Genet. Res., Camb. (2011), **93**, *pp.* 105–114. © Cambridge University Press 2011
doi:10.1017/S0016672310000558

Effective population size of current human population

LEEYOUNG PARK*
Natural Science Research Institute, Yonsei University, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea

(*Received 6 May 2010; revised 27 July and 2 November 2010; accepted 2 November 2010; first published online 31 March 2011*)

Summary

In order to estimate the effective population size (N_e) of the current human population, two new approaches, which were derived from previous methods, were used in this study. One is based on the deviation from linkage equilibrium (LE) between completely unlinked loci in different chromosomes and another is based on the deviation from the Hardy–Weinberg Equilibrium (HWE). When random mating in a population is assumed, genetic drifts in population naturally induce linkage disequilibrium (LD) between chromosomes and the deviation from HWE. The latter provides information on the N_e of the current population, and the former provides the same when the N_e is constant. If N_e fluctuates, recent N_e changes are reflected in the estimates based on LE, and the comparison between two estimates can provide information regarding recent changes of N_e. Using HapMap Phase III data, the estimates were varied from 622 to 10 437, depending on populations and estimates. The N_e appeared to fluctuate as it provided different estimates for each of the two methods. These N_e estimates were found to agree approximately with the overall increment observed in recent human populations.

1. Introduction

Effective population size (N_e) is a crucial parameter in population genetics. It plays a vital role in studying the mutation rate, recombination rate, selection pressure and genetic diversity (Palstra & Ruzzante, 2008; Charlesworth, 2009). Recent estimates based on linkage disequilibrium (LD) indicated that the effective population size of humans is much less than the previous estimates and the usually quoted value of 10 000 (Tenesa *et al.*, 2007). It has been suggested that the differences resulted from the different timescales and methods in N_e estimation. The fluctuation of N_e is important and the estimated harmonic mean (N_e) tends to be dominated by the smallest N_e (Hartl & Clark, 2007). Therefore, both long-term and short-term estimates of N_e are important for understanding evolution, and estimating the N_e of the current population could be an important step towards studying various past N_e values.

* Corresponding author: Natural Science Research Institute, Yonsei University, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea. Tel: (82)2-2123-3530. Fax: (82)2-313-8892. e-mail: lypark@yonsei.ac.kr

There are three different kinds of effective population size based on measuring magnitude: (1) the change in probability of identity by descent (inbreeding effective size); (2) the change in variance in allele frequency (variance effective size); and (3) the rate of loss of heterozygosity (eigenvalue effective size) (Hartl & Clark, 2007). Various estimation methods have been developed to determine effective population sizes (Wang, 2005; Waples, 2005; Nomura, 2009), many of which have proven useful for estimating the short-term N_e. The temporal method was developed based on the properties of the variance of gene frequency changes (Nei & Tajima, 1981; Waples, 1989). This method requires the collection of at least two samplings of a population, which would usually be more difficult than collecting a single sampling of a population. Therefore, two particular short-term methods, the LD method (Hill, 1981) and the heterozygote-excess method (Pudovkin *et al.*, 1996), were examined in this study, which can estimate N_e from a single sampling of a population.

Recent advances in genetic technologies have enabled the N_e estimation of human populations using LD (Tenesa *et al.*, 2007). The estimation of N_e from

LD is primarily based on the tightly linked loci and requires the separate or concurrent estimation of recombination rates. However, the N_e estimation using unlinked loci could be more useful, because it could provide more precise estimates without having to estimate recombination rates and could be useful for studying population history (Hill, 1981; Waples, 2005). In reality, unlinked loci provide the very recent information on N_e, which involves only current and several previous generations based on the Wright–Fisher model due to the fast decay of LD (Waples, 2005). The previous methods, using linked or unlinked loci, were based on two-locus inbreeding descent measures (Weir & Cockerham, 1969; Cockerham & Weir, 1973), from which the N_e estimate was the inbreeding effective population size. This study is focused primarily on the variances of allele and haplotype frequencies generated during random mating and sampling of a population, deriving different equations for estimating the variance effective population size from the previous methods for the inbreeding effective population size.

The heterozygote-excess method, useful for estimating a small N_e, is based on the differences of allele frequencies in males and females (Pudovkin *et al.*, 1996; Luikart & Cornuet, 1999; Nomura, 2009). As summarized previously (Wang, 2005), genetic drift generates not only the heterozygote-excess but also deviations from the Hardy–Weinberg equilibrium (HWE), which enables the N_e estimation of the current generation. These estimations result in the variance effective size, and could be comparable with the variance effective size based on LE. In this study, the N_e estimations for current and recent generations were conducted based on these two, newly developed methods for various populations provided by the HapMap consortium. Then, they were compared with the actual census of human population.

