Theor Appl Genet (2002) 105:1019–1026
DOI 10.1007/s00122-002-1038-6

**G.M.M. Bredemeijer · R.J. Cooke · M.W. Ganal
R. Peeters · P. Isaac · Y. Noordijk · S. Rendell
J. Jackson · M.S. Röder · K. Wendehake · M. Dijcks
M. Amelaine · V. Wickaert · L. Bertrand · B. Vosman**

Construction and testing of a microsatellite database containing more than 500 tomato varieties

Received: 16 November 2001 / Accepted: 17 December 2001 / Published online: 7 September 2002

Abstract The aim of this study was to evaluate the suitability of sequence tagged microsatellite site (STMS) markers for varietal identification and discrimination in tomato. For this purpose, a set of 20 STMS primer pairs was used to construct a database containing the molecular description of the most common varieties (>500) of tomato grown in Europe. The database was built and tested by a consortium of five European laboratories each using a different STMS detection system. In this way, it could be demonstrated that the STMS markers and database were suitable for use in network activities where a common database is being established on a continuing basis with data from different laboratories.

Microsatellite polymorphism in tomato was found to be relatively low. The number of alleles per locus ranged from 2 to 8 with an average of 4.7 alleles per locus. Nevertheless, more than 90% of the varieties had different microsatellite profiles. A "blind testing" exercise showed that in general, identification of unknown samples (or detecting the most similar variety) with the 20 markers and the database was relatively easy for homogeneous varieties but less certain with heterogeneous varieties when using pools of 6 individuals.

Communicated by G. Wenzel

G.M.M. Bredemeijer · Y. Noordijk · B. Vosman (✉)
Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands
e-mail: B.Vosman@plant.wag-ur.nl
Tel.: +31-317-476980, Fax: +31-317-418094

R.J. Cooke · S. Rendell · J. Jackson
NIAB, Huntingdon Road, CB3 OLE, Cambridge, United Kingdom

M.W. Ganal · M.S. Röder · K. Wendehake
Institute for Plant Genetics and Crop Plant Research, Correnstrasse 3, D-06466 Gatersleben, Germany

R. Peeters · M. Dijcks
Nunhems Zaden B.V., P.O. Box 4005, 6080 AA Haelen, The Netherlands

P. Isaac · M. Amelaine · V. Wickaert · L. Bertrand
Agrogene SA, 620 rue Blaise Pascal, Z.I., 77550 Moissy Cramayel, France

Keywords Microsatellites · STMS · *Lycopersicon esculentum* · Variety identification · Variety discrimination

Introduction

For registration and granting of Plant Breeders' Rights to new varieties of agricultural and horticultural crops, morphological and physiological markers are used to determine distinctness, uniformity and stability (DUS). New varieties have to be distinct from all existing varieties 'in common knowledge' by at least one character. In addition, they have to meet established standards with respect to uniformity and stability of the characteristics used to demonstrate distinctness. The morphological characters used for registration purposes can also be used for varietal identification. However, many of the morphological descriptors are multi-genic, quantitative or continuous characters and their expression can be altered by environmental factors, making it necessary to use extensive greenhouse or field trials. Furthermore, the number of registered varieties increases over time. Because of this, it is impossible for any testing authority to check efficiently each newly submitted variety against all existing varieties in common knowledge. To help to circumvent these problems, it has been suggested that modern methods, such as the use of computerised image analysis systems and various DNA profiling techniques, should be evaluated for variety identification and related uses (see Cooke 1999). Molecular markers have many advantages for plant variety identification over the more traditionally used morphological characters because of their independence from environmental influences, their generally high level of polymorphism, their almost unlimited availability, and their greater potential for automation.

There are several approaches to DNA profiling (Cooke 1999). Especially in a crop like tomato where genetic diversity is very limited (Miller and Tanksley 1990) the molecular marker of choice must be very informa-

tive. The sequence tagged microsatellite site (STMS) approach has proven to be particularly useful for variety identification and testing in several crops, e.g. wheat (Plaschke et al. 1995), soybean (Maughan et al. 1995), rice (Luce et al. 2001), rose (Vosman et al. 2001b) and potato (Corbett et al. 2001). The utility of this technique in tomato has been reported previously by Smulders et al. (1997), Bredemeijer et al. (1998) and Areshchenkova and Ganal (1999). However, in these studies only small numbers of varieties were analysed. Even so, the polymorphism of the microsatellite markers found in these studies provides the basis of a system for the identification of tomato varieties. For example, complete differentiation between 16 varieties could be achieved by genotyping with as few as four STMS markers (Bredemeijer et al. 1998). DNA typing of additional loci would almost certainly allow more varieties to be separated.

