Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome

Tatsuhiko Tsunoda[1], G. Mark Lathrop[2], Akihiro Sekine[3], Ryo Yamada[4], Atsushi Takahashi[1], Yozo Ohnishi[5], Toshihiro Tanaka[6] and Yusuke Nakamura[7,8*]

Laboratories for Medical Informatics[1], Genotyping[3], Rheumatic Diseases[4], SNP Analysis[5], Cardiovascular Diseases[6], Pharmacogenetics[7], SNP Research Center, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan

[2]Centre National de Genotypage, 2, rue Gaston Crémieux, CP 5721, 91057 Evry Cedex, France

[8]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

*To whom correspondence should be addressed.

Tel: +81-35449-5372

Fax: +81-35449-5433

Email: yusuke@ims.u-tokyo.ac.jp

ABSTRACT

A principal goal in human genetics is to provide the tools necessary to enable genome-wide association studies. Extensive information on the distribution of gene-based single-nucleotide polymorphisms (SNPs) and linkage disequilibrium (LD) patterns across the genome is required in order to choose markers for efficient implementation of this approach. To obtain such information, we have genotyped a large Japanese cohort for SNPs identified by systematic resequencing of more than 14,000 autosomal genes. Analysis of these data led to the conclusion that the Japanese population contains approximately 130,000 common autosomal gene haplotypes (frequency > 0.05), of which more than 35% are identified in the present study. We also examined allele frequencies and LD patterns according to the position of variants within genes, and their distribution across the genome. We found lower allele variability at exonic SNP sites (both non-synonymous and synonymous) compared to non-exonic SNP sites, and greater average LD between SNPs within exons of the same gene compared to other SNP combinations, both of which could be signals of selection. LD was correlated with the recombination rate per physical distance as estimated from the meiotic map, but the strength of the relationship varied considerably in different regions of the genome. Unique LD patterns, characterised by frequent instances of high LD between non-adjacent SNPs punctuated by blocks of low LD, were found in a 7 Mb region on chromosome 6p that includes the MHC (Major Histocompatibility Complex) locus and many non-MHC genes. These results demonstrate the complexity that must be taken into account when considering SNP variability and LD patterns, while also providing tools necessary for implementation of efficient genome-wide association studies.

INTRODUCTION

Historically, disease association studies have been limited by a number of factors, including lack of knowledge about the appropriate candidate genes to investigate and insufficient information on sequence variation to assure full assessment of genes (1-4). In some recent studies, the increasing knowledge of human genome and the availability of high-throughput genotyping at relatively low cost have been exploited to allow disease association investigations of single-nucleotide polymorphisms (SNPs) in many thousand genes simultaneously (5-8). In the near future, it will be possible to extend such studies to the whole genome. In principle, this strategy circumvents one of the limits of candidate gene investigations because it does not rely on *a priori* biological hypothesis about disease-gene relationships. Efficient implement of this approach will require detailed information on SNP variants and linkage disequilibrium (LD) patterns across the whole genome. As a prelude to obtaining whole genome data, SNPs have been studied at relatively high density in specific regions, and for a limited number of complete chromosomes (9-15).

A number of important observations and hypotheses are emerging from these studies. Genetic variation is non-randomly distributed within the human genome, with a higher frequency of polymorphisms observed in non-coding vs. coding regions; non-synonymous SNPs within coding regions are also reported to be less variable than other SNPs (16-19). It has been proposed that the human genome exhibits a haplotype block structure characterized by segments of high LD punctuated by regions of low LD. With knowledge of the block structure, markers that identified the major haplotypes can be selected for genotyping to obtain the essential information

for association studies. It has been hypothesized that the boundaries of haplotype blocks may be defined by recombination hotspots (3, 11, 14, 20), which could render haplotype blocks independent of the population under study and simplify studies in different ethnic groups. Since LD is negatively correlated with the frequency of recombination per physical distance (13, 15), the extent of block structure may be variable across the genome. Such issues have important implications for the use of LD as a genetic mapping tool (1, 3, 21), and they underscore the need for further investigations of LD patterns across the genome, such as that recently initiated in the international haplotype map program (22).

We have recently embarked on a national program of SNP discovery in Japan to systematically identify common gene variants as a tool to implement genome-wide gene-based association studies (18). To date, the database covers more than 15,000 genes and predicted exons (http://snp.ims.u-tokyo.ac.jp). In order to provide a comprehensive view of allele frequencies and LD patterns for gene-based variants across the human genome, we report here the analysis of more than 74,000 SNPs selected from this database and genotyped in a large cohort of Japanese volunteers. This provides an important contribution to enabling association studies by identifying more than 35% of the principal gene-based haplotypes in the Japanese population. The analysis also reveals complex patterns of allele variation and LD related to the position of SNPs within genes and within the genome.

RESULTS

Allele frequency distributions for gene-based SNPs

Sequencing of known genes and predicted exons in 24 Japanese control individuals led to the identification of more than 195,000 gene-based SNPs (18) (http://snp.ims.u-tokyo.ac.jp). The sequencing covers 17 Mb in coding regions, 91.4 Mb in intronic regions, 13.2 Mb in promoters, 14.8 Mb in 5' UTR, 3' UTR and 3' flanking regions. We genotyped a random selection of 80,788 autosomal SNPs from these in up to 564 additional, unrelated Japanese volunteers using the Invader assay and multiplex amplification as described (23). After the application of validation criteria, genotypes from 74,842 SNPs were retained for inclusion in the present study (see Methods). The markers are distributed across 14,271 genes, providing an average of 5 SNPs per gene. The SNP map spans a total of 680 Mb, with a median distance of <4 kb between neighbouring markers, and covers all regions of the autosomal genome that contain identified genes (Figure 1). There are 13,119 exonic SNPs and 61,723 non-exonic SNPs (e.g. within introns, 5' flanking region or 3' flanking region of the gene) included in the dataset. Previous analysis of a subset of the markers and DNA samples studied here showed no evidence of population heterogeneity based on the geographic origin within Japan (24).

We previously found variations to be less frequent in exonic regions compared to the other intragenic regions that were screened for SNPs (1 SNP / 1,298 bp in coding sequence; 1 SNP / 847 bp in introns; 1 SNP / 861 bp in promoter regions; 1 SNP / 969 bp in 3' flanking regions) (18). From the present data, we can conclude that different classes of variants also have different allele frequency spectra (Figure 2A). At variant sites within exons, the minor alleles were less frequent on average

than those at other sites, particularly for SNPs that introduced amino acid changes: 19.6% of non-synonymous exonic SNPs had a minor allele frequency <0.05 compared to 14.4% of synonymous exonic SNPs and 12.0% of non-exonic SNPs. These differences were significant (see Methods for statistical tests), not only for the comparison between non-synonymous and synonymous exonic variants ($P<10^{-9}$) or non-exonic variants ($P=10^{-31}$) but also between synonymous exonic and non-exonic variants (P=0.0002). Average heterozygosity was also significantly different for the three classes: 0.282 ± 0.007 for non-synonymous variants, 0.311 ± 0.006 for synonymous variants, and 0.321 ± 0.002 for non-exonic variants. Lower minor allele frequencies for non-synonymous exonic SNPs and other SNPs are generally attributed to differential effects of purifying selection (16, 19, 25). Our results are of particular interest in this light because they provide evidence of a similar but less prevalent pattern when synonymous exonic SNPs are compared to SNPs outside of exons.