2. Methods

(i) *Data*

The HapMap Phase III genotype data from the HapMap project were used for the estimations (International HapMap Consortium, 2003, 2005; Frazer *et al.*, 2007; Altshuler *et al.*, 2010), which included the original and expanded HapMap samples. The original HapMap samples were collected from four geographically diverse populations: the Yoruba in Ibadan, Nigeria (YRI); Japanese in Tokyo, Japan (JPT); Han Chinese in Beijing, China (CHB); and the CEPH (USA Utah residents with ancestry from northern and western Europe, CEU). And, additional samples were collected from seven populations: Maasai in Kinyawa, Kenya (MKK); Luhya in Webuye, Kenya (LWK); Chinese in metropolitan Denver, CO, USA (CHD); Gujarati Indians in Houston, TX, USA (GIH); Tuscans in Italy (TSI); African ancestry in the southwest USA (ASW); and Mexican ancestry in Los Angeles, CA, USA (MEX). The ASW, CEU, MEX, MKK and YRI were family samples, and only the parents were used in this data analysis (indicated as ASWp, CEUp, MEXp, MKKp and YRIp, respectively). A total of 988 samples were analysed, and the sample number of each population was indicated in Table 1. For the accuracy of estimates, single nucleotide polymorphisms (SNPs) with minor allele frequencies greater than 0·4 and without missing data were selected for analysis.

(ii) N_e *estimation based on unlinked loci between different chromosomes*

The LD observed between different chromosomes comes almost entirely from random mating and sampling of populations. In this study, the LD indicates r^2, the correlation coefficient of gene frequencies. For the completely unlinked loci in different chromosomes, if ideal Wright–Fisher population assumptions are held in a population, the variances due to genetic drifts are the only factors for LD between different chromosomes. In diploids, four different mating structures could be considered: (1) monoecious with selfing; (2) monoecious without selfing; (3) dioecious with random pairing (equal number of males and females); and (4) dioecious with a hierarchical mating structure (Weir & Hill, 1980). The r^2 based on descent measures was the same for (1), (2) and (3) in the previous study (Weir & Hill, 1980). The difference between monoecious and dioecious is the division of a population into two separate genders. Because half of the population is assigned to a certain gender randomly, there are negligible actual differences in estimates unless the gender ratios differ. And, if there is no preference for selfing, a monoecious population with selfing does not provide a different r^2 from a monoecious population without selfing. Simulations in this study also indicated that (1), (2) and (3) produced the same result. Therefore, for simplicity, the monoecious with selfing was considered to derive the relationship between N_e and r^2.

Each generation is discrete. Let N_{ec} be the population size of the current generation, and let N_{ep} be the population size of the previous generation. The population sampling procedure from one generation to the next is simplified as: (1) sampling of $2N_{ec}$ individuals from N_{ep}; (2) generating gametes from each individual of selected $2N_{ec}$; and (3) random pairing of the generated gametes. The expected r^2 due to random sampling of population (N) is $1/(2N)$ if each gamete can be observable and there is no disequilibrium (Weir, 1996). It should be noted that diploidy gives additional randomness to the existing LD when

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

Table 1. *Simulated* r^2 *from different effective population sizes compared with the expected values calculated from eqn (1) (10 000 SNP pairs were simulated; *: calculated from the expected* r^2*, assuming* N_e *to be constant).*

(a) When Ne was constant (the simulated estimates were obtained after 10 initial generations).

N_e	Simulation	Expected	*N_e estimate
100	0·00837	0·00833	100
1000	0·00083	0·00083	1007
2000	0·00042	0·00042	1997
5000	0·00017	0·00017	5049
10 000	0·00008	0·00008	9939

(b) When N_e was changed from N_{e1} to N_{e2} (the simulated estimates were obtained after 10 initial generations with N_{e1}).

N_{e1}	N_{e2}	Simulation	Expected	*N_e estimate
1000	2000	0·000575	0·000583	1448
2000	1000	0·000674	0·000667	1237
10 000	1000	0·00054	0·000533	1543
1000	10 000	0·000381	0·000383	2188
3000	8000	0·000176	0·000174	4734

(c) When N_e was changed from N_{e1} to N_{e5} (the simulated estimates were obtained after five initial generations with N_{e1}).

N_{e1}	N_{e2}	N_{e3}	N_{e4}	N_{e5}	Simulation	Expected	*N_e estimate
1000	2000	3000	4000	5000	0·000195	0·000196	4244
5000	4000	3000	2000	1000	0·000649	0·000651	1281
100	1000	1000	1000	1000	0·000931	0·000880	947
1000	10 000	10 000	10 000	10 000	0·000089	0·000088	9467
10 000	1000	2000	4000	10 000	0·000160	0·000160	5212
2000	3000	2000	4000	5000	0·000202	0·000202	4134
1000	1000	2000	5000	10 000	0·000151	0·000152	5479
1000	2000	3000	4000	3000	0·000264	0·000263	3168
20 000	20 000	20 000	20 000	2200	0·000262	0·000244	3416

sampling proceeds for unlinked loci. This is because the randomly paired gametes in N individuals generate variable haplotype frequencies instead of the expected haplotype frequencies from allele frequencies. For example, the haplotype frequency of AB is not the multiplication of the allele frequency of A and B. Therefore, sampling from N individuals provides additional departure from LE with the value of $1/(2N)$.

