stimulation for each eye. We identified the V1 blind-spot representation in individual subjects as those voxels in the left calcarine sulcus that showed both a significant response to ipsilateral stimulation and a significantly greater response to ipsilateral than blind-spot stimulation using a minimum statistical threshold of $t = 2.0$, $P < 0.05$. The blind-spot representation ranged in size from 3–5 voxels (voxel size 3.125×3.125×4 mm) across subjects, consistent with size estimates based on post-mortem neuroanatomical studies (J. C. Horton, personal communication).

Binocular rivalry and stimulus alternation scans

During these scans, two subjects viewed the red vertical grating and green horizontal grating with their left eye and right eye, respectively, whereas two subjects received the reverse eye assignment. Subjects performed 7–10 scans of rivalry and an equal number for stimulus alternation. Each scan lasted for 90 s. We discarded the first 10 s of fMRI activity to remove transient responses to the onset of the stimulus. We converted fMRI activity from the V1 blind-spot representation to per cent signal change from the mean level during the scan, and potential MR spikes and artefacts were minimized by reducing any outliers to lie within 3 s.d. of the mean.

We conducted an event-related fMRI analysis for reported switches between the blind-spot and ipsilateral grating. Previously, we found that rivalry responses increase as a function of percept duration and that very brief percepts led to unreliable fMRI responses[8]. Here, a switch was considered valid only if the percept immediately preceding and following the reported switch lasted longer than 2 s. An intervening blend response was allowed if it occurred within 1 s before the reported switch, in which case the blend duration was incorporated into the pre-switch period. fMRI responses of each subject were calculated by separately averaging the fMRI time course surrounding all occurrences of a reported switch to the ipsilateral grating or blind-spot grating for rivalry versus stimulus alternation, time-locked to each reported switch (rounded to the nearest second). Each average fMRI response function consisted of 39–55 observations.

Received 14 December 2000; accepted 12 February 2001.

1. Helmholtz, H. v. *Helmholtz's Treatise on Physiological Optics* (The Optical Society of America, Rochester, New York, 1924).
2. Levelt, W. J. M. *On Binocular Rivalry* (Royal VanGorcum, Assen, The Netherlands, 1965).
3. Blake, R. A neural theory of binocular rivalry. *Psychol. Rev.* **96,** 145–167 (1989).
4. Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379,** 549–553 (1996).
5. Logothetis, N. K. & Schall, J. D. Neural correlates of subjective visual perception. *Science* **245,** 761–763 (1989).
6. Sheinberg, D. L. & Logothetis, N. K. The role of temporal cortical areas in perceptual organization. *Proc. Natl Acad. Sci. USA* **94,** 3408–3413 (1997).
7. Tootell, R. B. H. *et al.* Functional analysis of primary visual cortex (V1) in humans. *Proc. Natl Acad. Sci. USA* **95,** 811–817 (1998).
8. Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21,** 753–759 (1998).
9. Walls, G. L. The filling in process. *Am. J. Optom. Arch. Am. Acad. Optom.* **31,** 329–341 (1954).
10. Stensaas, S. S., Eddington, D. K. & Dobelle, W. H. The topography and variability of the primary visual cortex in man. *J. Neurosurg.* **40,** 747–755 (1974).
11. Fox, R. & Rasche, F. Binocular rivalry and reciprocal inhibition. *Percept. Psychophys.* **5,** 215–217 (1969).
12. Lehky, S. R. An astable multivibrator model of binocular rivalry. *Perception* **17,** 215–228 (1988).
13. Polonsky, A., Blake, R., Braun, J. & Heeger, D. J. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neurosci.* **3,** 1153–1159 (2000).
14. Lansing, R. W. Electroencephalographic correlates of binocular rivalry in man. *Science* **146,** 1325–1327 (1964).
15. Cobb, W. A., Morton, H. B. & Ettlinger, G. Cerebral potential evoked by pattern reversal and their suppression in visual rivalry. *Nature* **216,** 1123–1125 (1967).
16. Brown, R. J. & Norcia, A. M. A method for investigating binocular rivalry in real-time with the steady-state VEP. *Vision Res.* **37,** 2401–2408 (1997).
17. Tononi, G., Srinivasan, R., Russell, D. P. & Edelman, G. M. Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proc. Natl Acad. Sci. USA* **95,** 3198–3203 (1998).
18. Lumer, E. D., Friston, K. J. & Rees, G. Neural correlates of perceptual rivalry in the human brain. *Science* **280,** 1930–1934 (1998).
19. Wade, N. J. Monocular and binocular rivalry between contours. *Perception* **4,** 85–95 (1975).
20. Logothetis, N. K., Leopold, D. A. & Sheinberg, D. L. What is rivalling during binocular rivalry? *Nature* **380,** 621–624 (1996).
21. Lee, S. H. & Blake, R. Rival ideas about binocular rivalry. *Vision Res.* **39,** 1447–1454 (1999).
22. Andrews, T. J. & Purves, D. Similarities in normal and binocularly rivalrous viewing. *Proc. Natl Acad. Sci. USA* **94,** 9905–9908 (1997).
23. Crick, F. & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375,** 121–123 (1995).
24. Friston, K. J. *et al.* Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2,** 189–210 (1994).

Acknowledgements

We thank A. Seiffert and K. Nakayama for comments on this manuscript; D. Tran for research assistance; and J. Mazziotta, M. Cohen, the UCLA Brain Mapping Medical Organization, the Ahmanson Foundation, the Pierson-Lovelace Foundation, the Tamkin Foundation, and the Jennifer Jones-Simon Foundation for support. This research was funded by a McDonnell-Pew Grant in Cognitive Neuroscience and by the National Institutes of Health.

Correspondence and requests for materials should be addressed to F. T. (e-mail: ftong@princeton.edu).

