Theor Appl Genet (2014) 127:261–271
DOI 10.1007/s00122-013-2215-5

ORIGINAL PAPER

Genome-wide marker development for the wheat D genome based on single nucleotide polymorphisms identified from transcripts in the wild wheat progenitor *Aegilops tauschii*

Julio Cesar Masaru Iehisa · Akifumi Shimizu · Kazuhiro Sato · Ryo Nishijima · Kouhei Sakaguchi · Ryusuke Matsuda · Shuhei Nasuda · Shigeo Takumi

Received: 20 July 2013 / Accepted: 14 October 2013 / Published online: 25 October 2013

Abstract

Key message **13,347 high-confidence SNPs were discovered through transcriptome sequencing of *Aegilops tauschii*, which are useful for genomic analysis and molecular breeding of hexaploid wheat.**

Abstract In organisms with large and complex genomes, such as wheat, RNA-seq analysis is cost-effective for discovery of genome-wide single nucleotide polymorphisms (SNPs). In this study, deep sequencing of the spike transcriptome from two *Aegilops tauschii* accessions representing two major lineages led to the discovery of 13,347 high-confidence (HC) SNPs in 4,872 contigs. After removing redundant SNPs detected in the leaf transcriptome from the same accessions in an earlier study, 10,589 new SNPs were discovered. In total, 5,642 out of 5,808 contigs with HC SNPs were assigned to the *Ae. tauschii* draft genome sequence. On average, 732 HC polymorphic contigs were mapped in silico to each *Ae. tauschii* chromosome. Based on the polymorphic data, we developed markers to target the short arm of chromosome 2D and validated the polymorphisms using 20 *Ae. tauschii* accessions. Of the 29 polymorphic markers, 28 were successfully mapped to 2DS in the diploid F_2 population of *Ae. tauschii*. Among ten hexaploid wheat lines, which included wheat synthetics and common wheat cultivars, 25 of the 43 markers were polymorphic. In the hexaploid F_2 population between a common wheat cultivar and a synthetic wheat line, 23 of the 25 polymorphic markers between the parents were available for genotyping of the F_2 plants and 22 markers mapped to chromosome 2DS. These results indicate that molecular markers that developed from polymorphisms between two distinct lineages of *Ae. tauschii* might be useful for analysis not only of the diploid, but also of the hexaploid wheat genome.

Communicated by X. Xia.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2215-5) contains supplementary material, which is available to authorized users.

J. C. M. Iehisa · R. Nishijima · K. Sakaguchi · R. Matsuda · S. Takumi (✉)
Graduate School of Agricultural Science, Kobe University, Nada-ku, Kobe 657-8501, Japan
e-mail: takumi@kobe-u.ac.jp

A. Shimizu
Department of Biological Resources Management, School of Environmental Science, University of Shiga Prefecture, Hikone 522-8533, Japan

K. Sato
Institute of Plant Science and Resources, Okayama University, Kurashiki 710-0046, Japan

S. Nasuda
Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan

Introduction

Common wheat (*Triticum aestivum* L.) is an allohexaploid species that originated by natural hybridization between tetraploid wheat (*Triticum turgidum* L.), containing the A and B genomes, and the wild diploid relative *Aegilops tauschii* Coss., containing the D genome (Kihara 1944; McFadden and Sears 1944). *Ae. tauschii* is widely distributed in Eurasia and shows abundant genetic variation (Dvorak et al. 1998; Dudnikov and Kawahara 2006; Matsuoka et al. 2007, 2008, 2009; Takumi et al. 2009a). Based on population structure analyses, *Ae. tauschii* has been divided into two major genealogical lineages, lineage

1 (L1) and lineage 2 (L2), and a minor lineage, HGL17 (Mizuno et al. 2010a; Wang et al. 2013). The gene pool of *Ae. tauschii* can be easily accessed in wheat breeding, but remains largely unexplored. Synthetic hexaploid wheat can be obtained through interspecific hybridization between tetraploid wheat and *Ae. tauschii* (Kihara and Lilienfeld 1949; Matsuoka and Nasuda 2004), and the resulting lines can be used as intermediates to exploit the natural variation in *Ae. tauschii* for improvement of common wheat (Trethowan and Mujeeb-Kazi 2008; Jones et al. 2013). It is believed that the *Ae. tauschii* populations involved in the origin of common wheat are limited to a narrow distribution range and restricted to L2, which has given rise to a founder effect in hexaploid wheat (Feldman 2001; Mizuno et al. 2010a; Wang et al. 2013). Therefore, *Ae. tauschii*, especially L1, has large genetic diversity that is not represented in common wheat (Feldman 2001; Mizuno et al. 2010a, 2010b).

Development of molecular markers is an important step for molecular breeding and map-based cloning. Although a draft genome sequence of *Ae. tauschii* based on a whole genome shotgun strategy has been published, only 1.72 Gb out of the 4.36 Gb genome was anchored to chromosomes (Jia et al. 2013). A physical map of *Ae. tauschii* covering 4 Gb has also been developed (Luo et al. 2013), while the availability of bacterial artificial chromosome sequences is limited. In a previous study, annotation-based genome-wide single nucleotide polymorphism (SNP) discovery has been developed to overcome the complex nature of *Ae. tauschii* genome (You et al. 2011). In this approach, Roche 454 shotgun reads could be annotated with low genome coverage for one genotype, and then genomic and cDNA shotgun reads of another genotype generated from SOLiD and Solexa platforms were used to identify around 500,000 putative SNPs. However, around 56 % of the sequence length, characterized as a repetitive region, was excluded from the analysis (You et al. 2011). RNA-seq, a next-generation sequencing technology for transcripts, is cost-effective for SNP discovery in organisms having large and complex genomes and insufficient reference information (Hansey et al. 2012; Iehisa et al. 2012). The transcript-based approach for SNP discovery has also been applied in hexaploid wheat (Allen et al. 2011; Cavanagh et al. 2013). We previously sequenced the leaf transcriptome of two *Ae. tauschii* accessions, each from L1 and L2 (Iehisa et al. 2012). After de novo assembly of the reads, 4,337 SNPs were discovered in at least 1,700 contigs, and around 200 polymorphic contigs per chromosome were mapped in silico to barley virtual chromosomes by the GenomeZipper approach (Mayer et al. 2011). More recent draft sequence information on the barley genome could be more helpful for in silico mapping [International Barley Genome Sequencing Consortium (IBSC) 2012]. Because almost all of the validated SNP markers were polymorphic between L1 and L2, it was assumed that these markers would be available for wheat D-genome genotyping (Iehisa et al. 2012). The objectives of the present study were to discover new SNPs from RNA-seq data from spikes of *Ae. tauschii* and to assess the *Ae. tauschii* SNP libraries for D-genome analysis of diploid and hexaploid wheat. We also evaluated the D-genome SNP markers using genomic information on *Ae. tauschii* and barley.

