Journal of
Computer Science & Systems Biology

Mezzavilla and Ghirotto, J Comput Sci Syst Biol 2015, 8:1
DOI: 10.4172/jcsb.1000168

Research Article Open Access

Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS

Massimo Mezzavilla[1,2] and Silvia Ghirotto[3]*

[1]*Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy*
[2]*University of Trieste, Italy*
[3]*Department of Life Sciences and Biotechnologies, University of Ferrara, Italy*

Abstract

Objective: Estimating the effective population size (*Ne*) is crucial to understanding how populations evolved, expanded or shrunk. One possible approach is to compare DNA diversity, so as to obtain an average Ne over many past generations; however as the population sizes change over time, another possibility is to describe this change. Linkage Disequilibrium (*LD*) patterns contain information about these changes, and, whenever a large number of densely linked markers are available, can be used to monitor fluctuating population size through time. Here, we present a new R package, NeON that has been designed to explore population's *LD* patterns to reconstruct two key parameters of human evolution: the effective population size and the divergence time between populations.

Methods: *NeON* starts with binary or pairwise-LD *PLINK* files, and allows *(a)* to assign a genetic map position using HapMap (NCBI release 36 or 37) *(b)* to calculate the effective population size over time exploiting the relationship between Ne and the average squared correlation coefficient of *LD* (r^2*LD*) within predefined recombination distance categories, and *(c)* to calculate the confidence interval about *Ne* based on the observed variation of the estimator across chromosomes; the outputs of the functions are both numerical and graphical. This package also offers the possibility to estimate the divergence time between populations given the Ne values calculated from the within-population *LD* data and a matrix of between-populations F_{ST}. These routines can be adapted to any species whenever genetic map positions are available.

Results and Conclusion: The functions contained in the R package *NeON* provide reliable estimates of effective population sizes of human chromosomes from *LD* patterns of genome-wide SNPs data, as it is shown here for the populations contained in the CEPH panel. The *NeON* package enables to accommodate variable numbers of individuals, populations and genetic markers, allowing analyzing those using standard personal computers.

Keywords: R package; Effective population size; Divergence time; Demographic parameter; Linkage disequilibrium; Recombination map; SNPs panel; Polymorphism data

Introduction

The effective population size (*Ne*) is at the same time one of the most important parameters of natural population, and one of the most difficult to evaluate directly [1,2]. Common approaches to estimate *Ne* involved temporal methods [1,3,4] that require at least two samples, separated in time, of the same population. Other single-sample methods are based on the heterozygote excess [5,6], on the amount of linkage disequilibrium in neutral, unlinked loci [7], or on measures of the extent of current genetic variation [8,9]. Recently, the considerable progresses in the field of population genetics, along with the development of methods based on the coalescent theory, have allowed to estimate the effective population size through time directly from a sample of gene sequences [10], or entire genomes [11,12]. Another way to study past populations dynamics exploits the information contained in the pattern of linkage disequilibrium (the non-random association between genetic loci, *LD* hereafter) between densely spaced single nucleotide polymorphisms (SNPs) data. The *Ne* that is usually calculated from genomic variation represents an estimate of the long-term *Ne*, that is an average of the effective population size over many past generations, disregarding of past demographic fluctuations. By contrast, the extent and the strength of linkage disequilibrium between two genetic loci contains information about population dynamics such as changes in the effective population size through time [13,14]. Indeed, levels of *LD* increase due to random genetic drift and decays due to recombination, according to a recombination rate between pairs of genetic markers that are positively correlated with the distance between genetic markers. This means that *LD* decreases at increasing physical distance between loci. Consequently, if we consider that levels of *LD* depend on both *Ne* and on the recombination rate between markers [14], *LD* between loci separated by large distances along the chromosome reflects relatively recent *Ne* whereas *LD* over short recombination distances depends on relatively ancient *Ne* [13]. This relationship between *LD* and *Ne*, detailed in the further section, can be exploited to monitor fluctuating population size through time.

As well as offering a gold opportunity to follow the dynamics of demographic events, in addition the estimation of *Ne* from *LD* can be used to date the time since two populations diverged from one another.

Materials and Methods

All the functions contained in the *NeON* package have been developed for the free statistical R environment (http://www.r-project.org) and run under the major operating systems (UNIX and OSX).

***Corresponding author:** Silvia Ghirotto, Department of Life Sciences and Biotechnologies, University of Ferrara, Italy, Tel: (+39) 0532-455312; Fax: (+39) 0532-249761; E-mail: ghrslv@unife.it

Received November 29, 2014; **Accepted** December 17, 2014; **Published** January 07, 2015

Citation: Mezzavilla M, Ghirotto S (2015) *Neon:* An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. J Comput Sci Syst Biol 8: 037-044. doi:10.4172/jcsb.1000168

Most of the *NeON* functions interact with the *PLINK* program [15] or relate with *PLINK* data files, a widely used data format in population genetic studies. The workflow of a complete *NeON* analysis consists of six steps as shown in (Figure 1). A detailed description of the functions available in *NeON* is reported below.

