E

BMT-TWV/Tomato/1/1
ORIGINAL: English
DATE: March 14, 2001

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

AD HOC CROP SUBGROUP ON MOLECULAR TECHNIQUES FOR TOMATO

First Session
Le Magneraud, France, March 19 to 21, 2001

CONSTRUCTION OF THE STMS DATABASE FOR TOMATO

Document prepared by experts from the Netherlands

Construction of the STMS database for tomato

Ben Vosman and Gerard Bredemeijer

Department of Biodiversity and Identity, Plant Research International, PO Box 16,

6700 AA Wageningen, The Netherlands

Introduction

To be useful for variety testing and identification, STMS marker databases have to satisfy several requirements. The most important are that the databases must contain enough information to discriminate, if possible, all varieties by means of unique marker/allele combinations and that the data generation must be sufficiently accurate so that the results can be obtained in a reproducible fashion by any laboratory experienced in marker analysis. In a previous document (BMT6/12) we have described several aspects of harmonised plant variety testing by STMS markers in tomato and wheat.

To be able to validate fully the STMS technology and to establish and demonstrate its usefulness it is essential to analyse a large number of varieties. Therefore, the goal of this project was the development of a microsatellite database for more than five hundred tomato varieties currently grown in Europe. The varieties were independently analysed with a defined set of microsatellite markers in the laboratories of at least two partners of the consortium to ensure reproducibility. The database was subsequently used to determine the discriminative power of the microsatellites, the degree of resolution between the investigated varieties, and the uniformity of the varieties as detected by the markers used.

Methods

Plant material
The seeds of the tomato varieties in the database were obtained from 15 plant breeding companies: Heinz (North America), Western Seed, Enza Zaden, Agricultural University of Athens, Tézier, Esasem, Saatzucht Aschersleben, Kleinwanzlebener Saatzucht, INRA, Harris Moran Seed Company, Rijk Zwaan, Seminis, De Ruiter Zn, Novartis and Nunhems.

A number of these companies have provided the seeds with the restriction that it is not permitted to publish with the variety names.

Duplicate varieties, as far as known, were investigated in order to test the quality of the results and the homogeneity of the seed material for these varieties.

Microsatellite markers and DNA isolation
The tomato varieties were analysed using 20 selected primers (Table 1). For each variety two bulks of 6 individuals were analysed in different laboratories (Table 2)

Construction of the final database
In total 521 tomato varieties were analysed with 20 STMS markers. Each variety was analysed by at least two laboratories as shown in Table 2 (Varieties 1-22 were analysed in a previous task of the project (see BMT6/12).

The allele data and sizes were entered in Excel spreadsheets and then communicated to PRI by each of the partners for further analysis and comparison. PRI made consensus tables

for each marker as shown in Table 3 and listed the discrepancies in scoring between partners.

The discrepancies for the first 6 loci of the list (Table 1) were analysed in detail. Discrepancies that persisted after correction of errors could be caused by heterogeneity (i.e. different seeds have different genotypes), but also by methodological problems that were not completely solved. In order to be able to explain these discrepancies, replicate experiments including exchange of DNA isolated by different partners and/or testing of individual plants were carried out. The results of these experiments were used to indicate the cause of the discrepancies between duplicate samples in the column 'Heterogeneity' of the tables with consensus data (example Table 3).

The system devised to indicate the cause of the discrepancies between duplicate samples was as follows:

1. Heterogeneity has been demonstrated by testing 6 individuals
2. Heterogeneity is likely to exist since it has been demonstrated that the discrepancy between DNA isolated by 2 partners persists when testing both samples at one laboratory with the same system. (In this case it can not be excluded that the discrepancies are caused by differences in the quality of the DNA isolated by the partners).
3. Heterogeneity is deduced from the fact that one or both partners detected three different alleles for a variety.
4. The cause of the discrepancies is not known because no replicate or other experiments have been performed. Discrepancies may be caused either by heterogeneity or by errors or methodological problems.

Data analysis

Data analysis including the evaluation of the blind test samples was performed in the Excel database and ACCESS (PRI). The analysis of the data with respect to genetic relatedness and ability to discriminate the individual varieties was performed with the program NTSYS.

Results

Allele numbers

The STMS markers amplified 2-8 alleles per locus, on average nearly 5 alleles per marker. In total 95 alleles were detected (Table 1). Only 2 markers had null alleles.

