Theor Appl Genet (2015) 128:2461–2470
DOI 10.1007/s00122-015-2601-2

ORIGINAL ARTICLE

Implications of using genomic prediction within a high-density SNP dataset to predict DUS traits in barley

Huw Jones[1] · **Ian Mackay**[1]

Received: 29 May 2015 / Accepted: 28 August 2015 / Published online: 8 September 2015

Abstract

Key message **Alternative methods for genomic prediction of traits and trait differences are compared and recommendations made. We make recommendations for implementing methods in the context of DUS testing.**

Abstract High-throughput genotyping provides an opportunity to explore the application of genotypes in predicting plant phenotypes. We use a genome-wide prediction model to estimate the contribution of all loci and sum over multiple minor effects to predict traits. A potential use is in plant variety protection to discriminate among varieties on distinctness. We investigate this use with alternate scenarios in a set of 431 winter and spring barley varieties, with trait data from UK DUS trials comprising 28 characteristics, together with SNP genotype data. Firstly, each trait is predicted from genotypes by ridge regression with discrimination among varieties using predicted traits. Secondly, squared trait differences between each pair of varieties are regressed on genetic distances between each variety by ridge regression, with discrimination among varieties using the predicted squared trait differences directly. This latter approach is analogous to the use of phenotype and marker differences introduced to human genetic linkage analysis by Haseman and Elston and to the analysis of heritability in natural populations of plants by Ritland. We compare correlations between methods, both trait by trait and summarised across all traits. Our results show wide variation among correlations for each trait. However, the aggregate distances calculated from values predicted by genotypes show higher correlations with distances calculated from measured values than any previously reported. We discuss the applicability of these results to implementation of UPOV Model 2 in DUS testing and suggest 'success criteria' that should be considered by testing authorities seeking to implement UPOV Model 2.

Introduction

The International Union for the Protection of New Varieties of Plants (UPOV) is an intergovernmental organisation whose system of plant variety protection offers *sui generis* protection of plant breeders' intellectual property rights (Jones et al. 2013). Variety registration and protection of crop varieties require distinctness, uniformity and stability (DUS) testing of new varieties. This is currently carried out by assessment of phenotypic characteristics where candidate varieties may be registered if they are found to be morphologically distinct when compared to existing varieties. UPOV has established the Biochemical and Molecular Techniques (BMT) Working Group tasked with exploring opportunities to apply molecular marker technology within variety registration. The BMT guidelines suggest three application models for molecular markers in variety registration (UPOV document INF/18/1, 2010). Here we consider Model 2: calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics.

An earlier study of winter and spring barley varieties considered genotype data comprising 3072 SNP markers

Communicated by E. A. Carbonell.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2601-2) contains supplementary material, which is available to authorised users.

✉ Huw Jones
huw.jones@niab.com

[1] NIAB, Huntingdon Road, Cambridge CB3 0LE, UK

and phenotype data from UK DUS trials comprising 33 characteristics, of which 28 were European Union Community Plant Variety Office (CPVO) characteristics (CPVO-TP/019/3 2012). High correlations were obtained between morphological and molecular distance; however, when we modelled distinctness decisions using these data, we found that decisions made using molecular distances did not fully reproduce decisions made using morphological distances (Jones et al. 2013b).

Genomic prediction by ridge regression offers the possibility of improving correlations between morphological and molecular distance and hence improving the quality of DUS decisions made under UPOV Model 2. In our previous study (Jones et al. 2013b), we predicted phenotypes from genotypes by ridge regressions and demonstrated high correlations between phenotypic distances and distances calculated from predicted phenotypes. In this study, we extend our previous by using ridge regression to implement genomic prediction of phenotypic *distances* directly from values from marker-based genetic *distances* and compare results with prediction of phenotypic trait *values* from marker data *values* followed by calculation of phenotypic distances. We reasoned that prediction of trait differences from multiple marker distances may show advantages over working directly on the traits and markers themselves. Moreover, working on trait differences and genetic distances fits well conceptually within the framework of defining variety distinctness. This approach was based on analogous approaches used in human genetic linkage analysis (Haseman and Elston 1972) and in the study of heritable variation in natural populations (Ritland 2000).

Haseman and Elston showed that for a quantitative trait, the squared difference in trait values among sib-pairs will be inversely proportional to their probability of identity by descent at a linked locus (Haseman and Elston 1972). A highly significant negative regression coefficient between the squared trait difference and estimates of identity by descent at a marker locus is evidence of linkage of the marker to a QTL. Methods for QTL mapping using squared phenotypic differences (distances) rather than working directly on phenotypic values have been reviewed by Feingold (2002). Ritland (2000) used a similar approach, regressing squared trait differences on marker-based estimates of kinship among pairs of individuals in natural populations to estimate heritability. For application in DUS, we have extended these approaches by using ridge regression of squared trait distances on multiple marker-based estimates of genetic distances, reasoning that this may better exploit the genetic relationship between varieties in the reference panel and new candidate varieties.