Linkage disequilibrium in the human genome

David E. Reich*, Michele Cargill*†, Stacey Bolk*, James Ireland*, Pardis C. Sabeti‡, Daniel J. Richter*, Thomas Lavery*, Rose Kouyoumjian*, Shelli F. Farhadian*, Ryk Ward‡ & Eric S. Lander*§

* *Whitehead Institute / MIT Center for Genome Research,*
Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
‡ *Institute of Biological Anthropology, University of Oxford, Oxford OX2 6QS, UK*
§ *Department of Biology, MIT, Cambridge, Massachusetts 02139, USA*

With the availability of a dense genome-wide map of single nucleotide polymorphisms (SNPs)[1], a central issue in human genetics is whether it is now possible to use linkage disequilibrium (LD) to map genes that cause disease. LD refers to correlations among neighbouring alleles, reflecting 'haplotypes' descended from single, ancestral chromosomes. The size of LD blocks has been the subject of considerable debate. Computer simulations[2] and empirical data[3] have suggested that LD extends only a few kilobases (kb) around common SNPs, whereas other data have suggested that it can extend much further, in some cases greater than 100 kb[4–6]. It has been difficult to obtain a systematic picture of LD because past studies have been based on only a few (1–3) loci and different populations. Here, we report a large-scale experiment using a uniform protocol to examine 19 randomly selected genomic regions. LD in a United States population of north-European descent typically extends 60 kb from common alleles, implying that LD mapping is likely to be practical in this population. By contrast, LD in a Nigerian population extends markedly less far. The results illuminate human history, suggesting that LD in northern Europeans is shaped by a marked demographic event about 27,000–53,000 years ago.

To characterize LD systematically around genes, each of the 19 regions that we studied was anchored at a 'core' SNP in the coding region of a gene. The core SNP was chosen from a database of more than 3,000 coding SNPs that had been identified by screening in a multi-ethnic panel (see Methods), subject to two requirements. First, 'finished' genomic sequence was available for 160 kb in at least one direction from the core SNP; second, the frequency of the minor (less common) allele was at least 35% in the multi-ethnic panel.

We focused on high-frequency SNPs for several reasons. First, they tend to be of high frequency in all populations[7], facilitating cross-population comparisons. Second, LD around common alleles represents a 'worst case' scenario: LD around rare alleles is expected to extend further because such alleles are generally young[8] and there has been less historical opportunity for recombination to break down ancestral haplotypes[2]. Third, LD around common alleles can be measured with a modest sample size of 80–100 chromosomes to a precision within 10–20% of the asymptotic limit (see Methods). Last, LD around common alleles will probably be particularly relevant to the search for genes predisposing to common disease[9].

To identify SNPs at various distances from the core SNP, we re-sequenced subregions of around 2 kb centred at distances 0, 5, 10, 20, 40, 80 and 160 kb in one direction from the core SNP using 44 unrelated individuals from Utah. Altogether, we screened 251,310 bp (see Methods) and found an average heterozygosity of $\pi = 0.00070$, consistent with past studies[1]. A total of 272 'high frequency' polymorphisms were identified (Table 1).

We measured LD between two SNPs using the classical statistic D' (see Methods)[10]. D' has the same range of values regardless of the frequencies of the SNPs compared[11]. Its sign (positive or negative)

† Present address: Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA.

depends on the arbitrary choice of the alleles paired at the two loci. We chose the pair of SNPs that caused $D' > 0$ in Utah so that, in comparisons with other populations, the sign of D' indicates whether the same or opposite allelic association is present. In a large sample, $|D'|$ of 1 indicates complete LD; 0 corresponds to no LD. The degree of LD needed for effective mapping depends on the details of a particular study[2]. A useful measure is the 'half-length' of LD (the distance at which the average $|D'|$ drops below 0.5).

Comparing the 19 randomly selected regions, LD has a half-length of about 60 kb (Fig. 1). Significant *P*-values for LD occur in greater than 50% of cases at distances of ≤ 80 kb. LD therefore extends much further than a previous prediction[2], and our data indicate that, in general, blocks of LD are large. Although the average extent is large, there is great variation in LD across the

Table 1 Distribution of regions and SNPs within regions

Gene identification*	Chromosome	Local recombination rate (cM Mb^{-1})†	Span of region in physical map (cR)‡	Number of high-frequency polymorphisms at distances (kb) from core SNPs§						
				0	5	10	20	40	80	160
BMP8‖	1	1.4	60.88–61.04	1	1	1	2	2	1	3
ACVR2B‖	3	1.1	37.27–37.43	1	1	2	3	0	3	1
TGFBI	5	1.4	152.16–152.00	2	3	1	0	1	7	0
DDR1‖	6	–	46.046–45.89	1	4	1	4	2	2	0
GTF2H4	6	–	46.059–46.22	1	3	3	6	2	0	0
COL11A2‖	6	–	48.27–48.43	3	0	2	2	3	1	1
LAMB1‖	7	2.3	106.58–106.42	1	0	5	3	3	2	4
WASL	7	0.5	122.99–122.83	0	1	0	4	2	2	1
SLC6A12	12	3.3	3.62–3.78	2	8	2	1	6	0	0
KCNA1	12	3.3	8.50–8.66	2	1	5	2	1	1	0
SLC2A3	12	2.1	16.33–16.49	1	2	1	1	5	0	2
ARHGDIB‖	12	1.2	21.84–22	1	9	1	2	2	1	2
PCI‖	14	4.3	98.41–98.57	3	7	0	0	4	14	1
PRKCBI	16	1.0	32.50–32.66	0	2	3	0	4	3	0
NFI	17	1.0	38.10–38.26	0	0	2	2	5	1	0
SCYA2‖	17	3.0	40.21–40.37	1	5	1	1	3	1	1
PAI2	18	2.7	64.17–64.01	0	0	2	1	5	2	10
IL17R‖	22	5.9	14.48–14.32	2	1	2	1	2	0	4
HCF2‖	22	2.0	17.91–17.75	1	1	2	2	1	0	3

* Abbreviations from LocusLink (www.ncbi.nlm.nih.gov/LocusLink/list.cgi).

