Human Molecular Genetics, 2006, Vol. 15, No. 5 ***789–795***
doi:10.1093/hmg/ddi493
Advance Access published on January 25, 2006

Extended tracts of homozygosity in outbred human populations

Jane Gibson*, Newton E. Morton and Andrew Collins

Human Genetics Research Division, School of Medicine, University of Southampton, Southampton SO16 6YD, UK

Received November 25, 2005; Revised and Accepted January 18, 2006

Long tracts of consecutive homozygous single nucleotide polymorphisms (SNPs) can arise in the genome through a number of mechanisms. These include inbreeding in which an individual inherits chromosomal segments that are identical by descent from each parent. However, recombination and other processes break up chromosomal segments over generations. The longest tracts are therefore to be expected in populations with an appreciable degree of inbreeding. We examined the length, number and distribution of long tracts of homozygosity in the apparently outbred HapMap populations. We observed 1393 tracts exceeding 1 Mb in length among the 209 unrelated HapMap individuals. The longest was an uninterrupted run of 3922 homozygous SNPs spanning 17.9 Mb in a Japanese individual. We find that homozygous tracts are significantly more common in regions with high linkage disequilibrium and low recombination, and the location of tracts is similar across all populations. The Yoruba sample has the fewest long tracts per individual, consistent with a larger number of generations (and hence amount of recombination) since the founding of that population. Our results suggest that multiple-megabase-scale ancestral haplotypes persist in outbred human populations in broad genomic regions which have lower than average recombination rates. We observed three outlying individuals who have exceptionally long and numerous homozygous tracts that are not associated with recombination suppressed areas of the genome. We consider that this reflects a high level of relatedness in their ancestry which is too recent to have been influenced by the local recombination intensity. Possible alternative mechanisms and the implications of long homozygous tracts in the genome are discussed.

INTRODUCTION

An individual is homozygous for a diallelic marker if both alleles at that marker are identical. In some individuals, long tracts can be found where homozygous markers occur in an uninterrupted sequence. The most obvious explanation for such tracts is autozygosity, where the same chromosomal segment has been passed to a child from parents who inherited it from a common ancestor. As recombination interrupts long chromosome segments over time, the length of a homozygous segment depends, in part, on the time since the last common ancestor of the parents. Therefore, in an inbred population we expect to see longer homozygous segments than in outbred populations.

The International HapMap project began in 2002 with the aim of cataloguing human variation by genotyping (Phase I) over a million single nucleotide polymorphisms (SNPs) in 269 individuals across four populations: CEPH Utah residents with ancestry from northern and western Europe (CEU), Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT) and Yoruba from Ibadan, Nigeria (YRI). These data were periodically released into the public domain (http://www.hapmap.org) and provide a useful resource for genetic research. The completed Phase I provides SNP data with an average density of 1 SNP every 5 kb. The second phase of data release comprises another 2.1 million SNPs with an average of 1 SNP every 600 bp (1).

We might expect to see relatively short segments of homozygosity in the apparently outbred HapMap populations. However, using the SNP genotype data we find regions of many megabases that are homozygous and therefore likely to be identical by descent. Long tracts of homozygosity have been recorded previously in CEPH individuals (2). Broman and Weber used 8000 short tandem-repeat polymorphisms (STRPs) in CEPH families and found several families with long homozygous segments >10 cM in length.

*To whom correspondence should be addressed. Tel: +44 2380796939; Fax: +44 2380794264; Email: jg@soton.ac.uk

They examined the role of possible typing error, back-mutation of STRPs, gene conversion events and the limitations imposed by locally low marker density in determining the limits of homozygous segments. They were able to determine relationships between 'unrelated' individuals in some pedigrees, but there remained a degree of autozygosity approaching or exceeding that expected in the progeny of a first cousin mating where relationships were not detected.

In contrast to STRPs, SNPs are thought to be of more ancient origin. We might therefore expect to see, in comparison, fewer and shorter homozygous tracts in SNP maps. However, this is partly offset by the relatively reduced mutation rate of these markers which might allow the longer tracts to remain unbroken over more generations. Various factors may influence the length, abundance and location of homozygous tracts including the mutation rate, population structure, uniparental disomy (UPD), natural selection, recombination and linkage disequilibrium (LD) patterns.

