DNA Research, 2018, 25(3), 317–326
doi: 10.1093/dnares/dsy004
Advance Access Publication Date: 21 February 2018
Full Paper

Full Paper

An efficient approach for the development of genome-specific markers in allohexaploid wheat (*Triticum aestivum* L.) and its application in the construction of high-density linkage maps of the D genome

Goro Ishikawa[1,2,*,†], Mika Saito[1,†], Tsuyoshi Tanaka[2], Yuichi Katayose[2], Hiroyuki Kanamori[2], Kanako Kurita[2], and Toshiki Nakamura[1,*]

[1]Tohoku Agricultural Research Center, National Agriculture and Food Research Organization (NARO), Morioka, Iwate 020-0198, Japan, and [2]Institute of Crop Science, National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki 305-8518, Japan

*To whom correspondence should be addressed. Tel. +81 19 643 3514. Fax. +81 19 643 3514. Email: goro@affrc.go.jp (G.I.), tnaka@affrc.go.jp (T.N.)
[†]Co-first authors.

Edited by Prof. Kazuhiro Sato

Received 23 October 2017; Editorial decision 16 January 2018; Accepted 16 February 2018

Abstract

In common wheat, the development of genotyping platforms has been hampered by the large size of the genome, its highly repetitive elements and its allohexaploid nature. However, recent advances in sequencing technology provide opportunities to resolve these difficulties. Using next-generation sequencing and gene-targeting sequence capture, 12,551 nucleotide polymorphisms were detected in the common wheat varieties ‘Hatsumochi’ and ‘Kitahonami’ and were assigned to chromosome arms using International Wheat Genome Sequencing Consortium survey sequences. Because the number of markers for D genome chromosomes in commercially available wheat single nucleotide polymorphism arrays is insufficient, we developed markers using a genome-specific amplicon sequencing strategy. Approximately 80% of the designed primers successfully amplified D genome-specific products, suggesting that by concentrating on a specific subgenome, we were able to design successful markers as efficiently as could be done in a diploid species. The newly developed markers were uniformly distributed across the D genome and greatly extended the total coverage. Polymorphisms were surveyed in six varieties, and 31,542 polymorphic sites and 5,986 potential marker sites were detected in the D genome. The marker development and genotyping strategies are cost effective, robust and flexible and may enhance multi-sample studies in the post-genomic era in wheat.

Key words: *Triticum aestivum*, allohexaploid, next-generation sequencing, amplicon sequencing, wheat D genome

1. Introduction

High-throughput genotyping platforms are essential tools for various genetic studies that involve genetic mapping, genome-wide association studies, phylogenetic analyses, marker-assisted selection (MAS) and genomic selection.[1,2] Recent advances in sequencing technologies have greatly facilitated the discovery of polymorphisms in the whole genome. The 9K iSelect[3] and 90K iSelect[4] single-nucleotide polymorphism (SNP) arrays have been developed for allohexaploid wheat (*Triticum aestivum* L.) by transcriptome sequencing using next-generation sequencing (NGS) technology. These arrays comprise many gene-based SNPs that allow an individual plant to be genotyped at all sites simultaneously and tend to be robust marker platforms. Therefore, these arrays have been widely used for germplasm characterization and quantitative trait locus (QTL) mapping.[5–12]

However, array-based markers are inflexible and have a relatively high per-sample cost. The number of polymorphisms detected by the markers on an array chip tends to substantially decrease in lines that are genetically distant from those used to design the array. When the 9K iSelect array, which was designed based on SNP information of American and Australian varieties, is used in Japanese varieties, a low polymorphic rate is detected, particularly in varieties grown in the southern regions (latitude 33–37°N).[13] In this region, the wheat harvest season overlaps with the rainy season, and mainly domestic genetic resources have been used to breed varieties with resistance to pre-harvest sprouting and *Fusarium* head blight. On the other hand, in northern regions from Hokkaido (42–45°N) to Tohoku (37–41°N), there is less rain during the harvest season and several North American varieties were introduced to improve flour quality. In many cases, the polymorphisms detected in American and Australian varieties are not conserved in Japanese varieties. Furthermore, when Axiom HD wheat genotyping arrays (Affymetrix, Santa Clara, CA, USA) were used in Japanese materials, only 3.1% of the markers were categorized as 'PolyHighResolution', which defines markers with good cluster resolution and at least two examples of the minor allele (unpublished). However, a different set of markers in this array were of high quality and were more polymorphic in Japanese materials than in samples from varieties processed by the WISP (http://www.wheatisp.org/ (5 February 2018, date last accessed)) in the design of the array. Thus, the usefulness of array-based markers largely depends on the source of the polymorphisms. Additionally, genotype calls remain complicated because of the polyploid nature of wheat, and allele data should be interpreted with caution. These disadvantages are deleterious for MAS in wheat breeding programs because a cost-effective genotyping platform with a rapid turnaround time, low per-sample cost and very low rate of missing data is required for efficient MAS. Furthermore, MAS usually relies on a set of specific markers for specific QTLs and genes using a medium throughput system, rather than random genes or markers.

Available array platforms are limited in their application to the D genome, because the genetic coverage of the D genome is highly inadequate.[14] This tendency is also observed using other whole-genome genotyping platforms, such as DArT and genotyping by sequencing (GBS), which also have comparatively fewer D genome markers.[15–17] Marcussen et al. reported that the origin of the D genome was 1–2 million years later than that of the A and B genomes.[18] Furthermore, the D subgenome of hexaploid wheat was established by polyploidization ~0.4 million years later than the other subgenomes. The shorter time for accumulation of nucleotide polymorphisms induced by natural mutations might influence the scarcity of markers in the D genome. Therefore, to conduct studies focused on chromosomes in the D genome, new markers must be developed. However, the discovery of polymorphisms in wheat is hampered by its large genome size (16 Gb) and high repeat content (approximately 80%).[19] Although the cost of sequencing is decreasing, sequencing the whole genome remains prohibitively expensive, particularly in species with large genomes. Henry et al.[20] reported that exome capture combined with NGS can be used to successfully and efficiently detect Ethyl methanesulfonate-induced mutations. These authors compared the cost effectiveness among plant species with various genome sizes and concluded that this approach is increasingly advantageous as the genome size increases. Therefore, wheat is a suitable material for sequence capture in terms of cost effectiveness. Furthermore, this technique is useful in controlling the distribution of targets along chromosomes.

