Theor Appl Genet (2013) 126:901–911
DOI 10.1007/s00122-012-2024-2

ORIGINAL PAPER

Evaluation of the use of high-density SNP genotyping to implement UPOV Model 2 for DUS testing in barley

Huw Jones · Carol Norris · David Smith · James Cockram · David Lee · Donal M. O'Sullivan · Ian Mackay

Received: 10 September 2012 / Accepted: 26 November 2012 / Published online: 12 December 2012

Abstract Developments in high-throughput genotyping provide an opportunity to explore the application of marker technology in distinctness, uniformity and stability (DUS) testing of new varieties. We have used a large set of molecular markers to assess the feasibility of a UPOV Model 2 approach: "Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics". We have examined 431 winter and spring barley varieties, with data from UK DUS trials comprising 28 characteristics, together with genotype data from 3072 SNP markers. Inter varietal distances were calculated and we found higher correlations between molecular and morphological distances than have been previously reported. When varieties were grouped by kinship, phenotypic and genotypic distances of these groups correlated well. We estimated the minimum marker numbers required and showed there was a ceiling after which the correlations do not improve. To investigate the possibility of breaking through this ceiling, we attempted genomic prediction of phenotypes from genotypes and higher correlations were achieved. We tested distinctness decisions made using either morphological or genotypic distances and found poor correspondence between each method.

Communicated by E. Carbonell.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-2024-2) contains supplementary material, which is available to authorized users.

H. Jones (✉) · C. Norris · D. Smith · J. Cockram · D. Lee · D. M. O'Sullivan · I. Mackay
NIAB, Huntingdon Road, Cambridge CB3 0LE, UK
e-mail: huw.jones@niab.com

Present Address:
C. Norris
Bayer CropScience Limited, 230 Cambridge Science Park, Milton Road, Cambridge CB4 0WB, UK

Introduction

The development of new crop varieties offers potential benefits, in terms of yield to growers and in quality improvements to end users. A new variety represents a considerable investment by plant breeders and this is sustained by commercial returns underpinned by sui generis protection of plant breeders' intellectual property rights. The International Union for the Protection of New Varieties of Plants (UPOV) is an intergovernmental organisation whose system of plant variety protection is intended to encourage innovation in the field of plant breeding. Variety registration and protection of crop varieties require distinctness, uniformity and stability (DUS) testing of new varieties. DUS testing is currently carried out by assessment of phenotypic characteristics where new varieties are compared with existing varieties. Developments in high-throughput genotyping have provided the opportunity to apply molecular marker technology within variety registration. The International Union for the Protection of New Varieties of Plants (UPOV) recognised the potential of these methods when it established the Biochemical and Molecular Techniques (BMT) Working Group. The BMT guidelines suggest three application models for molecular markers in variety registration (UPOV document INF/18/1 2011):

1. Molecular characteristics as a predictor of traditional characteristics: Use of molecular characteristics which are directly linked to traditional characteristics (gene specific markers),

2. Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics, and
3. development of a new system.

Success by the barley research community in cloning or fine-mapping a number of genes underlying variation in morphological traits has prompted the examination of how diagnostic polymorphisms may be employed in a UPOV BMT Model 1 approach to Distinctness in barley (Cockram et al. 2012). Although this study demonstrates how certain morphological traits may be assessed purely using molecular markers, it concluded that fully diagnostic markers are still too few to provide the necessary discrimination to inform distinctness decisions. UPOV BMT Model 2 requires "Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics". This requirement should ensure that decisions made under a molecular testing system would reproduce those made using phenotypic characteristics. Underlying UPOV BMT Model 2 is an expectation of a strong correlation between inter variety distances calculated using molecular characteristics and traditional characteristics. The UPOV BMT Model 2 has been investigated in clonally propagated (grapevine), open pollinating (maize and oilseed rape) and in predominantly self-pollinating (durum wheat and barley) crops. The outcomes of these investigations have been mixed. In grapevine (Ibáñez et al. 2009) it is possible to use microsatellites to calibrate a minimum shared allele distance among varieties produced by sexual reproduction in an accession set that included closely related varieties. Essentially derived varieties (EDVs) could not be differentiated in the same way. Variety pairs that exceeded a minimum molecular threshold could be declared distinct (D) but where differences in inter variety molecular profiles did not significantly exceed intra variety differences, further testing would be required in what amounts to a 'Super D' approach (Button 2008). In durum wheat (Noli et al. 2008) inter variety distances calculated using molecular characteristics (SSR and AFLP) and traditional characteristics were compared in a collection of 69 advanced lines from seven crosses. The correlation between the molecular distances (99 SSRs and AFLP) was good ($r = 0.89$) while the correlation between morphology and molecular markers was moderate (SSRs, $r = 0.66$; AFLP, $r = 0.62$) leading to the conclusion it was possible to describe variety pairs as distinct where molecular profiles differ greatly in a 'Super D' approach but field testing could not be eliminated. In maize (Gunjaca et al. 2008) the correlation between phenotypic and molecular distances, calculated using 28 SSR loci, was poor ($r = 0.21$). A study in a large, international set of oilseed rape varieties, genotyped using a suite of 29 SSR markers to calculate molecular distances and using records from DUS testing authorities to calculate phenotypic distances (CPV5766 Final Report 2008). The outcome of this study was far more disappointing, with the correlation between phenotypic and molecular marker-based distances falling between 0.03 and 0.08, depending on the methods used to calculate the distances. Taken together, these results offer little prospect for successfully implementing a UPOV BMT Model 2 approach.

