Recent human effective population size estimated from linkage disequilibrium

Citation for published version:

Digital Object Identifier (DOI):
10.1101/gr.6023607

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Genome Research

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Recent human effective population size estimated from linkage disequilibrium

Albert Tenesa, Pau Navarro, Ben J. Hayes, et al.

*Genome Res.* 2007 17: 520-526 originally published online March 9, 2007
Access the most recent version at doi:10.1101/gr.6023607

This article cites 38 articles, 15 of which can be accessed free at: [http://genome.cshlp.org/content/17/4/520.full.html#ref-list-1](http://genome.cshlp.org/content/17/4/520.full.html#ref-list-1)

Article cited in: [http://genome.cshlp.org/content/17/4/520.full.html#related-urls](http://genome.cshlp.org/content/17/4/520.full.html#related-urls)

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see [http://genome.cshlp.org/site/misc/terms.xhtml](http://genome.cshlp.org/site/misc/terms.xhtml)). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at [http://creativecommons.org/licenses/by-nc/3.0/](http://creativecommons.org/licenses/by-nc/3.0/).

Email alerting service: Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/subscriptions)
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. Visscher3,5,8

1Colon Cancer Genetics Group, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; 2MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; 3Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom; 4Institute of Animal Science, DPI, Atwood 3049, Australia; 5Queensland Institute of Medical Research, Royal Brisbane Hospital, Brisbane 4006, Australia; 6The Wellcome Trust Centre for Human Genetics, The University of Oxford, Oxford OX3 7BN, United Kingdom; 7Institute 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. Chrome-specific estimates of 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), 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 ~3,000, whereas the estimate for the YRI was ~7,500, 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.

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) = 1/(a + 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 a = 1 in the absence of mutation (Sved 1971) and a = 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

and

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

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

At equilibrium, \( E(r_{t+1}^2) = E(r_t^2) = E(r^2) \), so \( E(r^2) = (1 - c^2)E(r^2) + [1 - E(r^2)](1 - c^2)/2N_c \). Rearranging and approximating \((1 - c^2)^2 \) by \((1 - 2c)\) gives \( E(r^2) = 1/(1 + 4N_c) \). 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 \((3/2)N_c\), because females produce \( N_c \) X gametes and males produce \((1/2)N_c \) gametes. Hence,

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

\[
\text{var}(r_{t+1}^2) = \frac{2(1 - c)}{3(1 - c)^2}E(r^2)[(3/2)N_c],
\]

and

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

At equilibrium, and ignoring the smaller terms, \( E(r^2) = 1/(1 + 2N_c) \).

**Relationship between \( N_c \) and \( E(r^2) \) with mutation**

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

**Chance LD due to finite experimental sample size**

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

\[
E(r^2) = 1/(\alpha + kN_c) + 1/n,
\]

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, \( N_c \) 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_phase1/). 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. 2005). A combined EM/Lander-Green algorithm to estimate pairwise haplotype frequencies (as implemented in Haplov; Barrett et al. 2005; http://www.broad.mit.edu/mpg/haplov/) 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/eamv63/. SNPs were rejected if their P-value for Hardy-Weinberg equilibrium (HWE) was \(< 0.05 \) (the default setting in Haplov) 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 \( N_c \). To verify the robustness of our estimates of \( N_c \), 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_y = 1/(\alpha_y + \beta_yd_y + \epsilon_y \text{ with } y_y = (r^2 - 1)/m), \text{ that is, } r^2 \text{ adjusted for chromosome sample size, for SNP pair } i \text{ in window } j \text{ at physical distance } d_y \text{. Parameters } \alpha_y \text{ and } \beta_y \text{ were estimated iteratively using least squares. A restriction was imposed that } \alpha_y \geq 1 \text{ and } \beta_y \geq 0. \text{ If the recombination rate within window } i \text{ is constant, then } \beta_y = 4N_c\lambda/d = \rho_i/d \text{ 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_y \text{, where } 0.05 \text{ 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, \text{ was estimated using the number of pairs in a window as weight, i.e., } \hat{x} = m[S_n\rho_i]/[LS_n], \text{ with } m \text{ the number of windows and } n, \text{ the number of pairs in a window. For each window, the recombination rate per Mb}


was estimated as $c_i^* = \hat{\beta}_i / \hat{\alpha}$. 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., $c_{ij} = \hat{c}_i d_{jk}$). 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 $y_i = (r^2_i - 1/n)$, that is, $r^2$ adjusted for chromosome sample size, for SNP pair $i$ at recombination distance $c_i$ (in Morgans). Parameters $\alpha$ and $\beta$ 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 $\rho$. 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
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.

