E

BMT/9/12
ORIGINAL: English
DATE: June 15, 2005

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA PROFILING IN PARTICULAR

Ninth Session
Washington, D.C., June 21 to 23, 2005

ANALYSIS OF A DATABASE OF DNA PROFILES OF 734 HYBRID TEA ROSE (*ROSA HYBRIDA*) VARIETIES

Document prepared by experts from the Netherlands

ANALYSIS OF A DATABASE OF DNA PROFILES OF 734 HYBRID TEA ROSE (*ROSA HYBRIDA*) VARIETIES

M.J.M. Smulders, D. Esselink, R. E. Voorrips & B. Vosman
Plant Research International, P.O. Box 16, NL-6700 AA Wageningen, The Netherlands
Contact: ben.vosman@wur.nl

Introduction

1. Rose is the largest ornamental crop. Over 25,000 varieties of modern roses have been described (Cairns, 2000). The first hybrid tea rose was introduced in 1867 and since then more than 10,000 hybrid teas have entered the market. Such large numbers of varieties may cause problems in the DUS testing context. A major problem for all countries carrying out DUS tests is the requirement to compare new varieties to all other varieties in common knowledge. Clearly, strict adherence to this concept is logistically and financially impossible in a species such as rose, which is cultivated around the world. Thus DUS testing stations tend to take a somewhat pragmatic view of common knowledge, limiting it, for instance, to varieties that can be grown in similar climatic zones. Nevertheless, this still means that many hundreds of varieties may have to be taken into account for roses.

2. Another problem is the reference collections. At the moment, rose DUS testing on behalf of the CPVO is carried out by UK and Germany (outdoor types) and the Netherlands (greenhouse, cut flower types). The examination office for greenhouse roses does not currently hold a living reference collection – mainly because of the high costs associated with maintaining such a collection, and disease problems. Therefore, the examination office needs to request reference varieties from the breeders of the candidate varieties. It is important that the examination office can quickly verify the identity of the material submitted. For this aspect of quality assurance, molecular markers are ideally suited, as they are highly discriminating and can be assayed rapidly and relatively cheaply.

3. Several first generation molecular marker techniques have been applied to roses (Esselink et al 2003). All these marker systems have some drawbacks for variety identification and related activities. In some, there is a lack of high levels of polymorphism, whilst other methods are difficult to reproduce, laborious and/or provide complex patterns inconvenient for database building (Vosman 1998). In contrast, DNA microsatellites (simple sequence repeats, SSRs) are highly polymorphic and have the advantage of providing a co-dominant marker system based on a PCR technology. When analyzed as sequenced-tagged microsatellite site (STMS) markers, they provide simple banding patterns that are easy to record and are especially suitable for automated and objective analysis. In addition, the resultant data can be readily stored in a database. New varieties or new markers can be easily added to an existing database.

4. The application of the STMS approach was recently successfully demonstrated in roses by Esselink et al. (2003) and for the construction of databases in collaborative studies for tomato (Vosman et al. 2001, Bredemeijer et al. 2002) and wheat (Röder et al. 2002). The tomato and wheat studies aimed at the construction of central databases that were populated with data produced by different laboratories. They were designed in such a way that they were independent of the technology (equipment) used for detection of the polymorphisms.

The study of Esselink et al. (2003) was extended to the actual generation of a database containing the molecular profiles of as many varieties of rose as possible. The database now contains 734 entries of Hybrid tea varieties, including all new varieties of the last five years. Since for the first time a database of this size has been established, we set out to analyse the molecular data in detail. Specifically, we have looked at discriminative power of the markers, reproducibility of the results, genetic (sub) structure in the set of varieties analyzed, as well as correlation between molecular and DUS characteristics (option 2 approach).

Material and methods

5. Rose varieties included in the database were based on the list of applications for PBR from the years 1997-2004, plus additional varieties for comparison. DNA was extracted from frozen young leaves using the Qiagen DNA extraction kit. Rose microsatellites were analysed as described by Esselink et al. (2003). Results of the following 11 markers were used: RhE2b, RhAB15, RhAB22, RhD201, RhD221, RhAB40, RhB303, RhM405, RhEO506, RhO517, and RhP519.

