E

BMT/13/6
ORIGINAL: English
DATE: November 17, 2011

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

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

Thirteenth Session
Brasilia, November 22 to 24, 2011

A POTENTIAL UPOV OPTION 2 APPROACH FOR BARLEY USING HIGH DENSITY SNP GENOTYPING

Document prepared by experts from the United Kingdom

INTRODUCTION

1. This project aimed to use existing data from a collaborative research programme to investigate a potential Option 2 approach ("Calibrated molecular distances in the management of varieties and collections") in DUS testing of barley. The Option 2 approach requires "Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics" (see document TC/38/14 – CAJ/45/5). This requirement is intended to ensure that decisions made under a new molecular testing system would be the same as those made under the existing morphological testing system. The molecular testing system would, of course have to meet the quality criteria set out in the "GUIDELINES FOR DNA-PROFILING: MOLECULAR MARKER SELECTION AND DATABASE CONSTRUCTION" (document UPOV/INF/17/1).

2. The costs of genotyping plant material have fallen dramatically with the advent of capillary based DNA analysis equipment, SNP arrays and 'next generation' sequencing. By comparison the costs of phenotyping to determine plant morphologies for DUS testing have remained relatively high. As the costs of genotyping decline in relative terms the attractions of Option 2 will increase, provided the quality of variety protection remains similar or improves. Ideally there would be a perfect relationship between morphological and molecular distances such that the decisions made using a molecular system would exactly mirror those made under the current system. (Figure 1, upper graph). Should the relationship between the two testing methods be anything less than perfect, there would be a zone of 'uncertainty' where

ambiguous decisions might be made (Figure 1, lower graph). Quantifying these relationships and the extent of ambiguity were the objectives of this study.

3. The utility of Option 2 has been investigated in grapevine, maize, oilseed rape durum wheat and barley. The results of these investigations have been mixed. In a study of durum wheat lines (Noli *et al.*, 2008), a collection of 69 advanced lines from seven crosses were assessed for distinctness using 17 characteristics from the Community Plant Variety Office (CPVO) protocols selected as variable among the parental lines, a suite of 99 SSR markers and AFLP assays using combinations of two and three selective bases in seven primer combinations. The correlation between the molecular markers (SSRs and AFLP) was good (r = 0.89) while the correlation between morphology and molecular markers was moderate (SSRs, r = 0.66; AFLP, r = 0.62).

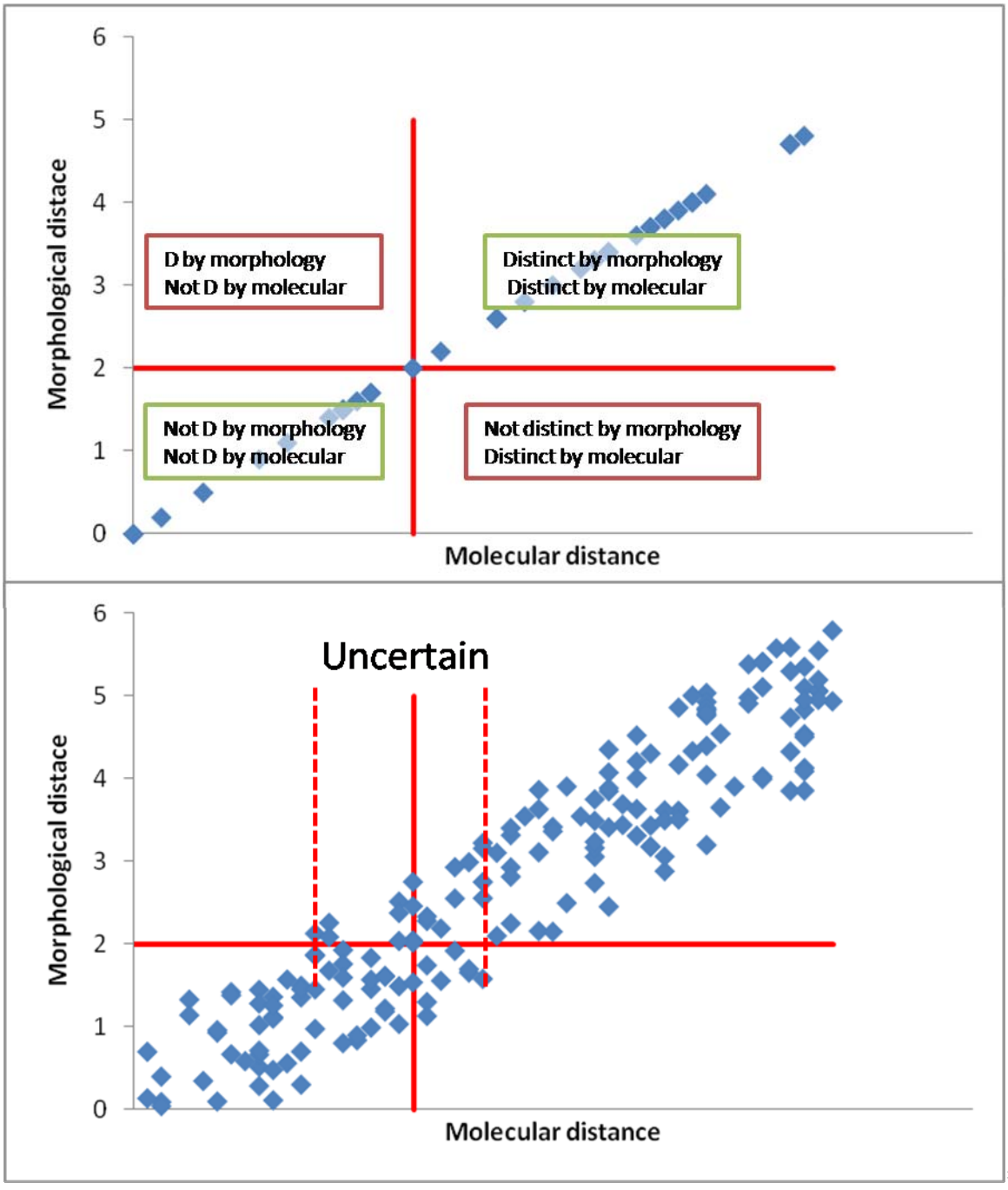


Figure 1: Calibration of molecular against morphological distances under Option 2. The

upper graph illustrates decision making under a perfect correlation between molecular and morphological distances. The lower graph illustrates possible uncertainty where the correlation between molecular and morphological distances is sub optimal

4. Investigations into the correlations between morphology and molecular based distances in maize (Gunjaca *et al*., 2008) examined a collection of 41 inbred lines comprising 13 publicly available varieties and 28 breeders' lines. Morphological descriptions were calculated using 34 characteristics from the UPOV Test Guidelines and molecular distances calculated using data for 28 SSR loci. In this instance the correlation between morphology and molecular markers was poor (r = 0.21). A large, international set of varieties was examined in a study of oilseed rape (CPV5766 Final Report) using 335 records from DUS testing authorities in Denmark, France, Germany and the United Kingdom. The collection was genotyped using a suite of 29 SSR markers. The outcome of this study was far more disappointing, with the correlation between morphological and molecular marker based distances falling between 0.03 and 0.08, depending on the methods used to calculate the distances.

OBJECTIVES

5. The main objective of the project was to assess the genetic and phenotypic distances between varieties using a combination of statistical methods and, for the purposes of DUS testing, determine whether sufficient correlation exists between the two to implement an Option 2 approach in barley. We addressed this objective by testing two hypotheses:

- Genotypic and phenotypic distance measures for a set of varieties will have a strong positive correlation to each other.

