Plant Biotechnology Journal

aab Association of Applied Biologists SEB Society for Experimental Biology

Plant Biotechnology Journal (2013), pp. 1–9 doi: 10.1111/pbi.12072

Sequence-based SNP genotyping in durum wheat

Remco M.P. van Poecke[1], Marco Maccaferri[2], Jifeng Tang[1], Hoa T. Truong[1], Antoine Janssen[1], Nathalie J. van Orsouw[1], Silvio Salvi[2], Maria C. Sanguineti[2], Roberto Tuberosa[2] and Edwin A.G. van der Vossen[1,]*

[1]*Keygene N.V., Wageningen, The Netherlands*

[2]*Department of Agricultural Sciences, University of Bologna, Bologna, Italy*

Received 21 December 2012;
revised 8 March 2013;
accepted 15 March 2013.

**Correspondence* (Tel. +31 (0) 317 466 866; fax +31 (0)317 424 939; email edwin.van-der-vossen@keygene.com)

Keywords: illumina sequencing, high-density mapping, recombinant inbred lines, *Triticum durum* Desf., genome complexity reduction, *Pst*I/*Taq*I.

Summary

Marker development for marker-assisted selection in plant breeding is increasingly based on next-generation sequencing (NGS). However, marker development in crops with highly repetitive, complex genomes is still challenging. Here we applied sequence-based genotyping (SBG), which couples AFLP®-based complexity reduction to NGS, for *de novo* single nucleotide polymorphisms (SNP) marker discovery in and genotyping of a biparental durum wheat population. We identified 9983 putative SNPs in 6372 contigs between the two parents and used these SNPs for genotyping 91 recombinant inbred lines (RILs). Excluding redundant information from multiple SNPs per contig, 2606 (41%) markers were used for integration in a pre-existing framework map, resulting in the integration of 2365 markers over 2607 cM. Of the 2606 markers available for mapping, 91% were integrated in the pre-existing map, containing 708 SSRs, DArT markers, and SNPs from CRoPS technology, with a map-size increase of 492 cM (23%). These results demonstrate the high quality of the discovered SNP markers. With this methodology, it was possible to saturate the map at a final marker density of 0.8 cM/marker. Looking at the binned marker distribution (Figure 2), 63 of the 268 10-cM bins contained only SBG markers, showing that these markers are filling in gaps in the framework map. As to the markers that could not be used for mapping, the main reason was the low sequencing coverage used for genotyping. We conclude that SBG is a valuable tool for efficient, high-throughput and high-quality marker discovery and genotyping for complex genomes such as that of durum wheat.

Introduction

Marker-assisted selection (MAS) facilitates breeding by increasing efficiency and accuracy. This is especially true for complex traits where individual loci have only limited effects on the trait. The use of markers in breeding ranges from mapping and selecting loci controlling valuable traits (e.g. disease resistance, phenology, quality, yield), to genomic selection and varietal identification (Collard and Mackill, 2008; Gupta *et al.*, 1999; Heffner *et al.*, 2009; Masojć, 2002; Tester and Langridge, 2010). The optimal/minimum marker density required in such studies varies greatly and depends on several factors such as the type of breeding scheme, the species, genome size, genome-wide distribution of markers, linkage disequilibrium decay, population size, and structure and type of analysis. In general, up to several thousands of markers can be required. In comparison with animal species, marker development in crop plants is often more difficult as crop plants tend to have larger genomes, more transposable elements, larger gene-families, and may have higher ploidy levels (Morrell *et al.*, 2011). Thus, within a plant genome there are on average more homologous loci compared to animal genomes, which greatly complicates marker development. Partially as a result of this, the development of high-density markers for GWA mapping is relatively straightforward in humans (Lewis *et al.*, 2011), while it is still a challenge in wheat (Fleury *et al.*, 2010; Maccaferri *et al.*, 2011).

To enhance the use of MAS in plant breeding, a marker discovery and genotyping method should have the potential to target thousands of loci besides being scalable, both in sample size and number of loci targeted, cost-effective, and not limited to species for which a sequenced reference genome is available. Next-generation sequencing (NGS) has ushered in such genotyping methods. One way of applying NGS to MAS is by sequencing the genomes of multiple genotypes of the same species to identify sequence polymorphisms such as single nucleotide polymorphisms (SNPs, Ossowski *et al.*, 2008; Wheeler *et al.*, 2008; Xu *et al.*, 2012). After this polymorphism identification phase, large numbers of samples can be genotyped using more cost-effective methods such as high-density SNP-genotyping arrays and fluorescent allelic discrimination assays (Cuppen, 2007; Fan *et al.*, 2006). Notwithstanding the spectacular progress in sequencing during the past decade, whole-genome sequencing in species with a large and complex genome is still very costly. In such cases, the polymorphism discovery phase can be simplified by reducing genome complexity, as allowed with the CRoPS® method (van Orsouw *et al.*, 2007). Reduction of genome complexity as a prerequisite to discover SNPs has been successfully applied to animal and plant species (Davey *et al.*, 2011). This notwithstanding, in polyploid species with a highly repetitive genome, marker conversion from polymorphism identification platform to genotyping platform has proven difficult (Trebbi *et al.*, 2011).

Please cite this article as: van Poecke, R.M.P., Maccaferri, M., Tang, J., Truong, H.T., Janssen, A., van Orsouw, N.J., Salvi, S., Sanguineti, M.C., Tuberosa, R. and van der Vossen, E.A.G. (2013) Sequence-based SNP genotyping in durum wheat. *Plant Biotechnol J.,* doi: 10.1111/pbi.12072

© 2013 Society for Experimental Biology, Association of Applied Biologists and John Wiley & Sons Ltd

Genotyping generally consists of two phases, a polymorphism discovery phase and a genotyping phase. However, with the advent of powerful NGS technologies both phases can be combined in a single experiment. In fact, in order to avoid marker conversion and to improve the efficiency and cost-effectiveness of genotyping workflow, genotyping itself can be performed by NGS as well. Examples are RAD (sequencing of restriction site-associated genomic DNA) (Baird *et al.*, 2008) and genotyping-by-sequencing (GBS) (Elshire *et al.*, 2011; Poland *et al.*, 2012) methods. Both methods rely on restriction enzyme-based complexity reduction of genomic DNA coupled to high-throughput short-read sequencing. By using methylation-sensitive restriction enzymes, these methods reduce the effect of repetitive sequences while, through complexity reduction, allowing for the genotyping of large genomes, without the need of a priori sequence information. Recently, a highly flexible genotyping method, belonging to the GBS-category and based on high-throughput sequencing, was introduced, called sequence-based genotyping (SBG) (Truong *et al.*, 2012). In SBG, different stringency settings are desired for the polymorphism discovery phase versus the genotyping phase. This is because, at the genotyping phase, both the position and the type (in case of SNPs e.g. a C-to-A or a G-to-C etc.) of each polymorphism are already known, whereas at the discovery phase they are not. In practice, this translates to higher sequencing coverage desired for the discovery phase compared to the genotyping phase. This simple, cost-effective method requires neither marker conversion nor a priori sequence information while being highly scalable, as it combines robust AFLP®-based complexity reduction with high-throughput short-read sequencing, thus allowing for an additional level of complexity reduction through the use of selective nucleotides in fragment amplification. In this paper, we describe the results of applying SBG to durum wheat (Triticum durum Desf.), a tetraploid species with a highly repetitive genome.