Materials and methods

Plant material and cDNA library construction

In total, 20 *Ae. tauschii* accessions from each sublineage and 10 hexaploid wheat lines were used (Table 1). Synthetic hexaploid wheats were obtained through crosses of tetraploid wheat cultivar Langdon (Ldn) and different *Ae. tauschii* accessions, followed by chromosome doubling of the interspecific ABD hybrids (Takumi et al. 2009b; Kajimura et al. 2011). For RNA-seq, the *Ae. tauschii* accessions PI476874 from the L1-2 sublineage and IG47182 from the L2-2 sublineage were selected. Total RNA was isolated from spikes (about 3–6 cm long) before heading

Table 1 Lineage, accession number and origin of *Ae. tauschii* and hexaploid wheat accessions used in this study

Lineage–sublineage	Accession number (country)
Ae. tauschii accessions	
L1–1	IG48508 (Turkmenistan), KU-2627 (Afghanistan)
L1–2	PI476874 (Afghanistan), IG126387 (Turkmenistan)
L1–3	KU-2826 (Georgia), KU-2087 (Iran)
L1–4	IG131606 (Kyrgyzstan), IG48559 (Tajikistan)
L1–5	IG48747 (Armenia), KU-2144 (Iran)
L1–6	AT47 (China), AT76 (China)
L2–1	KU-2069 (Iran), KU-2811 (Armenia)
L2–2	IG47182 (Azerbaijan), KU-2100 (Iran)
L2–3	KU-2159 (Iran), KU-2093 (Iran)
HGL17	AE454 (Georgia), AE929 (Georgia)
Hexaploid wheat accessions	
L1-derived synthetics	Ldn/IG131606 (Kyrgyzstan), Ldn/IG126387 (Turkmenistan), Ldn/PI476874 (Afghanistan)
L2-derived synthetics	Ldn/KU-2090 (Iran), Ldn/KU-2069 (Iran), Ldn/KU-2159 (Iran), Ldn/KU-2097 (Iran)
Common wheat cultivars	Norin 61 (Japan), Kitanokaori (Japan), Chinese Spring (China)

Underlining indicates the accessions used for RNA-seq

HGL haplogroup lineage

stage using an RNeasy Plant Mini Kit (QIAGEN, Hilden, Germany). mRNA was purified from 48 μg of RNA using an *Oligotex-dT30* mRNA Purification Kit (Takara Bio, Ohtsu, Japan). A cDNA Synthesis System (Roche Diagnostics, Mannheim, Germany) was used to fragment a 200 ng aliquot of mRNA and synthesize cDNA from it. Approximately, 10^8 adapter-ligated cDNA molecules from the samples were used for library preparation using a GS FLX Titanium Rapid Library Preparation Kit (Roche Diagnostics).

RNA-seq, sequence assembly and discovery of polymorphic sites

The cDNA libraries were sequenced with a GS FLX Titanium Sequencing Kit on a GS FLX System (Roche Diagnostics) according to the manufacturer's instructions. Files containing raw sequence data were deposited in the sequence read archive of the DNA Data Bank of Japan (DDBJ) (accession number DRA001014). Before assembly, raw sequence reads were trimmed to remove primer and adapter sequences and poly-A tails. All reads from both accessions were merged and assembled de novo with the GS de novo assembler algorithm (Newbler) version 2.6 (Roche Diagnostics) to generate reference contig sequences (minimum-overlap length of 40 bp and minimum-overlap identity of 90 %). SNPs and insertions/deletions (indels) were discovered by aligning all individual reads to the reference contig sequences using GS Reference Mapper version 2.6 software (Roche Diagnostics). Only the accession-specific sequence variants (supported by at least two reads) were extracted as true polymorphisms from "All" and "High-Confidence" sequence differences produced by the GS Reference Mapper.

Generation of non-redundant (NR) contig sequences

To obtain NR contig sequences from all the leaf- and spike-derived contigs, sequence clustering using the CD-HIT-EST web server (Li and Godzik 2006) with 95 % identity (parameters: -c 0.95, -n 8) was performed for each of the two tissues. Next, the NR contigs from both tissues were merged and sequence clustering was performed using CD-HIT-EST with 95 % identity (-c 0.95, -n 8).

Gene annotation and comparison of SNPs between the spike and leaf datasets

The reference sequences were searched against the National Center for Biotechnology Information (NCBI) NR protein database using the blastx algorithm with an *E*-value cutoff of 10^{-3}. Gene ontology (GO) terms were assigned using Blast2GO software (Conesa et al. 2005) based on blastx hits against the NCBI NR database.

For all the SNPs discovered in leaves (Iehisa et al. 2012) and spikes, 50 bp sequences were extracted from contigs, positioning the SNPs in the middle of the sequence. The sequences generated from the spike libraries were searched against those generated from the leaf libraries using the blastn algorithm with an *E*-value cutoff of 10 and hit length ≥25 bp. An SNP was classified as "common" when the 26th nucleotide (SNP) from the 5′ end of the query sequence matched the SNP of the subject sequence. For this, the BLAST output was parsed to extract SNPs in common between the two datasets using an in-house Perl script.

In silico mapping of new SNPs and NR contigs to genome sequences of *Ae. tauschii* and barley

Contigs with high confidence (HC), lower confidence (LC) SNPs and NR contigs were blastn searched against *Ae. tauschii* genome sequences (Jia et al. 2013) and barley genome sequences, among which fingerprinted contigs, whole genome shotgun assemblies and HC genes were included (IBSC 2012), with an *E*-value threshold of 10^{-5} and hit length ≥50 bp. The LC SNPs were here defined as all detected SNPs except the HC SNPs. Based on the blastn data, the SNPs of *Ae. tauschii* were plotted onto the barley genome using Circos version 0.63 software (Krzywinski et al. 2009).

Development of SNP and indel markers

The identified SNPs were converted to cleaved amplified polymorphic sequence (CAPS) or high-resolution melting (HRM) markers. The primer sequences of SNP and indel markers, product lengths and restriction enzymes are summarized in Table S1. PCR and analysis were performed according to our previous studies (Matsuda et al. 2012; Iehisa et al. 2012). The polymorphic information content (PIC) was calculated using Excel Microsatellite toolkit add-in software (Park 2001).