Here we report the use of molecular markers to assess the feasibility of a UPOV Model 2: "Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics".

Genotype data derived from whole genome association scans in barley (Waugh et al. 2009) from the AGOUEB project consisted of 3072 SNP markers from 490 UK barley varieties and phenotype data from UK DUS trials comprising 33 characteristics, of which 28 were CPVO characteristics from 579 winter and spring barley varieties. The final consolidated dataset, taking into account missing data points, consisted of 431 varieties with both genotypic and phenotypic data. The data were analysed to quantify correlations between phenotypic and genotypic distances and compare phenotypic and genotypic distances against a common standard derived from known pedigree relationships within the dataset.

Materials and methods

The project used data 3072 SNP loci collected in the course of the AGOUEB project (http://www.agoueb.org/) for a collection of 490 barley varieties selected from UK registration trials over the past 20 years (Cockram et al. 2010). These SNP 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 two markers/cM across the approximately 1,100-cM barley genome (Close et al. 2009). 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. 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 DUS field examinations at NIAB, though a small number of descriptions were obtained by bilateral purchase from DUS authorities in another country. We considered only those characteristics included in CPVO-TP/019/3 (2012). These datasets were united to produce a final set of 431 varieties with both phenotypic and genotypic data. The final data set was drawn from the molecular and phenotypic datasets 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 analysis was carried out using Microsoft Excel, ASReml (Gilmour et al. 1995) and the R Statistical Package (2010). The analysis required the R packages 'mice': Multivariate Imputation by Chained Equations (van Buuren and Groothuis-Oudshoorn 2011) and 'cluster': Cluster Analysis Extended (Struyf et al. 1997). These packages were used to calculate the simple genetic distance metrics: Manhattan and Euclidean Distances and simple phenotypic distances: Manhattan and Modified Manhattan Distances and Gower's Coefficient (1971). The Manhattan Distance was used to calculate phenotypic distances as it reflects the decision-making process used in DUS examinations. The Modified Manhattan Distance is a variation to the Manhattan Distance such that the value of the pair-wise comparison for a characteristic must meet or exceed a threshold value, termed the 'band width', if it is to be added to the inter variety distance. The value of the band width is set by experts at a level that ensures calculated differences are not an artefact of variation in the observation and recording system within and between years. Gower's coefficient was selected for its suitability when handling data sets that include binary, multistate and continuous data (Gower 1971). We predicted each phenotypic characteristic using ridge regression implemented in the 'R' 'penalized' package (Goeman 2010) within the R statistical package using linear regression. Linear regression was considered appropriate for the quantitative traits. The values used for the tuning parameter λ were determined by tenfold cross-validation. This empirically determined tuning parameter λ for each characteristic was used in the genomic prediction of phenotype datasets that were, in turn, used to calculate distance matrices.

There was a high proportion of missing phenotypic data in the final set. The risk of erroneously calculating low inter variety phenotypic distances, introduced by missing data, was reduced by creating data sets where missing values were replaced using the 'R' 'mice' package (van Buuren and Groothuis-Oudshoorn 2011). Missing phenotype data were replaced by values drawn at random from the existing data to generate 100 'complete' data-setsPhenotypic distance matrices were calculated for each data and the results pooled by averaging. Thus, phenotypic distances for two data sets were available for comparison, the raw phenotype data (P1) and a set where the missing values have been replaced in this way (P2).