When sampling of $2N_{ec}$ individuals from N_{ep}, the additional departure from the existing r^2 is $1/(2N_{ep})$. For unlinked loci, each locus act independently, and the original LD and the randomness from previous generation are reduced by a quarter during gamete transmission. In addition, the $2N_{ec}$ gametes generated from selected $2N_{ec}$ individuals are randomly paired to make N_{ec} individuals in the current generation, providing departure from LE ($1/(2N_{ec})$). Therefore, the LD between loci in different chromosomes is dependent on the effective population size of current and previous generations as well as the LD at the previous generation as indicated in eqn (1), in which r_0^2 is the LD of the previous generation, N_{ep} and N_{ec} represent the effective size of the previous and current generation, respectively, and c is 0·5. When the population sizes are constant for each generation, r_0^2 becomes equal to the expected r^2 and can be expressed as $5/(6N_e)$:

$$E(r^2) = \left(r_0^2 + \frac{1}{2N_{ep}}\right)(1-c)^2 + \frac{1}{2N_{ec}}. \quad (1)$$

Due to the same reason indicated in the explanation of the population sampling procedure, when the n samples were selected from a population, the expected LD in the sampling population was not exactly zero. Therefore, due to randomly paired gametes for two independent loci in each individual in the sampling population, $1/(2N_{ec})$ should be added to the original LD of the population as sampling variances. During the inference of haplotypes from genotypes, the maximum-likelihood estimation of haplotype frequencies makes the variances due to sampling of $1/n$ instead of $1/(2n)$ (Hill, 1974; Weir & Hill, 1980; Hill, 1981). Therefore, sampling provides an additional $1/(2N_{ec}) + 1/n$ to the original r^2 in eqn (1). In this case, assuming a constant effective population size of N, the expected r^2 is $4/(3N) + 1/n$. This approach provides an N_e estimate that is four times larger than the inbreeding N_e estimate using unlinked loci (Hill, 1981;

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

Waples, 2006). The r^2 of SNPs in different chromosomes in each HapMap population were calculated based on the parallelized C++ code.

(iii) N_e *estimation based on HWE*

Similar to the estimation of N_e based on LD between unlinked loci, the deviations from HWE through genetic drift provide information on effective population size. The estimation of deviation is intrinsically similar to the asymptotic HWE tests (Weir, 1996). In the HWE tests, the chi-squared distribution is entirely based on the sample number and the disequilibrium coefficient; however, it actually includes genetic drift of the population as well as the sampling. Due to the random mating of population, the deviation from HWE depends only on the current population. Therefore, through approximations, eqn (2) applies to the estimation of the current effective population size

$$E\left(\frac{\left(\tilde{P}_{AA}-\tilde{P}_A^2\right)^2}{\tilde{P}_A^2\left(1-\tilde{P}_A\right)^2}\right)=\frac{1}{N_e}+\frac{1}{n}. \quad (2)$$

In the HapMap data, the SNPs that deviated from the HWE for a significance level of 0·001 were removed (2005). However, in this study, all of the data should be used for estimating N_e. Therefore, corrections were necessary for the missing data, most of which might naturally deviate from the HWE due to the distributional property. To find the proportion of the mean of the data for the significance level higher than 0·001 to the mean of the total data, simulations were conducted by generating random numbers based on the chi-squared distribution as much as the number of total SNPs used for the estimation. The simulations were repeated 1000 times, and yielded the correction term, 0·9873 using the R statistical package (R Foundation for Statistical Computing, Vienna, http//www.R-project.org). Although the exact tests were conducted for the quality control of the HapMap data (International HapMap Consortium, 2005), the exact test and the asymptotic test showed similar results when the minor allele frequency was high and/or the sample size was large enough (Vithayasai, 1973; Hernandez & Weir, 1989; Wigginton *et al*., 2005). Thus, the actual cut-offs were the same for the smallest samples (49 for the parents of ASW and 50 for the parents of MEX), when the minor allele frequencies were higher than 0·4. Therefore, the simulations were based on an asymptotic chi-squared distribution. The deviations from the HWE were also calculated using the parallelized C++ code along with the r^2 calculations.

(iv) *Simulation and errors*

The derived eqn (1) becomes clearer in simulation studies. The simulation was conducted with 10 000 diallelic loci for various population sizes. The alleles for each locus at the initial generation were generated by binomial draw on the allele frequencies of 0·5, and random mating of the population proceeded for 5–10 generations after the initial generation of the population. The population sampling procedure was the same as that previously described for estimating N_e based on unlinked loci between different chromosomes. Similar to the previous notation, let N_{ep} be the population size at the previous generation and N_{ec} be the size at the current generation. First, $2N_{ec}$ individuals were selected from N_{ep} individuals for mating. Second, from each of the $2N_{ec}$ individuals, one transmitting gamete was generated. Since each locus is independent, random selection of one allele between two alleles from each locus in an individual was conducted. Third, random pairings of the gametes generated N_{ec} individuals. At the sixth or eleventh generations, the r^2 of the population was determined based on 49 995 000 pairs of loci for a simulation. The unequal gender ratios in a population were tested by dividing the N individuals randomly at each generation into two genders in fixed proportions.

