Theor Appl Genet (2011) 123:11–20
DOI 10.1007/s00122-011-1562-3

ORIGINAL PAPER

Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm

Delphine Van Inghelandt · Jochen C. Reif · Baldev S. Dhillon · Pascal Flament · Albrecht E. Melchinger

Received: 8 September 2010 / Accepted: 15 February 2011 / Published online: 15 March 2011

Abstract Association mapping is based on linkage disequilibrium (LD) resulting from historical recombinations and helps understanding the genetic basis of complex traits. Many factors affect LD and, therefore, it must be determined empirically in the germplasm under investigation to examine the prospects of successful genome-wide association mapping. The objectives of our study were to (1) examine the extent of LD with simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers in 1,537 commercial maize inbred lines belonging to four heterotic pools, (2) compare the LD patterns determined by these two marker types, (3) evaluate the number of SNP markers needed to perform genome-wide association analyses, and (4) investigate temporal trends of LD. Mean values of the squared correlation coefficient ($\bar{R}$) were almost identical for unlinked, linked, and adjacent SSR marker pairs. In contrast, $\bar{R}$ values were lowest for the unlinked SNP loci and highest for the SNPs within amplicons. LD decay varied across the different heterotic pools and the individual chromosomes. The SSR markers employed in the present study are not adequate for association analysis, because of insufficient marker density for the germplasm evaluated. Based on the decay of LD in the various heterotic pools, we would need between 4,000 and 65,000 SNP markers to detect with a reasonable power associations with rather large quantitative trait loci (QTL). A much higher marker density is required to identify QTL with smaller effects. However, not only the total number of markers but also their distribution among and along the chromosomes are primordial for undertaking powerful association analyses.

Communicated by A. Charcosset.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1562-3) contains supplementary material, which is available to authorized users.

D. Van Inghelandt (✉) · B. S. Dhillon · A. E. Melchinger
Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany
e-mail: vaninghe@uni-hohenheim.de

D. Van Inghelandt · P. Flament
Limagrain Europe, Site ULICE, BP 173, 63204 Riom Cedex, France

J. C. Reif
State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany

Introduction

Association mapping (AM) and joint linkage association mapping (JLAM) are powerful complements to linkage mapping (LM) for understanding the genetic basis of complex traits (Yu et al. 2008). All three methods exploit linkage disequilibrium (LD) between genes coding for a trait and closely linked markers. LD, also known as gametic phase disequilibrium, is the non-random association between alleles at different loci.

In LM studies, LD can be calculated theoretically because these are carried out with populations constructed from bi-parental crosses in a systematic manner. In contrast, AM and JLAM make use of the LD resulting from many generations of historical recombinations in germplasm with unknown relatedness. In this case, LD is affected by many genetic factors such as mutation, recombination, selection, migration, and mating pattern (Flint-Garcia et al. 2003) and can only be determined empirically. Therefore, the germplasm under investigation must be examined with respect to the extent of LD and its genomic distribution to determine the prospects of carrying

out successful genome-wide association mapping (Rafalski 2002; Kim et al. 2007).

The relationship of LD with physical or genetic distance is highly variable across species, marker systems, and genome regions (for review see Flint-Garcia et al. 2003; Rafalski and Morgante 2004; Yu and Buckler 2006). In maize, different germplasm sets such as indigenous landraces, exotic materials, and lines from public breeding programs have been examined for LD in certain genome regions. These studies revealed a variable LD decay with a decrease of r^2 from 0.95 to 0.33 within 100 bp (Tenaillon et al. 2001) to a r^2 smaller than 0.1 within 1,500 bp (Remington et al. 2001). Yan et al. (2009) used 632 inbred lines from temperate, tropical, and subtropical public breeding programs for a whole genome LD scan. In this study, LD decay distances differed among chromosomes and ranged from 1 to 10 kb, and was much higher in temperate than tropical and subtropical germplasm. Nevertheless, knowledge of LD over large map distances is still limited particularly in commercial maize germplasm (Veyrieras et al. 2007).

A high marker density is required for a whole genome association mapping approach in germplasm with a rapid LD decay. Simple sequence repeat (SSR) markers, which are widely used for diversity studies owing to their multiallelism and the resulting high information content, may lack the density required for association studies (Ching et al. 2002). In contrast, the bi-allelic single nucleotide polymorphism (SNP) markers, which are found in much larger number in the genome than SSRs, could provide the needed high marker density. Stich et al. (2006) compared multi-allelic (SSR) and bi-allelic (amplified fragment length polymorphism; AFLP) markers to investigate LD in 72 central European maize inbred lines and concluded that SSRs should be preferred over AFLPs except in populations having a long history of recombination. SNPs, however, have advantages over AFLP markers with regard to collision (co-migration of different fragments within a single lane) and homoplasy (Gort et al. 2009). Therefore, a direct comparison of SSR and SNP markers for the measurement of LD is of interest. In a set of 102 maize inbreds representing a broad cross-section of public breeding germplasm from temperate and tropical regions, Remington et al. (2001) compared the extent of LD in six genes based on SNPs versus genome-wide distributed SSRs and observed that the SSRs revealed stronger evidence of LD than the SNPs. Nevertheless, a direct comparison of the extent of LD between genome-wide distributed SSRs and SNPs in commercial maize breeding germplasm is still lacking.

Temporal trends in LD were investigated by Reif et al. (2005) with SSR markers. They observed globally a decrease of LD over time in the most important maize hybrids grown in Germany over the past five decades. However, no information is available on SNP-based temporal trends of LD in maize germplasm developed by a commercial breeding program.

