Mol Breeding (2009) 23:23–33
DOI 10.1007/s11032-008-9210-6

Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars

Shiaoman Chao · Wenjun Zhang · Eduard Akhunov · Jamie Sherman · Yaqin Ma · Ming-Cheng Luo · Jorge Dubcovsky

Received: 17 January 2008 / Accepted: 9 July 2008 / Published online: 29 July 2008

Abstract In this study, we developed 359 detection primers for single nucleotide polymorphisms (SNPs) previously discovered within intron sequences of wheat genes and used them to evaluate SNP polymorphism in common wheat (*Triticum aestivum* L.). These SNPs showed an average polymorphism information content (PIC) of 0.18 among 20 US elite wheat cultivars, representing seven market classes. This value increased to 0.23 when SNPs were preselected for polymorphisms among a diverse set of 13 hexaploid wheat accessions (excluding synthetic wheats) used in the wheat SNP discovery project (http://wheat.pw.usda.gov/SNP). PIC values for SNP markers in the D genome were approximately half of those for the A and B genomes. D genome SNPs also showed a larger PIC reduction relative to the other genomes ($P < 0.05$) when US cultivars were compared with the more diverse set of 13 wheat accessions. Within those accessions, D genome SNPs show a higher proportion of alleles with low minor allele frequencies (<0.125) than found in the other two genomes. These data suggest that the reduction of PIC values in the D genome was caused by differential loss of low frequency alleles during the population size bottleneck that accompanied the development of modern commercial cultivars. Additional SNP discovery efforts targeted to the D genome in elite wheat germplasm will likely be required to offset the lower diversity of this genome. With increasing SNP discovery projects and the development of high-throughput SNP assay technologies, it is anticipated that SNP markers will play an increasingly important role in wheat genetics and breeding applications.

Keywords EST · SSR · SNP · Wheat

Abbreviations

EST	Expressed sequence tag
FP	Fluorescence polarization
HRS	Hard red spring
HWS	Hard white spring
HRW	Hard red winter
HWW	Hard White Winter

Electronic supplementary material The online version of this article (doi:10.1007/s11032-008-9210-6) contains supplementary material, which is available to authorized users.

S. Chao (✉)
USDA-ARS Biosciences Research Lab, 1605 Albrecht Blvd., Fargo, ND 58105-5674, USA
e-mail: shiaoman.chao@ars.usda.gov

W. Zhang · Y. Ma · M.-C. Luo · J. Dubcovsky
Department of Plant Sciences, University of California, Davis, Davis, CA 95616, USA

E. Akhunov
Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA

J. Sherman
Department of Plant and Soil Sciences, Montana State University, Bozeman, MT 59717, USA

PIC	Polymorphism information content
SSR	Simple sequence repeat
SNP	Single nucleotide polymorphism
SRW	Soft red winter
SWS	Soft white spring
SWW	Soft white winter

Introduction

The discovery of a large number of single nucleotide polymorphisms (SNPs) in humans has revealed the power of this technology to generate high-resolution genetic maps (Brookes 1999). Though SNPs are generally biallelic and thus often less informative than multi-allelic simple sequence repeats (SSRs), their sheer abundance makes possible the development of high density SNP genetic maps, providing the foundation for subsequent population-based genetic analysis (Rafalski 2002). The utilization of multi-SNP haplotypes can offset the relatively low information content of single SNP loci (Brumfield et al. 2003). Another advantage of the SNP markers is that they do not depend on estimates of fragment lengths, a requirement of SSR markers that has limited the standardization of such data across different laboratories and equipment.

The large amount of plant sequencing data available in public databases represents a rich resource for SNP discovery using bioinformatics approaches. In rice, for example, thousands of candidate polymorphisms were identified by comparing the draft genome sequences from *indica* and *japonica* subspecies (Feltus et al. 2004). For other plant species, extensive expressed sequence tag (EST) collections provide an alternative source for the *in silico* detection of SNPs (referred to hereafter as eSNPs). Indeed, the alignment of ESTs from different cultivars of maize (*Zea mays* L.) has been useful to develop eSNPs for the functional part of the maize genome (Batley et al. 2003). In barley, too, over 3,000 EST-derived eSNPs have been identified, genetically mapped, and subsequently used to assess genetic diversity and the extent of genome-wide linkage disequilibrium (Rostoks et al. 2006; Hayes and Szucs 2006; Tim Close, personal communication). Genome-wide maps comprised of large numbers of SNP markers have also been reported in *Arabidopsis thaliana* (Cho et al. 1999), rice (*Oryza sativa* L.) (Nasu et al. 2002), and soybean (*Glycine max* L.) (Choi et al. 2007).

