Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Increased gene coverage and Alu frequency in large linkage disequilibrium blocks of the human genome

Y. Wang[1], P.Y. Fong[1], F.C.C. Leung[2], W. Mak[1] and P.C. Sham[1,3]

[1]Genome Research Centre and [2]Department of Zoology,
University of Hong Kong, Pokfulam, Hong Kong SAR, China
[3]Institute of Psychiatry, Denmark Hill, London, UK
Corresponding author: Y. Wang
E-mail: wangyong@hkucc.hku.hk

Genet. Mol. Res. 6 (4): 1131-1141 (2007)
Received July 16, 2007
Accepted November 16, 2007
Published December 11, 2007

ABSTRACT. The human genome has linkage disequilibrium (LD) blocks, within which single-nucleotide polymorphisms show strong association with each other. We examined data from the International HapMap Project to define LD blocks and to detect DNA sequence features inside of them. We used permutation tests to determine the empirical significance of the association of LD blocks with genes and Alu repeats. Very large LD blocks (>200 kb) have significantly higher gene coverage and Alu frequency than the outcome obtained from permutation-based simulation, whereas there was no significant positive correlation between gene density and block size. We also observed a reduced frequency of Alu repeats at the gaps between large LD blocks, indicating that their enrichment in large LD blocks does not introduce recombination hotspots that would cause these gaps.

Key words: Linkage disequilibrium blocks, Alu, Gene density, Gene coverage, Recombination

INTRODUCTION

It is now widely accepted that the linkage disequilibrium (LD) structure of the human genome is shaped by recombination; gaps between LD blocks are hotspots for recombination (Reich et al., 2002; Kauppi et al., 2003; Twells et al., 2003; Wall and Pritchard, 2003). The blocks are defined to represent a genomic region in which high LD is maintained. The gaps between two blocks are basically where the site breakage of high LD occurs. LD structure is also affected by recombination, mutation and population history (Phillips et al., 2003), as well as by genomic features such as the distribution and size of coding sequences, regulatory motifs, and repetitive elements (Dawson et al., 2002; Hinds et al., 2005; Smith et al., 2005). Actually, there is no mutual conflict among these actors, and the current LD structure seems to be a result of co-effects of all the related actors. In the human genome, recombination occurs more frequently in gene-poor than in gene-rich regions (McVean et al., 2003). This is apparent because recombination hotspots in genes must be restricted in order to avoid disruption of exons. The hotspots are also not freely located inside other genic regions due to the presence of ultra-conserved functional elements outside of exons (Dermitzakis et al., 2002, 2004). Nonetheless, hotspots are present in large introns as demonstrated by some studies on long genes (Kauppi et al., 2003; Twells et al., 2003). Consequently, recombination hotspots should reside in gene-poor regions and within large introns. Extension of genic regions is, therefore, hypothetically associated with extension of high LD, giving high gene density and high gene coverage in large LD blocks. In previous reports, such a conclusion was drawn on the basis of statistical analysis of gene count or coding region coverage (Hinds et al., 2005; Smith et al., 2005). However, the genomic-biased distribution of long genes and short genes may have introduced misleading results. For instance, a high LD region containing many long genes would lower the correlation between gene density and high LD. A permutation test in this study could eliminate the difference in gene length among genomic regions by first shuffling information on gene density and gene coverage among genomic regions and then performing statistical analyses. Gene coverage differs from gene density by measuring coverage percentage of transcribed regions of protein-coding genes in a genomic area.

We are also interested in Alu enrichment at the gaps between LD blocks. Alu repeats, covering 10.3% of the human genome, are more frequent in both gene-rich and high LD regions (International Human Sequencing Consortium, 2001; Smith et al., 2005). There is a conflict here since Alu repeats in gene-rich regions frequently induce homologous recombination and occasionally lead to breakage of LD structure. Therefore, gene-rich regions cannot co-exist with high LD regions. We also aimed to solve the conflict by detecting Alu frequency in the gaps between large LD blocks.

Availability of detailed information on human LD structure from The International HapMap Consortium (2005) provides an opportunity to examine the relationship between LD structure and sequence features. We examined the relationship of LD blocks to genic regions and Alu repeats in the human genome, using data from the HapMap Project.