Hence the main objective of the present study was to test STMS markers on a larger and, in a practical sense, a more realistic number of varieties to evaluate the discrimination power of this technique, and to build a database with a molecular description of most varieties of tomato that are currently grown in Europe. Genotyping of large numbers of varieties/accessions with STMS markers and establishing a database using results from one laboratory has been reported previously, e.g. for vines (Regner et al. 2001) and rice (Luce et al. 2001). On the other hand, collaborative databases have not been successfully built previously. However, for a broad use of STMS markers and databases, it is important that they are suitable for network activities in which a common database can be continually fed with data from different laboratories. Therefore, an important element of this research was the fact that the tomato microsatellite database was constructed and tested by a consortium of five European laboratories, each using a different STMS detection system. The standardisation and reproduction of genotyping techniques for tomato varieties with the STMS markers selected for this study was previously reported using a small set of varieties. The consistency of the results between the different laboratories was proven since the data were comparable between individual laboratories (Vosman et al. 2001a). In the present paper we describe the construction and characterisation of a database containing marker information for more than 500 tomato varieties.

Materials and methods

Plant DNA

Seeds of 521 tomato (*Lycopersicon esculentum*) varieties containing 26 duplicates, were obtained from 15 breeding companies: Heinz (North America), Western Seed, Enza Zaden, Agricultural University of Athens, Tézier, Esasem, Saatzucht Aschersleben, Kleinwanzlebener Saatzucht, INRA (France), Harris Moran Seed Company, Rijk Zwaan, Seminis, De Ruiter Zn, Syngenta and Nunhems.

DNA was extracted from pools of six seedlings of each variety, as described by Fulton et al. (1995) with the minor modification that the chloroform-isoamyl mixture was replaced by chloroform. For each variety, duplicate samples were analysed in different laboratories. Each of the participating laboratories analysed an agreed number of varieties.

PCR conditions

Twenty tomato microsatellite markers were selected from a previous study (Vosman et al. 2001a). These markers represented various repeat classes and at least one marker for each of the 12 chromosomes, except for chromosome 7. PCR was performed either with individual markers or in multiplex combinations of primer pairs. The forward primers were labelled by a fluorescent label, an IR label or 33P, depending on the detection system used.

The standard set of PCR conditions was as follows: reaction volume 25 µl, 0.2 µM of each primer, 0.25 mM of each dNTP, 2.5 mM of $MgCl_2$, 0.5–1.0 units of Ampli*Taq* Gold (Perkin Elmer) depending on the laboratory and approximately 10 ng of DNA. Standard cycling conditions included 45 cycles of 94 °C for 1 min, 50, 55, or 60 °C for 1 min and 72 °C for 1 min. After the 45 cycles, one cycle of 72 °C for 5 min or 10 min (in the case of pigtail primers; Brownstein et al. 1996) was added. The specific amplification conditions (annealing temperature, number of cycles and the use of pigtail primers) for each marker are listed in Table 2.

Detection of STMS polymorphisms

Various methods were used to detect the PCR products following denaturing polyacrylamide electrophoresis. At three of the five laboratories fluorescently labelled primers were used in combination with automated DNA sequencers: PRI used an ALFexpress (Pharmacia) as described by Bredemeijer et al. (1998) with some minor modifications (Gibco BRL denaturing polyacrylamide and the use of pigtail primers instead of a T4 treatment). IPK used an ALF or ALFexpress (Pharmacia) and Nunhems an ABI prism 377 sequencer according to the instructions of the manufacturer (Perkin Elmer). NIAB used IR-labelled primers in combination with a LI-COR DNA analyser 4,200 (MWG) as described previously (Corbett et al. 2001), and Agrogene used 33P labelled primers in combination with conventional sequencing gels and a Molecular Dynamics Storm 860 imager.

Gels were scored for the presence or absence (1/0) for each fragment/peak (allele) observed among all varieties at each STMS locus.

Allele calling and database establishment

The establishment of a database based on duplicate experiments and results from different laboratories required the use of a generally applicable standard for allele designation. A first set of alleles was defined by the initial 22 tomato varieties that were analysed in previous standardisation experiments (Vosman et al. 2001a). The alleles in this set were used as reference alleles during the establishment of the database by running them side by side with the analysed samples during subsequent electrophoretic runs. If in the analysed samples new alleles were identified, these alleles were placed according to their size into the preliminary database. During construction of the final database the preliminary allele codes were replaced by definitive ones and consecutively numbered.