Several types of molecular markers have been applied for durum wheat linkage map construction including RFLP, SSR, AFLP, and DaRT markers (Blanco *et al.*, 1998; Korzun *et al.*, 1999; Lotti *et al.*, 2000; Mantovani *et al.*, 2008; Röder *et al.*, 1998). Recently, it became evident that SNP markers are more amenable to high-throughput discovery and genotyping due to their abundance, uniform genome distribution, and cost-effectiveness (Gupta *et al.*, 2008). However, the highly repetitive, polyploid nature of wheat combined with a particularly low polymorphism has hampered SNP detection and genotyping in the Triticeae (Haudry *et al.*, 2007; Koebner and Summers, 2003; Somers *et al.*, 2003). In order to cope with the highly repetitive nature of the wheat genome, complexity reduction targeting repetitive sequences, either by using mRNA or by digestion with methylation-sensitive restriction enzymes has been applied within the workflow of SNP detection and/or genotyping protocols (Akbari *et al.*, 2006; Bernardo *et al.*, 2009). Complexity reduction coupled with 454 sequencing (i.e. CRoPS) has recently allowed for the discovery of over 2500 durum wheat SNPs (Trebbi *et al.*, 2011). A subset of these SNPs was used for genotyping with the Illumina Golden Gate assay, resulting in a 36% successful marker-conversion rate (Trebbi *et al.*, 2011). In such case, even though the SNP discovery phase can be considered as high throughput, the genotyping phase is not so, due to marker-conversion problems mainly caused by paralogous and homeologous sequences.

More recently, GBS was applied to bread-wheat, resulting in the incorporation of thousands of markers in the bread-wheat map (Poland *et al.*, 2012). Here, we demonstrate high-throughput SNP discovery and genotyping in durum wheat using SBG using as a proof of concept 92 RILs derived from a cross between the two elite cultivars Colosseo (CLS) and Lloyd (LLD) (Mantovani *et al.*, 2008).

Results

SNP discovery

For SNP discovery, we sequenced four samples (both the CLS and LLD parental genotypes as well as two randomly selected RILs) at relatively high coverage (four samples per lane). In total, 24 407 879 reads were generated, of which 23 315 715 contained one of the four sample identification tags as well as the expected restriction site motif. After trimming of the sample identification tags, these 23 315 715 reads were used for sequence assembly resulting in 221 836 contigs and 1 904 275 singletons. Contigs not starting with the restriction site motif or with a length not between 70 and 76 nucleotides were filtered out. The remaining 198 351 contigs were used as a reference for mapping. Of the 23 315 715 reads, 12 969 616 (56%) were uniquely mapped to one of the contigs, resulting in an average 15× coverage per sample.

The sequencing data of the two parents were used for SNP discovery. After filtering for average read quality per allele, allele-frequency, allele-constitution, and coverage (see Materials and Methods), 9983 putative SNPs were detected in 6372 contigs. Table 1 summarizes the types of putative SNPs detected.

The two RILs were added to assess whether the identified SNPs did indeed segregate in the population. For 7524 of 9983 SNPs, both RILs were either genotyped as A or B; 1841 (24%) were genotyped as A for both RILs; 2028 (27%) as B for both RILs; and 3655 (49%) were segregating. In addition, the number of heterozygous scored SNPs was assessed for both RILs. Considering unambiguously scored genotypes only (i.e. A, B, or H), 8916 SNPs were genotyped for one of the two RILs and 8852 SNPs for the other, of which respectively 261 (2.9%) and 453 (5.1%) were scored as heterozygous.

Genotyping a durum wheat RIL population

The 92 RILs were sequenced at relatively low coverage, namely 13–14 samples per GAII lane. RIL reads were mapped against the reference sequence generated in the SNP discovery experiment, and RILs were genotyped based on the 9983 previously identified SNP positions. For one sample, only very few reads were available. This sample was removed from further analyses. For the other samples, 1.5 ± 0.2 M filtered reads were used for mapping against the reference created in the SNP discovery phase. On average, 41% of the reads were uniquely mapped, resulting in a 3× average coverage per sample. As we sequenced at low coverage, genotyping thresholds were relaxed with respect to the number of reads per allele, to a threshold of 1.

In total, 9980 of 9983 putative SNPs in 6370 of 6372 loci were genotyped. Subsequent filtering on parental genotypes (parental duplicates identical to genotypes from the SNP discovery phase), redundancy (only one marker per contig), and presence across the RIL population (70% of the RILs genotyped per contig), resulted in 2721 of 6370 contigs (43%) that were useful for genotyping (see Table 1). At the threshold level of at least one read per allele, heterozygote scoring is not possible. However, for simple, Type-1 SNPs (see Table 1) this problem is mitigated as the RIL population herein considered has very low levels of residual

Table 1 Identified putative single nucleotide polymorphisms (SNPs) classified by SNP type

		Example					
SNP Type*	Description	P1	P2	Identified SNPs	Contigs represented	Contigs for mapping†	Contigs in map†
1	SNPs with one allele per parent	G	C	3890	2851	2021	1905
2	SNPs with one allele in one parent and two alleles in the other parent	G	G/C	6071	3590	756	500
3	SNPs with two alleles per parent	G/T	G/C	22	21	10	9
4	SNPs with more than two alleles in at least one parent	G	G/C/T	219	na	na	na

*Type-1 SNPs are considered as 'simple' SNPs; Type-2 and Type-3 SNPs are considered 'complex' SNPs; Type-4 SNPs were not included in further analyses.
†Some contigs may be counted twice as they contained multiple SNP types from which a consensus is created during genotyping.

