Mourad et al. BMC Genomics (2020) 21:434
https://doi.org/10.1186/s12864-020-06835-0

BMC Genomics

RESEARCH ARTICLE Open Access

Molecular genetic analysis of spring wheat core collection using genetic diversity, population structure, and linkage disequilibrium

Amira M. I. Mourad[1*], Vikas Belamkar[2] and P. Stephen Baenziger[2]

Abstract

Background: Wheat (*Triticum aestivium* L.) is an important crop globally which has a complex genome. To identify the parents with useful agronomic characteristics that could be used in the various breeding programs, it is very important to understand the genetic diversity among global wheat genotypes. Also, understanding the genetic diversity is useful in breeding studies such as marker-assisted selection (MAS), genome-wide association studies (GWAS), and genomic selection.

Results: To understand the genetic diversity in wheat, a set of 103 spring wheat genotypes which represented five different continents were used. These genotypes were genotyped using 36,720 genotyping-by-sequencing derived SNPs (GBS-SNPs) which were well distributed across wheat chromosomes. The tested 103-wheat genotypes contained three different subpopulations based on population structure, principle coordinate, and kinship analyses. A significant variation was found within and among the subpopulations based on the AMOVA. Subpopulation 1 was found to be the more diverse subpopulation based on the different allelic patterns (*Na*, *Ne*, *I*, *h*, and *uh*). No high linkage disequilibrium was found between the 36,720 SNPs. However, based on the genomic level, D genome was found to have the highest LD compared with the two other genomes A and B. The ratio between the number of significant LD/number of non-significant LD suggested that chromosomes 2D, 5A, and 7B are the highest LD chromosomes in their genomes with a value of 0.08, 0.07, and 0.05, respectively. Based on the LD decay, the D genome was found to be the lowest genome with the highest number of haplotype blocks on chromosome 2D.

Conclusion: The recent study concluded that the 103-spring wheat genotypes and their GBS-SNP markers are very appropriate for GWAS studies and QTL-mapping. The core collection comprises three different subpopulations. Genotypes in subpopulation 1 are the most diverse genotypes and could be used in future breeding programs if they have desired traits. The distribution of LD hotspots across the genome was investigated which provides useful information on the genomic regions that includes interesting genes.

Keywords: Linkage disequilibrium, Haplotype blocks, Genome-wide association study, Analysis of molecular variance, Genotype-by-sequencing

* Correspondence: amira_mourad@aun.edu.eg
[1]Department of Agronomy, Faculty of Agricultural, Assuit University, Asyut, Egypt
Full list of author information is available at the end of the article

Background

Wheat (*Triticum aestivum* L.) is one of the most important cereal crops globally. It feeds more than a third of the human population around the world. The genome of bread wheat is an allohexaploid which contains three different genomes A, B, and D [1–3]. Generally, the genetic analysis of the wheat genome is very complex due to the polyploidy nature and the large genome size. The wheat genome is larger than *Arabidopsis thaliana* (~ 120 times), and *Oryza sativa* L. (~ 40 times) [4–6]. To well understand the complexity of the wheat genome, it is required to use good type of molecular markers which reduces the size of this genome by digesting it to multiple parts using restriction enzymes.

Generally, there are many types of molecular markers which could be used in various genetic analysis such as genetic diversity, genome-wide association studies, fingerprinting, evolutionary origin, and breeding applications. The most common type of markers is single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) [7]. However, by comparing SNPs and SSR markers, it was found that SNPs are excellent markers for studies that require a high number of markers such as association studies, QTL mapping, population structure, and genomic selection [8–12]. Recently, new techniques of sequencing have been developed to produce high-density genome-wide markers. Genotyping-by-sequencing (GBS) is one of these techniques which uses two different types of restriction enzymes (*PstI/MspI*) to reduce the complexity of large genomes such as wheat [13, 14]. Using the GBS technique provides many advantages such as; low cost, fewer purification steps, and easy sample handling [15].

Understanding the linkage disequilibrium (LD) between marker pairs is very important in association mapping studies as it determines the resolution of the association [16]. For example, if the LD rapidly decays, the resolution of the association will be high and vice versa [17]. Many previous studies discussed the relationship between LD decay and the resolution of association mapping in the wheat genome using different kinds of markers such as SSR and DArT and found that the LD varied among different wheat populations [18–21]. To achieve a high-resolution association mapping, a large number of markers should be used. GBS method produces such a high number of markers distributed across the genome.

As wheat is one of the most important crops globally, it is very important to study the global genetic variation. This requires the collection of cultivars from different countries. The USDA-ARS national plant germplasm system is a good resource for plant breeders worldwide as it contains a large number of accessions of wheat (~ 58,000) which have been collected starting from 1897. In 1995, the number of NSGC core accessions has been reduced to only 10% of the total number of the collected accessions following Brown 1989 [22] outline as described in Bonman et.al [23].. Following this outline, a collection of wheat accessions from all countries has resulted. This core collection, or a sample from it, could be considered as an ideal collection to study the genetic diversity of worldwide wheat germplasm. Consequently, understanding the genetic diversity in wheat germplasm is critical in breeding programs as it enables the wheat breeders to select the appropriate parents for the different breeding purposes. It is also very important in further breeding studies such as marker-assisted selection (MAS), genome-wide association studies (GWAS), and genomic selection. In the current study 103 spring genotypes representing 14 countries were collected from USDA gene bank and tested for their agronomic traits under the Egyptian conditions to increase the genetic diversity of adapted wheat genotypes in Egypt.

The objectives from this study were to (1) understand the genetic diversity and population structure in spring wheat using 103-accessions representing different countries worldwide, (2) compare the genetic properties among subpopulations, and (3) determine the patterns of linkage disequilibrium (LD).

Results

Distribution of SNP markers across the different wheat genomes

The total number of GBS derived SNPs from the tested genotypes was 287,798 SNPs. After quality filtering, the total number of high-quality SNPs was 36,720 which were well distributed across the genome (Fig. 1). The highest number of SNPs was located on genome B with a percentage of 41% (15,172 SNPs) while, the lowest number of SNPs located on genome D with a percentage of 19% (7119 SNPs). There were 1161 SNPs located within scaffolds with an unknown chromosomal location. The number of SNPs/chromosome (Chro.) ranged from 367 SNPs (4D Chro.) to 2764 SNPs (2B Chro.).

Genetic diversity and the polymorphism information content (PIC)

The PIC value across chromosomes ranged from 0.1 (1598 SNPs) to 0.4 (6836 SNPs) with an average of 0.24 (Fig. 2a). Gene diversity (GD) ranged from 0.1 (829 SNPs) to 0.5 (10,554 SNPs) with an average of 0.29. The percentage of heterozygosity extended from 0% (842 SNPs) to 100% (18 SNPs) with an average of 0.15, respectively (Fig. 2b and c). Minor allele frequency ranged from 0.1 (10,286 SNPs) to 0.5 (4384 SNPs) with an average of 0.21 (Fig. 2d).