6. In the early years (2000-2002), the 11 microsatellite loci were amplified in separate PCR reactions in 96 wells microtiter plates, then combined and analysed in four runs on an ABI 3700. For the varieties from 2003 onwards, the amplifications were done in multiplex format, so that the total number of handlingsteps was greatly reduced. Hence, for error estimation we counted the number of different scores in identical genotypes (which consisted of differences in duplicate varieties included as references, in members of mutant groups, and in replicated samples due to bad amplification) separately for 2000-2002 and 2003-2004. Each year, also 18 extra varieties were included in the analyses, consisting of reference and standard varieties. For reference varieties always the same DNA extraction was used, but standard samples were analysed in duplo from independent DNA extractions.

Data analysis

7. Population genetic analysis is not straightforward for polyploid species, since most programs cannot handle more than two alleles per locus. One approach is to take the presence or absence of each allele as a dominant marker, as an ‘allelic phenotype’ (Esselink et al 2003). We used the dominant scores to calculate a Jaccard genetic distance with Genstat. As an alternative, we also applied SPAGeDi 1.2 (Hardy and Vekemans 2002), which can handle plants of various ploidy levels. SPAGeDi was also used to calculate the genetic differentiation (Fst) across years and across breeding companies.

8. Overall morphological Euclidean distances were calculated based on 44 DUS trait scores (UPOV guidelines for Rose) without any transformation or normalization, using the Genstat FSIMILARITY command with TEST=euclidean. The presence or absence of structure among the morphological or genetic distances was assessed using a PCO analysis. Further morphological distances were calculated based only on trait 11730.1 (flower color) using simple matching: 0 if the two color scores were equal, 1 if they were different. The correlation between genetic and morphological distances was assessed using both morphological distance measures separately. The association between pairwise genetic distances and the pairwise differences in morphological scores was tested by randomization (Mantel test), using 1000 permutations.

Results

Discriminative power of the markers

9. The loci amplified between 4 (RhM405) and 9 (RhAB40) different alleles. Figure 1 displays the occurrence of alleles across genetically different genotypes (i.e., all unique genotypes plus one representative each of mutant and duplicate groups). In terms of allelic phenotypes, the power of discrimination was even higher. Table 1 compares number of alleles and number of allelic phenotypes among the varieties. The observed number of allelic phenotypes can be as much as 8 times that of the number of alleles. Some allelic phenotypes were abundant in the set of varieties, up to 37% of the varieties containing the same allelic phenotype for marker RhM405. Some abundant allelic phenotypes were homozygous (RhE2b and RhAB22), one was completely heterozygous (RhM405), but this was generally consistent with what would be expected under independent inheritance of these alleles. The PIC values were high (between 0.52 and 0.77), indicating that the markers can easily distinguish all varieties.

Reliability of the database

10. To achieve a full set of molecular data, all samples with problematic data were repeated. Problems encountered included lanes with no signal (no sizer or unsuccessful PCR), signal too low or too high, and unsure patterns (bad separation, wrong sizing of the internal sizer). The sample repeat rate amounted to 18% for 2000-2001 and 15% for 2002 samples. For the 2003 set of varieties onwards, the markers were analysed in multiplex PCR, which greatly reduces the number of pipetting steps. Probably as a result hereof, the repeat rate was only 3.9% for 2003 and 3% for 2004.

11. The reliability was evaluated by looking at the error rate in the scores of the duplicate samples and in those of samples that turned out to belong to a mutant group. The error rate for 2003 samples was 0.26% of the loci (errors between mutants), for 2004 samples it was 0.30% of the loci (in duplo samples). Since the average number of alleles per variety across loci was 2.47, this translates into an error rate of about 1 in 1000 for any allele score in the database.