The effect of missing data, rare alleles and uneven distributions of markers across the genome within the genotype data were investigated by creating ten subsets of genotypic data (Table 1). The first set represented all available data (A). Two sets were created by excluding all monomorphic loci and including all loci with no missing data (B) or including all loci with 5 % or less missing data (E). In order to investigate whether data sets comprising loci with highly imbalanced allele frequencies (for example allele frequencies = 0.9:0.1) offered different correlations to data sets comprising loci with balanced allele frequencies (for example allele frequencies = 0.6:0.4) we created four further data sets. Two sets excluding loci with highly imbalanced allele frequencies were partitioned from data sets B and E by selecting only those loci where the minor allele frequency was between 0.101 and 0.499 (Sets C and F) and two sets including only loci with highly imbalanced allele frequencies by selecting only those loci where the minor allele frequency was 0.100 or less (Sets D and G). The markers used in this study have been mapped across the barley genome to 944 map positions over seven chromosomes. The markers are not evenly distributed across these map positions with 448 map positions represented by a single marker and one map position, on chromosome 3 harbouring 38 SNP loci. The markers were sampled repeatedly in 2000 replications. 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 makers (Set H) were used to calculate distance matrices and these distances were correlated with the morphological distances. An optimum set was selected by interrogating the data to identify markers at each marker position that were frequently associated with high correlations. The resulting set of 944 markers (Data set I) were then used to calculate distance matrices which were, in turn, correlated against morphological distances. A final marker set was simply chosen by randomly sampling a random number (constrained between 300 and 1400) of markers from within the full set of marker data in 50,000 replications. The selected makers (Set J) were used to calculate distance matrices and these distances were correlated with the morphological distances.

The relationships amongst the varieties selected investigated by researching their pedigrees. 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 (http://www.jic.ac.uk/germplas/bbsrc_ce/Pedb.txt) and Abstammungskatalog der Gerstensorten (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/), the US 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.

Table 1 Genotype datasets selected to calculate genotypic distances

Data set	Number of loci	Criterion
A	3072	Full data set
B	1274	Data set A with all loci with any missing data and all monomorphic loci removed
C	905	Data set B excluding highly imbalanced loci (loci with the minor allele frequency $\leq$0.10)
D	369	Data set B including highly imbalanced loci (loci with the minor allele frequency $\leq$0.10)
E	2262	Data set A with all loci with missing data at 5 % or more and all monomorphic loci removed
F	1554	Data set E excluding highly imbalanced loci (loci with the minor allele frequency $\leq$0.10)
G	708	Data set E including highly imbalanced loci (loci with the minor allele frequency $\leq$0.10)
H	944	Evenly distributed markers: Markers were selected at random to represent each of 944 map positions. Multiple sets of makers were generated
I	944	Optimised evenly distributed markers: The set of markers selected from among the multiple sets of evenly distributed markers (H) for optimum correlation with morphological distances
J	339	Optimised random markers: The set of markers selected from among full data set (A) for optimum correlation with morphological distances. Multiple sets of makers were generated

Data set	Genomic prediction	Training set	Test set
K	3072	All varieties	All varieties
L	3072	Half of varieties (216)	All varieties
M	3072	One quarter of varieties (108)	All varieties
N	3072	One eigth of varieties (54)	All varieties
O	3072	Varieties with complete phenotype data (196)	Varieties where phenotype data was incomplete for one or more characteristics (235)

Table 2 Correlations between phenotypic distances and genotypic distances calculated using alternative data sets and alternative methods of calculation