- Varieties shown as 'similar' using phenotypic distances will also be shown as 'similar' using genotypic distances.

MATERIALS AND METHODS

6. Barley was selected as the subject of this study as high density genotype data were available for a set of varieties that had been assessed in DUS examinations using a standardised set of phenotypic characteristics.

7. The project used genotypic data collected in the course of the AGOUEB project (http://www.agoueb.org/). The AGOUEB project used 3072 SNP marker loci developed from more than 1500 genes (one to three SNPs per gene) to genotype a collection of 500 barley varieties selected from United Kingdom registration trials over the past 20 years (Cockram *et al*., 2010). Phenotypic data originating from the DUS trials for the same period for 579 winter and spring barley lines were collated for this project. The majority of descriptions were derived from data collected by NIAB in the course of DUS examinations, though a small number of descriptions were obtained by bilateral purchase and therefore DUS tested in another country.

8. The data assessed in DUS testing comprised 33 characteristics assessed for 579 varieties. The number of characteristics was reduced to reflect only those characteristics included in CPVO-TP/019/2 (2010) (see Table 1).

Table 1: Characteristics used in DUS-test and preparation of descriptions. 'band width' represents a stringency criterion for each characteristic representing the minimum difference that may be used within the NIAB test system. * These quantitative characteristics appear in UPOV document TG/19/10 alongside qualitative characteristics for the same characteristics

Characteristic (CPVO-Nr.)	UPOV No	Details	Band width
1. Plant: growth habit	1	Quantitative characteristic measure coded as a 1-9 scale	3
2. Lowest leaves: hairiness of leaf sheaths	2	Grouping characteristic scored as Present (9) or Absent (1)	1
3. Flag leaf: intensity of anthocyanin coloration of auricles	3*	Quantitative characteristic measure coded as a 1-9 scale	3
4. Plant: frequency of plants with recurved flag leaves	5	Quantitative characteristic measure coded as a 1-9 scale	3
5. Flag leaf: glaucosity of sheath	6	Quantitative characteristic measure coded as a 1-9 scale	3
6. Time of ear emergence	7	Quantitative characteristic measure coded as a 1-9 scale	2
7. Awns: intensity of anthocyanin coloration of tips	9*	Quantitative characteristic measure coded as a 1-9 scale	3
8. Ear: glaucosity	10	Quantitative characteristic measure coded as a 1-9 scale	3
9. Ear: attitude	11	Quantitative characteristic measure coded as a 1-9 scale	3
10. Plant: length	12	Quantitative characteristic measure coded as a 1-9 scale	2
11. Ear: number of rows	13	Grouping characteristic scored as Two-rows (1) or More than two rows (2)	1
12. Ear: shape	14	Pseudo-qualitative characteristic scored as one of three character states (tapering (3), parallel (5) or fusiform (7)).	3
13. Ear: density	15	Quantitative characteristic measure coded as a 1-9 scale	3
14. Ear: length	16	Quantitative characteristic measure coded as a 1-9 scale	3
15. Awn: length	17	Quantitative characteristic measure coded as a 1-9 scale	2
16. Rachis: length of first segment	18	Quantitativeharacteristic measure coded as a 3-7 scale	3
17. Rachis: curvature of first segment	19	Quantitative characteristic measure coded as a 1-9 scale	3
18. Ear: development of sterile spikelets	-	Qualitative characteristic scored as one of two character states (none or rudimentary (1) or full (2)).	1
19. Sterile spikelet: attitude	20	Quantitative characteristic measure coded as a 1-3 scale	2
20. Median spikelet: length of glume and its awn relative to grain	21	Quantitative characteristic measure coded as a 1-3 scale	2
21. Grain: rachilla hair type	22	Grouping characteristic scored as short (1) or long (2)	1
22. Grain: husk	23	Qualitative characteristic scored as absent (1) or present (9)	1
23. Grain: anthocyanin coloration of nerves of lemma	24	Quantitative characteristic measure coded as a 1-9 scale	3
24. Grain: spiculation of inner lateral nerves of dorsal side of lemma	25	Quantitative characteristic measure coded as a 1-9 scale	3
25. Grain: hairiness of ventral furrow	26	Grouping characteristic scored as absent (1) or present (9)	1

Characteristic (CPVO-Nr.)	UPOV No	Details	Band width
26. Grain: disposition of lodicules	27	Qualitative characteristic scored as frontal (1) or clasping (2)	1
27. Kernel: colour of aleuron layer	28	Pseudo-qualitative characteristic scored as one of three character states (whitish (1), weakly coloured (2), strongly coloured (3)).	2
28. Seasonal type	29	Grouping characteristic scored as one of three character states (Winter type (1), alternative type (2), Spring type (3)).	2

9. The data comprise a mix of quantitative characteristics converted into scores (e.g. plant height), pseudo-qualitative characteristics converted into scores (e.g. ear shape) and qualitative characteristics (e.g. grain: husk). This data set includes five grouping characteristics, omitting a sixth found in document TG/19/10 (Awns: anthocyanin coloration of tips (characteristic 8): presence / absence). Within the NIAB implementation of the DUS test system a stringency criterion, 'band width', is used as a filter when making variety / candidate comparisons. The 'band width' represents a minimum difference threshold for each characteristic that must be met when calculating differences.

10. The genotypic markers were discovered using publicly available barley expressed sequence tags (ESTs) which were converted to a series of Illumina Golden Gate SNP arrays capable of generating 3072 assays, averaging more than 2 markers/cM across the approximately 1,100-cM barley genome (14, 17). This represents the most comprehensive resource of its kind currently available in barley and the highest density of markers used in an investigation of Option 2.

11. These disparate datasets were united for this study to produce a final set of 431 varieties with both phenotypic and genotypic data. The intersection between the genotypic and phenotypic datasets included 465 varieties. The final data set was drawn from among the 465 varieties by rejecting varieties where there were missing data for more than ten DUS test characteristics and varieties with more than 20% missing genotypic data.

The data were stored using a 'Microsoft Access' database. The data structures are shown in Figure 2.

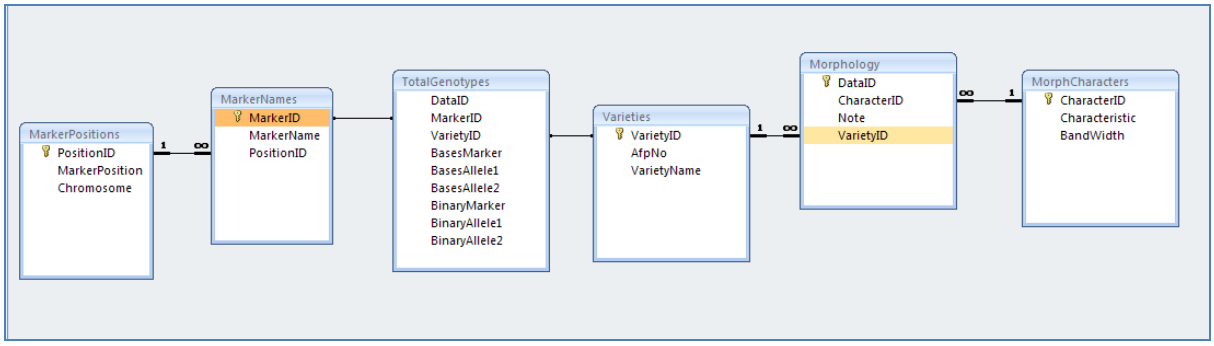


Figure 2: Database structures used to store and manage the data within the project

12. Further subsets were drawn from the genotype data by removing markers from among the full set (Table 2). The data sets were generated using a series of SQL statements within the RODBC package of the R statistics package.

