

# Human genome sequence variation and the influence of gene history, mutation and recombination

David E. Reich<sup>1\*</sup>, Stephen F. Schaffner<sup>1\*</sup>, Mark J. Daly<sup>1</sup>, Gil McVean<sup>2</sup>, James C. Mullikin<sup>3</sup>, John M. Higgins<sup>1</sup>, Daniel J. Richter<sup>1</sup>, Eric S. Lander<sup>1,4</sup> & David Altshuler<sup>1,5</sup>

\*These authors contributed equally to this work.

Published online: 5 August 2002, DOI: 10.1038/ng947

**Variation in the human genome sequence is key to understanding susceptibility to disease in modern populations and the history of ancestral populations. Unlocking this information requires knowledge of the patterns and underlying causes of human sequence diversity. By applying a new population-genetic framework to two genome-wide polymorphism surveys, we find that the human genome contains sizeable regions (stretching over tens of thousands of base pairs) that have intrinsically high and low rates of sequence variation. We show that the primary determinant of these patterns is shared genealogical history. Only a fraction of the variation (at most 25%) is due to the local mutation rate. By measuring the average distance over which genealogical histories are typically preserved, these data provide the first genome-wide estimate of the average extent of correlation among variants (linkage disequilibrium). The results are best explained by extreme variability in the recombination rate at a fine scale, and provide the first empirical evidence that such recombination ‘hot spots’ are a general feature of the human genome and have a principal role in shaping genetic variation in the human population.**

## Introduction

Each copy of the human genome is unique and differs in sequence from any other copy in the population by roughly 1 in 1,250 nucleotides<sup>1,2</sup>. This variation in DNA sequence influences individual characteristics such as physical appearance, susceptibility to disease and response to medical treatments. Sequence polymorphism also represents a fossil record of the history and structure of ancestral populations. Thus, a central goal of medical and population genetics is to understand the patterns and determinants of sequence variation in the human population.

When averaged across windows of 200 kb, rates of heterozygosity show up to a tenfold variation<sup>1</sup>. But the variation of polymorphism rates at the finer scale (<100 kb) typical of individual genes has not been described. To understand how the human genome varies at this finer scale, we have developed a framework for interpreting population genetic data and have applied it to two genome-wide data sets. This analysis provides three empirical results. First, we show that gene history is quantitatively the main force responsible for local patterns of human genome sequence variation. Second, we make the first genome-wide measurement of the correlations of nearby alleles (linkage disequilibrium). Last, we show that extreme variability in the recombination rate at a

fine scale (<100 kb) is a general and major determinant of local patterns of human genome sequence variation.

## Results

### Regions of high and low polymorphism

We analyzed data from two genome-wide polymorphism discovery projects<sup>1</sup> that together yielded a collection of 1.42 million single-nucleotide polymorphisms (SNPs). The SNP Consortium (TSC) selected clones of different sizes from libraries of genomic DNA (constructed from a multiethnic group of 24 individuals) and sequenced the cloned fragments from either one end ('single reads') or both ends ('paired-end reads'). SNPs were identified<sup>1,3,4</sup> by comparing the resulting sequence with that of the draft genome produced by the Human Genome Project<sup>5</sup>. The 'BAC overlap' projects<sup>1</sup> identified SNPs by comparing the genomic sequences<sup>5</sup> of different individuals that were cloned into bacterial artificial chromosomes (BACs). The data sets included in this analysis were each very large and spanned the human genome, with the TSC data comprising 729 Mb (571,500 heterozygous positions) and the BAC overlap data comprising 46 Mb (37,300 heterozygous positions). When analyzed by a single suite of validated computational methods<sup>1,3,4</sup>, the average polymorphism rate across the genome

<sup>1</sup>Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Cambridge, Massachusetts 02139, USA. <sup>2</sup>Department of Statistics, University of Oxford, Oxford, UK. <sup>3</sup>Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. <sup>4</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>5</sup>Departments of Genetics and Medicine, Harvard Medical School and Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. Correspondence should be addressed to D.A. (e-mail: altshuler@molbio.mgh.harvard.edu).

**Table 1 • Summary of statistics for the data sets**

Quantity <sup>a</sup>	Human polymorphism		Primate divergence
	Calculated from TSC data	Calculated from BAC data	Calculated from GenBank data
Base pairs examined	729 Mb	46.2 Mb	1.01 Mb
Number of variant sites	571,500	37,300	13,224
Mean $[\hat{\pi}]$	$8.01 \times 10^{-4}$	$7.78 \times 10^{-4}$	0.013
Standard deviation in $[\hat{\pi}]$	$1.47 \times 10^{-3}$	$1.59 \times 10^{-3}$	$8.72 \times 10^{-3}$
Coefficient of variation of $[\hat{\pi}]$	1.83	2.05	0.67
Standard deviation in $\pi$	$8.09 \times 10^{-4}$	$8.63 \times 10^{-4}$	$6.69 \times 10^{-3}$
Coefficient of variation of $\pi$	1.01	1.11	0.49
Mean $\mu$	—	—	$2.5 \times 10^{-8}$
Standard deviation in $\mu$	—	—	$1.07 \times 10^{-8}$
Coefficient of variation of $\mu$	—	—	0.54
Mean $\tau$	$1.60 \times 10^4$	$1.55 \times 10^4$	—
Standard deviation in $\tau$	$1.22 \times 10^4$	$1.34 \times 10^4$	—
Coefficient of variation of $\tau$	0.76	0.87	—
% of variance in $\pi$ due to history	0.57	0.61	—

<sup>a</sup>The observed variation in polymorphism rate,  $[\hat{\pi}]$ , is greater than the true underlying rate of variation across the genome  $\pi$ , owing to stochastic variation in the number of SNPs observed over a finite length of sequence (Methods). Both quantities ( $\pi$  and  $\hat{\pi}$ ) are presented here.

was similar between the two projects and in agreement with previous estimates<sup>2,3,6–10</sup>:  $8.01 \times 10^{-4}$  for the TSC data and  $7.78 \times 10^{-4}$  for the BAC overlaps (Table 1).