Linkage map construction

Three F_2 mapping populations with different ploidy levels were used for construction of D-genome linkage maps. The diploid mapping population consisted of 97 F_2 individuals derived from a cross between PI476874 (L1) and IG47182 (L2) (Iehisa et al. 2012); the synthetic hexaploid wheat population consisted of 117 F_2 individuals derived from a cross between Ldn/KU-2075 (L2) and Ldn/KU-2025 (L1) (Mizuno et al. 2011; Matsuda et al. 2012) and a set of 108 F_2 individuals derived from a cross between the common wheat cultivar Norin 61 (N61) and an L1-derived synthetic wheat (Ldn/PI476874) (Okamoto et al. 2012). The new

markers were assigned to chromosomes 2D of the established genetic maps. Linkage maps were constructed using the MAPMAKER/EXP version 3.0b package (Lander et al. 1987) and drawn in MapChart version 2.2 software (Voorrips 2002). The genetic distances were calculated with the Kosambi function (Kosambi 1944).

Results

Sequencing and assembly of expressed sequence tags (ESTs)

The sequencing of spike cDNA libraries from PI476874 and IG47182, respectively, produced 848,953 and 893,917 reads, which corresponded to 361 and 386 Mb per accession after trimming. Both libraries were merged and assembled de novo using Newbler to generate the reference sequences. During assembly, the Newbler algorithm constructs multiple alignments of overlapping reads and divides them into consistent sequences; i.e., contigs. When contig graphs contain branching structures, Newbler traverses paths through connected branches, generating isotigs that may represent splicing variants. In total, 21,208 contigs (length ≥100 bp) and 18,530 isotigs were generated. The nucleotide length of the majority of contigs and isotigs ranged between 500 and 1,000 bp, with an average of 903 bp for contigs and 1,231 bp for isotigs (Table 2).

To estimate how many of the contigs were expressed in spikes rather than leaves, NR contig sequences were first obtained for each of the two tissues, and then the number of spike contigs was reduced from 21,208 to 20,518 and the number of leaf contigs from 10,224 to 9,896. Next, both tissue-derived NR contigs were merged and second sequence clustering was performed again. Out of the 30,414 contigs, 24,875 NR contigs were obtained, and 15,739 contigs were found in spikes but not in leaves (Fig. S1).

For gene annotation, isotigs were reassembled using the CAP3 program (Huang and Madan 1999) with default parameters (identity ≥90 % and overlap of 40 bp), and longer NR sequences were obtained. As a result, the 18,530 isotigs from both tissues were reassembled into 17,598 NR sequences. These NR sequences were blastx searched against the NCBI NR protein database, and significant hits were obtained for 88.4 % of the query sequences. GO annotation was performed based on these blastx results, assigning one or more GO terms to 13,309 sequences. Of the assigned GO terms, 34,500 were under the biological process domain, 15,851 under molecular function and 19,017 under cellular component. Compared with the leaf isotigs, GO term enrichment was observed in the developmental process, cellular component organization, nucleus, plasma membrane and cytosol categories (Fig. S2).

Table 2 Contig and isotig sequence length distribution

Sequence length (bp)	Number of contig	Number of isotig
100–500	6,559	1,356
501–1,000	7,840	8,217
1,001–1,500	3,576	4,289
1,501–2,000	1,705	2,244
2,001–2,500	759	1,071
>2,500	769	1,353
Total	21,208	18,530
Average length (bp)	903	1,231

SNP and indel detection

To search for polymorphic sites, individual reads from the spike libraries of two accessions were aligned to the reference sequence using GS Reference Mapper. Of the 28,268 total polymorphic sites detected in contigs longer than 100 bp, 25,837 were SNPs, 2,108 indels and 323 variants with two or more nucleotide changes (Table 3). According to the read depth (≥3 non-duplicated reads) and quality (QV ≥20), 13,347 polymorphic sites were classified as HC SNPs, which were detected in 4,872 contigs (Table S2). On average, one SNP was found for every 741 bp of sequence, and an HC SNP appeared once per 1,435 bp of sequence.

The SNP dataset from the published leaf libraries contained SNPs derived from contigs shorter than 100 bp (Iehisa et al. 2012). Thus, we selected only SNPs detected in contigs ≥100 bp to compare the SNPs detected in the

Table 3 Polymorphisms detected between the two *Ae. tauschii* accessions

	Number of polymorphic sites in contigs ≥100 bp	
	Spike	Leaf*
Total	28,268 (8,592 contigs)	10,355 (3,571 contigs)
All SNPs	25,837 (8,325 contigs)	9,429 (3,441 contigs)
HC SNPs	13,347 (4,872 contigs)	4,335 (1,747 contigs)
Indel	2,108 (1,444 contigs)	794 (542 contigs)
Multiple nucleotide polymorphisms	323 (298 contigs)	112 (100 contigs)
Average bp per HC SNP	1,435	1,854

* Obtained after filtering the data from Iehisa et al. (2012)

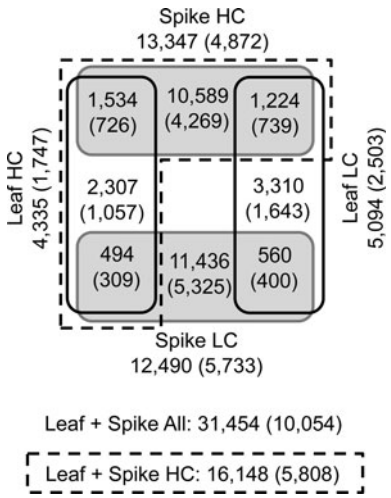

Fig. 1 Comparison of SNPs found in contigs of leaf and spike transcripts. The number of SNPs in leaves is indicated *inside black continuous lines*, of spikes in *shaded boxes* and HC SNPs of both tissues inside the *broken black line*. The number of contigs is indicated in *parentheses*. In the case of common SNPs, the number of contigs is represented by the number of spike contigs because of their longer sequences

spike libraries with those in the leaf libraries (Table 3). Out of the SNPs found in both tissues, SNPs considered as HC in either of the tissues were counted as HC SNPs. In total, 31,454 NR SNPs were detected in 10,054 contigs of both tissues (Fig. 1), and 16,148 SNPs (51 %) in 5,808 contigs were classified as HC. Consequently, SNPs were detected in 23 % of the NR contigs (5,808/24,875), and 10,589 of the 13,347 HC SNPs found in spikes were new.