As shown in Table 1, the simulated mean of LD with 10 000 SNP pairs was almost same as the expected mean of LD. In Table 1(a), the N_e estimates were similar to the actual N_e. When N_e fluctuated, it would not be possible to estimate N_e at each generation. In the case of Table 1(b) and (c), the current N_e was mostly reflected in the mean r^2, and recent changes of N_e were also reflected. Therefore, it could be used to infer the current and recent effective population sizes. The simulations also revealed that the mean of subtractions from the sampled r^2 to the original r^2 gives approximately $1/(2N_{ec})+1/n$. The same simulations were conducted for population sampling procedures, and then the sampling of n individuals among N_{ec} individuals was conducted. From the sampled genotypes, the haplotype frequencies for two loci were estimated based on the expected maximisation (EM) algorithm. For each simulation of 1000 or higher, the original r^2 and the sampled r^2 were collected, and the mean of the subtraction between the values was examined.

The square roots of mean-squared errors were obtained from simulations for the population size of 1000 and several sample sizes. The same procedure previously described was repeated for this simulation. Simulations were repeated 100 times for 10 000 polymorphisms to get the mean-squared errors, which provided 10 000 results for the deviation from HWE and 49 995 000 results for the deviation from LE in each simulation. The samplings with replacement and the constant N_e were applied. The N_e estimates from the simulations were subtracted from the actual N_e, and the square roots of mean-squared errors were obtained. As summarized in Table 2, the square roots

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

Table 2. *Square root of mean-squared errors* (*SRmse*) *from 100 simulations with 10 000 polymorphisms and a population size of 1000*

Sample size	100	200	500	1000	2000
Mean r_2 LE	0·01164	0·00639	0·00335	0·00235	0·00183
Mean r_2 HWE	0·01111	0·00600	0·00300	0·00200	0·00150
Mean N_e LE	877	980	992	993	999
Mean N_e HWE	1202	1051	1005	998	999
SRmse LE	293·94	160·68	64·75	38·85	17·95
SRmse HWE	1232·00	254·95	100·74	50·05	33·30

of the mean-squared errors were smaller for the N_e estimates using the deviation from LE than the estimates using the deviation from HWE. Sample sizes were critical for reducing mean-squared errors. Increasing the number of polymorphisms helped in reducing the errors to a certain extent, but the simulation numbers had very little influence. These results indicated that the interpretation of results acquired with small sample sizes should be conducted with caution. All the simulations in this study were executed using the R statistical package with additional C++ coding of the core computation. The generation of every random number was based on the R statistical package.

3. Results

In this study, two methods were proposed for estimating the variance effective population size, based on the deviation from linkage equilibrium (LE) and the deviation from the HWE. The estimate of the variance effective size using the deviation from LE was fourfold larger than the previous estimate of inbreeding effective size. Simulation studies exhibited excellent agreement between the estimates produced by the simulations and the theoretically expected values (Table 1). Unequal gender ratios reduce N_e estimates using LE, but, as the ratio increased, the estimates became larger than $4MF/(M+F)$, in which M is the number of males and F is the number of females. There was no difference in N_e estimates using HWE depending on gender ratios. More studies would be necessary for determining the effects of the gender ratios on these estimates. As shown in Table 2, if the sample sizes are not large enough, there are considerably large errors in the estimates, especially for the estimate using the deviation from HWE.

These new methods were used to estimate the effective population size of current human populations using the HapMap data. As summarized in Table 3, the estimated effective population sizes were varied (ranging from 622 to 10 437), depending on population and estimates. Since the sample sizes in the HapMap data were small, there were many chances for errors to occur in the estimates. However, the errors in the N_e estimates from LE were relatively small enough to allow the approximate determination of the true N_e, and, overall, the estimates from LE were similar to the estimates from HWE. Therefore, the estimates in this study could be reliable enough to interpret the results. There were three pairs of cryptic relatedness in the samples of YRIp and CEUp, one pair in YRIp and two pairs in CEUp (International HapMap Consortium, 2005). The relatedness could increase the departure from HWE, and decrease the N_e estimate. However, estimates of both YRIp and CEUp based on HWE were higher than other estimates in this study. More detailed studies are needed to explain the effects further.

In Table 3, three populations, CEUp, TSI and YRIp, showed relatively large N_e, whereas ASWp, LWK, MEXp and MKKp showed relatively small N_e. The polygamy culture could be responsible for the small N_e in the Maasai and Luhya population (Nomura, 2005). ASWp and MEXp might be admixed populations, and a recent study indicated that N_e estimation is often underestimated when subpopulations differ substantially in allele frequencies when the temporal method is used (Araki *et al*., 2007). It is likely that the current method could show a similar effect due to the admixture between two different populations. In addition, MEXp and ASWp were immigrants, whose populations might historically have experienced a bottleneck before or after immigration. In order to understand correctly, more studies are needed to determine the reason for the large N_e values of the three populations of CEUp, TSI and YRIp as well as others.