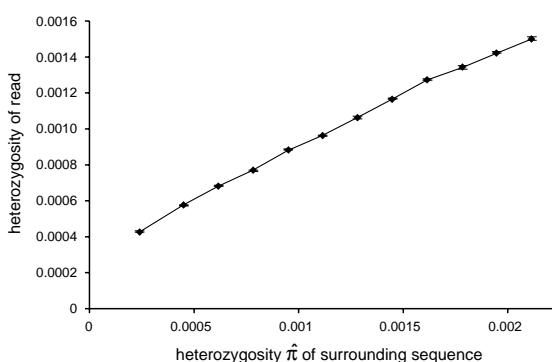
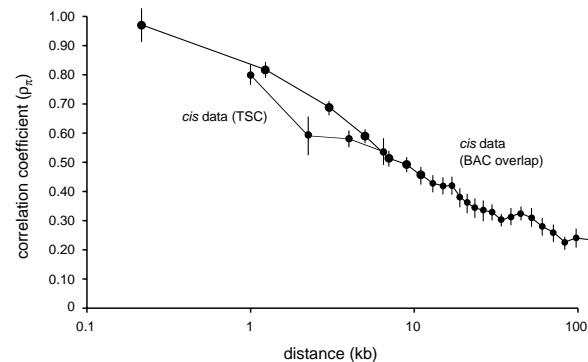
A previous analysis of the TSC data examined polymorphism rate averaged over windows of 200 kb. The polymorphism rate was found to vary up to tenfold across regions<sup>1</sup>. This observed variation could not be explained simply by random sampling; that is, the magnitude of fluctuation in the polymorphism rate across loci was much greater than would be expected if the underlying rate of sequence diversity were constant across all sites<sup>1</sup>.

To better understand this variation, we examined the fluctuation in polymorphism rates at a finer scale (Methods). Despite the significant variation in heterozygosity observed at a coarse scale, we found that nearby regions show very similar rates of polymorphism (Fig. 1a). For example, segments separated by a physical distance ( $d$ ) of only a few hundred base pairs show substantial correlation in the observed polymorphism rate,  $p([\hat{\pi}]_x, [\hat{\pi}]_{x+d}) = 0.28$ . (The quantity  $p(\pi_x, \pi_{x+d})$  is formally the ‘autocorrelation’ in polymorphism rate over distance  $d$ ; we refer to it here as the ‘correlation’ for simplicity.) When corrected for the imprecision in our estimate of  $[\hat{\pi}]$  at each location (resulting from stochastic variation

in the mutational process and the short length of sequence examined in a single sequencing read), the correlation in heterozygosity is nearly complete:  $p(\pi_x, \pi_{x+d}) = 0.97$  for  $d = 100$  bp. That is, the underlying rate of sequence variation is very similar for closely linked sites. The correlation declines with distance, however, falling to half its maximum over a distance of about 8 kb (Fig. 1b). The correlation remains significant even at distances of 100 kb ( $p(\pi_x, \pi_{x+d}) = 0.20 \pm 0.03$ ), although no correlation is observed for unlinked sites ( $p(\pi_x, \pi_{x+d}) < 0.001$ ). The size and length of the correlation are probably general, given the large amount of data examined and the agreement between the results from the two polymorphism discovery projects (Fig. 1b).

### Contribution of gene history and mutation rate

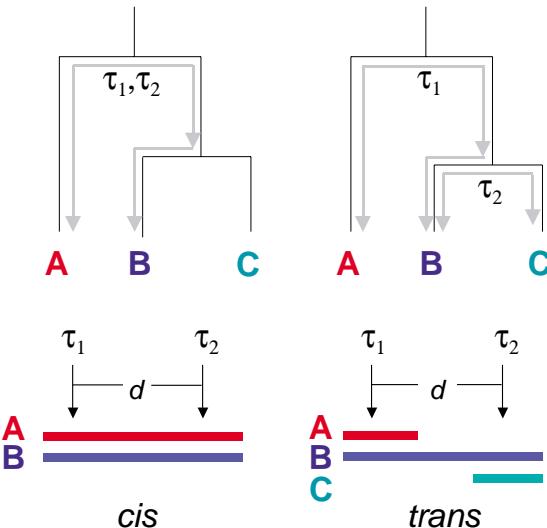
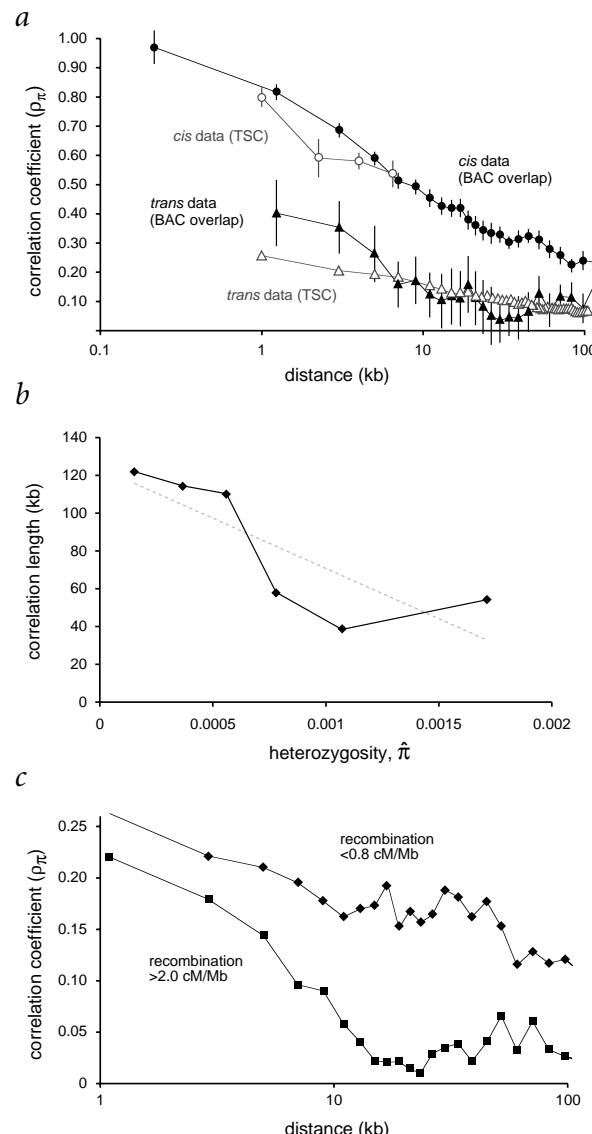
To explain why the tendency to be rich or poor in sequence variation persists for distances of 10–100 kb across the human genome, we considered the mechanisms that are responsible for sequence variation. At any locus in the genome, the rate of polymorphism is shaped by only two forces: the local gene history and the local mutation rate. (Natural selection, for example, acts by altering the history of a locus.) ‘Gene history’ refers to the genealogical relationships among copies of a locus in the current population, which can vary markedly across the genome. For any two copies of a given locus, the number of generations since their shared ancestor, and thus the historical opportunity for mutation, is  $\tau$ . The expected neutral rate of sequence differences between two copies of a locus<sup>11</sup> is simply  $2\tau$  (the number of generations in which mutations could have occurred) multiplied by  $\mu$ , the mutation rate per generation at that locus. Thus, the correlation in polymorphism rates across neighboring sites must be explained by persistence over each region of a similar gene history ( $\tau$ ), a similar mutation rate ( $\mu$ ), or both.