SNP densities in plants vary widely, as revealed by several SNP discovery studies. In out-crossing species, some average SNP densities include: 1 SNP/47 bp in non-coding regions of 36 inbred lines of maize (Ching et al. 2002); 1 SNP/73 bp in regions corresponding to 592 unigenes of 12 inbred maize lines (Vroh Bi et al. 2006); and 1 SNP/72 bp in 315 EST-derived loci of a panel of 13 lines of sugar beet (*Beta vulgaris* L.) (Schneider et al. 2007). The average SNP densities among self-pollinated species, however, tend to be lower: an estimated 1 SNP/270 bp in non-coding and random genomic regions of 25 soybean cultivars (Zhu et al. 2003) and 1 SNP/200 bp in 870 unigene-derived genomic regions of eight diverse accessions of barley (Rostoks et al. 2005). In wheat, one study comparing the sequences of 21 genes across 26 diverse germplasm accessions revealed an average of 1 SNP/330 bp in genic regions (Ravel et al. 2006), while a different study using a smaller and less diverse germplasm sample (12 genotypes) discovered an average of 1 eSNP/540 bp of wheat EST regions (Somers et al. 2003).

Further large-scale eSNP discovery in wheat is limited by both the polyploid nature of the organism and the high sequence similarity found among the three homoeologous wheat genomes (Somers et al. 2003). In an effort to complement the eSNP discovery strategy, a wheat SNP discovery project (http://wheat.pw.usda.gov/SNP) was funded by the National Science Foundation (NSF) to discover SNPs within intronic regions of wheat genes. The strategy used by this group of researchers was to sequence homoeologous gene regions from the three wheat genomes, develop genome-specific primers to amplify intronic regions, and use those primers to screen a diverse collection of 13 *Triticum aestivum* accessions originated from different parts of the world and eight synthetic wheats. Previously, a similar strategy of developing genome-specific primers based on intron sequences was used successfully to detect SNPs within starch biosynthesis genes in wheat (Blake et al. 2004).

In the current study, we used the genome-specific primers and SNP information generated by the NSF project to develop 359 new SNP-detection primers. Using the template-directed dye-terminator incorporation assay with fluorescence polarization detection

(FP-TDI) (Chen et al. 1999), we assessed the level of SNP polymorphism across 20 US adapted wheat genotypes representing seven market classes. The genetic diversity of these US cultivars was then compared to that of the same 13 diverse accessions used in the wheat SNP discovery project as a means of investigating the broad effects of modern breeding selection on wheat genome diversity.

Materials and methods

Plant materials

Twenty elite wheat (*Triticum aestivum* L.) cultivars and advanced breeding lines representing seven US wheat market classes were used in this study (Table 1). These twenty genotypes are the parents of mapping populations that, due to their segregation for a variety of important agronomic traits, play a central role in a current collaborative project among US public wheat breeding programs (http://maswheat.ucdavis.edu/). The pedigree information and year of release for these genotypes have been described previously (Chao et al. 2007). For comparison, we also included the panel of 13 diverse *T. aestivum* accessions used in the wheat SNP discovery project, referred to hereafter as the "diverse germplasm set". This panel included accession PI350731 from Austria; Yangxian Yangquanmai and Chinese Spring (CS) from China; IWA10993, IWA10940, and Iranian spelt (405a) from Iran; PI16698, PI166305, PI166792, and PI119325 from Turkey; PI410595 from Pakistan; a common wheat cultivar Yecora Rojo from the US; and Opata 85 from Mexico (http://wheat.pw.usda.gov/SNP).

SNP selection and genome-specific primer validation

The *T. aestivum* SNPs used in this study were ascertained during the NSF Wheat SNP project from a panel including the 13 common wheat accessions described above and eight synthetic wheats. There was no frequency cut-off applied to define SNPs. The SNP definition criterion and the use of a discovery panel that was more diverse than our target population (US commercial varieties) is expected to limit the impacts of ascertainment bias (Brumfield et al. 2003).

We first developed SNP-detection primers for a set of ESTs that contain at least one polymorphism among the 13 accessions of the diverse germplasm set defined above. A second group of SNP-detection primers was then developed from ESTs that are not polymorphic in the diverse germplasm set but that contain at least one polymorphism among the synthetic and durum wheats included in the NSF Wheat SNP project. These two sets of SNPs were analyzed separately to determine if pre-selection of SNPs polymorphic in the diverse *T. aestivum* germplasm set indeed enhances the chance of finding polymorphism among US cultivars.

To ensure that the PCR fragments amplified by the genome-specific primers were derived only from the

Table 1 Number and percentage of polymorphic SNP markers detected for each pair of parental lines included in this study

Parent 1	Parent 2	No. (%) polymorphic markers	No. markers scored
Rio Blanco (HWW[a])	IDO444 (HRW)	50 (15.1)	331
McNeal (HRS)	Thatcher (HRS)	69 (20.2)	342
Louise (SWS)	Penawawa (SWS)	63 (18.3)	344
Grandin/ND614 (HWS)	NY18/Clark's Cream 40-1 (SWW)	80 (23.4)	342
Platte (HWW)	CO940610 (HWW)	61 (17.7)	344
TAM 105 (HRW)	Jagger (HRW)	68 (19.7)	345
Harry (HRW)	Wesley (HRW)	51 (14.7)	346
P91193 (SRW)	P92201 (SRW)	45 (13.5)	333
Cayuga (SWW)	Caledonia (SWW)	26 (7.5)	345
USG 3209 (SRW)	Jaypee (SRW)	46 (13.4)	343
	Total	559 (16.4)	3,415