† For three regions the genetic and physical maps were inconsistent and no estimates were made.

‡ Span of region within a radiation hybrid map (http://www.ncbi.nlm.nih.gov/genome/seq/HsHome.shtml): 1 Mb ≈ 1 centirad. Position of the core SNP is listed first.

§ Number of SNPs discovered with at least 15 copies of the minor allele (successfully genotyped in at least 32 individuals) and in Hardy–Weinberg equilibrium using a significance criterion of $P < 0.02$.

‖ The ten regions selected for follow-up genotyping.

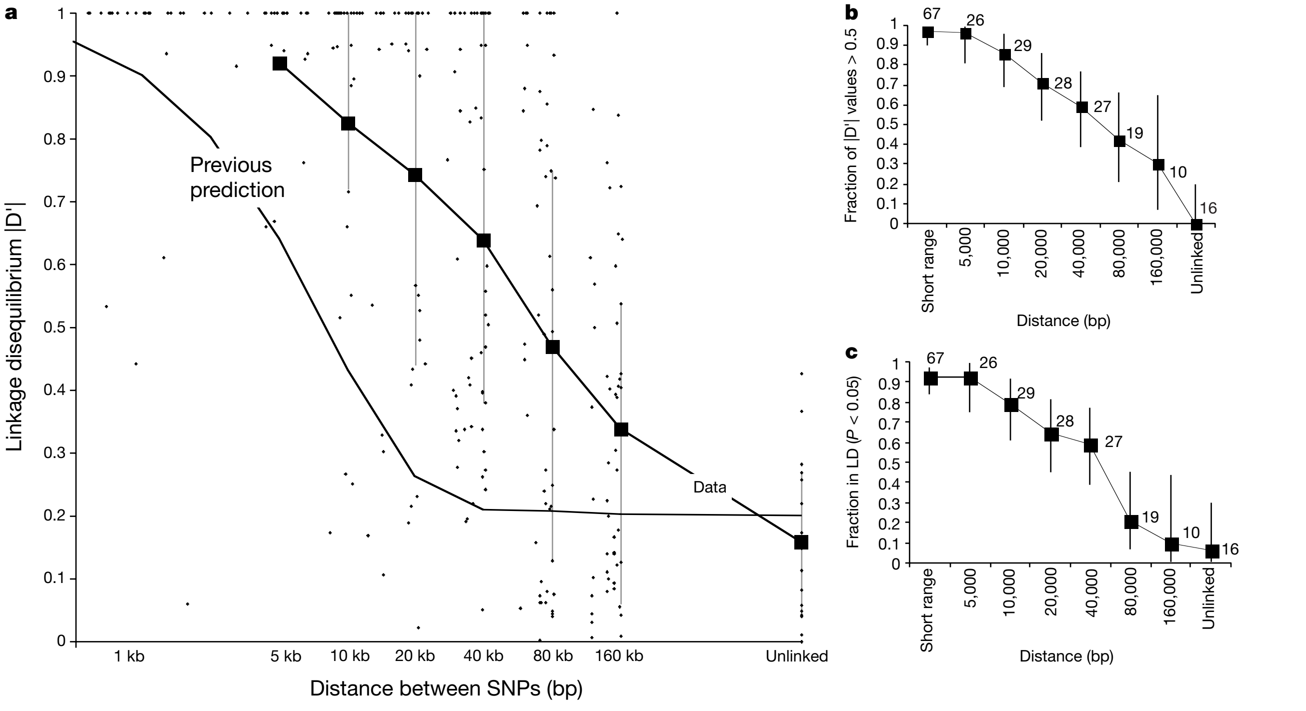

Figure 1 LD versus physical distance between SNPs. For each distance from the core SNP (Table 1), we chose the SNP with the largest number of copies of the minor allele for comparison to SNPs at other distances. At a given distance, all comparisons are independent. **a**, Average |D′| values for each distance separation ('Data'; dotted lines indicate the 25th and 75th percentiles), compared with a prediction[2] based on simulations (see Methods). |D′| values for shorter physical distances were calculated by looking within contiguously sequenced stretches of DNA containing at least two SNPs, and picking the two with the most minor alleles. Unlinked marker comparisons are obtained by comparing SNPs in the 40-kb bin in each row of Table 1 to those in the next row. **b**, **c**, Fraction of |D′| values greater than 0.5 (**b**) and proportion of significant ($P < 0.05$) associations (**c**) between two SNPs separated by a given distance (as assessed by a likelihood ratio test[10]). Bars indicate 95% central confidence intervals. The number of data points used to make the calculations are shown.

genomic regions (see also ref. 12), which is apparent in the different rates at which LD declines around the core SNP (Fig. 2). For example, $|D'| > 0.5$ for at least 155 kb around the *WASL* gene, but for less than 6 kb around *PCI* (Fig. 2). The variability across different genomic regions within the same population sample provides a context for explaining why past empirical studies, each based on one to three regions[3–5], have produced such different results. Large variations in LD are expected because of stochastic factors, such as different gene histories across loci[13]. Differences in recombination rates among regions can also affect the extent of LD. We observe a significant and important correlation ($P < 0.005$) between LD and the estimated local recombination rate (Fig. 2, inset).

