E

UPOV

BMT/8/4
ORIGINAL: English
DATE: August 13, 2003

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Eighth Session
Tsukuba, Japan, September 3 to 5, 2003

THE USE OF SSRs FOR DUS TESTING OF WHEAT

2. GENETIC AND PHENOTYPIC DISTANCES

Document prepared by experts from the United Kingdom

The Use of SSRs for DUS Testing of Wheat

2. Genetic and Phenotypic Distances

Shanhong Wang, Susan D Freeman, David Lee, John R Law, Paolo Donini, Vince Lea and Robert J Cooke

NIAB, Cambridge, UK

1. Introduction

We have previously reported the results of the detailed analysis of a set of 40 winter and spring wheat varieties at eight selected and well characterised microsatellite (simple sequence repeat, SSR) loci. The analysis included an investigation of both varietal uniformity and stability. As a result of this, we believe that it is possible to suggest a revised approach to DUS testing in wheat using SSRs (an Option 3 approach – see BMT Review Group meeting, April 2002), which could offer reductions in time and costs and/or form the basis of a means of managing large reference collections more effectively, without undermining current levels of protection. However, given the current view of the BMT Review Group (April 2002) that there are fewer difficulties with Option 2 approaches, we are also investigating the use of such a system in wheat, by extending the number and range of SSRs used to analyse this set of varieties and then making comparisons between various estimates of genetic and phenotypic distance in this variety set. A similar project is also being undertaken with oilseed rape varieties and SSR markers.

2. Marker Selection and General Approach

The first phase of this project utilised SSRs that had been previously evaluated in an EU-funded project (see Röder *et al*., Theor & Appl Genet 106, 67, 2002). Initially, 49 SSRs were screened, to produce a set of 23, which were highly polymorphic within a collection of 10 UK wheat varieties. These 23 were then assessed for their suitability for Uniformity assessment by analysing 48 individuals from within the 10 varieties. Factors investigated included the ability to analyse the distinctness and uniformity of varieties, but also the ease of analysis, accurate detection and scoring of products (i.e. the quality of the marker), simultaneous analysis (multiplexing), and map position. From this, a set of 8 SSRs were selected for detailed use. In the current second phase of the project, an additional c. 40 SSR markers were optimised and tested. These were obtained from a range of sources, and had mostly already been used at NIAB, both in phase 1, the EU project and in other projects. They also included markers that were derived from expressed regions of the genome, i.e. could be considered as "genic" markers (they are actually microsatellite sequences from un-translated, non-coding regions of genes). Thus the analytical conditions and expected product sizes of all of the SSRs were well known, as were the map locations of most of them.

These markers were used to analyse bulked DNA samples from the same 40 varieties as in phase 1, with the objective of producing a set of well characterised, largely mapped SSRs, which can be reliably scored independent of the detection platform used. The use of c. 50 SSR markers in total should allow a more equivalent comparison between the current UPOV

morphological markers (there are 26 characteristics in wheat, each with a number of states) and the molecular markers. This in turn should facilitate the estimation of the relationship between genetic and phenotypic distances, i.e. an Option 2 approach as favoured by the UPOV BMT Review Group.

It is planned that the genetic distances (from SSR data) between the 40 varieties will be assessed in various ways, including, but not limited to Rogers and City Block approaches, both with and without the use of available mapping information. It should also be possible to compare distances estimated using genic and non-genic markers, separately and combined. At the same time, the morphological data will be analysed to give phenotypic distances, again estimated in different ways. Once comparable pairwise distance estimates (phenotypic and genetic) are available, a range of "proximity" analyses (including Mantel statistics and Generalised Procrustes) will be conducted to assess the degree of association observed. The SSR and morphological datasets have also been provided to our colleagues at GEVES (France) for preliminary analysis using their "PREDIP" software. The overall objective is to analyse the relationship between the two estimates of distance (both assessed in a range of ways), and subsequently to investigate the potential for managing the reference collection of wheat varieties and determining most similar varieties prior to field testing.

3. Preliminary Results

(i) Markers: From the above process, a set of 47 SSR primer pairs were used to analyse the 40 wheat varieties. The SSRs comprised 13 that could be described as "genic" markers, whilst the rest ("genomic") are essentially anonymous (or of unknown function), although they may be linked to traits of interest. Table 1 summarises some features of the markers. There is at least one marker on each chromosome, although some chromosomes (e.g. 2B, 3A) are less well represented than others. The allele(s) at each locus were recorded in a standardised form (in accordance with procedures used previously). A total of 217 alleles were found in the 40 varieties. Eight of the SSR loci had null alleles, which in some cases were relatively abundant (e.g. with marker 33, 11 of the 40 varieties carried the null allele). The data were then used to produce pairwise distance estimates, along with similar estimates produced from morphological data (using the UPOV characteristics for all 40 varieties).