*a**b*

**Fig. 1** Correlation in heterozygosity. **a**, Heterozygosity ( $\hat{\pi}$ ) of individual sequence reads for TSC data compared with heterozygosity of the flanking sequence, which is defined as a region of 2.5 kb on each side of a read. **b**, Correlation in heterozygosity (corrected for stochastic variance; Methods) as a function of distance for the TSC and BAC overlap data.

**Fig. 2** *Cis* versus *trans* comparisons. Shown is a simple genealogical history with three samples chromosomes, A, B and C. In a *cis* comparison, two chromosomes are aligned (A and B) and heterozygosity is compared at two segments on each chromosome separated by a distance  $d$ . In a *trans* comparison, a single chromosome (in this case, B) is compared with two independently sampled chromosomes (A and C) at each of the two segments. In the absence of recombination in the history of either chromosome, the time to the common ancestor at each of the two segments are identical for *cis* comparisons ( $\tau_1 = \tau_2$ ) but not necessarily for *trans* comparisons.

The extent to which particular values of  $\tau$  persist is determined by the history of recombination (in the ancestors of the current sample) across each region. In the absence of meiotic recombination along a chromosome, genomic segments are inherited *en bloc* from generation to generation and thus share a single genealogical history ( $\tau$ ) across their length. By contrast, recombination events juxtapose neighboring chromosomal segments that have different histories, which disrupts the correlation of  $\tau$  with distance. That is, over short distances (such that historical recombination would be unusual), local values of  $\tau$  are expected to be nearly identical. With increasing physical distance, it becomes likely that one or more recombination events will have occurred in the history of the sample, and thus the correlation in  $\tau$  is expected to decline.

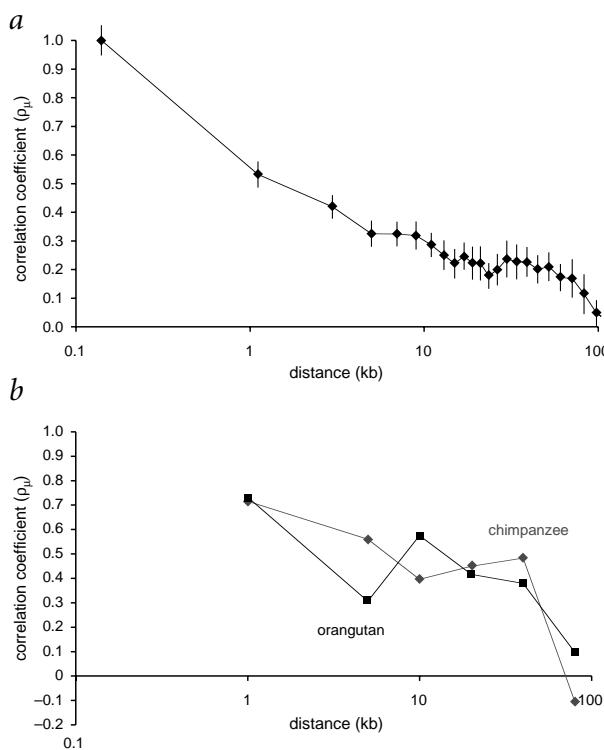


The properties of the correlation in  $\tau$  have been well described in theoretical population genetics under the assumption of a constant-sized, freely mixing population with a uniform recombination rate<sup>12–17</sup>. But previously it has not been possible to measure the empirical distribution of  $\tau$  over distance in any organism. Similarly, although it is known that mutation rates can vary for different classes of sites (such as CpG dinucleotides), the empirical distribution of mutation rate variation over large genomic regions has not been reported. Thus, to understand the observed correlation in polymorphism rates with distance, we needed to measure the underlying variation in both gene histories and mutation rate.

We first wanted to show directly that the long-range correlation in polymorphism rate is, at least in part, attributable to variation in gene history. We examined the correlation in polymorphism rates for nearby loci sampled according to two different protocols (Fig. 2). We reasoned that segments carried on the same chromosome in the current population would share more similar genealogical histories than would segments separated by the same distance but carried on different chromosomes in the population. We call these two types of comparison *cis* (if the two segments are carried on the same physical copy of the chromosome) and *trans* (if the two segments are carried on different physical copies; Fig. 2). If gene history is a principal factor in local differences in polymorphism rate, then the correlation in *cis* should be greater in magnitude than that in *trans*. By contrast, if variation in mutation rate is primarily responsible for the local persistence in heterozygosity, then *cis* and *trans* measurements should be very similar. The physical distance over which differences in *cis* and *trans* persist should indicate the span over which shared history persists in the human genome.

Empirically, we found that the correlation in *cis* was much stronger than that in *trans* for both data sets (Fig. 3a). This shows directly that gene history has a great effect on the observed correlation. In addition, the *cis* correlation was greater

**Fig. 3** Impact of gene history on the correlation in heterozygosity. **a**, *Cis* correlations are much greater in magnitude than are *trans*, as is expected if gene history has a principal role in heterozygosity. **b**, Local polymorphism rate versus extent of correlation. The rate of polymorphisms is measured using the TSC data, considering a 14-kb window of sequence around each read. The extent of correlation is defined as the distance over which the *trans* correlation (shown in a) drops to a quarter of its maximal value. **c**, Local recombination rate versus extent of correlation. The recombination rate in each region is estimated by comparing the physical<sup>5</sup> with the genetic<sup>45</sup> map (Methods).



**Fig. 4** Correlation in mutation rate (inferred from sequence divergence). Great ape sequences are compared with the corresponding sequence from the human genome for 9 large-insert chimpanzee clones from GenBank (**a**) and 19 previously studied regions<sup>28</sup> for both chimpanzee and orangutan sequence (**b**).

We found that there is substantial variation in the amount of sequence divergence across different loci (Table 1). Notably, the amount of sequence divergence at a given locus typically also persists over significant distances (tens of thousands of base pairs) and drops to zero only over distances of 70–100 kb (Fig. 4a,b). Known determinants of local mutation rate (such as GC content) explain less than 10% of the measured variation in  $\mu$  (data not shown), which shows that there must be major, as yet uncharacterized, determinants of mutation rate in the human genome. We note that there may be some variation in the coalescent age of different loci, even in a comparison across species<sup>19,20</sup>. Thus, the estimate from interspecies sequence divergence represents an upper limit rather than an exact measure of the variation in mutation rate (Methods). Below, we present estimates based on this upper limit and a lower limit of no variation in mutation rate across loci.

