Recent human effective population size estimated from linkage disequilibrium

Albert Tenesa,[1,2,3] Pau Navarro,[3] Ben J. Hayes,[4] David L. Duffy,[5] Geraldine M. Clarke,[6] Mike E. Goddard,[4,7] and Peter M. Visscher[3,5,8]

[1]Colon Cancer Genetics Group, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; [2]MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom; [4]Victorian Institute of Animal Science, DPI, Attwood 3049, Australia; [5]Queensland Institute of Medical Research, Royal Brisbane Hospital, Brisbane 4006, Australia; [6]The Wellcome Trust Centre for Human Genetics, The University of Oxford, Oxford OX3 7BN, United Kingdom; [7]Institute of Land and Food Resources, University of Melbourne, Parkville 3010, Australia

Effective population size (N_e) determines the amount of genetic variation, genetic drift, and linkage disequilibrium (LD) in populations. Here, we present the first genome-wide estimates of human effective population size from LD data. Chromosome-specific effective population size was estimated for all autosomes and the X chromosome from estimated LD between SNP pairs <100 kb apart. We account for variation in recombination rate by using coalescent-based estimates of fine-scale recombination rate from one sample and correlating these with LD in an independent sample. Phase I of the HapMap project produced between 18 and 22 million SNP pairs in samples from four populations: Yoruba from Ibadan (YRI), Nigeria; Japanese from Tokyo (JPT); Han Chinese from Beijing (HCB); and residents from Utah with ancestry from northern and western Europe (CEU). For CEU, JPT, and HCB, the estimate of effective population size, adjusted for SNP ascertainment bias, was ~3100, whereas the estimate for the YRI was ~7500, consistent with the out-of-Africa theory of ancestral human population expansion and concurrent bottlenecks. We show that the decay in LD over distance between SNPs is consistent with recent population growth. The estimates of N_e are lower than previously published estimates based on heterozygosity, possibly because they represent one or more bottlenecks in human population size that occurred ~10,000 to 200,000 years ago.

Effective population size (N_e) is an important population parameter that helps to explain how human populations evolved and expanded, and to improve the understanding and modeling of the genetic architecture underlying complex traits (Reich and Lander 2001). Traditionally, N_e has been estimated by comparing DNA sequences (i.e., from the distribution and divergence of polymorphisms). However, N_e is unlikely to have been constant during the evolution of humans, and so DNA sequence heterozygosity estimates some average N_e over a long period of time. N_e can also be estimated from linkage disequilibrium (LD) data (Hill 1981). This approach will estimate N_e over more recent history than DNA sequence heterozygosity (Hayes et al. 2003) and can therefore complement evolutionary studies of human populations. Until recently it has not been possible to estimate N_e from LD due to the large number of closely linked markers required to do so.

In this study we estimated genome-wide N_e from LD using data from ~1,000,000 SNPs (HapMap project [The International HapMap Consortium 2003], data release #16 [http://www.hapmap.org/genotypes/2005-03_16a_phaseI/]) in four different human populations of African, Asian, and European ascent. Ours is the first example of using LD to estimate the effective population size of human chromosomes.

LD between each pair of SNPs depends on both N_e and the recombination rate between the SNPs. The distances between SNPs that we used (5–100 kb) are too small to estimate recombination rate using pedigree-based linkage analysis, so we have used other methods. Since errors in estimates of recombination rates from population data might bias the estimate of N_e, we have used three different methods to estimate these recombination rates. Each method resulted in very similar estimates of effective population size.

Methods

All our analyses were based on the known approximate relationship between LD, as measured by r^2, the squared correlation of allele frequencies at a pair of loci, and N_e. In particular, we used $E(r^2) \approx 1/(\alpha + 4N_e c) + 1/n$ for markers on the same autosome, where c is the recombination rate between the SNPs and n is the chromosome experimental sample size. The constant $\alpha = 1$ in the absence of mutation (Sved 1971) and $\alpha \approx 2$ if mutation is taken into account (Hill 1975; Weir and Hill 1980; McVean 2002). We first describe how these formulae were derived and then how this theory was applied to the estimation of N_e from SNP data from multiple population samples. Although formulae for the expectation of r^2 have been published, for completeness we include succinct derivations.

