Plant Biotechnology Journal

Plant Biotechnology Journal (2009) **7**, pp. 364–374 doi: 10.1111/j.1467-7652.2009.00404.x

High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.)

Aurélie Bérard[1], Marie Christine Le Paslier[1], Mireille Dardevet[2], Florence Exbrayat-Vinson[2], Isabelle Bonnin[3], Alberto Cenci[4], Annabelle Haudry[4], Dominique Brunel[1] and Catherine Ravel[2,]*

[1]*INRA, UR1279 Etude du Polymorphisme des Génomes Végétaux, CEA-IG/Centre National de Génotypage, 2 rue Gaston Crémieux, CP5724, F-91057 Evry, France*
[2]*INRA, UMR1095 Génétique Diversité Ecophysiologie des Céréales, 234 avenue du Brézet, F-63100 Clermont-Ferrand, France*
[3]*UMR de Génétique Végétale, INRA/CNRS/UPS/INA-PG, Ferme du Moulon, F-91190 Gif/Yvette, France*
[4]*UMR Diversité et Génomes des Plantes Cultivées, INRA Domaine de Melgueil, F-34130 Mauguio, France.*

Received 26 September 2008;
accepted 24 December 2008.

**Correspondence* (fax 0033 47362 4453;
e-mail ravel@clermont.inra.fr)

Keywords: high-throughput genotyping, polyploidy, single nucleotide polymorphism (SNP), SNPlex™, wheat.

Summary

Over the past few years, considerable progress has been made in high-throughput single nucleotide polymorphism (SNP) genotyping technologies, largely through the investment of the human genetics community. These technologies are well adapted to diploid species. For plant breeding purposes, it is important to determine whether these genotyping methods are adapted to polyploidy, as most major crops are former or recent polyploids. To address this problem, we tested the capacity of the multiplex technology SNPlex™ with a set of 47 wheat SNPs to genotype DNAs of 1314 lines that were organized in four 384-well plates. These lines represented different taxa of tetra- and hexaploid *Triticum* species and their wild diploid relatives. We observed 40 markers which gave less than 20% missing data. Different methods, based on either Sanger sequencing or the MassARRAY® genotyping technology, were then used to validate the genotypes obtained by SNPlex™ for 11 markers. The concordance of the genotypes obtained by SNPlex™ with the results obtained by the different validation methods was 96%, except for one discarded marker. Furthermore, a mapping study on six markers showed the expected genetic positions previously described. To conclude, this study showed that high-throughput genotyping technologies developed for diploid species can be used successfully in polyploids, although there is a need for manual reading. For the first time in wheat species, a core of 39 SNPs is available that can serve as the basis for the development of a complete SNPlex™ set of 48 markers.

Introduction

Single nucleotide polymorphisms (SNPs) are the most frequent type of polymorphism in genomes. They can provide a huge number of useful markers for many genetic analyses (e.g. phylogenetic analysis, ultra-dense genetic mapping, genotype/phenotype association studies) and important applications (e.g. cultivar identification, marker-assisted selection), which are simplified as these markers are most often biallelic. In the last decade, the constitution of very large SNP collections for humans (http://www.hapmap.org), several animal species (http://www.livestockgenomics.csiro.au/ibiss) and even model plants (for *Arabidopsis* and rice, see http://www.arabidopsis.org and http://irfgc.irri.org respectively) has made possible the development of genome-wide analyses which require the genotyping of a large number of SNPs. As such genotyping methods are limited by the cost and time for scoring SNPs, there has been an increasing demand for the development of high-throughput and low-cost genotyping methods.

Automation and multiplexing enhance the effectiveness of SNP genotyping enormously. Several high-throughput platforms are now available for the genotyping of a variable number of genomic DNA (gDNA) samples for one to up to one million SNPs in parallel. These technologies were first developed for human genomic studies i.e. a diploid genome. They include, but are not limited to, TaqMan® and SNPlex™

technologies from Applied Biosystems (Foster City, CA, USA) and array-based technologies from Illumina (San Diego, CA, USA) (GoldenGate® and Infinium®) or Affymetrix (Santa Clara, CA, USA) (for a review, see Sobrino *et al*., 2005; Syvänen, 2005; Khlestkina and Salina, 2006; for a comparison, see De La Vega *et al*., 2005; Giancola *et al*., 2006).

Today, except for model plant species (*Arabidopsis* and rice) and a small number of other species (e.g. maize, barley, pine, grapevine), the number of SNPs available is still low. In wheat, SNP databases are available (Gupta *et al*., 2008), but with a limited number of SNPs for hexaploid material. This will certainly change as many SNP discovery projects in plants reach completion. Nevertheless, the utilization of SNPs in breeding programmes, such as marker-assisted selection, depends on high-throughput SNP genotyping methods. Recently, some of these methods have been applied in plants. For instance, SNPlex™ technology has been used in *Arabidopsis* (Drouaut *et al*., 2007; Simon *et al*., 2008) and grapevine (*Vitis vinifera*) (Lijavetzky *et al*., 2007; Pindo *et al*., 2008), and GoldenGate® technology has been used in spruce (*Picea glauca* and *Picea mariana*) (Pavy *et al*., 2008) and soybean (*Glycine max*) (Hyten *et al*., 2008). These current technologies have been applied to diploid species. However, many plant species are former or recent polyploids, and nearly all crop plants are polyploid (for a review, see Adams and Wendel, 2005), and therefore their genotyping represents an extra challenge.

Bread wheat (*Triticum aestivum* L.) is a major cereal crop in both human and animal nutrition. Its large genome, consisting of three highly related genomes (homoeologous A, B and D genomes), originated from two independent polyploidization events (for a review, see Dubcovsky and Dvorak, 2007). The first event involved the hybridization of two diploid progenitors, an ancestor of *Triticum urartu* (AA genome) and an unconfirmed species related to *Aegilops speltoides* (BB genome), which resulted in wild and cultivated allotetraploid wheats (*T. turgidum* ssp.). The second hybridization between an ancestor of the diploid *Aegilops tauschii* var. *strangulata* (DD genome) and an allotetraploid wheat formed the hexaploid. Few studies have been carried out on nucleotide diversity in wheat because of the presence of two or three homoeologous genome copies. Cultivated wheat species are reported to have a low level of nucleotide diversity, explained by their evolutionary history, including several demographic bottlenecks and selective events (Ravel *et al*., 2006; Haudry *et al*., 2007). Therefore, to date, SNP discovery in these species has been an arduous task.

SNPlex™ technology (De La Vega *et al*., 2005) allows the investigation of one to several hundred samples with sets of 48 SNP markers, and is a good compromise between very-high-throughput genotyping methods, such as array-based technologies (large number of SNPs but a limited number of samples), and simplex technologies (for example, the TaqMan® assay allows the analysis of one SNP at a time in thousands of samples). Therefore, the purpose of this work was to study whether a multiplex genotyping method could be used in polyploid wheat species. To address this problem, we simultaneously genotyped 47 SNPs by SNPlex™ using gDNAs from different taxa of wheat (tetra- and hexaploids) and their wild diploid relatives.