#### Measuring the correlation of gene history over distance

Using these measurements we assessed the quantitative contributions of variation in gene history and mutation rate to the local patterns of sequence variation in the human genome. For this calculation, we assumed that  $\tau$  and  $\mu$  are independent, which is equivalent to the common assumption of ‘neutrality’<sup>21</sup>—that is, that the vast majority of mutations are not subject to natural selection. Examined using this framework, the data showed that variation in local gene history is the main determinant of local rates of human polymorphism and accounts for at least 57% of the variability in the amount of sequence diversity ( $\pi$ ) across loci (Methods and Table 1). Variation in mutation rate ( $\mu$ ) across loci has a lesser role and explains at most 25% of the variation in sequence divergence at different loci (Table 1). These data represent the first empirical and quantitative evaluation of how human genome sequence variation is shaped by variation in the genealogy of genes versus the rate of mutation across the genome.

The data also show that each given genealogical history typically persists across a region of considerable size in the human genome. We compared our measurement of the correlation in gene histories with the theoretical prediction under the standard population genetic assumptions (Wright–Fisher) of a constant-sized, freely mixing population and a uniform rate of recombination (Fig. 5a)<sup>12–15</sup>. Theory predicts a much shorter persistence of genealogical age<sup>12–15</sup>, with the correlation decreasing to less than 0.5 over only 3 kb, and to less than 0.2 by 11 kb. By contrast, the empirically measured correlation in genealogy extends much further: similarity in gene history remains greater than 0.5 until 8–19 kb, and greater than 0.2 at 100 kb. These results translate into a high probability that any two modern copies of a locus have been inherited without recombination since their most recent common ancestor (Fig. 5b; see Methods for equation and assumptions). Specifically, we have estimated that there is a 38–50% chance that any two copies of a segment 10 kb in length have been inherited without historical recombination since their shared ancestor, as compared with an expectation of only 17% under the standard population genetic assumptions above. These results have significant implications for patterns of linkage disequilibrium in the human genome.

#### A genome-wide measurement of linkage disequilibrium

Linkage disequilibrium refers to nonrandom statistical associations between alleles at nearby sites and is a crucial tool for

than the *trans* correlation for all distances examined (Fig. 3a), which indicates that the same genealogical history is often preserved across substantial distances in the human genome.

The role of gene history in shaping the correlation in heterozygosity was also evident in two other analyses. First, we considered the relationship between the amount of polymorphism (magnitude of  $[\pi]$ ) and the distance over which a given value of heterozygosity persists. Theory predicts that ‘old’ loci—those with a large average value of  $\tau$ —should show both a high rate of polymorphism (because there have been more generations in which mutation could occur) and a shorter correlation over distance (because there has been more opportunity for recombination to scramble gene histories), whereas ‘young’ loci should show the opposite pattern. This prediction was confirmed by the data (Fig. 3b). Similarly, theory predicts that the persistence of polymorphism rate should be shorter in genomic regions where, on average, meiotic recombination (as measured at a scale of megabases on the genome-wide linkage map)<sup>18</sup> is more active. Comparison of the persistence of polymorphism rate for regions with high and low average rates of recombination bore out this prediction (Fig. 3c).

We characterized the distribution of variation in mutation rate ( $\mu$ ) across genomic segments and assessed its contribution to variation in heterozygosity ( $\pi$ ). The classical method of estimating the human mutation rate is to compare human DNA sequence with that of a great ape, assuming that in any genomic region the amount of sequence divergence that has occurred is proportional only to the local mutation rate<sup>11</sup>. (This should be true if most mutations in the genome are not subject to natural selection and the time since the common ancestor is about the same across all loci.) Using data from GenBank and our own laboratory, we obtained 1.3 Mb of human and great ape (chimpanzee and orangutan) sequence alignments drawn from a total of 28 loci (Methods). We then measured the correlation over physical distance in interspecies sequence divergence (formally, the autocorrelation  $\rho(\mu_x, \mu_{x+d})$  with distance).

**Fig. 5 Correlation in gene history.** *a*, The simple correlation in gene history  $\rho_{cis}(\tau_w \tau_{x,y})$  (Methods). Open squares represent an upper limit and filled squares a lower limit. The upper limit is obtained by assuming that all variation in interspecies sequence divergence is attributable to variation in mutation rate; the lower is based on the assumption of a constant mutation rate across the genome. For comparison, the expectation<sup>15</sup> for the *cis* correlation for a constant-sized population of  $N = 10,000$  and a uniform recombination rate<sup>5</sup> of 1.3 cM/Mb is shown. *b*, Probability of no historical recombination (since the most recent common ancestral chromosome) between two sites separated by a distance  $d$  (Methods). *c*, Comparison of the value of linkage disequilibrium extrapolated from the correlation in gene history to a direct measurement in Americans of European ancestry based on genotyping of 2,745 polymorphic SNPs discovered by TSC and distributed over 51 genomic regions<sup>32</sup>. The two assessments are in qualitative agreement, with both exceeding the predictions of population genetic theory.

mapping genes that contribute to disease<sup>22–24</sup>. To assess linkage disequilibrium in this genome-wide data set, we combined our empirical measurements with a recently described statistical framework (G.M., unpublished data; see below for URL). This approach relates the correlation in genealogy (measured above) to the average extent of linkage disequilibrium in the genome, without making any assumptions about population history or selection.