MATERIAL AND METHODS

Defining breakage point of high linkage disequilibrium extension

We downloaded single-nucleotide polymorphism (SNP) genotype data, release 16c.1, from the HapMap website (http://www.hapmap.org). The release was based on version 34 of the human genome and dbSNP b124. The four populations are Yoruba in Ibadan (YRI), Japanese in Tokyo (JPT), Han Chinese in Beijing (CHB), and CEPH (Utah residents with ancestry from northern and western Europe). We combined JPT and CHB into an Asian dataset in this study. Gabriel's algorithm, implemented in the HaploView program, was used to define LD blocks (Gabriel et al., 2002; Barrett et al., 2005). The setting of 95% for the confidence interval of D' was applied to define strong LD between SNPs. The output file gives the lists of SNPs which were grouped into blocks. The start and end SNPs for all of the LD blocks were extracted; the positions of the start and end SNPs were then used as the boundaries of the LD blocks. The site where a large LD block (>50 kb) stops is considered as a starting point of the gap between LD blocks.

Extracting sequences at the breakage site

The gap sequences between large LD blocks were examined by two approaches. Due to inadequate SNP coverage, some of the apparent gaps could be shorter than their true lengths because the start and end SNPs of an LD block may not accurately represent the true boundaries of the LD block. To reduce the deviation from the real boundaries, we only selected small gaps with sizes of <2 kb between two large LD blocks (>50 kb). In the other approach, a gap was also defined as a <2-kb sequence between the boundary SNP of a large (>50 kb) LD block and its nearest inter-block SNP. A position dataset was generated for the inter-block SNPs. Inter-block gaps of <2 kb next to large LD blocks (>50 kb) were identified. The sequences were extracted from the human genome from the positions of the gaps. We extracted 10 neighboring sequences flanking each gap, five on each side, of the same size as the gap. RepeatMasker (http://RepeatMasker.org) was used to search Alu repeats in the gaps and the flanking sequences.

Genes and Alu repeats in contigs of human autosomes

Gene and Alu locations (version 34) were obtained from the UCSC browser (http://genome.ucsc.edu/). Genic regions are demonstrated here as transcribed regions of protein-coding genes. We defined gene coverage as the proportion of the sequence of an LD block covered by genes. The counts of Alu repeats in the LD blocks were then converted to Alu frequency (count per kb LD blocks). In the test on genic regions, Alu repeats were counted if the beginning site of the repeats was within the genes. The LD blocks were classified into six groups according to their size: 1-10, 10-20, 20-50, 50-100, 100-200, and >200 kb.

Permutation-based simulation

To further test the association between LD blocks and genes, we created simulated datasets by permutation. The association between LD blocks and Alus was tested in the same approach. In order to preclude the influence of sequencing gaps and centromeres in the human genome, we did analyses on long contigs >10 Mb of all autosomes. They were in at least version 5 of the NCBI (http://www.ncbi.nlm.nih.gov/) in order to assure high annotation quality. The contigs were divided or truncated, and packed into a dataset that finally consisted of 203 sequences each of 10 Mb. Moreover, to avoid double gene counting, the boundaries of the 10-Mb sequences were not allowed to be within genes. LD blocks (>1 kb), genes and

Alu repeats were located on the sequences. The locations of the blocks, the genes and the Alu repeats served as a real dataset. Gene coverage, gene count and Alu count in all the blocks (>1 kb) were measured. The gene and Alu locations in each 10-Mb sequence were converted to a set of relative locations by setting the start site of the 10-Mb sequences to zero. Taking these relative positions, the genes and Alus in these 10-Mb sequences were then randomly shuffled into other 10-Mb sequences. In the process, the genes and Alus in the same 10-Mb sequences moved together. In a new 10-Mb sequence, they were assigned with new locations in reference to the start point of these 10-Mb sequences. The LD blocks were fixed throughout the shuffling process and again gene coverage, gene count and Alu count were calculated for each 10-Mb region. The permutation was carried out 10,000 times. Gene coverage, gene count and Alu count were recorded in each permutation cycle. In a given group of LD blocks, the average values for the gene coverage percents, gene counts and Alu counts were the results of the permutation tests. Empirical P values were calculated as a frequency of the permutation replicates for which $| Ni - \overline{N} | > | Nr - \overline{N} |$ (i = 1,2,3......10,000), where N denotes gene coverage, gene count or Alu count, r labels the actual value in the real data, and $\overline{N}$ is the average value from the 10,000 permutations.