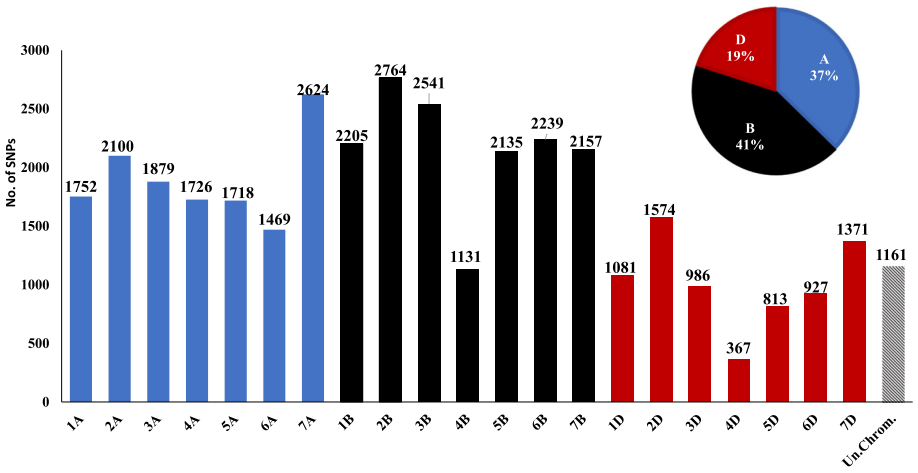

Fig. 1 The distribution of the 36,720 SNPs across the 21 chromosomes in the 103-spring wheat panel

Population structure and relationships

The STRUCTURE analysis software was used to identify the number of subpopulations in the tested 103 genotypes (Fig. 3). The number of clusters (K) was plotted against ΔK to identify the suitable number of subpopulations. The largest ΔK value was observed at $K = 3$ suggesting the presence of three subpopulations in the tested genotypes (Fig. 3a and b). As illustrated in Fig. 3c, there is a continuous-gradual increase in the assessed log-likelihood with the increase in the number of K confirming the presence of three subpopulations in the tested genotypes with the highest probability. The three groups consist of 48, 46, and nine genotypes for the red, blue, and green group, respectively (Fig. 3 and Table 1). By comparing the results of STRUCTURE software and the principle coordinate analysis, we found that both are in agreement and dividing the tested genotypes into three groups (Fig. 4 a and b). Based on both analyses, the first group (48 genotypes) contained all of the genotypes from Australia, Germany, Greece, and Kenya while, the second subpopulation (46 genotypes) contained the genotypes from Algeria, Ethiopia, and Tunisia. The genotypes from the remaining countries such as Egypt, Afghanistan, Canada, Iran, Kazakhstan, Morocco, Saudi Arabia, and Oman were distributed among the three groups. For example, most of the Egyptian genotypes belonged to the first group except for six genotypes that belonged to the third group. The percentage of the membership of each country in the three subpopulations is presented in Table 2.

Significant genetic differentiation was found among the three subpopulations and expected heterozygosity

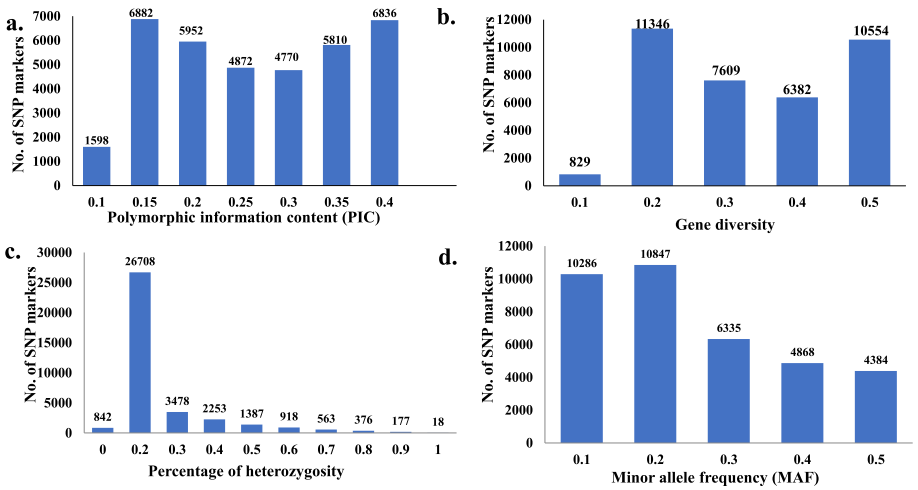

Fig. 2 The distribution of polymorphic information content (PIC) (**a**), gene diversity (**b**), percentage of heterozygosity (**c**), and minor allele frequency (**d**) for the 37,295 SNP markers in the 103-spring wheat panel

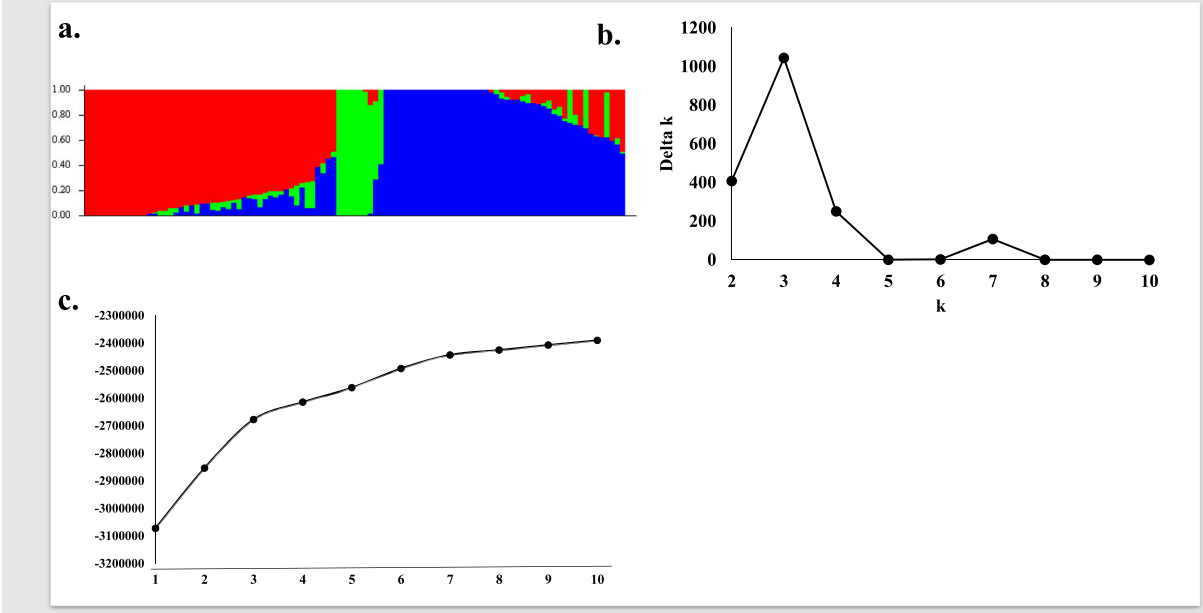

Fig. 3 Analysis of population structure using 36,720 SNP markers: (**a**) estimated population structure of 103-spring wheat genotypes (k = 3). The y-axis is the sub-population membership, and the x-axis is the genotypes, and (**b**) delta (**Δ**) K for different numbers of sub-populations, and (**c**) the average of log-likelihood value

