The Annals of Applied Statistics
2013, Vol. 7, No. 4, 1898–1916
DOI: 10.1214/13-AOAS671

AN ALGORITHM FOR DECIDING THE NUMBER OF CLUSTERS AND VALIDATION USING SIMULATED DATA WITH APPLICATION TO EXPLORING CROP POPULATION STRUCTURE

BY MARK A. NEWELL[1,*], DIANNE COOK[2,†], HEIKE HOFMANN[2,†]
AND JEAN-LUC JANNINK[‡]

The Samuel Roberts Noble Foundation[*], *Iowa State University*[†]
and Cornell University[‡]

A first step in exploring population structure in crop plants and other organisms is to define the number of subpopulations that exist for a given data set. The genetic marker data sets being generated have become increasingly large over time and commonly are of the high-dimension, low sample size (HDLSS) situation. An algorithm for deciding the number of clusters is proposed, and is validated on simulated data sets varying in both the level of structure and the number of clusters covering the range of variation observed empirically. The algorithm was then tested on six empirical data sets across three small grain species. The algorithm uses bootstrapping, three methods of clustering, and defines the optimum number of clusters based on a common criterion, the Hubert's gamma statistic. Validation on simulated sets coupled with testing on empirical sets suggests that the algorithm can be used for a wide variety of genetic data sets.

1. Introduction. In the field of plant breeding, a breeder often wants to cluster available genetic lines, characterized by a set of markers, to organize the lines based on attributes of the population such as structure and linkage disequilibrium [Newell et al. (2011)]. They may also want to cluster growing environments based on yield data of various lines to define a target set of environments best suited to the line [Cooper and DeLacy (1994)]. Clustering algorithms, where individuals or cases are assigned to groups based on their similarity, is used. In many fields of science where large amounts of data are being generated, clustering similar cases or variables is often useful to organize the data. As in plant breeding, cluster analysis is often used to answer specific questions. Whether the research question is largely exploratory or inferential, cluster analysis can contribute useful insight into the structure hiding in a data set. Due to the underlying variation that is generally unknown without genetic information, a major obstacle to cluster analyses is estimating the number of clusters, which for genetic data might be considered to be

Received July 2011; revised June 2013.

[1] Supported by USDA-NIFA Grant 2008-55301-18746.

[2] Supported in part by NSF Grant DMS-07-06949.

Key words and phrases. Cluster analysis, high dimensional, low sample size, simulation, genetic marker data, visualization, bootstrap, dimension reduction.

subpopulations. In fact, although current clustering methods, such as k-means and hierarchical, are quite useful, they do little to address the practical question of how many clusters exist [Fraley and Raftery (2003)]. Milligan and Cooper (1985) tested various procedures for determining the number of clusters using classical hierarchical methods, however, the simulated data was small and nonoverlapping, and therefore not practical for genetic data. The methods for estimating the number of clusters for k-means clustering has been reviewed and include algorithmic, graphical and formulaic approaches [Steinley (2006)]. Having insight into the number of clusters present for a genetic marker data set can aid in understanding population structure.

Model-based clustering provides some help on choosing the number of clusters by calculating some criterion based on the population distribution assumptions. The most widely used model-based clustering approach used in genetic studies is implemented in the computer software STRUCTURE [Pritchard, Stephens and Donnelly (2000)]. It decides the number of clusters by comparing variance-penalized log-likelihoods. STRUCTURE has been cited in many research manuscripts. Vähä et al. (2007) applied four separate rounds of STRUCTURE to Atlantic salmon (*Salmo solar*) genetic marker data and found that, although it seemed to work well in clustering the genetic structure appropriately, the computational time was intolerably long. Hamblin et al. (2010) applied the STRUCTURE model-based clustering to a large barley (*Hordeum vulgare* L.) data set consisting of 1816 individuals and 1416 variables (markers), wherein convergence did not occur after very lengthy runs, finally requiring the use of another algorithm. In addition to computational issues, STRUCTURE makes genetic assumptions that are rarely met in breeding populations: (1) marker loci are unlinked and in linkage equilibrium with one another within populations, and (2) Hardy–Weinberg equilibrium within populations. The first of these assumptions can be simply avoided by selection of markers that are unlinked and in linkage equilibrium. In contrast, the second assumption is rarely the case for plant breeding populations in which selection plays a major role in population development. An important result of these assumptions is that allele frequencies across loci must be relatively similar, which is rarely the case for genetic data.

For plant breeding, as in many other fields of science, the increasing availability of data also results in high-dimensional data sets that can be difficult computationally to cluster. The data that this paper uses is binary data, presence or absence of a genetic marker, for each unique line. There are commonly lots of missing values. High-dimensionality issues related to cluster analyses were originally described by Bellman (1961) as an exponential growth of hypervolume as a function of dimension. Clarifying this for clustering, Murtagh (2009) determined that in very high-dimensional space there is a simplification of structure, demonstrating that the distances within clusters become tighter, while between cluster distance expands, with an increase in dimensionality. Though the research presented by Murtagh (2009) makes a convincing argument to utilize all dimensions in high-dimensional data sets, this is often not done due to the computational burden. In addition, genetic data often times includes a high frequency of nuisance variables that do not contribute to the structure of the data. In order to overcome these possible issues, it may be appropriate to implement cluster analyses on low-dimensional projections such as the principal components (PCs) for some methods [Fraley and Raftery (2002)]. Hall, Marron and Neeman (2005) found that low-dimensional projections of such data sets, where the number of dimensions $d \to \infty$ while the number of observations n is fixed, tend to lie at vertices of a regular simplex, in agreement with Murtagh (2009) and Ahn et al. (2007). (Note that the supplementary material for this paper [Newell et al. (2013)] contains video of higher-dimensional views of plant breeding data that also support the claim that this simplified structure is present in these genetic data sets.)

HDLSS data can pose a challenge when applying principal component analysis (PCA) because the covariance matrix is not of full rank. This leads to a strong inconsistency in the lower eigenvectors, in which case the added variation obscures the underlying structure of the covariance matrix [Jung and Marron (2009)]. The first few eigenvectors are consistent if there is a large difference in size of the eigenvalues between these and the rest. Classic studies of dimension reduction and cluster analysis [e.g., Chang (1983)] caution against using PCA before clustering. The reason is that PCA is finding directions of maximum variance in the data, which does not always correspond to differences between clusters. This is well known and explored further in several other papers, although it is still mistakenly done. Projection pursuit, particularly with the holes index [Cook, Buja and Cabrera (1993), Steinley, Brusco and Henson (2012)], is a better approach for reducing dimension before clustering. Ideally, clustering is done without reducing the dimension, but some clustering methods that do not work well for high-dimensional data, such as model-based clustering that depends on estimating variance–covariance matrices, require a dimension reduction step.