RESULTS AND DISCUSSION

Very large linkage disequilibrium blocks show high gene coverage

We divided the human genome into LD blocks of various sizes. The blocks were classified into different size groups and their distribution was determined for all populations (Table 1). Gene coverage was measured in each block; there was no relationship between gene coverage and length of LD block for block size below 100 kb. However, a tendency for association between very large LD blocks (>100 kb) and gene coverage was clear (Table 2). The YRI population showed increased gene coverage beginning at a smaller block size of 50 kb. The reason is likely the subdivision of the large LD blocks defined in CEPH and Asian populations into smaller ones in the YRI population (Gabriel et al., 2002). Overall, YRI had about 10% more LD blocks than CEPH and Asian populations, and fewer large LD blocks (Table 1).

Table 1. Component of linkage disequilibrium (LD) blocks with different sizes in three populations.

Size	YRI	CEPH	Asian
1-10 kb	57.2%	43.4%	45.0%
10-20 kb	20.0%	21.1%	20.7%
20-50 kb	17.2%	23.0%	22.3%
50-100 kb	4.1%	8.4%	8.1%
100-200 kb	1.2%	3.2%	3.0%
>200 kb	0.2%	0.9%	0.8%
Total	78,956	71,129	71,126

The result was based on the data release of 16c.1 (June 2005) from the HapMap. YRI: Yoruba in Ibadan; Asian: JPT (Japanese in Tokyo) + CHB (Han Chinese in Beijing), and CEPH: Utah residents with ancestry from northern and western Europe. Gabriel's algorithm embedded in the HaploView was used to define LD blocks (Barrett et al., 2005). LD blocks smaller than 1 kb were not taken into account.

Table 2. Gene count and gene coverage in linkage disequilibrium (LD) blocks in real and simulated datasets.

	Asian		CEPH		YRI	
	Gene count	Gene coverage	Gene count	Gene coverage	Gene count	Gene coverage
1-10 kb	0.849 (0.831)	0.395 (0.39)	0.831(0.83)	0.39 (0.39)	0.8 (0.83)	0.376 (0.389)
10-20 kb	0.833 (0.863)	0.389 (0.4)	0.829 (0.862)	0.385 (0.4)	0.779 (0.864)	0.364 (0.4)
20-50 kb	0.884 (0.968)	0.385 (0.415)	0.884 (0.968)	0.386 (0.415)	0.853 (0.961)	0.377 (0.414)
50-100 kb	1.136 (1.294)	0.402 (0.442)	1.138 (1.293)	0.407 (0.441)	1.331 (1.274)	0.462 (0.441)
100-200 kb	1.992 (2.007)	0.498 (0.476)	2.03 (1.996)	0.495 (0.475)	2.362 (1.957)	0.586 (0.473)**
>200 kb	4.678 (3.905)	0.636 (0.524)*	4.979 (3.939)	0.628 (0.524)*	5.79 (3.854)*	0.753 (0.524)**

The datasets were LD blocks in 10-Mb unbroken genomic regions of the human autosomes (version 34); 203 such regions were investigated. The simulated datasets were from permutation-based simulations of the real datasets. The table shows gene coverage and gene count in real datasets and simulated datasets. The LD blocks had been classified into six groups according to their sizes: 1-10, 10-20, 20-50, 50-100, 100-200, and >200 kb. The simulated datasets (in parentheses) are the average values of 10,000 permutation circles. Empirical P values, as indicated by asterisks, were denoted by the occurrence frequency of the cases of $|Ni - \overline{N}|>|Nr - \overline{N}|$ (i = 1,2,3......10,000), where N denotes gene coverage or gene count, r represents the true value and $\overline{N}$ is the average value from the permutation circles. *P < 0.05; **P < 0.01. The abbreviated population names are as shown in Table 1.

Briefly, the results in <50-kb blocks did not follow the trend for the larger blocks in a given population. The three groups of small- and medium-size LD blocks (1-50 kb) did not appear to differ in tests of both gene coverage and Alu frequency, whereas an upward trend was clearly observed in the groups of large LD blocks (>50 kb).

