Genome-Wide Linkage Disequilibrium from 100,000 SNPs in the East Finland Founder Population

Pekka Uimari,[1] Outi Kontkanen,[1] Peter M. Visscher,[2] Mia Pirskanen,[1] Ricardo Fuentes,[1] and Jukka T. Salonen[1]

[1] *Jurilab Ltd, Kuopio, Finland*

[2] *Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, United Kingdom*

Information about linkage disequilibrium (LD) is important in understanding the genome structure and has its applications in association studies. Here we present the first genome-wide LD study based on a founder population (East Finland). The LD data consist of 118 unrelated individuals and around 480,000 SNP pairs genotyped with the Affymetrix 100K genotyping assay. Using the minor allele frequency (MAF) limit of .05, the squared correlation coefficient between two loci (r^2) was .48, .37, .28, and .20 for distances of 5, 10, 20, and 40 kb respectively. MAF had a significant effect on the mean r^2 so that the extent of useful LD ($r^2 > .3$) varied from 17 kb to 80 kb depending on the limit set for the MAF. For D' the effect of MAF was smaller but reflected the possible age of the mutation: SNPs with high MAF had lower D' than those with low MAF. The X chromosome showed higher D' values than autosomes and the extent of useful LD ($r^2 > .3$) was twice as long on the X chromosome than on the autosomes. Based on the results, LD varies across the genome and is correlated to local recombination rate between and within chromosomes. However, the recombination rate does not explain all the variation found in LD. We also report a number of long chromosomal regions where exceptionally high or low LD were detected.

After the successful period of mapping genes and mutations causing simple Mendelian disorders with family-based linkage methods, research groups around the world have now focused their interest on mapping loci predisposing to complex and common diseases. Currently the most promising design is the population-based case-control (or association study) design (Risch & Merikangas, 1996). Traditionally, association studies have been used in the fine mapping phase where the resolution of the family design is limited. Whole-genome association studies have not previously been possible because of the lack of dense enough marker maps and high-throughput genotyping platforms. Based on the first simulations, approximately 500,000 single-nucleotide polymorphisms (SNPs) would be required for whole-genome studies.[2] In the near future, new genotyping platforms will dramatically lower operating costs and increase throughput so that the required number of SNPs can be typed.

In addition to the effect size of a disease-predisposing allele, the power of an association study depends on linkage disequilibrium (LD), the nonrandom association of alleles at different loci. One of the LD measures, the squared correlation coefficient between two loci (r^2), is directly related to power. To achieve the same power at the marker locus as is achievable at the disease-predisposing locus, the sample size must be increased by a factor of $1/r^2$ (Pritchard & Preworski, 2001). The extent of useful LD (say, $r^2 > 0.3$) gives the resolution that a genotyping assay should have in order to be useful in association studies with a reasonable number of samples.

The extent of LD is population-specific and also varies across the genome. The factors that influence the extent of LD at the population level include genetic drift, population growth, admixture or migration, and population structure (Ardlie et al., 2002). Less is known, however, about the factors which create variation in LD between genomic regions. One such component obviously is the recombination rate (Greenwood et al., 2004; McVean et al., 2004). Also, natural selection can be an important factor maintaining LD whereas gene conversion and mutation might be involved in eroding LD in specific chromosomal regions (Ardie et al., 2002; Ardie et al., 2001). The extent of LD also depends on the markers used in a study. Generally, studies based on microsatellites report longer range LD than studies based on SNPs (Laan & Pääbo, 1997; Peterson et al., 1995; Varilo et al., 2003). This finding has been explained by higher information content and mutation rate in microsatel-

Received 30 March, 2005; accepted 4 April, 2005.

Address for correspondence: Pekka Uimari, Jurilab Ltd, Microkatu 1, 70210 Kuopio, Finland (www.jurilab.com).
E-mail: pekka.uimari@jurilab.com

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

lites when compared to SNPs (Kauppi et al., 2003; Pritchard & Preworski, 2001; Varilo et al., 2003).

The first population-based simulation studies estimated that a useful level of LD for SNPs extends only 3 kb while no level of LD was observed for distances of 30 kb and longer in general populations (Kruglyak, 1999). This estimate may be too conservative on the basis of other simulations (Pritchard & Preworski, 2001) and empirical findings. On the other hand, the International HapMap Project aims to create a map of 600,000 SNPs or more to cover most of the common haplotypes in the genome and aid the disease gene-mapping (The International HapMap Consortium, 2003).

A number of studies have been published that have investigated LD in the human genome and we summarize their findings, in particular those in which samples from Finns are used. Dunning et al. (2000) studied LD in four populations: Afrikaners, Ashkenazim, Finns and East Anglian British. The Finns were from the same region as the samples in our study. The decay of *D'* was rapid in all four populations. Significant LD was observed between marker pairs less than 20 kb apart in all four populations. However, only Finns showed significant LD for distances of more than 500 kb between SNPs. This is in agreement with theoretical calculations that LD for loci that are close together reflects the effective population size (N_e) in the distant past whereas LD for loci over long distances reflects more recent N_e (Hayes et al., 2003). Thus loci that are close together may have similar LD across populations whereas loci that are further apart exhibit more LD in small founder populations than in general populations with larger recent N_e. Taillon-Miller et al. (2000) studied LD on the X chromosome in three populations: CEPH (The Centre d'Etude du Polymorphisme Humain), Finns, and Sardinians. The Finnish population in this study was a small regional isolate. The extent of LD was similar in all three populations. This was explained by the choice of SNPs (more ancient SNPs than general random SNPs from the genome). The extent of useful LD was approximately 50 kb in all three populations. Kaessmann et al. (2002) studied LD in four populations: Evenki, Saami, Finns and Swedes. The Finns in this study were a random sample of individuals from all over the country. The authors found extensive LD in Evenki and Saami populations and weaker and similar LD in both Finns and Swedes. Varilo et al. (2003) compared LD in three different Finnish populations: the early settlement, the late settlement and a regional isolate. The results reflected the population history well: the highest LD was observed in the most isolated (and the youngest) population and the lowest LD in the early settlement population. The late settlement population is the same population from where our samples originated. Finally, Shifman et al. (2003) concluded that the extent of useful LD varies between 7 kb and 23 kb for 'the low LD regions' and 'the high LD regions' respectively in Caucasian populations. Only a slightly stronger LD was observed in the Ashkenazi population than in the Caucasian population. LD in the African American population was only half of that in the Caucasian population.

A more comprehensive review of the LD studies done before 2002 was made by Ardlie et al. (2002). These earlier studies were based on European or European descent populations with sample size varying from 10 to 1000 individuals, and the number of investigated SNPs varying from four to several thousands. In general, depending on efficiency and cost of the genotyping platform, the total number of genotypes was from a few thousand to tens of thousands. The review concludes that LD is stronger in European populations than in African populations, and that LD varies from one region of the genome to another. Also the extent of useful LD ranges from 10 to 30 kb for European populations.

In this article we present the characteristics of LD in the East Finland founder population. We present the first genome-wide LD study at high marker density in a founder population. We will show that the extent of useful LD in the East Finland population is from 17 kb to 80 kb, depending on the threshold used for the MAF (minor allele frequency). We also show that the variation in LD is in agreement with the ratio of chromosome physical and genetic lengths but does not explain all of the variation. Further, we list genomic regions that show exceptionally low or high LD.