Discrimination power

From 13 varieties it was known that they occurred twice in the database. Of the remaining 508 potentially different varieties, 468 had a unique molecular profile. Therefore, 92% of the varieties analysed could be identified uniquely. The 40 varieties that could not be identified by a 'unique'combination of polymorphic bands/peaks (perfect match of alleles) could be divided in 18 pairs and one group of four varieties. In most of the cases, these pairs are varieties from one particular breeding company. The larger breeding companies all had a few of these pairs, which presumably represent closely related varieties. Among these varieties are a few pairs that are morphologically different and a few pairs that have different disease resistances. All together, this is not a completely unexpected result since the plant material was only selected for morphological and physiological characters and it is very unlikely that the STMS markers used in this project are linked to all these characters.

A number of those varieties that could not be discriminated by a 'unique' combination of polymorphic peaks with the 20 markers, were analysed at PRI with two additional STMS

markers i.e. LEEF1Aa and LEE11. These two markers , although 'highly' polymorphic (7 and 8 alleles respectively in 16 varieties (Bredemeijer et al 1998)), were not included in the current database because of allele calling problems between the partners. However, when using a single detection system, in this case the ALFexpress, in one laboratory, few problems were caused by these markers. By using the additional information from LEE11 the varieties of two of the four pairs tested could be discriminated and the group of 4 varieties was reduced to a group of three varieties. The analysis of these varieties with LEEF1Aa did not discriminate these any varieties further.

Reliability of the database

The reliability of the database was studied by analysing the results from the duplicate samples. For 13 varieties it was known that they occur twice in the database. The results from the duplicate samples agreed very well with each other. In only one case were two identical varieties classified differently. One discrepancy was observed between scoring in the final database for LESATTAGA (cv 6: DFG, cv 518: F). The difference can in part be explained by known heterogeneity in varieties No. 6. For the other part (allele D) it is probably misscoring.

Heterogeneity in tomato varieties

For the first 6 markers of the list most discrepancies between duplicate samples could be resolved after performing replicate experiments. For the remaining markers, however, only some data entry mistakes were corrected and in a few cases replicate experiments were carried out.

For each locus a table was constructed containing the scoring data from the partners along with a consensus column. In the column 'final database' the scoring deduced from the consensus column is presented (Table 3). The scoring data of all individual markers were combined in the final database. This database was used for further analysis. Taking all of the data together, 70% of the varieties was uniform, and 25% showed heterogeneity with one or two markers (Fig. 1). Only 5% (24) of the varieties was polymorphic for more than 2 loci. There was a significant ($P<0.01$) correlation (X^2-test) between the level of heterogeneity detected for a locus and the number of alleles at that locus (compare Fig. 2 and Table 1). An exception was TMS33. This may indicate scoring problems with this marker.

In a more detailed analysis of homogeneity, a large number (approx. 36) of seeds for 10 varieties were analysed seperately. The selection of STMS markers was made on the basis of previous knowledge on the ease of scoring, polymorphism and chromosomal location (Table 4). These 10 varieties had been found homogeneous in the bulked sample experiment described above. Seven of the 10 varieties tested were uniform for all 6 STMS markers, whilst in the 3 other varieties, individuals were found which were not uniform (Table 5). This shows that in some varieties a low level of heterogeneity occurs which cannot be detected in the bulk experiments. The results also show that STMS markers may be used to detect differences in uniformity among tomato varieties.

Conclusions

A goal was to construct a database with the alleles of 20 STMS markers for about 500 tomato varieties and to demonstrate that this database may be used for the identification of varieties. The results clearly show that such a database can indeed be constructed. It also highlighted the difficulties in achieving this objective. There were a number of disagreements between partners in the first versions of the consensus database, due to such factors as data entry errors, technical problems in allele calling (see BMT6/12) and sample heterogeneity.

It would be very useful if every variety that has been utilised commercially in Europe in the last ten years were identifiable by a unique STMS genotype. With the 20 STMS markers chosen more than 90% of the tomato varieties (481 of a total of 521, 92%) could be discriminated by a unique STMS genotype. Even when the data of the four least informative markers (LELEUZIP, TMS1, ATTa, LEWIPIG) were not used, it was still 91%. Adding a few highly polymorphic markers is very effective. This was illustrated by the fact that by using one additional locus (LEE11) more varieties could be discriminated. This result indicates that when carefully selected STMS markers are used, a very high level of discrimination can be reached.