Shown in parentheses after the cultivar name is the market class

[a] HRS = Hard red spring; HWS = Hard white spring; SWS = Soft white spring; HRW = Hard red winter; HWW = Hard white winter; SRW = Soft red winter; SWS = Soft white winter

intended chromosomes, nullisomic-tetrasomic (NT) lines of CS (Sears 1954) were used to optimize the specificity of the PCR conditions to our PCR equipment. The optimized targeted regions were then amplified in the 20 wheat cultivars used in this study. All PCR reactions were performed using 50 ng of genomic DNA in 20 μl PCR reaction mix containing 1 unit of Taq polymerase, 1.5 mM $MgCl_2$, 100 μM of each of the four dNTPs, and 5 pmol each of forward and reverse genome-specific primer. The PCR cycling used for most of the primer pairs included an initial denaturation step at 95°C for 3 min, followed by 10 cycles of touch down at 95°C for 20 s, from 63 to 58°C for 20 s (0.5°C decrease per cycle), and 72°C for 80 s. The touch down was followed by 36 cycles of 95°C for 20 s, 58°C for 20 s, and 72°C for 80 s. When these general conditions did not work, they were modified by extending the number of cycles to 40 and by varying the annealing temperatures from 56 to 66°C.

SNP-detection primers

Whenever possible, SNP-detection primers were designed from regions ending one base immediately upstream from the polymorphic site on both DNA strands. The primers were designed with melting temperatures between 55 and 60°C and lengths between 25 and 30 bases. In cases where multiple SNPs were discovered within the same EST, SNP haplotypes among the wheat accessions were compared. If two haplotypes were detected, only one SNP was selected for assay design. However, if more than two haplotypes were observed within the discovery panel, SNP-detection primers were designed for assaying two different SNP sites informative for identifying different haplotypes among the accessions.

SNP detection and allele scoring

SNP detection was carried out using a single-base extension assay based on the method of template-directed dye-terminator incorporation assay with fluorescence polarization detection (FP-TDI) (Chen et al. 1999). An aliquot of the amplified genome-specific fragments was combined with primer extension reaction mix and 5 pmol SNP-detection primer from one DNA strand, following the protocols for the AcycloPrime II SNP detection kit provided by Perkin Elmer (Boston, MA) with two fluorescent dye-labeled nucleotides included allowing the two allelic variants of a specific SNP to be interrogated in a single assay. The primer extension reactions were carried out using an initial denaturation cycle at 95°C for 2 min, followed by 20 cycles of 95°C for 15 s and 60°C for 30 s. At the end of the assay, the reaction mix was subjected to fluorescence polarization (FP) measurements using a Perkin Elmer's Victor V plate reader. The data analysis and allele calls based on clustering FP values were performed using the Excel workbooks from http://www.snpscoring.com.

Diversity analysis

SNP marker diversity for the US cultivars and the diverse germplasm set was measured using the polymorphism information content (PIC) formula proposed by Weir (1996) and implemented in the PowerMarker software (Liu and Muse 2005). PIC values, which provide an estimate of the probability of finding a polymorphism between two random samples of the germplasm, were also calculated separately for each chromosome and genome. The ratio between the PIC values for the US cultivars and the diverse germplasm set was used to estimate the relative decrease in diversity among the three genomes in this modern group of cultivars. To test the significance of the differences in diversity reduction among genomes, the PIC ratios for individual loci were treated as replications in a non-parametric Kruskal–Wallis rank sum test (implemented using the *R* statistical package, http://www.r-project.org). SNP frequency distributions for the different genomes were compared using χ^2 tests.

The SNP PIC values were also compared with PIC values obtained from a previous study including 242 wheat genomic SSR markers and the same set of US cultivars (Chao et al. 2007). Distance matrices for the 20 cultivars were calculated for SSR and SNP markers separately using Rogers' distance, and the correlation between matrices was determined using the Mantel test (Mantel 1967). These calculations were performed using the PowerMarker software.

Results

SNP-detection primers

In this study, we assayed a total of 364 ESTs, including 145 for which we designed SNP primers for two different polymorphic sites. For each EST, the chromosome bin location, SNP position and DNA strand, and SNP-detection primer sequence are available in supplementary Table S1. Only the primer from the strand resulting in unambiguous allele discrimination is reported.

Primers from 350 ESTs yielded unambiguous genotype calls in the FP assays, with at least one of the two primers designed for each polymorphic site. The SNP assays for the remaining 14 ESTs failed to give clear genotype calls, which is more likely due to a technical failure than to a problem with the SNP primer specificity. Nevertheless, the conversion rate from discovered SNPs to working assays depends on many factors, such as the levels of sequencing errors, sequence compositions near the targeted SNPs, and the genotyping systems used. Based on our results, we estimate that the single-base extension method can yield an overall assay success rate of 96% in wheat cultivars when primers for both strands are tested.