Gene-based haplotypes

Often only SNPs with minor allele frequency exceeding a predetermined minimum value (e.g. 0.05) will be examined for disease association because of limited statistic power to detect relationships with less frequent variants. In most genes, LD is sufficiently widespread to restrict considerably the observed combinations, or haplotypes, corresponding to different allelic phases of these SNPs. Assuming that the haplotypes reflect an underlying block structure, the information required to assess disease association can be obtained by examining a subset of SNPs that are chosen to distinguish between these (haplotype tag SNPs or htSNPs), rather than exhaustively examining all variation within a gene. Implementation of this strategy requires knowledge of the gene haplotypes.

To obtain the principal haplotypes for the genes in our study, we applied the SNPHAP program (see Methods) to the 65,080 SNPs with minor allele frequencies >0.05. These SNPs were distributed across 13,419 genes, 9,416 of which contained two or more SNPs. We identified 46,558 frequent haplotypes (frequency > 0.05), and 21,338 haplotypes with frequencies in the range 0.01-0.05. The average number of frequent haplotypes per gene reached a limit of 4.3 for genes containing at least 6 SNPs (Figure 3A). This was largely independent of the size of genomic region covered when the markers in the gene spanned 5 kb or more (Figure 3B). Although slightly more frequent haplotypes were observed for intermediate span lengths (10 kb –100 kb), overall the correlation is negligible (0.01; n.s.). Ninety percent of genes had 6 or fewer frequent haplotypes irrespective of the number of SNPs that they contained (Figure 3A), and only a small number (74 / 13,419) that contained at least one SNP with minor allele frequency > 0.05 exhibited no frequent haplotypes. We found an average of 10 haplotypes per gene in the frequency range 0.01-0.05, for genes in which the number of SNPs investigated was 13 or more, and 90% of genes had 17 or fewer such haplotypes (Figure 3A). Haplotypes of the "yin yang" pattern (i.e. contrasting alleles at several consecutive sites) were found with frequency similar to that reported by Zhang et al. (26) (results not shown). As indicated by these authors, these patterns are expected to occur frequently under a neutral evolution model.

Once haplotypes had been estimated, we determined htSNPs needed to distinguish between them. The total number of htSNPs was 30,420 to distinguish the common haplotypes and 39,017 to distinguish haplotypes with frequency >0.01 in the 13,419 genes. The upper limit on average number of these htSNPs per gene was 3 and 5 respectively for frequencies of >0.05 and >0.01 (Figure 3C). Inclusion of the

9,762 SNPs with minor allele frequency <0.05 did not substantially alter the limiting behaviour as the number of SNPs in a gene increased (results not shown).

Pairwise linkage disequilibrium

We also undertook LD analyses between pairs of SNP markers separated by <200 kb. Unless otherwise indicated, the results discussed in this section are restricted to analyses involving SNPs with minor allele frequency > 0.2 (where the allele frequency distributions for exonic SNPs and non-exonic SNPs are similar), and separation distances of 5 kb - 200 kb (where the differences are most apparent because the overall frequency of high LD pairs is low). The frequency of complete or nearly complete LD was much higher for exonic SNP pairs compared to other pairs: 886 / 1,608 (55%) for exonic pairs and 82,234 / 366,790 (22%) for other pairs (Figure 2B). The overall frequency of |D'| < 0.1 was also lower for exonic pairs compared to other pairs in the 5 kb - 200 kb separation range (8.4% vs. 23%). Ninety-eight percent of these intragenic exonic pairs are formed by SNPs from different exons, so the patterns observed apply across the gene and not just within exons. Intragenic exonic SNP pairs separated <5 kb also show an excess of high LD compared to other pairs within this distance range (93% vs. 89%), although the differences appear less marked because LD is generally higher at this distance.

Table 1 shows a more detailed examination of LD within the 601 genes from which the exonic pairs at distance 5 kb - 200 kb were drawn. This confirms that exonic pairs exhibited higher LD overall compared to other SNP combinations within the same genes. We also compared the intragenic pairs to neighbouring SNP pairs to control for possible regional variation in LD as described below. Again the exonic SNP pairs exhibit much higher LD overall than the neighbouring pairs whereas the

latter are indistinguishable from the other intragenic SNP pairs (Table 1). To evaluate differences statistically, we accounted for correlation of LD between intragenic SNPs by selecting only the most widely separated exonic SNP pair in the range 5 kb - 200 kb within each gene for analysis. These were compared to a similar number of control SNP pairs that were matched to the exonic pairs for separation distance, regional recombination rates estimated from the microsatellite marker map, and allele frequencies. The Wilcoxon signed rank test showed that the exonic SNP pairs had significantly higher values of |D'| ($p=0.008$) and r^2 ($p<0.0001$).

The frequencies of complete or nearly complete LD (|D'|>0.9) were also compared using a logistic regression model. We found significant effects of separation distance ($\chi^2_1 = 234.5$; $P<10^{-53}$), recombination rate ($\chi^2_1 = 67.1$; $P<10^{-16}$) and allele frequencies ($\chi^2_1 = 5.1$; $P<0.024$) as parameters in the analysis. We obtained $\chi^2_1 = 12.2$ ($P=0.0005$) for differences in the outcome |D'|>0.9 between the SNP pair exonic and control categories with the other variables in the equation.

Linkage disequilibrium across the genome

The wide distribution of our gene-based SNPs allowed us to examine LD patterns across the genome (Figure 1, Table 2). A strong inverse correlation was found between the strength of LD and the frequency of recombination per physical distance as estimated from the microsatellite marker map (27), as previously described for some chromosomes (13, 15). Centromeric regions were generally found to exhibit greater LD than telomeric regions (Figure 4A), consistent with the fact that recombination per physical distance is higher in the latter. However, the strength of the relationship between LD and recombination varies markedly between chromosomes and, sometimes, between chromosome arms (Table 2), while some

chromosome arms (chromosomes 7p and 8q) did not conform to the tendency for greater LD in centromeric regions (Figure 4A). Reasons for such differences could include issues related to study design, such as inaccuracies in the sequence assembly used for constructing the physical map or a low density of gene-based SNPs in some regions, or simply the low variability in the recombination rate on a particular chromosome or arm. However, these factors seem unlikely to account for the totality of the differences seen in Table 2 and Figure 4A.

Unique linkage disequilibrium patterns on chromosome 6p

Chromosome 6p exhibited one of the highest correlations between recombination rate and LD. Of particular interest are the complex LD patterns observed on chromosome 6p within a large region (28 Mb - 35 Mb) that spans the MHC locus (29.8 Mb - 33.4 Mb) (28) and has low recombination. Extensive LD is a known feature of the MHC (20, 29-33), but our study reveals that exotic complex LD patterns extend to more than 1 Mb to each side of this locus, throughout a region that includes 274 genes and 783 SNPs in our map. In order to compare the complex LD patterns here with those in other regions of the genome, we calculated a quantitative LD complexity index (CI). This was defined as the number of SNP triplets of which two flanking markers have high LD with each other and simultaneously have low LD with the internal marker (precise criteria are given in Methods) normalized by the total number of SNP triplets within windows of 1 Mb.

While the average CI over all the autosomal genome was 2%, it was between 10% - 50% at many points between 28.5 Mb - 35.5 Mb. The highest complexity (CI=50%) was within the MHC class I region, whereas in the class III the complexity rose to 10%. In the regions flanking the extended MHC, the maximum values of the

CI index were 36% (28 Mb - 29.8 Mb) and 11% (33.4 Mb - 35 Mb). We found no other instances in the genome with complexity approaching this over such an extended region (Figure 4B), despite a similar density of SNPs in several regions of equivalent length. Strong LD could also be observed between SNPs in the 6p region at distances of more than 1 Mb, with lack of LD between intermediate markers, which is a further unique feature of these LD patterns.