heterozygosity. For more complex Type-2 SNPs, this is more problematic, as the parent-2 genotype from the example in Table 1 cannot be detected accurately with only one read. Thus, for true heterozygous genotypes of low sequencing depth, an artificial inflation of the measured homozygous genotype is expected, which is indeed observed (Figure 1). To avoid such imbalanced markers, markers with an A : B ratio more extreme than 1 : 3 or 3 : 1 were removed, leaving 2606 markers. Note that for Type-3 SNPs, the common allele (G in the example in Table 1) that is observed in both parents is not considered to belong to the segregating locus, whose alleles are represented by the T and C bases in the example in Table 1. For these loci, samples with only the common SNP allele (G) sequenced were designated as unknown. Of the 2606 remaining markers, 2365 (91%) were integrated in the pre-existing framework map containing 708 SSR, DArT, and SNP markers from CRoPS technology (Trebbi *et al.*, 2011; see Table 2; Figure 2; and Figure S1), with a map-size increase of 492 cM (23%).

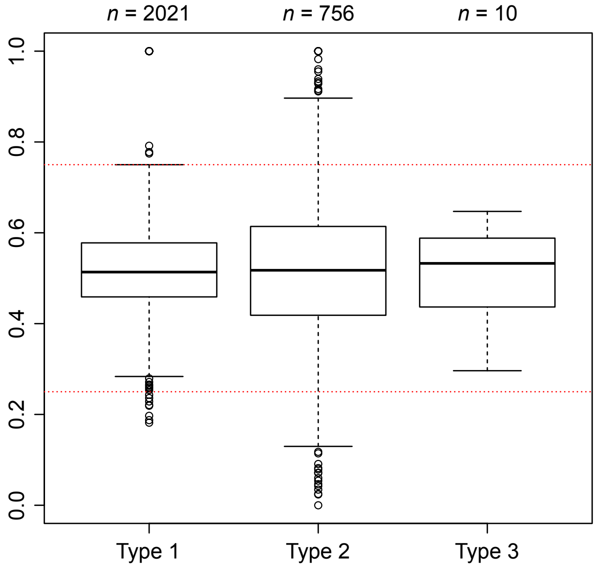

Figure 1 Segregation distortion in complex single nucleotide polymorphisms (SNPs). This figure shows boxplots of the fraction of A-genotyped samples over A-genotyped plus B-genotyped samples for Type-1, Type-2, and Type-3 SNPs before filtering out distorted SNPs. Clearly the Type-2 SNPs contain more distorted SNPs compared to Type-1 or Type-3 SNPs.

The robustness of the integration of the newly developed SNPs into the SSR- and DArT-based framework map was assessed by inspecting the corresponding likelihood of the odds ratio (LOD) scores. For each mapped locus, the marker pairs including the top ten most closely linked loci were sorted by their recombination frequency and LOD scores using the *maximum linkages* option of JoinMap, and the LOD scores were inspected. The top ten marker pairs involving only framework markers (SSR and DArT markers) were mapped at an average LOD score equal to 21.8 ± 3.8 (average over all linkage groups). The top ten marker pairs from the Type-1 SNP markers with <10% missing genotypes (high-quality SNPs) were mapped at an average LOD score equal to 21.6 ± 4.0, which is very close to the average value observed for the framework markers. For all the other mapped SNPs, the top ten maximum linkages had an average LOD score equal to 18.1 ± 4.0.

The global map-size increase observed upon the integration of the new SNPs was subdivided in two portions: (i) interstitial map-size increase, due to increased marker density within linkage groups whose boundaries were defined by framework markers, (ii) map-size increase due to broadening of the existing linkage groups towards either distal or proximal (centromeric) regions of the chromosomes. Between these two positions, interstitial mapping accounted for 39.8% of the global map increase (196 over 492 cM), while expansion of linkage groups accounted for the remaining 60.2%.

Features of the 2365 mapped SNPs such as the locus name, linkage group position, single-nucleotide base polymorphic alleles, and the corresponding contig sequences are reported in Table S2.

Genotyping consistency

Contigs with multiple SNPs can be used to assess the consistency of genotyping. Under the assumption that no recombination occurred within the 70–76 nt contigs, multiple SNPs from the same contig should result in the same genotype. By removing the redundancy filter as to allow multiple SNPs per contig, 3805 SNPs in 2775 contigs were identified in at least 70% of the RILs, of which 738 contigs contained multiple SNPs. In theory, 738 contigs genotyped in 91 RILs should result in 91 × 738 = 67 158 genotypes. Not including missing genotypes (i.e. contig/RIL combinations where SNP information of that contig was missing), 61 610 called genotypes remained. Of these, 2738 (4.4%) showed conflicting information between SNPs belonging to the same contig. Approximately half of these (1239) were 'conflicts' between an SNP or SNPs with missing or unknown data and an

Table 2 Overview of total number of markers per chromosome and map size

		−rSBG map			+rSBG map		
Ancestral Genome	Chromosome	# Markers	# Linkage groups	cM	# Markers	# Linkage groups	cM
A	I	32	1	125	95	1	148
A	II	53	2	147	311	1	217
A	III	39	2	185	179	1	221
A	IV	51	1	102	202	1	113
A	V	27	1	140	136	1	184
A	VI	62	1	141	252	1	177
A	VII	46	2	167	218	2	203
B	I	68	1	161	354	1	207
B	II	66	2	141	261	1	204
B	III	66	1	180	233	1	194
B	IV	25	1	101	188	1	116
B	V	24	1	195	103	1	200
B	VI	66	1	107	271	1	179
B	VII	83	1	223	270	1	244
Total		708	18	2115	3073	15	2607

SBG, sequence-based genotyping.

SNP or SNPs with an A, B, C, D, or H genotype (see example 2 in Table 3 and Figure S1). Thus these ‘conflicts’ did not represent inconsistent genotyping. The remaining 1499 conflicts (i.e. 2.4% of the total called genotypes), did represent inconsistent genotyping (see examples 3 and 4 in Table 3 and Figure S2). Note that the 2.4% inconsistent genotyping does not result in 2.4% error rate in genotyping: of the 1499 conflicts, 583 could be resolved based on a more predominant genotype, whereas 916 were set to U.