Relationship between N_e and $E(r^2)$ without mutation

Given the correlation of the frequency of alleles at two autosomal loci at generation t (r_t), the mean and variance of the correlation at generation $t + 1$ is

$$E(r_{t+1}) = (1 - c)r_t \quad \text{and}$$

$$\mathrm{var}(r_{t+1}) = [1 - E^2(r_{t+1})]/2N_e = [1 - (r_t^2)(1 - c)^2]/2N_e.$$

The latter expression uses the general expression for the sampling

[8]**Corresponding author.**
E-mail peter.visscher@qimr.edu.au; fax +61-7-3362-0101.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6023607.

variance of an estimate of a correlation coefficient r with sample size n; i.e., $\mathrm{var}(r) = [1 - \mathrm{E}^2(r)]/n$. Using $\mathrm{E}(x^2) = \mathrm{E}(x)^2 + \mathrm{var}(x)$ for a random variable x (for example, see Lynch and Walsh 1998) results in

$$\mathrm{E}(r_{t+1}{}^2) = (1 - c)^2 r_t{}^2 + [1 - (r_t{}^2)(1 - c)^2]/2N_e.$$

At equilibrium, $\mathrm{E}(r_{t+1}{}^2) = \mathrm{E}(r_t{}^2) = \mathrm{E}(r^2)$, so $\mathrm{E}(r^2) = (1 - c)^2\mathrm{E}(r^2) + [1 - \mathrm{E}(r^2)(1 - c)^2]/2N_e$. Rearranging and approximating $(1 - c)^2$ by $(1 - 2c)$ gives $\mathrm{E}(r^2) = 1/(1 + 4N_ec)$. This result was first reported by Sved (1971).

For the X chromosome, recombination occurs only in females. The X chromosome in males may have recombined (since it is a maternal chromosome), and only the maternal X chromosomes in females may have recombined. Hence, two-thirds of X chromosomes may have recombined and one-third may not. The sample size for the disequilibrium (correlation) coefficient is $(3/2)N_e$ because females produce N_e X gametes and males produce $(1/2)N_e$ gametes. Hence,

$$\mathrm{E}(r_{t+1}) = (2/3)(1 - c)r_t + (1/3)r_t = [1 - (2/3)c]r_t,$$

$$\mathrm{var}(r_{t+1}) = \{1 - [1 - (2/3)c]^2 r_t{}^2\}/[(3/2)N_e], \quad \text{and}$$

$$\mathrm{E}(r_{t+1}{}^2) = [1 - (2/3)c]^2 r_t{}^2 + \{1 - [1 - (2/3)c]^2 r_t{}^2\}/(3/2)N_e.$$

At equilibrium, and ignoring the smaller terms, $\mathrm{E}(r^2) = 1/(1 + 2N_ec)$.

Relationship between N_e and E(r^2) with mutation

For autosomal loci, Hill (1975) showed that, in the presence of mutation, $\mathrm{E}(r^2) \approx (10 + \rho)/(22 + 13\rho + \rho^2)$, with $\rho = 4N_ec$. Since $(22 + 13\rho + \rho^2)$ factors into $(11 + \rho)(2 + \rho)$, a further approximation is $\mathrm{E}(r^2) \approx 1/(2 + \rho) \approx 1/(2 + 4N_ec)$. For the X chromosome, following the same logic as before, $\mathrm{E}(r^2) \approx 1/(2 + 2N_ec)$.

Chance LD due to finite experimental sample size

Weir and Hill (1980) showed that experimental sampling introduces chance disequilibrium of $\mathrm{var}(r) = 1/n$ and they suggested the adjustment of $\mathrm{E}(r^2)$ for chromosome sample size. Taking both experimental and evolutionary sampling effects into account, we can summarize the relationship between LD and N_e in the general expression

$$\mathrm{E}(r^2) = 1/(\alpha + kN_ec) + 1/n,$$

where $\alpha = 1$ in the absence of mutation, $\alpha = 2$ if mutation is taken into account, $k = 4$ for autosomes, and $k = 2$ for the X chromosome. In data applications, we observe r^2 and, assuming that we know c or have a good estimate thereof, N_e can be estimated for autosomes and the X chromosome.

Data

HapMap samples from four different populations were available (http://www.hapmap.org/genotypes/2005-03_16a_phaseI/). Samples were 30 trios from CEPH (CEU) families; 30 trios from Yoruba in Ibadan, Nigeria (YRI); and 45 unrelated Japanese (JPT) and Han Chinese (HCB) individuals. Chromosome sample size (n) was ~120 for the CEU and YRI samples and 90 for the JPT and HCB samples. For more information, refer to The International HapMap Consortium (2003).

To compare LD across two samples from approximately the same population, data generated by Perlegen Sciences for European Americans ($n = 46$) were also obtained (http://genome.perlegen.com/browser/download.html).

Haplotype frequency and r^2 estimation

For each chromosome, pairwise r^2 was calculated (Hill and Robertson 1968) only for SNP pairs between 5 kb and 100 kb apart both to avoid the influence of gene conversion on observed LD at SNPs that are closer (Frisse et al. 2001) and to minimize the effect of a very recent expansion of the effective population size on LD (Hayes et al. 2003). A combined EM/Lander-Green algorithm to estimate pairwise haplotype frequencies (as implemented in Haploview; Barrett et al. 2005**;** http://www.broad.mit.edu/mpg/haploview/) was used to estimate r^2 for all autosomes for the HapMap data. For the X chromosome, an EM algorithm combining phase known and unknown data was used. The software for the X chromosome calculations is freely available at http://homepages.ed.ac.uk/eanv63/. SNPs were rejected if their P-value for Hardy-Weinberg equilibrium (HWE) was <0.001 (the default setting in Haploview) or if their minor allele frequency (MAF) was <0.05.