Advances in technology enable simulation of genetic data sets with known cluster classifications. This application allows better testing and evaluation of new algorithms on data sets with known properties. Comparisons can also be made between simulated and empirical data sets to gain insight into empirical data sets. The computer software GENOME [Liang, Zöllner and Abecasis (2007)], a coalescent-based whole genome simulator, offers just this by simulating sequences backward in time. Simulation of genetic sequences is conditional on chosen parameters including, but not limited to, population size, recombination rate and rates of migration between subpopulations. The software is particularly fast so it has the ability to generate a large number of data sets in a relatively short period of time. Most importantly, setting the available parameters enables the user to simulate data sets similar to empirical sets with respect to the number of clusters and the level of structure present. Milligan and Cooper (1985) evaluated different methods for determining the number of clusters, however, the simulated data was very limited in the number of observations, number of dimensions and, most importantly for genetic data, the lack of any cluster overlap.

The clustering methods that are currently available result in distinctive outcomes that are often compared by the researcher on some criterion and chosen accordingly. An approach that implements the array of clustering methods available and chooses the method that minimizes or maximizes a common criterion would be a useful approach that could capitalize on the positives associated with specific methods. This paper describes such an approach, that identifies the number of clusters for genetic marker data that incorporates model-based, k-means and hierarchical methods, and uses bootstrapping and cluster criterion to help decide the number of clusters. The algorithm is validated using GENOME simulated data and assessed on six empirical data sets. Outcomes of the research include evaluation of an algorithm to define the number of clusters using simulated data sets similar to our empirical sets, comparison of simulated data sets to empirical data sets, and development of graphical diagnostics to aid in the determination of the number of clusters. We expect that these contributions might be more generally applied to HDLSS data.

The paper is organized as follows. Section 2 describes the algorithm for choosing clusters. Section 3 describes the simulated and empirical data sets used to validate the algorithm. Section 4 describes the results. Supplementary material [Newell et al. (2013)] contains (1) the data sets, (2) R code for the analysis and (3) videos of the data sets, and resulting clusters, shown in more than two dimensions to better see the differences between clusters.

2. Algorithm for choosing the number of clusters.

The algorithm to determine the number of clusters has four steps: bootstrap sampling, clustering, calculation of a cluster validity statistic, and the computation of a permutation test for significance. Hubert's gamma statistic [Halkidi, Batistakis and Vazirgiannis (2001)], available in the R package `fpc` [Hennig (2011)], is the cluster validity statistic of choice, chosen heuristically from many criteria within the algorithm including the average distances within and between clusters and their ratio, the within clusters sum of squares, the Dunn index and entropy. Additionally, it is on a standard scale which makes comparison between methods simpler and calculation across clustering methods trivial. For consistency, matrices are denoted in bold typeface with the subscript representing the number of rows and columns, respectively. Let $\mathbf{X}_{n\times p}$ (n rows and p columns) be the data set to be clustered. In the genetic marker data, rows contain the lines and columns the marker information. For the empirical sets, missing data was imputed using the mean marker frequency for that marker, which is common practice for genetic data. In addition, the steps are graphically displayed using a $n = 150$ by $p = 100$ simulated data set using STRUCTURE composed of three clusters of equal size with a migration rate of 0.00001. The cluster structure is displayed using the first two principal components, which for this data is shown in Figure 1(left). The user sets a maximal number of clusters, $k_{\max}$, based on prior knowledge of a maximum. The steps are then as follows:

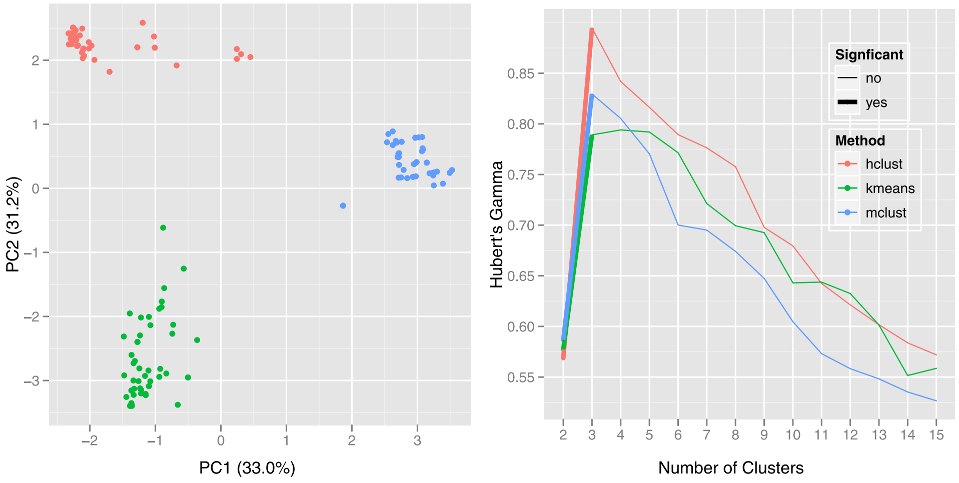

FIG. 1. *(Left) principal component one (PC1) versus PC2 with percent of the variation explained in parentheses for the example data set used to show the steps of the proposed method. (Right) Hubert's gamma values at each cluster number for the three methods of clustering on the example simulated data set generated to have three clusters. Thick lines represent significant* ($p < 0.01$) *increases in Hubert's gamma for pair-wise cluster numbers.*