Marker set		Data set P1: Phenotype data (with missing data)						Data set P2: Phenotype data (missing data replaced by sampling)					
		Genotypic distance: Manhattan			Genotypic distance: Euclidean			Genotypic distance: Manhattan			Genotypic distance: Euclidean		
		Gower	Manhattan	Modified Manhattan	Gower	Manhattan	Modified Manhattan	Gower	Manhattan	Modified Manhattan	Gower	Manhattan	Modified Manhattan
A		0.638	0.622	0.596	0.626	0.611	0.579	0.656	0.625	0.602	0.642	0.615	0.582
B		0.638	0.621	0.594	0.628	0.612	0.578	0.656	0.624	0.598	0.644	0.615	0.581
C		0.630	0.615	0.594	0.621	0.607	0.580	0.647	0.619	0.593	0.637	0.612	0.578
D		0.244	0.231	0.181	0.232	0.220	0.172	0.255	0.219	0.213	0.242	0.209	0.201
E		0.640	0.624	0.597	0.628	0.613	0.579	0.658	0.627	0.603	0.645	0.616	0.584
F		0.640	0.624	0.597	0.628	0.613	0.579	0.658	0.627	0.603	0.645	0.616	0.584
G		0.263	0.250	0.207	0.256	0.245	0.202	0.275	0.244	0.242	0.268	0.239	0.234
H	Mean correlation	–	–	–	–	–	–	0.637	0.604	0.576	0.624	0.593	0.557
H	Max correlation	–	–	–	–	–	–	0.665	0.630	0.603	0.652	0.620	0.586
I		0.696	0.681	0.650	0.686	0.673	0.636	0.716	0.688	0.670	0.705	0.680	0.657
J		0.675	0.659	0.634	0.670	0.656	0.626	0.698	0.673	0.652	0.692	0.671	0.642
K								0.855	0.842	0.842	0.816	0.819	0.819
L								0.812	0.789	0.789	0.785	0.772	0.772
M								0.765	0.735	0.735	0.743	0.724	0.724
N								0.725	0.683	0.683	0.715	0.675	0.675
O								0.488	0.512	0.506	0.485	0.504	0.5

As all varieties within this dataset have been granted Plant Breeders' Rights, they are distinct from each other, making it impossible to assess DUS decisions at the normal thresholds. In order to compare the decision making using morphology or genotype data we set an arbitrary threshold for phenotypic distances such that 10 % of the varieties (43 varieties) were 'not distinct' (non-D). This set of 'non-D' varieties was used as a benchmark 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.

Results and discussion

The correlations between phenotypic and genotypic distances are all positive. The correlations observed are greater than 0.57 with the exception of values obtained for genotype data sets D and G. Data sets D and G were selected to investigate whether correlations improve if genetic loci harbouring rare alleles were used to calculate the genetic distances. The results in Table 2 clearly show that this is not the case. It is possible that these low correlations are a consequence of selecting few markers

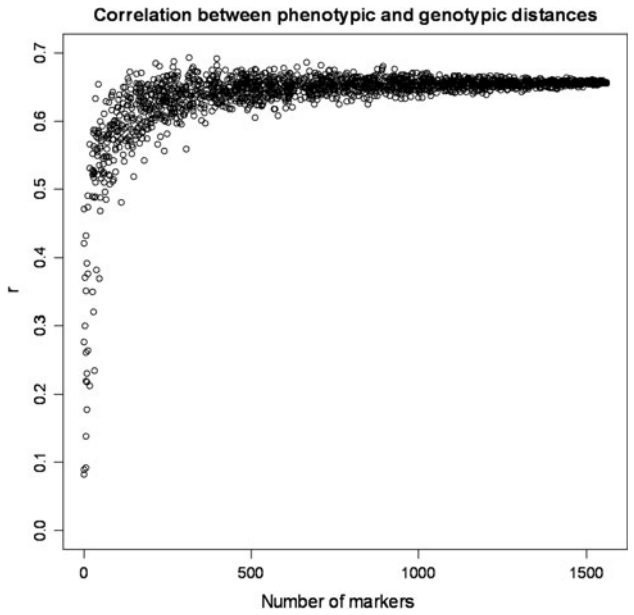

Fig. 1 Scatter plots of correlations between genotypic and phenotypic distances shows correlations improve as marker numbers increase until a ceiling is reached

(D = 369 markers, G = 708 markers). The correlations follow a pattern when considering the phenotypic distances, such that correlations using $r_{(\text{Gower Distance})} > r_{(\text{Manhattan Distance})} > r_{(\text{Modified Manhattan Distance})}$ ($p < 0.05$). The correlations follow a pattern when considering the genotypic distances such that $r_{(\text{Manhattan Distances})} > r_{(\text{Euclidean Distances})}$ ($p < 0.05$).