SNP marker polymorphism in US wheat cultivars

Of the 145 ESTs for which we designed SNP primers for two polymorphic sites, 31 were found to be polymorphic at both sites for the 20 genotypes used in the study. These two polymorphisms defined two haplotypes for 22 of the ESTs and more than two haplotypes for the other nine. Out of the remaining 114 ESTs, 57 were found to be polymorphic for only one site, 51 were monomorphic for both sites, and six failed. This result indicates that assaying two different SNPs per EST (each for both strands) increases the chance of finding polymorphic SNP markers by at least 16% (57/350), suggesting that an optimization step using primers from both strands is worthwhile. The results further showed that among the ESTs with more than one SNP assayed, only a small portion detected more than two haplotypes (9/145) among wheat cultivars, thereby providing evidence that the extent of linkage disequilibrium (LD) is likely extensive within the population of wheat cultivars used in this study. The extent of LD is expected to decrease within populations of more diverse germplasm such as those used for SNP discovery. Altogether, 359 SNPs (341 ESTs with 1 SNP and nine with 2 SNPs) yielded unambiguous genotype calls. Call quality was further assessed by determining the rate of high quality calls among the 20 genotypes assayed in the study. Over 70% of the 359 SNPs yielded high quality calls for all 20 genotypes. Among the other 30%, 29% gave high quality calls for 17–19 genotypes, whereas only 1% of the SNPs assayed exhibited high quality calls for 13 to 16 genotypes. None of the selected SNPs gave less than 13 high quality calls among the 20 genotypes examined.

Of the 359 SNPs selected, 253 revealed at least one polymorphism among the diverse germplasm set, and the remaining 106 were polymorphic only in synthetic or tetraploid wheat accessions. These last 106 SNPs were selected for a balanced representation of the three homoeologous genomes (31 from the A genome, 36 from the B genome, and 39 form the D genome). The sets of 253 and 106 SNPs were evaluated separately to quantify the effect of pre-selecting polymorphic SNPs in the diverse germplasm set on the level of polymorphism detected among US cultivars.

Results from the full SNP dataset showed that 212 markers (59%) detected at least one difference among the 20 US wheat cultivars. This percentage increased to 70% when only those 253 SNPs pre-selected for polymorphisms were considered and dropped to 33% when considering only the remaining 106 SNPs. The full SNP dataset was also used to calculate the number of polymorphic markers and the level of polymorphism in the parental lines of the 10 mapping populations in order to estimate the number of polymorphisms that can be expected in mapping populations based on crosses between adapted US cultivars (Table 1). On average, 16.4% of the pairs of parental lines revealed polymorphisms for the 359 SNPs tested. The percentage was slightly higher (20.4%) for the subset of 253 pre-selected SNPs and much lower (6.5%) for the remaining 106 SNPs. Considering all the SNPs, the highest level of polymorphism (23.4%) was found between two parental lines belonging to different growth habits (spring line Grandin/ND614 and winter line NY18/Clark's Cream 40-1) (Table 1).

Genetic diversity of SNP markers in US wheat cultivars

The average genetic diversity present among the 20 US wheat cultivars was evaluated using the three datasets described above. Although the 359 SNP loci were distributed along the 21 chromosomes, fewer markers were tested in the D genome due to the lower number of polymorphic D genome-SNPs found in the wheat SNP discovery project (excluding synthetic wheats). Table 2 shows the marker distribution, number of alleles, and PIC values calculated for each chromosome. The PIC values across chromosomes ranged from 0.04 (6D) to 0.29 (2B), with SNP markers in the D genome showing a higher proportion of non-polymorphic SNPs (62%) than the A (32%) or B genomes (39%).

Table 2 Allele number and PIC values calculated for 359 SNP and 242 SSR markers in a set of 20 US wheat cultivars

	SNP marker			SSR marker		
	No. marker	No. allele	PIC	No. marker	No. allele	PIC
Chromosomes						
1A	19	35	0.272	12	73	0.666
1B	17	25	0.074	8	48	0.675
1D	7	11	0.170	11	44	0.580
2A	23	42	0.253	13	78	0.699
2B	16	29	0.288	11	58	0.562
2D	11	16	0.144	10	76	0.722
3A	24	38	0.208	12	71	0.696
3B	24	42	0.266	12	57	0.617
3D	8	13	0.184	12	80	0.730
4A	10	18	0.238	11	63	0.596
4B	19	25	0.081	11	54	0.551
4D	6	9	0.048	11	58	0.643
5A	23	36	0.194	13	65	0.597
5B	22	37	0.191	13	69	0.556
5D	9	13	0.078	15	88	0.638
6A	17	25	0.177	8	40	0.623
6B	14	20	0.114	10	59	0.660
6D	19	23	0.037	13	53	0.531
7A	30	52	0.253	12	56	0.494
7B	22	38	0.195	12	71	0.655
7D	19	24	0.088	12	64	0.678
Genome						
A	146	246	0.228	81	446	0.624
B	134	216	0.173	77	416	0.611
D	79	109	0.107	84	463	0.646
Homoeologous group						
1	43	71	0.172	31	165	0.640
2	50	87	0.228	34	212	0.661
3	56	93	0.219	36	208	0.681
4	35	52	0.123	33	175	0.597
5	54	86	0.154	41	222	0.597
6	50	68	0.109	31	152	0.605
7	71	114	0.179	36	191	0.609
Total	359	571		242	1,325	
Overall mean		1.591	0.181		5.475	0.626

The average genetic diversity (mean PIC value) among US cultivars for the complete set of 359 SNP markers was 0.18. This value increased to 0.23 (Table 3) when only the subset of 253 pre-selected SNPs were considered and decreased to 0.07 in the subset of the remaining 106 SNPs. As expected, the average PIC value for the diverse germplasm set was higher (0.34) than for the US cultivars (0.23, Table 3).