Results

Assay design

We submitted 75 SNPs to Applied Biosystems (Table 1). Among these, 73 passed the design rules and were included in two sets of compatible markers for the multiplex reaction: one with 47 markers and one with 26 markers (sets 1 and 2 in Table 1). We considered that the second set had an insufficient number of markers, and so only used the complete set (set 1).

Data

In this study, wheat gDNAs with different ploidy levels were used (see Experimental procedures). Four plates were constituted and named Di-Tetra, CC1, CC2 and RILS-ReR for diploids and tetraploids, hexaploid core collection 1, hexaploid core collection 2 and hexaploid wheat recombinant inbred lines (RILs), respectively (Table 2).

Allele calling was performed by an automatic analysis of SNPlex™ signals with the software GeneMapper v3.7, followed by a manual reading. Samples giving signals which could not be discriminated from the negative control (water) were treated as missing data. We observed more than 20% missing data in all plates for all four markers developed in a hypothetical gene on the 5A chromosome (HG_A_Y_335, HG_A_M_489, HG_A_K_615 and HG_A_R_695), as well as for the marker PSY_B_M_640. No data were obtained for GSP_A_R_366 and Hd1_A_K_1545 (Tables 1 and 3). Thus, 40 of 47 markers gave less than 20% missing data for all plates and, henceforth, these are referred to as selected markers.

For all markers, the average number of missing data points (calculated as the sum of the number of missing data for each marker/total number of markers, i.e. 47) was higher in CC1 (67.7) than in CC2 (48.1). The correlation between the number of missing data points for each marker in these two

Table 1 Description of submitted single nucleotide polymorphisms (SNPs)

Code of submitted SNP*	Gene code (chromosome location)	Full gene name	Reference†	Set no
SPA_B_S_142	SPA (1B)	Storage protein activator	Ravel *et al*. (2006)	1
AAP_A_R_99	*AAP* (2A)	Amino acid permease	Haudry *et al*. (2007)	1
AAP_B_Y_335	*AAP* (2B)		Haudry *et al*. (2007)	2
AAP_B_Y_929	*AAP* (2B)		Haudry *et al*. (2007)	1
AAP_D_Y_450	*AAP* (2D)		Balfourier *et al*. (2006)	1
ZDS_A_K_745	ZDS (2A)	Lycopene synthase	Haudry *et al*. (2007)	1
ZDS_A_M_1312	ZDS (2A)		Haudry *et al*. (2007)	2
ZDS_B_Y_651	ZDS (2B)		Haudry *et al*. (2007)	1
GIb_A_S_197	*GI* (3A)	Gigantea	I. Bonnin (unpubl. data)	1
GId_A_R_64	*GI* (3A)		I. Bonnin (unpubl. data)	1
LDD_A_Y_317	*LD* (3A)	Lumini dependens	I. Bonnin (unpubl. data)	1
LDD_A_R_660	*LD* (3A)		I. Bonnin (unpubl. data)	2
LDD_A_Y_476	*LD* (3A)		I. Bonnin (unpubl. data)	1
LDD_B_Y_531	*LD* (3B)		I. Bonnin (unpubl. data)	1
LDD_B_M_924	*LD* (3B)		I. Bonnin (unpubl. data)	1
LDD_B_R_1071	*LD* (3B)		I. Bonnin (unpubl. data)	1
GDH_A1_Y_67	*GDH* (5A)	Glutamate dehydrogenase	Haudry *et al*. (2007)	1
GDH_A1_W_315	*GDH* (5A)		Haudry *et al*. (2007)	1
GDH_A1_R_891	*GDH* (5A)		Haudry *et al*. (2007)	1
GDH_A1_Y_970	*GDH* (5A)		Haudry *et al*. (2007)	1
GDH_A1_1326	*GDH* (5A)		Haudry *et al*. (2007)	2
GDH_B9_R_965	*GDH* (5B)		Balfourier *et al*. (2006)	2
GDH_B9_S_133	*GDH* (5B)		Balfourier *et al*. (2006)	1
GDH_B9_indel_1178	*GDH* (5B)		Balfourier *et al*. (2006)	Fail
GDH_B5_W_259	*GDH* (5B)		Balfourier *et al*. (2006)	1
GDH_A8_Y_576	*GDH* (5B)		Balfourier *et al*. (2006)	2
GSP_A_R_176	*GSP* (5A)	Grain softness protein	Haudry *et al*. (2007)	1
GSP_A_R_345	*GSP* (5A)		Haudry *et al*. (2007)	1
GSP_A_R_366	*GSP* (5A)		Haudry *et al*. (2007)	1
GSP_A_R_627	*GSP* (5A)		Haudry *et al*. (2007)	2
GSP_A_Y_789	*GSP* (5A)		Haudry *et al*. (2007)	1
GSP_A_R_689	*GSP* (5A)		Haudry *et al*. (2007)	Fail
GSP_A_Y_716	*GSP* (5A)		Haudry *et al*. (2007)	1
GSP_B_M_278	*GSP* (5B)		Haudry *et al*. (2007)	2
GSP_B_R_358	*GSP* (5B)		Haudry *et al*. (2007)	2
GSP_B_Y_395	*GSP* (5B)		Haudry *et al*. (2007)	1
GSP_B_Y_669	*GSP* (5B)		Haudry *et al*. (2007)	2
GSP_B_R_729	*GSP* (5B)		Haudry *et al*. (2007)	2
Hipl_A_Y_284	*Hipl* (5A)	Hedgehog interacting protein	Haudry *et al*. (2007)	2
Hipl_A_S_347	*Hipl* (5A)		Haudry *et al*. (2007)	1
Hipl_A_Y_410	*Hipl* (5A)		Haudry *et al*. (2007)	1
Hipl_A_R_530	*Hipl* (5A)		Haudry *et al*. (2007)	2
CHS_A_S_133	Chs (5A)	Chalcone synthase	Haudry *et al*. (2007)	1
CHS_A_Y_300	Chs (5A)		Haudry *et al*. (2007)	1
HG_A_Y_52	*HG* (5A)	Hypothetical gene	Haudry *et al*. (2007)	2
HG_A_Y_335	*HG* (5A)		Haudry *et al*. (2007)	1
HG_A_M_489	*HG* (5A)		Haudry *et al*. (2007)	1
HG_A_K_615	*HG* (5A)		Haudry *et al*. (2007)	1
HG_A_R_695	*HG* (5A)		Haudry *et al*. (2007)	1
PSY_A_Y_198	*Psy-2* (5A)	Phytoene synthase 2	Haudry *et al*. (2007)	1
PSY_A_Y_447	*Psy-2* (5A)		Haudry *et al*. (2007)	2
PSY_A_ Y_631	*Psy-2* (5A)		Haudry *et al*. (2007)	2
PSY_A_R_830	*Psy-2* (5A)		Haudry *et al*. (2007)	1
PSY_B_M_640	*Psy-2* (5B)		Haudry *et al*. (2007)	1
PSY_B_K_873	*Psy-2* (5B)		Haudry *et al*. (2007)	1
PSY_D_Y_66	*Psy-2* (5D)		C. Ravel (unpubl. data)	2