1. *Bootstrapping*: a number of bootstrap samples, b, are drawn at random from the rows of $\mathbf{X}$ with replacement. The resulting matrix is denoted as $\mathbf{X}^{*i}_{n\times p}$ for $i = 1, \dots, b$.
2. *Cluster analysis*: three methods of cluster analysis are implemented for $1, 2, \dots, k_{\max}$ clusters including model-based, hierarchical and k-means clustering.
   (a) Model-based (mclust) cluster analysis, available in the R package `mclust` [Fraley and Raftery (2011)], is applied to principal components $\mathbf{Y}_{n\times 1}$, $\mathbf{Y}_{n\times 2}$, $\mathbf{Y}_{n\times 3}, \dots, \mathbf{Y}_{n\times k_{\max}}$ where the number of clusters is set to k. Thus, the number of principal components is equal to the number of clusters. The principal components were used only for the mclust method.
   (b) Hierarchical (hclust) clustering is applied to the Manhattan distance matrix $\mathbf{D}_{n\times n}$ and cut at k clusters. The Manhattan distance was preferred to Euclidean distance, as it represents the absolute distance between lines based on their binary marker data.
   (c) k-means (kmeans) clustering is applied to the bootstrap sample $\mathbf{X}^{*i}_{n\times p}$ with the number of clusters set to k.
3. *Cluster validity*: for each $1, 2, \dots, k_{\max}$ clusters, Hubert's gamma is calculated for model-based, hierarchical and k-means clustering on the Manhattan distance matrix. This results in three Hubert's gamma statistics at each of $1, 2, \dots, k_{\max}$ number of clusters.
4. *Permutation test*: a paired permutation t-test is computed for each consecutive number of clusters across bootstrap samples for each method of clustering,

meaning between clusters 2:3, 3:4, ..., and $k - 1$:$k_{\max}$. A linear model is applied to each pair with Hubert's gamma as the response and the cluster number as the explanatory variable.

5. *Choosing the number of clusters*: the clustering method resulting in the highest Hubert's gamma is used. The algorithm returns the lowest number of clusters for which Hubert's gamma is significantly greater than the number below it, but not for the number above it. Results for the example data set are shown in Figure 1(right), with bold lines representing significant increases in Hubert's gamma between consecutive cluster pairs. For the example data set, all clustering methods would return three clusters; hierarchical clustering yielded the highest Hubert's gamma, so it would be used.

3. Data.

3.1. *Simulated.* In order to validate the proposed method, data sets were simulated with varying numbers of clusters and degree of separation between clusters. The coalescent whole genome simulator GENOME was used for all simulations and was chosen because it was able to simulate data sets covering the spectrum of variation in our empirical sets. The simulated sets ranged in the number of clusters including 1, 2, 3, 4, 5, 6, 9 and 12 clusters. The level of separation between clusters was specified by adjusting the migration rate per generation per individual, levels for this parameter were 0.00005, 0.0001 and 0.00015. High levels of migration resulted in less separated clusters, while low levels of migration resulted in more separated clusters. The number of clusters and migration rate were arranged as a factorial such that 100 simulations were tested at each cluster—migration rate combination. All simulated sets included 200 observations and 400 markers, with each cluster having equal numbers of observations. Because the number of observations per simulation was fixed at 200, as the number of clusters increased, the number of observations per cluster decreased. The simulated data sets were HDLSS, which is generally the case for genetic data sets. In order to gauge the variation in the simulated sets, Figure 2 shows the first two principal components for one random sample of each cluster—migration rate combination. Visually, the simulated sets cover a wide spectrum of variability with respect to the number of clusters and, most notably, the level of separation.

3.2. *Empirical.* Six empirical data sets were used from three small grain crops, including three oat (*Avena sativa* L.), two barley and one wheat (*Triticum aestivum* L.) data set. The first oat data set, referred to as newell2010, is a collection that includes varieties, breeding lines and landraces of worldwide origin originally used for analysis of population structure and linkage disequilibrium [Newell et al. (2011)]. The newell2010 data set has 1205 observations and 402 Diversity array technology (DArT) markers, which are binary, with 5.1% missing data. The second oat data set, referred to as tinker2009, is also a set of varieties, breeding lines

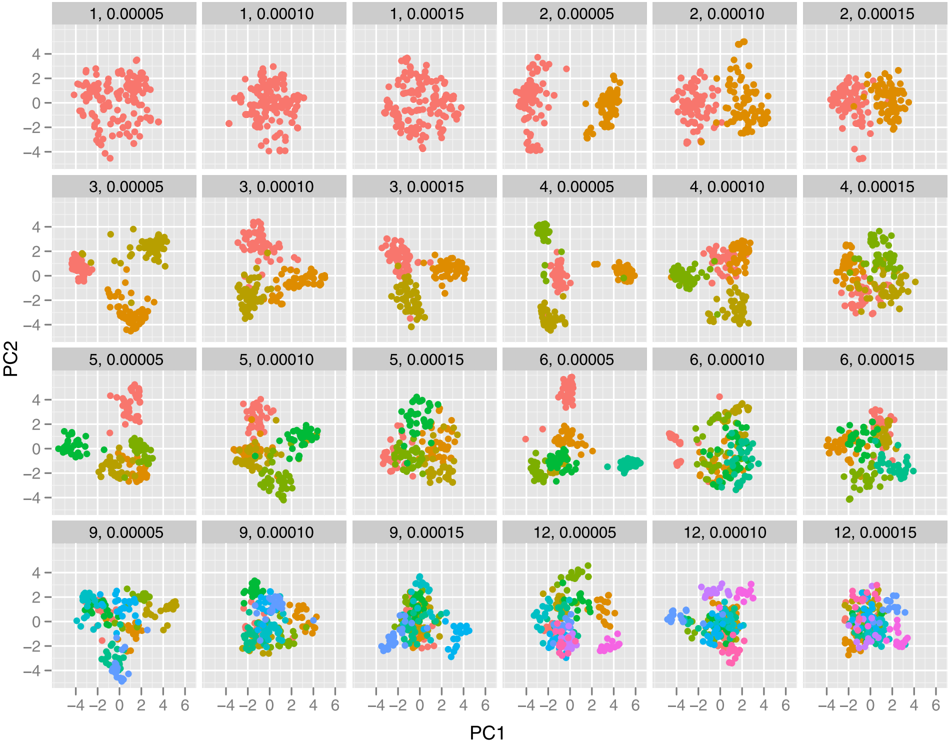

FIG. 2. *PC1 versus PC2 for a randomly selected simulated data set for each cluster—migration rate combination. Note that as the migration rate increases, clusters are generally more overlapped. With more clusters, the first two PCs are insufficient to capture the separation of clusters, but it can still be seen that the clusters are further apart with a lower migration rate.*

and varieties of global origin that was used by Tinker et al. (2009) in the initial DArT development work. The tinker2009 data set consists of 198 observations and 1958 DArT markers with 21.6% missing data. The third oat data set, referred to as asoro2011, consists of 446 North American elite lines scored for 1005 DArT markers with 5.8% missing data [Asoro et al. (2013)]. We note that there is some overlap between the newell2010 data set with both the tinker2009 and asoro2011 data sets. This is because the newell2010 data set combined data sets from independently assembled collections. Although some observations are duplicated from the two sets in newell2010, all three data sets have different combinations of marker data, thus they will cluster quite differently.