Table 2: Genotype datasets selected in order to calculate various genotypic distances

	Data set	Number of loci	Criterion
A	Full data set	3072	None
B	No missing data	1562	All loci with any missing data removed
C	No missing data, no monomorphic	1274	As above with all monomorphic loci removed
D	No missing data, no monomorphic, minor allele frequency >0.1	905	No missing data, no monomorphic, including loci with the minor allele frequency between 0.1 and 0.499
E	No missing data, no monomorphic, minor allele frequency <0.1	369	No missing data, no monomorphic, excluding loci with the minor allele frequency between 0.1 and 0.499
F	No missing data, no monomorphic, minor allele frequency >0.05	1021	No missing data, no monomorphic , including loci with minor allele frequency between 0.05 and 0.499
G	No missing data, no monomorphic, minor allele frequency <0.05	254	No missing data, no monomorphic, excluding loci with minor allele frequency between 0.05 and 0.499
H	5% missing data	2654	All loci with more than 5% missing data removed
I	5% missing data, no monomorphic	2262	As above with all monomorphic loci removed
J	5% missing data, no monomorphic, minor allele frequency >0.1	1554	5% missing data, no monomorphic Where only loci with the minor allele present at a frequency between 0.1 and 0.499
K	5% missing data, no monomorphic, minor allele frequency <0.1	708	5% missing data, no monomorphic Where only loci with the minor allele present at a frequency between 0.001 and 0.1
L	5% missing data, no monomorphic, minor allele frequency >0.05	1803	5% missing data, no monomorphic Where only loci with the minor allele present at a frequency between 0.05 and 0.499
M	5% missing data, no monomorphic, minor allele frequency <0.05	459	5% missing data, no monomorphic Where only loci with the minor allele present at a frequency between 0.001 and 0.05
N	Evenly distributed markers	944	Markers are clustered by map position, in groups of between 1 – 38 markers. Markers were selected at random to represent each map position

13. There was a high proportion of missing phenotypic data in this final set. The risk of low inter variety distances introduced by missing data was reduced by imputation. The methods for imputation of missing data were developed by medical statisticians to handle data-sets that include incomplete survey results. The imputed data used to replace missing values should not substantially change the results of analysis or the conclusions drawn from the results. Multiple imputed data-sets are therefore generated and the results of analysis of each data-set compared or pooled in order to ensure that the conclusions drawn from analysis are defensible. The work flow is described schematically below in Figure 3. The process starts with an incomplete data-set. Missing data were replaced by imputed values to generate a number of complete data-sets, each of which is analysed, generating a number of result sets. The multiple results sets are pooled and conclusions drawn. In this case, we imputed phenotype data by random sampling: for each characteristic, missing data were replaced by values drawn at random from the existing data. Multiple sets of phenotype data were generated in this way and distance matrices calculated for each of them and the results held in a three dimensional array. The distance matrices were pooled by taking an arithmetic mean over the third dimension to calculate a conventional two dimensional distance matrix.

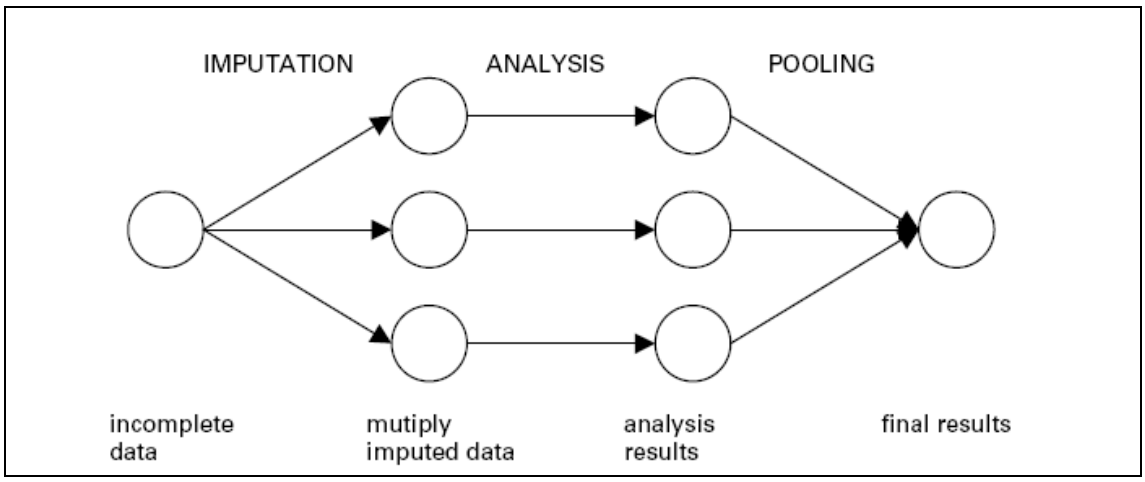


Figure 3: A schematic of the work flow through the imputation process. (Figure from van Buuren and Oudshoorn, 1999)

14. The data analysis was carried out using Microsoft Excel, ASReml (Gilmour *et al*., 1995) and the R Statistical Package (2010) including packages mice: Multivariate Imputation by Chained Equations (van Buuren and Oudshoorn, 2011) and cluster: Cluster Analysis Extended (Rousseeuw et al, 2011). These packages were used to calculate the simple genetic distance metrics: Manhattan and Euclidean Distances and simple phenotypic distances: Manhattan and Modified Manhattan Distances and Gower's Coefficient (1971). The Manhattan Distance was used to calculate phenotypic distances as it reflects the decision making process used in DUS examinations. The Modified Manhattan Distance is a variation to the Manhattan Distance such that the value of the pair-wise comparison for a characteristic must meet or exceed a threshold value, termed the 'band width', if it is to be added to the inter variety distance. The value of the band width is set by experts at a level that ensures calculated differences are not an artefact of variation in the observation and recording system within and between years. Gower's coefficient was selected for its suitability when handling data sets that include both binary, multistate and continuous data.

RESULTS

Validation of phenotypic datasets:

15. Two data sets were used to calculate phenotypic distances, the raw phenotype data (P1) and a set where the missing values have been replaced by imputation (P2). These data were, in turn, used to calculate three simple phenotypic distances: Manhattan Distances, Modified Manhattan Distance and Gower's Coefficient, generating six distance matrices. The data set with imputed missing data (P2) was validated by correlation with the raw phenotype data (P1). This validation showed the distance matrices calculated using P1: Raw phenotype data and P2: Phenotypes with imputed missing data correlated strongly with one another (Table 3). These correlations are represented graphically in Figure 4.