Results from previous studies suggest that better correlations between phenotypic and genotypic distances are obtained when genetic distances are calculated using many markers. The number of markers in this study is an order of magnitude greater than the number of markers used in previous studies. We investigated the effect of marker numbers on the correlation between phenotypic distance and genotypic distance by repeatedly selecting random sets of genotypic markers. Correlations between the genotypic distances and the phenotypic distances were calculated and tabulated with the number of markers for each random selection of markers. Scatter plots of calculated correlations against number of markers used show a clear pattern: initially, the correlations between the genotypic distances and the phenotypic distances increase as the number of markers used increases but 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 (Fig. 1). The low initial correlation values when small numbers of markers are used offer an explanation for the poor correlations observed in earlier studies. This result suggests that it is possible to determine the minimum number of markers needed to offer a reasonable prospect of achieving optimum correlations between phenotypic and genotypic distances. When the values obtained for correlations calculated using data sets with low minor allele frequencies (minor allele frequency <0.1, Set D and Set G) are compared with the scatters shown in Fig. 1, it can be seen that the calculated values are systematically lower than the values that would be obtained by drawing an equivalent numbers of markers at random.

The markers used in this study have been mapped across the barley genome to 944 map positions over seven chromosomes, but the markers are not evenly distributed across these map positions. We maximised the sampling of markers across the genome by selecting, at random, one marker to represent each map position for 2000 replications. The selected makers were used to calculate distance matrices and these distances were correlated with the morphological distances. The mean and maximum of the correlations obtained are shown in Table 2 and it may be seen that the mean correlation obtained in this way was comparable with correlations obtained for data sets A–C and E–F and the maximum correlations obtained were, in all cases, higher ($p < 0.05$). When an 'optimum' marker

set (I) was selected by interrogating the data to select markers at each marker position that were frequently associated with high correlations, the correlations obtained were consistently higher than those obtained for data sets A–C and E–F ($p < 0.05$). Finally, a set of markers was simply selected at random from within the full set of marker data in 50,000 replications. At the first step of each replication a random number was generated which would determine the number of markers drawn at random from the full set of marker data. The optimum correlations were obtained for a marker set comprising 339 markers.

Using these approaches we have calculated correlations between genotypic and phenotypic distance that exceed any previously reported, demonstrating the strong positive correlation between genotypic and phenotypic distance measures which 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 towards zero once an optimum number of markers have been used. This second conclusion is important as a guide to future research policy by DUS authorities; previously it had been hoped that increasing the number of markers would yield better correlations, but we have shown that beyond an empirically discovered point, simply increasing the number of markers will not improve results.

The genotypic distance calculations have given all markers an equal weight. 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". We have used genomic prediction to quantify the contribution of each marker within the genotype data to each characteristic within the phenotypic data on the assumption that expression of genotypes at all loci will, to a greater or lesser extent, result in the expression of a characteristic. Regression analysis in a 'training set' allows quantification of the contribution of each and every marker to expression of a characteristic, where phenotype is the sum of an effect contributed by each genetic locus

$$\text{Phenotype}_i = \sum_{j=1}^{n} m_{ij} g_i$$

where Phenotype_i is the predicted trait value for the ith line (equally the ith genotype), m_{ij} is the marker score for the jth marker for the ith line and is the regression coefficient for the jth marker. Variation in the regression coefficients (g_j) would, in effect, give the markers differing weights in the distance calculations. 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 was repeated for each characteristic that makes up the phenotypic data. Genomic prediction was implemented using linear regression. Logistic regression offered no improvement in correlations between predicted and measured phenotypes for those 'binary' characteristics within the morphological datasets. Initially, we tested genomic prediction for each phenotypic characteristic with tenfold cross-validation. On each of ten occasions the variety set was divided into a training set (90 %) and a test set (10 %) of accessions and the

Table 3 Correlations between predicted and measured characteristics achieved using genomic prediction

UPOV no.	Characteristic	Correlation (predicted vs. measured characteristics)
1	Plant: growth habit	0.661
2	Lowest leaves: hairiness of leaf sheaths	0.925
3	Flag leaf: intensity of anthocyanin coloration of auricles	0.459
5	Plant: frequency of plants with recurved flag leaves	0.250
6	Flag leaf: glaucosity of sheath	0.227
7	Time of ear emergence	0.295
9	Awns: intensity of anthocyanin coloration of tips	0.445
10	Ear: glaucosity	0.504
11	Ear: attitude	0.274
12	Plant: length	0.288
13	Ear: number of rows	0.954
14	Ear: shape	0.140
15	Ear: density	0.293
16	Ear: length	0.285
17	Awn: length	0.393
18	Rachis: length of first segment	0.329
19	Rachis: curvature of first segment	0.343
20	Sterile spikelet: attitude	0.682
21	Median spikelet: length of glume and its awn relative to grain	0.256
22	Grain: rachilla hair type	0.572
23	Grain: husk	0.201
24	Grain: anthocyanin coloration of nerves of lemma	0.698
25	Grain: spiculation of inner lateral nerves of dorsal side of lemma	0.773
26	Grain: hairiness of ventral furrow	0.746
27	Grain: disposition of lodicules	0.554
28	Kernel: colour of aleuron layer	0.764
29	Seasonal type	0.975
–	Ear: development of sterile spikelets	0.738

predicted phenotype correlated with the actual phenotype and the mean correlations tabulated (Table 3). 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 traits used in DUS testing.