A number of points require further considerations regarding the establishment and general use of the tomato database for variety identification. First, when possible, the analysis of PCR products should be performed on an automated DNA sequencer. This provides a higher level of precision than on conventional gels. Second, there is a structural problem with tomato varieties regarding homogeneity as heterogeneity was probably the main factor that lead to differences between scoring of duplicate tomato samples analysed at 2 laboratories.The level of heterogeneity in the tomato varieties was estimated on approx. 30%.The problem of heterogeneity has been studied in detail for only a few markers, by performing replicate experiments, by exchanging DNA and by testing individuals. A solution would be to accept a low level of heterogeneity when using molecular markers.

Acknowledgements
We wish to thank all other participants of the 'Molecular Identification' project Robert Cooke[2], Martin Ganal[3], Roger Peeters[4], Peter Isaac[5] , Marion Röder[3], Jenny Jackson[2], Sarah Rendell[2], Marcel Dijcks[4] , Yolanda Kleyn [1], Dirk Visser[1] , Katja Wendehake[3], Tatyana Areshchenkova[3], Victor Korzun[3], Marc Amelaine[5] and Vincent Wickaert [5] [[*1]Plant Research International, [2]NIAB, [3]Institute for Plant Genetics and Crop Plant Research, [4]Nunhems Zaden b.v., [5]Agrogene SA. The authors acknowledge the financial support of the European Commission DGXII Biotechnology grant BIO4-CT97-2377. We also thank the breeding companies for providing the tomato seeds.

Table 1 Characteristics of the tomato microsatellites selected for the construction of the database.

Marker	Repeat type	product sizes (bp)	Chrom. location	Number of Alleles[f]
TMS9[a]	(GATA)26 imperfect	337-354	12	5
LE20592[b]	(TAT)15-1(TGT)4	158-167	11	4
LEE6[c]	(GTT)28-3	201-207	1	4
LEMDDNa[b]	(TA)9	204-221	5	7
TMS34	(GA)19	180-205	9	4
LED4[c]	(TCT)32-1	150-188	10	5
LED10[c]	(TCT)29-2	197-307	6	5
LE21085[b]	(TA)2(TAT)9-1	98-113	4	4
LELEUZIP[b]	(AGG)6-1TT(GAT)7	96-98	8	2
TMS1	(GT)n	130-132	2	6
ATTa[d]	(TTA)5CT(ATT)8...	218-221	3	3
LEE102[c]	(GTT)88 imperfect	283-307	12	5
LELE25[b]	(TA)11	211-217	10	4
TMS33[a]	(GA)26 imperfect	268-276	12	4
LED112A[c]	(GAA)32-2	282-328	8	6
LEWIPIG[b]	(CT)4(AT)4	255-263	9	2
LESATTAGA[b]	(TA)11(GA)11	167-171	?	7
JACKP1[e]	(GATA)n,(GACA)n	371-389	11	8
TMS22[a]	(GT)9(AT)8(AC)13(GA)12 imperfect	152-156	4	4[g]
LED1A[c]	(TCT)21TCCTTCC(TCT)6	145-169	10	6

[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 523 varieties
[g] Alleles of the locus generating short fragments

Table 2: Varieties tested by the partners and detection systems used.

Partner	Detection system	No's of varieties tested
PRI	ALFexpress	23----------422
Agrogene	33P phosphoimager	23----------322
IPK	ALF	323----------524
NIAB	LI-COR	423----472
Nunhems	ABI	473-----524

Table 4: Tomato STMS markers used for uniformity study

Locus	Mix No	Chromosome	No. of alleles	Size range (bp)
LED4	1	10	6	150-188
LED10	1	6	5	197-307
TMS9	1	12	5	337-354
LE20592	2	11	4	158-167
LEMDDNa	2	5	8	204-221
LED112A	2	8	6	282-328

Table 5: STMS alleles in individual plants of tomato varieties

Variety	LED4	LED10	TMS9	LE20592	LEMDDNa	LED112A
A	36xD	36xA	36xCE	36xB	36xBD	36xD
B	36xD	36xAE	36xCE	36xB	36xBD	36xD
C	36xD	36xA	36xCE	36xAB	36xBD	36xD
D	35xCD 1xBD(6)	36xAE	36xE	36xB	36xB	34xCD 1xBD(6) 1xD(36)
E	36xD	36xA	36xDE	36xAB	36xD	36xD
F	36xD	36xA	35xCE 1xC(23)	36xB	35xBD 1xD(23)	36xD
G	30xD	30xA	27xA 1xAE(9) 2xE (20,30)	30xB	30xB	30xD
H	36xCD	36xDE	36xAC	36xB	36xA	36xCD
I	36xCD	36xA	36xCD	36xC	36xA	36xCD
J	36xBC	36xA	36xD	36xB	36xA	36xBC