Genotyping by multiplexing amplicon sequencing (GBMAS) (Lab protocols of Schnable Lab., Iowa State University: http://schnablelab.plantgenomics.iastate.edu/resources/protocols/ (5 February 2018, date last accessed)) and genome-tagged amplification (GTA)[21] are new genotyping platforms that combine multiplexed PCR and multiplexed samples using bar codes with NGS. These methods appear to be suitable for MAS due to higher flexibility in choosing combinations of marker number and sample number. However, designing PCR primers for wheat is hindered by the close homology of the three genomes (A, B and D) and the high sequence similarity among the genes and gene family members. Recently, the International Wheat Genome Sequencing Consortium (IWGSC) offered a solution for distinguishing genomes using the chromosome-arm sorting technique.[22] Using the differences among homoeologous sequences, genome-specific primers that span target polymorphic sites can be designed. Genotyping by amplicon sequencing using genome-specific amplicons could be highly beneficial for simplifying genotype calls and achieving robust analysis for MAS.

The objective of this study is to develop an efficient approach for locating polymorphisms across the wheat genomes using the advantages of a sequence capture technique. To evaluate this approach, we designed D genome-specific primers and constructed linkage maps using multi-sample genotyping by amplicon sequencing.

2. Materials and methods

2.1. Plant materials and DNA extraction

Genomic DNA was extracted from leaves of *Triticum aestivum* cv. 'Kitahonami' and cv. 'Hatsumochi' using a DNeasy Plant Mini kit (Qiagen, Hilden, Germany). 'Kitahonami' is a winter wheat variety adapted to Hokkaido and has superior properties, such as high yield, high flour yield and high noodle-making quality. In contrast, 'Hatsumochi' was the first registered waxy (glutinous) variety and is adapted to the Kanto region (Central Japan). Diversity analyses indicated that modern Japanese varieties were separated into three groups corresponding to adaptation regions, and these two varieties fall into separate groups[13,17]; thus, we expected to observe a reasonable level of polymorphism between them. To construct the genetic map, a population of 94 recombinant inbred lines (RILs) was developed by crossing 'Hatsumochi' and 'Kitahonami' then using the single-seed descent method to advance to the F_9 generation. The total genomic DNA was extracted from the leaves of the RILs using the automated DNA extracting machine PI-50α (Kurabo Industries Ltd., Osaka, Japan).

2.2. Capture probe design and preparation of libraries

A workflow of sequence capture, polymorphism detection and designing genome-specific primers is described in Fig. 1. To design the

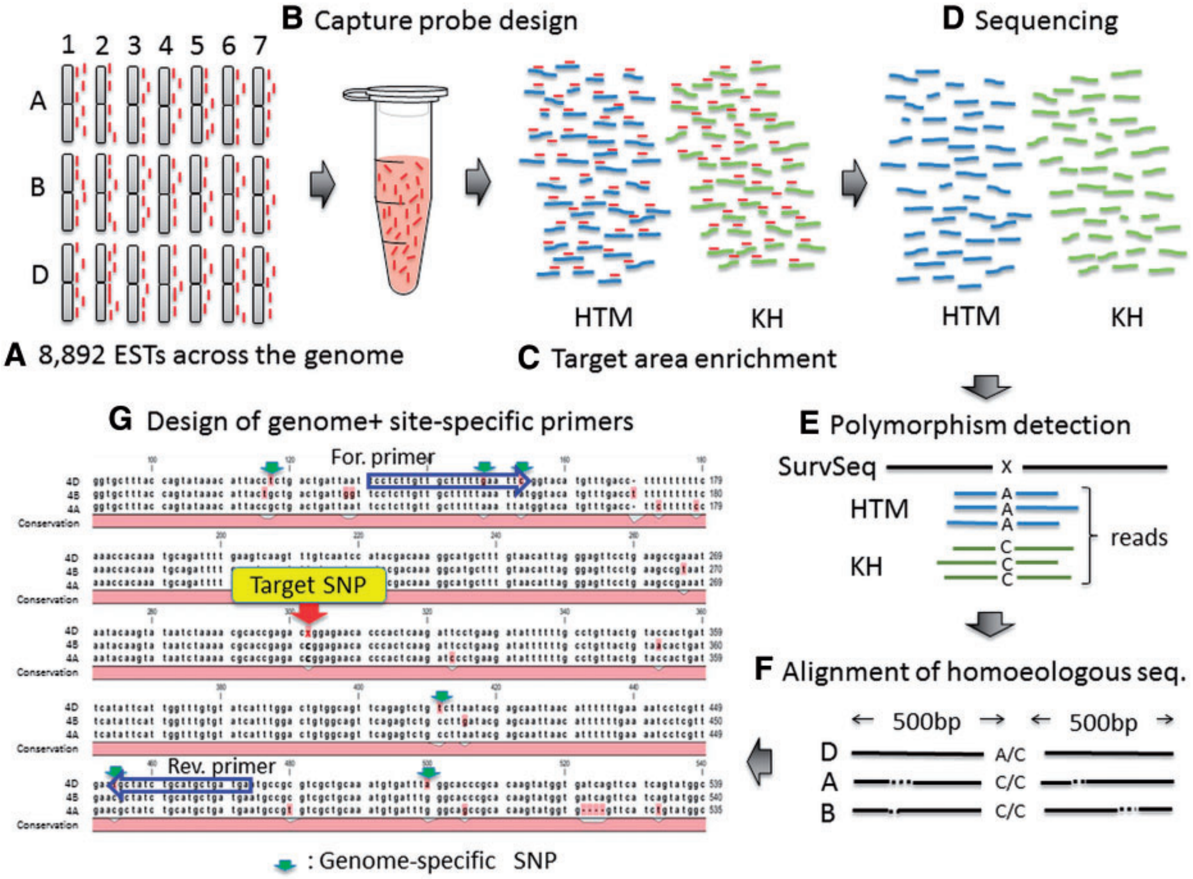

Figure 1. A workflow of sequence capture, polymorphism detection and designing genome-specific primers for amplicon sequencing. HTM; Hatsumochi, KH: Kitahonami, SurvSeq: IWGSC survey sequence.[22]

capture probes, we selected 8,892 wheat expressed sequence tag (EST) and full-length cDNA sequences from NCBI (https://www.ncbi.nlm.nih.gov/ (5 February 2018, date last accessed)) and TriFLDB[23] (Supplementary Table S1). Of these, 4,895 EST sequences were selected earlier during our design of PCR-based Landmark Unique Gene (PLUG) markers[24] and were known to be evenly distributed across the wheat chromosomes. Most of the remaining 3,997 sequences were EST sequences derived from mapped 9K iSelect probes[3] showing polymorphisms in Japanese varieties, and additional sequences to fill in gaps in coverage were based on bin-mapped wheat ESTs[25] that were selected based on the barley physical map.[26] Using the syntenic relationships among rice, barley and wheat, the selected sequences were estimated to be evenly distributed across the seven *Triticeae* chromosomes (Supplementary Table S1). All sequences were compared with one another, and only one representative homoeologous copy of each gene was selected for probe design. DNA capture of genomic DNA of ‘Hatsumochi’ and ‘Kitahonami’ was performed with a SeqCap EZ Developer Kit (Roche Diagnostics, Basel, Switzerland) following the manufacturer’s protocol.