Materials and Methods

Samples

A sample of 118 unrelated subjects (117 males and 1 female) were used in this study. To study LD on the X chromosome only 117 males were used. The sample size reflects the recommendation given by Weiss and Clark (2002). Subjects were selected from the Kuopio Ischaemic Heart Disease Risk Factor (KIHD) prospective population-based cohort study, which was designed to investigate risk factors for coronary heart disease (CHD), atherosclerosis, acute myocardial infarction (AMI) and related outcomes in the East Finland founder population (Salonen, 1988). The KIHD cohort includes almost 3000 subjects. The KIHD study protocol was approved by the Research Ethics Committee of the University of Kuopio. The subjects were a random subsample of the participants in an AMI nested case-control association study. From the 118 subjects, 77 were controls and 47 were cases. In some genome regions this may affect the results but those areas are only a fraction of the whole genome. All subjects were from the Kuopio region in East Finland. Their parents were also from the same region.

The East Finland Population

The East Finland (including North Savo) population is one of the late settlement populations in Finland. It was established in the 16th century (approximately 15 to 25 generations ago), mainly by people from the South Savo region (de la Chapelle & Wright, 1998).

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

The number of founders is difficult to estimate because some families left the region after the first settlement wave and new families settled down. A rough estimate of 400 to 800 effective founders is based on the numbers given in the history books of Savo. The population has been rather closed until the last century. Even today, immigration to the region is very limited, if it takes place at all.

Genotyping Assay

Genotyping of the SNP loci was performed using the Affymetrix early access Human 100K genotyping assay. The assay consists of two arrays, Xba and Hind, which denote the restriction digestion enzymes used in these assays, and yields theoretically more than 126,000 individual genotypes. A total of 250 ng of genomic DNA in reduced EDTA TE buffer (50 ng/μl) was used for each of the individual assays. The DNA was digested with either XbaI or Hind III (New England Biolabs, NEB) in the mixture of NE Buffer 2 (1x; NEB), bovine serum albumin (1x; NEB), and either Xba I or Hind III (0.5 U/μl; NEB) for 2 hours at +37°C followed by enzyme inactivation for 20 minutes at +70°C. Xba I or Hind III adapters were then ligated to the samples by adding Xba or Hind II adapter (0.25μM, Affymetrix), T4 DNA ligase buffer (1 x; NEB), and T4 DNA ligase (250 U; NEB) to the digested DNA samples. Ligation reactions were allowed to proceed for 2 hours at +16°C followed by 20 minute incubation at +70°C. Each ligated DNA sample was diluted with 75 μl of H_20 (BioWhittaker Molecular Applications/Cambrex). Diluted DNA samples were subjected to four identical 100 μl volume polymerase chain reactions (PCR) by implementing an aliquot of 10 μl of DNA sample with Pfx Amplification Buffer (1x; Invitrogen), PCR Enhancer (1x; Invitrogen), $MgSO_4$ (1 mM; Invitrogen), dNTP (300 μM, Takara), PCR primer (1 μM, Affymetrix), and Pfx Polymerase (0.05 U/μl; Invitrogen). The PCR was allowed to proceed for 3 minutes at +94°C, followed by 30 cycles of 15 seconds at +94°C, 30 seconds at +60°C, 60 seconds at +68°C, and finally, for the final extension, for 7 minutes at +68°C. The performance of the PCR was checked by standard 2% agarose gel electrophoresis in 1x TBE buffer for 1 hour at 120V. The PCR products were purified according to Affymetrix manual using MinElute 96 UF PCR Purification kit (Qiagen) by combining all four PCR products of an individual sample into the same purification reaction. The purified PCR products were eluted with 40 μl of EB buffer (Qiagen), and the yields of the products were measured at the absorbance of 260 nm. A total of 40 μg of each PCR product was then subjected to a fragmentation reaction consisting of 0.2 U/μl fragmentation reagent (Affymetrix) and 1x Fragmentation Buffer. The fragmentation reaction was allowed to proceed for 35 minutes at +37°C followed by 15 minutes incubation at +95°C for enzyme inactivation. The fragmented PCR products were then checked for completeness of fragmentation by running a 4% agarose gel electrophoresis in 1x TBE buffer (BMA Reliant precast) for 30 to 45 minutes at 120V. The fragmented PCR products were then labeled using 1x Terminal Deoxinucleotidyl Transferase (TdT) buffer (Affymetrix), GeneChip DNA Labeling Reagent (0.214 mM; Affymetrix), and TdT (1.5 U/μl; Affymetrix) for 2 hours at +37°C followed by 15 minutes at +95°C. The labeled DNA samples were combined with hybridization buffer consisting of 0.056 M MES solution (Sigma), 5% DMSO (Sigma), 2.5x Denhardt's solution (Sigma), 5.77 mM EDTA (Ambion), 0.115 mg/ml Herring Sperm DNA (Promega), 1x Oligonucleotide Control reagent (Affymetrix), 11.5 μg/ml Human Cot-1 (Invitrogen), 0.0115% Tween-20 (Pierce), and 2.69 M Tetramethyl Ammonium Chloride (Sigma). The DNA-hybridization mix was denatured for 10 minutes at +95°C, cooled on ice for 10 seconds and incubated for 2 minutes at +48°C prior to hybridization onto the GeneChip array. Hybridization was completed at +48°C for 16 to 18 hours at 60 rpm in an Affymetrix GeneChip Hybridization Oven. Following hybridization, the arrays were stained and washed in a GeneChip Fluidics Station 450 according to the recommended protocol Mapping10Kv1_450. The arrays were scanned with a GeneChip 2500 Scanner and the genotype calls for each of the SNP probes on the array were generated by using the Affymetrix Genotyping Tools (GTT) software.

Linkage Disequilibrium Analysis

Because only genotypes were available and the samples were unrelated, true haplotypes were not available except for the males on the X chromosome. Pairwise haplotype frequencies were estimated using the EM-algorithm (Weir, 1996) and these estimates were used to calculate D, D', and r^2. For all SNPs pair-wise LD was calculated for six adjacent markers.

Statistical significance was based on the likelihood ratio test (LRT) statistic with 1 *df*. The null hypothesis was no LD ($D = 0$) and the alternative hypothesis was that there is LD ($D \neq 0$). A test was considered as statistically significant if the LRT statistic was greater than 3.841 ($p < .05$).

Genomic regions with exceptionally low or high LD were determined based on 'the sliding window' approach. A 2 Mb window was moved by 100 kb steps across the chromosomes with 1.9 Mb overlapping between the windows. Average D' values were calculated for all the SNPs that were 20 to 40 kb from each other within a window and were reported for the middle point of the window. The minimum number of 20 SNP pairs for each window was required in order to calculate the average value. The average recombination rate (cM/Mb) for chromosome 5 were calculated with a 2 Mb sliding window based on the positions of the SNPs in the Affymetrix 100K early access genotyping assay (based on the NCBI human genome assembly Build 34.3).