Fig. 1: Number of loci showing heterogeneity in the 521 varieties tested

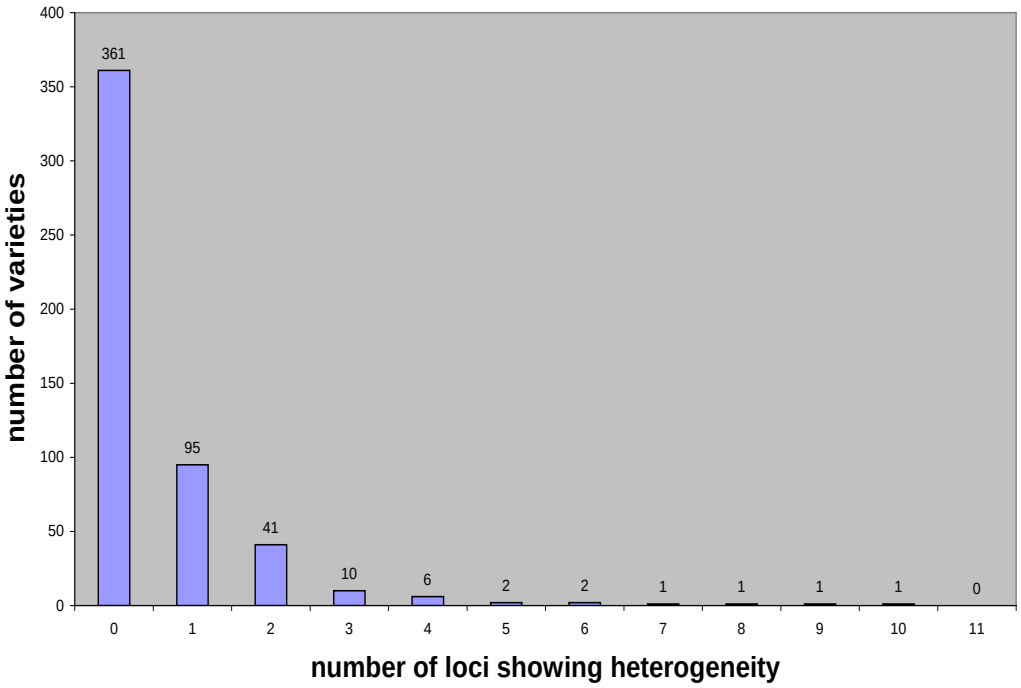

Fig.2: Number of heterogeneous varieties per microsatellite

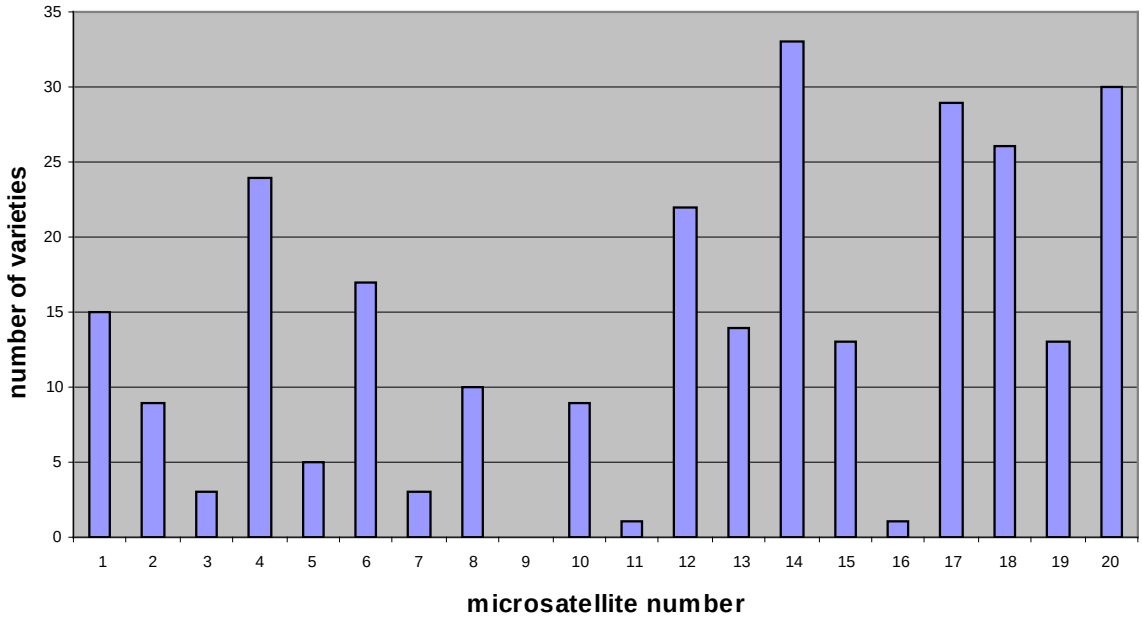

Table 3: Example of a database sheet

Variety	Variety name		TMS9 CPRO					TMS9 Agrogene					TMS9			IPK		TMS9 NIAB					TMS9 Nunhems					consensus					final database TMS9					hetero-	remarks
		final 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	geneity*	
number		inst. 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	Ailsa Craig					1					1					1				1	1					1		0	0	1	5	0			1	1		1	heterogeneity: CPRO tested 6 ind.: 1x C, 1x CD, 4x D
2	Aranca		1		1			1		1			1		1			1		1			1		1			5	0	5	0	0	1		1			
3	Aromata				1	1				1	1				1	1				1	1				1	1		0	0	5	5	0			1	1		
4	Durinta		1		1			1		1			1		1			1		1			1		1			5	0	5	0	0	1		1			
5	Isola				1		1			1		1			1		1			1		1			1		1	0	0	5	0	5			1		1	
6	Nun6328		1					1					1					1					1					5	0	0	0	0	1					
7	Trend					1					1					1					1					1		0	0	0	5	0				1		
8	VFNT Cherry						1					1					1					1					1	0	0	0	0	5					1	
23		1			1		1			1		1																0	0	2	0	2			1		1	
24		1			1					1																		0	0	2	0	0			1			Agrogene tested 6 ind. All C
25		1			1					1																		0	0	2	0	0			1			
26		1				1				1	1																	0	0	###	2	0			1	1		1	Agrogene tested 6 ind.: 3 x CD, 3 x C
27		1					1					1																0	0	0	0	2					1	
28		1			1					1																		0	0	2	0	0			1			
29		1			1					1																		0	0	2	0	0			1			
30		1			1	1				1	1																	0	0	2	2	0			1	1		