Relative reductions in SNP marker diversity per genome

To investigate the relative changes in SNP diversity among the different genomes, we compared diversity values obtained for the 20 US wheat varieties with those obtained for the diverse germplasm set. The subset of 253 SNP markers was used for this analysis. Since this subset includes only polymorphic markers, the PIC values cannot be used to estimate the natural diversity of the diverse germplasm set or the decrease in diversity among the US cultivars. However, the comparison of the reduction in genetic diversity among genomes is still valid because of its relative nature.

The ratios of PIC values from the panel of US cultivars and the diverse germplasm set for SNPs grouped by genome ($\mathrm{PIC}_{\mathrm{cultivar}}/\mathrm{PIC}_{\mathrm{diverse\ set}}$) revealed that a higher amount of SNP diversity was retained in both the A genome (77%) and the B genome (62%) than in the D genome (50%). Using the $\mathrm{PIC}_{\mathrm{cultivar}}/\mathrm{PIC}_{\mathrm{diverse\ set}}$ ratios of individual loci as replications in a non-parametric Kruskal–Wallis ANOVA, we confirmed the absence of a significant difference between the A and B genome values ($P > 0.05$). However, the $\mathrm{PIC}_{\mathrm{cultivar}}/\mathrm{PIC}_{\mathrm{diverse\ set}}$ ratios for the D genome SNPs were significantly lower than those in the other two genomes ($P < 0.05$).

We hypothesized that the larger reduction in diversity observed in the D genome could be the result of differences in the distribution of polymorphism frequencies in the different genomes. To explore this possibility, we compared the distribution of the minor allele frequency for the three genomes (Fig. 1). In the diverse germplasm set, the D genome showed a higher proportion of polymorphisms with low minor allele frequencies (lower quartile, frequency <0.125) and a lower proportion of polymorphisms with high minor allele frequencies (upper quartile, frequency between 0.375 and 0.500) than the A and B genomes (Fig. 1, lower panels).

To test if these differences were significant, we performed a χ^2 test comparing the number of SNPs per quartile among the different genomes. The A and B

Table 3 Comparison of Polymorphism Information Content (PIC) values among 20 US wheat cultivars and a set of 13 diverse common wheat accessions

	No. marker	US cultivars		Diverse accessions		PIC ratio
		No. alleles	Mean PIC	No. alleles	Mean PIC	Cultivar/Diverse set
Genome						
A	115	200	0.261	230	0.337	0.77
B	98	172	0.221	196	0.355	0.62
D	40	58	0.145	80	0.293	0.50
Homoeologous group						
1	32	56	0.225	64	0.336	0.67
2	38	69	0.279	76	0.396	0.70
3	38	66	0.273	76	0.346	0.79
4	20	34	0.182	40	0.298	0.61
5	40	66	0.202	80	0.362	0.56
6	27	43	0.186	54	0.325	0.57
7	58	96	0.217	116	0.296	0.73
Total	253	430		506		
Overall mean		1.70	0.227	2	0.337	

Values are the mean of 253 SNP markers pre-selected for their polymorphism in the diverse germplasm set