For the Perlegen data, standard EM algorithms were applied to estimate haplotype frequencies and these used to estimate r^2 for all autosomes. We filtered out markers with a minor allele frequency <0.05 and estimated r^2 for all pairs of markers formed by markers that were between 5 kb and 100 kb apart. A total of 866,949 pairwise r^2 estimates were in common with the CEU HapMap sample.

Estimation of recombination rates using three methods

The estimation of the recombination rate for pairs of markers is important because given the value of the recombination rate, effective population size can be estimated from the relationship between r^2 and N_ec. To verify the robustness of our estimates of N_e, we estimated recombination rates using three different methods.

Method 1

We obtained estimates of recombination rate from LD and the known map length of the chromosome. For each chromosome the pairwise r^2 was calculated for all pairs of SNP in sliding windows of 100 kb with a 50 kb overlap. The average recombination rate for each 100 kb window was estimated from the average LD within that window. These average recombination rates were scaled so that over a whole chromosome they add up to the known map length of the chromosome. The recombination rate for any pair of SNPs within a 100 kb window was estimated to be proportional to the physical distance between the SNPs and the average recombination rate over the window. This approach ignores variation in recombination rate within a 100 kb window.

More specifically, to obtain estimates of the recombination rate for any pair of SNPs, we fitted for each window the nonlinear model $y_{ij} = 1/(\alpha_i + \beta_i d_{ij}) + \mathrm{e}_{ij}$ with $y_{ij} = (r^2 - 1/n)$, that is, r^2 adjusted for chromosome sample size, for SNP pair j in window i at physical distance d_{ij}. Parameters α_i and β_i were estimated iteratively using least squares. A restriction was imposed that $\alpha_i \geq 1$ and $\beta_i \geq 0$. If the recombination rate within window i (c_i) is constant, then $\beta_i = 4N_ec_i/d = \rho_i/d$ is the scaled recombination rate per unit of physical distance. For each window, we calculated the scaled recombination rate of the entire window as $\rho_i = 0.05\beta_i$, where 0.05 corresponds to the length (in Mb) of the nonoverlapping part of the window. This quantity is summed over all windows and equated to the known map length (L in Morgan, from pedigree data; Kong et al. 2004; http://compgen.rutgers.edu/maps/index.shtml). A calibration constant, x, was estimated using the number of pairs in a window as weight, i.e., $\hat{x} = m[\Sigma n_i\hat{\rho}_i]/[L\Sigma n_i]$, with m the number of windows and n_i the number of pairs in a window. For each window, the recombination rate per Mb was estimated as $\hat{c}_i^* = \hat{\beta}_i/\hat{x}$. At least 25 pairs were required to estimate local recombination rate in a window.

Given these estimates of local recombination rates, a nonlinear least squares regression method (details below) was subsequently used to estimate N_e from recombination distance between all pairs of markers. For a given pair of markers, the recombination distance was calculated from the estimated recombination rate per unit of physical distance of the window that was the midpoint of the location of the pair and the physical distance between the pair (i.e., $\hat{c}_{\ i(jk)} = \hat{c}_i^* d_{jk}$). A pair was included only if the intermarker distance was <100 kb and if the number of pairs of observations that were used to estimate the local recombination rate for the window was at least 25. The number of pairs of SNPs that were used to estimate the recombination rate in the window was used as a weight in the regression analysis.

Method 2

The recombination fraction between all pairs of markers was estimated using the method described by Clarke and Cardon (2005), and the Kosambi map function was used to convert it to map distance.

Method 3

We obtained fine-scale recombination rates from two independent sets of publicly available data (Phase I HapMap [Altshuler et al. 2005] data [http://www.hapmap.org/downloads/recombination/latest/] and data generated by Perlegen Sciences [Myers et al. 2005; http://www.stats.ox.ac.uk/mathgen/Recombination.html]). We extracted marker pairs for which estimates of recombination rates were available in both data sets. This yielded 98,399 pairs that were formed from 169,545 individual markers. We used these data in four different ways to estimate N_e, using (1) r^2 estimated from HapMap data (HM) and recombination rates estimated from Perlegen data (PL), (2) r^2 and recombination rates estimated from HM, (3) r^2 and recombination rates estimated from PL, and (4) r^2 estimated from PL and recombination rates estimated from HM.

Estimation of chromosome effective population size

Given the formulae described in the theory section, and knowing r^2 and c, we estimated N_e for each chromosome by fitting the nonlinear regression model

$$y_i = 1/(\alpha + \beta c_i) + e_i,$$

with $y_i = (r^2 - 1/n)$, that is, r^2 adjusted for chromosome sample size, for SNP pair i at recombination distance c_i (in Morgans). Parameters α and β were estimated iteratively using least squares.

