Theor Appl Genet (2004) 109: 681–689
DOI 10.1007/s00122-004-1695-8

ORIGINAL PAPER

Mark Jung · Ada Ching · Dinakar Bhattramakki · Maureen Dolan · Scott Tingey · Michele Morgante · Antoni Rafalski

Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm

Received: 26 January 2004 / Accepted: 2 April 2004 / Published online: 6 August 2004

Abstract Linkage disequilibrium (LD) at the *adh1* locus was examined in two sets of maize inbreds. A set of 32 was chosen to represent most of the genetic diversity in the cultivated North American elite maize breeding pool. A second set of 192 inbreds was chosen to sample more deeply the two major heterotic groups in elite maize germplasm. Analysis of several loci in the vicinity of the *adh1* gene shows that LD as measured by D' and r^2 extends greater than 500 kbp in this germplasm. The presence of this exceptionally long segment of high LD may be suggestive of selection acting on one of the genes in the vicinity of *adh1* or of a locally reduced rate of recombination.

Introduction

Direct analysis of genetic variation at the DNA sequence level at many loci has become possible in recent years due to improvements in sequencing and genotyping technology. Advances in high-throughput SNP genotyping (Shi 2001) and automation (Gut 2001) have stimulated discussion regarding association mapping and its use for discovering the genetic determinants underlying common disease in humans (Jorde 2000; Weiss and Terwilliger 2000) and traits of agronomic importance in plants (Buckler IV and Thornsberry 2002; Rafalski 2002).

In order to determine the feasibility of the proposed association studies, analysis of the linkage disequilibrium (LD) patterns around sample loci in relevant germplasm would be helpful in deciding upon the parameters used during the association experiments (Goldstein and Weale 2001).

Recent reports show LD diminishing rapidly in maize; in some cases it is less than a few hundred base pairs (Tenaillon et al. 2001). If this situation is common, then whole genome-scan association studies clearly are not practical with current genotyping technologies because of a requirement for large numbers of markers. In contrast, candidate gene-based association studies would benefit from higher resolution due to a rapid decline of LD at positions flanking the causative mutation. However, individual loci may depart from the overall pattern of LD due to selection and other effects.

Different populations, especially those subject to genetic bottlenecks, may also exhibit higher levels of LD. North American elite maize germplasm may have been subject to such effects. To address this issue, several loci were selected in our laboratory for analysis in maize, including *adh1* reported here and *y1* (Palaisa et al. 2003).

The *adh1* locus was selected due to the availability of over 160 kb of genomic sequence (Tikhonov et al. 1999). Genotypes of a number of maize lines at the *adh1* gene itself had already been generated (Ching et al. 2002). Here, we analyzed SNP genotypes at varying distances from the *adh1* gene in two collections of germplasm in order to evaluate long-distance LD (Reich et al. 2001). The two most commonly used statistics, D' (Lewontin 1964) and r^2 (Hill and Robertson 1968), were computed for this study. These were recently reviewed by Flint-Garcia et al. (2003).

Communicated by F. Salamini

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-004-1695-8.

M. Jung (✉) · A. Ching · M. Dolan · S. Tingey · M. Morgante · A. Rafalski
DuPont Crop Genetics, Experimental Station,
P.O. Box 80353 Wilmington, DE, 19880-0353, USA
e-mail: mark.t.jung@cgr.dupont.com
Tel.: +1-302-6952618
Fax: +1-302-6952726

D. Bhattramakki
Pioneer Hi-Bred International, Inc.,
7300 NW 62nd Avenue,
P.O. Box 1004 Johnston, IA, 50131-1004, USA

M. Morgante
Dipartimento di Produzione Vegetale e Tecnologie Agrarie,
Universita' di Udine,
Via delle Scienze 208,
33100 Udine, Italy

Materials and methods

Plant material

Thirty-two maize lines were selected representing >95% of the variation present as defined by SSR and RFLP markers (Ching et al. 2002). Of those, 12 lines used in a previous study (Taramino and Tingey 1996) were obtained from G. Taramino and include lines 2, 4, 8, 24, and 26–32 (Electronic Supplementary Material, Table 1). Inbred maize lines representative of American elite public and proprietary stiff-stalk synthetic (SSS) and nonstiff-stalk synthetic (NSS) germplasm were selected by O. Smith (personal communication; Electronic Supplementary Material, Table 1). Leaves from 2-week-old greenhouse-grown plants were harvested for DNA extraction and frozen at −80°C or lyophilized.

DNA extraction

Leaf material was ground with glass beads (150 μm, Sigma-Aldrich G9018) into a fine powder, using mortar and pestle in the presence of liquid nitrogen. For SNP discovery, DNA was then extracted using Plant DNAzol (Invitrogen/Life Technologies, no. 10978021), following the manufacturer's recommendation with one modification: after the initial room temperature incubation the tissue homogenate was centrifuged at 10,000 *g* for 10 min, and the supernatant was collected and used for the chloroform extraction step. For genetic mapping, DNA was isolated as described (Palaisa et al. 2003).

Locus selection, repeat masking, and gene locations

Repetitive elements were identified in the published *adh1* sequence AF123535, using the Crossmatch program (Smith 1981; P. Green, University of Washington, http://www.genome.washington.edu/UWGC/analysistools/Swat.cfm; P.Green, unpublished) and a library of known maize repetitive DNA sequences (Myers et al. 2001). Sequences matching the repetitive elements were masked prior to analysis (Fig. 1). Gene locations were defined by several methods. Annotations provided in Tikhonov et al. (1999) were first used, then FGENESH gene-finding software (Solovyev 2001) was used to predict ORFs (http://www.softberry.com/berry.phtml?topic=g-find&prg=FGENESH). Blast (Altschul et al. 1997) homology to EST sequences was also used. In two cases, we were able to use massively parallel signature sequences (MPSS, Brenner et al. 2000) to identify putative 3′ ends (Fig. 1).