In silico mapping of polymorphic contigs

A draft genome sequence was recently reported in *Ae. tauschii*, and around 40 % scaffolds of the genomic sequence was anchored to the *Ae. tauschii* chromosomes (Jia et al. 2013). In addition, a 4.03 Gb physical map of *Ae. tauschii* has been published, and approximately 61 Mb of the genomic sequences obtained from the extended SNP marker sequences, which were used for anchoring bacterial artificial chromosome contigs to the linkage map, is now available (Luo et al. 2013). We first integrated both sets of genomic information by performing a blastn search of the extended marker sequences against the draft genome sequence of *Ae. tauschii*. Hits were obtained for all extended markers with an *E*-value $\leq 2 \times 10^{-152}$. In addition to 13,688 scaffolds (a total of 1.28 Gb) anchored to the Y2280/AL8/78 linkage map of *Ae. tauschii* (Jia et al. 2013), 3,188 scaffolds were anchored to the AL8/78/AS75 linkage map (Luo et al. 2013), resulting in a total of 1.49 Gb of sequence anchored to the *Ae. tauschii* maps (Table S3). Next, the polymorphic and NR contigs

Table 4 Number of contigs mapped to the draft genome of *Ae. tauschii* and barley

	Ae. tauschii genome	Barley genome
NR contigs with SNPs	10,054 (5,808)	
Aligned to the genome	10,022 (5,642)	9,569 (5,435)
Spike	7,621 (4,262)	7,292 (4,112)
Leaf*	2,253 (1,053)	2,142 (1,009)
Common	1,646 (1,260)	1,595 (1,227)
Anchored to the genetic map	8,848 (5,122)	9,516 (5,407)
Spike	6,812 (3,871)	7,251 (4,091)
Leaf*	1,415 (705)	2,131 (1,003)
Common	1,522 (1,170)	1,589 (1,223)

Number of HC contigs is indicated in parenthesis

* Data published by Iehisa et al. (2012)

Table 5 Number of polymorphic contigs assigned to *Ae. tauschii* and barley chromosomes

Anchored to D-genome genetic map			Barley physical map		
	All	HC		All	HC
1D	1,202	724	1H	1,179	695
2D	1,401	799	2H	1,504	839
3D	1,313	721	3H	1,446	775
4D	1,128	662	4H	1,176	687
5D	1,498	875	5H	1,551	907
6D	1,071	623	6H	1,255	708
7D	1,235	718	7H	1,405	796
Total	8,848	5,122	Total	9,516	5,407

obtained in the present study were mapped in silico to the *Ae. tauschii* draft genome sequence based on blastn searches. Hits were obtained for 24,633 (99.0 %) of the 24,875 NR contigs. For the contigs with SNPs, 10,022 (99.7 %) out of 10,054 were confirmed to be present in the D-genome draft sequence, and 6,179 (61.5 %) were successfully mapped to the anchored scaffolds. Additionally, it has been reported that genomic scaffolds can be anchored to known wheat linkage maps using simple sequence repeat marker and EST sequences (Jia et al. 2013). Using this information, the number of contigs with SNPs anchored to chromosomes of the D genome increased to 8,848 (Tables 4 and S4). Similarly, 5,642 of the contigs with HC SNPs were mapped to the *Ae. tauschii* genome, and 5,122 were anchored in silico to the *Ae. tauschii* maps. On average, 732 contigs with HC polymorphisms were mapped to each *Ae. tauschii* chromosome (Table 5).

The polymorphic and NR contigs were also aligned to the barley draft genome because of the extensive conserved synteny between barley and wheat chromosomes (Mayer et al. 2011). blastn hits were obtained for 22,711 NR

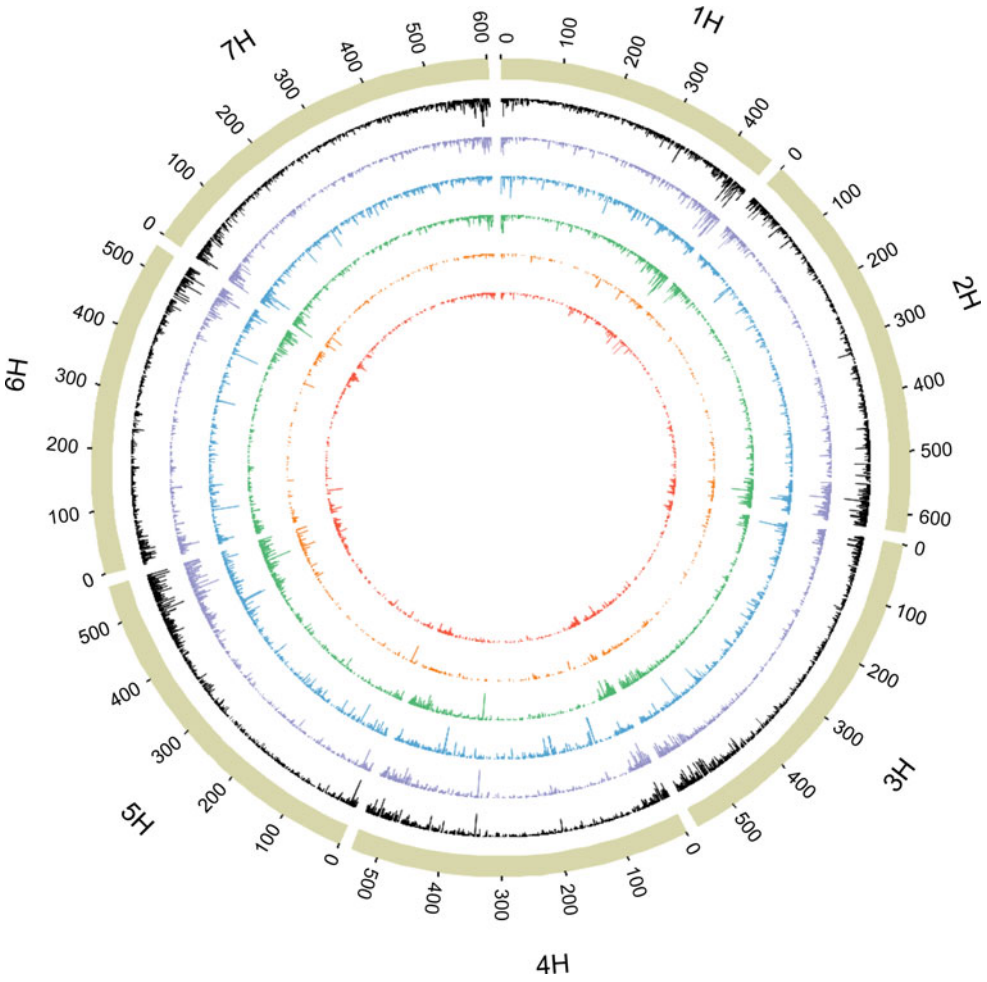

Fig. 2 Distribution of contigs and HC SNPs on the physical map of barley. From the *outer circle to inner*: track 1 shows the physical map of barley (scale in Mb), track 2 represents a histogram of the number of mapped NR contigs, track 3 the number of all HC SNPs, track 4 the number of HC SNPs per mapped NR contig in the same region, track 5 the number of HC SNPs detected only in spikes, track 6 the number of HC SNPs detected only in leaves and track 7 the number detected in both tissues. The distribution of contigs and SNPs is summarized in non-overlapping 250 kb intervals of the barley physical map

contigs, of which 22,549 were assigned to the barley chromosomes. In total, 9,516 polymorphic contigs were assigned to the barley chromosomes, which included 5,407 contigs with HC SNPs (Table 4). On average, 772 HC polymorphic contigs were mapped to each barley chromosome (Table 5). Because the genomic sequences anchored to the linkage map are longer in barley (>3.09 Gb) than in *Ae. tauschii*, and physical map information is available in barley (IBSC 2012), the HC SNPs of *Ae. tauschii* were plotted onto the barley chromosomes (Fig. 2). The number of both SNPs and NR contigs was higher in telomeric regions than in centromeric regions.

