Theor Appl Genet (2007) 114:1265–1275
DOI 10.1007/s00122-007-0517-1

ORIGINAL PAPER

A worldwide bread wheat core collection arrayed in a 384-well plate

François Balfourier · Valérie Roussel · Pjotr Strelchenko · Florence Exbrayat-Vinson · Pierre Sourdille · Gilles Boutet · Jean Koenig · Catherine Ravel · Olga Mitrofanova · Michel Beckert · Gilles Charmet

Received: 24 October 2006 / Accepted: 28 January 2007 / Published online: 21 February 2007

Abstract Bread wheat (*Triticum aestivum*), one of the world's major crops, is genetically very diverse. In order to select a representative sample of the worldwide wheat diversity, 3,942 accessions originating from 73 countries were analysed with a set of 38 genomic simple sequence repeat (SSR) markers. The number of alleles at each locus ranged from 7 to 45 with an average of 23.9 alleles per locus. The 908 alleles detected were used together with passport data to select increasingly large sub-samples that maximised both the number of observed alleles at SSR loci and the number of geographical origins. A final core of 372 accessions (372CC) was selected with this M strategy. All the different geographical areas and more than 98% of the allelic diversity at the 38 polymorphic loci were represented in this core. The method used to build the core was validated, by using a second set of independent markers [44 expressed sequence tag (EST)-SSR markers] on a larger sample of 744 accessions: 96.74% of the alleles observed at these loci had already been captured in the 372CC. So maximizing the diversity with a first set of markers also maximised the diversity at a second independent set of locus. To relate the genetic structure of wheat germplasm to its geographical origins, the two sets of markers were used to compute a dissimilarity matrix between geographical groups. Current worldwide wheat diversity is clearly divided according to wheat's European and Asian origins, whereas the diversity within each geographical group might be the result of the combined effects of adaptation of an initial germplasm to different environmental conditions and specific breeding practices. Seeds from each accession of the 372CC were multiplied and are now available to the scientific community. The genomic DNA of the 372CC, which can be entirely contained in a 384-deep-well storage plate, will be a useful tool for future studies of wheat genetic diversity.

Communicated by M. Bohn.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-007-0517-1) contains supplementary material, which is available to authorized users.

F. Balfourier (✉) · V. Roussel · F. Exbrayat-Vinson · P. Sourdille · G. Boutet · J. Koenig · C. Ravel · M. Beckert · G. Charmet
INRA, UMR1095 Amélioration et Santé des Plantes, 234, avenue du Brézet, 63100 Clermont-Ferrand, France
e-mail: balfour@clermont.inra.fr

P. Strelchenko · O. Mitrofanova
All-Russian Research Institute of Plant Industry (VIR), 42, Bolshaya Morskaya str., 190000 St. Petersburg, Russia

Introduction

Bread wheat (*Triticum aestivum*) is a major cereal crop for both human and animal food and shows great genetic diversity worldwide. There are many large collections of wheat-genetic resources whose agro-morphological traits have been evaluated in the past, thus conserving the genetic diversity that has contributed to the more recent progress in the selection and registration of improved cultivars. With the development and the use of molecular markers, methods of describing and assessing genetic diversity at the molecular level have advanced rapidly over the last decade. However, applying such methods to large collections can still be costly and time-consuming and may be inefficient if the collection shows a significant level of redundancy.

Assembling a core collection of accessions is one way of streamlining the process of evaluating genetic diversity. This is done by selecting a subset of a larger germplasm collection that contains the maximum possible genetic diversity of the species with a minimum of repetitiveness (Frankel 1984). For instance, having a small representative sample of accessions should make it easier to look for allelic variation in genes of interest and more efficient to assess genotype-phenotype associations.

There are many different approaches available to build such a core as has been shown for numerous species such as sweet potato (Huaman et al. 1999), maize (Malosetti and Abadie 2001), peanut (Upadhyaya et al. 2002), and rice (Li et al. 2002); In general, these approaches have three steps. The first step is to describe the genetic diversity of the whole collection. Several descriptors can be used possibly in combination (e.g. passport data giving the geographical origin, morphological and phenotypic traits, and biochemical and molecular markers). The second step is to analyse the genetic structure of the diversity expressed by the different descriptors calculating the genetic distance between accessions and using clustering methods. In the last step, samples of individual accessions are selected from the whole collection to form the core. The core can be obtained applying stratified versus non-stratified or deterministic versus un-deterministic sampling strategies (Spagnoletti and Qualset 1993; van Hintum 1994; Balfourier et al. 1998; Hu et al. 2000). These different approaches have been compared by Franco et al. (2006). These authors demonstrated that one of the best strategies is that one developed by Shoen and Brown (1993) who proposed the M (for maximisation) strategy in which genetic markers are used to sample the core collection while maximizing allele richness at each marker locus.

Here, we studied about 40% of the INRA bread wheat collection, which has more than 10,000 accessions covering a wide range of geographical origins while periods of production varies from the eighteenth century to the twentieth century. Preliminary analyses of sub-samples from the whole collection indicated that the diversity in this collection is not randomly distributed and could be partly explained by temporal and geographical trends in variation linked to breeding practices and agricultural policies in different countries (Roussel et al. 2004, 2005). In order to define a bread wheat core collection, genomic-SSR and EST-SSR molecular markers were used to (1) characterise and assess the genetic diversity of a set of lines taken from the INRA bread wheat accessions and (2) sample a bread wheat core collection of the right size that maximise the number of observed alleles at each locus.