Both CHB and CHD originated from China, and had Han Chinese ancestors. Both estimates of these two samples showed similar N_e values. In contrast, LWK and MKKp samples were selected from Kenya, but showed different N_e estimates. Individuals from CHB and CHD were unrelated and had at least three out of four Han Chinese grandparents, but MKKp and LWK differed in their ethnic groups, i.e. MKKp samples were the parent group of family data in which each individual had four Massai grandparents and LWK samples were unrelated individuals who identified themselves as having four Luhya grandparents.

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

Table 3. *Estimated effective population sizes from the mean of LD coefficients between different chromosomes (Mean* r_2*) and the mean disequilibrium coefficient of the HWE* ($\mathbf{M_{dc}}$)*;* (N_{ei}*: inbreeding effective population size;* N_e*: variance effective population size estimated from LE;* N_{ec}*: variance effective population size based on the deviation from HWE; ASWp: African ancestry in the southwest USA; CEUp: USA Utah residents with ancestry from northern and western Europe; CHB: Han Chinese in Beijing, China; CHD: Chinese in metropolitan Denver, CO, USA; GIH: Gujarati Indians in Houston, TX, USA; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEXp: Mexican ancestry in Los Angeles, CA, USA; MKKp: Maasai in Kinyawa, Kenya; TSI: Tuscans in Italy; YRIp: the Yoruba in Ibadan, Nigeria)*

Population	ASWp	CEUp	CHB	CHD	GIH	JPT	LWK	MEXp	MKKp	TSI	YRIp
No. of Samples	49	112	84	85	88	86	90	50	143	88	113
Total SNP No.	183 207	169 627	166 994	160 918	188 309	173 953	167 329	193 847	162 871	182 703	161 207
Total SNP pairs	1·59E + 10	1·36E + 10	1·32E + 10	1·23E + 10	1·68E + 10	1·43E + 10	1·32E + 10	1·78E + 10	1·25E + 10	1·58E + 10	1·23E + 10
Mean r_2	0·02255	0·00918	0·01228	0·01218	0·01188	0·01200	0·01200	0·02188	0·00824	0·01169	0·00911
M_{dc}	0·02162	0·00891	0·01209	0·01197	0·01145	0·01193	0·01165	0·02025	0·00748	0·01141	0·00887
M_{dc} corrected	0·02190	0·00902	0·01225	0·01212	0·01160	0·01208	0·01180	0·02051	0·00758	0·01156	0·00898
N_{ei}	155	1331	887	809	644	885	375	178	267	1020	1275
N_e	622	5325	3550	3238	2574	3541	1502	710	1067	4079	5101
N_{ec}	670	10 437	2926	2819	4319	2219	1451	1971	1710	5071	7729
Ratio (N_{ec}/N_e)	1·08	1·96	0·82	0·87	1·68	0·63	0·97	2·77	1·60	1·24	1·52

As such, the age group, as well as ethnicity, of MKKp differed from LWK because individuals of MKKp were the parent group. The differences found between MKKp and LWK are likely to be due to differences in age, ethnicity and/or culture (especially marriage). Particularly, it should be noted that each African ethnic group is known to practice specific marriage customs, including polygamous patterns, which could influence N_e to a great extent.

The estimates based on HWE presented the effective population size of the current generation, and the estimates based on LE indicated the N_e of current and recent generations. Therefore, the N_e based on HWE can be compared with the estimates based on LE to examine recent population changes. Most populations showed increased N_e values for current generations compared to the N_e estimates based on LE, which generally agreed with the recent population growth of humans. The increment was largest in MEXp, with a ratio of 2·77, and CEUp (1·96), GIH (1·68), MKKp (1·60) and YRIp (1·52) also showed relatively high ratios. The JPT sample showed a significantly decreased N_e value in the current generation. Since Japan is a highly aging society and has the nearly lowest and declining birth rates among countries, this estimation result might be reasonable.

As indicated by the HapMap consortium, the population data do not represent any specific nation or ethnic group. Nevertheless, the HapMap samples were selected from specific regions of countries with specific criteria, and they could at least partially represent the population of the country. To compare and examine the approximate reliability of the current methods, census data were examined. The possibly relevant United Nations (UN) census data were summarized in Fig. 1 (UNSD, 2010). Due to the policy guiding HapMap data construction (International HapMap Consortium, 2004), all members of the HapMap family data were adults only. As such, 1980 could be the most reproducibly representative year for the parent group, whose samples were collected in 2000, and 1950 could be the most reproductive year for the parent group of those parents. The year 2000 could be the best representative year for unrelated individuals who are all adults, and 1970 could be the most representative year for their parent groups. The completely tabulated censuses were used exclusively, with the exception of one U.S. (United States of America) data set for the female to male ratio in 1970. If the census for the targeted year did not exist for a particular country, the closest year was selected for use.