When testing correlations between phenotypic distances and phenotypic distances predicted from the genotypic data, 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. First, the 'training set' was selected to include all varieties with complete phenotype data (Dataset K). In the remaining three cases, the 'training set' was selected from among the varieties with complete phenotype data to include approximately one half (L), one quarter (M) and one-eighth (N) of the number of varieties in the complete data set. In this way we represented a scenario where the addition of candidate varieties, year on year, would increase the number of lines in the 'test set' relative to the number in the 'training set'. However, the requirement to calculate distances among all varieties, including both 'test set' and 'training set' varieties, compromises the independence of the 'training set' from the 'test set'. In the fifth instance the 'training set' to include only those varieties where phenotype data were complete for all characteristics (196 varieties) and the 'test set' selected to include only those varieties where phenotype data was incomplete for one or more characteristics (O). 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 2).

The correlations for data sets generated using genomic prediction (Sets K, L, M and N) are a clear improvement ($p < 0.05$) over any of the other correlations shown in Table 2, suggesting 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 and the correlations decrease as the 'test set' increases in number in comparison with the 'training set'. If this method were implemented with the 'training set' and 'test set' completely independent, (Dataset V) the calculated correlations are observed to be lower than the best among those shown in Table 2.

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; 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, averaged and tabulated by relationship (Table 4). The correlations between phenotypic and genotypic distances and the correlations between kinship and all distances suggest that UPOV Model 2 has potential to succeed in the absence of 'noisy' data.

Having optimised the correlations between phenotypic and genotypic distances we can consider the quality of

Table 4 Mean phenotypic or genotypic distances among sets of related varieties as shown by their kinship and correlations between distance measures for varieties grouped by their kinship

Average distances	Families	Pairs	Phenotypic distances			Genotypic distances		Kinship
			Gower	Manhattan	Modified Manhattan	Manhattan	Euclidean	
All varieties	NA	92665	0.25	38.87	29.31	1567.7	39.3	0
Full siblings	28	67	0.16	25.67	16.74	639.7	24.5	0.50
Half siblings	126	2676	0.19	31.58	22.24	1025.2	31.7	0.25
Quarter siblings	179	11975	0.20	33.04	23.60	1106.0	33.0	0.125
Parent–offspring pairs	115	365	0.18	28.41	19.29	755.8	27.0	0.50
Grandparent–offspring pairs	67	327	0.19	30.76	21.79	1024.4	31.7	0.25
Correlation to genotypic distance		Manhattan	0.98	0.99	1.00			
Correlation to genotypic distance		Euclidean	0.96	0.99	0.99			
Correlation to kinship			−0.89	−0.95	−0.95	−0.95	−0.97	