Another feature of the data is that, near the range of distances at which LD drops off, there is often considerable variability in $|D'|$ values at neighbouring SNPs (for example, around *IL17R*, *SCYA2*, *TGFBI* and *BMP8* in Fig. 2). Such a wide scatter of LD, even for markers close to each other, has been noted before[14], and is due to the underlying haplotypic structure of LD. SNPs marking sections of chromosome with short extents of correlation are likely to display much lower $|D'|$ values than SNPs marking long haplotypes. In regions of high haplotype diversity, several SNPs may have to be genotyped to have a good chance of tagging most haplotypes. LD-based gene mapping may therefore require clusters of closely spaced SNPs to have maximal power.

Why does LD extend so far? LD around an allele arises because of selection or population history—a small population size, genetic drift or population mixture—and decays owing to recombination, which breaks down ancestral haplotypes[15]. The extent of LD

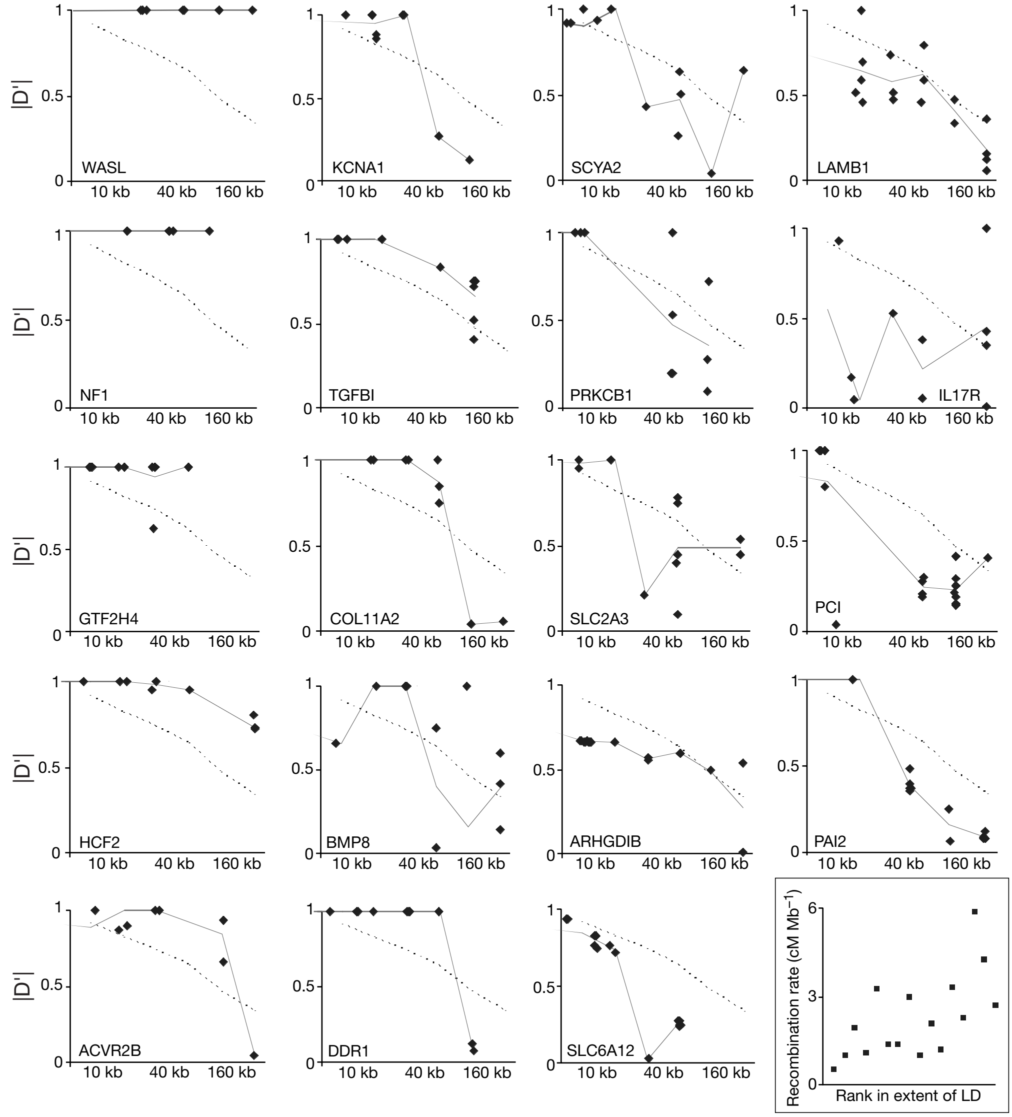

Figure 2 LD profiles for each genomic region. The chosen SNP is usually the core SNP itself, unless the core SNP could not be readily genotyped or another SNP with more minor alleles had been identified within 1 kb. In both cases, we substituted the closest high-frequency SNP. The chosen SNP is compared with every other high-frequency SNP in the same genomic region. Solid line indicates average $|D'|$ values for each distance for which SNPs were available; for comparison, the dashed line indicates the consensus LD curve from Fig. 1. The extent of LD was calculated by performing a least-squares linear regression to the average $|D'|$ values at each distance from the chosen SNP; more sloped lines indicate less LD. The regions are ordered according to the extent of LD (most extensive LD, top left; least extensive LD, bottom right). Inset (bottom right) shows the rank of each region in terms of LD extent versus the estimated recombination rate per unit of physical distance. For each 160-kb region of interest, we looked for the closest pair of flanking genetic markers from the Marshfield map[34] subject to the condition that they were separated by a non-zero genetic distance on the map. We divided genetic map distance by the physical map distance on the basis of the available draft genome sequence[35]. We analysed the 16 regions for which the genetic and physical map orderings of markers were locally consistent (one-sided Spearman rank correlation, $P < 0.005$).

decreases in proportion to the number of generations since the LD-generating event. The simplest explanation for the observed long-range LD is that the population under study experienced an extreme founder effect or bottleneck: a period when the population was so small that a few ancestral haplotypes gave rise to most of the haplotypes that exist today. Our simulations show that a severe bottleneck (inbreeding coefficient $F \geq 0.2$) occurring 800–1,600 generations ago (about 27,000–53,000 years ago assuming 25 years per generation) could have generated the LD observed (Fig. 3). In principle, long-range LD could also be generated by population mixture[16], but the degree of LD is much greater than would arise from the mixing of even extremely differentiated populations. An alternative explanation for the observed long-range LD is that the recombination rates in the regions studied might be markedly less than the genome-wide average. This could happen if recombination occurred primarily in well-separated hotspots and our regions fell between them (Fig. 3). However, under this hypothesis, the regions would be expected to show long-range LD in all populations, and this pattern is not observed (see below).