Heterozygosity and LD in a population depend on N_e over the history of the population. However, LD between SNPs a large distance apart reflects more recent N_e than LD between SNPs closer together (Hayes et al. 2003). We therefore investigated the change in population size over time, as LD between loci with a recombination rate of c reflects the ancestral effective population size $1/(2c)$ generations ago (Hayes et al. 2003) (under the assumption of linear growth). We compared N_e estimated by recombination estimation method 1 from SNPs 5–100 kb apart. This corresponds to a time between ~10,000 and 500 generations ago, i.e., between 200,000 and 10,000 yr ago (Hayes et al. 2003). For each chromosome, we estimated N_e from the mean r^2 for average SNP recombination distances in the range of 0.01–0.5 cM. For each distance, the corresponding number of generations in the past was calculated.

Estimates of the scaled recombination rate and effective population size per chromosome obtained by method 1 were compared with the number of genes and length of the chromosome (NCBI build 35; http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606&build=previous) using correlation and linear regression.

The frequency distribution of the SNPs ascertained by the HapMap project is different from the frequency distribution of SNPs that have been completely ascertained (Nielsen et al. 2004). To determine if the HapMap SNP ascertainment procedure was biasing our estimates of N_e, we carried out coalescent (Hudson 1983) simulations for a neutral model and constant population size. Simulations were performed using the program "ms" (Hudson 2002; http://home.uchicago.edu/~rhudson1/source/mksamples.html). For a segment length of 25 Mb and an effective population size of 2000, the chosen input parameters equate to a mutation rate of 10^{-8} per nucleotide and a recombination rate of 0.01 per Mb. For each replicate, the average r^2 was calculated for bins of 1 kb spacing and adjusted for chromosome sample size. A nonlinear regression model was used to relate the adjusted r^2 to physical distance. Under this model, the estimate of the regression coefficient is an estimate of the scaled recombination rate ρ. One thousand replicates were run.

Results

Figure 1 shows, as expected, a near linear relationship between $1/r^2$ and physical distance between pairs of SNPs. Linearity of the reciprocal of r^2 with physical distance at small values, and a concave relationship at larger values, is clearly shown, consistent with a population that has increased over time.

For the first method of estimating c, the average estimates of effective population size for CEU are similar to those for JPT and HCB, but lower than those for YRI (Table 1). This is expected under the out-of-Africa theory of ancestral human population expansion (Templeton 2002), because, when humans moved out of Africa, only a subset of the amount of genetic variation present in the African population at that time was represented in the migrants. An ANOVA on the estimates of N_e, fitting population

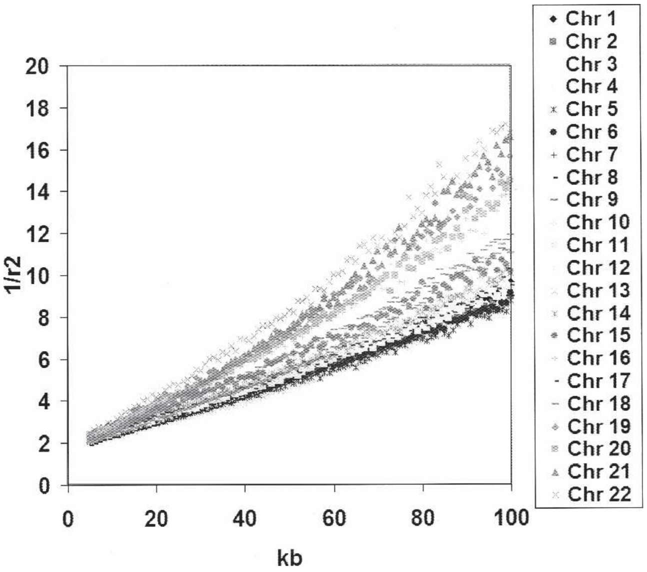

Figure 1. Reciprocal of r^2 plotted against physical distance for each of the 22 autosomes for the CEU population. The *X*-axis shows the physical distance in kilobases and the *Y*-axis shows 1/mean(r^2). The mean value of r^2 was calculated in 1 kb bins.