We also explore the differences among alternative methods for distance calculations and different prediction methods. In particular, we compare prediction of squared trait differences and their use in defining distinctness with a more conventional two-stage approach in which traits are first predicted and then predicted traits are used to estimate distance between variety pairs. We discuss the implications of each combination of methods for DUS decisions and implementation of UPOV Model 2.

Materials and methods

The project used data obtained from 3072 SNP loci collected in the course of the AGOUEB project (544 varieties) (http://www.agoueb.org/) for a collection of barley varieties selected from UK registration trials over the past 20 years (Cockram et al. 2010). Use of these data was previously reported by Jones et al. (2013b). In brief, SNP markers were discovered and reported by Close et al. (2009) (Supplementary Table S3). Data were discarded for 517 monomorphic loci and for 1510 loci with any missing data; the dataset created maps to Dataset B from Jones et al. (2013b). Data for 28 morphological characteristics originating from UK registration trials between 1997 and 2004 were collated (Supplementary Table S2). We considered only those characteristics included in CPVO-TP/019/2 (2010) (Table 1). Varieties with more than ten DUS test characteristics missing were excluded. These final datasets comprised 431 varieties with both phenotypic and genotypic data.

Data analysis was carried out using Microsoft Excel (Microsoft Corporation) and the R Statistical Package including the package 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 Euclidean Distances and Gower's Coefficient. Gower's coefficient was selected for its suitability when handling datasets that include binary, multistate and continuous data (Gower, 1971).

We compared phenotypic distances with a prediction for the phenotypic distance made using genotype data. We calculated predicted phenotypic distances by two alternative approaches. In the first approach, we used linear ridge regression as implemented in the 'R' 'penalized' package (Goeman 2010) to predict phenotypes scores, characteristic by characteristic, using the genotype data. Predicted phenotypic distance matrices were subsequently calculated from these predicted phenotypes. We describe this approach as ridge regression of phenotypic values (RRPV). In the second approach, we calculated both locus by locus genotypic distances and characteristic by characteristic phenotypic distances using the dist function within R. We calculated both Manhattan and Euclidean distances and created arrays of one-dimensional distances. In this special, one-dimensional case Manhattan, Euclidean or indeed Hamming distances

Table 1 Correlation of phenotypes and distances (predicted vs. observed) within the test set

UPOV no.	Characteristic	Ridge regression of phenotypic values (RRPV)		Ridge regression of phenotypic distance v genotypic distance	
		Predicted phenotype	Predicted phenotypic distance	RRPD	RRPDSq
1	Plant: growth habit	0.64	0.35	0.19	0.17
2	Lowest leaves: hairiness of leaf sheaths	0.91	0.82	0.82	0.80
3	Flag leaf: intensity of anthocyanin coloration of auricles	0.48	0.39	0.43	0.45
5	Plant: frequency of plants with recurved flag leaves	0.14	0.01	0.08	0.05
6	Flag leaf: glaucosity of sheath	0.21	0.02	0.00	0.02
7	Time of ear emergence	0.14	0.06	0.00	−0.03
9	Awns: intensity of anthocyanin coloration of tips	0.34	0.17	0.41	0.41
10	Ear: glaucosity	0.45	0.24	0.10	0.15
11	Ear: attitude	0.23	0.00	−0.03	−0.03
12	Plant: length	0.26	0.01	0.01	0.01
13	Ear: number of rows	0.93	0.95	0.96	0.97
14	Ear: shape	0.22	0.02	0.02	0.01
15	Ear: density	0.17	0.01	0.00	−0.01
16	Ear: length	0.26	0.05	0.04	0.07
17	Awn: length	0.39	0.15	0.12	0.04
18	Rachis: length of first segment	0.36	0.11	0.01	0.07
19	Rachis: curvature of first segment	0.18	0.01	0.05	0.05
20	Sterile spikelet: attitude	0.63	0.43	0.37	0.34
21	Median spikelet: length of glume and its awn relative to grain	0.21	0.06	−0.09	−0.01
22	Grain: rachilla hair type	0.52	0.38	0.23	0.26
23	Grain: husk	0.63	0.35	−0.02	−0.02
24	Grain: anthocyanin coloration of nerves of lemma	0.61	0.34	0.38	0.36
25	Grain: spiculation of inner lateral nerves	0.69	0.55	0.53	0.52
26	Grain: hairiness of ventral furrow	0.67	0.56	0.42	0.42
27	Grain: disposition of lodicules	0.34	0.42	0.73	0.96
28	Kernel: colour of aleurone layer	0.69	0.57	0.66	0.55
29	Seasonal type	0.96	0.93	0.92	0.92
–	Ear: development of sterile spikelets	0.58	0.34	0.53	0.53