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

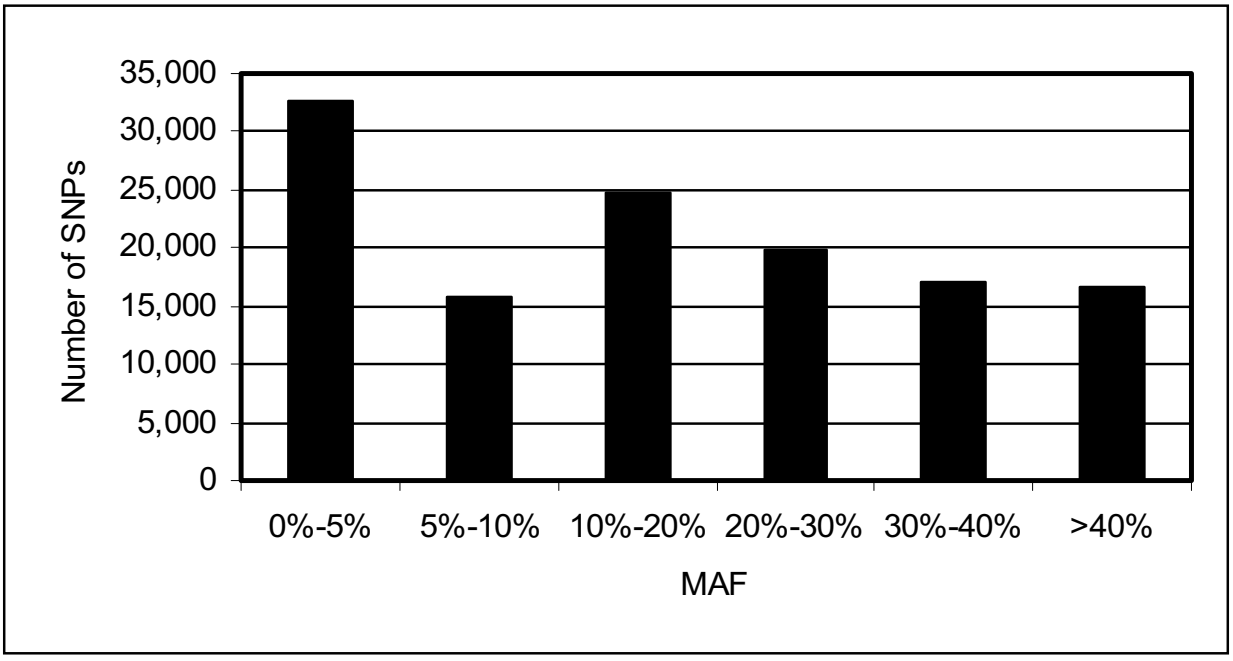

Figure 1

Distribution of the minor allele frequency (MAF) of the Affymetrix 100K genotyping assay with 126756 SNPs in 118 samples.

Results

SNP Quality Characteristics

For the statistical analysis we used only SNPs that were in Hardy–Weinberg equilibrium ($\chi^2 < 6.635$), had a call rate over 50%, and MAF over 5%. Out of the 126,757 SNPs, 1735 SNPs (1.4%) had a call rate less than 50%, whereas for most of the SNPs (69%), the call rate was over 80%. Call rate can be affected by changing the allele calling parameters in the Affymetrix Genotyping Tools (GTT) software. We selected very stringent criteria, increasing the accuracy but at the same time decreasing the call rate. The orig-

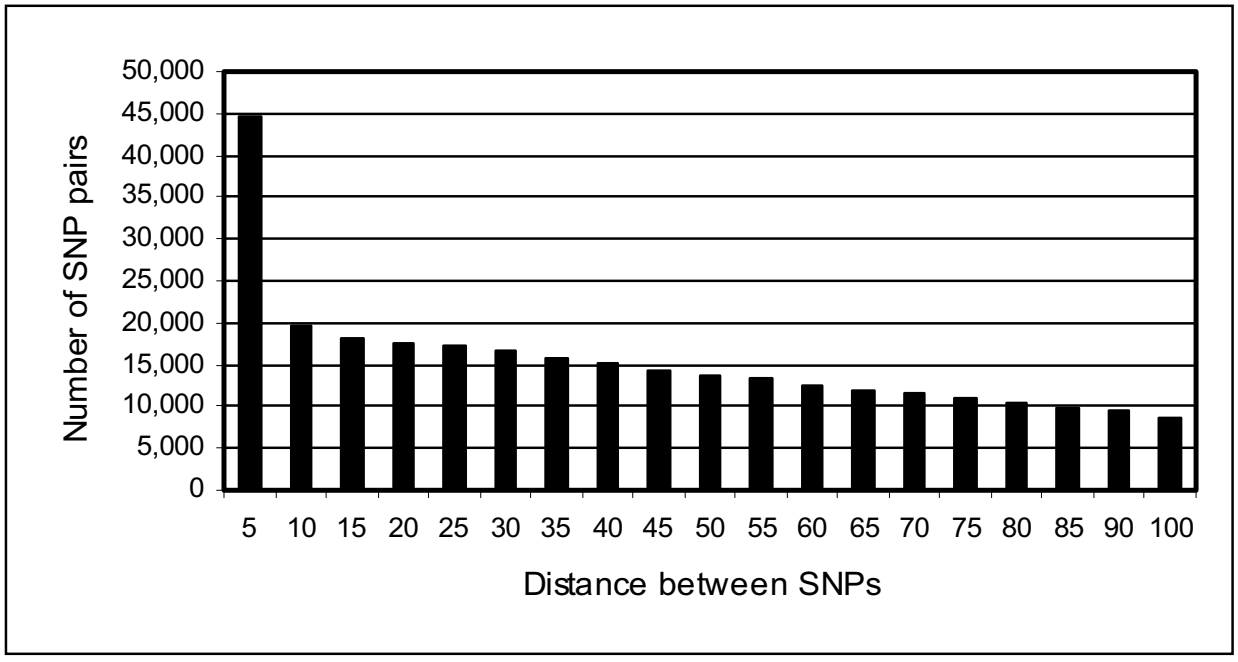

Figure 2

Distribution of the distances (kb) between the SNP pairs that fulfilled the selection criteria.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

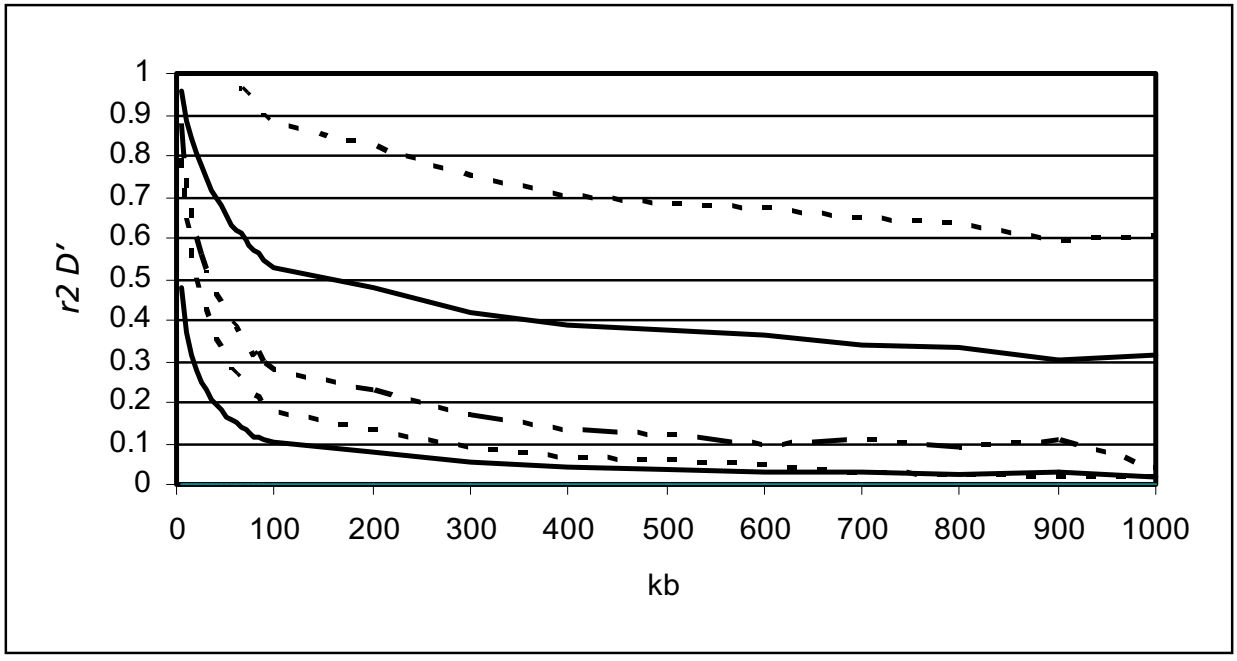

Figure 3

Average *D'* and r^2 for different distances between SNPs with a minor allele frequency of .05 (solid lines), one unit of standard deviation around the mean values are presented by dotted lines