Data analysis

The analysis of the data with respect to genetic relatedness and the ability to discriminate between the individual varieties were performed with the program NTSYS. For each locus, the genetic diversity was calculated using the formula $D = 1 - \sum Pi^2$ where Pi is the frequency of the I^{th} allele in the 521 varieties examined.

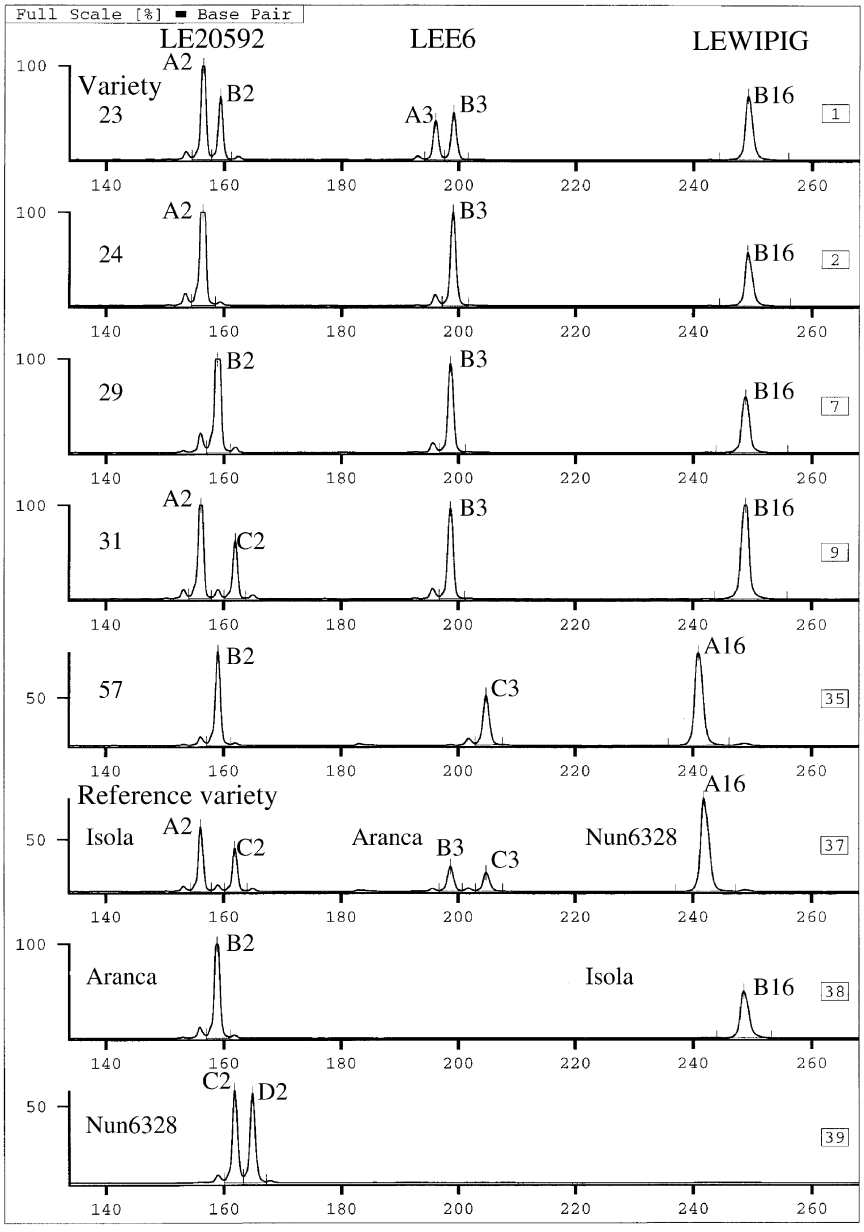

Fig. 1 Allele patterns generated by using fluorescently labelled primers in combination with an ALFexpress DNA sequencer. In this example, PCR products of three STMS markers (LE20592, LEE6 and LEWIPIG) were mixed prior to loading of the gel

Results

Construction of the STMS database

The main objective of this work was to examine the potential of STMS markers for discrimination between and identification of tomato varieties and to build a consensus database. Twenty primer pairs were used to analyse 521 tomato varieties. PCR products were analysed with various types of detection systems, including isotopic, infrared and fluorescent labelling. An example of tomato microsatellite profiles generated with fluorescent labelling is shown in Fig. 1. Data such as these are relatively easy to 'score', since each variety is analysed individually and the allele pattern can be recorded either by numbering according to the refence alleles or by size.

