

# FINAL REPORT

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

Agreement number EPM 7501705, co-funded by the **Community Plant Variety Office (CPVO)** Research and Development Section and NIAB Innovation Fund, 20 December 2010 to 20 December 2011.

NIAB, Huntingdon Road, Cambridge, CB3 0LE, UK

Telephone: +441223342200  
Fax: +441223342303  
Email: [info@niab.com](mailto:info@niab.com)  
Internet: [www.niab.com](http://www.niab.com)

**Co-ordinator & Responsible Scientist:** Dr Carol Norris

**Other major contributors:** Dr Huw Jones, Dr James Cockram, David Smith and Dr Ian Mackay



## EXECUTIVE SUMMARY

Variety registration and protection of barley varieties is carried out in several European Member States (EU MS), and requires distinctness, uniformity and stability (DUS) testing of new varieties. Developments in high throughput genotyping have provided the opportunity to explore the application of marker technology in this process. The overall objective of this project was to examine the potential uses of DNA molecular markers (specifically SNPs) to assess the feasibility of a UPOV Model 2 approach: ‘Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics’.

The experimental approaches were to statistically analyse available phenotypic and genotypic data to:

- Identify whether a correlation exists between phenotypic and genotypic distances
- Quantify distances measured from markers
- Phenotype against a common standard derived from known pedigree relationships within the dataset
- Adopt approaches from genomic selection to predict phenotype from the genome-wide marker set.

For this purpose, a subset of data from a previous collaborative project on barley was used. This consisted of 431 winter and spring varieties with phenotype data from UK DUS trials comprising 33 characteristics, together with genotype data from 3072 SNP markers.

Distance estimates were calculated using both the molecular and morphological data sets and compared. For Model 2 to succeed, good correlation between molecular and morphological distances is required: it should be possible to calibrate distances from molecular markers to set a threshold for Distinctness such that the same decisions are made using morphological distances. Results are more positive than previous studies. All correlations between phenotypic and genotypic distances were large, ranging from correlation coefficient ( $r$ ) = 0.55 to  $r$  = 0.66 (the closer to +1 or -1, the more closely the two variables are related). Comparison of phenotypic and genotypic distances amongst varieties grouped by kinship showed that the phenotypic and genotypic distances of these groups correlated well. Examination of data

sub-sets with increasing numbers of markers showed that there is a ceiling after which the correlations do not improve. To investigate the possibility of breaking through this ceiling, genomic prediction was used and correlations of up to  $r = 0.86$  were achieved.

To test how the positive correlations between phenotypic and genotypic distances affect decision making for Distinctness, an arbitrary threshold was set in order to simulate 10% of varieties as 'non distinct' using the morphological data (only listed varieties were used in the project and are all distinct). This set of 'non distinct' varieties was used for comparisons with varying thresholds for the genotypic data. When 43 'non distinct' varieties were identified using genetic distances, fewer than half these varieties were non distinct in the morphological distance set. When larger variety sets were used to explore this result further, it was still possible to include varieties that were 'similar' and one variety that was non distinct by phenotypic distance among the varieties selected as distinct using genotypic distance. Complete convergence with the total that are non distinct by morphology is not achieved even when 93% of the variety set are selected.

Results have demonstrated that the quality and quantity of molecular data now available can produce good correlations between molecular and morphological distances. However in practical terms the UPOV Model 2 approach could not be adopted without a level of risk. Nevertheless, correlations between morphological and marker based estimates of distance are greater than reported and we have demonstrated the promise of approaches based on high density SNP markers. To identify a better model, it is suggested that further work in this area should include: i) a large range of European barley varieties; ii) varieties that have been found to be non distinct using traditional methods; iii) validation and harmonisation of scoring of characteristics; iv) varieties genotyped with an expanded marker set to fill in gaps; v) assessment of methods to combine markers with phenotypic scores within the testing systems operational with Europe.

# TABLE OF CONTENTS

<b>INTRODUCTION AND BACKGROUND</b>	<b>5</b>
<i>State of the art – molecular markers and DUS testing</i>	<b>7</b>
<b>OBJECTIVES</b>	<b>12</b>
<b>METHODS AND RESULTS</b>	<b>15</b>
<i>Validation of phenotypic datasets</i>	<b>19</b>
<i>Minimum number of markers</i>	<b>21</b>
<i>Correlations between phenotypic and genotypic distances</i>	<b>21</b>
<i>Marker optimisation</i>	<b>26</b>
<i>Use of genomic prediction to calculate predicted morphological distances</i>	<b>28</b>
<i>Relationships within the variety set</i>	<b>32</b>
<i>Comparison of decision making using morphology or genotype data</i>	<b>36</b>
<b>CONCLUSIONS</b>	<b>40</b>
<b>PROPOSAL FOR FUTURE WORK</b>	<b>44</b>
<b>REFERENCES</b>	<b>44</b>
<b>ACKNOWLEDGEMENTS</b>	<b>47</b>

## INTRODUCTION AND BACKGROUND

The use of molecular markers for Distinctness, Uniformity and Stability (DUS) testing has been discussed by the International Union for the Protection of New Varieties of Plants (UPOV) and other interested parties for several years now and three options were recognised by the Working Group for Biochemical and Molecular Techniques (BMT) working group in 2002, and more recently revised in the document ‘Possible Uses of Molecular Markers in the Examination of Distinctness, Uniformity and Stability’ (BMT/DUS) in which the Options are referred to as Models:

- Molecular characteristics as a predictor of traditional characteristics. Use of molecular characteristics which are directly linked to traditional characteristics (gene specific markers). Model 1
- Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics. Model 2
- Development of a new system. Model 3

The UPOV Convention requires all new varieties to be compared with existing varieties of ‘common knowledge’. Within the European Union (EU) context, this should at the very least include all relevant varieties with European rights and/or listed on the Common Catalogue and should be as comprehensive as is practically possible. In the case of barley, and many other agricultural crop species, this results in a large variety reference collection which is increasing every year as more varieties are listed in the country of testing and within the EU. In order to maintain the strength of protection offered by Plant Breeder’s Rights (PBR), the principle of comparing new varieties with those of common knowledge must be upheld, and therefore some means of ‘managing’ reference collections is highly desirable to avoid the logistical and financial implications of having to include and directly compare all common knowledge varieties with candidate varieties. One means of such management would be to use molecular markers (DNA profiling) to compare new varieties with the profiles of those in a database, eliminate those which do not need to be compared in a field trial (according to pre-defined criteria) and then only grow the most similar varieties.

According to the UPOV BMT Guidelines, ‘Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics’ would be an acceptable method for use in the management of reference collections, provided there would

be no significant shift in the typical minimum distances as measured by traditional characteristics. Previous research aimed at assessing this approach has shown little or no correlation between phenotypic and genetic distances <sup>(1,2)</sup>. A reason for this lack of correlation could be that these previous studies have used very small numbers of markers (mostly less than 30). Low genome coverage means that the studies were much less likely to identify significant correlations than if a large number of markers with very good genome coverage were used.

At the UPOV Working Group for Biochemical and Molecular Techniques (BMT) meeting in Ottawa in May 2010, it was reported that the UPOV Technical Committee had considered the conclusions of the BMT Review Group and recognized the need for further work to examine the assumptions made for this approach and to improve the knowledge of the relationship between morphological and molecular distances. There is currently no working model acceptable to UPOV for ‘Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics’ due to the lack of correlation seen in previous studies with oilseed rape and potato. The aim of this project was to test for a correlation between morphological distances and molecular distances in barley while employing methods with higher genome coverage than those previously used.

DUS testing of barley within the European Union (EU) countries follows CPVO-TP 019/2 with additional National guidelines Guideline characteristics for National Listing. Although the characteristics to be recorded in barley are thus harmonised, there are varying approaches to the testing adopted in different Member States (MS), and various sets of ‘national’ characteristics used. Testing follows Article 7 of the 1991 UPOV Convention which says that a variety shall be considered Distinct ‘...if it is clearly distinguishable from any other variety whose existence is a matter of common knowledge at the time of the filing of the application’. Common knowledge is broadly defined as including all known varieties, i.e. any variety entered into or subject to an application for PBR, varieties grown commercially, varieties held in publicly accessible reference collections, or of which there is a published description.

A major problem for all countries carrying out DUS tests is the requirement to compare new varieties with an increasing number of existing varieties. In order to maintain the efficacy of the system for granting PBR, the reference collection should be as large as possible. Whilst

in theory, the full reference collection to be used for comparison purposes for any candidate variety is the known world-wide collection of varieties of the species, in practice, the number of varieties to be included in a growing test can be reduced. UPOV TG/1/3 (2002) allows that ‘a systematic individual comparison may not be required with all varieties of common knowledge. For example, where a candidate variety is sufficiently different, in the expression of its characteristics, to ensure that it is distinct from a particular group (or groups) of varieties of common knowledge, it would not be necessary for a systematic individual comparison with the varieties in that group (or those groups).’ UPOV TG/1/3 (2002) continues by indicating that the selection can usually be further narrowed down by using documented variety descriptions and the information on the most similar varieties supplied by the breeder in the Technical Questionnaire which accompanies the application for testing. Thus a testing authority can use a range of sources of information to limit the number of varieties from the reference collection which must be used in the field growing test <sup>(3)</sup>.

One possible way of limiting the number of reference varieties to be grown is to use DNA profiling as a management tool. By comparing the profiles of candidate varieties with those of existing varieties maintained in a central database, it might be possible both to eliminate from further testing those varieties which do not require comparison in a field trial (according to an agreed set of criteria) and to select the varieties most similar to the candidate for close comparison in field tests <sup>(4,5)</sup>.

### ***State of the art – molecular markers and DUS testing***

A whole range of studies on markers within the variety registration process has been carried out on different species. Potential uses of molecular technology include their application in the management of reference collections, for variety identification, infringement cases and examining essential derivation.

In a study of grapevine <sup>(6)</sup>, 991 cultivars were assayed with nine microsatellite loci. Pair wise comparisons showed these markers offered unique identification for 352 accessions. The remaining 639 accessions were assayed with a further 16 loci. The authors conclude that it is possible to calibrate a minimum distance between varieties using microsatellites in a variety set that included closely related varieties (parents, progeny, full sibs, half sibs, grandparents etc.) with a difference greater than four alleles in all but 10 out of 119,316 pair wise comparisons. However, essentially derived varieties (EDVs) could not be differentiated to the

same degree, and a difference of only two alleles was observed between varieties. The robustness of decisions made using an inter variety distance of two alleles needs to be tempered by the observation of intra variety differences of one allele. The authors concluded that variety pairs that exceed a minimum threshold using molecular methods may be declared Distinct (D) but where no or few differences in molecular profiles exist, further testing is required either by the use of additional markers or by comparing morphologies. This equates to an approach allowing an initial screen that would increase efficiency of DUS testing by eliminating the number of comparisons that would need to be made in the field but would not allow full replacement of the current test system. The authors recommend that minimum distances using markers can only be established experimentally, on a crop by crop basis, taking into consideration the inter and intra variety variability of the test system used.

An alternative experimental approach was used to study durum wheat lines (<sup>7</sup>) where a collection of 69 breeding lines from seven crosses were assessed for distinctness using 17 morphological markers from 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,  $r = 0.62$ ). Notwithstanding these correlations, the authors recognised difficulties in assessing 'D' using a 'Model 2' approach because of the wide range of variation for molecular marker differences among varieties around or beneath the 'D' threshold using morphological markers. Once more, the authors concluded that the calibration of molecular and morphological methods would allow a declaration of 'D' where molecular profiles differ greatly in the style of approach but that field testing could not be eliminated.

Investigations into the correlations between morphology and molecular based distances in maize (<sup>8</sup>) examined a collection of 41 inbred lines comprising 13 publicly available varieties and 28 breeders' lines. Morphological descriptions were calculated using 34 characters from the UPOV guidelines and molecular distances calculated using data for 28 SSR loci. In this instance the correlation between morphology and molecular markers was poor ( $r = 0.21$ ). Once more the authors conclude that molecular markers are a possible addition to the DUS testing procedures but their implementation depends upon deciding on the type and number of markers to be used as well as setting the threshold values for distinctness.



A large, international set of varieties was examined in a CPVO co-funded study of oilseed rape (CPV5766 Final Report) using 335 records from DUS testing authorities in Denmark, France, Germany and the UK. The collection was genotyped with 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. Clearly these results offer little prospect for successfully implementing a UPOV BMT Model 2 approach.