Nemap *(bim.file, map.file)*

Since the method implemented here to estimate *Ne* and divergence time is based on the recombination or genetic distance between SNPs, it is fundamental to have available correct genetic map information. *Nemap* actually prepares the file to update the genetic map information of the markers in your *PLINK* data file (binary format), based on the recombination rates and hotspots compiled file that can be downloaded from the HapMap website http://hapmap.ncbi.nlm.nih.gov/downloads/recombination (e.g. NCBI36/hg18 and GRCh37/hg19). The genetic map information is then extracted 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. *Nemap* requires as arguments the name of the .bim data file and the path to the recombination rates and hotspots compiled files (a single file for each chromosome), along with the prefix used in each file name before the chromosome number, paying attention to use the map that matches the build of your data. We provide the properly formatted recombination map files for the two last releases of human variation data, i.e. hg18 and hg19. Given so, to map your bim file to hg18 the right call of the *Nemap* function would be Nemap ("./mydata.bim","./genetic_map/genetic_map_b36_chr"), whereas to get genetic map information for hg19 it would be Nemap ("./mydata.bim","./genetic_map/genetic_map_GRCh37").

Ne Update *(plink.file, snp.list, outfile)*

This function relies on the snp.list file created by the previous function (*Nemap*) to update your *PLINK* .bim data file with the correct genetic map information. *NeUpdate* requires as arguments the prefix of your *PLINK* data files (i.e. without the .bim, .bed or .fam extension), the name of the file obtained from the *Nemap* function, and the prefix of the updated *PLINK* data files that will be created. The *PLINK* executable has to be in the same folder of the data files.

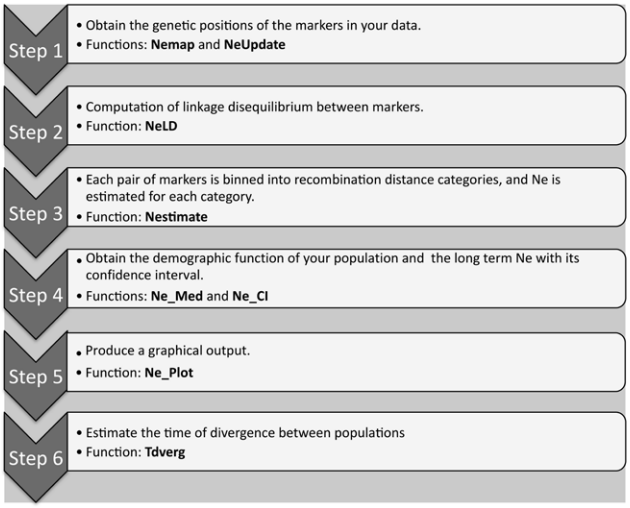

Figure 1: Workflow of a complete NeON analysis.

Ne LD *(plink.file, geno=0.02, mind=0.9, ld.window.kb=500, ld.window=9999, outfile="output.ld")*

This function estimates the squared correlation coefficient of linkage disequilibrium (r^2_{LD} [16]) 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). These parameters are also detailed in the *PLINK* tutorial online at http://pngu.mgh.harvard.edu/~purcell/plink/ld.shtml#ld1, and can be changed to fit your purpose. Other than these parameters, *NeLD* requires the prefix of your *PLINK* data files (i.e. without the .bim, .bed or .fam extension) and the name of the output file. The *PLINK* executable has to be in the same folder of the data files.

Nestimate *(file.ld, sample.size, min.R2=0.001,max.R2=0.999, method="MG",min.cfr=5)*

This function estimates the effective population size. It requires the *output.ld* obtained from the *NeLD* function and applies the formula $Ne \approx 1/(4c) * [(1/r^2)-2]$, where *c* is the distance between genetic markers in Morgan. *Nestimate* creates several categories of recombination distance, with incremental upper boundaries of 0.005 centiMorgan (cM) up to 0.25 cM, and calculates the r^2_{LD} for each pair 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 [17], with 50 not overlapping bin sizes 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. The *Ne* value calculated in each bin corresponds to the effective population size at a specific moment in the past, i.e.*1/(2c)* generation ago [13], with *c* calculated as the mean value in each recombination distance category. Other parameters of the function are: *sample size*, that is the size of your sample to allow the r^2 value to be corrected according to the formula $r^2=r^2 -1/n$ (where *n* is the sample size); *min.R2* and *max.R2* that are the minimum and the maximum r^2 allowed for the *Ne* estimation (very high and very low r^2 values, e.g. equal to 0 and 1, may lead to untreatable results), and *min.cfr* that is the minimum number of comparisons in each recombination category to allow the bin to be considered. *Nestimate* returns a data frame with the values of the effective population size and the correspondent time in the past (in generations), for each bin, for each chromosome.