The extremely dense SNP typing in the HapMap sample provides a unique opportunity to examine the distribution, size and location of homozygous tracts and their relationship to recombination/LD patterns and also to consider the other mechanisms.

LD is the tendency for alleles to be inherited together more often than would be expected under random segregation. In the human genome, there are regions of strong LD broken up by small (punctate) regions of intense recombination (3). The pattern of LD can be represented by maps in which each marker is given a location in LD units (LDUs) (4). The LDU map is analogous to the linkage map in centiMorgans in that distances are additive and can be plotted against the physical map to reveal strikingly similar contours (5). The main determinant of the LD structure is recombination, shown by a comparison of genome-wide LDU and linkage map lengths which gave R^2 values of 0.97 for chromosome arms and 0.94 for their deciles (6). The contour of the LDU map identifies steps which correspond to regions of high recombination and plateaus which reflect recombination cold regions (7). In addition to recombination, lesser effects, which distort the simple relationship between the LDU and linkage maps, involve selection, drift and mutation amongst others. The overall length of the LDU map reflects duration since an effective population bottleneck (8). LDU maps of the 22 autosomes and the X chromosome, based on HapMap release 16, can be found in the LD database (LDDB) (http://cedar.genetics.soton.ac.uk/public_html/).

In this paper, we characterize long tracts of homozygosity in the human genome represented by the HapMap sample and examine the relationship between the location and size of long tracts of homozygosity, recombination and LD patterns and also examine evidence for recent inbreeding in certain HapMap individuals.

RESULTS

Among the 209 unrelated individuals in four populations, we found a total of 1393 homozygous tracts greater than 1 Mb in length with a minimum SNP density of 1 SNP every 5 kb (Table 1), the 20 longest tracts are presented in Table 2.

Table 1. The population sample and the number and maximum length of tracts examined

HapMap population sample	No. unrelated individuals	No. tracts	Max. tract length (Mb)
CEPH Utah residents with ancestry from northern and western Europe (CEU)	60	498	6.48
Han Chinese Beijing (CHB)	45	263	2.63
Japanese Tokyo (JPT)	44	370	17.91
Yoruba Ibadan Nigeria (YRI)	60	262	11.14
All	209	1393	17.91

The longest tract over all populations is 17.9 Mb in an individual from the JPT sample. This tract comprises 3922 consecutive homozygous SNPs. The CHB sample has the smallest maximum tract length at 2.63 Mb.

To determine whether the locations of homozygous tracts are related to the LD structure, we obtained the correlation between tract coverage in each 1 Mb segment of the genome (see Materials and Methods) and LDU/Mb. We also performed linear regression using LDU/Mb as the dependent variable and tract coverage as the independent variable. Analyses were carried out for each population sample separately and for the concatenated sample. All the correlations were significant at $P < 0.0001$ with correlation coefficients of around -0.3 for all samples. The regression analyses were also significant ($P < 0.0001$) with R^2 values ranging from 8 to 10% (Table 3). To confirm that the location of homozygous tracts is directly related to the recombination pattern, we carried out the same analysis but with cM/Mb replacing LDU/Mb. This is important because the structure of the LDU map must partly reflect the presence of homozygous tracts in the sample. The linkage map is independent and therefore avoids any circularity. The linkage map used (9) comprises 14 759 polymorphic markers. Again all the results were significant ($P < 0.0001$) with correlation coefficients of -0.2 and R^2 values ranging from 4 to 5% (Table 4).

We then considered whether a region of the genome with a large number of tracts in one population also had a large number of tracts in other populations. We examined the correlations between tract coverage in each megabase for all populations. The results show (Table 5) that all correlations are significant ($P < 0.0001$) and the correlation coefficients range from 0.27 to 0.68. The YRI and CEU samples show the least similarity and CHB and JPT show the most similarity.

Individuals were then examined within each population by counting the total number of homozygous tracts in each individual. The average tract count per individual was 4.4, 5.3, 8.3 and 8.4 for the YRI, CHB, CEU and JPT samples, respectively (Fig. 1). We found three individuals with particularly high tract counts, one in the CEPH sample (NA12874) and two in the Japanese sample (NA18992 and NA18987).