(average distance) among genotypes in each subpopulation (Table 1). Subpopulation 1 had the highest value of expected heterozygosity with a value of 0.2671, followed by the third subpopulation (0.23526) and the second subpopulation (0.1776). The Fixation index (Fst) could be considered as the best index for the determination of the overall genetic variation among subpopulations. In our studied materials, the highest genetic variation was found in subpopulation 2 with the Fst value of 0.6142. While subpopulation 1 showed lower genetic variation among its genotypes with the Fst value of 0.1984 (Table 1). The analysis of kinship is illustrated as a genetic clustering and indicated that the current panel of genotypes was divided into three possible subgroups, with considerable genetic differences among the genotypes (Fig. 5).

Table 1 STRUCTURE analysis of 103-spring wheat genotypes for the fixation index (Fst) (significant divergences), average distance (expected heterozygosity) and number of genotypes in each subpopulation

Subpopulation	Fst [a]	Exp. Hetero [b]	No of genotypes
Subpopulation 1	0.1984	0.2671	48
Subpopulation 2	0.6142	0.1776	46
Subpopulation 3	0.3090	0.2325	9

[a]Fst is a measure of genetic differentiation; [b]Expected heterozygosity

Genetic differentiation of populations

The three subpopulations identified based on STRUCTURE analysis were used to calculate the AMOVA and genetic diversity indices in GenAlex 6.41 software. A significant variation within and among the subpopulations was found based on the AMOVA results. The total variation between the tested genotypes could be classified into two parts; variation among subpopulations with a percentage of 15%, and variation within subpopulations with a percentage of 85% (Table 3). The haploid number of migrants (Nm) was 2.90 indicating that there is a high gene exchange among subpopulations.

The allelic pattern across the populations

The average number of different alleles (Na) and effective alleles (Ne) were 2.528 and 1.781, respectively (Table 4). The Shannon index (*I*), the diversity index (*h*), and the unbiased diversity index (*uh*) had average values of 0.636, 0.384, and 0.403 based on the average of the three subpopulations (Table 4). Based on all allelic patterns, subpopulation 1 was the most diverse subpopulation when compared to subpopulations 2 and 3 as it has higher numbers of all the diversity indices. Subpopulation 3 was the least diverse subpopulation based on all indices as might be expected with its low number of lines. The percentage of polymorphic loci within subpopulations was 99.71, 99.39, and 64.84 for the first,

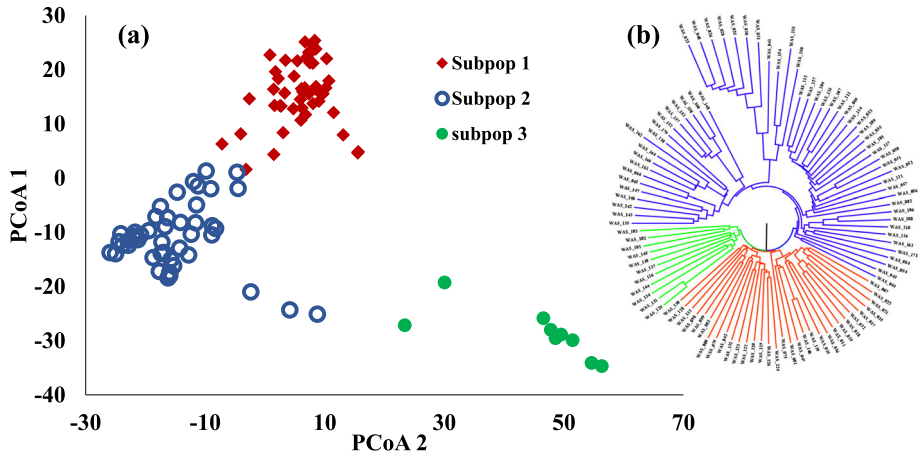

Fig. 4 **a** Principle coordinates analysis (PCoA) based on genetic distance (SNPs), **b** Dendrogram analysis based on the genetic distance calculated by UPGMA

second, and third subpopulation, respectively with an average of 87.99%.

Evaluation of linkage disequilibrium

The analysis of linkage disequilibrium showed that the LD decayed with the genetic distance (Supplementary Fig. 1). The values of R^2 revealed that there is no high LD among the 36,720 SNP pairs in the tested genotypes with an average value of 0.138 (Table 5). However, it was more useful to test the LD between each pair of SNPs located on the same chromosome and determine the average of the LD in each genome to identify the pattern of LD in the three genomes. Table 5 represents the average LD/chromosome and the number of significant and nonsignificant LD between each pair of SNPs located on the same chromosome. At the genome level, the highest LD was found in the D genome with an average of 0.1853, while the LD on both A and B genomes was almost the same with an average of 0.1189 and 0.1124, respectively. The LD within each genome ranged from 0.106 (1A) to 0.125 (4A), 0.098 (6B) to 0.122 (4B) and 0.167 (4D) to 0.241 (2D). The significance of LD between each SNP pair located on the same chromosome was tested using Bonferroni correction ($\alpha = 0.01$). The D Genome contained the highest significant LD based on the average of chromosomes with $R^2 = 0.887$ followed by genomes A and B with an average R^2 of 0.818 and 0.815, respectively. Likewise, the highest

Table 2 The percentage of the membership of each country in the three subpopulations