Genetic structure within the set of varieties

12. To exclude the possibility that the set of varieties we analyzed contains a substructure that needs to be taken into account when making the comparisons between molecular and morphological data we first checked for this using population-genetic analysis tools. We took from the database those varieties that had been submitted for PBR in the period 2000-2004 (2000: 42 varieties; 2001: 69; 2002: 98; 2003: 120; and 2004: 78). In total, 420 varieties were included in the database for these 5 years. Among these there were 407 different genotypes. 13 additional varieties belonged to 12 groups of mutants (consisting of 2-3 identical genotypes each).

13. The genetic differentiation among years for this set of 407 varieties was estimated at Fst=0.0007 +/- 0.0005, indicating that every year a similar set of varieties is submitted for PBR.

14. Across the years, we analyzed the differentiation among breeders. There were 45 different breeders, but 12 were present with only one variety, and others with only a few

varieties. To optimize the sensitivity of the analysis, the varieties from breeders with less than 5 varieties and the group of unknown breeders was removed. The resulting 299 varieties were grouped together in 17 breeding companies. Among these, Fst= 0.0056 +/- 0.0011. Apparently, also these 17 companies use basically the same gene pool, although the allele frequencies differ slightly among the companies (examples of two of the markers in Fig 2).
Not surprisingly, a PCO analysis of the main variation among the molecular data does not show any obvious structure (result not shown).

Correlation with DUS characteristics

15. A PCO plot of the separate DUS characteristics does show three groups of varieties on the first axis (22% explained variation), but not for the second axis. The PCO analysis indicated that the distinction in three groups is based mainly on the scores of two of the flower color-related traits: 11732 (spot on the inside) and 11737 (spot on the outside). Both traits have a large effect since they are either 1 or 9. This indicates the problem when trying to combine various measures on different scales. Two strategies can be followed: an aggregate measure combining all data for the standard set of DUS characteristics, or a focus on one or a few most important traits only.

16. Using an aggregate morphological distance, we correlated the pairwise genetic distance to the pairwise DUS morphological distance. This produces a large group of samples (Fig. 3). Clearly separate are the mutant pairs, which are genetically identical. There is a large gap in genetic similarities between mutants and seed-derived varieties, since the latter have a genetic similarity that is always less than 0.90 (see Figure 3). There is no obvious relationship between pairwise genetic and aggregate morphological similarities, although this might be influenced by the way the DUS characteristics were treated.

17. The most important distinguishing trait in cut rose is flower color. Therefore, we focused on the flower color, which is scored in color classes (1-19, 34, 40, 46-47, 50; UPOV colour grouping according to the RHS Colour Chart (2001)). In particular we considered the question whether a higher genetic similarity between two varieties increases the probability that these varieties are in the same color group. Table 2 shows the frequency of color matches in variety pairs at different levels of genetic similarity. The number of matches in the classes above 0.7 genetic similarity (**bold** in Table 2) is significantly higher than the background level of correct matches that is generated by chance alone (Mantel test, $p<0.001$). This may be due to the fact that these variety pairs have a common ancestry. Alternatively, some colors may occur only in a specific genetic background.

18. However, predicting the color of a variety based on its genetic similarity with another variety is not reliable. Even at a genetic similarity above 0.7, only 17% of the variety pairs have matching color, which is hardly useful although higher than the overall frequency of matches (8%). Further, the number of pairs with a similarity above 0.7 is very small: only 0.8% of all pairs.

Conclusions

19. The microsatellite markers used show a high discriminative power. All seedling varieties can be distinguished and in pairwise comparisons the genetic similarities between the pairs of varieties are always lower than 0.9 using the Jaccard index. Original varieties and

mutants thereof show a genetic similarity of 1. This indicates that it is no problem to differentiate between mutants and seedling varieties.

20. Reliability of the data stored in the database is high, with an error rate of about 1 in 1000.

21. Our data clearly indicate that when fewer steps are needed (in our case by combining reactions into multiplexes) the number of samples that needs to be repeated is lower.