Ne_CI (Nestimate.output,ci=c(0.05,0.5,0.95))

This function estimates the long-term *Ne* and its confidence intervals. The long term *Ne* is calculated as the harmonic mean [18] of the effective population sizes along the generations in the past. The confidence interval of the long term *Ne* is calculated using each chromosome as a replicate (default 5th, 50th and 95th percentile of the distribution of the *Ne* over each chromosome). *Ne_CI* requires as input the output of the previous function (*Nestimate*).

Ne_Med *(Nestimate.output, method="MG", ci=FALSE, ci.int=c (0.05,0.5,0.95))*

This function calculates the demographic function (effective population size over time) of a population along with its confidence interval, calculated as above, for each bin. *Ne_Med* requires as input the output of the *Nestimate* function and the method used to bin the data in recombination distance categories ("McEvoy" or "MG", MG as

default). The default confidence interval is the 90%; once again, it can be modified changing the values of the *ci.int* parameter of the function. This function returns a data frame with the first three columns indicating the quantiles of the distribution of the effective population size for each bin over all chromosomes (the default values are the 90% confidence interval and the median value) and the last column with the moment in time to which the effective population size is referred (*1/2c* generation ago).

Ne_Plot (Ne.file, approx=TRUE, ylim=c (0, 15000), xlim=c (200, 6000), main="Ne from linkage disequilibrium", xlab="Generation ago", ylab="Ne",ci=TRUE)

This function is useful to obtain a graphical representation of the changes in the effective population size of a population over time. *Ne_Plot* takes as input the data frame with the effective size and temporal information obtained from the *Ne_Med* function for each bin, and plots the demographic function of the population. We indicated some default values for the standard R graphical parameters; obviously they can be modified, if needed.

Tdverg (Fst,All_H)

This function returns a matrix of the time of divergence between populations in generation following: $T=ln\ (1-F_{ST})/ln\ (1-1/2Ne)$. *Tdverg* requires a matrix of pairwise F_{ST} between populations (*Fst*) and a text file with a list of the long-term *Ne* for each population, with header that match the population labels reported in the F_{ST} matrix (*All_H*).

We validated the efficiency of the *Tdverg* function in correctly estimating the time of split between two populations using a forward simulation-based approach through the python library simuPOP [19]. The basic model is a scenario in which an ancestral population give rise to two different lineages evolving independently for 2000 generations. Starting from this model, we simulated four different scenarios in which: 1) the two daughter populations have the same effective size (i.e. 5,000 individuals), 2) one population underwent a bottleneck reducing his size to 2,500 individuals 3) the two populations have the same effective size but including bi-directional migration, and 4) one population underwent a bottleneck (as in case 2) and including bi-directional migration (Figure 2A). As for the scenarios with migration [3,4], we considered the two populations exchanging one migrant per generation; this value has been used as the lowest gene flow to avoid panmixia [20]. A detailed description of the parameters used in the simulations is reported in (Table 1). Each scenario was replicated 100 times, and we evaluated the power in the divergence time estimation every 100 generations sampling 25 individuals per population.

Results

To show how to perform a complete analysis using *NeON*, we analyzed the populations contained in the CEPH panel [21]. Since the method implemented in *NeON* to calculate *Ne* relies on *LD* patterns, to avoid any bias from small sample size, we decided to consider only the CEPH populations with a sample size >20 individuals. We started from the rough data, represented by the *PLINK* binary files that can be downloaded at http://www.hagsc.org/hgdp/files.html. The map files we used to build the genetic map were downloaded at http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/ and needed a little editing to

Scenario	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Ne ancestral	10000	10000	10000	10000
Ne1	5000	2500	5000	2500
Ne2	5000	5000	5000	5000
Sample size	25	25	25	25
N SNPs	2200	2200	2200	2200
N chromosome	22	22	22	22
SNP mutation rate	2.00E-08	2.00E-08	2.00E-08	2.00E-08
Recombination rate between SNPs	0.01	0.01	0.01	0.01
Migration	0	0	1 migrant per generation	1 migrant per generation

Table 1: Simulation parameters used in the forward simulations.