Marker development and mapping to chromosome 2DS of *Ae. tauschii*

To assess the usefulness of the SNP dataset for construction of wheat linkage maps, 2DS-specific markers were designed based on the in silico mapping of polymorphic contigs. This chromosome arm was chosen because several agriculturally important loci, such as *Net2*, a causal D-genome gene for hybrid necrosis in interspecific crosses between tetraploid wheat and *Ae. tauschii* (Mizuno et al. 2011; Matsuda et al. 2012), and some of grain shape-related QTLs (Okamoto et al. 2012), have been reported. In total, 40 markers were new and 8 were discarded because the PCR products were longer than expected. Of the remaining primer sets, 3 were indel markers, 9 were CAPS and 20 were HRM markers. Polymorphisms using the 32 markers were examined in 20 accessions of *Ae. tauschii*, which were selected from each of the sublineages reported by Mizuno et al. (2010a) to cover the species' genetic diversity. Two HRM markers and one CAPS marker were excluded from further study because we failed to detect any polymorphism in these markers. Based on the allelic diversity within the 20 accessions, PIC values were calculated, which ranged from 0.09 to 0.57, with an average of 0.29 (Fig. 3). Out of the 29 markers, 2 were IG47182-specific SNP alleles and 1 was a PI476874-specific SNP allele (PIC = 0.09) (Fig. S3). In 21 of the markers (72 %), the PIC value was greater than 0.20, indicating that they were polymorphic among most of the accessions; these included 6 markers that clearly distinguished L1 from L2 accessions (Fig. S3).

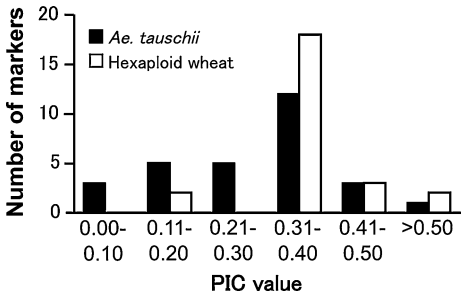

Fig. 3 Frequency distribution of PIC values of the markers developed. PIC values of 29 markers among the 20 *Ae. tauschii* accessions (*black bars*) and 25 markers among the 10 hexaploid wheat *lines* (*open bars*)

To confirm the 2DS-specific assignment of the markers developed, an F_2 population of PI476874/IG47182 was used for genotyping. One of the HRM markers (*Xctg10285*) failed in the genotyping because of difficulty distinguishing the alleles. All of the remaining 28 markers (96.6 %) were successfully assigned to chromosome 2DS (Fig. 4). Segregation distortion has been reported for chromosome 2D of this population (Iehisa et al. 2012). Segregation distortion was observed in the chromosomal region between the marker bags20g23 (HRM13) (220.3 cM) and *Xgwm539* (347.1 cM). In this region, the frequency of the homozygous PI476874-allele was significantly lower than expected, whereas the homozygous IG47182-allele appeared at high frequency (Fig. S4; Table S5).

Polymorphism and mapping of the markers in hexaploid wheat

To examine the usefulness of the markers developed for detection of polymorphisms in hexaploid wheat, three common wheat cultivars, three L1-derived synthetic wheat lines and four L2-derived synthetics were genotyped. All the synthetic hexaploids were obtained by crossing the tetraploid wheat cultivar Ldn with the respective L1 or L2 accessions of *Ae. tauschii* (Takumi et al. 2009b; Kajimura et al. 2011). The synthetic wheat line derived from the L1 accession used in the RNA-seq analysis, Ldn/PI476874, was also genotyped, but no synthetic line derived from the L2 accession was genotyped because triploid hybrids of Ldn/IG47182 exhibited hybrid lethality (Hatano et al. 2012). In addition to the 29 markers, 14 previously developed HRM markers were also used (Table S6). Of the 43 markers, 25 (58 %) were polymorphic, defined as showing a distinct genotype in at least one of the 10 hexaploid wheat lines. Of the three marker types, 75 % of the CAPS markers (six out of eight) were polymorphic, followed by 67 % of indel (two out of three) and 53 % of HRM markers (17 out of 32). The low polymorphism of HRM markers in hexaploid wheat might be explained by a decrease in sensitivity of HRM analysis due to the presence of the A and B genomes. PIC values ranged from 0.16 to 0.56 (average 0.36), and the markers with the lowest PIC values detected specific alleles in Ldn/PI476874 (Fig. 3). Six of the polymorphic markers clearly distinguished the L1-derived hexaploids from the L2-derived synthetics and common wheat cultivars (Fig. S5). In most of the markers (92 %), the PIC values were greater than 0.20, indicating high polymorphism, especially between L1 and L2 synthetics and between L1 and common wheat, as supposed previously (Iehisa et al. 2012). The HRM marker *Xctg216781*, which recognized an allele specific to IG47182 among the 20 *Ae. tauschii* accessions, distinguished common wheat cultivars from the wheat synthetics. For three HRM markers, polymorphisms were observed even among common wheat cultivars.

To examine whether these markers could be mapped to 2DS in the allohexaploid background, F_2 individuals of two mapping populations were genotyped. In the N61//Ldn/PI476874 population, 23 of the 25 polymorphic markers between the parents were available for genotyping of the F_2 plants, and no segregation distortion was observed (Table S7). Four of the 5 (80 %) CAPS markers and 3 of the 16 (19 %) HRM markers were used as dominant markers, although 6 of these 7 markers mapped as co-dominant markers in the diploid mapping population (Fig. S6). In total, 22 markers were assigned to 2DS (Fig. 4) and the assignment of 1 marker (*Xctg05269*) was ambiguous between chromosomes 2B and 5D.