Discussion

In many crops, MAS has significantly accelerated plant breeding, especially for complex traits (Tester and Langridge, 2010; Tuberosa *et al.*, 2011). Due to their abundance, SNP markers have become increasingly popular in plant breeding. However, SNP discovery in complex genomes of highly repetitive and/or polyploid nature, such as wheat, remains challenging (Ganal *et al.*, 2009). For such reasons, the initial SNP discovery phase in wheat requires greater efforts as compared to other crops with less complex genomes, and most of the studies reported so far have targeted the expressed portion of the genome. For instance, for SNP discovery, Winfield *et al.* (2012) designed a NimbleGen array, capture-based re-sequencing experiment targeted to 56.5 Mb of transcripts of eight bread-wheat varieties and were able to identify more than 500 000 putative SNPs, a sample of which were validated using KASPar technology. Saintenac *et al.* (2011) and Bundock *et al.* (2012) used the solution-based hybridization method (Agilent SureSelect, Agilent Technologies, Santa Clara, CA) for SNP discovery in coding sequences among two tetraploid wheat and two sugarcane genotypes, respectively. Lai *et al.* (2012) used the 454 sequencing technology to investigate the transcriptome of three bread-wheat varieties and identified 38 928 putative SNPs. Similarly, You *et al.* (2011) and Iehisa *et al.* (2012) have recently investigated the transcriptome of *Aegilops tauschii* to identify and map SNPs and insertion/deletions.

In addition to the effort required in the initial SNP discovery phase, SNP assay conversion following the discovery phase can result in the loss of many potentially useful markers, especially in wheat (Kaur *et al.*, 2012; Trebbi *et al.*, 2011). The development of high-throughput sequencing technologies has greatly enhanced SNP discovery and can also be used for direct genotyping without assay conversion, for example using restriction site–associated DNA sequencing (RAD-seq), GBS, or sequenced-based genotyping (SBG) methods (Baird *et al.*, 2008; Elshire *et al.*, 2011; Miller *et al.*, 2007; Truong *et al.*, 2012). All these methods combine DNA complexity reduction with high-throughput sequencing and are especially valuable for complex crops such as *Brassica napus*, *Mischantus sinensis*, and wheat (Bus *et al.*, 2012; Ma *et al.*, 2012; Poland *et al.*, 2012; Trebbi *et al.*, 2011). Accordingly, we have applied SBG for SNP discovery and genotyping in durum wheat. Using a robust AFLP-based complexity reduction with a *Pst*I/*Taq*I enzyme combination, we targeted low-repetitive sequences for SNP discovery and genotyping (Fellers, 2008; Trebbi *et al.*, 2011).

Genotyping output

In total, 9983 putative SNPs were detected in 6372 contigs. Four types of SNPs were identified (Table 1). Type-1 (i.e. simple) SNPs are straightforward and do not require further explanation. Type-2 and 3 (i.e. complex) SNPs likely represent a mixture of two homeologous or paralogous sequences, with only one of the homeologs/paralogs containing a polymorphism at the SNP position. As no marker conversion is required for genotyping, these SNPs can be used as regular markers when applying the Euclidean/Bayesian genotyping analysis (see Materials and Methods). Type-4 SNPs likely represent a mixture of more than two homeologs and/or paralogs. Genotypes based on Type 4 SNPs will combine information of more than one segregating homeologs/paralog and thus will not be useful for mapping. Consequently, they were discarded from further analysis and are not included in the 9983 putative SNPs. As can be deduced from Table 1, most of the contigs containing simple SNPs were used for mapping

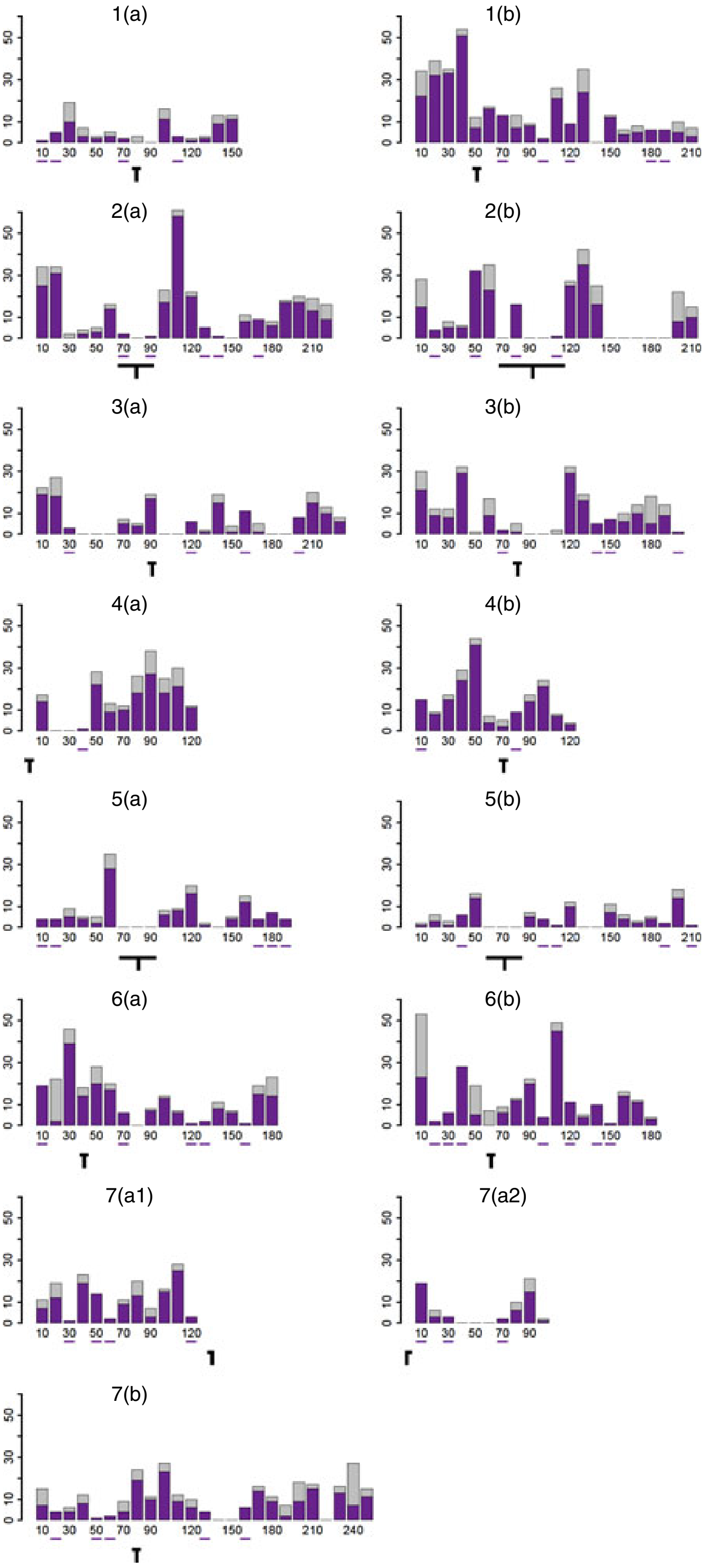

Figure 2 Distribution of markers in 10 cM bins over the 15 linkage groups. Codes of the linkage groups refer to chromosome numbers. Chromosome 7A consists of two linkage groups. SSR, DArT, and CRoPS-derived markers from the pre-existing framework map are shown in grey; sequence-based genotyping (SBG) markers are shown in purple. Black T-bars indicate estimated centromere positions (Trebbi *et al.*, 2011). Purple-underlined bins indicate bins with only SBG markers.