inal data included 30,948 SNPs with MAF < .05 (around 25% of the SNPs). Twelve per cent of the SNPs were fixed in the population. The reason for the high proportion of the nonpolymorphic SNPs is the homogeneity of the study population and the layout of the test version of the Affymetrix 100K genotyping assay. Also by chance some of the low frequency SNPs did not come out in the 118 samples. Otherwise, MAF distribution was quite even (Figure 1). Ninety-five thousand eight hundred and forty-six SNPs (86% of the SNPs with MAF > 0) fulfilled the Hardy–Weinberg equilibrium requirement at the .01 significance level. Using the criteria given above, the data set included 80,639 autosomal SNPs giving 481,187 SNP pairs (some of the pairs were discarded because of the ambiguous position information) plus an additional 9366 pairs on the X chromosome. Mean MAF for the selected pairs was 0.24 (*SD* = 0.13).

Figure 2 gives the distribution of SNP pairs according to their physical distance from each other. Almost 10% (44,070 pairs) of SNPs were closer than 5 kb from each other and over 11,600 SNPs were further than 1 Mb apart from each other. The mean distance was 120 kb, the median was 69 kb, and the standard deviation 322 kb.

Overall LD Across the Genome

Decline of the *D'* and r^2 with physical distance for all SNP pairs with MAF ≥ .05 is shown in Figure 3. Decline of LD is very rapid up to 100 kb and then slower. For the distances of 5, 10, 20, 40 and 80 kb between SNPs, r^2 was .48, .37, .28, .19, and .12, respectively. Similar values have been presented for general Caucasian populations and much lower values for African American population (Ke et al., 2004; Shifman et al., 2003). The estimates are also in concordance with the values presented in review by Ardlie et al. (2002).

Effect of the Minor Allele Frequency

Because r^2 depends on the allele frequency (Guo, 1997), its effect was studied using MAF thresholds of .05, .1, .2, .3 and .4. MAF had a strong effect on r^2, especially in short distances (Figure 4a). For example, for SNPs closer than 5 kb from each other, r^2 was .48 with MAF threshold of 5% but as high as .66 for MAF ≥ 0.2. With the most heterozygous SNPs (MAF ≥ 0.4), r^2 reached .90. Similarly, the extent of useful LD (r^2 > .3) varied from 17 kb (MAF ≥ .05) to 80 kb (MAF ≥ .4) depending on MAF. This big difference can be partly explained by the fact that with an increased threshold of MAF, the number of SNP pairs with similar allele frequencies increases, thus increasing r^2. If the SNPs were grouped according to MAF in both SNPs in a pair, the average level of r^2 increased for the low MAF groups (Figure 4c). However, the group with the highest MAF had highest r^2 values, the only exception was the category of .05 < MAF ≤ .1. The latter was again due to more similar allele frequencies in this narrower MAF category (a 5% window) compared to other categories (10% windows). If a line for .1 < MAF ≤ .15 (a 5% window) had also been added into Figure 4c, it would have been above that of the lowest MAF category.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

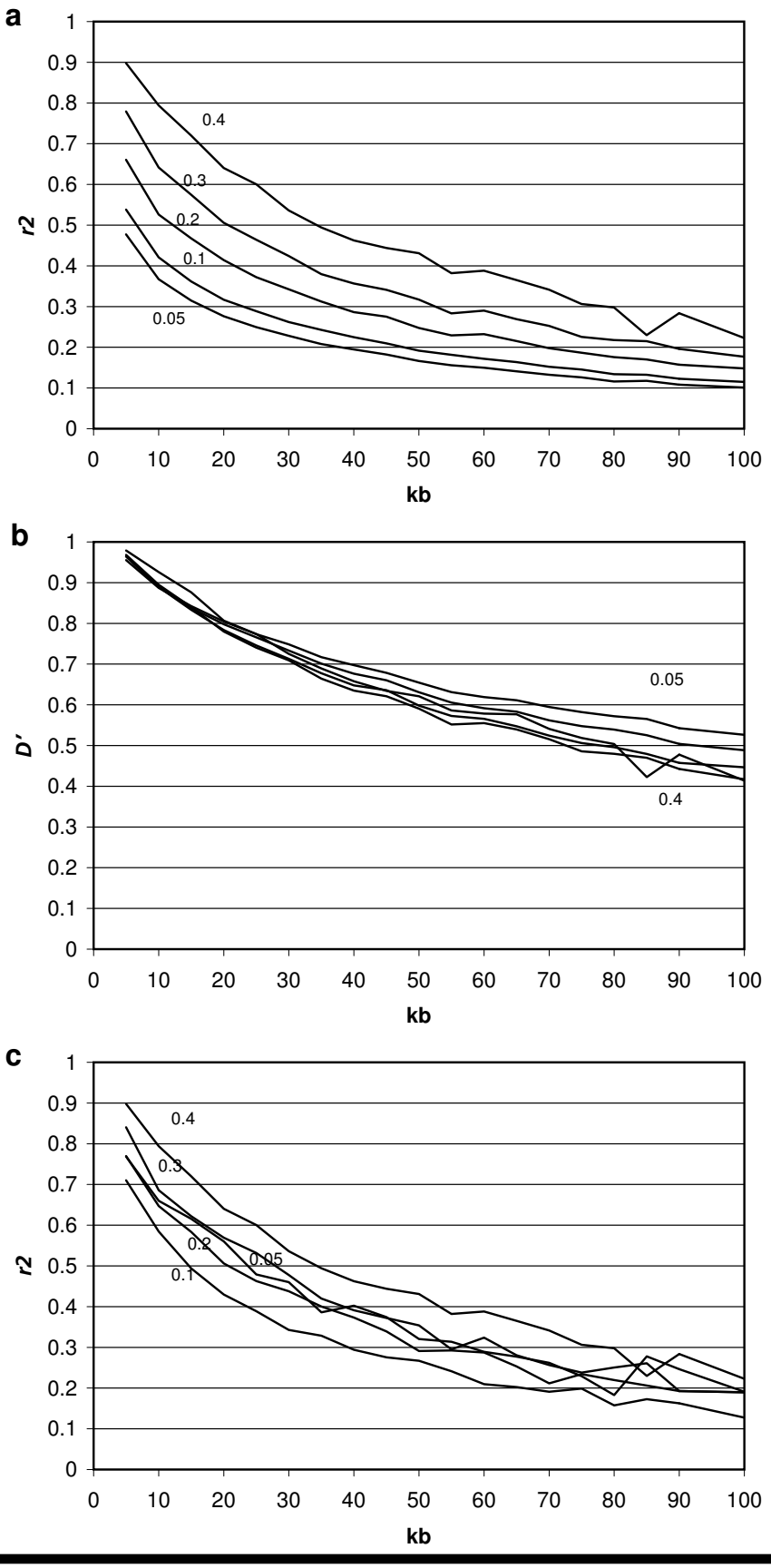

Figure 4

Effect of the minor allele frequency (MAF ≥ .05, .1, .2, .3, and .4) on r^2 (a) and D' (b) for different distances between SNPs. Mean r^2 for MAF categories (.05 ≤ MAF < .1, .1 ≤ MAF < .2, .2 ≤ MAF < .3, .3 ≤ MAF < .4, and MAF ≥ .4; c).

