GONE TUTORIAL (4th April 2022)

In this tutorial we will show how to use the software GONE to obtain historical estimates of the effective population size and will mention some of its limitations. We will consider an example set with 100 individuals, 10 chromosomes of 25 Mb each, and about 100,000 SNPs.

FILES NEEDED

The files needed to run the software in Linux or in Mac are available in the directories **Linux** and **MacOSX**.

Copy the whole directory **Linux** (or **MacOSX**) in your equipment. This directory has a subdirectory called **PROGRAMMES** with all necessary executable files. Make sure these files have the permission to be run. If not, use command "chmod r+x ExecutableFileName" to grant permission. There is also a file called **INPUT_PARAMETERS_FILE** with the different parameters considered in the analysis, the **script_GONE.sh** to run the programme, and an example data with the usual Plink files, **example.map** and **example.ped**.

Thus, in your running directory you must have:

```
$ ls
PROGRAMMES
example.map
example.ped
INPUT_PARAMETERS_FILE
script_GONE.sh
```

If you prefer to compile the programmes by yourself, the codes are available in the directories CODES_LINUX and CODES_MACOSX, and there is a **COMPILE&LINK.txt** file to do it. GONE and GONEaverage require gcc/7.2.0 to be compiled. The compiled files should be included in the directory **PROGRAMMES**.

INPUT FILES

The **data.map** should have the following columns separated by spaces:
> First column: chromosome or linkage group number **starting from 1**. The maximum number of chromosomes allowed is **200**.
> Second column: SNP name or number, e.g. SNP1.
> Third column: if available, the genetic location in centimorgans.
> Fourth column: The SNP genomic position, e.g. 423.

```
$ more example.map
1 SNP1 0.000423 423
1 SNP2 0.005214 5214
1 SNP3 0.005887 5887
1 SNP6 0.008360 8360
1 SNP7 0.009697 9697
1 SNP8 0.009992 9992
1 SNP9 0.010005 10005
1 SNP10 0.014364 14364
1 SNP12 0.016301 16301
```

```
$ tail example.map
10 SNP119103 249.959541 249959541
10 SNP119104 249.959596 249959596
10 SNP119106 249.968239 249968239
10 SNP119107 249.969003 249969003
10 SNP119108 249.980688 249980688
10 SNP119110 249.985945 249985945
10 SNP119111 249.986972 249986972
10 SNP119113 249.996743 249996743
```

If the genetic locations are not available, the third column should contain **zeroes**, as shown below.

```
$ more example.map
1 SNP1 0 423
1 SNP2 0 5214
1 SNP3 0 5887
1 SNP6 0 8360
1 SNP7 0 9697
1 SNP8 0 9992
1 SNP9 0 10005
1 SNP10 0 14364
1 SNP12 0 16301
1 SNP13 0 19207
1 SNP14 0 20904
```

In this case, the software will assume an average rate of recombination for the species (e.g. 1 cM per Mb by default) which is set up in the **INPUT_PARAMETERS_FILE** (see below).

The **data.ped** file has the genotypes for each individual (one individual in each line) with this format:

```
1 IND1 0 0 1 -9 T A A T ...
1 IND2 0 0 1 -9 T T A T ...
```

Genotypes are assigned by pairs of letters for each marker. Thus, individual IND1 is heterozygote TA (the first two letters in the line) for the first SNP and individual IND2 is homozygote TT for the first SNP.

The line **must have a -9** just before the genotypes; the columns before the -9 are not considered.

It also admits 1 and 2 as SNP alleles, i.e.:

```
1 nameofsnp1 0 0 1 -9 2 1 1 2 ...
1 nameofsnp2 0 0 1 -9 2 2 1 2 ...
```

To indicate **missing genotyping data use 0** (not -9).