Data for Manhattan distances are presented, though these correlations are typical of all distance measures

should be identical. We then used linear ridge regression to predict phenotypic squared distances for each characteristic by regressing the square of the phenotypic distance against the genotypic distances (RRPDSq). For completeness, we then also used linear ridge regression to predict phenotypic distances for each characteristic by regressing phenotypic distance (i.e. not the squared distances) against the genotypic distances. We describe this approach as ridge regression of phenotypic distances (RRPD).

Practical application of ridge regression analysis estimates a prediction equation from a 'training set' (or reference set) with known phenotypes and genotypes. This equation is then used to predict phenotypes (or phenotypic distances) in a test set (or set of candidates), where only genotypes are known. In RRPV, the training set comprises a phenotype score and the genotype data for all loci for each variety. In RRPDSq, the training set comprises squared trait distances calculated for one phenotype and an array of locus by locus genotypic distances for every pair of varieties. In RRPD, the training set comprises distances calculated for one phenotype and an array of locus by locus genotypic distances for every pair of varieties. The prediction equation is an array of regression coefficients for each locus and the predicted phenotype or distance is the sum of an effect contributed by each genetic locus.

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

where phenotype$_i$ is the predicted trait value (or distance) for the ith variety, m_{ij} is the marker score for the jth marker for the ith line and g_{j} 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. In matrix form, to assess marker effects, ordinary least squares regression solves the equation

$$\mathbf{P} = \mathbf{Mg}$$

as

$$\mathbf{g} = \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{P}$$

where $\mathbf{g} = [g_0\ g_1\ g_2 \ldots g_{\mathrm{n}}]$ is a vector of fixed marker effects with g_0 the mean and $g_1 \ldots g_{\mathrm{n}}$ the effects for each marker, $\mathbf{P}$ is a vector of phenotypes and $\mathbf{M}$ is the design matrix for markers and assigns alleles at each locus to the individual phenotypes in $\mathbf{P}$. Ridge regression modifies the ordinary least squares estimates as

$$\mathbf{g} = \left(\mathbf{M}'\mathbf{M} + \mathbf{I}\lambda\right)^{-1}\mathbf{M}'\mathbf{P},$$

where $\mathbf{I}$ is a unit matrix with the same dimensions as $\mathbf{M}$, λ is a positive number which acts to shrink the estimates of elements of $\mathbf{g}$ back towards zero. The value of the tuning parameter λ was determined for each characteristic on the tenfold cross-validated partial likelihood that minimised the residuals between predicted and observed phenotypes within the training set.

To model DUS testing procedures where candidate varieties are required to be distinct from existing varieties in the reference collection, we divided the variety set by the chronological order of application for protection; the 200 varieties with the earliest applications were selected as the training set and the 231 later varieties placed in the test set. This allowed us to treat the training set as equivalent to a DUS reference collection of varieties and the test set as equivalent to a set of candidate varieties without compromising the independence of the training set from the test set. We compared results among Euclidean or Manhattan distance matrices calculated from predicted phenotype data with observed phenotypic distance matrices. Similarly, we compared results obtained when Euclidean or Manhattan marker-based distances were regressed against measured phenotypic distances in the training set and the regression used to predict phenotypic distances in the test set.

The training and test sets represent barley varieties submitted for DUS testing and the whole set is ordered by the date of submission. We modelled predictions over time by comparing measured and predicted phenotypic distances and phenotypic distances for new varieties as they enter the test system. We compared distances between test set and the training set and among test set. For simplicity, the comparisons were made for successive groups of 20 varieties as they entered the test set. We compared our results against randomised selections for the training set ($n = 200$) and test set ($n = 231$).

As all varieties within this dataset have been granted Plant Breeders Rights, they are defined as distinct from each other, making it impossible to assess DUS decisions at the usual thresholds. To compare the decision making using phenotype or genotype data, we set arbitrary thresholds for phenotypic distances, predicted and measured, such that 10 % of the varieties in the test set were ‘not distinct’ (non-D). These sets of ‘non-D’ varieties were used in comparisons between measured phenotypic distances and phenotypic distances derived from genomic prediction. The decision making, using measured or predicted phenotypic distances, could be compared by simply counting the number of varieties that were described as ‘non-D’ by either method. We used the same approach to model the ‘super-D’ method (Button 2008) of managing reference collections, where new varieties shown by genomic prediction to differ to a high degree from the reference collection and from each other would be eliminated from growing trials. Ideally, no novel variety shown to be ‘super-D’ genomic prediction would be shown as non-D by phenotypic distances. We estimated this by counting the coincidence of varieties among the least distant by observed phenotypic distance and the most distant by predicted phenotypic distance.