Gene sequences and primer design

Fourteen DNA segments from the 32 inbred set and 13 from the 192-inbred set were polymerase chain reaction (PCR) amplified from the set of maize inbred lines. Gene-specific primer pairs were designed using the Primer3 program (Rozen and Skaletsky 2000, http://www-genome.wi.mit.edu/genome_software/other/primer3.html). For 12 of the 14 loci, two GenBank entries (M32984 and AF123535) were used to design primers. For amplicon 1 (bacm.pk14b8.f), a BAC end sequence was used. For amplicon 11 (cluster22280–1.2), an EST consensus sequence (accession no. AY111936) was used. Expected product sizes were 300–500 bp. A T3 tag (5′-AATTAACCCTCACTAAAGGG-3′) was added to the 5′ end of the forward primer, and a T7 tag (5′-GTAATACGACTCAC-TATAGGGC-3′) was similarly added to the reverse primer to facilitate direct PCR-product sequencing.

PCR amplification

DNA amplifications were performed in a 50 μl volume containing 100 ng genomic DNA, 10 pM (0.2 μM) each primer, 200 μM each dNTP, 2 mM $MgCl_2$, 5% DMSO, 0.5 U AmpliTaq Gold (Applied Biosystems, Foster City, Calif., USA, no. 4311806) and 1X PE Buffer II (Applied Biosystems). PCR reactions for genotyping were done in a 25 μl volume containing 50 ng genomic DNA, 10 pM

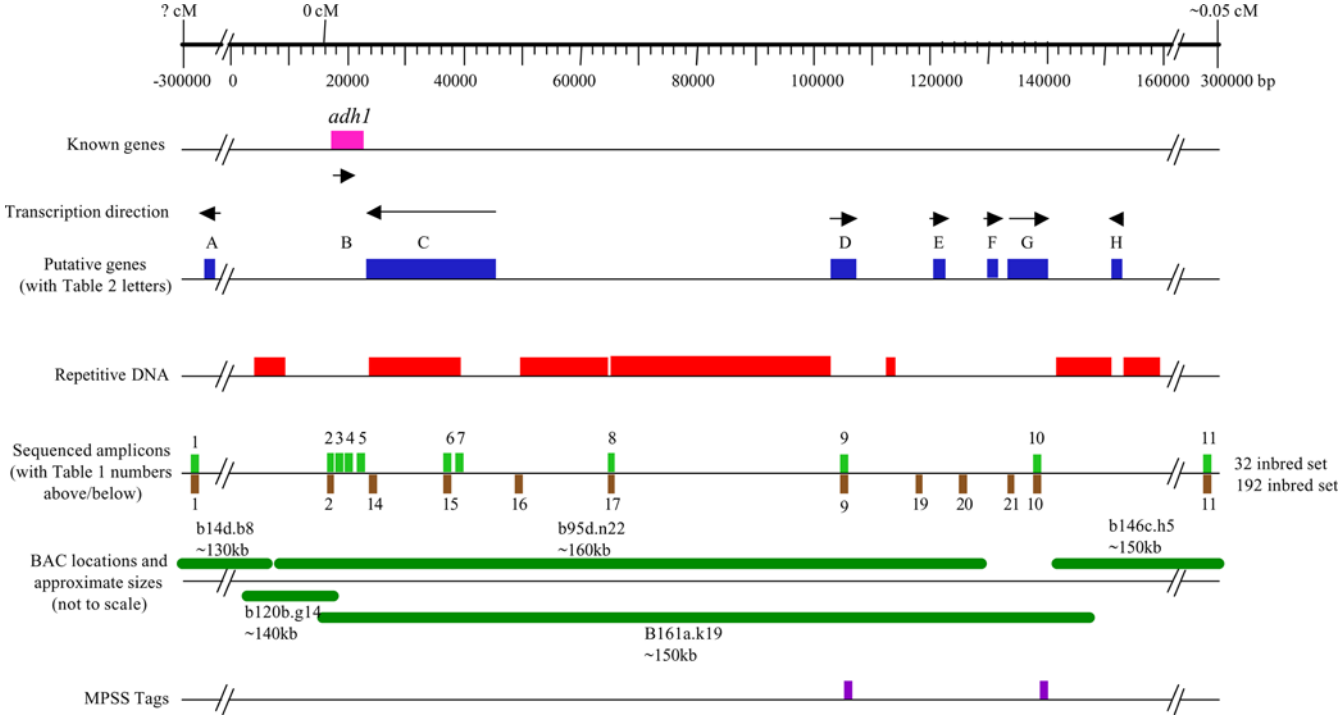

Fig. 1 Graphical representation of the *adh1* locus and surrounding features. Amplicon numbers correspond to those in Table 1. Gene designations *A–H* correspond to those used in Table 2

Table 1 List of amplicons in the *adh1* region. Numbering of amplicons is the same as in Fig. 1. The *adh1* coding sequence is located at nucleotides 17943–20930 (GenBank AF123535). *Indel* Insertion–deletion polymorphism