In total, 1,000 ng of each DNA sample was sheared using a Covaris LE220 (Covaris Inc., Woburn, MA, USA) focused ultrasonicator to fragments that averaged 600 bp. An NEBNext Quick DNA Library Prep Master Mix set for 454 (New England Biolabs Inc., Beverly, MA, USA) was used to construct two genomic libraries. These libraries were amplified and fractionated from 500 to 700 bp before hybridization using a SeqCap EZ reagent Kit. To avoid hybridization with repetitive regions in the wheat genome, 1,500 ng of each amplified genomic library was mixed with 10 μL of SeqCap EZ reagent Kit in the presence of 10 μg PCE (plant capture enhancer, Roche Diagnostics) in a 0.2-mL thin-wall tube.

2.3. Detection of nucleotide polymorphisms between ‘Hatsumochi’ and ‘Kitahonami’

The enriched genomic DNAs were sequenced using the next-generation sequencer GS FLX plus (Roche Diagnostics). Reads were mapped to the wheat survey sequences[22] using 454 Sequencing System Software 2.7 (option: -mi 98). The detected polymorphisms against the survey sequences, including SNPs and Indels, were supported with at least three reads in each variety. From them, polymorphic sites between ‘Hatsumochi’ and ‘Kitahonami’ were extracted. The predicted effects of the polymorphisms were analysed using SnpEff[27] with IWGSC ver. 2.2 annotation data (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Genes-annotations (5 February 2018, date last accessed)).

2.4. Design of genome-specific primers

The flanking sequences of target polymorphism sites were obtained and blasted against the IWGSC survey sequences using PSI-BLAST,[28] and three homoeologous sequences in terms of their chromosomal locations were identified. The three sequences were aligned by MAFFT ver. 6.864[29] with a default parameters and imported into an in-house Java pipeline. This pipeline requires an alignment file of the homoeologous sequences in a target region. Based on the polymorphisms among the genomes, the pipeline automatically detects genome-specific primer pairs of 18–22 nucleotides at an amplicon length of approximately 300 bp. The primers contained nucleotides specific to the target sequence at the first and/or second positions from their 3′ end. The forward and reverse primers for the first PCR were tailed with the common sequence tags CS1 (5′-ACACTGACGACATGGTTCTACA-3′) and CS2 (5′-TACGGTAGC AGAGACTTGGTCT-3′), respectively, to allow for the addition of adapters during the second round of PCR according to the protocol of the Access Array for Ion Torrent PGM Sequencing System (Fluidigm Corporation, San Francisco, CA, USA).

2.5. Library preparation and amplicon sequencing

The reaction mix for the first PCR contained a total volume of 10 μL consisting of 1× Multiplex PCR Master Mix (Qiagen), a primer mix (described below) and 20–50 ng of template DNA. The primer mix contained 48 sets of randomly selected locus-specific primers, with CS1 or CS2 tails attached. The first PCR program consisted of a denaturation step at 95 °C for 15 min, followed by 32 cycles of 30 s at 94 °C, 90 s at 60 °C and 1 min at 72 °C, and a final extension for 10 min at 72 °C. The product from the first PCR of each sample was diluted 100-fold with sterilized distilled water, and 2 μL of the diluted product was used as the template for the second PCR. To perform bidirectional amplicon tagging in the second PCR, a forward fusion primer containing Ion A adapter, barcode, and the CS1 sequence and a reverse fusion primer containing Ion P1 adapter and the CS2 sequence were used with a portion of the template, while a second portion was amplified with an A-adaptor–barcode–CS2 and P1 adaptor–CS1 primer combination.

The 10-μL second PCR mix contained 1× Multiplex PCR Master Mix (Qiagen) and 400 nM forward and reverse fusion primers, and PCR was run using the following profile: 15 min at 95 °C, followed by 15 cycles of 30 s at 94 °C, 90 s at 60 °C and 1 min at 72 °C, and a final extension step of 10 min at 72 °C. All second PCR products were mixed in equivalent volumes (2 μL of PCR product per sample). The pooled product was purified using Agencourt AMPure XP Reagent beads (Beckman-Coulter, Fullerton, CA, USA) as follows: 12 μL of pooled products, 24 μL of TE buffer and 36 μL of well-mixed AMPure XP beads were vortexed. After a 10-min incubation at room temperature, the sample was placed onto a magnetic separator for 1 min, and the supernatant was discarded. The beads with sample attached were washed twice with 180 μL of freshly prepared 70% ethanol. Finally, the purified PCR products were suspended in 40 μL of low TE buffer. The quality of the amplicon library was assessed using an Agilent 2100 Bioanalyzer, and a high sensitivity kit (Agilent Technologies, Santa Clara, CA, USA) was used to define the region covering all PCR library peaks (300–450 bp). The purified library was quantified using a Qubit dsDNA HS assay kit (Thermo Fisher Scientific), with dilution to a concentration of 5 pM. Sequencing was performed using an Ion Torrent PGM system with an Ion PGM 400 sequencing kit and 318 chips (Thermo Fisher Scientific). A schematic diagram of the experimental procedure is shown in Supplementary Fig. S1.

2.6. Data processing and genotype calling

The removal of the Ion Torrent sequencing adaptor sites and demultiplexing of the barcodes to separate the different samples were automatically performed by Torrent Suite ver. 5 (Thermo Fisher Scientific). Further analysis was conducted using a custom Java script. First, the CS1 and CS2 tail sequences were removed. Second, the sequences were aligned to reference sequences using a BLAST-like alignment tool. The reference sequences consisted of internal sequences in the primers of each marker. Third, based on the alignment results of each sample-marker combination, the total number of aligned reads containing A, C, G and T bases, null alleles or missing values (NaN) and nucleotide deletions (“–”) in the base position of interest were counted. Fourth, the alleles were defined according to the bases with the highest read counts. The minimum read count was set to 5. Heterozygous site (H) was called when two alleles each had more than 40% of the total reads. Finally, the results of the genotype calls were summarized and exported into a tabular form.