Results

For each characteristic, we optimised λ by tenfold cross validation within the training set, and used this value within the complete training set to estimate a prediction equation for that characteristic. Having optimised λ, we carried out ridge regression for each characteristic using the training set to estimate a prediction equation for that characteristic. This was then used to predict values for varieties in the test set for that characteristic from their genotypic data. Similarly, ridge regression was used to predict phenotypic distances from marker distances.

Distances for each predicted morphological characteristic

Correlation coefficients were calculated for each characteristic in turn between measured and predicted phenotypes and between measured and predicted phenotypic distances (Table 1). We investigated the effect of differing methods used to calculate observed phenotypic distances (Euclidean, Manhattan and Gower distances) for variation in calculated correlations. The correlations between measured and predicted phenotypes, calculated characteristic by

Table 2 Correlations among distances (predicted vs. observed) using differing calculation methods

Distance measure used in association with ridge regression	Euclidean			Manhattan		
Observed phenotypic distance	Euclidean	Manhattan	Gower	Euclidean	Manhattan	Gower
RRPV						
Distance among all varieties	0.73	0.71	0.71	0.73	0.71	0.72
Distance among training set varieties	0.75	0.72	0.72	0.75	0.73	0.74
Distance of test set from training set	0.73	0.71	0.71	0.72	0.71	0.72
Distance within test set	0.71	0.70	0.71	0.71	0.71	0.72
RRPD						
Distance among all varieties	0.75	0.74	0.72	0.78	0.77	0.75
Distance among training set varieties	0.87	0.87	0.83	0.87	0.87	0.82
Distance of test set from training set	0.78	0.77	0.75	0.78	0.77	0.74
Distance within test set	0.72	0.71	0.71	0.72	0.71	0.70
RRPDSq						
Distance among all varieties	0.76	0.72	0.69	0.76	0.72	0.69
Distance among training set varieties	0.83	0.78	0.72	0.83	0.78	0.72
Distance of test set from training set	0.78	0.75	0.72	0.78	0.75	0.72
Distance within test set	0.68	0.65	0.62	0.68	0.63	0.62

characteristic, were in the range 0.14–0.96 in the test set. When the measured and predicted phenotypes were used to calculate distances (RRPV), the correlations between distances calculated characteristic by characteristic were in the range of 0.00–0.95 in the test set; the correlations obtained were similar whether Manhattan, Euclidean or Gower's distance was used to calculate observed phenotypic distance. The distance-by-distance correlations for 13 out of 28 characteristics were less than 0.2 by this method.

When RRPD was used to predict phenotypic distances, the correlations between observed and predicted phenotype, calculated characteristic by characteristic, were in the range from −0.09 to 0.96 in the test set. When RRPDSq was used to predict phenotypic distances, the correlations between observed and predicted phenotype, calculated characteristic by characteristic, were in the range from −0.03 to 0.97 in the test set. Similar correlations were obtained whether Manhattan, Euclidean or Gower's distances was used to calculate observed phenotypic distance. The correlations for 15 out of 28 characteristics were less than 0.2 by these methods.

Distances using all predicted morphological characteristics

Distinctness decisions are made using the sum of character by character distances. We compared observed distance matrices with predicted distance matrices and calculated correlations for distances among varieties in the training set, varieties in the test set and distances between varieties in the training and test sets (Table 2). As expected, the highest correlations are seen for distances among the varieties in the training set and the lowest for those within the test set with intermediate values for distances between the training and test set varieties. The method used to calculate the genotypic distances used for prediction made little difference in the correlations. Looking at the observed phenotypic distances, Manhattan and Euclidean distances gave similar correlations, while correlations for Gower distances were slightly lower. Considering the different methods used to implement ridge regression, the highest correlations were obtained by using RRPD, the lowest by using RRPV. Closer examination of distances between test set and training set using William's test (Steiger 1980) shows that the differences between the correlations are significant ($p < 0.05$). The correlations within all varieties, including both the test set and training set, calculated by ridge regression of distances, exceed the highest correlation (0.69) obtained by other methods during our previous study (Jones et al. 2013b). The results for the 'ordered' selection for training and test set were typical of those obtained from randomised selections.

