GONE USER´S GUIDE (29 August 2021)

OVERVIEW

This program calculates and uses linkage disequilibrium at genomic marker loci to infer the effective population size trajectories over a period of about 100-200 hundred generations back in time. It is not reliable for further generations.

CITATION

Santiago, E., Novo, I., Pardiñas, A. F. Saura, M., Wang, J., Caballero, A. (2020). Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Molecular Biology and Evolution 37: 3642–3653.

FILES NEEDED

All executable files to run the software in Linux or in Mac are available in the directories "Linux" and "MacOSX".

The input files are the usual Plink files **data.map** and **data.ped** (see below).

You need also an **INPUT_PARAMETERS_FILE** with the parameters desired and the **script_GONE.sh** to run the programme.

Copy all these files as well as the directory **PROGRAMMES** into your running directory. Make sure the executable files have the permission to be run. If not, use command "chmod r+x ExecutableFileName" to grant permission.

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.

INPUT FILES

The **data.map** has the following columns:
> 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 distance in centimorgans. If this is not available leave it as zeroes, and then give the average rate of recombination for the species (e.g. 1 cM per Mb) to use it instead as a constant rate in the INPUT_PARAMETERS_FILE
> Fourth column: The SNP genomic position, e.g. 34456.

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 ...

Thus, individual IND1 is heterozygote TA 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 IND1 0 0 1 -9 2 1 1 2 ...
1 IND2 0 0 1 -9 2 2 1 2 ...
```

The INPUT_PARAMETERS_FILE includes the parameters that can be set up:

```
PHASE=2 ### Phase = 0 (pseudohaploids), 1 (known phase), 2 (unknown phase).
cMMb=1  ### CentiMorgans per Megabase assumed if the genetic distance between markers is unknown. If the third column of the data.map file contains only zeros, this entrance is used. If not, the entrance is ignored and the distances of the data.map file are used.
DIST=1  ### none (0), Haldane correction (1) or Kosambi correction (2). Default 1.
NGEN=2000 ### Number of generations for which linkage data is obtained in bins. Default 2000 - There will be data analysed for 2000 generations, i.e. only pairs of SNPs with recombination fraction c > 1/4000 will be considered.
NBIN=400  ### Number of bins (e.g. 1000, so that each bin includes NGEN/NBIN generations). Default 400 - There will be 400 bins of pairs of SNPs, thus each with 2000/400 = 5 generations gaps. However, the first 10 generations are analysed with 2-gen gaps and the rest for 5-gen gaps.
MAF=0.0   ### Minimum allele frequency (0-1). Default 0.0 - No MAF pruning will be applied to the data.
ZERO=1    ### 0: Remove SNPs with zeroes i.e. ungenotyped SNPs (1: allow for them). Default 1 - SNPs with zero values will be considered.
maxNCHROM=-99  ### Maximum number of chromosomes to be analysed (-99 = all chromosomes; maximum number is 200). Default -99 - All chromosomes are analysed.
maxNSNP=50000 ### Approximated number of SNPs per chromosome to be analysed (maximum is 100,000; default 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 Ne to be obtained. If you do so, do not sent all runs at the same time, as the initial random seed is taken from the computer clock time.
hc=0.05   ### Maximum value of c analysed. Default and recommended 0.05. If the population has recent migrants from another population, the estimation of Ne will be biased. A typical artefact observed is a very recent drastic drop and a previous increase (see Fig. 2f of article). This can be partly corrected by using a maximum value of c lower than that recommended above, for example hc=0.01.
REPS=40   ### Number of internal replicates run by GONE. Default and recommended 40.
```

threads=-99 ### Number of threads (if -99 it uses all possible processors). Default and recommended -99.

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 maximum sample size is 1800 individuals.

TO RUN THE SOFTWARE

The ideal is to run the script in a scratch directory with qsub 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 available in the LINUX and MACOSX directories, the running command is:

```
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 214 seconds, and for GONE 141 seconds. Note that for data with many more SNPs and individuals the time of processing can be much larger.

OUTPUT FILES

A **TEMPORARY_FILES** directory is made to include all temporary files.

An OUTPUT_FILE file (e.g. OUTPUT:EXAMPLE) will show the total number of SNPs used:

```
TOTAL NUMBER OF SNPs
73240
```

And the deviation from Hardy-Weinberg proportions in the sample and in the population:

```
HARDY-WEINBERG DEVIATION
-0.007088       Hardy-Weinberg deviation (sample)
-0.002063       Hardy-Weinberg deviation (population)
```

And other information per chromosome. For example, for chromosome 1:

```
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
```

Which includes information on the number of individuals analyses (NIND=SAM), the total number of SNPs of the chromosome (NSNP), those used in the calculations and those removed because have more than 2 alleles, have ungenotyped SNPs, or those which are monomorphic. The next number is the actual number of individuals (NIND_corrected) considering only the SNPs analysed, and is used in the corrections for sample size. The next number (freq_MAF) is the minimum allelic 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), showing the number of pairs of SNPs, the average c and the average d2 for each bin.

The Output_Ne_$FILE, e.g. *Output_Ne_example*, shows the following output:

```
Ne averages over 40 independent estimates.
Generation Geometric_mean
1     64.0348
2     64.0348
3     64.0348
…
```

Where 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 d2 (linkage disequilibrium) values for different bins of recombination rates (c).

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