The amount of LD in regions where homozygous tracts occur was investigated for each individual. From analysis of the HapMap-derived LDU maps in the LDDB, the genome-wide averages for LDU/Mb are between 20.2 and 22.6 for the CEU, JPT and CHB populations and 28.4 for the YRI population (Table 6). We then computed, for all tracts and

Table 2. The 20 longest homozygous tracts in the sample

Population	Chromosome	Start of tract		End of tract		Tract length (kb)	Tract length (LDU)	No. SNPs in tract	Tract SNP density	Person ID
		SNP	kb	SNP	kb					
JPT	2	rs13395300	18043.375	rs3914801	35948.498	17905.123	308.462452	3922	0.219	NA18992
JPT	4	rs4689486	6675.164	rs2946399	23664.462	16989.298	387.022169	3984	0.235	NA18992
JPT	16	rs3789027	3784.641	rs4985167	15049.304	11264.663	390.773261	2466	0.219	NA18992
YRI	18	rs4800285	23613.506	rs4583326	34754.867	11141.361	277.694017	5446	0.489	NA19140
JPT	3	rs7643092	95503.475	rs6437591	106502.781	10999.306	118.560068	2540	0.231	NA18992
JPT	3	rs1963442	75710.872	rs13316811	86312.69	10601.818	86.764281	2496	0.235	NA18992
JPT	14	rs2269304	63244.095	rs11849858	72838.845	9594.75	120.380145	2064	0.215	NA18987
YRI	2	rs4954220	136083.415	rs12474604	144069.64	7986.225	155.929054	2390	0.299	NA18501
JPT	8	rs12542501	111413.758	rs2514751	119178.091	7764.333	76.407253	3191	0.411	NA18987
YRI	5	rs9784722	71295.611	rs6868351	78046.512	6750.901	246.526579	1531	0.227	NA19201
JPT	3	rs6766522	78998.788	rs12493621	85603.115	6604.327	37.095983	1514	0.230	NA18994
CEU	2	rs4592792	192734.56	rs1978890	199212.407	6477.847	65.13898	1418	0.219	NA07056
JPT	3	rs7647011	106506.516	rs2173318	112590.137	6083.621	93.074236	1417	0.233	NA18992
JPT	8	rs7828769	72455.634	rs12677157	78368.814	5913.18	98.779363	2452	0.415	NA18987
JPT	3	rs276117	81402.93	rs2043663	86959.055	5556.125	30.504793	1357	0.244	NA18964
JPT	3	rs4855928	118219.228	rs1877378	123690.638	5471.41	63.743805	1306	0.239	NA18972
JPT	14	rs2073346	49219.163	rs999817	54302.95	5083.787	114.422156	1124	0.221	NA18987
YRI	5	rs2301079	83403.686	rs304132	88299.667	4895.981	69.271019	1097	0.224	NA19201
CEU	1	rs6425093	184632.813	rs4276896	189395.178	4762.365	45.055776	1253	0.263	NA12874
JPT	14	rs11627778	54312.032	rs1955700	59036.522	4724.49	84.524193	979	0.207	NA18987

Table 3. The relationship between LDU/Mb and tract coverage

Population sample	Correlation coefficients	Regression R^2
CEU	−0.32	0.10
CHB	−0.29	0.09
JPT	−0.30	0.09
YRI	−0.28	0.08
All	−0.29	0.08

*$P < 0.0001$.

Table 4. The relationship between cM/Mb and tract coverage

Population sample	Correlation coefficients	Regression R^2
CEU	−0.23	0.05
CHB	−0.21	0.04
JPT	−0.21	0.04
YRI	−0.21	0.04
All	−0.21	0.04

*$P < 0.0001$.

all individuals, the LDU and Mb lengths of the genome covered by tracts. As expected, from the co-localization of long homozygous tracts and regions of strong LD (low LDU/Mb) shown above, we find that the mean LDU/Mb in regions containing homozygous tracts is much lower than the genome average, being in the range 6.7–8.3 for the CEU, JPT and CHB populations and 15.4 for the YRI sample (Table 6). However, the same three individuals from the previous analysis are outliers (Fig. 2) as they have tracts in areas that do not have particularly high LD and, in fact, have levels of LD approaching the genome average for their populations, ranging from 15.1 to 17.6 LDU/Mb. To examine if this is significantly different from the LDU density in tract regions generally, we used a regression model weighted by physical size in Mb. LDU/Mb was the dependent variable and x was the independent variable with $x = 1$ for individual NA12874 and $x = 0$ for other individuals in the CEU sample. The same model was used for the JPT sample with two variables to represent the two outliers. We find that the three individuals NA12874, NA18992 and NA18987 have tracts in regions of significantly higher LDU/Mb (less LD) than is typical for homozygous tract regions generally. Regression analysis shows that NA12847 explains 42% of the variance in LDU/Mb in the CEU sample and NA18992 and NA18987 together explain 89% of the variance in LDU/Mb in the JPT sample.