2.7. Construction of the linkage map

The wheat *Pst*I (*Taq*I) v.3 DArT array (Diversity Array Technology Pty Ltd., Bruce, Australia) and 9K SNP array[3] were used to genotype 94 RILs. Publicly available SSR markers (GrainGenes 2.0) and functional markers[30] of the *Wx-A1*, *Wx-D1* and *Pina-D1* genes were also used. The data obtained using DArT, 9K, SSR and the newly developed markers (TARC) were merged, and polymorphic markers between parents were extracted. Poor-quality genotypes and genotypes with more than 10% missing data or segregation distortion were removed. Because redundant markers are completely correlated or identical and cannot provide additional information, only one of a redundant set of markers was used in map construction. MapDisto 1.7.5[31] was used to identify linkage groups. Mapped markers corresponding to those on the hexaploid wheat consensus map[3,15,32] were used as anchors to assign each linkage group to a particular chromosome and to orient linkage groups on short and long chromosome arms.

2.8. Polymorphism survey using multiple genotypes

To increase the resources for nucleotide polymorphism detection, four additional varieties, namely ‘Shunyou’, ‘Tohoku224’, ‘Kinuhime’ and ‘Yumechikara’, were subjected to capture sequencing. The first three varieties are adopted to the Kanto and Tohoku regions (Central and Northeastern Japan), while the ‘Yumechikara’ variety is adopted to the Hokkaido region. The procedures used for DNA extraction, library preparation, sequencing by GS FLX plus (Roche Diagnostics) and mapping to the reference sequences[22] were identical to those described above. To identify highly polymorphic sites, we performed pair-wise comparisons among the six varieties used. The detected polymorphisms, including SNPs and Indels, were supported with at least one read as either “Reference” or “SNP” in a given variety. If reads supported both “Reference” and “SNP”, we identified that variety as heterozygous at that site.

3. Results

3.1. Detection of polymorphisms between ‘Hatsumochi’ and ‘Kitahonami’

Using the next-generation sequencer GS FLX plus, 1,114,867 (536,055,413 bp) and 1,304,168 (615,634,987 bp) reads were obtained from the enriched genomic DNA of ‘Hatsumochi’ and ‘Kitahonami’, respectively. The average lengths of the sequences were 481 bp for ‘Hatsumochi’ and 472 bp for ‘Kitahonami’. Based on the criteria described above, 12,551 nucleotide polymorphisms were detected between these two varieties (Supplementary Table S2). Using the survey sequences,[22] we localized these polymorphisms to

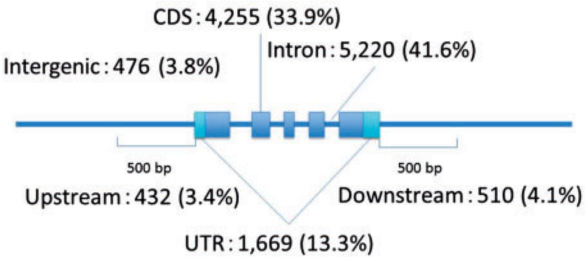

Figure 2. Cumulative number of polymorphic sites between ‘Hatsumochi’ and ‘Kitahonami’ in each genic region. The IWGSC gene models were used to define the genic regions.[22]

each chromosome or chromosome arm. As shown in Table 1, the number of polymorphisms varied among the chromosomes and genomes. The number of polymorphisms on the D genome were approximately one-sixth and one-fifth of those on the B and A genomes, respectively. Relatively fewer polymorphisms were observed in the Group 4 chromosomes than in the other groups. The cumulative numbers of polymorphic sites according to the gene models are shown in Fig. 2. Most polymorphisms were found in genic regions, including both the 500-bp upstream and downstream regions (Fig. 2, Supplementary Table S2). In the intron regions, 5,220 polymorphisms were detected, meaning the intron regions carried the highest percentage (41.6%) of the total polymorphisms.

3.2. Validation of polymorphisms by amplicon sequencing

Three hundred and ninety-six D genome-specific primers were designed using the primer-picking pipeline described above. The details regarding the primers are provided in Supplementary Table S3. Preliminary analysis of primers using gel electrophoresis of PCR products indicated that single PCR products with the expected fragment sizes were obtained using 380 of the 396 primer sets. Multiple products or a lack of bands were observed using the remaining primer sets. The PCR products from the 380 successful primer sets were mixed and sequenced using GS FLX plus. In total, 442,564 (average length 322.7 bp) and 640,147 (average length 327.2 bp) reads were obtained from the amplicons of ‘Hatsumochi’ and ‘Kitahonami’, respectively. The sequences derived from 312 markers were mapped to reference sequences without interference from off-target or homoeologous sequences (Table 2). Mixtures of homoeologous sequences were observed using 44 markers, which were classified according to the degree of the mixture as follows: low (<30%), medium (30–50%) and high (>50%). The alleles in markers with a low level of mixture could automatically be defined; in contrast, manual examination of the data was required to define alleles in markers with medium and high levels of mixture. The polymorphisms detected with 334 of the markers were consistent with those obtained using sequence capture (Supplementary Table S3). Thirty-six markers failed to validate the polymorphisms because of mixtures of homoeologous sequences, while 17 markers showed low read coverage. Nine markers did not show polymorphisms at the target sites.

3.3. Genotyping by amplicon sequencing in 96 multiplexed samples

The ‘Hatsumochi’/‘Kitahonami’ RILs were genotyped using an Ion PGM (Thermo Fisher Scientific) according to the procedure described above. The variations in the number of reads per sample and per maker are shown in Fig. 3. The number of reads per sample was relatively stable, and the average read count was 34,873. However, the variation in the number of reads per marker was high. The difference between the most and least sequenced markers was more than 150-fold. The markers with large product sizes tended to show a low number of reads (data not shown). Based on mapping to the reference sequences, we classified markers into eight classes (Table 2). The alleles in markers in classes 1–3 could automatically be defined, while those in markers in classes 4–6 had to be defined manually. Class 7 indicates markers with mixtures of sequences and indels, and class 8 represents no or quite low reads. The genotypes of 96 samples using 359 markers were obtained, and missing values accounted for only 1.2% (392 of 33,746 data points).

Table 1. Number of polymorphisms between ‘Hatsumochi’ and ‘Kitahonami’ in each wheat chromosome arm

		Group							
Genome	Arm	1	2	3	4	5	6	7	Total
A	S	293	409	137	160	150	287	490	5,293
	L	718	651	529	218	513	292	446	
B	S	174	519	1,197[a]	123	241	428	151	6,316
	L	648	929		398	723	512	273	
D	S	62	143	23	13	16	63	64	942
	L	56	126	86	17	146	75	52	
Total		1,951	2,777	1,972	929	1,789	1,657	1,476	12,551

[a] Total number of short and long arms.