The second F_2 population derived from the cross between L1 and L2 synthetics (Ldn/KU-2075//Ldn/KU-2025) varied only in the D genome, in contrast to the N61//Ldn/PI476874 population. First, we checked the polymorphism of the 29 markers developed in this study and 6 HRM markers developed by Iehisa et al. (2012) between the parental synthetics. Only 14 % of the markers (5/35) were polymorphic and could thus be used for genotyping of the F_2 plants. The low polymorphism implied that the markers based on polymorphisms between PI476874 and IG47182 might be less useful in some cross combinations of L1 and L2 accessions. It might also be explained in part by the low sensitivity of the HRM procedure in allopolyploid species. Two CAPS markers and two HRM markers were used as dominant markers in the Ldn/KU-2075//Ldn/KU-2025 population. All five markers mapped to chromosome 2DS, and no segregation distortion was observed (Fig. 4; Table S8).

Using the PI476874/IG47182 and N61//Ldn/PI476874 linkage maps, in which more than 20 ESTs were assigned, the relationship between genetic and physical distances was evaluated. The physical map of barley was used in this analysis because the number of genomic scaffolds

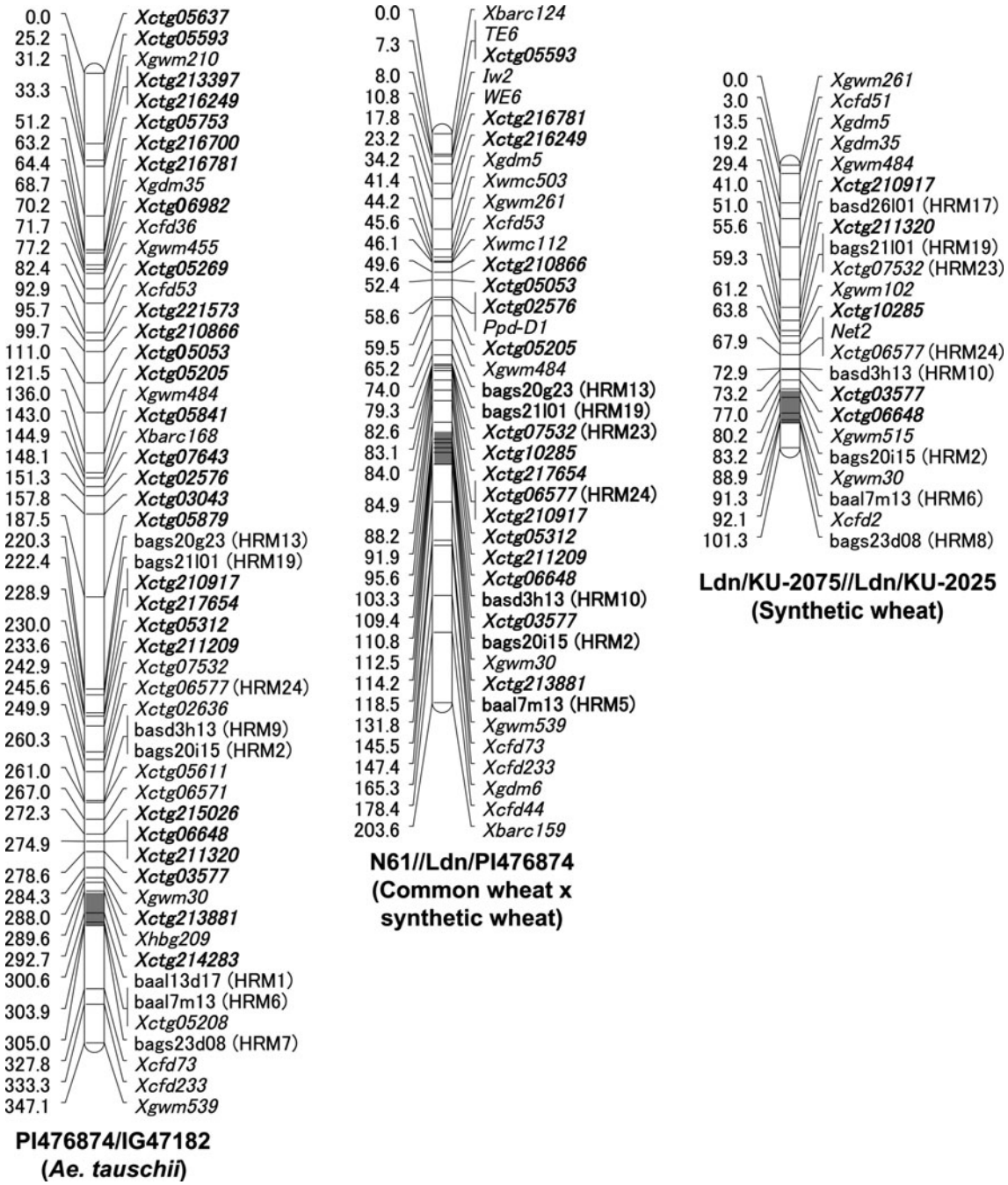

Fig. 4 Genetic maps of chromosome 2D of the *Ae. tauschii* and hexaploid wheat F_2 populations. The diploid PI476874/IG47182 map (*left*) and hexaploid maps of N61//Ldn/PI476874 (*middle*) and Ldn/KU-2075//Ldn/KU-2025 (*right*) populations are shown. *Gray bars* indicate the putative centromere positions. The markers mapped in this study (indicated in *bold*) were placed on previously constructed linkage maps. Map distances are shown in centimorgans (cM)

aligned to the map was larger than in *Ae. tauschii*. The recombination frequency on 2DS greatly decreased at distances farther than 100 Mb from the distal end of the physical map in both the *Ae. tauschii* and hexaploid wheat maps (Fig. 5). This result was consistent with previous reports in barley and *Ae. tauschii* (IBSC 2012; Luo et al. 2013).