However, there is an expectation of improved correlations between morphological and molecular marker based distances using high density polymorphism data, such as SNP markers generated using a SNP array. An SSR study conducted in a set of 40 winter wheat varieties showed that pair-wise discrimination increased as more SSR loci were considered (NIAB unpublished data). The initial rate of increase in discrimination was rapid but tailed off as marker numbers increased until additional markers offered no advantage. This can be explained by linkage between markers and population structure within the variety set. It is to be expected that the correlations between morphological and molecular marker based distances would improve in a similar way, reaching a plateau when an optimum number of markers have been used to calculate molecular distances. While the minimum number of markers required should be determined empirically for each species, it is possible that the marker numbers used in previous studies may be sub optimal.

DUS testing of barley is carried out in several EU MS according to the CPVO technical protocol for barley (CPVO-TP 019/2), however slightly different approaches are taken by different EOs. A total of 28 characters are routinely observed or measured in the UK. All of these are phenotypic characteristics, however electrophoresis (a characteristic in CPVO-TP 019/2) is sometimes used to establish distinctness where there is an indication of a small difference in phenotype between similar varieties. To date, research on barley using molecular markers as an aid to DUS testing has been promising. Research into the use of diagnostic markers for the vernalization requirement in barley has been successful in isolating a diagnostic assay for winter and spring seasonal types. At the BMT meeting in 2008 a paper was presented on the use of molecular distances in combination with phenotypic characteristics within GAIA (pre-selection software developed by GEVES, France) which showed that molecular markers can contribute to the management of the spring barley reference collection (<sup>9</sup>).

Also at the BMT meeting in 2008, a similar paper was presented on a method for combining phenotypic data with molecular data in maize, using GAIA. This was further considered at the BMT Review Group meeting in April 2009. It was concluded that this proposal ‘System for combining phenotypic and molecular distances in the management of variety collections’, for the management of variety collections, was acceptable within the terms of the UPOV Convention and would not undermine the effectiveness of protection offered under the UPOV system. It was also agreed that the proposal represented a model that might be applicable to other crops provided that the elements of the proposal were equally valid. The BMT Review Group concluded that it was important to consider on a case-by-case basis whether the model would be applicable, and noted that some of the elements of the proposal were similar to the previously-named Model 2 approach ‘Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics’. However, the BMT Review Group concluded that it would not be appropriate to classify the proposal under Model 2 and agreed that the proposal should be referred to as the ‘System for combining phenotypic and molecular distances in the management of variety collections’. These conclusions were presented at the BMT meeting in May 2010.

Recently, in barley, Food and Environment Research Agency (Fera), PVS funding permitted NIAB to explore the possibility of using a Model 1-type approach solely for the purpose of predicting seasonal growth type. A series of publications and reports established that while winter and spring types are easily recognized, the much rarer alternative type required more complex assays (<sup>10-13</sup>). The Fera project defined a protocol which correctly identifies most alternative types from molecular genotype and flags them for field evaluation to unambiguously class them as either ‘winter’ or ‘alternative’, thus providing the option to avoid most vernalization trials in most years. Results of the project were presented at the UPOV BMT in May 2010. The UPOV position was that National authorities could decide to implement if the new system complied with the criteria set out in document BMT/DUS Draft 3.

Significantly, seasonal growth habit is not the only DUS characteristic for which underlying genetic variation has been described at the gene level in barley. The row number locus, *Vrs1*, on chromosome 2H, was cloned in 2007 (<sup>14</sup>). Just three independent mutations in the gene have abolished *Vrs1* suppression of lateral spikelet fertility, and therefore the row number can

be unambiguously assigned using a diagnostic molecular marker assay. The nud gene was more recently cloned (<sup>15</sup>). For all lines investigated, the naked (hulless) phenotype in barley is governed by a single mutation in an ERF transcription factor. Recent work at NIAB has shown a 16bp deletion in the barley homologue of the maize R/B anthocyanin regulatory genes is completely diagnostic for the ability to produce anthocyanin coloration in awns and auricles (<sup>16</sup>), bringing the number of DUS characters (including one 'growing' character) which can be directly predicted from genotype to four.

NIAB was recently a partner in a collaborative project called Association Genetics of UK Elite Barley (AGOUEB) alongside the Scottish Crop Research Institute (SCRI), University of Birmingham, barley breeding companies and industrial partners from end-user industries. The AGOUEB project used association genetics to dissect the genetic control of characteristics by looking at the variation (at the DNA level) at sites across the whole barley genome and by taking a retrospective look at the variants of genes that exist within UK varieties and the genes that control important characteristics. Association of phenotypic and genotypic data was used to determine patterns of genetic control of quantitative, qualitative and pseudo-qualitative characteristics.

Furthermore, the AGOUEB project has defined very precise locations (and hence closely linked flanking SNP markers) in the barley genome for a further 3 characteristics (rachilla hair length, hairiness of leaf sheath, and ventral furrow hair) and we predict that the identities of these three major genes and the pertinent allelic variants will be identified in the very near future. If these studies come to a successful conclusion, the characteristics covered will be exclusively qualitative (2- or 3-state) characteristics and their low number will allow an accurate but limited classification of barley germplasm.

Model 2 approaches do not require complete understanding of the genetic architecture and variation of each individual DUS characteristic, and are therefore more likely to enter productive use in addressing the issues that face DUS testing systems such as:

1. The identification of the most similar reference varieties for comparison prior to the growing trial
2. The rationalisation of trial design
3. The harmonisation of DUS systems at the international level.

## OBJECTIVES

Clearly, better techniques for the use of molecular markers within the variety testing process have become available over the last two or three years. The overall rationale of this project was thus to test an alternative method of calibrating marker distances against phenotypic distances, potentially on a characteristic by characteristic basis using new techniques and data available.

Due to the specific developments within the AGOUEB project outlined above, namely,

1. Identification of genes underlying DUS characters and their variants;
2. Generation of a genome-wide SNP genotyping platform and a database which includes 479 UK barley varieties with high quality SNP data at 1,111 loci as well as their full DUS descriptions, a major opportunity presented itself to re-open the study of how well genetic distance, as measured by molecular markers, can predict phenotypic distance. The main objective of this project was to calculate the genetic and phenotypic distances between varieties using a combination of statistical methods and, for the purposes of DUS testing, to determine whether a correlation exists between the two to evaluate the Model 2 approach in barley. We addressed this objective by testing two hypotheses:

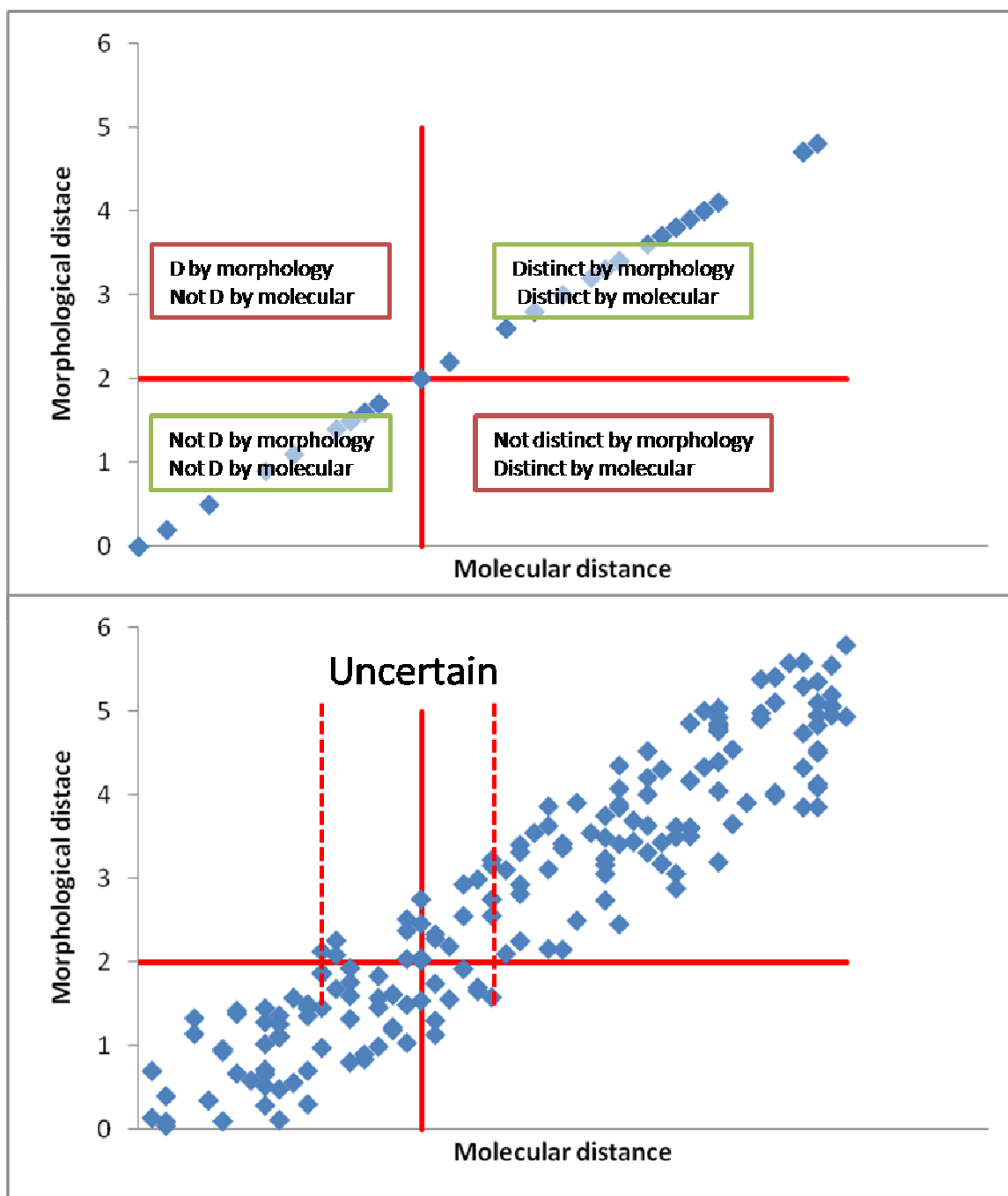
- Genotypic and phenotypic distance measures for a set of varieties will have a strong positive correlation to each other.
- Varieties shown as ‘similar’ using phenotypic distances will also be shown as ‘similar’ using genotypic distances.

This project used existing data from the collaborative AGOUEB research programme to investigate the UPOV Model 2 approach in DUS testing of barley with the aim of testing whether 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 must meet the quality criteria set out by UPOV in their ‘GUIDELINES FOR DNA-PROFILING’.

Ideally 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.

However, it is important not to overemphasise the importance of simple correlation between phenotypic and genotypic distances. The correlations already obtained may be ‘fit for purpose’. The success of UPOV Model 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. Plots of model data with correlation coefficients are shown in Figure 1. Both cases use a minimum phenotypic distance and molecular threshold of two. It is clear that the distribution of variation results in better decisions by molecular methods where the scatter and uncertainty of the correlation is greater. This is an area that was explored within this study.



**Figure 1: Calibration of molecular against morphological distances under UPOV BMT Model 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**

## METHODS AND RESULTS

The AGOUEB data used within this project was made up of 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 UK registration trials over the past 20 years (<sup>28</sup>). 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 and obtained from the examination office of that country.

The morphological data derived from DUS testing comprised 33 characteristics assessed for 579 varieties. The number of characteristics was reduced to 28 to reflect only those characteristics included in CPVO-TP/019/2 (2010) (see Table 1).

The morphological data was made up of quantitative characteristics converted into notes (e.g. plant height), pseudo-qualitative characteristics converted into notes (e.g. ear shape) and qualitative characteristics (e.g. grain: husk presence). This data set includes five grouping characteristics, omitting a sixth found in UPOV/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 comparisons of candidate varieties with other varieties. The ‘band width’ represents a minimum difference threshold for each characteristic that must be met when calculating differences. The variety comparisons must meet a certain threshold (a combination of minimum differences for each characteristic) in order to be considered as distinct.

