NeON Tutorial

Introduction

This tutorial provides hints and examples about how to use the NeON package.

Before using NeON functions

NeON R functions require PLINK v1.07 [1] binaries into the working folder; we provide compiled binaries for both MacOSX and Linux environment.
To test whether the PLINK program works try totype *./plink* from the bash.
Theexamples in this tutorial require also a "genetic_map" and a"data" folder containing respectively the recombination maps and the data examples. These two folders have to be located in the working directory

Install package

Download the add-on R package and type the following command

```
$ R CMD INSTALL NeON_1.0.tar.gz
```

After that, in R environment, load the NeON library typing

```
>library(NeON)
```

Preparing input files

NeON requires input files in PLINK binaryformat. For each population three files are expected (.bed, .bim, .fam extensions); look at the PLINK website for further information.
To obtain information about the level of Linkage Disequilibrium, it is necessary to know the exact genetic map position for each marker (usually indicated in the 3rd column of the.bim file).
If this information is not already contained in the data files, it can be updated by means of the NeON functions. Within the package we provide a method to obtain the genetic map information from the recombination rates and hotspots maps retrieved from hapmap website (http://hapmap.ncbi.nlm.nih.gov)

Example

As example, we consider the CEU population from the CEPH panel; the files in PLINK format are located into the *data* folder.

The genetic map information files have to be stored in the *genetic_map* folder, a single map file for each chromosome, each file having a similar prefix (e.g *genetic_map_b36_chr1.txt*, *genetic_map_b36_chr2.txt* and so and so forth).

The header of each file has to be as follow:

Position(bp) COMBINED_rate(cM/Mb) Map(cM)

(It is really important to check that the genetic map match the release of your data; within the package we provide properly formatted recombination map files for the two last releases of human variation data, i.e. hg18 and hg19)

At this point we could update the genetic position of the markers using the function *Nemap* and *NeUpdate*.

Nemap function prepares the file to update the genetic map information of the markers in your dataset, basing on the recombination rates and hotspots compiled file (that can be downloaded from the HapMap website) of the same release of your data (e.g. NCBI36/hg18 and GRCh37/hg19), and extract the genetic map information by matching the physical positions of the SNPs contained in the two files (the dataset and the recombination map). *Nemap* returns a list of SNP identifiers (snp.list), that can be used by the following function, *NeUpdate*, to actually update the genetic map information. The PLINK executable has to be in the same folder of the data files.

```
>library(NeON)
>snplist<-Nemap("data/hapmap-ceu-NeON.bim","genetic_map/genetic_map_b36_chr")
>write.table(snplist, "data/snplist.txt", row=F, col=F, quote=F)
>NeUpdate(plink.file="data/hapmap-ceu-NeON", snp.list="data/snplist.txt",
outfile="data/CEU_gen")
```

The new PLINK file CEU_gen.bim now contains the genetic position for each of the markers. If some markers are not present in the recombination map files (i.e. their genetic map position is missed), they are no more considered. Indeed, the *Neupdate* function runs the PLINK program to extract from you dataset just the markers for which an update of the genetic map is possible (generating new CEU_gen.bed, CEU_gen.bim and CEU_gen.fam files).

Linkage disequilibrium estimation

The *NeLD* function estimates the squared correlation coefficient of linkage disequilibrium (r^2_{LD}) between markers. The default parameters of the function are a genotyping rate higher than 98% (geno=0.02), a rate of individual missing data lower than 10% (mind=0.9), a window of 500 kilobases (ld.window.kb=500) and 9999 SNPs (ld.window=9999). For further information about these parameters see the PLINK website.

Example

CEU_gen.bedCEU_gen.bimCEU_gen.fam are the PLINK files obtained with the functions *NeMap* and *NeUpdate*.

```
>library(NeON)
>NeLD(plink.file="data/CEU_gen", outfile="data/CEU_gen", geno=0.02, ld.window.kb=500, ld.window=9999)
```

The output file is CEU_gen.ld, which contains the r^2 information.

Effective population size estimation

The function *Nestimate* allows to estimate the effective population size from linkage disequilibrium. *Nestimate* is based on the method described in Tenesa [2] and McEvoy [3].

Example

```
>library(NeON)
>samplesize<-read.table("data/CEU_gen.fam")
>samplesize<-length(samplesize$V1)
>myNe<-Nestimate(file.ld="data/CEU_gen.ld",sample.size=samplesize, min.R2=0.001, max.R2=0.999, method="MG", min.cfr=5)
>write.table(myNe, “data/CEU_Ne_output.txt”, row=F, quote=F)
```

Nestimate function estimates the effective population size. It requires the output.ld obtained from the *NeLD* function and applies the well-known formula Ne≈ 1/(4c) * [(1/r2) - 2], where c is the distance between genetic markers in Morgan. *Nestimate* creates several categories of recombination distance, with incremental upper boundaries of 0.005 cM up to 0.25 cM, and calculates the r^2_{LD} for each pairs of markers in each recombination distance category. To do this, we implemented two different methods: one (method="Mcevoy") is the same method that has been used in McEvoy et al 2011, with 50 not overlapping bin from 0.005 up to 0.25 cM, the other (method="MG", the default) is the Mezzavilla-Ghirotto method, which consider 250 overlapping bins with a step of 0.001 cM from 0.005 to 0.25 cM. *Nestimate* calculates a value of effective population size, according to the formula above, within each of the 50 or 250 identified bins.