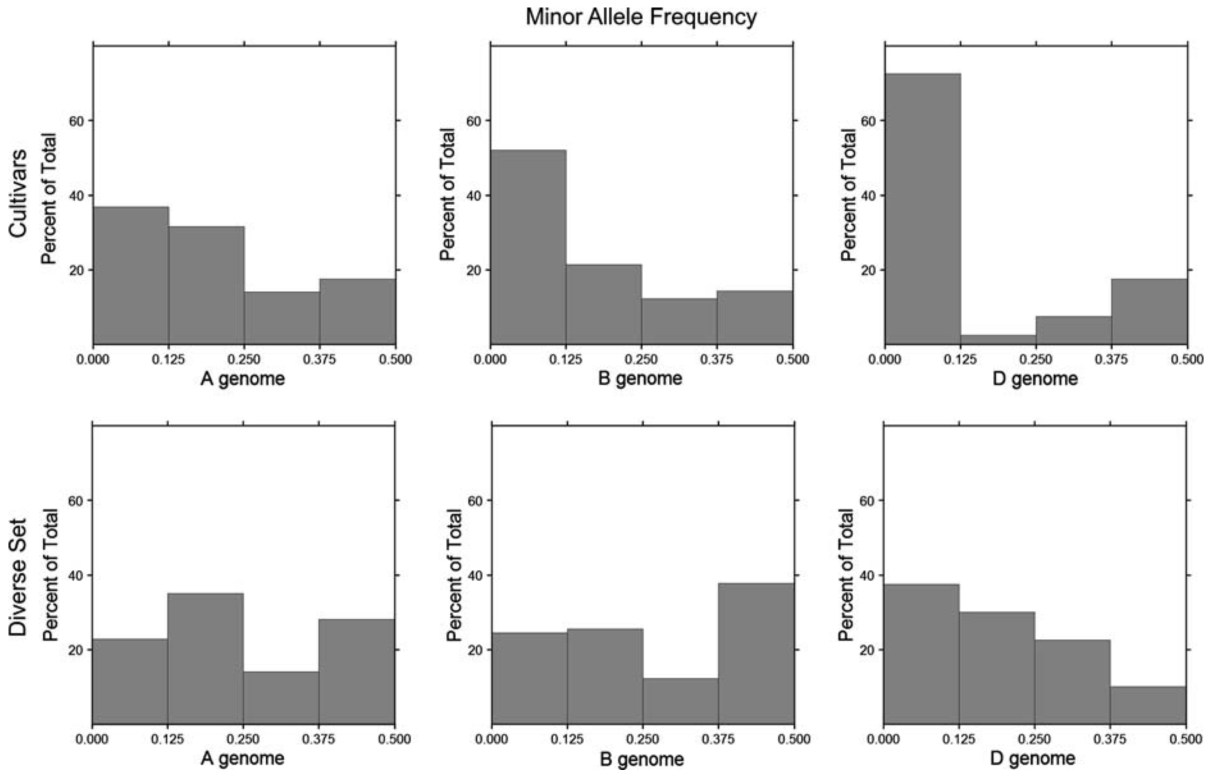

Fig. 1 Distribution of minor allele frequencies for 253 polymorphic SNP markers pre-selected for polymorphism among the diverse germplasm set in each of the three genomes. Frequencies distributions among US cultivars (upper panel) and a set of 13 diverse germplasm set excluding synthetic wheats (lower pane)

genomes showed no significant differences ($P > 0.05$), whereas the D genome differs significantly from the other two genomes ($P < 0.05$) (Fig. 1). These results confirm that, in the diverse germplasm set, SNPs in the D genome have a different frequency distribution than those in the A and B genomes.

Comparison between SSR and SNP markers in wheat

In a previous genetic diversity analysis we characterized 43 wheat genotypes with a set of 242 SSR markers distributed across all 21 chromosomes (Chao et al. 2007). From this data set we selected the 20 US cultivars included in this study and recalculated the diversity values. The average number of alleles per SSR marker was 5.5 and the mean PIC value of 0.63 (Table 2), more than 3-fold larger than the value obtained for the SNP markers (PIC = 0.18).

We also compared the genetic relationships among the 20 US cultivars inferred from the two sets of markers. The Rogers' distance matrices based on SSR and SNP markers were moderately correlated (Mantel test $R = 0.42$, $P < 0.0001$), indicating that the two matrices contain common information. However, the intermediate R value indicates that the relationships among germplasm inferred from the two different marker systems can be quite different.

Discussion

SNP polymorphism in wheat

The quantification of the SNP polymorphism among elite wheat cultivars is important to estimate the utility of SNP markers in commercial wheat breeding programs. The level of polymorphism determines the proportion of useful SNPs and, therefore, affects the cost of using SNPs to develop genetic maps, perform association studies, or use them in marker assisted selection.

The average PIC value of 0.18 found in our study indicates that, on average, one in five to one in six of the intron-derived SNPs are expected to be polymorphic between any two US common wheat cultivars. This value is considerably lower than the mean PIC value of 0.27 estimated by Somers et al. (2003), a discrepancy most likely due to the more diverse genotypes used in their study. Somers et al. (2003) included wheat cultivars from different countries and, more importantly, a synthetic wheat. Synthetic wheats are developed by hybridizing tetraploid wheat with *Aegilops tauschii* accessions, capturing greater diversity than the one currently present in the *T. aestivum* germplasm. The germplasm set in this study, by comparison, was comprised of elite cultivars adapted to different production regions in the US and included no synthetic wheats.

The average PIC values of 0.18 found with the complete SNP set increased to 0.23 when calculations were based only on the 253 SNPs pre-selected for polymorphisms among the diverse germplasm set. In contrast, the PIC values found using the 106 SNPs that were polymorphic only in synthetic and durum wheats in the SNP discovery project were less than one-third (PIC = 0.07) of the previous values. From these results, we conclude that studies using commercial wheat cultivars and a limited number of SNPs can benefit greatly from selecting only those SNPs that are polymorphic among the diverse germplasm dataset. That being said, studies limited by the number of available SNPs can certainly find some additional polymorphism within the SNPs showing polymorphisms in synthetic and tetraploid wheat; for within this set of 106 SNPs, 36 (34%) showed at least one polymorphism among the US cultivars.