Country	Subpopulation 1	Subpopulation 2	Subpopulation 3	Number of genotypes
Afghanistan	11.11	88.89	0.00	9
Algeria	0.00	100.00	0.00	1
Australia	100.00	0.00	0.00	1
Canada	80.00	20.00	0.00	5
Egypt	64.71	0.00	35.29	17
Ethiopia	0.00	100.00	0.00	1
Germany	100.00	0.00	0.00	2
Greece	100.00	0.00	0.00	3
Iran	7.14	92.86	0.00	14
Kazakhstan	75.00	25.00	0.00	8
Kenya	100.00	0.00	0.00	5
Morocco	64.29	35.71	0.00	14
Oman	0.00	87.50	12.50	8
Saudi Arabia	28.57	71.43	0.00	7
Tunisia	0.00	100	0.00	1
Unknown countries	42.86	28.57	28.57	7

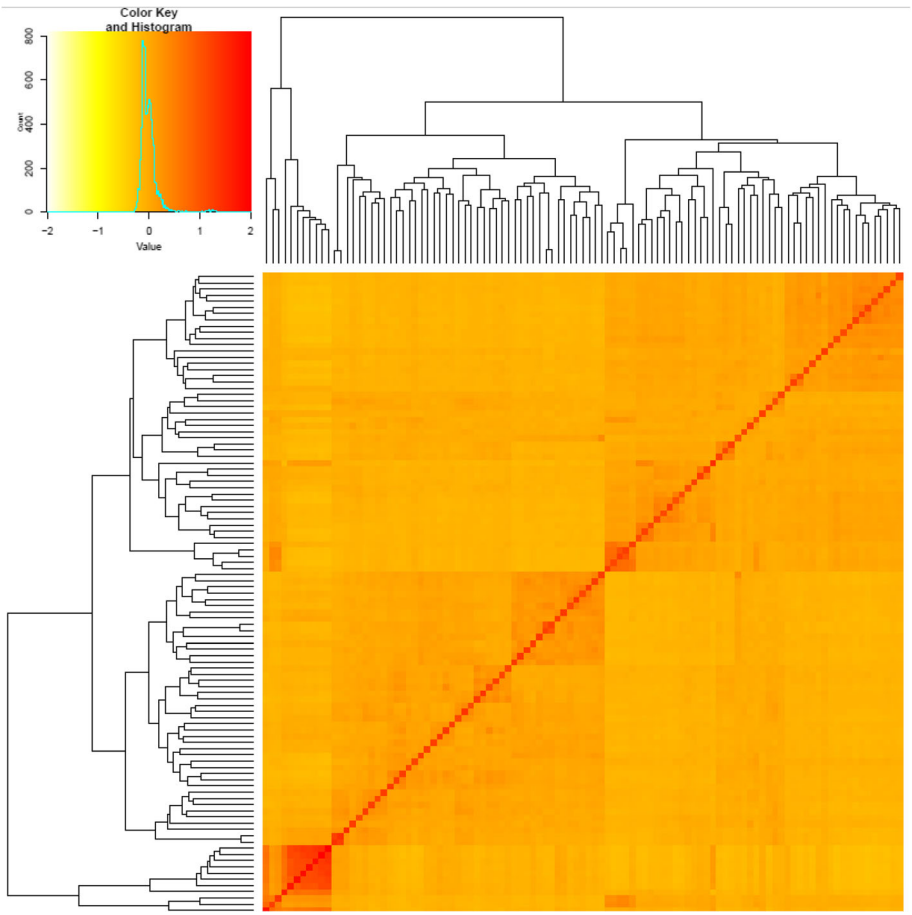

Fig. 5 Heat map of kinship matrix with the dendogram shown on the top and left based on the 36,720 SNP markers

LD as an average of all SNP pairs with non-significant LD was found in genome D (0.149), while the LD average of non-significant markers was approximately the same in genome A and B with an average of ~ 0.084.

The ratio between the number of significant LD and the number of nonsignificant LD could be arranged from higher to lower as follows; 0.06, 0.05, and 0.04 for genome D, genome A, and genome B respectively. At the chromosome level, chromosomes 2D, 5A, and 7B had the highest ratios between the number of significant and non-significant LD with values of 0.08, 0.07, and 0.05, respectively. The R^2 between each pair of markers was plotted against genetic distance (kb). The LD decay in each genome is illustrated in Fig. 6 and whole-genome in Supplementary Figure 1. The LD decay in the D genome was slower than the LD decay in A and B genomes. The LD decay in A genome was slower than the B genome (Fig. 6a-d). The number of haplotype blocks was investigated for the highest three chromosomes. Chromosome 2D was found to contain 28 haplotype blocks followed by

Table 3 Analysis of molecular variance using 36,720 SNPs and the genetic differentiation among the three subpopulations of the 103-spring wheat panel

Source	df	SS	MS	Est. Var.	%	*P* value
Among Pops	2	47,935.156	23,967.578	676.092	15	0.001
Within pops	100	392,111.058	3921.111	3921.111	85	0.001
Total	102	440,046.214		4597.203	100	0.001
Nm (haploid)	2.900					

Table 4 Mean of different genetic parameters including number of different alleles (*Na*), number of effective allele (*Ne*), Shannon's index (*I*), diversity index (*h*), unbiased diversity index (*uh*), and percentage of polymorphic loci (*PPL*) in each subpopulation of the 103-genotypes

Subpopulations	*Na*	*Ne*	*I*	*h*	*uh*	*PPL*
Subpopulation 1	2.897	1.994	0.782	0.471	0.482	99.71
Subpopulation 2	2.869	1.921	0.002	0.445	0.457	99.39
Subpopulation 3	1.816	1.429	0.380	0.236	0.271	64.87
Mean	2.528	1.781	0.636	0.384	0.403	87.99

Table 5 Linkage disequilibrium between SNP markers located on the same chromosome and genome

Chromosome	R^2	Number sig. LD	Average Sig. LD	Percentage of sig. R^2	Number non sig. LD	Average non sig. LD	No. of sig. LD/ No. of non sig. LD
1A	0.106696275	2673	0.773570652	4.6	55,965	0.074845024	0.05
2A	0.117889775	2973	0.79594235	4.8	58,919	0.083675849	0.05
3A	0.112887651	1876	0.852327861	3.4	54,032	0.087214164	0.03
4A	0.125257515	2590	0.862161693	4.6	53,148	0.089346816	0.05
5A	0.125428444	3419	0.809304824	6.2	52,153	0.080595484	0.07
6A	0.120074851	2829	0.794986633	5.9	44,846	0.077499696	0.06
7A	0.124468994	3482	0.835755958	3.9	86,668	0.095892112	0.04
mean	0.118957644	19,842	0.817721425	4.7	405,731	0.084152735	0.05
1B	0.114425037	2767	0.804864224	3.9	68,196	0.086411	0.04
2B	0.108414675	2979	0.821397105	3.4	85,494	0.083571122	0.03
3B	0.115633024	3343	0.797272582	4.0	80,596	0.087359648	0.04
4B	0.122410098	1520	0.837638581	4.0	36,103	0.092297717	0.04
5B	0.106133555	3151	0.828076529	4.2	72,350	0.074692834	0.04
6B	0.098446778	2543	0.799670654	3.4	72,669	0.073907947	0.03
7B	0.121483303	3397	0.814441598	4.8	67,784	0.086755649	0.05
mean	0.112420924	19,700	0.814765896	3.9	483,192	0.083570845	0.04
1D	0.186308632	1559	0.859202458	6.3	23,371	0.141422172	0.07
2D	0.240878007	2986	0.929159518	7.7	36,041	0.183853824	0.08
3D	0.206075532	1320	0.901496541	6.4	19,422	0.158811824	0.07
4D	0.16616349	239	0.905766016	3.7	6244	0.137853911	0.04
5D	0.178633759	505	0.89830723	3.6	13,699	0.152103713	0.04
6D	0.145398046	606	0.818621715	3.0	19,755	0.124746388	0.03
7D	0.173585725	1134	0.893450395	3.7	29,693	0.146093503	0.04
mean	0.185291884	8349	0.886571982	5.3	148,225	0.149269334	0.06
Genome mean	0.137984						