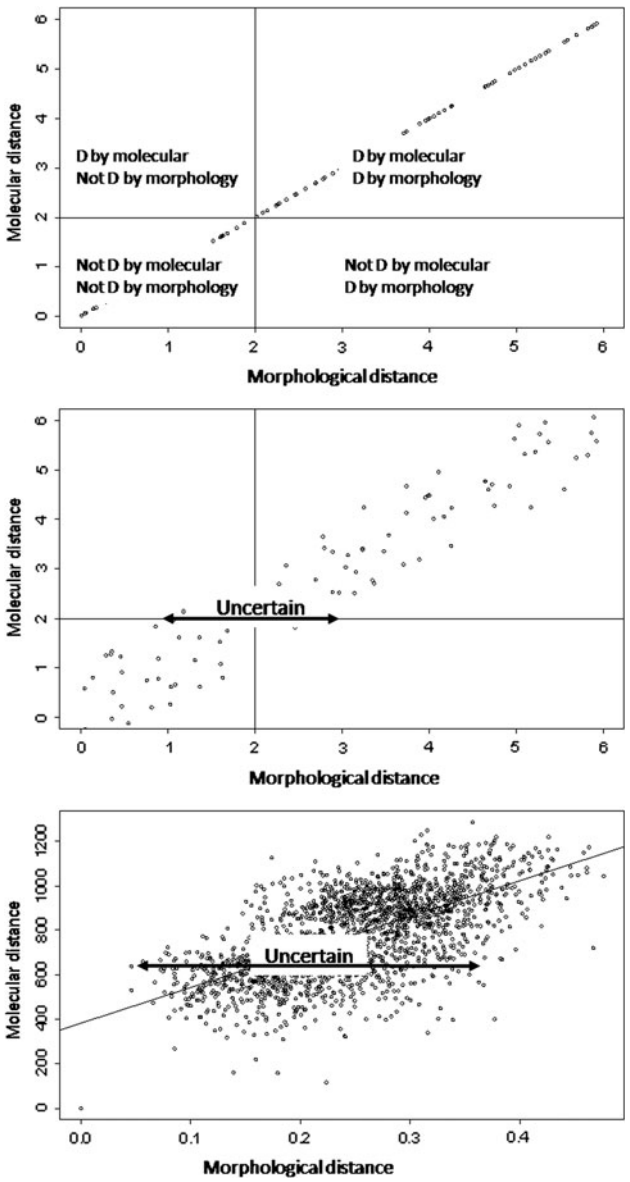

Fig. 2 Calibration of molecular against morphological distances under UPOV BMT Model 2. The *upper graph* illustrates decision making under a perfect correlation between molecular and morphological distances. The *middle graph* illustrates possible uncertainty where the correlation between molecular and morphological distances is sub optimal. The *lower graph* illustrates uncertainty seen within the data used in this study

distinctness decisions made using morphological characteristics or genotype data. We examine the hypothesis that 'Varieties shown as 'similar' using phenotypic distances will also be shown as 'similar' using genotypic distances'. The 'typical' data shown in Fig. 2 illustrate the issues that need to be resolved. Despite the positive correlation between phenotypic and genotypic distances, there may be ambiguity when comparing decisions made using morphological and genotypic data.

As all varieties within this dataset have been granted Plant Breeders' Rights we arbitrarily declared 10 % of varieties (43 varieties) as non-distinct using morphological characteristics and used this set of 'non-D' varieties 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. 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. 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, Manhattan Distance or Modified Manhattan Distance and compared with sets of 'non-D' varieties identified using genetic distances (Table 5). 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 clearly 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 correspond most closely with those made using measured characteristics but these results remain unsatisfactory. The possibility of adopting a 'super-D' approach 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 'low-D' varieties using the genotype data before we could be confident that that we would not include varieties that are 'D' by morphology among the genotypic 'super-D' varieties. Once more, the results are disappointing. Among the genotypic dataset tested, it is possible to select 400 (out of 431) varieties as 'low-D' and still include one variety that is 'non-D' by morphology among the 'super-D' indentified by genotypic distances.

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. 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

Table 5 Comparisons of Distinctness decisions made using either morphological or genotypic distances

		Number of genotypic 'non-D' varieties						
		43	100	200	250	300	350	400
Gower: 43 'non-D' varieties								
A	Full data set	26	56	91	100	100	100	100
B	No missing data	28	67	95	98	100	100	100
E	5 % missing data	30	58	93	100	100	100	100
I	Optimised mapped markers	33	60	95	98	100	100	100
K	Genomic prediction Training set: all varieties	47	81	95	100	100	100	100
Manhattan: 43 'non-D' varieties								
A	Full data set	23	44	79	91	95	98	98
B	No missing data	21	47	81	93	95	95	98
E	5 % missing data	23	44	81	95	95	95	98
I	Optimised mapped markers	26	47	86	91	95	95	100
K	Genomic prediction Training set: all varieties	35	72	86	93	93	100	100
Modified Manhattan: 43 'non-D' varieties								
A	Full data set	23	47	77	93	93	98	100
B	No missing data	19	44	79	91	93	95	98
E	5 % missing data	23	44	79	93	93	95	100
I	Optimised mapped markers	23	47	86	93	95	95	98
K	Genomic prediction Training set: all varieties	49	79	86	95	100	100	100

The percentage concordance between methods is shown

as percentage missing data, the optimum correlations with phenotypic distance measures were $r = 0.58$–0.66. Better correlations were achieved by selecting the 'best marker' at each mapped position across the genome ($r = 0.65$–0.72). 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 much above 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 LD within each crop genome.