Comparison between SNP and SSR diversity values

The SNP PIC values were found to be approximately three times lower than those based on SSR markers (Table 2). This agrees well with published results suggesting that for linkage studies approximately three times as many SNPs are needed in comparison to SSRs (Kruglyak 1997). SNPs are bi-allelic markers and, therefore, are limited to maximum PIC values of 0.5 (when both alleles have identical frequencies), whereas multi-allelic markers (e.g. SSR) do not have this limitation. In addition, nucleotide mutation rates in intronic regions of the wheat genome (average 5.5×10^{-9} substitutions nt^{-1} $year^{-1}$, Dvorak and Akhunov 2005) are several orders of magnitude slower than the mutation rate of SSRs (2.4×10^{-4} repeats $allele^{-1}$ $generation^{-1}$, Thuillet et al. 2002). Finally, most of the SSRs used in the previous study (Chao et al. 2007) were derived from genomic regions, whereas the SNPs included in this study were selected exclusively from genic regions (introns), which tend to evolve more slowly than wheat intergenic regions (Dubcovsky and Dvorak 2007). Our previous study revealed significantly lower levels of polymorphism among EST-derived SSRs (~22%) than among genomic SSRs (>50%, Chao et al. 2007). A similar result has been reported in barley (Russell et al. 2004).

The different mutation rates, as well as the different genome regions represented in the SNP and SSR datasets, may contribute to the low correlation observed between the genetic distance matrices for US cultivars derived from these different types of markers. The higher mutation rate of the SSR markers may also explain the less severe reduction in diversity observed in the D genome compared with the SNP data (Table 2).

A way to increase SNP diversity values is to combine the information of multiple SNPs for a single locus by haplotype analysis (Brumfield et al. 2003). A two-fold increase of haplotype diversity over individual SNPs was found in maize (Ching et al. 2002). Higher haplotype diversities were also observed in sugar beet (Schneider et al. 2007), soybean (Zhu et al. 2003), and barley (Russell et al. 2004). As shown in humans, haplotype analysis offers the advantage of capturing most of the genetic variation across a region, and a minimal set of SNPs can be selected and used to distinguish common haplotypes in a block (Cardon and Abecasis 2003). Knowledge of haplotype structure in genic regions will also help to assess the extent of LD across genes in cultivated wheat.

Low level of polymorphism in wheat D genome

In this study, we found a lower level of SNP polymorphism in markers located in the D genome relative to those located in the A and B genomes in both the US cultivars and the diverse germplasm set.

Using the 253 pre-selected SNPs, the ratio of the average PIC value for the A and B genomes to that of the D genome (PIC_{AB}/PIC_D) was found to be 1.7 for the US cultivars and 1.4 for the diverse germplasm set (Table 3). These ratios increased to 1.9 and 1.7, respectively, when the complete set of 359 SNP was considered (Table 2, data not shown), suggesting that the low PIC_{AB}/PIC_D value for the diverse germplasm set may be a result of the exclusion of non-polymorphic SNPs from this dataset. When the same calculation was made using a set of 1,228 genes from the NSF wheat SNP discovery project, including both polymorphic and non-polymorphic genes, the PIC_{AB}/PIC_D ratio for the diverse germplasm set increased to 2.2 (data not shown), a value more similar to the ones found in our study. These values suggest that the average PIC values for the A and B genomes are approximately two-fold higher than the PIC values for the D genome in hexaploid wheat.

The low level of diversity found in the D genome is expected from the evolutionary history of hexaploid wheat. *Triticum aestivum* originated less than 10,000 years ago from the hybridization of tetraploid wheat with a limited number of *A. tauschii* accessions (Dvorak et al. 1998; Talbert et al. 1998; Caldwell et al. 2004). After these few polyploidization events, limited gene flow occurred between *A. tauschii* and *T. aestivum,* whereas frequent hybridization and good fertility of the pentaploid hybrids allowed for continuous gene flow between *T. aestivum* and tetraploid wheat species, increasing the diversity of the A and B genomes relative to the D genome (reviewed in Dubcovsky and Dvorak 2007).

This study suggests that the selection of modern cultivars from the original hexaploid landraces resulted in an additional diversity bottleneck that had a stronger effect on the D genome than on the A and B genomes (Fig. 1, Table 3). Because it is unlikely that those differences arose by differential selection of genes located in a particular genome, the most likely explanation is differential effect of genetic drift during the diversity bottleneck that accompanied the development of the modern adapted germplasm due to preexisting differences on the proportion of low-frequency alleles among genomes. The reduction of effective population size during the bottleneck imposed by selection in modern breeding programs increased the chance of genetic drift, which in turn increased the probability of losing low frequency alleles from the population. The discovery that the proportion of low-frequency alleles in the D genome is higher than that in the A and B genomes in the diverse germplasm set may have determined a higher loss of allelic variants in the D genome (Fig. 1).

Conclusion

In summary, the characterization of wheat SNPs in commercial US cultivars is encouraging and suggests that SNP markers have adequate levels of polymorphisms to make them useful in genetic and breeding studies. This study indicates that additional SNP discovery efforts targeted to the D genome would likely be required to offset the lower diversity of this genome among the elite wheat cultivars. With increasing SNP discovery projects and the development of SNP assay technologies that can assay thousands of SNPs in parallel, it is anticipated that SNP markers will play an increasingly important role in wheat genetics and breeding applications.