Table 5. Correlation coefficients between populations for tract coverage of each megabase

	CHB	JPT	YRI
CEU	0.51	0.46	0.27
CHB		0.68	0.30
JPT			0.30

*$P < 0.0001$.

DISCUSSION

Our studies show that homozygous tracts are remarkably common and long even in the unrelated individuals from the apparently outbred populations represented in the HapMap

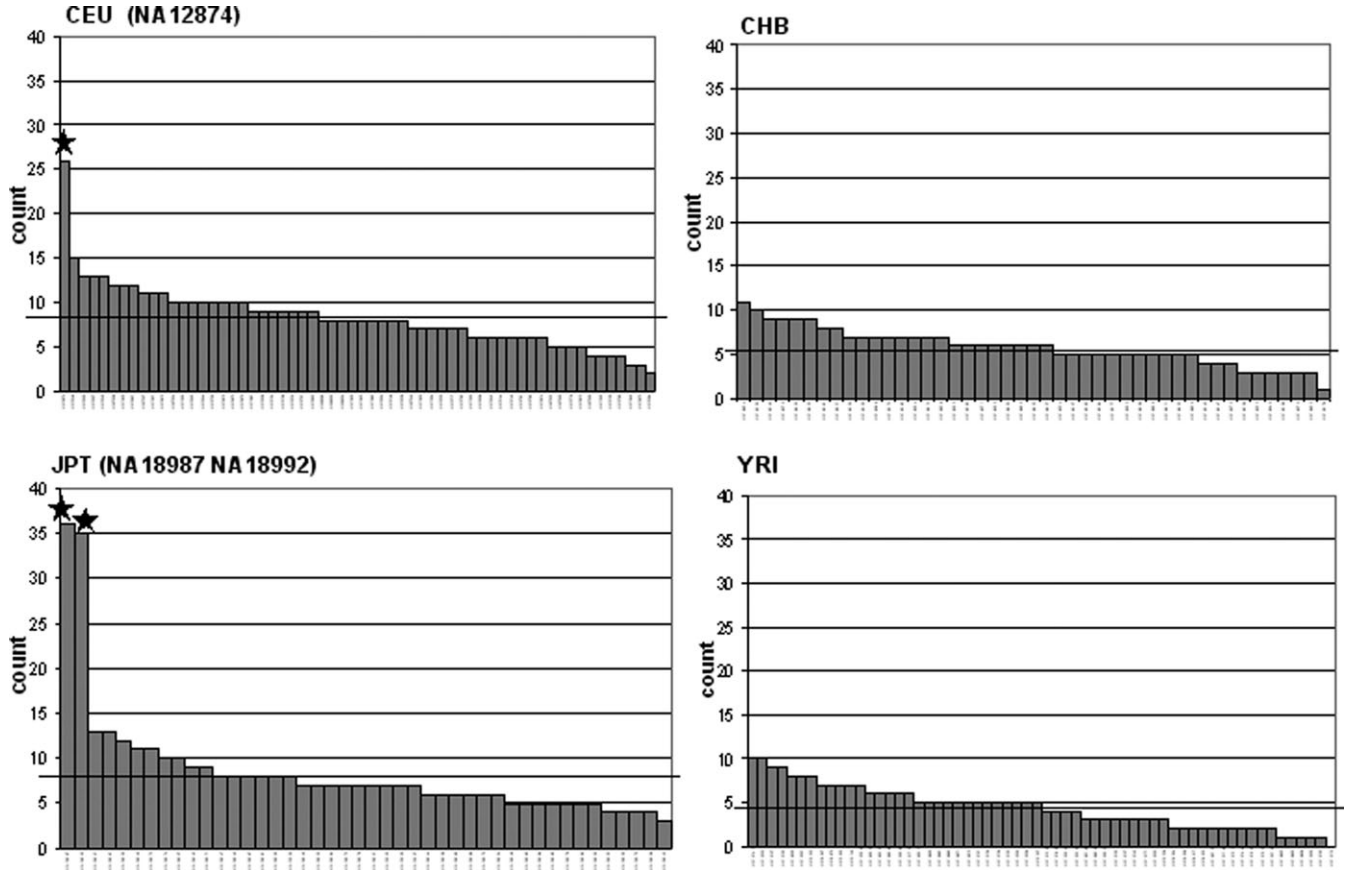

Figure 1. Graphs of the number of tracts (count) per individual for each population sample. The individuals have been ordered by number of tracts. The horizontal line indicates the means of 8.3 for CEU, 8.4 for JPT, 5.8 for CHB and 4.4 for YRI. The stars highlight the three individuals noted to have particularly high tract counts.