Materials and methods

Plant material and DNA extraction

The INRA cereal crop collections are kept in the Clermont-Ferrand Genetic Resources Center (http://www.clermont.inra.fr/umr-asp). The germplasm collection of hexaploid wheat consists of more than 10,000 accessions. One-third of the collection originates from France, one-third from Europe and the remaining third from the rest of the world. Based on preliminary observations of the geographical basis of diversity in this collection (Roussel 2005), 45 geographical areas, representing more than 70 different countries, were considered here. From this collection, 3,942 accessions (approximately 40% of the total collection) were chosen according to the number of accessions in each geographical group (total sample in Table 1). Included in this sample are approximately 350 accessions mainly from Asian regions (RUS-Central Asia, Caucasus and Asian countries) that were contributed by the Vavilov Institut (VIR). This whole sample of 3,942 accessions, which included landraces and both old and recently registered cultivars, was used for microsatellite analysis. Some landraces (about 5% of the whole sample) were used in the present analysis to represent geographical areas where there were little or no registered varieties in total collection. All the seeds used for DNA isolation were obtained from self-pollinated ears. Fresh leaves of five to six plants per accession were pooled and bulk genomic DNA was extracted using a cetyltrimethylammonium bromide (CTAB) protocol as previously described (Tixier et al. 1998).

Primary SSR analysis

To assess the genetic diversity in the whole sample of 3,942 accessions, a first set of 37 genomic SSR markers was used revealing 38 polymorphic loci. All these microsatellite markers are from IPK-Gatersleben (WMS SSRs from Röder et al. 1998), except CFD71, which was developed at INRA Clermont-Ferrand (Guyomarc'h et al. 2002) and shows two polymorphic loci. Markers were chosen according to their location on the wheat genetic map and their suitability for high throughput genotyping (Roussel et al. 2004, 2005). Conditions for PCR amplification and fragment size analyses were those described by Roussel et al. (2004, 2005). Polymorphisms were visualised using an ABI

Table 1 Number and percentage (in parenthesis) of accessions per geographical area in the different collections

Geographical area [a]	Total sample	Core collection	Validation sample	Geographical area [a]	Total sample	Core collection	Validation sample
FRA	1312 (33.3)	101 (27.2)	103 (13.8)	AUS–NZL	111 (2.82)	13 (3.49)	15 (2.02)
NLD	76 (1.93)	5 (1.34)	19 (2.55)	RUS–Central Asia (TJK–TKM–KAZ–KIR–UZB)	125 (3.17)	11 (2.96)	13 (1.75)
DEU	89 (2.26)	6 (1.61)	17 (2.28)	Caucasus (ARM–GEO–AZE)	40 (1.01)	10 (2.69)	10 (1.34)
GBR–IRL	95 (2.41)	6 (1.61)	17 (2.28)	TUR	58 (1.47)	7 (1.88)	10 (1.34)
BEL	69 (1.75)	3 (0.81)	17 (2.28)	NPL	73 (1.85)	24 (6.45)	24 (3.23)
SWE	75 (1.90)	2 (0.54)	16 (2.15)	CHN–KOR–MNG	116 (2.94)	17 (4.57)	17 (2.28)
NOR–DNK	18 (0.46)	1 (0.27)	15 (2.02)	JPN	67 (1.70)	12 (3.23)	12 (1.61)
FIN	26 (0.66)	6 (1.61)	17 (2.28)	PAK–KSM	30 (0.76)	5 (1.34)	10 (1.34)
CHE	81 (2.05)	7 (1.88)	20 (2.69)	SYR	34 (0.86)	4 (1.08)	9 (1.21)
POL	78 (1.98)	7 (1.88)	17 (2.28)	AFG–IRN–IRQ	16 (0.41)	1 (0.27)	10 (1.34)
CZE	57 (1.45)	6 (1.61)	17 (2.28)	IND	44 (1.12)	5 (1.34)	9 (1.21)
AUT	52 (1.32)	6 (1.61)	17 (2.28)	DZA–MAR	16 (0.41)	2 (0.54)	9 (1.21)
ROM	68 (1.73)	3 (0.81)	16 (2.15)	EGY–TUN	25 (0.63)	5 (1.34)	15 (2.02)
BGR	80 (2.03)	5 (1.34)	17 (2.28)	ETH–NER	15 (0.38)	3 (0.81)	14 (1.88)
UKR–BLR	69 (1.75)	5 (1.34)	19 (2.55)	KEN	30 (0.76)	2 (0.54)	10 (1.34)
YUG–HRV	77 (1.95)	2 (0.54)	15 (2.02)	ISR–LBN–PAL	56 (1.42)	7 (1.88)	12 (1.61)
HUN	80 (2.03)	7 (1.88)	17 (2.28)	ZAF–ZWE	21 (0.53)	3 (0.81)	10 (1.34)
ESP	65 (1.65)	11 (2.95)	21 (2.82)	BRA	54 (1.37)	4 (1.08)	10 (1.34)
PRT	33 (0.84)	4 (1.08)	16 (2.15)	CHL	33 (0.84)	1 (0.27)	9 (1.21)
GRC–ALB–MAD	15 (0.38)	2 (0.54)	15 (2.02)	COL–PER	12 (0.30)	2 (0.54)	9 (1.21)
ITA	78 (1.98)	4 (1.08)	17 (2.28)	MEX–GTM	103 (2.61)	9 (2.42)	10 (1.34)
USA	115 (2.92)	12 (3.23)	21 (2.82)	ARG–URY	76 (1.93)	5 (1.34)	11 (1.48)
CAN	79 (2.00)	9 (2.42)	20 (2.69)				
				Total	3,942 (100.00)	372 (100.00)	744 (100.00)