The objectives of our study were to (1) examine the extent of LD with SSRs and SNPs in 1,537 commercial maize inbred lines belonging to four heterotic pools, (2) compare the LD patterns determined by using these two marker types, (3) evaluate the number of SNP markers needed to perform genome-wide association analyses, and (4) investigate temporal trends of LD in each heterotic pool.

Materials and methods

Plant materials and molecular markers

Our investigation was based on a set of 1,537 maize inbred lines representing elite European and North American germplasm and described in details in a companion study (Van Inghelandt et al. 2010). Briefly, the inbred lines are proprietary to the plant breeding company Limagrain (France), and were assigned by breeders to four heterotic pools, namely Flint (396 inbreds), Lancaster (399 inbreds), Stiff Stalk (SSS; 377 inbreds), and Iodent (365 inbreds). The inbreds of each heterotic pool were grouped into four periods according to the year of release: Period 1 (<1990), Period 2 (1990–1995), Period 3 (1996–2000), and Period 4 (2001–2008).

The lines were analyzed with 359 SSR and 8,244 SNP markers. The SSRs (80% public and 20% proprietary) were selected over years by Limagrain with respect to their polymorphism information content (PIC) value (Botstein et al. 1980) in various sets of maize inbreds. The SNPs (100% proprietary) were discovered by sequencing 2,973 amplicons in a development set of 30 diverse maize inbreds. From these, SNPs which showed an Illumina designability score >0.4 and were not in complete LD in the development set were selected for genotyping the 1,537 lines. The proportion of missing data was 5.1% for the SSRs and 2.7% for the SNPs. On average, the amplicons had a size of 477 bp and contained three SNPs (supplementary material S1).

All markers were mapped in the IBM population (Lee et al. 2002). Chromosomes 1–10 carried 59, 42, 41, 34, 36, 31, 36, 31, 27, and 22 of the SSR markers, respectively. Further, 1,456, 858, 902, 898, 1,002, 633, 578, 632, 699, and 586 of the SNPs were mapped to chromosomes 1–10, respectively. The frequency distribution of marker-pairs distances are shown in supplementary material S2.

Distances between SNP loci pairs within an amplicon were 0. The total map length was 4,265 cM for the SSRs and 4,378 cM for the SNPs.

Genotyping of the SSRs was performed by Limagrain Verneuil Holding (Riom, France) using standard protocols. Genotyping of the SNPs was performed by Biogemma (Clermont-Ferrand, France, unpublished data) using an Illumina Infinium iSelect chip.

Statistical analyses

The square of Pearson's correlation coefficient, commonly noted by r^2 for bi-allelic loci (Hill and Robertson 1968) and by R for its extension to multi-allelic loci, was used to determine the extent of LD and calculated following Maruyama (1982):

$$R = \frac{\sum_{i=1}^{a_i}\sum_{j=1}^{b_j} D_{ij}^2}{\left(1-\sum_{i=1}^{a_i} p_i^2\right)\left(1-\sum_{j=1}^{b_j} p_j^2\right)} \quad (1)$$

were a_i and b_j are the number of alleles occurring at the two loci under consideration, p_i and p_j the observed frequencies of the ith and jth allele at these two loci, respectively, and $D_{ij} = p_{ij} - p_i p_j$ the deviation between the observed haplotype frequency p_{ij} and its expected frequency $p_i p_j$. For a matter of simplicity, we used R for both bi-allelic and multi-allelic markers.

Each loci pair was categorized as unlinked (marker loci located on different chromosomes), linked (marker loci located on the same chromosome), or adjacent (neighbor marker loci). For the SNP markers, the adjacent loci included neighbor markers within and between amplicons (for distance distribution, see supplementary material S2). In addition, we defined a further category termed "amplicon" for SNP marker pairs within amplicons. To reduce computational efforts, LD for unlinked SNP marker pairs was computed only between the markers of successive chromosomes (1 vs. 2, 2 vs. 3,…, 10 vs. 1) rather than all chromosome pairs.

Mean R values ($\overline{R}$) were calculated for all categories of loci pairs. The percentage of loci pairs with R values higher than δ ($P_{R>\delta}$) were determined for linked and adjacent pairs of SSRs and SNPs as well as for SNP pairs within amplicons. Three different thresholds (Q_{95}, 0.1, 0.8) were chosen for δ. Threshold Q_{95} refers to the 95% quantile of R values between unlinked loci and was taken as a population specific critical value of R beyond which LD was likely to be caused by genetic linkage (Breseghello and Sorrells 2006). Threshold 0.1 was the minimal R value to detect associations for rather large quantitative trait loci (QTLs) (Ersoz et al. 2007). This R threshold is based on the reasoning that detection of markers that explain <1% of the phenotypic variance requires an exponentially increasing population size. Therefore, such small effects would be considered undetectable in a population of reasonable size (300). If a marker explaining 1% of the phenotypic variation is in LD with a QTL explaining 10% of the phenotypic variation, which—for a complex trait—is considered as large, the LD correlation between the marker and the QTL is $R = 0.1$ (Zhu et al. 2008). Threshold 0.8 was chosen following the current practice for genome-wide association studies in human genetics (Barett and Cardon 2006).

Estimates of LD are influenced by the sample size. As the four heterotic pools were of similar size, no adjustment was necessary on this account. However, in order to compare estimates from individual heterotic pools to the whole set of 1,537 inbreds, a sampling procedure with sample size $n = 384$ (mean number of inbred lines in one heterotic pool) was carried out. Following Stich et al. (2005), 50 random samples were drawn and their LD measures were averaged to be used as estimates for the whole set sample (hereafter referred to as Mixed sample).