Table 6. Comparisons of the strength of LD across the genome with the corresponding values in regions with homozygous tracts

Population	Genome-wide LDU/Mb	Tract regions LDU/Mb
CEU	20.2	8.3
CHB	22.6	6.7
JPT	20.4	7.3
YRI	28.4	15.4

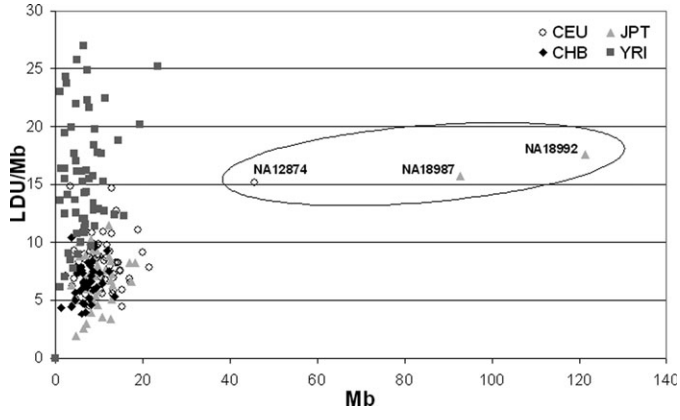

Figure 2. A graph of average LDU/Mb in the regions covered by homozygous tracts in each individual against the total length of homozygous tracts in that individual in Mb. The symbols/shades represent the four population samples. The three individuals circled are individuals noted to have high LDU/Mb (less LD) in the regions covered by tracts and much longer total tract length than the other individuals in their respective population samples.

sample. The evidence indicates that homozygous tracts are generally found in regions of relatively extensive LD and locally low rates of recombination. The presence of relatively short haplotype 'blocks', regions of low haplotype diversity, has been well known for several years (10). Indeed it is the characterization of such regions and subsequent selection of haplotype tagging SNPs, which underpins the International HapMap project (1). However, the presence of much longer homozygous tracts (lengths greatly in excess of 1 Mb) was perhaps not widely anticipated.

Homozygous tracts can occur when a child inherits the same chromosomal segment from both parents, who themselves inherited it from a common ancestor. There are two broad mechanisms by which this could happen. One explanation is that the parents have a relatively recent common ancestor, so there has been little opportunity for recombination to break up the segment. A second possibility is that any relationship between the parents is distant but a

lack of recombination in the region (and therefore high LD) has enabled the ancestral segment to persist intact.

Our analysis considers only homozygous tracts that exceed 1 Mb in length, but there are numerous smaller segments many of which must contribute to the low haplotype diversity characterized in blocks. Given this, the true level of homozygosity in the genome is likely to be much greater than indicated by our analysis of the more extreme examples. Our results show that extensive LD correlates strongly with a higher proportion of homozygous tracts in that region. This is because genomic regions with low recombination allow particularly long chromosome segments to remain intact over time, increasing the chance that they come together in an individual as a homozygous tract.

Analysis of tract coverage between populations shows that tracts tend to be co-localized in all populations. Patterns of LD are also highly similar across human populations (5) which presumably reflects, in part, the close similarity between historical and current recombination patterns for all populations. The co-localization of recombination hot-spots/regions in all human populations allows long homozygous tracts to persist in the shared intervening regions which have low meiotic activity. As might be expected, the YRI population has the fewest long tracts per individual (4.4), reflecting the longer time over which recombination has been breaking haplotypes in this sub-Saharan African population. Among 'out-of-Africa' populations, the CHB sample has the fewest long homozygous tracts and the shortest maximum tract length at 2.63 Mb (Table 1), consistent with lower genome-wide average LD (higher LDU/Mb) than the JPT and CEU samples (Table 6). This may be the consequence of a variety of mechanisms including differences in population history, for example, in the timing, number and intensity of population bottlenecks and/or differences in social conventions such as reduced acceptance of first-cousin marriages.