399		14			1		1								1		1											0	0	2	0	2			1		1		CPRO tested IPK.DNA:C is 13% of E; CPRO.DNA: C is 45% of E
400		14			1										1													0	0	2	0	0			1			
402		14				1										1												0	0	0	2	0				1		
403		14			1	1									1	1												0	0	2	2	0			1	1		
404		14			1										1													0	0	2	0	0			1			
405		14				1										1												0	0	0	2	0				1		
406		14												1														0	1	0	0	0		1					CPRO: no germination; rely on one dataset
407		14			1										1		1											0	0	2	0	1			1		1	2	CPRO.DNA: E is 13-16% of C(lower than 20%) IPK.DNA: E is 28-34% of C
440		15													1	1				1	1							0	0	2	2	0			1	1		
441		15													1		1			1		1						0	0	2	0	2			1		1	
442		15													1					1								0	0	2	0	0			1			
443		15													1		1			1		1						0	0	2	0	2			1		1	
444		15														1					1							0	0	0	0	2					1	
445		15														1	1			1		1						0	0	1	1	2			1	1	1	4	
446		15													1	1				1	1							0	0	2	2	0			1	1		
447		15													1					1								0	0	2	0	0			1			
473		15															1										1	0	0	0	0	2					1	
474		15													1		1								1		1	0	0	2	0	2			1		1	
475		15														1	1									1	1	0	0	0	2	2				1	1	
476		15													1	1									1	1		0	0	2	2	0			1	1		
477		15													1	1									1	1		0	0	2	2	0			1	1		
478		15															1											0	0	0	0	1					1		Nunhems no germination; rely on one dataset
479		15														1	1								1	1	1	0	0	1	2	2			1	1	1	3	Nunhems sample heterogeneous. IPK cannot exclude C1 as weak allele
480		15														1	1									1	1	0	0	0	2	2				1	1		

[End of document]