(*AFG* Afghanistan, *ALB* Albania, *ARG* Argentina, *ARM* Armenia, *AUS* Australia, *AUT* Austria, *AZE* Azerbaijan, *BEL* Belgium, *BGR* Bulgaria, *BLR* Belarus, *BRA* Brazil, *CAN* Canada, *CHE* Switzerland, *CHL* Chile, *CHN* China, *COL* Colombia, *CSK* Czech and Slovak Republics, *DEU* Germany, *DNK* Denmark, *DZA* Algeria, *EGY* Egypt, *ESP* Spain, *ETH* Ethiopia, *FIN* Finland, *FRA* France, *GEO* Georgia, *GBR* Great Britain, *GRC* Greece, *GTM* Guatemala, *HUN* Hungary, *HRV* Croatia, *IND* India, *IRL* Irleland, *IRN* Iran, *IRQ* Iraq, *ISR* Israel, *ITA* Italy, *JPN* Japan, *KAZ* Kazakhstan, *KEN* Kenya, *KIR* Kyrgyzstan, *KOR* Korea, *KSM* Kashmir, *LBN* Lebanon, *MAD* Macedonia, *MAR* Morocco, *MEX* Mexico, *MNG* Mongolia, *NER*:Niger, *NLD* Netherlands, *NOR* Norway, *NPL* Nepal, *NZL* New Zealand, *PAL* Palestine, *PAK* Pakistan, *POL* Poland, *POR* Portugal, *PER* Peru, *ROM* Romania, *RUS* Russia, *SYR* Syria, *SWE* Sweden, *TJK* Tajikistan, *TKM* Turkenistan, *TUN* Tunisia, *TUR* Turkey, *URY* Uruguay, *UKR* Ukraine, *USA* United States, *UZB* Uzbekistan, *YUG* Yugoslavia, *ZAF* South Africa, *ZWE* Zimbabwe)

PRISM 3100 Genetic Analyser (Applied Biosystems, Foster City, CA, USA).

The total number of alleles observed per locus, the number of alleles per polymorphic locus (i.e. effective alleles), the number of rare alleles per locus (i.e. alleles with frequency lower than 5%) and the Nei's genetic diversity index (*H*) (Nei 1973) per locus were calculated for the whole set of 3,942 accessions, using GENETIX software (Belkhir et al. 2004).

Core collection sampling

The M strategy described by Schoen and Brown (1993) was used for sampling core collections to maximise both the number of observed alleles at SSR loci and the number of geographical origins. The superiority of this method is based on expected linkage disequilibrium between neutral marker loci and other selected loci associated through hitchhiking; alleles at a SSR locus (marker locus) might be related to alleles at another locus (target locus) because the accessions share common ancestries or because of their mating system (e.g. self-fertilization) that favours gametophytic disequilibrium. The allele richness of a core subset formed by the M strategy was defined in terms of the number of allele classes represented in the sample. The MSTRAT software algorithm (Gouesnard et al. 2001) was used to generate core subsets according to this strategy. As increasing the number of iterations (p) did not improve the values of the core subset diversity measurements, the final number of iterations run on MSTRAT was $p = 100$ for each process, while the number of repetitions (q) for core sampling was $q = 50$.

The efficiency of the strategy was assessed by comparing the total number of alleles captured using MSTRAT to the number of alleles captured in randomly chosen collections of equal size, with sample sizes increasing from 1 to 1,053 in increments of 55 accessions. Fifty independent core collection samplings were made in each case ($q = 50$).

In MSTRAT software, the sampling procedure allows the user to specify a compulsory set of accessions (the "kernel" core) that will always be included in the core collection. As our objective here was to capture a maximum of alleles from the 3,942 accessions, a kernel core of all the accessions with unique alleles was specified in the sampling process. For this, an allele at a single locus was considered as unique when it occurred in only one of the 3,942 accessions.

Building and validating a core collection based on a second SSR analysis

The size of the core was arbitrarily fixed at 372 accessions to facilitate future analysis of the whole core subset with 12 controls on a single 384-well plate. So 100 cores of 372 accessions, including the kernel, were generated by MSTRAT ($p = 100$, $q = 100$) and the core exhibiting the greatest allelic richness and the best Nei's diversity index was selected as the final 372 core collection (372CC).

To see whether the method for selecting the final 372CC also captured unknown diversity, a second independent set of SSR markers were tested on 744 accessions. This sample of 744 accessions comprised the 372CC plus a second set of 372 accessions chosen from the remaining accessions analysed with the first set of SSR markers. The 372 supplementary accessions were chosen in order to balance, as far as possible, the sizes of the 45 different geographical groups of accessions (except for the French group; Table 1).

The second set of 44 markers, GPW (Nicot et al. 2004) and CFE EST-SSRs markers (Zhang et al. 2005) were selected in order to give good coverage of the whole genome map. Sequences, melting temperature and details of primers are given in the GrainGenes database (http://www.wheat.pw.usda.gov). Each forward primer was M13-tailed [M13: 5′-CACGACGTTGTAAAACGAC-3′] and synthesised by MWG (Germany). PCR analyses using the M13 protocol were performed as described by Nicot et al. (2004) with an annealing temperature of 60°C for 30 cycles (30 s at 94°C, 30 s at 60°C, 30 s at 72°C) and 56°C for 8 cycles.

Geographical component of bread wheat collection diversity

In order to get a representation of the genetic structure of the INRA wheat germplasm collection related to its geographical origins, genomic-SSR and EST-SSR results from the 744 accessions were merged and accessions were grouped according to their geographical origins in order to calculate individual allelic frequencies in the 45 groups. Then the full group frequency matrix was used to compute Cavalli-Sforza and Edwards (1967) dissimilarity coefficients, between pairs of geographical groups.

Finally, to illustrate the geographical components of the wheat germplasm diversity, a Neighbour-Joining tree was created from the Cavalli-Sforza and Edwards dissimilarity matrix, using DARwin software (Perrier et al. 2003).

