2010, **11** 727
// . . /1471-2164/11/727

Genomic

RESEARCH ARTICLE Open Access

- -

(Triticum aestivum .)

S C [1], J D [2], J D [2], M -C L [2], S P B [3], R M [4,18], D R C [5], L E T [6], J A A [7], S D [8], K G [9], J C [10], K C [11], P L B [12], J C R [13], S H [14], B F C [15], S P [16], M E S [17], E D A [4*]

Abstract

. . : S (SNP) -
, -
. H , - 1536 SNP (LD)
478 (/ / / — / /) 17
S M .
, M (OPA) SNP 478
. H
. A
, - .
E (F_{ST}) .
I , F_{ST}
. S QTL. A ,
LD D- , LD A-
B- . T LD
BMC G

Background

In crops, the level of genetic diversity and linkage disequilibrium (LD) can be affected by various factors including demography and inbreeding [1-6], selection for favorable alleles [7,8], domestication [2,9,10], outcrossing of crop cultivars with genetically distinct lines of wild ancestors and landraces [1,11,12] and admixture [13,14]. Genetic diversity of domesticated crops is usually reduced compared to wild ancestors [2,6,9,15,16]. In tetraploid

The SNPs discovered in a panel of 32 lines of tetraploid and hexaploid wheat were downloaded from the Wheat SNP Database [36]. SNP selection and assay design were performed according to previously described procedures [32]. The following criteria were applied for SNP selection: no more than 2 SNPs were selected per locus, with preference being given to SNPs present in at least two lines in the discovery panel. Additional SNPs were discovered by sequencing the transcriptomes of *T. aestivum* cv. Chinese Spring and Jagger. Repetitive elements were detected and masked by comparing sequences with the TREP [37] and GIRI [38] databases. The masked sequences were submitted to Illumina for processing by Illumina® Assay Design Tool (ADT). The ADT generates designability rank scores for each SNP that can vary from 0 to 1. The SNPs with scores above 0.6 have a high probability of being converted into a successful genotyping assay. A total of 1536 SNPs were selected for developing the wheat OPA (Additional File 2.xls). Genotyping was performed at the USDA-ARS genotyping laboratory in Fargo, North Dakota according to standard Illumina GoldenGate assay protocols [39]. Subsequent genotype calling was carried out using Illumina's BeadStudio software v.3. The accuracy of the genotype call was manually evaluated for the misclassification of homozygous and heterozygous clusters using the software's clustering algorithm. This step proved critical for reducing the genotyping error rate associated with peculiarities of clustering patterns in polyploid wheat. Following the removal of loci with low-quality clustering, the previously estimated genotyping error rate for hexaploid wheat was a mere 1% [32].

Genetic diversity was evaluated by calculating the polymorphism information content ($PIC = 1 - \sum_{i}^{n} p_i^2$, where p_i is the frequency of the i-th allele [40]) for the number of alleles across and within breeding programs using PowerMarker software [41]. Analyses were performed separately on four datasets: three datasets included SNPs grouped by genome and one dataset included complete set of SNPs.

For analysis of population structure, the SNP dataset was divided into the three genome-specific datasets and one combined dataset. To reduce the effect of frequency correlation between linked alleles, we selected SNP loci located approximately 4 cM or farther apart from each other. The A-genome dataset included 91 SNP loci while the B-genome and D-genome dataset included 89 and 39 SNP loci, respectively (Additional File 3.xls). We assumed that each individual in the population was homozygous for all loci, and heterozygous loci were treated as missing data. The proportion of heterozygous

l-2164/11/727.078355nl

number of lines. The mean F_{ST} values in a sliding window of 5 consecutive linked SNPs were calculated to identify genomic regions genetically differentiated between spring and winter wheat lines. A 95% confidence interval (CI) for mean F_{ST} values was estimated by sampling 1,000 times with replacement the sets of 5 SNPs randomly selected from the 849-SNP dataset and taking the 95th percentile of the distribution of means.

Regions of the wheat genome showing elevated F_{ST} levels were compared with the positions of previously mapped or cloned flowering time QTL. The sequences of genes containing SNPs included in the wheat OPA were compared with the sequences of gene-derived flanking molecular markers (ESTs, cDNA) used in QTL mapping studies. The syntenic relationship between wheat, rice and Brachypodium genomes was used to compare and validate map positions.