DISCUSSION

Our study provides information on the distribution of intragenic variation, LD and haplotype patterns as inferred from SNPs in 14,271 genes genotyped on up to 1,128 chromosomes. These SNPs are representative of the variants obtained by systematic resequencing of exons, 3' flanking regions, promoters and portions of the introns of these genes in 24 individuals (48 chromosomes). Because of the relatively large number of samples resequenced for SNP discovery phase, the database from which our study SNP are drawn contains most of common variants and larger representation of less frequent variants than available in most other datasets, which are strongly biased towards the most frequent polymorphisms, as discussed by Phillips et al. (15) amongst others.

We have previously shown that genetic variants occur less frequently in exons compared to other regions (18). Here, we have demonstrated that allele frequency spectra differ for both synonymous and non-synonymous variants within exons compared to those in other intragenic regions. Lower variability of non-synonymous exonic variants has been predicted because purifying selection is expected to act against alleles that introduce deleterious changes of amino acids (17, 19). We found that the average variability at synonymous exonic SNPs is intermediate between that of the other SNP groups (non-synonymous and non-exonic), and significantly different from non-exonic SNPs. This suggests that synonymous variants generally may also be subject to purifying selection, although this would appear to be less prevailing than that for non-synonymous variants. A possible explanation for this observation is selectively driven codon usage (34), which may affect translation/transcription efficiency, although association of this phenomenon with

other causes (e.g. three dimensional structural changes; changes in transcript factor binding properties; functional RNA, e.g. micro RNA, which may affect transcription/translation regulation or splicing efficiency) cannot be excluded.

We identified an average of 3.5 frequent haplotypes (frequency > 0.05) per gene, and more than 99% the genes studied exhibited at least one frequent haplotype. Interestingly, the average number of frequent haplotypes was largely independent both of the number of SNPs and the length of the genome region that they spanned, when the number of SNPs was sufficiently large. For genes that contained 6 or more SNPs, we found averages of 4.3 frequent haplotypes and 2.8 htSNPs, and 90% had 6 or fewer frequent haplotypes. Because we genotyped a large cohort, it was also possible to have reliable estimates for less frequent haplotypes. To eliminate the effect of rare variants, which are under-represented in the SNP database, we examined all haplotypes in the frequency range >0.01 that were formed by SNPs with minor allele frequencies > 0.05. Although the limits are somewhat less evident than those for frequent haplotypes alone (Figure 3), the average number was approximately 14 per gene once 14 or more SNPs were studied, and 90% of genes exhibited 23 or fewer haplotypes.

Our results can be compared with observations in Crawford et al. (35) who have studied haplotypes of 100 genes in a small sample of individuals of European and African origin. In 23 individuals of European descent, they found an average 4-5 frequent haplotypes and 13 total haplotypes from SNPs with minor allele frequency >0.05. This suggests a similar degree of haplotype diversity in European and Japanese. Since Crawford et al. examined all the common SNPs from the entire genomic sequence of the genes they studied, this also suggests that genotyping additional markers or undertaking frequent genomic resequencing would be unlikely

to reveal substantially more frequent haplotypes for genes containing 6 or more SNPs in our sample. The 24 individuals of African descent that Crawford et al. studied revealed a similar average number of frequent haplotypes (5) but a larger average number of total haplotypes (23), reflecting the expected higher genetic diversity compared to Europeans or Japanese. These results are consistent with the hypothesis of European and Asian populations arising from migration out of Africa (36-38).

Extrapolation from an estimate of 30,000 genes (39, 40) in the human genome leads us to suggest that the Japanese population contains a total of 130,000 frequent gene haplotypes, and that around 85,000 htSNPs may be required to distinguish these on an individual gene basis. In the present study, we have identified about 35% of the haplotypes and the htSNPs. We can also estimate that the Japanese population will contain at least 300,000 gene haplotypes with frequency 0.01-0.05 based on SNPs with minor allele frequencies > 0.05, and that additional 60,000 htSNPs will be required to distinguish these. However, the number of haplotypes in this frequency range would be much larger if SNPs with minor allele frequency 0.01-0.05 were included. At present, it is not possible to have a systematic survey of such variants because they are under-represented in databases.

A detailed analysis of LD patterns revealed that disequilibrium was stronger between exonic variants within the same gene compared to other combinations of SNPs (e.g. intronic or intergenic SNPs). This was most evident when examining pairs of SNPs separated by 5 kb or more, where LD is generally modest. For example, 55% of within-gene exonic SNP pairs were in complete or nearly complete LD compared to 22% of other SNP pairs in this separation range. In this comparison, most of the intragenic exonic pairs consist of SNPs from different exons, showing that the observed patterns are related to the whole gene rather than to variants within the

same exon. The differences in LD patterns were statistically significant after taking into account allele frequencies, separation distances, and the relationship of LD to recombination rates.

A number of previous studies have reported higher than expected LD in specific genes that were examined because of *a priori* evidence that they contained a selected variant (41, 42). However, to our knowledge, our results provide the first evidence of a widespread pattern of high LD specifically between intragenic exonic SNPs. High LD within a gene may arise due to hitchhiking, in which positive selection for a specific variant increases the frequency of neutral SNPs on the same haplotype (41, 42). High intragenic LD could also be due to non-random distribution of recombination events if, for example, recombination hotspots reside preferentially outside of genes. However, neither of these explanations would appear to account for a pattern of high LD specifically between intragenic exonic SNP pairs compared to other intragenic SNP pairs, such as observed in the 601 genes with widely separated exonic variants from our dataset. A possible explanation would be selection acting to maintain particular combinations of exonic SNPs. Irrespective of the interpretation given, it would now be of great interest to examine the same genes in other populations to compare LD patterns.

Across the genome, LD was highest in regions in which the recombination frequency per physical distance as estimated from the meiotic map was low, and in centromeric regions where recombination frequency per physical distance is also generally reduced. This is expected because LD is partially a reflection of historical recombination events. However, we found that the correlation between LD and recombination varies considerably for different chromosomes or chromosome arms. In the most extreme instances, recombination rate accounts for 72% of the variability

in LD on chromosome 20p but less than 2% on chromosome 7p. Some chromosome arms (chromosomes 7p and 8q) did not conform to the overall tendency for greater LD in centromeric regions. In addition, the chromosome 6p (28 Mb - 35 Mb) region, containing MHC genes but also many others, was shown to have a much more complex patterns of LD than other regions in our map, suggesting that unique factors contribute to maintaining LD here.

It is of interest to determine if similar patterns of LD variability will be found in other populations, as this is an important issue for understanding the origin of haplotype patterns and for the design of trans-ethnic disease mapping studies. Published LD maps on Caucasians based on a similar density of SNPs are available for two chromosomes, 19 and 22 (13, 15). Although formal comparison is difficult because of differences in the choice of markers, visual inspection shows a high degree of global resemblance in the gross patterns of LD in the two populations (Figure 4C). For example, both exhibit regions of extended high LD around 19 Mb - 23 Mb on chromosome 19 and 39 Mb - 41 Mb on chromosome 22. Variation in recombination accounted for a similar proportion of variance of LD in the Japanese and Caucasian samples on these chromosomes. Additional comparative studies of these and other chromosomes, particularly those for which we found a lower correlation between LD and recombination, are required to determine the degree of LD conservation between Japanese and Caucasian populations.