We assume that all four HapMap samples are representative of relatively outbred human populations and homozygosity is not particularly exaggerated due to a limited number of haplotypes or 'atypical' individuals represented in the sample. However, there is little information about the individuals represented in the data and sample sizes of 44–60 unrelated individuals are fairly small. It is therefore conceivable that the samples are not truly representative of their respective populations. Three individuals stand out as having particularly long tracts and high tract counts; one in the CEPH sample and two in the Japanese sample (Fig. 2). We have shown that the tracts in these individuals are not associated with regions of elevated LD, in contrast to tracts in other individuals. Therefore, it is reasonable to suggest that a different mechanism is involved and we speculate that recent inbreeding is likely. That is, the parents of these individuals have an unknown relationship which goes back a comparatively small number of generations. In the most extreme case (NA18992), long homozygous tracts cover ~4% of the genome of that individual. As only contiguous tracts longer than 1 Mb are included here, the total proportion of identical-by-descent homozygosity is likely to be much higher. However, this is difficult to quantify without detailed pedigree information over a number of generations. Autozygosity of 6% is to be expected in the offspring of a first cousin mating (2). We note that the same two Japanese individuals have been identified independently by the HapMap Consortium (11). The same analysis did not identify the CEPH individual (NA12874) however. The authors surmised that the two Japanese individuals showing 'an above average degree of cryptic relatedness' are likely to have a relatively unimportant influence on the characterized LD structure as they form only a small proportion of the individuals in the sample. It also seems reasonable to assume that the impact on the LDU map by including these individuals is modest although we did not test this directly.

We have considered the role of inbreeding and the persistence of ancestral segments indicated by homozygosity in regions of low recombination. However, there are other mechanisms which might contribute to the observed extent of homozygosity. There are different types of UPD; isodisomy is the form where a child inherits two copies of the same chromosome from the same parent. This results in the child being homozygous at all loci. Segmental isodisomy can occur when a part, but not the whole, chromosome is affected. UPD can cause various diseases when it occurs in a region with imprinted genes and can also cause rare recessive disorders. A case of maternal UPD of chromosome 1 was found by chance (12) as there were no apparent phenotypic effects. This suggests that as well as isodisomy for rare recessive genes or UPD in imprinted regions, some cases of UPD may be asymptomatic and perhaps quite common. The phenomenon has been little studied where not associated with a disease; however, scanning methodologies to detect UPD are being developed, initially as a diagnostic tool, but could be used to answer this question in the future (13,14).

Heterozygous deletions can sometimes be detected by apparent homozygosity over an extended region. For example, a novel 8 kb deletion in a patient with glycogen storage disease type II was found using this method (15). Cell line transformations are also known to result in 'loss' of parts of the genome. Apparent homozygosity may serve as an indicator of the presence of a deletion but other molecular techniques are required to be definitive. SNPs with missing data were not included in this analysis, however, when a run of markers with one allele missing is detected, this may also indicate a deletion. Deletions might account for a few homozygous tracts but we presume that none of the long tracts we have examined here reflect cryptic deletion.

Particularly long haplotypes that are common in a population may be evidence of a region that has undergone selection. The HapMap Consortium (11) found evidence for selective sweeps in 19 genomic regions (their supplementary Table 4). The top-scoring result from the Consortium screen was a selective sweep in the CEU population for a region between 136.75 and 137.25 Mb of chromosome 2 around the lactase (LCT) gene. This gene is known to influence ability to digest lactose in European-derived populations. Bersaglieri *et al.* (16) found that two alleles that are associated with lactase persistence mark a long common haplotype found at high frequency in European populations. The reduced haplotypic diversity associated with regions that have undergone a selective sweep will be expected to be reflected in a greater

proportion of individuals with homozygous tracts in the population undergoing selection. For the 500 kb LCT region (11), we find that 16 individuals (26.7%) in the CEU sample have homozygous tracts compared with fewer individuals: 8 (17.8%), 3 (6.8%) and 1 (1.7%) in the HCB, JPT and YRI samples, respectively. This is consistent with the evidence for a selective sweep in this region as presented by HapMap, although, of course, our evidence is not independent as it uses the same data. Bersaglieri *et al*. (16) noted long stretches of homozygosity around the haplotype containing the lactase-persistence alleles that did not appear to simply reflect a locally low recombination rate. Selection may contribute to the homozygous tracts characterized here but low levels of recombination are likely to have had a greater influence for the majority of tracts.