Results

Overall diversity of 3,942 accessions

The genomic diversity of a collection of 3,942 wheat accessions was assessed using molecular markers corresponding to 38 polymorphic loci and the resulting statistics are summarised in Table 2. A total of 908 alleles were detected from the 38 amplified loci. Primer pairs WMS312 and WMS664 detected, respectively, the most and the fewest alleles (45 and 7), while the average number of alleles detected per locus was 23.9. The total number of rare alleles (730) represented about 80% of the total number of alleles. The number of rare alleles per locus ranged from 5 for locus Xgwm664-3D to 39 for Xgwm312-2A. Of the 3942 accessions, a set of 111 accessions (the kernel core) has unique alleles. The microsatellite markers used showed different levels of genetic diversity: Nei's genetic diversity index (H) ranged from 0.223 to 0.903, with an average of 0.742 for all markers.

Generation of core collections

The 908 alleles detected at 38 genomic SSR loci and the 45 geographical groups were both used as "alleles" in MSTRAT to generate increasingly large core collections. First, samples were made without a kernel core. Figure 1 shows that, in this case, the sampling efficiency (i.e. the ability to capture genetic diversity) using the M strategy was always better than a random sampling. Furthermore, the relative efficiency of the M strategy was highest for smaller samples: for instance, the M strategy outperformed a random sample by 50% when sample sizes were in the 111–215 range (i.e. a core size that is 3–6% of the whole sample size). However, both strategies failed to capture the whole set of 908 SSR alleles and the 45 geographical "alleles" in core samples with less than 1,053 accessions.

Table 2 Total number of effective alleles, number of rare alleles and Nei's diversity index (*H*) for 38 genomic SSR loci

SSR locus	Total number of alleles in 3,942 accessions	Number of rare alleles in 3,942 accessions	*H*	Total number of alleles in 372 core
Xgwm99-1A	26	22	0.688	24
Xgwm135-1A	36	32	0.713	36
Xgwm11-1B	26	22	0.791	24
Xgwm413-1B	20	15	0.789	20
Xgwm642-1D	20	17	0.624	19
Xgwm337-1D	23	19	0.846	23
Xgwm312-2A	45	39	0.874	45
Xgwm372-2A	36	30	0.903	36
Xgwm257-2B	13	10	0.644	13
Xgwm120-2B	30	26	0.864	29
Xgwm539-2D	40	35	0.888	40
Xgwm261-2D	28	25	0.721	28
Xgwm2-3A	11	7	0.579	11
Xgwm480-3A	26	24	0.317	26
Xgwm566-3B	14	8	0.801	14
Xgwm664-3D	7	5	0.223	7
Xgwm341-3D	34	28	0.885	33
Xgwm610-4A	26	23	0.642	24
Xcfd71-4A	11	8	0.459	10
Xgwm251-4B	25	19	0.844	25
Xgwm149-4B	15	12	0.565	14
Xcfd71-4D	23	16	0.885	23
Xgwm415-5A	10	7	0.587	10
Xgwm186-5A	27	22	0.863	26
Xgwm408-5B	28	22	0.821	27
Xgwm234-5B	28	21	0.881	27
Xgwm272-5D	18	14	0.65	17
Xgwm190-5D	25	19	0.743	25
Xgwm427-6A	24	19	0.847	23
Xgwm219-6B	30	24	0.869	30
Xgwm626-6B	20	18	0.549	20
Xgwm469-6D	21	16	0.833	21
Xgwm325-6D	18	11	0.768	18
Xgwm260-7A	30	26	0.824	30
Xgwm400-7B	18	12	0.828	18
Xgwm46-7B	27	21	0.865	27
Xgwm44-7D	21	14	0.855	21
Xgwm437-7D	28	22	0.861	28
Total	908	730		892
Mean/locus			0.742	

In a second step, we used the set of 111 accessions with unique alleles as a kernel core in MSTRAT to compare random and M strategies. Again, the sampling efficiency of the M strategy was always better than a random sampling (Fig. 1). However, the capture of all alleles was reached with about 400 accessions using the M strategy, while this result was not achieved in randomly selected samples.

Final 372-core collection

The size of the core is arbitrary but was fixed here at 372 to facilitate future analysis of the whole core by arraying it on a 384-well plate. The final 372-core collection (372CC), chosen from 100 core collections generated by MSTRAT software, captured 892 out of 908 SSR alleles and all 45 geographical origins. The total number of effective alleles observed per locus in the 372CC is given in Table 2; compared to the distribution in the whole sample of 3,942 accessions, there was no apparent distribution bias. With 892 SSR alleles, 372CC captured more than 98% of all the allelic diversity observed at 38 polymorphic loci in the sample of 3,942 accessions. These 372 accessions and their geographical origins are listed as electronic supplementary material (see ESM), and the numbers of accessions grouped by geographical origin are given in Table 1. Compared to the proportional representation in the whole sample of 3,942 accessions, we observed that some countries of origin, such as Sweden, Norway,

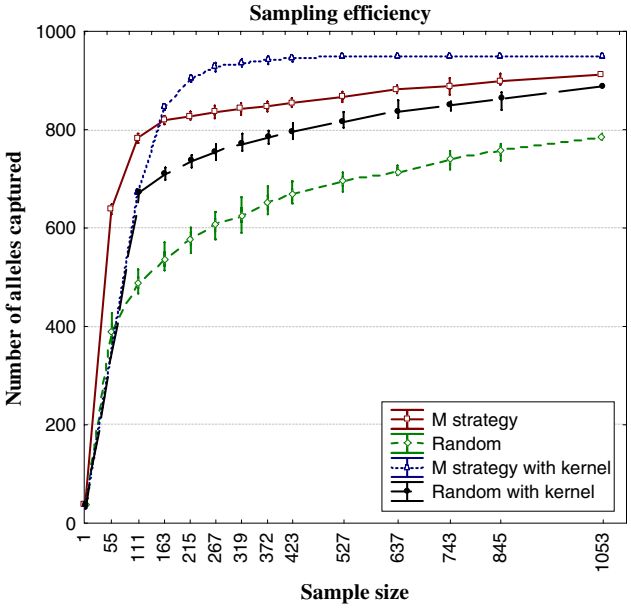

Fig. 1 Number of alleles captured with respect to accession sample size in four sampling strategies generated in MSTRAT software

Yugoslavia and Romania were under-represented, while others, such as Spain and Nepal, were over-represented in 372CC.