All datapoints were generated in at least two independent laboratories using different STMS detection techniques. For each locus an allele table was constructed containing the scoring data of the duplicate samples analysed in two laboratories along with a consensus column in which discrepancies in scoring between duplicate samples could be detected (see example in Table 1). After rechecking and repetition of all discrepancies between the different laboratories out of 10,420 datapoints, 287 discrepancies remained initially unresolved. The distribution of these discrepancies within the varieties was as follows: 361 varieties (70%) had no discrepancies, 136 varieties (25%) showed discrepancies within one or two loci and 24 varieties (5%) revealed discrepancies with more than two loci (Fig. 2).

Table 1 A part of the consensus data of the tomato marker TMS9. *H is heterogeneity (1: Heterogeneity established by testing 6 individuals. 2: Heterogeneity confirmed by testing duplicate DNA. 3: Three alleles indicate heterogeneity. 4: Discrepancies not investigated). **Final allele code ***Preliminary allele code

Var No.	Lab.	PRI					Agrogene					IPK					NIAB					Nunhems					Consensus data					Final database					H*
	Code**	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	A1	B1	C1	D1	E1	
	Code***	A1	B1	C1	D1	E1	A		C	D1	E	A1	A2	B1	C1	D1	a14		b14	c14	d14																
	Size bp	335	339	345	348	353	345		353	358	360	337	341	346	351	354						332	335	341	345	348											
1					1					1					1				1	1					1		0	0	1	5	0			1	1		1
2		1		1			1		1			1		1			1		1			1		1			5	0	5	0	0	1		1			
3				1	1				1	1				1	1				1	1				1	1		0	0	5	5	0			1	1		
4		1		1			1		1			1		1			1		1			1		1			5	0	5	0	0	1		1			
5				1		1			1		1			1		1			1		1			1		1	0	0	5	0	5			1		1	
6		1					1					1					1					1					5	0	0	0	0	1					
7					1					1					1					1					1		0	0	0	5	0				1		
8						1					1					1					1					1	0	0	0	0	5					1	
52				1	1				1	1																	0	0	2	2	0			1	1		
53						1			1		1																0	0	1	0	2			1		1	1
54				1					1																		0	0	2	0	0			1			
55		1			1		1			1																	2	0	0	2	0	1			1		
56		1			1		1			1																	2	0	0	2	0	1			1		
57					1					1																	0	0	0	2	0				1		
58		1		1			1		1																		2	0	2	0	0	1		1			
400				1										1													0	0	2	0	0			1			
402					1										1												0	0	0	2	0				1		
403				1	1									1	1												0	0	2	2	0			1	1		
404				1										1													0	0	2	0	0			1			
405					1										1												0	0	0	2	0				1		
406													1														0	1	0	0	0		1				
407				1										1		1											0	0	2	0	1			1		1	2
441														1		1			1		1						0	0	2	0	2			1		1	
442														1					1								0	0	2	0	0			1			
443														1		1			1		1						0	0	2	0	2			1		1	
444																1					1						0	0	0	0	2					1	
445															1	1			1		1						0	0	1	1	2			1	1	1	4
446														1	1				1	1							0	0	2	2	0			1	1		
447														1					1								0	0	2	0	0			1			
474														1		1								1		1	0	0	2	0	2			1		1	
475															1	1									1	1	0	0	0	2	2				1	1	
476														1	1									1	1		0	0	2	2	0			1	1		
477														1	1									1	1		0	0	2	2	0			1	1		
478																1											0	0	0	0	1					1	
479															1	1								1	1	1	0	0	1	2	2			1	1	1	3
480															1	1									1	1	0	0	0	2	2				1	1	