(67%), whereas contigs with complex SNPs more often did not pass our filtering steps (21% used for mapping). This is likely due to the fact that the Euclidean/Bayesian genotyping method requires more sequencing depth for genotyping compared to simple SNP genotyping. Thus, genotypes of complex SNPs are more likely to be classified as ‘unknown’. With the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA), a fivefold increase in sequencing output can easily be obtained without increasing

Table 3 Examples of consensus genotypes for contigs with multiple single nucleotide polymorphisms (SNPs)

SNP	1	2	3	4	5
1	A	A	A	A	M
2	A	U	H	B	M
3	A	U	H	M	M
Consensus	A	A	H	U	M

sequencing costs compared to the GAII platform used here. With higher sequencing output, many of the complex SNPs could likely be used for mapping.

Quality assessment

In the two RILs added in the SNP discovery phase to assess segregation of the SNPs, the following segregation is expected, not considering residual heterozygosity: 25% of the SNPs genotyped as A in both RILs, 25% as B in both RILs, and 50% of the SNPs should be segregating between these two RILs (A for one RIL and B for the other). The observed percentages, that is, 24% genotyped as A in both RIls, 27% as B in both RILs, and 49% segregating show that the putative SNPs are behaving as expected, strengthening the confidence that these putative SNPs are true SNPs. In addition, the observed levels of heterozygosity in the two RILs of 2.9% and 5.1% are well within the expected range of 1.6–6.3% heterozygosity in a RIL $F_{6:8}$ population.

Contigs with multiple SNPs can also be used for quality assessment, as multiple SNPs from the same contig should result in the same genotyping under the assumption that no recombination occurred within the 70–76 nt contigs. Of the 61 610 genotypes available for this assessment, 97.6% showed no inconsistencies in genotyping, demonstrating the high consistency of the genotyping assay. The remaining 2.4% are mainly due to conflicts in certainties (i.e. A and C or B and D conflicts) and only very few are A to B conflicts (Figure S2). These latter can be, for example due to sequencing errors or due to mixed homeologs/paralogs with one SNP representing one homeolog/paralog and another SNP representing the other homeolog/paralog.

A final assessment of genotyping quality is the integration of the SBG markers in the linkage map previously established with SSR and DArT markers. The high percentage of mapped SBG markers (91%) as well as the good integration of the SBG markers with those from the pre-existing framework map (Figure 3, Figure S1) demonstrate the high quality of the SBG markers.

A few linkage maps based on RAD-seq and GBS have been recently reported in crop species. For instance, Bus *et al.* (2012) used the RAD-seq method in *Brassica napus*, a tetraploid species with a complex genome, and more than 20 000 SNPs and 125 insertion/deletions were found among eight different germplasm types. However, Bus *et al.* (2012) limited their investigation to the SNP discovery phase. In *Mischanthus sinensis*, a highly heterozygous diploid species with genome size similar to maize (2.5 Gbp), a composite linkage map has been obtained based on progenies obtained from a 'two-way pseudo-testcross' (Ma *et al.*, 2012). The authors, using 192 progenies, multiplexing 12 samples per lane and sequencing with two Illumina flow cells, were able to map 3745 SNP markers on 19 linkage groups with a 0.64 cM average resolution, a result quite similar to the one herein reported. In grape, the original RAD-seq technique based on a single restriction enzyme has been used by Wang *et al.* (2012) to genotype an F1 population of 100 individuals and to assemble an integrated male–female linkage map of 1646 SNPs. In cotton, a genome reduction experiment with a double digestion, similar to the AFLP concept, has identified 11 834 SNPs between *Gossypium hirsutum* and *G. barbadense* accessions (Byers *et al.*, 2012). In this case, a medium-density linkage map of 367 (267 nongenic and 100 genic) SNPs, genotyped with the Fluidigm system, has been produced. In wheat, a two-enzyme-based GBS experiment combining a rare and a frequent cutter has been performed by Poland *et al.* (2012) for mapping SNPs in the Synthetic x Opata mapping population; specifically designed common reverse *Y*-adapters were used to ensure the amplification of only those fragments bounded by one restriction site from the rare-cutting enzyme on one side and by one restriction site from the frequent-cutting enzyme on the other side. Forty-eight-plex libraries were made for genotyping the lines with the Illumina GAII or the HiSeq2000 sequencers, and a highly consistent linkage map of 1485 SNPs with low percentage of missing data was produced. Additionally, 19 720 SNPs with higher% of missing data were positioned to the framework map using a bin-mapping approach.

Map comparisons

Sequence-based genotyping markers were found on all durum wheat chromosomes (Figure 2). The distribution of SBG markers over linkage groups was moderately correlated with linkage group size in cM ($r^2 = 0.3$) and similarly compared to the markers from the pre-existing framework map ($r^2 = 0.6$). In the pre-existing framework map, chromosomes 2A, 2B, and 3A were represented by two linkage groups each. With the aid of the SBG markers, these linkage groups could be merged to one linkage group per chromosome. Only chromosome 7A remained divided in two linkage groups. Looking at the binned marker distribution (Figure 2), 63 of the 268 10 cM bins contained only SBG markers, showing that the SBG markers are filling in gaps in the framework map. In total, 38 bins remained without any markers. Some of these bins may represent highly methylated parts of the chromosomes, where there are few restriction-sites of the methyl-sensitive *Pst*I enzyme used in SBG, CRoPS-derived as well as DArT markers. Indeed, 15 of these unrepresented bins are at or near the estimated centromere positions, known to be hypermethylated. Alternatively, some of these bins may represent segments that are identical by descent (IBD), with no polymorphisms present.