MATERIALS AND METHODS

DNA samples, marker selection and genotyping Written informed consent was provided by all participants in the genotyping study following procedures approved by the Ethical Committee at SNP Research Center, RIKEN, Tokyo. Genomic DNAs were prepared from white blood cells by standard methods. Markers were selected from the database at http://snp.ims.u-tokyo.ac.jp. Multiplexed amplification and genotyping were performed as described (23).

Data review and quality check Of the 80,788 original markers, we excluded 3,612 markers that could not be aligned uniquely to finished sequences in build33 (April 2003) of the human genomic sequence from National Center for Biotechnology Information, and 2 that were not polymorphic. Next, we considered a SNP to be validated only if the genotype data showed no significant deviation from Hardy-Weinberg equilibrium ($p>0.01$). The Hardy-Weinberg criteria led to elimination of 2,332 markers. Although this will include a number of markers that have exceeded the critical limit by chance, in many instances Hardy-Weinberg deviations were found to be due to technical issues (e.g. neighbouring SNPs causing imbalance of the polymerase chain reaction (PCR) products or affecting the genotyping assay).

Allele frequency comparisons SNPs were classified as non-synonymous exonic variants, synonymous exonic variants and non-exonic variants and categorised by minor allele frequency <0.05 or otherwise. To account for possible effects of sequence quality towards the extremities of sequenced fragments, which could lead to bias against detection of rare polymorphisms in these regions, we also classified the

SNPs by distance from sequencing primers (<80 bp, 80 bp - 270 bp, >270 bp). Although global differences of minor allele frequencies were detected between the distance categories, the frequencies were not correlated with distance within each category. Three-way contingency table analysis was performed using log-linear models, and the significance of the interaction term for SNP category/frequency is reported with the inclusion of the other second-order interaction terms included under the null and alternative hypotheses.

LD calculations and haplotypes We calculated |D'| and r^2 using standard methods (43, 44). Moving averages of |D'| were calculated in sliding windows of length 1 Mb using all markers within 500 kb of each SNP within the window. To estimate recombination rate by LD of our data for comparison, we used a decay model for fitting: $D' = (1-r)^n D_0'$, where n is the number of generation since $D' = D_0'$, and r is recombination fraction between two loci. Generation times ranging from 550-1250 were applied, and the conclusions were found to be unaffected by the choice. For estimating haplotypes we applied the program SNPHAP obtained from http://www-gene.cimr.cam.ac.uk/clayton/software, except that for reanalysis of the Caucasian chromosome 19 data (15) in which haplotypes were reconstructed using the programs Merlin (45) and Fugue (obtained from G. R. Abecasis).

Assessment of LD differences We assessed LD differences between intragenic exonic pairs of SNPs and other pairs using the non-parametric Wilcoxon signed rank test for |D'| and r^2, and logistic regression with the outcome variable |D'|>0.9. We selected the most widely spaced exonic SNPs with minor allele frequencies >0.2 for these tests. The exonic pairs were matched to the closest other SNP pair (excluding

other intragenic exonic pairs) that had similar separation distance, allele frequencies, and associated recombination. When a matched pair could not be found on the same chromosome, a pair meeting the other matching criteria was selected from a different chromosome. P-values using the normal approximation for the Wilcoxon signed rank test statistic were found to be conservative in simulations involving 50,000 replicates in which SNP pairs with similar characteristics to these intragenic exonic pairs were selected randomly from all pairs and matched as above. Continuous variables that were included in the assessment outcome frequencies in the logistic regression were: the average minor allele frequencies for the SNPs in the pair, the physical distance between the pair, and the recombination rate assigned to the region using the microsatellite marker map.

LD complexity index We defined LD complexity based on combinations of 3 SNPs where: a) LD between flanking markers is high, specifically 95% upper confidence limit for |D'| estimate > 0.98 and 95% lower confidence limit for |D'| estimate > 0.7; b) LD between the interior marker and the flanking markers is low, specifically with 95% upper confidence limit < 0.9 in both instances. The complexity index was calculated as the frequency within windows of 1 Mb length of triplets meeting the criteria in a) and b) amongst all triplets that met the criteria in a).

ACKNOWLEDGEMENTS

The authors thank G.R. Abecasis for providing and instructing the haplotype-phasing program Merlin and Fugue. We also thank M. Yamaguchi, H. Kawakami, and members of SNP Research Center, RIKEN for support.

REFERENCES

1. Boehnke,M. (2000) A look at linkage disequilibrium. *Nat. Genet.,* **25,** 246-247.
2. McCarthy,J.J. and Hilfiker,R. (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.,* **18,** 505-508.
3. Goldstein,D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.,* **29,** 109-111.
4. Cardon,L.R. and Bell,J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.,* **2,** 91-99.
5. Ozaki,K., Ohnishi,Y., Iida,A., Sekine,A., Yamada,R., Tsunoda,T., Sato,H., Hori,M., Nakamura,Y. and Tanaka,T. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.,* **32,** 650-654.
6. Takei,T., Iida,A., Nitta,K., Tanaka,T., Ohnishi,Y., Yamada,R., Maeda,S., Tsunoda,T., Takeoka,S., Ito,K. *et al.* (2002) Association between single-nucleotide polymorphisms in selectin genes and immunoglobulin A nephropathy. *Am. J. Hum. Genet.,* **70,** 781-786.
7. Suzuki,A., Yamada,R., Chang,X., Tokuhiro,S., Sawada,T., Suzuki,M., Nagasaki,M., Nakayama-Hamada,M., Kawaida,R., Ono,M. *et al.* (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.,* **34,** 395-402.
8. Tokuhiro,S., Yamada,R., Chang,X., Suzuki,A., Kochi,Y., Sawada,T., Suzuki,M., Nagasaki,M., Ohtsuki,M., Ono,M. *et al.* (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.,* **35,** 341-348.

9. Taillon-Miller,P., Bauer-Sardina,I., Saccone,N.L., Putzel,J., Laitinen,T., Cao,A., Kere,J., Pilia,G., Rice,J.P. and Kwok,P.Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.*, **25,** 324-328.
10. Abecasis,G.R., Noguchi,E., Heinzmann,A., Traherne,J.A., Bhattacharyya,S., Leaves,N.I., Anderson,G.G., Zhang,Y., Lench,N.J., Carey,A. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68,** 191-197.
11. 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.
12. Reich,D.E., Cargill,M., Bolk,S., Ireland,J., Sabeti,P.C., Richter,D.J., Lavery,T., Kouyoumjian,R., Farhadian,S.F., Ward,R. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411,** 199-204.
13. Dawson,E., Abecasis,G.R., Bumpstead,S., Chen,Y., Hunt,S., Beare,D.M., Pabial,J., Dibling,T., Tinsley,E., Kirby,S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418,** 544-548.
14. Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296,** 2225-2229.
15. Phillips,M.S., Lawrence,R., Sachidanandam,R., Morris,A.P., Balding,D.J., Donaldson,M.A., Studebaker,J.F., Ankener,W.M., Alfisi,S.V., Kuo,F.S. *et al.* (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.*, **33,** 382-387.
16. Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al.* (1999) Characterization of single-

nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.,* **22,** 231-238.