All methods where predictions are made on the basis of the diversity of the training set are vulnerable to error should novel diversity be introduced into the test set. The training and test sets represent barley varieties submitted for DUS testing and the whole set is ordered by the date of submission. Looking at the distances between individual members of the test set and the training set, we calculated the correlations between the observed distances and predicted distances. In general, the correlations were high and positive but there were a small number of varieties where

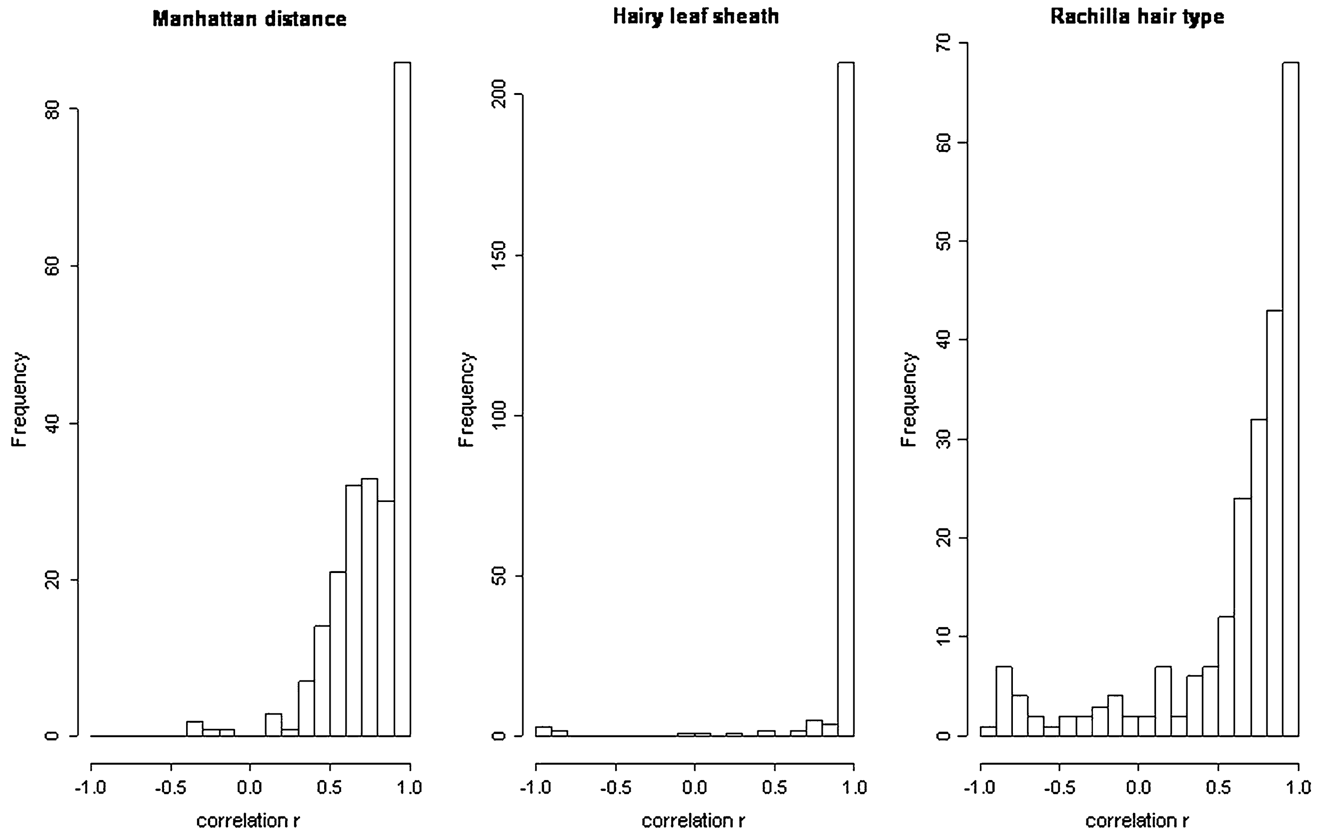

Fig. 1 Distributions of correlations for the distances between the training set and individual members of the test set. The correlation for all characteristics, calculated as Manhattan distances by RRPV, is the sum of distances for characteristics with a strong, positive correlation (e.g. hairy leaf sheath, $r = 0.80$) and characteristics with a low correlation (e.g. rachilla hair type, $r = 0.23$)

the correlations were low or negative; Fig. 1 shows correlations between observed Manhattan phenotypic distances and predicted Manhattan phenotypic distances produced by RRPD using Manhattan genetic distances. Included among the varieties with low correlations were pairs where one variety was reported as a parent of the other. The distances between the training set and individual members of the test set were dissected characteristic by characteristic. For some characteristics, e.g. hairy leaf sheath, a majority of varieties were predicted correctly but a small number of outliers were observed. For other characteristics, e.g. rachilla hair type, a broader range of correlations was observed.

