INSTRUCTIONS TO RUN script_GONE.sh

Recent demographic history inferred by high-resolution analysis of linkage disequilibrium.

Enrique Santiago, Irene Novo, Antonio Pardiñas, María Saura, Jinliang Wang and Armando Caballero.

This program calculates and uses linkage disequilibrium at genomic marker loci to infer the effective population size trajectories over a period of hundreds of generations (from now).

STEPS:

(0) Copy necessary files

Copy the directory **PROGRAMMES**. Make sure the executable files have the permission to be run. If not, use command "chmod r+x ExecutableFileName" to grant permission. Make sure your data files (data.ped and data.map) and the present script file are also in your working directory.

The data.ped file must have a -9 just before the genotypes, e.g. 1 IND1 0 0 1 -9 A A A A …

The ideal is to run the script in a scratch directory.

(1) The necessary executable programmes available in the directory **PROGRAMMES** are:

- MANAGE_CHROMOSOMES2 (C programme)
- LD_SNP_REAL3 (C programme)
- SUMM_REP_CHROM3 (C programme)
- GONE (C++ programme that requires gcc/7.2.0)
- GONEaverage (C++ programme that requires gcc/7.2.0)
- GONEparallel.sh (bash script)

The input parameters file is **INPUT_PARAMETERS_FILE**.

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 chromosomes.

(2) Run the bash script for the analysis

The command is:

bash script_GONE.sh <FILE> &

With the following argument:

FILE = Data file name (prefix of files .ped and .map)

In the **INPUT_PARAMETERS_FILE** there are several parameters that can be set up:

```
PHASE=2 ### Phase = 0 (pseudohaploids), 1 (known phase), 2 (unknown phase)
cMMb=1  ### CentiMorgans per Megabase (if distance is not available). For humans set cMMb=0 and distances are obtained from genetic data file
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 (e.g. 1000, 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 chromosomes)
maxNSNP=-99 ### Approx number of SNPs per chromosomes to be analysed (-99 = all SNPs)
hc=0.05   ### Maximum value of c analysed
REPS=40   ### Number of replicates to RUN GONE
threads=-99  ### Number of threads (if -99 it uses all possible processors)
```

For example, using the simulation data from example.map and example.ped

the running command can be:

bash script_GONE.sh example &

In the example there are 10 chromosomes, a total of about 100000 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.

Example of parameters:

- **PHASE=2**. The phase is not known

- **cMMb = 1**. One centiMorgan per Megabase is assumed as the genetic distance between markers is unknown.
- **DIST = 1**. Haldane´s correction is applied to obtain the genetic distances.
- **NGEN = 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**. 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**. No MAF pruning will be applied to the data.
- **ZERO = 1**. SNPs with zero values will be considered.
- **maxNCHROM=-99**. All 10 chromosomes are analysed.
- **maxNSNP=-99**. All SNPs will be used. If, for example, SNP = 10000, a maximum of 10000 SNPs per chromosome will be used. If the number of SNPs in the chromosome is lower than this number, all SNPs of the chromosome are analysed. If the value is larger, however, a random sample of 10000 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 to be obtained. If the number of SNPs in a chromosome is much larger than 100,000 the memory may be insufficient, so using this maximum is advised.
- **hc=0.05**. The maximum value of c to be analysed is 0.05.
- **REPS=40**. The number of replicates to RUN GONE is 40.
- **threads=-99**. All processors will be used in the analysis.

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

(3) The script first divides the files into chromosomes

The analysis requires the chromosomes to be analysed one by one, so first it is necessary to divide the files into chromosomes with the programme MANAGE_CHROMOSOMES2:

```
./PROGRAMMES/MANAGE_CHROMOSOMES2<<@
$maxNSNP
@
```

This will generate files chromosome1.map and chromosome1.ped for all chromosomes. A maximum number of SNPs per chromosome can be set up (maxNSNP). If this is smaller than the number of SNPs available a random sample is obtained. Every time the script is run a different random sample is obtained, allowing for different replicates of the estimates.

(4) For each chrom file the programme LD_SNP_REAL3 is run

This programme obtains the linkage disequilibrium values (d^2) for all pairs of SNPs analysed and accumulate them in bins.

The input of this programme is:

```
for ((n=1; n<=$NCHR; n++)); do echo $n; done | xargs -I % -P
$threads bash -c "./PROGRAMMES/LD_SNP_REAL3 %
$options_for_LD"
```

The output files of this programme are called *outfileLD$n* and *parameter$n*, where *$n* is the chromosome number.