The first barley data set, referred to as hamblin2010, was originally used to explore population structure and linkage disequilibrium [Hamblin et al. (2010)]. This set is the largest used in this study and consists of 1816 observations from ten barley coordinated agricultural project (CAP) participating breeding programs throughout the US and scored for 1416 single nucleotide polymorphisms (SNPs), with only 0.2% missing data. Unlike the oat data sets, hamblin2010 has strong population structure, thus enabling testing of a wide variety of cluster separation in the empirical sets. The second barley data set, referred to as zhang2009, was originally used to assess barley population structure and linkage disequilibrium [Zhang et al. (2009)]. The data set is comprised of 169 lines consisting of mainly Canadian cultivars and breeding lines scored on 971 DArT markers. The zhang2009 data set has about 2.6% missing data. The last empirical data set, referred to as chao2010, is a wheat data set also originally used to explore population structure and linkage disequilibrium [Chao et al. (2010)]. The data set consists of 849 SNPs scored on 478 spring and winter wheat cultivars from 17 breeding programs across the US and Mexico. The chao2010 data set contains 0.9% missing data.

Taken together, the empirical data sets used in this study cover a wide range of variation with respect to the level of separation between clusters. The variation across empirical data sets can easily be seen from their first two principal components (Figure 3). The oat data sets, newell2010, asoro2011 and tinker2009, have relatively weak structure with less distinct clusters. In contrast, the barley data sets, hamblin2010 and zhang2009, and wheat data set, chao2010, show relatively strong structure, and clusters can easily be seen in the principal component plots. These differences in the level of structure across crops are most likely explained by the breeding processes implemented for the specific crops. For example, oats include

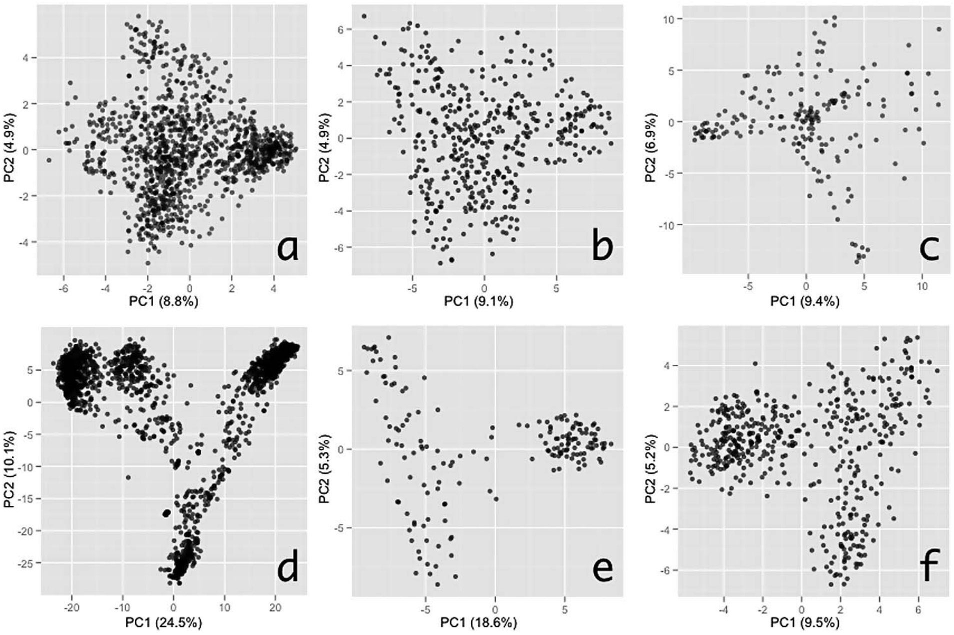

FIG. 3. *The large amount of variation across the empirical sets by visualization of the first two principal components with amount of variation explained by each axes in parentheses for* (a) newell2010, (b) asoro2011, (c) tinker2009, (d) hamblin2010, (e) zhang2009, *and* (f) chao2010.

TABLE 1

Summary of the six empirical data sets used in this study including the assigned name, source of the original publication, crop, origins of lines included, types of lines included, the dimensions designated rows x columns, and the marker type designated as DArT or SNP for Diversity Array Technology or single nucleotide polymorphism, respectively

Source (name)	**Crop**	**Origins**	**Line types**	**Dimensions**	**Marker type**
Newell et al. (2011) (newell2010)	Oat	World	Varieties, breeding lines, landraces	1205 × 403	DArT
Asoro et al. (2013) (asoro2011)	Oat	North American	Elite cultivars	446 × 1005	DArT
Tinker et al. (2009) (tinker2009)	Oat	World	Varieties, breeding lines, landraces	198 × 1958	DArT
Hamblin et al. (2010) (hamblin2010)	Barley	US	Elite cultivars	1816 × 1416	SNP SNP
Zhang et al. (2009) (zhang2009)	Barley	Canada	Cultivars, breeding lines	169 × 971	DArT
Chao et al. (2010) (chao2010)	Wheat	US, Mexico	Spring/winter wheat cultivars	478 × 219	SNP

hulled, naked, spring and winter types, but breeding generally occurs across the major types and for the most part, lines are usually spring, hulled types. In contrast, barley includes 2-row, 6-row, spring and winter types in all combinations in which it is common practice to cross individuals within the same type but not between types. This leads to the strong structure seen for the first two principal components relative to the oat data sets. Similarly, the first two principal components for the wheat data set separated spring and winter types and further split spring types into two based on their region of development. This indicates that for wheat, crossing does not occur between spring and winter types and crossing most likely does not occur across major regions within the spring types. The principal component plots also allow comparison of the empirical and simulated sets. Comparison of Figures 3 and 4 demonstrate how the low-dimensional projections from PCA are quite similar between the simulated and empirical sets and, more importantly, the simulated sets cover the range of possibilities encountered in real data. A summary of the empirical data sets used in this study is shown in Table 1.