Conclusions

We report the integration of 2365 SBG markers into a pre-existing framework RIL map in durum wheat. By using SBG, we have more than quadrupled the number of markers in this map, which was based on SSRs, DArT markers, and CRoPS SNPs. Compared to methods that use one platform for SNP discovery and another for genotyping, the omission of an assay conversion both reduces costs and increases efficiency. For example, Trebbi *et al.* (2011) reported a validation rate of only 36% after assay conversion, likely due to the repetitive and polyploid nature of the durum wheat genome. Thus, the omission of an assay conversion is especially valuable in complex genomes with much repetitive content and/or increased ploidy levels. Based on the results herein presented, SBG can be considered a powerful tool for high-throughput marker discovery and genotyping, also in complex,

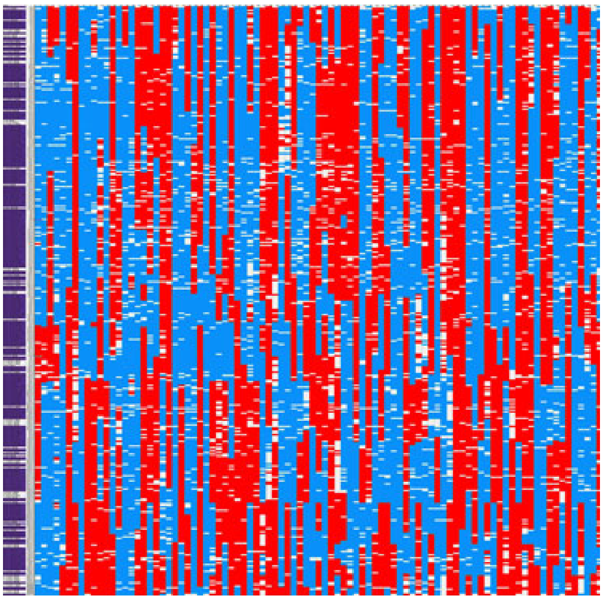

Figure 3 Genotyping the recombinant inbred lines (RIL) population for chromosome I_B. Markers are in rows; RILs in columns; CLS-derived loci in red; LLD-derived loci in blue. Marker codes in purple background indicate added sequence-based genotyping markers.

polyploid crops. The findings herein presented corroborate the most recent reports on the validity of SBG methods while providing 2365 novel SNPs in durum wheat, a species notoriously characterized by a low level of polymorphism.

Experimental procedure

DNA samples

DNA samples were extracted from the North American elite durum wheat cv. Lloyd (LLD), the Italian elite cv. Colosseo, as well as from 92 $F_{6\,:\,8}$ recombinant inbred lines (RILs) derived from a cross between these parents (Mantovani *et al.*, 2008). DNA was extracted following the method described by Maccaferri *et al.* (2005).

Library preparation

Libraries were prepared for Illumina single-end sequencing as described previously (Truong *et al.*, 2012), adapted for the restriction enzymes described below and using a + 2 selective touch-down amplification. In short, 250 ng total genomic DNA was digested using 5 units *Taq*I (1 h at 65 °C) followed by 5 units *Pst*I (1 h at 37 °C). Adapter ligation was performed using a universal P7 *Taq*I adapter (top oligo: 5′-CAAGCAGAAGACGGCA-TACGAG-3′; bottom oligo: 5′-CGCTCGTATGCCGTCTTCTGCT-TG-3′) and a sample-specific tagged *Pst*I adapter (top oligo: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxATGCA-3′; bottom oligo: 5′-TxxxxxAGATCGGAAGAGCGTCGT-3′-NH_2; xxxx = sample identification tag). PCR amplification was performed on 5 μL of a 10-fold diluted restriction-ligation mixture using 5 ng Illumina P5 primer (5′-AATGATACGGCGACCACCG-3′) and 30 ng P7 *Taq*I + CT primer (5′-CAAGCAGAAGACGGCATACGAGCGACT-3′) in 20 μL total volume. A touch-down cycle profile was used: 2 min 72 °C; followed by 13 cycles of 30 s 94 °C, 2 min 67–0.7 °C/cycle, 2 min 72 °C; followed by 37 cycles of 30 s 94 °C, 2 min 58 °C, 2 min 72 °C. For each sequencing library, equal volumes of amplification mixture were pooled and purified using the MinElute PCR Purification Kit (Qiagen, Hilden, Germany), thus producing eight sequencing libraries. One library contained four samples: the two parents as well as two randomly selected RILs. This library was used for SNP discovery; the other seven libraries each contained 13–14 samples, covering in total 92 RILs and two replicates of each of the parents, these libraries were used for genotyping. Notably, the two randomly selected RILs in the SNP discovery library were also present in the genotyping libraries. Single-end sequencing (76 nt) was performed using the Illumina Genome Analyzer II, one library per lane.

Read processing

Illumina reads were filtered based on presence of sample identification tags and the *Pst*I restriction site motif: reads without sample identification tags or *Pst*I restriction site motif were removed. From the remaining reads, sample identification tags were removed. These reads were clustered based on 100% sequence similarity in order to produce a condensed, nonredundant data set.

SNP discovery

The nonredundant data from the SNP discovery library were used as input for CAP3 clustering (Huang and Madan, 1999). The resulting contigs were filtered on restriction site motif and contig length: only contigs containing the restriction site motif and with a length between 70 and 76 nucleotides were used as a reference for SNP discovery. The same data were subsequently used for mapping against the reference using BWA (Li and Durbin, 2009). Only sequences with a mapping score of at least 20 were used for SNP discovery. Identified polymorphisms were filtered based on average base quality and allele coverage, with each allele sequenced at least seven times in one of the two parents and not found in the other parent. Note that for coverage calculations, the condensation step based on 100% similarity was taken into account. SNPs at positions with more than two alleles per parent were discarded. SNP calling also included single-nucleotide insertions/deletions.