R values for pairs of linked loci were plotted against their genetic map distances for SSR and SNP markers, and curves of decay of LD with distance were fitted by nonlinear regression (NLR; Bates and Watts 1988) and locally weighted regression smoothing (Cleveland 1981). For NLR, our data were fitted to the expectation of R between adjacent sites using the model of Hill and Weir (1988):

$$E(R) = \left(\frac{10+C}{(2+C)\times(11+C)}\right) \times \left(1+\frac{(3+C)\times(12+12C+C^2)}{n\times(2+C)\times(11+C)}\right) \quad (2)$$

where n is the sample size and C the population recombination parameter (with $C = 4N_e c$; N_e being the effective population size and c the recombination fraction between the loci pair considered).

Estimates of LD decay for each chromosome in each heterotic pool were obtained by determining the abscise values of the intersection of the NLR curves with the base lines $R = Q_{95}$ or $R = 0.1$. To calculate the number of SNP markers required for association studies with regard to these two thresholds, we divided the length of each chromosome by the LD decay distance for the chromosome in each heterotic pool and summed the values over the 10 chromosomes.

To evaluate temporal trends of LD, $\overline{R}$, $P_{R>0.1}$, and $P_{R>0.8}$ were calculated for inbreds of each heterotic pool grouped in four time-periods. As the numbers of inbreds were different for each heterotic pool-period combination, we used the above described sampling strategy to draw samples of the size of the smallest group ($n = 20$).

To study the distribution of LD along the chromosomes, we calculated $\tilde{R}$ values for every given SNP locus x by averaging its R values with all other loci located within a window of 1 cM. The averaged $\tilde{R}$ values thus obtained were plotted against the map position of the locus x on the chromosome.

All analyses were performed using R software (R Development Core Team 2006).

Results

For each heterotic pool, $\overline{R}$ values were almost identical for unlinked, linked, as well as adjacent SSR marker pairs (Table 1). In contrast, $\overline{R}$ values based on SNPs differed among the four groups of loci and were lowest for the unlinked and highest for the amplicon loci. The range of the $\overline{R}$ values among the pools was narrow for both marker types, particularly the SSRs. For the Mixed sample, $\overline{R}$ based on both types of markers were within the range of those of the heterotic pools except that the value between unlinked SNP loci pairs was larger.

For the SSRs, the Q_{95} values of unlinked loci ranged between 0.043 and 0.055 for the heterotic pools and Mixed sample. For the SNPs, Q_{95} varied from 0.036 to 0.058 among the pools and were lower than for the Mixed sample (0.078). The proportion $P_{R > Q_{95}}$ of linked and adjacent SSR pairs had a narrow range 4.64–6.59% for all groups. For the SNPs, the $P_{R > Q_{95}}$ values in the heterotic pools ranged from 11.93 to 14.38% for the linked pairs, from 46.88 to 50.45% for the adjacent pairs, and from 53.67 to 61.08% for amplicon pairs. In comparison with the heterotic pools, the values for the Mixed sample were lower for linked and amplicon SNP pairs, and higher for adjacent loci pairs.

The proportion $P_{R > 0.1}$ of linked SSR marker pairs ranged between 0.16 and 0.65% in all groups, and was only slightly higher (0.37–1.43%) for the adjacent pairs. Furthermore, no SSR loci pairs had R values higher than 0.8 ($P_{R > 0.8}$). For SNP markers, $P_{R > 0.1}$ was considerably lower for linked (3.59–7.94%) than for adjacent (38.20–41.39%) and amplicon (46.40–48.67%) loci pairs in the heterotic pools. The proportions $P_{R > 0.8}$ were much lower (0.11–0.18%) for linked than for amplicon (13.24–18.44%) loci pairs, but the adjacent SNP pairs were even slightly higher (14.49–18.54%). $P_{R > 0.8}$ for the Mixed group was lower than for the different heterotic pools, and was higher for the Iodent and SSS pools than for the Flint and Lancaster pools.

There was a rapid reduction in the magnitude of R as genetic map distance between the SNP markers increased, whereas in the case of the SSR markers, R remained consistently at the same low level (Fig. 1). LD decay varied across the different heterotic pools (supplementary material S3). Across the 10 chromosomes, the LD decay distances inferred as the intersection of NLR curves with $R = 0.1$ varied from 0.11 cM for the Flint pool to 2.74 cM for the SSS pool (Table 2). LD decay varied also for the individual chromosomes. LD decay varied between 0.04 and 4.74 cM among heterotic group × chromosome combinations. A different ranking of the chromosomes was observed for LD decay in the various heterotic pools (Fig. 2). In the Flint pool, the most rapid LD decay was observed for chromosome 7 and the slowest for chromosome 10, whereas in the SSS pool the most rapid and slowest LD decay was found for chromosome 9 and 2, respectively (Table 2).

Compared to Period 1, the values of $\overline{R}$, $P_{R > 0.1}$, and $P_{R > 0.8}$ for adjacent SNP pairs in the Lancaster and SSS pools increased significantly in Period 2 and 4 after a decrease (significant in the case of SSS) from Period 2 to 3 (Table 3). In the Flint and Iodent pools, all LD measures showed a distinct decrease from Period 1 to Period 2. In the Flint pool, the measures then increased significantly from Period 2 to 3 and then from Period 3 and 4, reaching again the level of Period 1. By comparison, the measures in the Iodent pool generally did not show any significant differences among Period 2, 3, and 4.