Stability of predictions over time

We explored the stability of the predictions over time by examining the correlations between the training set and sequential groups of 20 individuals within the test set, testing the hypothesis that correlations would decline over time. For example, comparing observed Manhattan distances with distances predicted by ridge regression of Manhattan distances, the correlations for each tranche varied considerably (0.69–0.87 against a correlation of 0.77 for the whole set). However, when the values were plotted by order of submission, there was no suggestion of trend (Fig. 2). The results for the 'ordered' selection for training and test set were typical of those obtained from randomised selections.

The correlations for individual characteristics varied widely (e.g. 0.14–0.96 when distances are predicted by ridge regression of phenotypic values (Table 1)). We found that these correlations followed the relationship seen by Cockram et al. (2010) for heritability calculated for each characteristic ($r = 0.79$). Taken together, these results raise a question of whether those morphological traits with low correlations can be considered to "result from a genotype or combination of genotypes" (UPOV 1991). We investigated whether a small number of characteristics with high correlations were responsible for the overall correlations by dropping highly correlated characteristics, one after another, out of the calculations. In each case, the measured phenotypic distance for all characteristics was compared with the predicted phenotypic distance with one characteristic omitted from the calculation of total distance, then two characteristics omitted, and so on. Looking at Euclidean or Manhattan distances calculated by

RRPV, the results showed that it was possible to remove up to three highly correlated characteristics without reducing the overall correlation between measured and predicted phenotypes (Table 3). When the 12 most highly correlated characteristics were removed from the calculation, the correlation between predicted and observed phenotypic distance was 0.53. This suggests that all predicted characteristics are making a positive contribution to the overall correlation.

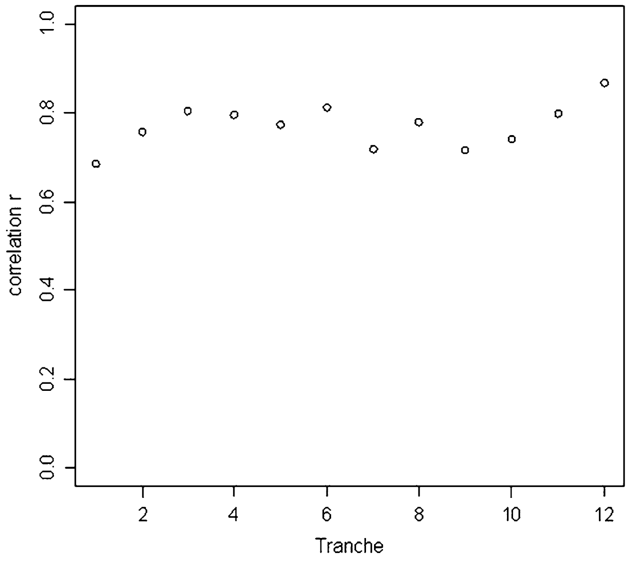

Fig. 2 The correlation for observed and predicted Manhattan distances calculated by RRPD between candidate ('test set') and reference collection ('training set') varieties shows no suggestion of trend when the values were plotted in subsets (tranches) of varieties ($n = 20$) by order of submission

Use of predicted phenotypes to make distinctness decisions

As all members of this dataset are named varieties that are distinct from each other, we adjusted thresholds for distinctness to arbitrarily declare 10 and 20 % of varieties (23 and 46 varieties within the test set) as non-distinct from the training set using morphological characteristics. This set of 'non-D' varieties was then used as a bench mark for comparisons with thresholds set at 10 and 20 % for the genotypic data. The decision making using phenotypic or genotypic data could then be compared by counting the number of varieties that were described as 'non-D' by both methods (Table 4). The results obtained show that when phenotypic distances are predicted by RRPV, identical distinctness decisions are made for 6 out of 23 and 18 out of 46 varieties, respectively, at the 10 and 20 % thresholds for non-D. The results for RRPD show that identical distinctness decisions are made for 7 out of 23 and 23 out of 46 of varieties. A similar comparison for RRPDSq shows that identical distinctness decisions are made for 5 out of 23 and 17 out of 46 of varieties. Taken together, these results suggest that

Table 3 The data show the effect of removing highly correlated characteristics, one after another, from the distance calculations

UPOV no.	Characteristics dropped from distance calculations	Distance among all varieties	Distance among training set varieties	Distance of test set from training set	Distance within test set
No characteristics dropped from distance calculations		0.73	0.75	0.73	0.71
13	Ear: number of rows	0.73	0.75	0.73	0.71
29	Seasonal type	0.72	0.75	0.73	0.72
2	Lowest leaves: hairiness of leaf sheaths	0.74	0.77	0.75	0.73
26	Grain: hairiness of ventral furrow	0.68	0.71	0.69	0.68
25	Grain: spiculation of inner lateral nerves	0.60	0.61	0.59	0.61
20	Sterile spikelet: attitude	0.60	0.61	0.59	0.61
27	Grain: disposition of lodicules	0.60	0.61	0.59	0.61
22	Grain: rachilla hair type	0.60	0.61	0.59	0.61
3	Flag leaf: intensity of anthocyanin coloration of auricles	0.59	0.59	0.58	0.60
1	Plant: growth habit	0.55	0.55	0.53	0.57
24	Grain: anthocyanin coloration of nerves of lemma	0.53	0.53	0.51	0.55
–	Ear: development of sterile spikelets	0.56	0.59	0.55	0.57