For measuring LD, the locations of gene loci harboring SNPs on the *Ae. tauschii* genetic map reported by Luo et al. [51] were used. Pair-wise linkage disequilibrium (LD) was measured using the squared allele-frequency correlations, r^2 , according to Weir [50]. In order to reduce the variation of LD estimates generated by the inclusion of rare alleles, only SNP alleles with minor allele frequency (MAF) higher than 0.05 were used in these calculations. LD levels and the rate of LD decay were assessed by calculating r^2 for pairs of SNP loci and plotting them against genetic distance. The relationship between LD decay and genetic distance was summarized by fitting a locally-weighted linear regression (loess) line to r^2 data. The statistical significance of individual r^2 estimates was calculated by the exact test following Weir [50]. The false discovery rate (FDR) was established at 0.01 using the Benjamini & Hochberg method [52]. Chromosome specific r^2 values were plotted using the R package LDheatmap [53]. Blocks of SNPs showing elevated levels of LD were identified using the method described by Gabriel et al. [54] and implemented in the program Haploview [55]. Background LD was estimated as the 95th-percentile of the distribution of r^2 values for unlinked SNP loci [25].

Results

The genotyping of 478 spring and winter wheat lines with multiplexed 1,536 Illumina Golden Gate SNP assay generated 734,208 genotyping data points (Table 1). After the removal of SNPs failing to generate clear genotype clustering, 1,299 SNPs with high quality genotype calls were obtained with a 85% success of SNP conversion into the working genotyping assays. Considering these SNPS, 849 were polymorphic among the 478 lines included in this study. Most genotypes were homozygous (400,328 = 98.6%) with only a small fraction showing residual heterozygocity (1,961 = 0.5%) or no amplifilule (340(1)0(,([)33.1(s)-356.5(=)-340.4(])0fi)00.3(%)0(e)-1-.84(])0

Table 1 Wheat OPA evaluation.

				(%)		
A	642	93	549	368 (67%)	1.67	0.165
B	675	109	566	374 (66%)	1.67	0.170

population subdivision. The inferred population structure was consistent across multiple simulation runs. The majority of winter and spring wheat breeding populations were assigned to separate clusters (Additional File 5.xls). Only the A-genome data in the NY winter wheat population showed an equal proportion of ancestry in the two clusters (Additional File 5.xls). Grouping of varieties at $K = 2$ using the D-genome data did not result in clear separation of spring and winter lines. Only 7 out of 17 breeding populations derived more than 80% of their D-genome's genetic ancestry from only one of the two clusters.

To identify the optimal number of K clusters in genome-specific datasets, we calculated the posterior probability Pr (K|X) [42] and K [45] for each simulation run. The posterior probability in structure runs was constantly increasing with increasing the values of K ranging from 2 to 21 providing little guidance in selecting the optimal number of clusters. The InStruct software [47] showed a similar trend (data not shown) for the same range of K values. These observations were consistent with previously reported analyses of population structure in barley and maize breeding populations using multi-locus SNP data [56,57]. Therefore, the selection of the optimal value of K in this study was based on the analysis of relationship between Pr (X|K), value of K and the variation of Pr (X|K) among multiple independent runs of Gibb's sampler.

The probability of data for K from 2 to 5 for the A-genome SNP dataset was consistent among multiple independent runs of Structure (Figure 2 and Additional File 4.ppt). For the B-genome SNP dataset, we obtained consistent Pr (X|K) for K values varying from 2 to 4. For values of K above 6 for the A-genome dataset and above 5 for the B-genome dataset the simulation runs could not converge to a single mode (Additional File 4. ppt). The ambiguity of clustering solutions was also accompanied by smaller increase in the mean Pr (X|K). The cluster analysis of both the A- and B-genome datasets showed that the rate of change of Pr (X|K) with