For the example,

```
$ more example.ped
1 IND1 0 0 1 -9 A A A A A T A A T A A A T A A A A A T A A A A A A A A A A A A
A A T A T A A A A A A A A A T A A A A A A A T A A A A A A A A A A A A T A A A A A
A A T A A A A A A A T A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A A T A A A A A A A T T A A A A A A A A A T A A A A A A A T A T A A
A A A A A A A A A A A A A A A A A A A A A A A A A T A A A A A A A A A A A T A T
A A A A A A A A A A A A A A A A A T A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A T A A A A A T T A A A A A A A A A A A T T A A A A A A A A A A A
```

The **INPUT_PARAMETERS_FILE** includes the parameters that can be set up:

```
$ more INPUT_PARAMETERS_FILE
#INPUT_PARAMETERS_FILE
###############################################################
PHASE=2 ### Phase = 0 (pseudohaploids), 1 (known phase), 2 (unknown phase)
cMMb=1  ### CentiMorgans per Megabase (if distance is not available).
DIST=1  ### none (0), Haldane correction (1) or Kosambi correction (2)
NGEN=2000 ### Number of generations for which linkage data is obtained in bins
NBIN=400  ### Number of bins (so that each bin includes NGEN/NBIN generations)
MAF=0.0   ### Minimum allele frequency (0-1)
ZERO=1    ### 0: Remove SNPs with zeroes (1: allow for them)
maxNCHROM=-99  ### Maximum number of chromosomes to be analysed (-99 = all chrom
osomes; maximum number is 200)
maxNSNP=50000 ### Maximum approx number of SNPs per chromosomes to be analysed
hc=0.05   ### Maximum value of c analysed
REPS=40   ### Number of GONE internal replicates
threads=-99  ### Number of threads (if -99 it uses all possible processors)
#####################################################################
```

PHASE indicates if the phase is that for pseudohaploids (PHASE=1), diploids with known phase (PHASE=1) or unknown phase (PHASE=2). The two letters of each pair must be the same for each marker in the case of pseudohaploids. For phased genomes, the first letter of each pairs always corresponds to the same chromosome (say maternal) and the second letter corresponds to the other chromosome.

cMMb indicates the average rate of recombination (in centiMorgans per Megabase) assumed if the genetic distance between markers is unknown. If the third column of the data.map file contains only zeroes, this entrance is used. If not, the entrance is ignored and the distances of the data.map file are used.

DIST indicates if no genetic distance correction is applied (0), or Haldane´s correction (1) or Kosambi´s correction (2) are applied.

NGEN is the number of generations for which linkage data is obtained in bins. For example, NGEN=2000 means that there will be data analysed for the equivalent to 2000 generations, i.e. only pairs of SNPs with recombination fraction c > 1/4000 will be analysed.

NBIN is the number of bins, so that each bin includes NGEN/NBIN generations. For example, if NBIN=400, there will be 400 bins of pairs of SNPs, thus each with 2000/400 = 5 generations interval. However, the first 10 generations are always analysed with 2-generations intervals.

MAF is the minor allele frequency to be applied (0-1). If MAF=0, no MAF pruning will be applied to the data. This is the recommended action.

ZERO This is a variable to indicate if ungenotyped SNPs are removed (0) or not (1). The default option should be ZERO=1, i.e. allow for ungenotyped SNPs. In any case, if for a given SNP, less than half the total number of individuals are not genotyped, the SNP will not be considered in the analysis. The option ZERO=0 discards all SNPs for which at least one individual has not been genotyped (allele 0 in the ped file).

maxNCHROM is the variable used to set up a maximum number of chromosomes to be analysed. If maxCHROM=-99, all chromosomes will be analysed. The maximum number of chromosomes to be analysed is 200. This is useful for example for human data, where the map file may include up to 26 entrances (X and Y chromosomes, mitochondrial DNA, etc.) If you want to analyse only the autosomal chromosomes, then maxCHROM=22 will only consider the first 22 chromosomes.

maxNSNP gives the approximated number of SNPs per chromosome to be analysed. The maximum is 100,000 but the process may fail because of lack of memory and the recommended maximum is 50,000. If the number of SNPs in the chromosome is lower than this number, all SNPs of the chromosome will be analysed. If the value is larger, however, a random sample of 50,000 SNPs will be used. This is useful to avoid too lengthy estimations. In addition, different

runs may include different random subsets of SNPs allowing for empirical errors of temporal N_e (accounting only for errors attached to SNP variation) to be obtained. If you do so, do not send all runs at the same time, as the initial random seed is taken from the computer clock time.

hc gives the maximum value of recombination rate (c) analysed. A pair of SNPs very far away in the chromosome are expected to have a recombination rate of c=0.5 (independent SNPs). Thus, the first bin of pairs of SNPs is that with c=0.5 or lower, and the following bins include lower values of c (SNPs closer and closer). If the population has recent migrants from another population, the estimation of N_e will be generally biased. A typical artefact observed is a very recent drastic drop and a previous increase (see Fig. 2F of Santiago et al. 2020, and below) when the analysis corresponds to a mixture of populations. This can be partly corrected by using a maximum value of c lower than hc=0.5. A value of hc=0.05 is the recommended one for empirical data. For simulation data, where sampling is random and there is no admixture of other populations you can use hc=0.5, which will use all recombination rate bins.