(ii) Initial Analyses: Initial statistical analysis has been concerned with deriving genetic and phenotypic distances using various approaches. Thus far, Euclidean and City Block methods have been used for both the SSR and morphological data, as in previous work (e.g. Law *et al.* Plant Varieties & Seeds 12, 335, 1999). Within a data type, a high level of comparability between similarity coefficients based on City Block and Euclidean approaches exists (correlation 0.9703 for morphology, 0.958 for SSRs). With the available SSR data it was also possible to use the original band presence/absence (1/0) scores as a genetic profile and to estimate genetic similarity using Jaccard's method. Correlation coefficients between sets of pair-wise genetic similarity estimates from the SSR data were higher for the comparison of Jaccard v. City Block (0.8181) than for Jaccard v. Euclidean (0.6557).

The correlations by data type and similarity estimation approach are given in Table 2. Irrespective of the method used, the correlations between similarity estimates based on morphology and molecular are relatively weak (< 0.42, rising to approximately 0.5 when the

Jaccard method is applied to the SSR data). Figure 1 illustrates the general relationship observable thus far between genetic and phenotypic distances.

(iii) Comparison of Methods: We are also examining an approach to the assessment of data type and similarity algorithm in terms of similarity distribution. The rationale for this is to remove the scale effect and to focus on the cases of interest where similarity values (however obtained) are high and thus potential distinctness difficulties are likely to arise. With 40 varieties there are 1560 off-diagonal pair-wise elements of each similarity matrix. It is informative to assess if the phenotypic distance and the genetic distance estimators identify the same pairs of varieties with specific similarity properties. As a starting point, the top 5% of ranked similarities were taken as an arbitrary cut-off point. Thus the top 78 ranked similarities were studied. This upper 5 percentile for morphology for Euclidean estimation corresponds to a similarity value of 0.959 (0.897 if City Block is used), whilst for SSRs the similarity value that marks this 5 percentile is 0.855 (City Block), 0.904 (Euclidean) and 0.524 (for band profile Jaccard approach).

(a) Morphological data: If the Euclidean and City Block approaches to the estimation of phenotypic similarity are operating in a same way, then both would identify the same 78 variety pairs in the top 5% of ranked similarities for the respective methods. Observed data shows that 58 variety pairs (out of 78) were in fact identified by both methods.

(b) SSR data: The corresponding result for SSRs showed that, when comparing Euclidean and City Block estimation, 61 of 78 variety pairs were 'flagged' by both methods as being within the top 5% of ranked similarities. When all three SSRs distance methods were compared, 38 variety pairs were in the top 5% of ranked similarities with all three approaches; and a further 20 agreed in two of the three estimation methods.

(c) Between Data Types: Having established a degree of conformity in terms of the upper portion (5%) of the similarity distributions within data types, it is also possible to compare the phenotypic similarity estimation and genetic similarity. Using the Euclidean method for morphology and City Block for SSRs, ten variety pairs were 'flagged' as being in the upper 5% of ranked similarities by both phenotypic and genetic distance estimates. When all five distance methods utilised thus far were compared 8 variety pairs (of the 78) were 'flagged' by all five methods, and a further 18 by three of the five methods.

(iv) PREDIP: At the 21st TWC meeting (2003) an approach using a software system being produced by GEVES (PREDIP) was discussed, which seeks to improve the relationship between phenotypic and genetic distances. The wheat data have been kindly processed through PREDIP by GEVES colleagues. PREDIP is a statistical predictive method. It needs a "learning" set of varieties to take into account phenotypic and molecular variability and to design the links between both types of data. In the practical case, the learning set would be the reference collection. With the wheat data, the sample size (40 varieties) was too small to work with effectively using qualitative data (200 varieties would be needed). Thus the data have been considered as if they were quantitative, and 7 characteristics that did not exhibit sufficient variability had to be omitted. Dendrograms were constructed from the SSR data (using Nei and Li distance), from the phenotypic data (Euclidean) and from the predicted phenotypic distance, and the resultant groupings compared. Whilst the correlation between Nei and Li distance and phenotypic distance was low (0.36), that between the predicted and

observed phenotypic distances was much better (0.89). It has to be remembered, however, that this is not a validation test of the method, as the same data are being used for the learning set and for the test set. Nonetheless, the initial results are encouraging. As in the previous studies to validate PREDIP, the results are consistent when pairs of varieties from the learning set are considered, i.e. the PREDIP distance is a pseudo-genetic distance that is well correlated with the phenotypic distance.

4. Other Crops

A similar project is being undertaken in oilseed rape. In this case, for distinctness and uniformity, 45 varieties (all from the UK NL), with 48 individuals from each variety, are being analysed using 16 SSRs. This work is then being extended by the analysis of a further approximately 150 varieties from other EU countries, using the same 16 SSRs. It is planned that morphological data will be available for all of these varieties, and that this will allow a similar approach to be taken as above, i.e. a comparative study of the relationship between various estimates of genetic and phenotypic distance, with the objective of "pre-screening" varieties before field testing.