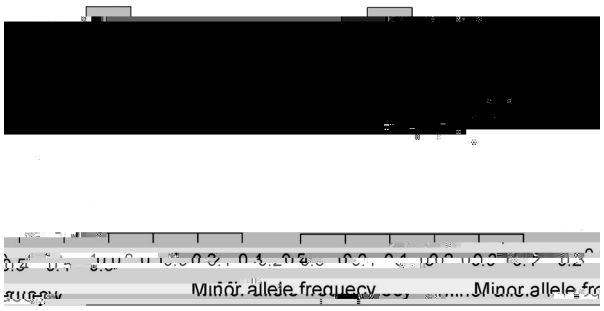
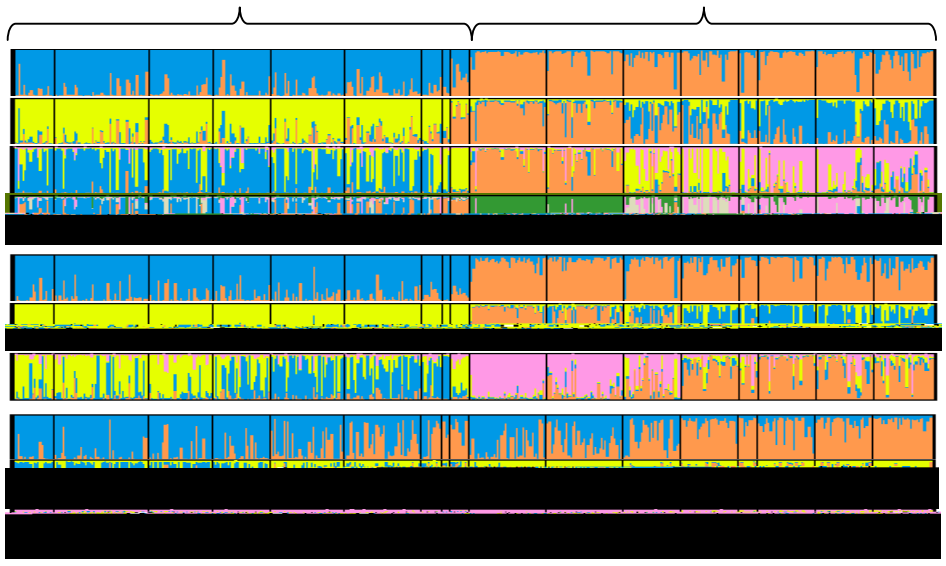


Figure 1

increase in K reached more or less stationary value (~200) for values of K = 4 or higher (Additional File 4. ppt). The maximum likelihoods of clustering obtained for correlated and uncorrelated allele frequency models suggested different values of K for the D-genome dataset. The likelihood of the correlated allele frequency model for the D-genome dataset reached its maximum at K = 9. However, the rate of likelihood gain decreased for K values above 7. The likelihood of independent allele frequency model showed that the improvement of the likelihood of clustering dropped dramatically for K above 5.

The genome-wide set of 219 SNPs was first used for assigning each of the 17 pre-defined populations to separate clusters. It is expected that each population should have maximum membership in only one cluster if the allele frequencies among populations are significantly different. However, the clustering analysis demonstrated that in several cases more than one population had membership in the same cluster (Figure 3A). There were also at least five clusters for which none of the 17 pre-defined populations showed a maximum membership coefficient. The maximum values of population-specific membership coefficients Q suggested that only NY and OK winter wheat populations and SD, CA, CM, MN, and MT spring wheat populations derived the majority of their alleles from a single cluster. These results indicated that 17 clusters exceeded the actual



number of genetically distinct populations in our sample. A number of cultivars from the CA spring wheat population share ancestry with the lines from the CIMMYT population and nearly all SD and MN spring cultivars were assigned to the same cluster. The winter wheat populations showed lower levels of genetic differentiation with the majority of cultivars in SD, NE, CO, and MT populations having membership coefficient