**Table 1: Characteristics used in the 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 TG/19/10 alongside qualitative characteristics for the same character.**

Characteristic	UPOV No	Details	Band width
Plant: growth habit	1	Quantitative characteristic measure coded as a 1-9 scale	3
Lowest leaves: hairiness of leaf sheaths	2	Grouping characteristic scored as Present (9) or Absent (1)	1
Flag leaf: intensity of anthocyanin coloration of auricles	3*	Quantitative characteristic measure coded as a 1-9 scale	3
Plant: frequency of plants with recurved flag leaves	5	Quantitative characteristic measure coded as a 1-9 scale	3
Flag leaf: glaucosity of sheath	6	Quantitative characteristic measure coded as a 1-9 scale	3
Time of ear emergence	7	Quantitative characteristic measure coded as a 1-9 scale	2

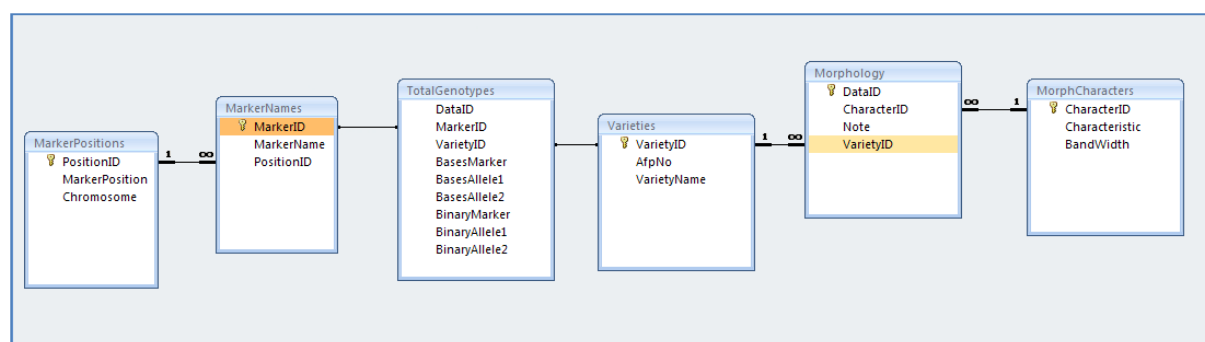
Characteristic	UPOV No	Details	Band width
Awns: intensity of anthocyanin coloration of tips	9*	Quantitative characteristic measure coded as a 1-9 scale	3
Ear: glaucosity	10	Quantitative characteristic measure coded as a 1-9 scale	3
Ear: attitude	11	Quantitative characteristic measure coded as a 1-9 scale	3
Plant: length	12	Quantitative characteristic measure coded as a 1-9 scale	2
Ear: number of rows	13	Grouping characteristic scored as Two-rows (1) or More than two rows (2)	1
Ear: shape	14	Pseudo-qualitative characteristic scored as one of three character states (tapering (3), parallel (5) or fusiform (7)).	3
Ear: density	15	Quantitative characteristic measure coded as a 1-9 scale	3
Ear: length	16	Quantitative characteristic measure coded as a 1-9 scale	3
Awn: length	17	Quantitative characteristic measure coded as a 1-9 scale	2
Rachis: length of first segment	18	Quantitative characteristic measure coded as a 3-7 scale	3
Rachis: curvature of first segment	19	Quantitative characteristic measure coded as a 1-9 scale	3
Ear: development of sterile spikelets	-	Qualitative characteristic scored as one of two character states (none or rudimentary (1) or full (2)).	1
Sterile spikelet: attitude	20	Quantitative characteristic measure coded as a 1-3 scale	2
Median spikelet: length of glume and its awn relative to grain	21	Quantitative characteristic measure coded as a 1-3 scale	2
Grain: rachilla hair type	22	Grouping characteristic scored as short (1) or long (2)	1
Grain: husk	23	Qualitative characteristic scored as absent (1) or present (9)	1
Grain: anthocyanin coloration of nerves of lemma	24	Quantitative characteristic measure coded as a 1-9 scale	3
Grain: spiculation of inner lateral nerves of dorsal side of lemma	25	Quantitative characteristic measure coded as a 1-9 scale	3
Grain: hairiness of ventral furrow	26	Grouping characteristic scored as absent (1) or present (9)	1
Grain: disposition of lodicules	27	Qualitative characteristic scored as frontal (1) or clasping (2)	1
Kernel: colour of aleuron layer	28	Quantitative characteristic scored as one of three character states (whitish (1), weakly coloured (2), strongly coloured (3)).	2
Seasonal type	29	Grouping characteristic scored as one of three character states (Winter type (1), alternative type (2), Spring type (3)).	2

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 UPOV Model 2.

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**

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 RODB package of the R statistics package.

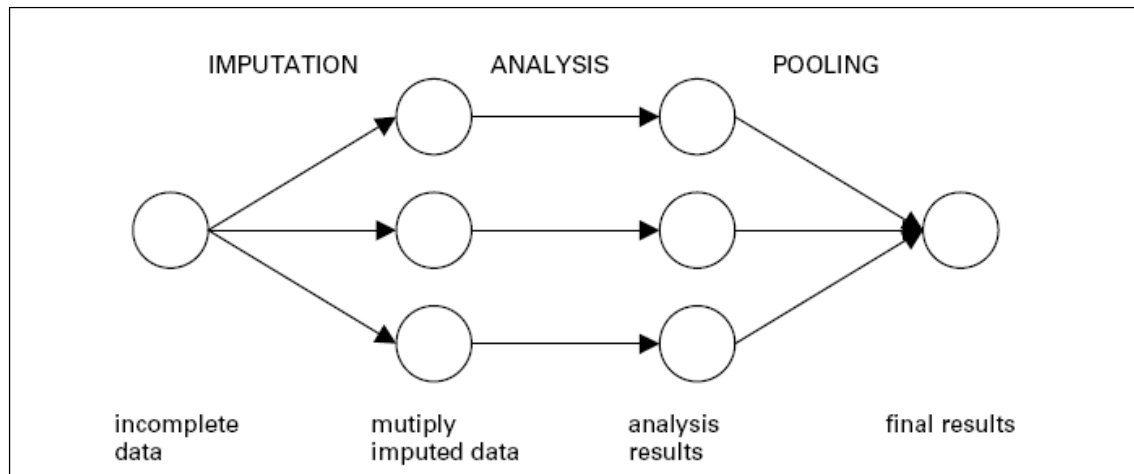
**Table 2: Genotype datasets selected in order to calculate various genotypic distances**

	<b>Data set</b>	<b>Number of loci</b>	<b>Criterion</b>
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. Multiple sets of makers were generated

	<b>Data set</b>	<b>Number of loci</b>	<b>Criterion</b>
O	Optimised evenly distributed markers	944	The set of markers selected from among the multiple sets of evenly distributed markers (N) for optimum correlation with morphological distances
Q	Optimised random markers	339	The set of markers selected from among full data set (A) for optimum correlation with morphological distances

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 results sets. The multiple results sets are pooled and conclusions drawn. In this case, we imputed phenotype data by random sampling and 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.

The data analysis was carried out using Microsoft Excel, ASReml <sup>(29)</sup> and the R Statistical Package (2010) including packages mice: Multivariate Imputation by Chained Equations <sup>(30)</sup> and cluster: Cluster Analysis Extended <sup>(31)</sup>. These packages were used to calculate the simple genetic distance matrices: 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 qualitative, pseudo qualitative and quantitative data.



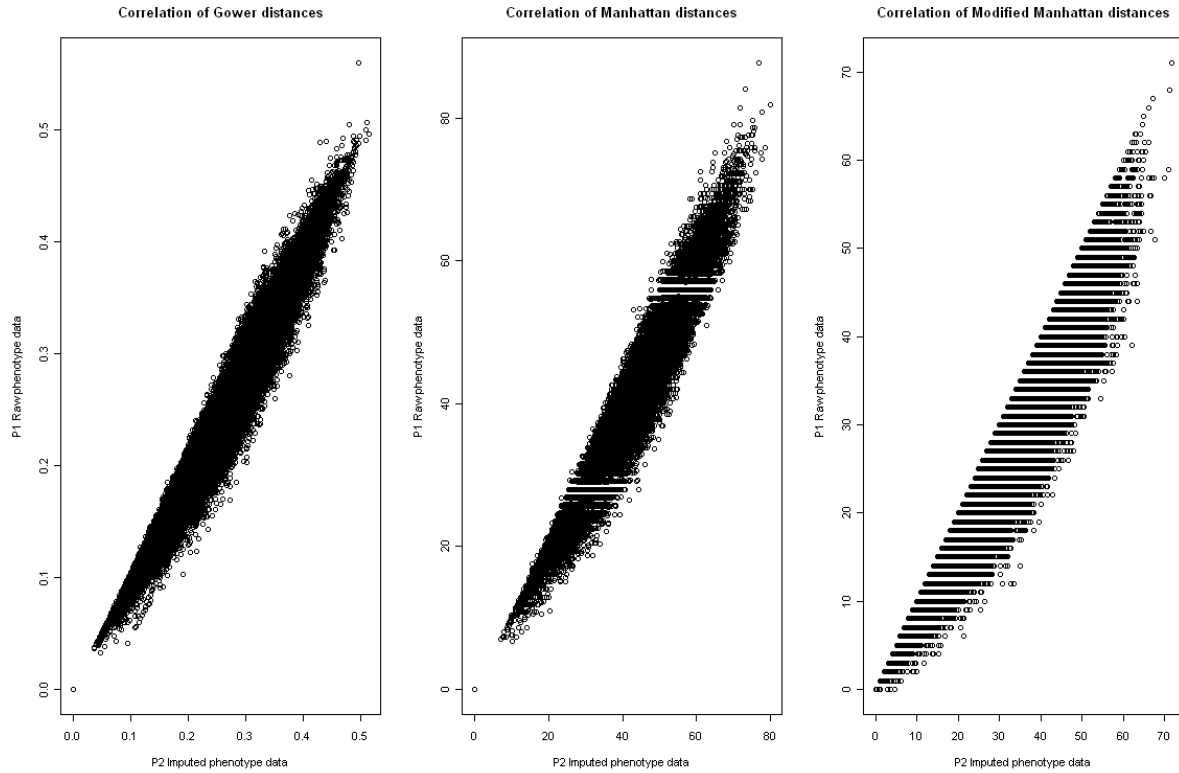
**Figure 3: A schematic of the work flow through the imputation process. (Figure from van Buuren and Oudshoorn, 1999)**

### *Validation of phenotypic datasets*

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		
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)**

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.

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, and 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***

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 was 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).

Figure 5 shows a clear pattern in every case. Initially, the correlations between the genotypic distances and the phenotypic distances increase with the number of markers. 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 the UPOV Model 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.

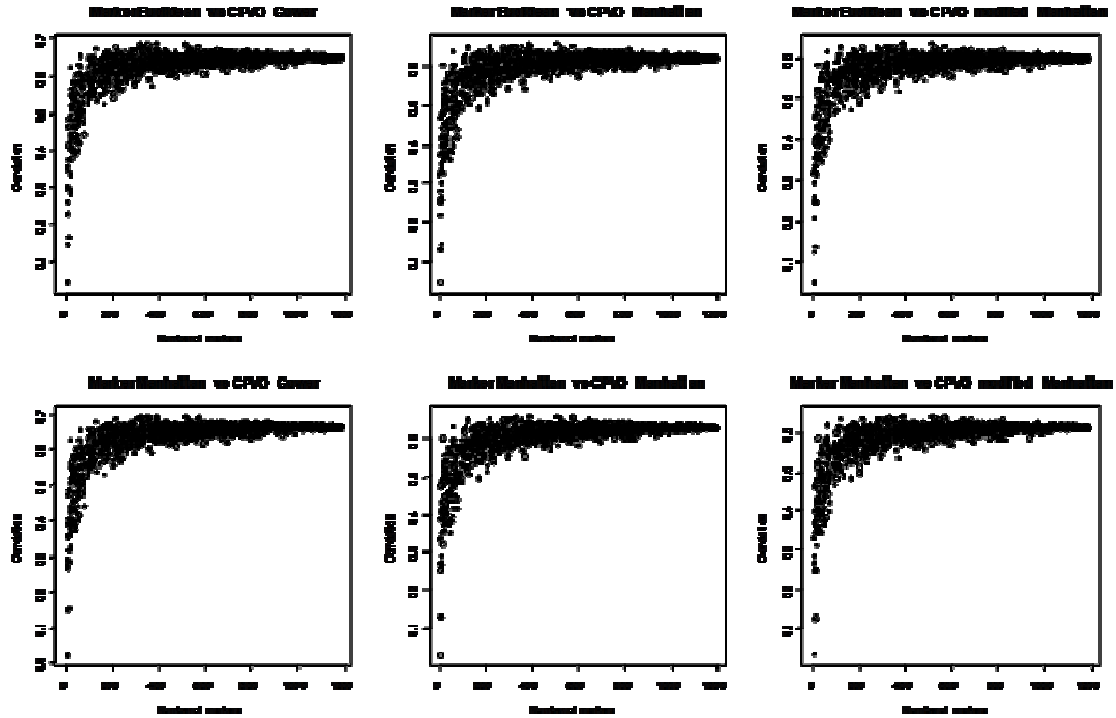
Here we present data showing the extent of correlation between the subsets of phenotypic and genotypic data using different methods to calculate distance matrices. The sets have been chosen to allow an investigation of factors that may affect the quality of the distance measures. We have used the raw 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.