Two functions estimate the long-term Ne and the demographic function (i.e the effective population size over time), together with the confidence interval (see the main paper for further details). These functions are respectively *Ne_CI* for the long-term and *Ne_Med* for the demographic function.

Example

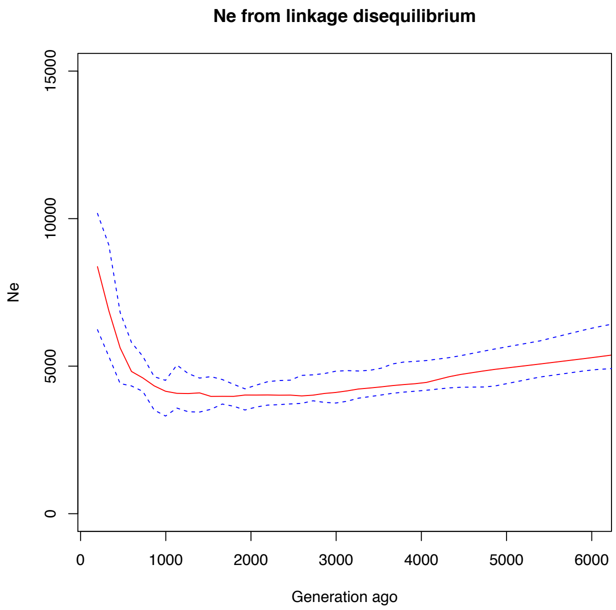
```
>library(NeON)
>myNe=read.table("data/CEU_Ne_output.txt",h=T)
>myCI<-Ne_CI(myNe, ci = c(0.05, 0.5, 0.95))
>write.table(myCI, "data/CEU_CI_output.txt",quote=F,row=F)
>myMed<-Ne_Med(myNe,method = "MG", ci=TRUE, ci.int = c(0.05, 0.5, 0.95))
>write.table(myMed, "data/CEU_Med_output.txt", row=F, quote=F)
```

Ne_Med function calculates the demographic function of the population, users must specify the method that has been used to estimate Ne ("MG" or "McEvoy"); if ci=TRUE confidence intervals are also calculated.

Plotting functions

The function *Neplot* allows the user to plot the output of the *Ne_Med* function.

```
>library(NeON)
>pdf("./data/NeON_plot.pdf")
>Neplot("data/CEU_Med_output.txt", approx=TRUE, ylim=c(0, 15000), xlim=c(200, 6000), main="Ne from linkage disequilibrium", xlab="Generation ago", ylab="Ne", ci=TRUE)
>dev.off()
```



This figure represents the output of *Neplot* function

Users can customize the parameters of the plot. If ci=TRUE the confidence interval of the effective population size is plotted (dotted blue lines) along with the median Ne values (red line); if ci=FALSE the plot shows just the median Ne values. We set the interpolation between temporal points equal to TRUE as default (approx=TRUE).

Estimation of divergence time

Using the *Tdverg* function is it possible to estimate the divergence time between populations, as reported in McEvoy [3].

Tdverg requires a matrix of pairwise F_{ST} between populations and a text file with a list of the long-term Ne for each population, with an header that match the population labels reported in the F_{ST} matrix; an example is reported into the data folder.

Example

```
>library(NeON)
>myTime<-Tdverg(Fst="data/Fst_3pops.txt", All_H="data/Ne_3pops.txt")
>write.table(myTime, "data/NeON_Time.txt", quote=F)
```

The output of the function is a matrix where each value represents the divergence time of a specific pair of populations. To visualize the evolutionary relationships among populations, it is possible to calculate an unrooted UPGMA from the divergence time matrix using the function upgma of the phangorn R package [4]:

```
>library(phangorn)
>plot.phylo(upgma(myTime), type="unrooted")
```

Contacts

Massimo Mezzavilla email: massimo.mezzavilla@burlo.trieste.it
Silvia Ghirotto email: ghrslv@unife.it

References

[1]Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575

[2] Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., &Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome research*, *17*(4), 520-526.

[3]McEvoy, B. P., Powell, J. E., Goddard, M. E., &Visscher, P. M. (2011). Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research*, *21*(6), 821-829.

[4] Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592-593.