single SNP loci is supported by the standard deviations of chromosome- and genome-specific F_{ST} estimates exceeding the values of means (Table 4). Similar observation for single-locus F_{ST} estimates was previously noted in human populations [58]. We found that the F_{ST} estimates for individual genomes were consistent with the results of population structure analysis. A lower level of genetic differentiation between the spring and winter wheat populations was observed in the D-genome (mean F_{ST} = 0.07) relative to the A- (mean F_{ST} = 0.1) and B-genomes (mean F_{ST} = 0.11). Up to 50% reduction in variation of single-locus F_{ST} estimates can be achieved by calculating the group means of adjacent SNP markers (Table 4). As shown in the Additional File 7.tif, the mean F_{ST} values for windows of 5 SNP loci exhibited a lower proportion of extreme values and a narrower distribution. However, the fact that significant

of its initial value at about 5 cM in winter wheat compared to 6.3 cM in spring wheat (Figure 5A). In the D-genome, LD decayed faster in winter wheat declining to 50% of its initial value at about 6 cM whereas in spring wheat a similar level of LD was reached at 7 cM (Figure 5A). Both spring and winter wheat populations showed an identical rate of LD change in the B-genome decaying to 50% of its initial value over 7 cM (Figure 5A).

In order to investigate population-specific recombination processes we studied the rate of LD decay within the populations of different origin (Figure 5B and 5C). Due to the limited number of lines, WB, KS and NY winter wheat and WA spring wheat populations were excluded from this analysis. As expected, compared to LD estimates in the combined population dataset (Figure 5A), the estimates of LD within populations were higher (Figure 5B and 5C). The level of initial LD (the highest value of LD on Figure 5) in spring wheat populations, except for the CIMMYT population, varied from 0.43 to 0.49 (Figure 5B). The CIMMYT population had the lowest level of initial LD (r^2

polymorphic SNPs shared between spring and winter populations.

The proportion of genetic differentiation explained by

be optimally clustered at the same value of *K* using the A- and D- genome SNP sets, the proportion of genetic ancestry of cultivars in these clusters was variable for dif-

studies in wheat would require a smaller number of markers per unit of genetic distance than needed in cultivated barley.

Variation in the extent of LD along the chromosome affect the number of tagSNPs (subset of SNPs that capture a large fraction of the allelic variation of all SNP loci [76]) required in each genomic region to ensure that causal mutations are in LD with neighboring SNPs. The interaction of many factors affecting the rate of LD decay in the different parts of the genome complicates the determination of the number of tagSNPs required to gain sufficient power for genome-wide association map-

. T
LD, SNP
(MAF) LD . T -
LD - [2]
50 . T [2]

50 . T (FDR)
0.01 B & H 52 .

T SDA AFRI
CRIS0219050, KS A E S E.A. SDA
AFRI -CAP 2006-55606-16629.
M G
.

[1] SDA ARS G L , B R L , F ,
ND, SA. [2]D P S , C , D , CA,
SA. [3]P S B , N , L , NE, SA.
[4]D P P , K S , M , KS, SA.
[5] B , LLC, B , MT, SA. [6]D P S , M
S , B , MT, SA. [7]D . A & P G ,
M , S . P , MN, SA. [8]G R
E , CIMM T, M , D.F., M . [9]P S
D , S D S , B , SD, SA. [10]
I A R & E C , A ID, SA. [11] SDA-
ARS G , Q , P & D R ,
S , P A, SA. [12]P S P
P , B , MT, SA. [13]T A L R E
C , A , T , SA. [14]S C S D , C
S , F C , CO, SA. [15]O S ,
D P S S , S , OK, SA. [16] B , LLC,
H , KS, SA. [17]P B G , C , I , N ,
SA. [18]I B G , RAS, R .

EDA JD ; EDA ; EDA
OPA;
JD , MCL, EDA ; SC ;
EDA SC ; JD , SPB, RM, DRC, LET, JAA, SD, KG,

70. D J, D JG
. 2007, **316**:1862-1866.
71. RL, M -K A, F -D G, R S:
. . *C* 1996, **36**:218.
72. M -K A, G LI, RL, D R: **10**
. *C* 2000, **40**:590-591.
73. ML, C J, F J, K M, T R, R S, P , P, D S, G M:
. *E* —2006, **14** :289-301.
74. P JK, P M:
. *A J H G* 2001, **6** :1-14.
75. D J, L MC, L, HB: **Aegilops tauschii** . *A G* 1988, **7**:657-670.
76. I H M C :