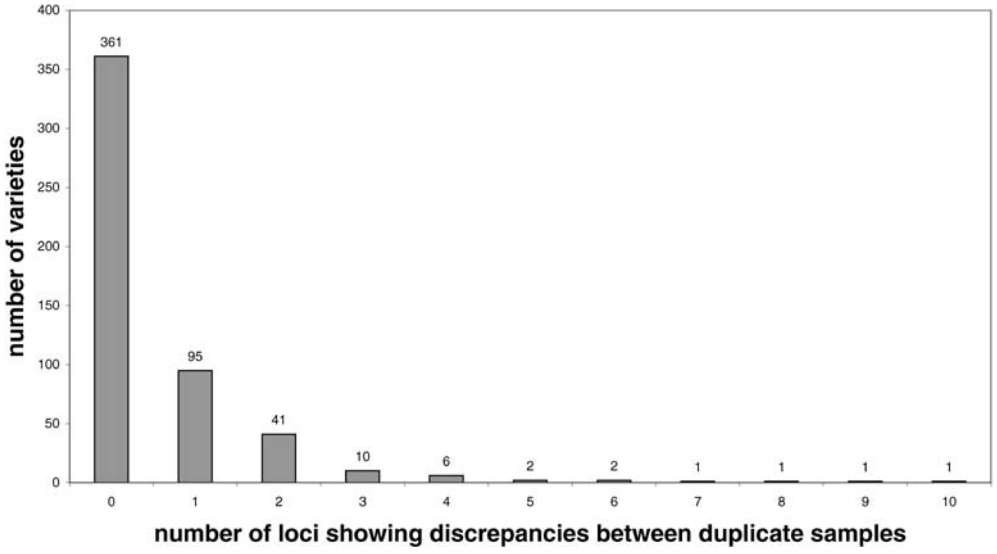

Fig. 2 Diagrammatic representation of the number of discrepancies between data from different laboratories for duplicate samples

Analysis of these discrepancies was performed in many cases by testing two duplicate DNA extractions using one detection system in one laboratory and/or by testing six individuals from a variety. The main reason for the occurrence of discrepancies between data from duplicate samples was heterogeneity of the seed sample (1, 2 or 3 in column 'Heterogeneity' in Table 1). In a few cases the discrepancies were a consequence of methodological differences (e. g. selection thresholds for allelic peaks). It is reasonable to assume that this is also true for the discrepancies that were not tested in this form (indicated by a 4 in the column 'Heterogeneity' in Table 1). Hence, the data shown in Fig. 2 were probably a reflection of heterogeneity within the seed samples.

Uniform scoring of duplicate samples in this study did not entirely exclude heterogeneity of a seed sample as it was demonstrated in a separate study. In this more detailed analysis of heterogeneity, a larger number (approximately 36) of individual seeds from ten varieties were analysed separately using six informative STMS markers (D values from 0.41 to 0.70). Seven of the ten tested varieties were uniform for all six STMS markers, while in the three other varieties, heterogeneity was detected in a few cases (data not shown). Interestingly, one of the heterogeneous varieties appeared to be uniform based on the analysis of two bulks of six individuals, indicating that low levels of heterogeneity might be present in a considerable number of varieties.

The final scoring data of all individual markers (column 'Final database' in Table 1) within the consensus data, and the results from the replicated experiments, were combined into the final database and used for further analysis.

The reliability of the database was studied by comparing the results of the duplicate varieties. Only one discrepancy was observed between scorings in the final database for two entries (No. 6 and No. 518).This difference with marker LESATTAGA, could in part be explained by a known heterogeneity in variety No. 6 and in part by potential mis-scoring.

STMS polymorphism and discrimination power

The results from the analyses with the 20 primer pairs are summarised in Table 2. The STMS markers amplified 2–8 alleles per locus with an average of 4.7 alleles per locus. In total 94 alleles were detected. As a measure of the polymorphism information content, gene diversity values were calculated for each of the tomato markers. The diversity values, which integrate allele frequency as well allele number, ranged from 0.01 to 0.70 (Table 2). The locus ATTa had a diversity value of only 0.01 revealing three alleles with a distribution of 520-2-1; the locus TMS9 with the highest D value (0.70) has five alleles with a distribution of 298-282-183-60-4. Comparison of the gene diversity values for the large collection of varieties with those previously reported for a set of 16 other varieties (Bredemeijer et al. 1998) showed that diversity values are a direct result of the choice of the varieties used (Table 2).