	Amplicon	Status	Location	Amplicon length	Contig length	Number of inbreds	Base pairs screened	SNP	Indel
Locus for 32 inbred set									
bacm.pk14b8.f	1	OK	−250,000	463	426	32	13,632	1	0
Block1	2	OK	16,415	437	431	30	12,930	7	2
PS37	3	OK	16,756	346	346	27	9,342	5	0
PS38	4	OK	17,776	360	360	32	11,520	3	1
ADH1		–	17,943	–	–	–	–	–	–
PS39–40	5	OK	20,944	834	834	32	26,688	7	7
Block2		Failed	23,110	595	–	–	–	–	–
Block3	6	OK	37,605	551	482	23	11,086	1	0
Block3.1	7	OK	38,657	541	504	30	15,120	0	–
Block4	8	OK	63,977	333	295	26	7,670	0	–
Block5		Failed	99,838	567	–	–	–	–	–
MPSS2	9	OK	104,887	571	727	32	23,264	4	4
Block6		Failed	138,021	531	–	–	–	–	–
MPSS1	10	OK	137,440	511	493	30	14,790	16	12
cluster22280–1.2	11	OK	300,000	250	294	31	9,114	10	2
Total used for analysis		–	–		4,472		132,410	54	28
Locus for 192-inbred set									
bacm.pk14b8.f above	1	OK	−250,000	463	445	188	83,660	1	0
Block1	2	OK	16,415	437	431	189	81,459	7	2
ADH1		–	17,943	–	–	–	–	–	–
LD76	14	OK	23,782	500	505	187	94,435	1	0
LD77	15	OK	37,703	509	470	176	82,720	5	0
LD78	16	OK	44,286	693	440	187	82,280	3	0
LD79	17	OK	63,993	724	380	186	70,680	2	1
MPSS2	9	OK	104,887	571	690	191	131,790	4	4
LD81		Failed	111,400	–	–	–	–	0	–
LD82	19	OK	117,833	484	475	187	88,825	3	0
LD83	20	OK	124,200	488	450	188	84,600	2	0
LD84	21	OK	130,315	722	190	190	36,100	1	0
MPSS1	10	OK	137,440	511	460	177	81,420	12	14
cluster22280–1.2	11	OK	300,000	250	275	160	44,000	9	1
Total used for analysis		–	–		5,211		961,969	50	22

Table 2 List of putative coding regions identified in the *adh1* region. Gene labels *A–H* correspond to those in Fig. 1. *E* Expectation

Gene ID[a]	Fig. 1 reference	Public EST	DuPont EST	Score[b]	*E*[b]	Direction	Location (kb)
334B7.2	A	AF124736	–	805	0	+	~70 upstream from ADH1 (not on AF123535)
334b7.3 (*adh-1*)	B	AF050457	–	1,025	0	+	16–22
334B7.4	C	AF124740	–	482	7.90E-43	−	27–43
334B7.5	D	–	cef1f.pk001.f15	140	1.80E-06	+	102–106
334B7.6	E	–	cmst1 s.pk001.e18	626	1.00E-177	+	119–121
334B7.7	F	–	cco1n.pk073.b23	176	1.00E-43	+	130
334B7.8	G	–	cco1n.pk054.o14	315	2.00E-85	+	133–138
334B7.9	H	AA979993	–	1,185	0	−	154

[a]The public ID numbers 334B7.2–334B7.9 refer to Tikhonov et al. (1999)
[b]Refers to the results of Blast analysis

each primer, 5% DMSO, and 0.1275 U HotStarTaq Master Mix (Qiagen, Valencia, Calif., USA, no. 203445). The reactions were incubated using a Perkin Elmer 9700 thermocycler in MicroAmp Optical 96-well Reaction Plates and MicroAmp Full Plate Covers (Applied Biosystems, nos. N801-0560 and N801-0550) with the following cycling conditions: 95°C for 10 min; ten cycles of 1 min at 94°C, 1 min at 55°C, 1 min at 72°C; 35 cycles of 30 s at 95°C, 1 min at 68°C; followed by a final extension of 7 min at 72°C. PCR products were analyzed for DNA quality and quantity on 1% agarose gels (Invitrogen/Life Technologies, no. 10975035), and treated with Exo/SAP-it (USB, Cleveland, Ohio, USA, no. 78200) prior to sequencing.

A panel of ten oat–maize addition lines was used to screen primer pairs to verify that they are specific to chromosome 1, on which the *adh1* gene is located. The panel consists of ten oat–maize addition lines, each containing an individual maize chromosome (Ananiev et al. 1997). The panel also includes positive and negative controls. The positive control is the maize donor line S60, and the negative control is the oat acceptor line. Primer pairs were accepted if products were amplified as single bands from the chromosome 1 line and the maize donor line positive control (but not the negative control), as well as from B73 and Mo17.

DNA sequencing

PCR products were sequenced directly using T3 and/or T7 primers. Sequencing reactions were performed using one-quarter strength ABI PRISM Big Dye (version 2.0 or 3.0) Terminator Cycle Sequencing Ready Reaction kits with AmpliTaq FS DNA polymerase (Applied Biosystems, no. 4390236) and analyzed on ABI 377 or ABI3700 DNA analyzers. Base calls and quality scores were assigned by Phred, and the sequences were assembled with Phrap and viewed/edited in Consed (http://www.phrap.org/ and http://depts.washington.edu/ventures/uwtech/license/express/ppccombo.htm- license) or Sequencher (http://www.genecodes.com/). Polymorphic positions were identified by inspection in Consed, tagged, and catalogued in a Sybase (version 11.5.1) database (http://www.sybase.com/home). Base calls with a Phred quality value below 15 were not considered. GenBank accession numbers are listed in Supplementary Table 4.