REPS is the number of internal replicates run by GONE. In order to get a consensus estimate of the historical N_e, GONE makes replicates and provides the geometric mean of estimates from these replicates. The default and recommended value is REPS=40.

Threads indicates the number of threads to be used to run the processes in parallel. If Threads=-99 the computer will use all possible processors. To know the number of processors in your equipment type getconf _NPROCESSORS_ONLN.

The number of chromosomes (NCHR) and the sample size (SAM) (number of diploid individuals) are taken from ped and map files. For human data we suggest to use only the first 22 (autosomal) chromosomes. The **minimum sample size is two individuals**, and the **maximum sample size is 1,800 individuals**.

Do not try to run data sets with more than, say, **ten million SNPs**, as the software may crash, and the recommended maximum number of SNPs used per chromosome in each analysis is **50,000**. An analysis of human data with 9.1 million SNPs in the map and ped files, but analysing 50,000 SNPs for each of the 22 chromosomes in a cluster with 22 processors, took 546 minutes to get the linkage disequilibrium data for 1.1 million SNPs, which implied 27×10^9 pairs of SNPs, and only 209 seconds for GONE.

TO RUN THE SOFTWARE

The ideal is to run the script in a scratch directory with a command “qsub”, or equivalent, for which the script should be accommodated depending on the machine. But you can run it directly in your home directory as

```
bash script_GONE.sh <FILE>
```

For example, using the simulation data from example.map and example.ped, the running command would be:

```
bash script_GONE.sh example
```

In the example there are 10 chromosomes, a total of about 100,000 SNPs and a sample of 20 individuals. For this example, using 8 processors in parallel the time needed for the chromosomal analysis is 209 seconds, and for GONE 138 seconds. Note that for data with many more SNPs and individuals the time of processing can be much longer as it increases approximately with the square of the number of SNPs.

OUTPUT FILES

A **TEMPORARY_FILES** directory is made to include all temporary files. This includes the map and ped files for each of the chromosomes, and an output file (outfileLD1 … 10) also for each chromosome. This output file is identical to the final output file (see below) and can be used to run GONE for specific chromosomes. There is also a directory outfileLD_TEMP which includes the temporary files obtained by GONE. These give the estimates obtained for the different prediction replicates (REPS), which are later averaged as a geometric mean.

```
$ ls -l
total 11048
-rw-rw-r-- 1 TEMPORARY_FILES 288528 Mar 18 12:17 chromosome1.map
-rw-rw-r-- 1 TEMPORARY_FILES 819471 Mar 18 12:17 chromosome1.ped
-rw-rw-r-- 1 TEMPORARY_FILES 280530 Mar 18 12:17 chromosome2.map
-rw-rw-r-- 1 TEMPORARY_FILES 748431 Mar 18 12:17 chromosome2.ped
-rw-rw-r-- 1 TEMPORARY_FILES 281310 Mar 18 12:17 chromosome3.map
-rw-rw-r-- 1 TEMPORARY_FILES 750511 Mar 18 12:17 chromosome3.ped
……..
-rw-rw-r-- 1 TEMPORARY_FILES 754511 Mar 18 12:17 chromosome8.ped
-rw-rw-r-- 1 TEMPORARY_FILES 332770 Mar 18 12:17 chromosome9.map
-rw-rw-r-- 1 TEMPORARY_FILES 815151 Mar 18 12:17 chromosome9.ped
-rw-rw-r-- 1 TEMPORARY_FILES 332770 Mar 18 12:17 chromosome10.map
-rw-rw-r-- 1 TEMPORARY_FILES 815151 Mar 18 12:17 chromosome10.ped
-rw-rw-r-- 1 TEMPORARY_FILES   9024 Mar 18 12:20 outfileLD1
-rw-rw-r-- 1 TEMPORARY_FILES   9010 Mar 18 12:20 outfileLD2
…….
-rw-rw-r-- 1 TEMPORARY_FILES   9009 Mar 18 12:20 outfileLD9
-rw-rw-r-- 1 TEMPORARY_FILES   9025 Mar 18 12:20 outfileLD10
drwxrwxr-x 2 TEMPORARY_FILES 12288 Mar 18 12:22 outfileLD_TEMP
```