Table 1 *Continued*

SPA2_A_Y_1292	*SPA2* (5A)	SPA heterodimerizing protein	C. Ravel (unpubl. data)	1
SPA2_A_S_1361	*SPA2* (5A)		C. Ravel (unpubl. data)	1
SPA2_A_W_1630	*SPA2* (5A)		C. Ravel (unpubl. data)	2
SPA2_B_R_126	*SPA2* (5B)		C. Ravel (unpubl. data)	2
SPA2_B_Y_453	*SPA2* (5B)		C. Ravel (unpubl. data)	1
SPA2_B_R_623	*SPA2* (5B)		C. Ravel (unpubl. data)	2
SPA2_B_S_1540	*SPA2* (5B)		C. Ravel (unpubl. data)	2
Hd1_A_K_1545	*Hd3a* (7A)	Flowering time locus T (*Hd3a*)	Bonnin *et al.* (2008)	1
Hd3a_A_Y_391	*Hd3a* (7A)		Bonnin *et al.* (2008)	2
Hd1_B_INDEL_382	*Hd3a* (7B)		Bonnin *et al.* (2008)	2
Sal1_A_S_149	*Sal1* (7A)	Supernumerary aleurone layer 1	Balfourier *et al.* (2006)	1
Sal1_B_S_140	*Sal1* (7B)		Balfourier *et al.* (2006)	2
Sal1_B_Y_167	*Sal1* (7B)		Balfourier *et al.* (2006)	2
Sal1_B_Y_302	*Sal1* (7B)		Balfourier *et al.* (2006)	1
Sal1_B_S_416	*Sal1* (7B)		Balfourier *et al.* (2006)	1
Sal1_B_Y_452	*Sal1* (7B)		Balfourier *et al.* (2006)	1
Sal1_B_S_476	*Sal1* (7B)		Balfourier *et al.* (2006)	2
Sal1_B_M_509	*Sal1* (7B)		Balfourier *et al.* (2006)	1
Sal1_D_M_445	*Sal1* (7D)		Balfourier *et al.* (2006)	1

*The SNP code corresponds to the code of the gene followed by the homoeologous genome, the type of SNP using the IUB code and its position in the consensus sequence that we obtained. For GDH genes, the genome is followed by a number indicating the location of the fragment (fragments A1 and B9, from exon 1 to exon 4; fragments A8 and B5, from exon 4 to exon 6).

†The flanking sequences of unpublished SNPs are given in Table S1 (see Supporting information).

Table 2 Number of accessions per wheat taxon genotyped by SNPlex™ technology

Plate name	Species	Ploidy	Number
Di-Tetra	*Triticum urartu*	Diploid	2
	Aegilops speltoïdes	Diploid	16
	Ae. tauschii	Diploid	2
	Other*	Diploid	30
	T. turgidum spp. *dicoccoides*	Tetraploid	62
	T. turgidum spp. *dicoccum*	Tetraploid	97
	T. turgidum spp. *durum*	Tetraploid	71
	T. turgidum spp. *carthlicum*	Tetraploid	12
	T. turgidum spp. polonicum	Tetraploid	31
	T. turgidum spp. *timopheevi*	Tetraploid	21
	T. turgidum (*)	Tetraploid	30
	Total diploids		50
	Total tetraploids		324
CC1, CC2	*T. aestivum*	Hexaploid	744
RILs-ReR	*T. aestivum*	Hexaploid	196
	Total hexaploids		940
	Total		1314

*Indeterminate diploid species or *T. turgidum* subspecies.

plates was high (0.93). CC1 contained more DNAs that were unable to give analysable signals than did CC2. Twenty-one lines were detected which gave no signal regardless of the marker in CC1. The technical repeat gave similar results. This suggests that the DNA in CC1 is of low quality, which may be critical for a few markers (AAP_A_R_99, GSP_A_Y_716, PSY_A_R_830 and SPA2_A_Y_1292) (Table 3). The correlations between CC1 or CC2 and the Di-Tetra plate were lower (0.85 and 0.79, respectively). This lower correlation was probably a result of the fact that several wheat species were included in the Di-Tetra plate.

For the 40 selected markers, the success rates were 92.1%, 91.1% and 98.1% for plates containing diploids and tetraploids, CC1 and CC2, respectively (Table 4).

Validation

No discordance was detected between independent technical repeats, i.e. the same genotype was observed in each analysis.

Three independent methods (M1, M2, M3) were used to validate the SNPlex™ data for the CC1 plate (see Experimental procedures). M1 corresponded to the Sanger sequencing of

Table 3 Percentage of missing data. Validation markers are indicated in bold and discarded markers are indicated in italic

Marker	Diploids and tetraploids	CC1*	CC2†
SPA_B_S_142[c]	15.5	7.8	2.9
AAP_A_R_99	40.4	68.4	6.2
AAP_B_Y_929	1.9	5.6	0.5
AAP_D_Y_450	1.1	5.9	0.5
ZDS_A_K_745	2.1	5.9	0.3
ZDS_B_Y_651	0.3	0.3	0.5
Glb_A_S_197	1.6	6.2	0.5
Gld_A_R_64	8.8	6.4	0.5
LDD_A_Y_317	10.2	9.1	0.5
LDD_A_Y_476	7.5	2.7	0.3
LDD_B_Y_531	6.4	5.6	0.5
LDD_B_M_924‡	12.3	6.7	0.3
LDD_B_R_1071	0.0	5.9	5.4
GDH_A1_Y_67	9.4	9.9	0.5
GDH_A1_W_315	10.4	6.4	0.5
GDH_A1_R_891‡	8.6	6.4	0.5
GDH_A1_Y_970[c]	6.7	5.6	0.3
GDH_B9_133_S	16.3	5.9	0.5
GDH_B5_259_W‡	0.3	5.6	0.0
GSP_A_R_176	0.0	5.6	0.5
GSP_A_R_345	5.4	5.6	0.3
GSP_A_R_366	*100.0*	*100.0*	*100.0*
GSP_A_Y_789	2.1	6.2	0.5
GSP_A_Y_716	17.4	21.2	6.7
GSP_B_Y_395	20.1	9.7	2.4
Hipl_A_S_347	13.1	2.7	4.5
Hipl_A_Y_410	1.3	5.6	5.4
CHS_A_S_133	8.0	6.4	5.4
CHS_A_Y_300	11.5	12.9	23.8
HG_A_Y_335	*28.6*	*54.4*	*50.3*
HG_A_M_489	*27.3*	*55.5*	*49.7*
HG_A_K_615	*27.5*	*55.8*	*50.3*
HG_A_R_695	*26.2*	*56.0*	*50.3*
PSY_A_Y_198	8.8	6.7	0.3
PSY_A_R_830	21.9	24.1	4.3
PSY_B_M_640	*49.7*	*70.5*	*65.0*
PSY_B_K_873	1.3	3.8	0.5
SPA2_A_Y_1292	23.0	20.6	10.4
SPA2_A_S_1361	6.4	5.6	0.3
SPA2_B_Y_453	8.6	5.4	0.5
Hd1_A_K_1545	*28.6*	*100.0*	*100.0*
Sal1_A_S_149	10.7	29.5	0.3
Sal1_B_Y_302	2.7	5.9	0.5
Sal1_B_S_416‡	11.5	17.7	0.8
Sal1_B_Y_452	0.8	5.9	0.5
Sal1_B_M_509	0.3	4.3	2.7
Sal1_D_M_445	13.4	5.9	0.5

*CC1 corresponds to the 372 lines of the core collection defined by Balfourier *et al*. (2007).