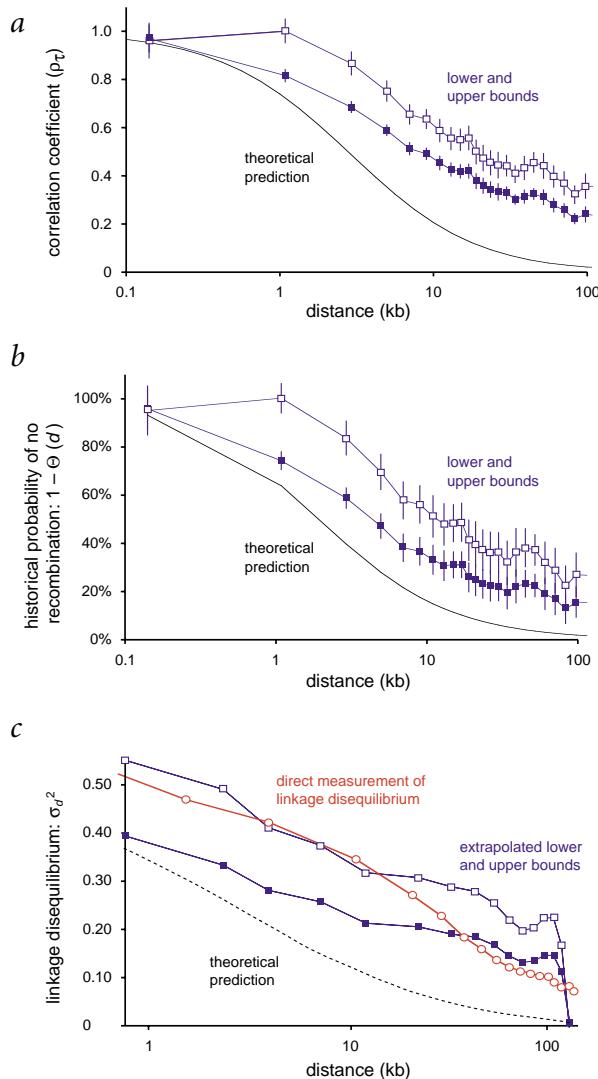
A common measure of linkage disequilibrium is the  $r^2$  statistic, which is particularly relevant for gene mapping because its magnitude can be translated directly to the sample size that is required for an association study<sup>25</sup>. Our statistical framework provides an estimator of  $r^2$ , which is accurate at predicting  $r^2$  for common (>10% frequency) alleles (G.M., unpublished data; see also refs 26, 27). Applying this approach to the empirical data, we obtain the first genome-wide estimate of the average extent of linkage disequilibrium in the human genome (for the samples used in polymorphism discovery). Linkage disequilibrium extends for significant distances (Fig. 5c), which are much longer than predicted under the standard assumptions described above, as has been shown in empirical studies based on direct genotyping across a much smaller fraction of the human genome sequence<sup>28–32</sup>. To our knowledge, these results represent the first truly genome-wide estimate of this crucial quantity for disease gene mapping.

This approach to estimating linkage disequilibrium has a significant advantage: a genome-wide estimate of linkage disequilibrium can be derived in any organism from only a measurement of heterozygosity. By sequencing DNA from several chimpanzees, for example, it should be possible to obtain simultaneously a genome sequence, a polymorphism map and a genome-wide assessment of linkage disequilibrium.

### Inhomogeneous recombination in the human genome

We wanted to understand the mechanisms responsible for the unexpectedly long persistence of gene histories in the human genome. Previous work has suggested the involvement of population bottlenecks<sup>28,33–36</sup>, mixing of populations<sup>37</sup> and hot spots of recombination<sup>38–41</sup>. We used coalescent computer simulations to understand which of these might be responsible for the observed long-range correlation in  $\tau$ .

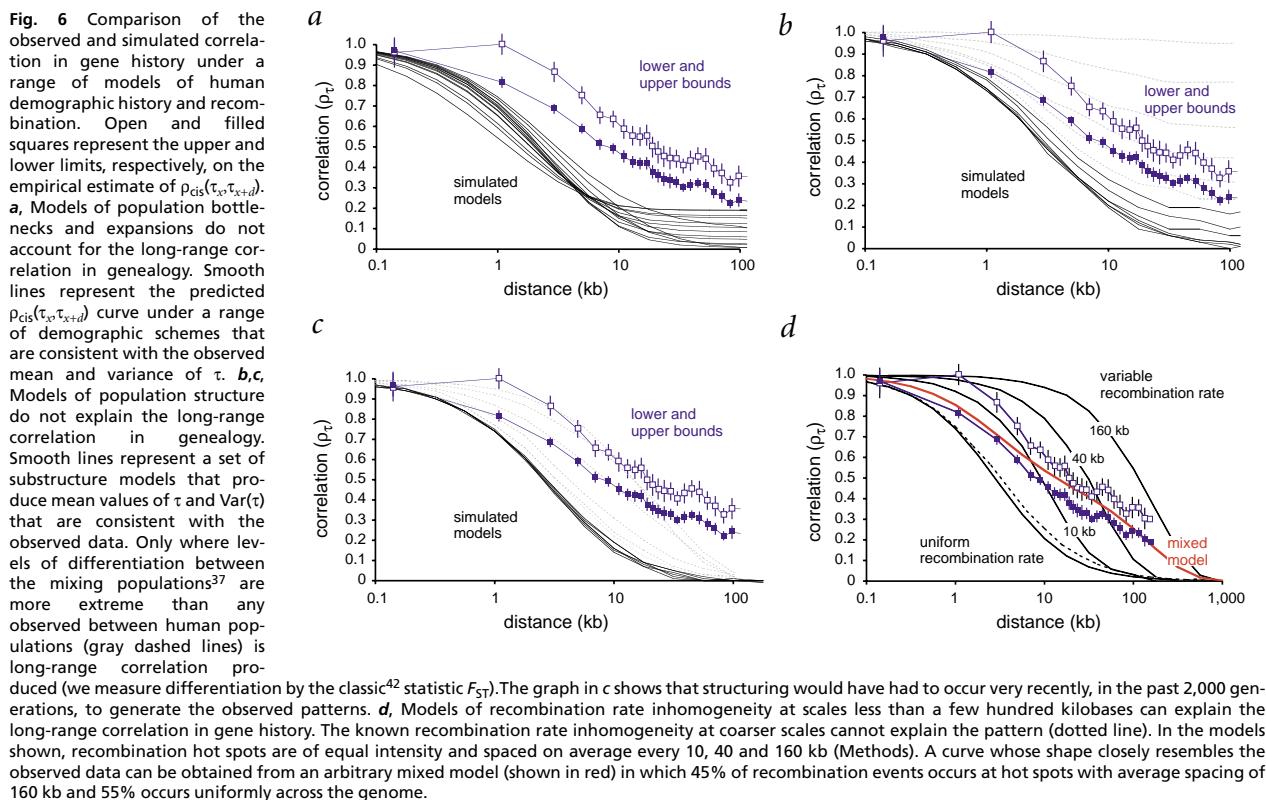
First, we compared our empirical results with those obtained by exploring a wide range of population demographic models (Methods). These allow for expansions and contractions in population size of various magnitudes at different times in the past (Fig. 6a). We found that changes in population size have only a modest effect on the correlation in gene history. In fact, we could not identify a model of population expansion or contraction that was consistent with the measured mean and variance in the age of alleles ( $\tau$ ), and that could generate the observed persistence in  $\tau$  (Fig. 6a). This



does not mean that expansions and contractions have not occurred in human history (they surely have<sup>28,33–35</sup>), but rather that population bottlenecks cannot be the sole explanation for the long regions of shared gene history that we observe.