Three highly correlated characteristics could be omitted without reducing the overall correlation between measured and predicted phenotypes. The correlations presented are Euclidean distances calculated by RRPV

Table 4 The error rate for distinctness decisions made using observed versus predicted phenotypic distance is shown as the number of varieties where the decisions differ

	No of varieties	Predicted phenotypic distance						
		10% non-D	20% non-D	10% super-D	20% super-D	30% super-D	40% super-D	50% super-D
		23	46	23	46	69	92	115
Observed phenotypic distance								
RRPV								
10% non-D	23	19		0	0	0	1	4
20% non-D	46		28	0	0	1	4	6
RRPD								
10% non-D	23	16		1	1	1	1	1
20% non-D	46		23	1	2	2	4	5
RRPDSq								
10% non-D	23	19		1	1	1	1	1
20% non-D	46		31	1	1	2	2	4

The number of varieties at 10 or 20 % non-D threshold is 23 of 46, respectively. The error rate for 'super-distinctness' is the number of varieties classified as non-D using observed distances and super-D using predicted distances. The number of super-D varieties in each class is shown in the table

even at the high distance to distance correlations achieved in this study, the same decision will not be made using observed morphological distances and morphological distances predicted by genetic distances.

We investigated the possibility of adopting a 'super-D' approach using these data. A 'Super-D' approach is one where distances characteristics are used at a high threshold to identify candidates that differ in a high degree from reference varieties and eliminate them from further field testing. Ideally, there would be no varieties in common between a set shown to be non-D by observed phenotypic distances and a set selected as super-D using predicted distances. The threshold for super-D was varied for predicted distances and the number of varieties that were both super-D by predicted phenotypic distances and non-distinct by observed phenotypic distances was counted (Table 3). To achieve the ideal outcome of no super-D varieties by predicted phenotypic distances identified as non-distinct by measured phenotypic differences, the threshold for super-D was set at 28 % for distances predicted by RRPV, 5 % for ridge regression of RRPD and 3 % for distances predicted by RRPDSq. This would allow 65, 12 or 5 varieties, respectively, out of 231 to be safely declared as super-D by their predicted phenotypic distances and eliminated from growing trials. For the three methods used to predict phenotypic distances, it is possible to increase the super-D threshold to 35 % before the error rate starts to rise rapidly. Typically, 81 varieties (out of 231) could be eliminated from growing trials if a low error rate rather than a zero error rate was deemed acceptable. At a 35 % super-D threshold and a 20 % non-D threshold there would be two varieties incorrectly assigned using ridge regression of phenotypic values, four varieties incorrectly assigned by RRPD and two varieties incorrectly assigned by RRPDSq (predicted Manhattan distances vs. observed Manhattan distances). We present the error rate for super-D thresholds of 10, 20, 30, 40 and 50 % for predicted Manhattan distances versus observed Manhattan distances in Table 4. The error rates for other combinations of distance measures are similar with between one and four varieties shown as erroneous, depending on the distance measures used (at a 35 % super-D threshold and a 20 % non-D threshold). The varieties shown as erroneous differ, with seven varieties appearing among the subsets of varieties belonging to both a 35 % super-D and a 20 % non-D set.

Discussion

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 represent the whole barley genome; the marker set is an order of magnitude larger than any dataset used in an exploration of UPOV BMT Model 2 previously reported. In our previous paper (Jones et al. 2013b), we demonstrated a positive correlation between genotypic and phenotypic distance measures for this set of varieties, demonstrated methods to optimise correlations and demonstrated the potential of genomic prediction by ridge regression. Application of genomic prediction in plant breeding is becoming more widespread. However, we believe this is the first occasion, where ridge regression among genotype distances has been applied to predict crop phenotype differences. We have demonstrated improvements in correlation by implementing ridge regression between phenotype values and genotypes and between characteristic by characteristic phenotypic distances with genome-wide genotypic distances.

We note that, when considered on a characteristic by characteristic basis, there was considerable variation in the correlations between predicted and measured characteristics and between predicted and measured distances. This suggests that there is considerable variation in the heritability of the characteristics and hence considerable variability in the quality of information when the characteristics are used in distinctness testing under the current system. Nonetheless, we demonstrate a small improvement in correlations between predicted and measured distances and show an improvement in the quality of decision making should molecular-derived distance be applied in distinctness testing.