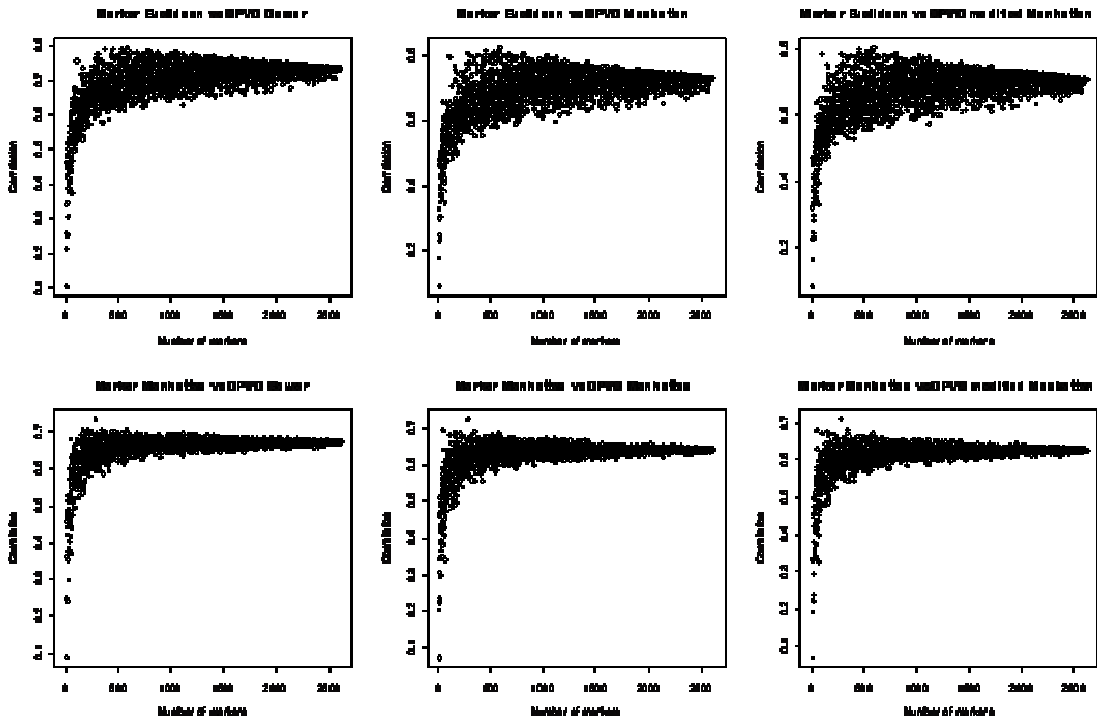
The correlations between phenotypic and genotypic distances are all positive. The correlations observed are greater than 0.55 with the exception of values obtained for genotype data sets E, G, K and M (defined in Table 2). These four data sets were selected to investigate whether correlations between phenotypic and genotypic distances improve if genetic loci harbouring rare alleles were used to calculate the genetic distances. The results in Table 4 and Table 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 correlations calculated using these data sets are compared with the scatters shown in Figure 5, the calculated values are systematically lower than the values that would be obtained by drawing an equivalent numbers of markers at random.

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.

These observed correlations in Table 4 and Table 5 are all positive but may not be described as strong. 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 are not 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.’



## 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
<b>Geonotypic distance: Manhattan</b>				
<b>A</b>	Full data set	0.638	0.622	0.596
<b>B</b>	No missing data	0.638	0.621	0.594
<b>C</b>	No missing data, no monomorphic	0.638	0.621	0.594
<b>D</b>	No missing data, no monomorphic, minor allele frequency >0.1	0.630	0.615	0.594
<b>E</b>	No missing data, no monomorphic, minor allele frequency <0.1	0.244	0.231	0.181
<b>F</b>	No missing data, no monomorphic, minor allele frequency >0.05	0.638	0.621	0.596
<b>G</b>	No missing data, no monomorphic, minor allele frequency <0.05	0.151	0.142	0.103
<b>H</b>	5% missing data	0.639	0.623	0.597
<b>I</b>	5% missing data, no monomorphic	0.640	0.624	0.597
<b>J</b>	5% missing data, no monomorphic, minor allele frequency >0.1	0.640	0.624	0.597
<b>K</b>	5% missing data, no monomorphic, minor allele frequency <0.1	0.263	0.250	0.207
<b>L</b>	5% missing data, no monomorphic, minor allele frequency >0.05	0.637	0.621	0.596
<b>M</b>	5% missing data, no monomorphic, minor allele frequency <0.05	0.224	0.210	0.169
<b>Geonotypic distance: Euclidean</b>				
<b>A</b>	Full data set	0.626	0.611	0.579
<b>B</b>	No missing data	0.628	0.612	0.578
<b>C</b>	No missing data, no monomorphic	0.628	0.612	0.578
<b>D</b>	No missing data, no monomorphic, minor allele frequency >0.1	0.621	0.607	0.580
<b>E</b>	No missing data, no monomorphic, minor allele frequency <0.1	0.232	0.220	0.172
<b>F</b>	No missing data, no monomorphic, minor allele frequency >0.05	0.628	0.613	0.581
<b>G</b>	No missing data, no monomorphic, minor allele frequency <0.05	0.161	0.151	0.111
<b>H</b>	5% missing data	0.627	0.612	0.579
<b>I</b>	5% missing data, no monomorphic	0.628	0.613	0.579
<b>J</b>	5% missing data, no monomorphic, minor allele frequency >0.1	0.628	0.613	0.579
<b>K</b>	5% missing data, no monomorphic, minor allele frequency <0.1	0.256	0.245	0.202
<b>L</b>	5% missing data, no monomorphic, minor allele frequency >0.05	0.626	0.611	0.579
<b>M</b>	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
<b>Geonotypic distance: Manhattan</b>				
<b>A</b>	<b>Full data set</b>	0.656	0.625	0.602
<b>B</b>	<b>No missing data</b>	0.656	0.624	0.598
<b>C</b>	<b>No missing data, no monomorphic</b>	0.656	0.624	0.598
<b>D</b>	<b>No missing data, no monomorphic, minor allele frequency &gt;0.1</b>	0.647	0.619	0.593
<b>E</b>	<b>No missing data, no monomorphic, minor allele frequency &lt;0.1</b>	0.255	0.219	0.213
<b>F</b>	<b>No missing data, no monomorphic, minor allele frequency &gt;0.05</b>	0.656	0.625	0.599
<b>G</b>	<b>No missing data, no monomorphic, minor allele frequency &lt;0.05</b>	0.158	0.127	0.120
<b>H</b>	<b>5% missing data</b>	0.657	0.627	0.603
<b>I</b>	<b>5% missing data, no monomorphic</b>	0.658	0.627	0.603
<b>J</b>	<b>5% missing data, no monomorphic, minor allele frequency &gt;0.1</b>	0.658	0.627	0.603
<b>K</b>	<b>5% missing data, no monomorphic, minor allele frequency &lt;0.1</b>	0.275	0.244	0.242
<b>L</b>	<b>5% missing data, no monomorphic, minor allele frequency &gt;0.05</b>	0.655	0.625	0.601
<b>M</b>	<b>5% missing data, no monomorphic, minor allele frequency &lt;0.05</b>	0.234	0.205	0.204
<b>Geonotypic distance: Euclidean</b>				
<b>A</b>	<b>Full data set</b>	0.642	0.615	0.582
<b>B</b>	<b>No missing data</b>	0.644	0.615	0.581
<b>C</b>	<b>No missing data, no monomorphic</b>	0.644	0.615	0.581
<b>D</b>	<b>No missing data, no monomorphic, minor allele frequency &gt;0.1</b>	0.637	0.612	0.578
<b>E</b>	<b>No missing data, no monomorphic, minor allele frequency &lt;0.1</b>	0.242	0.209	0.201
<b>F</b>	<b>No missing data, no monomorphic, minor allele frequency &gt;0.05</b>	0.644	0.616	0.582
<b>G</b>	<b>No missing data, no monomorphic, minor allele frequency &lt;0.05</b>	0.167	0.134	0.125
<b>H</b>	<b>5% missing data</b>	0.644	0.616	0.583
<b>I</b>	<b>5% missing data, no monomorphic</b>	0.645	0.616	0.584
<b>J</b>	<b>5% missing data, no monomorphic, minor allele frequency &gt;0.1</b>	0.645	0.616	0.584
<b>K</b>	<b>5% missing data, no monomorphic, minor allele frequency &lt;0.1</b>	0.268	0.239	0.234
<b>L</b>	<b>5% missing data, no monomorphic, minor allele frequency &gt;0.05</b>	0.642	0.615	0.582
<b>M</b>	<b>5% missing data, no monomorphic, minor allele frequency &lt;0.05</b>	0.234	0.206	0.203

### ***Marker optimisation***

The experiments run to investigate the effect of marker numbers on the correlation between phenotypic distance and genotypic distance suggest that good correlations could be obtained by selecting markers at random. Two sampling strategies were adopted to test this. In the first, markers were selected, at random, to represent each ‘mapped position’ within the full set of marker data. This strategy resulted in a relatively uniform distribution of markers across the genome. The second strategy simply sampled markers at random from the full set of marker data. The second method could, without constraint, sample co-located markers resulting in uneven sampling of markers across the genome.

The markers used in this study have been mapped across the barley genome to 944 map positions over seven chromosomes and are not evenly distributed across these map positions (Table 6)

**Table 6: Distribution of markers across the seven barley chromosomes**

<b>Chromosome</b>	<b>Length in cM</b>	<b>Number of map postions</b>	<b>Map positions with a single marker</b>	<b>Maximum no markers at a map position</b>
1	140.53	121	64	17
2	160.29	156	74	23
3	173.17	144	67	38
4	123.29	109	51	23
5	196.85	175	93	24
6	129.38	106	46	23
7	166.56	125	53	33

Markers were selected for each map position. Where a map position was represented by a single marker, that marker was always selected. Where a map position was represented by more than one marker, one marker was selected, at random, to represent that map position. The selected markers were used to calculate distance matrices and these distances were correlated with the morphological distances. This process was carried out for 2000 replications and a summary of the data obtained is shown in Table 7.

The optimum marker set was selected by interrogating the data to identify markers at each marker position that were frequently associated with high correlations. The upper quartile of the correlations was collated and, for each map position, the most frequently occurring marker was selected. The resulting set of 944 markers (Data set O: Optimised evenly distributed markers) were then used to calculated distance matrices which were, in turn, correlated against morphological distances (Table 8). The results for Data set O show a clear

improvement over the randomly selected spaced markers and over the correlations tabulated in Table 4 and Table 5.

**Table 7: Summary of correlations between marker and morphological distances obtained by randomly sampling markers at every map position**

Correlations using random spaced marker set (Data set N)						
Morphological Distance	Marker Manhattan distance			Marker Euclidean distance		
	Gower	Manhattan	Modified Manhattan	Gower	Manhattan	Modified Manhattan
Minimum	0.602	0.566	0.537	0.588	0.555	0.518
Median	0.638	0.604	0.576	0.624	0.593	0.558
Mean	0.637	0.604	0.576	0.624	0.593	0.557
Maximum	0.665	0.630	0.603	0.652	0.620	0.586
Correlations using optimised spaced marker set (Data set O)						
Data set P1: Raw phenotype data	0.696	0.681	0.650	0.686	0.673	0.636
Data set P2: Phenotype data with imputed missing values	0.716	0.688	0.670	0.705	0.680	0.657

The second strategy simply required random sampling of markers from within the full set of marker data. At the first step of each replication a random number was generated which would determine the number of markers drawn from the full set of marker data. In light of the information gathered while determining the minimum number of markers (Figure 5) the number of markers was constrained between 300 and 1400. Random markers sets were drawn in 50,000 replications with the set yielding the optimum correlations between marker and morphological distances recorded at each replication. The optimum correlations were obtained for a marker set comprising 339 markers.