3.4. Extension of D genome maps using the newly developed markers

By combining the data obtained using the newly developed markers (defined as TARC markers) with those obtained using DArT, 9K array and SSR, genotyping data for 3,956 markers was obtained (Supplementary Table S4). According to the grouping of redundant markers, 1,408 markers were considered non-redundant and were used for map construction using MapDisto software.[31] After filtering by redundancy, the number of DArT and 9K markers was reduced to 42.9% and 25.9% of the initial number, respectively. In contrast, of the 359 TARC markers, 261 (72.7%) remained after filtering. Using the default settings of MapDisto, 32 linkage groups were extracted from the 1,399 markers. Nine markers were unassigned. Three linkage groups were not assigned to chromosomes because of the lack of sequence information for the markers. The linkage map covered 3,994.7 cM, with an average chromosome length of 189.6 cM, and the lengths of the individual chromosomes varied from 95.3 cM (6D) to 275.4 cM (3B). Chromosomes 1A, 2A, 3A, 5D, 7B and 7D consisted of two linkage groups, and chromosome 6D consisted of three linkage groups. Only maps of the D genome are illustrated in Fig. 4. The complete set of linkage maps is shown in Supplementary Fig. S2. The number of loci per chromosome varied from 29 (4D) to 112 (2B), with an average of 66 loci per chromosome. Using a maximum of 0.61 for chromosome 1A and a minimum of 0.21 for chromosome 4D, the overall marker density was 0.36 markers per cM. The D genome linkage maps were compared with previous maps without the TARC markers. The total length of the D genome increased from 878.4 cM to 1333.1 cM, and the average distance between markers changed from 1.45 to 2.09 cM. In comparison, the A genome map length was 1222.7 cM with 2.58 cM between markers on average and the B genome 1425.4 cM with

Table 2. Classification of markers based on the mapping results of the amplicon sequencing

		Validation by parents		Genotype of RILs	
Class	Description	No. of markers	% in total	No. of markers	% in total
1	Genome specific	312	78.8	247	62.4
2	Mix off-target	6	1.5	11	2.8
3	Mix low (<30%)	16	4.0	34	8.6
4	Mix medium (30–50%)	21	5.3	43	10.9
5	Mix high (>50%)	7	1.8	9	2.3
6	Indel	2	0.5	15	3.8
7	Other[a]	11	2.8	24	6.1
8	Low or no reads	21	5.3	13	3.3
	Total	396		396	

[a]Markers with both mixtures of sequences and indels.

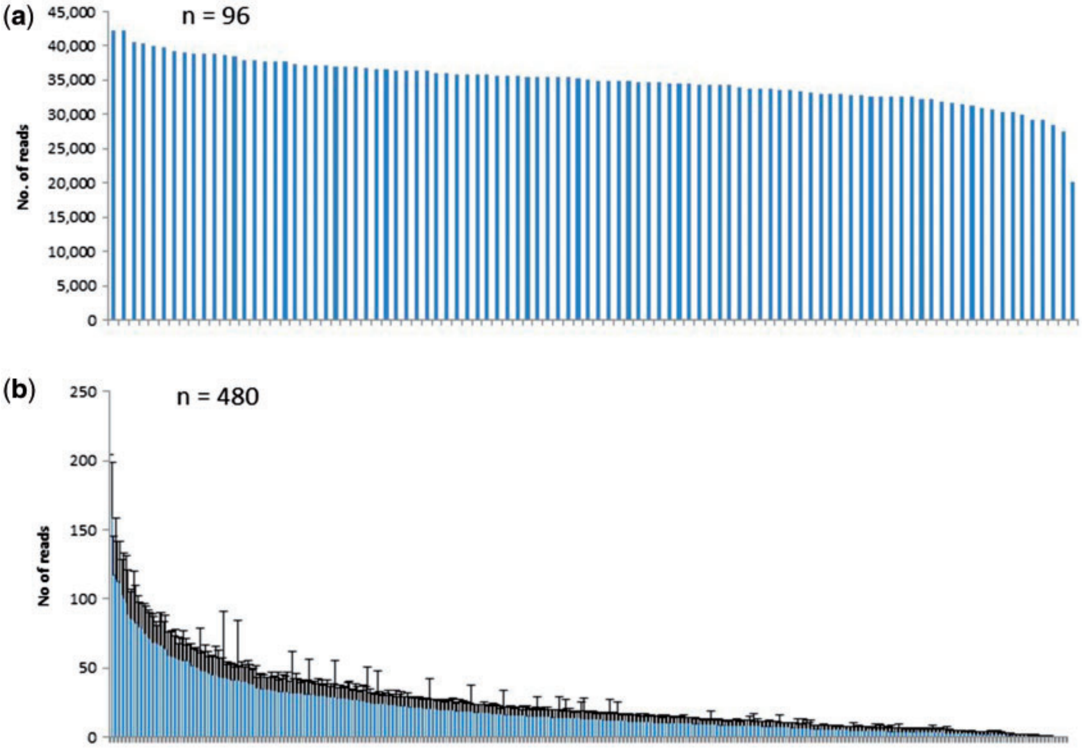

Figure 3. Variations in the number of reads per sample (a) and per marker (b). Error bars in the lower figure indicate standard deviations among 96 samples.

2.80 cM between markers. The TARC markers were distributed across the seven wheat chromosomes and significantly filled the previously observed gaps (Fig. 4). Furthermore, the TARC markers successfully extended the long arms of 1D, 2D, 4D and 6D and the short arms of 2D and 4D.

3.5. Detection of polymorphisms among the six varieties

Using the GS FLX plus, 8,253,381 reads were obtained from the six varieties. Of these reads, 5,473,010 reads were uniquely aligned to the IWGSC survey sequences.[22] Using the criteria described above, 31,542 polymorphic site candidates were detected (Supplementary Table S5). The pair-wise comparisons among the six varieties are shown in Table 3. The number of polymorphisms among the varieties ranged from 9,491 ('Tohoku224' vs. 'Kinuhime') to 19,926 ('Hatsumochi' vs. 'Yumechikara') and averaged 15,498 sites. Based on the mapping of the results to the IWGSC survey sequences, 5,986 polymorphic sites were predicted to be on the D genome and, therefore, can be considered potential markers for the D genome (Supplementary Table S5).