Validation of the core

Table 1 shows the number of accessions per geographical area within the sample of 744 accessions used to validate the core. Except for France, the number of accessions per geographical region ranged from 9 to 24. Statistics describing the allelic diversity of these 744 accessions for 44 additional EST-SSR markers are summarised in Table 3. A total of 184 alleles were detected with the 44 EST-SSRs markers. The total number of alleles per locus ranged from 1 to 14. Nine of these loci were monomorphic. When considering only the 35 polymorphic EST-SSR loci, the average number of effective alleles per locus was 5. For these markers, Nei's genetic diversity index (H) ranged from 0.006 to 0.782, with an average of 0.347 for the 35 markers. A total of 178 alleles were detected within the 372CC. Table 4 summarises the total number of alleles detected for the two types of markers and the different sample sizes. Most of the genomic SSR alleles were present in the 744 accessions, as 98.24% of these alleles were captured in the 372CC when using the M strategy. For the second independent set of 44 markers, 96.74% of the alleles observed in the 744 accessions were captured in the 372CC. Therefore, maximizing the diversity of the core with the first set of genomic markers maximised the diversity at an independent set of loci more likely to be associated with adaptive traits.

Geographical components of diversity in bread wheat germplasm

Figure 2 shows the Neighbour-Joining tree based on a Cavalli-Sforza and Edwards dissimilarity matrix between the 45 geographical groups. This hierarchical tree clearly separates European, Oceanic and North American countries of origin (lower part of the figure) from the others (upper part). In the lower part of the figure, one sub-cluster groups North-West European countries (NLD, DEU, GBR-IRL, BEL, SWE, NOR-DNK, POL, CZE, AUT, CHE and FRA); South-East European countries (UKR-BLR, ROM, YUG, HRV, BGR and HUN) form a second sub-cluster and Mediterranean European countries (ITA, GRC, ALB, MAD, PRT and ESP) a third sub-cluster. North American countries (USA and CAN) and Oceanic countries (AUS-NZL) are grouped together in the lower part of the tree. By contrast, in the upper part of the figure, some Near Eastern and Central-Asian areas are grouped in the same sub-cluster (TUR, Caucasus, RUS, Central Asia, AFG-IRN-IRQ). Far Eastern countries (JPN, CHN-KOR-MNG, NPL, PAK, KSM, IND) form a second sub-cluster while African, and South-American countries form a third sub-cluster.

Discussion

Here, we describe the fingerprinting of 3,942 bread wheat accessions with SSR markers, the largest and most extensive study of this kind to date. Huang et al. (2002) genotyped 998 bread wheat accessions originating from 68 countries, detecting 470 alleles from 26 microsatellite loci (i.e. on average, 18.1 alleles per locus). More recently, Hao et al. (2006), genotyping 340 Chinese wheat accessions with 78 SSR loci, detected 967 alleles (13.6 alleles per polymorphic locus). The allelic richness of 23.9 observed in our study was much higher. This result can be directly compared to the allelic richness observed in 559 French accessions and 480 European accessions using the same set of markers—14.5 and 16.4, respectively—as the same accessions are included in the 3,942 sample (Roussel et al., 2004, 2005). This gives an idea of the reserves of diversity present in bread wheat germplasm originating from outside Europe. Furthermore, the high proportion (80%) of rare alleles found in our sample indicates that, conversely, there is a low proportion of non-informative alleles in the wheat collection (i.e.

Table 3 Total number of alleles, number of rare alleles and Nei's diversity index (*H*) for 44 EST-SSR loci

EST-SSR locus	Total number of alleles in 744 accessions	Number of rare alleles in 744 accessions	*H*	Total number of alleles in 372 core
Xcfe189-1A	1			1
Xgpw7072-1A	8	3	0.759	8
Xcfe167-1A	1			1
Xgpw7443-1B	4	2	0.18	3
Xgpw7577-1B	6	4	0.433	6
Xgpw7082-1D	5	3	0.275	5
Xcfe78-1D	1			1
Xgpw7570-2A	5	4	0.105	4
Xcfe175-2A	4	2	0.481	4
Xgpw7438-2B	2	2	0.491	2
Xcfe52-2B	7	4	0.511	7
Xgwp7325-2D	1			1
Xcfe68-2D	3	2	0.101	3
Xgpw7213-3A	3	1	0.128	3
Xgpw7335-3B	1			1
Xgpw7452-3B	10	7	0.693	10
Xgpw7586-3D	2	1	0.006	2
Xgpw7553-3D	2	1	0.012	2
Xcfe172-3D	3	1	0.182	3
Xcfe300-4A	9	4	0.745	9
Xgpw7241-4B	2	1	0.018	2
Xcfe8-4B	6	4	0.518	6
Xgpw7666-4D	1			1
Xgpw7795-4D	7	6	0.125	7
Xcfe186-5A	3	0	0.463	3
Xgpw7218-5A	6	3	0.553	5
Xgpw7574-5B	1			1
Xgpw7425-5B	2	0	0.172	2
Xgpw7107-5D	1			1
Xcfe301-5D	5	1	0.568	5
Xgpw7384-6A	4	2	0.297	4
Xgpw7592-6A	4	2	0.206	4
Xcfe273-6A	3	1	0.207	3
Xcfe214-6B	2	1	0.017	2
Xgpw7433-6D	5	3	0.172	5
Xcfe95-6D	3	1	0.106	3
Xgpw7386-7A	9	3	0.782	9
Xgpw7185-7A	2	1	0.052	2
Xgpw7288-7A	9	6	0.636	9
Xcfe248-7A	1			1
Xgpw7596-7B	2	0	0.176	2
Xgpw7320-7B	14	11	0.744	13
Xgpw7342-7B	11	6	0.763	10
Xcfe135-7D	3	1	0.457	2
Total	184	94		178
Mean/locus			0.347	

