Lab 2 – Simulation of genomic data

Prepared by I. Aguilar, D. Lourenco, A. Legarra, and Z. Vitezica

We will use the software QMSim for data simulation (Sargolzaei & Schenkel, 2009, Bioinformatics 25:680-681). The software and its manual can be found in http://www.aps.uoguelph.ca/~msargol/qmsim/

The files for these exercises are in:
`/data1/RFN201908/daniela/labs/lab2`
For running QMSim, use: `echo ex01.par | ./QMsim`

Exercise 1

1. Run the QMSim program. An example of the parameter file is *ex01.par*. Note that historical population was generated by mutation and drift over 100 generations (t) with an effective population size of 100 (t = 1 to 95) and gradually expanded to 3,000 offspring (t = 100).

2. Now change *ex01.par* and simulate a base population of 200 males and 2,600 females, 5 generations of selection for a trait (i.e. live weight) with a phenotypic variance of 1.

    a) How many SNPs did you simulate?
    b) How many QTLs might potentially affect the phenotype?
    c) How many animals do you have in the recent population?
    d) Answer the previous question assuming a litter size equal to 12.
    e) Which is the mean of the TBVs after 5 generations?
    f) Use selection and culling based on EBVs, does the mean of the TBV change?
    g) Include positive assortative mating.
    h) Which is the value of the polygenic variance?

3. Look at the simulated file *p1_data_001.txt*, you will have the following columns:

1: animal id
2: sire id
3: dam id
4: sex
5: generation
6: number of males' progenies
7: number of females' progenies
8: inbreeding
9: homozygosity
10: phenotype
11: simulated residual (e)
12: individual true breeding value for polygenic effect
13: individual true breeding value for direct effect (QTL)
14: EBV from QMSim internal BLUP

4. Check the file p1_freq_mrk_001.txt. What does it show?

5. Take a look at script *edit_data.sh*. Run it by typing `bash edit_data.sh` This script creates the pedigree, the phenotype, and the genotype files for

BLUPf90 from the QMSim simulated data. Which is the pedigree file, the phenotype file, and the marker file?
Note that `edit_data.sh` uses the directory `r_ex01b/`

6. Using Unix commands, check the number of animals in the pedigree file:
   `wc -l ped.txt`
   Do the same for the phenotype file.

7. Look at the genotype file
   `less -S snp.txt`

8. Before continuing the analysis, it is important to check the "quality" of the files for some typical errors. For example, are there duplicated animals in the pedigree? Check it using
   ```
   awk '{print $1}' ped.txt | sort +0 -1 | uniq -c |
   awk '$1>1'
   ```

9. What is the number of progeny for each sire?
   ```
   awk '{print $2}' ped.txt | sort +0 -1 | uniq -c >
   sire.prog
   ```

10. How many genotyped animals are in the SNP file?
    `wc -l snp.txt`

11. How many SNP?
    `awk '{print length($2)}' snp.txt`

12. Does everyone is genotyped for the same number of SNP?
    `awk '{print length($2)}' snp.txt | sort -u`

13. Let's extract the first 200 individuals *while keeping the format.* This is a one-line command, beware of simple and double quotes!

    ```
    awk 'NR <= 200' snp.txt > anim200.temp
    ```

    OR

    ```
    awk 'NR<=200 {printf("%-10s %s\n", $1,$2)}' snp.txt
    > anim200a.temp
    ```

    OR

    ```
     awk 'NR <= 200 {printf("%10s%1s%" length($2) "s\n",
    $1, " ", $2) }' snp.txt > anim200b.temp
    ```

Note that the format is defined by "%10s%1s%" length($2) "s\n" which means "10 positions, 1 position (for the space in " " later), as many positions as SNPs we have for each individual (in %" length($2) "s), and the line return in \n.

14. Extract SNP number 50 for all animals:

```
awk ' {printf("%10s%1s%1s\n", $1, " ",
substr($2,50,1)) }' snp.txt > snp50.temp
```

    Note that it returns 1 SNP from column 2, starting at position 50.

15. Create a pedigree file only for genotyped animals:

```
sort +0 -1 ped.txt > s_ped.temp
awk '{print $1}' snp.txt | sort +0 -1 > s_gen.temp
join -1 +1 -2 +1 s_gen.temp s_ped.temp > ped.gen
```

16. Create a pedigree file only for ungenotyped animals:

```
join -v2 -1 +1 -2 +1 s_gen.temp s_ped.temp  >
ped.ungen
```

17. Compute average phenotypic value and true breeding value for genotyped animals:

```
sort +0 -1 data.txt > s_data.temp
join -1 +1 -2 +1 s_gen.temp s_data.temp > data.gen
awk '{ sumf += $10 } END { print "Average Phenotype
= " sumf/NR }' data.gen
awk '{ sumf += $13 } END { print "Average TBV = "
sumf/NR }' data.gen
```

Optional Exercise

1. Run the QMSim program with the parameter file: **'ex02.par'**. Note that the population structure involves an F2 design produced from inbred lines with divergent phenotypes.

    1. How many SNPs did you simulate?
    2. How many animals do you have in the cross between line 1 and line 2 after 2 generations?
    3. Which are the values of inbreeding in lines 1 and 2?

2. Edit phenotypes, pedigree, and genotypes. Be aware you need tom combine data from lines 1, 2, and F1.

3. Write a parameter file to simulate a backcross between the F1 and the line 1.