Table 3: Comparisons of correlations between phenotypic distances calculated using Dataset P1: Raw phenotype data and Dataset P2: Phenotype with imputed missing data

		P1: Raw phenotype		
		Gower	**Manhattan**	**Modified Manhattan**
P2: Phenotype with imputed missing data	**Gower**	0.981	0.929	0.851
	Manhattan	0.920	0.977	0.920
	Modified Manhattan	0.865	0.937	0.961

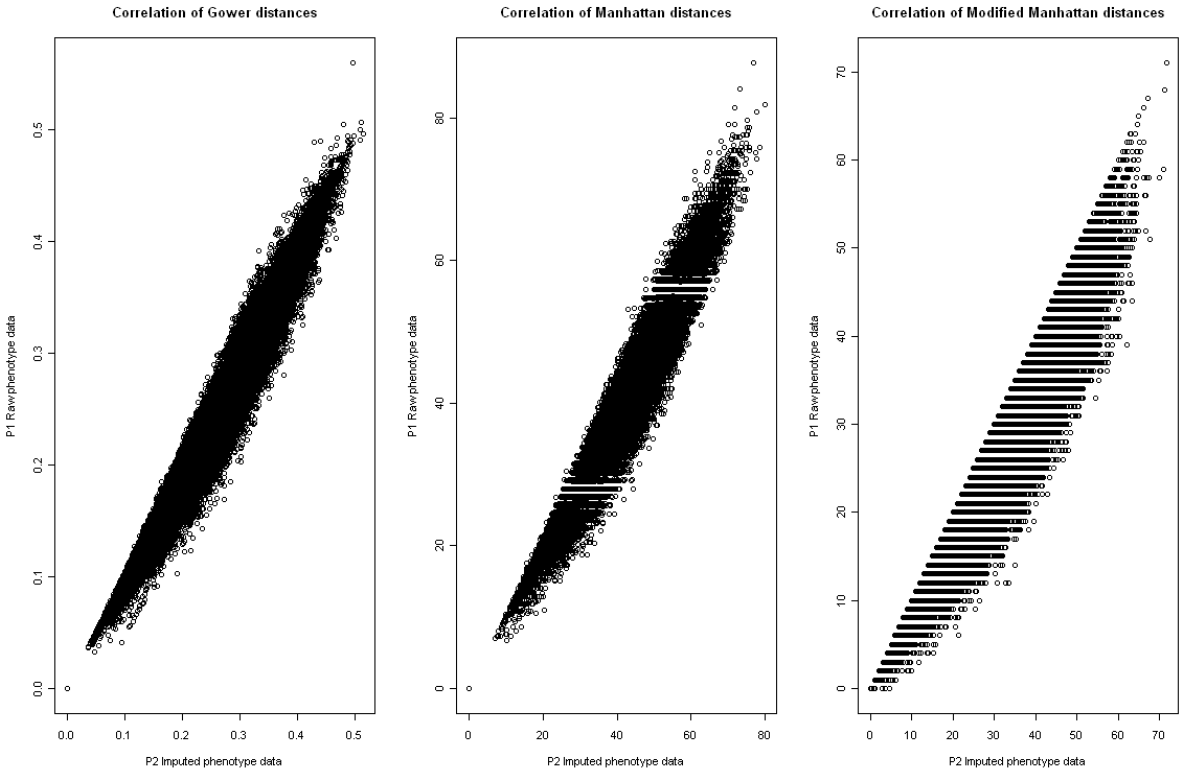


Figure 4: Scatter plots comparing distances calculated using data sets P1Raw phenotype data and P2 Phenotype with imputed missing data using Gower's coefficient (left), Manhattan distance (centre) and Modified Manhattan distance (right)

16. The average of the distances calculated using P1 Raw phenotype data (Gower = 0.239, Manhattan = 37.3, Modified Manhattan = 22.9) are consistently lower than those calculated using P2 phenotype with imputed missing data (Gower = 0.248, Manhattan = 38.5, Modified Manhattan = 26.1) and these differences were significant ($p < 0.001$). The pattern seen in the three scatter plots suggests that the difference between the distances calculated using the two data sets is least for either high or low distances.

17. Internal validation tests were designed to assess the number of imputations needed to produce a robust data set. Four values were tested for the number of imputations (5, 10, 20, 100) and the deviation among data sets created using these values by carrying out this process in 99 iterations. The results of this validation test showed that the mean distances computed were the same in all cases though the precision around that mean improved as the number of imputations increased. One hundred imputations were used in practice.

Minimum number of markers

18. Results from previous studies have shown a range of correlations between phenotypic and genotypic distances. Here we report the results of a study where the number of available markers is at least an order of magnitude greater than the number of markers used in previous studies. In order to investigate the effect of marker numbers on the correlation between phenotypic distance and genotypic distance, a random set of genotypic markers were selected from among Data Set B (No missing data) and Data Set H (5% missing data) in turn. Correlations were calculated between the genotypic distances (Euclidean and Manhattan distance) and the phenotypic distances ((Gower, Manhattan and Modified Manhattan distance) for each random selection. The number of random selections used was 15620 for Data Set B: No missing data (1562 markers) and 26540 for Data Set H (5% missing data (2654 markers). The calculated correlations were tabulated with the number of markers selected and the results were plotted (Figure 5).

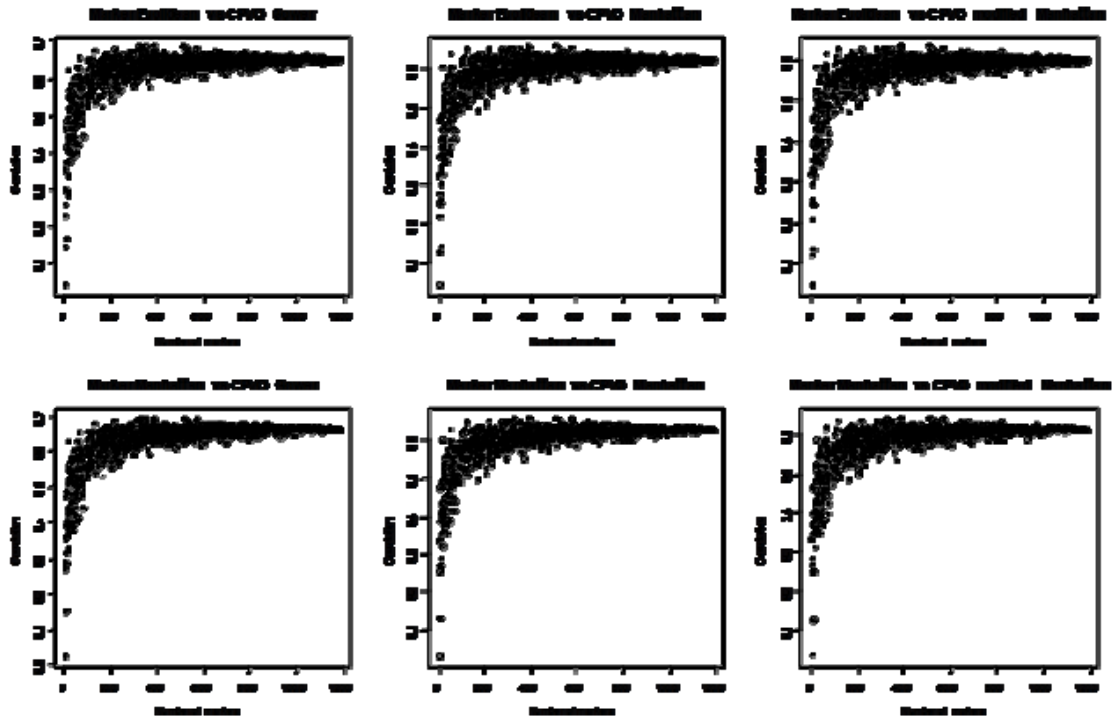
19. Figure 5 shows a clear pattern in every case. Initially, the correlations between the genotypic distances and the phenotypic distances increase as the number of markers used to calculate the genotypic distances increase. As the number of markers increases further, the correlation values plateau. Once the correlation has reached a plateau, the scatter of correlations around a central value reduces with increasing marker numbers. The low initial correlation values when small numbers of markers are used to calculate genetic distances offers an explanation for the poor correlations observed in earlier studies. The data presented in Figure 5 suggests that a minimum of 300 - 400 markers should be selected from Dataset A (No missing data) and 800 – 1000 from Dataset H (5% missing data) in order to achieve acceptable accuracy when calculating correlations.