common alleles present in the majority of accessions are non-informative in building a core collection with the M strategy).

In addition to their use in describing genetic diversity, molecular markers can be useful for forming core collections (Shoen and Brown 1993). Very recently, Franco et al. (2006) used genetic markers on three maize data sets to study 24 stratified sampling strategies and to investigate which strategy conserved the maximum diversity in the core subset compared to the M strategy. They concluded that for most of the diversity indices they used the M strategy outperformed the D method, a strategy based on maximising genetic distance rather than allelic richness in the core. We used information about SSR alleles to generate core collections with the M strategy using MSTRAT software. McKhann et al. (2004) used this software to form core collections from 256 accessions of *Arabidopsis thaliana*, using single nucleotide polymorphisms as markers. The M strategy has been shown to be more

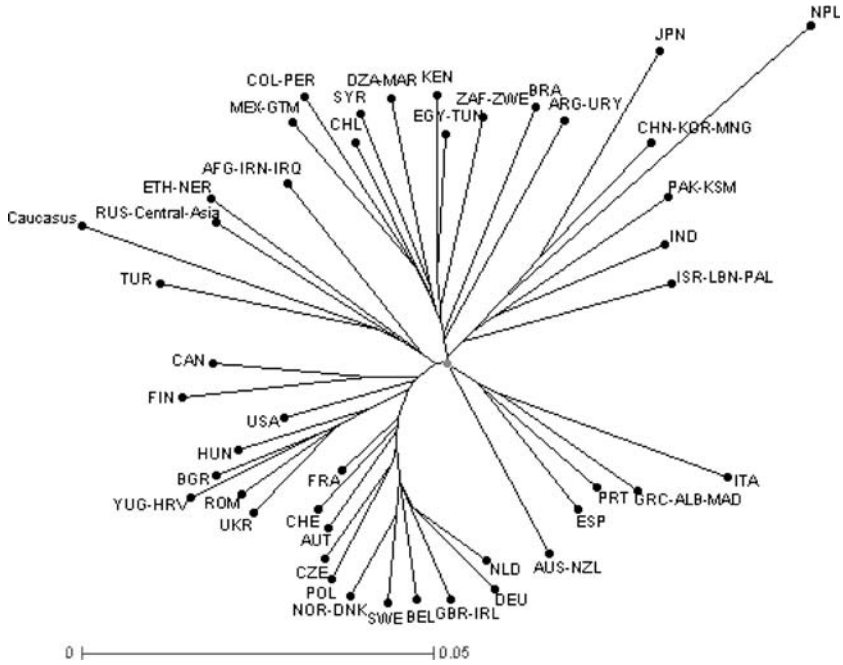

Fig. 2 Neighbour-Joining tree based on a Cavalli-Sforza and Edwards distance matrix between 45 geographical origins of bread wheat germplasm (abbreviations used are those from Table 1)

Table 4 Total number and proportion of alleles as a function of the two types of markers and the different sample sizes of the collections

Collection set of markers	Whole collection (3,942 accessions)	744 accessions	Core collection (372 accessions)
38 genomic SSRs	908	903	892
Percentage of total	100%	99.45%	98.24%
44 EST-SSRs	–	184	178
Percentage of total	–	100%	96.74%

effective for retaining widespread and low frequency neutral alleles than other sampling strategies (Bataillon et al. 1996). The M strategy is expected to perform particularly well when accessions come from populations with restricted gene flow or when they are primarily autogamous. Bread wheat is a selfing species and its germplasm diversity includes a large number of rare neutral alleles. However, core subsets can be formed to either include rare and localised (unique) alleles, maximizing the total allele diversity in the core (the M strategy favoured by geneticists), or to include a broad range of adapted accessions, maximizing the representativeness of the genetic diversity in the core (the D strategy, which is more aligned to a breeder's perspective; Marita et al. 2000). It has been demonstrated that geographical origin is a significant factor structuring genetic diversity in wheat germplasm collections and local adaptation can explain some differences between accessions (Roussel et al. 2004, 2005). To combine the objectives of both geneticists and breeders, information about SSR alleles and geographical origins were used jointly here as variables in generating core collections using MSTRAT. This method appears very powerful compared to random sampling and, by using a kernel core, the majority of the diversity at 38 polymorphic SSR loci evenly distributed throughout the genome was captured with a limited number of accessions (about 10% of the whole collection). This sampling intensity is similar to those suggested by other authors (see Franco et al. 2006) ranging from 5 to 20% of the total number of accessions and indicates quite a high level of redundancy in the entire collection.