†CC2 corresponds to the 372 supplementary accessions chosen by Balfourier *et al*. (2007) to balance the size of the geographical groups which structured the core collection CC1.

‡Markers used for linkage mapping.

Table 4 Success rate of the assay, excluding the seven discarded markers: number and percentage of called lines for each plate

Success rate	Di-Tetra	CC1	CC2
Average	344.4	340.3	366.7
	92.1%	91.1%	98.1%
Range	223.0–374.0	119.0–373.0	285.0–374.0
	59.6%–100.0%	31.8%–99.7%	76.2%–100.0%

Table 5 Percentage of discordant values observed for validation markers using three methods of validation in CC1. First, alleles obtained by SNPlex^{TM} and Sanger sequencing of 42 reference lines were compared (method M1). The two other methods involved the comparison of SNPlex^{TM} data and sequences (M2) or data generated by MassARRAY® technology (M3) for all of the lines in CC1

	Percentage of discordant SNP values		
Validation marker	Method M1	Method M2	Method M3
AAP_D_Y_450	0	–	0
LDD_B_M_924	9.5	–	4.0
GDH_A1_Y_67	2.4	–	0.6
GDH_A1_R_891	0	–	1.7
GDH_B5_259_W	0	–	0
Sal1_A_S_149	11.9	–	6.6
Sal1_B_Y_302	0	0	1.1
Sal1_B_S_416	0	0	0
Sal1_B_Y_452	0	0	–
Sal1_B_M_509	36.8	30.0	30.0
Sal1_D_M_445	0	–	0.6

–, method was not used; SNP, single nucleotide polymorphism.

all markers in 42 lines, M2 to the Sanger sequencing of all lines for four validation markers, and M3 to the MassARRAY® technology on all lines for 11 validation markers.

For 42 lines, the data obtained by SNPlex™ genotyping were compared with the data obtained by sequencing (M1 in Table 5). The marker AAP_A_R_99, which showed a lot of missing data in the CC1 plate that included the 42 references lines, was discarded. We observed no discordance for 34 of the 39 selected markers. For two markers (GDH_A1_Y_67 and GDH_B9_S_133), we detected only one discordance. The markers LDD_B_M_924 and Sal1_A_S_149 gave four and five (about 10%) discordant values, respectively. The worst results were obtained for Sal1_B_M_509 (M2 in Table 5). It appeared to be monomorphic for allele C, whereas the frequency of the minor allele (A) was about 35% in sequenced lines.

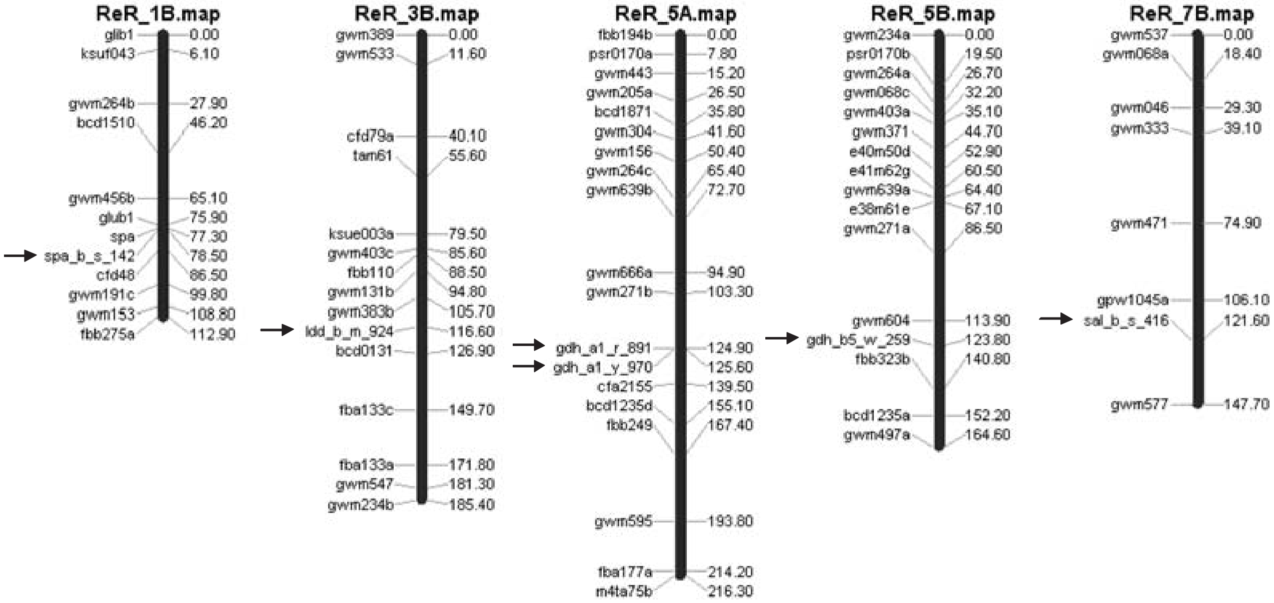

Figure 1 Genetic map positions of six polymorphic markers in Renan × Recital population. The markers genotyped in this study are indicated by black arrows. Map distances are indicated in centiMorgans. The marker called spa, previously mapped, is identical to spa_b_s_142.

For the 11 validation markers, the concordance of the genotypes given by SNPlex™ technology with those obtained by Sanger sequencing and MassARRAY® technology approached 96%. The genotype calls given by SNPlex™ and MassARRAY® technology were in perfect agreement for three markers (AAP_D_Y_450, GDH_B5_259_W and Sal1_B_S_416), and we observed greater than 5% discordance for only two markers (Sal1_A_S_149 and Sal1_B_M_509 with 6% and 30%, respectively) (M3 in Table 5). For the latter marker, we observed many discordant values, as described above for the M1 sequencing results. Sanger sequencing and MassARRAY® technology showed that 114 lines contained the A allele; all were classified as the C allele by SNPlex™.

The number of discordant genotypes observed for the M1 and M3 validation methods based on 11 markers was highly correlated (0.98). The correlation between the number of discordant genotypes observed for the M1 and M2 validation methods was based on only four markers and was 0.99. These high correlations suggest that the results obtained with a set of reference lines are representative for the whole sample, and thus can be used for validation.

We also mapped the six markers showing a polymorphism between Renan and Récital cultivars (SPA_B_S_142, LDD_B_M_924, GDH_A1_R_891, GDH_A1_Y_970, GDH_B5_W_259 and Sal1_B_S_416). The results shown in Figure 1 are in accordance with previous mapping results for three genes, *SPA* (storage protein activator), *LD* (lumini dependens) and *GDH* (glutamate dehydrogenase) (Guillaumie *et al*., 2004; Fontaine *et al*., in press), and present the first mapping of the supernumerary aleurone layer 1 gene (*Sal1*).