Correlations between phenotypic and genotypic distances

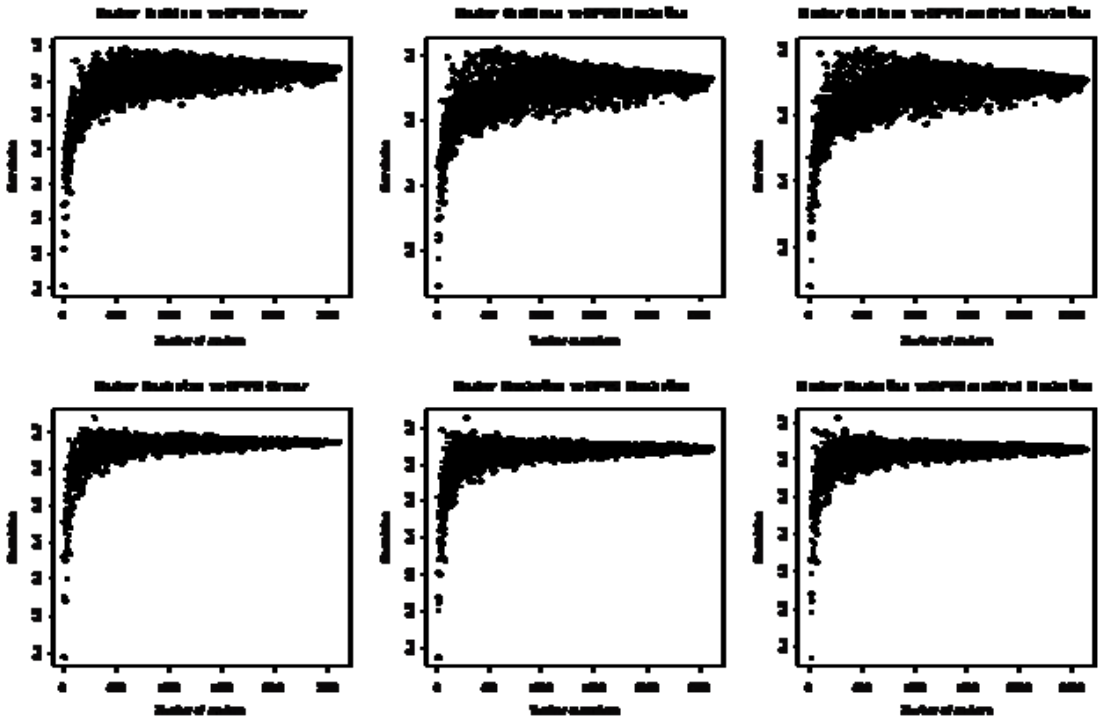
The success or failure of Option 2 depends, in part, on upholding the hypothesis which states:

> Genotypic and phenotypic distance measures for a set of varieties will have a strong positive correlation to each other.

20. Here we present data showing the extent of correlation between the subsets of phenotypic and genotypic data using different methods to calculate distance matrices. The sets have been chosen to allow an investigation of factors that may affect the quality of the distance measures. We have used the raw phenotype data without modification from the data abstracted from our ‘live’ DUS examination database. Concerns that the extent of missing data within this set might introduce errors into the analysis were addressed by creating a second data set where missing values were replaced with imputed data.



Dataset B: No missing data



Dataset H: 5% missing data

Figure 5: Scatter plots of correlations between genotypic and phenotypic distances for Data sets B and H. For each data set the Euclidean genotypic distances are represented on the top row, the Manhattan distances on the second row. The Gower phenotypic distances are represented in the first column, the Manhattan distances in the second column and the Modified Manhattan distances in the third column
Table 4: Correlations between phenotypic and genotypic distances, raw phenotype data

		Data set P1: Raw phenotype data		
		Gower	**Manhattan**	**Modified Manhattan**
	Geonotypic distance: Manhattan			
A	**Full data set**	0.638	0.622	0.596
B	**No missing data**	0.638	0.621	0.594
C	**No missing data, no monomorphic**	0.638	0.621	0.594
D	**No missing data, no monomorphic, minor allele frequency >0.1**	0.630	0.615	0.594
E	**No missing data, no monomorphic, minor allele frequency <0.1**	0.244	0.231	0.181
F	**No missing data, no monomorphic, minor allele frequency >0.05**	0.638	0.621	0.596
G	**No missing data, no monomorphic, minor allele frequency <0.05**	0.151	0.142	0.103
H	**5% missing data**	0.639	0.623	0.597
I	**5% missing data, no monomorphic**	0.640	0.624	0.597
J	**5% missing data, no monomorphic, minor allele frequency >0.1**	0.640	0.624	0.597
K	**5% missing data, no monomorphic, minor allele frequency <0.1**	0.263	0.250	0.207
L	**5% missing data, no monomorphic, minor allele frequency >0.05**	0.637	0.621	0.596
M	**5% missing data, no monomorphic, minor allele frequency <0.05**	0.224	0.210	0.169
	Geonotypic distance: Euclidean			
A	**Full data set**	0.626	0.611	0.579
B	**No missing data**	0.628	0.612	0.578
C	**No missing data, no monomorphic**	0.628	0.612	0.578
D	**No missing data, no monomorphic, minor allele frequency >0.1**	0.621	0.607	0.580
E	**No missing data, no monomorphic, minor allele frequency <0.1**	0.232	0.220	0.172
F	**No missing data, no monomorphic, minor allele frequency >0.05**	0.628	0.613	0.581
G	**No missing data, no monomorphic, minor allele frequency <0.05**	0.161	0.151	0.111
H	**5% missing data**	0.627	0.612	0.579
I	**5% missing data, no monomorphic**	0.628	0.613	0.579
J	**5% missing data, no monomorphic, minor allele frequency >0.1**	0.628	0.613	0.579
K	**5% missing data, no monomorphic, minor allele frequency <0.1**	0.256	0.245	0.202
L	**5% missing data, no monomorphic, minor allele frequency >0.05**	0.626	0.611	0.579
M	**5% missing data, no monomorphic, minor allele frequency <0.05**	0.224	0.212	0.170

Table 5: Correlations between phenotypic and genotypic distances, phenotype data with imputed values