chromosome 5A and 7B which contain 12 and 11 blocks, respectively (Supplementary figure 2).

Discussion

The studied wheat genotypes were collected from different countries representing five of the world continents (Africa, Europe, Asia, North America, and Australia) which enable us to estimate wheat genetic diversity in the studied countries. The study was conducted using 36,720 SNPs which were well distributed across the three hexaploid wheat genomes (A, B, and D). The highest number of SNPs were found on genome B (41%), while the lowest number of SNPs were found on genome D (19%) indicating that genome D is the least diverse wheat genome (Fig. 1). The D genome was reported to be the least diverse genome in previous studies which used different types of markers such as GBS-SNPs, RFLP, SSR, AFLP, and DArT markers [24–30]. Dubcovsky and Dvorak [1] concluded that the proportion of diversity in *Triticum aestivum* L. resulted in the polyploid nature of its tetraploid ancestor with AABB. This conclusion could be a good explanation of the high level of diversity among hexaploid wheat genotypes and the high number of SNPs in the A and B genomes.

The PIC values and genetic diversity are very helpful parameters to measure the polymorphism between the genotypes used in breeding programs. Generally, for multi-locus markers such as SSR markers, the PIC values range from 0 to 1.0. According to Botstein et.al [31, 32], multi-allelic markers could be classified into three categories based on their PIC values. These three categories are: (1) highly informative markers with PIC values higher than 0.5, (2) moderately informative marker with PIC value ranging from 0.25 to 0.5, and (3) slightly informative markers with PIC values less than 0.25. However, for the bi-allelic markers like SNPs, the highest PIC value is 0.5. As a result of this bi-allelic nature, SNP markers could be considered as moderate to low informative markers.

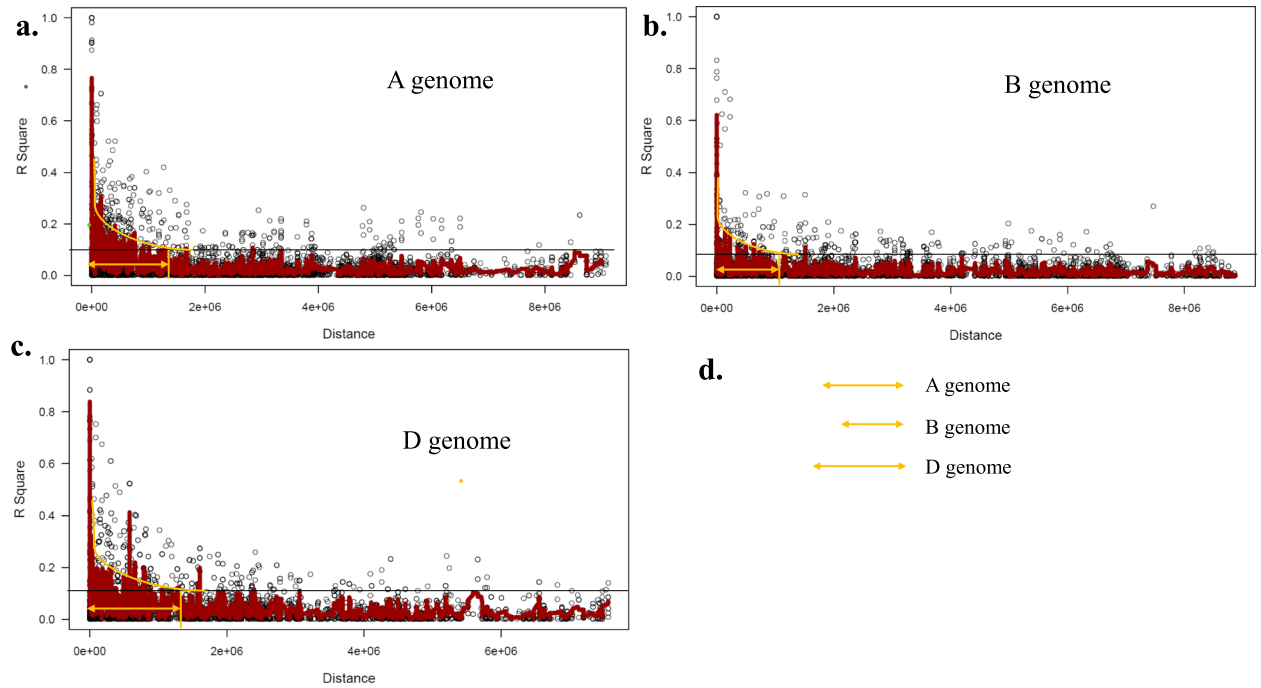

Fig. 6 The rate of linkage disequilibrium (LD) decay of the genome A (**a**.), genome B (**b**.) and genome D (**c**.) of the 103-spring wheat based on the 36,720 SNP markers. (**d**.) comparison between LD decay (distance) among the three genomes

The average PIC value obtained in this study is 0.24 which is similar to PIC values in previous studies [24, 33]. This PIC value was reported as a good indicator of informative markers which could be used in studying the genetic diversity in the different organisms [34]. Based on the PIC values in our tested population and the good distribution of the studied SNP markers, we can conclude that these markers explained the genetic diversity in spring wheat and could be used in other genetic studies such as genome-wide association study to identify alleles controlling target traits.