Discussion

Seventy-five pre-selected SNPs were submitted and all but two were included in two sets of markers. This good result can probably be attributed to the manual pre-selection step (see Experimental procedures), which is required to avoid SNPs which cannot be developed by this technology.

The software GeneMapper v3.7 allows the automatic calling of alleles and is well adapted to diploids. For each SNP, the software creates a genotype plot (two neighbouring peaks), which shows the intensity of each allele, and a Cartesian plot using a clustering algorithm to assign genotypes (Figure 2a) (De La Vega *et al*., 2005; Tobler *et al*., 2005). However, the polyploidy of the wheat genome complicates the interpretation of these plots. The allele calling of each SNP is hindered by the presence of the homoeologous genomes. Consequently, we observed a modification of the location of clusters relative to the diploid plot and a mis-assignment of genotypes, as the software considers one of the two homozygous genotypes as heterozygous. Thus, additional manual reading and corrections are required (Figure 2b). This manual genotype reading is based on the semi-quantitative nature of the SNPlex™ method. Figure 3 gives an example of a tetraploid and hexaploid genetic context for an SNP on one homoeologue: assuming

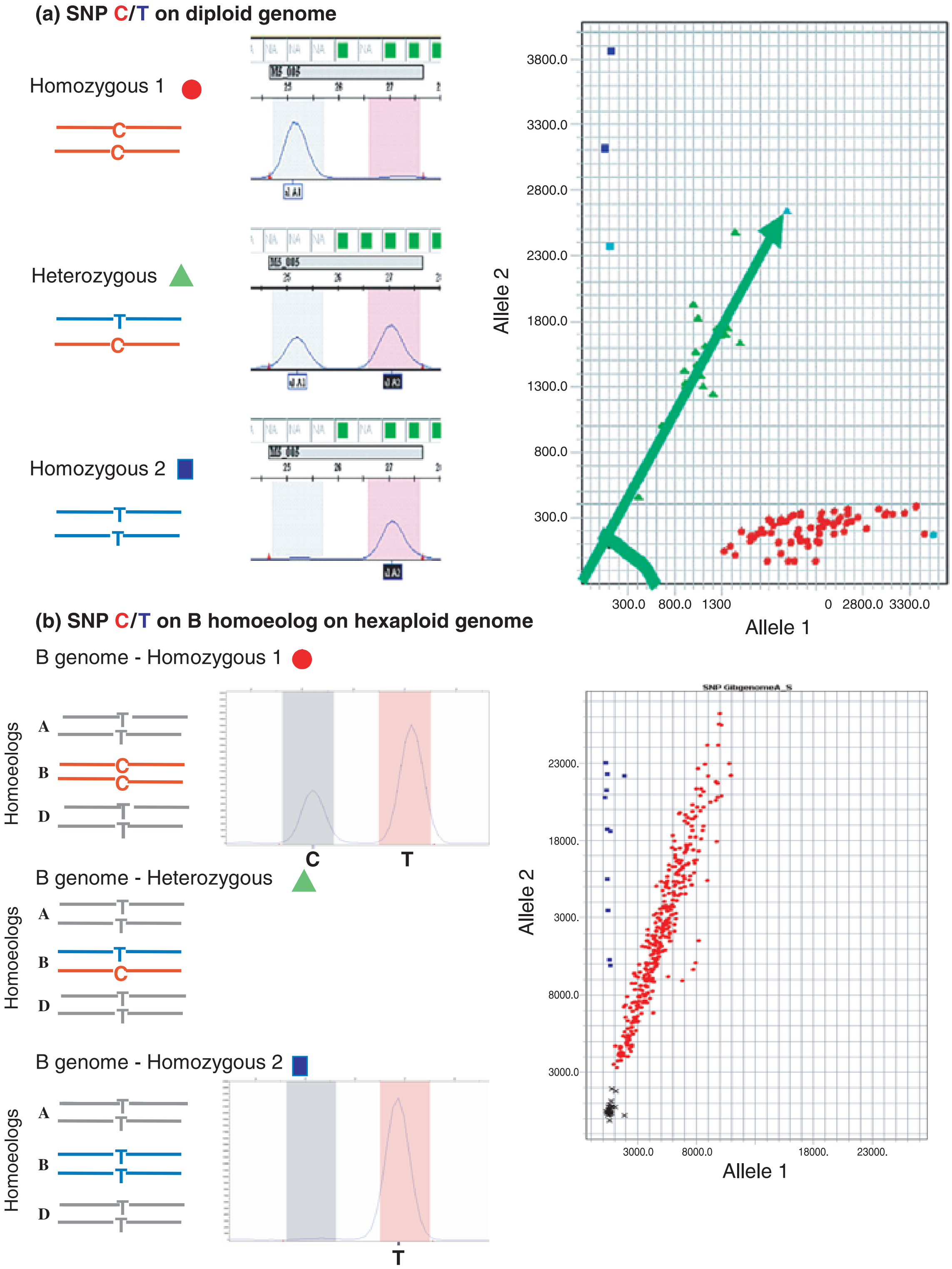

Figure 2 GeneMapper® analysis. The genetic context of a particular single nucleotide polymorphism (SNP) is represented on the left. The two peaks indicating the genotype are shown in the middle. The Cartesian plots with the intensities of each peak on the *x* and *y* axis are on the right. (a) Analysis in a diploid species: three clusters, one for each homozygous genotype (blue and red dots) and one for the heterozygous genotype (green dots). (b) Analysis in hexaploid wheat for an SNP on the B homoeologue after manual correction of calling [without correction, the homozygous genotypes (red dots) were considered as undetermined points or heterozygous, and shown as black or green dots, respectively]. In this case, all samples were homozygous at the SNP position, and thus no green dots are seen.

