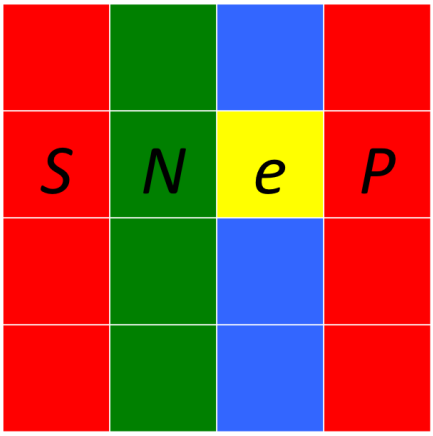

V1.1 (AUGUST 2015)

User Manual

SNeP - A tool to determine

trends in recent effective population size trajectories

using genome-wide SNP data.

1 Contents

2 Introduction

SNeP is a piece of software written in C++ that performs historical Effective Population Size (Ne) trajectories estimation through Linkage Disequilibrium (LD). Calculations are based on genome-wide genotype data.

The theory behind *SNeP* has been thoroughly investigated in several publications and therefore will not be discussed in this manual. If you are interested in the details of the theory you can refer to a few key publications that describe the theory behind *SNeP* in further detail (see Sved,1971, Hill,1981 and Corbin et al. 2012).

We strongly suggest reading the aforementioned literature or at least the latter *SNeP* paper (Barbato et al. 2015), in order to optimise the tools functionality in a case specific manner.

2.1 Disclaimer

The software is freely available for fellow researchers interested in investigating demographic trends. However, development of *SNeP* is ongoing and therefore the software is provided "as is" without warranty of any kind.

2.2 Contacts

We are more than willing to try and help you with any problems concerning the software. If you encounter a problem or want to report a bug, please write an email with detailed information on the issue experienced, the log file of the last run, a description of the data submitted to *SNeP* and if possible a screenshots or a transcription of the error to:

`barbatom@cardiff.ac.uk`

3 How to run *SNeP*

SNeP works from command line using flags to access different options, all the flags are listed and described within section 4.

Some flags require an argument, while others operate under a presence or absence behaviour, and in turn switch some of *SNeP's* capabilities on or off.

Flags and arguments have to be separated using spaces as in:

```
$ ~/SNeP -ped path/file.ped -map path/file.map -ld -svedf
```

There is no particular order to include the flags, but for those flags that require an argument the latter has to follow the relative flag.

Make sure your GCC is updated to at least version 4.8.2.

3.1 Quick start

SNeP's default parameters should fit most of the general requirements. To give *SNeP* a quick try, ensure that your .ped and .map files have the same file path and name, open your system's terminal and type the following:

```
$~/SNeP -ped filepath/filename.ped
```

If there are no problems with the input files, *SNeP* will quickly perform the analysis and create a `filepath/filename.NeAll` output file and a `filepath/filenameSNeP.log` logging file.

4 *SNeP* Options

4.1 File Control

The flag **`-help`** will output a brief description of *SNeP's* options (displayed in *italic* in this section).

4.1.1 File control flags:

`-ped file.ped`

A valid .ped file (for example: ../infile.ped)

The correct format for .ped files is described in section 5.1.

`-map file.map`

A valid .map file. If not provided SNeP will assume it has the same root file name and location as .ped file (for example: ../infile.map)

If this option is not applied, *SNeP will* look for a .map file with the same path and name of the .ped file provided.

The correct format for .map files is described in section 5.2.

`-distld filename`

Root file name of the distance-LD file

When using an already computed distance-LD file (either produced by *SNeP* or third party software) this flag should be used followed by the distance-LD file path/name. The format of this file is explained in section 4.3.

`-out filename`

Root file name of the output files

SNeP normally saves all the output files using the .ped file path and filename as a template. This flag allows the user to select a customised path and filename that will then be used for all the output files.