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

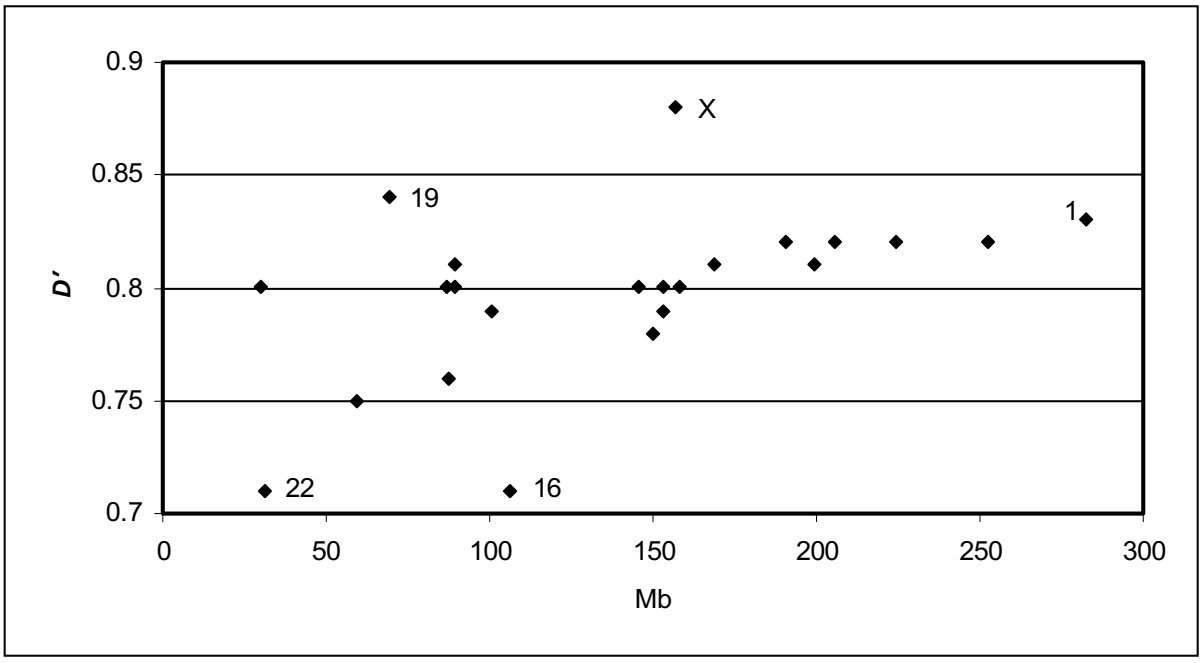

Figure 5

Average *D'* versus physical length (Mb) for different chromosomes (MAF≥0.05, 15-20kb distance between SNPs). The physical lengths are from Kong et al. (2002).

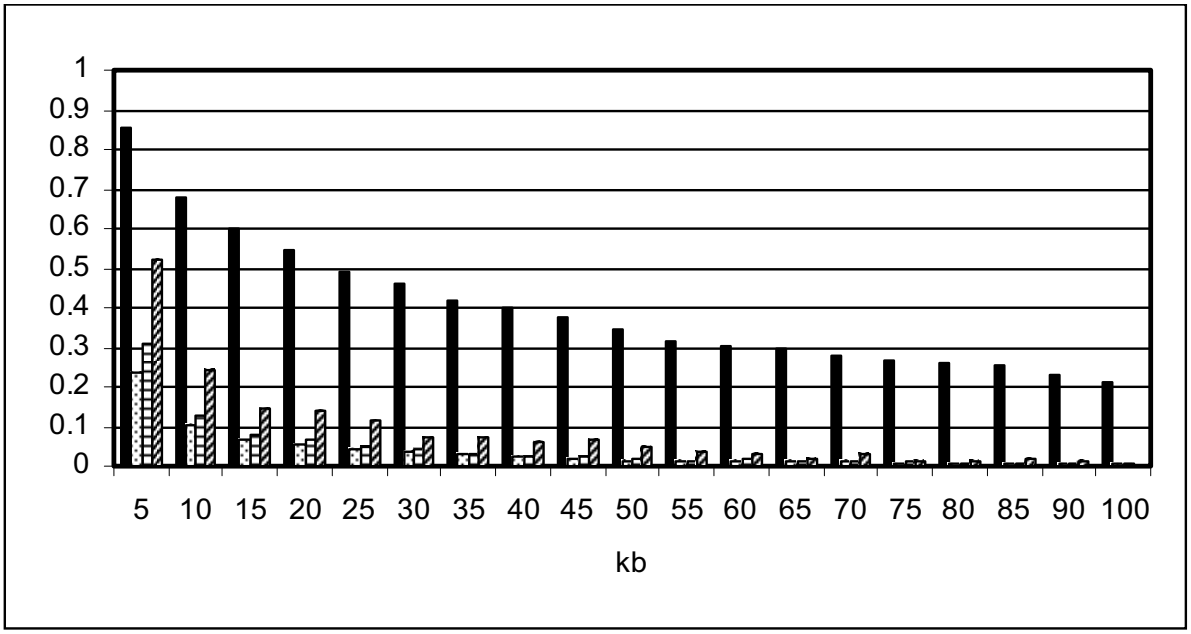

Figure 6

Proportion of SNPs in complete (*D'* = 1.0) and in absolute (r^2 = 1.0) LD for different distances between SNP pairs: *D'* = 1.0 and MAF ≥ .05 (solid black bars), r^2 = 1.0 and MAF ≥ .05 (dotted bars), r^2 = 1.0 and MAF ≥ 0.2 (horizontally striped bars), r^2 = 1.0 and MAF ≥ .4 (diagonally striped bars).

As was expected, the effect of MAF on *D'* was smaller than for r^2 (Figure 4b). Also, opposite to r^2, MAF had biggest effects on *D'* with large distances or small *D'* values. The highest *D'* was achieved with low MAF (≥ .05) and the lowest *D'* with a high MAF (≥ .4). For a distance of 100 kb between SNPs, MAF ≥ .05 resulted in *D'* of .53 and MAF ≥ .4 resulted in *D'* of .41. This can be explained by the age of the SNP, those with a high allele frequency are presumable older and showing less LD that those with a low allele frequency.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