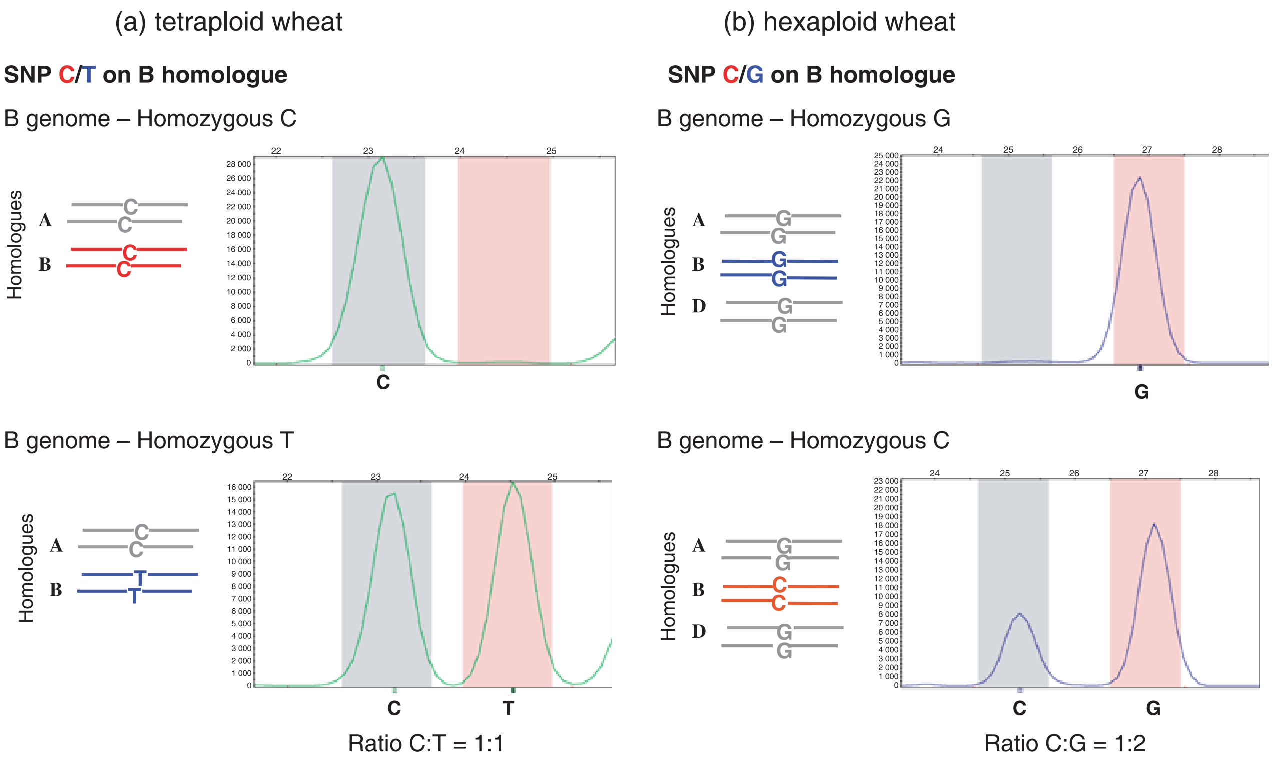

Figure 3 GeneMapper® genotype plots. The genetic context of a particular single nucleotide polymorphism (SNP) on the B homoeologous genome is represented on the left. Allelic ratios in a homozygous tetraploid species (a) and in a homozygous hexaploid species (b).

that all genomes are amplified, for a homozygous sample, we expect a ratio of 1 : 1 and 1 : 2 for the size of both allelic peaks for tetraploids and hexaploids, respectively. Although manual reading is feasible, it is time consuming, decreases the throughput and is likely to generate errors.

As the taxa under study are produced by self-fertilization, we expected very few heterozygous genotypes (Enjalbert and David, 2000; Charlesworth and Wright, 2001). However, we can expect that heterozygosity will complicate the interpretation of the Cartesian plot by producing a third cluster between the two homozygous ones. For the future use of this genotyping method in plant breeding, this difficulty will need to be solved. The addition of heterozygous control samples (natural or artificial) in the assay will facilitate the separation of the three expected clusters. Moreover, for an efficient use of SNPlex™ technology with polyploid genomes, an adaptation of the GeneMapper® clustering algorithm needs to be developed.

We obtained reliable genotypes for 35–39 SNPs of the 47 simultaneously genotyped by SNPlex™ (as previously described in Results, four markers gave reliable results on only one of the three plates). Two markers failed to give genotypes for all the lines under study. This is in agreement with the results reported by Pindo *et al*. (2008) and our experience in other species, as we often obtain up to four failed markers per set of 48 markers. Some markers failed because they gave too much missing data, making them unusable for genetic studies. We observed more than 50% missing data for five markers. Four are located in a hypothetical gene and one in a gene coding for phytoene synthase (*Psy-2*). The sequence variations in this hypothetical gene were probably too frequent, thus preventing hybridization of the probes. We also observed a high level of polymorphism in the B homoeologue of the *Psy-2* gene. These results highlight the difficulty in developing SNPlex™ probes (i.e. allelic and locus-specific probes) in highly polymorphic genomic regions. It is expected that the same difficulty will be encountered in other genotyping technologies based on hybridization probes (VeraCode™, GoldenGate®, Infinium®).

Validation by Sanger sequencing and MassARRAY® genotyping technology showed that the number of lines that were well called was high. For most markers, we observed more than 95% of correctly called lines. This is less than the concordance rate given by the manufacturer (> 99.2%) (http://www.appliedbiosystems.com). However, this rate is given for human samples with automated allele calling. In our context (polyploidy and manual reading), the rate of 5% misclassified lines appears to be acceptable. This assay failed to call the correct allele for two markers (Sal1-A_S_149 and Sal1_B_M_509). The discordance observed for the first marker (11.9% and 6.6% for M1 and M3 validation methods, respectively) may be caused by the presence of the same polymorphism in its D homoeologous sequence. Indeed, Sanger sequencing and MassARRAY® genotyping technology use a specific homoeologous amplification, whereas SNPlex™ technology amplifies the three genomes. This underlines the importance of knowing the sequence of each homoeologue and of pre-selecting SNPs for SNPlex™ technology. For the second marker, the result cannot be explained by the two other homoeologous sequences which have no polymorphism at this locus. Despite a good representation of the A allele, we detected only the C allele, as if the A allele could not be recognized by the probe. This may suggest a problem of synthesis of the probe. It also could be a result of a dilution of the A allele, if we suppose gene duplication of one homoeologue (the homozygous C): the ratio between the allelic peaks would then be 1 : 3 and the clusters would be more difficult to distinguish.

As an example of how the data can be used, we undertook the genetic mapping of six markers. The marker SPA_B_S_142 was mapped at 1.2 cM from the previous mapping position found by Guillaumie *et al*. (2004) (*SPA*). The genetic locations of the two markers on the A homoeologue of *GDH* (GDH_A1_R_891, GDH_A1_Y_970) are 124.90 cM and 125.60 cM, respectively. These different values are included in the confidence interval. We mapped the B homoeologue of *GDH* at 37.3 cM from gwm271, which is in agreement with the results reported by Fontaine *et al*. (in press). The B copy of *Sal1* has not yet been mapped in wheat. We mapped this gene in the telomeric region of the 7BL chromosome. In barley, *Sal1* was mapped in a syntenic position (Jestin *et al*., 2008). Therefore, these mapping results confirmed that the genotyping data used are reliable (Figure 1).

This study demonstrates the possibility of using existing high-throughput multiplex technologies, such as SNPlex™, for the analysis of polyploid genomes. For the first time, we have generated a set of 39 SNPs that can be used routinely and could serve as the basis for the development of a set of 48 markers.

Nevertheless, these results have highlighted some important points to be considered. It is evident that studying individuals from related species for which the polymorphism is largely or completely unknown should be avoided. Indeed, the larger number of lines giving missing data in the Di-Tetra plate, as well as the lower correlation between the number of missing data in CC1 or CC2 and Di-Tetra plates, suggests that the allelic and locus probes, which are the specific components, probably do not hybridize to their genomic target sequence in all species studied. Ideally, it is necessary to obtain sequences for several genotypes in order to validate markers that function in the assay. Moreover, the importance of having reference controls of known genotypes must be emphasized.