BAC contig analysis

The DuPont–Pioneer maize BAC physical map is composed of 125,266 BAC clones clustered into contigs and was made from two libraries that contain *Hin*dIII or *Eco*RI partially digested genomic DNA from the inbred Mo17 (M. Morgante, unpublished data). To generate more data for LD calculations at larger distances from *adh1* (i.e., beyond the AF123535 sequence), a Mo17 BAC contig containing the *adh1* gene was selected based on oligonucleotide hybridization (Gardiner et al. 2004). The presence of the *adh1* locus was confirmed by PCR amplification, using block1 (amplicon 2) and mpss1 (amplicon 10) (Table 1). BACs in the contig were isolated as follows: Inoculation was from glycerol stocks into 2XYT (12.5 μg/ml chloramphenicol); growth was for 16H 37°C, 250 rpm. DNA was isolated using an Autogen 740 Automatic DNA Isolation System (http://www.autogen.com). BAC insert sizes were estimated by *Not*I (New England Biolabs, no. R0189S) digestion of ~400 ng DNA and electrophoresis on a 1% agarose (0.5× TBE) pulse-field gel, using a low-range PFG marker (NEB no. N0350S). Approximate physical distances (base pairs) are shown in Fig. 1. Two loci were selected from the *adh1* BAC contig: bacm.pk14b8.f (GenBank accession number see Supplementary Table 4), a BAC end sequence approximately 300 kb upstream of *adh1* (amplicon 1), and cluster22280–1 (GenBank AY111936), an EST-homologous region approximately 300 kbp downstream (amplicon 11). Distances to these two loci from *adh1* were estimated from the BAC sizes above (Fig. 1). Amplification results were also used to approximate BAC locations within the contig and their sizes (Fig. 1).

Population structure

Structure software (version 1.0, http://pritch.bsd.uchicago.edu/software/structure.html) was used to define population structure in the 192-inbred data set, using molecular marker data for 185 SSRs (data not shown, Pritchard et al. 2000).

Data analysis

Polymorphism data was extracted using a UNIX extraction tool (M. K. Hanafey, personal communication). Data was converted into an Access (Microsoft, Redman, Wash., USA http://www.microsoft.com/office/access/default.asp) database, using SAS (Cary, N.C., USA http://www.sas.com/) routines (S. Wall, personal communication) for analysis in TASSEL (http://www.maizegenetics.net/bioinformatics/index.htm) or a NEXUS text file for analysis in DNAsp (version 3.53, http://www.ub.es/dnasp, Rozas and Rozas 1999). In general, insertions–deletion polymorphisms (indels) were excluded from the analysis (Tenaillon et al. 2002). If indels were included, each indel was treated as a single mutational event. Nucleotide diversity π and θ are reported on a per-site basis. Confidence interval of θ was determined by coalescent simulation without recombination. To evaluate distribution of allele frequencies in the population, DNAsp was used to calculate Tajima's *D* statistic (Tajima 1989a).

Genetic mapping

Three amplicons were selected for mapping: bacm.pk14b8.f (amplicon 1), block1 (amplicon 2), and cluster22280–1.2 (amplicon 11). Sequence data were inspected for polymorphisms between B73 and MO17, the inbred parents used to generate the recombinant inbred mapping (IBM) population (http://www.maizemap.org/index.htm). This population is derived from a cross between B73 and MO17, followed by four generations of sib mating and selfing to F_8 (Lee et al. 2002). When possible, indels were used so that scores could be generated by genotyping on 3% Metaphor (Cambrex, Rockland, Minn., USA no. 50181) agarose gels (amplicon 11). If no indels were available, a SNP was scored by direct sequencing (amplicons 1 and 2). MAPMAKER/EXP, version 3.0b (http://www-genome.wi.mit.edu/genome_software/other/mapmaker.html), was used to map the three selected loci independently.

LD analysis

Calculation of LD measures D′ and r^2 and plotting of LD values (Figs. 2, 3) was done using TASSEL software (E.S. Buckler IV, http://www.maizegenetics.net/bioinformatics/tasselindex.htm. Rare SNPs, with allele frequency <0.1 were excluded from LD analysis.

Results

To characterize patterns of DNA-sequence polymorphisms around the *adh1* locus, primers were designed for use on the panel of 32 maize inbreds (Electronic Supplementary Material, Table 1) at 14 separate loci (Fig. 1; Table 1) in a ~600-kbp window around *adh1* and at three locations within the transcription unit (5′ UTR, 3′ coding, and 3′ UTR). In a separate experiment, 13 loci were analyzed on the panel of 192 maize inbreds at similar locations (Electronic Supplementary Material, Table 1; Fig. 1). All sequenced amplicons are numbered and given positions

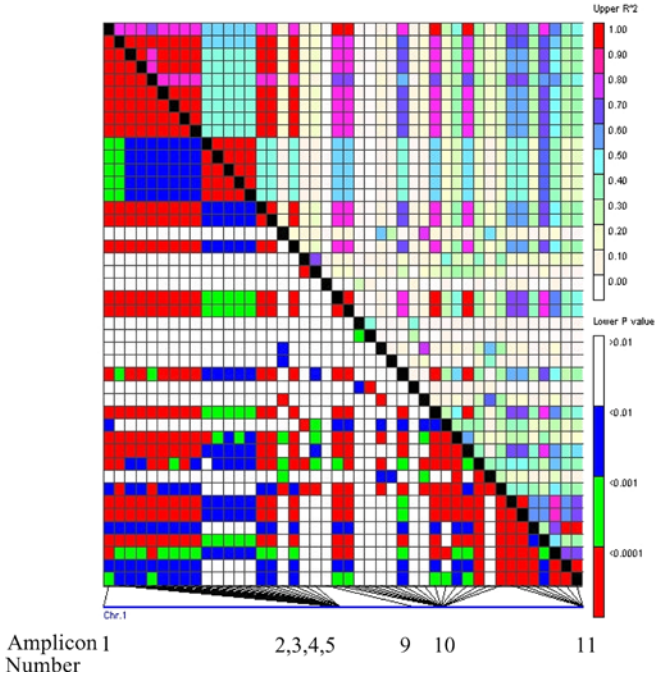

Fig. 2 Linkage disequilibrium (LD) measure (r^2, *above diagonal line*) and probability value (*P*, *below diagonal line*) for the 32 inbred dataset. The picture represents all pairwise comparisons of identified polymorphic loci. The physical map locations of SNPs and the corresponding amplicon numbers (Table 1) are shown at the *bottom* of the graph