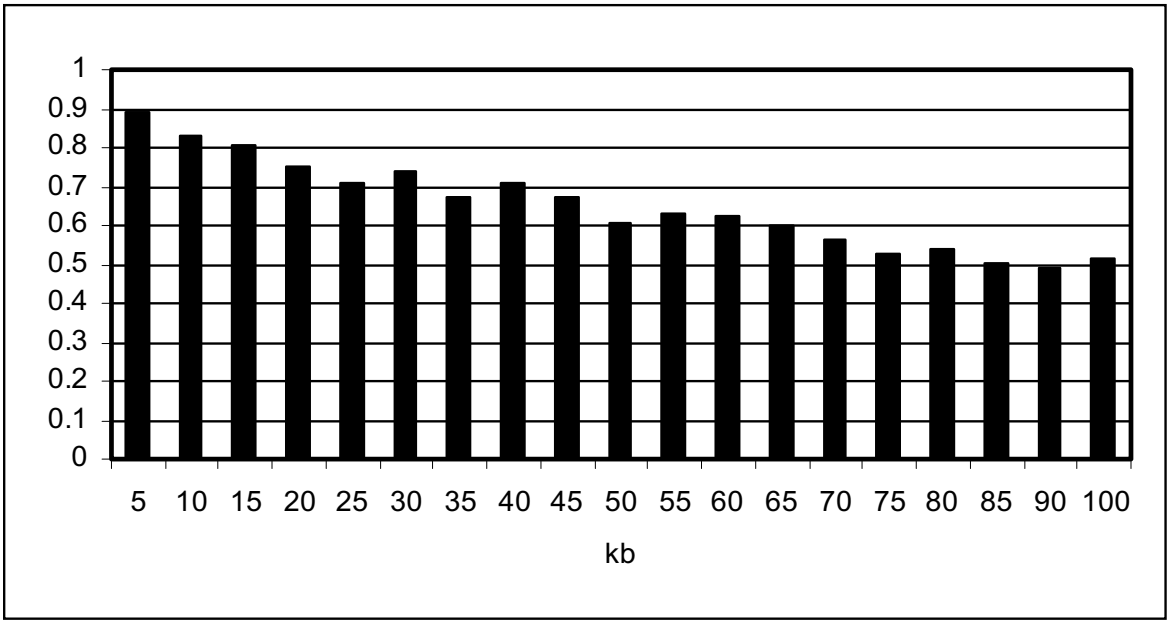

Figure 7

Proportion of SNP pairs that show statistically significant ($p < .05$) LD for different distances between SNP pairs on chromosome 1q.

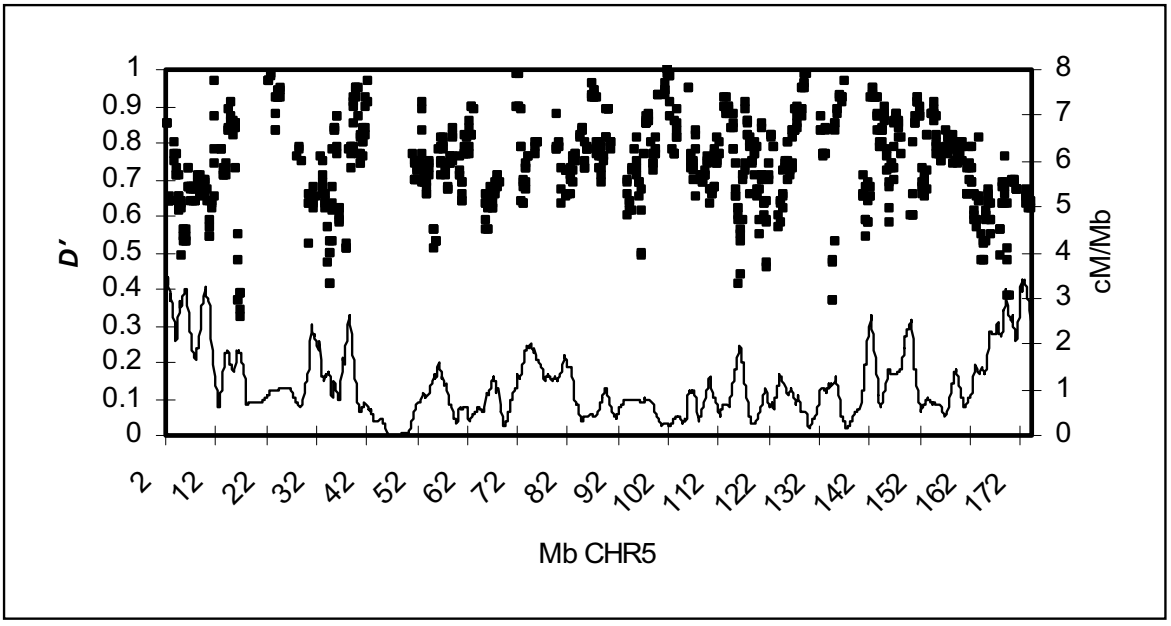

Figure 8

Average D' and recombination rates on chromosome 5 based on the sliding window approach (2 Mb window, 1.9 Mb overlapping). Only SNPs with 20 to 40 kb distance apart were used (MAF ≥ .05). D' values are shown by squares and recombination rate (cM/Mb) with a solid line. Distances are based on the NCBI human genome assembly Build 34.3.

LD in Different Autosomal Chromosomes

Autosomal chromosomes differed in their average LD levels. In Table 1 the mean *D'* and r^2 values are given for the 22 autosomes and for the X chromosome with distances of 5, 20, 40 and 80 kb between SNPs. The average LD was weaker in short chromosomes compared to long chromosomes. This corresponds to the previous observation of a higher recombination rate in short chromosomes when compared to long ones (Kong et al., 2002). For example, *D'* (r^2) on chromosome 1 for the distance of 20 kb between SNPs was 0.83 (.29) compared to 0.71 (.20) on chromosome 16. Correlation between physical length and average LD can also be seen in Figure 5 where *D'* values versus

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

Table 1

Average *D'* and r^2 for Different Chromosomes and Distances Between SNP Pairs (MAF ≥ .05)

	Distance between SNPs							
	5 kb		20 kb		40 kb		80 kb	
Chromosome	*D'*	r^2	*D'*	r^2	*D'*	r^2	*D'*	r^2
1	.96	.46	.83	.29	.71	.19	.55	.12
2	.96	.48	.82	.30	.74	.22	.61	.13
3	.96	.47	.82	.27	.69	.18	.55	.12
4	.96	.48	.82	.31	.72	.20	.59	.13
5	.95	.48	.81	.28	.70	.20	.59	.11
6	.96	.48	.82	.27	.71	.18	.59	.13
7	.95	.49	.81	.27	.71	.22	.59	.11
8	.95	.49	.80	.27	.70	.20	.54	.11
9	.94	.45	.78	.26	.67	.18	.52	.10
10	.96	.48	.80	.29	.72	.21	.60	.10
11	.96	.50	.80	.29	.69	.21	.59	.13
12	.96	.47	.79	.25	.68	.18	.57	.10
13	.96	.48	.79	.27	.68	.19	.59	.13
14	.95	.46	.80	.26	.67	.18	.61	.13
15	.95	.44	.76	.23	.64	.15	.55	.10
16	.94	.42	.71	.20	.64	.14	.48	.07
17	.97	.52	.80	.25	.66	.21	.53	.13
18	.96	.49	.81	.26	.67	.19	.57	.10
19	.97	.48	.84	.28	.70	.24	.50	.08
20	.94	.47	.75	.30	.61	.15	.46	.07
21	.97	.57	.80	.28	.70	.18	.50	.08
22	.96	.48	.71	.22	.61	.16	.52	.08
X	.97	.60	.88	.47	.77	.32	.62	.18

physical lengths in Mb (Kong et al., 2002) are plotted for all chromosomes. Interestingly, the results also revealed some variation in the extent of LD among short chromosomes. For example, LD on chromosome 19 (average *D'* 0.84, *SD* = 0.27, *N* = 46) resembles more that seen for long chromosomes (Figure 5; e.g., chromosome 1, average *D'* 0.83, *SD* = 0.29, *N* = 1316) than that of the short chromosomes (e.g., chromosome 16, average *D'* 0.71, *SD* = 0.34, *N* = 361) even though the recombination rate per physical distance has been estimated to be higher for chromosome 19 (1.58 cM/Mb) than for example for chromosome 16 (1.21 cM/Mb) or for chromosome 1 (0.96 cM/Mb; Kong et al., 2002). A simple linear regression between average *D'* values (distance 20 kb between SNP pairs) and average values of the recombination rate per physical distance (cM/Mb) on autosomes was significant (p = .024). The recombination rate per physical distance explained 23% (R^2) of the variation seen in *D'* values between chromosomes (*D'* = .865 – .054 * [cM/Mb]). Regression between r^2 and recombination rate per physical distance was not significant (p = .148) and explained only 10% of the variation seen in r^2 values.

