```bash
#!/bin/bash

## 1) simulating data with QMSim
echo ex01.par | QMSim

## 2) simulating data with QMSim
## base population of 200 males and 2,600 females
## 5 generations of selection for a trait with a phenotypic variance of 1
echo ex01b.par | QMSim

#a)      How many SNPs did you simulate?
grep "begin_chr" ex01b.par
grep "nmloci" ex01b.par
#b)      How many QTLs might potentially affect the phenotype?
grep "begin_chr" ex01b.par
grep "nqloci" ex01b.par
#c)      How many animals do you have in the recent population?
  #2600*5 + 2600 + 200 = 15800
#d)      Answer the previous question assuming a litter size equal to 12.
  # (2600*12)*5 + 2600 + 200 = 158800
#e)      Which is the mean of the TBVs after 5 generations?
awk '{ sumf += $13 } END { print " Average TBV = " sumf/NR }' r_ex01b/p1_data_001.txt
  #Average TBV = 0.0567695
#f)      Use selection and culling based on EBVs, does the mean of the TBV change?
#g)      Include positive assortative mating.
  #md = p_assort/ebv;
#h)      Which is the value of the polygenic variance?

## 4) Check the file p1_freq_mrk_001.txt. What does it show?
  # allele frequencies

## 5) Run edit_data.sh
bash edit_data.sh

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

## 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
```