Genomic prediction was attempted 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 heritability 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, a strong correlation was 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 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 unreasonably diminished if a novel genetic threshold is set at too high a level.

If these issues cannot be resolved, it is likely that the rapidly reducing costs of high-throughput DNA sequencing will make UPOV Model 3 more attractive. Under this model there would be complete replacement of the current system by the use of molecular markers. Variety registration could be completed in a matter of weeks or months with field inspections becoming a matter of historical interest. There can be no fuller description of a variety than its entire DNA sequence. However, the ability to describe a variety based on its DNA sequence may pose as many problems as it addresses: how should uniformity (U) be treated given that polymorphisms exist between monozygotic (identical) twins? How should stability (S) be addressed when the probability of mutation at any base is of the order on 10–8 per base per generation? Even if a satisfactory outcome can be agreed, where would the boundaries for minimum distance and essential derivation be set for distinctness (D) testing?

Such a change would have impacts on other areas of statutory testing such as seed certification. Currently seed lots are certified by reference to their variety description; thus the use of variety descriptions based wholly, or in part, on molecular data in DUS testing would impose molecular testing on seed certification authorities. There would be clear advantages to this change: using the reference sequence as the varietal description, seed lots could be certified as pure and true to type by assaying samples of seeds without the need for repeated field inspections and the purity of hybrid seed lots with respect to the hybrid formula could be put beyond doubt. This revised system would require a review of the sampling techniques used in seed certification. A revised system may place small-scale seed producers at a disadvantage and it could discourage seed production in nations without an infrastructure of sophisticated laboratory facilities.

A radical revision of PVP to utilise the data production potential of 'next generation sequencing' is almost inevitable. There should be urgency in the discussions to redefine 'varieties' with reference to the available data types and a managed transition to a new system that can be implemented in all nations, regardless of their economic status.

Acknowledgments The authors gratefully acknowledge the Community Plant Variety Office (EPM.7501705) and NIAB Trust for funding this work. They are thankful to the Association Genetics of UK Elite Barley Consortium (work supported by Defra, the Scottish Government, and the Biotechnology and Biological Sciences Research Council through Sustainable Arable LINK Program Grant 302/BB/D522003/1) for making their data available. They would also like to thank Dr Robert Cooke and Dr John Law for help and advice over many years.

References

Button P (2008) Situation in UPOV concerning the use of molecular techniques in plant variety protection. Presented at symposium on the application of molecular techniques for plant breeding and in plant variety protection, Seoul, Korea

Close TJ et al (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genom. doi: 10.1186/1471-2164-10-582

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D et al (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci USA 107:21611–21616

Cockram J, Jones H, Norris C, O'Sullivan DM (2012) Assessment of diagnostic molecular markers for DUS phenotypic assessment in the cereal crop, barley (Hordeum vulgare). Theor Appl Genet 125:1735–1749. doi:10.1007/s00122-012-1950-3

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

CPVO-TP/019/3 (2012) Protocol for Distinctness, Uniformity and Stability tests *Hordeum vulgare* L. sensu lato: Barley. Published by Community Plant Variety Office, 3, boulevard Maréchal Foch, FR-49000 ANGERS. [Available from http://www.cpvo.europa.eu/main/en/home/technical-examinations/technical-protocols/tp-agricultural-species]

Gilmour AR, Thompson R, Cullis BR (1995) Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51:1440–1450

Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. Biometrical J 52(1):70–84

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–874

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

Heslot N, Yang H-P, Sorrells ME (2012) Jannink J-L (2012) genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160. doi:10.2135/cropsci2011.06.0297

Ibáñez J, Vélez MD, de Andrés MT, Borrego J (2009) Molecular markers for establishing distinctness in vegetatively propagated crops: a case study in grapevine. Theor Appl Genet 119: 1213–1222

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

Struyf A, Hubert M, Rousseeuw PJ (1997) Integrating robust clustering techniques in S-PLUS. Comput Stat Data Anal 26:17–37

van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. J Stat Softw 45:1–67

Waugh R, Jannink J-L, Muehlbauer GJ, Ramsay L (2009) The emergence of whole genome association scans in barley. Curr Opin Plant Biol 12:218–222

UPOV document INF/18/1 (2011) Possible use of Molecular Markers in the Examination of Distinctness, Uniformity and Stability (DUS)