Second, we investigated the effects of population substructure (deviations from a freely mixing population) and found that this is also unlikely to explain the long stretches of shared gene history (Fig. 6b,c). Extreme models of structuring in the current population could produce long-range correlation according to our simulations (Fig. 6b and Methods), but they required a degree of substructure much greater than any observed in modern-day human populations<sup>42</sup>. It is possible that severe structuring might have occurred in the past; however, our simulations (Fig. 6c) indicated that it would have had to persist until relatively recently (within the past 2,000 generations or ~50,000 years) to produce the observed long-range correlation. As there is no genetic or archaeological evidence for such extreme structure having existed so recently (and then disappeared through mixing) in the period since the appearance of modern humans outside Africa<sup>42</sup>, we consider this possibility unlikely although we cannot rule it out.

Last, we explored whether non-uniform recombination could explain the long-range similarity in polymorphism rates. The rate of recombination in the human genome varies at large scales



(>1 Mb)<sup>18</sup>, and in one part of the genome it has been observed to vary markedly at a fine scale<sup>41</sup>. (Specifically, across a 200-kb region of the major histocompatibility complex locus, ~95% of all recombination events are restricted to six hot spots of <2 kb that together cover <5% of the whole region<sup>41</sup>.) The general pattern of human recombination, however, has not been characterized at a fine scale. To understand the impact of recombination patterns on the correlation in gene history, we first examined the well-known pattern of variation at multi-megabase scales<sup>18</sup>. We found that variation at such a coarse scale has only a slight effect on the correlation curve (Fig. 6d) and cannot explain the long persistence that we observe. When we modeled fine-scale variation in recombination rate, however, we found a marked effect on the correlation (Fig. 6d). Fine-scale variation in recombination rate essentially corresponds to the presence of hot spots of recombination, which can greatly increase the average extent of correlation in gene history because ‘colder’ regions between hot spots are subject to relatively little historical recombination. Thus, of the schemes examined in this analysis, only an extremely inhomogeneous recombination rate at short distance scales is compatible with the long persistence of shared gene history in the human genome.

## Discussion

In this genome-wide, empirical exploration of the fine-scale pattern and underlying causes of sequence variation in the human genome, we have shown that the human genome is composed of large regions, as long as 100 kb, that have intrinsically different rates of sequence polymorphism. The characteristics of these regions are determined largely by differences in gene history and less by differences in local mutation rate.

Our analysis also shows that shared gene history and linkage disequilibrium typically extend over much longer distances than

would be expected under standard population genetic models and assumptions. Although this result has been suggested in previous studies of a few regions<sup>28–32</sup>, to our knowledge this study is the first truly genome-wide assessment. Of a range of schemes examined, inhomogeneous recombination offers the best explanation for the long correlation in gene history. Variation in recombination rate at multi-megabases scales has been described<sup>18,43</sup> and hot spots of recombination have been reported anecdotally<sup>38,40,41</sup>, but we have now shown that recombination rate inhomogeneity at a fine scale is a general feature of the genome and has a major impact on human variation. Although our results cannot specify the exact architecture of this fine-scale variation, we propose that previously described hot spots<sup>41</sup> offer the most likely explanation.

Our results have implications for genetic association studies of human disease. The inheritance of chromosomal regions without recombination from shared ancestors (Fig. 3c), which are also known as haplotypes, freezes particular combinations of alleles in the population. Haplotypes are valuable for medical genetic studies because they allow mapping of disease-susceptibility alleles without the need to discover and test every SNP across each chromosomal region<sup>22–24</sup>. If recombination rate inhomogeneity is indeed a defining feature of linkage disequilibrium in the human genome, then SNPs interspersed in any region of low recombination will track together in the population<sup>32,39,41</sup>. Our study provides fundamental evidence that gene mapping techniques that take advantage of this phenomenon will be broadly applicable across the human genome.

## Methods

**TSC and BAC overlap data sets.** The data sets were subsets of those used to obtain a map of 1.42 million SNPs<sup>1</sup>. The TSC data were obtained by comparing random sequencing reads, averaging 514 bp (range 400–700 bp) of high-quality sequence, with single BAC sequences from

version OO18 of the public human genome assembly<sup>5</sup>. For the subset of TSC data that included ‘paired-end’ reads with sequence available from two ends of cloned human inserts, it was possible to make *cis* comparisons; the remainder of comparisons were *trans*. The BAC overlap data were obtained by comparing 500-bp segments of finished sequence from BAC clones in the RPCI-11 library (which was used for most of the public sequence<sup>5</sup>) with a finished BAC from another library. Most of the BAC overlap results were therefore *cis* comparisons, but *trans* comparisons were obtained by examining triplets of overlapping BACs: one from RPCI-11, one from another RPCI or Caltech clone library, and one from any third library.

To minimize contamination by low-copy paralogous repeats (which can generate spurious stretches of high measured heterozygosity), we eliminated reads in which five or more SNPs were observed<sup>3,4</sup>. To eliminate regions near identifiable repeats, we removed from our genome-wide analysis any 200-kb window in which more than 2% of reads aligned to different genomic locations (this criterion was based on a detailed comparison of heterozygosity and repeat content). The more stringent filtering explains why the TSC heterozygosity that we report differs by 4% from a previous estimate<sup>1</sup>. (The two analyses also differ in that reads aligning to more than one BAC were omitted from the current analysis and that the contribution of each read to heterozygosity was weighted by the number of times that it was used in the calculation of the correlation.)

**Heterozygosity and correlation statistics.** We calculated heterozygosity by dividing the number of SNPs observed by the number of bases for which high-quality sequence was available (defined by the neighborhood quality standard<sup>3</sup>). Correlation in polymorphism rate  $p(\hat{\pi}_x \hat{\pi}_{x+d})$  was calculated as described below. *Cis* correlation at very short distances (for the BAC overlap data) was calculated by splitting 500-bp reads into two perfectly overlapping 250-bp reads of alternating base pairs. The TSC data did not provide a pure *trans* comparison: 1 of 48 comparisons in the TSC data set involved (by chance) a *cis* comparison of two reads from the same chromosome in the pool. The *trans* curves shown (Fig. 3a) are therefore corrected for roughly a 2% admixture of *cis* data.