**Table 8: Correlations using optimised random marker set**

	Marker Manhattan distance			Marker Euclidean distance		
	Gower	Manhattan	Modified Manhattan	Gower	Manhattan	Modified Manhattan
Data set P1: Raw phenotype data	0.675	0.659	0.634	0.670	0.656	0.626
Data set P2: Phenotype data with imputed missing values	0.698	0.673	0.652	0.692	0.671	0.642

Using these approaches we have calculated correlations between genotypic and phenotypic distance that exceed any previously reported in support of 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 UPOV Model 2. We have also shown that increasing marker numbers initially improves the correlation between genotypic and phenotypic distances but the rate of improvement in correlation decreases toward zero. This second conclusion is important as a guide to future research policy by DUS authorities; previously it has been hoped that increasing the number of markers would yield even better correlations, however we have shown that beyond an empirically discovered point more markers will not improve results.

### *Use of genomic prediction to calculate predicted morphological distances*

The 1991 Act of the UPOV Convention defines a variety as a group of plants that can be 'defined by the expression of the characteristics resulting from a given genotype or combination of genotypes' and can be 'distinguished from any other plant grouping by the expression of at least one of the said characteristics.' While there is an ideal that underlies UPOV BMT Model 1 that a characteristic will be the expression of genotypic variation at one locus in the genome, genomic prediction assumes that expression of genotypes at all loci will, to a greater or lesser extent, result in the expression of a characteristic. Genomic prediction requires a 'training set' of varieties where both genotypic and phenotypic data are available. Regression analysis within the training set allows quantification of the contribution of each marker to the expression of a characteristic, where phenotype is the sum of an effect contributed by each genetic locus.

$$Phenotype_i = \sum_{j=1}^n m_{ij}g_j$$

Where  $Phenotype_i$  is the predicted trait value for the  $i$ th line (equally the  $i$ th genotype),  $m_{ij}$  is the marker score for the  $j$ th marker for the  $i$ th line,  $g_j$  is the regression coefficient for the  $j$ th marker.

The results of this regression can be used to predict the expression of that characteristic in a 'test set' of varieties where genotypic data are available but phenotypic data are not. The coefficients of the quantitative contribution of each genetic locus may be applied subsequently to genetic variation at each locus in the test set to predict the expression of the

characteristic for each member of the test set. The process is repeated for each characteristic that makes up the phenotypic data.

We predicted the phenotype of each characteristic using ridge regression implemented in the ‘penalized’ package (<sup>32</sup>) within the R statistical package using linear regression. Linear regression was considered appropriate for the quantitative traits. The values used for the tuning parameter  $\lambda$  were determined by ten-fold cross-validation, repeating the analysis using a range of values for  $\lambda$ . On each of ten occasions the variety set was divided into a training set (90%) and a test set (10%) of varieties. Phenotype was regressed against genotype in the training set using each value for  $\lambda$  and calculated coefficients used to predict the phenotype from the genotype data. The optimum value for  $\lambda$  was that value for which the residual differences between the predicted and measured trait values were minimised (Figure 6). This empirically determined tuning parameter  $\lambda$  for each characteristic was used in the genomic prediction of phenotype datasets that were, in turn, used to calculate distance matrices.

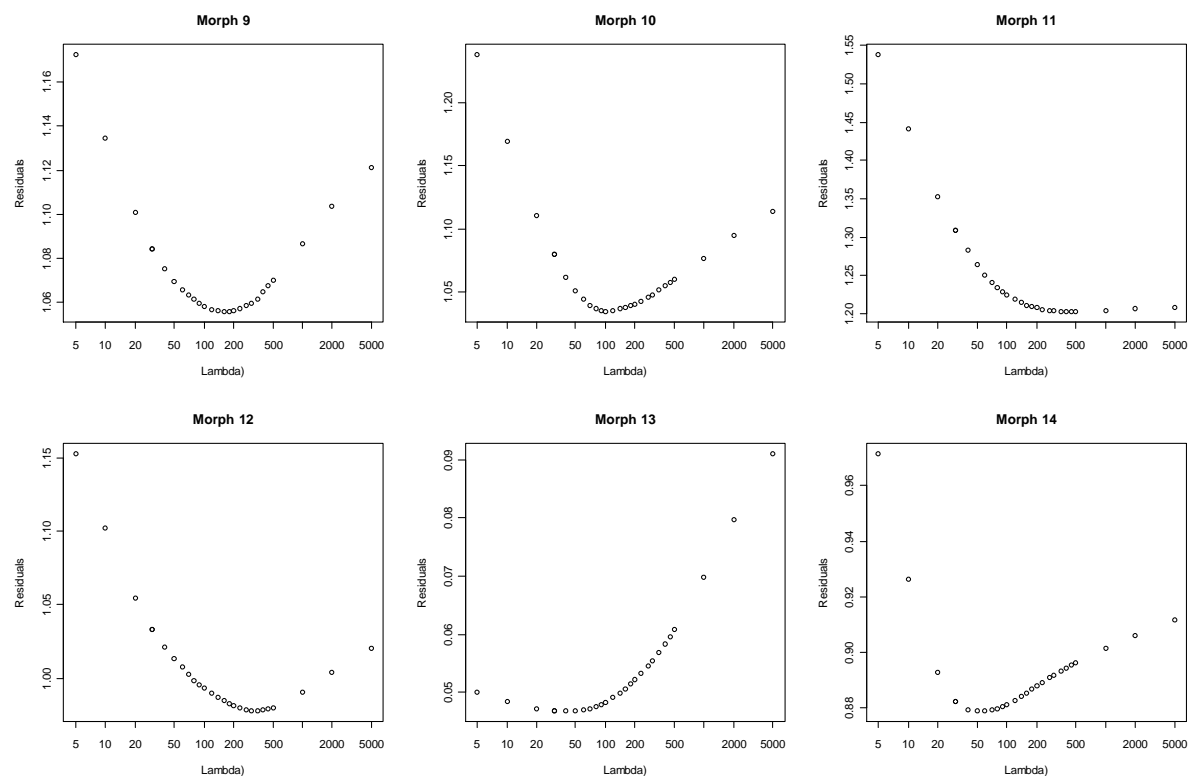
The correlations between predicted and measured characteristics were averaged over the ten iterations for the optimum value of  $\lambda$  (

Table 9). The correlations ranged between  $r = 0.140$  and  $r = 0.975$ . The UPOV convention states that characteristics must fulfil certain criteria to be selected for use in the DUS examination. ‘Characteristics should be a result of a given genotype or combination of genotypes; ....’ While we cannot assume that we have selected markers close the loci responsible for all of the characteristics in the morphology data set, the extent of linkage disequilibrium (LD) in elite barley suggests that many characteristics should correlate with at least some members of this dense set of markers. This makes it all the more surprising that we have not obtained better results for genomic prediction of individual characteristics and may open questions regarding the heritability of the characteristics used in DUS testing.

Genomic prediction was implemented using linear regression. We investigated whether logistic regression offered improved correlations between predicted and measured phenotypes for those ‘binary’ characteristics within the morphological datasets (2. Lowest leaves: hairiness of leaf sheaths, 13. Ear: number of rows, 22. Grain: rachilla hair type, 23. Grain: husk, 26. Grain: hairiness of ventral furrow and 27. Grain: disposition of lodicules). The correlations were higher when ‘penalized’ linear regression was implemented in all cases (Table 10).

**Table 9: Empirically derived optimum values for  $\lambda$  and the correlations between predicted and measured characteristics for that value of  $\lambda$ .**

UPOV No	Characteristic	Optimum values for $\lambda$	Average correlation (predicted vs measured characteristics)	St. Deviation of correlation
1	Plant: growth habit	300	0.661	0.054
2	Lowest leaves: hairiness of leaf sheaths	200	0.925	0.041
3*	Flag leaf: intensity of anthocyanin coloration of auricles	1000	0.459	0.153
5	Plant: frequency of plants with recurved flag leaves	200	0.250	0.112
6	Flag leaf: glaucosity of sheath	200	0.227	0.074
7	Time of ear emergence	200	0.295	0.142
9*	Awns: intensity of anthocyanin coloration of tips	200	0.445	0.202
10	Ear: glaucosity	100	0.504	0.217
11	Ear: attitude	200	0.274	0.140
12	Plant: length	300	0.288	0.106
13	Ear: number of rows	50	0.954	0.023
14	Ear: shape	50	0.140	0.098
15	Ear: density	300	0.293	0.077
16	Ear: length	300	0.285	0.141
17	Awn: length	200	0.393	0.118
18	Rachis: length of first segment	1000	0.329	0.089
19	Rachis: curvature of first segment	200	0.343	0.188
20	Sterile spikelet: attitude	100	0.682	0.080
21	Median spikelet: length of glume and its awn relative to grain	1000	0.256	0.108
22	Grain: rachilla hair type	100	0.572	0.190
23	Grain: husk	1000	0.201	0.120
24	Grain: anthocyanin coloration of nerves of lemma	50	0.698	0.055
25	Grain: spiculation of inner lateral nerves of dorsal side of lemma	50	0.773	0.084
26	Grain: hairiness of ventral furrow	50	0.746	0.071
27	Grain: disposition of lodicules	1000	0.554	0.219
28	Kernel: colour of aleuron layer	50	0.764	0.065
29	Seasonal type	100	0.975	0.023
-	Ear: development of sterile spikelets	100	0.738	0.094



**Figure 6: Plots for residual values (predicted – measured characteristics) vs tuning parameter ( $\lambda$ ). Optimum values for  $\lambda$  were identified when residuals were minimised.**

**Table 10: Comparison of correlations between predicted and measured phenotypes offered by logistic or linear regression implemented within the ‘penalized’ package.**

UPOV No	Characteristic	Logistic regression (r)	Linear regression (r)
2	Lowest leaves: hairiness of leaf sheaths	0.932	0.960
13	Ear: number of rows	0.946	0.971
22	Grain: rachilla hair type	0.699	0.833
23	Grain: husk	0.572	0.801
26	Grain: hairiness of ventral furrow	0.766	0.864
27	Grain: disposition of lodicules	0.518	0.800

Characteristic 14: Ear: shape, with three states (tapering (1,0), parallel (1,0) or fusiform (1,0)) was analysed using both linear and logistic regression and the characteristic re-composed from the results of analysis. This analysis offered no improvement in the correlations between predicted and measured phenotypes.

Genomic prediction was implemented selecting the training set and test sets in five different ways. In the first four instances the ‘training set’ was selected on a characteristic by characteristic basis and the ‘test set’ included all varieties. Firstly, the ‘training set’ was selected to include all varieties with complete phenotype data (Dataset R). In the next three

cases, the ‘training set’ was selected from among the varieties with complete phenotype data to include approximately one half (216, Dataset S), one quarter (108, Dataset T) and one eighth (54, Dataset U) of the number of varieties in the complete data set. In the fifth instance the ‘training set’ to include only those varieties where phenotype data was complete for all characteristics (196 varieties) and the ‘test set’ included only those varieties where phenotype data was incomplete for one or more characteristics (Dataset V). In all cases, Euclidean and Manhattan distance matrices were calculated from the predicted phenotype data calculated for each ‘test set’ and these matrices were, in turn, correlated against the three phenotypic distance matrices (Table 11). The Dataset P2: Phenotype data with imputed missing values was used for all correlations.

The results for data sets R, S, T and U are a clear improvement over any shown in Table 4 and Table 5 and this suggests that improved correlations have been obtained by novel statistical approaches. However, the ‘training set’ is a subset of the ‘test set’ for each of these data sets rather than being completely independent. If this method were implemented in the future then the ‘training set’ and ‘test set’ would be independent in the same way that they are independent for Dataset V, where the calculated correlations are no better than the best among those shown in Table 4 and Table 5.