4. Results.

4.1. *Simulated data.* Results for 100 simulated data sets at each cluster—migration rate combination are summarized in Table 2. The mean estimated number of clusters at the lowest migration rate was within 0.09 of the true number of clusters across all combinations, excluding the case when one cluster was sim-

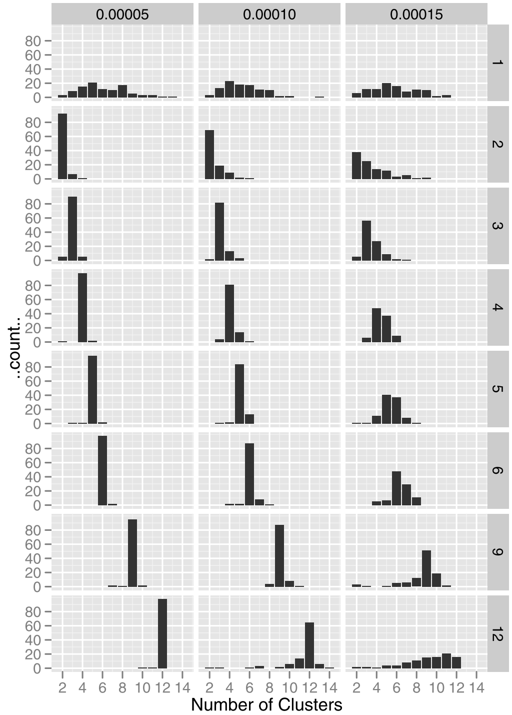

FIG. 4. *Bar graph showing the resulting number of clusters implementing the proposed algorithm for* 100 *simulations at each cluster—migration rate combination. The numbers at the top and right of each facet represent the migration rate and the true number of clusters simulated*, *respectively.*

ulated. Hclust was the preferred method of clustering, followed by kmeans and mclust which were chosen based on the algorithm in 69%, 28% and 3% of the simulations. As expected for the largest migration rate, the mean estimated number of clusters deviated the most from the true simulated number of clusters across all combinations. Overall, the proportion of times the algorithm chose the correct number of clusters ranged from 0.16 when 12 clusters were simulated at the largest migration rate and 0.98 when six and 12 clusters were simulated at the lowest migration rate. In general, the proportion of times the algorithm chose the correct number of clusters decreased as the migration rate was increased. This was expected given the fact that as the migration rate is increased, the clusters become less distinct with more overlapping. These results are also shown in Figure 4 at each cluster—migration rate combination. Because the true classifications are known, a comparison between the true and estimated Hubert's gamma across bootstrap samples can be made. Across all simulations the true and estimated values of the Hubert's gamma statistic decreased as the migration rate was increased. Likewise, in all cases the estimated Hubert's gamma was larger than the true value; this trend

TABLE 2

Summary of results for the simulated data sets including the true number of clusters and migration rate simulated, mean estimated number of clusters, true Hubert's gamma, Hubert's gamma, and the proportion of times the correct number of clusters was chosen

True no	Migration rate	Mean est no	True Hubert's Gamma	Hubert's Gamma	Proportion correct
1	0.00005	6.09	1.00	0.48	–
	0.00010	5.37	1.00	0.47	–
	0.00015	5.79	1.00	0.48	–
2	0.00005	2.09	0.69	0.73	0.92
	0.00010	2.47	0.49	0.55	0.69
	0.00015	3.46	0.36	0.48	0.38
3	0.00005	3.00	0.77	0.79	0.90
	0.00010	3.17	0.58	0.61	0.82
	0.00015	3.50	0.46	0.52	0.56
4	0.00005	4.00	0.78	0.79	0.97
	0.00010	4.12	0.59	0.62	0.81
	0.00015	4.49	0.46	0.51	0.48
5	0.00005	4.99	0.76	0.78	0.96
	0.00010	5.09	0.59	0.63	0.84
	0.00015	5.40	0.47	0.51	0.41
6	0.00005	6.02	0.75	0.77	0.98
	0.00010	6.04	0.58	0.62	0.87
	0.00015	6.34	0.46	0.51	0.48
9	0.00005	8.97	0.72	0.75	0.95
	0.00010	9.06	0.55	0.59	0.87
	0.00015	8.53	0.43	0.46	0.51
12	0.00005	11.97	0.68	0.71	0.98
	0.00010	11.36	0.51	0.55	0.65
	0.00015	9.20	0.41	0.43	0.16

can also be seen in the mean estimated number of clusters where this tends to overestimate the true number of clusters.

Like other methods of clustering, the algorithm does not directly have the ability to detect the case when no structure exists, as is the case when one cluster is simulated. The predicted number of clusters when one cluster was simulated covered the range of possible values with no definitive result across simulations (Figure 4). For this reason, it is important to have a diagnostic to determine when this is in fact the case. The shape of the Hubert's gamma statistic relative to the number of clusters can distinguish the case when no structure is present. Figure 5 shows the shape of the Hubert's gamma statistic for the three methods of clustering when one, two and six clusters were simulated at the lowest migration rate for each simulation. As depicted, when one cluster is simulated the shape of the

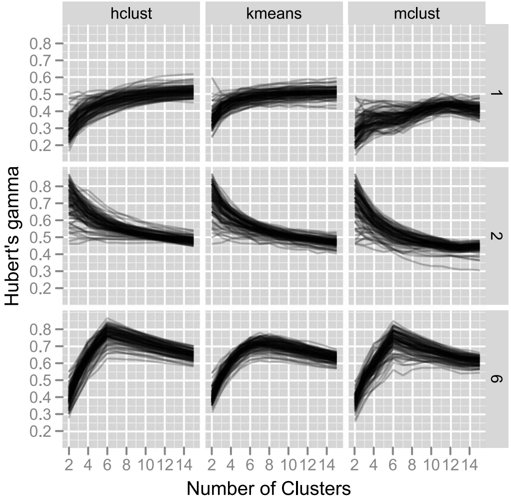

FIG. 5. *Diagnostic plot of Hubert's gamma for varying numbers of clusters to distinguish between the case when either one or two clusters are present, illustrated on simulated data. One, two and six cluster results are shown in the rows, and cluster method, hierarchical, k-means and model-based clustering methods in the columns. Of most importance is the different pattern between* 1 *and* 2 *clusters: in the case of just one cluster, Hubert's gamma increases from* 0.2, *but in the case of two clusters there is a gradual decline in Hubert's gamma from* 0.7 *with increasing number of clusters. For six clusters, a distinctive peak occurs at* 6.