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

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>CEU</th>
<th></th>
<th>YRI</th>
<th></th>
<th>JPT</th>
<th></th>
<th>HCB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$N_e$</td>
<td>$n$</td>
<td>$N_e$</td>
<td>$n$</td>
<td>$N_e$</td>
<td>$n$</td>
<td>$N_e$</td>
</tr>
<tr>
<td>1</td>
<td>1,350,336</td>
<td>2824</td>
<td>1,602,447</td>
<td>6197</td>
<td>1,55,365</td>
<td>2549</td>
<td>1,70,837</td>
<td>2694</td>
</tr>
<tr>
<td>2</td>
<td>1,770,347</td>
<td>3197</td>
<td>1,958,708</td>
<td>6675</td>
<td>1,382,444</td>
<td>3200</td>
<td>1,404,237</td>
<td>3294</td>
</tr>
<tr>
<td>3</td>
<td>1,318,345</td>
<td>3085</td>
<td>1,300,983</td>
<td>6897</td>
<td>936,268</td>
<td>2534</td>
<td>949,450</td>
<td>2743</td>
</tr>
<tr>
<td>4</td>
<td>1,052,614</td>
<td>3232</td>
<td>1,086,591</td>
<td>6351</td>
<td>772,233</td>
<td>3226</td>
<td>790,450</td>
<td>3312</td>
</tr>
<tr>
<td>5</td>
<td>1,074,385</td>
<td>3116</td>
<td>974,226</td>
<td>6932</td>
<td>795,412</td>
<td>2554</td>
<td>803,800</td>
<td>2762</td>
</tr>
<tr>
<td>6</td>
<td>1,323,089</td>
<td>3031</td>
<td>1,359,998</td>
<td>6745</td>
<td>1,101,243</td>
<td>2679</td>
<td>1,114,544</td>
<td>2854</td>
</tr>
<tr>
<td>7</td>
<td>946,353</td>
<td>2196</td>
<td>816,514</td>
<td>6745</td>
<td>600,416</td>
<td>2739</td>
<td>613,743</td>
<td>2897</td>
</tr>
<tr>
<td>8</td>
<td>2,292,984</td>
<td>3029</td>
<td>2,531,320</td>
<td>7004</td>
<td>2,036,130</td>
<td>2650</td>
<td>2,061,705</td>
<td>2802</td>
</tr>
<tr>
<td>9</td>
<td>1,779,571</td>
<td>3048</td>
<td>1,928,452</td>
<td>7019</td>
<td>1,564,337</td>
<td>2679</td>
<td>1,572,613</td>
<td>2830</td>
</tr>
<tr>
<td>10</td>
<td>919,144</td>
<td>3227</td>
<td>1,044,438</td>
<td>6868</td>
<td>770,465</td>
<td>3036</td>
<td>782,584</td>
<td>2860</td>
</tr>
<tr>
<td>11</td>
<td>789,664</td>
<td>3027</td>
<td>771,232</td>
<td>6181</td>
<td>613,884</td>
<td>2520</td>
<td>614,225</td>
<td>2669</td>
</tr>
<tr>
<td>12</td>
<td>987,085</td>
<td>3157</td>
<td>993,895</td>
<td>7007</td>
<td>733,950</td>
<td>2812</td>
<td>740,215</td>
<td>2919</td>
</tr>
<tr>
<td>13</td>
<td>660,098</td>
<td>2031</td>
<td>795,844</td>
<td>4806</td>
<td>547,950</td>
<td>1928</td>
<td>554,326</td>
<td>2041</td>
</tr>
<tr>
<td>14</td>
<td>521,596</td>
<td>2232</td>
<td>495,815</td>
<td>4719</td>
<td>390,544</td>
<td>1983</td>
<td>397,645</td>
<td>2134</td>
</tr>
<tr>
<td>15</td>
<td>436,105</td>
<td>1933</td>
<td>465,322</td>
<td>4143</td>
<td>346,811</td>
<td>1830</td>
<td>354,237</td>
<td>1878</td>
</tr>
<tr>
<td>16</td>
<td>417,032</td>
<td>3017</td>
<td>419,655</td>
<td>7334</td>
<td>324,468</td>
<td>2825</td>
<td>324,315</td>
<td>2935</td>
</tr>
<tr>
<td>17</td>
<td>411,015</td>
<td>2955</td>
<td>387,225</td>
<td>6384</td>
<td>307,220</td>
<td>2843</td>
<td>312,735</td>
<td>2752</td>
</tr>
<tr>
<td>18</td>
<td>1,223,437</td>
<td>2409</td>
<td>1,459,903</td>
<td>5771</td>
<td>1,008,782</td>
<td>2315</td>
<td>1,029,745</td>
<td>2365</td>
</tr>
<tr>
<td>19</td>
<td>270,281</td>
<td>3126</td>
<td>259,975</td>
<td>7416</td>
<td>208,095</td>
<td>2656</td>
<td>218,568</td>
<td>2608</td>
</tr>
<tr>
<td>20</td>
<td>393,509</td>
<td>3338</td>
<td>355,544</td>
<td>6864</td>
<td>270,347</td>
<td>2534</td>
<td>275,690</td>
<td>2795</td>
</tr>
<tr>
<td>21</td>
<td>656,672</td>
<td>1485</td>
<td>748,128</td>
<td>3853</td>
<td>663,630</td>
<td>1440</td>
<td>672,430</td>
<td>1574</td>
</tr>
<tr>
<td>22</td>
<td>679,770</td>
<td>1459</td>
<td>669,718</td>
<td>3246</td>
<td>587,244</td>
<td>1217</td>
<td>584,544</td>
<td>1081</td>
</tr>
<tr>
<td>X</td>
<td>924,274</td>
<td>3613</td>
<td>1,117,100</td>
<td>9421</td>
<td>738,824</td>
<td>3141</td>
<td>788,186</td>
<td>3458</td>
</tr>
<tr>
<td>Mean</td>
<td>962,509</td>
<td>2772</td>
<td>1,023,620</td>
<td>6286</td>
<td>776,366</td>
<td>2517</td>
<td>788,297</td>
<td>2620</td>
</tr>
<tr>
<td>SD</td>
<td>513,402</td>
<td>398</td>
<td>581,811</td>
<td>1357</td>
<td>447,337</td>
<td>524</td>
<td>452,066</td>
<td>557</td>
</tr>
</tbody>
</table>

($n$) Number of marker pairs; ($N_e$) effective population size. Intermarker distance was in the range of 5 kb to 100 kb for all SNP pairs.
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-

<table>
<thead>
<tr>
<th>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)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>Mean</td>
</tr>
</tbody>
</table>

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.

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 \(E(\sigma^2) = \frac{1}{4(N_e+4\theta)} \). The number of generations in the past was calculated as \(1/(4\theta)\), 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).
explicitly 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 isolated (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