**Table 11: Correlations between predicted and measured phenotypic distance matrices**

Genomic Predicted Phenotype Data set	Measured Phenotype	Correlations (r)				
		R	S	T	U	V
Euclidean distance	Gower distance	0.816	0.785	0.743	0.715	0.485
Euclidean distance	Manhattan distance	0.819	0.772	0.724	0.675	0.504
Euclidean distance	Modified Manhattan distance	0.819	0.772	0.724	0.675	0.500
Manhattan distance	Gower distance	0.855	0.812	0.765	0.725	0.488
Manhattan distance	Manhattan distance	0.842	0.789	0.735	0.683	0.512
Manhattan distance	Modified Manhattan distance	0.842	0.789	0.735	0.683	0.506

### ***Relationships within the variety set***

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](http://www.jic.ac.uk/germplas/bbsrc_ce/Pedb.txt)) and information taken from the Abstammungskatalog der Gerstensorten ([www.lfl.bayern.de/ipz/gerste/09740/gerstenstamm.php](http://www.lfl.bayern.de/ipz/gerste/09740/gerstenstamm.php)). Additional information was taken



from passport data held by germplasm collections including the Genebank of IPK Gatersleben ([http://gbis.ipk-gatersleben.de/gbis\\_i/](http://gbis.ipk-gatersleben.de/gbis_i/)) , 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.

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 (Table 12 and Table 13); 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 12: Mean phenotypic distances among sets of related varieties**

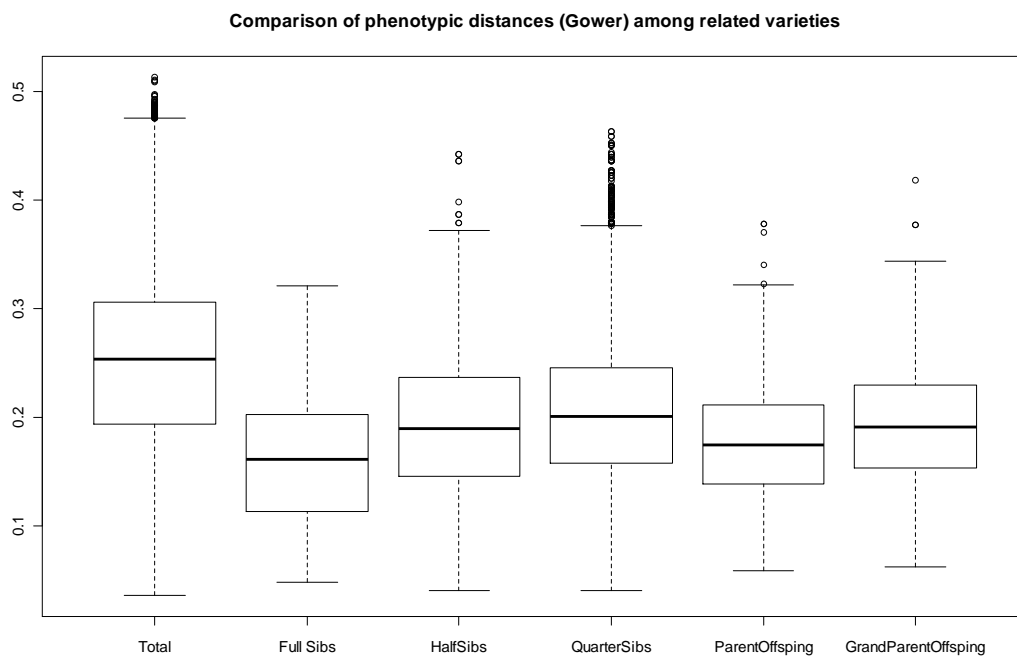
Average distances	Families	Pairs	Gower	Manhattan	Modified Manhattan
All varieties	NA	92665	0.25	38.87	29.31
Full siblings	28	67	0.16	25.67	16.74
Half siblings	126	2676	0.19	31.58	22.24
Quarter siblings	179	11975	0.20	33.04	23.60
Parent - offspring pairs	115	365	0.18	28.41	19.29
Grandparent - offspring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

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 distribution of the mean distances for the related sets is illustrated in Figure 7, Figure 8 and Figure 9.

**Table 13: Mean genotypic distances among sets of related varieties**

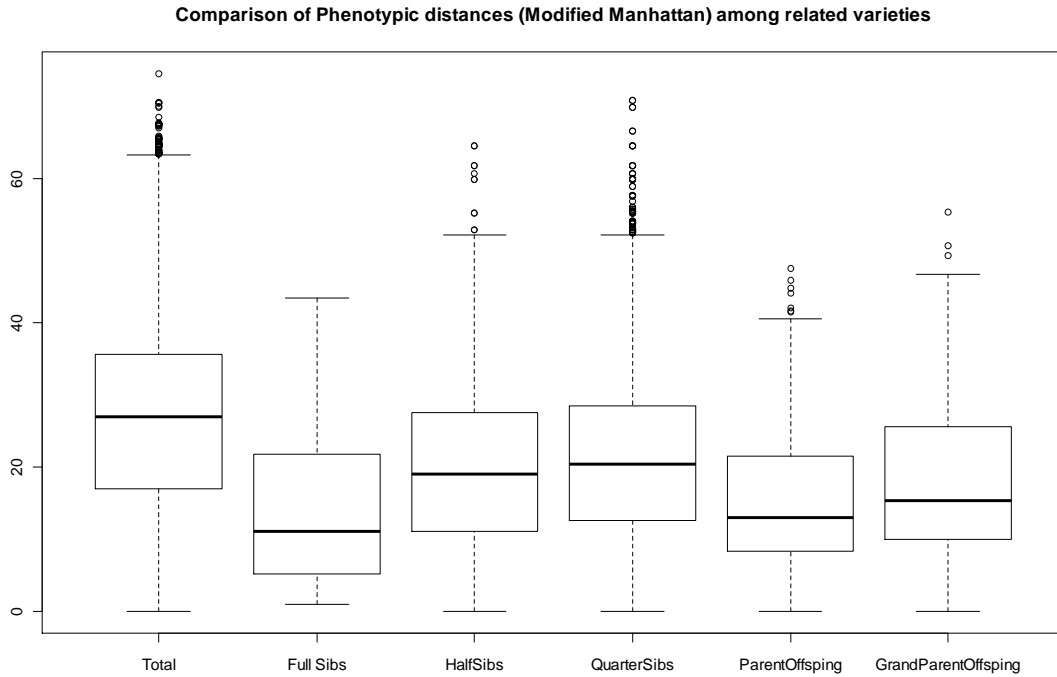
Average distances	Families	Pairs	Manhattan	Euclidean
All varieties	NA	92665	1567.7	39.3
Full siblings	28	67	639.7	24.5
Half siblings	126	2676	1025.2	31.7
Quarter siblings	179	11975	1106.0	33.0

Parent - offspring pairs	115	365	755.8	27.0
Grandparent - offspring pairs	67	327	1024.4	31.7

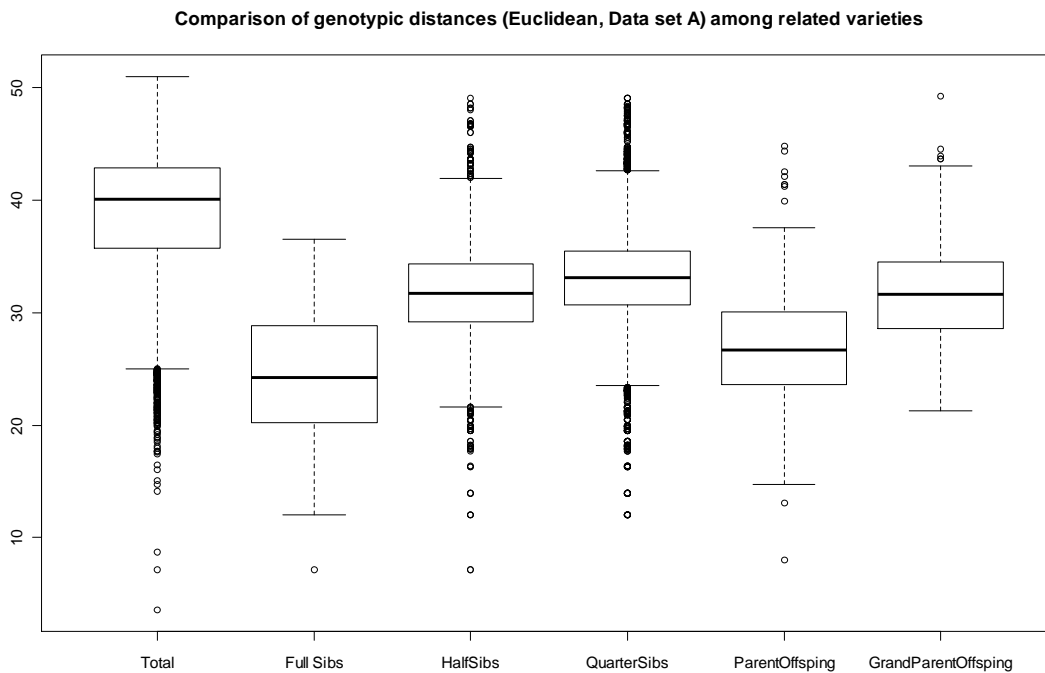


**Figure 7: Distribution of Gower's distances among the related sets**

In Figure 7, and Figure 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.

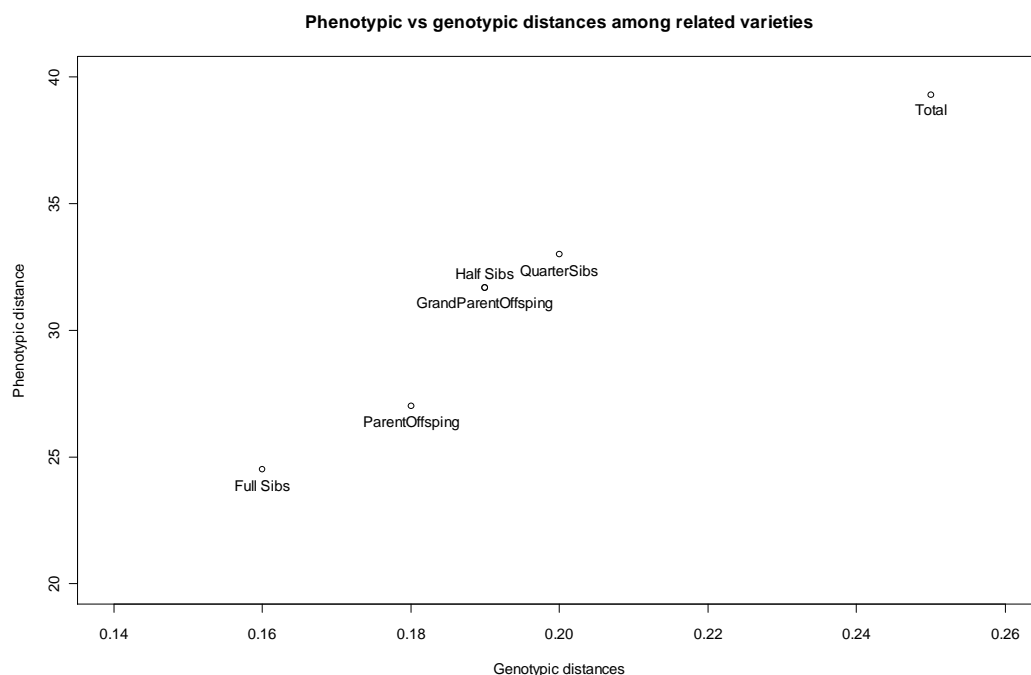


**Figure 8: Distribution of Modified Manhattan distances among the related sets**



**Figure 9: Distribution of genetic distances among the related sets**

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 UPOV Model 2 in the absence of 'noisy' data.