The distribution of LD along each chromosome, presented as an example for chromosome 3 and 9 in the Flint and SSS pools, revealed different numbers and positions of $\tilde{R}$ values higher than 0.8 in the two pools (Fig. 3). The SSS pool had 13 such peaks whereas the Flint pool had three on chromosome 3. On chromosome 9, the Flint pool had three peaks and the SSS pool four at different positions.

Discussion

Several statistical parameters derived from D have been described in the literature to estimate the extent of LD. The most commonly used are R and D' (Maruyama 1982; Hedrick 1987). In this study, R was preferred for various reasons. First, the sampling properties of D' are poorly understood when $|D'| < 1.0$. In addition, R is less influenced by small sample size than D', which results in less erratic behavior when comparing loci having low allele frequencies (Flint-Garcia et al. 2003). Furthermore, R can be interpreted more readily in the context of association mapping because it relates the amount of variance explained by the marker to the amount of variance explained by the associated QTL (Zhu et al. 2008). Finally, Ducrocq et al. (2008) showed in an empirical example that, unlike D', the R plot of SNPs in the vicinity of the gene followed closely the significance levels of associations between markers and phenotype.

Table 1 Mean R values for unlinked, linked, adjacent, and amplicon loci pairs of simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers for the four heterotic pools and a sample of the whole set (Mixed). Q_{95} the 95% quantile of the R distribution for the unlinked loci pairs, $P_{R>\delta}$ the percentage of linked loci pairs with R values higher than δ, with $\delta = Q_{95}$, $\delta = 0.1$, and $\delta = 0.8$, N_{pairs} the number of loci pairs

Marker	Pool	n	R					$P_{R>\delta}$								
			Mean				Q_{95}	$\delta = Q_{95}$ (%)			$\delta = 0.1$ (%)			$\delta = 0.8$ (%)		
			Unlinked	Linked	Adjacent	Amplicon	Unlinked	Linked	Adjacent	Amplicon	Linked	Adjacent	Amplicon	Linked	Adjacent	Amplicon
SSR																
	Flint	396	0.024	0.024	0.025		0.055	4.64	4.87		0.65	1.43		0.00	0.00	
	Lancaster	399	0.022	0.022	0.022		0.049	4.71	5.44		0.30	0.57		0.00	0.00	
	SSS	377	0.021	0.021	0.022		0.047	4.83	6.59		0.33	0.57		0.00	0.00	
	Iodent	365	0.020	0.020	0.020		0.043	4.70	5.73		0.16	0.57		0.00	0.00	
	Mixed[a]	384	0.020	0.020	0.020		0.044	4.70	5.74		0.26	0.37		0.00	0.00	
SNP																
	Flint	396	0.009	0.020	0.236	0.253	0.036	12.88	50.43	61.08	3.59	38.20	46.40	0.11	14.49	13.24
	Lancaster	399	0.012	0.023	0.248	0.267	0.048	11.93	50.45	59.14	4.51	41.39	48.67	0.12	15.55	14.41
	SSS	377	0.013	0.032	0.275	0.301	0.058	14.38	46.88	53.67	7.94	41.29	48.33	0.18	18.28	18.20
	Iodent	365	0.011	0.029	0.273	0.304	0.051	13.46	47.55	55.32	7.15	40.63	47.71	0.18	18.54	18.44
	Mixed[a]	384	0.019	0.029	0.272	0.257	0.078	9.36	54.06	53.56	6.34	49.63	49.06	0.10	14.65	12.53

[a] 50 random samples were drawn from the whole set, and their results were averaged

N_{pairs} SSR: Unlinked: 57,536; Linked: 6,725; and Adjacent: 349

N_{pairs} SNP: Unlinked: 6,770,678; Linked 3,394,130; Adjacent: 8,207; and Amplicon 17,770

Comparison of the extent of LD assessed with SSR and SNP markers

The increasing $\overline{R}$ values from unlinked to linked, adjacent and then to amplicon SNP markers pairs (Table 1), and more generally, with decreasing distance between SNP loci (Fig. 1, supplementary material S4), suggested a decay of LD with distance. On the contrary, $\overline{R}$ values did not differ among adjacent, linked and unlinked SSR loci. This greater power of SNPs to examine LD was almost certainly attributable to the higher marker density of the SNPs (supplementary material S2).