An OUTPUT_FILE (e.g. OUTPUT_example) will show the total number of SNPs used, the deviation from Hardy-Weinberg proportions in the sample and in the population, and other information per chromosome as shown below.

```
$ more OUTPUT_example

TOTAL NUMBER OF SNPs
73240

HARDY-WEINBERG DEVIATION
-0.013669      Hardy-Weinberg deviation (sample)
0.011976       Hardy-Weinberg deviation (population)

CHROMOSOME 1
 NIND(real sample)=20
 NSNP=10239
 NSNP_calculations=7442
 NSNP_+2alleles=0
 NSNP_zeroes=0
 NSNP_monomorphic=2797
 NIND_corrected=20.000000
 freq_MAF=0.025000
 F_dev_HW (sample)=-0.020159
 F_dev_HW (pop)=0.005485
 Genetic distances available in map file

etc.
```

This information includes information on the number of individuals analyses (NIND), the total number of SNPs of the chromosome (NSNP), those used in the calculations (NSNP_calculations) and those removed because have more than 2 alleles (NSNP_+2alleles), have non-genotyped SNPs (NSNP_zeroes), or those which are monomorphic (NSNP_monomorphic). The next line is the corrected number of individuals (NIND_corrected) considering only the SNPs analysed, and is used in the corrections for sample size. The next line (freq_MAF) is the minimum allele frequency of the SNPs analysed in the chromosome. The next one (F_dev_HW (sample)) is the average deviation from Hardy_Weinberg proportions in the sample for that chromosome. And, finally, the last one (F_dev_HW (pop)) is its corrected value for the population. A final warning is given if the genetic distances are available in the map file or they are assumed from a given cM per Mb rate.

The last part of the file shows the input used by GONE (INPUT FOR GONE), which is the average of the outfileLDn for all chromosomes,

weighting the values of each chromosome by their number of pairs of SNPs.

```
INPUT FOR GONE
2       Phase (0: pseudohaploids; 1: known phase; 2: unknown phase)
20.000000       sample size (individuals; corrected for zeroes)
-0.013669       Hardy-Weinberg deviation
48978984 0.148248 0.066297 4
60071956 0.101803 0.073132 6
35649335 0.072247 0.078820 8
22991196 0.055970 0.086197 10
32140719 0.041013 0.103651 15
16820876 0.028953 0.121714 20
10287768 0.022402 0.134635 25
6919246 0.018281 0.148388 30
4984760 0.015443 0.160851 35
3773479 0.013372 0.170688 40
2946868 0.011791 0.183412 45
2366374 0.010545 0.194279 50
1942615 0.009538 0.204093 55
…….
1474 0.000255 0.435856 1965
1458 0.000254 0.437269 1970
1497 0.000253 0.431380 1975
1461 0.000253 0.469789 1980
1473 0.000252 0.434803 1985
```

First, the phase assumed, the sample size and the overall HW deviation in the sample are given. Then, the first column shows the number of pairs of SNPs for each bin of c values, the second column gives the average c of the bin, and the third gives the average d^2. The last column would be an equivalent number of generations to which bin corresponds assuming that the generation is equal to 1/(2c). For example, the second bin:

60071956 0.101803 0.073132 6

indicates that there are 60071956 pairs of SNPs in a bin which includes pairs of SNPs with a recombination rate c between 1/8 = 0.125 and 1/12 = 0.083, with an average c = 0.101803 and an average square correlation of SNP frequencies of d^2 = 0.073132.

Although the file shows all bins, those actually included in the GONE analysis are set up by the hc option above. If hc=0.05, this means that the first bin to be considered in the estimation is that with average c just below 0.05, i.e. the bin with entrance:

32140719 0.041013 0.103651 15

The lowest bin to be considered is, by default, the last one with c > 0.001.

