SGIG Brisbane Module 7

Practical Mo 13 Feb afternoon: **Genomic Prediction accuracy**

1: Design parameters to predict accuracy

Use the spreadsheet '*GSaccuracy.xls*'.

Note that this program uses several different formulas, according to different references. These mainly differ in how they predict Me. But you can check how important that is to determine prediction accuracy

Use the program to investigate and explain the impact of the following parameters on the effective number of chromosome segments (M_e), the proportion of variance explained by markers (q^2), the accuracy with which marker effects are estimated ($r_Q$hat), and the accuracy of the genomic prediction of breeding value, GBV (r_{MBV}):

- number of markers (M)
- Effective population size (N_e)
- Heritability of phenotypes (h^2)
- Number of training individuals (N)

Set L=1 and k=30 for a genome of 30 chromosomes of 1 Morgan

You can compare different formulas and contrast those

a) What is the minimum number of markers that is needed to achieve near maximum genome coverage (q^2=0.99) when N_e=100 versus 340 versus 1000 versus 10,000? Enter this in the table below.

b) Set the number of markers M = 1,000,000 to get nearly complete coverage regardless of N_e. Set h^2=0.9.
Now evaluate the size of the training set (N) needed to reach an MBV accuracy of 0.8 for N_e=100, versus 340 versus 1,000 versus 10,000. Enter the results in the table below.

c) Repeat b) for heritabilities equal to 0.5 and 0.2

	N_e = 100	**N_e = 340**	**N_e = 1000**	**N_e = 10,000**
Min. # markers				
h^2 = 0.9				
h^2 = 0.5				
h^2 = 0.2				

d) Test the genome scaling argument that if the size of the simulated genome is reduced by a factor C, then the size of the training population also has to be reduced by the same factor C in order to maintain the same accuracy of MBV.

2: Combining information sources

Use the spreadsheet GSaccuracyHeteroSources.xls

This allows you to look at designs of reference populations with varying numbers of more and less related individuals, and its effect on overall prediction accuracy.
For each source we have a number observed (N) and an effective size

It uses three sources of information	N	Ne
1) Wider population	usually many	large
2) More local data, e.g. herd mates, or a local community	fewer	lower
3) Direct relatives, e.g. half sibs	few	very small

Explore the overall prediction accuracy by varying N1, N2 and N3, as well as Ne1, Ne2 and Ne3.
You can draw a graph showing accuracy versus total number in the reference, with and without closer relatives.
Set h2=0.25, and Ne of the population is 1000:
Compare accuracy with and without closer relatives, for N = 2000, 5000, 10,000, 20,000 (or simply look at graph)
Repeat for h2=0.05

Repeat for Ne of the population is 100:

Set again h2=0.25, and Ne of the population is 1000: Compare Nmarkers = 10k, 50k, 500k

Using the mtg2 program (written by Sang Hong Lee)
Use the following website
https://sites.google.com/site/honglee0707/mtg2
you can download a windows or a linux version

- Download the zip file
- Unzip the zip file
- Install: ww_ifort_redist_ia32_2016.1.146.msi" by double clicking the filename OR
- Install: "ww_ifort_redist_intel64_2016.1.146.msi" and restart
- Open a DOS window and run the mtg2.exe program by simply typing..
-

```
mtg2
```

the program will tell you that it needs some files for analysis.

```
-p fam file –d dat file –g grm file …
```

However, we want to use the program for calculation of Me, given certain parameters:

```
mtg2 –Me
```

the program will then respond

```
Effective number of chromosome segment given the following parameters *****
Ne, length and number of chromosome should be specified
```

So we can try:

```
mtg2 -Me 100 1 30
```

```
Effective number of chromosome segment given the following parameters *****
Effective pop. size          :   100
Genomic length per each chr   :    1.000000
Number of chr                :   30

Eq. 10 :   246.4847
Eq. 11 :   253.5588
```

The program can also provide the theoretical prediction accuracy, given certain parameters

```
mtg2 -pred_acc
```

```
Expected prediction accuracy given the following parameters *****
h2, N, M, Me, k, p, p2 and ckv should be specified
```

The paramters are:

h2 : Proportion of variance due to SNPs on the liability scale

N   : Sample size

M   : Number of SNPs

Me : Effective number of chromosome segments

k   : Population prevalence

p   : Proportion of cases in training sample

p2 : Proportion of cases in validation sample

ckv : Top prop. of genetic profile scores in validation sample

The last 4 parameters are only relevant if you are interested in case control type predictions
If that is not relevant, use a value of 0.001 for each of those paramters

mtg2 -pred_acc 0.3, 1000, 30000, 250, 0.001, 0.001, 0.001, 0.001

and the output will be:

```
************************************************************


 Cor(g,g-hat) in quantitative trait       :   0.7371541
 Cor(y,g-hat) in quantitative trait       :   0.4020841
 Cor(u,u-hat) in case-control data        :   0.2804853
 Cor(y,u-hat) in case-control data        :   4.1185454E-03
 Cor(y,u-hat) on the liability scale      :   0.1529920
 AUC in case-control data                 :   0.6152011
 OR contrasting top/bottom percentile     :            NaN
 OR contrasting top/general population    :            NaN
```

If you are interested in case control studies, you can explore these parameters as well.