The explanation for the inconsistency in small LD blocks lies in the problematic definition of what is an LD block, particularly for those <20 kb. Different algorithms yield different patterns of LD blocks. The dynamic block structures at this scale make it difficult to interpret our results. We believe that it is nearly impossible to develop an algorithm that is capable of precisely defining small LD blocks on a 0- to 20-kb scale. This is because of the complex linkage relationships among SNPs in these regions; consequently such small-size LD blocks cannot be unambiguously defined. Potential small LD blocks could more likely overlap or cover others. That is why we preferred large LD blocks, which could be easily and reliably identified by any algorithms. Although the boundaries of large blocks could be blurred due to the interference of some small blocks, their large size was retained. Therefore, regardless of what algorithm was used for LD block identification, our results on large LD blocks can demonstrate an association between genes and extension of high LD. Hence we classified the LD blocks according to size. Our conclusions were drawn from the results on large LD blocks.

The discrepancy among the three populations that we found in our study is irrelevant to the algorithm. Many reports have shown that YRI LD blocks are smaller than those of the other two populations, because many YRI LD blocks are derived from splitting of the large LD blocks in CEPH and Asian populations into two or more (Gabriel et al., 2002). This accounts for nearly all the differences among the populations including gene coverage and Alu frequency. We found that the large LD blocks that were divided into two or more in YRI had low gene coverage in the CEPH and Asian populations. This has been suggested to explain the tendency for higher gene coverage in YRI >200-kb LD blocks (Table 2).

Permutation test confirmed the correlation between linkage disequilibrium block size and gene coverage

To eliminate any existing correlation between genes and LD blocks across the human genome, we made extensive permutation-based simulations to randomly shuffle the sequence data among 203 unbroken genomic contigs (10 Mb). Evidence of significant correlation was derived by comparing the summary statistics obtained from the real dataset with the data from simulated datasets. Gene count and gene coverage in real LD blocks and the average values in the corresponding blocks after 10,000 permutation cycles were calculated (Figure 1). The frequency of genes in real blocks was not clearly correlated with block size. The permutation test also showed no significant correlation. The gene count in the real dataset was not significantly different from those of the simulated datasets in all the populations. The only exception was LD blocks of >200 kb in the YRI population, which had a significant excess in gene count ($P < 0.05$), different from the finding by Smith et al. (2005). In contrast, gene coverage appears to correlate to block size in the blocks (Figure 1C). According to the statistics, the gene coverage in the largest LD blocks (>200 kb) was significantly ($P < 0.05$) greater in the real than in the simulated data (Table 2). This was true in 100- to 200-kb YRI LD blocks as well ($P < 0.01$). In sum, large LD blocks had a high percentage of gene coverage, but not a correspondingly high gene density, inferring an excess of long genes in large LD blocks. We found that genic regions occupied more than 60% of large LD blocks (Table 1). Diverse types of human genes yield a large variety of gene lengths and gene conservation degrees, making it difficult to calculate gene density in LD blocks. A previous report showed a positive correlation between gene density and LD pattern on chromosome 22 (Dawson et al., 2002). We found the significant positive correlation only on chromosome 3p and chromosome 15 (data not shown). However, direct correlation between gene density and LD block size is not found in every human chromosome. This is probably due to different proportions of long and short genes. Long genes in a chromosomal region will make gene density decline. On the other hand, the associated large LD blocks are expected to hold more genes. Taken together, the presence of long genes provokes a poor correlation between gene density and LD block size.

A relevant question is whether the gaps will frequently interrupt the coding regions of the genes? Our results on chromosome 3p showed that gaps of LD blocks present mostly restrictively to long genes (>20 kb), and only rarely cases were observed in short genes (<10 kb). As expected, LD blocks basically did not interrupt the short genes, among which most housekeeping genes are included. The genes of LRP5 and MHC in previous similar reports were long genes as well (Kauppi et al., 2003; Twells et al., 2003). Thus, it seems that only long genes can accept the gaps between LD blocks due to their longer introns and larger intragenic spaces (Vinogradov, 2004). In addition, some genes, for instance those acting involved in the immune response and in sensory perception, seem to favor the presence of the gaps (Smith et al., 2005). To understand the influence of gaps on the function of the long genes is a long-term objective. We interpret the positive correlation between gene coverage and LD block size as a result of the presence of excessive long genes and sometimes because of high gene density in large LD blocks.