The HapMap data undergo extensive QC procedures but genotyping or reporting errors are still possible. Errors in the original Phase I data were corrected in the 16b, 16c and 16c.1 versions. We did not repeat this analysis with those data, however we did confirm that the longest tract detected (17.9 Mb) was still present and was not the result of a problem with the data. After this analysis was completed, the HapMap consortium released data for Phase II of the project with over three times as many SNPs in total as the Phase I data, including 3 902 623 genotypes that passed QC for the JPT sample. We traced the 17.9 Mb homozygous tract in individual NA18992 to see if it remained intact. We found that the same region which had 3922 SNPs in Phase I had 12 778 SNPs in Phase II. We found 11 heterozygotes breaking the tract into 12 pieces, the largest of which was 5.619 Mb. No two heterozygotes were adjacent. One marker of the 11 was present in the Phase I data as 'CC' but in Phase II is recorded as 'CT'. The presence of only 11 heterozygotes in a contiguous tract of 12 778 otherwise homozygous SNPs spanning 17.9 Mb is interesting. We suggest that these 11 comprise typing errors and/or relatively recent mutations. We assume therefore that the much higher density genotyping in Phase II will break some of the very long tracts but the strong relationships to the LD structure and evidence for inbreeding will be preserved. However, it is likely that other long homozygous tracts are interrupted by only small numbers of heterozygotes which are either typing errors or more recent mutations. This supports our assertion that the degree of homozygosity we have characterized is conservative, and ancestrally derived homozygous tracts are likely to be more numerous and longer than we have described.

We conclude that homozygous tracts are common, in some cases very long and in a few cases are a signature of cryptic relatedness within an individual. Otherwise, long homozygous tracts reflect the presence of long ancestral haplotypes that remain intact because of locally low rates of recombination or, more rarely, other mechanisms such as UPD, deletions and selection. It is conceivable that the abundance of homozygous regions and their contribution to long regions of high LD will significantly reduce our ability to fine map disease genes using association. However, further analysis and perhaps larger population samples will be required to determine the potential impact on genome-wide association studies.

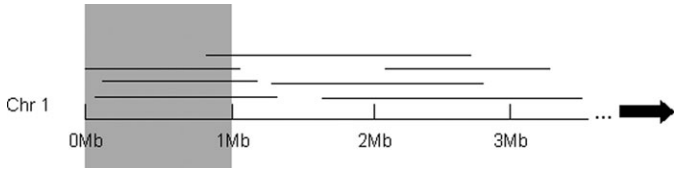

Figure 3. The method for computing the mean tract coverage of each megabase, for each population sample beginning on chromosome 1. The grey box is the first 1 Mb segment of the chromosome and each horizontal line represents a long homozygous tract. The position of tracts in four individuals is shown and all four have tracts covering or overlapping the first 1 Mb segment. The total lengths of the parts of tracts overlapping or containing the 1 Mb segment are summed (in kilobases) and divided by the number of individuals in the population sample to give the mean tract coverage of the megabase segment for that population. This continues along chromosome 1 and for all the other autosomes.

MATERIALS AND METHODS

The data sample

We examined the genome-wide SNP genotypic data from the 209 unrelated individuals from the four HapMap populations. These were the parents of the trios from the CEU and YRI populations and the unrelated individuals from the JPT and CHB populations. The HapMap release 16 (Phase I) data were used (non-redundant files that have passed quality controls downloaded from http://www.hapmap.org) and include 3 970 277 genotypes across the 22 autosomes for all four populations. These data were further filtered to remove genotypes with significant deviation from Hardy–Weinberg ($\chi^2 > 10$) and SNPs with minor allele frequencies below 0.05. The total number of individuals and genotypes from each population comprised 60 individuals and 728 353 genotypes from the CEU sample, 60 individuals and 744 006 genotypes from the YRI sample, 45 individuals and 644 060 genotypes from the CHB sample and 44 individuals and 639 460 genotypes from the JPT sample.