Genotyping

The nonredundant sequences from the genotyping libraries were mapped against the reference sequence using BWA. Only sequences with a mapping score of at least 20 were used for genotyping. Genotyping was performed only at the polymorphic positions identified in the SNP discovery phase. Per sample per position, genotypes (A = LLD; B = CLS, H = heterozygous; C = A or H; D = B or H; U = unknown; M = missing) were determined based on Bayesian theory and a Euclidean distance calculation. Euclidean distance was used to calculate the similarity between each sample and a given genotype (A, B, and H), based on the proportion of reads per allele as well as base quality. These distances were used as input for Bayesian calculation of probabilities. The prior probabilities for genotypes A, B, and H used for Bayesian probability calculation were estimated based on the theoretical probabilities for an F6 RIL population, that is, A : B : H = 0.4995 : 0.4995 : 0.001. If the most likely genotype had a probability that was at least five times that of the second-most likely genotype, the genotype was set to the most likely genotype. If not, there are three possibilities: if the most likely and second-most likely genotypes were A and B or *vice versa*, the genotype was set to U; if they were A and H or *vice versa*, the genotype was set to C; if they were B and H or *vice versa*, the genotype was set to D. After this first round of genotyping, all SNPs

© 2013 Society for Experimental Biology, Association of Applied Biologists and John Wiley & Sons Ltd,

for which more than 90% of the RILs were genotyped were used to assess the actual A : B : H probabilities. These new probabilities were used as priors for a second round of Bayesian genotyping. These final genotyping results were used for further selection based on several criteria: (i) if available, the genotypes of the parental duplicates genotyped at low coverage should match the genotypes of the parents genotyped as high coverage; (ii) many contigs contain several SNPs. As multiple SNPs on a contig of 70–76 nt provide redundant information for mapping, the genotyping information was condensed to one marker per contig. If multiple SNPs on the same contig gave exactly the same genotyping results for all RILs, the data set was reduced by taking the genotyping information of only one of the SNPs. If for some RILs, multiple SNPs from the same contig gave conflicting information, a consensus genotype was extracted using the following rules (see Table 3 for examples): A, B, H, C, or D genotypes are always preferred over unknown (U) or missing (M) genotypes. If there is a conflict between A, B, or H genotypes, the consensus genotype is the predominant one. If no predominant genotype can be called, the genotype is set to unknown (U); (iii) all H, C, and D as well as missing genotypes were set to U. For H, C, and D genotypes, this was done as heterogeneous scores were not recorded for the initial SSR and DART markers that were used in the anchor map (see below). Only SNPs for which 70% of the RILs were genotyped (i.e. not missing or unknown) were used; (iv) for a given SNP, A : B ratios over the population had to be in between 1 : 3 and 3 : 1.

Mapping

Mapping has been carried out with JoinMap 4.0 (van Ooijen, 2006) using the SSR and DArT marker-based map previously assembled with the complete RIL population of 176 lines as a framework of markers of fixed mapping position. Grouping of the newly developed SNPs was performed using the log10 of the LOD method (incremental LOD thresholds from 2 to 10 with LOD 1.0 steps) by selecting the grouping nodes with a stable (nonvariant) number of markers in the LOD range between 6 and 10. The newly defined linkage groups, including the framework markers and the new SNPs, were used to calculate the corresponding maps based on the maximum-likelihood (ML) mapping algorithm, by assuming the framework marker order as *fixed*. The mapping process was based on the repeated rounds of (i) simulated annealing Monte Carlo map order optimization search, followed by (ii) Gibbs sampling (Monte Carlo Expectation Maximization algorithm) that is used to obtain the multipoint recombination frequency estimates. For map building, the number of map optimization rounds was set equal to five. The linkage group maps were gradually constructed by considering spatial samples of loci using the five default recombination frequency spatial sampling thresholds (0.10, 0.05, 0.03, 0.02, and 0.01). In the simulated annealing marker-ordering phase, a chain of 5000 trials and error steps with constant acceptance probability was used, and the system-stop was set after reaching 1000 chains without further improvement; other parameters were set as default. In the Gibbs sampling recombination frequency estimation phase, the length of the burn-in chain was set equal to 5000 iterations and the chain length per Monte Carlo Expectation Maximization cycle was set to 1000 iterations; other parameters were as default.

Mapping was carried out in two phases, with some manual curation performed after each phase: In the first phase, mapping integration was carried out using SNPs for which at least 90% of the RILs were genotyped. The marker order in the resulting integrated map was considered as fixed for the second phase, integrating all the considered SNP data (i.e. at least 70% of the RILs genotyped). After curation, a final mapping round was performed to refine the map distances among loci.

Description of additional data files

The following additional data are available with the online version of this paper. Figure S1 shows the genetic map of the CLS × LLD RIL population with the integrated SBG markers. Figure S2 shows the distribution of genotyping conflicts in contigs with multiple SNPs over different conflict types. Table S1 gives an overview of marker types in the genetic map of the CLS × LLD RIL population. Table S2 gives the features and sequence information of the 2365 novel SNP mapped in the CLS × LLD RIL population.

Acknowledgements

The AFLP®, CRoPS®, and SBG technologies are covered by patents and patent applications owned by Keygene N.V. AFLP and CRoPS are registered trademarks of Keygene N.V. Other trademarks are the property of their respective owners. The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2002-2006) under grant agreement n° FOOD-CT-2005-513959 BIOEXPLOIT.

References

Akbari, M., Wenzl, P., Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., Hayden, M.J., Howes, N., Sharp, P., Vaughan, P., Rathmell, B., Huttner, E. and Kilian, A. (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor. Appl. Genet.* **113**, 1409–1420.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Bernardo, A.N., Bradbury, P.J., Ma, H., Hu, S., Bowden, R.L., Buckler, E.S. and Bai, G. (2009) Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. *BMC Genomics*, **10**, 251.

Blanco, A., Bellomo, M.P., Cenci, A., De Giovanni, C., D'Ovidio, R., Iacono, E., Laddomada, B., Pagnotta, M.A., Porceddu, E., Sciancalepore, A., Simeone, R. and Tanzarella, O.A. (1998) A genetic linkage map of durum wheat. *Theor. Appl. Genet.* **97**, 721–728.

Bundock, P.C., Casu, R.E. and Henry, R.J. (2012) Enrichment of genomic DNA for polymorphism detection in a nonmodel highly polyploidy crop plant. *Plant Biotechnol. J.* **10**, 657–667.

Bus, A., Hecht, J., Huettel, B., Reinhardt, R. and Stich, B. (2012) High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics*, **13**, 281.

Byers, R.L., Harker, D.B., Yourstone, S.M., Maughan, P.J. and Udall, J.A. (2012) Development and mapping of SNP assay in allotetraploid cotton. *Theor. Appl. Genet.* **124**, 1201–1214.

Collard, B.C.Y. and Mackill, D.J. (2008) Marker-assisted selection, an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 557–572.