Population structure and relationships

Studying the population structure is very helpful in understanding the genetic diversity of the tested genotypes. This is the first step in conducting the association mapping studies. In our tested materials, STRUCTURE analysis, as well as the PCoA, confirmed the presence of three subpopulations. In each subpopulation, there were genotypes from different countries and continents. This result was expected due to the continuous gene flow of wheat genotypes among the different countries historically to the present. This exchange resulted in the presence of diverse genetic backgrounds in the same country, and thus the presence of genotypes from different countries in the same subpopulation. The majority of genotypes from Afghanistan, Iran, Oman, and Saudi Arabia were clustered together in subpopulation 2, while, the majority of genotypes from Egypt, Canada, Kazakhstan, and Morocco were grouped in subpopulation 1. Genotypes from Germany, Greece, and Kenya were grouped in only one subpopulation. This information is very important in selecting the candidate parents for target traits in breeding programs as genetic distance should be highly considered. Genetic diversity between two genotypes from two different countries representing two different continents may be very low and not useful to use such parents in breeding programs. Understanding this presence of population structure in the tested 103-spring wheat genotypes is very important. It must be taken into account before conducting genome-wide association studies (GWAS) as it could result in a superior association between the studied trait and the GBS-derived SNPs [35].

Genetic differentiation of populations

The result of AMOVA indicated the presence of highly significant genetic diversity among the three subpopulations (Table 2). The high level of genetic diversity within the subpopulations could be due to the selection for specific traits that have been done by the wheat breeders in the different countries for specifically targeted traits. In addition, each subpopulation had wheat genotypes from different countries. While the low level of genetic diversity among the populations (15%) could be due to gene flow resulted from the wheat germplasm flow among the different regions. Therefore, it could be more useful to

select genotypes as parents, in the breeding programs for improving target traits, from the same subpopulation than selecting from different subpopulations. However, this may change depending on the breeding goals. Making crosses among genotypes from different subpopulations may be required to incorporate haplotypes from different founder populations. Similar results of high genetic diversity within the subpopulations and low diversity among the subpopulations were found in winter and synthetic wheat genotypes [24, 36]. In order to identify the level of gene flow among the subpopulation, Nm (haploid) was calculated. It was reported that, if the Nm (haploid) value was 1.00 or lower, this indicates the low level of gene flow [37]. In our tested materials, Nm (haploid) was 2.900 which is much higher than 1.00 indicating the high level of gene flow between the subpopulations. This result supports the distribution of the genotypes from one country in the three subpopulations in the tested plant material.

Based on all the allelic pattern indices (*Na*, *Ne*, *I*, *h*, *uh*, and *PPL*) among the three subpopulations, subpopulation 1 is the most diverse subpopulation as it shows the highest values of all the indices. This result is expected as this subpopulation contains genotypes from 11 different countries compared with the other two subpopulations which contain genotypes from ten and two different countries, respectively (Supplementary Table 1). Based on these results, we can conclude that the studied 103-spring wheat genotypes, especially subpopulation 1, provide a broad and useful source of genetic diversity in wheat. This set of genotypes could be used in future breeding programs to increase the genetic diversity among wheat genotypes. Increasing genetic diversity is very useful in conducting genome-wide association studies (GWAS) and marker-assisted selection (MAS) for identifying genes controlling important traits in wheat. Moreover, selection among the genotypes in subpopulation 1 for target traits will be fruitful for the genetic improvement of wheat.

Linkage disequilibrium and kinship between the studied genotypes

The determination of the LD magnitude and decay is very important as they affect the resolution of association mapping and the SNP markers required for conducting association studies [16]. The extent of LD differs across genomes in many species. As wheat has three different genomes, we analyzed the LD decay in each genome. The LD decay was estimated when the values of LD declined below 0.1 based on the curve of the nonlinear logarithmic trend line. The LD decayed in genome D at higher distances than genomes A and B. The lowest rate of LD decay was observed in genome B. This result suggested that fewer markers are needed to detect target QTLs located on genome D using GWAS than those needed for detecting QTLs on the other genomes [38]. By looking at the number of markers in each genome, the D genome had the lowest number of SNPs (20%) followed by genomes A and B, respectively. This indicates that the current set of material and SNPs are very appropriate to conduct GWAS to identify alleles associated with target traits in wheat. The high and low LD found across the three genomes provide a high chance to detect target QTL with large and small effects in the current materials [39]. The same results of LD decay pattern across the three genomes of wheat were reported by Liu et al. (2017) and Ayana et al. (2018) [38, 40].

Interestingly, high LD regions at a high genetic distance were observed in each genome. These high LD regions which were among low LD regions are called LD hotspots. Visibly, LD hotspot regions in genome A and D were higher than those in genome B. According to the LD significance level between the markers, the ratio of the number of significant to non-significant markers was higher in genome D (6%) and A (5%) than in genome B (4%) which means that genome B had the highest number of markers in non-significant LD (Table 5). Therefore, it is very important to understand the structure of LD in the wheat population and the distribution of LD hotspot regions in each genome. Understanding the LD structure enables to identify the genetic regions associated with agronomic traits and determining the density of markers needed to associate the genotypes with the studied traits [16].

The pattern and number of LD hotspots in the genome provide useful information in determining marker density. The greater the recombination rate, the greater the need for higher marker density as the greater chance for the LD to be broken by a recombination event when QTL and the marker are close together [41]. By looking at the LD plot including the three genomes (supplementary Figure 2), hotspots genomic regions were clearly found at a high genetic distance and separated the low LD regions.

Conclusion

In conclusion, the analysis of population structure and LD decay were genetically dissected in a set of 103 wheat core collection genotypes from different countries. The current material was divided into three possible subpopulations. The most diverse genotypes were found in subpopulation 1 and they can be used in the future breeding program by crossing among parents with target traits. The population structure was also very useful to determine the appropriate GWAS statistical methods that can be used to detect QTLs in these populations. Moreover, the genetic diversity of markers in the current population suggested that the markers are informative and polymorphic. The genetic properties of this population including the number of genotypes and SNP

markers allow this population to be used for further genetic studies to genetically improve spring wheat through advanced breeding programs. We identified the distribution of LD hotspots across each genome and the whole genome which provided useful information on the possibility of genomic regions that may include interesting genes.

Methods

Plant materials

A set of 103 spring bread wheat genotypes were obtained from the USDA-ARS worldwide core collection and used in this study. The genotypes are representing fourteen different countries (Table 2). Out of the 103 tested genotypes, fifteen are local varieties that are usually planted in Egypt. The remaining 88 genotypes were from other countries, evaluated in Egypt, and found to perform well. Hence those 88 genotypes were adapted to the Egyptian environmental conditions and would be used as global parents for cultivar improvement.