		Data set P2: Phenotype data with imputed missing values		
		Gower	**Manhattan**	**Modified Manhattan**
	Geonotypic distance: Manhattan			
A	**Full data set**	0.656	0.625	0.602
B	**No missing data**	0.656	0.624	0.598
C	**No missing data, no monomorphic**	0.656	0.624	0.598
D	**No missing data, no monomorphic, minor allele frequency >0.1**	0.647	0.619	0.593
E	**No missing data, no monomorphic, minor allele frequency <0.1**	0.255	0.219	0.213
F	**No missing data, no monomorphic, minor allele frequency >0.05**	0.656	0.625	0.599
G	**No missing data, no monomorphic, minor allele frequency <0.05**	0.158	0.127	0.120
H	**5% missing data**	0.657	0.627	0.603
I	**5% missing data, no monomorphic**	0.658	0.627	0.603
J	**5% missing data, no monomorphic, minor allele frequency >0.1**	0.658	0.627	0.603
K	**5% missing data, no monomorphic, minor allele frequency <0.1**	0.275	0.244	0.242
L	**5% missing data, no monomorphic, minor allele frequency >0.05**	0.655	0.625	0.601
M	**5% missing data, no monomorphic, minor allele frequency <0.05**	0.234	0.205	0.204
	Geonotypic distance: Euclidean			
A	**Full data set**	0.642	0.615	0.582
B	**No missing data**	0.644	0.615	0.581
C	**No missing data, no monomorphic**	0.644	0.615	0.581
D	**No missing data, no monomorphic, minor allele frequency >0.1**	0.637	0.612	0.578
E	**No missing data, no monomorphic, minor allele frequency <0.1**	0.242	0.209	0.201
F	**No missing data, no monomorphic, minor allele frequency >0.05**	0.644	0.616	0.582
G	**No missing data, no monomorphic, minor allele frequency <0.05**	0.167	0.134	0.125
H	**5% missing data**	0.644	0.616	0.583
I	**5% missing data, no monomorphic**	0.645	0.616	0.584
J	**5% missing data, no monomorphic, minor allele frequency >0.1**	0.645	0.616	0.584
K	**5% missing data, no monomorphic, minor allele frequency <0.1**	0.268	0.239	0.234
L	**5% missing data, no monomorphic, minor allele frequency >0.05**	0.642	0.615	0.582
M	**5% missing data, no monomorphic, minor allele frequency <0.05**	0.234	0.206	0.203

21. The correlations between phenotypic and genotypic distances are all positive. The correlations observed are greater than 0.55 with the exception of values obtained for genotype data sets E, G, K and M. These four data sets were selected to investigate whether correlations between phenotypic and genotypic distances improve if genetic loci harbouring rare alleles were used to calculate the genetic distances. The results in tables 4 and 5 clearly show that this is not the case. It is possible that these low correlations are a consequence of selecting a small number of markers (E = 369 markers, G = 254 markers, K = 708 markers, M = 459 markers). When the values obtained for correlations calculated using these data sets are compared with the scatters shown in Figure 5, it can be seen that the calculated values are systematically lower than the values that would be obtained by drawing an equivalent numbers of markers at random.

22. The correlations follow a pattern when considering the phenotypic distances, such that correlations using Gower Distance > Manhattan Distance > Modified Manhattan Distance and the correlations calculated using P2 (Phenotype data with imputed missing values) are greater than those obtained by using P1 (Phenotype raw data). The correlations when considering the genotypic distances such that Manhattan Distances > Euclidean Distances though this pattern breaks down for the small data sets G and M.

23. These observed correlations in tables 2 and 3 are all positive but may not be described as strong correlations. Excepting genotypic data sets E, G, K and M, the correlations fall into the range 0.62 – 0.66 when Gower's Distance is used as the phenotypic distance, 0.61 – 0.63 when Manhattan Distance is used and 0.58 – 0.60 when Modified Manhattan Distance is used. While these correlations obtained cannot be described as weak, they offer only equivocal support for the hypothesis which states: "Genotypic and phenotypic distance measures for a set of varieties will have a strong positive correlation to each other."

Marker optimisation using spaced markers or genomic selection methods

24. The markers used in this study have been mapped across the barley genome to 944 map positions over seven chromosomes. In the last phase of the project we will investigate whether the use of a genotype data set selected to give spaced markers (Data set N) offers an improvement in the correlations between genotypic and phenotypic distance measures.

25. Population based association mapping has been applied to these data sets to identify QTLs for individual traits such as the DUS characteristics in turn. In the last phase of the project we will extend this approach and adopt methods used in trait prediction for genomic selection to the collection of 28 DUS characteristics under consideration in this study.

26. The success of these approaches may be critical to upholding the hypothesis "Genotypic and phenotypic distance measures for a set of varieties will have a strong positive correlation to each other" which in turn is fundamental to successfully implementing Option 2.

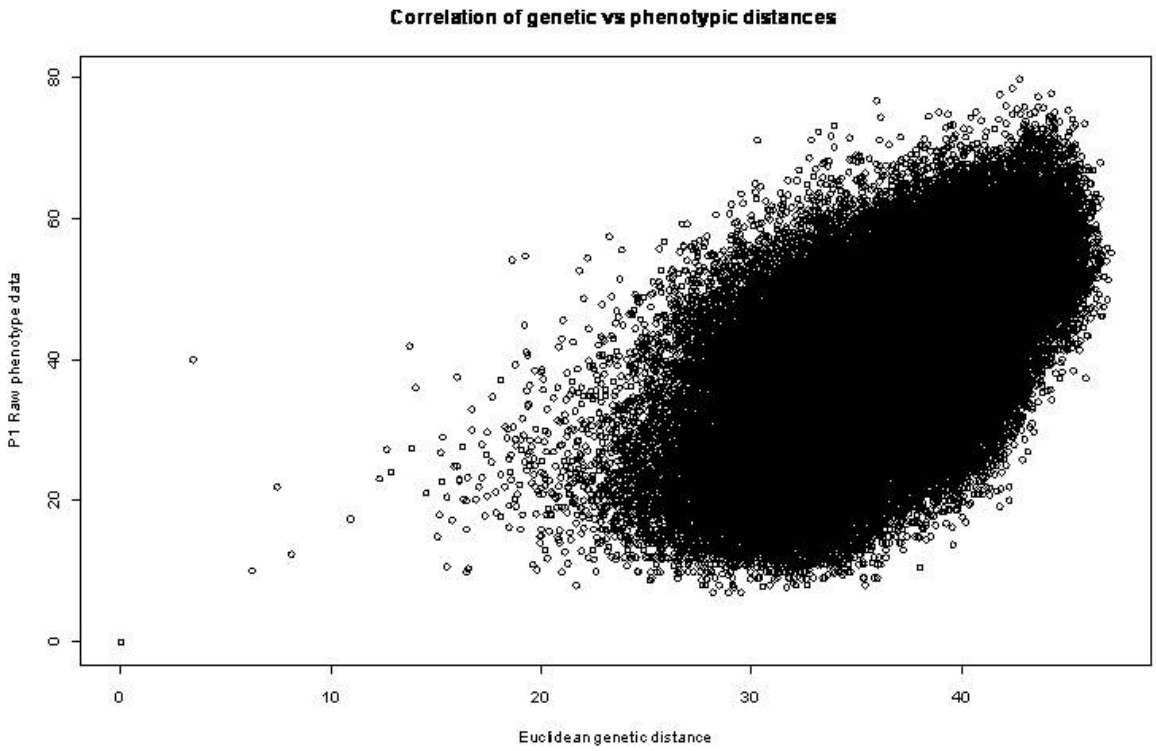


Figure 6: 'Typical' scatter of genetic vs phenotypic distances

27. The 'typical' data shown in Figure 6 illustrates the issue that needs to be resolved. Despite the positive correlation between phenotypic and genotypic distances, the noise is such that there will be ambiguity in decision making unless the correlations can be improved.