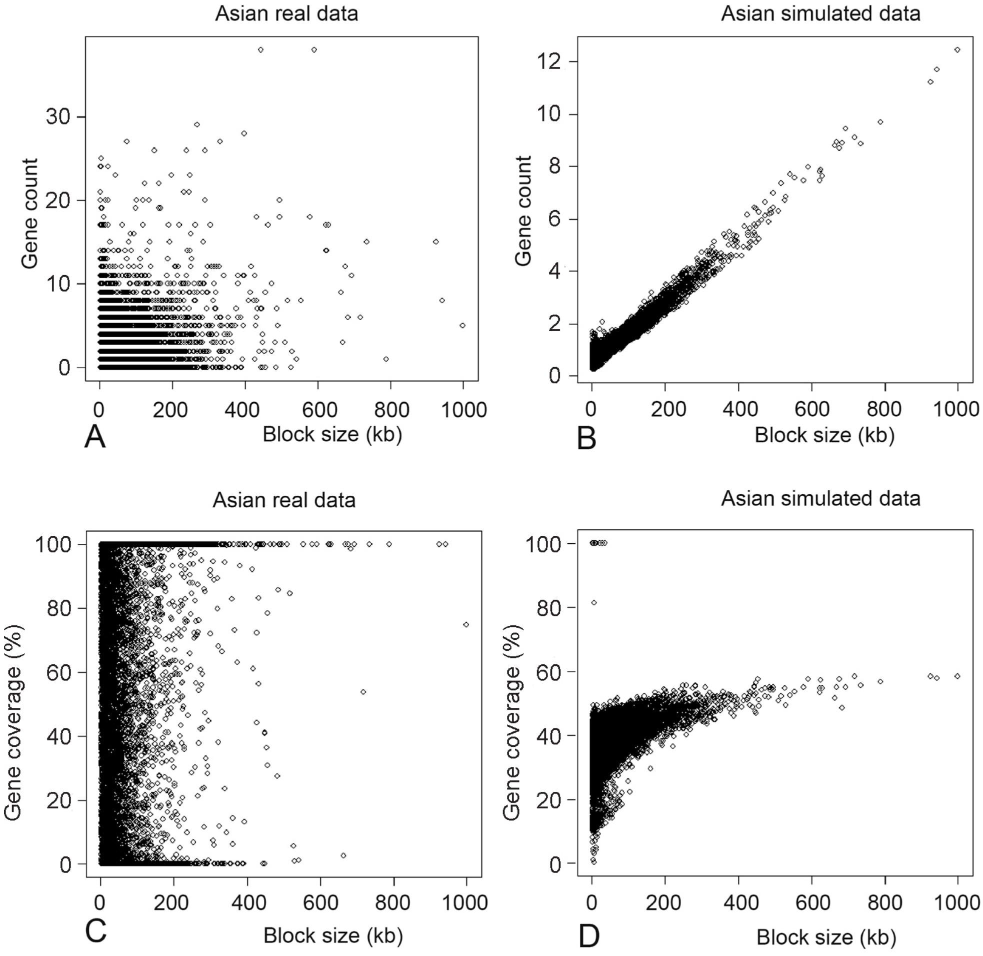

Figure 1. Gene count and gene coverage in linkage disequilibrium (LD) blocks from a real and a simulated dataset. The results are from the Asian population. **A, B.** Gene count in real and simulated datasets. **C, D.** Gene coverage. The simulation is permutation-based (see Methods). The dots in Figure 1B and D represent the average values in individual LD blocks after 10,000 permutation cycles.

Alu frequency in large linkage disequilibrium blocks and gaps

Co-occurrence of recombination hotspots and LD block gaps (Reich et al., 2002; Kauppi et al., 2003) was confirmed by a recent report demonstrating a negative correlation between LD block size and recombination rate (Greenwood et al., 2004). We attempted to link the high Alu frequency with the recombination force creating the LD block gaps. Similar tests have been performed on Alu repeats. The relationship between LD block length and Alu repeat frequency, measured as number of repeats per kb (Figure 2), was rather similar to the relationship between LD block size and gene coverage. For instance, there was no relationship among small to modest length haplotypes, whereas there was a significant positive correlation ($P < 0.05$) for very large LD blocks (>100 kb). The significant correlation among large blocks was also demonstrated by permutation testing. Only extremely large blocks (>200 kb) and 100- to 200-kb YRI LD blocks gave the significant correlations (Table 3).