4. Discussion

In allohexaploid wheat, marker development is hampered by a large genome size, a high proportion of highly repetitive sequence (>80%), and the presence of three different genomes in which corresponding genes share a high level of sequence similarity. Because of these characteristics of the wheat genome, nucleotide polymorphisms have been obtained primarily using transcript sequences[33,34]

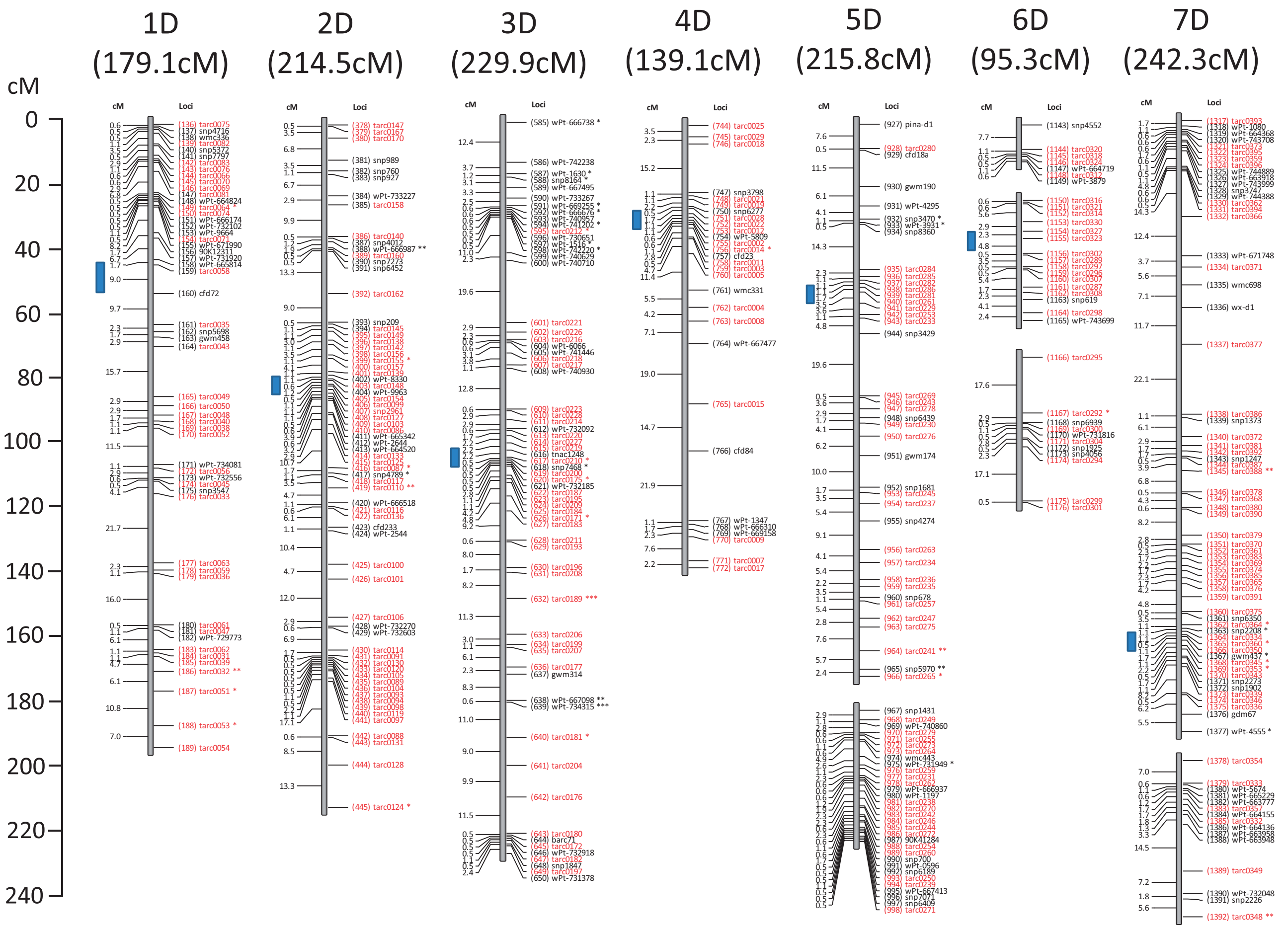

Figure 4. Linkage maps of the D genome using the 'Hatsumochi'/'Kitahonami' RILs. Newly developed markers are prefixed with 'tarc'. Box on the left of each chromosome indicates the putative position of the centromere.

Table 3. Number of polymorphic site candidates detected in pair-wise comparisons of six wheat varieties

	Kitahonami	Yumechikara	Tohoku224	Hatsumochi	Kinuhime	Shunyou
Kitahonami	—	15,638	16,934	16,800	16,670	15,774
Yumechikara		—	18,526	19,926	19,216	18,219
Tohoku224			—	12,518	9,491	13,542
Hatsumochi				—	11,561	14,008
Kinuhime					—	13,661
Shunyou						—

or restriction-site flanking sequences.[16,17] Although these methods have greatly contributed to increasing the number of markers, they do not provide a method of controlling the chromosomal locations of the detected polymorphisms. Therefore, although the number of polymorphisms is high, a significant bias occurs in the distribution of polymorphisms across the genome. Thus, targeted development of markers is required to saturate linkage maps. In this study, gene-enriched libraries were prepared using custom capture probes. To design the probes, we used positional information from sources such as the consensus linkage map (International Triticeae Mapping Initiative, ITMI),[32] bin-mapped ESTs,[25] PLUG markers[35] and the barley physical map.[26] Through this process, the map length in the D genome increased dramatically from 878.4 cM to 1333.1 cM (Fig. 4), and we were able to successfully develop markers in regions that were not covered by commercially available marker platforms, including DArT and 9K arrays. Because of the large genome size of wheat (16 Gb), the extraction of reliable polymorphic sites supported by sufficient read depth was believed to require a high-performance sequencer. However, in this study, using the GS FLX plus (Roche diagnostics), we obtained only 536 and 616 Mb sequences from the genomic DNA of 'Hatsumochi' and 'Kitahonami', respectively, indicating that less than 4% of the genome was sequenced in each variety. Despite this low coverage, we detected 12,551 polymorphic sites with an average read depth of 5.4 and 5.9 in 'Hatsumochi' and 'Kitahonami', respectively (Supplementary Table S2). These observations indicated that the SeqCap EZ reagent kit (Roche Diagnostics) enriched the target sites by approximately 160-fold compared with whole genome analysis and greatly contributed to the effectiveness of polymorphism detection. In reference to the gene models of the IWGSC survey sequences,[22] most polymorphic sites were located near genes, indicating that we successfully eliminated repetitive sequences and effectively enriched the targets. The elimination of repetitive elements is important for genome mapping because these elements usually distort the mapping results. In this study, more than 40% of the polymorphic sites were in intron regions, and thus could not be found using transcriptome analyses. Because intron regions have a higher level of polymorphisms than exon regions, enriched genomic sequencing is a beneficial approach for developing gene-related markers. Furthermore, based on SnpEff analysis,[27] 60 and 1,630 polymorphisms were predicted to have high and moderate effects on the corresponding gene functions, respectively (Supplementary Table S2). Therefore, the strategy used in this study effectively identified polymorphisms that could potentially be related to agronomically important traits.