22. Finally it appears that correlation between genetic similarities based on morphological characters and molecular characters is absent. Only for high genetic similarities (above 0.7) there is some correlation but this only refers to a small number of varieties. From this we conclude that with the 11 microsatellite markers used and the way the DUS characteristics were treated, an option 2 approach is not realistic for rose.

Acknowledgements

23. We thank Yolanda Noordijk for skillfully handling DNA extraction and PCR reactions, and Gerrit van de Wardt for help with analysis of the DUS characteristics. This research was funded under the method development scheme of CGN.

References

Bredemeijer GMM, Cooke RJ, Ganal MW, Peeters R, Isaac P, Noordijk Y, Rendell S, Jackson J, Röder MS, Wendehake K, Dijcks M, Amelaine M, Wickaert V, Bertrand L, Vosman B (2002) Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theoretical and Applied Genetics 105: 1019-1026

Cairns T (ed) (2000) Modern roses XI. Academic Press, London

Esselink D, Smulders MJM, Vosman B (2003) Identification of cut-rose (Rosa hybrida) and rootstock varieties using robust Sequence Tagged Microsatellite markers. Theoretical and Applied Genetics 106: 277-286

Hardy OJ, X Vekemans (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620

Röder MS, Wendehake K, Korzun V, Bredemeijer G, Laborie D, Bertrand L, Isaac P, Rendell S, Jackson J, Cooke RJ, Vosman B, Ganal MW (2002) Construction and analysis of a microsatellite-based database of European wheat varieties. Theoretical and Applied Genetics 106: 67-73

Vosman B, D Esselink, R Smulders (2001) Microsatellite markers for identification and registration of rose varieties. UPOV document BMT-TWO/Rose/1/1

Vosman, B. (1998) The use of molecular markers for the identification of tomato cultivars. In: Molecular tools for screening Biodiversity: In: Molecular tools for screening Biodiversity: Plants and Animals. eds: Karp, A., Isaac, P. G Ingram D.S. Publishers: Chapman and Hall, pages 283-287.

Table 1. Power of discrimination of the markers used, in a set of 407 different varieties.

Locus	Number of alleles	Number of allelic phenotypes	PIC value based on allelic phenotypes	Frequency of most common allelic phenotype	Number of different alleles in allelic phenotype with highest frequency
RhAB15	6	28	0.72	0.29	2
RhAB201	4	15	0.67	0.23	2
RhAB22	7	23	0.52	0.31	2
RhAB40	9	79	0.76	0.19	2
RhB303	6	37	0.76	0.12	3
RhD221	6	32	0.67	0.31	2
RhE2b	7	32	0.54	0.37	1
RhEO506	6	34	0.72	0.20	2
RhM405	4	9	0.73	0.4	4
RhO517	5	27	0.77	0.12	3
RhP519	6	32	0.71	0.22	3

Table 2: Correlation between genetic similarity and identity of flower color class for any pair of varieties from the period 2000-2004 (mutants with similarity 1.00 excluded). The number of correct assignments in the classes above 0.7 genetic similarity (**bold**) is significantly higher than the background level of correct matches by chance (Mantel test, $p<0.001$). However, it only represents (last column) 1.5% of all pairs, or 1% above the level that is obtained by change (which is 48).

Genetic similarity (Jaccard) above	Total number of pairs of varieties	number of matches (same color class)	% matches in the same similarity class
0.90	0	0	
0.85	4	**0**	0
0.80	16	**4**	25
0.75	98	**17**	17
0.70	504	**82**	16
0.65	1957	216	11
0.60	4805	484	10
0.55	10899	921	8
0.50	14609	1181	8
0.45	21951	1626	7
0.40	14062	944	7
0.35	9334	615	7
0.30	3272	208	6
0.25	929	72	8
0.20	174	7	4
0.15	7	0	0
0.10	0	0	
0.05	0	0	
0	0	0	
total	82621	6377	

Figure 1. Allele occurrence per marker (alleles sorted in order of decreasing occurrence) in all genetically different varieties. Y-as should read 0 tot 70% in stead of 0 tot 0.7