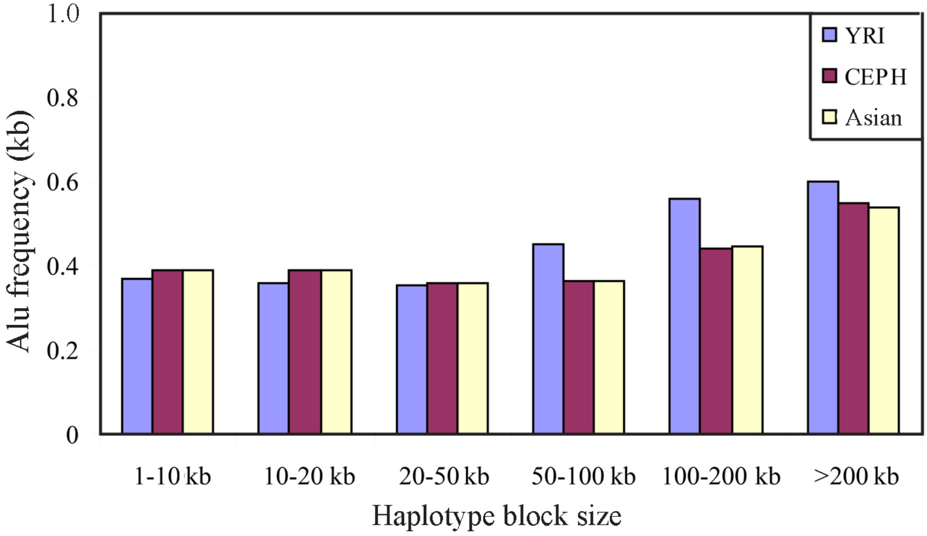

Figure 2. Alu frequency in linkage disequilibrium (LD) blocks. The LD blocks from the three populations were classified on the basis of their sizes. The large standard errors are not shown.

Table 3. Comparison of Alu repeats in linkage disequilibrium (LD) blocks in real and simulated datasets.

	Asian	CEPH	YRI
1-10 kb	1.9 (2.0)	1.9 (2.0)	1.7 (1.9)
10-20 kb	5.5 (5.8)	5.5 (5.9)	4.9 (5.8)
20-50 kb	11 (12.6)	10.9 (12.7)	10.4 (12.4)
50-100 kb	24 (27.7)	24.1 (27.9)	28.8 (27.2)
100-200 kb	58.8 (54.3)	57.7 (54.6)	73.4 (53.8)**
>200 kb	157.9 (119.1)*	165.3 (120.7)*	179.1 (117.7)**

The datasets and abbreviated population names are described in Tables 1 and 2. The table shows the average numbers of Alu repeats in LD blocks classified according to their sizes. The numbers in parentheses are from the simulated datasets. *P < 0.05; **P < 0.01.

The gaps in the LD blocks were studied to clarify the role of Alu repeats in segregating LD blocks. The gaps between pairs of long LD blocks (>50 kb) were extracted and examined for Alu repeats. The Alu frequency was reduced on average by 0.15 per kb in the gaps between long LD blocks, as compared to the flanking regions (Figure 3A). The difference between the gaps and their matched flanking regions was significant in all populations ($P < 0.001$; Wilcoxon test). The Alu frequencies were also slightly reduced at the two closest flanking regions of the gaps, possibly because the boundary SNPs only approximately stand for the actual boundaries of LD blocks. A similar pattern of results was obtained from examining the gaps located in the genic regions (Figure 3B). The gaps and flanking regions had the same statistical significance level in Alu frequency difference ($P < 0.001$; Wilcoxon test). The Alu repeats in LD blocks were not more frequently observed at the gaps, indicating that they were not involved in the recombination hotspots at the gaps.