DNA extraction, genotype-by-sequencing (GBS), and SNP calling

DNA was extracted from all the tested genotypes from 2 to 3 leaves of 2 weeks old seedlings using BioSprint 96 DNA Plant Kits (Qiagen, Hombrechtikon, Switzerland). The extracted DNA was digested for GBS purpose using two different restriction enzymes, *PstI* and *MspI*. The Illumina, Inc. NGS platforms were used to generate the sequencing of the pooled libraries. TASSEL 5.0 v2 software GBS pipeline was used to identify SNPs [42]. Chinese Spring genome v1.0 from the International Wheat Genome Sequencing Consortium (IWGSC) was used as a reference genome for SNP calling and GBS tags were aligned using Burrows-Wheeler Aligner [43]. Generated SNPs were filtered for minor allele frequency (MAF) less than 5%, missing data less than 20%, missing genotypes less than 30%, and maximum heterozygosity 35%.

Data analysis

Genetic properties of markers

Genetic diversity statistics of all the 36,720 SNP markers such as polymorphic information content (PIC), gene diversity, percentage of heterozygosity, and minor allele frequency (MAF) were calculated using PowerMarker software V 3.25 [44]. The following formula was used to calculate the PIC according to [31].

$$PIC = 1 - \sum_{j=1}^{n} P_{ij}^2 - \sum_{j=1}^{n=1} \sum_{k=j+1}^{n} 2P_{ij}^2 P_{ik}^2$$

Where P_{ij} and P_{ik} are the frequencies of j_{th} and k_{th} alleles for marker i, respectively.

Analysis of population structure

To estimate the number of subpopulations in the current tested genotypes, a model-based (Bayesian) method with the filtered SNPs (36,720) was used. STRUCTURE 3.4.0 software was used to analyze population structure [45]. Burn-in iteration was 100,000 followed by 100,00 Markov chain Monte Carlo (MCMC) replications after burn-in for each run. In this analysis, allele frequencies and the admixture correlated models were considered. Five impended iterations were used in the STRUCTURE. The hypothetical number of the subpopulation (k) extended from 1 to 10. STRUCTURE HARVESTER [46] was used to identify the best k, where k is the number of subpopulations [47]. The genetic distance among the tested genotypes was calculated using TASSEL v.5.2.5 software [42]. Based on this genetic distance, principal coordinate analysis (PCoA) was performed.

Analysis of molecular variance (AMOVA) and genetic diversity indices

For this analysis, 14,400 SNPs based on the highest PIC values (from 0.282 to 0.375) were used. The number of subpopulations based on the STRUCTURE analysis was considered in the AMOVA. The genetic indices such as fixation index (F_{st}), different alleles (*Na*), number of effective alleles (*Ne*), Shannon's index (*I*), the diversity index (*h*), the unbiased diversity index (*uh*), and percentage of polymorphic loci (*PPL*) were calculated. The AMOVA and estimation of genetic indices were performed using GenAlex 6.41 [48].

Linkage disequilibrium (LD) structure

The LD between each pair of the 36,720 SNPs was calculated as the squared allele frequency correlation coefficient (R^2) using TASSEL v.5.2.5 software [42]. The LD was calculated separately for each chromosome in each genome (A, B, and D) in order to understand the structure of LD in the current population. To identify the significant LD, Bonferroni correction ($\alpha = 0.01$) was applied [12]. The kinship matrix between the tested genotypes as well as the LD decay for each genome was calculated using GAPIT, R package [49].

Haplotype block analysis

In each genome, the chromosome contains the highest significant LD percentage was tested for the number of haplotype blocks using Haploview 4.2 software [50]. To perform this, the SNP data for the target chromosome was used to calculate the pair-wise LD between SNPs. The haplotype blocks were constructed using the four-gamete method and a cutoff 1% was used [51, 52].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06835-0.

Additional file 1: Table S1. List of genotypes in each subpopulation and their country.

Additional file 2: Figure S1. The rate of linkage disequilibrium (LD) decay of the 103-spring wheat based on the 36,720 SNP markers.

Additional file 3: Figure S2. Haplotype block analysis represents the number of haplotype blocks on: (a) chromosome 2D, (b) chromosome 5A, and chromosome 7B.

Abbreviations

GBS: Genotyping-by-sequencing; GWAS: Genome-wide association study; MAS: Marker-assisted selection; LD: Linkage disequilibrium

Acknowledgments

The authors would like to thank Dr. Ahmed Sallam, Associate Professor, Department of Genetics, Faculty of Agriculture, Assuit University, Asyut, Egypt, for his help and support in data analysis and discussing the results.

Authors' contributions

AM conducting all the data analysis and drafting the manuscript. VB performed the SNP calling from GBS data. PSB helped in drafting the manuscript. All the authors agreed to be accountable for the content of the work and revised the manuscript.

Funding

This work was financially partial supported by the cultural affairs and mission sector, the Egyptian government.

Availability of data and materials

Sequence data is available with the authors. The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Agronomy, Faculty of Agricultural, Assuit University, Asyut, Egypt. [2]Department of Agronomy and Horticulture, Plant Science Hall, UNL, Lincoln, NE, USA.