Although several loci had a low gene diversity value, and thus were not very informative, combining data from the different primer pairs increased considerably the discrimination capabilities of the microsatellites. The discrimination power of the database was investigated by determining the genetic similarity of all varieties using the NTSYS program. Although most varieties could be discriminated from each other based on the information obtained from the 20 markers, it was not possible to achieve 100% discrimination between all 521 varieties. It was known that 13 varieties occurred twice in the database. Of the remaining 508 potentially different varieties, 468 (=92%) had a unique molecular STMS profile. The 40 varieties that could not be identified by a unique combination of polymorphic fragments could be divided into 18 pairs and one group of four varieties. In most cases, these pairs were varieties from the same breeding company. Breeding companies with many varieties all had such pairs of indistinguishable varieties, presumably being closely related or unknown duplicates. When the data from the four least informative markers (LELEUZIP, TMS1, ATTa and

Table 2 Description of the tomato microsatellites selected for the construction of the database. Product sizes are based on fragments detected with an ALFexpress DNA sequencer (PRI)

Marker	Repeat type	PCR Ta (°C), cycles	Product sizes[i] (bp)	Chromosome	Number of alleles[f]	Diversity index[h]
1. TMS9[a]	(GATA)26 imperfect	55C-45x	337–354	12	5	0.70
2. LE20592[b]	(TAT)15-1(TGT)4	55C-45x	158–167	11	4	0.58 (0.44)
3. LEE6[c]	(GTT)28-3	55C-45x	201–207	1	4	0.37 (0.50)
4. LEMDDNa[b]	(TA)9	55C-45xpt[j]	204–221	5	7	0.61 (0.44)
5. TMS34	(GA)19	55C-45x	180–205	9	4	0.10
6. LED4[c]	(TCT)32-1	50C-45x	150–188	10	5	0.41 (0.17)
7. LED10[c]	(TCT)29-2	55C-45x	197–307	6	5	0.52 (0.40)
8. LE21085[b]	(TA)2(TAT)9-1	50C-45x	98–113	4	4	0.36 (0.34)
9. LELEUZIP[b]	(AGG)6-1TT(GAT)7	55C-45xpt	96–98	8	2	0.05 (0.12)
10. TMS1	(GT)n	60C-45x	130–132	2	6	0.05
11. ATTa[d]	(TTA)5CT(ATT)8…	50C-45x	218–221	3	3	0.01
12. LEE102[c]	(GTT)88 imperfect	55C-45x	283–307	12	5	0.55 (0.28)
13. LELE25[b]	(TA)11	50C-45x	211–217	10	4	0.36 (0.55)
14. TMS33[a]	(GA)26 imperfect	60C-45xpt	268–276	12	4	0.60
15. LED112A[c]	(GAA)32-2	50C-45x	282–328	8	6	0.42
16. LEWIPIG[b]	(CT)4(AT)4	60C-45x	255–263	9	2	0.06 (0.06)
17. LESATTAGA[b]	(TA)11(GA)11	50C-45x	167–171	?	7	0.69 (0.61)
18. JACKP1[e]	(GATA)n, (GACA)n	55C-45x	371–389	11	8	0.59
19. TMS22[a]	(GT)9(AT)8(AC)13(GA)12 imperfect	55C-45x	152–156	4	4[g]	0.38
20. LED1A[c]	(TCT)21TCCTTCC(TCT)6	50C-45xpt	145–169	10	6	0.50 (0.56)

[a] Areshchenkova and Ganal (1999)
[b] Smulders et al. (1997)
[c] STMS isolated by Arens, P. (PRI)
[d] Broun and Tanksley (1996)
[e] Phillips et al. (1994)
[f] Number of alleles found in the database varieties
[g] Alleles of the locus generating short fragments
[h] Diversity index calculated from all varieties of the database; in parentheses D values from 16 other varieties (Bredemeijer et al. 1998)
[i] Product sizes observed with an ALFexpress (PRI)
[j] pt = pigtail primer

Table 3 Blind test data from PRI

Sample code	Variety expected	Variety found	Discrepancies	Alleles of Expected Variety (database)	Alleles found in blind test	Data of six individuals
1	24	24[a] or 37[a]	TMS22	D	CD	nt
			TMS33	BC	B	nt
2	39	39[a]	TMS9	CD	C	4×C, 2×CD
			LE20592	AB	A	5×A, 1×AB
3	51	51	None	–	–	–
4	53	53[a]	TMS9	CE	E	2×CE, 4×E
5	57	57[a]	LEMDDNa	nullA	null	1×A,1×D,4×null
6	170	170	None	–	–	–
7	386	386	None	–	–	–
8	2	2 or 60[b]	None	–	–	–

[a] Match not perfect; [b] Variety 60 is a duplicate of variety 2

LEWIPIG) were not used in this analysis, 90% of the varieties were discriminable. A number of varieties that could not be discriminated by a unique combination of alleles with the 20 markers were analysed further with additional STMS primer pairs, i.e. LEEF1Aa and LEE11 (Bredemeijer et al. 1998) and LEH228. By using the additional information from LEE11 and LEH228 another four pairs could be discriminated and the group of four indistinguishable varieties was reduced to three.