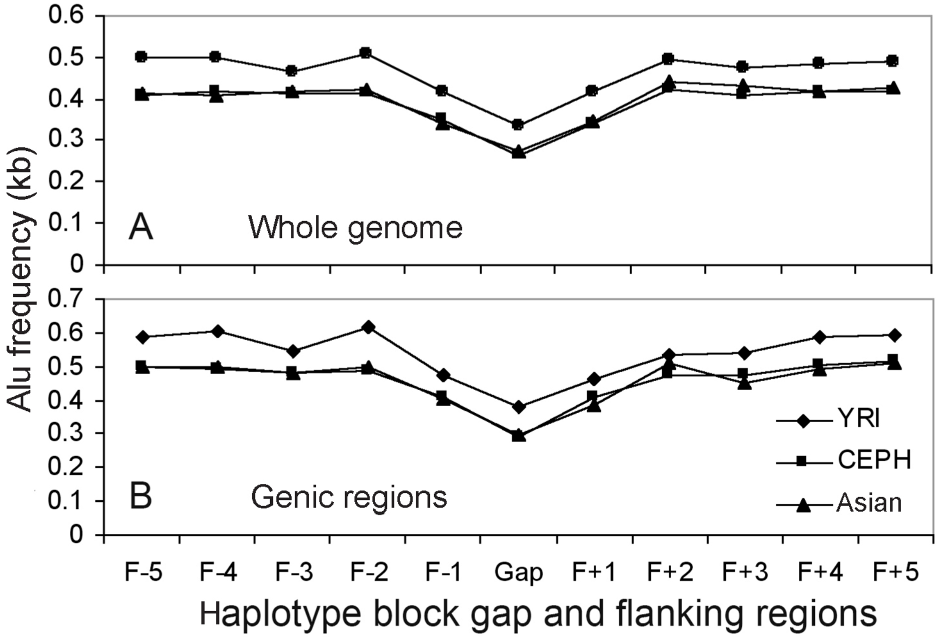

Figure 3. Reduced Alu frequency in linkage disequilibrium (LD) block gaps. The gaps were <2-kb regions outside of LD blocks in size of >50 kb (see Methods). The F ± n (n = 1-5) denotes flanking sequences of the same size to the gap. **A.** Alu numbers were cumulated from all the gaps and their flanking regions and divided by the total length to gain Alu frequency in every region. **B.** Results of the gaps and flanking regions located in genic regions.

The Alu repeat is one of the most widespread repeats in primates. Numerous human SNPs have been located in Alu repeats (Ng and Xue, 2006). The highly similar ~300-bp repeats mediate numerous homologous recombination events, resulting in gene duplication, genome reshuffling and genome expansion (Harteveld et al., 1997; Deininger and Batzer, 1999; Babcock et al., 2003; Bailey et al., 2003; Liu et al., 2003). All of these have been found to give rise to a variety of genetic disorders. Unequal homologous recombination between Alu repeats is responsible for 0.3% of human genetic diseases (Deininger and Batzer, 1999). Alu repeats spread in the human genome through three models (Bailey et al., 2003). The models for Alu duplication suggest that an Alu-inserted region was prone to be Alu-rich. Core consensus sequences in Alu repeats are critical to Alu-mediated recombination. Homologous recombination is, therefore, more likely to occur between Alu-rich regions. We found that the gaps of large LD blocks were relatively Alu-poor regions, and thus were not Alu-related recombination hotspots. The flanking regions of the gaps had even more Alu repeats. We, thus, conclude that the correlation between Alu frequency and LD block size is a result of the close association between Alu repeats and genes, and that recombination hotspots at the gaps are not mainly attributable to Alu repeats due to their low frequency at the gaps.

In humans, Alu repeats are reported to be able to serve as the intronic elements involved in alternative splicing of genes (Sorek et al., 2002; Kreahling and Graveley, 2004), by providing a site for intron splicing at the mRNA level (Lev-Maor et al., 2003). Additionally, Alu repeats also have numerous sites for the regulatory function of some hormone genes (Vansant and

Reynolds, 1995; Babich et al., 1999). With cumulated mutations and deletions under positive selection, early Alu families were deprived of mobility. These old and fixed Alu repeats are functional and somewhat different from the newly inserted Alu repeats, which are capable of performing free recombination with other Alus before subsequently being fixed (Shankar et al., 2004). Thus, divergence time from their ancestral consensus is probably essential to recombination activity of Alu repeats. This is an apparent answer to the low recombination potentials of Alu repeats. The low Alu frequency in LD block gaps means that some motifs other than Alu repeats fill in the space. What the motifs are and how they facilitate the formation of the gaps need to be studied.

CONCLUDING REMARKS

Up to now, the hypothesis of recombination hotspots has been widely accepted as an explanation to LD pattern and LD block. Recombination hotspots are hypothetically caused by brokage of double-strand DNA at weak sites, by gene conversion and by homologous recombination (Gerton et al., 2000; Reich et al., 2002; Kauppi et al., 2003; Matthew et al., 2004; Shankar et al., 2004). If the fundamental mechanism involves in physical properties of the DNA molecule, the genomic regions with fewer recombination hotspots therefore would serve as scaffolds for implantation of the functional elements. They dominate the distributions of the functional units such as genes. Hence, analogous to large muscles attaching specially to long bones, long LD blocks can harbor more and longer genes. Nevertheless, the model seems weak to cope with the exceptions. Quite a few genes still have LD block gaps (Smith et al., 2005).