The final 372-core collection included more than 98% of the 908 observed SSR alleles. It is particularly noteworthy that although the INRA collection is very diverse, most of the diversity can be represented by a subset of 372 accessions. All the different geographical areas of origin were represented in the core, although some were under or over-represented compared to the total sample (Table 1). This could be explained by differences in allelic richness between germplasm from different geographical areas and by the status of the accessions. For instance, Spanish accessions have significantly more alleles than Swedish or Yugoslavian accessions (Roussel et al. 2005). Similarly, the higher proportion of Nepalese accessions in the core is due to the higher level of diversity among accessions from this

country where landraces are more prevalent than cultivars; landraces are known to be richer in alleles than cultivars (Roussel et al. 2004). However, the proportions of accessions from France, Europe and the rest of the world in the total sample of 3,942 accessions were roughly maintained in the 372-core collection. Most modern cultivars currently used in genetic and genomic programs in France figure in the 372 CC (e.g. cvs. Courtot, Chinese Spring, W7984, Opata, Arche, Récital, Renan, Apache, and Ornicar). These accessions will be useful in relating information from future analyses of the core to published data. Furthermore, as the 372CC can be entirely arrayed in a 384-deep-well DNA storage plate, twelve additional accessions of particular interest (e.g. controls) may be added to the plate.

The method used to build the core was validated by using a second set of independent markers on a larger sample of accessions. The 372 supplementary accessions were chosen with the aims of (1) validating the M strategy and (2) showing the geographical components of diversity in the INRA collection. By considering both more loci and more accessions per geographical group, significant allelic frequencies could be calculated for each group and then used for distance matrix and hierarchical tree analysis. The 44 EST-SSR markers used to analyse the validation sample of 744 accessions generally showed lower diversity indices (allelic richness and Nei's index) than the 38 genomic SSRs. It has been demonstrated on the same set of accessions that differences between the two types of SSRs are highly significant (Balfourier et al. 2006), which is in agreement with similar studies (Eujayl et al. 2002; Gupta et al. 2003). This could be explained by the fact that EST-SSRs are derived from expressed sequences, which are probably more conserved than the DNA segments containing genomic SSRs. So, with the lower allelic richness and fewer rare alleles, the use of such a set of EST-SSR markers to validate the method may have diminished the effectiveness of the validation of the core.

Maximizing the diversity of a first set of markers (38 genomic SSRs) at the same time maximises unknown diversity, here expressed by a second set of makers evenly distributed throughout the genome (Table 4). We can thus imagine that this method would also capture phenotypic variation in traits that are not controlled a priori by the few genes used to sample the core collection. This hypothesis of capturing phenotypic traits by accumulating allelic diversity at neutral loci would strongly depend on the number and type of markers used. The hypothesis was tested for four agronomic traits (plant height, heading date, frost susceptibility and 1000-kernel weight) measured on the entire sample of accessions. The results indicate that the mean and the range of variation in these traits were similar to those of the core collection (data not shown). This might be explained by high levels of linkage disequilibrium in the collection, so that every allele of any gene is captured by linkage drag when maximizing allele diversity in a subset of markers. Further work is in progress to check this hypothesis by analysing the extent of linkage disequilibrium in the collection.

A complete cross-validation of the method would have required the total sample of 3,942 accessions to be tested with the second set of markers. Although such a detailed analysis was unfeasible, the fact that a set of 744 accessions captures the majority of the alleles from the first set of markers validates the method used to build the 372CC.

The Neighbour-Joining tree shows a clear separation between the lower and upper clusters (Fig. 2), which can be initially interpreted as corresponding to the European and the Asian wheat germplasm, respectively. This confirms a previous analysis of 78 wheat landraces originated from 22 countries (Strelchenko et al. 2005). Wheat species could therefore have been spread by the first farmers from the Fertile Crescent both westwards (to Europe) and eastwards (to Asia). The germplasm from European countries (lower part of Fig. 2) forms three sub-clusters (N–W Europe, S–E Europe and Mediterranean) in agreement with a previous study of European wheat (Roussel et al. 2005). The contrast between NW and SE Europe has been explained by the adaptation of the initial wheat germplasm to different climatic and environmental conditions between northern and southern areas defined by the arc formed by the Alps and the Carpathian Mountains. A second explanation is linked to the spread of agriculture from the Middle East to Western Europe: the differences between the two sub-clusters could reflect the oldest migration pathways of the Neolithic farmers bringing the initial wheat germplasm. Compared to the other European countries, it was suggested that the Mediterranean wheat germplasm forms a sub-cluster because the geography of these countries (isolation) and the specific climatic conditions (such as drought) may have played a major role in the differentiation of the germplasm (Roussel et al. 2005). Finally, the relative position of USA, Canada, and Australia–New Zealand cultivars clearly confirms the European origins of North American and Oceanic wheats, which were introduced there during the sixteenth and the nineteenth centuries, respectively.

By contrast, the upper part of Fig. 2 clusters all the other geographical areas considered in this study. One

sub-cluster was composed only of Asian wheats as, except for Ethiopia and Niger, all the geographical areas represented are in the Near and the Central East. Wheats from Far Eastern countries made up a second sub-cluster. The third sub-cluster comprised germplasm from all the South American countries and most of the African areas, which clustered with that from Syria in the Middle East. This sub-cluster might reflect recent breeding practices developed by CIMMYT in South America and ICARDA in the Middle East. In fact, in recent decades, these two international centres of agricultural research crossbred Asian wheat germplasm in order to improve newly adapted bread wheat cultivars for South America and Africa.

So, the entire hierarchical tree indicates that worldwide wheat diversity is not randomly distributed but distinctly divided into two groups, which can be explained by historical events or processes. Within these groups, diversity might be the result of the combined effects of adaptation of the initial germplasm to different environmental conditions and of specific breeding practices.