Table 1. *N_e* **estimates from the nonlinear model (Method 1)**

	Population							
	CEU		YRI		JPT		HCB	
Chromosome	*n*	N_e	*n*	N_e	*n*	N_e	*n*	N_e
1	1,350,336	2824	1,602,447	6197	1,155,365	2549	1,170,837	2694
2	1,770,347	3197	1,958,708	6675	1,382,444	3200	1,404,237	3294
3	1,318,345	3085	1,300,983	6897	936,628	2534	949,450	2743
4	1,052,614	3232	1,086,591	6351	772,233	3226	790,459	3312
5	1,074,385	3116	974,226	6932	795,412	2554	803,800	2762
6	1,323,089	3031	1,359,998	6745	1,101,243	2678	1,114,544	2854
7	946,353	2196	816,514	6745	600,416	2739	613,743	2897
8	2,292,984	3029	2,531,320	7004	2,036,130	2650	2,061,705	2802
9	1,779,571	3048	1,928,452	7019	1,564,337	2678	1,572,613	2830
10	919,144	3227	1,044,438	6868	770,465	3036	782,584	2860
11	789,664	3027	771,232	6181	613,884	2520	614,225	2669
12	987,085	3157	993,895	7007	733,950	2812	740,215	2919
13	660,098	2031	795,844	4806	547,950	1928	554,326	2041
14	521,596	2232	495,815	4719	390,544	1983	397,645	2134
15	436,105	1933	465,532	4143	346,811	1830	354,237	1878
16	417,032	3017	419,655	7334	324,468	2825	324,315	2935
17	411,015	2935	387,225	6384	307,220	2843	312,735	2752
18	1,223,437	2409	1,459,903	5771	1,008,782	2315	1,029,745	2365
19	270,281	3126	259,975	7416	208,095	2656	218,568	2608
20	393,509	3338	355,554	6864	270,347	2534	275,690	2795
21	656,672	1485	748,128	3855	663,630	1440	672,430	1574
22	619,770	1459	669,718	3246	587,244	1217	584,544	1081
X	924,274	3613	1,117,100	9421	738,824	3141	788,186	3458
Mean	**962,509**	**2772**	**1,023,620**	**6286**	**776,366**	**2517**	**788,297**	**2620**
SD	**513,402**	**598**	**581,811**	**1357**	**447,337**	**524**	**452,066**	**557**

(*n*) Number of marker pairs; (N_e) effective population size. Intermarker distance was in the range of 5 kb to100 kb for all SNP pairs.

and chromosome as factors, resulted in significance (P < 0.001) for both population and chromosome. The correlation between samples of N_e obtained from different chromosomes ranged from 0.87 (JPT, YRI) to 0.98 (JPT, HCB). Figure 2 shows the estimates of N_e from each of the 23 chromosomes for the JPT plotted against those estimated for the HCB. The remarkable similarity between JPT and HCB estimates most likely indicates common ancestry.

There were significantly different from zero and positive correlations between N_e and chromosome length in Mb (with significance values ranging from $P = 0.002$ for HCB to $P = 0.06$ for YRI) and number of genes (with significance values ranging from $P = 0.02$ for CEU to $P = 0.05$ for YRI), but not between N_e and gene density (with significance values $P > 0.6$ for all four populations). The significant correlations were driven by the low estimate of N_e for the short chromosomes 21 and 22.

The second method, which estimates recombination rates between each pair of adjacent HapMap markers from a model-free method that detects recombination hotspots from LD (Clarke and Cardon 2005; Visscher and Hill 2006), changed the estimate of N_e between +33% and −45% with an average reduction of 27% (mean $N_e = 1901$) when compared with that obtained from the first method (results not shown in Tables).

The third method used estimates of fine-scale recombination rates (rather than using physical distance as a proxy for recombination rate) from coalescent models from either Phase I HapMap (Altshuler et al. 2005) data or data generated by Perlegen Sciences (Myers et al. 2005). These estimates are shown in Table 2 and are consistent with estimates presented in Table 1.

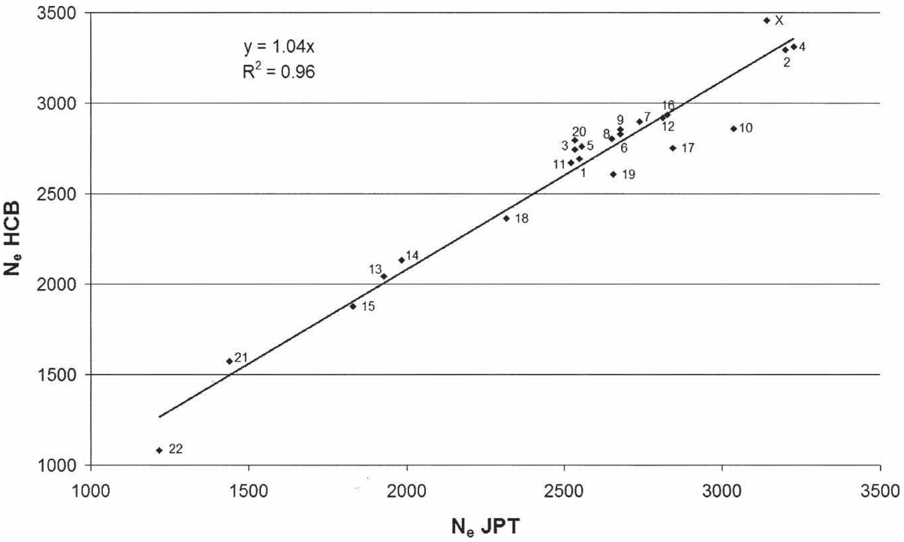

Figure 2. Comparison of the estimate of N_e from each of the labeled 23 chromosomes for the JPT and HCB population (Method 1). The *X*-axis shows the effective population size for each autosome and the X chromosome estimated from the JPT sample using a nonlinear regression of r^2 on physical distance between pairs of SNPs. The *Y*-axis shows the estimates of N_e for each chromosome from HCB sample. The fitted linear regression line fits nearly perfectly, presumably reflecting common ancestry of the two populations.