To confirm our findings of long-range LD and to investigate the reasons for its occurrence, we next examined a representative subset of SNPs in two additional samples. We first studied another north-European sample (48 southern Swedes) and found LD in a nearly identical pattern to that observed in Utah, both in terms of the overall magnitude of LD and the particular alleles that were associated (indicated by the sign of D′) (Fig. 4). The similarity in LD patterns may be due to the same historical event, which occurred deep in European prehistory before the separation of the ancestors of these two groups. This suggests that the long-range LD pattern is general in northern Europeans[3,17].

We next studied 96 Yorubans (from Nigeria), believed to share common ancestry with northern Europeans about 100,000 years ago[18]. At short distances, the Nigerian and European-derived populations typically show the same allelic combinations (Fig. 4): D′ has the same sign and a similar magnitude, indicating a common LD-generating event tracing far back in human history. However, the half-length of LD seems to extend less than 5 kb (Fig. 4) in the Yorubans. Markedly shorter range LD in sub-Saharan Africa has also been observed in several studies of single regions[19,20] (although

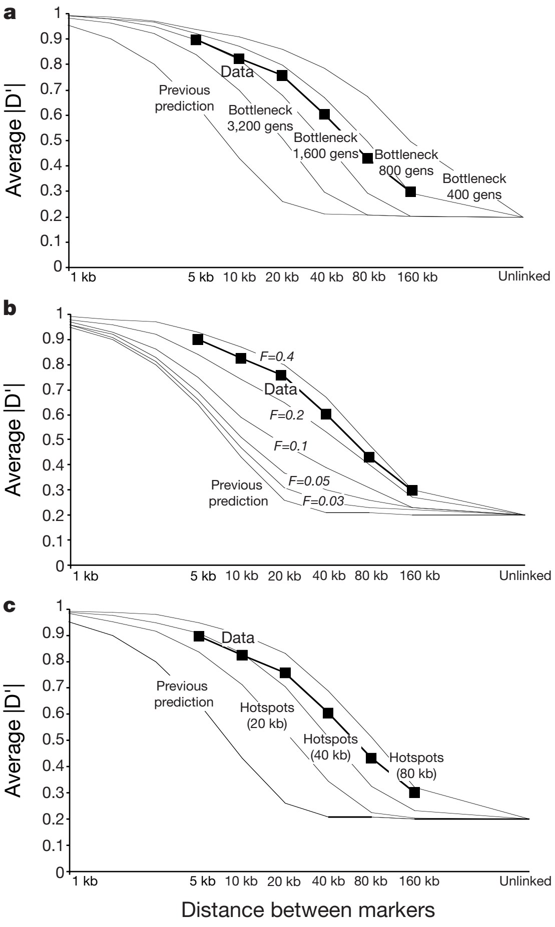

Figure 3 Effect on LD of assumptions about population history and recombination. **a**, The effect of a population bottleneck instantaneously reducing the population to a constant size of 50 individuals for 40 generations ($F = 0.4$) and occurring 400, 800, 1,600 or 3,200 generations ago. **b**, The effect of bottleneck intensity (F) for a bottleneck that occurred 800 generations ago. **c**, The effect of variation in recombination rate, assuming that all recombination in the genome occurs at hotspots randomly distributed according to a Poisson process with an average density of one every 20, 40 and 80 kb, and with a genome-wide average rate of 1.3 centiMorgans per megabase per generation[35]. Results are compared to a no-hotspot model for the same historical hypothesis. ('Data' refers to the mean |D′| values at each physical distance separation, obtained from Fig. 1a).