Received: 13 March 2020 Accepted: 16 June 2020
Published online: 26 June 2020

References

1. Dubcovsky J, Dvorak J. Genome plasticicty a key factor in the success of polyploid wheat under domestication. Science (80-). 2007;316:1862–6. https://doi.org/10.1126/science.1094305.
2. Shewry PR. Wheat. J Exp Bot. 2009;60:1537–53. https://doi.org/10.1093/jxb/erp058.
3. Matsuoka Y. Evolution of polyploid triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol. 2011;52:750–64.
4. Bennett MD, Smith JB. Nuclear DNA amounts in angiosperms. Philos Trans R Soc B Biol Sci. 1976;274:227–74. https://doi.org/10.1098/rstb.1976.0044.
5. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Report. 1991;9:208–18. https://doi.org/10.1007/BF02672069.
6. Sorrells ME, La Rota M, Bermudez-kandianis CE, Greene RA, Kantety R, Munkvold JD, et al. Comparative DNA sequence analysis of wheat and rice genomes. Genome Res. 2003;13:1818–27.
7. Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 2002;5:94–100.
8. Kumar S, Banks TW, Cloutier S. SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics. 2012. https://doi.org/10.1155/2012/831460.
9. Mourad AMI, Sallam A, Belamkar V, Mahdy E, Bakheit B, El-wafaa AA, et al. Genetic architecture of common bunt resistance in winter wheat using genome- wide association study. BMC Plant Biol. 2018;18:1–14.
10. Mourad AMI, Alomari DZ, Alqudah AM, Sallam A, Salem KFM. Recent Advances in wheat (Triticum spp.) breeding. In: Advances in planr breeding strategies: cereals; 2019. p. 559–93.
11. Hussain W, Baenziger PS, Belamkar V, Guttieri MJ, Jorge P, Easterly A, et al. Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. Sci Rep. 2017;7 April:1–15. https://doi.org/10.1038/s41598-017-16006-z.
12. Sallam A, Martsch R. Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (*Vicia faba* L.). Genetica. 2015;143:501–14. https://doi.org/10.1007/s10709-015-9848-z.
13. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379. https://doi.org/10.1371/journal.pone.0019379.
14. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome. 2012;5:103–13.
15. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510. https://doi.org/10.1038/nrg3012.
16. Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003;54:357–74.
17. Rafalski AJ. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 2002;162:329–33.
18. Breseghello F, Sorrells ME. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics. 2006;172:1165–77. https://doi.org/10.1534/genetics.105.044586.
19. Maccaferri M, Sanguineti MC, Noli E, Tuberosa R. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breed. 2005;15:271–89.
20. Chao S, Zhang W, Dubcovsky J, Sorrells M. Evaluation of genetic diversity and genome-wide linkage disequilibrium among U.S. wheat (Triticum aestivum L.) germplasm representing different market classes. Crop Sci. 2007;47:1018–30.
21. Somers DJ, Banks T, DePauw R, Fox S, Clarke J, Pozniak C, et al. Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome. 2007;50:557–67.
22. Brown AHD. Core collections: a practical approach to genetic resources management. Genome. 1989;31:818–24.
23. Bonman JM, Babiker EM, Cuesta-Marcos A, Esvelt-Klos K, Brown-Guedira G, Chao S, et al. Genetic diversity among wheat accessions from the USDA national small grains collection. Crop Sci. 2015;55:1243–53.
24. Eltaher S, Sallam A, Belamkar V, Emara HA, Nower AA, Salem KFM, et al. Genetic diversity and population structure of F3:6 Nebraska Winter wheat genotypes using genotyping-by-sequencing. Front Genet. 2018;9 MAR:1–9.
25. Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo M-C, et al. Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol Breed. 2009;23:23–33. https://doi.org/10.1007/s11032-008-9210-6.
26. Nielsen NH, Backes G, Stougaard J, Andersen SU, Jahoor A. Genetic diversity and population structure analysis of European hexaploid bread wheat (Triticum aestivum L.) varieties. PLoS One. 2014;9:e94000. https://doi.org/10.1371/journal.pone.0094000.
27. Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, et al. A microsatellite map of wheat. Genetics. 1998;149:2007–23.
28. Y-G LIU, TSUNEWAKI K. Restriction fragment length polymorphism (RFLP) analysis in wheat. II. Linkage maps of the RFLP sites in common wheat. Japanese J Genet. 1991;66:617–33. https://doi.org/10.1266/jjg.66.617.
29. Peng J, Korol AB, Fahima T, Röder MS, Ronin YI, Li YC, et al. Molecular genetic maps in wild emmer wheat, Triticum dicoccoides: genome-wide

coverage, massive negative interference, and putative quasi-linkage. Genome Res. 2000;10:1509–31. https://doi.org/10.1101/gr.150300.
30. Alipour H, Bihamta MR, Mohammadi V, Peyghambari SA, Bai G, Zhang G. Genotyping-by-sequencing (GBS) revealed molecular genetic diversity of Iranian wheat landraces and cultivars. Front Plant Sci. 2017;8:1293. https://doi.org/10.3389/fpls.2017.01293.
31. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32:314–31 http://www.ncbi.nlm.nih.gov/pubmed/6247908. Accessed 15 June 2019.
32. Robbana C, Kehel Z, Ben Naceur M, Sansaloni C, Bassi F, Amri A. Genome-wide genetic diversity and population structure of tunisian durum wheat landraces based on DArTseq technology. Int J Mol Sci. 2019;20:1352.
33. Allen AM, Barker GLA, Berry ST, Coghill JA, Gwilliam R, Kirby S, et al. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). Plant Biotechnol J. 2011;9:1086–99.
34. Salem KFM, Sallam A. Analysis of population structure and genetic diversity of Egyptian and exotic rice (Oryza sativa L.) genotypes. C R Biol. 2015;339:1–9. https://doi.org/10.1016/j.crvi.2015.11.003.
35. Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN, editors. Association mapping in plants. New York, NY: Springer New York; 2007. https://doi.org/10.1007/978-0-387-36011-9.
36. Bhatta M, Morgounov A, Belamkar V, Poland J, Baenziger PS. Unlocking the novel genetic diversity and population structure of synthetic Hexaploid wheat. BMC Genomics. 2018;19:1–12.
37. Wright S. The interpretation of population structure by F-statistics with special regard to system of mating. Evolution (N Y). 1965;19:395–420. https://doi.org/10.1111/j.1558-5646.1965.tb01731.x.
38. Liu J, He Z, Rasheed A, Wen W, Yan J, Zhang P, et al. Genome-wide association mapping of black point reaction in common wheat (Triticum aestivum L.). BMC Plant Biol. 2017;17:1–12.
39. Würschum T, Maurer HP, Kraft T, Janssen G, Nilsson C, Reif JC. Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet. 2011;123:1121–31.
40. Ayana GT, Ali S, Sidhu JS, Gonzalez Hernandez JL, Turnipseed B, Sehgal SK. Genome-wide association study for spot blotch resistance in hard winter wheat. Front Plant Sci. 2018;9 July:1–15.
41. Larmer SG, Sargolzaei M, Schenkel FS. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J Dairy Sci. 2014;97:3128–41.
42. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
43. Li H, Durbin R. Fast and accurate short read alignment with burrows – wheeler transform. Bioinformatics. 2009;25:1754–60.
44. Liu K, Muse SV. PowerMaker: an integrated analysis environment for genetic maker analysis. Bioinformatics. 2005;21:2128–9.
45. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
46. Earl DA, von Holdt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4:359–61.
47. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol. 2005;14:2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x.
48. PEAKALL R, SMOUSE PE. Genalex 6: genetic analysis in excel. Population genetic software for teaching and research. Mol Ecol Notes. 2006;6:288–95. https://doi.org/10.1111/j.1471-8286.2005.01155.x.
49. Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. A SUPER powerful method for genome wide association study. PLoS One. 2014;9:e107684.
50. Barrett JC, Fry B, Maller J, Daly MJ. Haploview : analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
51. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks : the interplay of population history , recombination , and mutation. Am J Hum Genet. 2002;71:1227–34.
52. Mourad AMI, Sallam A, Belamkar V, Wegulo S, Bowden R, Jin Y, et al. Genome-wide association study for identification and validation of novel SNP markers for Sr6 stem rust resistance gene in bread wheat. Front Plant Sci. 2018;9 March:1–12. https://doi.org/10.3389/fpls.2018.00380.

Publisher's Note