Proportion of SNPs in Complete LD

The proportion of SNP pairs showing complete LD also declines with distance (Figure 6). Eighty-six percent of the SNPs that were closer than 5 kb from each other were in complete LD (*D'* = 1.0). If the distance is 40 to 50 kb, 34% of the SNP pairs are in complete LD. MAF had an effect on the proportion of SNP pairs that are in absolute LD (r^2 = 1.0; the effect of MAF on *D'* was not as strong as for r^2 and the results are not shown here). For the short distances (0 to 5 kb between the SNP pairs), the proportion of SNPs in absolute LD varied from 24% to 52% for MAF .05 and .40, respectively. Shifman et al. (2003) estimated that 13% of SNPs are in absolute LD in general Caucasian populations for SNP pairs separated less than 5kb. For distances longer than 10 kb the proportion varied from 7% (MAF ≥ .05) to 15% (MAF ≥ .4). The effect of MAF on r^2 was stronger with short distances than with long distances (Figure 4a).

Proportion of SNPs Showing Statistically Significant LD

Following the results presented by Varilo et al. (2003), the proportion of SNP pairs that show statistically significant LD for chromosome 1q are presented in Figure 7. The proportion of SNPs in LD was 84% for the distance of 20 kb or less between SNPs, compared with 59% presented by Varilo et al. (2003) for the late settlement region. Overall, the proportion of SNPs showing statistically significant LD was over 50% for SNPs as far as 100 kb from each other. The discrep-

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

Table 2

Genomic Regions That Show Exceptionally High LD ($D' \geq .99$)

Cytogenetic band	Length of the region (kb)[a]	Defining SNPs	cM/Mb
1p21.2	940	rs10493921–rs10493943	0.54 (56%)
1q42.13	200	rs3010192–rs4128390	0.90 (94%)
2p12	190	rs10520343–rs10496270	0.68 (67%)
2p11.2	240	rs7600315 - rs10520366	0.74 (73%)
2q32.1	120	rs1593661–rs7581009	0.17 (17%)
3p12.2	440	rs4856497–rs1406780	0.16 (16%)
5p14.3	80	rs9293044–rs1980221	1.00 (97%)
5q13.2	170	rs9293437–rs7711353	1.18 (114%)
5q21.1	300	rs1562960–rs10515332	0.27 (26%)
5q23.3	90	rs10520061–rs10491274	0.11 (11%)
6p12.3	570	rs10484405–rs10485288	0.35 (35%)
6q11.1	260	rs1577630–rs1075171	0.16 (16%)
6q16.2	100	rs330842–rs2397429	0.42 (42%)
7q31.2	110	rs10500052–rs2402002	0.36 (34%)
8q22.3	520	rs10505047–rs2340768	0.67 (66%)
8q23.3	100	rs1397371–rs320497	0.20 (20%)
9p21.1	210	rs3900585–rs2502193	0.38 (36%)

Note: Proportions of the local recombination rate per physical distance (cM/Mb) to the average for a given chromosome are given in brackets.

[a]Based on the dbSNP build 121.

ancy between our results and those presented by Varilo et al. (2003) is mainly due to the smaller data size (54 samples) used by Varilo et al. (2003).

LD on the X Chromosome

The X chromosome was analyzed separately from autosomes because of its different recombination history. True haplotypes were available from the male samples (117 males). SNPs were selected to fulfil MAF ≥ .05 and a call rate over 50%. No restriction was set for Hardy–Weinberg equilibrium because no heterozygous genotypes were expected except for the pseudo autosomal region of the Xp-ter. There were 1864 SNPs in total (196,793 called genotypes, average call rate 90%) available for the analysis. The X chromosome also gave an opportunity to estimate the error rate in genotype calling. From 196,793 genotypes 5901 were heterozygous (249 of those were in pseudo autosomal region, Xp22.3), so the error rate was approximately 3%. Most of the wrong calls were from SNPs that gave heterozygous genotypes for all the samples. Thus errors were more regular in nature than random. However, erroneous SNPs were randomly distributed across the X chromosome. Removing the SNPs that gave heterozygous genotypes for all the samples decreased the error rate to 1.2%. Erroneous genotypes were excluded from the analysis. Data for the other chromosomes also contain some genotyping errors resulting in the possible dilution of LD results.

The mean level of LD was higher on the X chromosome than on the autosomes (Figure 5 and Table 1). However, the difference was smaller for D' values than for r^2 values. D' values for the X chromosome were similar to chromosome 2 and are in accordance with results presented by Taillon-Miller et al. (2000). Depending on MAF the extent of useful LD ($r^2 > .3$) on the X chromosome was 30 kb, 70 kb and 200 kb for MAF values of .05, .2 and .4, respectively. These distances are twice as long as those observed for an average autosomal chromosome (17 kb, 35 kb, 80 kb, respectively).

Low and High LD Regions

The sliding window approach revealed several genomic regions with exceptionally low or high LD (Tables 2 and 3). Based on the SNP selection (MAF ≥ .05 and 20 to 40kb between markers) 17 regions with an average D' over .99 were observed. Similarly 11 regions with very low LD ($D' \leq .4$) were observed. For example, there was a very long region (.9 Mb) on chromosome 1 where only a few historical recombinations have occurred. This region includes the genes DBT, RTCD1, CDC14A, GPR88, VCAM1, EXTL2, SLC30A7 and CGI-30. In contrast, there is a long region on chromosome 12 (.8 Mb) with evidence of several historical recombinations or recombination hotspots. The observed number of high and low LD regions depends, of course, on the distribution of the SNPs on chromosomes. Some regions did not have enough SNPs to calculate the average D' values with the used sliding window approach. On average, 40% of the genome was covered by the sliding window approach and, for example, only a small proportion of the short chromosome was screened.

A clear tendency was observed with the recombination rate per physical distance (cM/Mb) and the mean D' in the high and low LD regions. However, some exceptions exist, for example, a region on 10q11.22 (Table 3) where low LD was observed but the recombination rate was also low based on the genetic map. This can be the result of some local recombination hotspots that are not captured in the family-based genetic maps or simply due to fact that the genetic map is an approximation between the markers that are included in the published genetic maps.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

Table 3

Genomic Regions That Show Exceptionally Low LD ($D' \leq .4$)

Cytogenetic band	Length of the region (kb)[a]	Defining SNPs	Mean D'	cM/Mb
1p21.3	180	rs10493871–rs696619	.32	0.81 (84%)
3q13.31	750	rs980944–rs9283564	.30	1.05 (108%)
5p15.1	620	rs248917–rs297193	.37	3.04 (295%)
5p15.1	190	rs10515467–rs2652072	.37	1.21 (117%)
5q35.1	100	rs876303–rs314143	.38	5.20 (505%)
8p23.1	200	rs10503381–rs10503380	.37	1.39 (137%)
9p23	880	rs1185261–rs10511593	.39	1.56 (149%)
10q11.22	770	rs477761–rs2137917	.38	0.60 (50%)
12q24.21	790	rs59336–rs10507261	.26	1.76 (157%)
13q12.12	230	rs1475488–rs1575784	.37	2.78 (217%)
14q11.2	2290	rs10498268–rs2284992	.37	4.32 (317%)

Note: Proportions of the local recombination rate per physical distance (cM/Mb) to the average for a given chromosome are given in brackets

[a]Based on the dbSNP build 121.