Identification of varieties

A "blind test" was performed to investigate whether the markers and database could be used to identify varieties under practical conditions. At each laboratory the same eight varieties, randomly chosen from the database, were genotyped in comparison to known reference varieties representing allele ladders for all markers. Subsequently, the data were entered in the database and analysed. In most cases identification of the blind test varieties was correct and the matching with the expected variety was perfect. In a number of cases, however, there was no perfect match with any of the database varieties, i.e. the scoring for one or two markers was different. Nevertheless, this still produced genetic similarity values between the varieties in the database and the blind test samples of more than 0.95, and the most similar variety identified was the expected one. Since these differences could be due to heterogeneity in the cases of non-perfect match-

ing, six individuals were analysed. As an example, the blind test data and the results of individual testing are shown in Table 3. Comparison of the data demonstrates that the discrepancies between the scoring of the blind test varieties and expected varieties in the database can indeed be explained by heterogeneity of the seed samples. In one case, unequivocal identification was impossible as this sample clustered with two varieties in the dendrogram. Careful analysis of the fragment patterns revealed that this problem was caused by unreliable scoring of TMS22 due to poor amplification of this locus in the blind test experiment.

Discussion

In the present study, a microsatellite database containing the molecular description of the most common European varieties of tomato was constructed and tested by a consortium of five laboratories. It was shown that the database could be used in a reproducible way for variety discrimination and identification. Thus we have demonstrated that the selection of STMS markers and construction of a consensus database as carried out in this work constitute a suitable system for use in network activities in which a common, centrally held database is continually fed with data from different laboratories.

The existence of such DNA databases would facilitate the testing of new varieties of crops against all existing ones, reducing the need for individual laboratories or testing centres to maintain their own large reference collections (Donini et al. 2000). Ideally, all tomato varieties would be identifiable by a unique STMS genotype. However, it was not possible to achieve 100% discrimination between all 508 different varieties using the combined data of the 20 markers selected. Nevertheless, 92% of the varieties analysed could be identified uniquely in spite of the fact that several markers used had a very low informative value. The gene diversity values of the tomato STMS markers ranged from 0.01 to 0.70, which is low compared to values reported for several other crops (e.g. Corbett et al. 2001). Among the varieties that could not be discriminated were pairs that are morphologically different and a few pairs that had different disease resistances. This is not completely unexpected considering the low number of markers used and the fact that new varieties are selected on the basis of morphological and physiological characters which are frequently determined by single genes. It is very unlikely that the STMS markers used in this study are linked to all these characters.

It was shown that the use of data from additional STMS loci could enhance the number of discriminated varieties. On the other hand, when the four least informative markers were not included in the database, this had little effect on the discriminatory power (decreased from 92 to 90%). Therefore, if the markers with low gene diversity values were replaced by carefully selected (i.e. easy to score and reproducible) more informative markers, it could be anticipated that a very high level of discrimination could be reached. This work, along with previous reports (Vosman et al. 2001a), has clearly shown that the selection of reliable markers that can be analysed in different laboratories is vital to the construction of consensus databases.

The ability of identifying unknown samples with such a database was examined by genotyping eight randomly chosen varieties and their subsequent comparison to known reference varieties. In general, identification was easy to achieve for homogeneous varieties, but less certain for heterogeneous ones, since heterogeneity sometimes led to differences in allele scoring between different samples of the same variety. Among the 508 varieties analysed, a considerable number (30%) were heterogeneous with respect to one or more markers. In most cases where this heterogeneity was investigated by the analysis of individual plants, two or three STMS genotypes occurred among only six seeds. Probably, this intra-variety variability could be attributed to heterogeneity of one of the parental lines, as was confirmed for the breeding line Nun 6328, or to mixing of samples. In contrast, the much lower level of heterogeneity (number of off-types less than 10% of 36 individuals) observed in the detailed study on heterogeneity (data not published) may be due to residual heterozygosity, selfing or contamination. It is not entirely surprising that a number of varieties displayed degrees of non-uniformity at certain STMS loci, since the varieties have not been selected consciously for homogeneity at these loci. This is important in the context of variety identification using molecular markers more generally. Because of the occurrence of heterogeneity, it may not be practical to use a perfect match as the sole criterion for identity when using STMS markers, since in many cases it is possible that even the same variety analysed in duplicate will show slight differences. Samples showing high similarity (e.g. higher than 95%) should be analysed further using individual seedlings in order to clarify whether they are indeed the same. A duplicate analysis, possibly in another laboratory, may also be very helpful for identifying heterogeneities in comparison to scoring mistakes. Nevertheless, such a database could significantly reduce the number of varieties to be analysed in such questionable cases.