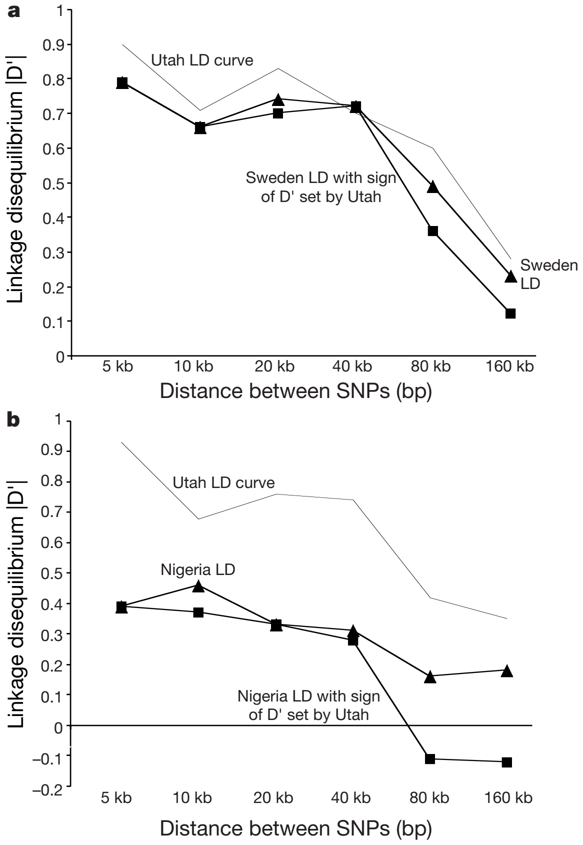

Figure 4 LD curve for Swedish and Yoruban samples. To minimize ascertainment bias, data are only shown for marker comparisons involving the core SNP. Alleles are paired such that $D' > 0$ in the Utah population. $D' > 0$ in the other populations indicates the same direction of allelic association and $D' < 0$ indicates the opposite association. **a**, In Sweden, average D′ is nearly identical to the average |D′| values up to 40-kb distances, and the overall curve has a similar shape to that of the Utah population (thin line in **a** and **b**). **b**, LD extends less far in the Yoruban sample, with most of the long-range LD coming from a single region, *HCF2*. Even at 5 kb, the average values of |D′| and D′ diverge substantially. To make the comparisons between populations appropriate, the Utah LD curves are calculated solely on the basis of SNPs that had been successfully genotyped and met the minimum frequency criterion in both populations (Swedish and Yoruban).

two other studies did not show a clear trend[21,22]). Our results indicate that the pattern of shorter LD in sub-Saharan African populations may be general.

Notably, LD in the Nigerians is largely a subset of what is seen in the northern Europeans. The Yoruban haplotypes are generally contained within the longer Utah haplotypes, and there is little Yoruban-specific LD (85% of observations of substantial LD ($|D'| > 0.5$) in Yorubans are also substantial and of the same sign in Utah). The vast difference in the extent of LD between populations points to differences owing to population history, probably a bottleneck or founder effect that occurred among the ancestors of north Europeans after the divergence from the ancestors of the Nigerians. The short extent of LD in Nigerians is more consistent with the predictions of a computer simulation study assuming a simple model of population expansion[2].

Could the apparent differences in the extent of LD among populations be due to 'ascertainment bias' in the identification of the SNPs? The core SNPs are probably not subject to bias because they were identified in a multi-ethnic population. The neighbouring SNPs were identified in the Utah population and subsequently studied in the other populations, and thus they may be susceptible to ascertainment bias. However, we selected only SNPs with high frequency in Utah and most of these satisfied the high-frequency criterion for use in the other populations (87% in Sweden and 71% in Nigeria). Thus, the inferences about LD are not likely to be much different from what would have been obtained had we used SNPs ascertained in the Yoruban sample. Moreover, the cross-population comparisons (Fig. 4) minimize ascertainment bias because they involve only the core SNP, and because they calculate LD in each population using only the SNPs present in both.

What was the nature of the population event that created the long-range LD? The event could be specific to northern Europe, which was substantially depopulated during the Last Glacial Maximum (30,000–15,000 years ago), and subsequently recolonized by a small number of founders[23,24]. Alternatively, the long-range LD could be due to a severe bottleneck that occurred during the founding of Europe or during the dispersal of anatomically modern humans from Africa[19,20,25,26] (the proposed 'Out of Africa' event) as recently as 50,000 years ago. Under the first hypothesis, the strong LD at distances ≥ 40 kb would be absent in populations not descended from northern Europeans. Under the second hypothesis, the same pattern of long-range LD could be observed in a variety of non-African populations. Regardless of the timing and context of the bottleneck, the severity of the event (in terms of inbreeding) can be assessed from our data. To have a strong effect on LD, a substantial proportion of the modern population would have to be derived from a population that had experienced an event leading to an inbreeding coefficient of at least $F = 0.2$ (Fig. 3). This corresponds to an effective population size (typically less than the true population size[15]) of 50 individuals for 20 generations; 1,000 individuals for 400 generations; or any other combination with the same ratio.

Our results have implications for disease gene mapping, suggesting a possible two-tiered strategy for using LD. The presence of large blocks of LD in northern European populations suggests that genome-wide LD mapping is likely to be possible with existing SNP resources[1]. Although the large blocks should make initial localization easier, they may also limit the resolution of mapping to blocks of DNA in the range of 100 kb[27]. Populations with much smaller blocks of LD (for example, Yorubans) may allow fine-structure mapping to identify the specific nucleotide substitution responsible for a phenotype[12]. Our study also has implications for LD as a tool to study population history[19–22]. Simultaneous assessment of LD at multiple regions of the genome provides an approach for studying history with potentially greater sensitivity to certain aspects of history than traditional methods based on properties of a single locus. □

Methods

Core SNPs were identified by screening more than 3,000 genes in a multi-ethnic panel of 15 European Americans, 10 African Americans, and 7 East Asians (see ref. 28 for details; a full description of this database will be presented elsewhere). DNA used for sequencing was obtained from the Coriell Cell Repositories. Identification numbers for these Utah samples from the CEPH mapping panel were NA12344, 06995, 06997, 07013, 12335, 06990, 10848, 07038, 06987, 10846, 10847, 07029, 07019, 07048, 06991, 10851, 07349, 07348, 10857, 10852, 10858, 10859, 10854, 10856, 10855, 12386, 12456, 10860, 10861, 10863, 10830, 10831, 10835, 10834, 10837, 10836, 10838, 10839, 10841, 10840, 10842, 10843, 10845 and 10844. We did follow-up genotyping in 48 Swedes (healthy individuals from a case/control study of adult-onset diabetes) and in 96 Yoruban males from Nigeria (healthy individuals from a case/control study of hypertension).