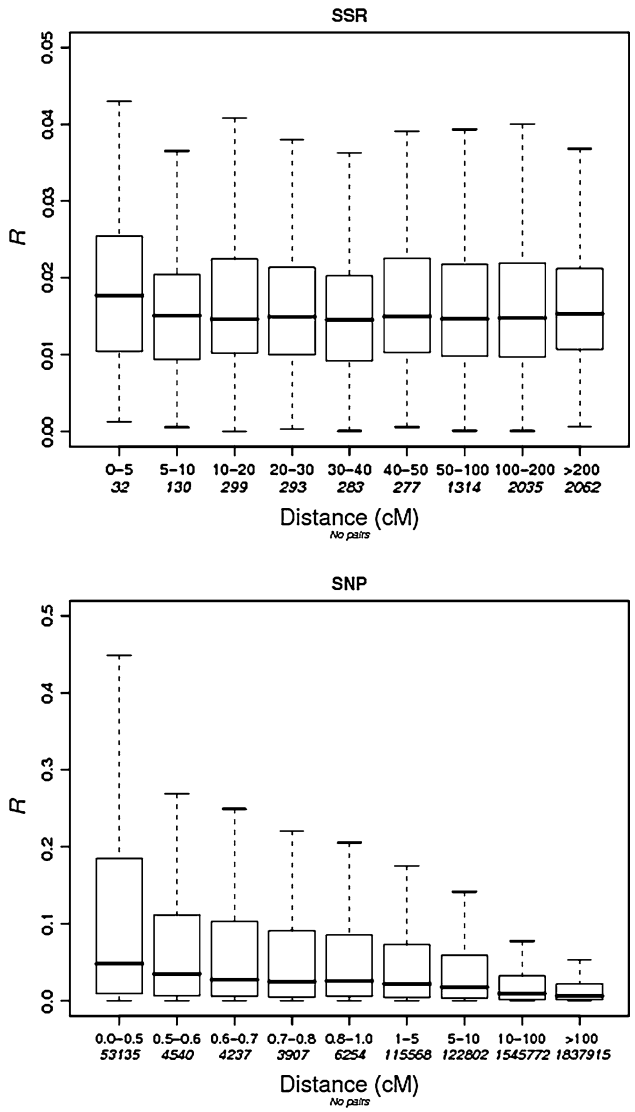

Fig. 1 Distribution of linkage disequilibrium measure *R* over different map distance classes between linked simple sequence repeat (SSR) or single nucleotide polymorphism (SNP) marker pairs over the whole set of genotypes. *No pairs* represent the number of marker pairs in each distance class. *Axes scales* are different for each plot

Furthermore, the $P_{R > Q_{95}}$ of adjacent and linked SSR (about 5%), contrarily to SNP loci pairs (up to 61%), indicated no evidence of LD caused by genetic linkage. This results in a high ratio of LD between unlinked and adjacent loci for our SSR markers, which is a major disturbing force in gene mapping. It indicates association between markers and genes located on different chromosomes and will result in a high rate of false positive detections of marker-trait association. Nevertheless, we observed for unlinked SSR markers higher $\overline{R}$ than for SNPs (Table 1). This finding might be explained on the basis that multi-allelic SSRs (the mutation rate of which is about four to five orders of magnitude higher than of SNP's), have a higher power to detect LD than bi-allelic SNPs, if marker density is ignored. This is in accordance with the study of Stich et al. (2006), which showed a clear advantage of SSRs over AFLPs to detect LD in a population with short history of recombination. Consequently, SSRs could be more powerful for association mapping if they were available in the genome with the same density as SNP markers. However, the SSR markers employed in the present study would not be adequate for association analysis, because of insufficient marker density for the germplasm evaluated. Therefore, all trends regarding LD presented below were based on SNP markers.

Influence of sample size, genetic diversity and population structure on the extent of LD

Besides providing fair comparisons of the whole set versus each individual heterotic pool, the estimates obtained for the Mixed sample with the sampling procedure enabled us to examine the influence of sample size n on the R values for the whole set. Slightly higher $\overline{R}$ values for the Mixed sample than for the whole set (0.02, data not shown) indicated a slight upward bias due to small sample size. The graphic presentation of $E(R)$ versus n based on Eq. 2 showed that R is always upwardly biased for small n, but the bias is negligible for $n > 400$, independent of N_e (supplementary material S5).

As predicted, the LD in the investigated breeding germplasm was considerably higher than previously reported in more diverse germplasm collections. Yan et al. (2009) examined 632 inbreds using 1,229 SNP markers and reported a lower amount of LD ($\overline{R} = 0.009$) than that found in our study for the whole set ($\overline{R} = 0.027$; data not shown). The higher LD in our study was most likely attributable to the marker density and the genetic diversity in the germplasm. We employed 8,244 SNP markers compared to 1,229 used by Yan et al. (2009), who examined temperate, tropical, and subtropical public inbred lines that were more diverse and contained less related materials than the elite

Table 2 Genetic map distance (LD decay) for which the trend line of the non-linear regression falls below $R = Q_{95}$ (Q_{95} of R values between unlinked loci), or below $R = 0.1$, and number of single nucleotide polymorphism (*Nb* SNP) markers required for association studies

Pool	Threshold	LD decay (cM)											*Nb* SNP
		All chr	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7	Chr 8	Chr 9	Chr 10	
Flint	$R > Q_{95}$ (0.036)	0.39	1.11	0.59	0.41	2.09	0.77	0.83	0.05	0.17	0.21	2.88	17,420
	$R > 0.1$	0.11	0.31	0.17	0.12	0.58	0.22	0.23	0.01	0.04	0.06	0.80	65,580
Lancaster	$R > Q_{95}$ (0.048)	1.01	2.32	0.49	1.96	0.80	2.72	1.15	2.10	0.31	0.14	3.94	6,750
	$R > 0.1$	0.40	0.92	0.19	0.77	0.32	1.07	0.46	0.81	0.12	0.05	1.54	17,070
SSS	$R > Q_{95}$ (0.058)	5.61	4.52	9.93	8.29	3.78	8.06	2.93	6.28	2.99	0.26	7.47	2,030
	$R > 0.1$	2.74	2.22	4.74	3.99	1.86	3.88	1.45	3.06	1.46	0.13	3.60	3,980
Iodent	$R > Q_{95}$ (0.051)	5.50	4.25	2.25	6.32	4.01	8.25	9.50	8.89	2.99	0.17	7.75	2,885
	$R > 0.1$	2.29	1.78	0.95	2.62	1.68	3.38	3.86	3.64	1.26	0.07	3.17	6,790
Whole set	$R > Q_{95}$ (0.071)	0.45	1.35	0.42	0.49	0.83	1.42	0.76	1.31	0.09	0.06	2.01	14,230
	$R > 0.1$	0.29	0.87	0.28	0.32	0.53	0.92	0.49	0.85	0.06	0.04	1.29	22,400

lines evaluated by us. The effect of diversity was also evident from the magnitude of $\overline{R}$ in the different pools. Flint and Lancaster had smaller $\overline{R}$ values than SSS and Iodent (Table 1). The SSS and Iodent pools have a narrower genetic base (Hallauer and Miranda 1988; Troyer 1999) and less genetic diversity than Flint and Lancaster (Van Inghelandt et al. 2010), which is expected to lead to higher LD, as shown by Stich et al. (2005) in a comparison of LD between related versus unrelated Flint lines.