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

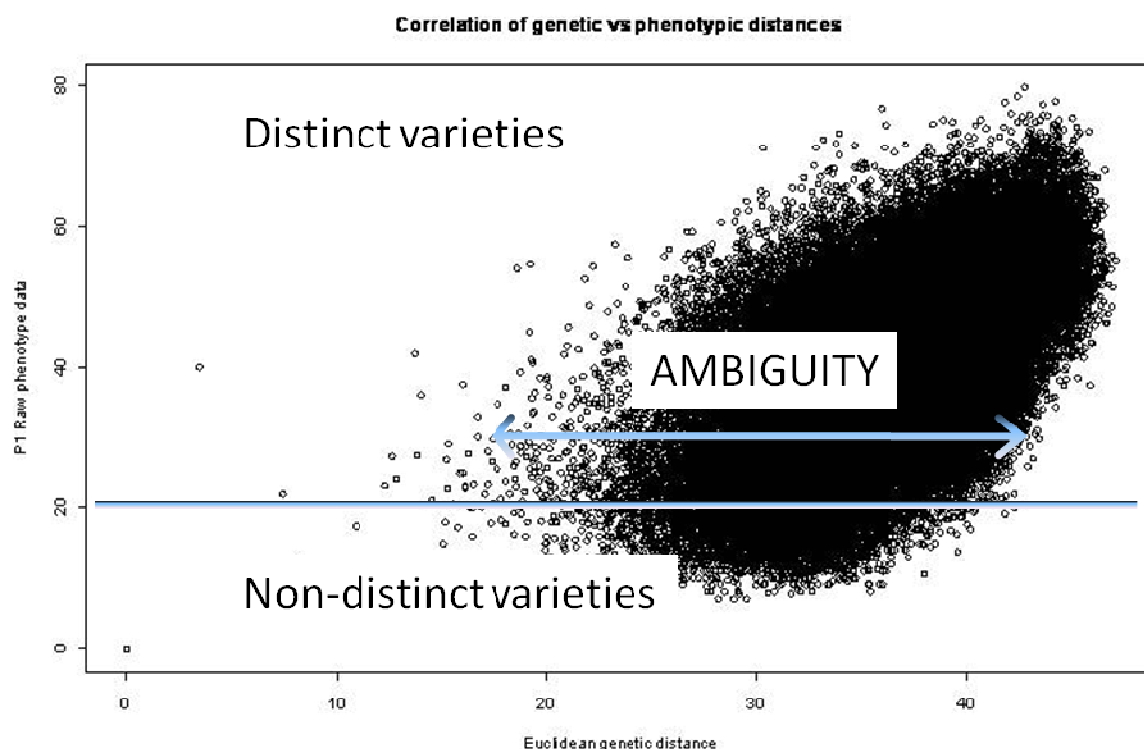
When the mean distances for each kinship group are correlated to a simple coefficient of relatedness (full siblings: 0.5, half siblings: 0.25, quarter siblings: 0.125, parent - offspring pairs: 0.5, grandparent - offspring pairs: 0.25), the correlations with morphological distanced are  $r = -0.94$  (Gower's Distance),  $r = -0.94$  (Manhattan Distance) and  $r = -0.90$  (Modified Manhattan Distance). The correlations for genetic distances fall into the range  $r = -0.96$  to  $r = -0.97$ .

### *Comparison of decision making using morphology or genotype data*

Here we examine the hypothesis that 'Varieties shown as 'similar' using phenotypic distances will also be shown as 'similar' using genotypic distances'. The 'typical' example data shown in Figure 11 illustrates the issue that needs to be resolved. Despite the positive correlation between phenotypic and genotypic distances, there will be ambiguity when comparing decisions made using morphological and genotypic data.

As all varieties within this dataset are distinct from each other it is not possible to assess DUS decisions at the conventional thresholds. An alternative approach was taken where an arbitrary threshold was set in order to declare 10% of varieties (43 varieties) as non distinct using the morphological data. This set of 'non-D' varieties was used as a bench mark for

comparisons made by setting thresholds for the genotypic data in an attempt to reproduce the decisions made using the morphological data. A series of threshold values were applied to the genetic distance matrices that would generate a series of ‘non-D’ variety sets with 43, 100, 200, 250, 300, 350 and 400 members. The decision making using phenotypic or genotypic data could be compared by simply counting the number of varieties that were described as ‘non-D’ by both methods.



**Figure 11: 'Typical' scatter of genetic vs phenotypic distances illustrating ambiguity when attempting to reproduce morphological distinctness decisions made using genotype data**

The ability to use genotype data to reproduce distinctness decisions made using morphology is shown when 43 ‘non-D’ varieties are identified using Gower’s Distance (Table 14), Manhattan Distance (Table 15) or Modified Manhattan Distance (Table 16) and compared with sets of ‘non-D’ varieties identified using genetic distances. When 43 ‘non-D’ varieties are identified using genetic distances, fewer than half the varieties appear in both the genotypic ‘non-D’ set and the morphology ‘non-D’ set. This clearly shows that the same decision will not be made using genetic distances or morphological distances. This is a setback regarding implementation of UPOV BMT Model 2 molecular methods as a direct replacement for the current system should the success criterion be that genotypic and morphological decisions correspond exactly. The decisions made using genomic prediction of

morphology (Datasets R, S and T) correspond most closely with those made using measured morphology but these results remain unsatisfactory.

**Table 14: Comparison of distinctness decisions: Gower's Distance vs genotypic distances**

			Number of genotypic 'non-D' varieties						
			43	100	200	250	300	350	400
Morphological distance :Gower. 43 'non-D' varieties									
		Genetic distance							
A	Full data set	Euclidean	11	24	39	43	43	43	43
		Manhattan	11	24	39	43	43	43	43
C	No missing data, no mono-morphic loci	Euclidean	12	29	41	42	43	43	43
		Manhattan	12	29	41	42	43	43	43
D	No missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	12	25	36	40	43	43	43
		Manhattan	12	25	36	40	43	43	43
E	No missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	6	11	32	37	40	41	43
		Manhattan	6	11	32	37	40	41	43
I	5% missing data, no mono-morphic loci	Euclidean	13	25	40	43	43	43	43
		Manhattan	13	25	40	43	43	43	43
J	5% missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	13	25	40	43	43	43	43
		Manhattan	13	25	40	43	43	43	43
K	5% missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	14	23	33	38	40	41	42
		Manhattan	14	23	33	38	40	41	42
O	Optimised evenly distributed markers	Euclidean	14	26	41	42	43	43	43
		Manhattan	14	26	41	42	43	43	43
R	Training set: all varieties with complete data	Euclidean	13	24	41	42	43	43	43
		Manhattan	20	35	41	43	43	43	43
S	Training set: 216 varieties with complete data	Euclidean	11	27	41	42	43	43	43
		Manhattan	18	32	41	43	43	43	43
T	Training set: 108 varieties with complete data	Euclidean	9	24	40	42	42	43	43
		Manhattan	14	29	41	42	42	43	43

**Table 15: Comparison of distinctness decisions: Manhattan Distance vs genotypic distances**

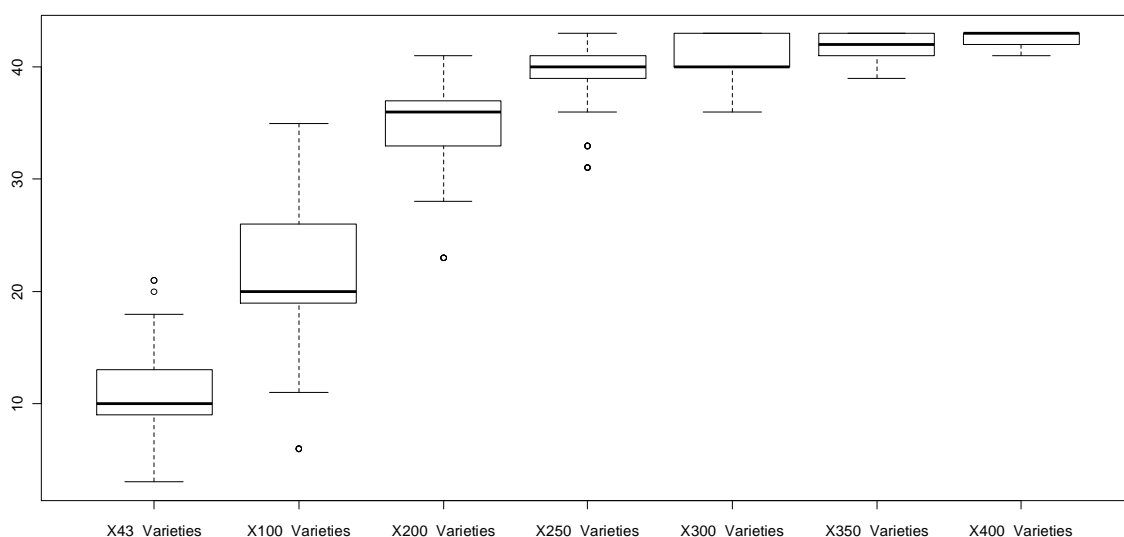
			Number of genotypic 'non-D' varieties						
			43	100	200	250	300	350	400
Morphological distance: Manhattan. 43 'non-D' varieties									
		Genetic distance							
A	Full data set	Euclidean	10	19	34	39	41	42	42
		Manhattan	10	19	34	39	41	42	42
C	No missing data, no mono-morphic loci	Euclidean	9	20	35	40	41	41	42
		Manhattan	9	20	35	40	41	41	42
D	No missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	9	15	31	37	40	40	43
		Manhattan	9	15	31	37	40	40	43
E	No missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	4	8	26	30	35	39	41
		Manhattan	4	8	26	30	35	39	41
I	5% missing data, no mono-morphic loci	Euclidean	10	19	35	41	41	41	42
		Manhattan	10	19	35	41	41	41	42
J	5% missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	10	18	35	41	41	41	42
		Manhattan	10	18	35	41	41	41	42
K	5% missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	10	18	29	33	37	41	41
		Manhattan	10	18	29	33	37	41	41
O	Optimised evenly distributed markers	Euclidean	11	20	37	39	41	41	43
		Manhattan	11	20	37	39	41	41	43
R	Training set: all varieties with complete data	Euclidean	15	24	37	39	40	43	43
		Manhattan	15	31	37	40	40	43	43

S	Training set:	Euclidean	10	25	37	38	40	43	43
	216 varieties with complete data	Manhattan	13	27	37	40	41	43	43
T	Training set:	Euclidean	9	21	36	39	39	42	43
	108 varieties with complete data	Manhattan	11	26	36	39	39	43	43

**Table 16: Comparison of distinctness decisions: Modified Manhattan Distance vs genotypic distances**

			Number of genotypic 'non-D' varieties						
			43	100	200	250	300	350	400
Morphological distance: Modified Manhattan. 43 'non-D' varieties									
Genetic distance									
A	Full data set	Euclidean	10	20	33	40	40	42	43
		Manhattan	10	20	33	40	40	42	43
C	No missing data, no mono-morphic loci	Euclidean	8	19	34	39	40	41	42
		Manhattan	8	19	34	39	40	41	42
D	No missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	9	17	32	36	39	40	41
		Manhattan	9	17	32	36	39	40	41
E	No missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	3	6	23	31	36	39	42
		Manhattan	3	6	23	31	36	39	42
I	5% missing data, no mono-morphic loci	Euclidean	10	19	34	40	40	41	43
		Manhattan	10	19	34	40	40	41	43
J	5% missing data, no mono-morphic loci, minor allele frequency >0.1	Euclidean	10	19	34	40	40	41	43
		Manhattan	10	19	34	40	40	41	43
K	5% missing data, no mono-morphic loci, minor allele frequency <0.1	Euclidean	7	17	28	33	38	41	42
		Manhattan	7	17	28	33	38	41	42
O	Optimised evenly distributed markers	Euclidean	10	20	37	40	41	41	42
		Manhattan	10	20	37	40	41	41	42
R	Training set: all varieties with complete data	Euclidean	18	28	37	39	42	43	43
		Manhattan	21	34	37	41	43	43	43
S	Training set: 216 varieties with complete data	Euclidean	13	29	37	40	42	43	43
		Manhattan	17	32	37	41	43	43	43
T	Training set: 108 varieties with complete data	Euclidean	11	25	36	39	40	42	43
		Manhattan	12	29	36	40	40	43	43

The possibility of adopting an approach using varieties identified by genotype as 'very different' was investigated by identifying further, larger sets of varieties using the genotypic data. Here we sought to determine what proportion of the variety set had to be identified as 'similar but distinct' using the genotype data before we could be confident that that we would not include varieties that are 'D' by morphology among the genotypic 'very different' varieties. Among the genotypic dataset tested, it is possible to select 400 (out of 431) varieties as 'similar but distinct' and still include one variety that is 'non-D' by morphology among the 'very different' identified by genotypic distances. Figure 12 show how the distribution of varieties that are 'non-D' by morphology and identified as 'non-D' or 'similar but distinct' by the genotypic data approaches the total that are 'non-D' by morphology as the number of varieties selected using the genotypic data increases. It is notable that complete convergence with the total that are 'non-D' by morphology is not achieved even when 93% of the variety set are selected.



