

'Automatic' GWAS analyses using WOMBAT

WOMBAT now provides an option for GWAS analyses, processing a large number of SNPs one by one in a single analysis. This uses an efficient computing strategy which avoids inversion of large matrices and splits computations into a part to be performed once and a much smaller part specific to individual SNPs.

- The GWAS feature is invoked with the run time option **--snap** (standing for 'snappy').
- It provides estimates of multiple SNP effects together with their standard errors for a mixed model with specified (co)variance components for residuals and random effects fitted.
- SNP effects are expected to be fitted as linear covariables.
- The other effects in the model are arbitrary, in such that all models usually available in WOMBAT can be specified.
- Allele counts are expected to be read sequentially from a separate file:
 - This file has the default name **QTLAllels.dat**
 - The file, or a symbolic link to it, must be in the working directory.
 - There should be one long row with allele counts for each SNP to be analyzed.
 - Counts should be a single digit for each individual without spaces between them. For example, if there are 1000 genotyped individuals, a row comprised of 1000 digits – usually 0, 1 or 2 – is expected to be read (in FORTRAN: the input format would be (1000 i1)).
 - Currently, there is no provision for missing information, i.e. any missing counts should be imputed prior to analysis or the data should be edited accordingly.
 - The file with allele counts should not contain any blank line – due to the formatted read, WOMBAT can not distinguish between a blank line and a line with all allele counts of "0"!
 - Any SNP with all allele counts the same for all individuals (i.e. a SNP which is not polymorphic) is reported. This is likely to yield an undetermined equation – WOMBAT checks for this and sets corresponding estimates and standard errors to zero. In addition, a warning message is written to `WOMBAT.log`.
- Estimates are written out to a file called **QTLSolutions.dat**; this file has a line per SNP with columns 1 to 3 giving the estimate, its standard error and the respective *t*-value.

Toy Example

This is Example14 as distributed with WOMBAT.

Consider records for 16 animals given in file Simdata.dat. Column 1 gives the animal code, columns 2 and 3 give the code for two cross-classified fixed effects and column 4 codes their interaction. The SNP allele count is given in column 5. Column 6 gives another covariable and column 7 represents the trait measured.

Corresponding pedigree information is given in file SimPed.dat. Including parents without records themselves yields a total of 26 animal genetic effects in the model of analysis.

The SNP allele counts given in the data file are not analyzed, but merely act as place holders. The actual counts are expected to be read from the file QTLAllels.dat. For this example, we have a total of 4 SNPs and 16 animals with records. Hence the file QTLAllels.dat has 4 rows with 16 digits each (of 0, 1 or 2) giving the allele counts for the animals in the data in the *same* order as the records in the data file. For instance, the "1" in column 1 of row 1 is the allele count for animal "104", the "2" in column 2 pertains to animal "105" and the "1" in column 16 is the count for animal "213".

QTLAllels.dat

```
1211001011111101
1011001011211101
1211001022011101
1211001011011122
```

SimData.dat

```
104 3 21 321 2.0 15.0 237.6668
105 3 20 320 3.0 14.0 261.7078
108 3 21 321 2.0 15.0 236.7068
109 2 21 221 2.0 15.0 215.2132
110 4 22 422 1.0 14.0 242.9927
111 4 21 421 1.0 10.0 225.6397
112 4 22 422 2.0 13.0 248.1895
113 3 20 320 1.0 11.0 187.5965
204 4 22 422 2.0 12.0 228.1260
205 2 20 220 2.0 14.0 206.7631
208 4 21 421 2.0 14.0 249.6117
209 3 21 321 2.0 12.0 221.0490
210 2 21 221 2.0 12.0 206.4229
211 2 21 221 2.0 15.0 209.7781
212 3 21 321 1.0 15.0 225.9217
213 3 21 321 2.0 14.0 229.8008
```

SimPed.dat

```
101 0 0
102 0 0
103 0 0
104 101 102
105 101 102
106 0 0
107 0 0
108 101 103
109 101 103
110 105 106
111 105 106
112 107 108
113 107 108
201 0 0
202 0 0
203 0 0
204 201 202
205 201 202
206 0 0
207 0 0
208 201 203
209 201 203
210 205 206
211 205 206
212 207 208
213 207 208
```

The parameter file – `wombat.par` – for this analysis specifies the analysis type, data and pedigree file and model of analysis as for any other WOMBAT analysis. Specific features for the GWAS analysis are:

- The first line: `RUNOP --snap`. This gives the run time option in the parameter file. This is convenient so as not to forget it; alternatively it could be given on the command line.
- The special instruction: `COVZER snp(1) FIT. WOMBAT` is fussy about covariables which have a value of zero – too often are these missing values! This line specifies that any allele counts of “0” are indeed valid covariable values
- The special instruction: `QTLEFF snp(1)`. This specifies which of the covariables fitted represents the allele counts which are subsequently to be replaced with values read from `QTLAllels.dat`.

Finally, estimates for the 4 SNPs analyzed are given in the file `QTLsolutions.dat`, with column 1 containing the estimate, column 2 the corresponding standard error and column 3 the ratio of the two, i.e. the t -value for a significance test.

QTLsolutions.dat			
17.7955	6.12671	2.90457	QTLeffect1
-6.67386	7.59983	-0.878159	QTLeffect2
11.6144	4.51428	2.57281	QTLeffect3
12.1597	5.09893	2.38476	QTLeffect4

```

wombat.par
RUNOP --snap
ANAL UNI
PEDS ../SimPed.dat
DATA ../SimData.dat
  animal 0
  kfix 10
  mfix 33
  kxl 199
  snp
  age
  wgt
END

MODEL
  COV snp(1)
  COV age(2)
  FIX kfix
  FIX mfix
  FIX kxl
  RAN animal nrm
  trait wgt
END MOD

VAR animal 1
30
VAR residual 1
70

SPECIAL
  COVZER snp(1) FIT
  QTLEFF snp(1)
END

```