LDU map construction

Genome-wide LDU maps (4,6) were constructed from these data. LDU maps are constructed from multiple pairwise association data using a model which describes the decline in association, ρ, with distance $\rho = (1 - L)Me^{-\varepsilon d} + L$. The three parameters in the model are: M, the association at zero distance reflecting association at the last major bottleneck, L, the association at large distance reflecting 'background' LD and ε which measures the exponential decline of LD with distance. The LDU distance is calculated as $\varepsilon_i d_i$ for each interval i of d kilobases between a pair of SNPs and LDU locations are computed by summation over intervals. The LDMAP program used is available at http://cedar.genetics.soton.ac.uk/public_html/.

Characterization of homozygous tracts

An extended homozygous tract was defined as an uninterrupted sequence of homozygous SNPs spanning at least 1 Mb in a single individual. SNPs with one or two missing alleles were ignored. The average (kb) SNP density across all populations is 0.23, i.e. approximately 1 SNP every 5 kb.

As a locally low SNP density may artificially extend a homozygous tract, tracts with an average SNP density of less than 0.2 were excluded. Also omitted were the centromeric regions and acrocentric p-arms for the same reason.

Tract coverage

To calculate 'tract coverage', the 22 autosomes were split into 1 Mb segments. The shorter segments remaining at the ends of the chromosomes were included in the last segment. For each population sample, the mean tract coverage of each megabase was computed by summation of the length (kb) of all homozygous tracts overlapping the 1 Mb segment. This value was then divided by the number of individuals in that population sample (Fig. 3). The LDU/Mb ratio was also calculated for each segment using the LDU map for that population.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Biotechnology and Biological Sciences Research Council (UK).

Conflict of Interest statement. None declared.

REFERENCES

1. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
2. Broman, K.W. and Weber, J.L. (1999) Long homozygous chromosomal segments in reference families from the Centre D'etude Du Polymorphisme Humain. *Am. J. Hum. Genet.*, **65**, 1493–1500.
3. Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.
4. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X. and Morton, N.E. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.
5. De La Vega, F.M., Isaac, H., Collins, A., Scafe, C.R., Halldorsson, B.V., Su, X., Lippert, R.A., Wang, Y., Laig-Webster, M., Koehler, R.T. *et al*. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res.*, **15**, 454–462.
6. Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S. and Morton, N.E. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA*, **102**, 11835–11839.
7. Zhang, W., Collins, A., Maniatis, N., Tapper, W. and Morton, N.E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl Acad. Sci. USA*, **99**, 17004–17007.
8. Zhang, W., Collins, A., Gibson, J., Tapper, W.J., Hunt, S., Deloukas, P., Bentley, D.R. and Morton, N.E. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl Acad. Sci. USA*, **101**, 18075–18080.
9. Kong, X., Murphy, K., Raj, T., He, C., White, P.S. and Matise, T.C. (2004) A combined linkage-physical map of the human genome. *Am. J. Hum. Genet.*, **75**, 1143–1148.
10. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
11. Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. and Donnelly, P. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
12. Field, L.L., Tobias, R., Robinson, W.P., Paisey, R. and Bain, S. (1998) Maternal uniparental disomy of chromosome 1 with no apparent phenotypic effects. *Am. J. Hum. Genet.*, **63**, 1216–1220.
13. Bruce, S., Leinonen, R., Lindgren, C.M., Kivinen, K., Hlman-Wright, K., Lipsanen-Nyman, M., Hannula-Jouppi, K. and Kere, J. (2005) Global analysis of uniparental disomy using high density genotyping arrays. *J. Med. Genet.*, **42**, 847–851.
14. Altug-Teber, O., Dufke, A., Poths, S., Mau-Holzmann, U.A., Bastepe, M., Colleaux, L., Cormier-Daire, V., Eggermann, T., Gillessen-Kaesbach, G., Bonin, M. *et al*. (2005) A rapid microarray based whole genome analysis for detection of uniparental disomy. *Hum. Mutat.*, **26**, 153–159.
15. Huie, M.L., Nyane-Yeboa, K., Guzman, E. and Hirschhorn, R. (2002) Homozygosity for multiple contiguous single-nucleotide polymorphisms as an indicator of large heterozygous deletions: identification of a novel heterozygous 8-kb intragenic deletion (IVS7-19 to IVS15-17) in a patient with glycogen storage disease type II. *Am. J. Hum. Genet.*, **70**, 1054–1057.
16. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, **74**, 1111–1120.