As well as being useful in its own right and serving as a tool for variety identification, the constructed database of more than 500 different European tomato varieties might be of interest for breeding companies and variety registration agencies for additional purposes such as description of the gene pool represented in the currently grown tomato varieties. To maximise the value of the current database, however, it would be necessary to include additional characters, such as disease resistance information. Furthermore, such a database only retains its value if it is permanently updated with newly released varieties.

In conclusion, we have shown that it is possible to construct databases containing molecular marker information for varieties of crop species, and that these databases can be populated with data from different laboratories. Moreover, it is valuable to apply the methodology

adopted in this work for the construction of databases. By using a consortium of laboratories using different analytical equipment, as opposed to all of the analytical work being carried out in a single place, the information in the database is arguably more useful to potential users in the future because of its general applicability. There are some important issues that arise from this methodology, such as the need for a careful selection of markers (rejecting any that cause difficulties in any laboratory, for instance), duplication of analyses in at least two laboratories, and having information about the possible heterogeneity of varieties for identification and related purposes. Last but not least, to be of any practical value, there must be information on a sufficiently large number of varieties. Given attention to these factors, it is clearly possible to construct consensus databases that have a wide range of practical applications in the variety and seeds area and for genetic resource analysis generally.

Acknowledgements The authors acknowledge the financial support of the European Commission DGXII Biotechnology grant BIO4-CT97-2377. We also thank N. van Marrewijk for providing information concerning the selection of tomato varieties and the breeding companies for providing the seeds.

References

Areshchenkova T, Ganal MW (1999) Long tomato microsatellites are predominantly associated with centromeric regions. Genome 42:536–544

Bredemeijer GMM, Arens P, Wouters D, Visser D, Vosman B (1998) The use of semi-automated microsatellite analysis for tomato cultivar identification. Theor Appl Genet 97:584–590

Broun P, Tanksley SD (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet 250:39–49

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. BioTechniques 20:1004–1010

Cooke RJ (1999) Modern methods for cultivar verification and the transgenic plant challenge. Seed Sci Technol 27:669–680

Corbett G, Lee D, Donini P, Cooke RJ (2001) Identification of potato varieties by DNA profiling. Acta Hort 546:387–390

Donini P, Cooke RJ, Reeves JC (2000) Molecular markers in variety and seed testing. In: Arencibia AD (ed) Plant genetic engineering: towards the third millennium. Elsevier Science B.V. pp 27–34

Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol Biol Rep 13:207–209

Luce C, Noyer JL, Tharreau D, Ahmadi N, Feyt H (2001) The use of microsatellite markers to examine the diversity of the genetic resources of rice (*Oryza sativa*) adapted to european conditions. Acta Hort 546:221–235

Maughan PJ, Saghai Maroof MA, Buss GR (1995) Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38:715–723

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor Appl Genet 80:437–448

Phillips WJ, Chapman CGD, Jack PL (1994) Molecular cloning and analysis of one member of a polymorphic family of GACA-hybridising DNA repeats in tomato. Theor Appl Genet 88:845–851

Plaschke J, Ganal MW, Röder MS (1995) Detection of genetic diversity in closely related bread wheat using microsatellite markers. Theor Appl Genet 91:1001–1007

Regner F, Stadlbauer A, Eisenheld C (2001) Molecular markers for genotyping grapevine and for identifying clones of traditional varieties. Acta Hort 546:331–341

Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor Appl Genet 94:264–272

Vosman B, Cooke R, Ganal M, Peeters R, Isaac P, Bredemeijer G (2001a) Standardization and application of microsatellite markers for variety identification in tomato and wheat. Acta Hort 546:307–316

Vosman B, Esselink D, Smulders R (2001b) Microsatellite markers for identification and registration of rose varieties. Document for UPOV Working group on biochemical and molecular techniques and DNA-profiling in particular (BMT-TWO/Rose/1/1)