Two opposing trends are expected when comparing the Mixed sample versus heterotic pools: the extent of LD could be higher in the Mixed sample than in the heterotic pools due to population structure (Yu and Buckler 2006), or it might be lower in the former than in the latter because of the higher genetic diversity. We observed that the extent of LD in the Mixed sample was in the range of the heterotic pools for linked, adjacent, and amplicon loci, but slightly higher for the unlinked SNP loci. This result is in accordance with the fact that population structure creates LD especially between unlinked loci (Stich et al. 2005), and, therefore, leads to false positive associations.

Temporal trends of LD

The LD between adjacent loci present in the Period 1 of the four heterotic pools reflected the influence of the relatedness on LD as discussed above. Across time periods, the Lancaster and the SSS pools behaved similarly and showed an increase of LD from Period 1 to Period 4 (Table 3). This increase might be attributable to high selection intensity for favorable epistatic combinations of alleles at different loci (Falconer and Mackay 1996). The small decrease in both pools from Period 2 to Period 3 can be explained by the successive generations of intermating carried out by Limagrain breeders within the pools since their establishment.

In the Flint pool, a strong decrease in R between adjacent loci took place earlier (between Period 1 and 2), so that there were finally no differences between Period 1 and 4. This could be explained by the introgression of new material that counterbalanced the effect of selection. Indeed, Limagrain acquired during this period several plant breeding companies with a strong base in the early flint maize germplasm. A strong decrease was also observed for the Iodent pool after Period 1. This was probably due to the use of distinct germplasm that was introgressed into the Iodent pool by Limagrain.

LD decay

Quantifying the extent and patterns of LD in germplasm of interest will facilitate the design of a SNP chip for association mapping studies. The observed association of the R values with the genetic map distances based on SNP markers suggested a decay of LD with distance (Fig. 2). This finding was expected and also reported by Remington et al. (2001) and Yan et al. (2009) for physical distances.

The LD decay was variable among the heterotic pools as shown by the distance for which the NLR curves dropped below the pool specific threshold Q_{95} and below 0.1. It was more rapid in Flint and Lancaster, the pools with higher diversity, than in SSS and Iodent, the pools with higher relatedness (Table 2). These results are in accordance with those reported by Yan et al. (2009). Further, LD decay also varied among chromosomes (Fig. 2). This reflects the known complex genome structure of maize. Based on the LD decay per chromosome calculated with threshold 0.1, we would need for the various heterotic pools between 4,000 (SSS) and 65,000 (Flint) SNP markers (Table 2) to detect associations with QTL of complex traits explaining at least 10% of the phenotypic variation in a population of 350–400 inbreds like in the present study.

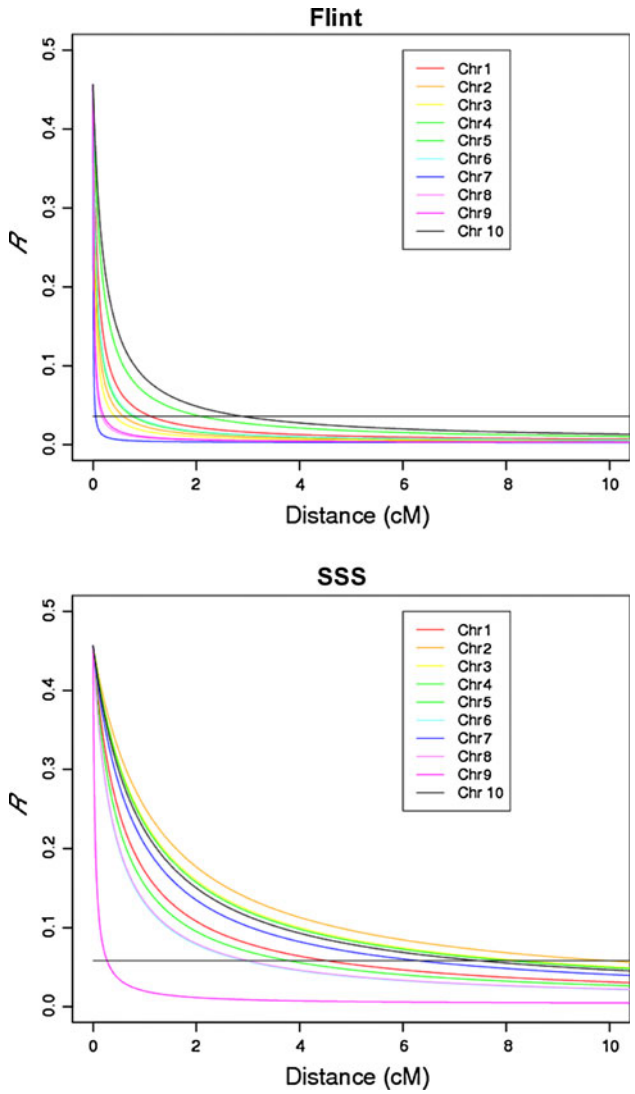

Fig. 2 Trend lines of the non-linear regression of the linkage disequilibrium measure R versus genetic map distance between single nucleotide polymorphism (SNP) marker pairs for each chromosome in the Flint and SSS pools. *Horizontal lines* indicate the Q_{95} of the distribution of R between unlinked marker pairs