17. Halushka,M.K., Fan,J.B., Bentley,K., Hsie,L., Shen,N., Weder,A., Cooper,R., Lipshutz,R. and Chakravarti,A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.,* **22,** 239-247.
18. Haga,H., Yamada,R., Ohnishi,Y., Nakamura,Y. and Tanaka,T. (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. *J. Hum. Genet.,* **47,** 605-610.
19. Hughes,A.L., Packer,B., Welch,R., Bergen,A.W., Chanock,S.J. and Yeager,M. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc. Natl. Acad. Sci. U S A,* **100,** 15754-15757.
20. 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.
21. Pritchard,J.K. and Przeworski,M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.,* **69,** 1-14.
22. The International HapMap Consortium (2003) The International HapMap Project. *Nature,* **426,** 789-796.
23. Ohnishi,Y., Tanaka,T., Ozaki,K., Yamada,R., Suzuki,H. and Nakamura,Y. (2001) A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.,* **46,** 471-477.
24. Yamada,R., Kawakami,H., Yamaguchi,M., Tatsu,E., Sekine,A., Yamamoto,K., Nakamura,Y. and Tsunoda,T. (2001) Analysis of population structure of Japanese

population using genotype data of genome wide hundreds of SNPs. *51th Annual meeting, The American Society of Human Genetics.* San Diego, CA, **69** supplement, Nr.1413.

25. Kimura,M. and Ota,T. (1974) On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U S A,* **71,** 2848-2852.
26. Zhang,J., Rowe,W.L., Clark,A.G. and Buetow,K.H. (2003) Genomewide Distribution of High-Frequency, Completely Mismatching SNP Haplotype Pairs Observed To Be Common across Human Populations. *Am. J. Hum. Genet.,* **73,** 1073-1081.
27. Kong,A., Gudbjartsson,D.F., Sainz,J., Jonsdottir,G.M., Gudjonsson,S.A., Richardsson,B., Sigurdardottir,S., Barnard,J., Hallbeck,B., Masson,G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.,* **31,** 241-247.
28. Mungall,A.J., Palmer,S.A., Sims,S.K., Edwards,C.A., Ashurst,J.L., Wilming,L., Jones,M.C., Horton,R., Hunt,S.E., Scott,C.E., *et al.* (2003) The DNA sequence and analysis of human chromosome 6. *Nature,* **425,** 805-811.
29. Begovich,A.B., McClure,G.R., Suraj,V.C., Helmuth,R.C., Fildes,N., Bugawan,T.L., Erlich,H.A. and Klitz,W. (1992) Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *J. Immunol.,* **148,** 249-258.
30. Bugawan,T.L., Klitz,W., Blair,A. and Erlich,H.A. (2000) High-resolution HLA class I typing in the CEPH families: analysis of linkage disequilibrium among HLA loci. *Tissue Antigens,* **56,** 392-404.

31. Klitz,W., Stephens,J.C., Grote,M. and Carrington,M. (1995) Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the HLA class II region. *Am. J. Hum. Genet.,* **57,** 1436-1444.

32. Kochi,Y., Yamada,R., Kobayashi,K., Takahashi,A., Suzuki,A., Sekine,A., Mabuchi,A., Akiyama,F., Tsunoda,T., Nakamura,Y. *et al.* (2004) Analysis of single-nucleotide polymorphisms in Japanese rheumatoid arthritis patients shows additional susceptibility markers besides the classic shared epitope susceptibility sequences. *Arthritis Rheum.,* **50,** 63-71.

33. Walsh,E.C., Mather,K.A., Schaffner,S.F., Farwell,L., Daly,M.J., Patterson,N., Cullen,M., Carrington,M., Bugawan,T.L., Erlich,H. *et al.* (2003) An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.,* **73,** 580-590.

34. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.,* **53,** 290-298.

35. Crawford,D.C., Carlson,C.S., Rieder,M.J., Carrington,D.P., Yi,Q., Smith,J.D., Eberle,M.A., Kruglyak,L. and Nickerson,D.A. (2004) Haplotype Diversity across 100 Candidate Genes for Inflammation, Lipid Metabolism, and Blood Pressure Regulation in Two Populations. *Am. J. Hum. Genet.,* **74,** 610-622.

36. Jorde,L.B., Bamshad,M. and Rogers,A.R. (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays,* **20,** 126-136.

37. Ingman,M., Kaessmann,H., Paabo,S. and Gyllensten,U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature,* **408,** 708-713.

38. Underhill,P.A., Shen,P., Lin,A.A., Jin,L., Passarino,G., Yang,W.H., Kauffman,E., Bonne-Tamir,B., Bertranpetit,J., Francalacci,P. *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nat. Genet.,* **26,** 358-361.

39. Ewing,B. and Green,P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.,* **25,** 232-234.

40. Roest Crollius,H., Jaillon,O., Bernot,A., Dasilva,C., Bouneau,L., Fischer,C., Fizames,C., Wincker,P., Brottier,P., Quetier,F. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.,* **25,** 235-238.

41. Sabeti,P.C., Reich,D.E., Higgins,J.M., Levine,H.Z., Richter,D.J., Schaffner,S.F., Gabriel,S.B., Platko,J.V., Patterson,N.J., McDonald,G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature,* **419,** 832-837.

42. Toomajian,C., Ajioka,R.S., Jorde,L.B., Kushner,J.P. and Kreitman,M. (2003) A method for detecting recent selection in the human genome from allele age estimates. *Genetics,* **165,** 287-297.

43. Lewontin,R.C. (1988) On measures of gametic disequilibrium. *Genetics,* **120,** 849-852.

44. Hartl,D.L. and Clark,A.G. (1997) *Principles of population genetics.* 3rd ed. Sinauer Associates, Sunderland, Mass.

45. Abecasis,G.R., Cherny,S.S., Cookson,W.O. and Cardon,L.R. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.,* **30,** 97-101.

LEGENDS TO FIGURES

Figure 1. Distribution of genes, SNPs and LD. LD was evaluated by calculating average pairwise |D'| within sliding windows of 2 Mb length. To alleviate effects of low allele frequencies on |D'|, we made these calculations twice, first, with the 65,080 SNPs having minor allele frequency > 0.05 and, second, with the 42,116 SNPs having minor allele frequency> 0.2. The results of the two calculations were similar. Black lines: sliding window plots of |D'| coefficients for SNPs with minor allele frequency > 0.2; dashed lines indicate segments where SNP density was judged to insufficient for these calculation. Green lines: Meiotic recombination rates estimated from the microsatellite marker map (27). Blue squares: regions of high LD with the length of the region represented by height at which the square is placed. Red triangles: normalized frequency of low LD segments. The values of |D'|, recombination rates, regions of high LD, and normalized frequency of low LD segments range from (0, 1.0), (0, 3 cM/Mb), (0, 500 kb), and (0, 0.5), respectively. The sections below the LD graph show the number of SNPs and genes per sliding window, respectively.

Figure 2. Allele frequencies and pairwise linkage disequilibrium. (**A**) The observed distributions of minor allele frequencies of all exonic (black bar), non-synonymous (purple bar), synonymous (yellow bar) and non-exonic SNPs (white bar). (**B)** LD as a function of separation distance in the range 5 kb - 200 kb for SNPs with minor allele frequencies >0.2. The plot shows a moving average of the frequency of |D'|>0.9 for within gene exonic SNP pairs (red line), between gene exonic SNP pairs (green line) and other SNP pairs (blue line).

Figure 3. Characteristics of haplotypes and haplotype tag SNPs in gene. (**A**) Average number (red line) and 90% limit (green line) of haplotypes with frequencies > 0.05 and average number (blue line) and 90% limit (purple line) of haplotypes with frequencies > 0.01 per gene as a function of the number of SNPs in gene. (**B**) Average number of haplotypes per gene frequency > 0.05 as a function of the distance spanned by the SNPs for all genes (red line) and genes with 6 or more SNP (green line). **(C**) Average number (red line) and 90% limit (green line) of haplotype tag SNPs for haplotypes with frequencies > 0.05 and average number (blue line) and 90% limit (purple line) of haplotype tag SNPs for haplotypes with frequencies > 0.01 per gene according to the number of SNPs in gene.