We also detected other problems, which are more specifically related to the polyploid structure of the wheat genome. A previous knowledge of the sequences of the three homoeologues allows the construction of a set of SNPlex™ markers that is more likely to succeed. Unfortunately, this situation is rare in wheat because of the difficulties presented by this species for Sanger sequencing. Thus, efforts are needed to develop SNP markers for this species. The major handicap for the routine use of this technique is the obligation to perform a manual verification of the results, and the lack of software that takes into account the polyploidy of genomes. These factors increase both the time required for the analysis and the risk of errors. We expect that the same limitations will hold for other SNP genotyping platforms, which, like SNPlex™, were developed for the analysis of the human genome.

In conclusion, we have shown the efficiency of the high-throughput genotyping technique SNPlex™ for species that are tetraploid or hexaploid, such as wheat. Our results suggest that other multiplex genotyping techniques (VeraCode™, GoldenGate® and even Infinium®) may also give good results for polyploid genomes. The routine use of such techniques awaits the development of software that is adapted to polyploid genomes. For the first time, a core of 39 SNPs is available for genotyping the complex of wheat species. This core can serve as the basis for the development of a set of 48 markers.

Experimental procedures

Plant material

Plant material consisted of 1314 lines of wheat with three different ploidy levels (2×, 4× and 6×).

gDNA from these lines was distributed in four 384-well plates. In each plate, we included four positive controls of known genotype and two negative controls (water). Table 2 describes the samples of *Triticum* taxa used. For each taxon, we used a core set of individuals representing the highest allelic diversity. These individuals were chosen on the basis of simple sequence repeat (SSR) polymorphism data using MSTRAT (Gouesnard *et al*., 2001). One plate, called Di-Tetra, contained 374 gDNAs from diploid progenitors of polyploid wheat species ($n = 50$) and tetraploid species ($n = 324$). Among the latter, tetraploid wheats (*T. turgidum*) were represented by the wild *T. turgidum* spp. *dicoccoïdes* ($n = 62$), the primary domesticated *T. turgidum* spp. *dicoccum* ($n = 97$), the cultivated durum wheat *T. turgidum* spp. *durum* ($n = 71$) and additional species and subspecies (*T. turgidum* ssp. *carthlicum* and ssp. *polonicum*, and *T. timopheevii* ssp. *timopheevii*). In this plate, four positive controls were included: one line per diploid progenitor (*T. urartu*, *Aegilops speltoïdes* and *T. tauschii*) and the cultivar Langdon (durum wheat).

Two plates contained 2 × 372 accessions of the hexaploid *T. aestivum*. This sample of 744 accessions comprised the 372 lines of the core collection defined by Balfourier *et al*. (2007), called CC1, and 372 supplementary accessions (CC2) chosen by these authors to balance the size of the geographical groups which structured the core collection. In these two plates, two of these 744 accessions (Récital and Redman cultivars) were each spotted twice and considered as positive controls. Among the 372 lines in the CC1 plate, 42 were used as reference lines for validation: data from other genotyping methods are available.

Seeds were obtained from the Biological Resource Centre for Cereal Crops (INRA, Clermont-Ferrand, France) and from Pierre Roumet (INRA, Montpellier, France; tetraploid wheats and *Aegilops* species). All seeds were obtained from a single self-pollinated head. Fresh leaves of five to six plants per accession were pooled, and bulk gDNA was extracted using a cetyltrimethylammonium bromide (CTAB) protocol, as described previously (Tixier *et al*., 1998).

A fourth plate, RILs-ReR, consisted of 194 F_7 RILs of the mapping population developed from the cross between Renan and Récital plus the two parental lines (Groos *et al*., 2002).

SNPlex™ genotyping

The SNPlex™ genotyping system uses an Applied Biosystems oligonucleotide ligation assay (OLA) combined with multiplex polymerase chain reaction (PCR) to achieve allelic discrimination and target amplification. The specificity of the ligation probes (allelic and locus-specific probes) for the target is critical.

All SNPs used came from French genomic projects which focused on SNP discovery. As we expected that polyploidy would engender difficulties for reading allele values obtained by SNPlex™, to the extent possible we used SNPs present in gene fragments for which we had sequenced the three homoeologues. We sent 75 SNPs to Applied Biosystems Assays-by-Design Service^{SM} (http://www.appliedbiosystems.com) in order to form a set of 48 SNPs (Table 1). We initially examined and pre-selected all these SNPs, following the SNPlex™ design rules, with the advice of Applied Biosystems. Generally, polymorphisms that were too close to one another were discarded, as were long insertions–deletions. In rare cases, we detected the same polymorphism in homoeologous genomes. Such SNPs were usually discarded.

The SNPlex™ assay was carried out in 2 days using the manufacturer's instructions (http://www.appliedbiosystems.com), taking care to perform pre-PCR and post-PCR steps in different locations. To facilitate the reading of genotyping data, it is necessary to have a uniform DNA concentration among samples on the same plate. All gDNAs were initially quantified by the Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen™, Carlsbad, CA, USA) in order to obtain an overview of the homogeneity of the DNA concentration. One modification has been made in the manufacturer's protocol: after testing different concentrations of gDNA, we performed the SNPlex™ reaction with a concentration of gDNA around 50 ng/μL, instead of 18.5 ng/μL, as is used for human gDNA. The need for a higher concentration of DNA could be explained by the different size of the wheat (16.5 Gb for *T. aestivum*) and human (3.5 Gb) genomes. Thus, in a given quantity of DNA, the target locus is about fourfold less represented in wheat than in humans.

Samples were run on a 3730*xl* DNA Analyser (Applied Biosystems). GeneMapper® software version 3.7 (Applied Biosystems) was used to analyse the raw data and corrected by a manual reading. For each SNP, this software analyses the raw capillary electrophoresis data, creates different plots and provides automated allele calling.

Validation

To validate SNPlex™ data in *T. aestivum*, one independent technical repetition was carried out. For all SNPs, sequence data from a set of 42 reference lines from the CC1 plate were available. Genotypes from these sequences were compared with SNPlex™ data generated in both assays (validation method 1 – M1). Furthermore, for all CC1 samples, 11 of the 47 SNPlex™ markers (23%) were characterized previously either by Sanger sequencing (validation method 2 – M2) or by MassARRAY® technology (Sequenom Inc., San Diego, CA, USA), a simplex genotyping method based on matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry (validation method 3 – M3) (Balfourier *et al*., 2006) (Table 5). These 11 markers are called validation markers. For three of these, localized in the supernumerary aleurone layer 1 gene, *Sal1* (Shen *et al*., 2003), both sets of genotyping data were available for the B homoeologous fragment.

Linkage analyses of six polymorphic markers in Renan and Récital cultivars were performed using Mapmaker/exp 3.06 (Lander *et al*., 1987). The Kosambi mapping function (Kosambi, 1944) was applied to transform recombination frequencies into additive distances in centiMorgans (cM).

Acknowledgements

We are grateful to Heather McKhann for critical reading of the manuscript. We thank Steven McGinn from Ivo Gut's group at the CEA-IG/CNG (Evry, France), directed by Mark Lathrop, for providing advice on the SNPlex™ technology.