Relationships within the variety set

28. The varieties selected for this study have differing degrees of relatedness. We abstracted information from the technical questionnaires submitted with each candidate variety identifying their parents. We integrated this information with pedigree data from the BBSRC Barley Pedigree Report (www.jic.ac.uk/germplas/bbsrc_ce/Pedb.txt) and information taken from Abstammungskatalog der Gerstensorten (www.lfl.bayern.de/ipz/gerste/09740/gerstenstamm.php?loeschen=zum+Auswahl-Men%FC). Additional information was taken from passport data held by germplasm collections including the Genebank of IPK Gatersleben (http://gbis.ipk-gatersleben.de/gbis_i/home.jsf;jsessionid=c25e8cb830d6a706b961cd894c4fb52be608e066f60c) , the U.S. Department of Agriculture's Agricultural Research Service Germplasm Resources Information Network (http://www.ars-grin.gov/), and the ECPGR Barley Database (http://barley.ipk-gatersleben.de/ebdb/). The pedigree data were tabulated and interrogated in Excel.

29. The varieties within the study showed some surprising degrees of relatedness; for example, the variety 'Igri' features in the pedigree of 217 varieties, either as a parent, grandparent, great grand parent or great – great grandparent. We identified all possible full, half and quarter siblings, and those varieties related as parent – offspring or grandparent – offspring (Tables 6 and 7); for example, 65 varieties were full siblings of at least one other variety, organised into 28 families of between two and four siblings in 47 pairs. The pair wise phenotypic and genotypic distance for all related pairs were extracted and tabulated by relationship.

Table 6: Mean phenotypic distances among sets of related varieties

Average distances	Families	Pairs	Gower	Manhattan	Modified Manhattan
All varieties	NA	92665	0.25	38.87	29.31
Full siblings	28	67	0.16	25.67	16.74
Half siblings	126	2676	0.19	31.58	22.24
Quarter siblings	179	11975	0.20	33.04	23.60
Parent - offspring pairs	115	365	0.18	28.41	19.29
Grandparent - offspring pairs	67	327	0.19	30.76	21.79

30. The phenotypic data ranked the related sets differently with Gower's Distance placing the sets in order of full siblings, parent – offspring, half siblings, grandparent – offspring then quarter siblings while the Manhattan and Modified Manhattan distances placed the sets in order full siblings, parent – offspring, grandparent – offspring, half siblings then quarter siblings. The genotypic distances rank the sets in the same order as the Manhattan and Modified Manhattan phenotypic distances. The variance among the mean distances for the related sets is illustrated in Figures 7 to 9.

Table 7: Mean genotypic distances among sets of related varieties

Average distances	**Families**	**Pairs**	**Manhattan**	**Euclidean**
All varieties	NA	92665	1567.7	39.3
Full siblings	28	67	639.7	24.5
Half siblings	126	2676	1025.2	31.7
Quarter siblings	179	11975	1106.0	33.0
Parent - offspring pairs	115	365	755.8	27.0
Grandparent - offspring pairs	67	327	1024.4	31.7

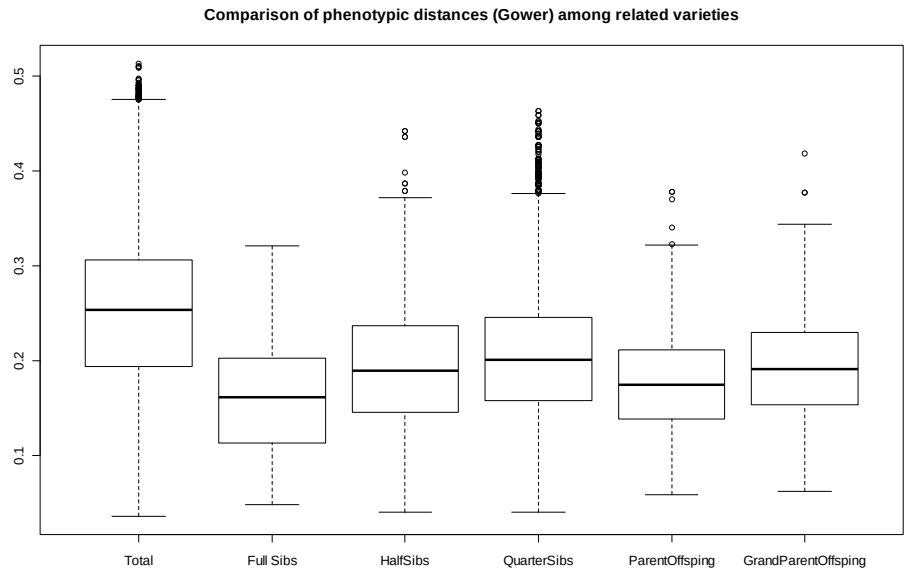


Figure 7: Variance of Gower's distances among the related sets

31. In Figures 7 and 8 an overlap in the distribution of distances can be seen between the different related sets. In contrast, the distributions of genetic distances appear to be more distinct (Figure 9). This is encouraging as it suggests that genetic distances may offer greater resolution so there may be solutions that will allow a reasonable calibration of genetic distances against phenotypic distances.

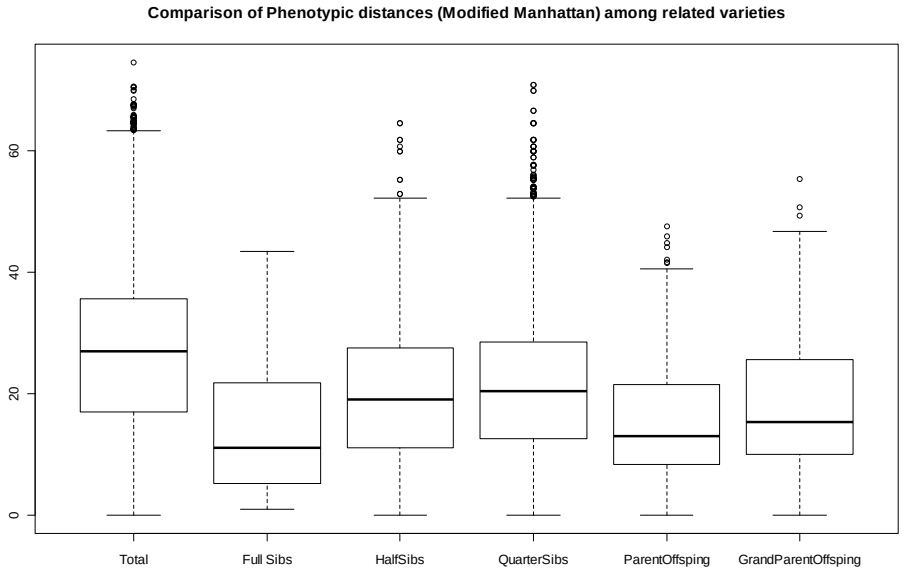


Figure 8: Variance of Modified Manhattan distances among the related sets

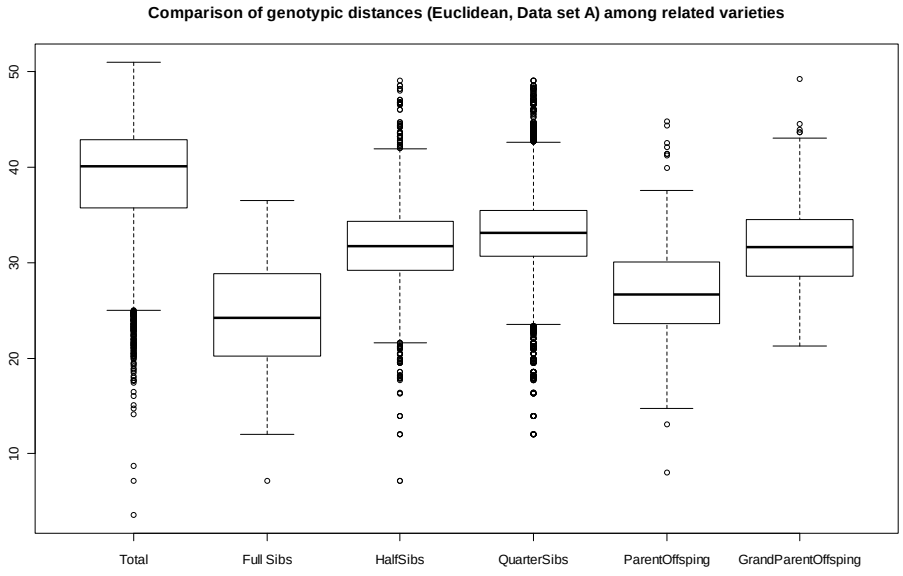


Figure 9: Variance of genetic distances among the related sets

32. When the means for each related set using phenotypic and genotypic data are plotted (Figure 10) they show a clear relationship (r = 0.977). This result confirms the potential for Option 2 in the absences of ‘noisy’ data.