In allohexaploid wheat, the presence of three homoeologous genomes poses a challenge in SNP detection. Although a difference in degree is observed, many markers on the commercially available SNP arrays are affected by interference from the other two homoeologous sequences.[36] In this study, the genome-specific primers were highly beneficial for defining alleles. Approximately 80% of the marker loci were successfully genotyped using the locus-specific approach when we validated the polymorphisms using the two parental varieties (Table 2). When we used the RIL populations, more than 70% of the markers were grouped into classes 1–3, indicating that these markers could be genotyped similarly to diploid species. Based on these results, our strategy of designing genome-specific primers was effective and demonstrated the importance of obtaining a priori knowledge of the polymorphisms among genomes by comparing the homoeologous sequences of interest.

The variations in the read number among samples and markers were investigated (Fig. 3). In this study, we did not equilibrate the multiplex PCR samples from individual wells before mixing samples. Despite this simplification of the process, the read numbers among the samples were relatively uniform, except for one outlier sample that had an extremely low number of reads. Therefore, the protocol used in this study is beneficial for processing many samples. In contrast, the read numbers varied among the markers, with the read number ranging from nearly zero to more than 200 per sample. Markers with longer amplicons tended to have fewer reads (data not shown); however, other factors, such as the annealing efficiencies of the primers, also affected the read numbers. Because the multiplex levels of the samples and markers were determined according to the minimum number of reads required for each marker, a uniform distribution of read numbers among the markers is important. Further optimization of the multiplex PCR conditions will be beneficial in minimizing missing genotype values.

Because amplicon sequencing primers are typically designed to flank the site of interest, other polymorphisms in addition to the site of interest can sometimes be identified. When applying the method described in this paper to breeding or genetic analysis, the detection of new polymorphisms that are independent of the target sites could provide additional haplotype information for the samples of interest, and such haplotype information substantially improves power in the detection of marker-trait associations.[37] Informative markers that contain more than two polymorphic sites could reduce the marker number required for an investigation, thereby reducing the cost for genotyping, and help in performing various genetic studies, such as association mapping and genomic selection in breeding programs.

Because the existing SNP arrays are substantially deficient in the number of D genome markers, we developed markers for the D genome to demonstrate the effectiveness of our approach. The polymorphism rate of the D genome is lower than that of the other two genomes, which is attributed to the evolutionary history of the D genome. According to recent studies, interploidy natural hybridization and subsequent introgression played a significant role in the diversification of common wheat,[38] and therefore, the D genome had fewer opportunities to exchange genetic material than the A and B genomes. Although the polymorphic rate is low, comparable numbers of important genes and QTLs on the D genome are described in the catalog of gene symbols for wheat.[39] Recent genome-wide association studies for pre-harvest sprouting using Chinese wheat landraces[40] and European winter wheats[41] reported QTLs on 1D, 3D and 5D chromosomes. Additionally, resistant genes against wheat yellow mosaic virus and soil-borne wheat mosaic virus were found on chromosomes 2D and 5D, respectively.[6,42] For the fine mapping of these agronomically important genes or QTLs, the number of markers must be increased around the regions of interest. From the polymorphism survey using six varieties, around 6,000 polymorphic sites were detected on the D genome, and some sites have been successfully used to develop markers (data not shown). Therefore, the polymorphic information obtained in this study is a useful resource for the further development of markers across the genome.

In this study, we proposed efficient strategies for the detection of nucleotide polymorphisms among varieties of interest and the design of locus-specific primers to achieve robust high-throughput genotyping. The IWGSC's continuous effort to obtain the first reference sequence of the spring wheat variety 'Chinese Spring' is opening the post-genomic era (http://www.wheatgenome.org/ (5 February 2018, date last accessed)). In this era, the motivation for collecting genomic resources using in-house materials is likely to increase. By comparing the polymorphic sites found in this study with probe sequences in publicly available SNP arrays, at least two-thirds of the total sites did not show any similarity with those probes in BLASTn searches (Supplementary Tables S2 and S5). The high percentage of new polymorphic sites indicates the importance of polymorphism surveys using materials of interest. In addition to the number of polymorphic sites, the polymorphic frequencies among materials are also important. Compared with the existing SNP arrays, the polymorphic sites among the six varieties in this study provided highly polymorphic markers among Japanese materials, particularly among varieties from Central and Western Japan (data not shown). To date, the germplasm used in genomic studies has been limited to the leading varieties in developed countries. However, for specific traits, such as resistance to disease or abiotic stress, many sources of germplasm from around the world remain unexamined. Because the method described in this study is less expensive, more flexible and more reliable than previous methods, this method is suitable for pan-genome studies that must process many haplotypes.

Data Availability

All sequences analysed in the present study were deposited into the DDBJ/GenBank/MMBL database with accession numbers DRA006270. Sample indices are prefixed to sequence names as follows: 'Kitahonami', H7U3MUP; 'Hatsumochi', H7YSFDH; 'Shunyou', IY4LCZI; 'Kinuhime', IZHNKMJ; 'Tohoku224', IZO1F6G; and 'Yumechikara', HKAWHI.