Conclusion

Our study of worldwide bread wheat diversity shows that microsatellite markers are a very effective tool both for evaluating large collections of genetic resources and building core collections. The 372CC will be a useful tool for genetic diversity studies such as SNP discovery and phenotyping agronomic traits in wheat germplasm. Initial work on SNP discovery using this core has already been published (Balfourier et al. 2006). The core could also be useful for studying the association of complex traits, although rare alleles are generally not suitable for such studies. As such, it would have been better not to include the kernel of 111 accessions with unique alleles in the core. The analysis of the population genetic structure of the core (Pritchard et al. 2002) in order to define ancestor groups and to avoid spurious associations is in progress. More markers (SSRs, DArTs) are being genotyped on the core with the aim of analysing the level of linkage disequilibrium in detail. Finally, the core is currently being evaluated in field crop design experiments to characterise different traits such as plant height, lodging, flowering time, pre-harvest sprouting, grain colour, kernel weight, and quality traits (such as protein content). Each accession of the 372CC has been multiplied and the seeds deposited in the INRA Clermont-Ferrand Cereal Crop Genetic Resource Centre and are available upon request and for collaborative projects.

Acknowledgments This research was supported by the French Ministry of Research and Technology and the Ministry of Finance (ASG programme: Cereal Genotyping and Food Quality). Part of this work was also funded by INRA in the framework of INRA–VIR cooperative programs.

References

Balfourier F, Charmet G, Prosperi JM, Goulard M, Monestiez P (1998) Comparison of different spatial strategies for sampling a core collection of natural populations of fodder crops. Genet Sel Evol 30(Suppl 1):215–235

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

Bataillon TM, David JL, Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics 144:409–417

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. (2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France)

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Franco G, Crossa J, Warburton M, Taba S (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci (in press). doi:10.2135/cropsci2005.07-0201

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber W, limensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–171

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) Mstrat: an algorithm for building germplasm core collections by maximising allelic or phenotypic richness. J Hered 92(1):93–94

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Theor Appl Genet 270:315–323

Guyomarc'h H, Sourdille P, Charmet G, Edwards KJ, Bernard M (2002) Characterisation of polymorphic markers from *T. tauschii* and transferability to the D-genome of bread wheat. Theor Appl Genet 104:1164–1172

Hao CY, Zhang XY, Wang LF, Dong YS, Shang XW, Jia JZ (2006) Genetic diversity and core collection evaluations in common wheat germplasm from the northwestern spring wheat region in China. Mol Breed 17:69–77

van Hintum ThJL (1994) Comparison of marker systems and construction of a core collection in a pedigree of European spring barley. Theor Appl Genet 89:991–997

Hu J, Zhu J, Xu MH (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies base on the genotypic values of crops. Theor Appl Genet 101:264–268

Huaman Z, Aguilar C, Ortiz R (1999) Selecting a Peruvian sweet potato core collection on the basis of morphological, ecogeographical, and disease and pest reaction data. Theor Appl Genet 98:840–844

Huang XQ, Börner A, Röder MS, Ganal MW (2002) Assessing genetic diversity of wheat (*Triticum aestivum* L.) germplasm using microsatellite markers. Theor Appl Genet 105:699–707

Li Z, Zhang H, Zeng Y, Yang Z, Shen S, Sun C, Wang X (2002) Studies on sampling schemes for establishment of core collection of rice landraces in Yunnan, China. Genet Res Crop Evol 49:67–74

Malosetti M, Abadie T (2001) Sampling strategy to develop a core collection of Uruguayan maize landraces based on morphological traits. Genet Res Crop Evol 48:381–390

Marita JM, Rodriguez JM, Nienhuis J (2000) Development of an algorithm identifying maximally diverse core collections. Genet Resour Crop Evol 47:515–526

McKhann HI, Camilleri C, Berard A, Bataillon T, David JL, Reboud X Le Corre V, Caloustian C, Gut IG, Brunel D (2004) Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J 38:193–202

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy F, Bernard M, Sourdille P (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet 109:800–805

Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) Genetic diversity of cultivated tropical plants. Enfield, Science Publishers, Montpellier, pp 43–76

Pritchard JK, Stephens M, Donnelly P (2002) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW (1998) A microsatellite map of wheat. Genetics 149:2007–2023

Roussel V (2005) Analyse de la diversité et de la structuration génétique d'une collection de blés tendres (*Triticum aestivum*) à l'aide de marqueurs agro-morphologiques, biochimiques et moléculaires, Thèse de doctorat de l'ENSAR, 127pp

Roussel V, Koenig J, Beckert M, Balfourier F (2004) Molecular diversity in French bread wheat accessions related to temporal trends and breeding programmes. Theor Appl Genet 108:920–930

Roussel V, Leisova L, Exbrayat F, Stehno Z, Balfourier F (2005) SSR allelic diversity changes in 480 European bread wheat varieties released from 1840 to 2000. Theor Appl Genet 111:162–170

Shoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

Spagnoletti PL, Qualset CO (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theor Appl Genet 87:295–304

Strelchenko P, Street K, Mitrofanova O, Mackay M, Balfourier F (2005) Genetic diversity among hexaploid wheat landraces with different geographical origins revealed by microsatellites: comparison with AFLP, and RAPD data. In: Proceedings of 4th International Cropscience Congress, Brisbane, Australia, 26 Sep–1 Oct 2004. | ISBN 1 920842 20 9 |

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

Upadhyaya HD, Bramel PJ, Ortiz R, Sing S (2002) Developing a mini core of peanut for utilisation of genetic resources. Crop Sci 42:2150–2156

Zhang LY, Bernard M, Leroy P, Feuillet C, Sourdille P (2005) High transferability of bread wheat EST-derived SSRs to other cereals. Theor Appl Genet 111:677–687