The listed population sizes in Fig. 1 represent the total population. The most appropriate actual population number comparable to N_e would be the number of adults, which includes breeding and senescent adults (Frankham, 2007). However, the HapMap data themselves do not represent the population

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

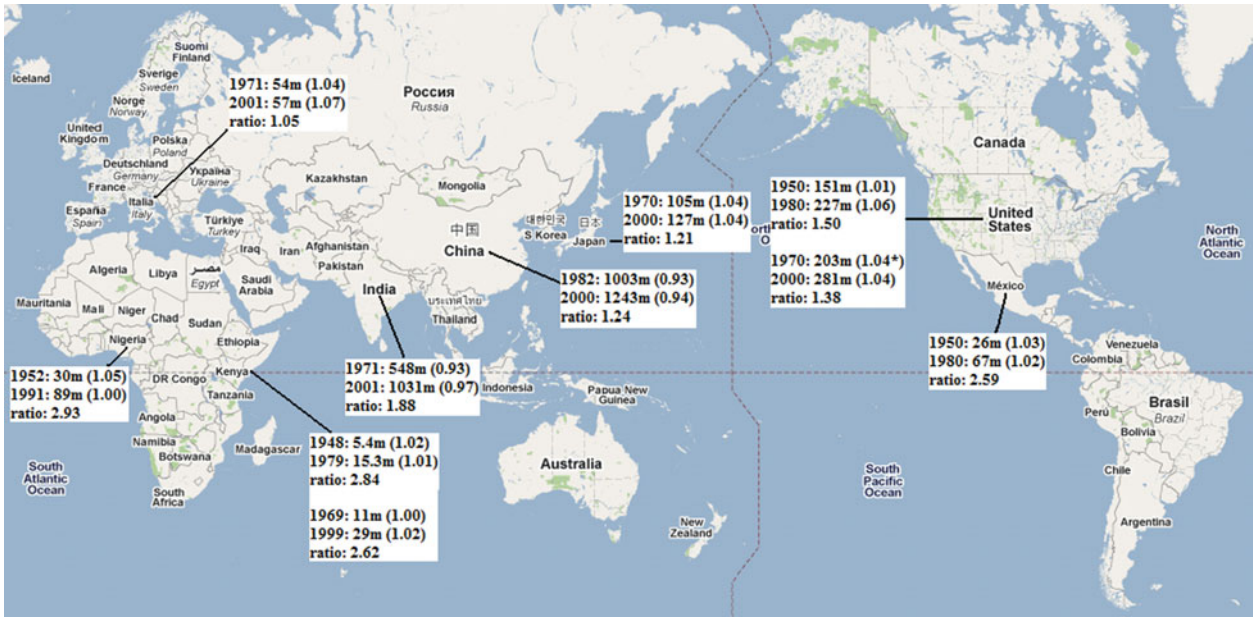

Fig. 1. World Population Census from the United Nations Statistics Division Demographic Statistics (Parenthesis: female to male ratio; *: estimates, not the complete census; the map was obtained through the courtesy of Google Maps). The studied populations were listed below, and the relevant population censuses were indicated in the parenthesis: ASWp: African ancestry in the southwest USA (1970–2000 in USA); CEUp: USA Utah residents with ancestry from northern and western Europe; CHB: Han Chinese in Beijing, China (1982–2000 in China); CHD: Chinese in metropolitan Denver, CO, USA (1982–2000 in China and 1970–2000 in USA); GIH: Gujarati Indians in Houston, TX, USA (1971–2001 in India and 1970–2000 in USA); JPT: Japanese in Tokyo, Japan (1970–2000 in Japan); LWK: Luhya in Webuye, Kenya (1969–1999 in Kenya); MEXp: Mexican ancestry in Los Angeles, CA, USA (1950–1980 in Mexico and 1950–1980 in USA); MKKp: Maasai in Kinyawa, Kenya (1948–1979 in Kenya); TSI: Tuscans in Italy (1971–2001 in Italy); YRIp: the Yoruba in Ibadan, Nigeria (1952–1991 in Nigeria).

of specific countries, and there are problems in determining the appropriate age range for the adult groups, which might differ depending on culture. In addition, there are many missing age-classified data in the UN population census data. Therefore, in this study, approximate comparisons based on the total census population were conducted. Since unequal gender ratios influence N_e estimations, the gender ratios of each country were also examined; however, there were no significant deviations in the gender ratios.

As expected, the N_e estimates using HapMap data did not correlate well with the population size of the countries in which the corresponding HapMap samples were collected. For an example, the CHB and JPT samples showed similar N_e estimates, even though China has almost 10 times more population than Japan. This might be due to the fact that the collected samples are not the representative population of a country. They are a partial, limited selection of the population. The correlation between N_e based on LE and real population data was better at 0·15 for the previous generation and 0·14 for the current generation of real human population than the correlation between N_e based on HWE and real population data, which showed almost no correlation. This could be due to the higher errors in the estimates based on HWE. These discordances might arise from the fact that the samples may not be the representative population of the country, because the samples were collected based on specific regions and ethnicity.

A higher correlation of 0·5 was observed between the ratios of estimates of N_e based on HWE to N_e based on LE and the ratios of current population size to the population size of previous generation. The increment of N_e of ASWp and CEUp data was determined to represent the population growth of the U.S. roughly well, because these people had resided in the U.S. for a relatively long period of time. Other populations collected in the U.S. that originated from other countries (CHD, GIH and MEXp) showed trends that were more similar to their nations of origin, agreeing with the fact that their immigration was relatively recent.