Table 2. N_e **estimates from the European ancestry samples from HapMap (HM,** $n = 120$**) and Perlegen (PL,** $n = 46$**) estimates of fine-scale recombination rates (Method 2)**

Chromosome	HMPL	HMHM	PLPL	PLHM	Mean
1	3383	2909	3362	2863	3129
2	3390	3432	3429	3405	3414
3	3383	3070	3399	3044	3224
4	3531	3172	3869	3416	3497
5	3177	3205	3332	3272	3247
6	3864	3477	3811	3397	3637
7	3000	3223	2760	3006	2997
8	2709	2619	2776	2664	2692
9	2576	2108	2521	2033	2310
10	2776	2591	2701	2487	2639
11	2656	2809	2512	2635	2653
12	4023	3059	3769	2851	3426
13	2443	2473	2467	2460	2461
14	2720	3007	2777	2973	2869
15	1901	1957	1808	1813	1870
16	2141	2320	2223	2325	2252
17	2789	2767	2643	2612	2703
18	3109	2806	3097	2731	2936
19	2193	2135	2163	2077	2142
20	2239	1873	2170	1792	2019
21	2496	2102	2450	1987	2259
22	2322	1813	2496	1863	2124
Mean	**2856**	**2679**	**2843**	**2623**	**2750**

The first two letters of each header indicate the sample (HM or PL) from which the r^2 were estimated, and the last two letters indicate from which sample the fine-scale recombination rates were estimated.

Hence, using three different methods to estimate recombination rate and using two different samples of individuals from European descent gave estimates of the effective population size ranging from 1901 (Method 2) to 2843 (Method 3).

From the simulation study we found that the estimation method was not biased when SNPs were simulated as if they had been completely ascertained (data not shown). We then simulated SNPs to mimic the SNP frequency distribution from the HapMap data. For this, SNPs with minor allele frequencies between 0.05 and 0.5 were ascertained with equal probability; i.e., the frequency distribution of the SNPs was uniform. The estimates of N_e obtained from mimicking the HapMap data were biased downward by ~18% compared with the complete ascertainment data. Hence, the HapMap ascertainment strategy may bias our estimates by approximately one-fifth, giving adjusted estimates of effective population sizes of ~3100 (non-African populations) and ~7500 (African population).

For the CEU and YRI samples we estimated effective population size as a function of time in the past. Results for the CEU data support recent dramatic population growth (Fig. 3A). This is in agreement with the likely demographic history of the ancestral population of the non-African samples; a population bottleneck, following an out-of-Africa expansion, followed by rapid growth (Watkins et al. 2001). Results for the YRI data indicate an ancestral population size of ~7000, followed by expansion in the last ~20,000 yr (Fig. 3B).

Discussion

Overall, the estimates of N_e appear to be much lower than the usually quoted value of 10,000 (Takahata 1993). Earlier studies using mtDNA data suggested an N_e in the range of 1000–6000 (Rogers and Harpending 1992; Harpending et al. 1993; Sherry et al. 1994), for a population ~200,000 yr ago (~10,000 generations ago). Erlich et al. (1996) estimated a recent population size of ~10,000 from HLA polymorphisms. Sherry et al. (1997) estimated an ancestral population size of ~17,800 during the last one to two million yr from *Alu* repeats evolution. Our estimates of N_e were reasonably consistent across chromosomes and methods, and similar to estimates of N_e obtained from LD in 10 small genomic regions in a sample of 15 Italians (Frisse et al. 2001) and from three Y chromosome genes from a worldwide sample of Y chromosomes (Thomson et al. 2000). Relative to estimates of N_e from polymorphism levels, these estimates are approximately two to three times smaller. Why is this the case? We propose that the most likely explanation is that different studies implicitly or ex-