Cuppen, E. (2007) Genotyping by allele-specific amplification (KASPar). *CSH Protoc.* **9**, doi: 10.1101/pdb.prot4841.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Fan, J.-B., Chee, M.S. and Gunderson, K.L. (2006) Highly parallel genomic assays. *Nat. Rev. Genet.* **7**, 632–644.

Fellers, J.P. (2008) Genome filtering using methylation- sensitive restriction enzymes with six base pair recognition sites. *Plant Genome*, **1**, 146.

Fleury, D., Jefferies, S., Kuchel, H. and Langridge, P. (2010) Genetic and genomic tools to improve drought tolerance in wheat. *J. Exp. Bot.* **61**, 3211–3222.

Ganal, M.W., Altmann, T. and Röder, M.S. (2009) SNP identification in crop plants. *Curr. Opin. Plant Biol.* **12**, 211–217.

Gupta, P.K., Mir, R.R., Mohan, A. and Kumar, J. (2008) Wheat genomics, present status and future prospects. *Int. J. Plant Genomics*, **2008**, 1–36.

Gupta, P.K., Varshney, R.K., Sharma, P.C. and Ramesh, B. (1999) Molecular markers and their applications in wheat breeding. *Plant Breed.* **118**, 369–390.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glémin, S. and David, J. (2007) Grinding up wheat, a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506–1517.

Heffner, E.L., Sorrells, M.E. and Jannink, J.L. (2009) Genomic selection for crop improvement. *Crop Sci.* **49**, 1–12.

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

Iehisa, J.C., Shimizu, A., Sato, K., Nasuda, S. and 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.

Kaur, S., Francki, M.G. and Forster, J.W. (2012) Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant Biotechnol. J.* **10**, 125–138.

Koebner, R.M.D. and Summers, R.W. (2003) 21st century wheat breeding, plot selection or plate detection? *Trends Biotechnol.* **21**, 59–63.

Korzun, V., Röder, M.S., Wendehake, K., Pasqualone, A., Lotti, C., Ganal, M.W. and Blanco, A. (1999) Integration of dinucleotide microsatellites from hexaploid bread wheat into a genetic linkage map of durum wheat. *Theor. Appl. Genet.* **98**, 1202–1207.

Lai, K., Duran, C., Berkman, P.J., Lorenc, M.T., Stiller, J., Manoli, S., Hayden, M.J., Forrest, K.L., Fleury, D., Baumann, U., Zander, M., Mason, A.S., Batley, J. and Edwards, D. (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol. J.* **10**, 743–749.

Lewis, S.N., Nsoesie, E., Weeks, C., Qiao, D. and Zhang, L. (2011) Prediction of disease and phenotype associations from genome-wide association studies. *PLoS One*, **6**, e27175.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Lotti, C., Salvi, S., Pasqualone, A., Tuberosa, R. and Blanco, A. (2000) Integration of AFLP markers into an RFLP-based map of durum wheat. *Plant Breed.* **119**, 393–401.

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Jones, S.T., Farrar, K., Clifton-Brown, J., Donnison, I., Swaller, T. and Flavell, R. (2012) High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Mischantus sinensis*. *PLoS One*, **7**, e33821.

Maccaferri, M., Sanguineti, M.C., Demontis, A., El-Ahmed, A., Garcia del Moral, L., Maalouf, F., Nachit, M., Nserallah, N., Ouabbou, H., Rhouma, S., Royo, C., Villegas, D. and Tuberosa, R. (2011) Association mapping in durum wheat grown across a broad range of water regimes. *J. Exp. Bot.* **62**, 409–438.

Maccaferri, M., Sanguineti, M.C., Noli, E. and Tuberosa, R. (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Mol. Breed.* **15**, 271–290.

Mantovani, P., Maccaferri, M., Sanguineti, M.C., Tuberosa, R., Catizone, I., Wenzl, P., Thomson, B., Carling, J., Huttner, E., DeAmbrogio, E. and Kilian, A. (2008) An integrated DArT-SSR linkage map of durum wheat. *Mol. Breed.* **22**, 629–648.

Masojć, P. (2002) The application of molecular markers in the process of selection. *Cell. Mol. Biol. Lett.* **7**, 499–509.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A. and Johnson, E.A. (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248.

Morrell, P.L., Buckler, E.S. and Ross-Ibarra, J. (2011) Crop genomics, advances and applications. *Nat. Rev. Genet.* **13**, 85–96.

van Ooijen, J. (2006) *JoinMap 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen, Netherlands: Kyazma BV.

van Orsouw, N.J., Hogers, R.C.J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H. and van Eijk, M.J.T. (2007) Complexity reduction of polymorphic sequences (CRoPS), a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033.

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.

Röder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. (1998) A microsatellite map of wheat. *Genetics*, **149**, 2007–2023.

Saintenac, C., Jiang, D. and Akhunov, E.D. (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* **12**, R88.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome*, **46**, 431–437.

Tester, M. and Langridge, P. (2010) Breeding technologies to increase crop production in a changing world. *Science*, **327**, 818–822.

Trebbi, D., Maccaferri, M., de Heer, P., Sørensen, A., Giuliani, S., Salvi, S., Sanguineti, M.C., Massi, A., van der Vossen, E.A.G. and Tuberosa, R. (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.). *Theor. Appl. Genet.* **123**, 555–569.

Truong, H.T., Ramos, A.M., Yalcin, F., de Ruiter, M., van der Poel, H.J.A., Huvenaars, K.H.J., Hogers, R.C.J., van Enckevort, L.J.G., Janssen, A., van Orsouw, N.J. and van Eijk, M.J.T. (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One*, **7**, e37565.

Tuberosa, R., Graner, A. and Varshney, R.K. (2011) Genomics of plant genetic resources, an introduction. *Plant Genet. Res.* **9**, 151–154.

Wang, N., Fang, L., Xin, H., Wang, L. and Li, S. (2012) Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing. *BMC Plant Biol*. **12**, 148.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. and Rothberg, J.M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L.A., Coghill, J.A., Burridge, A., Hall, A., Brechley, R.C., D'Amore, R., Hall, N., Bevan, M.W., Richmond, T., Gerhrdt, D.J., Jeddeloh, J.A. and Edwards, K. (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J. and Wang, W. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111.

You, F.M., Huo, N., Deal, K.R., Gu, Y.Q., Luo, M.C., McGuire, P.E., Dvorak, J. and Anderson, O.D. (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.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Genetic map of the Colosseo × Lloyd RIL population.
Figure S2 Distribution of genotyping conflicts in contigs with multiple single nucleotide polymorphisms (SNPs).