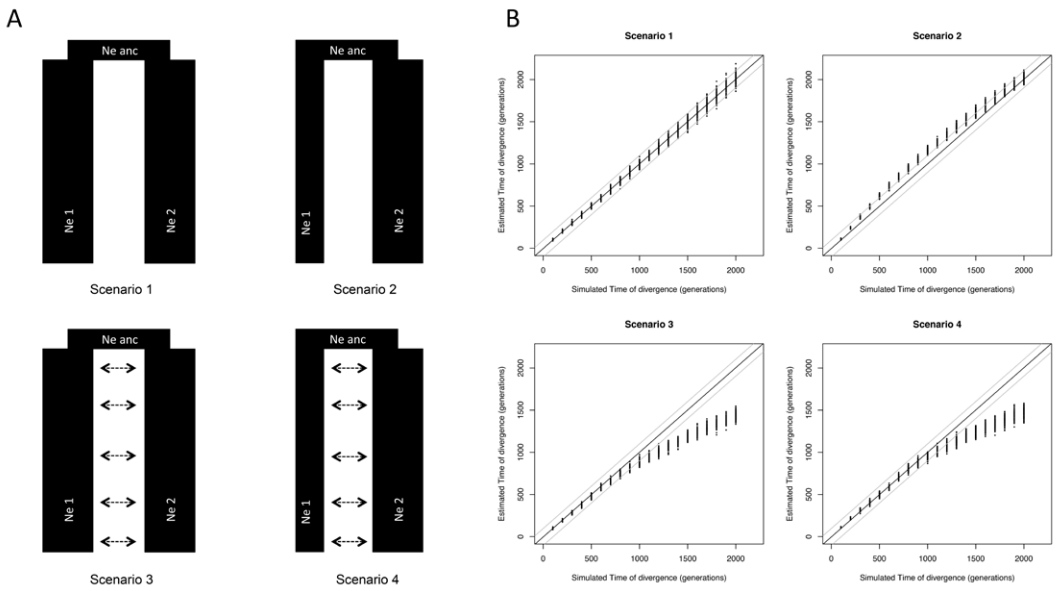

Figure 2: Forward simulation analisys. A) Simulated scenarios B) Results of the simulations; the black line represents the real values, the grey lines represent a range spanning 200 generations around the real value and the dots represent the values we estimated.

fit the requirement of the *Nemap* function: the recombination map has to be in a single folder, with separate files for each chromosome, and each of these files needs to be structured with five columns with an header (Chromosome; Position (bp); Rate (cM/Mb); Map (cM)). It is fundamental to pay attention to consider the genetic map corresponding to the correct release of the data. Within the package we already provide the edited genetic map corresponding to the human genome reference NCBI36/hg18 and GRCh37/hg19. If your .bim file already contains information about the genetic map, you should skip this step and proceed with the effective population size estimation using *NeLD* and *Nestimate*. For each population we ran *NeLD* function that exploit the *PLINK* program to estimate the strength of the linkage disequilibrium between markers (r_{LD}^2) and to arrange the data in a specific output file for the subsequent analysis (*output.ld*). Through *Nestimate*, the r_{LD}^2values were binned into distance categories, averaged, and related to *Ne* as $E(r_{LD}^2) \sim 1/(2 + 4Nec)$, where c is the genetic distance between loci in Morgans [14]. The so calculated *Ne* for each bin corresponds to an estimate of the effective population size~*1/(2c)* generations ago [13]. This function returns a dataframe with the estimate of *Ne* for each bin for each chromosome separately, along with the time to which the estimate is referred (in generations). For each bin, and hence for a specific moment in the past, a single estimate of the value of *Ne* is obtained using the *Ne_Med* function, that calculates the median of the distribution of the *Ne* estimates across all chromosomes, together with user specified quantiles of this distribution (default: 0.05-0.95). The demographic functions describing the variation of the effective population size through time were visualized by means of *Ne_Plot*, a function that takes the output of *Ne_Med* and plots the demography of the corresponding population. Figure 3 shows the functions of the effective population size through time for the CEPH populations as resulting from *Ne_Plot*. The x-axis represents the time (in generations) from the present (on the left) to the past (on the right). The time depth depends on the minimum distance categories chosen to bin the data, here corresponding to the default value of 0.005-0.01 cM (~6,500 generation ago). This range of values also represents the lower boundary allowed by our method, since smaller marker distances may have been particularly affected by gene conversion, for which the presented method does not account for [14]. The *y-axis* represents the effective population size values. It is possible to follow the demographic history of a population going from left to right, with solid lines representing the median value of effective population size over all chromosomes in each bin, and dotted lines representing the value of the 5th and 95th quantile of the distribution of the effective population size over all chromosomes in each bin. As it is shown in (Figure 3) the demographic history of populations clearly reflects their geographical localization: African populations have remained stable and larger over time (with a slight decrease in recent times, especially for Biaka Pygmies), whereas most non-African populations start to expand around 1,000 generations ago (corresponding to 25,000 years ago, considering a generation time of 25 years). This is particularly evident for European (e.g. French) and Asian (e.g. Han) populations. Other populations, like Yakut in Siberia or Maya in South America, show relatively low and constant population sizes over time. With the *Ne_CI* function we calculated the long term *Ne* for each considered population, along with its confidence interval estimated as the harmonic mean of *Ne* at the 5th and 95th quantile of the distribution of the effective population sizes for each bin. Figure 4 shows the long term *Ne* for the considered populations. The range estimated spans from 10,000 (Africans) to 4,000 (Maya and Yakut), consistent with previous estimates obtained by different methods [22-24]. We also compared, by means of a Mantel test, our CEPH *Ne* estimates with those reported in a previous study [25] for the same populations, obtaining a correlation coefficient of 0.946 (p-value<2.2e-16).