Hubert's gamma increases and levels off for hclust and kmeans with no decrease in Hubert's gamma; this occurs in the opposite direction in the case when two clusters were simulated. Likewise, for the case when six clusters were simulated, the Hubert's gamma increased to a peak followed by a decrease. Although these shapes can help distinguish the case when one or two clusters are present, this can become quite difficult when the migration rate is increased due to the fact that the peak becomes less profound as the clusters become less distinct. In addition to this diagnostic plot, the Hubert's gamma is positively correlated with the proportion of times the correct number of clusters was chosen with a value of 0.84 (Table 2). Therefore, a low Hubert's gamma statistic for a data set gives an indication into the confidence that the correct number of clusters was called. Thus far, the results have been presented as if no prior information is known. For genetic data sets this is rarely the case and can also be exploited in choosing the final number of clusters.

4.2. *Empirical data.* The empirical data sets imposed more variability with respect to the degree of separation between clusters, number of lines per cluster and the number of markers per data set. Table 3 summarizes the algorithm results for the empirical sets used in this study for 50, 100 and 200 bootstrap samples. Results presented throughout will be for 200 bootstrap samples unless otherwise stated. The final numbers of clusters for the six data sets ranged from one to six and

TABLE 3

Summary of results for the six empirical data sets in this study including the final number of clusters, method and Hubert's gamma shown in parentheses for 50, 100 *and* 200 *bootstrap samples, and previous results*

	No of clusters, method (Hubert's Gamma)			**Previous results**	**Previous method**
Data set	**50**	**100**	**200**		
newell2010	4, kmeans (0.481)	4, kmeans (0.481)	5, kmeans (0.526)	6 (0.438)	mclust on PCA
asoro2011	3, kmeans (0.434)	4, kmeans (0.424)	5, kmeans (0.431)	3 (0.434)	kmeans
tinker2009	1, – (1)	1, – (1)	1, – (1)	None specified	PCA and hclust
hamblin2010	6, hclust (0.816)	6, hclust (0.816)	6, hclust (0.816)	7 (0.590)	STRUCTURE
zhang2009	2, kmeans (0.786)	2, kmeans (0.786)	2, kmeans (0.786)	2 (0.774)	PCA, prior knowledge
chao2010	3, kmeans (0.581)	4, hclust (0.579)	4, hclust (0.579)	9 (–)	STRUCTURE

are also represented as the number of clusters versus the Hubert's gamma statistic in Figure 6. This plot is the diagnostic plot presented in the simulation results. As shown, the starting value for the Hubert's gamma statistic at two clusters covered a wide range across the three clustering methods. zhang2009 has a unique

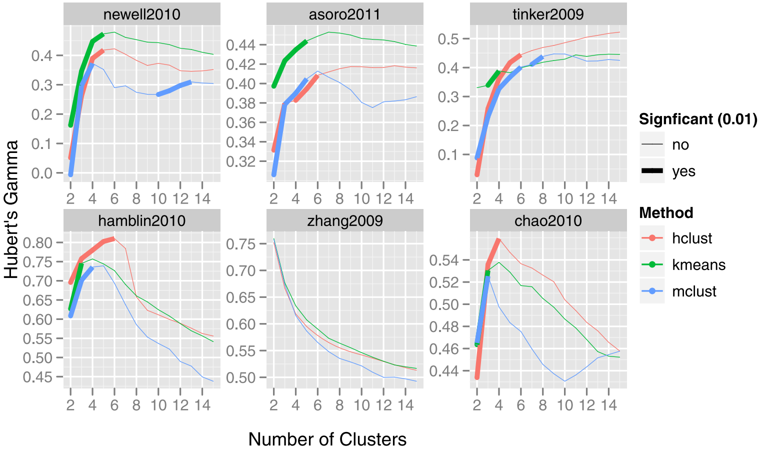

FIG. 6. *Number of clusters versus the Hubert's gamma statistic for the six empirical sets in the study for* 200 *bootstrap samples. Colors refer to the three clustering methods and bold lines represent significant increase of Hubert's gamma for each consecutive cluster pair at* $p < 0.01$.

shape that is characteristic of the case when two clusters are present. Newell2010, hamblin2010 and chao2010 all have distinct peaks for all methods of clustering, indicating that greater than one cluster is present. Asoro2011 shows an increase in Hubert's gamma for kmeans until about seven clusters, at which time it decreases, also indicating that there is greater than one cluster. In contrast, tinker2009 is the only data set that is characteristic of the situation in which only one cluster exists. If greater than one cluster was true, the algorithm would identify six clusters using hclust. Due to the fact that the Hubert's gamma using hclust does not show a peak but a continuous increase, it is concluded that tinker2009 has only one cluster.

In order to assess the appropriate number of bootstrap samples required for the empirical sets, the algorithm was applied using 50, 100 and 200 bootstrap samples (Table 3). Results for two of the data sets, hamblin2010 and zhang2009, did not change beyond 50 bootstrap samples, indicating that this was sufficient for these data sets. The results for chao2010 did not change beyond 100 boostrap samples, in which case this would be sufficient for this data set. Newell2010 required 200 bootstrap samples to reach equilibrium with respect to the number of clusters; data is not shown for 300. The results for asoro2011 are unusual in the sense that the number of clusters is still changing up to 200 bootstrap samples. The algorithm was further tested for this data set using 300 and 400 bootstrap samples, where the number of clusters identified was six and five, respectively. This outcome can be justified by the nature of the data set, where the lines included are all North American elite oats with a narrow genetic base. For a data set such as this it would be concluded that the true number of clusters would be in the range of five to six; in this case any prior information about the data set would be helpful in a final decision. Interestingly, the number of bootstrap samples required is negatively related to the Hubert's gamma statistic for all of the data sets. Asoro2011 requires the most bootstrap samples and has the lowest Hubert's gamma, and hamblin2010 and zhang2009 require the fewest number of bootstrap samples and have the highest Hubert's gamma statistics. Application of the results of the Hubert's gamma statistics at 50 bootstrap samples can be used as an indicator for the number of bootstrap samples required for a particular data set. For example, data sets with a Hubert's gamma in the range of 0.786 to 0.816 only require 50 bootstrap samples, those in the range of 0.581 require 100, those in the range of 0.481 require 200, and less than 0.434 require greater than 200 bootstrap samples, although, with a sample size of only six, application to a greater number of empirical sets would be required to solidify this claim. In summary, data sets resulting in larger Hubert's gamma statistics require less bootstrap samples and, from the simulation results, are more likely to determine the correct number of clusters.