5. Future Work

This project will continue to evaluate a number of phenotypic and genetic distance estimators and a wide range of statistical tools to assess the proximity/association between these distance measures, in both wheat and oilseed rape. It is likely that more data (both varieties and SSRs) will be required in order to be able to validate fully the various approaches. With PREDIP, once the model is fitted with a learning set, it must be validated on a (different) validation set. This further step should give us an idea of the ability of the model to predict phenotypic distances for candidate varieties (i.e. ones that are not in the learning set). It would also be advantageous to have sufficient data (varieties and SSRs) to enable the morphological characteristics to be analysed as qualitative traits. This in turn would allow the separation of the varieties into winter and spring types, and enable the investigation of other potential groupings (e.g. feed vs. bread-making types). We hope that collaborative work will continue with our French colleagues (and others if possible), as we all strive for scientifically sound and practically operable solutions.

Acknowledgements – we are grateful to Defra for funding and support for this work. The co-operation of colleagues at GEVES is also gratefully acknowledged.

Table 1. SSRs used to analyse 40 winter and spring wheat varieties.

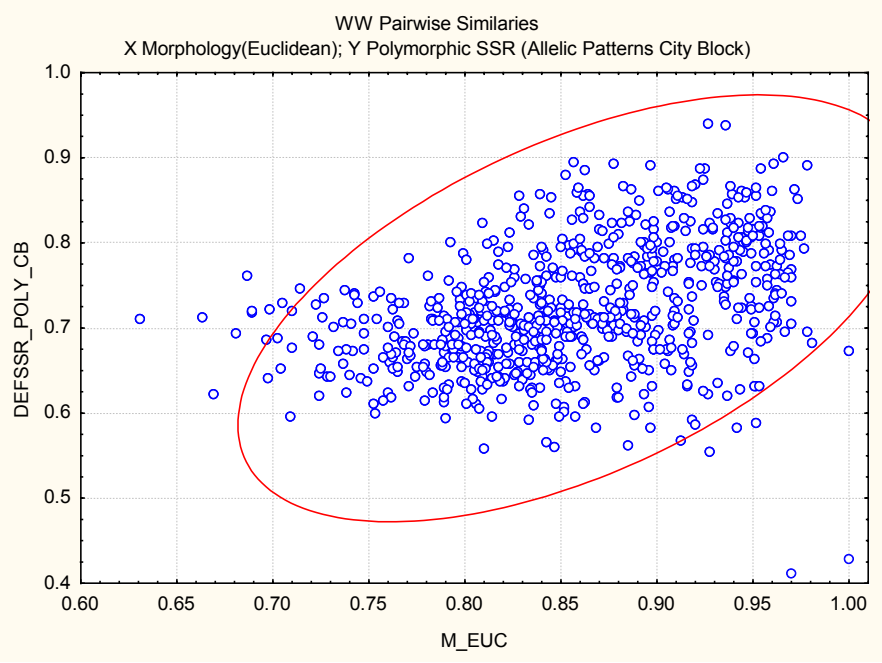
SSR no.	SSR Type	Chromosome	No. of alleles/40 vars
1	Genic	4A	4
2	Genic	n.d.	2
3	Genic	4A	1
4	Genic	5B	3
5	Genic	n.d.	3
6	Genic	n.d.	3
7	Genic	n.d.	4
8	Genic	5B	3
9	Genic	n.d.	2
10	Genic	6B	5
11	Genic	n.d.	3
12	Genic	n.d.	7
13	Genic	1BS	8
14	Genomic	3AL	4
15	Genomic	3DS	4
16	Genomic	2DS	5
17	Genomic	6DS	6
18	Genomic	5BL	5
19	Genomic	1DS	3
20	Genomic	6BS	3
21	Genomic	1BS	5
22	Genomic	2AS	5
23	Genomic	5AL	5
24	Genomic	3DL	2
25	Genomic	4AL	7
26	Genomic	4B,4D	5
27	Genomic	5DS	4
28	Genomic	3BS	7
29	Genomic	4BL	4
30	Genomic	7AL	3
31	Genomic	1AL	2
32	Genomic	7DL	5
33	Genomic	7BL	9
34	Genomic	5AL	8
35	Genomic	2DL	8
36	Genomic	7BS	7
37	Genomic	1AL	7
38	Genomic	5DS	5
39	Genomic	5DL	4
40	Genomic	1DL	8
41	Genomic	6A	2
42	Genomic	1BL	2
43	Genomic	6AL	4
44	Genomic	2D,4D	1
45	Genomic	4AL	3
46	Genomic	2AS	11
47	Genomic	2BL	6

n.d. = not determined

Table 2. Similarity correlations by data type and method

		SSR		
		Euclidean	City Block	Jaccard
Morphology	Euclidean	0.355	0.428	0.510
	City Block	0.339	0.411	0.493

Figure 1. A preliminary analysis of the correlation between genetic and phenotypic distances



[End of document]