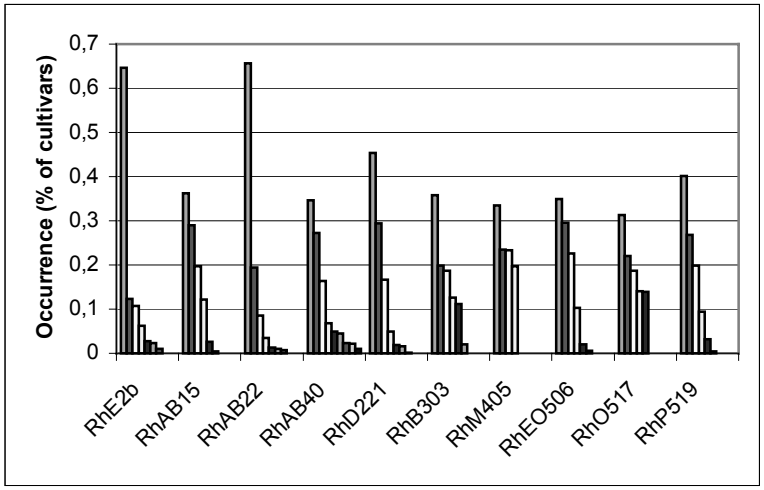
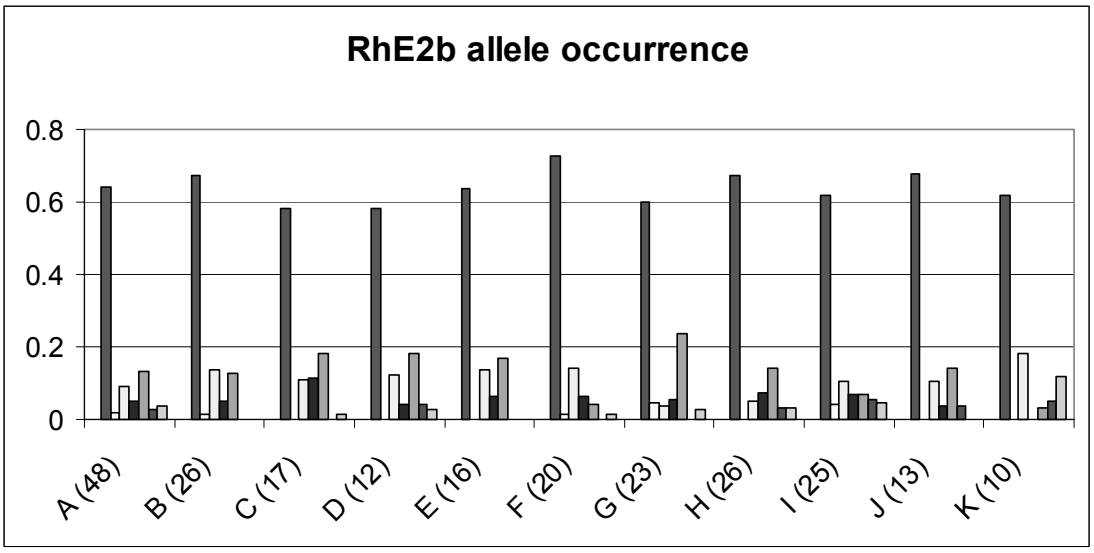
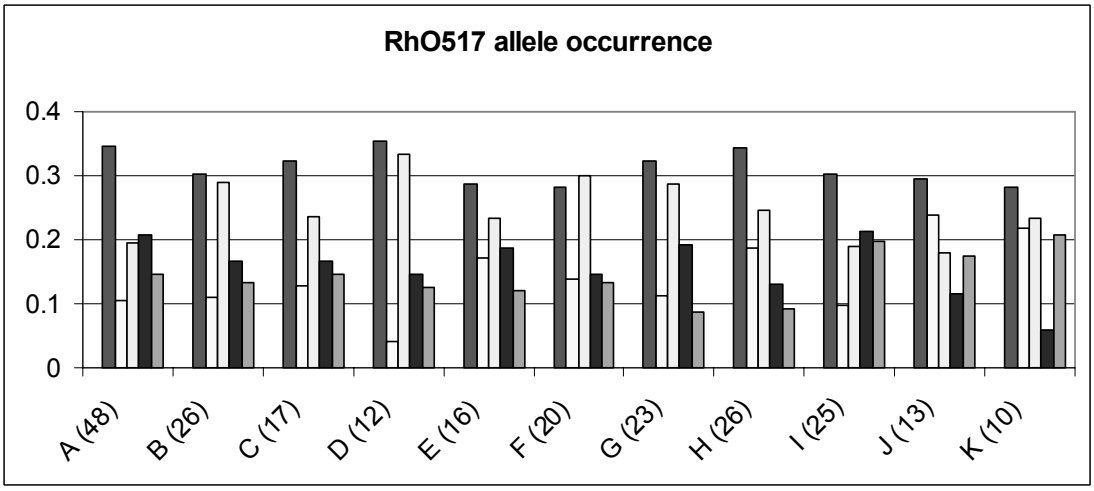