References

Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol*. **8**, 135–141.

Balfourier, F., Ravel, C., Bochard, A.M., Exbrayat-Vinson, F., Boutet, G., Sourdille, P., Dufour, P. and Charmet, G. (2006) Développement, utilisation et comparaison de différents types de marqueurs pour étudier la diversité parmi une collection de blé tendre. *Les Actes du BRG*, **6**, 129–144.

Balfourier, F., Roussel, V., Strelchenko, P., Exbrayat-Vinson, F., Sourdille, P., Boutet, G., Koenig, J., Ravel, C., Mitrofanova, O., Beckert, M. and Charmet, G. (2007) A worldwide bread wheat core collection restricted to a full 384 deep well storage plate. *Theor. Appl. Genet*. **114**, 1265–1275.

Bonnin, I., Rousset, M., Madur, D., Sourdille, P., Dupuits, C., Brunel, D. and Goldringer, I. (2008) FT genome A and D polymorphisms are associated with the variation of earliness components in hexaploid wheat. *Theor. Appl. Genet*. **116**, 393–394.

Charlesworth, D. and Wright, S.I. (2001) Breeding systems and genome evolution. *Curr. Opin. Genet. Dev*. **11**, 685–690.

De La Vega, F.M., Lazaruk, K.D., Rhodes, M.D. and Wenz, M.H.

(2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan® SNP Genotyping Assays and the SNPlex™ Genotyping System. *Mutat. Res*. **573**, 111–135.

Drouaut, J., Mercier, R., Chelysheva, L., Bérard, A., Falque, M., Martin, O., Zanni, V., Brunel, D. and Mézard, C. (2007) Sex-specific crossover distributions and variations in interference level along *Arabidopsis thaliana* chromosome 4. *PLoS Genet*. **3**, 1096–1107.

Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Enjalbert, J. and David, J.L. (2000) Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics*, **156**, 1973–1982.

Fontaine, J.X., Ravel, C., Pageau, K., Heumez, E., Dubois, F., Hirel, B. and Le Gouis, J. (in press) A quantitative genetic study for elucidating the contribution of glutamine synthetase, glutamate dehydrogenase and other nitrogen-related physiological traits to the agronomic performance of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet*.

Giancola, S., McKhann, H.I., Bérard, A., Camilleri, C., Durand, S., Libeau, P., Roux, F., Reboud, X., Gut, Y.G. and Brunel, D. (2006) Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and TaqMan, in diploid and polyploid plants. *Theor. Appl. Genet*. **112**, 1115–1124.

Gouesnard, B., Bataillon, T., Decoux, G., Rozale, C., Schoen, D.J. and David, J. (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J. Hered*. **92**, 93–94.

Groos, C., Robert, N., Bervas, E. and Charmet, G. (2002) Analysis of genetic, environmental and genetic × environmental components for grain protein content, grain yield and thousand kernel weight in bread wheat. *Theor. Appl. Genet*. **104**, 39–47.

Guillaumie, S., Charmet, G., Linossier, L., Torney, V., Robert, N. and Ravel, C. (2004) Co-location between a gene encoding for the bZip factor SPA and an eQTL for a high-molecular-weight glutenin subunit in wheat (*Triticum aestivum*). *Genome*, **47**, 705–713.

Gupta, P.K., Mir, R.R. and Kumar, J. (2008) Wheat genomics: present status and future prospects. *Int. J. Plant Genomics*, **2008**, 896451.

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.

Hyten, D.L., Song, Q., Choi, I.Y., Yoon, M.S., Specht, J.E., Matukumalli, L.K., Nelson, R.L., Shoemaker, R.C., Young, N.D. and Cregan, P.B. (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theor. Appl. Genet*. **116**, 945–952.

Jestin, L., Ravel, C., Auroy, S., Laubin, B., Perretant, M.R., Pont, C. and Charmet, G. (2008) Inheritance of the number and thickness of cell layers in barley aleurone tissue (*Hordeum vulgare* L.): an approach using F2–F3 progeny. *Theor. Appl. Genet*. **116**, 991–1002.

Khlestkina, E.K. and Salina, E.A. (2006) SNP markers: methods of analysis, ways of development, and comparison on an example of common wheat. *Genetika*, **42**, 725–736.

Kosambi, D.D. (1944) The estimations of map distance from recombination value. *Ann. Eugeni*, **12**, 172–175.

Lander, E.S., Geen, P., Abrahanson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newbur, L. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **1**, 174–181.

Lijavetzky, D., Cabezas, J.A., Ibáñez, A., Rodríguez, V. and Martínez-Zapater, J.M. (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera L.*) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics*, **8**, 424.

Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I., Lamothe, M., Isabel, N. and Bousquet, J. (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics*. **9**, 21.

Pindo, M., Vezzulli, S., Coppola, G., Cartwright, D.A., Zharkikh, A., Velasco, R. and Troggio, M. (2008) SNP high-throughput screening in grapevine using the SNPlex genotyping system. *BMC Plant Biol*. **8**, 12.

Ravel, C., Praud, S., Murigneux, A., Canaguier, A., Sapet, F., Samson, D., Balfourier, F., Dufour, P., Chalhoub, B., Brunel, D., Beckert, M. and Charmet, G. (2006) Single-Nucleotide Polymorphisms (SNPs) frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). *Genome*, **49**, 1131–1139.

Shen, B., Li, C., Min, Z., Meeley, R.B., Tarczynski, M.C. and Olsen, O.A. (2003) *Sal1* determines the number of aleurone cell layers in maize endosperm and encodes a class E vacuolar sorting protein. *Proc. Natl. Acad. Sci. USA*, **10**, 6552–6557.

Simon, M., Loudet, O., Durand, S., Bérard, A., Brunel, D., Sennesal, F.X., Durand-Tardif, M., Pelletier, G. and Camilleri, C. (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics*, **178**, 2253–2264.

Sobrino, B., Brión, M. and Carracedo, A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic. Sci. Int*. **154**, 181–194.

Syvänen, A.C. (2005) Toward genome-wide SNP genotyping. *Nat. Genet*. **37**(Suppl.), S5–S10.

Tixier, M.H., Sourdille, P., Charmet, G., Gay, G., Jaby, C., Cadalen, T., Bernard, S., Nicolas, P. and Bernard, M. (1998) Detection of QTLs for crossability in wheat using double-haploid population. *Theor. Appl. Genet*. **97**, 1076–1082.

Tobler, A.R., Short, S., Andersen, M.R., Paner, T.M., Briggs, J.S., Lambert, S.M., Wu, P.P., Wang, Y., Spoonde, A.Y., Koehler, R.T., Peyret, N., Chen, C., Broomer, A.J., Ridzon, D.A., Zhou, H., Hoo, B.S., Hayashibara, K.C., Leong, L.N., Ma, C.N., Rosenblum, B.B., Day, J.P., Ziegle, J.S., De La Vega, F.M., Rhodes, M.D., Hennessy, K.M. and Wenz, H.M. (2005) The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J. Biomol. Tech*. **16**, 398–406.

Supporting information

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

Table S1 Flanking sequences of unpublished SNPs.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.