However, plant breeders and geneticists are interested in explaining as much of the phenotypic variation with molecular markers as possible, and, therefore, in detecting also QTLs explaining <10% of the phenotypic variation. In a population of 300 genotypes, the power to detect a marker explaining 1% of the phenotypic variance is only about 0.2 (Yu et al. 2006; Stich and Melchinger 2009). To detect a QTL explaining 1% of the phenotypic variation, we need a marker explaining 1% of the phenotypic variation which is in LD with the QTL with a correlation higher than 0.8. The calculations with this threshold were not possible in our case, because none of the NLR showed $R > 0.8$. Nevertheless, we can assume that with this threshold we would need about one million of markers

Table 3 $\overline{R}$ values for adjacent single nucleotide polymorphism (SNP) markers for four time-periods: Period 1 (<1990), Period 2 (1990–1995), Period 3 (1996–2000), and Period 4 (2001–2008) for the four heterotic pools with pool size standardized to that of the smallest group. $P_{R>\delta}$ represents the percentage of adjacent loci pairs with R values higher than δ, with $\delta = 0.1$ and $\delta = 0.8$

Pool	Period	n	$\overline{R}$		$P_{R>\delta}$ (%)			
					$\delta = 0.1$		$\delta = 0.8$	
Flint	Period 1*	20	0.344	na	60.05	na	19.84	na
	Period 2[A]	20	0.320	a	56.29	a	17.57	a
	Period 3[A]	20	0.338	b	58.53	b	19.55	b
	Period 4[A]	20	0.344	b	59.30	c	20.01	b
Lancaster	Period 1[A]	20	0.314	a	57.90	a	16.93	a
	Period 2[A]	20	0.356	b	61.63	b	20.84	b
	Period 3[A]	20	0.349	b	61.20	b	20.38	b
	Period 4[A]	20	0.364	c	62.51	c	21.66	c
SSS	Period 1[A]	20	0.418	a	66.96	a	26.25	a
	Period 2[A]	20	0.459	c	68.93	b	30.82	c
	Period 3[A]	20	0.438	b	67.29	a	28.16	b
	Period 4[A]	20	0.457	c	68.90	b	30.09	c
Iodent	Period 1*	16	0.473	na	68.27	na	31.48	na
	Period 2[A]	20	0.441	a	66.53	a	29.54	a
	Period 3[A]	20	0.441	a	65.11	b	30.16	a
	Period 4[A]	20	0.440	a	65.66	b	29.75	a

[A] 50 random samples were drawn from the pool–period combination, and their results were averaged

[abc] represents differences at the 0.05 significance level according to a pairwise t test with Bonferroni correction

* No sampling possible

(Gore et al. 2009). Such a high number of markers is currently not available in chip format, but maybe developed soon, and next generation sequencing will probably meet this demand (Lai et al. 2010).

Genome-wide distribution of LD

In our study, the fine scale SNP coverage (average marker distance = 0.5 cM) was adequate to reveal a good picture of the structure of LD in maize at the whole genome level. To design a marker chip for association analyses, not only the variation of LD among chromosomes, but also among loci on a chromosome is important. We observed $\tilde{R}$ to vary for markers at different positions on a chromosome (Fig. 3, supplementary material S6). The marker density, though very high, was not uniform on the chromosomes, which could explain the observed differences.

The variation of $\tilde{R}$ within a chromosome did not follow the same pattern in the various heterotic pools. The non-uniform density of markers cannot explain these differences because we used the same SNPs in all pools. The variation among pools may be due to the variation in