Figure 4. Comparison of LD patterns. (**A**) Average |D'| within telomeric and centromeric region on each chromosome arm. (**B**) Comparison of LD complexity index around the MHC region with other regions of the genome. (**C**) Comparison of LD patterns in Japanese (solid lines) and Caucasians (broken lines) for chromosomes 19 and 22 including all SNPs with minor allele frequency > 0.05.

Table 1. LD patterns for SNP pairs in 601 genes compared to neighbouring SNP pairs

		All SNPs in 601 selected genes and control pairs		
Distance range	Type of pair	LD range (|D'|)		
		0-0.10	0.10-0.90	0.90-1.0
0-5000	Exonic within	3	44	561
	%	0.4%	7.2%	92.2%
	Others within	77	846	6650
	%	1.0%	11.1%	87.8%
	Neighbours	24	290	3426
	%	0.6%	7.7%	91.6%
5000-10000	Exonic within	15	99	285
	%	3.7%	24.8%	71.4%
	Others within	116	1433	3578
	%	2.2%	27.9%	69.7%
	Neighbours	23	379	933
	%	1.7%	28.3%	69.8%
10000-20000	Exonic within	10	135	314
	%	2.1%	29.4%	68.4%
	Others within	383	3148	4279
	%	4.9%	40.3%	54.7%
	Neighbours	65	614	922
	%	4.0%	38.3%	57.5%
20000-30000	Exonic within	14	91	120
	%	6.2%	40.4%	53.3%
	Others within	464	2993	2178
	%	8.2%	53.1%	38.6%
	Neighbours	111	618	469
	%	9.2%	51.5%	39.1%
30000-50000	Exonic within	18	104	117
	%	7.5%	43.5%	48.9%
	Others within	1033	4529	1817
	%	13.9%	61.3%	24.6%
	Neighbours	236	947	320
	%	15.7%	63.0%	21.2%
50000-100000	Exonic within	40	117	35
	%	20.8%	60.9%	18.2%
	Others within	2763	6847	934
	%	26.2%	64.9%	8.8%
	Neighbours	393	1300	184
	%	20.9%	69.2%	9.8%
100000-200000	Exonic within	38	41	15
	%	40.4%	43.6%	15.9%
	Others within	4098	4358	260
	%	47.0%	50.0%	2.9%
	Neighbours	564	717	28
	%	43.0%	54.7%	2.1%

Table 2. Correlation between LD and meiotic recombination rates on different autosomal chromosomes and chromosome arms

Chr.	per chromosome				p arm				q arm			
	r^2	windows	sd(m)	sd(e)	r^2	windows	sd(m)	sd(e)	r^2	windows	sd(m)	sd(e)
1	0.21	376	0.22	0.26	0.17	199	0.20	0.26	0.24	175	0.24	0.26
2	0.31	374	0.21	0.31	0.29	143	0.17	0.33	0.29	230	0.23	0.29
3	0.34	285	0.28	0.31	0.64	129	0.32	0.33	0.10	156	0.25	0.30
4	0.28	230	0.23	0.31	0.31	60	0.28	0.33	0.23	169	0.19	0.29
5	0.45	244	0.25	0.34	0.19	55	0.21	0.30	0.46	191	0.25	0.34
6	0.31	280	0.27	0.31	0.57	111	0.29	0.35	0.16	168	0.26	0.28
7	0.25	259	0.23	0.32	0.01	103	0.12	0.27	0.30	154	0.27	0.32
8	0.24	141	0.19	0.27	0.31	71	0.18	0.29	0.12	73	0.18	0.23
9	0.29	167	0.18	0.32	0.39	56	0.19	0.28	0.22	109	0.18	0.34
10	0.25	201	0.26	0.31	0.09	61	0.22	0.32	0.21	135	0.26	0.29
11	0.26	199	0.32	0.33	0.47	67	0.34	0.38	0.11	132	0.29	0.27
12	0.24	193	0.23	0.29	0.09	60	0.23	0.23	0.23	133	0.23	0.28
13	0.14	99	0.19	0.31	-	-	-	-	0.14	99	0.19	0.31
14	0.21	125	0.22	0.24	-	-	-	-	0.21	125	0.22	0.24
15	0.44	135	0.28	0.36	-	-	-	-	0.44	135	0.28	0.36
16	0.47	125	0.24	0.32	0.42	57	0.18	0.32	0.56	69	0.28	0.33
17	0.42	141	0.28	0.31	0.45	42	0.31	0.29	0.38	96	0.25	0.31
18	0.49	96	0.27	0.33	0.45	26	0.27	0.26	0.34	69	0.23	0.31
19	0.21	101	0.28	0.22	0.17	44	0.31	0.22	0.28	57	0.25	0.23
20	0.68	105	0.26	0.33	0.73	49	0.22	0.28	0.57	55	0.24	0.28
21	0.32	46	0.14	0.24	-	-	-	-	0.32	46	0.14	0.24
22	0.55	64	0.25	0.31	-	-	-	-	0.55	64	0.25	0.31

r correlation coefficient calculated for SNPs with minor allele frequency > 0.05

sd(m) standard deviation of $\log_{10}$(measured meiotic recombination rate)

sd(e) standard deviation of $\log_{10}$(recombination rate expected by LD)

Figure 1.