Error bars in all figures correspond to 1 s.d. and were calculated by bootstrapping: that is, new correlation curves were generated by re-sampling a random subset of the original data set<sup>44</sup>. The TSC data were re-sampled 50 times, the BAC overlap data 100 times, and the human–chimpanzee data 400 times. Because the heterozygosity of nearby reads was correlated, we carried out re-sampling by partitioning the genome into contiguous sectors. Each sector spanned 5% of the total data set for BAC overlap and GenBank human–chimpanzee comparisons, and 1% of the total data set for the TSC comparisons.

**Mathematical formulae.** The expected polymorphism rate between two samples, with an observed rate  $\hat{\pi}$  and true underlying polymorphism rate  $\pi$ , is:

$$E[\hat{\pi}] = E[\pi] = 2\tau\mu.$$

The variation in observed polymorphism rate,  $\text{Var}(\hat{\pi})$ , can be parsed into its determinants  $\tau$ ,  $\mu$  and  $L$  (the length of a sequencing read), according to the following equation derived in Web Note A online:

$$\text{Var}(\hat{\pi}) \approx E[\pi/L] + 4\text{Var}(\tau)E[\mu]^2 + 4\text{Var}(\mu)E[\tau]^2 + 4\text{Var}(\mu)\text{Var}(\tau)$$

The ‘stochastic variance’ term  $E[\pi/L]$  arises because the observed value,  $\hat{\pi}$ , is not a perfect estimate of the underlying value  $\pi$  because of the limited length of sequence examined. We therefore estimated  $\text{Var}(\pi)$  as  $\text{Var}(\hat{\pi}) - E[\pi/L]$  (see Web Note A for details).

The similarity (covariance) in observed polymorphism rate  $\hat{\pi}$  between two sites separated by a distance  $d$  is defined by an equation of a form similar to the variance equation (see Web Note A for derivation):

$$\begin{aligned} \text{Cov}(\hat{\pi}_x, \hat{\pi}_{x+d}) &= \text{Cov}(\pi_x, \pi_{x+d}) \\ &= 4\text{Cov}(\tau_x, \tau_{x+d})E[\mu]^2 + 4\text{Cov}(\mu_x, \mu_{x+d})E[\tau]^2 + 4\text{Cov}(\tau_x, \tau_{x+d})\text{Cov}(\mu_x, \mu_{x+d}) \end{aligned}$$

To obtain  $\text{Cov}(\mu_x, \mu_{x+d})$ , we used the great ape–human comparisons, setting  $\text{Cov}(\tau_x, \tau_{x+d}) = 0$ .

Correlations are equal to covariances divided by variances:

$$\rho(\alpha_x, \alpha_{x+d}) = \text{Cov}(\alpha_x, \alpha_{x+d})/\text{Var}(\alpha)$$

where  $\alpha \hat{\pi} = \pi, \mu$  or  $\tau$ .

One of us has derived a relationship between correlation in genealogical history ( $\tau$ ) and linkage disequilibrium (as measured by an estimator of  $r^2$  called  $\sigma_d^2$ ; G.M., unpublished data):

$$\sigma_d^2 = \frac{\rho_{\text{cis}}(\tau_x, \tau_{x+d}) - 2\rho_{\text{trans}}(\tau_x, \tau_{x+d}) + \rho_{\text{dis}}(\tau_x, \tau_{x+d})}{E[\tau]^2/\text{Var}(\tau) + \rho_{\text{dis}}(\tau_x, \tau_{x+d})}.$$

The term ‘dis’, or disjoint correlation, refers to cases where all four segments that are used in an assessment of similarity in gene history are from different individuals (two at position  $x$  in the genome and two others at position  $x+d$ ). The disjoint curve we used in our analysis was obtained by comparing two different TSC reads that mapped to two different BAC clones.

For a constant-sized population, we derive a relationship (see Web Note A) between  $\rho_{\text{cis}}(\tau_x, \tau_{x+d})$ ,  $\rho_{\text{trans}}(\tau_x, \tau_{x+d})$ , and the probability that the current haplotype between two sites a distance  $d$  apart has been inherited without recombination from the most recent common ancestor ( $1 - \Theta(d)$ ).

$$1 - \Theta(d) = \frac{\rho_{\text{cis}}(\tau_x, \tau_{x+d}) - \rho_{\text{trans}}(\tau_x, \tau_{x+d})}{1 - \rho_{\text{trans}}(\tau_x, \tau_{x+d})}.$$

**Stochastic variance correction.** In our analysis, we eliminated reads with five or more SNPs (see above), which affects the stochastic variance estimate,  $E[\pi/L]$ . Therefore, we slightly modified the variance estimate on the basis of computer simulations of the effect of truncating reads. We made an additional small adjustment to address the 5% false-positive rate in SNP identification<sup>1</sup> (details are available on request).

**Comparison of genetic to physical map.** For assessment of the effect of recombination rate in reducing the extent of correlation of heterozygosity (Fig. 3c), we estimated recombination rates by comparing genetic distances between markers obtained from the Marshfield map<sup>45</sup> with their physical distance separations. (We compared markers approximately 5 Mb apart.) The first and last markers on each chromosome, and regions for which there was disagreement between physical and genetic map marker order, were discarded.

To determine how the known inhomogeneity in recombination rate at scales of more than 1 Mb would affect the theoretically predicted  $\rho_{\text{cis}}(\tau_x, \tau_{x+d})$  curve (Fig. 6d), we compared the physical map of the human genome<sup>5</sup> with the Marshfield genetic map<sup>45</sup> and built up a histogram of recombination rates by examining all marker spacings greater than 1 Mb. We then selected values randomly from this distribution and combined them with the theoretically predicted  $\rho_{\text{cis}}(\tau_x, \tau_{x+d})$  curve to obtain the dashed line in Fig. 6d.

**Great ape–human comparisons and autocorrelation in  $\mu$ .** From GenBank we obtained nine sequences from chimpanzees (accession numbers below). We aligned these to the human sequence and identified divergences through the same computational and filtering steps as above (allowing up to 9% sequence divergence across any 200-bp stretch). To test the generality of the observed patterns, we examined a second data set consisting of 180 kb of chimpanzee sequence and 110 kb of orangutan sequence that we compared with human sequence at 19 independent loci<sup>28</sup>. Altogether, our assessment was based on 18,350 sequence divergences between great apes and humans.