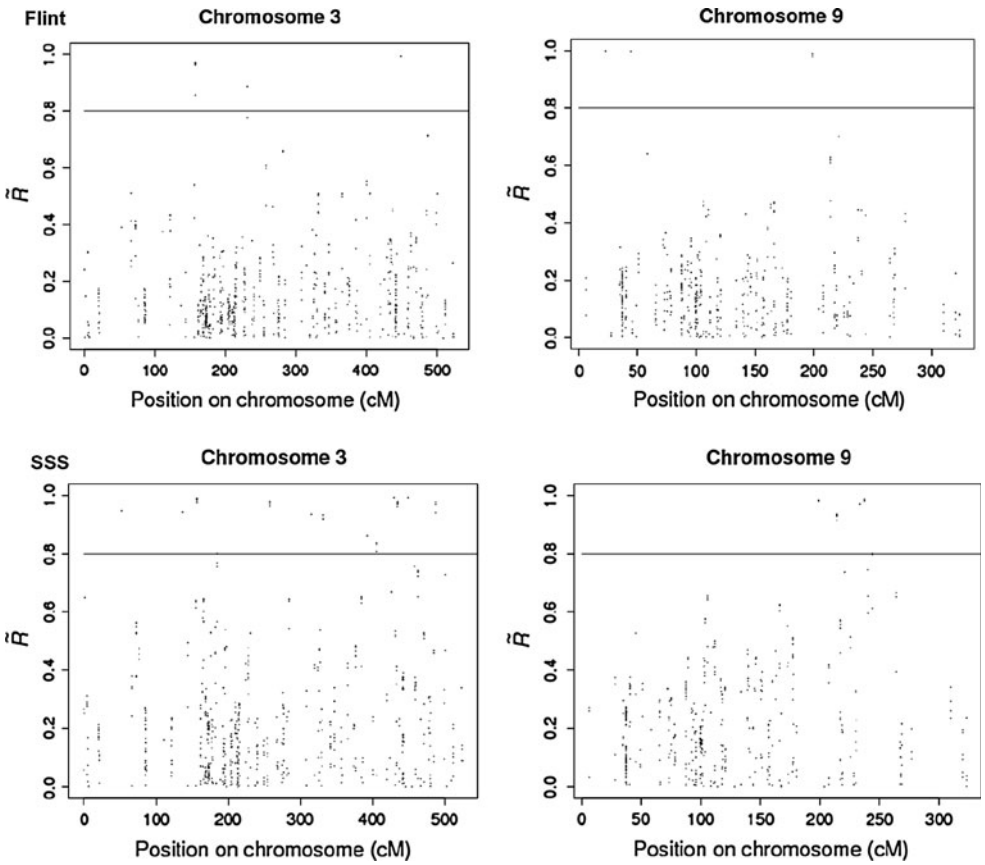

Fig. 3 Linkage disequilibrium measure ($\tilde{R}$) plotted against the position of the single nucleotide polymorphism (SNP) markers on chromosome 3 and 9 for the Flint and SSS pools. $\tilde{R}$ represents the averaged R for loci pairs within a window of 1 cM distance. For details see "Materials and methods"

recombination rate and in the history of recombination for specific chromosome regions within specific families (McMullen et al. 2009). Moreover, some markers may be monomorphic in certain chromosomal regions in one heterotic pool. Since no measure including D or any parametric representation of D is unaffected by differences in allele frequencies (Lewontin 1988), this could also have influenced the distribution of LD along the chromosomes. Consequently, our findings suggested that a universally applicable marker chip should have a marker coverage high enough to capture the fastest LD decay found in the chromosome regions of all considered heterotic pools.

Conclusions

In the present study, SNP markers of the examined density, unlike SSR markers present at a lower density, can be used effectively for association studies in commercial maize germplasm. The 60 K SNPs chip, currently available for maize, seems appropriate to identify QTL that explain at least 10% of the phenotypic variance. However, to identify QTL with smaller effect, which is a realistic situation for most traits of interest to maize breeders, a much higher marker density is required. Furthermore, not only the number of markers but also their distribution among and along the chromosomes in accordance with the extent of LD are primordial for undertaking powerful association analyses.

Acknowledgments We are indebted to Limagrain Europe for providing the genotyping data, J.P. Martinant for information about the amplicons, and S. Chauvet, S. Ducrocq, C. Lebreton, and Z. Karaman for inspiring discussions. The associate editor and two anonymous reviewers are thanked for their valuable suggestions.

References

Barett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662

Bates DM, Watts DG (1988) Nonlinear regression analysis and its applications. Wiley, New York

Botstein D, White RL, Skalnick MH, Davies RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. Am J Hum Genet 32:314–331

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:1–14

Cleveland WS (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54

Ducrocq S, Madur D, Veyrieras JB, Camus-Kulandaivelu L, Kloiber-Maitz M, Presterl T, Ouzunova M, Manicacci D, Charcosset A (2008) Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178:2433–2437

Ersoz ES, Yu J, Buckler ES (2007) Application of linkage disequilibrium and association mapping in crop plants. In: Genomics-assisted crop improvement, vol 1. Genomics Approaches and Platforms, pp 97–119

Falconer DS, Mackay TF (1996) Introduction to quantitative genetics, 4th edn. Longman Group Ltd, London

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Gore MA, Chia J-M, Elsaire LJ, Sun Q, Ersoz ES et al (2009) The first-generation haplotype map of maize. Science 326: 1115–1117

Gort G, van Hintum T, van Eeuwijk F (2009) Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation. Theor Appl Genet 119:397–416

Hallauer AR, Miranda JBF (1988) Quantitative genetics in maize breeding, 2nd edn. Iowa State University Press, Ames

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 39:1151–1155

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030

Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol Biol 48:453–461

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Maruyama T (1982) Stochastic integrals and their application to population genetics. Molecular evolution, protein polymorphism and the neutral theory. Springer, Berlin, pp 151–166

McMullen MD, Kresovich S, Sanchez Villeda H, Bradbuy P, Li H, Sun Qi, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Solvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111

Reif JC, Hamrit S, Heckenberger M, Schipprack W, Maurer HP, Bohn M, Melchinger AE (2005) Trends in genetic diversity among European maize cultivars and their parental components during the past 50 years. Theor Appl Genet 111:838–845

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci 98:11479–11484

Stich B, Melchinger AE (2009) Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics 10:94

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, van der Voort JR, Peleman J, Sørensen AP, Reif JC (2006) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol Breed 17:217–226

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphim along chromosome 1 of maize (*Zea mays* ssp. Mays L.). Proc Natl Acad Sci 98:9161–9166

Troyer AF (1999) Background of U.S. hybrid corn. Crop Sci 39:601–626

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Veyrieras JB, Camus-Kulandaivelu L, Gouesnard B, Manicacci D, Charcosset A (2007) Bridging genomics and genetic diversity: Linkage disequilibrium structure and association mapping in maize and other cereals. Crop Sci 47:60–71

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One 4:e8451

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160

Yu J, Pressoir G, Briggs WH, Bi IV, Yamsaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–2008