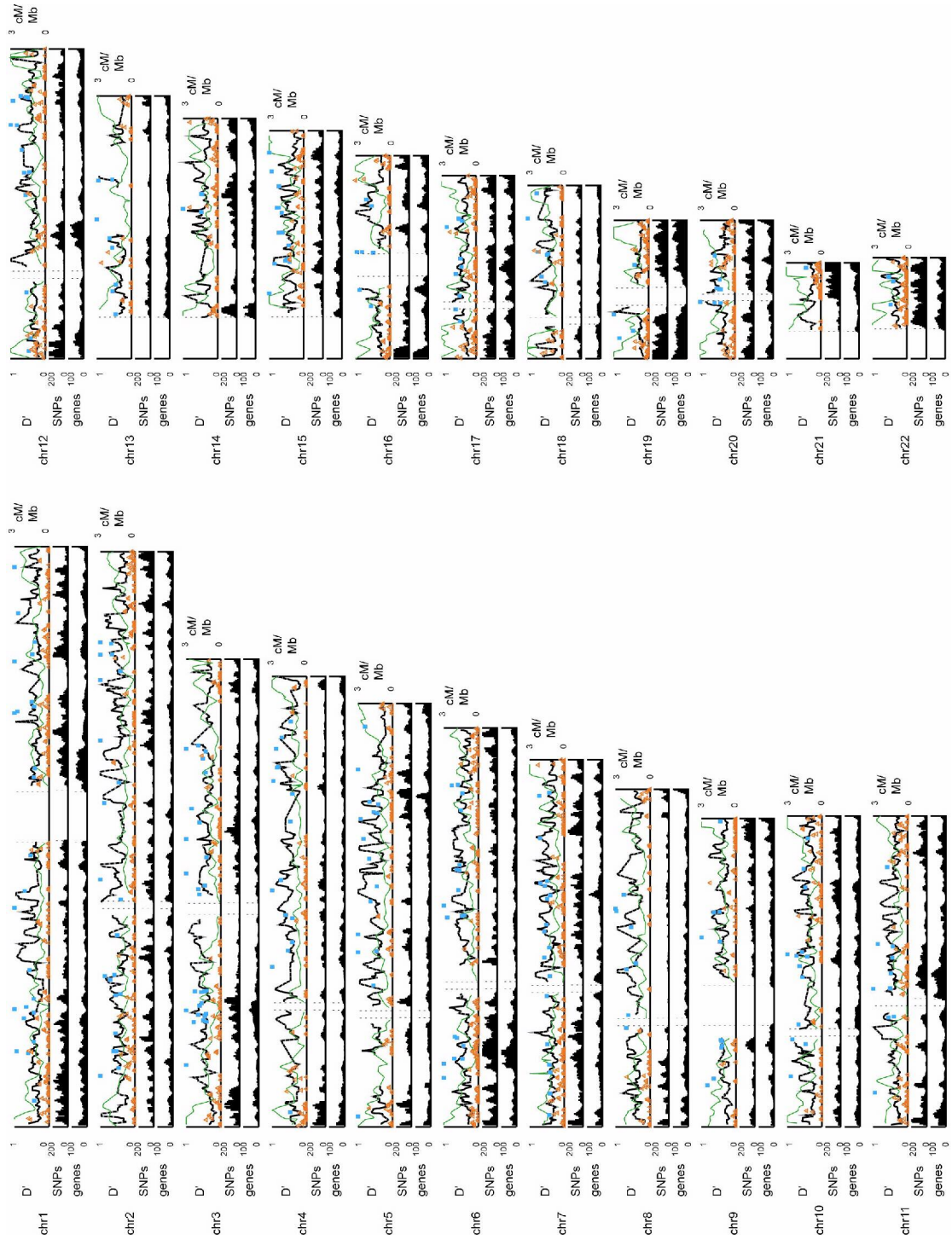

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
D'
SNPs
genes
cM/Mb

Figure 2.

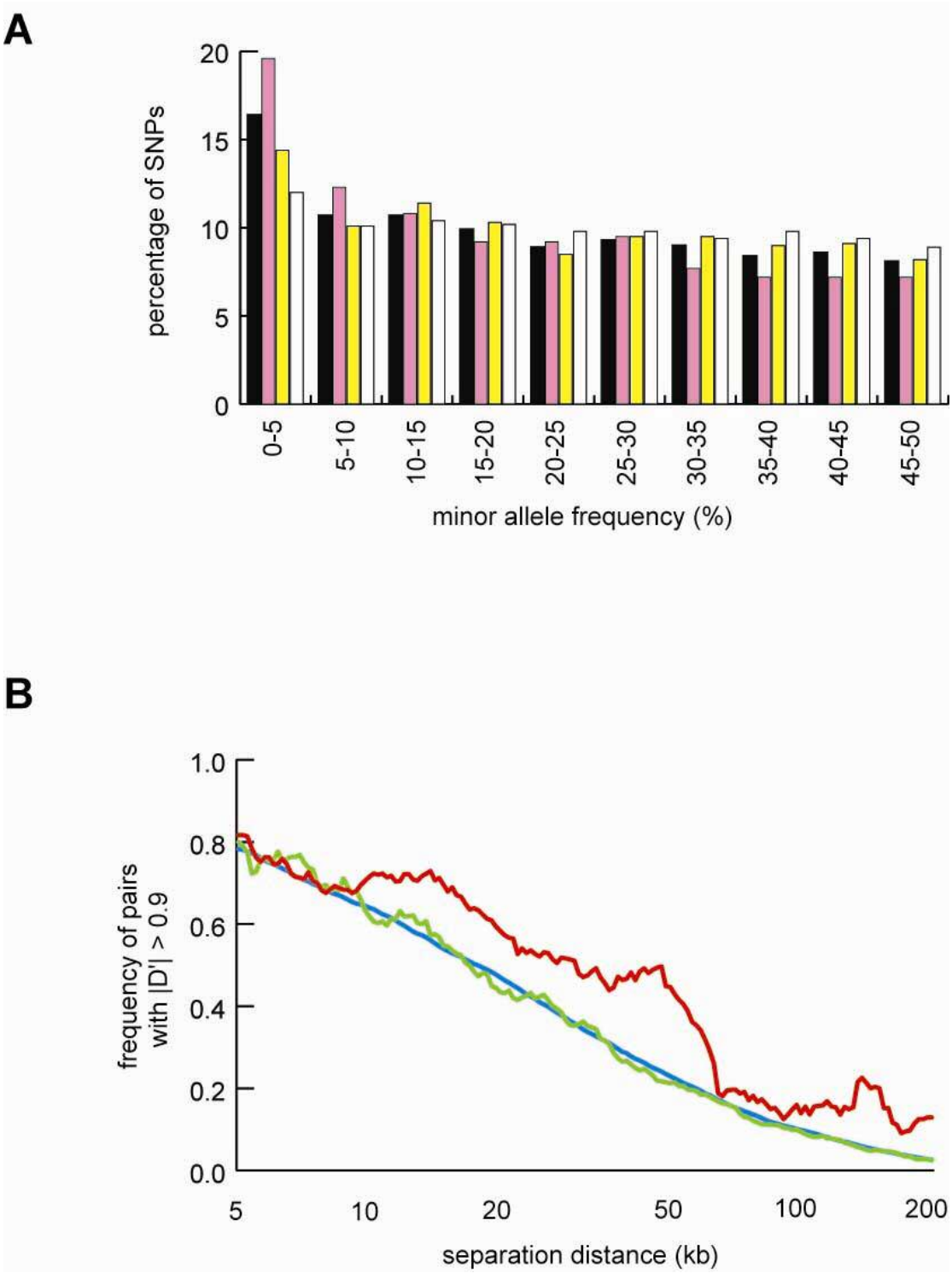

A
percentage of SNPs
20
15
10
5
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
minor allele frequency (%)
B
frequency of pairs with |D'| > 0.9
1.0
0.8
0.6
0.4
0.2
0.0
5
10
20
50
100
200
separation distance (kb)

Figure 3.