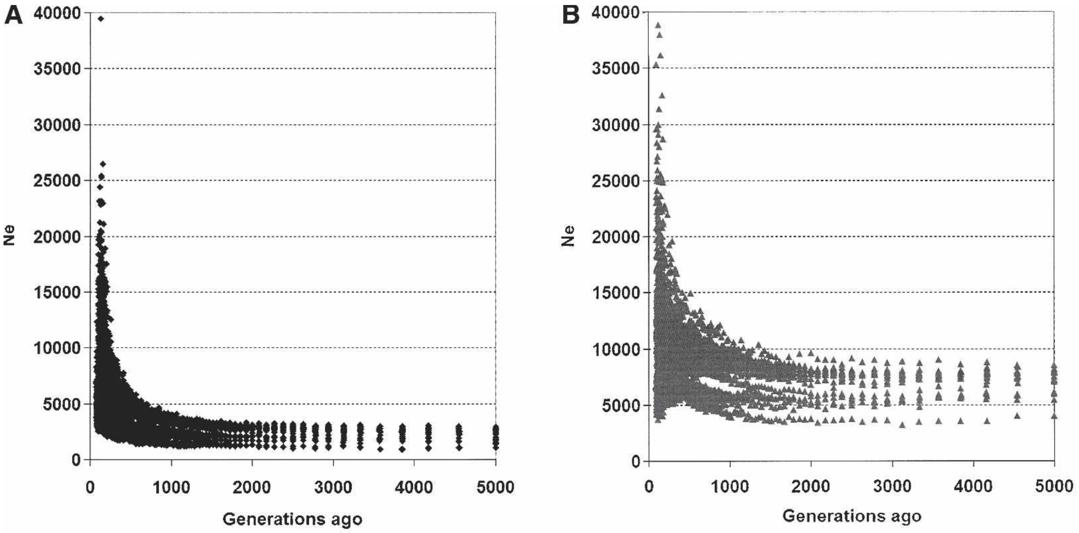

Figure 3. Ancestral population size estimated from 22 autosomes of the CEU (*A*) and YRI (*B*) data (Method 1). For each chromosome and for each SNP recombination distance bin (0.001 cM spacing), N_e was estimated from the mean r^2, using $E(r^2) = 1/(\alpha + 4N_ec)$. The number of generations in the past was calculated as $1/(2c)$, and was truncated at 5000. Effective population size has increased dramatically in the last ~1000 generations (20,000 yr), from a fairly constant ancestral size of ~2500 (CEU) and ~7000 (YRI).

plicitly have estimated N_e at a different point in time and that the estimates can be reconciled by taking the time element into account. The human population size has not been constant in the last few 100,000 yr and, in addition to an increase in population size in recent times, there has been evidence for population bottlenecks following the out-of-Africa expansion (Reich et al. 2001; Zhang et al. 2004). Population growth and bottlenecks both have an effect on the estimates of effective population size using either marker heterozygosity or LD, but to a different degree. Service et al. (2006) reported heterozygosity and LD for chromosome 22 markers in 12 human populations, of which 11 were isolates (i.e., populations that had recently experienced increased levels of inbreeding). The average level of heterozygosity varied very little across all populations, with a range of 0.359–0.373. However, the amount of LD per unit of physical distance varied nearly twofold. These observations are consistent with heterozygosity reflecting average population size over a long period of time including before bottlenecks (inbreeding) and LD reflecting a more recent population size. Reich et al. (2001) found, using simulations, that in order to explain their European data, the population had to go through a bottleneck with a size substantially smaller than 10,000 individuals. Estimates of N_e from variation at Y chromosome and autosomal microsatellite markers in populations of African, European, and Asian descent were similar to ours (Pritchard et al. 1999; Zhivotovsky et al. 2003). Microsatellites, due to their high mutation rate, reflect more recent population history than SNPs. Using a model that incorporates geographic distances among populations, as well as genetic data, Liu et al. (2006) inferred an N_e of ~1000 for the founding population from which modern humans derive. Other anthropological and genetic evidence has also suggested that the long-term N_e has been about three times larger in African populations than in non-African populations (Relethford and Harpending 1994; Relethford and Jorde 1999; Eller 2001), which is what we observed.

Our estimate of N_e for the X chromosome was 30%–50% larger than that for the autosomes. The X chromosome in humans has a number of unusually long haplotypes (Altshuler et al. 2005). However, estimates from coalescent methods using the same data also give an increase of ~50% in the estimate of N_e for the X chromosome compared with autosomes, although this difference disappears when using HapMap Phase II data (G. McVean, pers. comm.). It is not clear why the average LD, when adjusted for the absence of recombination in males, is smaller for the X chromosome.

We determined by simulation how the approximate ascertainment of SNPs in HapMap Phase I could bias our estimates of N_e, and adjusted these accordingly. Recently, Pe'er et al. (2006) have reported small upward biases in the estimation of LD from HapMap I data, consistent with a downward bias in the estimate of N_e. These biases would also affect our analyses and would not have been fully corrected for by our adjustment, which was based upon the allele frequencies of the ascertained SNPs. In addition, if there is variation in recombination rate that has not been reflected in our estimates of c using the three different methods, then our estimate of N_e would be biased downward.

In populations in which effective population size has changed over time, such as human populations, it is not meaningful to discuss effective population size without reference to a point in time (Hayes et al. 2003). For example, assuming a constant N_e when it has increased over time and estimating it from data on marker heterozygosity will result in an estimate of an average N_e over long periods of time, before bottlenecks if these have occurred recently. Methods, including the coalescent-based ones, that fail to take into account that N_e has changed over time will produce biased population parameter estimates, in particular when inference depends on the observed relationship between recombination distance and linkage disequilibrium.