Dependency between LD and a local recombination rate is further illustrated for chromosome 5 (Figure 8) using a 2 Mb sliding window approach. Overall, the recombination rate is higher in the telomeres than in the centromere located around 48 Mb from the p-ter. In regions with low D' values one would expect high recombination rate (cM/Mb; Greenwood et al., 2004). Based on the linear regression (984 data points, with mean D' of .74 and mean recombination rate of 1.20 cM/Mb), the recombination rate explained 20% of the variation (R^2) seen in D' values (D' = .833 – .076 * (cM/Mb), $p < .001$, 95% CI for the regression coefficient –.086 to –.067). These values are similar to those obtained with chromosomal averages.

Discussion

In this article we have studied LD on a very large data set with a total of 481,187 SNP pairs representing all 22 autosomes from 118 unrelated individuals, and 9266 SNP pairs for the X chromosome from 117 unrelated males, using the Affymetrix 100k genotyping assay. This is by far the largest data set used for a linkage disequilibrium study. Moreover, for the first time, the current study also gives reliable estimates of LD in different chromosomes based on the same study population and indicates genomic regions with exceptionally low or high LD in extended regions. The studied samples were from the East Finland population that is a large founder population established 15 to 25 generations ago by approximately 400 to 800 individuals. The population has not been entirely closed after establishment. However, based on historical records, most of the families currently living in the region have ancestors who were living in the very same region centuries ago. In addition, most marriages took place between inhabitants of the municipality until the last 50 years or so. Thus, even though the ancestry of study participants was not studied beyond the parent level (all from East Finland), it can be expected that the study material represents the original North Savo founders well.

For short distances, the proportion of SNPs that are either in complete (D' = 1.0) or in absolute (r^2 = 1.0) LD was twice as high in the East Finland population as in a general Caucasian population reported by Shifman et al. (2003). LD in the study was also higher than presented by Kaessmann et al. (2002) because this sample was from the late settlement region whereas the Finns in the study by Kaessmann et al. (2002) represented a random sample of Finns from all over the country. For the chromosome 19q13.2 (APOE region), the D' values were far higher than those presented by Dunning et al. (2000) even though the Finnish samples were from the same region. For the distance of 20 to 40 kb, the average D' was around .8 in this study compared to .2 in that of Dunning et al. (2000). No simple explanation for such a large difference is available.

LD differed between chromosomes. As was expected, LD was higher on the X chromosome than on the autosomes (Pritchard & Preworski, 2001). However, the difference was larger in r^2 values than in D' values. Useful LD ($r^2 > 0.3$) extended over twice as long on the X chromosome than on the autosomes. For autosomal chromosomes there was a clear tendency that the longer chromosomes showed stronger LD than the shorter chromosomes. This finding has been explained by the notion that there are less recombinations for a given physical distance in long chromosomes than in short chromosomes (Kong et al., 2002). However, based on the chromosomal averages, recombination rate per physical distance (cM/Mb) explained only one quarter of the variation seen in D' values on different chromosomes and even less of the variation in r^2 values. As an example, the higher recombination rate on chromosome 19 compared to chromosome 16 did not correspond to a weaker LD on chromosome 19. On the contrary, LD was stronger on chromosome 19 than on chromosome 16. Thus, recombination rate does not explain all of the variation seen in LD between different chromosomes. More detailed evidence about variation in LD and recombination rate was shown for chromosome 5 (Figure 8). Again, recombination rate per physical distance explained only one fifth of the variation seen in D' values. However, most of the exceptionally low and high LD areas are located in regions where the recombination rate is either high or

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

low (Tables 2 and 3). The question of the characteristic behind the chromosomal variation in LD has not been addressed here in detail and requires more future research. It has been shown recently that the recombination hotspot may occur as frequently as one in every 60 to 200 kb (Crawford et al., 2004; McVean et al., 2004) and may occur either outside genes (McVean et al., 2004) or inside genes (Crawford et al., 2004).

Based on the large data set we have shown that useful LD varies from 17 kb to 80 kb in the East Finland founder population depending on the threshold used for the MAF. This is a longer region than those reported for general populations elsewhere. We have also shown that the recombination rate explains some of the variation in LD but not exhaustively. Further, we have listed some interesting long genomic regions with exceptionally low or high LD. These regions require more research if they share some common properties.

Acknowledgments

We thank Tuula Kaikkonen and Anu Ojala for skilful technical help, Christopher Devine for helpful comments, and the Research Institute of Public Health, University of Kuopio for the availability of the samples. PMV acknowledges support from the Biotechnology and Biological Sciences Research Council (UK).

References

Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S., & Kruglyak, L. (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *American Journal of Human Genetics, 69,* 582–589.

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics, 3,* 299–309.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., & Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics, 36,* 700–706.

de la Chapelle, A., & Wright, F. A. (1998). Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *Proceedings of the National Academy Science USA, 95,* 12416–12423.

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C. F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D., & Ponder, B. A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics, 67,* 1544–1554.

Greenwood, T. A., Rana, B. K., & Schork, N. J. (2004). Human haplotype block sizes are negatively correlated with recombination rates. *Genome Research, 14,* 1358–1361.

Guo, S. W. (1997). Linkage disequilibrium measures for fine-scale mapping: A comparison. *Human Heredity, 47,* 301–314.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research, 13,* 635–643.

Kaessmann, H., Zollner, S., Gustafsson, A. C., Wiebe, V., Laan, M., Lundeberg, J., Uhlen, M., & Paabo, S. (2002). Extensive linkage disequilibrium in small human populations in Eurasia. *American Journal of Human Genetics, 70,* 673–685.

Kauppi, L., Sajantila, A., & Jeffreys, A. J. (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Human Molecular Genetics, 12,* 33–40.

Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A. P., Bentley, D., Cardon, L. R., & Deloukas, P. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics, 13,* 577–588.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., & Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics, 31,* 241–247.

Laan, M., & Pääbo, S. (1997).Demographic history and linkage disequilibrium in human populations. *Nature Genetics, 17,* 435–438.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science, 304,* 581–584.

Peterson, A. C., Di Rienzo, A., Lehesjoki, A. E., de la Chapelle, A., Slatkin, M., & Freimer, N. B. (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Human Molecular Genetics, 4,* 887–894.

Pritchard, J. K., & Preworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics, 69,* 1–14.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273,* 1516–1517.

Salonen, J. T. (1988). Is there a continuing need for longitudinal epidemiologic research? The Kuopio Ischaemic Heart Disease Risk Factor Study. *Annals of Clinical Research, 20,* 46–50.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., & Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics, 12,* 771–776.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P., & Kwok, P. Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics, 25*, 324–328.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature, 426,* 789–796.

Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger, J. D., & Peltonen, L. (2003). The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers on chromosomes of Finnish populations with different histories. *Human Molecular Genetics, 12,* 51–59.

Weir, B. S. (1996). *Genetic data analysis* (Vol. 2). Sunderland, MA: Sinauer Associates.

Weiss, K. M., & Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics, 18,* 19–24.

https://doi.org/10.1375/twin.8.3.185 Published online by Cambridge University Press