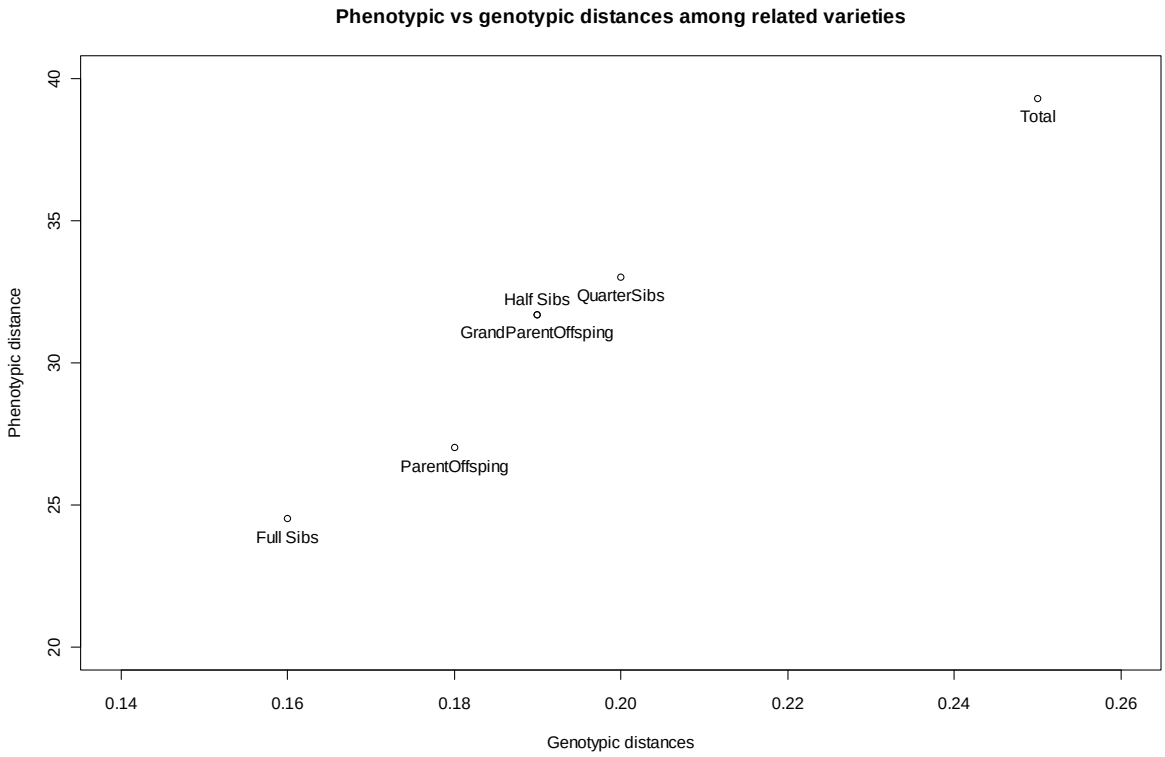


Figure 10: Mean phenotypic vs genotypic distances among the different classes of related varieties

Decision making

33. Once the investigations of marker spacing and trait prediction methods are complete, a comparison of decision making between use of phenotypic distances and fully optimized genotypic distances will be undertaken. We will attempt to quantify the risks of implementing a system based on Option 2 and make recommendations on methods for using high density markers to streamline DUS examinations.

Concluding remarks

34. Option 2 requires “Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics”. We have shown that we can use high density genotype data to calculate genotypic distances that have correlations to phenotypic distance between 0.58 and 0.66. We expect to improve on these correlations once we have optimized marker selection using spaced markers or adapted methods used in genomic selection. However, it is important not to overemphasize the importance of simple correlation between phenotypic and genotypic distances. The correlations already obtained may be ‘fit for purpose’. The success of Option 2, which depends on setting a molecular threshold that would replace the current minimum phenotypic distance, depends on the correlation in the region around the minimum phenotypic distance rather than on the overall correlation. This area will be explored further in the final phase of this study.

A project co-funded by the Community Plant Variety Office (CPVO)
Huw Jones, James Cockram, David Smith, Ian Mackay, Carol Norris; NIAB, United Kingdom

REFERENCES

UPOV document INF/17/1: Guidelines For DNA-Profiling: Molecular Marker Selection and Database Construction ("BMT Guidelines"). [online] Available at: http://www.upov.int/export/sites/upov/en/publications/pdf/upov_inf_17_1.pdf

Noli E, Teriaca MS, Sanguineti MC, Conti S (2008). Utilization of SSR and AFLP markers for the assessment of distinctness in durum wheat. Mol. Breeding **22**: 301–313.

Gunjaca J, Buhinicek I, Jukic M, Sarcevic H, Vragolovic A, Kozic Z, Jambrovic A, Pejic I (2008). Discriminating maize inbred lines using molecular and DUS data. Euphytica 161: 165–172.

CPV5766 Final Report (2008). Management of winter oilseed rape reference collections. NIAB, Cambridge, CB3 0LE on behalf of Community Plant Variety Office (CPVO), Anger, France.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties, *Biometrics* **27**, 857–874.

Cockram, J., White, J., Zuluaga, D.L., Smith, D., Comadran, j. et al (2010). Genome-wide association mapping to candidate polymorphism resolution in the un-sequenced barley genome. Proc. Natl. Acad. Sci. U.S.A., 107, 21611-21616.

UPOV TG/19/10 GUIDELINES FOR THE CONDUCT OF TESTS FOR DISTINCTNESS, UNIFORMITY AND STABILITY: Barley 94-11-04

CPVO TP/019/2 Rev English Date: 11/03/2010 PROTOCOL FOR DISTINCTNESS, UNIFORMITY AND STABILITY TESTS Hordeum vulgare L. sensu lato BARLEY

R version 2.12.0 (2010-10-15) Copyright (C) 2010 The R Foundation for Statistical Computing ISBN 3-900051-07-0

van Buuren S, Oudshoorn K (1999) Flexible multivariate imputation by MICE. TNO Prevention and Health, Leiden, The Netherlands. TNO report: PG/VGZ/99.054

Cluster Analysis, extended original from Peter Rousseeuw, Anja Struyf and Mia Hubert. Version 1.14.0 Date 2011-06-07

Gilmour, A.R., Thompson, R. and Cullis, B.R. (1995). Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51, 1440-50

[End of document]