CEU samples were collected in 1980 to carry out studies on the human genome using large sibships (Dausset *et al.*, 1990; Cann, 1992). Therefore, the active reproductive period could be anticipated to be 1950 for the selected grandparents and 1920 for the great grandparents, the data that were not listed in the UN data. The U.S. Census Bureau indicated that the U.S. population in 1920 was 106 million and, in 1950, 151 million (U.S. Department of Commerce, 2010). The number in 1950 is slightly different from

https://doi.org/10.1017/S0016672310000558 Published online by Cambridge University Press

the UN data. Since the increment ratio was determined to be 1·43, the larger difference between two N_e estimates of CEUp than the actual population change might not be due to the difference of time. Large families were the target of the CEPH collection, and the mean sibship size for initially collected families was 8·3. It is possible that these families could have been larger than the average for several recent generations, and these large family sizes could influence the N_e estimation.

4. Discussion

This study provides new potentially useful methods for estimating the effective population sizes of current and/or recent generations from genetic data. The estimates were more similar to the estimates from LD, which ranged approximately from 3100 to 7500 (Tenesa *et al*., 2007), than the estimate from nucleotide diversity, which was 10 400 (Yu *et al*., 2004). The estimates using LD were based on the same HapMap database used in this study, although larger sample sizes were used in this study and the estimated time point was closer to this study than the estimates from nucleotide diversity. Compared to the previous estimates based on LD (Tenesa *et al*., 2007), CHB and JPT showed similar estimates, but YRI and CEU were slightly different in this study. The estimates based on LD were the inbreeding N_e, but the current estimates based on LE were the variance N_e. It should be noted that, for unlinked loci, the variance N_e is four times larger than the inbreeding N_e, as indicated in the Method section. More studies would be necessary to compare the N_e estimates based on LE in this study to the inbreeding N_e estimates.

As shown in Table 2, the increase of sample sizes dramatically reduced the errors. With sufficient sample sizes, the methods suggested in this study would be very useful. As the sample size increased, the N_e estimates from LE increased to converge into the actual population size, but the N_e estimates from HWE decreased to converge into the actual population size. This indicated that there might be intrinsic bias in sampling similar to the previous study for the inbreeding N_e estimation using unlinked loci (Waples, 2006). Further studies could reduce the bias. The simulated and real distributions of deviations from HWE and LE looked like a chi-squared distribution with one degree of freedom, although the scales differed. There were small deviations in most of the Quantile–Quantile plots between the observed or simulated values and the random numbers that were generated based on the chi-squared distribution. However, more studies might be needed to understand their exact distributional properties and to correctly interpret the estimates. For estimates based on HWE, non-random mating induces more or less deviations from the expected values. When there are epistatic effects between unlinked loci, more LD than expected values would be observed. These effects should be studied carefully based on theoretical explanations of the distribution and real data with sufficiently large sample sizes.

In Table 3, the N_e estimates of both CHB and CHD indicated that there was a slight decrease in the N_e at the current generation. The one-child policy was introduced in China in 1978, and the CHB samples were collected in 2000. If the samples included young individuals in their early 20s, the decreased N_e of CHB could be observable. Immigration to U.S. from China increased until 2001 (U.S. Immigration and Naturalization Service, 2001), and the individuals are more likely to have been born in China. The less decreased N_e of CHD could also be the result of the one-child policy, and the effect might be smaller. A very slight decrease in LWK population could be explained by the pandemic of HIV/AIDS in Kenya, experiencing one of the world's harshest HIV/AIDS epidemics, in which the prevalence peaked at 13·4% in 2000 (National AIDS/STD Control Programme (NASCOP), 2006). Webuye, Kenya is an urban region in which big roads penetrate. Webuye is located near the region where the high prevalence of AIDS occurred (UNAIDS, 2004), and these roads provided easy access to the city and its people by those who were infected with AIDS (UNGASS, 2008). However, Kinyawa is a rural area, and the Maasai people have preserved their traditional lifestyle, having possibly less contact with outsides. Those conditions could have protected somewhat from HIV infections, possibly resulting in the increased N_e estimate at the current generation.

By drawing comparisons with real population data, the methods applied in this study were confirmed to be useful since they generally agreed with real population growth or could be explained historically or sociologically. However, considering the high errors in the estimates, a more systematic, large-scale study design would be necessary for studying the relationship between the effective population size and the actual population size, based on selecting the most representative samples. For further application to population genetic studies, such as population history, more accurate estimates of effective population sizes would be helpful, involving adjustments for the minor contributions of non-random mating, family structure, unequal gender ratios and/or other possibly contributing factors. Caution should be applied when interpreting the estimates obtained in this study. First, sample sizes were small, and there is the possibility of substantial errors in the estimates. Second, although corrections were applied, the analysed HapMap data were composed of data that excluded the SNPs that deviated from the HWE. Third, the HapMap samples

do not represent any particular populations. To study precise human demographic history or human genome evolution, an elaborate research design would be necessary. Nevertheless, it is expected that the current estimation approaches would be useful for revealing various population genetic aspects of diverse species based on well-designed research.