We have used a relatively small sample of individuals, combined with high-density genome-wide marker genotyping, to infer ancestral population size based upon the observed amounts of LD. Our study has shown that human effective population size estimated from entire human chromosomes is considerably lower than previously suggested, at least during a bottleneck up to ~20,000 yr ago when a large expansion began.

Acknowledgments

We thank W.G. Hill for helpful discussions and help in the derivation of X-linked N_e, and N. Barton, B. Weir, J. Taylor, G. McVean, T. Johnson, and A. Morris for helpful discussions. A.T. acknowledges Cancer Research UK; P.N. was supported by the Genes to Cognition Project; and P.M.V. acknowledges the UK Biotechnology and Biological Sciences Research Council, the Wellcome Trust, and the Australian National Health and Medical Research Council for funding.

References

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., and Donnelly, P. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Clarke, G.M. and Cardon, L.R. 2005. Disentangling linkage disequilibrium and linkage from dense single-nucleotide polymorphism trio data. *Genetics* **171:** 2085–2095.

Eller, E. 2001. Estimating relative population sizes from simulated data sets and the question of greater African effective size. *Am. J. Phys. Anthropol.* **116:** 1–12.

Erlich, H.A., Bergstrom, T.F., Stoneking, M., and Gyllensten, U. 1996. HLA sequence polymorphism and the origin of humans. *Science* **274:** 1552–1554.

Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J., and Di Rienzo, A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69:** 831–843.

Harpending, H.C., Sherry, S.T., Rogers, A.R., and Stoneking, M. 1993. The genetic structure of ancient human populations. *Curr. Anthropol.* **34:** 483–496.

Hayes, B.J., Visscher, P.M., McPartlan, H.C., and Goddard, M.E. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13:** 635–643.

Hill, W.G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor. Popul. Biol.* **8:** 117–126.

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

Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38:** 226–231.

Hudson, R.R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23:** 183–201.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337–338.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

Kong, X., Murphy, K., Raj, T., He, C., White, P.S., and Matise, T.C. 2004. A combined linkage–physical map of the human genome. *Am. J. Hum. Genet.* **75:** 1143–1148.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. 2006. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79:** 230–237.

Lynch, M. and Walsh, B. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA.

McVean, G.A.T. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics* **162:** 987–991.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.

Nielsen, R., Hubisz, M.J., and Clark, A.G. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168:** 2373–2382.

Pe'er, I., Chretien, Y.R., de Bakker, P.I.W., Barrett, J.C., Daly, M.J., and Altshuler, D.M. 2006. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hum. Genet.* **78:** 588–603.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., and Feldman, M.W. 1999. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16:** 1791–1798.

Reich, D.E. and Lander, E.S. 2001. On the allelic spectrum of human disease. *Trends Genet.* **17:** 502–510.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199–204.

Relethford, J.H. and Harpending, H.C. 1994. Craniometric variation, genetic theory, and modern human origins. *Am. J. Phys. Anthropol.* **95:** 249–270.

Relethford, J.H. and Jorde, L.B. 1999. Genetic evidence for larger African population size during recent human evolution. *Am. J. Phys. Anthropol.* **108:** 251–260.

Rogers, A.R. and Harpending, H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9:** 552–569.

Service, S., DeYoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A., et al. 2006. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.* **38:** 556–560.

Sherry, S.T., Rogers, A.R., Harpending, H., Soodyall, H., Jenkins, T., and Stoneking, M. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.* **66:** 761–775.

Sherry, S.T., Harpending, H.C., Batzer, M.A., and Stoneking, M. 1997. Alu evolution in human populations: Using the coalescent to estimate effective population size. *Genetics* **147:** 1977–1982.

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2:** 125–141.

Takahata, N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10:** 2–22.

Templeton, A.R. 2002. Out of Africa again and again. *Nature* **416:** 45–51.

Thomson, R., Pritchard, J.K., Shen, P.D., Oefner, P.J., and Feldman, M.W. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci.* **97:** 7360–7365.

Visscher, P.M. and Hill, W.G. 2006. Estimation of recombination rate and detection of recombination hotspots from dense single-nucleotide polymorphism trio data. *Genetics* **173:** 2415–2417.

Watkins, W.S., Ricker, C.E., Bamshad, M.J., Carroll, M.L., Nguyen, S.V., Batzer, M.A., Harpending, H.C., Rogers, A.R., and Jorde, L.B. 2001. Patterns of ancestral human diversity: An analysis of Alu-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* **68:** 738–752.

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

Zhang, W.H., Collins, A., Gibson, J., Tapper, W.J., Hunt, S., Deloukas, P., Bentley, D.R., and Morton, N.E. 2004. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl. Acad. Sci.* **101:** 18075–18080.

Zhivotovsky, L.A., Rosenberg, N.A., and Feldman, M.W. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72:** 1171–1186.

Received October 9, 2006; accepted in revised form January 17, 2007.