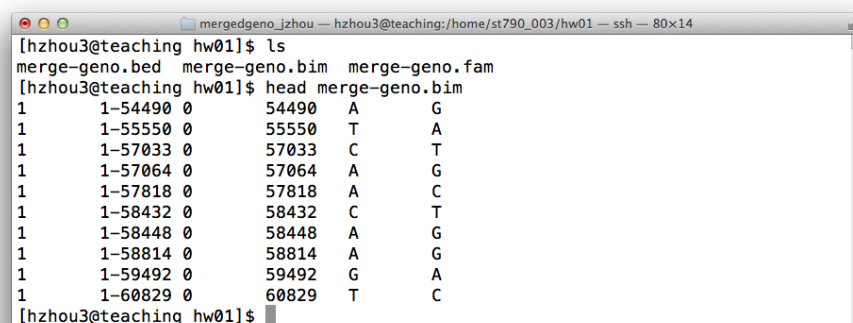
This is a *solo* homework. No group work is allowed.

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.



```
mergedgeno_jzhou — hzhou3@teaching:/home/st790_003/hw01 — ssh — 80x14
[hzhou3@teaching hw01]$ ls
merge-geno.bed merge-geno.bim merge-geno.fam
[hzhou3@teaching hw01]$ head merge-geno.bim
1      1-54490 0      54490  A      G
1      1-55550 0      55550  T      A
1      1-57033 0      57033  C      T
1      1-57064 0      57064  A      G
1      1-57818 0      57818  A      C
1      1-58432 0      58432  C      T
1      1-58448 0      58448  A      G
1      1-58814 0      58814  A      G
1      1-59492 0      59492  G      A
1      1-60829 0      60829  T      C
[hzhou3@teaching hw01]$
```

- `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.

```
mergedgeno_jzhou — hzhou3@teaching:/home/st790_003/hw01 — ssh — 80x13
[hzhou3@teaching hw01]$ head merge-geno.fam
2 T2DG0200001 0 0 1 0
2 T2DG0200002 0 0 2 0
2 T2DG0200003 0 0 2 0
2 T2DG0200004 0 0 2 0
2 T2DG0200005 0 0 1 0
2 T2DG0200006 0 0 1 0
2 T2DG0200007 0 0 2 0
2 T2DG0200008 0 0 2 0
2 T2DG0200009 0 0 2 0
2 T2DG0200012 0 0 1 0
[hzhou3@teaching hw01]$
```

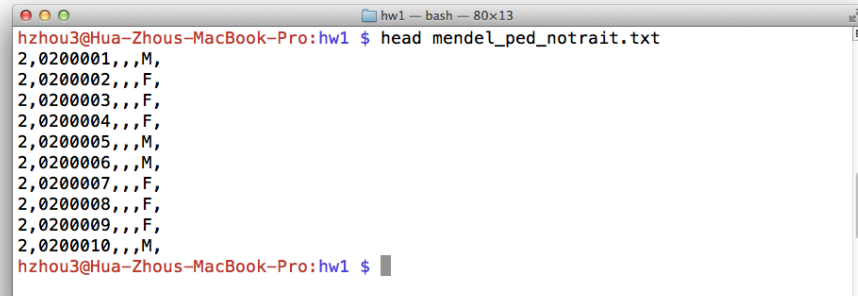
Please, do *not* put these data files into Git. They are *huge*. You even don't need to copy them into your directory. Just read from the data folder `teaching.stat.ncsu.edu:/home/st790_003/hw01` directly.

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

```
realpheno — hzhou3@teaching:/home/st790_003/hw01 — bash — 80x14
hzhou3@Hua-Zhous-MacBook-Pro:realpheno $ head mendel_snpdef.txt
2.40 = FILE FORMAT VERSION NUMBER.
8348674 = NUMBER OF SNPS LISTED HERE.
1-54490,1,54490
1-55550,1,55550
1-57033,1,57033
1-57064,1,57064
1-57818,1,57818
1-58432,1,58432
1-58448,1,58448
1-58814,1,58814
hzhou3@Hua-Zhous-MacBook-Pro:realpheno $
```

- 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



```
hzhou3@Hua-Zhous-MacBook-Pro:hw1 $ head mendel_ped_notrait.txt
2,0200001,,,M,
2,0200002,,,F,
2,0200003,,,F,
2,0200004,,,F,
2,0200005,,,M,
2,0200006,,,M,
2,0200007,,,F,
2,0200008,,,F,
2,0200009,,,F,
2,0200010,,,M,
hzhou3@Hua-Zhous-MacBook-Pro:hw1 $
```

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