As well as offering the possibility to study fluctuating population size across time and space, *LD* pattern can be used to explicitly date population divergence times (*T*). Under neutrality, the level of population differentiation is determined by genetic drift, the extent of which depends on *Ne* and on the time since the populations diverged. Having an estimate of *Ne*, and knowing the amount of differentiation (measured by F_{ST}) between a pair of populations, it is possible to estimate their separation time in generations according to the formula $T=ln(1-F_{ST})/ln(1-1/2Ne)$ [26] embedded in the *Tdverg* function. We estimated the pairwise Weir and Cockheram F_{ST} [27], using the software 4P [28], available online at (www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software). 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, we calculated an unrooted UPGMA from the divergence time matrix exploiting the *upgma* function of the *phangorn* R package (Figure 5). From the tree it is clear that separations happened more recently for populations from the same geographical area (e.g. between Central South Asian, European, and East Asian populations), whereas a long branch (namely a longer separation time) separate Africans from non-African populations. The Mozabite, a population from North Africa, falls next to Near East populations (Druze, Palestinian and Bedouin), highlighting a genetic resemblance already reported in previous studies [29]. A distinctive pattern of separation arises within Africa, with Biaka Pygmy that separated in ancient time from Yoruba and Mandenka. This is interesting because, even thought Mandenka, Yoruba and Biaka Pygmy experienced quite different historical dynamics and lifestyle, when compared with other worldwide populations, they cannot be genetically distinguished from each other [29]. The separation pattern that is shown here for African populations depicts what have already been reported in previous works using simulations methods [30-32]. The divergence times we obtained are also in agreement with what have been estimated by Gronau et al. [11]; they estimated indeed a separation time between Europe and Africa of 38-64 Kya (our estimate is ~62 Kya) and between Europe and Asia of 31-40 Kya (our estimate is ~36 Kya).

The results of the simulation framework we developed to test the power of this method in correctly estimating the divergence time between populations are shown in (Figure 2B). In general, the forward simulations show that the method implemented in *NeON* exhibit a reasonably good power in estimate the real time of divergence. This is particularly true when the two populations have the same size through time (scenario 1); in this case indeed all the estimates (dots) fall in a range spanning 200 generations (grey lines) around the real value (black line), with an extremely precise estimation for recent splits. When one of the two populations experienced a bottleneck (scenario 2) the divergence time is a bit overestimated, except for recent and ancient separations (namely below 500 and above 1700 generations). On the contrary, when migration is taken into account (scenarios 3 and 4) the divergence time is underestimated, but only for splits more ancient than 1,000 generations.

Discussion

Human effective population size represents the average effect of drift across generation and so it is related with the level of population differentiation, and allows one to understand how populations evolved through time, whether they expanded or experienced drastic