The Output_Ne_$FILE, e.g. Output_Ne_example, looks like this:

```
$ more Output_Ne_example
Ne averages over 40 independent estimates.
Generation     Geometric_mean
1      61.1594
2      61.1594
3      61.1594
4      61.1594
5      106.27
6      108.986
7      109.131
8      109.091
9      124.609
10      125.37
11      127.766
12      128.352
13      141.136
14      142.894
15      142.326
    ……..
```

The first column is the generation backward in time, and the second one is the geometric mean values of N_e over GONE estimation replicates.

There is also an Output_d2_$FILE, e.g. Output_d2_example, which shows the values of observed and estimated d^2 (linkage disequilibrium) values for different bins of recombination rates (c).

```
$ more Output_d2_example
Sample d2 values. Average (d2obs-d2prd)^2 = 6.34483e-07
rec.rate_c     observed_d2     adjusted_d2
0.041013      0.103651      0.104474
0.028953      0.121714      0.120067
0.022402      0.134635      0.134545
0.018281      0.148388      0.148136
0.015443      0.160851      0.160902
0.013372      0.170688      0.172843
0.011791      0.183412      0.184011
0.010545      0.194279      0.194443
……..
```

A file *timefile* shows the progress of the whole script:

```
$ more timefile
DIVIDE .ped AND .map FILES IN CHROMOSOMES
RUNNING ANALYSIS OF CHROMOSOMES
CHROMOSOME ANALYSES took 209 seconds
Running GONE
GONE run took 138 seconds
END OF ANALYSES
```

CAUTION NOTES

Population structuring

As all methods to estimate N_e, unbiased estimates require that the population is a closed one without recent admixture from other populations. If this is not the case, some artefacts may arise. For example, Figure 2F of Santiago et al. (2020) (shown below) illustrates a situation in which there are two subpopulations of N=1000 individuals between which there is a continuous gene flow of 0.2% per generation. The two subpopulations are regarded as a single one, however, and a sample of 100 individuals is analysed. When all recombination rate bins are considered in the analysis (option hc=0.5) there is a large recent increase in the estimated N_e, followed by a huge drop (black line). If the first bins are discarded (option hc=0.05; red line), the estimation is improved (N_e about 2,000), but the recent drop still persists.

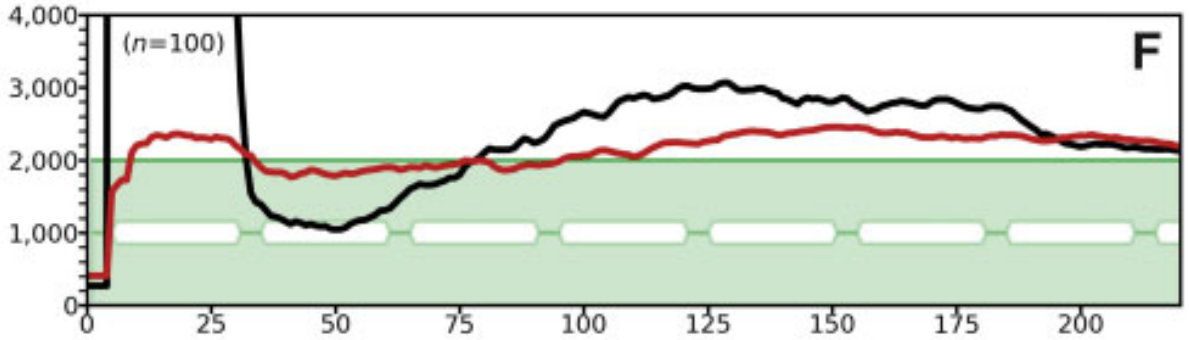

Figure 2F of Santiago et al. (2020). Mol. Biol. Evol. 37: 3642-3653.

When this type of result is observed, caution should be taken. A population structure analysis could be carried out to check if there is a population subdivision.

Number of generations for which the estimated N_e is reliable

Although the output N_e may give results for more than 600 generations, the reliable estimates to be considered should be for a maximum of 100 or 200 generations, perhaps less if the number of SNPs in the low recombination bins is low.

Genetic map.

A good genetic map is essential to obtain reliable estimates of historical N_e. If the map is not correct, perhaps it is better to disregard it and assume a constant average recombination rate (cMMb option) for the whole genome.