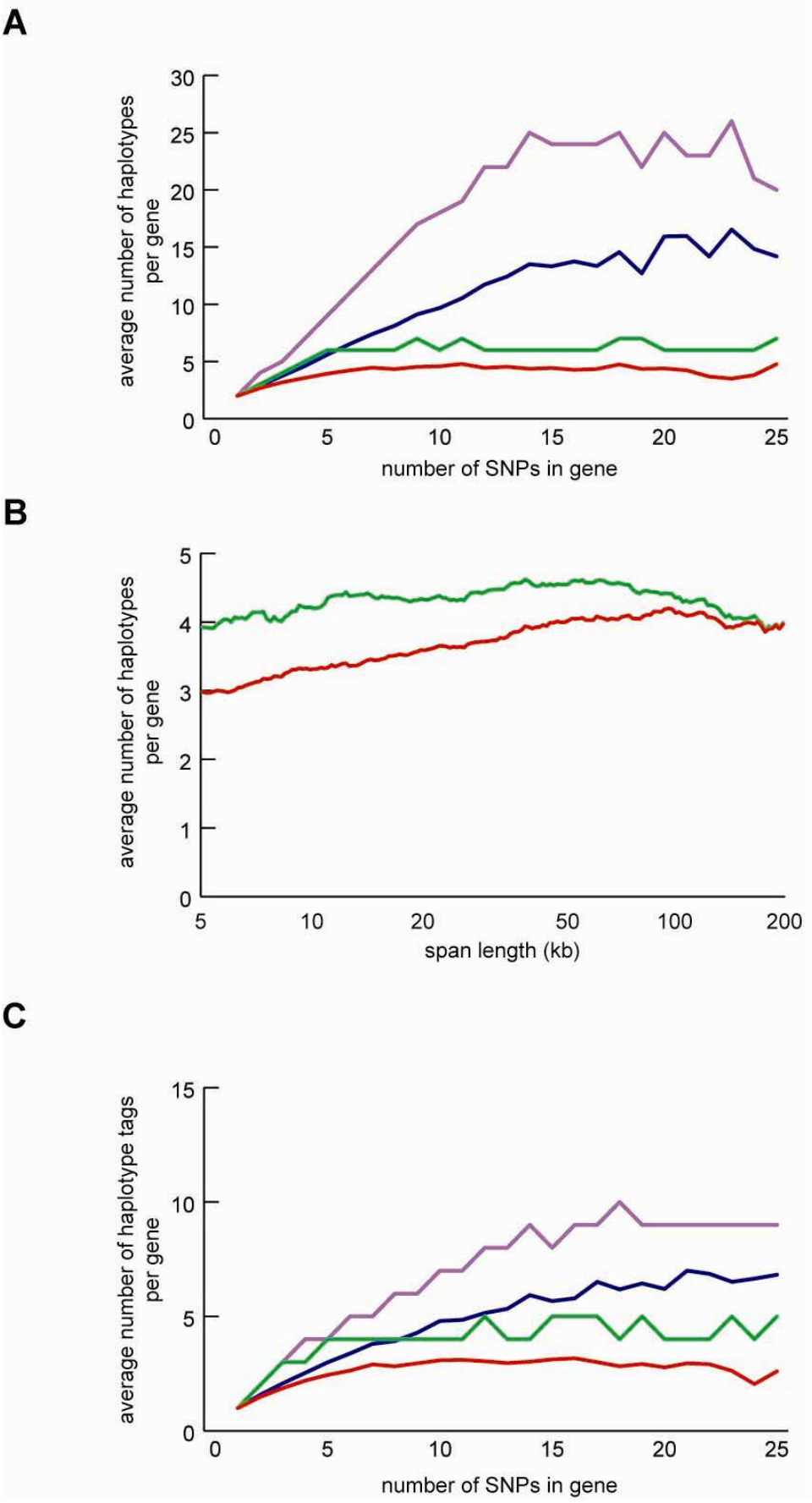
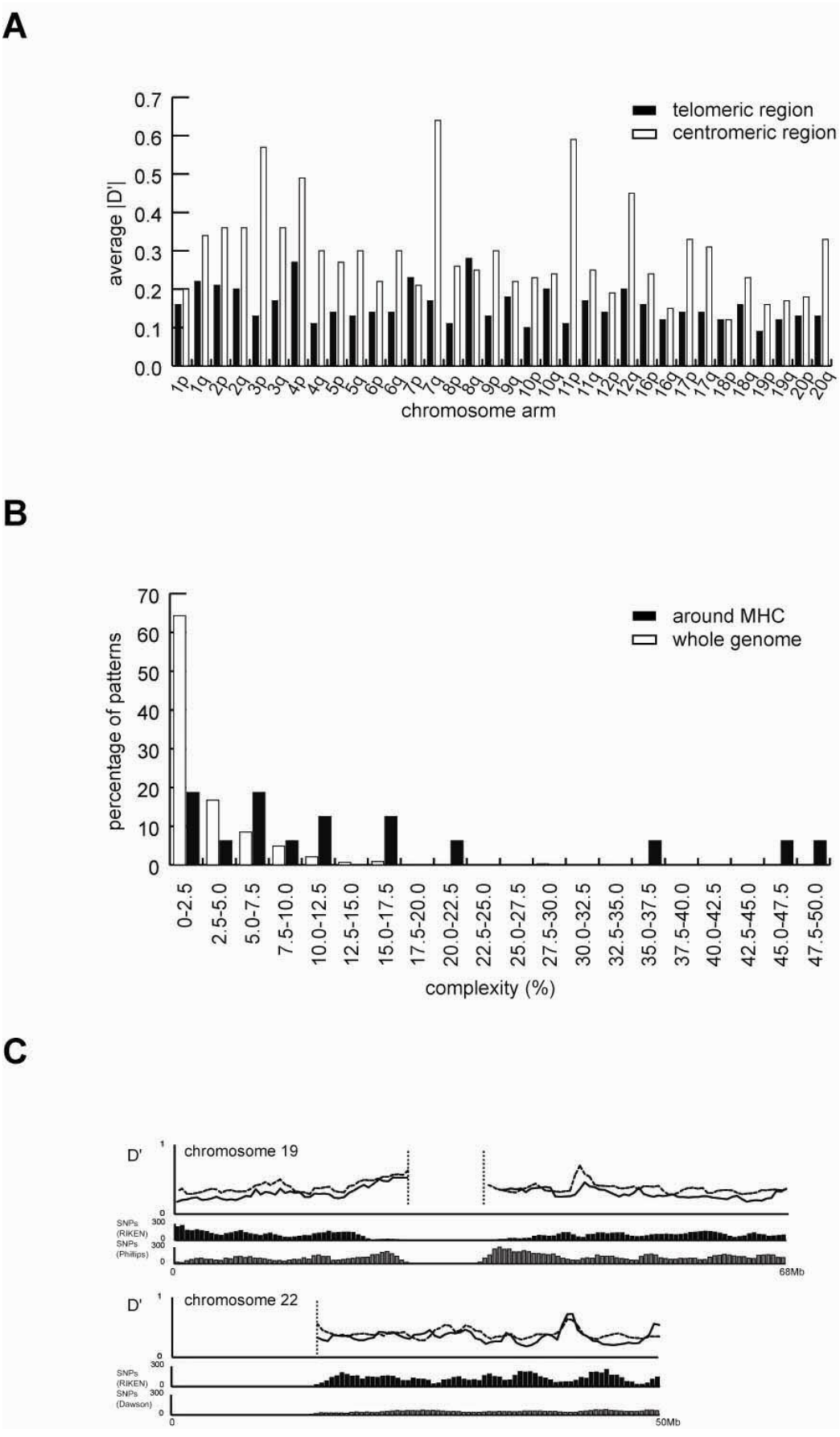

A
average number of haplotypes per gene
30
25
20
15
10
5
0
0
5
10
15
20
25
number of SNPs in gene
B
average number of haplotypes per gene
5
4
3
2
1
0
5
10
20
50
100
200
span length (kb)
C
average number of haplotype tags per gene
15
10
5
0
0
5
10
15
20
25
number of SNPs in gene

Figure 4.


A
telomeric region
centromeric region
average |D'|
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1p 1q 2p 2q 3p 3q 4p 4q 5p 5q 6p 6q 7p 7q 8p 8q 9p 9q 10p 10q 11p 11q 12p 12q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q
chromosome arm
B
around MHC
whole genome
percentage of patterns
0
10
20
30
40
50
60
70
0-2.5 2.5-5.0 5.0-7.5 7.5-10.0 10.0-12.5 12.5-15.0 15.0-17.5 17.5-20.0 20.0-22.5 22.5-25.0 25.0-27.5 27.5-30.0 30.0-32.5 32.5-35.0 35.0-37.5 37.5-40.0 40.0-42.5 42.5-45.0 45.0-47.5 47.5-50.0
complexity (%)
C
D'
chromosome 19
SNPs (RIKEN)
SNPs (Phillips)
68Mb
D'
chromosome 22
SNPs (RIKEN)
SNPs (Dawson)
50Mb

ABBREVIATIONS

SNP	single nucleotide polymorphism
LD	linkage disequilibrium
MHC	major histocompatibility complex
UTR	untranslated region
htSNP	haplotype tag SNP
CI	complexity index
PCR	polymerase chain reaction