SNPs were discovered by DNA sequencing[28] in the 44 individuals from Utah; we sequenced about 2 kb centred at each distance from the core SNP ≥ 5 kb, with about 1 kb sequenced around the core SNP itself. When no polymorphism of sufficiently high frequency was found, a nearby subregion of about two further kilobases was resequenced; this occurred in only 18% of the cases. Polymorphisms were identified and genotypes were scored automatically using Polyphred[29] and checked manually by at least two different scorers. SNPs in Hardy–Weinberg disequilibrium or showing evidence of breakdown of LD over short physical distances (< 2.5 kb) were triple-checked. Of the 275 high-frequency SNPs (that is, SNPs with at least 15 observed copies of the minor allele), three were discarded because of a Hardy–Weinberg P value of < 0.02; one of the SNPs used in the analysis had a nominally significant P value ($P < 0.05$). To assess the accuracy of scoring, we rescored 26 randomly chosen high-frequency SNPs; only seven discrepancies were found among 1,144 genotypes. For cases in which follow-up genotyping was done, the discrepancy rate was 47 out of 1,484 (3%) between genotypes obtained by both methods.

Genotyping of SNPs was performed by single-base extension followed by mass spectroscopy (Sequenom)[30], fluorescence polarization (LJL Biosystems)[31] or detection on a sequencing gel (Applied Biosystems)[32]. For the ten regions selected for follow-up genotyping (Table 1), we chose at most one 'representative' SNP at each distance from the core SNP (each column in Table 1) according to the criterion that it had the highest number of minor alleles of all SNPs at that distance from the core SNP. For other populations, only those SNPs that, when genotyped, had a minimum number of minor alleles were included in studies of LD. For Yorubans, the cutoff was 25 alleles (76% of SNPs met the criterion); for Swedes, the cutoff was 15 alleles (89%). The fact that most of the SNPs we studied in Utah are also present in high frequency in these other populations indicates that the assessment of LD is not likely to be subject to large ascertainment bias.

Heterozygosity[15] (π) was calculated as the average of $2jk/n(n-1)$ for all base pairs screened, with j and k equal to the number of copies of the minor and major alleles, respectively ($n = j + k$). A base was considered screened if it had Phred quality scores[29] of ≥ 15 in ≥ 10 individuals. D' values between markers with alleles A/a and B/b (allele frequencies, c_A, c_a, c_B and c_b; haplotype frequencies, c_{AB}, c_{Ab}, c_{aB} and c_{ab}) were obtained by dividing $c_{AB} - c_A c_B$ by its maximum possible value: $\min(c_A c_b, c_a c_B)$ if $D > 0$ and $\min(c_A c_B, c_a c_b)$ otherwise. An implementation of the expectation maximization algorithm was used to infer haplotype frequencies for pairs of SNPs both for actual and simulated data[33]. A likelihood ratio test was used to assess significance of associations between pairs of SNPs[10].

Computer simulations were based on a model related to that in ref. 2, assuming a population that was constant at an effective size of 10,000 individuals until 5,000 generations ago, when it expanded instantaneously to a size of 100,000,000 (an arbitrary value). This model captures many of the features of more complicated growth, as the effect of population growth on LD is not dependent on the precise details of population growth or the final population size when the growth is moderately fast[2]. Bottlenecks were modelled as described, with the population crashing to a constant size for a fixed number of generations before re-expanding. (The effect of a bottleneck on LD depends primarily on the F-value, the inbreeding coefficient, which is defined as the probability that two alleles randomly picked from the population after the bottleneck derive from the same ancestral allele just before the bottleneck.) Coalescent simulations were used to generate gene genealogies under these models for markers separated by a specified recombination distance (see ref. 13 for a more detailed description of the theory behind these simulations). Simulations were run 2,000 times with sample-size distributions mimicking our data. SNPs were generated by distributing mutations on the simulated gene genealogies at a mutation rate of 6×10^{-5} per generation, under an 'infinite alleles' mutation model. The mutation rate was chosen such that the probability of high frequency SNPs in a 2-kb stretch of DNA sequenced in 44 samples (for the model of a simple expansion 5,000 generations ago) was similar to what we observed (about 70%). We also tested mutation rates ten times higher and found that inferences about LD were essentially unchanged.

An extreme hypothesis of population mixture and its effect on LD were assessed in a simulated, mixed population of European Americans and sub-Saharan Africans. For the first simulated mixture, we constructed a mixture of 22 Yorubans and 22 samples from Utah, and used data from the ten core SNPs genotyped in both populations. For the second simulation, we used 26 SNPs from a previous study[7] that had been found to have a minor allele frequency of at least 15% in either African Americans or European Americans; we chose at most one SNP per gene, picking the first listed SNP (in Table A1 of ref. 7) that met our minimum frequency criterion. We found much stronger LD even at 40 kb (56% with $|D'| > 0.5$) (Fig. 1) in our actual data than in the simulated, admixed populations. For the 45 possible pairwise comparisons of the ten core SNPs in the simulated mix of Yoruban and Utah samples, no values of $|D'| > 0.5$ were observed. For the 325 possible pairwise comparisons from the second study, only 11% showed $|D'| > 0.5$. This suggests that admixture probably did not generate the strong signal of LD at long physical distances seen in Utah.

Received 11 December 2000; accepted 13 March 2001.