**Figure 12: Distribution of varieties that are ‘non-D’ by morphology and indentified as ‘non-D’ or ‘low-D’ by the genotypic data**

## CONCLUSIONS

We have explored the interactions between morphological and genetic distances in a set of 431 elite UK barley varieties. We have used a set of high density SNP genotype data that broadly represents the whole barley genome. With 3072 loci, the marker set is an order of magnitude larger than any data set used in an exploration of UPOV BMT Model 2 previously reported. We used these data to test the 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

In all cases we demonstrated a positive correlation between genotypic and phenotypic distance measures for this set of varieties. When we selected genotype data on the basis of simple criteria such as % missing data, the optimum correlations with phenotypic distance measures were  $r = 0.64$  (Gower’s Distance),  $r = 0.62$  (Manhattan Distance) and  $r = 0.60$  (Modified Manhattan Distance). We achieved better correlations by selecting the ‘best marker’ at each mapped position across the genome ( $r = 0.72$  (Gower’s Distance),  $r = 0.69$  (Manhattan Distance) and  $r = 0.67$  (Modified Manhattan Distance)). However, we



demonstrated, by repeated sampling, that there was a ceiling to the correlations achievable by simple calculation of genetic distance measures such that the addition of additional markers is unlikely to offer a prospect of correlations above  $r = 0.70$ . This analysis would have to be tested for each crop species considered and the ceiling is likely to vary according to the extent of linkage disequilibrium within each crop genome.

Genomic prediction was attempted in order to investigate the possibility of breaking through this ceiling. The results reported, at first sight, offer considerable encouragement, achieving correlations of  $r = 0.86$  (Gower's Distance),  $r = 0.84$  (Manhattan Distance) and  $r = 0.84$  (Modified Manhattan Distance). This apparent success must be tempered by the lower results calculated when the 'training set' and 'test set' were truly independent. It is also notable that, when considered on a characteristic by characteristic basis there was considerable variation in the correlations between predicted and measured characteristics. This suggests there is considerable variation in the predictability of the characteristics and hence considerable variability on the quality of information when the characteristics are used in distinctness testing under the current system. Genomic prediction using methods such as ridge regression are relatively new and there are few published software packages available. There is considerable active research in this area with an expectation that novel methods are being developed and implemented in new software (Heslot *et al.*, 2012).

When varieties were grouped according to their pedigree relationships, the strong correlation observed between a coefficient of relatedness and genetic or morphological distances, offering support for both or either type of data as suitable for use in resolution of issues regarding Essentially Derived Varieties (EDV).

The essence of UPOV BMT Model 2 requires calibration of genetic distance measures to reproduce the decisions made using morphological distances. We have demonstrated that a one to one correspondence of distinctness decisions is not possible, even at the high levels of correlation between genetic and morphological distances achieved in this study. This result raises a question. What level of correspondence between distinctness decisions made using genetic and morphological distances would be required before UPOV BMT Model 2 could be implemented? This cannot be answered by simply addressing technical issues but is a question that can only be addressed by the plant breeders and DUS testing authorities. Any result other than a one to one correspondence of decisions results in risk to plant breeders where the quality of existing protection by Plant Breeders Rights is diminished if a novel

genetic threshold is set at too low a level or the ‘distinctness’ needed to acquire protection of a new variety is unreasonable diminished if a novel genetic threshold is set at too high a level.

We cannot advocate the immediate adoption of an UPOV BMT Model 2 approach on the basis of the results reported here. Nevertheless, we have shown greater correlations between morphological and marker based estimates of distance and a greater concordance between decisions made on the basis of trait and marker scores than have been reported previously. We feel, therefore, that these results are promising and that the approaches put forward here merit development. The current study was only possible because of the availability of trait and markers data collated as part of the AGOUEB study. The dataset has been adequate to demonstrate the promise of approaches based on high density SNP marker sets, but for several reasons it is not ideal for the following reasons:

- Based predominantly on UK varieties only
- Only varieties which had been National Listed were included.
- Number of varieties tested was relatively small
- Limited number of methods for phenotype prediction tested
- The marker panel used contains gaps.

#### ***Predominantly UK varieties***

Genetic diversity at both the marker and trait level is reduced. Aside from any direct effect this may have on estimating the merit of alternative methods, one advantage of methods based on cheap high density SNP platforms is the uniformity and consistency of scoring over countries, compared with measurement of traits and other marker systems such as SSRs where there is often considerable variation from lab to lab. The merit of the SNP platform is therefore underestimated.

#### ***Only varieties which had been national listed were included***

To judge UPOV BMT Models 1, 2, and 3, correlations or regression coefficients between the existing system and the system under test are compared. However, only a selected sample of varieties is included: national listed varieties which have already been found to be distinct *inter se*. This selection biases the estimates of correlation and regression downwards. It is

important to reduce this source of bias. Extending the range of varieties sampled will do this to some extent, but including varieties which have failed National Listing on distinctness would be of greatest benefit. Breeder cooperation or collaboration may be required for this. (Breeders may be able to supply varieties were they never submitted because they knew they would fail.)

### ***Small number of varieties***

The AGOUEB sample of varieties was 500. Compared to most studies this is large. However it is still underpowered. Cockram *et al* (2010) failed to detect loci for 14 characteristics in their study of DUS phenotypes, and this was attributed in part to low power. An increased population size will increase power directly, and this can be simple to achieve by expanding the AGOUEB study to include varieties from across Europe. An additional benefit is that if a repetition of the Cockram study on a larger scale is possible, it should result in greater success, and may make an UPOV BMT Model 1 approach more likely.

### ***Limited number of methods for phenotype prediction tested***

For genome-wide prediction methods, we have focussed on ridge regression, since this is a simple and easy to implement. More complex Bayesian methods are available and are more usual in animal breeding, and methods of penalized regression other than ridge should also be tested. Increasingly, software implementing these methods is becoming available. Moreover, where several methods exist among which it is hard to discriminate, methods of Model Averaging may give greater improvements. Related to this, it is possible that approaches can be developed in which marker information is incorporated into the existing system rather than being used as an alternative, and this may give rise to improved efficiencies while maintaining all the benefits of the current trait based system. An alternative approach in which morphological *distances* between varieties are predicted directly from markers, rather than indirectly from marker based predictions of characteristics may also offer considerable advantages. For example it may offer a simple means of selecting or weighting markers towards those which discriminate among varieties in the same manner as the morphological scores.

### ***Gaps in the marker panel***

The 3072 markers genotyped on the AGOUEB panel have been superseded by larger panels of markers. While the present study used re-sampling methods to demonstrate that the markers could be thinned with little loss, there are gaps in the panel (maximum 8.05 cM) which could be detrimental but which cannot be assessed with the existing data. Including markers within these gaps may increase the accuracy of the systems tested. Since we propose to genotype more lines, we should take the opportunity to achieve a more uniform coverage. This could be achieved by genotyping with the higher density SNP platforms which are now available, or by creating our own panel of markers specifically for DUS purposes.

### **PROPOSAL FOR FUTURE WORK**

We have demonstrated that good correlations between phenotypic and genotypic distances can be achieved on this regional set of varieties, but that Model 2 now needs to be developed on a larger scale in a European collaborative project. We propose that the current project be expanded to sample as large a range of European barley varieties as possible, to include varieties which have failed DUS (on D) in addition to varieties which have been successfully listed. This may require the co-operation of breeders to supply varieties which have not been National Listed. A European-wide dataset of DUS characteristics data will need to be created and collated. There may also be a need for additional characteristic scoring to resolve inconsistencies and to incorporate any varieties included specifically for this study. All varieties will be genotyped with a markers set broadly based on those used in the current study, but expanded to fill gaps. A greater range of characteristic prediction methods will be explored and methods to combine markers with characteristic scores will also be considered. This project has laid the foundation for a larger research project by demonstrating that better correlations than previously reported can be obtained from good quality molecular data.

### **REFERENCES**

1. Van der Linden, G, van Kaauwen, M, Voorrips, R, Kochieva, E, Vosman, B (2005)  
The Potential of SNP Markers in expressed genes for identification of potato varieties and determination of distinctness. BMT/9/13
2. Norris, C. *et. al* (2008) Management of winter oilseed rape reference collections.  
BMT/11/11

3. Barendrecht, C. J. (1999). The concept of grouping in UPOV Guidelines focused on ornamentals. *Plant Varieties & Seeds* 12: 143-148.
4. Tommasini L, Batley J, Arnold GM, Cooke RJ, Donini P, Lee D, Law JR, Lowe C, Moule C, Trick M, Edwards KJ (2003). The development of multiplex simple sequence repeats (SSR) markers to complement distinctness, uniformity and stability testing of rape (*Brassica napus L.*) varieties.) *Theoretical and Applied Genetics* 106, 1091-1101.
5. Jones, H, Jarman, R J, Austin, L, White, J and Cooke, R J (2003). The management of variety reference collections in distinctness, uniformity and stability testing of wheat. *Euphytica* 175-184.
6. Ibáñez J, Dolores Vélez M., Teresa de Andrés M, Borrego J (2009). Molecular markers for establishing distinctness in vegetatively propagated crops: a case study in grapevine. *Theor Appl Genet* **119**, 7. 1213-1222.
7. 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.
8. 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.
9. Collonier, C, Lallemand, J, Zhang, D, Godin, C, Jaudeau, B, Duriez, J, Belliard, A, Bouhedi, B, Guiard, J (2008) BMT/11/21
10. Cockram, J. et al. (2007) Haplotype analysis of vernalization loci in European barley germplasm reveals novel *VRN-H1* alleles and a predominant winter *VRN-H1/VRN-H2* multi-locus haplotype. *Theor Appl Genet* **115**, 993-1001.
11. Cockram, J. et al. (2008) Association mapping of partitioning loci in barley. *BMC Genet* **9**, 16.
12. Cockram, J., Norris, C, O'Sullivan, D.M. (2009) PCR-Based Markers Diagnostic for Spring and Winter Seasonal Growth Habit in Barley. *Crop Science* **49**, 403-410.
13. Cockram, J., Cooke, R.J., O'Sullivan, D.M. (2006) Functional SNP markers for the vernalisation requirement in barley. 9.
14. Komatsuda, T. et al. (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc Natl Acad Sci U S A*, 0608580104.

15. Taketa, S. et al. (2008) Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc Natl Acad Sci U S A* **105**, 4062-4067.
16. Heffner EL, Sorrells M.E , Jannink J.L; Genomic Selection for Crop Improvement *Crop Sci.* 2009 **49**:1–12)
17. Doveri S, Lee D, Maheswaran M, Powell W (2008). Molecular Markers - History, Features and Applications. In: Principles and Practices of Plant Genomics, Vol I Genome Mapping, Chapter 2. Editors: Kole C and Abbott AG, Science Publishers Inc., Enfield, New Hampshire, USA.
18. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**: 487-491.
19. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1992) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.* **18**: 6531-6535.
20. Jones H, Bernole A, Jensen LB, Horsnell RA, Law JR, Cooke RJ, Norris CE (2008a) Minimising inter-laboratory variation when constructing a unified molecular database of plant varieties in an allogamous crop. *Theor. Appl. Genet.* **117**:1335–1344.
21. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.* **23**: 4407-4414.
22. Law JR, Donini P, Koebner RMD, Reeves JC, Cooke RJ (1998) DNA profiling and plant variety registration. III: The statistical assessment of distinctness in wheat using amplified fragment length polymorphisms *Euphytica*, 102: 335-342.
23. Kalendar R, Grob T, Regina M, Suoniemi A, Schulman A (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theor. Appl. Genet.* **98**: 704-711.
24. Flavell AJ, Knox MR, Pearce SR, Ellis THN. (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *The Plant Journal* **16**: 643–650.
25. Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* **29**: e25.

26. Leigh F, Lea V, Law J, Wolters P, Powell W Donini P (2003) Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. *Euphytica* **133**: 359-366. DOI: 10.1023/A:1025778227751.
27. Close TJ, *et al.* (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*. 10.1186/1471-2164-10-582.
28. 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.
29. 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
30. van Buuren S, Oudshoorn K (1999) Flexible multivariate imputation by MICE. TNO Prevention and Health, Leiden, The Netherlands. TNO report: PG/VGZ/99.054
31. Cluster Analysis, extended original from Peter Rousseeuw, Anja Struyf and Mia Hubert. Version 1.14.0 Date 2011-06-07
32. Goeman, J. J.(2010) L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 52(1), 70-84.

## ACKNOWLEDGEMENTS

The authors would like to thank AGOUEB for allowing the use of its data for this project.