On the other hand, if the recombination hotspots are mainly stimulated by gene conversion and homologous recombination, our conclusion in this study focusing on distribution of long genes will provoke an argument on which factor makes the largest contribution to LD blocks, the profile of genic regions or the recombination hotspots. Given that LD is stemmed from site linkage in genic regions, long genes are naturally associated with large LD blocks. Gene density and gene size could be more critical in determining LD blocks. They may affect LD blocks directly, as well as by virtue of affecting local recombination rates, inasmuch as gene-rich regions and long genic regions cannot have many recombination hotspots. Probably, studying the factors relating to gene density and gene size is also an alternative promising approach for exploring the mechanisms generating LD blocks.

Currentlly, reports have shown reciprocal influence between recombination and genes: 1) gene conversion can contribute much to the formation of recombination hotspots (Matthew et al., 2004), and 2) gene order is related to recombination rate in yeasts (Pal and Hurst, 2003). Therefore, at this level, the two explanations can be somewhat mutually consistent. We will study gaps between large LD blocks and conservative non-coding regions to learn more about how LD blocks form in the human genome, in an effort to cast light on the primary mechanisms of many genetic diseases.

ACKNOWLEDGMENTS

Hong Kong Research Grants Council Competitive Earmarked Research Grant HKU 7669/06M.

REFERENCES

Babcock M, Pavlicek A, Spiteri E, Kashork CD, et al. (2003). Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res.* 13: 2519-2532.

Babich V, Aksenov N, Alexeenko V, Oei SL, et al. (1999). Association of some potential hormone response elements in human genes with the Alu family repeats. *Gene* 239: 341-349.

Bailey JA, Liu G and Eichler EE (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73: 823-834.

Barrett JC, Fry B, Maller J and Daly MJ (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, et al. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418: 544-548.

Deininger PL and Batzer MA (1999). Alu repeats and human disease. *Mol. Genet. Metab.* 67: 183-193.

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, et al. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420: 578-582.

Dermitzakis ET, Kirkness E, Schwarz S, Birney E, et al. (2004). Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* 14: 852-859.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.

Gerton JL, DeRisi J, Shroff R, Lichten M, et al. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 97: 11383-11390.

Greenwood TA, Rana BK and Schork NJ (2004). Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res.* 14: 1358-1361.

Harteveld KL, Losekoot M, Fodde R, Giordano PC, et al. (1997). The involvement of Alu repeats in recombination events at the alpha-globin gene cluster: characterization of two alphazero-thalassaemia deletion breakpoints. *Hum. Genet.* 99: 528-534.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, et al. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science* 307: 1072-1079.

International Human Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Kauppi L, Sajantila A and Jeffreys AJ (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* 12: 33-40.

Kreahling J and Graveley BR (2004). The origins and implications of Aluternative splicing. *Trends Genet.* 20: 1-4.

Lev-Maor G, Sorek R, Shomron N and Ast G (2003). The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 300: 1288-1291.

Liu G, Zhao S, Bailey JA, Sahinalp SC, et al. (2003). Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* 13: 358-368.

Matthew EH, Willey D, Matthews L and Syed SH (2004). Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biol.* 5: 55r.

McVean GA, Myers SR, Hunt S, Deloukas P, et al. (2003). The finescale structure of recombination rate variation in the human genome. *Science* 304: 581-584.

Ng SK and Xue H (2006). Alu-associated enhancement of single nucleotide polymorphisms in the human genome. *Gene* 368: 110-116.

Pal C and Hurst LD (2003). Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* 33: 392-395.

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, et al. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* 33: 382-387.

Reich DE, Schaffner SF, Daly MJ, McVean G, et al. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* 32: 135-142.

Shankar R, Grover D, Brahmachari SK and Mukerji M (2004). Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol. Biol.* 4: 37.

Smith AV, Thomas DJ, Munro HM and Abecasis GR (2005). Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* 15: 1519-1534.

Sorek R, Ast G and Graur D (2002). Alu-containing exons are alternatively spliced. *Genome Res.* 12: 1060-1067.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* 437: 1299-1320.

Twells RC, Mein CA, Phillips MS, Hess JF, et al. (2003). Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. *Genome Res.* 13: 845-855.

Vansant G and Reynolds WF (1995). The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proc. Natl. Acad. Sci. USA* 92: 8229-8233.

Vinogradov AE (2004). Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20: 248-253.

Wall JD and Pritchard JK (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* 73: 502-515.