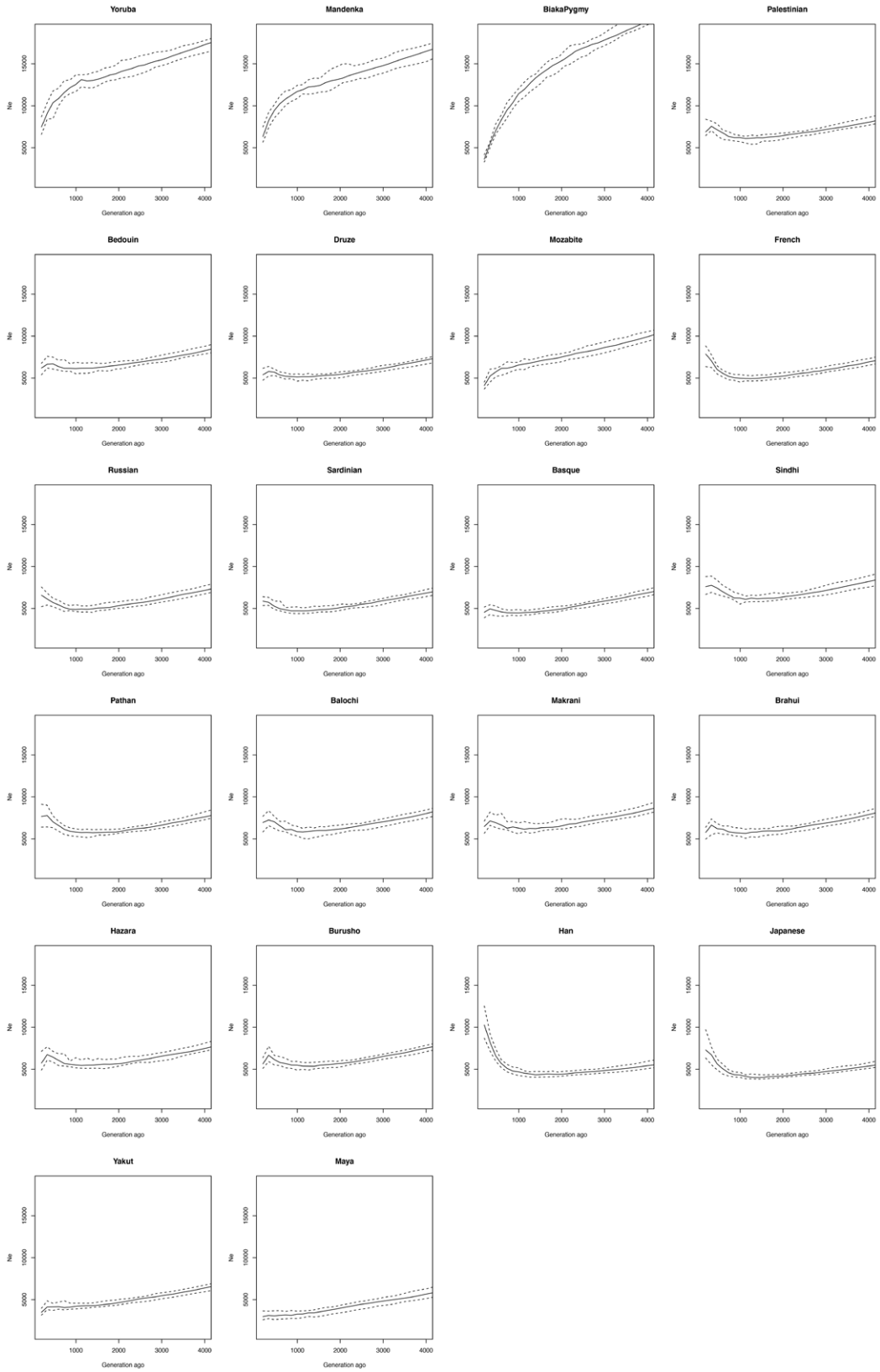

Figure 3: Plots of the effective population size trough time. The x-axis represents the time (measured in generations) from the present (on the left) to the past (on the right). The y-axis represents the effective population size values. The continuous lines correspond to the median values of the *Ne*, dashed lines correspond to the 5th and 95th percentile of the *Ne* distribution.

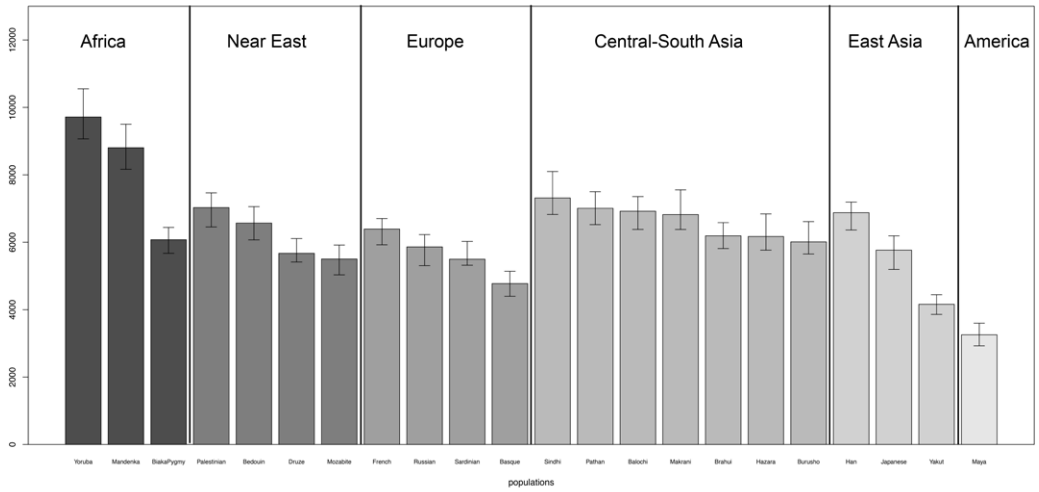

Figure 4: Long term Ne, calculated with Ne_CI function. Error bars indicate the 5th and 95th percentile of the distribution. Populations are grouped according to their geographical location