relative to the first nucleotide of Af123535 for easier reference (Table 1). Five loci were common to the two sets (amplicons 1, 2, and 9–11). For LD calculations only, a 2-bp indel at the 3′ end of amplicon 2 was selected as a central point for reference; positions of specific SNPs in Figs. 2 and 3a–f are calculated in reference to this point as "0 bp" (see also full-size Supplementary Figs. 2, 3, available at http://bioinf.dimi.uniud.it; Electronic Supplementary Material, Table 3).

Amplification of single loci at equal intervals proved impossible due to the presence of large retrotransposon blocks containing repetitive DNA (Fig. 1). Therefore, single-copy sequences and putative 3′ UTRs (Table 2) were selected for analysis. Putative genes were defined by Blast homology to public ESTs (Tikhonov et al. 1999). After masking repeats, about 104 kbp of the 160-kbp AF123535 sequence was found to be repetitive (65%). Of the 14 loci initially selected for amplification, 11 yielded acceptable PCR products on the 32-inbred set. Amplification from the panel of oat–maize addition lines confirmed chromosome 1 origin of these amplicons. While not every locus yielded complete sequencing data for every inbred, 30 of the 32 inbreds have at least 66.7% complete genotypes. Two of the amplicons contained no variants, and two contained one SNP. In the remaining amplicons, there were 82 sequence variants (Table 1), with 54 nucleotide substitutions and 28 indels ranging from 1 bp to 193 bp in length. Summing up the total consensus length of the analyzed amplicons and dividing by the number of sequence variants identified in this germplasm set results in one SNP per 83 bp and one indel per 160 bp. Of the SNPs, there were 34 transitions and 20 transversions yielding a transitions/transversions ratio of 1.7:1.

Of the 13 loci (Table 1; Fig. 1) selected from the 192-inbred dataset (Electronic Supplementary Material, Table 1), 12 were successfully amplified from chromosome 1; 167 of the 192 inbreds are represented in more than 66.7% of the genotypes. Three amplicons contained only one SNP. In the others, there were 72 variants total (Table 1), with 50 nucleotide substitutions and 22 indels ranging from 1 bp to 193 bp in length. Of the SNPs, there were 34 transitions and 16 transversions yielding a transitions/transversions ratio of 2:1. The ratio of the consensus length of analyzed amplicons to the number of sequence variants observed is 1 SNP per 104 bp and 1 indel per 236 bp.

The overall level of sequence polymorphism in the amplicons sequenced was lower than previously observed. The frequency of SNPs reported here is 1 in 83–104 bp in contrast to 1 in 61 (Ching et al. 2002) and 1 in 28 bp (Tenaillon et al. 2001). A frequency of 1 in 186 bp for indels was found in contrast to 1 in 126 found previously (Ching et al. 2002).

To calculate genetic diversity of the 32 inbred dataset, DNA sequences of all amplified loci were concatenated. Presence of some missing data for some of the inbred lines necessitated their exclusion due to the limitations of the analysis software. Therefore 15 of the 32 sequences were used in the final analysis (sequences 3, 7, 10, 12–14, 18–21, 26, 28, and 30–32), and a subset of the loci was used in the calculations: 14 polymorphic positions (25.9% of total detected) and 1,337 nonpolymorphic positions (25.6% of all analyzed). There was no bias resulting from the selection of sequence data containing no missing data points, as indicated by identical fraction of included polymorphic and nonpolymorphic sites (25.9%, 25.6%). However, selection of 15 of the 32 inbreds for final analysis on the basis of completeness of data set may have introduced a downward bias on diversity estimates. The average π was 1.42×10^{-3} per site (Table 1). θ was estimated to be 2.51×10^{-3} per site (95% CI 0.93×10^{-3}~ 5.0×10^{-3}). Both of these values are lower than the ranges previously reported of 1.49×10^{-2} to 2.1×10^{-2} for θ (Gaut and Clegg 1993; White and Doebley 1999) and 1.36×10^{-2} for π (White and Doebley 1999), possibly indicating selection in the region. Tajima's *D* statistic was calculated to be −1.823 ($P<0.05$), consistent with excess of rare alleles, which may occur during recovery from a selective sweep (Tajima 1989b). Diversity measures π and θ were not calculated in the 192-inbred data set due to the presence of at least one missing data point at each of the SNP positions.