Discussion

In the present study, RNA-seq analysis of transcripts from the spikes of two *Ae. tauschii* accessions was performed for SNP discovery and marker development. Although the number of reads and total read length did not differ greatly from values for leaves, the number of contigs and isotigs

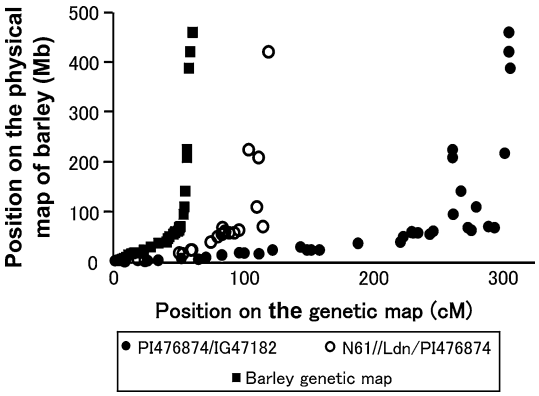

Fig. 5 Relationship between physical and genetic distances. Based on BLAST hits of the marker sequences, physical positions on the barley genome were determined. Squares indicate the positions of markers on the PI476874/IG47182 map, diamonds the positions on the N61//Ldn/PI47182 map and triangles the positions on the barley genetic map (IBSC 2012)

was twice as high and the number of HC SNPs was three times higher in the spike transcripts, and at least 10,589 SNPs were new (Tables 2, 3). These results indicated that sequencing of the more diverse set of genes expressed in spikes yielded higher detection of SNPs than in leaves, and also suggested that the SNP detection rate was slightly higher for genes expressed in spikes than in leaves (Table 3; Iehisa et al. 2012).

The draft genome sequence (Jia et al. 2013) and the physical map (Luo et al. 2013) of *Ae. tauschii* will facilitate the development of molecular markers in the wheat D genome. However, the number of genomic sequences anchored to the linkage map remains small compared with barley (IBSC 2012), which is supported by the in silico mapping results of the present study. Although the number of polymorphic contigs assigned to the *Ae. tauschii* genome was slightly higher than that of barley, the number of contigs anchored to the linkage map was somewhat higher for barley (Table 5). These results imply that *Ae. tauschii* genomic information may be complemented by that of barley for development of molecular markers in specified regions of the D genome.

The in silico mapping analysis revealed that both contigs and SNPs derived from leaves and spikes were predominantly distributed in the distal portions of chromosomes (Fig. 2), consistent with recent reports in barley (IBSC 2012) and *Ae. tauschii* (Luo et al. 2013). Genome analyses in barley and *Ae. tauschii* show that distal portions of chromosomes are more gene rich (IBSC 2012; Luo et al. 2013). In the case of the short arm of chromosome 2D, the number of contigs and SNPs that mapped to the barley physical map of 2HS was larger within the region of the first 100 Mb from the distal end, and this 100 Mb region coincided with high occurrence of recombination events (Figs. 2, 5; Luo et al. 2013). Thus, SNP discovery through RNA-seq was especially efficient in the distal portion of barley and wheat chromosomes, where gene density and recombination rate are correlated (Luo et al. 2013).

In the present study as well as in our previous report (Iehisa et al. 2012), molecular markers developed using information generated through RNA-seq are generally polymorphic between L1 and L2 accessions of *Ae. tauschii* (Fig. S3). Because a restricted population within L2 appears to be the main source of the D genome of common wheat (Mizuno et al. 2010a, b; Wang et al. 2013), these markers were also expected to be polymorphic between the D genomes of L1-derived synthetic hexaploid wheat and common wheat (Fig. S5). SNPs can be converted to PCR-based markers such as CAPS and HRM. In a hexaploid background, the leaky amplification of A- and B-genome sequences interferes with accurate SNP typing of the D genome. The interference results in an increase in the frequency of dominant markers, especially for conversion to a CAPS marker, and a decrease in the sensitivity of SNP detection in HRM analysis (Matsuda et al. 2012). Failure of SNP detection in CAPS markers occurs only when the fragmentation patterns of the A and B genomes mask both the digested and undigested alleles of the D genome. In such cases, the allopolyploid nature of the common wheat genome may make genotyping by CAPS and HRM markers difficult.

Many experimental platforms for SNP typing have also been applied to hexaploid wheat (Akhunov et al. 2009; Allen et al. 2011, 2013; Cavanagh et al. 2013). However, due to the comparatively low genetic diversity of the D genome of common wheat (Caldwell et al. 2004; Chao et al. 2009), the number of markers mapped to the D genome is usually three- to fivefold lower compared to the A and B genomes (Allen et al. 2011, 2013; Cavanagh et al. 2013). For wheat breeding, *Ae. tauschii* is a promising germplasm because of its abundant variation in various traits (Jones et al. 2013). Although nonadditive expression of homeologous genes is often observed during wheat synthetic formation (Pumphrey et al. 2009), the natural variation found in *Ae. tauschii* is significantly expressed in many traits under the allohexaploid genetic background of wheat synthetics (Kajimura et al. 2011; Iehisa and Takumi 2012; Okamoto et al. 2012). Thus, the wide natural variation found in *Ae. tauschii*, especially in the L1 accessions, can be exploited to introduce agronomically beneficial alleles, which may have been left behind during the allopolyploidization event (Ogbonnaya et al. 2005). Our results suggested the potential usefulness, at least in part, of the SNPs discovered between the L1 and L2 accessions of *Ae. tauschii* for genome analysis and molecular breeding of hexaploid wheat. SNPs might be valuable for marker development of the D genome, particularly in the progeny obtained through crosses between common wheat cultivars

and L1-derived synthetics. On the other hand, the low polymorphism of the markers between Ldn/KU-2075 and Ldn/KU-2025 suggests that finding more SNPs in the D genome between and within the two lineages will be necessary to apply them widely to *Ae. tauschii* and hexaploid wheat lines.

Acknowledgments The authors would like to thank Ms. Yuka Motoi at Okayama University for technical assistance. This work was supported by a grant from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan (Grant-in-Aid for Scientific Research (B) No. 25292008) to ST, and by MEXT as part of the Joint Research Program implemented at the Institute of Plant Science and Resources, Okayama University in Japan.

Conflict of interest The authors declare that they have no conflict of interest.

References

Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet 119:507–517

Allen AM, Barker GLA, Berry ST, Coghill JA, Gwilliam R, Kirby S, Robinson P, Brenchley RC, D'Amore R, McKenzie N, Waite D, Hall A, Bevan M, Hall N, Edwards KJ (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnol J 9:1086–1099

Allen AM, Barker GLA, Wilkinson P, Burridge A, Winfield M, Coghill J, Uauy C, Griffiths S, Jack P, Berry S, Werner P, Melichar JPE, McDougall J, Gwilliam R, Robinson P, Edwards KJ (2013) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). Plant Biotechnol J 11:279–295

Caldwell KS, Dvorak J, Lagudah ES, Akhunov E, Luo MC, Wolters P, Powell W (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167:941–947

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenacc C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, Lopez da Silva M, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci USA 110:8057–8062

Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo MC, Dubcovsky J (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol Breed 23:23–33

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Dudnikov AJ, Kawahara T (2006) *Aegilops tauschii*: genetic variation in Iran. Genet Resour Crop Evol 53:579–586

Dvorak J, Luo MC, Yang ZL, Zhang HB (1998) The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. Theor Appl Genet 97:657–670