`-ld`

Saves a .ld file with the distance and LD value for each valid SNP comparison (Beware: can produce huge files)

This tells *SNeP* to save a copy of the distance-LD values into a .ld file with all the pairwise comparisons that have been included in the final Ne estimation. The format of the .ld file produced is described in section 5.3.

The number of pairwise comparisons can be extremely large depending on the size of the dataset used. However, the use of this flag is extremely convenient whenever the user wants to test different model or binning optimisations. The user can perform the full calculation once adding the "`-ld`" flag to the command line. Then use the "`-distld filename`" flag and different model or binning parameters to ask the software to not calculate the LD values again and just perform the Ne estimation using the computed values saved into the file "`filename`". However if the user wants to apply a different LD metric, the full analysis has to be repeated.

4.2 Input Control

4.2.1 Input control flags:

`-chr list`

Select the chromosomes to consider in the analysis (default is use all chromosomes available)

This parameter allows the user to select any kind of chromosome combinations to be analysed.

The user can select:

single chromosomes:

```
-chr 1,2,5
```

This will analyse chromosomes 1, 2 and 5

a range of chromsomes:

```
-chr 1-5
```

Which will analyse chromosomes 1, 2, 3, 4 and 5 (to use ranges of chromosomes the IDs have to be numeric)

or a combination of both:

```
-chr 1,2,4-7,14
```

will analyse chromosomes 1, 2, 4, 5, 6, 7 and 14

The chromosome list cannot contain spaces:

```
-chr 1,2, 4-7,14
```

In the former example there is a space character before "`4-7`", this will produce an error while parsing the command line.

`-mindist dist`

Minimum distance in bp between SNPs to be analysed (default is 50000)

This value cannot be larger than the maximum distance between SNPs.

`-maxdist dist`

Maximum distance in bp between SNPs to be analysed (default is 4000000)

This value cannot be smaller than the minimum distance between SNPs.

`-maf f`

Minimum allele frequency for a locus to be analysed (default is 0.05)

Sets a minimum allele frequency (MAF) and therefore removes loci with allele frequencies less than the selected value.

A MAF filtering (with a default threshold of 0.05) **is always performed by *SNeP***, if by any reason the user does not want to apply the MAF filter, by applying "`-maf 0`" to the command line, one can remove the filter.

-maxsnp s

Maximum number of SNPs per chromosome (default is 100000)

If the number of SNPs in the chromosome under analysis is larger than s, a sufficient number of SNPs is randomly removed from the analysis to reach the desired threshold.

-seed s

Seed to initiate the random engine for SNP thinning (default is random)

Allows the user to select the seed that will prompt the SNP thinning. If the flag is not used or the value s is set to 0, a random seed will be used and its value stored in the log file (section 6.4).

4.3 Model Control

The formula used to estimate Ne from LD (Corbin et al., 2012) is:

$$(1) \qquad N_{T(t)} = \frac{1}{(4f(c_t))}\left(\frac{1}{E\left[r^2_{adj}|c_t\right]} - \alpha\right)$$

Where:

$N_{T(t)}$ is the effective population size estimated t generations ago in the past

c_t is the recombination rate t generations ago in the past

r^2_{adj} is the linkage disequilibrium estimation, adjusted for sampling bias

α is a constant.

4.3.1 Model control flags:

-phased

Indicates phased data and the Hill and Robertson's estimation is used. Otherwise genotypic data is assumed and r2 estimations are based on the correlation trend method.

-alpha a

Modifies the formula alpha value (default is 1)

-recrate r

Uses r as a recombination rate (default is 1e-008)

This value modifies the ratio of conversion between physical distance and linkage distance when the recombination rate between two SNP is first estimated, read section 4.5 for more details.