The three approaches implemented in this study give broadly similar high correlations. The correlations of distances between test set and training set rank in order RRPD > RRPDSq > RRPV and William's test (Steiger 1980) show that the differences between the correlations are significant.

Conclusions

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 morphological distances and predicted phenotypic distances achieved in this study and our previous study. However, we have also shown in this study that a small increase in correlation between distances from measured phenotypic distances and molecular-derived distances can leverage a useful gain in the quality of decision making.

If molecular methods are to be implemented within the UPOV BMT Model 2, either as a replacement for field testing or as a grouping tool for management of reference collections, the results from this study and our previous study suggest criteria for success. Firstly, we suggest that the marker density needs to be high in order to achieve useful correlations between measured phenotypic distances and molecular-derived distances (Jones et al. 2013b). Our previous study in barley suggests that when SNPs are deployed, over 500 markers are needed to reliably achieve correlations that exceeds $r = 0.60$. We also show that as markers are added, the correlations between genotypic distances and phenotypic distances eventually arrive at a plateau value. In this study, we show that correlations can exceed this plateau by application of genomic prediction. The highest correlations were obtained by regression of phenotypic distances against multi-locus genotypic distances (RRPD).

Secondly, we suggest that any proposal to implement molecular methods within UPOV BMT Model 2 should demonstrate correlations in excess of $r = 0.60$. A comparison of correlation values with the error rate in decision making suggests that when correlations are below 0.60, the discrepancy rate in distinctness decisions made using phenotypic and molecular distance exceeds 80 %; this is clearly unacceptable (Jones et al. 2013b). As the correlations improve using ridge regression, we have demonstrated that the discrepancy rate in distinctness decisions falls to below 50 %. This is an appallingly high value and highlights the difficulty of implementing UPOV Model 2 when the correlations between phenotypic and genetic distances are anything less than perfect. The effect of increasing correlation has more value when applied to a super-D approach. If a super-D threshold is set at 35 %, and a non-D threshold is set at 20 %, we can count the varieties that would be both non-distinct using morphological distances and be classed as super-D using predicted distance matrices. At the correlations achieved using our ridge regression approach, we demonstrate that the error rate falls below 2 %.

Thirdly, we suggest that any proposal to implement molecular methods within UPOV BMT Model 2 should model decision making for both distinctness and super-D decisions. It is relatively simple to set arbitrary thresholds for distinctness and discuss whether the proposed method changes the quality of variety protection. This would be particularly useful where part of the process requires an expert judgement to be placed on relative weightings for genotypic and phenotypic distances when they are used in combination. However, without clear guidance from UPOV on acceptable levels of error when a proposal is made to implement molecular methods within UPOV BMT Model 2, it may be difficult to make progress.

The rapidly reducing costs of high-throughput DNA marker generation and sequencing and the increased efficiencies of data processing make UPOV Model 2 implementation achievable and attractive. A revision of PVP to utilise the data production potential of 'next generation sequencing' is almost inevitable. There should be urgency in the discussions to define 'success criteria' for novel methods that would allow testing authorities to reduce the costs of DUS testing without diminishing the quality of variety protection.

Author contribution statement Huw Jones carried of the data analysis, wrote code and drafted the manuscript.

Ian Mackay proposed the original idea (use of ridge regression of trait differences vs. marker differences), supervised the research, advised on methods and edited the manuscript.

Acknowledgments The authors gratefully acknowledge the NIAB Trust for funding this work.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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 (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Natl Acad Sci USA, Proc

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. http://www.cpvo.europa.eu/main/en/home/technical-examinations/technical-protocols/tp-agricultural-species

Feingold E (2002) Regression-based quantitative-trait-locus mapping in the 21st century. Am J Hum Genet 71:217–222

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

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

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Jones H, Norris C, Cockram J, Lee D (2013a) Variety protection and plant breeders' rights in the 'DNA era'. In: Lübberstedt T, Varshney RK (eds) Diagnostics in plant breeding. Springer, Netherlands

Jones H, Norris C, Smith D, Cockram J, Lee D, O'Sullivan DM, Mackay I (2013b) Evaluation of the use of high-density SNP genotyping to implement UPOV Model 2 for DUS testing in barley. Theor Appl Genet 126:901–911

Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. Mol Ecol 9:1195–1204

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251

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

UPOV (1991) International convention for the protection of new varieties of plants: Article 1(vi)

UPOV document INF/18/1: Guidelines For DNA-Profiling: Molecular Marker Selection and Database Construction ("BMT Guidelines")