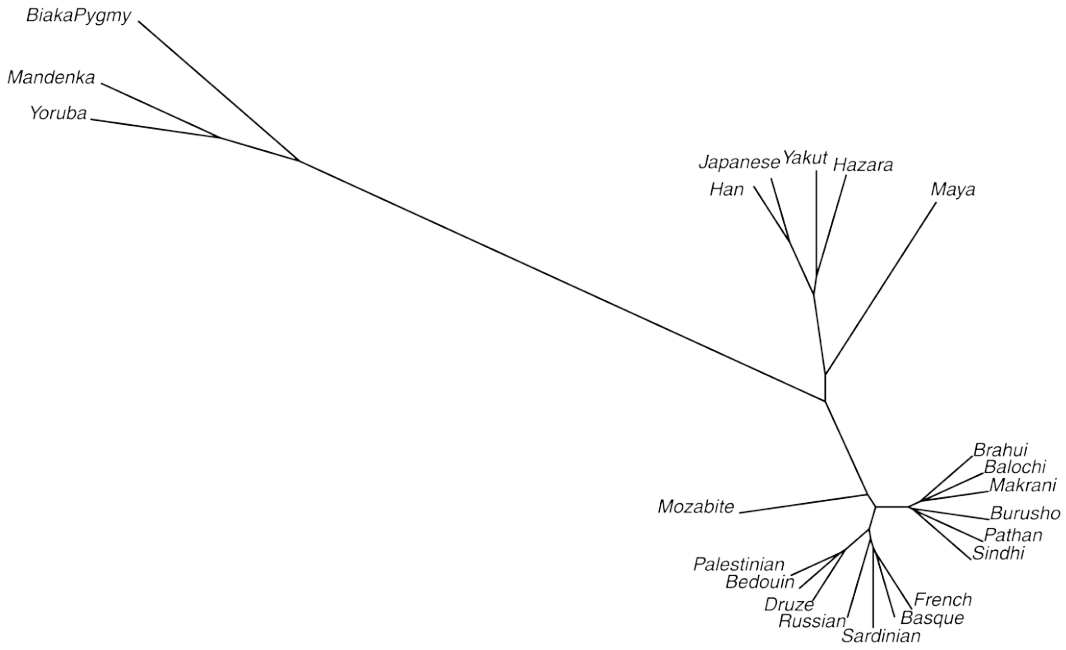

Figure 5: UPGMA tree based on the divergence time. The time is measured in generations.

reductions [33]. Because the effect of drift accumulated through time, a direct measure of *Ne* from census data is problematic. Advances in genome technology have facilitated the extensive genome-wide survey of densely spaced single nucleotide polymorphisms (SNPs), now available for many human populations [17,34,35]. This high-density genetic information can be used to estimate population genetics and evolutionary parameters that played a role in shaping today's genome variation (including recombination rate [36], level of population differentiation [37], both useful to infer past populations dynamics or demographic events [38]. A way to study these historical processes exploits the information contained in the pattern of *LD* between markers, which depends both on intrinsic cellular factors as mutation, recombination or gene conversion and on extrinsic evolutionary aspects of populations as selection, migration, and effective population size [39]. In this paper we have introduced a new tool to infer the history of effective population size from patterns of linkage disequilibrium of genome-wide single nucleotide polymorphisms data. Using the functions contained in the *NeON* package we showed how it is possible to estimate the past demographic dynamics and the divergence time of the populations genotyped in the CEPH-panel; moreover, our simulation framework showed that the divergence times so calculated can be generally considered well estimated, especially when the two populations diverged quite recently.

Other than clarify aspects of the biological evolution, the inference

of populations' demographic parameters as the degree of relatedness between human populations can also help to assess the presence and the extent of the interaction between biological and phenotypic or cultural variables. To give some examples, the so calculated divergence times can be correlated with those estimated from polymorphisms and cranial shape variables of human populations, to find evidence supporting a specific process of dispersal of early modern humans out of Africa [40], or compared with an estimation of linguistic split times, to test whether the parallelism between biological evolution and language diversification, firstly proposed by Charles Darwin [41] and subsequently verified with empirical data [42,43], has been originated by the same demographic dynamics.