Acknowledgements

This study was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, NGB-1002 and NGB-1007).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Gupta, P. K., Rustgi, S. and Mir, R. R. 2008, Array-based high-throughput DNA markers for crop improvement, *Heredity*, **101**, 5–18.
2. Xu, Y. B., Lu, Y. L., Xie, C. X., Gao, S. B., Wan, J. M. and Prasanna, B. M. 2012, Whole-genome strategies for marker-assisted plant breeding, *Mol. Breed.*, **29**, 833–54.
3. Cavanagh, C. R., Chao, S., Wang, S., et al. 2013, Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars, *Proc. Natl. Acad. Sci. USA.*, **110**, 8057–62.
4. Wang, S., Wong, D., Forrest, K., et al. 2014, Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array, *Plant Biotechnol. J.*, **12**, 787–96.
5. Iehisa, J. C., Ohno, R., Kimura, T., et al. 2014, A high-density genetic map with array-based markers facilitates structural and quantitative trait locus analyses of the common wheat genome, *DNA Res.*, **21**, 555–67.
6. Liu, S., Yang, X., Zhang, D., Bai, G., Chao, S. and Bockus, W. 2014, Genome-wide association analysis identified SNPs closely linked to a gene resistant to soil-borne wheat mosaic virus, *Theor. Appl. Genet.*, **127**, 1039–47.
7. Naruoka, Y., Garland-Campbell, K. A. and Carter, A. H. 2015, Genome-wide association mapping for stripe rust (*Puccinia striiformis* F. sp. *tritici*) in US Pacific Northwest winter wheat (*Triticum aestivum* L.), *Theor. Appl. Genet.*, **128**, 1083–101.
8. Wu, Q. H., Chen, Y. X., Zhou, S. H., et al. 2015, High-density genetic linkage map construction and QTL mapping of grain shape and size in the wheat population Yanda1817 x Beinong6, *PLoS One*, **10**, e0118144.
9. Bulli, P., Zhang, J., Chao, S., Chen, X. and Pumphrey, M. 2016, Genetic architecture of resistance to stripe rust in a global winter wheat germplasm collection, *G3-Genes Genomes Genet.*, **6**, 2237–53.
10. Lin, M., Zhang, D., Liu, S., et al. 2016, Genome-wide association analysis on pre-harvest sprouting resistance and grain color in U.S. winter wheat, *BMC Genomics*, **17**, 794.
11. Yu, L. X., Chao, S., Singh, R. P. and Sorrells, M. E. 2017, Identification and validation of single nucleotide polymorphic markers linked to Ug99 stem rust resistance in spring wheat, *PLoS One*, **12**, e0171963.
12. Chao, S. M., Dubcovsky, J., Dvorak, J., et al. 2010, Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.), *BMC Genomics*, **11**, 727.
13. Ishikawa, G., Nakamura, K., Ito, H., et al. 2014, Association mapping and validation of QTLs for flour yield in the soft winter wheat variety Kitahonami, *PLoS One*, **9**, e111337.
14. Zhai, H., Feng, Z., Li, J., et al. 2016, QTL analysis of spike morphological traits and plant height in winter wheat (*Triticum aestivum* L.) using a high-density SNP and SSR-based linkage map, *Front. Plant Sci.*, **7**, 1617.
15. Huang, B. E., George, A. W., Forrest, K. L., et al. 2012, A multiparent advanced generation inter-cross population for genetic analysis in wheat, *Plant Biotechnol. J.*, **10**, 826–39.
16. Poland, J. A., Brown, P. J., Sorrells, M. E. and Jannink, J. L. 2012, Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach, *PLoS One*, **7**, e32253.
17. Kobayashi, F., Tanaka, T., Kanamori, H., Wu, J., Katayose, Y. and Handa, H. 2016, Characterization of a mini core collection of Japanese wheat varieties using single-nucleotide polymorphisms generated by genotyping-by-sequencing, *Breed Sci.*, **66**, 213–25.
18. Marcussen, T., Sandve, S. R., Heier, L., et al. 2014, Ancient hybridizations among the ancestral genomes of bread wheat, *Science*, **345**, 1250092.
19. Brenchley, R., Spannagl, M., Pfeifer, M., et al. 2012, Analysis of the bread wheat genome using whole-genome shotgun sequencing, *Nature*, **491**, 705–10.
20. Henry, I. M., Nagalakshmi, U., Lieberman, M. C., et al. 2014, Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing, *Plant Cell*, **26**, 1382–97.
21. Bernardo, A., Wang, S., St Amand, P. and Bai, G. 2015, Using next generation sequencing for multiplexed trait-linked markers in wheat, *PLoS One*, **10**, e0143890.
22. IWGSC 2014, A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome, *Science*, **345**, 1251788.
23. Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. and Shinozaki, K. 2009, TriFLDB: a database of clustered full-length coding sequences from *Triticeae* with applications to comparative grass genomics, *Plant Physiol.*, **150**, 1135–46.
24. Ishikawa, G., Yonemaru, J., Saito, M. and Nakamura, T. 2007, PCR-based landmark unique gene (PLUG) markers effectively assign homoeologous wheat genes to A, B and D genomes, *BMC Genomics*, **8**, 135.
25. Qi, L. L., Echalier, B., Chao, S., et al. 2004, A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat, *Genetics*, **168**, 701–12.
26. Mayer, K. F., Waugh, R., Brown, J. W., et al. 2012, A physical, genetic and functional sequence assembly of the barley genome, *Nature*, **491**, 711–6.
27. Cingolani, P., Platts, A., Wang le, L., et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso, *Fly*, **6**, 80–92.
28. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
29. Katoh, K. and Toh, H. 2008, Recent developments in the MAFFT multiple sequence alignment program, *Brief Bioinform.*, **9**, 286–98.
30. Liu, Y., He, Z., Appels, R. and Xia, X. 2012, Functional markers in wheat: current status and future prospects, *Theor. Appl. Genet.*, **125**, 1–10.
31. Lorieux, M. 2012, MapDisto: fast and efficient computation of genetic linkage maps, *Mol. Breed.*, **30**, 1231–5.
32. Sorrells, M. E., Gustafson, J. P., Somers, D., et al. 2011, Reconstruction of the synthetic W7984 x Opata M85 wheat reference population, *Genome*, **54**, 875–82.
33. Allen, A. M., Barker, G. L., Berry, S. T., et al. 2011, Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.), *Plant Biotechnol. J.*, **9**, 1086–99.
34. Akhunov, E. D., Akhunova, A. R., Anderson, O. D., et al. 2010, Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes, *BMC Genomics*, **11**, 702.
35. Ishikawa, G., Nakamura, T., Ashida, T., et al. 2009, Localization of anchor loci representing five hundred annotated rice genes to wheat chromosomes using PLUG markers, *Theor. Appl. Genet.*, **118**, 499–514.
36. Akhunov, E., Nicolet, C. and Dvorak, J. 2009, Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay, *Theor. Appl. Genet.*, **119**, 507–17.
37. Lu, Y. L., Xu, J., Yuan, Z. M., et al. 2012, Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize, *Mol. Breeding*, **30**, 407–18.

38. Matsuoka, Y. 2011, Evolution of polyploid *Triticum* wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification, *Plant Cell Physiol*., **52**, 750–64.
39. McIntosh, R. A., Yamazaki, Y., Dubcovsky, J., et al. 2013, Catalogue of gene symbols for wheat. https://shigen.nig.ac.jp/wheat/komugi/genes/symbolClassList.jsp (5 February 2018, date last accessed)
40. Zhou, Y., Tang, H., Cheng, M. P., et al. 2017, Genome-wide association study for pre-harvest sprouting resistance in a large germplasm collection of Chinese wheat landraces, *Front. Plant Sci*., **8**, 401.
41. Albrecht, T., Oberforster, M., Kempf, H., et al. 2015, Genome-wide association mapping of preharvest sprouting resistance in a diversity panel of European winter wheats, *J. Appl. Genet*., **56**, 277–85.
42. Nishio, Z., Kojima, H., Hayata, A., et al. 2010, Mapping a gene conferring resistance to Wheat yellow mosaic virus in European winter wheat cultivar 'Ibis' (*Triticum aestivum* L.), *Euphytica*, **176**, 223–9.
43. Mayer, K. F., Taudien, S., Martis, M., et al. 2009, Gene content and virtual gene order of barley chromosome 1H, *Plant Physiol*., **151**, 496–505.