Figure 2. Differences in allele occurrence among 11 companies (coded A-K) for (a) RhE2b, an example of a locus with very large difference in allele occurrence, and (b) RhO517, an example of a locus with a relatively even distribution of allele occurrence. Between brackets the number of varieties included for each company. Y-as should read 0 tot 80% in stead of 0 tot 0.8

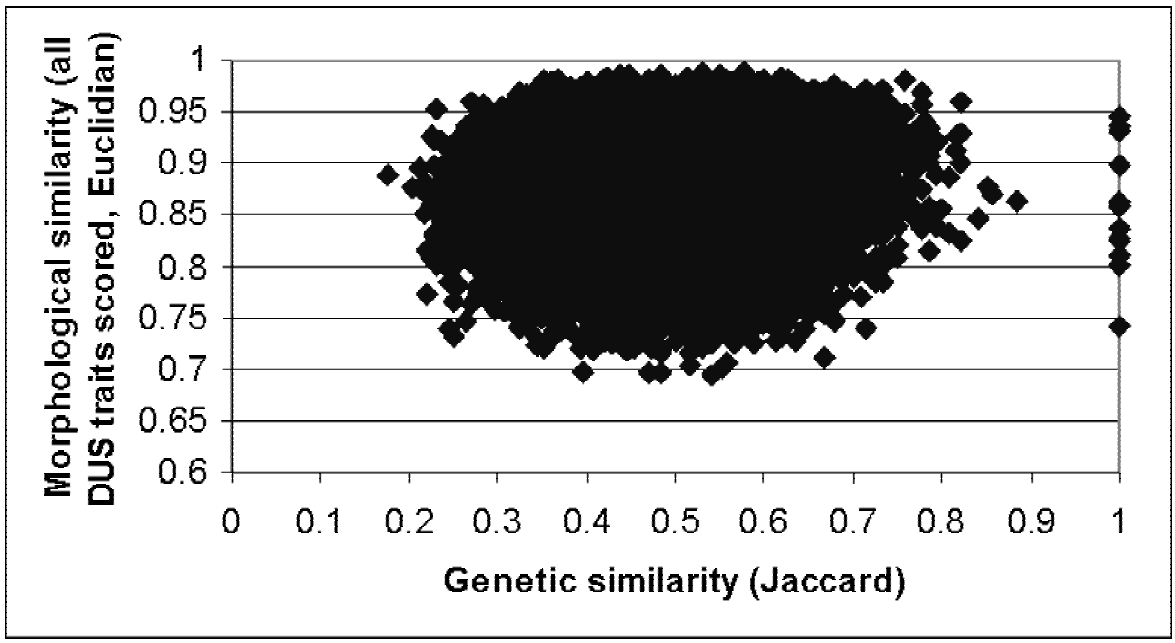


A



B

Figure 3.
Genetic versus overall morphological similarity.



[End of document]