The B73 × Mo17 recombinant IBM population was used in an attempt to determine genetic distances around *adh1*. Segregation of the following polymorphisms was scored: a 13-bp indel in amplicon 2; several indels in amplicon 11, adding to a total of 24-bp difference between the B73 and Mo17 alleles and a single SNP in amplicon 1. A total of three recombinants were found in 279

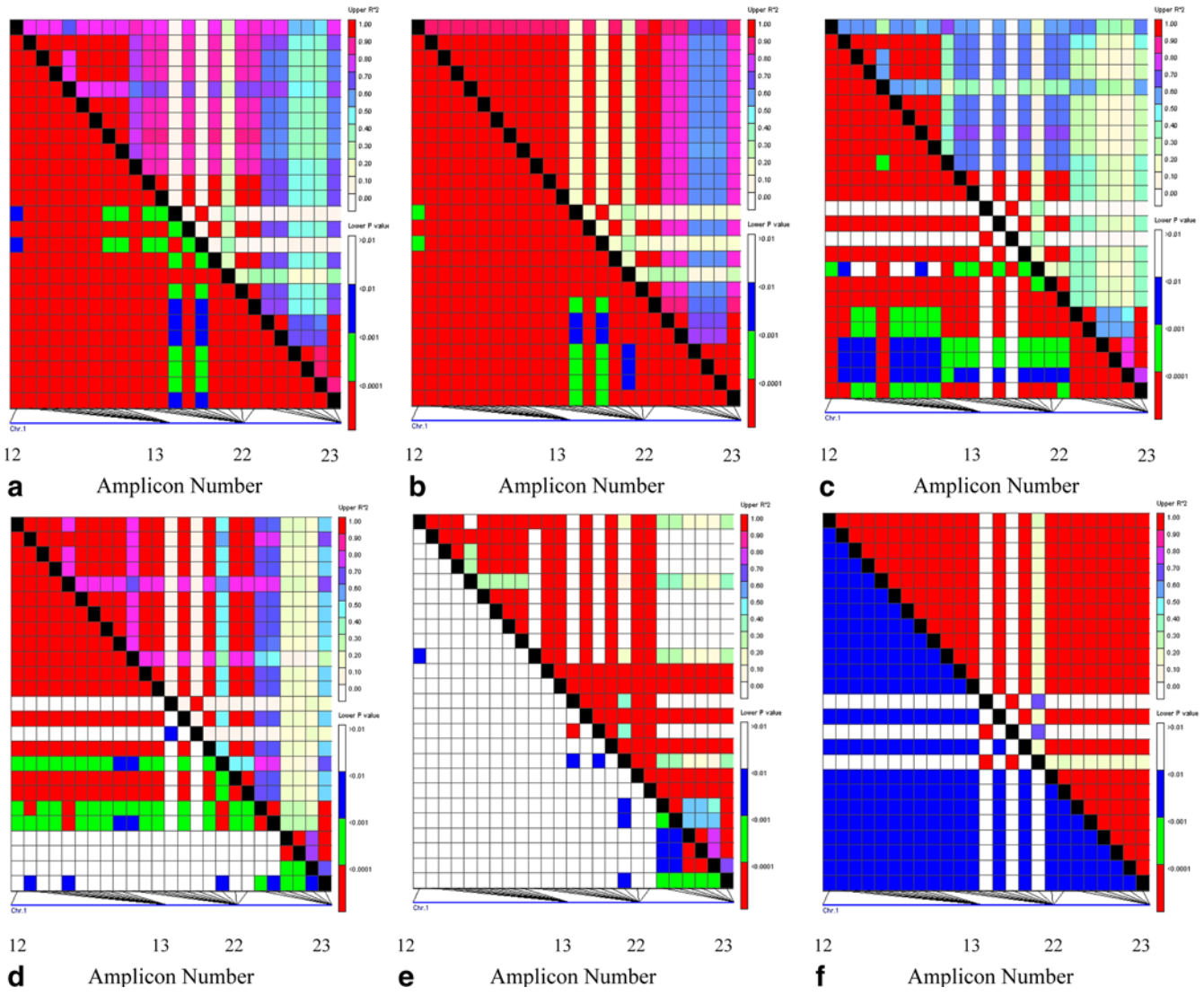

Fig. 3a–f r^2 and P for the 192-inbred dataset. Each graph represents all pairwise comparisons at each locus for r^2 (*above diagonal line*) and P (*below diagonal line*), using the set of individuals indicated. Physical map positions and corresponding amplicon numbers (Table 1) are shown at the *bottom* of the graph. **a** 167 Nonstiff-stalk synthetic (NSS) and stiff-stalk synthetic (SSS) individuals, **b** 82 SSS individuals, **c** 85 NSS individuals, **d** 26 Structure group 1 NSS inbreds, **e** 38 Structure group 2 NSS inbreds, and **f** 34 Structure group 3 NSS inbreds. Structure group 4 (not shown) includes a subset (57) of the 82 SSS lines of **b**

genotyped individuals. Two of these individuals carried apparent double recombinants in amplicons 1, 2, and 11. These genotypes were found to be correct after rescoring. The apparent double recombinants, which most likely result from the resolution of recombination intermediate without the exchange of flanking markers (noncrossover events), have been previously observed in maize (Dooner 2002). They are more likely to be identified in among the recombinants alleles that are not highly divergent in sequence, as is the case here (Dooner 2002). The single observed reciprocal event is insufficient to obtain accurate map distance. The average genetic-to-physical-distance ratio for maize is ~1.2 cM/Mbp (3,000 cM/2,500 Mbp); so, taking into account ca. fourfold map expansion in the IBM population (Lee et al. 2002), 2.4% recombination would be expected in the ca. 600-kb distance between the markers used here (block1 and cluster 22280_1.2, amplicons 1 and 11), which is not significantly different from the observed value of 0.36% at P=0.05.