These three estimates (chimpanzee–human comparisons from both GenBank and the 19 loci, and orangutan–human comparisons from only the 19 loci) agreed qualitatively, with autocorrelation persisting over tens of kilobases and dropping to zero only after 70–100 kb (Fig. 4a). We used the largest data set (from GenBank) to obtain  $\rho_{\text{cis}}(\tau_x, \tau_{x+d})$  for use in subsequent calculations, fitting a straight line to the log plot in Fig. 4a (omitting the data point at the shortest distance to minimize possible correlations owing to shared history at neighboring loci).

# article

To translate our results on primate divergence into  $\mu$  and  $\tau$ , we required an estimate of the mean mutation rate per nucleotide per generation (we used a value of  $2.5 \times 10^{-8}$ , obtained by calibration from the fossil record<sup>46</sup>). Errors in this estimate would not affect the conclusions of this paper and would change only four lines of Table 1 (estimates of the mean and standard deviation of  $\mu$ , and of the mean and standard deviation of  $\tau$ ).

**Computer simulations.** We first considered a freely mixing population with a constant effective size of  $N$  individuals until  $G$  generations ago, when there was a bottleneck that produced an inbreeding coefficient of  $F$  followed by a rapid expansion to very large size (Fig. 6a). We used coalescent computer simulations<sup>16</sup> to generate  $p_{\text{cis}}(\tau_x, \tau_{x+d})$  curves for a full range of models ( $N$ ,  $G$  and  $F$ ) that were consistent with the observed mean  $\tau$  and range of the coefficient of variation (0.76–1.11; Table 1). Specifically, to identify values of  $N$ ,  $G$  and  $F$  that were consistent with our data, we wrote equations for the mean  $\tau$  and the coefficient of variation of  $\tau$  in terms of the parameters  $N$ ,  $G$  and  $F$  in our model (available on request). Solving for the equations left only one free parameter, which we varied over its full range for coefficients of variation of  $\tau$  of 0.76, 1 and 1.11, to explore the full range of demographic schemes that were consistent with our model and data.

We next considered a population that was freely mixing and of constant size  $N$  until  $G$  generations ago, when it split into two populations of equal size  $N$  that remained separate until a recent mixing event in which a proportion  $p$  of samples came from the first population (Fig. 6b). We generated curves for a full range of models ( $N$ ,  $G$  and  $p$ ) that were consistent with the observed mean  $\tau$  and range of the coefficient of variation of  $\tau$  and constrained the parameters as for the simulations of bottlenecks and expansions (details of model available on request). We also modified the simulations to explore schemes in which the population substructure ended in the more distant past, that is, the populations mixed together again  $H = 2,000$  generations ago (Fig. 6c).

Last, we considered models in which recombination occurs only at hot spots, which were assumed to be of equal intensity and distributed randomly (according to a Poisson process) with mean spacing of  $k$  kilobases, for  $k = 10, 40$  and  $160$  kb (Fig. 6d). Computer simulations<sup>16</sup> were used to calculate the expected curve for different hot spots spacings and a constant-sized, freely mixing population.

**URL:** Further information on the statistical method for obtaining linkage disequilibrium from correlation data can be obtained from <http://www.stats.ox.ac.uk/~mcvean/lbgene.pdf>.

**GenBank accession numbers.** Chimpanzee sequences, AC087834.1, AC087835.1, AC087778.1, AC087736.1, AC087777.1, AC087568.1, AC087513.1, AC087264.2, AC087602.2.

**Note:** Supplementary information is available on the *Nature Genetics* website.

## Acknowledgments

We thank T. Lavery, A. Rachupka and J. Platko for assistance with great ape sequencing; B. Gilman for computer support; D. Cutler, P. Donnelly, J. Hirschhorn, L. Kruglyak, S. Myers, J. Pritchard and J. Wakeley for discussions and advice; and the laboratory of E. Green and the Baylor Sequencing Center for depositing large-insert chimpanzee sequences into GenBank. D.E.R. was supported in part by a National Defense Science and Engineering fellowship. D.A. is a Charles E. Culpeper Scholar of the Rockefeller Brothers Fund and a Burroughs Wellcome Fund Clinical Scholar in Translational Research. This work was supported by grants from The SNP Consortium to E.S.L. and D.A., the Massachusetts General Hospital to D.A. and the National Institutes of Health to E.S.L..

## Competing interests statement

The authors declare that they have no competing financial interests.

Received 8 April; accepted 26 June 2002.

1. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
2. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).

3. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
4. Mullikin, J.C. et al. An SNP map of human chromosome 22. *Nature* **407**, 516–520 (2000).
5. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
7. Cambien, F. et al. Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65**, 183–191 (1999).
8. Halushka, M.K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
9. Wang, D.G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
10. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
11. Li, W.-H. *Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1997).
12. Griffiths, R.C. in *Selected Proceedings of the Sheffield Symposium on Applied Probability*, IMS Lecture Notes Vol. 18 (eds I.V. Basawa & R.L. Taylor) 100–117 (Institute of Mathematical Statistics, 1991).
13. Griffiths, R.C. Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.* **19**, 169–186 (1981).
14. Kaplan, N. & Hudson, R.R. The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theor. Popul. Biol.* **28**, 382–396 (1985).
15. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
16. Hudson, R.R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D.J. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
17. Sved, J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2**, 125–141 (1971).
18. Yu, A. et al. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
19. Hudson, R.R. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217 (1983).
20. Takahata, N. & Satta, Y. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc. Natl. Acad. Sci. USA* **94**, 4811–4815 (1997).
21. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
22. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
23. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
24. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
25. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
26. Strobeck, C. & Morgan, K. The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88**, 829–844 (1978).
27. Ohta, T. & Kimura, M. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite populations. *Genetics* **68**, 571–580 (1971).
28. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
29. Abecasis, G.R. et al. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
30. Dunning, A.M. et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1550 (2000).
31. Taillon-Miller, P. et al. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
32. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **297**, 2225–2229 (2002); published online 23 May 2002 (10.1126/science.1069424).
33. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends. Genet.* **16**, 296–302 (2000).
34. Reich, D.E. & Goldstein, D.B. Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. USA* **95**, 8119–8123 (1998).
35. Kimmel, M. et al. Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930 (1998).
36. Tishkoff, S.A. et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
37. Wakeley, J. Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871 (1999).
38. Chakravarti, A. et al. Nonuniform recombination within the human  $\beta$ -globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).
39. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
40. Jeffreys, A.J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
41. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
42. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, New Jersey, 1994).
43. Kong, A. A high resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002); advance online publication, 10 June 2002 (doi:10.1038/ng917).
44. Lui, B.H. in *Statistical Genomics: Linkage, Mapping, and QTL Analysis* (CRC Press, Boca Raton, Florida, 1998).
45. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
46. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).