The method embedded in the functions of this R package can improve and/or integrate other methods developed to estimate the effective population size or the degree of relatedness among populations from genomic data (e.g. Treemix [44], and the PSMC [12] or its extension MSMC [45]. Respect to Treemix, NeON has the advantage of estimate the time of separation between populations other than their relationships, but does not take into account migrations, whereas, respect to the sophisticated MSMC, the method presented here does not require phased data from multiple genomes, and can hence be suitable when only SNP panels are available.

Although being aware that the SNP panels suffer of ascertainment bias (resulting even in biased estimates of 18% downward [14]), the method we proposed here has already been successfully applied to the study of human evolution [17]. However, the lack of user-friendly programs strong limited its application to the study of real populations. With the *NeON* package, developed for the widely used R environment, this method can now be easily applied to analyze past population dynamics, giving the opportunity to shed light to different aspects of human population history [40]. The package NeON, together with tutorial and examples, is available for download and installation from CRAN website (http://www.r-project.org) with the license of GPL (>=2), or from University of Ferrara, Population Genetics group's website (http://www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software). It requires the package psych (http://CRAN.R-project.org/package=psych Version=1.3.10) and *PLINK* executable [15].

Acknowledgement

This study was supported by the European Research Council ERC-2011-AdvG_295733 grant (Langelin). We thank Guido Barbujani for valuable suggestions and for critically reading the manuscript, Serena Tucci that helped us finding a name for the package, and the Population Genetics group of the University of Ferrara.

References

1. Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. Genetics 98: 625-640.
2. Waples RS (1988) Estimation of allele frequencies at isoloci. Genetics 118: 371-384.
3. Williamson EG, Slatkin M (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics 152: 755-761.
4. Waples RS, Yokota M (2007) Temporal estimates of effective population size in species with overlapping generations. Genetics 175: 219-233.
5. Pudovkin A, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics 144: 383-387.
6. Luikart G, Cornuet JM (1999) Estimating the effective number of breeders from heterozygote excess in progeny. Genetics 151: 1211-1216.
7. Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. Genetical Research 38: 209-216.
8. Ewens WJ (1971) The sampling theory of selectively neutral alleles. Theor Popul Biol 3: 87-112.
9. Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.
10. Drummond AJ, Rambaur A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol biol evol 22: 1185-1192.
11. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nature genetics 43: 1031-1034.
12. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493-496.
13. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research 13: 635-643.
14. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, et al. (2007) Recent human effective population size estimated from linkage disequilibrium. Genome research 17: 520-526.
15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
16. Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage Disequilibrium and Recombination in Hominid Mitochondrial DNA. Science 286: 2524-2525.
17. McEvoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome research 21: 821-829.
18. Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97-159.
19. Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21: 3686-3687.
20. Waples RS, Gaggiotti O (2006) Invited review: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular ecology 15: 1419-1439.
21. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V et al. (2002) A human genome diversity cell line panel. Science (New York, NY) 296: 261-262.
22. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature genetics 38, 1251-1260.
23. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. (2011) Demographic history and rare allele sharing among human populations. Proc Nati Acad Sci 108: 11983-11988.
24. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, et al. (2012) Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am J Hum Genet 91: 660-671.
25. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. science 319: 1100-1104.
26. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics 10: 639-650.
27. Cockerham CC, Weir BS (1984) Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics 40: 157-164.
28. Benazzo A, Panziera A, Bertorelle G (2014) 4P: fast computing of population genetics statistics from large DNA polymorphism panels. Ecology and Evolution in press.
29. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381-2385.
30. Patin E, Laval G, Barreiro LB, Salas A, Semino O, et al. (2009) Inferring the demographic history of African farmers and Pygmy hunter–gatherers using a multilocus resequencing data set. PLoS genetics 5: e1000448.
31. Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, et al. (2011)

An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol biol and evol 29: 617-630.

32. Verdu P, Austerlitz F, Estoup A, Vitalis R, Georges M, et al. (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Current Biology 19: 312-318.

33. Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I (2009) Genetic recombination and molecular evolution. In Cold Spring Harbor symposia on quantitative biology: Cold Spring Harb Symp Quant Biol 74: 177-186.

34. Consortium IH (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

35. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83: 347-358.

36. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304: 581-584.

37. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.

38. Sved JA (2009) Linkage disequilibrium and its expectation in human populations. Twin Research and Human Genetics 12: 35-43

39. Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299-309.

40. Reyes-Centeno H, Ghirotto S, Détroit F, Grimaud-Hervé D, Barbujani G, et al. (2014) Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. Proc Nati Acad Sci 111: 7248-7253.

41. Darwin C (1859) On the origins of species by means of natural selection. London: Murray

42. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Nati Acad Sci 85: 6002-6006.

43. Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Nati Acad Sci 87: 1816-1819.

44. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8: e1002967.

45. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. bioRxiv.