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. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22,** 139–144 (1999).
3. Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67,** 1544–1554 (2000).
4. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68,** 191–197 (2001).
5. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25,** 324–328 (2000).
6. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96,** 15173–15177 (1999).
7. Goddard, K. A. B., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66,** 216–234 (2000).
8. Watterson, G. A. & Guess, H. A. Is the most frequent allele the oldest? *Theor. Pop. Biol.* **11,** 141–160 (1977).
9. Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).
10. Schneider, S., Kueffler, J. M., Roessli, D. & Excoffier, L. Arlequin (ver. 2.0): A software for population genetic data analysis (Genetics and Biometry Laboratory, Univ. Geneva, Switzerland, 2000).
11. Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120,** 849–852 (1988).
12. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10,** 1435–1444 (2000).
13. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. J. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
14. Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63,** 595–612 (1998).
15. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer, Massachusetts, 1997).
16. Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85,** 9119–9123 (1988).
17. Eaves, I. A. *et al.* The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25,** 320–322 (2000).
18. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl Acad. Sci. USA* **92,** 6723–6727 (1995).
19. Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271,** 1380–1387 (1996).
20. Tishkoff, S. A. *et al.* Short tandem-repeat polymorphism/*Alu* haplotype variation at the *PLAT* locus: Implications for modern human origins. *Am. J. Hum. Genet.* **67,** 901–925 (2000).
21. Kidd, J. R. *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. *Am. J. Hum. Genet.* **66,** 1882–1899 (2000).
22. Mateu, E. *et al.* Worldwide genetic analysis of the *CFTR* region. *Am. J. Hum. Genet.* **68,** 103–117 (2001).
23. Housley, R. A., Gamble, C. S., Street, M. & Pettitt, P. Radiocarbon evidence for the Late glacial human recolonisation of northern Europe. *Proc. Prehist. Soc.* **63,** 25–54 (1994).
24. Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67,** 1251–1276 (2000).
25. 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).
26. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–713 (2000).
27. Altshuler, D., Daly, M. & Kruglyak, L. Guilt by association. *Nature Genet.* **26,** 135–137 (2000).
28. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
29. Nickerson, D. B., Tobe, V. O. & Taylor, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based sequencing. *Nucleic Acids Res.* **25,** 2745–2751 (1997).
30. Ross, P. Hall, L., Smirnov, I. & Haff, L. High level multiplex genotyping by MALDI-TOF mass spectroscopy. *Nature Biotech.* **16,** 1347–1351 (1998).
31. Chen, X., Levine, L. & Kwok, P. Y. Fluorescence polarization in homogenous nucleic acid analysis. *Genome Res.* **9,** 492–498 (1999).
32. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24,** 381–386 (2000).
33. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12,** 921–927 (1995).
34. 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,** 861–689 (1998).
35. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Acknowledgements

We thank L. Groop for the Swedish samples; R. Cooper and C. Rotimi for the Yoruban samples; and D. Altshuler, M. Daly, D. Goldstein, J. Hirschhorn, C. Lindgren and S. Schaffner for discussions. This work was supported in part by grants from the National Institutes of Health, Affymetrix, Millennium Pharmaceuticals, and Bristol-Myers Squibb Company, and by a National Defense Science and Engineering Fellowship to D.E.R.

Correspondence and requests for materials should be addressed to D.E.R. (e-mail: reich@genome.wi.mit.edu) or E.S.L. (e-mail: lander@genome.wi.mit.edu).

The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association

Enrique Martinez-Perez, Peter Shaw & Graham Moore

John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK

The correct pairing and segregation of chromosomes during meiosis is essential for genetic stability and subsequent fertility. This is more difficult to achieve in polyploid species, such as wheat, because they possess more than one diploid set of similar chromosomes. In wheat, the *Ph1* locus ensures correct homologue pairing and recombination[1]. Although clustering of telomeres into a bouquet early in meiosis has been suggested to facilitate homologue pairing[2,3], centromeres associate in pairs in polyploid cereals early during floral development[4]. We can now extend this observation to root development. Here we show that the *Ph1* locus acts both meiotically and somatically by reducing non-homologous centromere associations. This has the effect of promoting true homologous association when centromeres are induced to associate. In fact, non-homologously associated centromeres separate at the beginning of meiosis in the presence, but not the absence, of *Ph1*. This permits the correction of homologue association during the telomere-bouquet stage in meiosis. We conclude that the *Ph1* locus is not responsible for the induction of centromere association, but rather for its specificity.

We previously showed that centromeres associate in pairs before meiosis in polyploid cereals, but not until the beginning of meiosis in their diploid progenitors[4]. Using fluorescence *in situ* hybridization on intact root sections, we now report that centromeres also associate in pairs in developing xylem vessel cells of bread wheat (AABBDD, $2n = 6x = 42$) but not in those of its diploid progenitors (Fig. 1b, e, and Table 1). Moreover, we show that during this developmental process the chromosomes endoreplicate, becoming polytene. This is indicated by the substantial increase in size of the interphase chromosomes (and nucleus), as compared with the surrounding tissues (Fig. 1a, d, f, g).

The level of centromere association in xylem vessel cells of wheat is unaffected by the presence of *Ph1*, as in floral development[5] (Fig. 1e, h). Thus, neither endoreplication nor the *Ph1* locus can induce centromere association. Although centromeres associate in the xylem vessel cells in the presence and absence of *Ph1*, they are not associated in other root tissues (Fig. 1c). Polyploidy is therefore necessary but not sufficient to induce centromere association—a specific developmental context is also required, as in meiosis, floral development or xylem vessel development.

We have assessed homologue association in these vessel cells by labelling specific pairs of rye chromosomes in wheat–rye addition lines. These rye homologues associate at a high level through their centromeres during vessel development in the presence of *Ph1* (25/25

Table 1 Statistics of the number of centromeres

		Ph1+	*Ph1–*	*t*-test
Wheat–rye	Premeiosis	19.5 (2.7)	16.3 (2.9)	$P < 0.001$
	Telomere bouquet	23.5 (1.5)	13.7 (2.1)	$P < 0.001$
	Xylem vessel	20 (1.7)	16.3 (1.8)	$P < 0.001$
	Non-polytene root	24.9 (1.2)	25.2 (1.3)	$P = 0.6$
Wheat	Xylem vessel	21.7 (2.1)	21.5 (1.8)	$P = 0.798$
T. monococcum	Xylem vessel	12.56 (0.8)		

The s.d. is given in parentheses. Student's *t*-test was used to test the null hypothesis that the two means in the presence and absence of *Ph1* are the same. The null hypothesis can be discounted in all comparisons except the wheat xylem and the wheat–rye non-polytene root. All centromere sites were counted on the original three-dimensional confocal stacks.