Previous results for the six empirical sets are shown in Table 3 along with the method used for each result. As expected, the number of clusters determined by the proposed algorithm differs in most cases from previous results given the varying selection criteria across methods. The previous method implemented for newell2010 identified six clusters using model-based cluster analysis implemented on the first five principal components. In that study, the number of clusters was based largely on visual representation of principal components, thus, it was largely user defined. In contrast, the proposed algorithm defined five clusters using k-means clustering. Asoro et al. (2013) identified three clusters for the asoro2011 data set, but also indicated that this number was chosen based on the research objectives for that study; six clusters were initially identified. Previous results for tinker2009 did not necessarily identify a certain set number of clusters but used clustering more as a general guide to study the diversity of lines. The lines used in Tinker et al. (2009) were initially chosen to represent the diversity of oat on a worldwide scale; this can be seen in the first two PCs where lines tend to spread from a point resembling a bull's-eye (Figure 3). The algorithm identified only one cluster for this data set, which does conform to how the data was initially chosen. Similar results were found for the hamblin2010 data set by implementation of the proposed algorithm and STRUCTURE [Hamblin et al. (2010)], where six and seven clusters were identified, respectively. Results presented by Zhang et al. (2009) were the same for the proposed algorithm, with identification of two clusters. Last, the results for the chao2010 were largely different, with four and nine clusters identified for this algorithm and Chao et al. (2010), respectively. The four clusters identified by the algorithm respond to the group of winter and spring lines split into three groups. Overall, the proposed algorithm identifies a similar number of clusters to previous methods but is different considering the criterion for which the number of clusters is chosen.

In order to gain insight into where the empirical sets fall with respect to the simulated sets, Hubert's gamma statistics for each are shown simultaneously in Figure 7. The variation in the true Hubert's gamma for the simulated data sets at each

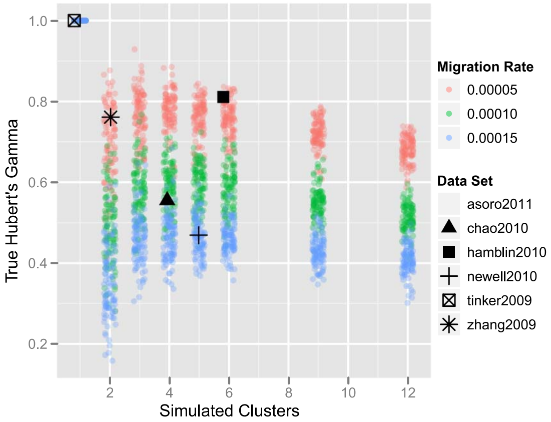

FIG. 7. *True Hubert's gamma values for all simulated data sets colored by migration rate overlaid with empirically determined Hubert's gamma values for the empirical data sets. This plot gives some suggestions for the migration rate observed with the empirical data*: *low for* zhang2009 *and* hamblin2010, *high for* asoro2011 *and* newell2010, *and medium for* chao2010.

cluster—migration rate combination covers a range of about 0.8, in which case a lower migration rate has a higher Hubert's gamma. Zhang2009 and hamblin2010 fall within the range of the lowest migration rate at two and six clusters, respectively. Chao2010 falls within the range of the middle migration rate, 0.0001, with four clusters. Newell2010 and Asoro2011 fall within the range of the largest migration rate of 0.00015, both at five clusters. Both the Newell2010 and Asoro2011 data sets, in addition to falling within the range of the largest migration rate, also have the smallest Hubert's gamma statistics. Last, tinker2009, having only one cluster, has a Hubert's gamma statistic of one. These comparisons can provide some information into the confidence of the correct number of clusters for the empirical sets. Empirical sets that fall within the range of the smallest and largest migration rates would have relatively more and less confidence, respectively.

5. Conclusion. This paper has proposed an algorithm that provides assistance in choosing the number of clusters and the clustering algorithm for HDLSS data. The algorithm uses bootstrap samples to quantify the cluster variation and permutation tests on Hubert's gamma statistics to test for significance of the chosen number of clusters. Validation of the algorithm on HDLSS data simulated by GENOME with varying numbers of clusters and level of separation indicates that the algorithm operates well on data of this sort. As clusters get more overlapped, if the migration rate is large, the accuracy in estimating the correct number of clusters declines. For the case when no cluster structure is present in a data set, a diagnostic plot of the change in Hubert's gamma across varying numbers of clusters can be used to indicate the lack of clusters.

The results from this algorithm on six empirical data sets vary slightly from the reported number of clusters in previous studies, but are not wildly different. The empirical data sets vary less uniformly than the simulated data sets, which might be expected. In most cases, the change in Hubert's gamma across the number of clusters in the simulated data resulted in significant peaks at the true simulated number of clusters. The three clustering methods did result in largely different Hubert's gamma statistics, with no one method being better than the others on all data sets, demonstrating the importance of including multiple clustering methods in the algorithm. However, it should be pointed out that mclust was the preferred clustering method in only 3% of the simulated data sets and was never the preferred clustering method in the empirical data sets. Previous research has highly recommended against clustering the principal components [Hubert and Arabie (1985)] and termed it "tandem clustering." Although the proposed algorithm does in fact implement mclust on the principal components, it rarely is actually selected as the best method.

In agreement with two previous studies [Hall, Marron and Neeman (2005), Murtagh (2009)], all of the empirical sets, and the simulated data, exhibit a simplex shape in the first few PCs. The different clusters form the vertices of the simplex. A comparison of the empirical to simulated sets illustrates that the Hubert's gamma statistics of the empirical sets are within the range of values observed for the simulated sets. This, along with the visualization of the PCs, supports a conclusion that the GENOME software is able to adequately simulate data sets that match well with the empirical sets. By plotting the Hubert's gamma of the empirical data sets in comparison to those of the simulated data sets for different migration rates, a reasonable sense of the migration rate observed by the empirical data sets can be determined.

Finally, we expect the cluster selection algorithm might be applicable to other HDLSS data. For other types of problems, where the data is not binary as is for the genetic data used here, comparison data might be simulated from a Gaussian mixture distribution for validation purposes.