LD measure r^2 is around 0.4 between the two end points of the locus, amplicons 1 and 11, in the 32-inbred set (Fig. 2). LD declines slowly as a function of distance over the entire ~600-kbp distance, between amplicons 1 and 11, in the 192 inbred population (Fig. 3a). Significance levels were at P<0.0001 for most loci in the 192-inbred set and less significant (P>0.01) for the 32-inbred set, reflecting the number of observations. The D' statistic gives similar results in the region (Fig. 4a, b). This result is striking considering reports of LD in maize to date (Remington et al. 2001; Tenaillon et al. 2001). When the individuals in the 192 dataset are separated into SSS and NSS groups, some differences in levels of LD are seen with higher LD observed in SSS than in NSS groups (Fig. 3b, c).

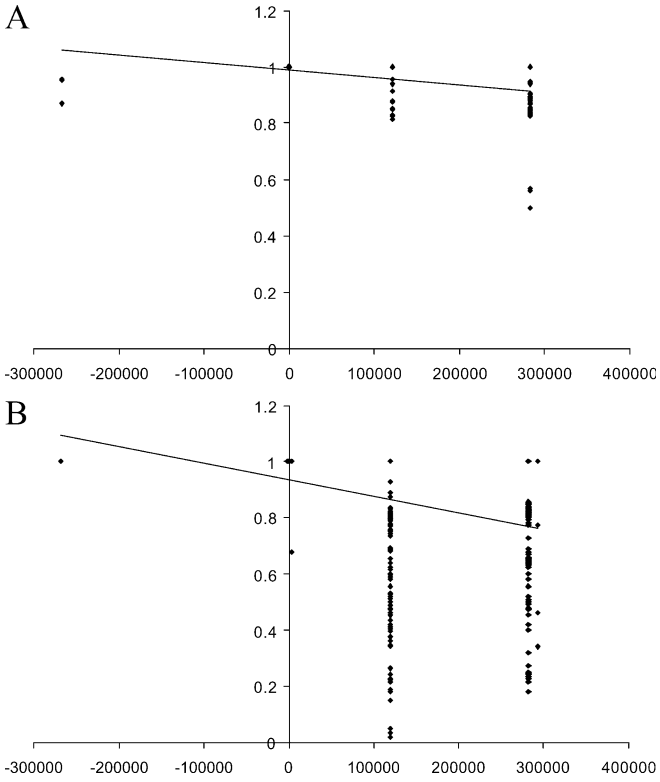

Fig. 4 Plot of LD measure D' (*yaxis*) versus distance (in base pairs, *x axis*) for the 192-inbred dataset (**a**) and the 32 inbred dataset (**b**). Each graph represents D' for all pairs of loci for the set of individuals indicated. Linear regression trend line is shown

We used the method of Pritchard et al. (2000) to estimate possible population structure effects in the 192-inbred dataset. The analysis was completed using 186 SSR markers and 157 inbreds for which SSR data were available. All but two of the inbreds were grouped (Electronic Supplementary Material, Table 1). The Structure analysis was consistent with four population subgroups of 26, 38, 34, and 57 individuals. Group 1 consists of 12 NSS and 14 SSS individuals. Group 2 includes 37 NSS and 1 SSS inbred line. Group 3 has 29 NSS and 5 SSS individuals, and group 4 is composed of 57 SSS inbred lines. These four new population groups were used for a new LD analysis in TASSEL (Fig. 3d, f). If population structure is largely responsible for the large blocks of LD, there should be less LD within the separate populations. Overall, the levels of LD in the three subpopulations (Fig. 3d, f) are more differentiated than those found in the completed data set (Fig. 3a). The NSS group 2 (Fig. 3e) shows a rapid decline in LD between amplicon 11 and amplicons 1 and 2, with very low overall significance values, while group 3 (Fig. 3f) shows strong and significant LD across the whole region.

As a control, we computed LD between the *adh1* locus and two unlinked loci on chromosome 6 for the 192-inbred dataset (data not shown). Out of 275 pairwise comparisons, only one showed significant LD at the $P<0.01$ level. In contrast, comparisons done within the *adh1* locus (Fig. 3a) all show significant LD at the $P<0.01$ level.

Discussion

DNA sequence diversity, recombination, and LD have been the subject of several recent studies in maize (Tenaillon et al. 2001, 2002); a rapid decrease of LD with distance was found. At most loci, LD decreased to undetectable levels within several hundred bases, and no regions of extended LD were detected despite some heterogeneity in the rate of LD decline with distance (Tenaillon et al. 2001). In a different set of germplasm, we have recently observed LD across distances of several kilobase pairs at the *adh1* locus of maize (Ching et al. 2002), and this prompted us to examine patterns of diversity at this locus across much larger distances.

We found significant LD extending beyond 600 kbp in a germplasm collection that provides an excellent representation of maize elite lines, but not of landraces and ancestral populations. Several scenarios may explain the large region of LD around *adh1*. First, it may be due in part to the relatively narrow genetic base of germplasm used. Many current elite maize inbred lines have some common ancestry in relatively recent history. This hypothesis is supported by the fact that LD breaks down more quickly in the more diverse set of 32 inbreds, as well as in the NSS set versus the more homogeneous SSS set population admixture, which may also increase apparent LD. Analysis of LD in subpopulations identified using Structure (Pritchard et al. 2000) reveals reduced LD in NSS group 2 (Fig. 3e). Significant LD is detected between amplicons 1, 2, and 10, but not between these and amplicon 11. In NSS group 1 (Fig. 3d), LD is detected again between amplicons 1, 2, 10, and some loci in amplicon 11. In contrast, NSS group 3 (Fig. 3f) shows LD across the whole 600-kbp region. Some of the effect in NSS group 2 is likely due to a reduced ability to detect significant LD in smaller datasets with lower number of informative SNPs. We conclude that the subpopulation-specific effects can be detected in the three NSS-derived germplasm groups identified by Structure (Fig. 3d–f). The LD observed in the NSS group of germplasm (Fig. 3c) is an aggregate of effects seen in smaller groups, with some admixture effects.