-samplesize x

Adjusts r2 due to limited sample size (n). The input parameter is the x in r2[adj] = r2 - (1/xn).(default is no correction)

4.4 Binning Control

Bins are generated for the user's requirements according to the following formulae:

$$(2a) \quad b_i^{min} = minD + (maxD - minD)\left(\frac{b_i - 1}{totBins}\right)^x$$

$$(2b) \quad b_i^{max} = minD + (maxD - minD)\left(\frac{b_i}{totBins}\right)^x$$

Where b_i is the i^{th} bin of the total number of bins *totBins*, the maximum b^{max} and minimum values b^{min} for each bin are defined according to the formulae.

4.4.1 Binning control flags:

-binwidth n

Uses n as a binwidth in bp to determine bins (default is 50000)

Parameter for the fixed range binning. Not interfaced yet.

-numBINS n

Defines the number of bins in the .NeAll file (default is 30)

Defines the *totBins* parameter in formulae 2a-b.

This parameter has to be set to a positive integer value greater than 0.

-expB n

Defines the distance distribution for bins in the .NeAll file (default is 3)

Defines the *x* parameter in formulae 2a-b.

This parameter has to be set to a positive integer value greater than 0.

-itemsTH n

Defines minimum number of items for a bin to be included in the final Ne computation (default is 500)

If a bin is built with less than *n* pairwise comparisons it will not be shown in the final results file.

This parameter has to be set to a positive integer value greater than 0.

-minr2 dist

Minimum r2 for a pair of SNPs to be analysed (default is 0)

The user can decide to keep only those pairwise comparisons whose LD value is greater than dist.

This parameter cannot be set to values smaller than 0.

-maxr2 dist

Maximum r2 for a pair of SNPs to be analysed (default is 1)

The user can decide to keep only those pairwise comparisons whose LD value is smaller than dist.

This parameter cannot be set to values larger than 1.

4.5 Recombination modifiers

SNeP first infers the recombination rate between a pair of SNPs considering the relation between physical distance (δ) and linkage distance (d) as directly proportional:

$$\delta = kd \quad (3)$$

Where k has a default value of 10^{-8} (this value can be modified using the flag "`-recrate r`", section 4.3.1).

The user can choose to use the approximation $d = c$ or to apply one of the mapping function provided in section 4.5.1.

4.5.1 Recombination modifiers flags:

`-sved`

uses Sved(1971) as recombination rate modifier

$$c = d\frac{1-\left(\frac{d}{2}\right)}{(1-d)^2} \quad (4)$$

Uses Sved's approximation (Sved, 1971).

`-kosambi`

uses Kosambi as recombination rate modifier

$$c = \frac{\tanh 2d}{2} \quad (5)$$

Uses Kosambi's mapping function (Kosambi, 1943).

`-svedf`

uses Sved & Feldman(1973) as recombination rate modifier

$$c = d\left(1 - \frac{d}{2}\right) \quad (6)$$

Uses Sved & Feldman approximation (Sved & Feldman, 1973).

`-haldane`

uses Haldane as recombination rate modifier

$$c = \frac{1-e^{-2d}}{2} \quad (7)$$

Uses Haldane mapping function (Haldane, 1919).

4.6 Optimization

SNeP can use multiple processors to perform heavy load calculations. This can be extremely time saving when large datasets are analysed. The user can allow *SNeP* to use multiple cores, if no option is specified *SNeP* will use a single processor for heavy duty computations, read section 4.6.1 for more details.

4.6.1 Optimisation flags:

`-threads t`

Uses t threads to perform the analysis (t must be an integer number, default is 1)

If *t* is larger than the number of threads available in the machine (*T*) minus one, *SNeP* forces *t = T - 1*.

5 Input files

SNeP uses a standard PLINK ped/map combination as input files.

A detailed description of ped and map file format can be found at these URLs:
http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped
http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#map

5.1 Ped file (filename.ped)

The .ped format implies a file with no header and one line per individual.

Each line has to follow this pattern:

```
ID-1  ID-2  0     0     0     -9    A     T     C     G     A     A
```

…

- First column: a textual string mostly used to store Family/Breed information, ignored by *SNeP*.
- Second column: a textual string mostly used to store individual information, ignored by *SNeP*.
- Four columns with alphanumerical values to identify parentage and phenotype status, ignored by *SNeP*.
- Two columns for each locus, the first with the 1st allelic variant, the second with the 2nd allelic variant.