Small sample sizes compared to population size would be a universal problem in most studies including this one, and the current method could still provide more accurate estimates since it has less dependency on other population genetic parameters that are usually unknown. Because many previous estimates required the estimation of additional parameters, such as the recombination rate and the mutation rate, the current methods are advantageous since they have nearly no dependence on other population genetic parameters. The methods in this study can also be applied to the estimation of the effective population sizes of endangered species. By obtaining the N_e of both current and recent generations from a single cohort sample, the status of the target species can obviously be observed. Moreover, the independence of the other population genetic parameters enables the estimation of other interested population genetic parameters by applying the estimated effective population size.

The author greatly appreciates the reviewers' comments, which critically improved the quality of this study. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-532-C00017) and by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2009-353-C00061). The key calculations were performed using the supercomputing resource at the Korea Institute of Science and Technology Information (KISTI), which provided support through grant No. KSC-2009-S01-0003. The author greatly appreciates the help from the optimization/parallelization support team at KISTI, which optimized and parallelized the C++ code for the supercomputing system. The author also thanks Eunhee Choi at the Department of Biostatistics, Yonsei University College of Medicine for being involved in a preliminary study using the HapMap Phase II data, which facilitated the establishment of the current study.

References

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Bonnen, P. E., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Gonzaga-Jauregui, C., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Zhang, Q., Ghori, M. J., McGinnis, R., McLaren, W., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D. & McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58.

Araki, H., Waples, R. S. & Blouin, M. S. (2007). A potential bias in the temporal method for estimating Ne in admixed populations under natural selection. *Mol Ecol* **16**, 2261–2271.

Cann, H. M. (1992). CEPH maps. *Curr Opin Genet Dev* **2**, 393–399.

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195–205.

Cockerham, C. C. & Weir, B. S. (1973). Descent measures for two loci with some applications. *Theor Popul Biol* **4**, 300–330.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J. M. & White, R. (1990). Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577.

Frankham, R. (2007). Effective population size/adult population size ratios in wildlife: a review. *Genet Res* **89**, 491–503.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M.,

Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R. & Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

Hartl, D. L. & Clark, A. G. (2007). Principles of Population Genetics. Sunderland, MA: Sinauer Associates Inc.

Hernandez, J. L. & Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics* **45**, 53–70.

Hill, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–239.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet Res* **38**, 209–216.

International HapMap Consortium. (2003). The International HapMap Project. *Nature* **426**, 789–796.

International HapMap Consortium. (2004). Integrating ethics and science in the International HapMap project. *Nat Rev Genet* **5**, 467–475.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Luikart, G. & Cornuet, J. M. (1999). Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* **151**, 1211–1216.

National AIDS/STD Control Programme (NASCOP), M. o. H., Government of Kenya (2006). Sentinnel Surveillance 2006 Report.

Nei, M. & Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640.

Nomura, T. (2005). Effective population size under random mating with a finite number of matings. *Genetics* **171**, 1441–1442.

Nomura, T. (2009). Interval estimation of the effective population size from heterozygote-excess in SNP markers. *Biomed J* **51**, 996–1016.

Palstra, F. P. & Ruzzante, D. E. (2008). Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Mol Ecol* **17**, 3428–3447.

Pudovkin, A. I., Zaykin, D. V. & Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144**, 383–387.

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 Res* **17**, 520–526.

United Nations Statistics Division (UNSD). (2010). UNSD Demographic Statistics. Available from the United Nations data site, http://data.un.org

U.S. Immigration and Naturalization Service. (2001). Statistical Yearbook of the Immigration and Naturalization Service. Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Commerce. (2010). U.S. Census Bureau. Census of Population and Housing. Available from the U.S. Census Bureau publication site: http://www.census.gov/prod/www/abs/decennial/index.html

UNAIDS (2004). Kenya epidemiological fact sheets. Available from the UNAIDS site: http://data.unaids.org/publications/Fact-Sheets01/kenya_en.pdf

UNGASS (2008). Country report – Kenya (M. o. S. P. Kenya Office of the President, Ed.). Available from the UNAIDS data site: http://data.unaids.org/pub/Report/2008/kenya_2008_country_progress_report_en.pdf

Vithayasai, C. (1973). Exact critical values of the Hardy-Weinberg test statistic for two alleles. *Commun Stat* **1**, 229–242.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci* **360**, 1395–1409.

Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379–391.

Waples, R. S. (2005). Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Mol Ecol* **14**, 3335–3352.

Waples, R. S. (2006). A bias correction for estimate of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet* **7**, 167–184.

Weir, B. S. (1996). Genetic Data Analysis II. Sunderland, MA: Sinauer Associates, Inc.

Weir, B. S. & Cockerham, C. C. (1969). Group inbreeding with two linked loci. *Genetics* **63**, 711–742.

Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.

Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. (2005). A note on exact tests of Hardy–Weinberg equilibrium. *Am J Hum Genet* **76**, 887–893.

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O. & Li, W. H. (2004). Nucleotide diversity in gorillas. *Genetics* **166**, 1375–1383.