The *outfileLD* file is the input of the N_e estimator programme. For example, for chromosome 1:

```
2
20.000000
-0.021697
0 0 0 2 0 0
4504119 0.147412 0.075161 4 0.000937 0.012468
6199951 0.101481 0.071979 6 0.000951 0.013206
3792367 0.072223 0.065569 8 0.000892 0.013607
……
```

The first series of numbers are:
- PHASE tag (0, 1 or 2)
- Sample size corrected for ungenotyped SNPs
- Deviation from Hardy-Weinberg equilibrium

The next columns include the linkage disequilibrium estimates for different bins (in this case, the first bin class has no elements). Only the first 3 columns are necessary for GONE, the rest are optional.

The columns are:

(1) Number of pairs of SNPS included in the bin
(2) Harmonic mean of recombination fraction (c) for the bin (already corrected by Haldane´s function if this option has been chosen)
(3) Average d^2 value for the bin (without correction for sample size; see below)
(4) Gap of the bin equivalent in generations (g = 1/2c)
(5) Average D^2 value for the bin
(6) Average Variance in allele frequencies for the bin

The *parameter* file will show the total number of SNPs used:

```
TOTAL NUMBER OF SNPs
73240
```

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=-0.021697
 G_var_bet_ind=0.489151
```

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) is the average deviation from Hardy_Weinberg proportions. And, finally, the last one (G_var_bet_ind) is the proportion of variation between individuals, which gives an idea of the heterogeneity of the sample analysed.

(5) The programme SUMM_REP_CHROM3 is run to pool results from all chromosomes

The estimation of N_e can be made for each chromosome separately, but it is more appropriate to accumulate all chromosome results in a single file. This is done by the programme **SUMM_REP_CHROM3**. The averages for each bin are made weighted by the number of pairs of SNPs in the different chromosomes. A single file called *outfileLD* and a *PARAMETERS_$FILE* (i.e. *PARAMETERS_example*) are generated. The latter gives the total number of SNPs analysed, the parameters for all chromosomes in sequence and the outfileLD file, which is the input for GONE.

(6) The estimation of temporal N_e is made by the programme GONE

The estimator programme (GONE) carries a genetic algorithm to find the best estimate of temporal N_e and has different options which can be changed in the script:

-ng : simulates the optimization for a number of generations (default 50000).
-bs : merges small bins into larger bins with at least n_pairs SNP. If not specified, the program optimizes the merging.
-bn : merges consecutive bins until the number is reduced to n_of_bins. If not specified, the program optimizes the number of bins.
-lc : lowest recombination frequency to be considered. By default is set to 0.003.
-hc : highest recombination frequency to be considered. By default is set to 0.5.
-ma : number of contiguous values of c and d2 values of input file to compute moving averages: 3, 5, 7 or 9. By default, no moving averages are carried out.
-sd : integer to seed random number generator.
-sr : Sampling with replacement. By default sampling without replacement.

GONE corrects d^2 values for sample size.

To run the programme, for example:

```
./PROGRAMMES/GONEparallel.sh -hc $hc outfileLD $REPS
```

Which uses all values by default except that the maximum c to be considered is 0.05.

Result files are appended in a directory outfileLD_TEMP in the TEMPORARY_FILES directory:

_GONE_Nebest Estimated N_e (col 2) at generations backward in time
_GONE_d2 c values (col 1), Observed d2 (col 2), and estimated d2 in the sample(col 3)
_GONE_input Input if bins are rearranged. This is the input file **_GONE_log Log file**

The output N_e file is recalled Output_Ne_$FILE, e.g. *Output_Ne_example*, which 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 estimation replicates.

There is also an Output_d2_$FILE, e.g. *Output_d2_example*, which shows the values of observed and estimated d2 values for different bins of recombination rates (c).

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

If the estimation of temporal N_e needs to be repeated with other parameters it is not necessary to run the whole script but only the GONE programme with the outfile input.

For example:

```
./PROGRAMMES/GONEparallel.sh -lc 0.001 -hc 0.2 -ng 50 -bs
1000 outfileLD 40
```

If estimates need to be obtained for a single chromosome, the outfileLD$n are available in the TEMPORARY_FILES directory. Then run GONE on the corresponding chromosome file. For example:

```
./PROGRAMMES/GONEparallel.sh outfileLD8 40
```