Acknowledgments. We would like to thank Eduard Akhunov, Franco Asoro, Francois Belzile and Nick Tinker for providing empirical data sets for testing.

SUPPLEMENTARY MATERIAL

Supplement A: Videos of High-Dimensional Views of Empirical and Simulated Data (DOI: 10.1214/13-AOAS671SUPPA; .pdf). Video footage of tours of the empirical and simulated data sets.

Supplement B: Empirical and Simulated Data (DOI: 10.1214/13-AOAS671SUPPB; .zip). Data sets used in this paper.

Supplement C: R Code (DOI: 10.1214/13-AOAS671SUPPC; .zip). Software used to make the calculations for this paper.

REFERENCES

AHN, J., MARRON, J. S., MULLER, K. M. and CHI, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766. MR2410023

ASORO, F. G., NEWELL, M. A., BEAVIS, W. D., SCOTT, M. P. and JANNINK, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* **4** 132–144.

BELLMAN, R. (1961). *Adaptive Control Processes*: *A Guided Tour*. Princeton Univ. Press, Princeton, NJ. MR0134403

CHANG, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **32** 267–275. MR0770316

CHAO, C., DUBCOVSKY, J., DVORAK, J., LUO, M. C., BAENZIGER, S. P., MATNYAZOV, R., CLARK, D. R., TALBERT, L. E., ANDERSON, J. A., DREISIGACKER, S., GLOVER, K., CHEN, J., CAMPBELL, K., BRUCKNER, P. L., RUDD, J. C., HALEY, S., CARVER, B. F., PERRY, S., SORRELLS, M. E. and AKHUNOV, E. D. (2010). Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat. *BMC Genomics* **11** 727.

COOK, D., BUJA, A. and CABRERA, J. (1993). Projection pursuit indexes based on orthonormal function expansions. *J. Comput. Graph. Statist.* **2** 225–250. MR1272393

COOPER, M. and DELACY, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet.* **88** 561–572.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

FRALEY, C. and RAFTERY, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *J. Classification* **20** 263–286. MR2019797

FRALEY, C. and RAFTERY, A. E. (2011). mclust: Model-based clustering/normal mixture modeling. Available at http://cran.r-project.org/web/packages/mclust/index.html.

HALKIDI, M., BATISTAKIS, Y. and VAZIRGIANNIS, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* **17** 107–145.

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. MR2155347

HAMBLIN, M. T., CLOSE, T. J., BHAT, P. R., CHAO, S., KLINK, J. G., ABRAHAM, K. J., BLAKE, T., BROOKS, W. S., COOPER, B., GRIFFEY, C. A., HAYES, P. M., HOLE, D. J., HORSLEY, R. D., OBERT, D. E., SMITH, K. P., ULLRICH, S. R., MUEHLBAUER, G. J. and JANNINK, J. L. (2010). Population structure and linkage disequilibrium in U.S. barley germplasm: Implications for association mapping. *Crop Science* **50** 556–566.

HENNIG, C. (2011). fpc: Flexible procedures for clustering. Available at http://cran.r-project.org/web/packages/fpc/index.html.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 192–218.

JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454

LIANG, L., ZÖLLNER, S. and ABECASIS, G. R. (2007). GENOME: A rapid coalescent-based whole genome simulator. *Bioinformatics* **23** 1565–1567.

MILLIGAN, G. W. and COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50** 159–179.

MURTAGH, F. (2009). The remarkable simplicity of very high dimensional data: Application of model-based clustering. *J. Classification* **26** 249–277. MR2587798

NEWELL, M. A., COOK, D., TINKER, N. A. and JANNINK, J. L. (2011). Population structure and linkage disequilibrium in oat (Avena sativa L.): Implications for genome-wide association studies. *Theor. Appl. Genet.* **122** 623–632.

NEWELL, M. A., COOK, D., HOFMANN, H. and JANNINK, J.-L. (2013). Supplement to "An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure." DOI:10.1214/13-AOAS671SUPPA, DOI:10.1214/13-AOAS671SUPPB, DOI:10.1214/13-AOAS671SUPPC.

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

STEINLEY, D. (2006). K-means clustering: A half-century synthesis. *British J. Math. Statist. Psych.* **59** 1–34. MR2242281

STEINLEY, D., BRUSCO, M. J. and HENSON, R. (2012). Principal cluster axes: A projection pursuit index for the preservation of cluster structures in the presence of data reduction. *Multivariate Behavioral Research* **47** 463–492.

TINKER, N. A., KILIAN, A., WIGHT, C. P., HELLER-USZYNSKA, K., WENZL, P., RINES, H. W., BJØRNSTAD, Å., HOWARTH, C. J., JANNINK, J. L., ANDERSON, J. M., ROSSNAGEL, B. G., STUTHMAN, D. D., SORRELLS, M. E., JACKSON, E. W., TUVESSON, S., KOLB, R. L., OLSSON, O., FEDERIZZI, L. C., CARSON, M. L., OHM, H. W., MOLNAR, S. J., SCOLES, G. J., ECKSTEIN, P. E., BONMAN, J. M., CEPLITIS, A. and LANGDON, T. (2009). New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics* **10** 39.

VÄHÄ, J. P., ERKINARO, J., NIEMELÄ, E. and PRIMMER, C. R. (2007). Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. *Molecular Ecology* **16** 2638–2654.

ZHANG, L. Y., MARCHAND, S., TINKER, N. A. and FRANÇOIS, B. (2009). Population structure and linkage disequilibrium in barley assessed by DArT markers. *Theor. Appl. Genet.* **119** 43–52.

M. A. NEWELL
THE SAMUEL ROBERTS NOBLE FOUNDATION
2510 SAM NOBLE PARKWAY
ARDMORE, OKLAHOMA 73401
USA
E-MAIL: manewell@noble.org

D. COOK
H. HOFMANN
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
2413 SNEDECOR
AMES, IOWA 50011
USA
E-MAIL: dicook@iastate.edu
hofmann@iastate.edu

J.-L. JANNINK
USDA-ARS
ROBERT W. HOLLEY CENTER
FOR AGRICULTURE AND HEALTH
DEPARTMENT OF PLANT BREEDING AND GENETICS
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853
USA
E-MAIL: jeanluc.jannink@ars.usda.gov