Acknowledgements This research was supported in part by the funds from the U.S. Department of Agriculture, Cooperative State Research, Education and Extension Service, Coordinated Agricultural Project grant number 2006-55606-16629 and NSF Grant No. DBI-0321757. We thank Dr. Jan Dvorak for facilitating early access to the SNP data generated by the NSF project and for his useful suggestions and ideas and to Iago Lowe for his thorough revision of the manuscript.

References

Batley J, Barker G, O'Sullivan J, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertion/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91. doi:10.1104/pp.102.019422

Blake NK, Sherman JD, Dvorak J, Talbert LE (2004) Genome-specific primer sets for starch biosynthesis genes in wheat. Theor Appl Genet 109:1295–1302. doi:10.1007/s00122-004-1743-4

Brookes A (1999) The essence of SNPs. Gene 234:177–186. doi:10.1016/S0378-1119(99)00219-X

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256. doi:10.1016/S0169-5347(03)00018-1

Caldwell KS, Dvorak J, Lagudah ES, Akhunov E, Luo MC, Wolters P et al (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167:941–947. doi:10.1534/genetics.103.016303

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140. doi:10.1016/S0168-9525(03)00022-2

Chao S, Zhang W, Dubcovsky J, Sorrells M (2007) Evaluation of genetic diversity and genome-wide linkage disequilibrium among US wheat (*Triticum aestivum* L.) germplasm representing different market classes. Crop Sci 47:1018–1030. doi:10.2135/cropsci2006.06.0434

Chen X, Levine L, Kwok P-Y (1999) Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res 9:492–498

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S et al (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19. doi:10.1186/1471-2156-3-19

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E et al (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 23:203–207. doi:10.1038/13833

Choi I-Y, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV et al (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696. doi:10.1534/genetics.107.070821

Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316:1862–1866. doi:10.1126/science.1143986

Dvorak J, Akhunov ED (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *aegilops-triticum* alliance. Genetics 171:323–332. doi:10.1534/genetics.105.041632

Dvorak J, Luo MC, Yang ZL (1998) Genetic evidence on the origin of *Triticum aestivum* L. In: Damania AB, Valkoun J, Willcox G, Qualset CO (eds) The origins of agriculture and crop domestication. ICARDA, Aleppo, Syria, pp 235–251

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies *Indica* and *Japonica* genome alignments. Genome Res 14:1812–1819. doi:10.1101/gr.2479404

Hayes P, Szucs P (2006) Disequilibrium and association in barley: Thinking outside the glass. Proc Natl Acad Sci USA 103:18385–18386. doi:10.1073/pnas.0609405103

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24. doi:10.1038/ng0997-21

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Mantel NA (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N et al (2002) Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers. DNA Res 9:163–171. doi:10.1093/dnares/9.5.163

Rafalski A (2002) Applications of single nucleotide polymorphism in crop genetics. Curr Opin Plant Biol 5:94–100. doi:10.1016/S1369-5266(02)00240-6

Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D et al (2006) Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). Genome 49:1131–1139. doi:10.1139/G06-067

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A et al (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274:515–527. doi:10.1007/s00438-005-0046-z

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656–18661. doi:10.1073/pnas.0606133103

Russell J, Booth A, Fuller J, Harrower B, Hedley P, Machray G et al (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome 47:389–398

Schneider K, Kulosa D, Soerensen TR, Möhring S, Heine M, Durstewitz G, Polley A, Weber E, Jamsari, Lein J, Hohmann U, Tahiro E, Weisshaar B, Schulz B, Koch G, Jung C, Ganal M (2007) Analysis of DNA polymorphisms in sugar beet (*Beta vulgaris* L.) and development of an SNP-based map of expressed genes. Theor Appl Genet doi:10.1007/s00122-007-0591-4

Sears ER (1954) The aneuploids of common wheat. Mo Agric Exp Stn Res Bull 572:1–59

Somers DJ, Kirkpatrick R, Moniwa M, Walsh A (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 49:431–437. doi:10.1139/g03-027

Talbert LE, Smith LY, Blake NK (1998) More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. Genome 41:402–407. doi:10.1139/gen-41-3-402

Thuillet AC, Bru D, David J, Roumet P, Santoni S, Sourdille P et al (2002) Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, *Triticum turgidum* (L.) Thell. ssp. *durum* desf. Mol Biol Evol 19:122–125

Vroh Bi I, McMullen MD, Sanchez-Villeda H, Schroeder S, Gardiner J, Polacco M et al (2006) Single nucleotide polymorphism and insertion-deletion for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci 46:12–21. doi:10.2135/cropsci2004.0706

Weir BS (1996) Genetic data analysis II. Sinauer Associate, Inc., Sunderlands

Zhu YL, Song QJ, Hyten DL, van Tassell CP, Matukumalli LK, Grimm DR et al (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134