Second, selection acting on *adh1* or any gene in the vicinity would be expected to reduce diversity with an increase in LD in the entire region due to selective sweep, assuming that the region is not highly recombinogenic. The footprint of selection would vary in size, depending upon selection intensity, time since sweep, and recombination rate. This selection effect has been demonstrated at the *y1* locus (Palaisa et al. 2003). Selection has not been found at *adh1* in the past (Tenaillon et al. 2001); however, the methodology used by these authors may not detect selection acting upon noncoding segments of the genes or more recent selection by breeders in the elite gene pool.

Measurable reduction in genetic diversity would be expected to accompany recent strong selection, and the amount of DNA sequence diversity (excluding indels) in *adh1* is significantly lower ($P<0.0001$ for θ) in the elite germplasm than the average in maize (Tenaillon et al. 2001). The full complement of genes in this region, which extends beyond the sequenced segment of 160 kbp (Tikhonov et al. 1999; Table 2), cannot be unambiguously defined at present. However, there is evidence for a disease-resistance QTL (Lubberstedt 1998) in the region and indirect evidence for quality and agronomic traits (Araki 1999), which supports selection occurring in the region. Tajima's *D* statistic indicates increased frequency of rare alleles, consistent with recovery from a selective sweep.

Third, the history of recombination in this region could have a significant effect on LD patterns. The *adh1* locus appears to lie in the region of relatively high recombination on a macro scale (Tenaillon et al. 2002). We were unable to measure recombination accurately in the mapping population available. It has been postulated that recombination in maize occurs primarily in genes (Fu and Dooner 2002a, b) and may be reduced in regions rich in indels (Dooner and Martinez-Ferez 1997). There is evidence for at least seven genes (Table 2) in the region in addition to *adh1* and the incidence of repetitive DNA (65%) is average for maize (Myers et al. 2001).

Since no large-scale multilocus investigations of long-range LD have been reported in maize elite germplasm to date, it is difficult to say whether or not regions of long-range LD are common. In humans, widely variable LD structure, with regions of high LD punctuated by recombinogenic hot spots, has been reported (Bonnen et al. 2002; Rafalski and Morgante 2004; Reich et al. 2001). Regions of high LD appear to coexist in maize with regions of low LD, and dependence on the choice of experimental populations is very strong, making broad generalizations difficult.

Acknowledgements We thank Steve Wall, Mike Hanafey, Stan Luck, Romeo Hubner, Nancy Caraher, Howie Smith, David Meyer, Marianna Faller, and Michael Gore for helpful advice and analysis support. We also thank Ed Buckler help with data analysis and comments on the manuscript.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ananiev EV, et al (1997) Oat–maize chromosome addition lines: a new system for mapping the maize genome. Proc Natl Acad Sci USA 94:3524–3529

Araki E (1999) Identification of genetic loci affecting amylose content and agronomic traits on chromosome 4A of wheat. Theor Appl Genet 98:977–984

Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res 12:1846–1853

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634

Buckler ES IV, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107–111

Ching A, Caldwell K, Jung M, Dolan M, Smith O, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Dooner HK (2002) Extensive interallelic polymorphisms drive meiotic recombination into a crossover pathway. Plant Cell 14:1173–1183

Dooner HK, Martinez-Ferez IM (1997) Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell 9:1633–1646

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

Fu H, Dooner HK (2002a) Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci USA 99:9573–9578

Fu H, Dooner HK (2002b) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99:1082–1087

Gardiner J, Schroeder S, Polacco ML, Sanchez-Villeda H, Fang Z, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, Chou H, Wing R, Soderlund C, Coe EH Jr (2004) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. Plant Physiol 134:1317–1326

Gaut BS, Clegg MT (1993) Molecular evolution of the *adh1* locus in the genus *Zea*. Proc Natl Acad Sci USA 90:5095–5099

Goldstein DB, Weale ME (2001) Population genomics: linkage disequilibrium holds the key. Curr Biol 11:R576-R579

Gut IG (2001) Automation in genotyping single nucleotide polymorphisms. Hum Mutat 17:475–492

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

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435–1444

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 R (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

Lubberstedt T (1998) Comparative quantitative trait loci mapping of partial resistance to *Puccinia sorghi* across four populations of European flint maize. Phytopathology 88:1324–1329

Myers B, Tingey S, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795–1806

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

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

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

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

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa

Shi MM (2001) Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. Clin Chem 47:164–172

Smith TFAW (1981) Identification of common molecular subsequecnes. J Mol Biol 147:195–197

Solovyev VV (2001) Statistical approaches in Eukaryotic gene prediction. In: Balding D, et al (eds) Handbook of statistical genetics. Wiley, New York, pp 83–127

Tajima F (1989a) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123:229–240

Tajima F (1989b) The effect of change in population size on DNA polymorphisms. Genetics 123:585–595

Taramino G, Tingey S (1996) Simple sequence repeats for germplasm analysis and mapping in maize. Genome 39:277–287

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

Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS (2002) Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp.*mays* L.). Genetics 162:1401–1413

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Weiss K, Terwilliger J (2000) How many diseases does it take to map a gene with SNPs? Nat Genet 26:151–157

White SE, Doebley JF (1999) The molecular evolution of *terminal ear 1*, a regulatory gene in the genus *Zea*. Genetics 153:1455–1462