Feldman M (2001) Origin of cultivated wheat. In: Bonjean AP, Angus WJ (eds) The world wheat book: a history of wheat breeding. Lavoisier Publishing, Paris, pp 3–53

Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Buell CR (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7:e33071

Hatano H, Mizuno N, Matsuda R, Shitsukawa N, Park P, Takumi S (2012) Dysfunction of mitotic cell division at shoot apices triggered severe growth abortion in interspecific hybrids between tetraploid wheat and *Aegilops tauschii*. New Phytol 194: 1143–1154

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Iehisa JCM, Takumi S (2012) Variation in abscisic acid responsiveness of *Aegilops tauschii* and hexaploid wheat synthetics due to the D-genome diversity. Genes Genet Syst 87:9–18

Iehisa JCM, Shimizu A, Sato K, Nasuda S, Takumi S (2012) Discovery of high-confidence single nucleotide polymorphisms from large-scale de novo analysis of leaf transcripts of *Aegilops tauschii*, a wild wheat progenitor. DNA Res 19:487–497

International Barley Genome Sequencing Consortium et al (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

Jia J, Shancen Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KFX, Dong Li D, Pan S, Fengya Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Gao J, Middleton C, Quan Z, Liu G, Wang J, IWGSC, Yang H, Xu Liu X, He Z, Mao L, Wang J (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95

Jones H, Gosman N, Horsnell R, Rose GA, Everest LA, Bentley AR, Tha S, Uauy C, Kowalski A, Novoselovic D, Simek R, Kobiljski B, Kondic-Spika A, Brbaklic K, Mitrofanova O, Chesnokov Y, Bonnett D, Greenland A (2013) Strategy for exploiting exotic germplasm using genetic, morphological, and environmental diversity: the *Aegilops tauschii* Coss. example. Theor Appl Genet 126:1793–1808

Kajimura T, Murai K, Takumi S (2011) Distinct genetic regulation of flowering time and grain-filling period based on empirical study of D-genome diversity in synthetic hexaploid wheat lines. Breed Sci 61:130–141

Kihara H (1944) Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare*. Agric Hortic 19:889–890 (In Japanese)

Kihara H, Lilienfeld F (1949) A new-synthesized 6x-wheat. Hereditas 35(Suppl):307–319

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172–175

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:178–181

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, Martis MM, Mayer KFX, Sehgal SK, Li W, Gill BS, Bevan MW, Šimková H, Doležel J, Weining S, Lazo GR, Anderson OD, Dvorak J (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of

Aegilops tauschii, the wheat D-genome progenitor. Proc Natl Acad Sci USA 110:7940–7945

Matsuda R, Iehisa JCM, Takumi S (2012) Application of real-time PCR-based SNP detection for mapping of *Net2*, a causal D-genome gene for hybrid necrosis in interspecific crosses between tetraploid wheat and *Aegilops tauschii*. Genes Genet Syst 87:137–143

Matsuoka Y, Nasuda S (2004) Durum wheat as a candidate for the unknown female progenitor of bread wheat: an empirical study with a highly fertile F_1 hybrid with *Aegilops tauschii* Coss. Theor Appl Genet 109:1710–1717

Matsuoka Y, Takumi S, Kawahara T (2007) Natural variation for fertile triploid F_1 formation in allohexaploid wheat speciation. Theor Appl Genet 115:509–518

Matsuoka Y, Takumi S, Kawahara T (2008) Flowering time diversification and dispersal in central Eurasian wild wheat *Aegilops tauschii* Coss.: genealogical and ecological framework. PLoS ONE 3:e3138

Matsuoka Y, Nishioka E, Kawahara T, Takumi S (2009) Genealogical analysis of subspecies divergence and spikelet-shape diversification in central Eurasian wild wheat *Aegilops tauschii* Coss. Plant Syst Evol 279:233–244

Mayer KFX, Martis M, Hedley PE, Simková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubaláková M, Suchánková P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Dolezel J, Waugh R, Stein N (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263

McFadden ES, Sears ER (1944) The artificial synthesis of *Triticum spelta*. Rec Genet Soc Am 13:26–27

Mizuno N, Yamasaki M, Matsuoka Y, Kawahara T, Takumi S (2010a) Population structure of wild wheat D-genome progenitor *Aegilops tauschii* Coss.: implications for intraspecific lineage diversification and evolution of common wheat. Mol Ecol 19:999–1013

Mizuno N, Hosogi N, Park P, Takumi S (2010b) Hypersensitive response-like reaction is associated with hybrid necrosis in interspecific crosses between tetraploid wheat and *Aegilops tauschii* Coss. PLoS ONE 5:e11326

Mizuno N, Shitsukawa N, Hosogi N, Park P, Takumi S (2011) Autoimmune response and repression of mitotic cell division occur in inter-specific crosses between tetraploid wheat and *Aegilops tauschii* Coss. that show low temperature-induced hybrid necrosis. Plant J 68:114–128

Ogbonnaya FC, Halloran GM, Lagudah ES (2005) D genome of wheat—60 years on from Kihara, Sears and McFadden. In: Tsunewaki K (ed) Frontiers of wheat bioscience, the 100th memorial issue of Wheat Information Service. Kihara Memorial Foundation for the Advancement of Life Sciences, Yokohama, pp 205–220

Okamoto Y, Kajimura T, Ikeda TM, Takumi S (2012) Evidence from principal component analysis for improvement of grain shape- and spikelet morphology-related traits after hexaploid wheat speciation. Genes Genet Syst 87:299–310

Park SDE (2001) Trypanotolerance in west African cattle and the population genetic effects of selection. Dissertation, University of Dublin

Pumphrey M, Bai J, Laudencia-Chingcuanco D, Anderson O, Gill BS (2009) Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics 181:1147–1157

Takumi S, Nishioka E, Morihiro H, Kawahara T, Matsuoka Y (2009a) Natural variation of morphological traits in wild wheat progenitor *Aegilops tauschii* Coss. Breed Sci 59:579–588

Takumi S, Naka Y, Morihiro H, Matsuoka Y (2009b) Expression of morphological and flowering time variation through allopolyploidization: an empirical study with 27 wheat synthetics and their parental *Aegilops tauschii* accessions. Plant Breed 128:585–590

Trethowan RM, Mujeeb-Kazi A (2008) Novel germplasm resources for improving environmental stress tolerance of hexaploid wheat. Crop Sci 48:1255–1265

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Wang J, Luo MC, Chen Z, You FM, Wei Y, Zheng Y, Dvorak J (2013) *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. New Phytol 198:925–937

You FM, Huo N, Deal KR, Gu YQ, Luo MC, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59