All the columns are separated by spacers, either spaces or tabs are allowed, the chosen spacer has to be consistently used throughout the file.

Allelic variants have to be coded using the `A`, `T`, `C` and `G` nomenclature, in capital letters; missing data should be coded as `0`.

Loci with missing data or coded with different characters than `ATCG` are automatically excluded from the analysis.

5.2 Map file (filename.map)

The map file complements a .ped file, storing loci information.

It has no header and contains one line for each locus.

Each line follows this pattern:

```
23    SNP1  0.1   1000000
```

- First column: chromosome identifier
- Second column: locus identifier
- Third column: linkage distance in Morgans (M).
- Fourth column: physical distance in base pairs (bp).

All the columns are separated by spacers, either spaces or tabs are allowed, the chosen spacer has to be consistently used throughout the file.

5.3 Ld file (filename.ld)

This file stores information regarding the LD values for each pairwise comparison.

It has a header with the following information:

```
CHR   dist(bp)    r2
```

The header is mandatory and meant for readability but its content is ignored by the software.

Following the headers, is one line for each pairwise comparison with the following information:

1. Chromosome in which the SNP pair resides.
2. Distance between the SNPs in base pairs.
3. Linkage disequilibrium value.

The three columns are tab delimited.

6 Output files

6.1 NeAll file (filename.NeAll)

This is the main output file that *SNeP* produces.

It holds 6 fields divided in tab-delimited columns, each line represents one bin and the header titles are:

`GenAgo`: average Generations Ago for that bin

`Ne`: Ne estimate for that bin

`Dist`: average distance between the loci used to build that bin

`r2`: average LD for that bin

`r2SD`: LD Standard deviation for that bin

`items`: number of pairwise comparisons that contributed to the estimates for that bin

This output file can be used to plot both the historical Ne trajectories against generations in the past or the LD decay against loci distance.

6.2 NeChr file (filename.NeChr)

This output file has been implemented but not interfaced yet.

6.3 Ld file (filename.ld)

This file follows the same format rules as in section 4.3.

6.4 Log file (filenameSNeP.log)

This file is produced by *SNeP* as a reminder of the options setup used for a specific analysis.

For each different path/filename.ped submitted to analysis, *SNeP* produces a new path/filenameSNeP.log file, but for two analyses on the same path/filename.ped the log information are appended into the same file.

7 How to cite *SNeP*

In publications which use results obtained in part using *SNeP*, please cite:

Barbato M, Orozco-terWengel P, Tapio M and Bruford MW (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* **6**:109. doi: 10.3389/fgene.2015.00109

8 Acknowledgments

Logo for *SNeP* designed by Dr. David WG Stanton.

Thanks to Dr. Isa-Rita Russo and Luke Chrimes for useful comments on v1.0 of this user manual.

9 References

Corbin, L. J., Liu, A. Y. H., Bishop, S. C., & Woolliams, J. A. (2012). Estimation of historical effective population size using linkage disequilibria with marker data. *Journal of Animal Breeding and Genetics = Zeitschrift Für Tierzüchtung Und Züchtungsbiologie*, *129*(4), 257–70. doi:10.1111/j.1439-0388.2012.01003.x

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, *8*, 299–309.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, *38*(03), 209–216. doi:10.1017/S0016672300020553

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenetics*, *12*(1), 172–175.

Sved, J. a. (1971). Linkage Disequilibrium and Homozygosity of Chromosome Segments in finite Populations. *Theoretical Population Biology*, *141*(2), 125–141. doi:10.1016/0040-5809(71)90011-6

Sved, J. a, & Feldman, M. W. (1973). Correlation and probability methods for one and two loci. *Theoretical Population Biology*, *4*(1), 129–132. doi:10.1016/0040-5809(73)90008-7